EN PRODUCCIÓN - GALERADAS



Círculo Rojo

EN PRODUCCIÓN - GALERADAS

MORTALIDAD EN ICTUS ISQUÉMICO

EPIDEMIOLOGÍA, MODELOS PREDICTIVOS
Y NUEVAS TENDENCIAS EN INVESTIGACIÓN

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

MORTALIDAD EN ICTUS ISQUÉMICO

EPIDEMIOLOGÍA, MODELOS PREDICTIVOS Y NUEVAS TENDENCIAS EN INVESTIGACIÓN



María del Carmen Lea Pereira, en representación del Grupo NEUROTICs
Juan Manuel García Torrecillas



Círculo Rojo
EDITORIAL

EN PRODUCCIÓN - GALERADAS

Primera edición: octubre 2024

Depósito legal: AL 961-2024

ISBN: 978-84-1073-242-1

Impresión y producción: Editorial Círculo Rojo

© Del texto: María del Carmen Lea Pereira, en representación del Grupo NEUROTICs / Juan Manuel García Torrecillas

© Maquetación y diseño: Equipo de Editorial Círculo Rojo

Editorial Círculo Rojo

www.editorialcirculorojo.com

info@editorialcirculorojo.com

Impreso en España - Printed in Spain

Editorial Círculo Rojo apoya la creación artística y la protección del copyright. Queda totalmente prohibida la reproducción, escaneo o distribución de esta obra por cualquier medio o canal sin permiso expreso tanto de autor como de editor, bajo la sanción establecida por la legislación.
Círculo Rojo no se hace responsable del contenido de la obra y/o de las opiniones que el autor manifieste en ella.

El papel utilizado para imprimir este libro es 100% libre de cloro y, por tanto, ecológico.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

AGRADECIMIENTOS

Queremos expresar nuestro más sincero agradecimiento a la **Fundación Andaluza Progreso y Salud** por su financiación y a **FIBAO** por su gestión. Su apoyo ha sido fundamental para que este proyecto se haya podido llevar a cabo.

De otro lado, este proyecto ha sido sostenido desde el amplio parasol de un hospital comarcal, otro de referencia y la indispensable red de Atención Primaria. Diferentes niveles asistenciales y todos ellos con una participación y un nivel de implicación encomiables.

No menos importante es agradecer todo el apoyo y soporte que desde la **Unidad de Investigación Biomédica del Hospital Universitario Torrecárdenas** y desde las **Direcciones, tanto hospitalarias como de Atención Primaria,** nos han brindado en todo momento.

Gracias al **Dr. Serrano Castro** por su cariñoso y admirable prólogo. Un amigo, un compañero y gran impulsor de nuestra debilidad por el estudio de este potente enemigo que es el ictus. En su buen hacer nos miramos para continuar trabajando en la senda correcta.

Agradecemos a nuestras familias el soporte y la paciencia, para lidiar con nuestras vaivenes e incursiones en este fascinante pero exigente mundo. Estamos seguros de que han sido muchos los desvelos que nos han ayudado a seguir y perseverar en esta empresa.

EN PRODUCCIÓN - GALERADAS

Finalmente, gracias a todos nuestros pacientes, porque ellos son el motor inicial y final de esta batalla, y por ellos siempre lo es todo.

Los coordinadores de la obra

EN PRODUCCIÓN - GALERADAS

AUTORES DE LA OBRA

Luis Andrade Zumárraga
UGC de Neurología del Hospital Universitario Torrecárdenas.
Cristina del Toro Pérez
UGC de Neurología del Hospital Universitario Torrecárdenas.
Laura Amaya Pascasio*
UGC de Neurología del Hospital Universitario Torrecárdenas.
Antonio Arjona Padillo
UGC de Neurología del Hospital Universitario Torrecárdenas.
María del Mar Iglesias Espinosa*
UGC de Neurología del Hospital Universitario Torrecárdenas.
María del Mar Rodríguez Salvador*
Diplomada en Enfermería. Distrito Sanitario Almería.
Francisco José Muñoz Ronda*
Antropología y Enfermería. Distrito Sanitario Almería
José Galván Espinosa*
Gestor de Proyectos. Fundación FIBAO. Hospital Universitario Torrecárdenas
Juan José López Ramos*
Técnico en Informática. Unidad TIC provincial Almería.
María Ortiz Puertas
Diplomada en Enfermería. Hospital Universitario Torrecárdenas
José Antonio Morales Molina
UGC de Farmacia Hospitalaria. Unidad de Investigación Biomédica. Hospital Universitario Torrecárdenas

EN PRODUCCIÓN - GALERADAS

Luis Téllez Ramírez
Becario de Investigación. Hospital Universitario Torrecárdenas
Fernando Reche Lorite*
Departamento de Matemáticas. Universidad de Almería
Carmen Rosa Garrido
Técnica en Metodología de la Investigación y Bioestadística. Fundación FIBAO. Hospital Universitario de Jaén
Regina Rodríguez Martínez
UGC de Urgencias. Unidad de Investigación Biomédica. Hospital Universitario Torrecárdenas
Adrián Aparicio Mota
Técnico en Metodología de la Investigación y Bioestadística. Fundación FIBAO.
Cristina Ortiz Puertas
Colaboradora. Facultad de Medicina de Lleida. UDL.
María del Carmen Lea Pereira*
Servicio de Medicina Interna. Hospital Universitario de Poniente.
En representación del grupo NEUROTICs
Bernabé Ortiz Puertas
UGC de Urgencias. Unidad de Investigación Biomédica. Hospital Universitario Torrecárdenas.
Juan Manuel García Torrecillas*
UGC de Urgencias. Unidad de Investigación Biomédica. Hospital Universitario Torrecárdenas.
***Grupo NEUROTICs**

EN PRODUCCIÓN - GALERADAS

PRÓLOGO

Hace poco leí un prólogo de un buen libro que comenzaba de una manera un tanto provocadora, diciendo que reconocía que no le gustaba el libro que iba a prologar. Durante un breve instante, después de ver las características y estructura del libro que el lector tiene entre sus manos he sentido la tentación de hacer lo mismo. Y es que, al igual que al prologuista al que me refiero, **no me gusta este libro**… por la sencilla razón de que tanto a él como a mi realmente lo que nos hubiera gustado seria que la idea se nos hubiera ocurrido a nosotros.

En el vasto y complejo paisaje de la medicina, pocas condiciones han desafiado tanto en los últimos años los principios clásicos de la destreza clínica y la agilidad terapéutica como el ictus isquémico. No en vano, el cambio en el tratamiento emergente de esta patología en los últimos 20 años y sus consecuencias prácticas en términos de morbi-mortalidad ha sido de tal magnitud que me atrevería a decir que ha obligado a alterar el sentido e incluso la propia idiosincrasia de una especialidad como la mia, la neurología. La Neurología, como disciplina, ha transitado en este tema desde la actitud contemplativa cercana al nihilismo tradicional a la frenética actividad que rodea a todo lo que sucede cuando alguien evoca el término "código ictus".

Hoy sabemos en el transcurso de un ictus, por término medio se pierden, ya de forma irreversible, nada menos que 2 millones de neuronas, 14 billones de sinapsis y 12 km de fibras mielini-

EN PRODUCCIÓN - GALERADAS

zadas... **por cada minuto**. Multiplíquense estas cifras por 60 y veremos la importancia de una hora en el transcurso de un ictus. La simple magnitud de estas cifras hará entender al lector por qué ha sido tan revolucionaria la llegada de terapias eficaces tiempo-dependientes en esta materia y por qué la neurología vascular ha mutado en su propio ADN desde la pasividad hasta el intervencionismo más exacerbado.

Este punto de partida constituye el espejo en el que este libro adquiere su verdadero sentido. Es necesario comprender como esta revolución interna se ha traducido en mejora para nuestros pacientes tanto en términos de morbilidad como de mortalidad. Y también mostrar, desde postulados basados en la evidencia, el camino para maximizar dichos beneficios a través de la elaboración de modelos predictivos capaces de guiar nuestra propia actividad asistencial. Ya adelanto que las conclusiones de los autores van en la línea esperada. No se parece en nada, afortunadamente, el pronóstico vital actual de una patología que, como el ictus isquémico, se encuentra en pleno crecimiento de incidencia poblacional en nuestro entorno, del que existía hace dos décadas. Hemos conseguido construir un dique de salud gracias al advenimiento de estos recursos y la consolidación de las infraestructuras asistenciales.

Pero, claro, no todo podría ser negativo. Hay otra razón por la que **sí me gusta este libro** y compete directamente a sus autores. Muchos de ellos han sido durante años compañeros de trabajo en una experiencia que, me atrevo a decir, ha sido consustancial con mi propia esencia de médico. Los médicos de urgencia, especialistas de diferentes áreas y los enfermeros que constituye el elenco de autores del libro son un ejemplo de los niveles de excelencia que el trabajo colaborativo puede llegar a conseguir en Medicina. Me siento orgulloso de haber participado de alguna manera en la puesta en marcha de la asistencia moderna al ictus isquémico en Almería allá por los albores del nuevo mileno. Y me congratula

EN PRODUCCIÓN - GALERADAS

comprobar que dichos esfuerzos han cristalizado de la mejor manera posible.

Los coordinadores de la obra, sin ir más lejos, atesoran tres características que rara vez he visto tan firmemente aunadas la misma persona: tesón, compromiso y conocimiento. Tesón para superar las dificultades que a veces las circunstancias acarrean. Compromiso con sus pacientes en el más puro sentido hipocrático de la palabra. Y por fin, conocimiento que vierten con generosidad en las páginas que siguen.

En definitiva, creo que se trataba de una obra necesaria, que ve la luz en el momento oportuno y que está elaborada por los mejores autores. Disfrútenla. Les merecerá la pena y aprenderán mucho.

Fdo. Dr. Pedro J. Serrano Castro. Jefe de Servicio de Neurología. Hospital Regional Universitario de Málaga. Instituto de investigación biomédica de Málaga (IBIMA-Plataforma Bionand)

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

FUNDING

Esta obra ha sido posible gracias a la financiación del proyecto AP-0013-2020-C1-F2 a través de la convocatoria de proyectos de investigación e innovación en el ámbito de la atención primaria y hospitales comarcales y centros hospitalarios de alta resolución del Sistema Sanitario Público de Andalucía, de la Fundación Pública Progreso y Salud, M. P (Consejería de Salud y Consumo-Junta de Andalucía)

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

A MODO DE INTRODUCCIÓN...

Unidos para mejorar un problema común

El ictus, también conocido como accidente cerebrovascular, es una de las principales causas de muerte y discapacidad en el mundo. La investigación sobre el ictus constituye un campo en constante evolución tanto desde el punto de vista del abordaje clínico y terapéutico como, qué duda cabe, la aproximación desde las nuevas metodologías de la investigación y la irrupción de métodos estadísticos cada vez más novedosos.

La investigación sobre el ictus ha avanzado considerablemente en los últimos años. Se han desarrollado nuevos métodos de diagnóstico y tratamiento que han permitido mejorar significativamente las tasas de supervivencia y recuperación. Aun así, el esfuerzo investigador debe ser constante para poder mantener las tasas de excelencia que se vienen alcanzado. Sin duda alguna, el grupo multidisciplinar, heterogéneo y al tiempo totalmente compenetrado de autores de esta monografía, hace de su día a día un nuevo reto para mejor la investigación en esta enfermedad tan prevalente.

UN LIBRO QUE MARCA LA DIFERENCIA

Este libro, **Mortalidad en ictus isquémico. Epidemiología, modelos predictivos y nuevas tendencias en investigación**, es una obra que reúne los últimos avances en este campo y las apor-

EN PRODUCCIÓN - GALERADAS

taciones de los miembros del equipo de investigación. En él, los autores ofrecen una visión completa y actualizada del ictus, desde sus causas y mecanismos fisiopatológicos hasta su diagnóstico, tratamiento y prevención. Además, se dedican extensos pasajes a desarrollar las nuevas metodologías que nos permiten en la actualidad aprender cada día más de las fuentes de información utilizando novedosos abordajes metodológicos y estadísticos.

La elaboración de este libro ha sido posible gracias a la colaboración de un gran número de personas e instituciones. En primer lugar, queremos agradecer a los autores su dedicación y esfuerzo por compartir sus conocimientos y experiencia. Profesionales sanitarios de las más diversas disciplinas, y de diferentes niveles asistenciales, junto a estadísticos, epidemiólogos, gestores, matemáticos, informáticos…, todos aunados en pro de optimizar las nuevas herramientas de abordaje en la investigación del ictus. Se han combinado para hacer de esta una obra, cuando menos, muy especial.

EN PRODUCCIÓN - GALERADAS

1. NEUROEPIDEMIOLOGÍA DEL ICTUS AVANZADA EN NUESTRO PAÍS. CONDICIONANTES Y MEDIOS.

AUTORES: LUIS ANDRADE ZUMARRAGA. CRISTINA DEL TORO PÉREZ. LAURA AMAYA PASCASIO.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - CAPTURADAS

1. INTRODUCCIÓN

El ictus es una enfermedad de primer orden debido al alto grado de discapacidad y el elevado coste socio-sanitario asociado. A pesar de las importantes mejoras en prevención primaria y el desarrollo de tratamientos reperfusores en fase aguda, el ictus sigue siendo una enfermedad devastadora: 17 millones de personas sufren anualmente un ictus en el mundo (Béjot et al., 2016). En España cada año se producen 110.000-120.000 ictus, tratándose de la primera causa de muerte en las mujeres y la segunda entre los hombres. Las enfermedades cerebrovasculares son la octava causa de hospitalización y representan el 70% de los ingresos neurológicos en España (Sociedad Española de Neurología, 2019).

Cuando hablamos de ictus, nos referimos a dos subtipos fisiopatológicamente distintos, pero que comparten mecanismos etiopatogénicos y comorbilidades: el ictus isquémico o infarto cerebral y el ictus hemorrágico o hemorragia cerebral intraparenquimatosa. Así mismo, el ataque isquémico transitorio (AIT) hace referencia a la presencia de déficits neurológicos atribuibles a un territorio vascular con una duración menor de 24 horas sin que haya evidencia de lesión isquémica establecida en neuroimagen (Easton et al., 2009).

El ictus reúne muchos de los criterios que justifican el establecimiento de un sistema de vigilancia epidemiológica: es un importante problema de salud pública, se puede prevenir en gran medida, si se aplican estrategias eficaces para la corrección y tratamiento de los factores de riesgo modificables, y es una

enfermedad que tiene un gran impacto en todos los países. Teniendo en cuenta estas observaciones, es esencial la elaboración de estudios epidemiológicos de calidad que permitan organizar la política sanitaria con el fin de mejorar la prevención y el tratamiento del ictus, determinando las necesidades específicas de cada población.

La mayoría de los informes de incidencia de ictus disponibles proceden de países desarrollados. No obstante, incluso en estas comunidades, con estructuras de vigilancia epidemiológica bien establecidas, hay una escasez de datos sobre las tendencias temporales de incidencia del ictus. Otro inconveniente es la heterogeneidad metodológica entre los diferentes estudios, que dificulta la comparación de las tasas de incidencia entre países y momentos temporales (Béjot et al., 2016; Sudlow & Warlow, 1996).

La fuente más fiable de información epidemiológica proviene de los registros de base poblacional, que garantizan tanto la calidad como la exhaustividad de los datos y permiten su comparación. Estos estudios están limitados por su coste y complejidad metodológica, por lo que se han diseñado otros modelos epidemiológicos, de base no poblacional, que permiten una estimación de la incidencia, prevalencia o mortalidad de la patología cerebrovascular con una correlación adecuada a las reales pero con un menor coste y complejidad (Béjot et al., 2016; VL Feigin et al., 2019). En este capítulo trataremos algunos de los requisitos metodológicos fundamentales para el diseño y desarrollo de un estudio epidemiológico de base poblacional y presentaremos algunas de las herramientas disponibles para realizar estimaciones en materia de incidencia, prevalencia y mortalidad. Así mismo, analizaremos los principales datos de la epidemiología del ictus en España.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN. GALERADAS

2. METODOLOGÍA DE LA EPIDEMIOLOGÍA DEL ICTUS.

2.1 ¿POR QUÉ MEDIR LA INCIDENCIA DEL ICTUS?

Tradicionalmente, las tasas de mortalidad por ictus se han utilizado para realizar estimaciones de la incidencia de esta enfermedad. Esto es debido a que los datos de mortalidad son los más fácilmente disponibles en cualquier país, aunque dependen de la exactitud de los certificados de defunción y únicamente proporcionan información sobre los casos mortales. Si consideramos únicamente los casos registrados de fallecimiento por ictus, un gran porcentaje de casos que sobreviven no son cuantificados. Además, el porcentaje de mortalidad por ictus es muy variable en función de las condiciones sociosanitarias de cada país, así como del acceso de cada población a tratamientos reperfusores o unidades de ictus. Una estimación de la incidencia de ictus basada exclusivamente en las cifras de mortalidad sería muy variable en función de las estrategias de codificación del fallecimiento y de la calidad asistencial existente en cada país analizado.

Otro inconveniente es la falta de información acerca del subtipo de ictus o de la etiología subyacente. Los estudios que miden la incidencia de ictus y que se basan en la selección de los casos de una comunidad cuentan con el potencial de estudiar y responder otras cuestiones como son la tasa de letalidad, el porcentaje de recurrencias y de discapacidad, así como infor-

mación sobre comorbilidades y factores de riesgo. Ampliar el rendimiento de estos estudios más allá de la mera identificación de casos ayuda a justificar el intenso esfuerzo realizado en el proceso de búsqueda de casos y permite identificar factores de riesgo asociados.

2.2 REQUISITOS PARA UNA CORRECTA EVALUACIÓN DE LA INCIDENCIA

A la hora de plantear un estudio que evalúe la incidencia del ictus en una región concreta, un diseño metodológico adecuado es fundamental para garantizar que se alcancen los objetivos propuestos, evitando la recogida de datos innecesarios. En 1987 Malmgrem et al establecieron los que denominaron criterios ideales para el análisis de la epidemiología del ictus, que fueron posteriormente revisados por Sudlow y Warlow (Malmgren et al., 1987; Sudlow & Warlow, 1996):

- Definiciones estandarizadas:

a) Definición de ictus. La Organización Mundial de la Salud (OMS) define el ictus como la presencia de signos clínicos de afectación focal (a veces global) de la función cerebral que dura más de 24 horas o que conduce a la muerte sin otra causa aparente que la de origen vascular (Thomas Truelsen et al., n.d.). Según la definición, deben excluirse las presentaciones de hemorragia subaracnoidea con cefalea de inicio repentino y sin alteración focal o global de la función cerebral. Sin embargo, generalmente la mayoría de los investigadores incluyen todas las hemorragias subaracnoideas, y los informes de los estudios deben dejarlo claro.

En segundo lugar, es importante diferenciar entre ictus y AIT, basándose en criterios clínicos y de imagen cerebral,

EN PRODUCCIÓN - GALERADAS

dado que el AIT incluye únicamente aquellos casos de afectación neurológica focal sin lesión establecida en neuroimagen. Existe controversia sobre cómo deben clasificarse los síndromes neurológicos transitorios que presentan un infarto cerebral agudo en neuroimagen. Una limitación es la gran variabilidad en la disponibilidad y el tipo de técnicas de imagen cerebral utilizadas en cada región, lo que puede modificar el diagnóstico final de los casos analizados, creando una variabilidad espuria entre los estudios y agravando las dificultades para comparar sus resultados. Además, la adjudicación de posibles eventos de AIT en un entorno comunitario es especialmente difícil debido a la gran cantidad de entidades que pueden mimetizar un AIT y el sesgo de la hospitalización.

b) Primer caso de ictus en la vida del sujeto: Es recomendable que se incluyan únicamente aquellos eventos cerebrovasculares que ocurren por primera vez en la vida del sujeto. No obstante, algunos estudios basan sus estimaciones de incidencia en el número de primeros ictus ocurridos dentro del periodo de estudio, sin excluir a los que tuvieron un ictus antes del inicio del estudio, incluyendo así un número de ictus recurrentes que no son verdaderos casos incidentes.

- Metodología estandarizada

a) Comprobación de casos comunitarios: Con frecuencia se han realizado estudios de incidencia que evalúan exclusivamente poblaciones hospitalarias. Estos no pueden ser considerados como representativos de toda la comunidad, ya que siempre habrá algunos casos que no ingresen en el hospital, porcentaje que varía en gran medida entre países y momentos temporales a medida que cambian las estrategias sanitarias.

EN PRODUCCIÓN - GALERADAS

b) *Exhaustividad en la determinación de los casos*: Debido a que muchos casos de ictus se dan entre personas mayores y discapacitadas, que pueden estar confinadas en casa, se requieren métodos de evaluación intensivos, para garantizar que no se excluyan dichos casos. Sin embargo, no sería eficiente concentrar los recursos en una fuente con un rendimiento insignificante, por lo que la clave está en identificar en cada población todas las fuentes posibles de información y justificar la exclusión de las que no se utilizan. Por ejemplo, en sociedades en las que la tasa de hospitalización por ictus es cercana al 100%, con una tasa de notificación comunitaria por parte de atención primaria insignificante, sería razonable centrarse prioritariamente en las fuentes hospitalarias para la búsqueda de casos.

c) *Estudio prospectivo*. El ictus es un diagnóstico clínico, por lo que la evaluación clínica precisa por parte de un especialista debidamente cualificado es una parte esencial de cualquier estudio de incidencia. Sólo los estudios verdaderamente prospectivos que persiguen los casos a medida que se producen, siendo valorados por especialistas, proporcionan un alto grado de precisión. La recopilación retrospectiva de información es más propensa a desdibujar el margen entre AIT e ictus o a incluir de manera menos exhaustiva las comorbilidades del paciente. La evaluación de ictus previo como antecedente en cada paciente, es de igual forma, menos precisa. Los estudios basados totalmente en revisiones retrospectivas de los registros raramente cuentan con registros completos para toda la comunidad, aunque pueden existir excepciones, como el registro de ictus de la Clínica Mayo o el de Dijon.

d) *Población de estudio:* Conviene elegir poblaciones bien definidas y estables. Las poblaciones inestables, con frecuentes cambios permanentes de residencia dentro y fuera del área

EN PRODUCCIÓN - GALERADAS

de estudio, son menos fiables. Sin embargo, cuanto más aislada esté una población, menos representativa de su país puede ser, debido a contar con una variabilidad genética restringida y unos hábitos alimentarios concretos. No existe un tamaño de población perfecto, las poblaciones con una elevada proporción de personas mayores producirán más accidentes cerebrovasculares para el mismo tamaño total de población que las poblaciones más jóvenes. Las mediciones de la incidencia deberían basarse en la población de riesgo, lo que implica que debemos eliminar del denominador los casos prevalentes, es decir, los que ya han sufrido un ictus. En la práctica, esta información no está disponible para la gran mayoría de los investigadores, por lo que no se realizan ajustes por prevalencia.

e) Presentación de datos: Los datos recogidos a lo largo de años completos evitan la confusión derivada del posible efecto de las variaciones estacionales en la incidencia. Las tasas de incidencia pueden variar a lo largo de periodos de aproximadamente 5 años. Esto implica que no deben promediarse en intervalos de tiempo superiores a 5 años y que las comparaciones entre diferentes estudios deben limitarse a puntos similares en el tiempo. Debido a las posibles diferencias entre los sexos, las tasas deben ser estandarizadas por sexo y por edad.

2.3. ESTUDIOS EPIDEMIOLÓGICOS DE BASE POBLACIONAL

Como hemos descrito previamente, los estudios de base poblacional proporcionan la información epidemiológica más fiable, permitiendo así la comparación internacional de los estudios. Los verdaderos estudios basados en la comunidad deben utilizar varias fuentes de información que se superponen para garantizar la determinación completa de los casos. El sistema sanitario local

EN PRODUCCIÓN - GALERADAS

determinará cuáles de ellas son aplicables. Las fuentes deben ser encuestadas con la frecuencia adecuada, cubriendo un amplio espectro de posibles presentaciones del ictus para que no se pierdan casos. Por ejemplo, etiquetas como "colapso", "mareo", "ataques" o "confusión" pueden revelar casos de ictus. Las fuentes externas al hospital, en particular los médicos de atención primaria, deben participar en el estudio y ser contactadas regularmente para garantizar la detección de todos los casos de la comunidad. Los registros administrativos de altas hospitalarias proporcionan una comprobación útil de los pacientes no identificados y la revisión de los certificados de defunción permite encontrar a los pacientes que fallecen sin que se haya realizado una derivación.

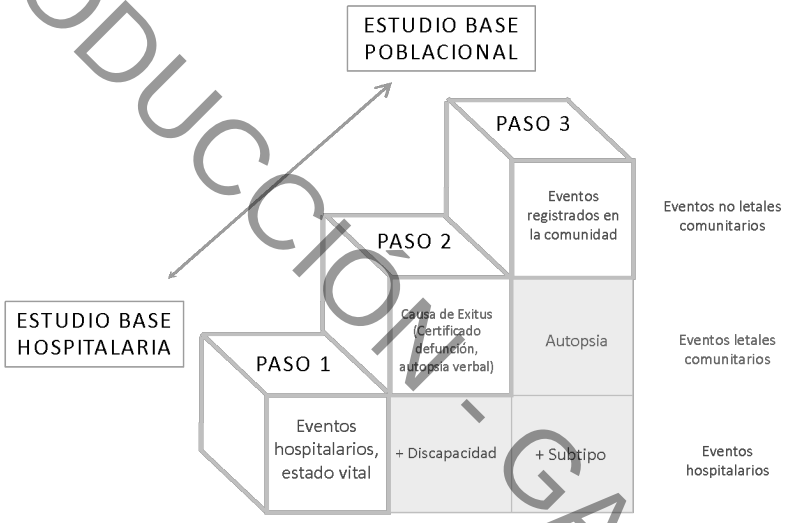
Un modelo destacado para la realización de estudios poblacionales es el enfoque por pasos STEPS (*STEPwise Approach to Surveillance*) de la OMS para la vigilancia de los accidentes cerebrovasculares (T Truelsen et al., 2001). Así mismo, algunos registros hospitalarios nacionales y modelos basados en encuestas poblacionales, principalmente dirigidos a evaluar la prevalencia del ictus, estarían a la altura de los criterios propuestos para los estudios de base poblacional.

El método STEPS de la OMS (T Truelsen et al., 2001) sugiere un enfoque gradual para la vigilancia de los accidentes cerebrovasculares, teniendo en cuenta tres grupos diferentes de pacientes con ictus: los ingresados en el hospital (u otros centros sanitarios), los sucesos mortales que ocurren fuera del hospital y los sucesos no mortales gestionados fuera de un centro hospitalario (figura 1). Los pacientes con ictus hospitalizados son los más fáciles de identificar, mientras que los casos no mortales no hospitalizados son los más difíciles. El paso 1 del modelo proporciona la frecuencia de ingresos hospitalarios por ictus. La inclusión del paso 2 permite calcular las tasas de mortalidad, mientras que la realización del paso 3 proporciona las tasas de incidencia y de letalidad. Aquellos estudios que incluyen los tres pasos de vigi-

EN PRODUCCIÓN - GALERADAS

lancia propuestos en el modelo STEPS (el registro hospitalario, los certificados de defunción y la determinación de casos no hospitalizados en el área de captación) pueden considerarse de base poblacional.

Figura 1. Modelo de vigilancia de la incidencia de ictus por pasos (STEPS). Adaptado de Truelsen et al (T Truelsen et al., 2001).



Con el objetivo de garantizar la calidad de los estudios epidemiológicos de base poblacional se han elaborado recientemente una serie de directrices que se recogen en una *checklist*, según se muestra en la tabla 1 (Bennett et al., 2015; V Feigin et al., 2018). Cuando se cumplen estos criterios, un estudio epidemiológico podría ser considerado de base poblacional y de alta calidad.

EN PRODUCCIÓN - GALERADAS

Tabla 1. Normas para informar de los estudios de incidencia de ictus y AIT basados en la población. Adaptado de Bennet et al., 2015

Sección	Recomendación
Título	Indique el tipo de diseño de estudio utilizando
Abstract	El resumen debe dar una descripción precisa de cómo se llevó a cabo el estudio y de los principales resultados.
Introducción	
Antecedentes	Deben informar de la justificación científica del estudio.
Objetivos:	Indique los objetivos específicos del estudio.
Métodos	
Diseño del estudio	Describa completamente el diseño del estudio y los criterios que cumple para denominarse estudio de base poblacional.
Contexto	Claramente definido (normalmente, pero no siempre, sobre una base geográfica), y estable, con información fiable sobre la migración de entrada y salida.
Población de origen (fuente)	Descripción de cómo se identificaron todos los miembros elegibles de la población y a través de qué fuentes de datos; fuente de datos utilizada para el estudio. Si incluye bases de datos administrativas deben describirse los algoritmos de extracción de datos utilizados. Descripción de la tasa de ingresos hospitalarios. Detalles del sistema sanitario en el país (región del estudio) donde se realizó el estudio. Descripción de cómo una persona con ictus o AIT en el país (región del estudio) es derivada al centro de referencia donde se realizó el estudio. Descripción y características de la tasa de respuesta/abandono y tasa de exclusión, si procede.

EN PRODUCCIÓN. GALERADAS

Participantes / población incluida en el estudio	La definición de ictus y AIT se identifica claramente y se presenta con suficiente detalle. Se describen los detalles del método de muestreo (los participantes son representativos de la población de origen). Fuente de diagnóstico plenamente validada o criterios estándar de referencia aplicados. Definición y justificación de la gravedad de la enfermedad (preferiblemente utilizando una escala estandarizada). Descripción de cómo se distinguen los tipos/subtipos de ictus/AIT (si procede). Descripción de cómo se ha evaluado la exhaustividad de la determinación de los casos. Descripción de si la exhaustividad de la determinación de los casos fue adecuada.
Aprobación ética	Deben informarse los detalles de la aprobación ética/ consentimiento informado/tratamiento de los datos.
Determinaciones	Dar detalles de cómo se determinó la incidencia (según el momento de la recogida de datos, ya sea prospectiva o retrospectivamente). Definición y justificación del momento de las mediciones. Especificar el período de tiempo (normalmente años enteros o tiempo-persona). Las cifras brutas se presentan con suficiente detalle para calcular las tasas adecuadas (por ejemplo, por edad o sexo).
Métodos estadísticos	Si las tasas se han estandarizado, deben darse los detalles de la población estándar utilizada. Si es posible, deben utilizarse dos poblaciones estándar, una con relevancia local y otra que facilite las comparaciones internacionales. Cualquier asunción en la realización de cálculos debe reportarse. Explicar cómo se abordaron los datos perdidos.
Resultados	
Resultados principales	Considere un diagrama de flujo que describa cómo se incluyeron los participantes en el estudio. Indique las tasas adecuadas con sus IC al 95% asociados. Informar de los resultados de cualquier análisis de sensibilidad.

EN PRODUCCIÓN - GALTERADAS

Discusión	
Resultados clave	Resumir las principales conclusiones en relación con los objetivos del estudio.
Interpretación	Interpretar los resultados en el contexto de la evidencia de otros estudios bien realizados con diseños y objetivos similares. Fiabilidad de las estimaciones (es decir, basada en el informe de la metodología estadística, y el diseño del estudio, la medición de la información clave).
Limitaciones	Discutir las posibles limitaciones del estudio. Incluya detalles sobre el riesgo de sesgo (por ejemplo, el sesgo de selección), la exhaustividad de la determinación de los casos y la calidad de los datos (evaluación de su probabilidad, tamaño e importancia potencial). Calidad de los datos (evaluación de su probabilidad, tamaño e importancia potencial).
Generalización de resultados	Discutir la validez externa de los resultados del estudio. ¿Son los resultados coherentes con las revisiones de estudios epidemiológicos descriptivos sobre el mismo tema que abarcan diferentes entornos?
Implicaciones	Discutir las implicaciones de los resultados de la investigación para la práctica y la investigación futura.

A pesar de los grandes esfuerzos por definir la metodología adecuada, la complejidad de esta y los costes que conlleva han hecho que haya muy pocos estudios poblacionales considerados como ideales. La mayoría presenta una gran heterogeneidad en cuanto a la población incluida, el periodo de tiempo evaluado y el momento de establecer el diagnóstico previo a ser registrado (Bennett, 2012).

Además, estos estudios se realizan principalmente en zonas urbanas, donde la incidencia de ictus difiere de la observada en las zonas rurales, esencialmente debido a las variaciones en la distribución de los factores de riesgo vascular. Así mismo, cubren poblaciones relativamente pequeñas que pueden no reflejar la

EN PRODUCCIÓN-GALERADAS

composición de toda la población de un país en términos de nivel socioeconómico de la población o acceso a los servicios médicos y son de difícil acceso en países con pocos recursos. Por lo tanto, la verdadera incidencia de accidentes cerebrovasculares puede estar subestimada y la generalización de los resultados de estos registros puede ser inadecuada (Béjot et al., 2016).

Con respecto a los estudios de prevalencia, una de las principales herramientas utilizadas es la encuesta autoinformada realizada puerta a puerta en grandes grupos poblacionales, representativos de la población de cada comunidad. Estas encuestas están destinadas a interrogar a la población general sobre la presencia de enfermedades cerebrovasculares recientes o en el pasado (V. L. Feigin, 2007).

Estos métodos de encuesta requieren el uso de instrumentos que permitan la comprobación clínica de eventos de ictus y AIT no hospitalizados y no mortales e incluyen una técnica validada de autopsia verbal de los eventos mortales (entrevista a los familiares y los profesionales de la salud pertinentes sobre las circunstancias de un fallecimiento ocurrido hasta dos años antes de la encuesta).

Las ventajas de este método son el coste relativamente bajo y la simplicidad, por lo que son el diseño de estudio epidemiológico de ictus más utilizado en entornos con pocos recursos. Las desventajas incluyen la posibilidad de subestimar de incidencia y prevalencia por la omisión de casos de ictus muy leves, el sesgo de recuerdo (aunque esto puede minimizarse restringiendo el periodo de recuerdo a dos años), y la necesidad de estudiar una muestra de gran tamaño (V. L. Feigin, 2007).

Woodfield et al realizaron una revisión sistemática evaluando la precisión de las encuestas autoinformadas en la prevalencia del ictus (Woodfield & Sudlow, 2015). Reportaron un valor predictivo positivo (VPP) entre el 22 y el 87% y una sensibilidad

EN PRODUCCIÓN. GALERADAS

variable (rango del 36 al 98%) objetivando una especificidad y un valor predictivo negativo (VPN) consistentemente altos (del 88 al 99%). Demostraron que en poblaciones con baja prevalencia de ictus (<10%), una gran proporción de ictus reportados por los sujetos fueron falsos positivos. Del mismo modo, descubrieron que entre el 6 y el 25% de las personas que declararon haber sufrido un ictus tenían un diagnóstico estándar de AIT. La agrupación de casos de ictus y AIT podría ser aceptable para algunas preguntas de investigación (por ejemplo, las que exploran los factores de riesgo comunes para el ictus/AIT). En base a estos resultados, los autores sugieren que las escalas autoinformadas que pretenden evaluar la prevalencia del ictus pueden ser una herramienta de cribado útil en estudios prospectivos para identificar posibles casos, pero no lo suficientemente precisa por sí misma para confirmarlos (Woodfield & Sudlow, 2015).

2.4. EPIDEMIOLOGÍA BASADA EN ESTIMACIONES POBLACIONALES.

Los registros de ictus basados en la población pueden proporcionar una incidencia de ictus válida y exhaustiva. Sin embargo, sus resultados son difíciles de extrapolar porque cubren una población pequeña y conllevan un elevado coste. Existen métodos alternativos a los estudios de base poblacional para la estimación de la incidencia, prevalencia y el pronóstico de los accidentes cerebrovasculares a partir de registros administrativos nacionales, encuestas y registros de mortalidad.

El Estudio *Global disease burden* desarrolló un modelo para proporcionar estimaciones a nivel nacional de la incidencia de accidentes cerebrovasculares, basado en registros de mortalidad y registros administrativos adaptados a los recursos disponibles en cada país. Se demostró que las estimaciones se correlacionaban correctamente con las tasas de incidencia reales proporcionadas por los registros poblacionales, y confirmaron que las tasas de in-

EN PRODUCCIÓN - GALERADAS

cidencia de ictus más altas se observan en Europa del Este (Bennett, 2012).

Los registros administrativos hospitalarios, de obligado cumplimiento en la mayoría de los países desarrollados, son aquellos que se recogen de forma rutinaria, a menudo con fines de facturación, sobre todos los pacientes de una población determinada. Se basan en la codificación de los diagnósticos al alta según la clasificación internacional de enfermedades (CIE), ya sea en su novena o décima versión (CIE 9, CIE 10). Las ventajas de la investigación basada en datos administrativos incluyen el gran número de casos registrados, la consistencia de los datos entre los pacientes y el escaso gasto necesario. Las desventajas incluyen la disponibilidad limitada de variables recogidas y las inexactitudes de la codificación

Otras fuentes disponibles para la estimación de la incidencia son las tasas de mortalidad cerebrovascular, si bien solo informan de los ictus mortales, y las encuestas de salud realizadas a la población por los institutos nacionales de estadística.

2.4.1 Registros administrativos en España

En España, el conjunto mínimo básico de datos (CMBD) de hospitalización y Atención Ambulatoria Especializada de los hospitales del Sistema Nacional de Salud es de obligado cumplimiento en todos los hospitales. Procede de las historias clínicas y es cumplimentado por las unidades de admisión. Se compone de 31 variables, entre ellas múltiples datos socio-demográficos como sexo, fecha nacimiento, lugar de residencia, datos hospitalarios, etc. También incluye diagnóstico principal y hasta una veintena de diagnósticos secundarios, procedimientos tanto diagnósticos como terapéuticos realizados durante el ingreso, complicaciones asociadas a ingreso y mortalidad intrahospitalaria (*BOE.es - BOE-A-2015-1235 Real Decreto 69/2015, de 6 de febrero, por el que se regula el Registro de Actividad de Atención Sanitaria Espe-*

EN PRODUCCIÓN - GALERADAS

cializada., n.d.). Este registro se encuentra disponible bajo justificación de la petición, en el portal web del Ministerio de Sanidad. En el caso del ictus, los códigos diagnósticos en base a la CIE9 o la CIE 10 (según el periodo analizado) se recogen en la tabla 2.

Tabla 2. Codificación de las enfermedades cerebrovasculares según la clasificación internacional de enfermedades, versiones 9 y 10 (CIE 9, CIE 10).

Código CIE-9	Código CIE-10
430: Hemorragia subaracnoidea	I60: Hemorragia subaracnoidea
431: Hemorragia intracraneal	I61: Hemorragia intracerebral no traumática
432: Otras hemorragias intracraneales no especificadas	I62: Otra hemorragia intracraneal no traumática y las no especificadas
433: Oclusión y estenosis de arterias precerebrales	I63: Infarto cerebral
434: Oclusión de arterias cerebrales	I64: Accidente cerebro vascular agudo no especificado
435: Ataque isquémico transitorio	I65: Oclusión y estenosis de las arterias precerebrales sin ocasionar infarto cerebral
436: Enfermedad cerebrovascular aguda mal definida	I66: Oclusión y estenosis de las arterias cerebrales sin ocasionar infarto cerebral
437: Otra enfermedad cerebrovascular y enfermedad cerebrovascular mal definida	I67: Otras enfermedades cerebrovasculares
438: Efectos tardíos de enfermedad cerebro vascular	I68: Trastornos cerebrovasculares en enfermedades clasificadas en otra parte
	I69: Secuelas de enfermedad cerebrovascular

EN PRODUCCION-GALERADAS

Una de las limitaciones del CMBD es el riesgo de la infra codificación de diagnósticos o procedimientos dado que dependen de su notificación por parte de los facultativos responsables en los informes de alta. Este sesgo podría resultar en la obtención de resultados paradójicamente protectores para algunas variables tradicionalmente reconocidas como nocivas o de riesgo. Condiciones como obesidad o tabaquismo podrían presentar este sesgo. Así mismo, en estos registros solo están representados los ictus hospitalarios e incluyen todos los episodios de ictus, dado que no se establece que los pacientes que hayan presentado varios ingresos, o ingresos previos al año analizado, hayan sido excluidos. Por último, este método de estimación no permite conocer la epidemiología de los subtipos de ictus, que es una información relevante para establecer comparaciones, dada la diferente influencia de los factores de riesgo.

Sin embargo, en las últimas décadas ha mejorado exponencialmente la concienciación poblacional frente al ictus así como la atención a esta entidad. En el siglo XX, la frecuencia de ingreso de los ictus era baja y sesgada, de forma que las cifras de incidencia procedentes de los registros eran muy bajas. En la actualidad, la mayoría de los pacientes con ictus es hospitalizada, y probablemente solo los casos de ictus de repetición, con causa conocida, edad muy avanzada o mal pronóstico puedan no estar incluidos en los registros hospitalarios. También ha mejorado de manera llamativa la calidad de la notificación diagnóstica y codificación (Aboa-Eboulé et al., 2013). En este sentido, la utilización del CMBD debe considerarse como un recurso útil y de mucho menor coste que un estudio de base poblacional (Matías-Guiu, 2007).

Aboa-EBoulé et analizaron si la base de datos de altas hospitalarias francesa, similar al CBMD español, que recoge habitualmente datos administrativos, podría ser una herramienta útil

EN PRODUCCIÓN - GALERADAS

para proporcionar datos sobre la incidencia nacional de los ictus. Todos los registros de pacientes con diagnóstico de ictus entre 2004 y 2008 obtenidos del registro administrativo francés fueron comparados con los recogidos en el Registro de Ictus de Dijon, de base poblacional (Aboa-Eboulé et al., 2013). La sensibilidad global y el VPP del registro administrativo para el diagnóstico de ictus fueron, respectivamente, del 77,1% y del 69,2%. La concordancia con el registro de ictus de base poblacional fue moderada (kappa 52,8; IC del 95 %: 46,8-58,9). Sin embargo, su precisión (VPP) aumentó con el tiempo, con una codificación de mejor calidad con el paso de los años, y fue mayor para algunos subtipos de ictus como el cardioembólico. Esto indica que, con el tiempo, los casos de ictus identificados en los registros administrativos deberían ser similares a los registrados en las bases de datos poblacionales (Aboa-Eboulé et al., 2013). Los resultados fueron consistentes con estudios previos realizados en otros países europeos y americanos (Palmieri et al., 2007; Tirschwell & Longstreth, 2002).

2.4.2 Institutos de Estadística

El Instituto Nacional de Estadística (INE) es un organismo autónomo de carácter administrativo, adscrito al Ministerio de Asuntos Económicos y Transformación Digital, que elabora y distribuye estadísticas a una parte representativa de la población española. Este registro público es fácilmente accesible vía web.

Desde el punto de vista epidemiológico, recoge tres fuentes de datos fundamentales para conocer información sobre incidencia, prevalencia y mortalidad del ictus:

- Población censada anual: Recogida a principio o a mitad de año, por grupos de edad y sexo y para los niveles geográficos de comunidad autónoma y provincia. Esta información es de vital importancia para cualquier estudio epidemioló-

EN PRODUCCIÓN - CAFERADAS

gico, ya sea poblacional o basado en estimaciones a partir de registros administrativos. Permite conocer el número de personas censadas en cada comunidad en un periodo concreto, en riesgo de padecer un ictus.

- Datos de mortalidad: el INE otorga información de la posición ocupada por el ictus con respecto a otras enfermedades. Los datos de mortalidad por ictus y la tasa de mortalidad ajustada por edad, se recogen periódicamente por parte del Ministerio de Sanidad, Consumo y Bienestar Social.
- Encuestas de salud: Permiten obtener datos de prevalencia de enfermedades, hábitos de vida, dependencia física en una muestra representativa de la población española. Destaca la Encuesta Nacional de Salud de España 2017 (ENSE 2017), realizada por el Ministerio de Sanidad, Consumo y Bienestar Social con la colaboración del Instituto Nacional de Estadística, recoge información sanitaria relativa a la población residente en España en 23.860 hogares. Es una investigación de periodicidad quinquenal que permite conocer numerosos aspectos de la salud de los ciudadanos a nivel nacional y autonómico, y planificar y evaluar las actuaciones en materia sanitaria. Consta de 3 cuestionarios, hogar, adulto y menor, que abordan 4 grandes áreas: sociodemográfica, estado de salud, utilización de los servicios sanitarios y determinantes de la salud. Fue realizada durante el periodo 2016-2017 con un tamaño muestral de 29.195 entrevistas realizadas, 23.089 a personas adultas (15 y más años), y 6106 a menores (0-14 años), estas últimas mediante entrevista a sus progenitores/ tutores (*Ministerio de Sanidad, Consumo y Bienestar Social - Portal Estadístico del SNS - Encuesta Nacional de Salud de España 2017*, n.d.).

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - OBRA RETIRADA

3. EPIDEMIOLOGÍA DEL ICTUS EN ESPAÑA

La incidencia de las enfermedades cerebrovasculares en nuestro medio ha sido evaluada en distintos estudios que presentan una gran variabilidad entre ellos, tanto en la metodología como en los límites de edad considerados. En los años 90s los datos epidemiológicos sobre ictus se basaban principalmente en registros hospitalarios o en encuestas poblacionales. Un estudio llevado a cabo en Cantabria, publicado en 1993, analizó la incidencia hospitalaria de ictus en adultos jóvenes, objetivando una tasa bruta de incidencia anual de 13,9 por 100.000 en el grupo de edad de 11 a 50 años, y de 12 por 100.000 en el grupo de edad de 16 a 45 años (Leno et al., 1993). En Alcoy (Alicante) se llevó a cabo un estudio mediante un diseño de encuestas puerta a puerta que se propuso estimar las tasas de prevalencia e incidencia del AIT y del ictus. La tasa de prevalencia de AIT e ictus fueron 13/1.000 habitantes y 21,5/1.000 habitantes respectivamente y la tasa de incidencia para ambas entidades fue de 2,8/1.000 habitantes (Matías-Guiu et al., 1994). Así mismo, fueron numerosos los estudios de prevalencia del ictus basados en encuestas puerta a puerta realizadas a la población mayor de 70 años en diversas regiones españolas. El estudio ZARADEMP evaluó la prevalencia del ictus en mayores de 70 años en Zaragoza mientras que el BIDASOA se llevó a cabo en Irún. Similares diseños epidemiológicos fueron aplicados en Pamplona y El Prat de Llobregat. El estudio NEDICES contó con un diseño multicéntrico, en el que las encuestas de prevalencia se realizaron en Madrid, Getafe y Ávila. En 2006 se publicó

un metaanálisis que recogió y analizó los datos de todos estos estudios, reportando una prevalencia de ictus del 6,7% entre la población mayor de 70 años residente en regiones del centro y noreste de España. Se observó un aumento de la prevalencia a mayor edad, así como una menor prevalencia en mujeres que en hombres. La prevalencia fue mayor (8,7%) en zonas urbanas y menor (3,8%) en la población rural (Boix et al., 2006).

Uno de los primeros estudios epidemiológicos prospectivos de base poblacional, según las recomendaciones de Malmgren et al (Malmgren et al., 1987) fue el realizado en Segovia en los 90s por *Sempere et al,* que tuvo como objetivo estimar la incidencia de AIT e ictus minor y para ello evaluaron todos los casos identificados en el ámbito hospitalario y de atención primaria, evidenciando una tasa de 80 casos/100.000 habitantes/año (Sempere et al., 1996).

Cayuela et al se basaron en los datos de mortalidad registrados en el INE para analizar la tendencia de la mortalidad en España por enfermedad cerebrovascular durante el periodo 1980-2011. La tasa de mortalidad hasta 2011 mostró una marcada reducción tanto en hombres como en mujeres, pasando de 143 a 35/100.000 habitantes y de 121 a 27/100.000 habitantes, respectivamente. Concluyeron que la variación de la tasa de mortalidad podría deberse a una disminución en la tasa de incidencia, junto con un aumento de la supervivencia de los pacientes con ictus (Cayuela et al., 2019). Estos datos son comparables a los obtenidos en el estudio Oxford Vascular Study (OXVASC), que analizó los cambios en la incidencia de AIT, ictus y el porcentaje de factores de riesgo asociado y de uso de tratamiento preventivo durante el periodo 1981-2004. Los resultados mostraron una disminución en la incidencia del primer ictus isquémico un 29% entre el inicio y el final del periodo analizado (Rothwell et al., 2004).

El proyecto RECENT consistió en un registro de base poblacional en el que numerosos facultativos de atención primaria

EN PRODUCCIÓN - GALERADAS

de tres comunidades autónomas, Castilla y León, Extremadura y Comunidad Valenciana, recogieron durante el año 2005 los casos de primer ictus de cada paciente con el objetivo de estimar su incidencia y describir los principales factores de riesgo. La tasa de incidencia estimada de ictus fue de 141 casos por 100.000 habitantes (intervalo de confianza [IC] 95%: 125-158), 134 (IC 95%: 112-157) en mujeres y 148 (IC 95%: 124-172) en hombres. Así mismo, reportaron un aumento significativo de incidencia a partir de los 65 años. El 12% de pacientes con ictus fallecieron en las primeras 24 horas (Vega et al., 2009).

Posteriormente, en el año 2012, se publicó el estudio IBERICTUS (Díaz-Guzmán et al., 2012), considerado el primer estudio realizado en España, que cumple los estándares óptimos atribuidos a estudios epidemiológicos de base poblacional. El objetivo principal fue la estimación de la incidencia de ictus en el periodo de 2006 en cinco localidades españolas: Almería, Lugo, Talavera, Segovia y Mallorca. La población de estudio la constituyeron todos los casos nuevos de ictus y AIT que cumplían los criterios diagnósticos de la OMS, diagnosticados en el período de estudio y residentes en las provincias participantes. Para poder estimar las tasas de incidencia, se utilizó como denominador la población mayor de 17 años en riesgo de padecer un primer ictus en cada zona geográfica durante el año 2006. Se revisaron las siguientes fuentes de datos:

- Archivo central hospital, a través del conjunto mínimo básico de datos (CMBD) y los informes de alta e historias clínicas.
- Registros clínicos de los servicios de urgencias correspondientes.
- Registros de atención primaria (red de médicos centinela, bases de datos de atención primaria).
- Como fuentes secundarias, los registros de mortalidad de la provincia/área.

EN PRODUCCIÓN - GALERADAS

En cada una de las cinco zonas de estudio, se utilizó el mismo procedimiento y cuestionario para la identificación de casos. Los informes de alta se obtuvieron de los archivos centrales de los hospitales a partir de los listados del CMBD. Así mismo, consultaron mensualmente los registros clínicos de los servicios de urgencias de los hospitales y los servicios de urgencias de atención primaria y revisaron trimestralmente los registros de atención primaria y de radiología de los hospitales participantes. Una vez finalizada la identificación de casos en el año 2006, solicitaron al INE de cada provincia participante en el estudio el listado de sujetos fallecidos con el diagnóstico de enfermedad cerebrovascular como causa fundamental del fallecimiento, y cruzaron esta información con la base de datos de casos identificados en la zona (Díaz-Guzmán et al., 2008).

Las definiciones usadas para el ictus, AIT e insuficiencia cardiaca fueron acordes con las establecidas por la OMS. Los casos incluidos fueron revisados por neurólogos y estudiados por neuroimagen para una clasificación más homogénea, identificando los siguientes subgrupos etiológicos:

- Aterotrombótico de gran arteria (35%).
- Cardioembólico (20%).
- Ictus lacunar u oclusión de pequeña arteria (18%).
- Causa inhabitual (3%).
- Causa indeterminada (24%).

La incidencia anual fue de 187 por 100.000 habitantes, siendo mayor en hombres que en mujeres (202 frente a 187). La incidencia de los AIT estimada fue de 30 casos por 100.000 habitantes y la de los ictus establecidos, 150 casos por 100.000/habitantes/año, siendo el 80% isquémicos frente al 20% hemorrágicos. Se registraron un total de 2.700 primeros episodios de ictus (*first-ever strokes*). Estas cifras fueron comparables con el res-

EN PRODUCCIÓN - GALERADAS

to de los países europeos, y se estableció una clara relación de la incidencia con la edad.

En el mismo año se publicó el estudio EPICES, un registro observacional, multicéntrico y prospectivo de la totalidad de los pacientes con ictus ingresados en los hospitales públicos participantes que recibieron una atención neurológica. Su objetivo principal fue determinar la frecuencia de los subtipos de ictus. Se incluyeron 6.197 pacientes entre abril de 2008 y enero de 2009. Un 87.6% presentó ictus isquémico, de los cuales un 28,4% fue aterotrombótico, 26,2% cardioembólico, 22,5% lacunar, 20,7% indeterminado y un 2,4% de causa poco frecuente. El 12.4% de los pacientes presentó hemorragia intraparenquimatosa con un 56,4% de origen hipertensivo, un 12,7% secundarias, 13,9% patología amiloide y un 17% de etiología indeterminada. En cuanto a los factores de riesgo, se evidenciaron diferencias en la prevalencia entre los distintos tipos de ictus y hemorragias, aun así la hipertensión fue el factor de riesgo más prevalente tanto en el caso de ictus isquémico como hemorrágico (Arias-Rivas et al., 2012).

EN PRODUCCIÓN - GALERADAS

4. DISCUSIÓN

El ictus es un problema sanitario de primer orden, potencialmente prevenible si se adecuan las estrategias de prevención sanitaria a los requerimientos poblacionales. Una de las limitaciones a tal efecto es el bajo número de estudios epidemiológicos de calidad que permitan conocer información exhaustiva acerca de la incidencia del ictus y los factores de riesgo asociados.

En este capítulo se han tratado los principales requisitos que ha de cumplir un estudio de base poblacional para que sea óptimo y pueda ser comparable con los datos procedentes de otros países. Es fundamental tener en cuenta no solo los pacientes ingresados o fallecidos sino también aquellos que permanecen en la comunidad, sin llegar a la esfera hospitalaria, garantizando de esta forma la total inclusión de casos.

A pesar de la validez de los estudios de base poblacional, su elevado coste y complejidad hace que cada vez sean más frecuentes los trabajos basados en estimaciones epidemiológicas a partir de registros administrativos y de mortalidad. El CMBD es una fuente hospitalaria ampliamente disponible que recoge un elevado porcentaje de casos de ictus de la comunidad. Si bien está limitado por un número finito de variables, no modificable, y un posible sesgo de infra-registro diagnóstico, la calidad de la codificación y la exhaustividad de la notificación diagnóstica están mejorando día a día, así como el número de ictus que se benefician de tratamiento hospitalario, lo que hace que sea una herramienta de gran valor en la actualidad.

EN PRODUCCIÓN - GALERADAS

En España, hasta la publicación del estudio Ibericctus en el año 2012 no disponíamos de estudios ideales que analizaran la incidencia y prevalencia del ictus, así como información relativa a su etiología. La mayoría de la información de que se dispone se basa en registros hospitalarios, en encuestas puerta a puerta o en el registro de defunciones, mostrando resultados dispares y no comparables con otras poblaciones.

EN PRODUCCIÓN - GALERADAS

5. CONCLUSIONES

Es necesario disponer de información epidemiológica de calidad para abordar una enfermedad de primer orden, en parte previsible, como es el ictus.

Los estudios de base poblacional son la herramienta más útil para conocer información epidemiológica exhaustiva y comparable internacionalmente. Su coste y complejidad limitan su realización.

Los estudios basados en estimaciones epidemiológicas mediante registros administrativos, encuestas o registros de defunciones presentan una mayor probabilidad de sesgos, pero otorgan una aproximación a los datos reales de una manera más sencilla y eficiente.

El estudio Iberictus es un estudio epidemiológico que garantiza las recomendaciones establecidas para la realización adecuada de estudios de base poblacional, demostrando una incidencia anual de 187 casos de ictus por 100.000 habitantes en España.

EN PRODUCCIÓN - GALERADAS

6. BIBLIOGRAFÍA

1. Aboa-Eboulé C, Mengue D, Benzenine E, Hommel M, Giroud M, Béjot Y, Quantin C. How accurate is the reporting of stroke in hospital discharge data? A pilot validation study using a population-based stroke registry as control. J Neurol. 2013;260(2):605-613.
2. Arias-Rivas S, Vivancos-Mora J, Castillo J. Epidemiology of the subtypes of stroke in hospitalised patients attended by neurologists: Results of the EPICES registry (I). Rev Neurol. 2012;54(7):385–393.
3. Béjot Y, Bailly H, Durier J, Giroud M. Epidemiology of stroke in Europe and trends for the 21st century. Presse Med. 2016;45(12 Pt 2):e391–e398.
4. Bennett D. Methodology of the global and regional burden of stroke study. Neuroepidemiology. 2012;38(1):30–40.
5. Bennett D, Brayne C, Feigin V, Barker-Collo S, Brainin M, Davis D, Gallo V, Jetté N, Karch A, Kurtzke J, Lavados P, Logroscino G, Nagel G, PM, P, PM, R, LW, S. Explanation and Elaboration of the Standards of Reporting of Neurological Disorders Checklist: A Guideline for the Reporting of Incidence and Prevalence Studies in Neuroepidemiology. Neuroepidemiology. 2015;45(2):113–137.
6. BOE.es - BOE-A-2015-1235. Real Decreto 69/2015, de 6 de febrero, por el que se regula el Registro de Actividad de Atención Sanitaria Especializada. [Fecha de consulta: 25 de octubre de 2021].
7. Boix R, del Barrio J, Saz P, Reñé R, Manubens J, Lobo A, Gascón J, de Arce A, Díaz-Guzmán J, Bergareche A, Bermejo-Pa-

EN PRODUCCIÓN - GALERADAS

reja F, de Pedro-Cuesta J. Stroke prevalence among the Spanish elderly: an analysis based on screening surveys. BMC Neurol. 2006;6.
8. Cayuela A, Cayuela L, Rodríguez-Domínguez S, González A, Moniche F. Analysis of cerebrovascular disease mortality trends in Andalusia (1980-2014). Neurología. 2019;34(5):309-317.
9. Díaz-Guzmán J, Egido-Herrero JA, Gabriel-Sánchez R, Barberà G, Fuentes B, Fernández-Pérez C, Abilleira S, Serrano-Castro P, Piñeiro-Bolaño R, Jiménez C, Tur-Campos S, Guerrero P, Prieto-Jurczynska C, Sánchez Del Valle O, Acosta-Vergara T, Ahomar-Millán M, Alonso M, Barallobre B, Barturén-Fernández F, Vega-Quiroga S. Incidence of strokes in Spain. Methodological foundations of the Ibericctus study. Rev Neurol. 2008;47(12):617–623.
10. Díaz-Guzmán J, Egido JA, Gabriel-Sánchez R, Barberá-Comes G, Fuentes-Gimeno B, Fernández-Pérez C. Stroke and transient ischemic attack incidence rate in Spain: The IBERICTUS study. Cerebrovasc Dis. 2012;34(4):272–281.
11. Easton J, Saver J, Albers G, Alberts M, Chaturvedi S, Feldmann E, Hatsukami T, Higashida R, Johnston S, Kidwell C, Lutsep H, Miller E, Sacco R. Definition and evaluation of transient ischemic attack: a scientific statement for healthcare professionals from the American Heart Association/American Stroke Association Stroke Council; Council on Cardiovascular Surgery and Anesthesia; Council on Cardio. Stroke. 2009;40(6):2276–2293.
12. Feigin VL. Stroke in developing countries: can the epidemic be stopped and outcomes improved? Lancet Neurol. 2007;6(2):94–97.
13. Feigin V, Norrving B, Sudlow C, Sacco R. Updated Criteria for Population-Based Stroke and Transient Ischemic Attack Incidence Studies for the 21st Century. Stroke. 2018;49(9):2248–2255.
14. Feigin VL, et al. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019;18(5):439–458.
15. Leno C, Berciano J, Combarros O, Polo JM, Pascual J, Quintana F, Merino J, Sedano C, Martín-Durán R, Alvarez C, Llorca

EN PRODUCCIÓN - CAFERADAS

J. A prospective study of stroke in young adults in Cantabria, Spain. Stroke. 1993;24(6):792-795.
16. Malmgren R, Warlow C, Bamford J, Sandercock P. Geographical and secular trends in stroke incidence. Lancet. 1987;2(8569):1196–1200.
17. Matías-Guiu J. La investigación en epidemiología del ictus en España. ¿Estudios de base poblacional o utilización de aproximaciones a partir del CMBD? Rev Esp Cardiol. 2007;60(6):563-564.
18. Matias-Guiu J, Oltra A, Falip R, Martin R, Galiano L. Occurrence of transient ischemic attacks in Alcoi: descriptive epidemiology. Neuroepidemiology. 1994;13(1–2):34–39.
19. Ministerio de Sanidad, Consumo y Bienestar Social - Portal Estadístico del SNS - Encuesta Nacional de Salud de España 2017. [Internet]. [Fecha de consulta: 30 de octubre de 2021].
20. Palmieri L, Barchielli A, Cesana G, de Campora E, Goldoni C, Spolaore P, Uguccioni M, Vancheri F, Vanuzzo D, P C, S G. The Italian register of cardiovascular diseases: attack rates and case fatality for cerebrovascular events. Cerebrovasc Dis. 2007;24(6):530–539.
21. Rothwell P, Coull A, MF G, Howard S, Silver L, Bull L, Gutnikov S, Edwards P, Mant D, CM S, A F, PA S, MS D, CP W, JM B, P A. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford Vascular Study). Lancet. 2004;363(9425):1925–1933.
22. Sempere A, Duarte J, Cabezas C, Clavería L. Incidence of transient ischemic attacks and minor ischemic strokes in Segovia, Spain. Stroke. 1996;27(4):667–671.
23. Sociedad Española de Neurología. El Atlas del Ictus. [Internet]. 2019.
24. Sudlow CLM, Warlow CP. Comparing Stroke Incidence Worldwide What Makes Studies Comparable? Stroke. 1996;27(3):550–558.
25. Tirschwell D, Longstreth W. Validating administrative data in stroke research. Stroke. 2002;33(10):2465–2470.

EN PRODUCCIÓN - GALERADAS

26. Truelsen T, Bonita R, Jamrozik K. Surveillance of stroke: a global perspective. Int J Epidemiol. 2001;30 Suppl 1:S11–S16.
27. Truelsen T, Begg S, Mathers C. The global burden of cerebrovascular disease.
28. Vega T, Zurriaga O, Ramos JM, Gil M, Álamo R, Lozano JE, López A, Miralles MT, Vaca P, del Mar Alvarez M. Stroke in Spain: Epidemiologic Incidence and Patterns; A Health Sentinel Network Study. J Stroke Cerebrovasc Dis. 2009;18(1):11–16.
29. Woodfield R, Sudlow C. Accuracy of Patient Self-Report of Stroke: A Systematic Review from the UK Biobank Stroke Outcomes Group. PLoS One. 2015;10(9).

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN GALERADAS

2. MORTALIDAD DEL ICTUS ANTES Y DESPUÉS DE LA ERA DEL TRATAMIENTO REPERFUSOR (FIBRINOLISIS/ TROMBECTOMÍA)

AUTORES: ANTONIO ARJONA PADILLO. MAR IGLESIAS ESPINOSA.

EN PRODUCCIÓN - GALERADAS

RESUMEN

El ictus es un proceso complejo con múltiples fases y factores implicados. Las tres principales medidas terapéuticas que podemos ofrecer en la actualidad a nuestros pacientes con ictus, y que disminuyen la discapacidad funcional, son el ingreso en una Unidad de Ictus, la fibrinolisis intravenosa y la trombectomía mecánica. La mayoría de los estudios valoran como pronóstico la mortalidad a los 3 meses (Rankin modificado 6) y de estas estrategias la Unidad de Ictus es la que ha demostrado con mayor evidencia (nivel I) una disminución de la mortalidad con un NNT de 33, además de ser la medida que puede aplicarse a un mayor número de pacientes. La fibrinolisis solo se ha observado que disminuye la mortalidad en algunos estudios que valoran sus resultados a los 36 meses sin observarse este efecto a los 3 meses. Finalmente, con respecto a la trombectomía mecánica, y a la espera de resultados a largo plazo, únicamente en dos estudios y en meta-análisis se aprecia su efecto en la mortalidad a los 3 meses con un NNT de 29, aunque con un número menor de pacientes en los que puede intervenirse.

EN PRODUCCIÓN - ALTERADAS

1. INTRODUCCIÓN

En España mueren por ictus actualmente unas 27000 personas al año (datos del Instituto Nacional de Estadística) y se estima que en el año 2035 esta cifra se incremente en un 39% debido al envejecimiento de la población. Argumentos como este justifican este libro y capítulos específicos como el que vamos a desarrollar a continuación.

El ictus tiende a ser más grave en mujeres con una mayor mortalidad en este género. Otros factores de riesgo son la edad, la gravedad del ictus, la discapacidad previa y otros factores de riesgo concomitantes que serán tratados en otros capítulos de este libro.

El ictus es un proceso complejo con múltiples fases y factores implicados. La mortalidad asociada se puede clasificar en función del momento en el que se produce:

1. **Fase aguda**, que incluye el periodo del ingreso hospitalario en Neurología y posteriormente en muchas ocasiones Neurorrehabilitación. Durante esta fase el tratamiento activo es clave en la disolución de agregados trombóticos.
2. **Fase secuelar**: más dependiente de las secuelas y de las enfermedades asociadas. En el caso de ictus leves, esta fase, de duración muy variable, puede durar entre 3 y 6 meses.

La primera es la más estudiada y de ella se ocupan la mayoría de las publicaciones. Aquí tiene especial relevancia la mortalidad y evolución durante el primoingreso y está especialmente imbricada con los apartados dedicados a la fibrinolisis y a la trombectomía.

EN PRODUCCIÓN - GALERADAS

Desde otro enfoque podemos clasificar los factores relacionados con la muerte en el ictus en otros dos grupos:

- En primer lugar, los que podríamos denominar **factores clínicos** y asociados de la enfermedad, esto es localización, severidad, edad y comorbilidad previa del paciente
- En segundo lugar, los relacionados con **tipo del tratamiento** del ictus, es decir, los asistenciales que nos permiten intervenir e intentar revertir las deletéreas consecuencias del proceso ictal, y que se abordan en este capítulo.

Las tres principales medidas terapéuticas que podemos ofrecer en la actualidad a nuestros pacientes con ictus son:

1. **Tratamiento médico general/Unidades de Ictus**
2. **Fibrinolisis**
3. **Trombectomía mecánica**

La cronología de la implantación de estas terapias ha sido progresiva en el tiempo. El primer ensayo que demostró el beneficio de la Unidad de Ictus data de 1991, etapa a la que denominaremos pre-Unidad de Ictus. A partir de esta se produce la introducción progresiva de estas Unidades, muy retrasada en algunas Comunidades como la de Andalucía, a la que sigue una tercera etapa determinada por los **tratamientos reperfusores**, trombolisis intravenosa y posteriormente trombectomía mecánica combinada o no con la primera. Como consecuencia del retraso en la implantación de las Unidades de Ictus en Andalucía y otras comunidades, esta tercera etapa, en contra de la evidencia científica, se implantó previamente a la segunda.

La **etapa pre-Unidad de Ictus** fue desesperanzadora para los neurólogos que fuimos observando cómo la principal línea de investigación, "fármacos denominados neuroprotectores", fraca-

EN PRODUCCIÓN - GALERADAS

saron uno tras otro paralelamente a un aumento de la incidencia del ictus, y a su vez, al envejecimiento de la población, y por tanto, a un incremento absoluto de su mortalidad. Mientras tanto, los avances científicos se enfocaban en otras enfermedades, como por ejemplo las infecciosas, ¿a la tendencia contraria?.

Finalmente, otro pilar importante del tratamiento del ictus, la neurorrehabilitación, ejerce una mayor influencia en las secuelas que en la mortalidad por lo que no será tratada.

En este capítulo nos referiremos principalmente al ictus isquémico ya que solo las medidas terapéuticas de las Unidades de Ictus pueden aplicarse a los ictus hemorrágicos, precisando este tipo de enfermedad vascular cerebral nuevas terapias.

EN PRODUCCIÓN - GALERADAS

2. LA MORTALIDAD DEL ICTUS EN LA ERA PREFIBRINOLISIS. EL ROL DE LAS UNIDADES DE ICTUS: EL PRIMER PASO

Aunque las Unidades de ictus serán estudiadas de manera pormenorizada en otro capítulo, no podemos dejar de hacer en este un esbozo sobre ellas.

Las Unidades de Ictus son unidades de cuidados intermedios, ubicadas en un espacio físico delimitado donde se aplican cuidados y tratamientos estandarizados a los pacientes a través de estrictos protocolos médicos y enfermeros. Con este planteamiento simple, se ha conseguido disminuir la mortalidad y la morbilidad asociada al ictus, isquémico o hemorrágico, de forma espectacular en los países que han conseguido implantarlas.

El nivel de evidencia de las Unidades de Ictus para disminuir la mortalidad es máximo, nivel I, ya que en múltiples trabajos de perfil más metodológico se han observado los mismos resultados positivos.

Un paciente que ingresa en una Unidad de Ictus tiene una mortalidad menor del 17% con respecto a los que no acceden a ellas, y de forma análoga la dependencia funcional disminuye hasta un 25%. El número necesario a tratar (NNT) es de 33 para evitar una muerte y de 20 para evitar dependencia.

Un dato más a favor de estas unidades es que prácticamente todos los pacientes con ictus agudo se benefician de ellas, e in-

EN PRODUCCIÓN - GALERADAS

dependientemente de si el ictus es isquémico o hemorrágico, es decir, su aplicación es casi universal a los pacientes con ictus.

FIGURA 1: Ingresos en el primer año de funcionamiento de la Unidad de Ictus del Hospital Torrecárdenas (Octubre 2018 a Septiembre 2019)

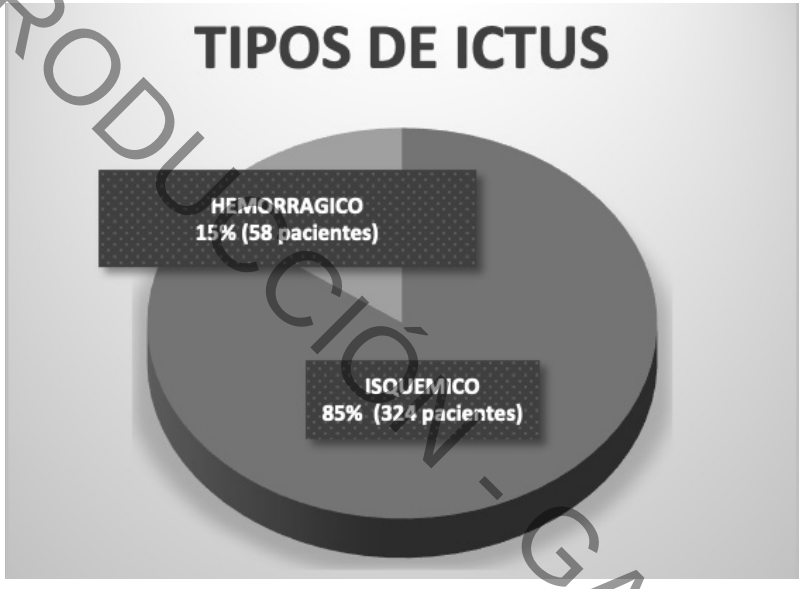


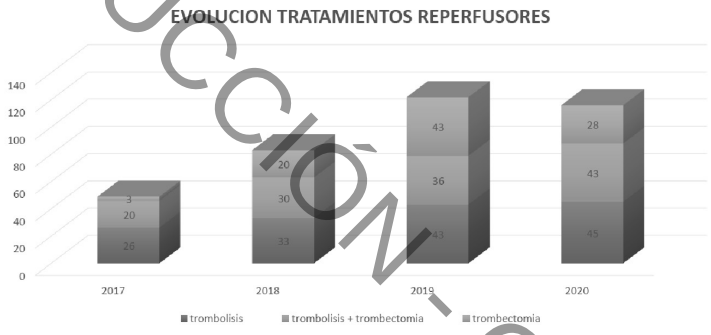
Tabla 1: Ingresos en el período indicado por sexo, edad con datos de estancia media

	TOTAL	HOMBRES	MUJERES
Nº pacientes	382	233	149
Edad media	66,17	64,80	68,31
Estancia media	2,53	2,61	2,41

EN PRODUCCIÓN - GALERADAS

Dentro del ictus isquémico, se estimó en 2018 que un 85% de los pacientes se beneficiaron del ingreso en una Unidad de Ictus, mientras que el tratamiento trombolítico sólo puede aplicarse a un porcentaje inferior al 30% y la trombectomía a un 10%, aunque estas cifras están evolucionando como podemos ver en la figura 2 donde se muestra el número y la evolución de los tratamientos reperfusores en el Hospital Universitario Torrecárdenas de Almería (ver efecto mínimo del año Covid 2020).

Figura 2: Evolución de los tratamientos reperfusores



Primera conclusión: el conocimiento y avances en los cuidados médicos de los pacientes con ictus llegaron a su culminación con la instauración y demostración de los beneficios en mortalidad de las Unidades de Ictus.

EN PRODUCCIÓN - GALERADAS

3. LA INFLUENCIA DE LA FIBRINOLISIS EN LA MORTALIDAD DEL ICTUS

En el año 2007 concluyó en Europa el ensayo clínico fase IV denominado SITS-MOST. Se trataba de un estudio prospectivo iniciado unos años antes y que supuso la demostración definitiva en la práctica clínica habitual de la eficacia de la trombólisis intravenosa en el ictus isquémico.

Atrás quedaron múltiples estudios negativos (ECASS 1, ECASS 2, ALTANTIS A, ATLANTIS B) y algunos positivos (NINDS) qué condujeron a las autoridades sanitarias de Europa a exigir este ensayo post comercialización para su aprobación definitiva.

EL SITS-MOST pretendía demostrar cómo centros con poca experiencia que no habían participado en los ensayos clínicos previos estaban preparados para esta terapia. Sus objetivos fueron determinar si se obtenía una disminución de la morbilidad a los 3 meses y estimar la frecuencia de hemorragias intracraneales, la complicación más temida y grave de la fibrinolisis. Estos objetivos se cumplieron, aunque, como ya se conocía, el margen de beneficio del fármaco era estrecho por lo que su aplicación debe ser bajo condiciones estrictas y no universales. Inicialmente se aprobó para su utilización en las primeras 3 horas tras el ictus y posteriormente un nuevo trabajo demostró que con criterios similares podía ampliarse la ventana terapéutica hasta las **4 horas y media**. Recientemente se ha demostrado, mediante la eviden-

EN PRODUCCIÓN - GALERADAS

cia de la denominada área de penumbra isquémica, que puede aplicarse hasta las 9 horas, previa realización de pruebas complementarias a esta zona de penumbra. Actualmente su uso en este intervalo de tiempo es escaso.

El lector echará en falta en este apartado que aún no se ha hablado de mortalidad. En este sentido, en un paciente subsidiario de tratamiento fibrinolítico, es preciso explicar detalladamente al paciente y/o familiares que los objetivos que se consiguieron en los ensayos clínicos indicados se limitaron a una disminución de las secuelas, sin repercusión en la mortalidad a los 90 días, tras balancear beneficio y complicaciones. Tampoco revisiones sistemáticas como las de la colaboración Cochrane condujeron a otros resultados más esperanzadores.

¿A qué se debe esta aparente contradicción? Se ha atribuido al llamado **efecto de los 10 días**. Este consiste en un paradójico aumento de complicaciones por el tratamiento en los primeros días de su aplicación, la temida hemorragia intracraneal, que incluso podría haber aumentado la mortalidad antes de ese décimo día (séptimo para otros autores). A partir de ese punto, el efecto "deletéreo" de la fibrinolisis va disminuyendo hasta conseguir, a los 90 días, una disminución de la morbilidad.

Para demostrar una disminución de la mortalidad, es preciso realizar un seguimiento a más largo plazo. No obstante, la evidencia actual, aunque no en todas las investigaciones, ha detectado que la mortalidad es menor una vez transcurridos 36 meses, consecuencia de las secuelas inferiores con respecto a los no tratados.

Evidentemente, el efecto de los 7-10 primeros días está muy influenciado por otras variables y factores de confusión como la severidad del infarto cerebral; además, las complicaciones hemorrágicas son más frecuentes si la extensión del ictus isquémico es mayor. Por consiguiente, precisamos análisis estadísticos más complejos y avanzados para alcanzar conclusiones y confirmar hipótesis previas derivadas de la amplia literatura al respecto.

EN PRODUCCIÓN - GALERADAS

Segunda conclusión: la fibrinolisis probablemente solo consigue una disminución de la mortalidad en el ictus isquémico a medio y largo plazo. A corto plazo, 3 meses, su efecto beneficioso no repercute de forma directa en la mortalidad por la presencia de complicaciones potencialmente graves. De hecho, la mera demostración de un menor resultado secuelar -aún con mortalidad sostenida en los supervivientes- no ha sido cuestión baladí y fue motivo de controversias hasta su demostración final en los ensayos clínicos pre y post-comercialización del fibrinolítico alteplasa.

EN PRODUCCIÓN - GALERADAS

4. LA REVOLUCIÓN DE LA TROMBECTOMÍA MECÁNICA EN EL ICTUS

El 1 de enero del año 2015 se publicó en la revista New England Journal of Medicine el resultado del ensayo clínico con trombectomía denominado con el acrónimo "Mr. Clean" que supuso un antes y después en el tratamiento del ictus agudo, ya que fue el primer estudio metodológicamente adecuado que demostraba el efecto positivo de este tratamiento. Pero tampoco el camino hasta ese día fue fácil ya que hubo que superar innumerables dificultades.

La demostración del efecto positivo de la trombectomía se publicó unos años antes de otros trabajos negativos en cuanto a su eficacia. La perspicacia de los investigadores fue rápida y acertada y se esgrimieron distintos factores metodológicos, tipos de dispositivos y tiempos, para argumentar que no había que abandonar este camino y únicamente había que retomar la senda correcta con unas técnicas de trombectomía más modernas y afinando en la metodología de los ensayos. Seguro que en pocos años nuevas tecnologías mejorarían los resultados de la trombectomía mecánica actual.

Por tanto, el año 2015 supuso una revolución en el tratamiento del ictus isquémico e inauguró una nueva era, ya que la trombectomía es la medida terapéutica que consigue los mejores resultados funcionales, de todos los tratamientos disponibles y que en este capítulo hemos repasado.

EN PRODUCCIÓN - GALERADAS

De nuevo recordar que el número potencial de pacientes a los que puede aplicarse es menor al que pueden alcanzar las Unidades de Ictus y la fibrinolisis, ya que precisa la demostración, habitualmente mediante un angioTAC con contraste, de una oclusión de un gran vaso intracraneal accesible para poder emplearla.

Tras la publicación de "Mr Clean", en un breve período, el resto de los ensayos clínicos desarrollados de forma paralela vieron la luz con idénticos y excelentes resultados respecto a las secuelas funcionales a los 90 días, a pesar de algunas diferencias metodológicas. En los años siguientes otras dos investigaciones concluyeron también su exitoso camino, y otras posteriores extendieron la ventana terapéutica de la trombectomía mecánica a las 24 horas, superando rápidamente las 6-8 horas de los primeros ensayos.

¿Qué ocurre con la mortalidad de los ictus tratados con trombectomía mecánica? De forma similar a lo mostrado por la fibrinolisis, la discapacidad secuelar a los 90 días está claramente disminuida en el grupo tratado, pero de nuevo el efecto no es tan claro en la mortalidad. Por ello, demostrar si el beneficio en las secuelas repercute en la mortalidad requiere de estudios a largo plazo, y hasta la fecha no hemos encontrado ningún análisis, con matices que ahora señalaremos, que muestre de forma irrefutable una disminución de las muertes en los pacientes tratados con trombectomía mecánica. Probablemente haya que esperar algunos años para ello y para que ese beneficio en la discapacidad conlleve una disminución de la mortalidad. No obstante, meta-análisis como el realizado por McCarthy y colaboradores en el año 2019 tampoco pudieron demostrar a largo plazo este beneficio, aunque la tendencia era claramente en esa línea.

En el año 2020 otro meta-análisis de 10 estudios ha tratado de forma específica el asunto de la mortalidad en pacientes sometidos a trombectomía. Bien es cierto que dos ensayos clínicos publicados, ESCAPE y DEFUSE III, mostraron una leve disminución de mortalidad en los pacientes sometidos a ese procedimiento.

EN PRODUCCIÓN - GALERADAS

Es conveniente señalar que la valoración del efecto de las terapias en el ictus se realiza con la valoración de la escala de Rankin modificada a los 3 meses, y que la escala codifica con una puntuación de 6 al paciente fallecido por ictus y de 5 a un paciente con dependencia grave. Algunos estudios incluyen este Rankin 5 y 6 objetivo conjunto a evaluar. En el meta-análisis del año 2020, Yaeger y colaboradores, la mortalidad a los 3 meses presentó un intervalo entre 8,6% y 22,2% en el grupo tratado y de un 12,4% al 25,6% en el grupo control con una media de 14,9% y 18,03% respectivamente. El riesgo relativo presentaba un rango entre 0,67 y 0,98. Si consideramos el objetivo a evaluar la muerte o discapacidad severa y el Rankin modificado 5-6, el meta-análisis también muestra resultados estadísticamente significativos, aunque con unos rangos aún más amplios que varían entre el 8,6 y el 30%, y un riesgo relativo de 0,69 comparado con el tratamiento médico estándar. Los resultados de este análisis son homogéneos y por tanto bastante consistentes y concluyen con una reducción en la mortalidad con un NNT de 29, replicando otros trabajos negativos.

Es interesante señalar que cuando se incluyeron solo los pacientes tratados en las primeras 6 horas, los resultados no fueron significativos, y solo tras la ampliación de la ventana hasta las 24 horas se observó el efecto positivo, que debe aplicarse a ictus isquémicos de la circulación anterior incluidos en los ensayos revisados.

Como **conclusión final,** con respecto a la trombectomía mecánica, únicamente con investigaciones amplias y actualizadas, y con una evaluación del pronóstico a largo plazo, se conseguirá demostrar que la trombectomía además de disminuir la discapacidad funcional, mejora la mortalidad del ictus isquémico.

EN PRODUCCIÓN - GALERADAS

BIBLIOGRAFÍA

1. Chwojnicki K, Kozera G, Sobolewski P, Fryze W, Nyka WM. Intravenous thrombolysis and three-year ischemic stroke mortality. Acta Neurol Scand. 2017 May;135(5):540-545.
2. Langhorme P. The Stroke Unit Story: Where Have We Been and Where Are We Going? Cerebrovasc Dis. 2021 Sep 21:1-8.
3. McCarthy DJ, Diaz A, Sheinberg DL, Snelling B, Luther EM, Chen SH, Yavagal DR, Peterson EC, Starke RM. Long-Term Outcomes of Mechanical Thrombectomy for Stroke: A Meta-Analysis. ScientificWorldJournal. 2019 May 2;2019:7403104.
4. Martí-Vilalta JL. Enfermedades vasculares cerebrales. Ediciones Mayo SA, 2012.
5. Díez Tejedor E. Ictus. Una cadena asistencial. Ediciones Mayo SA, 2004.
6. Weber. El atlas del ictus. Andalucía. 2019.
7. Yaeger KA, Martini ML, Hardigan T, Ladner T, Hao Q, Singh IP, Mocco J. Mortality reduction after thrombectomy for acute intracranial large vessel occlusion: meta-analysis of randomized trials. J Neurointerv Surg. 2020 Jun;12(6):568-573. doi: 10.1136/neurintsurg-2019-015383.
8. Yu AYX, Fang J, Kapral MK. One-Year Home-Time and Mortality After Thrombolysis Compared With Nontreated Patients in a Propensity-Matched Analysis. Stroke. 2019 Dec;50(12):3488-3493.

EN PRODUCCIÓN · GALERADAS

EN PRODUCCIÓN - GALERADAS

3. CUIDADOS POST ICTUS EN ATENCIÓN PRIMARIA DE SALUD

AUTORES: Mª DEL MAR RODRÍGUEZ SALVADOR, FRANCISCO JOSÉ MUÑOZ RONDA. JOSÉ GALVÁN ESPINOSA.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

1. INTRODUCCIÓN

El ictus causa en las personas que lo padecen problemas físicos duraderos, deterioro cognitivo y trastornos emocionales, lo que le convierte en una de las enfermedades crónicas más incapacitantes (1). Las consecuencias del ictus pueden afectar negativamente en el nivel de participación social y la calidad de vida de estas personas (2,3) además de suponer un importante impacto económico en la sociedad (4). A nivel mundial, el coste medio por país representa entre el 3% y el 4% del gasto total en salud (5) y se prevé que aumente en el futuro debido al incremento de la prevalencia de esta enfermedad como consecuencia del envejecimiento creciente de la población.

En los años posteriores al episodio, se pueden presentar problemas que afectan a la calidad de vida de la persona, como pérdida de memoria, dolor, espasticidad, fatiga, incontinencia urinaria, deterioro cognitivo, trastornos de la comunicación, discapacidad y limitaciones de la actividad, aislamiento social y alteraciones emocionales (6). El 33% de las personas que han sufrido un ictus no se sienten preparadas para gestionar estos problemas cuando vuelven a su domicilio tras la hospitalización; entre el 18%-46% experimentan problemas sociales y entre 19%-62% problemas emocionales (7). Las necesidades de estos pacientes pueden ser complejas según la fase temporal, la causa y la gravedad del ictus además de otros factores, como la presencia de comorbilidades asociadas. Aproximadamente de 50% a 80% padecerán hipertensión arterial, de 20% a 30% diabetes mellitus y entre el 10%

y el 30% enfermedad cardíaca, fibrilación auricular, enfermedad pulmonar, depresión, ansiedad, enfermedad renal o artritis (8). El ictus isquémico relacionado con COVID-19 provoca peores resultados funcionales y mayor mortalidad no sólo por las características de la COVID-19, sino también por la falta de actividad física debida al aislamiento o a la restricción de acceso a los tratamientos (9). Debido a la limitación de recursos y a las políticas de aislamiento, el acceso a los servicios se ha reducido considerablemente, por lo que la discapacidad ha aumentado aún más. En 2017, se realizaron cerca de 98 millones de visitas relacionadas con ictus en las consultas de Atención Primaria en Europa que supusieron un coste de 3.3 billones de euros; aproximadamente la mitad de este gasto, fue generado por Alemania y España (10).

Tras superar la fase aguda del proceso, Atención Primaria de Salud debe dar respuesta a estos problemas y garantizar las actuaciones necesarias para asegurar una atención centrada en la persona afecta y familia además de prevenir complicaciones y ofrecer la mayor calidad de vida posible.

EN PRODUCCIÓN - GALERADAS

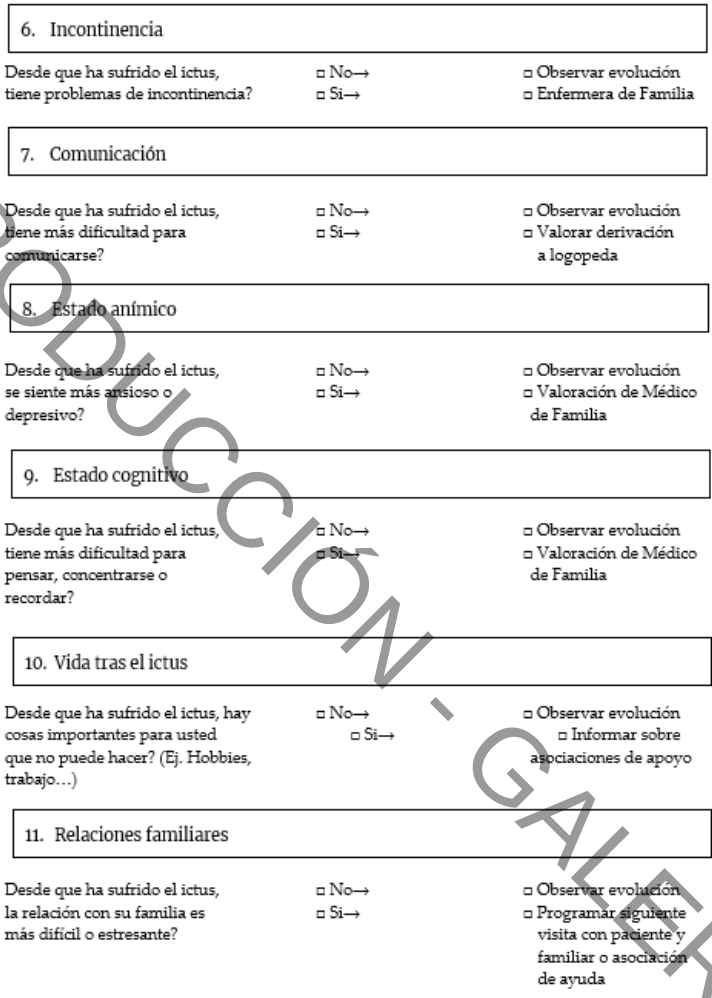
2. VALORACIÓN EN EL ENTORNO COMUNITARIO

Los equipos de Atención Primaria, deben realizar una valoración integral del paciente que ha sido dado de alta hospitalaria y sus cuidadores en las primeras 48 horas de su estancia en domicilio; la visita precoz puede disminuir la probabilidad de reingreso y aparición de complicaciones (11). Los principales objetivos de la atención desde el entorno comunitario se centran en la rehabilitación precoz del paciente, la identificación y el control de factores de riesgo, control de comorbilidades y prevención de complicaciones, la valoración física y emocional del paciente y familia, así como del entorno y recursos disponibles (8). El fin principal es incorporar a la persona a su entorno social en las mejores condiciones posibles y con el máximo grado de autonomía. Existen múltiples herramientas que facilitan la valoración de estos pacientes, como es el Poststroke checklist (PSC) (Tabla 1)(7,11); se trata de un instrumento sencillo y fácil de usar, que identifica las áreas más problemáticas y plantea las intervenciones con mayor impacto en la calidad de vida tras sufrir un ictus.

EN PRODUCCIÓN - GALERADAS

PREGUNTA	RESPUESTA	INTERVENCIÓN
1. Prevención secundaria		
Desde que ha sufrido el ictus, ¿ha recibido consejo sobre cambios en el estilo de vida o cómo manejar su tratamiento?	▫ No→ ▫ Si →	▫ Valorar riesgo y adherencia al tratamiento ▫ Observar evolución
2. Actividades de la vida diaria (AVD)		
Desde que ha sufrido el ictus, ¿ha más dificultad para su autocuidado?	▫ No→ ▫ Si → Tiene dificultad para vestirse, o bañarse? Tiene dificultad para prepararse la comida? Tiene dificultad para salir a la calle?	▫Observar evolución tenido ▫ Potenciar las habilidades existentes y valorar apoyo de fisioterapia o terapia ocupacional
3. Movilidad		
Desde que ha sufrido el ictus, ¿tiene más dificultad para moverse de la cama al sillón?	▫ No→ ▫ Si→ ▫¿Está recibiendo rehabilitación?	▫ Observar evolución ▫ No. Derivar a Unidad de Ictus ▫ Si. Registro en historia clínica y reevaluar en próxima visita
4. Espasticidad		
Desde que ha sufrido el ictus, ¿ha aumentado la rigidez en brazos, manos o piernas?	▫ No→ ▫ Si→ ▫ ¿Interfiere para realizar AVD?	▫ Observar evolución ▫ No. Registro en historia clínica y reevaluar en próxima visita ▫ Si. Derivar a Unidad de Ictus
5. Dolor		
Desde que ha sufrido el ictus, ha aparecido un nuevo dolor?	▫ No→ ▫ Si→	▫ Observar evolución ▫ Derivar a Unidad de Ictus

EN PRODUCCIÓN - GALERADAS



6. Incontinencia		
Desde que ha sufrido el ictus, tiene problemas de incontinencia?	▫ No→ ▫ Si→	▫ Observar evolución ▫ Enfermera de Familia
7. Comunicación		
Desde que ha sufrido el ictus, tiene más dificultad para comunicarse?	▫ No→ ▫ Si→	▫ Observar evolución ▫ Valorar derivación a logopeda
8. Estado anímico		
Desde que ha sufrido el ictus, se siente más ansioso o depresivo?	▫ No→ ▫ Si→	▫ Observar evolución ▫ Valoración de Médico de Familia
9. Estado cognitivo		
Desde que ha sufrido el ictus, tiene más dificultad para pensar, concentrarse o recordar?	▫ No→ ▫ Si→	▫ Observar evolución ▫ Valoración de Médico de Familia
10. Vida tras el ictus		
Desde que ha sufrido el ictus, hay cosas importantes para usted que no puede hacer? (Ej. Hobbies, trabajo…)	▫ No→ ▫ Si→	▫ Observar evolución ▫ Informar sobre asociaciones de apoyo
11. Relaciones familiares		
Desde que ha sufrido el ictus, la relación con su familia es más difícil o estresante?	▫ No→ ▫ Si→	▫ Observar evolución ▫ Programar siguiente visita con paciente y familiar o asociación de ayuda

Tabla 1. Checklist de valoración e intervención Post Ictus

Además de los ítems recogidos en la PSC, resulta imprescindible valorar otras necesidades en pacientes que han sufrido ictus y sus familias, como son (13):

EN PRODUCCIÓN - GALERADAS

- La educación y la capacidad de integrar conocimientos, la formación y el apoyo psicosocial.

- Capacidad para manejar nuevas tecnologías que permitan el acercamiento de los servicios socio-sanitarios al paciente y cuidador (Ej., video consulta, grupos de apoyo online…)

- Estado físico y psicológico del cuidador principal: signos de estrés, ansiedad, depresión, cansancio en el rol, recursos personales para afrontar la situación.

EN PRODUCCIÓN - GALERADAS

3. INTERVENCIONES POST ICTUS EN ATENCIÓN PRIMARIA

Tras el alta hospitalaria, la transición del paciente a su domicilio puede ser compleja según las secuelas que sufra, comorbilidades, recursos sociosanitarios disponibles y complicaciones que puedan surgir (9). Las intervenciones que asume el equipo multidisciplinar de Atención Primaria son principalmente las que se describen a continuación *(Tablas 2, 3, 4, 5)*

3.1. Seguimiento y control de los factores de riesgo para evitar recaídas
3.1.1. Intervenir sobre estilos de vida no saludables (Ej. consumo de alcohol, tabaco, sedentarismo
3.1.2. Exploración de peso y talla y cálculo de índice de masa corporal (IMC)
3.1.3. Control de tensión arterial, cifras de glucemia y colesterol
3.1.3. Control de aparición y/o seguimiento de arritmias (Ej. Fibrilación auricular) y manejo de tratamiento antiagregante en el caso de que hubiera
3.1.4. Valorar la adherencia al tratamiento periódicamente

Tabla 2. Intervenciones para control de factores de riesgo

EN PRODUCCIÓN - GALERADAS

3.2. Prevención de complicaciones
3.2.1. Intervenir para prevenir secuelas y complicaciones físicas: Déficits motores, alteraciones sensitivas, del lenguaje, fatiga, dolor, caídas, espasticidad, incontinencia urinaria, disfunción sexual…
3.2.2. Alteraciones del humor, como depresión, ansiedad, labilidad emocional
3.2.3. Alteraciones cognitivas, como demencia, déficits de atención o alteración de la memoria
3.2.4. Dificultades para realizar actividades de la vida diaria

Tabla 3. Complicaciones más frecuentes post-ictus (9,14)

3.3. Explorar estado general del paciente
3.3.1. Nivel de conciencia, déficits visuales, auditivos. Ofrecer recursos para minimizar el impacto y potenciar la autonomía
3.3.2. Estado nutricional y valorar presencia de problemas de deglución
3.3.3. Permeabilidad vía aérea, movilización de secreciones, reflejo tusígeno
3.3.4. Hidratación de piel y mucosas. Medidas preventivas para el deterioro de la integridad cutánea
3.3.5. Movilidad y entorno: Actividad física adaptada a las capacidades de la persona, riesgo de caídas, uso de ayudas para moverse, adaptación del domicilio a las limitaciones de movilidad. Tratamiento rehabilitador
3.3.6. Control del dolor
3.3.7. Ofrecer alternativas si existen problemas de comunicación (Ej. Uso de nuevas tecnologías)
3.3.8. Intervenciones educativas orientadas al autocuidado sobre su proceso

Tabla 4. Exploración del paciente (15,16)

EN PRODUCCIÓN · GALERADAS

3.4. Valoración del cuidador y entorno sociofamiliar
3.4.1. Identificar síntomas de cansancio del cuidador. Aplicar cuestionario Zarit
3.4.2. Intervenir en caso de claudicación o sobrecarga. Ofrecer recursos de apoyo
3.4.3. Explorar comunicación con la familia y apoyo sociofamiliar
3.4.4. Identificar signos de aislamiento social

Tabla 5. Valoración del cuidador del paciente post-ictus (13)

EN PRODUCCIÓN - GALERADAS

4. CONCLUSIONES

Los sistemas de salud actuales han de ofrecer la oportunidad y los recursos para prevenir el ictus, y que aquellas personas que lo sufran reciban tratamiento lo más rápidamente posible en unidades especializadas, una rehabilitación precoz para maximizar su recuperación, además del apoyo a largo plazo para recuperar la mayor independencia posible. La prevención, el diagnóstico, el tratamiento, la rehabilitación especializada, los cuidados, la integración social, la participación en la vida comunitaria y los cuidados al final de la vida deben estar asegurados. En este proceso, el papel del equipo de salud de Atención Primaria resulta de gran relevancia para garantizar la implementación de todos estos elementos y ofrecer una atención integral y de calidad a pacientes y cuidadores.

EN PRODUCCIÓN - GALERADAS

BIBLIOGRAFÍA

1. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. JAMA. 2002;288:1775–1779.
2. Day CB, Becker CC, Oliveira N, et al. Nursing home care educational intervention for family caregivers of older adults post stroke (SHARE): study protocol for a randomized trial. BioMed Central 2018; 19:96.
3. Edwards JD, Kapral MK, Fang J, Swartz RH. Long-term morbidity and mortality in patients without early complications after stroke or transient ischemic attack. CMAJ 2017; 189: E954-61.
4. Evers SMAA, Struijs JN, Ament André J.H.A., Ament AJ, et al. International comparison of stroke cost studies. Stroke 2004;35:1209–15.
5. Haacke C, Althaus A, Spottke A, Siebert U, Back T, Dodel R. Long-term outcome after stroke: evaluating health-related quality of life using utility measurements. Stroke 2006;37:193-198.
6. Jackson GL, Powers BJ, Chatterjee R, Bettger JP, Kemper AR, Hasselblad V, Dolor RJ, Irvine RJ, Heidenfelder BL, Kendrick AS, et al. The patient centered medical home. A systematic review. Ann Intern Med. 2013; 158:169–178.
7. Johnson CO, Nguyen M, Roth GA, et al. Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol 2019;18:439–58.
8. Kernan WN, Viera AJ, Billinger SA, Bravata DM, Stark SL, Kasner SE, Kuritzky L, Towfighi A; on behalf of the American

EN PRODUCCIÓN - GAFERADAS

Heart Association Stroke Council; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology and Intervention; and Council on Peripheral Vascular Disease. Primary care of adult patients after stroke: a scientific statement from the American Heart Association/American Stroke Association. Stroke. 2021;52:e558–e571.
9. Luengo-Fernández R, Candio P, et al. At what cost: The economic impact of stroke in Europe. Stroke Action Plan for Europe 2018-2030.
10. Murray J, Ashworth R, Forster A, Young J. Developing a primary care–based stroke service: a review of the qualitative literature. Br J Gen Pract 2003;53:137-142.
11. Ntaios G, Michel P, Georgiopoulos G, et al. Characteristics and outcomes in patients with COVID-19 and acute ischemic stroke: the global COVID-19 stroke registry. Stroke. 2020;51(9):e254–e258.
12. Philp I, Btainin M, Walker MF, et al. Development of a poststroke checklist to standardize follow-up care for stroke survivors. Journal of Stroke and Cerebrovascular Diseases 2013; 22: 173-80.
13. Rajsic S, Gothe H, Borba HH, et al. Economic burden of stroke: a systematic review on post-stroke care. Eur J Health Econ 2019;20:107–34.
14. Terman SW, Reeves MJ, Skolarus LE, Burke JF. Association between early outpatient visits and readmissions after ischemic stroke. Circ Cardiovasc Qual Outcomes. 2018;11:e004024.
15. Törnbom K, Persson HC, Lundälv J, et al. The impact of physical function on participation in the first year post-stroke. Acta Neurol Scand 2017;135:649–55.
16. Verberne DPJ, Van Mastrigt GAPG, Ponds RWHM, et al. Economic evaluation of nurse-led stroke aftercare addressing long-term psychosocial outcome: a comparison to care-as- usual. BMJ Open 2021;11:e039201.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN GALERADAS

4. ANÁLISIS DE LA CAPTACIÓN DE FINANCIACIÓN DE PROYECTOS SOBRE EL ICTUS EN ANDALUCÍA EN EL PERÍODO 2019-2021.

AUTORES: JOSÉ GALVÁN ESPINOSA. JUAN MANUEL GARCÍA TORRECILLAS. REGINA RODRÍGUEZ MARTÍNEZ. ADRIÁN APARICIO MOTA.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - CAFERADAS

INTRODUCCIÓN

El análisis de los datos internos de España y de comparativas internacionales con países de nuestro entorno geográfico y económico indican que la sanidad pública española tiene un claro déficit de financiación (1).

España está todavía a gran distancia de Europa en inversión en I+D. Según los datos publicados por el INE, el gasto en I+D interna alcanzó los 17.249 millones de euros en 2021, el 1,43% del PIB. Si lo comparamos con el año 2020 supone un incremento del 9,4%, teniendo en cuenta que la inversión se paralizó a consecuencia del impacto de la pandemia.

Sin embargo, esta cifra permanece lejos del compromiso europeo de dedicar al menos el 2% del PIB en I+D. España destina menos inversión que otros países europeos. Alemania invierte más del 3% de su PIB a I+D y Francia, el 2,3% [Fuente Eurostat] (2).

Conscientes de ello, el Ministerio de Ciencia e Innovación, mediante la Estrategia Española de Ciencia, Tecnología e Innovación 2021-2027, pretende duplicar la suma de inversiones pública y privada, hasta alcanzar la media europea en 2027 (desde el 1,24% del PIB en inversión en I+D+I registrado en 2018, hasta el 2,12 % en 2027).

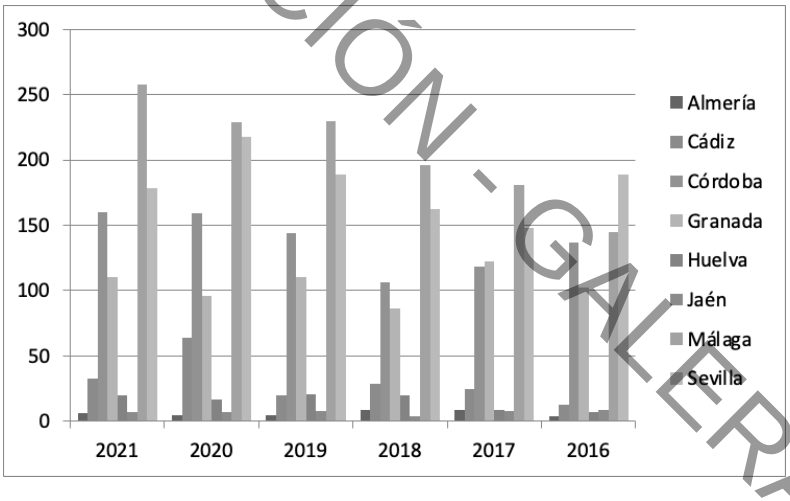
En Andalucía, la inversión en investigación es un tema relevante como motor clave para el desarrollo económico, social y cultural de una región.

La inversión en investigación en Andalucía ha aumentado de manera sostenida en los últimos años, alcanzando cifras récord

en 2021. Según el informe de la Agencia Andaluza del Conocimiento, en 2021 se destinaron más de 1.200 millones de euros a investigación, lo que representa un aumento del 6,5% con respecto al año anterior.

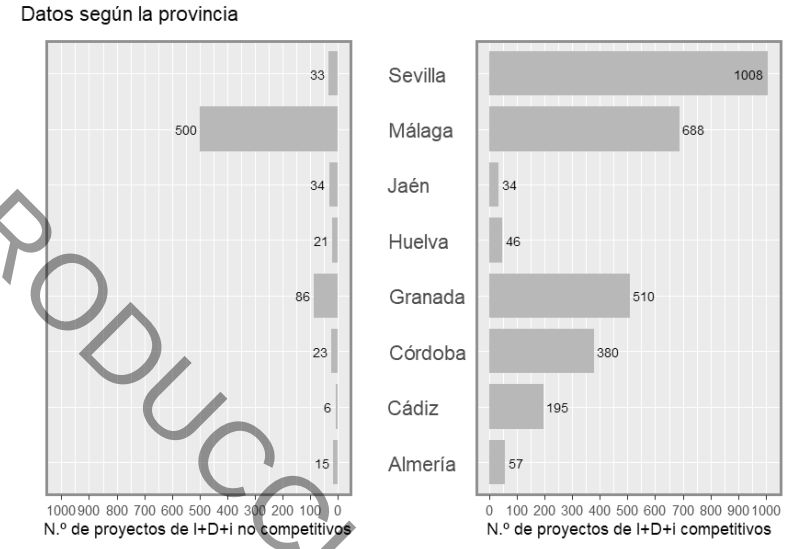
Con respecto a la investigación en salud, tanto a nivel nacional como autonómico, los resultados en I+D+i en Andalucía han experimentado un crecimiento en la mayoría de los indicadores con los que se miden los resultados, en parte, gracias a la inversión de la Consejería de Salud y Consumo en I+D+i en los últimos años con una concesión media anual de más de 17 M€ y 200 ayudas.

N.º de ayudas de I+D+i captadas durante el período 2016-2021 por provincias



Atendiendo al origen de la financiación, público o privado, es decir, si se ha obtenido a través de convocatorias públicas de financiación en concurrencia competitiva o no, la distribución por provincias es la que se muestra en la siguiente figura.

EN PRODUCCIÓN - GALERADAS



FUENTE: BOJA

Territorio	2021	2020	2019	2018	2017	2016	Total
Almería	204.611,30 €	100.848,25 €	128.297,60 €	142.531,90 €	223.041,03 €	115.078,53 €	914.408,61 €
Cádiz	1.782.148,42 €	4.075.920,48 €	1.679.845,47 €	1.154.101,58 €	5.272.843,91 €	770.538,86 €	14.735.398,72 €
Córdoba	10.420.530,27 €	14.826.766,57 €	10.530.417,39 €	6.161.707,26 €	5.479.851,61 €	9.453.395,56 €	56.872.668,66 €
Granada	11.832.244,06 €	9.100.075,72 €	20.257.728,11 €	6.901.927,34 €	8.558.490,66 €	9.354.717,94 €	66.005.123,83 €
Huelva	333.698,31 €	441.926,73 €	395.817,01 €	443.204,78 €	41.511,27 €	61.097,23 €	1.717.255,33 €
Jaén	248.658,52 €	303.885,75 €	468.498,80 €	314.659,54 €	147.785,96 €	600.302,06 €	2.083.790,63 €
Málaga	13.926.586,19 €	13.921.191,90 €	12.571.770,98 €	12.845.706,12 €	10.179.500,40 €	10.475.764,59 €	73.920.520,18 €
Sevilla	25.881.288,56 €	26.674.781,02 €	19.974.909,66 €	16.677.489,07 €	11.891.150,69 €	18.260.717,80 €	119.360.336,80 €
Andalucía	64.629.765,63 €	69.445.396,42 €	66.007.285,02 €	44.641.327,59 €	41.794.115,53 €	49.091.612,57 €	**335.609.502,76 €**

Fuente: Estadísticas de investigación biomédica en el Sistema Sanitario Público de Andalucía. Investiga+

Como puede observarse las provincias de Córdoba, Málaga y Sevilla han sido las que más proyectos y por ende financiación han captado.

En este marco estratégico juega un papel fundamental la Red de Fundaciones Gestoras de la Investigación del Sistema Sanitario Público de Andalucía (RFGI-SSPA). Su principal

EN PRODUCCIÓN NO ALTERADAS

objetivo es prestar apoyo a los profesionales del SSPA en el desarrollo de su actividad investigadora. Esta actividad está regulada principalmente en el Convenio de Colaboración entre el SAS y las Fundaciones Gestoras de la I+D+i del SSPA firmado en junio de 2019 y el Convenio Específico de colaboración para el desarrollo del marco de cooperación entre el SAS y la FPS en el ámbito de la protección y transferencia de los resultados derivados de la investigación, firmado en diciembre de 2021.

La RFGI-SSPA está compuesta por siete fundaciones: FCÁDIZ (Cádiz), FIBICO (Córdoba), FABIS (Huelva), FIMABIS (Málaga), FISEVI (Sevilla), FIBAO (con carácter interprovincial, dando cobertura a Granada, Jaén y Almería) y la Fundación Pública Andaluza Progreso y Salud (FPS) como entidad coordinadora de la Red.

De la Cartera de Servicios de Gestión y Apoyo a la I+D+i que la RFGI ofrece a los profesionales del SSPA nos vamos a enfocar en el servicio de asesoramiento para la captación de financiación y establecimiento de colaboraciones.

Centrándonos en el ictus y p*artiendo de los resultados de la captación de financiación para el desarrollo de proyectos de investigación de la RFGI-SSPA*, *se ha analizado si existe* proporcionalidad *entre la preocupación*, *la atención* e importancia *que recibe esta patología,* como se ha ido demostrando en los capítulos precedentes, *con la inversión que recibe* para su investigación.

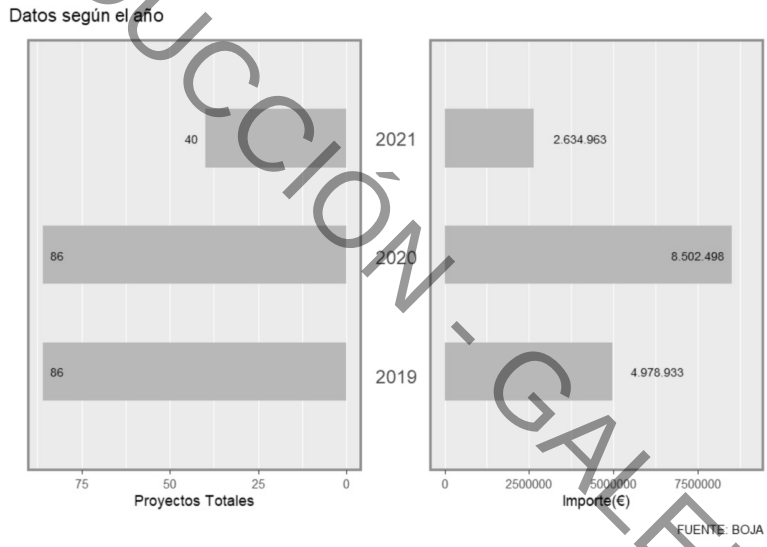
*Para ello se ha*n *realizado* dos estudios, que se describen en lo que sigue.

El primero *es un análisis* descriptivo a partir de 21*2 registros extraídos de las resoluciones definitivas de financiación publicadas en la aplicación Gestión de Convocatorias de la Consejería de Salud y* Consumo *de la Junta de Andalucía*, e*n el período 2019-2021*. De este análisis se obtuvo que el mon-

EN PRODUCCIÓN - GALERADAS

tante total destinado a financiar proyectos de investigación en el periodo indicado ha sido ***16.116.394***€ para un total de 212 proyectos.

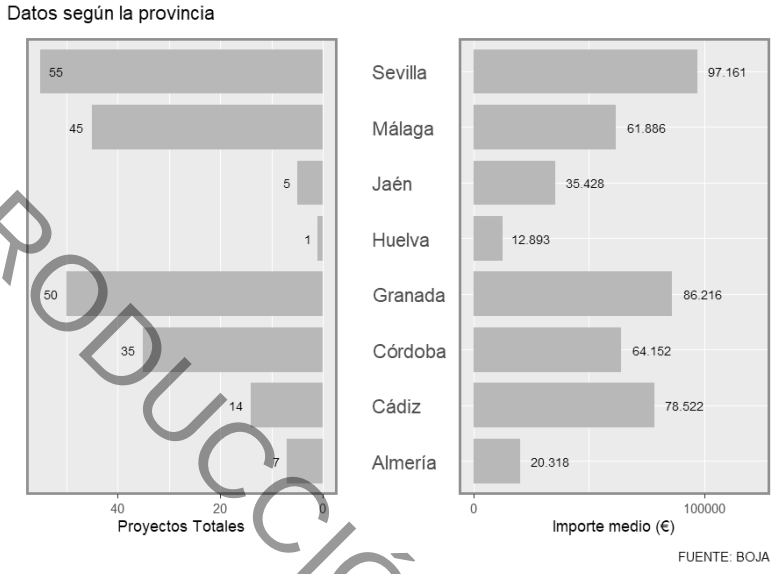
El número máximo de proyectos financiados fue de 86, en los años 2019 y 2020, sin embargo, con una importante diferencia de inversión entre ambos. En 2019 fue de 4.978.933€ mientras que en 2020 fue de 8.502.498€. En cambio el número mínimo de proyectos realizados fue de 40 en el año 2021 con tan solo 2.634.963€ de inversión.



El análisis por provincias muestra datos de interés, observándose que la tasa de éxito en la captación de financiación es mucho mayor en unas provincias con respecto a otras.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

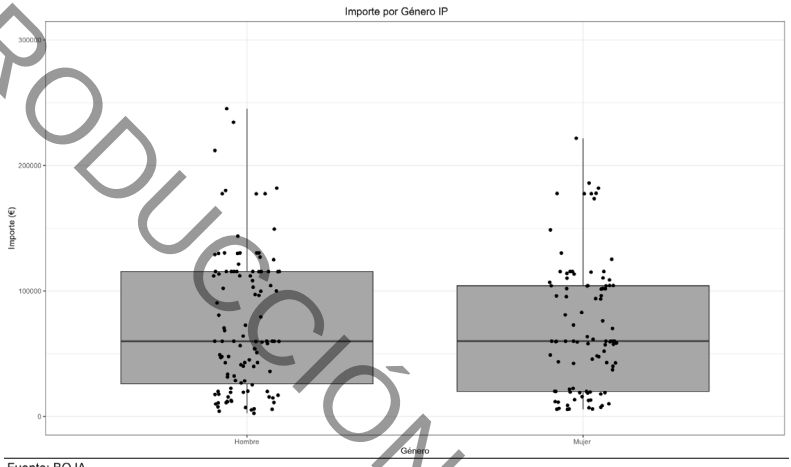


El número de proyectos financiados por provincias en el intervalo del estudio fue el siguiente:

	Importe (Euros)			
	Media	Suma	Número	%
Almería	20.317,75 €	142.224,27 €	7	3%
Cádiz	78.522,24 €	1.099.311,34 €	14	7%
Córdoba	64.151,84 €	2.245.314,40 €	35	17%
Granada	86.215,71 €	4.310.785,70 €	50	24%
Huelva	12.892,89 €	12.892,89 €	1	0%
Jaén	35.427,63 €	177.138,16 €	5	2%
Málaga	61.885,87 €	2.784.864,11 €	45	21%
Sevilla	97.161,16 €	5.343.863,64 €	55	26%
Total		16.116.394,52 €	212	100%

Se observa que sólo el 3,3% del total de proyectos financiados en Andalucía, corresponde a Almería, concretamente **142.224,26€,** mientras que Granada, Málaga y Sevilla reciben un total de **4.310.785,70€, 2.784.864,11€ y 5.343.863,64**, respectivamente.

Perspectiva de género y financiación:



Fuente: BOJA

Este gráfico box-plot muestra el resultado del análisis del importe según el género del investigador principal.

Podemos observar que apenas existen diferencias entre ambos grupos, ya que las medianas de cada uno son muy similares, tomando un valor cercano a los 60.000€. Además en los hombres el valor máximo se encuentra en torno a 150.000€ mientras que en las mujeres en torno a 125.000€.

Los datos demuestran que la Consejería de Salud y Consumo cumple con el compromiso adquirido a través de las políticas de igualdad de género y los planes de igualdad definidos en el artículo 46 de la Ley Orgánica 3/2007 de 22 de marzo, para la igualdad efectiva de mujeres y hombres. La diferencia entre hombres y mujeres en cuanto la obtención de financiación en las convocatorias indicadas no es significativa.

EN PRODUCCIÓN - GALERADAS

En el segundo estudio se analizaron un total de 321 proyectos solicitados a la convocatoria de financiación de la Consejería de Salud y Familias en la línea estratégica de fomento de la investigación de la AP y CHARES y Hospitales Comarcales. Este trabajo evidenció la siguiente distribución por provincias en Andalucía: Málaga (24,6%), Sevilla (19,9%), Córdoba (14,3%) y Almería (11,5%), resto acumulado de provincias 29,5 %. Financiaciones mayoritarias: Córdoba (34,78%), Almería (32,43%), Málaga (31,65%), Sevilla (31,25%). La financiación total ascendió a 1.571.797,18€. Las tres provincias con mayor media de financiación fueron Cádiz (32.755,83€), Granada (26.415,15€) y Sevilla (21.024,01€). De las 321 solicitudes, 4 proyectos estuvieron relacionados con neurología (3 financiados). Un único proyecto relacionado con el Ictus fue financiado (Hospital de Poniente/Almería). Los otros 3 proyectos relacionados con neurología fueron de la provincia de Málaga. La financiación media para Neurología fue de 27.051,54 Euros.

Como conclusión general de estos trabajos podemos inferir que la financiación de esta línea de investigación sigue siendo deficiente con respecto a otras patologías.

Todos los datos analizados se han ceñido al ámbito autonómico y no se ha tenido en cuenta la financiación de otras líneas de actuación, tales como infraestructuras, recursos humanos, captación del talento, estancias formativas, etc.. Sin embargo, consideramos relevante poner de manifiesto los siguientes logros por su impacto positivo para el estudio del Ictus en la provincia de Almería.

1.- La creación de la Red NEURORECA (www.neuroreca.com), la Red Andaluza de Investigación Clínica y Traslacional en Neurología, cuya misión es articular una red de investigación básica y clínica en Neurología que permita incrementar la capacidad de realización de EECC en Andalucía, traduciéndose en

EN PRODUCCIÓN. GALERADAS

beneficio para el paciente y ahorro para el Sistema Andaluz de Salud y en la que participa activamente investigadores de todas las provincias andaluzas.

2.- La obtención de financiación para la participación como grupo en las Redes de Investigación Cooperativa Orientadas al Resultado en Salud (RICORS), https://ricors-ictus.es/, que pretende, promover la I+D+i entre grupos de investigación procedentes de diferentes comunidades autónomas orientando su actividad investigadora hacia objetivos comunes que permitan la obtención de resultados trasladables a la población.

3.- El Hospital Universitario Torrecárdenas recibió el máximo galardón de la Organización Europea del Ictus obteniendo el nivel 'Diamante' por su "excelencia" en el abordaje de esta dolencia. Éste ha sido el primer centro público andaluz en conseguir este galardón que en esta edición también han recibido los hospitales de Bellvitge, Hospital del Mar y Hospital Vall d'Hebron.

4.- Qué duda cabe de la importancia de la multidisciplinariedad en la investigación. Por tal hecho destaca el grupo de investigación Neurotics (www.neurotics.es), compuesto por clínicos de diferentes especialidades, pero también por epidemiólogos, matemáticos, gestores e informáticos que han logrado el primer registro de ictus de la provincia pese a no contar con financiación.

Para concluir a la luz de los resultados analizados podemos decir que Málaga, Sevilla y Córdoba no son sólo referentes en investigación en otras patologías sino también en el estudio del ictus y en la captación de financiación para la Investigación. Almería, pese a no compartir ese status, se encuentra en posición destacable, impulsando, consolidando y aportando nuevas herramientas para su investigación.

EN PRODUCCIÓN - GALERADAS

A pesar de la preocupación por parte de la Consejería de Salud y Consumo ante esta enfermedad, todo ello manifestado y recogido en el Plan Andaluz de Atención al Ictus, es de las líneas de investigación que obtiene menos financiación pública para su investigación, lo que provoca que la generación de nuevos conocimientos para su tratamiento y mejor abordaje, sea más complejo de alcanzar.

Esta inequidad se presenta como una oportunidad para los organismos públicos de analizar cuáles son los factores que la provocan y poner los mecanismos correctivos o de ajustes necesarios.

EN PRODUCCIÓN - GALERADAS

BIBLIOGRAFÍA

1. Evolución del gasto en servicios sanitarios antes y después del ictus isquémico: análisis de base poblacional. DOI: 10.1016/j.nrl.2018.11.002. Disponible online el 20 de Marzo de 2019. [Open Access]
2. Consejo de Europa, Estrategia de Igualdad de Género 2018-2023; Directiva 2006/54/CE, de 5 de Julio de 2006; La Carta de los Derechos Fundamentales de la Unión Europea (artículos 21 y 23).
3. En el Estado Español: Artículo 14 de la Constitución Española. La Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de hombres y mujeres.
4. En la Comunidad Autónoma Andaluza: Artículo 15 del Estatuto de Autonomía para Andalucía. Ley 9/2018, de 8 de octubre, de modificación de la Ley 12/2007, para la promoción de la igualdad de género en Andalucía.

REFERENCIAS ADICIONALES:

- Gestión de convocatorias: http://fps.junta-andalucia.es/fundacionprogresoysalud/gestionconvocatorias/
- Investiga+: https://investigamas.junta-andalucia.es/fundacionprogresoysalud/investigamas/solucion/recRecursosProyectos/1260
- Instituto de Estadística y Cartografía de Andalucía: https://www.ieca.junta-andalucia.es/institutodeestadisticaycartografia/badea/informe/anual?idNode=26598

EN PRODUCCIÓN GALERADAS

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

5. MANEJO DEL ICTUS DE MODO PROTOCOLIZADO EN LOS SERVICIOS DE URGENCIAS

AUTORES: BERNABÉ ORTIZ PUERTAS. REGINA RODRÍGUEZ MARTÍNEZ. MARÍA ORTIZ PUERTAS. PATRICIA MARTÍNEZ SÁNCHEZ.

EN PRODUCCIÓN - GALERADAS

El ictus en fase aguda, a pesar del uso de protocolos consensuados de actuación, las nuevas terapias y la ampliación del periodo ventana para la realización de trombectomía, continúa siendo una patología compleja y grave.

Las cifras hablan por sí solas, en Europa fallecen unas 650.000 personas al año y cada 40 segundos, alguien sufre un ictus. En España alrededor de 120.000 personas sufren un ictus cada año, de las cuales el 50% quedan con secuelas discapacitantes o fallecen (el número de fallecimientos es de 40.000 personas al año). De cada 6 españoles 1 tendrá un ictus a lo largo de su vida, doblándose esta proporción si supera los 65 años.

En el caso de Andalucía la incidencia del ictus está en torno a los 200 casos por 100.000 habitantes/año, es decir, más de 17.000 personas sufren un ictus cada año.

A pesar de estas cifras tan contundentes, es un hecho que la mortalidad por enfermedad cerebrovascular ha descendido a lo largo de los últimos 20 años. Sin embargo, el envejecimiento de la población en Europa, cuyo incremento se estima en un 35% entre 2017-2050, conllevará también un incremento esperable de la incidencia del ictus.

A pesar de los avances en el conocimiento, manejo e investigación de esta enfermedad, el ictus representa actualmente en España, según datos del Instituto Nacional de Estadística (INE) la primera causa de defunción entre las mujeres y la segunda entre los hombres después de las enfermedades isquémicas del corazón.

El ictus es además el principal causante de deterioro cognitivo, demencia y epilepsia. Es por tanto una enfermedad asociada a un elevado coste socio-sanitario, representando la primera causa de discapacidad debido a sus secuelas en la población adulta. Dos tercios de los pacientes que sobreviven presentan algún tipo de secuela física o psíquica. En la actualidad más de 330.000 españoles tienen alguna limitación funcional tras haber sufrido un ictus.

EN PRODUCCIÓN - GALERADAS

El coste estimado ocasionado por el ictus en España supone 8,5 millones de € por año. El coste directo sanitario es de 2 millones de euros por año; el indirecto, por las secuelas ocasionadas, de 6,5 millones de euros.

NECESIDAD DE UN PROTOCOLO DE ACTUACIÓN COMÚN

La publicación en New England Journal of Medicine del ensayo clínico NINDS rt-PA Stroke Trial, hace ya más de dos décadas, que demostraba beneficios del tratamiento con trombolisis sistémica, supuso un cambio conceptual de esta enfermedad, enfatizando en la precocidad del tratamiento y convirtiéndola en una enfermedad tiempo dependiente, que debe ser considerada y manejada como una emergencia médica.

Es así como se definió en la Estrategia en Ictus del Sistema Nacional de Salud, publicada en 2009, que realizó un profundo análisis de la situación epidemiológica, la prevención y el tratamiento de las enfermedades cerebrovasculares en España. Teniendo como referente la Oficina Regional Europea de la OMS, en la Declaración de Helsinborg, publicada en 1995, junto con organismos científico-técnicos, estableció unos objetivos de mejora en todos los niveles de la atención sanitaria del ictus. En él se planteaba la necesidad de recursos humanos, materiales y formación específica de profesionales de emergencias para consolidar el objetivo principal: disminuir la incidencia y la morbimortalidad de los pacientes con ictus.

Es en este contexto donde se desarrolla un protocolo de actuación, que iba a denominarse Código Ictus (el término código hace referencia a todas las patologías tiempo-dependientes) que buscaba una actuación rápida y eficiente en el tratamiento de la fase aguda del ictus.

El Código Ictus aboga tanto por el reconocimiento precoz de los signos y síntomas sugerentes de ictus por parte de la población

EN PRODUCCIÓN - GALERADAS

general y del personal sanitario en todos los niveles asistenciales, como por la actuación inmediata por parte de los servicios de urgencias y emergencias, y la puesta en marcha de todos los recursos disponibles para dar un tratamiento inmediato, ya que el conjunto de estas actuaciones aumenta la posibilidad de supervivencia y la disminución de las secuelas.

Con el consenso, el esfuerzo y la acción conjunta de todos los agentes implicados (comunidades autónomas, Instituto Nacional de Gestión Sanitaria, Ministerio de Sanidad, sociedades científicas y asociaciones de pacientes) su implantación ha redundado en una notable mejora en el abordaje de esta enfermedad que ha llevado acompañado una significativa disminución de la mortalidad y una mayor equidad en el tratamiento del ictus en España.

Además, con el avance del conocimiento, el protocolo se ha enriquecido con la mejora de las recomendaciones encaminadas a la neuroprotección en la fase aguda mediante el control de factores de riesgo vascular, observándose un impacto en el pronóstico. También ha quedado demostrado los beneficios que el neurointervencionismo, dentro de unas indicaciones bien definidas, puede ofrecer. Por último tanto la rehabilitación precoz como las intervenciones no farmacológicas, dirigidas a la modificación del estilo de vida, son en la actualidad recomendaciones contrastadas en el tratamiento del ictus.

CÓDIGO ICTUS EN URGENCIAS

En el organigrama del proceso de atención al ictus en fase aguda, el papel de los servicios de Urgencias resulta clave, tanto en la fase prehospitalaria, donde la detección precoz y activación del código ictus es de capital importancia, como en la fase hospitalaria, donde se realiza valoración, estabilización del paciente, solicitud de pruebas complementarias y se participa en la toma de decisiones terapéuticas.

EN PRODUCCIÓN. GALERADAS

Son diferentes especialidades las que intervienen en el código ictus, siendo fundamental una actuación multidisciplinar, coordinada y eficaz en aras de un objetivo común: disminuir la alta morbimortalidad asociada a esta patología.

El ictus constituye un motivo habitual de consulta en los servicios de urgencias hospitalarios, siendo la causa más frecuente de demanda asistencial por enfermedad neurológica grave y la primera causa de ingreso por enfermedad neurológica.

El ictus se manifiesta como un déficit neurológico focal de inicio agudo. El diagnóstico es esencialmente clínico y depende de una buena anamnesis y exploración clínica. Las pruebas complementarias van destinadas a descartar una hemorragia intracraneal u otras enfermedades que puedan simular un ictus.

En la actualidad existen dos tratamientos aprobados en la fase aguda del ictus isquémico cuyo objetivo es la recanalización, lo más precoz posible, de la arteria ocluida responsable de la clínica del paciente, uno es intravenoso y, el otro, intraarterial.

La primera de las opciones terapéuticas sería la fibrinólisis, que consiste en la lisis química del trombo mediante la administración intravenosa de un fármaco y está indicada en las primeras 4,5 horas del inicio de la clínica (las mayores tasas de éxito se producen en la primera hora y media). Tradicionalmente el fármaco utilizado ha sido la Alteplasa (Actilyse®). En los últimos años se han publicado varios ensayos clínicos con Tecneteplasa (Metalyse®) aunque por el momento el grado de evidencia es menor. Los fármacos trombolíticos están contraindicados en pacientes con alto riesgo de sangrado y además sólo consiguen la recanalización precoz en arterias principales intracraneales en aproximadamente un 10-20% de los pacientes. Es por ello que en los últimos años se han desarrollado técnicas endovasculares como la trombectomía mecánica que consiste en la extracción del trombo utilizando dispositivos recuperables (stent-retriever) o mediante tromboaspiración. Estas técnicas están indicadas en las primeras 6 horas

EN PRODUCCIÓN - GALERADAS

del inicio de la clínica aunque, en un subgrupo de pacientes seleccionados por neuroimagen, el beneficio podría alcanzar hasta las primeras 24 horas. La ventaja de la trombectomía mecánica es que consigue la recanalización arterial en más del 70% de los casos y que se puede realizar en pacientes con alto riesgo de sangrado.

Ambos tratamientos demuestran mayor eficacia cuanto antes se administren, y esta es la razón por la que optimizar los tiempos de instauración del tratamiento endovascular y trombolítico es uno de los objetivos principales para mejorar el pronóstico de los pacientes con ictus.

CADENA ASISTENCIAL DEL ICTUS

Urgencias Extrahospitalarias

Se define cadena asistencial del ictus, a todos aquellos procedimientos y actuaciones que se realizan de forma secuencial, desde la activación inicial del sistema sanitario, de aquellas personas que sufren un cuadro clínico compatible con un evento cerebrovascular, se continúa con un proceso diagnóstico, hasta que se les administran los cuidados y tratamientos definitivos. El objetivo principal de esta cadena asistencial es acortar al máximo los tiempos de respuesta desde el inicio de los síntomas, hasta un tratamiento definitivo.

La Cadena Asistencial del Ictus se inicia en el reconocimiento de los síntomas y signos sugerentes de que un paciente sufre un Ictus Agudo. La puerta de entrada al sistema sanitario sería:

- La familia o paciente acude por sus propios medios a un servicio de urgencias.
- Médico de atención primaria (aviso domiciliario).
- Centro de coordinación de emergencias: 061/112. Identificación de pacientes que están sufriendo un Ictus tanto

EN PRODUCCIÓN · GALERADAS

desde la sala de coordinación de Urgencias y Emergencias (mediante cuestionario telefónico específico, basado en la escala de Cincinnati, riesgo vital y valoración calidad de vida). También desde los diferentes recursos sanitarios móviles que atienden in situ a estos pacientes.

Existen herramientas validadas en el medio prehospitalario que pueden facilitar la detección de ictus y su gravedad. Ejemplo de ellas son la escala Cincinnati o FAST. La escala Cincinnati mediante tres ítems evalúa el riesgo de haber sufrido un ictus. Valora la asimetría facial, la fuerza en brazos y el lenguaje. La presencia de uno de los tres parámetros (asimetría facial, descenso del brazo, lenguaje anormal) se corresponde con un 72% de probabilidades de ictus y la existencia de 2 los eleva hasta un 85%.

Además se emplearán cuestionarios para identificar la situación basal del paciente, como la escala de Rankin modificada (ERm):

0. Asintomático.

1. Sin incapacidad significativa. Síntomas y síntomas. Realiza actividades laborales y sociales.

2. Incapacidad leve. Incapaz de realizar todas las actividades previas. Atiende sus asuntos sin ayuda.

3. Incapacidad moderada. Camina sin ayuda de otra persona. Precisa ayuda para algunas tareas. Necesidad de cuidador al menos 2 veces en semana.

4. Incapacidad moderadamente grave. Ayuda para caminar y actividades básicas. Cuidador 1 vez/día.

5. Incapacidad grave: encamado, incontinente. Atención constante.

6. Exitus.

EN PRODUCCIÓN - GALERADAS

¿CUÁNDO ACTIVAR UN CÓDIGO ICTUS?

- Presencia de déficit neurológico agudo.
- No existe límite de edad. En caso de pacientes menores de 14 se debe contactar con el servicio de pediatría.
- Situación basal previa del paciente: funcionalmente independiente, con escala de Rankin <3.
- Tiempo menor de 4,5 horas desde el comienzo de los síntomas para trombolisis endovenosa y hasta 24 horas para intervencionismo.
- Ictus del despertar o de inicio desconocido: no excluye el código ictus.
- Ausencia de enfermedad terminal o demencia.
- Nivel de consciencia: la situación de coma no invalida el código ictus, pero en estos casos, el paciente debe ser trasladado a un hospital con unidad de ictus, con servicio de neuroradiología y neurocirugía.

Una vez detectado el paciente candidato a activación de Código Ictus, será el Centro Coordinador de Emergencias quién se comunicará con el centro receptor, facilitando los datos relevantes del paciente y su situación clínica al Neurólogo de guardia (filiación y antecedentes personales relevantes; hora de inicio de los síntomas; focalidad neurológica y NIHSS: situación clínica general y tratamiento farmacológico empleado; tiempo estimado de llegada al hospital y recurso de traslado).

Una vez activado el Código Ictus se gestionará el traslado al hospital movilizando los recursos que aseguren una derivación rápida al hospital con una neuroprotección adecuada y sin detrimento significativo en el tiempo de traslado.

La activación del recurso más adecuado, se hará en función de la prioridad otorgada al paciente y de la distancia geográfica entre el lugar de asistencia y el hospital útil, así como del tipo de ictus que

EN PRODUCCIÓN - GALERADAS

se sospecha. El objetivo será disminuir el tiempo desde el primer contacto con el sistema sanitario hasta la atención hospitalaria (< 45 minutos), de los pacientes catalogados como Código Ictus.

SERVICIO DE URGENCIAS HOSPITALARIAS

En la cadena asistencial del código ictus, los servicios de urgencias hospitalarios constituyen la puerta de entrada al circuito intrahospitalario. La activación del código ictus en urgencias implica la actuación coordinada de todos los componentes del equipo, cada uno de los cuales tiene protocolizadas sus tareas y debe cumplirlas sin interferir en las acciones realizadas por el resto de miembros. Los servicios de urgencias tienen procesos y vías para el manejo de estos pacientes, que incluyen la capacidad de recibir, identificar, evaluar, tratar, así como coordinar las actividades y transferir la información con los diferentes servicios implicados en la asistencia de dichos pacientes a nivel hospitalario: radiología, laboratorio y neurología.

El código ictus hospitalario se activará en las siguientes situaciones:

- Tras activar un código ictus extrahospitalario.
- Desde la consulta de triaje al identificar un paciente con síntomas de focalidad neurológica aguda, con un inicio de los síntomas inferior a 24 horas.
- En pacientes con una mRS ≤3 con clínica de ictus.
- Pacientes con ictus del despertar o de inicio incierto o desconocido.

Al activar el código ictus, se pondrá en marcha una sistemática protocolizada de actuación y una serie de procedimientos que permitirán la rápida realización de una valoración clínica y las pruebas diagnósticas oportunas.

EN PRODUCCIÓN - GALERADAS

Cuando la ambulancia llega a Urgencias, el circuito de admisión hospitalaria del paciente con sospecha de ictus agudo consiste en la atención inmediata en un box de críticos donde se transfiere la información sobre el inicio de los síntomas, la clínica y los antecedentes patológicos al médico de Urgencias.

Se comprueban y facilitan los datos administrativos al servicio de Admisión de Urgencias Hospitalarias.

Ante un paciente con ictus, no se debe retrasar la valoración en el área de Triaje. Se confirmarán los datos del paciente y se le asignará la prioridad correspondiente. Se evitará la toma de constantes, ya que se realizará de forma inmediata al iniciar la asistencia.

En el supuesto de haber sido derivado con preaviso, se realizará la preparación para la recepción avisando y pre-alertando al servicio de Neurología y Radiología.

MANEJO INICIAL DEL PACIENTE CON ICTUS.

Ante la sospecha de Código Ictus es importante optimizar el protocolo ABC a la vez que se realiza la anamnesis y exploración completas.

La anamnesis debe ser exhaustiva, describir los síntomas que presenta, establecer con la mayor certeza posible la hora de inicio de los síntomas y su cronopatología (cómo y en qué orden se fueron instaurando), indicar la forma de inicio (súbita o progresiva) y recoger otras circunstancias asociadas (fiebre, convulsiones, traumatismo craneal previo, cefalea, etc.). También debe incluir los antecedentes personales, factores de riesgo (HTA, diabetes mellitus, dislipemias, fibrilación auricular, obesidad, cardiopatía previa, ictus previos), hábitos tóxicos, existencia de neoplasias (localización, estirpe, estadio, tratamiento, presencia de metástasis), cirugías previas y las fechas de las mismas, situación funcional

EN PRODUCCIÓN - GALERADAS

previa (puntuación en la escala ERm previa al evento, tratamiento habitual (sobre todo si toma anticoagulantes) y alergias.

El médico de urgencias además se encargará de la estabilización del paciente, usando la valoración ABCDE con medidas de soporte vital avanzado si es necesario. Se trata de una valoración secuencial y sistemática que incluye (lo primero al valorar un paciente): A - Vía Aérea libre-; B- Respiración; C-Circulación; D-Neurológico; E-Exposición.

Se llevará a cabo exploración física y neurológica básica, utilizando la escala de NIHSS, del inglés National Institutes of Health Stroke Scale que consta de 11 parámetros y un rango de puntuación de 0 a 42 (nivel de conciencia, orientación, obediencia a órdenes sencillas, mirada conjugada, campos visuales, paresia facial, fuerza en las cuatro extremidades, ataxia, sensibilidad, lenguaje, disartria y atención), aplicando puntuación a cada ítem según exploración, estableciendo una puntuación total, tanto por el médico de urgencias como por el neurólogo. La escala NIHSS tiene una fuerte correlación pronóstica, siendo un ictus grave si> 20 puntos. Además se recogerá la puntuación en la escala de coma de Glasgow (utiliza tres parámetros: la respuesta verbal, la respuesta ocular y la respuesta motora, para determinar el nivel de conciencia) y la situación funcional previa con la escala Rankin.

Enfermería por su parte, realizará la toma de constantes (presión arterial, frecuencia cardiaca, temperatura, saturación basal y glucemia capilar). Además, canalizará dos vías venosas periféricas, preferentemente una en cada miembro superior, con suero salino fisiológico (SSF 1.500 ml/24h) y llave de tres vías (dos vías preparadas ante la necesidad de administrar contraste para completar el estudio diagnóstico y otra para la administración del fibrinolítico en el posible tratamiento de reperfusión). No usar suero glucosado por aumento del riesgo de edema cerebral.

El enfermero se encarga de la extracción de las diferentes muestras, si no las aporta previamente y las identificará como CÓDIGO

EN PRODUCCIÓN - GALERADAS

ICTUS cursándolas con personal de celadores en mano para no retrasar los resultados. Se solicitará hemograma con recuento celular y plaquetas, bioquímica con glucosa, urea, creatinina, iones y pH, estudio básico de coagulación (tiempo de trombina, tromboplastina parcial activada, INR, protrombina), otras determinaciones según el caso clínico (marcadores de lesión miocárdica, osmolaridad, etc.)

También se realizará el ECG. En el caso de que ya se haya realizado por los servicios de emergencia extrahospitalarios, no se repetirá para no retrasar los tiempos de asistencia. Se realizará monitorización cardiaca y pulsioximetría.

Todas las actuaciones descritas se realizarán en <10 minutos desde llegada a Urgencias.

Tras la valoración del paciente se decide si existen criterios para realizar tratamientos de reperfusión. Durante dicha valoración son precisas unas medidas de soporte para evitar un deterioro neurológico precoz, es el "ABC terapéutico":

- El traslado se realizará en decúbito supino con elevación de la cabecera de la cama a 30º con el fin de evitar una posible broncoaspiración.
- Oxigenoterapia si la "saturación de oxígeno" es inferior al 94%. En los pacientes con tendencia a la hipercapnia (EPOC), se consideran cifras inferiores al 92%, como límite para administrar oxígeno.
- IOT en Glasgow <8 y <9 (intubación preventiva), con la menor FiO2 posible y garantizar una CO2 normalizada. Aislamiento de la vía aérea en caso de una puntuación inferior a 9 de la escala de Glasgow.

También deben iniciarse medidas de Neuroprotección no farmacológica:

EN PRODUCCIÓN - GALERADAS

- Manejo de la TA. El tratamiento antihipertensivo difiere según el perfil clínico del paciente. Por regla general hay que evitar descensos bruscos de la presión arterial en los casos que se encuentre elevada para mantener una adecuada presión de perfusión cerebral, sin administrar fármacos hipotensores por vía sublingual. Si el paciente no es candidato a la activación del código ictus, se tratará si la tensión es de TAS 220 y/o TAD 120 mmHg. Si el paciente es candidato a la activación del código + fibrinólisis se tratará si la TAS 180 y TAD 105 mmHg. Se usarán Labetalol iv y, si existe alguna contraindicación para el mismo, se podría utilizar nicardipino o clevidipino.
  En caso de hipotensión arterial se debe buscar la causa (cardiopatía isquémica, disección de aorta, embolia pulmonar, hemorragia digestiva) y tratarla en consecuencia. En caso de no tener diagnóstico se puede emplear expansores del plasma como suero salino fisiológico a razón de 10ml/kg.
- Control de la glucemia. Se debe tratar la hiperglucemia cuando supere cifras de 155 mg/dl, con insulina rápida según glucemia capilar. La hipoglucemia (<70 mg/dl) se trata con soluciones glucosadas.
- Control de la fiebre. El control se realizará con temperatura timpánica cada 4 horas durante las primeras 48 horas de ingreso, y posteriormente cada 8 h.
  Si la temperatura es > 37,5ºC se usarán medidas físicas junto a la administración de Paracetamol. Si la Temperatura es > 38,5ºC se deben realizar hemocultivos, urocultivos, sedimento de orina y Rx tórax.
- Control de náuseas y vómitos. Asociar antieméticos endovenosos tipo metoclopramida (10 mg) u ondansetrón (8 mg). En pacientes en coma, con vómitos importantes o disfagia, está indicada la colocación de una sonda nasogás-

EN PRODUCCIÓN - GALERADAS

trica. Sin embargo, no se debe usar sonda nasogástrica si el paciente es candidato a fibrinolisis.

- Tratamiento de las convulsiones: benzodiacepinas por vía endovenosa como diazepam (2,5-10 mg) o midazolam (2-5 mg). En el caso de crisis repetidas o estatus se utilizarían fármacos antiepilépticos endovenosos (ej: levetiracetam).
- Control de la agitación: midazolam (2-5 mg) o haloperidol (5 mg), ambos por vía endovenosa.
- Si el paciente presenta signos de hipertensión intracraneal, se administrará Manitol al 20%, 250 ml a pasar en 30 minutos.
- Por el contrario nunca deberemos hacer:

    - No administrar soluciones glucosados, salvo en casos de hipoglucemia.
    - No administrar antiagregantes, anticoagulantes, heparina subcutánea antes de la realización de las pruebas de neuroimagen
    - Evitar sondaje vesical, sonda nasogástrica y punciones arteriales, salvo en aquellos casos en los que se considere estrictamente necesario.

DIAGNÓSTICO DIFERENCIAL EN LA SOSPECHA DE ICTUS

Existe un grupo de patologías que se confunden con ictus y se denominan "simuladores de ictus" o "stroke mimics", las cuales pueden confundirse con una enfermedad vascular cerebral. Estos simuladores de ictus suponen un 20-30% de los pacientes ingresados con un diagnóstico de ictus en los primeros momentos. Por ello ante todo evento cerebrovascular agudo se debe plantear un diagnóstico diferencial con las siguientes patologías:

EN PRODUCCIÓN - GALERADAS

- Convulsiones: una historia previa de convulsiones, una crisis comicial presenciada y periodo postcrítico inclinan la balanza hacia el diagnóstico de convulsión.
- Migraña: la existencia de episodios previos similares, un cuadro precedido por aura y el dolor de cabeza, hacen sospechar esta patología.
- Hipoglucemia: El control glucémico siempre estará indicado en un paciente con clínica neurológica aguda, sobre todo en pacientes diabéticos. Es fácil de descartar, con un tratamiento rápido y eficaz.
- Encefalopatía hipertensiva: dolor de cabeza, delirium, hipertensión arterial significativa, edema cerebral.
- Trastorno por conversión: la ausencia de afectación de los pares craneales, hallazgos neurológicos sin distribución vascular y un examen físico con inconsciencia.
- Otros: tumores, hematoma subdural, infecciones, esclerosis múltiple, delirio.

¿SE HA ACTIVADO EL CÓDIGO ICTUS?

Una vez activado el Código Ictus, se procede a la activación del dispositivo de diagnóstico de la filiación del mismo; el aviso urgente al neurólogo y radiólogo de guardia.

Una vez el paciente valorado por el médico de urgencias en colaboración con el neurólogo, será trasladado para la realización de pruebas de imagen. A continuación, y en el menor tiempo posible, se realiza la tomografía computarizada (TC) cerebral simple, sin contraste, técnica rápida y eficiente para el diagnóstico de hemorragia cerebral, que permite el diagnóstico precoz del ictus isquémico. Si tras esta primera prueba el paciente es candidato para la terapia endovascular, se iniciará lo antes posible en la sala TC. A continuación se realizará un angio-TC de arterias cerebrales intra y extracraneales. En el caso de ictus del despertar, de

EN PRODUCCIÓN - GALERADAS

inicio desconocido o fuera de ventana temporal, se valorará la realización de una TC de perfusión que nos ayudará a decidir la indicación de tratamientos reperfusores en esos casos.

Técnicas de Neuroimagen:

- TC craneal sin contraste:

- Permite excluir la hemorragia.
- La fibrinólisis no estará recomendada si hay una un infarto establecido > del 33% de un hemisferio cerebral (riesgo transformación hemorrágica).
- Signos precoces de isquemia: pequeñas áreas de hipodensidad, pérdida de distinción entre sustancias gris y blanca, oscurecimiento del núcleo lenticular, signo de la arteria hiperintensa.
- ASPECTS (Alberta Stroke Program Early CT Score) - valoración de ictus isquémicos de la circulación anterior- divide territorio de la ACM en 10 regiones (cada un 1 punto) - restar 1 punto por cada región con cambio isquémico precoz - 10 puntos: TC normal- <= 7 puntos: alta M-M y mala recuperación funcional. Una puntuación de <6 en la escala ASPECTS contraindica la TM (aunque no se utiliza para decidir la fibrinolisis).

- TC de perfusión (con contraste): informa de la perfusión cerebral distingue entre:

- núcleo del infarto o Core (irreversible)
- área de penumbra (tejido potencialmente recuperable)
- área de oligohemia "benigna" (raramente se infarta)
- tejido sano diferencia o "mismatch" entre Core y tejido hipoperfundido: área de penumbra si > 20%: indicación de TIV Angio-TC (incluir troncos supra aórticos) para la localización vaso obstruido

EN PRODUCCIÓN - GALERADAS

Una vez realizadas las pruebas de imagen cerebral se pueden dar las siguientes circunstancias:

Paciente con ictus isquémico, pero sin evidencia de obstrucción de gran vaso:

- Candidato a tratamiento fibrinolítico intravenoso:
  - En las primeras 4,5 horas desde el inicio de los síntomas (6 h. si clínica de territorio basilar).
  - Hora de inicio desconocida si existe imagen de "mismatch" en el TC-Perfusión.
  - Hora de inicio de 4,5 a 9 horas, podría considerarse si imagen de "mismatch" en el TC-Perfusión.

  ° Será tratado con alteplasa y sería deseable, si las circunstancias lo permiten, iniciar tratamiento fibrinolítico en la sala TC y posteriormente subir a la Unidad de Ictus para continuar tratamiento y cuidados, con asistencia del personal de la Unidad de Ictus y neurólogo de guardia.

- No candidato a tratamiento fibrinolítico:
  ° Sube directamente a la unidad de ictus

Paciente con ictus isquémico con evidencia de obstrucción de gran vaso con Equipo de neurointervencionismo presente en el hospital:

Terapia endovascular:

- En las primeras 6 horas (12 horas si oclusión de arteria basilar y 24 horas si oclusión basilar con clínica progresiva).

EN PRODUCCIÓN - GALERADAS

- Hora de inicio desconocido o 6-24 horas si imagen de "mismatch" en TC-P y edad ≤90 años.

- Candidato a tratamiento fibrinolítico:
  - Sería deseable, si las circunstancias lo permiten, iniciar tratamiento fibrinolítico con tenecteplase en el mismo TAC y mientras se prepara la sala de intervencionismo, continuar el tratamiento en la misma área de intervencionismo (antesala acondicionada para tal), con asistencia del personal de enfermería de urgencias y el neurólogo de guardia.
- No candidato a fibrinolisis
  - Mientras se prepara la sala de intervencionismo, esperar en la misma área de intervencionismo (antesala acondicionada para tal), con asistencia del personal de enfermería de urgencias y el neurólogo de guardia.

EN PRODUCCIÓN - GALERADAS

BIBLIOGRAFÍA

1. Caplan LR. Etiology, classification, and epidemiology of stroke. In: UpToDate, Post TW (Ed), UpToDate, Waltham, MA. Accessed.
2. Instituto Nacional de Estadística. Defunciones según causa de muerte 2018. Disponible en: http:// www.ine.es/.
3. Instituto Nacional de Estadística. Encuesta Nacional de morbilidad Hospitalaria. Disponible en: http:// www.ine.es/
4. Ministerio de Sanidad y Política Social. Estrategia en Ictus del Sistema Nacional de Salud. Madrid: Ministerio de Sanidad y Política Social; 2009. Disponible en: https://www.mscbs.gob.es/organizacion/sns/planCalidadSNS/docs/EstrategiaIctusSNS.pdf
5. Díaz-Guzmán J, Egido JA, Gabriel-Sánchez R, Barberá-Comes G, Fuentes-Gimeno B, Fernández-Pérez C; IBERICTUS Study Investigators of the Stroke Project of the Spanish Cerebrovascular Diseases Study Group. Stroke and transient ischemic attack incidence rate in Spain: the IBERICTUS study. Cerebrovasc Dis. 2012;34(4):272-81.
6. Kjellstrom T, Norrving B, Shatchkute A. Helsingborg Declaration 2006 on European stroke strategies. Cerebrovasc Dis, 2007. 23(2-3): p. 231-241.
7. Powers WJ. Acute Ischemic Stroke. N Engl J Med. 2020;383(3):252–60.
8. Powers WJ, Rabinstein A, Ackerson T, et al. Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management

EN PRODUCCIÓN - GALERADAS

of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 2019;50:e344-e418.
9. Simal-Hernández P, Matías J , Hernández-Meléndez T, Aparicio-Azcárraga P. Comité Institucional de la Estrategia en ICTUS del SNS. Logros y retos en la atención del ictus en España: desde la estrategia del sistema nacional de salud al plan de acción europeo 2018-2030. Rev. Esp. Salud Pública vol.95 Madrid 2021 Epub 04-Jul-2022
10. Gómez-Angelats E, Bragulat Obach V, Gómez-Choco M, Sánchez M, Miró O. Resultados alcanzados con la puesta en marcha del circuito "Código Ictus" en un gran hospital: papel de urgencias y análisis de la curva de aprendizaje. Emergencias. 2009;21(2):105-113.
11. Jiménez M, Alcazar P, Alés E, Aranda F, Arraez M, Carrillo E, et al. Plan Andaluz de atención al Ictus: 2011-2014. Sevilla: Junta de Andalucía. Consejería de Salud; 2014.
12. Comité Local de Atención al Ictus. (2021). Protocolo asistencial del ictus del Hospital Universitario Torrecárdenas (v.11). Servicio Andaluz de Salud. Consejería de Salud y Familias.
13. Montaner J, Àlvarez-Sabín J. La escala de ictus del National Institute of Health (NIHSS) y su adaptación al español. Neurología. 2006;21(4):192-202
14. Gállego J, Herrera M, Jericó I, Muñoz R, Aymerich N, Martínez-Vila E. El ictus en el siglo XXI: Tratamiento de urgencia. En: Anales del Sistema Sanitario de Navarra. Gobierno de Navarra. Departamento de Salud. 2008;31:15-29.
15. Aguilar MJB, Oviedo AM. Código ictus en urgencias del Hospital Obispo Polanco de Teruel. Atalaya Médica Turolense. 2013;(4):25-35.
16. Jiménez Fàbrega X, Espila JL. Códigos de activación en urgencias y emergencias: La utilidad de priorizar. En: Anales del Sistema Sanitario de Navarra. Gobierno de Navarra. Departamento de Salud. 2010;33:77-88.

EN PRODUCCIÓN - GALERADAS

17. Leal JS. Protocolo de diagnóstico y atención del ictus en urgencias. Medicine-Programa de Formación Médica Continuada Acreditado. 2019;12(70):4124-4129.
18. Meza HT, Inda IS, Ponz MS, Ara JR, Moreno JM. Impacto en los tiempos puerta-aguja de un conjunto de medidas para optimizar la atención hospitalaria del código ictus. Neurología. 2020.
19. Aguilera Manrique G. Unidades de rehabilitación de ictus. Situación en España. Aproximación al ROL enfermero. 2017.
20. Jiménez Murillo L, Montero Pérez FJ. Medicina de Urgencias y Emergencias. Guía Diagnóstica y Protocolos de Actuación. 4th ed. Editorial: Elsevier España, SL, Madrid. 2012.
21. Aranda Aguilar F, Borja Padilla J, Ramírez Castillo A, Saldaña Ruiz JL, Lerma Dorado FJ, Molina Delgado RI, López Aguilar M, Lobato Martínez, Martín Pérez P, Montero González IM. Procesos Neurológicos EPES 2019.
22. Sociedad Española de Medicina de Urgencias y Emergencias, Grupo de Trabajo Neuro SEMES. ICTUS: una emergencia del pasado con una mirada al futuro. Mayo 2023.
23. Valle J, et al. "Imitadores de ictus: un reto para el médico de emergencias". Sist Sanit Navar. 2014;37(1):117-128.
24. Jauch EJ, Saver J, Adams H, et al. Guidelines for the early management of patients with acute ischemic stroke. A Guideline for healthcare professionals from the American Heart Association / American Stroke Association. Stroke. March 31, 2013.
25. Berge E, Whiteley W, Auderbert H, et al. European Stroke Organisation (ESO) guidelines on intravenous thrombolysis for acute ischemic stroke. Eur Stroke J. 2021;6(1):I-LXII.
26. Mendelson SJ, Prabhakaran S. Diagnosis and management of transient attack and acute ischemic stroke. JAMA. 2021;325(11):1088-1098.

EN PRODUCCIÓN. CAMBIADAS

EN PRODUCCIÓN. GALERADAS

6. APLICACIONES MÓVILES COMO MEDIO DE OBTENER ESTIMACIONES RÁPIDAS Y FIABLES A PARTIR DE MODELOS COMPLEJOS EN ICTUS

AUTORES: JUAN JOSÉ LÓPEZ RAMOS. REGINA RODRÍGUEZ MARTÍNEZ. JOSÉ GALVÁN ESPINOSA. JOSÉ ANTONIO MORALES MOLINA.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

1.- INTRODUCCIÓN

La potencia de los smartphones y el uso de aplicaciones móviles enfocadas a la salud están creciendo exponencialmente. Estas aplicaciones han transformado y mejorado muchos aspectos de la práctica clínica, ayudando al diagnóstico, monitorización y tratamiento de enfermedades. De hecho, la Unión Internacional de Telecomunicaciones (ITU) y la OMS recomiendan el uso de tecnología móvil como aplicaciones y mensajería para el combate de enfermedades no transmisibles desde 2012.

Actualmente, hay más de 97000 aplicaciones de salud disponibles para descargar, de las cuales el 70% están destinadas al bienestar y deporte, mientras que el 30% restante son exclusivas para pacientes y profesionales de la salud. Además, se ha detectado que el 50% de los médicos utilizan aplicaciones móviles en su práctica.

Si nos enfocamos en el software, observamos que Apple se ha asociado con universidades para comercializar un framework (marco de trabajo) de desarrollo enfocado a las apps de salud, llamado Health Kit. Google está haciendo lo propio en Android y, de este modo, muchas otras grandes multinacionales del sector.

En cuanto al hardware, también está evolucionando para que los sanitarios cada vez dispongan de más y mejor información del paciente y esta sea de fácil acceso. Es el caso de smartwatches, que se han revelado como dispositivos ideales para recopilar información, debido a que incorporan sensores como pulsómetros, saturímetros, tensiómetros, frecuencia cardiaca, o también pequeños

dispositivos como sensores cutáneos para la diabetes. Todo ello aumenta la cantidad de datos que se pueden recoger del paciente, permitiendo la monitorización domiciliaria.

Así mismo, las aseguradoras están apostando cada vez más por este tipo de apps, sobre todo desde la contingencia ocasionada por el SARS-Cov-2, que ha provocado dificultad en los desplazamientos y en las atenciones sanitarias, favoreciendo la teleasistencia y acceso a historial desde aplicaciones móviles.

Esta fuerte inversión en tecnología se debe principalmente a que la combinación de estos dispositivos con software especializado permite crear aplicaciones, destinadas a pacientes y sanitarios, con las siguientes características fundamentales:

- Recogida de datos en tiempo real.
- Incrementa el volumen de datos recogidos.
- Favorece la comunicación del paciente con su médico.
- Evita desplazamientos.

EN PRODUCCIÓN - GALERADAS

2. DESARROLLO DE UNA APP MÓVIL PARA DETERMINAR EL RIESGO DE FALLECIMIENTO DE UN PACIENTE POR ICTUS ISQUÉMICO

En 2017, en el seno de un proyecto de investigación del Hospital Universitario Torrecárdenas se desarrolló un modelo predictivo de la mortalidad por ictus isquémico; este modelo predictivo fue implementado dentro de una aplicación que se detallará posteriormente. Precisamente la característica que ha permitido que se haya gestado dicha aplicación ha sido la accesibilidad de los móviles, la cual permite recoger datos por un médico en cualquier parte y momento de su trabajo y trasladarlos a una app móvil, así como la capacidad de procesamiento de estos dispositivos.

2.1.- PRIMER CONTACTO

Lo que el equipo de investigación precisaba fue el desarrollo de una aplicación móvil que permitiera el cálculo rápido, y a pie de cama, de un modelo de riesgo para la estimación de la probabilidad de fallecer en pacientes con ictus isquémico durante el primoingreso.

En el abordaje de dicho proyecto, los dos aspectos fundamentales que necesitaron un mayor esfuerzo fueron la planificación y adaptación al lenguaje técnico sanitario. Esto último se desarrolló

EN PRODUCCIÓN - GALERADAS

fácilmente debido a una interacción ágil y fluida con el equipo de investigación. La planificación, en cambio, llevó algo más de trabajo.

2.2.- PLANIFICACIÓN

Se requería desarrollar una aplicación para Android y considerar su desarrollo para alguna otra plataforma. Dicha aplicación debía ser intuitiva, de fácil uso, rápida y cuyo código pudiera adaptarse a plataformas como la web o iOS.

Partiendo de estos requerimientos, de entrada se descartó el desarrollo en lenguaje nativo, y se elaboró un informe con las tecnologías y framework de desarrollo de aplicaciones híbridas que había disponibles en aquel momento. Se analizaron y valoraron alternativas como Ionic, Xamarin, React Native, Cordova, Nativescript, entre otros.

En el momento de abordar el proyecto, se disponía de conocimientos de desarrollo Web general y de Angular. Finalmente, teniendo en cuenta todos estos aspectos, se decidió hacer el desarrollo con Ionic (versión 3). Ionic v3 es un framework potente, agradable de abordar si se ha hecho desarrollo web previamente, en continuo crecimiento y mejora. Además, Ionic goza de una gran comunidad, estupenda documentación y una enorme versatilidad para desarrollar proyectos multiplataforma (app y web) con nulas o mínimas modificaciones en el código. Está orientado a PWA (Aplicación Web Progresiva), que son un conjunto de técnicas de desarrollo web que permite, entre otras funcionalidades, que la página pueda cargar a gran velocidad y seguir funcionando a pleno rendimiento sin conexión a internet y que el usuario la sienta como una aplicación nativa o de escritorio.

EN PRODUCCIÓN GALERADAS

2.3.-DESARROLLO

Typescript y Visual Studio Code

Una de las bondades de trabajar con Ionic es que permite usar el lenguaje Typescript, un superconjunto de Javascript que dispone, entre otras cosas, del tipado estático y objetos basados en clases. Esto, combinado con un buen editor de código como Visual Studio Code, que proporciona excelentes plugins, permite una velocidad de desarrollo mayor, al evitar errores de tipado en tiempo de desarrollo y proponiendo código útil al desarrollador, como por ejemplo los atributos de un objeto o rutas de importación.

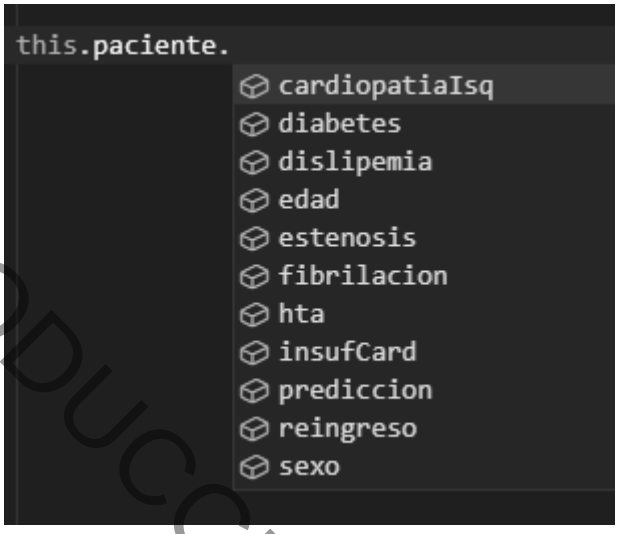
En el código, se aplicó esto creando el modelo o clase "Paciente", donde fueron determinados los atributos del paciente así como el tipo de datos de cada uno. Estos atributos se encargarían de recoger los datos que el usuario final introducirá en el formulario, como se verá más adelante.

src > paciente > TS paciente.ts > ...

```
1   export class Paciente{
2     edad:number;
3     sexo:number;
4     reingreso:number;
5     cardiopatiaIsq:number;
6     hta:number;
7     diabetes:number;
8     fibrilacion:number;
9     dislipemia:number;
10    insufCard:number;
11    estenosis:number;
12    prediccion:number;
13  }
14
```

Gracias a esto, el editor puede proponer los atributos de la clase, evitando errores de tipología.

EN PRODUCCIÓN GALERADAS



Es una ayuda que agiliza el desarrollo.

Otra característica de Typescript que ayudó, fue la posibilidad de declarar variables con la palabra reservada "**let**", la cual permite a esa variable vivir únicamente en el scope local del bloque donde está contenida. Por ejemplo, si dentro de la función donde se realiza la predicción, en el interior de un bloque, se declara una variable de esta forma, `let fx1 = 0;` y fuera de ese bloque existe otra homónima "fx1", para Typescript, esas variables son distintas. De hecho, como Typescript es un lenguaje compilado, si inspeccionamos el código Javascript plano resultante de la compilación, observaremos que internamente ambas han sido declaradas con la palabra reservada "var", pero con nombres diferentes.

Aquí se muestra cómo se realiza el cálculo de la predicción, una vez se han recogido todos los valores del formulario:

EN PRODUCCIÓN - GALERADAS

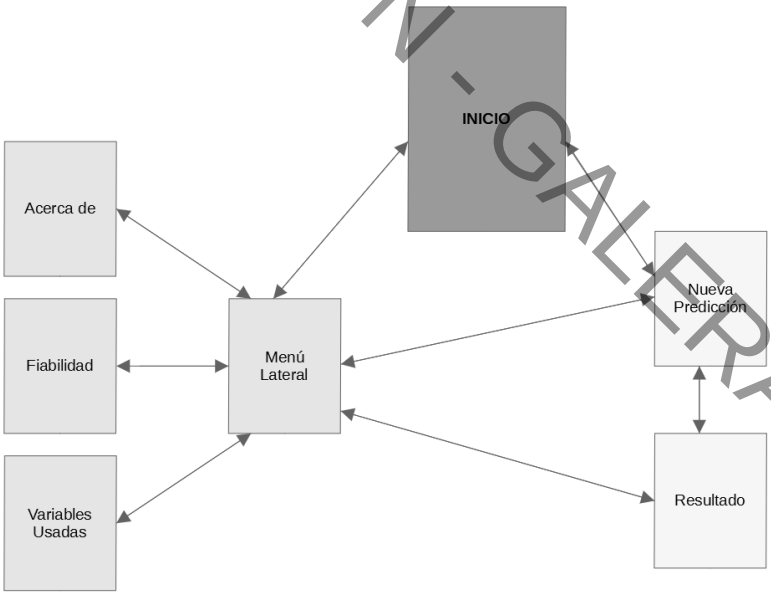
```
let fx1 = 0;
const e = Math.E;

fx1 = -7.852475 + 0.0666109*this.paciente.edad +
0.1888656*this.paciente.sexo +
0.6794386*this.paciente.reingreso +
0.2873691*this.paciente.cardiopatiaIsq -
0.3355499*this.paciente.hta +
0.0900304*this.paciente.diabetes +
0.4281204*this.paciente.fibrilacion -
0.4405003*this.paciente.dislipemia +
0.3973937*this.paciente.insufCard +
1.039144*this.paciente.estenosis;

let efx1 = Math.pow(e,fx1);

this.paciente.prediccion = ((efx1/(1+efx1))*100);
```

Figura 1. Flujo de la aplicación

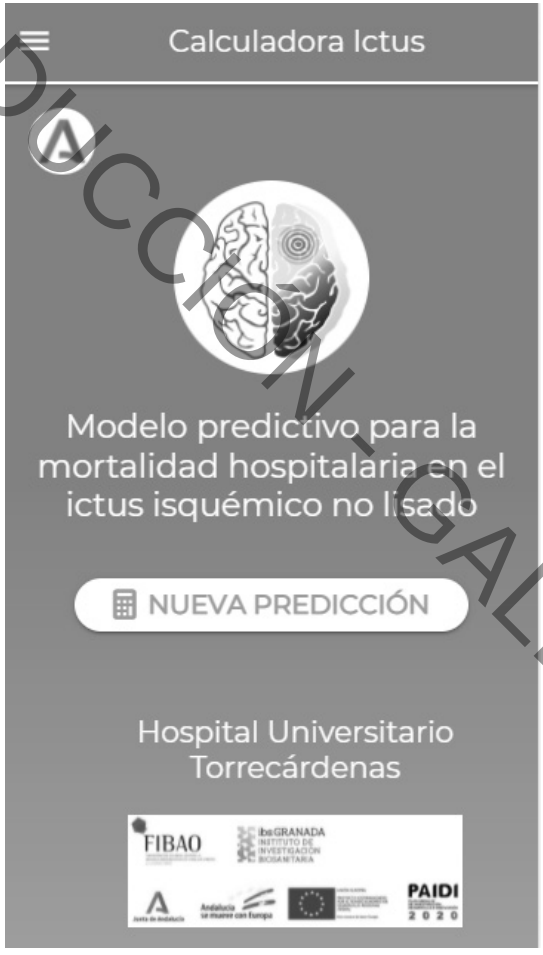


EN PRODUCCIÓN - GALERADAS

Como se desprende del flujo, el menú lateral es accesible desde cualquier página de la app, bien pulsando el botón del menú o haciendo el gesto de arrastrar desde el lateral izquierdo hacia la derecha.

Interfaces y Usabilidad

La **aplicación** consta de una página inicial desde donde se accede al menú lateral y a realizar una predicción:



El **menú lateral** *consta de las siguientes secciones:*

EN PRODUCCIÓN - GALERADAS

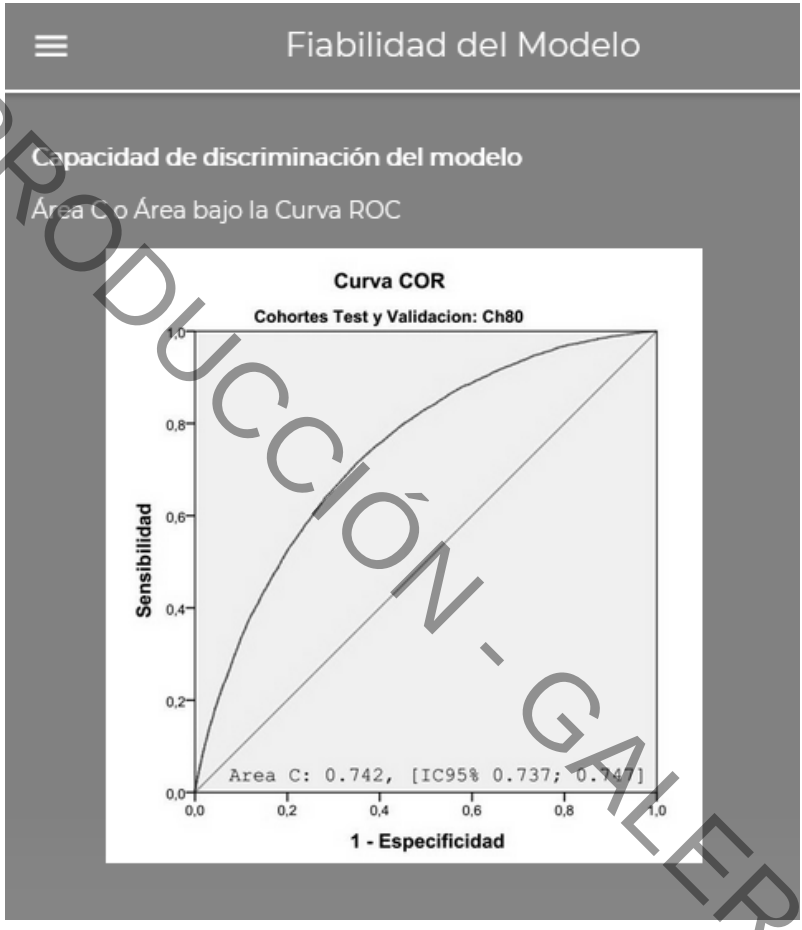


*La sección "***Variables Usadas***" detalla el significado de los datos que requiere el formulario:*



EN PRODUCCION - GALERADAS

La sección "**Fiabilidad**" muestra la capacidad de discriminación del modelo predictivo:



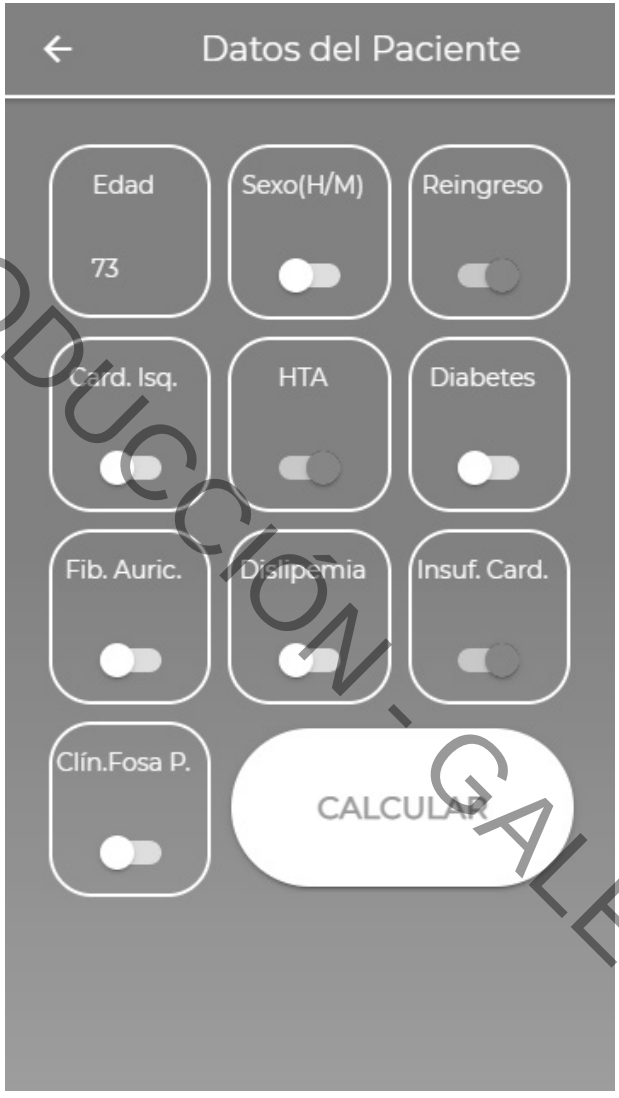
La sección "**Acerca de**" da detalles del proyecto y de sus autores:

EN PRODUCCIÓN - GALERADAS



Si accedemos a "**Nueva Predicción**" se nos mostrará un formulario intuitivo y rápido, a través del cual, en apenas unos segundos, el usuario introduce datos del paciente que se trasladarán a la función de estimación:

EN PRODUCCIÓN - GALERADAS



La app revisará si hemos introducido una edad válida (positiva y de 0 años hasta 120) y nos notificará mediante una alerta si hemos introducido una edad no válida:

EN PRODUCCIÓN - GALERADAS

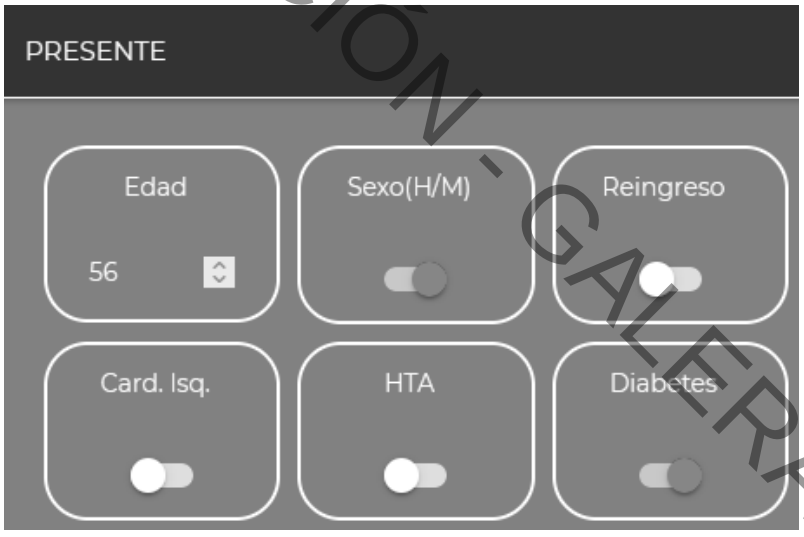


También nos informará mediante notificaciones *toast* cada vez que interactuemos sobre cada uno de los elementos de entrada con un deslizador interactivo tipo *toggle*. Este tipo de notificaciones aparecen en pantalla, permanecen unos segundos y desaparecen sin que se requiera acción del usuario. En este caso, al pulsar sobre el atributo sexo, vemos que el toggle cambia a color rojo y la notificación toast muestra "**Mujer**".

EN PRODUCCIÓN - GALERADAS

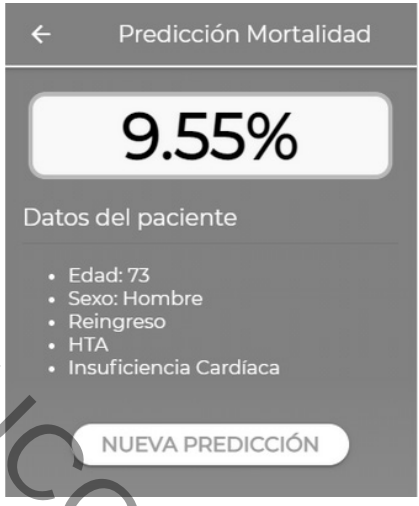


En el siguiente caso se refleja que el paciente tiene diabetes, por lo que se pulsa sobre el toggle "**Diabetes**", apareciendo un toast con el texto "**PRESENTE**".

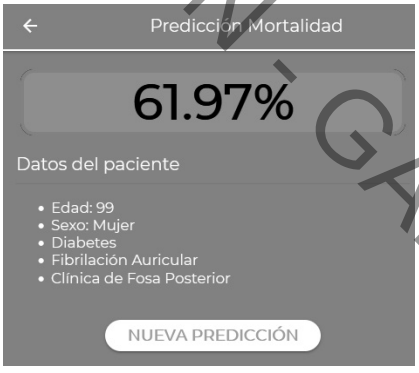


Cuando se haya terminado de completar el formulario, el resultado de la estimación se mostrará en pantalla, con un texto limpio y un código de colores que ayuda a una comprensión inmediata.

EN PRODUCCIÓN - GALERADAS



Esta predicción en forma de porcentaje aportará agilidad en la detección y facilitará la toma de decisiones.



Resultado

Se ha conseguido desarrollar una solución sencilla para un problema complejo. Una aplicación móvil/web, que permite el cálculo rápido y a pie de cama de un modelo de riesgo para la estimación de la probabilidad de fallecer en paciente con ictus

EN PRODUCCION - GALERADAS

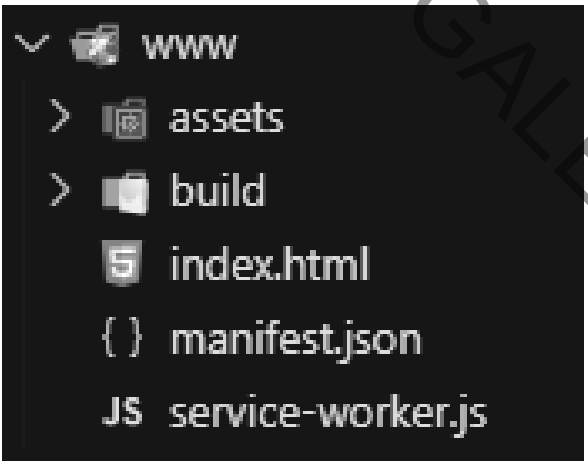
isquémico. Dicha aplicación es además muy versátil pudiéndose usar desde cualquier dispositivo con navegador (android, iOS, windows, linux, etc). Es preciso recoerdar y destacar **que se crea como soporte para la toma de decisiones, no como una sustitución del diagnóstico médico**.

Crear la aplicación web progresiva (PWA)

Unos meses después, a partir del código de la app móvil que se desarrolló para Android y con mínimas modificaciones, se consigue desarrollar una PWA en unas horas, con exactamente la misma funcionalidad y estética que la app móvil.

Aunque es cierto que las versiones actuales de Ionic generan una pwa con más facilidad, consideramos que puede ser útil mostrar aquí cómo se llevó a cabo en un proyecto de **Ionic 3**, ya que hay aplicaciones en esa versión que aún se mantienen en producción y a sus desarrolladores podría interesarle este conocimiento.

Una vez que en la consola de comandos se ejecuta *"ionic serve"* o *"ionic build --prod"* generamos dentro del proyecto la carpeta "www", donde se encuentran entre otras cosas los archivos service-worker.js, manifest.json e index.html.



EN PRODUCCIÓN - GALERADAS

Service Worker:

Un **service worker** es un script que permite *interceptar y controlar* las *peticiones de la red* y el *almacenamiento del cache del navegador web.* Esto permitió dotar a la versión web de características propias de una aplicación de escritorio o de móvil, como instalarse en el pc, generar un acceso directo en el escritorio y mantener funcionalidad sin acceso a internet.

www > JS service-worker.js > ...

```
'use strict';
importScripts('./build/sw-toolbox.js');

self.toolbox.options.cache = {
  name: 'ionic-cache'
};

// pre-cache our key assets
self.toolbox.precache(
  [
    './build/main.js',
    './build/vendor.js',
    './build/main.css',
    './build/polyfills.js',
    'index.html',
    'manifest.json'
  ]
);

// dynamically cache any other local assets
self.toolbox.router.any('/*', self.toolbox.fastest);

// for any other requests go to the network, cache,
// and then only use that cached resource if your user goes offline
self.toolbox.router.default = self.toolbox.networkFirst;

```

Manifest.json:

Desde este archivo se pueden referenciar los iconos en las distintas resoluciones del dispositivo, ya sea pc o móvil. Antes deberíamos haber generado los iconos en sus distintas resoluciones.

EN PRODUCCIÓN - GALERADAS

www > { } manifest.json > ...

```
{
  "name": "Calculadora Ictus",
  "short_name": "Calculadora Ictus",
  "start_url": "index.html",
  "display": "standalone",
  "icons": [
    {
      "src": "assets/imgs/logo.png",
      "sizes": "512x512",
      "type": "image/png"
    },
    {
      "src": "assets/imgs/icon-192x192.png",
      "sizes": "192x192",
      "type": "image/png"
    },
    {
      "src": "assets/imgs/icon-128x128.png",
      "sizes": "128x128",
      "type": "image/png"
    },
    {
      "src": "assets/imgs/icon-96x96.png",
      "sizes": "96x96",
      "type": "image/png"
    },
    {
      "src": "assets/imgs/icon-72x72.png",
      "sizes": "72x72",
      "type": "image/png"
    }
  ],
  "background_color": "#019288",
  "theme_color": "#019288"
}
```

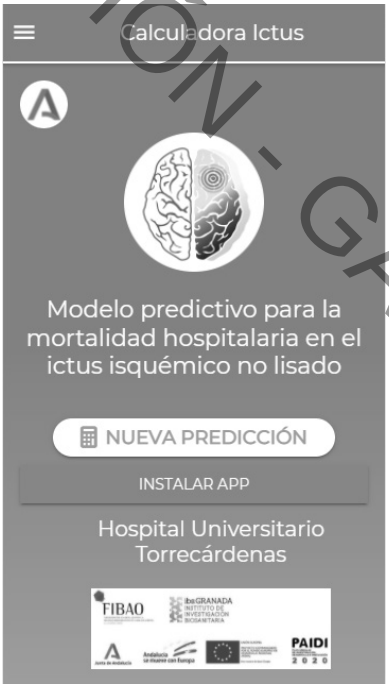
EN PRODUCCIÓN - GALERADAS

Index.html:

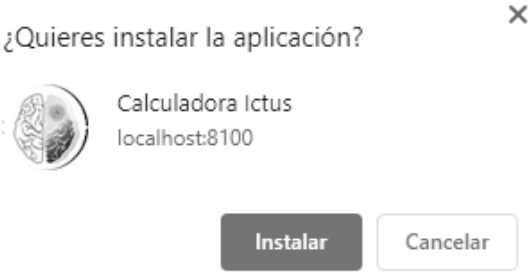
Adicionalmente se tuvieron que añadir estas líneas en el index.html

```
www > index.html > html > head > link
22
23    <script>
24      if ('serviceWorker' in navigator) {
25        navigator.serviceWorker.register('service-worker.js')
26          .then(() => console.log('service worker installed'))
27          .catch(err => console.error('Error', err));
28      }
29    </script>
30
31    <noscript>
32      Esta aplicación necesita Javascript para funcionar. Por favor, activa Javascript en tu navegador
33    </noscript>
```

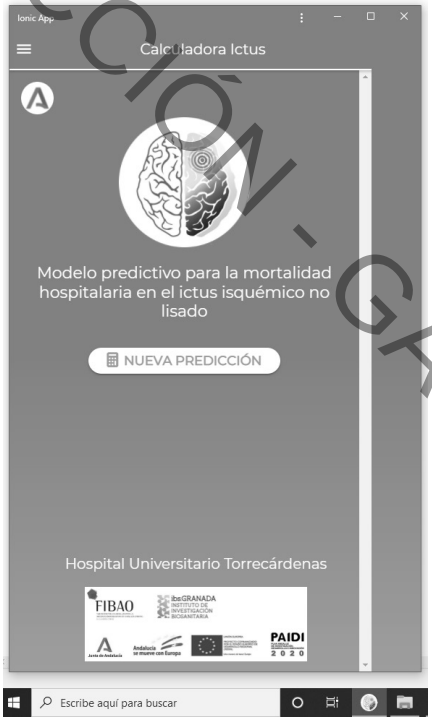
Todo esto permite que al ejecutar la aplicación en un navegador web (ya sea móvil o pc) detecte si la misma está instalada en el dispositivo. De no estarlo, mostrará un botón, rojo en este caso, desde donde podremos instalarla.



EN PRODUCCIÓN - GALERADAS



Una vez instalada, se visualiza exactamente igual que una aplicación nativa del dispositivo en el que se instale. En este caso, se ha instalado en un pc Windows:



Url de la app web: https://calculadora-ictus.web.app/

EN PRODUCCIÓN - GALERADAS

Publicar la PWA

Para alojar la aplicación en un servidor, se optó por Firebase, una plataforma en la nube de Google que ofrece una amplia gama de herramientas y servicios para el desarrollo de aplicaciones móviles y web, entre ellos alojamiento web (Firebase Hosting) y que se ajustaba perfectamente a este proyecto por su fácil integración con Ionic, su gran rendimiento y su rapidez en el desarrollo.

Los pasos seguidos fueron los siguientes (en cursiva comandos de consola):

- Crear un proyecto en Firebase Hosting.
- Instalar el paquete firebase-tools: *npm install -g firebase-tools*
- Iniciar sesión en las herramientas de Firebase: *firebase login*
- Hacer un build del proyecto: *ionic build --prod*
- Desde la raíz del proyecto: *firebase init*
- Nos hará una serie de preguntas donde debemos decirle que queremos usar como directorio público la carpeta "www", que queremos que lo configure como SPA y que no queremos reescribir el index.html
- Finalmente, realizamos el despliegue en el servidor, que nos devolverá la url de nuestra pwa alojada en Firebase: *firebase deploy*

Configurar Proxy en Visual Studio Code y npm

En ocasiones, tuvimos que desarrollar Ionic en un pc conectado a una red que estaba detrás de un proxy corporativo. Esto presenta el inconveniente de que si npm y Visual Studio no están bien configurados, no podrán funcionar.

Npm es el gestor de paquetes que se usa con Node.js y por tanto con Ionic para desarrollo. Para crear un nuevo proyecto,

EN PRODUCCIÓN - GALERADAS

descargar plugins o incluso para ejecutarse, se conecta a internet, usualmente a repositorios ubicados en GitHub. Si no se configura el proxy en ambos programas, no podremos acceder a ninguno de esos repositorios.

Los pasos que se siguieron con la versión 3 de Ionic fueron los siguientes:

- Configurar proxy en Visual Studio Code:

En el fichero settings.json, que está ubicado en la ruta C:\Users*Nombre_de_usuario*\AppData\Roaming\Code\User

Deberemos añadir las siguientes líneas:

{
"http.proxy": "http://X.X.X.X:puerto",
"https.proxy": "http://X.X.X.X:puerto",
"http.proxyStrictSSL": false,
}

- Configurar proxy en npm tecleando las siguientes líneas en la consola de comandos:

npm set proxy=http://X.X.X.X:puerto
npm set https-proxy=http://X.X.X.X:puerto

- Instalar el plugin del proxy para Ionic, tecleando la siguiente línea en la consola de comandos:

npm install -g @ionic/cli-plugin-proxy (para mí fue necesario tener instalado el Ionic CLI 3.0 o superior)

- Crear variables de sistema en las variables de entorno de Windows:

HTTP_PROXY=http://X.X.X.X:puerto
HTTPS_PROXY=https://X.X.X.X:puerto

EN PRODUCCIÓN - GALERADAS

• A la hora de generar el apk, deberemos asegurarnos de tener, y si no lo tenemos crearlo, el archivo gradle.properties en C:/Usuarios/carpeta_usuario/.gradle , con cuidado de no meter espacios en blanco en el archivo:

systemProp.http.proxyHost=X.X.X.X
systemProp.http.proxyPort=puerto
systemProp.https.proxyHost=X.X.X.X
systemProp.https.proxyPort=puerto

EN PRODUCCIÓN - GALERADAS

3.- BIBLIOGRAFÍA

1. Martín Fernández, A. (Tesis de maestría/doctorado). Aplicaciones móviles relacionadas con la salud. Un estudio sobre las aplicaciones con funcionalidad para el recordatorio de la toma de medicamentos. Universidad de Zaragoza.
2. Chavira-García, J., Arredondo-López, A. A. Aplicaciones móviles como herramientas en los servicios de salud.
3. Santamaría-Puerto, G., & Hernández-Rincón, E. Aplicaciones Médicas Móviles: definiciones, beneficios y riesgos.
4. Morony, J. Building Mobile Apps With Ionic & Angular.

EN PRODUCCIÓN - GALERADAS

7. EL PAPEL DE LA CIENCIA DE DATOS EN EL MODELADO PREDICTIVO: APLICACIÓN AL ICTUS ISQUÉMICO Y MORTALIDAD.

AUTORES: LUIS TÉLLEZ RAMÍREZ. JUAN MANUEL GARCÍA TORRECILLAS. FERNANDO RECHE LORITE.

EN PRODUCCIÓN GALERADAS

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

RESUMEN

La Ciencia de Datos es un campo interdisciplinario que combina las matemáticas (aprendizaje automatizado), la informática (requiere cierto nivel de programación) y conocimiento especializado acerca de la temática donde se quiere aplicar dicha rama. Técnicas como el Machine Learning o el Deep Learning (ramas de la Ciencia de Datos) han revolucionado el paradigma actual de la investigación y están suponiendo una auténtica transformación en diversos ámbitos. En este capítulo se pretende contextualizar y explicar algunas de estas técnicas, así como comentar el "estado del arte" (lo más puntero) relacionado con ello. Se pretende hacer una introducción a la Ciencia de Datos, a la vez que se muestra contenido del **Trabajo Fin de Máster "Aplicación de Ciencia de Datos en la Optimización de modelos predictivos para la estimación de mortalidad en Ictus Isquémico"** en el último capítulo de esta obra (Luis Téllez Ramírez 2021).

1. INTRODUCCIÓN

Vivimos en la *"era de la información"*, tanto es así que se ha hecho popular la frase: *"Data is the new oil"* (los datos son el nuevo petróleo). ¿Esto es realmente así? Realmente los datos en sí mismos no tienen ningún valor intrínseco, más allá de su mera recopilación, lo que resulta un problema complejo es la extracción de la información contenida en esos datos, de esta tarea se encarga la Ingeniería de Datos. Es dicha información, que se extrae de estos datos, lo que los hace valiosos, lo que ayuda a tomar decisiones.

Para ello, como se ha destacado, es importante cuidar la parte del almacenamiento de datos. Al igual que un chef no puede hacer un buen guiso si la materia prima no es buena, los científicos de datos no podrán crear un buen producto si los datos no gozan de un mínimo de calidad. Por tanto, es importante tener esto siempre en cuenta, y se recordará la importancia de ello más adelante.

Se dice que la ciencia de datos es un proceso que va de principio a fin y que podemos describir en los siguientes pasos:

1. Se parte de unos datos recopilados (**raw data**).
2. Se realiza un **preprocesamiento**, que se corresponde con la limpieza de datos.
3. Creamos **vistas** o **datos estructurados** para que se pueda realizar correctamente el proceso de analítica.
4. Se construyen **modelos** que nos sirvan para clasificar elementos (si estamos ante un problema de clasificación) o

EN PRODUCCIÓN - GALERADAS

para predecir (en caso de estar ante un problema de regresión).
5. Finalmente se procede al proceso de **encapsulación** del desarrollo y puesta en producción, creando el **producto**.

Se dice también que la analítica de datos es un proceso:

- **Orientado a producto/cliente**, puesto que se va a validar en escenarios reales.
- **Iterativo**, experimental y orgánico, con mejoras y adaptaciones continuas.
- **Data Driven**, puesto que está basado en el estado de los datos.

Además, existen distintas metodologías de aproximación a un problema con las características previamente definidas, como son Berkeley, CRISP-DM, Microsoft´s TDSP. Ni que decir tiene que el proceso por el cual se mueven estas metodologías debe estar basado principalmente en el Método Científico (Preguntas - Investigación - Hipótesis - Experimentos - Conclusiones ...). Se dice también que el 80% de lo que se considera como Análisis de Datos no lo es realmente, ya que la mayor parte del tiempo se pasa transformando y trasladando los datos.

Una de las partes más importantes de la Ciencia de Datos es el modelado, y para ello, se sustenta en el Machine Learning (ML) y Deep Learning (DL), pero, ¿qué es esto realmente? ¿en qué se diferencia del modelado estadístico? En primer lugar, hay que entender que tanto el Machine Learning como Deep Learning suponen un cambio de paradigma en la programación funcional tal y como se venía trabajando hasta la actualidad. En la informática clásica la estrategia es:

EN PRODUCCIÓN - GALERADAS

- Input Data ➜ Reglas de programación ➜ Output Data

Es decir, se parte de una serie de datos que, dada una programación sobre un conjunto de reglas, transforman dichos datos para producir una salida. Es el programador el que forma parte mayormente de los resultados obtenidos.

Tanto en el Machine Learning como en el Deep Learning se cambia esta perspectiva.

- Input Data & Output Data ➜ Machine Learning Model ➜ Reglas de programación
- Input Data & Output Data ➜ Deep Learning Model ➜ Reglas de programación

Es decir, el modelo es capaz de "aprender" y tratar de inferir lo que está ocurriendo en los datos. En un modelado estadístico se busca la interpretabilidad de parámetros, explicación de resultados y la obtención de la mejor aproximación del modelo generativo. Mientras que el Machine Learning está más centrado en los resultados, tanto que suele producir modelos "caja negra" (difíciles o imposibles de interpretar) en los que se busca la automatización e integración.

¿Cuál es la diferencia entre el Machine Learning y el Deep Learning? Ambas técnicas forman parte del ámbito Inteligencia Artificial, ya que son técnicas que permiten a un ordenador imitar el comportamiento humano. El Machine Learning, lo hace mediante un conjunto de algoritmos entrenados, mientras que el Deep Learning es un tipo de Machine Learning inspirado en la estructura del cerebro. Requiere menor intervención que el Machine Learning, y cuantos más datos tengamos, es muy posible que los resultados serán bastante mejores, por eso su popularidad en la era del Big Data.

EN PRODUCCIÓN-GALERADAS

1.1. INTRODUCCIÓN AL PROBLEMA

Bien, el problema que nos servirá de ejemplo para presentar lo descrito en la sección anterior tiene que ver con la estimación del riesgo de mortalidad tras padecer un ictus isquémico. Este problema, desde una perspectiva clínica, se describe con profundidad a lo largo de este libro y por tanto no vamos a hacer demasiado hincapié en dichas cuestiones. Simplemente destacaremos la justificación, propósito e hipótesis de dicho trabajo.

1.1.1 Fundamentos

El ictus isquémico supone la segunda causa de mortalidad en nuestro país en la población general y la primera causa de mortalidad en la mujer (Ministerio de Sanidad y Política Social 2009). A nivel mundial el ictus es la segunda causa de mortalidad y la tercera más común en los países industrializados. Por tanto, nos enfrentamos a un auténtico problema de salud pública. La mortalidad hospitalaria según fuentes españolas procedentes de registros de carácter clínico se sitúa en torno al 12,9% (Arboix A & García-Eroles L & Comes E & Oliveres M & Targa C & Balcells M 2008).

Un porcentaje elevadísimo de la misma son ictus isquémicos no lisables, esto es, no subsidiarios de tratamiento fibrinolítico (un tipo de fármaco) por no cumplir los estrictos criterios para la aplicación de este tratamiento (García-Torrecillas et al, 2019). Hasta la fecha se han detectado una serie de factores de riesgo tanto para el desarrollo de un evento isquémico como para estimar la probabilidad de fallecer o presentar secuelas (Arboix A & García-Eroles L & Comes E & Oliveres M & Targa C & Balcells M 2008). Los trabajos de Smith et al (Smith EE & Shobha N & Dai D & Olson DM & Reeves MJ & Saver JL 2013) permitieron la obtención de modelos predictivos para la mortalidad hospitalaria, tanto por ictus isquémico como hemorrágico.

EN PRODUCCIÓN. GALERADAS

1.1.2 Hipótesis del trabajo

Se ha demostrado que la existencia de un plan integral de actuaciones que desde la llegada del paciente al hospital maximice y optimice su atención repercute de manera beneficiosa en aquellos que han padecido un ictus agudo, aumentando las probabilidades de recuperación (CS, 2003).

El conocimiento de las variables susceptibles de ser recogidas a pie de cama al ingreso o en los primeros días de estancia del paciente y que, integradas en un modelo predictivo nos permitan estimar la probabilidad de fallecer durante el ingreso, tiene importantes repercusiones en la atención al paciente, puesto que:

- Permitiría modular la intensidad y velocidad del esfuerzo diagnóstico terapéutico en aquellos pacientes con scores de riesgo de puntaje más elevado.
- Facilitaría la integración en circuitos asistenciales específicos de los pacientes de alto riesgo.
- Sería un elemento clave para la optimización del coste-eficiencia en la atención a pacientes con ictus no lisable.

1.1.3 Inconvenientes

Ante esta situación y, dadas las fuentes de información disponibles en la actualidad, nos encontramos con las siguientes problemáticas a tener en cuenta:

- Los datos que manejamos están extraídos de la base de datos clínico-administrativa (Conunto Mínimo Básico de Datos o CMBD) que no goza de una calidad extraordinaria. Veremos más adelante cómo tratar algunos problemas derivados de ello.
- La base de datos, como es frecuente en medicina, es una base de datos desbalanceada, es decir, el número de perso-

EN PRODUCCIÓN - GALERADAS

nas que fallecen va a ser mucho menor que el número de personas que sobreviven. Esto es un éxito rotundo desde un punto de vista clínico-médico, pero supone una serie de problemas en la estimación del riesgo a partir de los datos recopilados.

1.2. PREPROCESAMIENTO DE LOS DATOS

Nuestro punto de partida es la base de datos ya filtrada por aquellos pacientes que han padecido un ictus isquémico en alguno de los diagnósticos. Dicha base de datos se nos ha suministrado en formato spss (.sav). Una vez que se eliminan aquellas variables puramente administrativas, nos quedan 36 variables. En este momento consideraremos dos escenarios de trabajo:

- **Escenario A**: Se trabajarán únicamente variables seleccionadas por criterio experto. Variables que son *fácilmente recopilables* por los médicos con el objetivo de conseguir una estimación rápida.
- **Escenario B**: Se trabajará con todas las variables que están a disposición, usando algoritmos de selección de variables exclusivos de Ciencia de Datos. El objetivo en este caso es mejorar los resultados obtenidos en el Escenario A.

Nota: El método de entrenamiento de los modelos pasa por la división del conjunto de datos en dos partes, una de ellas para entrenar el modelo y el resto de los datos para validarlo. Nuestro modelo no utilizará los datos de validación para construirse ya que éstos están destinados a estudiar la calidad predictiva del modelo estimado.

Previamente a la división de la base de datos en las dos partes anteriormente mencionadas, se *limpia* eliminando las observaciones que no tengan un mínimo de calidad. De esta forma,

EN PRODUCCIÓN - GALERADAS

podemos hacer la misma división para ambos escenarios. En consecuencia, entrenaremos y validaremos siempre sobre las mismas observaciones, pero con diferente número de variables.

En las variables, debemos distinguir el tipo para proceder a su tratamiento.

a. **Variables Cuantitativas:** En este caso, únicamente tenemos la variable EDAD, que podemos considerarla continua. Una posible opción sería discretizarla en estratos de edad, y otra, escalarla, pero esto último se haría únicamente si el modelo a entrenar es sensible a variables sin escalar.
b. **Variables Booleanas:** 9 variables booleanas, que están codificadas como "Sí" (1) y "No" (0).
c. **Variables Cualitativas:** En este caso no hay, pero sería interesante distinguir entre cualitativas ordinales (donde importa el orden) y categóricas.

Puesto que son datos médicos y vamos a hacer una estimación del riesgo de fallecimiento, no vamos a realizar imputación de datos faltantes. Por tanto, eliminaremos cualquier observación que tenga al menos un valor faltante en alguna de las 36 variables iniciales.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

2. ANÁLISIS EXPLORATORIO DE DATOS

"La principal diferencia entre un Data Scientist Junior y un Senior radica en que, mientras uno prueba distintos modelos para cercar el problema, el otro comprueba qué está ocurriendo en los datos."

Esta frase nos muestra un principio muy interesante, y es que, por mucho que se quiera, usando técnicas muy avanzadas de estadística y Ciencia de Datos, no mejorarán los resultados si hay una limitación intrínseca en ellos, es decir, si los datos no son suficientemente informativos, los resultados no serán buenos, por muy sofisticada que sea la técnica que se utilice. Es cierto que dependiendo de los datos unas técnicas funcionan mejor que otras, pero si en un momento no vemos mejora en los modelos construidos, es muy posible que sea por una limitación en los datos.

En la gráfica que se presenta en la Figura 1 podemos observar cómo la Edad va a ser una variable fundamental, ya que existe una diferencia significativa de Edad entre los dos grupos. Esto último podemos comprobarlo mediante muestreo por bootstrapping[1].

1 El Bootstrap es una técnica de computación intensiva propuesta por Bradely Efron a principios de los años 80 del siglo pasado. Esencialmente consiste replicar la muestra en el ordenador muchas veces (miles o millones) y aplicar la técnica concreta en muestras extraídas de esa replicación.

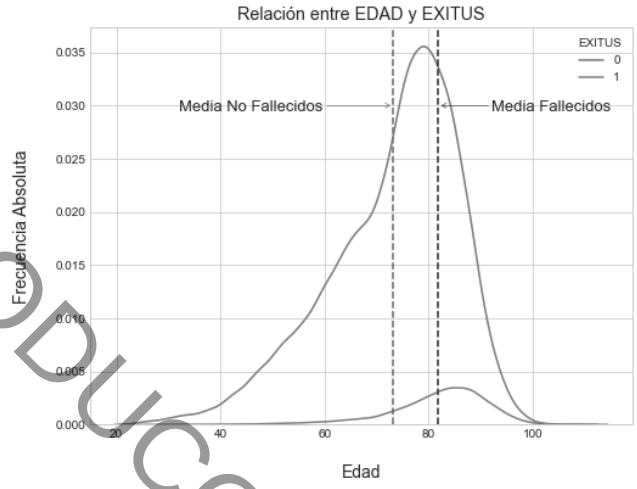


Figura. 1 Función de densidad de edad por grupos.

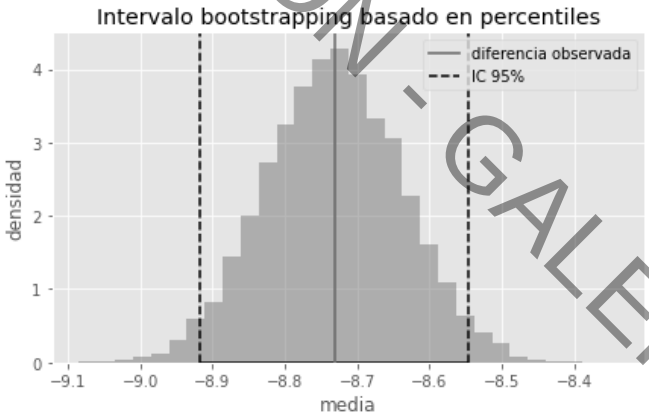


Figura.2 Intervalo Bootstrapping para diferencia de Edad.

En la Figura 3 se evidencia la relación directa existente entre la edad y el riesgo de fallecer por Ictus isquémico.

EN PRODUCCIÓN - GALERADAS

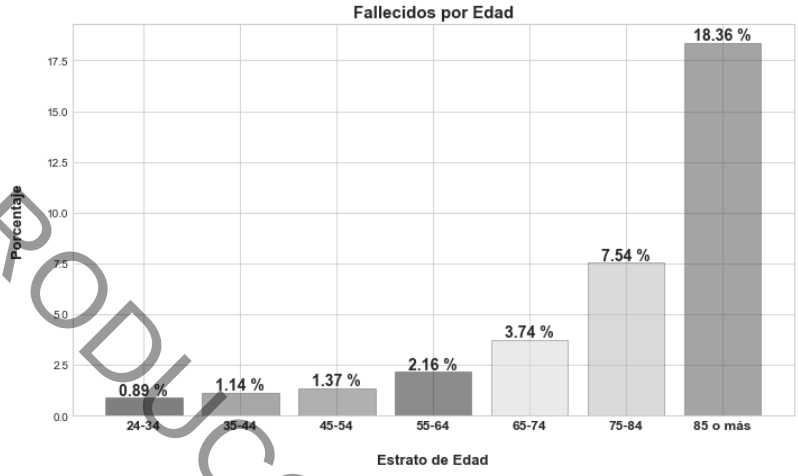


Figura.3 Fallecidos por estrato de Edad.

Véase (Figura 4) qué ocurre con el resto de predictores.

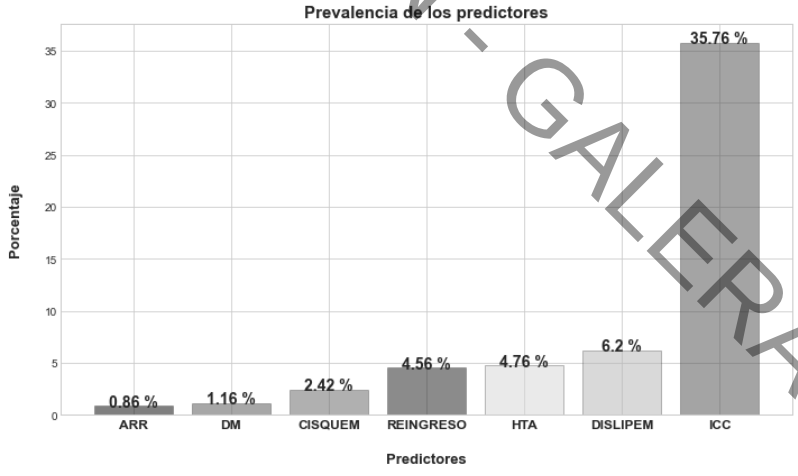


Figura.4 Prevalencia de los Predictores.

EN PRODUCCIÓN - GALERADAS

Si observamos el grupo de fallecidos, podemos ver (Figura 5) el comportamiento de la prevalencia de los predictores:

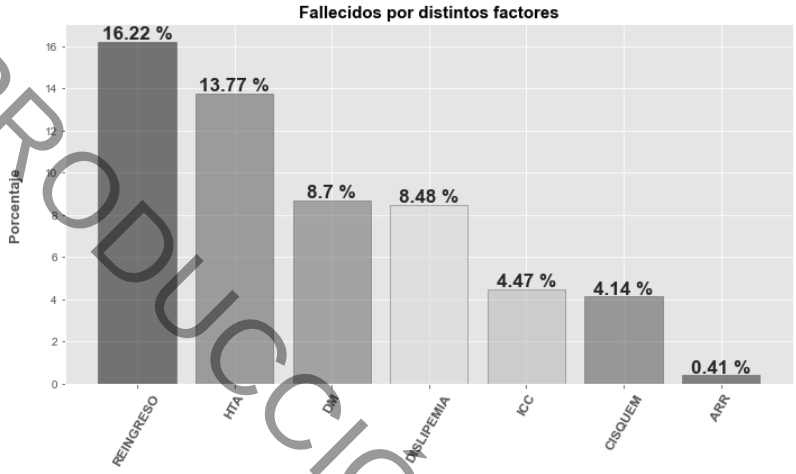


Figura.5 Prevalencia de los predictores en los fallecidos.

En cuanto al sexo, observamos que hay un 53,41% de hombres frente a un 46,59% mujeres. Mientras que en fallecimientos tenemos un 5,05% de hombres y un 8,72% de mujeres. Por tanto, esta variable va a ser otro factor predisponente.

2.1 DISCUSIÓN DE LA MÉTRICA A UTILIZAR

Podemos decir que este estudio lo podemos considerar como un problema de clasificación, puesto que, en última instancia, buscamos saber si el modelo predice el fallecimiento o no del paciente. Aunque dicha clasificación, no es otra cosa que una máscara que se pasa sobre la predicción (también conocido como umbral). Si supone un modelo f, y sea $X = (X_1, \ldots, X_n)$ una observación y $f(X)$ una predicción del modelo. Entonces:

EN PRODUCCION - GALERADAS

$$f(X) = p \implies h_p(X) = \{1\ si\ p > 0.5\ 0\ si\ p \leq 0.5$$

Una **distancia** o **métrica** es una aplicación (en el sentido matemático) que cumple unas determinadas propiedades y que nos ayudará a ver cuánto nos estamos equivocando en nuestras predicciones. Por tanto, a la hora de entrenar el modelo, buscaremos minimizar o maximizar (depende de la métrica seleccionada) el valor de dicha métrica.

Puesto que nos encontramos ante un problema desbalanceado, es decir, aproximadamente el 93% de los datos corresponden a personas que no han fallecido, debemos tener cierto cuidado a la hora de seleccionar la métrica. Por ejemplo, no tiene sentido usar la métrica *Accuracy*[2], ya que, si un modelo clasifica todas las observaciones como **No pacientes de riesgo**, estará obteniendo un 93% de acierto (que es bastante), y, sin embargo, estará obviando el riesgo del otro 7%.

Además, como el lector podrá comprender, no supone lo mismo decirle a un paciente con ictus que no está grave cuando sí lo está, a decirle que sí está grave cuando no lo está. Es decir, se deben evitar los falsos negativos sin que ello suponga una disminución drástica en la precisión (que no es otra cosa que la habilidad de clasificar un verdadero positivo como positivo).

Por tanto, la métrica seleccionada es la *F1-Score* (una media armónica entre la precisión[3] y recall[4]).

Sería también apropiado el uso de la variante F_β.

2 Porcentaje de aciertos en el modelo predictivo.

3 Razón entre los aciertos en la predicción de los positivos (fallecidos) de entre todos los fallecidos predichos.

4 Razón entre los aciertos en la predicción de los positivos (fallecidos) de entre los que realmente han fallecido.

EN PRODUCCIÓN - GALERADAS

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$F_\beta = (1 + \beta^2) \frac{Precision \cdot Recall}{(\beta^2 Precision) + Recall}$$

Otras medidas interesantes son AUROC, que mide el área bajo la curva ROC, y en particular la habilidad del modelo para distinguir entre clases. AUPRC área bajo la curva precision-recall, o Balanced-Accuracy.

2.2 DIVISIÓN EN TRAIN-TEST

El método de entrenamiento y evaluación será la validación cruzada + Holdout, es decir, lo primero que realizaremos será una división entre train y test (85%-15%), pero hemos de resaltar algo importante, lo haremos de manera estratificada. Es decir, se aprenderán los parámetros e hiperparámetros de nuestro modelo en la muestra train mediante una validación cruzada exhaustiva, y se validarán los resultados sobre la muestra test. El modelo no debe "ver" nunca los datos test, ya que se incurre en una falsificación (produciendo fuga de datos).

Al entrenar el modelo, este ajusta una serie de parámetros, sin embargo, hay algunos que permanecen por defecto y hay que modificarlos manualmente. Estos son conocidos como "hiperparámetros", al hacerlo mediante una validación cruzada exhaustiva, se probarán todas las combinaciones posibles de los parámetros declarados y devolverá la mejor respecto de la métrica indicada.

EN PRODUCCIÓN - GALERADAS

3. MODELOS A UTILIZAR

Este es un punto importante para tratar, ya que la selección del modelo depende de diversos factores. Algunos de estos son:

1. El tipo de datos que estemos utilizando para entrenar el modelo.
2. La interpretabilidad del modelo.
3. La complejidad del problema.

Por ejemplo, si nuestros datos son imágenes, lo mejor será una red neuronal, mientras que, para datos tabulados, árboles de decisión, random forest o modelos similares, son métodos que producen mejores resultados en la mayoría de los casos, además, de manera más sencilla y con una información más rica.

Puesto que este capítulo es una consecuencia de un trabajo ya previamente publicado, en que se ha utilizado la regresión logística para construir el modelo predictivo, proponemos hacer uso de este mismo modelo y tratar de mejorarlo usando la metodología de random forest.

Random forest es una metodología que forma parte de un conjunto de técnicas denominadas métodos de **ensembles**. Esta metodología consiste en la construcción de un modelo formado a partir de un conjunto de modelos "base" teniendo estos (en la mayoría de las ocasiones), parámetros diferentes.

Para clasificar un nuevo ejemplo, se utilizarán en todos los clasificadores base y se **agregarán** las predicciones individuales de cara a obtener una predicción resultado, como una especie de democracia de modelos. Hay distintas maneras de agregar los resultados, como el voto por mayoría, voto ponderado o un promediado de modelos.

En pocas palabras, la metodología de random forest es un conjunto de árboles de decisión estimados mediante *bagging*

EN PRODUCCIÓN - GALERADAS

(Bootstrap aggregation). En cada muestreo se utiliza una selección aleatoria de variables. Se introduce aleatoriedad tanto en los datos de entrada (se estiman los parámetros a partir de muestras bootstrap), como en la construcción propia del clasificador. No hay proceso de poda[5].

Este proceso requiere un apunte previo. Como se ha comentado a lo largo de este trabajo, el hecho de que la muestra sea desbalanceada supone algunos problemas a la hora de estimar un modelo sobre ellos. Usualmente, hay tres maneras de abordar un problema desbalanceado:

a. **Ignorar el problema:** conlleva entrenar el modelo igualmente, y normalmente acaba provocando la obtención de clasificadores triviales o cerca de la trivialidad (consideran todos los casos pertenecientes a la clase mayoritaria).
b. **Submuestrear la clase mayoritaria:** Suele ser una buena técnica implementar un submuestreo aleatorio para igualar las clases y una posterior aplicación del modelo.
c. **Sobremuestrear la clase minoritaria:** Crear nuevas instancias "duplicando" algunas de la clase minoritaria (o usando algoritmos más complejos como "s*mote*"). Puede conllevar problemas de sobreajuste si no se hacen las cosas correctamente.

En nuestro caso, haremos uso de las dos primeras técnicas mencionadas, aunque con una variante de la primera, implementando una penalización por clasificación incorrecta de un elemento de la clase minoritaria proporcional a la proporción de casos entre las clases.

5 Metodología que se aplica a los árboles de decisión para evitar sobreaprendizajes.

EN PRODUCCIÓN - GALERADAS

3. RESULTADOS

En primer lugar, hemos de resaltar que un modelo construido utilizando la metodología de random forest no es tan fácil de interpretar como podría ser una regresión logística, es decir, tiene tantos parámetros que muchas veces no es posible lograr entender al 100% cómo funciona el modelo internamente. Sin embargo, hay ciertas técnicas por las que se puede intuir cómo toma dichas decisiones.

Random forest implementa una técnica que asigna importancia a los predictores, lo que nos puede aportar una primera idea del mecanismo de predicción del modelo.

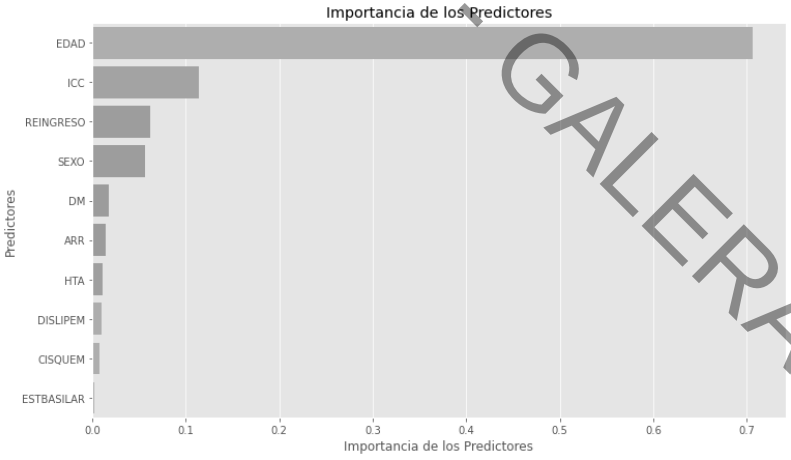


Figura.6 Importancia de los Predictores en Random Forest.

EN PRODUCCIÓN - GALERADAS

Véase cómo se comportan los modelos en el escenario A.

	Accuracy	Balanced accuracy	F1-Score	AUROC	Precision	Recall
Balanced Logistic Regression	0.633756	0.671532	0.208851	0.732837	0.122282	0.715214
Random Undersample + Logistic Regression	0.637426	0.671994	0.209760	0.732947	0.123003	0.711966
Balanced random forest	0.651886	0.673818	0.213537	0.733811	0.126013	0.699178
Under-sampling + Random forest	0.639648	0.661089	0.204715	0.712233	0.120337	0.685882

Figura.7 Métrica de los modelos.

Se observa cómo el modelo que mejores resultados da es Random Forest Balanceado. Se aprecia una limitación en cuanto a la métrica AUROC.

Ambos modelos, siendo uno bastante más complejo, no tiene mayor capacidad de discriminación, como se apuntaba en (2), si encuentra una limitación en los datos intrínseca al CMBD.

EN PRODUCCIÓN - GALERADAS

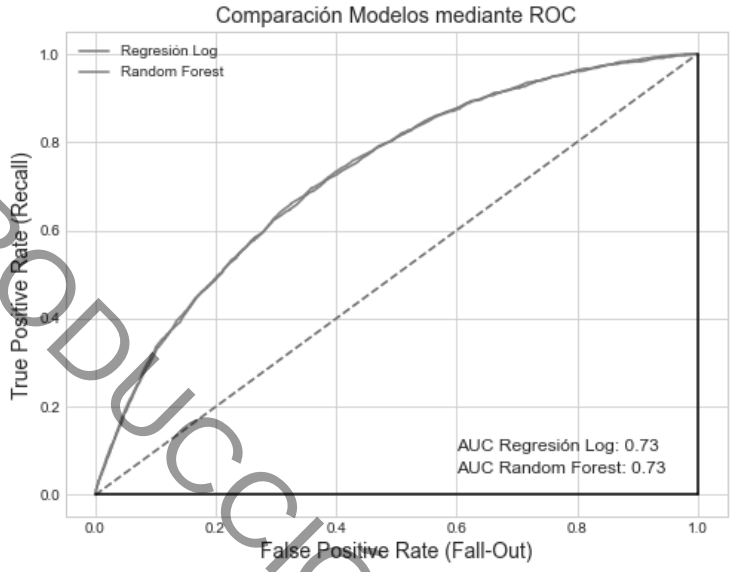


Figura.8 Comparación ROC de los modelos.

3. SHAP VALUES (SHAPLEY ADDITIVE EXPLANATIONS)

Lo que hace **SHAP** es cuantificar la contribución que cada característica aporta a la predicción realizada por el modelo. Véase una predicción interpretada por el algoritmo SHAP. En rojo vemos las características que contribuyen más al aumento de probabilidad (riesgo), y en azul, las que contribuyen a la disminución.

EN PRODUCCION - GALERADAS

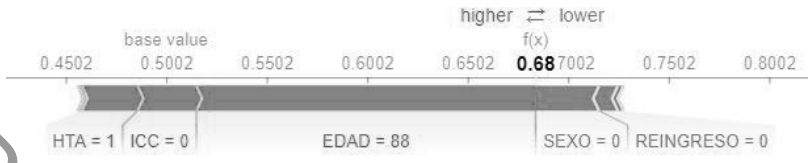


Figura.9 SHAP values.

El modelo ha sabido captar el riesgo asociado a las observaciones estimadas. Sin embargo, si tomamos como umbral 0.5, aquí se habría predicho que ambos casos fallecen, y, sin embargo, se sabe a priori, que estos pacientes no fallecieron. ¿Eso significa que nuestro modelo no funciona? Es una pregunta interesante, y en parte tiene que ver con que la medicina funcione y que estas personas que presentan factores de riesgo (que han sido detectados) hayan sobrevivido.

Podría preguntarse qué ocurriría si elimina algunos de los predictores que presentan menos relevancia para el modelo, y vea que no obtiene un cambio significativo en la evaluación de las métricas.

EN PRODUCCIÓN - GALERADAS

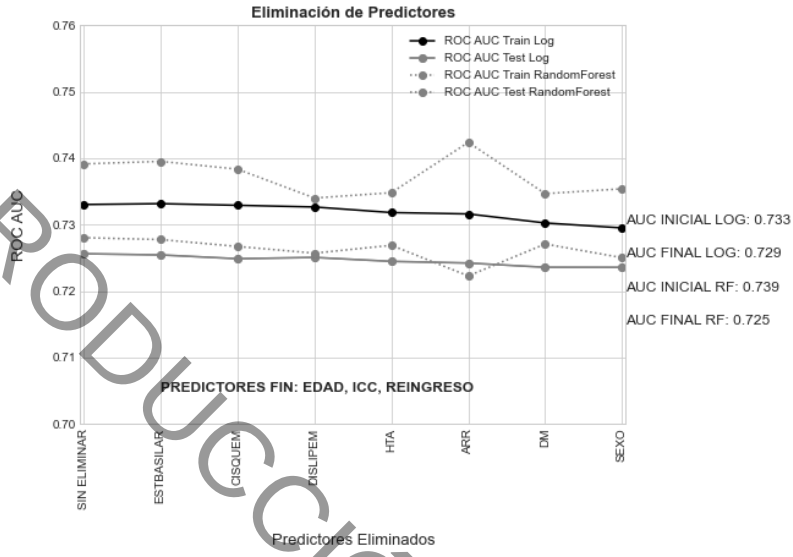


Figura.10 Eliminación recursiva de predictores.

Como se puede apreciar, eliminando algunos de los factores no varía mucho la métrica AUC, habría que pararse detenidamente a estudiar qué ocurre con el resto. Todo esto se debe a que los factores SEXO, REINGRESO y EDAD, se llevan la mayor importancia del modelo.

Respecto al **escenario B,** los resultados mejoran algo, que en parte era esperable debido a que se utilizan datos obtenidos a posteriori. Es decir, el objetivo es que cuando llegue un paciente del cual no se tienen demasiados datos, se pueda estimar un riesgo de mortalidad. Y no que una vez que se tengan bastantes datos del paciente (como el número de días que lleva ingresado) se realice la estimación.

EN PRODUCCIÓN - GALERADAS

	Accuracy	Bal-Acc	F1-Score	AUROC	Precis.	Recall
Balanced Logistic Regression	0.680313	0.703465	0.235936	0.770848	0.140699	0.730236
Random Undersample + Logistic Regression	0.683414	0.703810	0.236999	0.770645	0.141567	0.727394
Balanced random forest	0.688579	0.716228	0.245163	0.788986	0.146609	0.748200
Under-sampling + Random forest	0.671052	0.679860	0.220933	0.737112	0.131527	0.690045

En la siguiente imagen podemos apreciar, cómo el modelo le da una gran importancia a la variable estancia y también a la Edad.

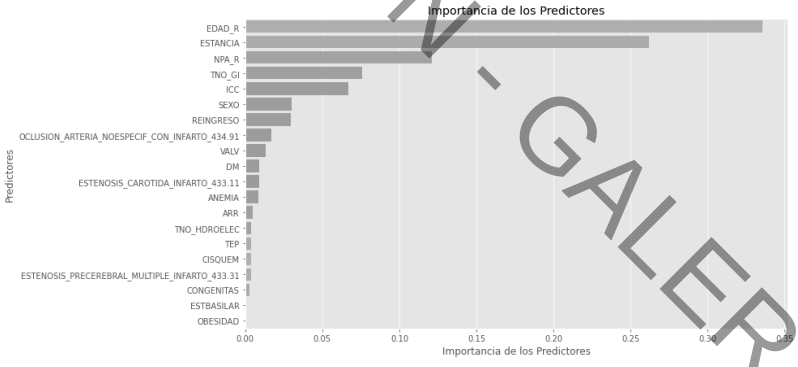


Figura.11 Importancia Predictores Escenario B.

De nuevo, podríamos plantearnos la eliminación de algunos predictores que no tengan demasiada importancia, y vemos cómo va afectando al modelo (óptimo con 9 predictores):

EN PRODUCCION - GALERADAS

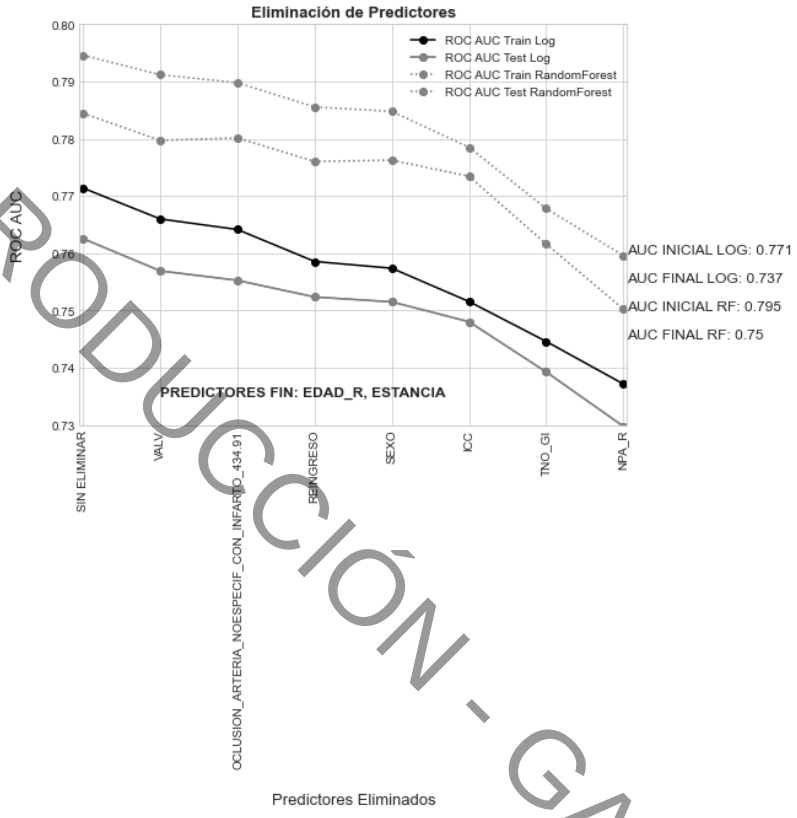


Figura.12 Eliminación de predictores escenario B.

En esta imagen puede apreciarse cómo el random forest mejora los resultados de la logística, esto se debe a una mayor complejidad y capacidad de distinción por parte de los datos, proporcionado en su mayoría por la variable "estancia".

EN PRODUCCIÓN - GALERADAS

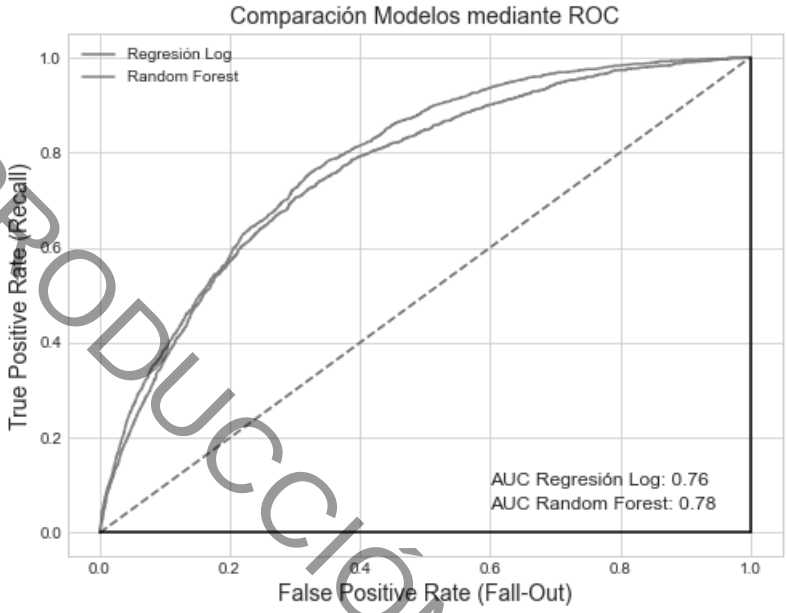


Figura.13 Comparación modelos Escenario B

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

4. DISCUSIÓN

Como era de esperar, se ha conseguido mejorar los resultados del **escenario A**, usando únicamente criterios de Ciencia de Datos. ¿Realmente es necesario aplicar Machine Learning? Lo cierto es que debemos examinar cuidadosamente los problemas en los que queramos aplicar dicha tecnología, ya que como comentaba Elon Musk en una de sus recientes charlas: "El Machine Learning es una materia muy complicada de aplicar correctamente, y en la mayoría de los casos no se necesita".

Este problema se podría haber abordado desde una perspectiva puramente estadística, pero nuestra misión en este caso era ampliar el horizonte y conocer que existen otras técnicas.

Incluso si tratásemos de aproximar este problema desde el Deep Learning, costaría bastante trabajo, ya que, al ser un problema con tantos datos, pero a su vez bastante desbalanceado y sin una frontera de decisión clara, las redes neuronales acaban cayendo en clasificadores triviales si no se hacen un buen uso de estas. Por esta razón, se requiere una gran atención a la hora de entrenar dichos modelos.

5. CONCLUSIÓN

La Ciencia de Datos es una rama que se encuentra en pleno auge y que trata de complementar algunas visiones que ya se venían haciendo hasta ahora. Sin embargo, hemos visto que, si existe una limitación intrínseca a los datos, estos modelos no conseguirán hacer "magia", y es por ello por lo que hay que ser muy disciplinado y luchar porque los datos se almacenen de manera óptima.

También, al ser una disciplina en constante desarrollo, lleva a quien quiera utilizar estas técnicas a verse involucrado en todo el proceso y estudiar constantemente los problemas y soluciones que van surgiendo. Sin ir más lejos, con el incremento tan grande al que se ha visto sometida la capacidad de cómputo y la democratización de estas tecnologías (pasa de ser un conocimiento muy especializado a algo que todo el mundo puede estudiar), ha conllevado tanto un aumento de la inversión en los últimos años, como un continuo progreso de los recursos.

Esto ha desembocado en la aparición de algoritmos muy potentes de recomendación, en la clasificación de imágenes a un nivel casi humano, incluso un análisis del lenguaje y sentimientos.

No olvide que estamos en este punto de desarrollo:

EN PRODUCCIÓN - ALERADAS

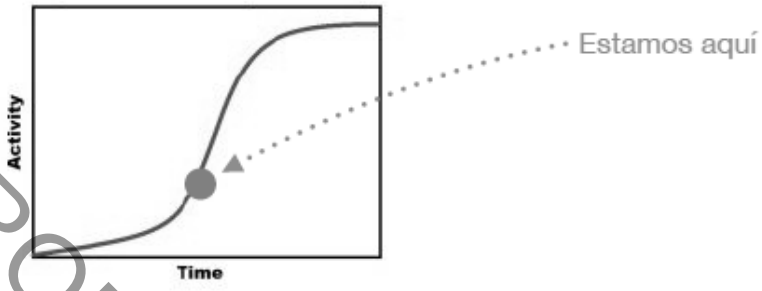


Figura.14 Situación actual de la IA.

EN PRODUCCIÓN - GALERADAS

BIBLIOGRAFÍA

1. Arboix A, García-Eroles L, Comes E, Oliveres M, Targa C, Balcells M, et al. (2008). Importancia del perfil cardiovascular en la mortalidad hospitalaria de los infartos cerebrales. Revista española de cardiología, 61, 1020-1029.
2. Feigin VL, Lawes CM, Bennett DA, Anderson CS. (2003). Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case fatality in the late 20th century. The Lancet Neurology, 2, 53–53.
3. Ministerio de Sanidad y Política Social. (2009). Estrategia en ictus del sistema nacional de salud. 163 p. España.
4. Luis Téllez Ramírez, Jesús Martínez Gómez y Juan Manuel García Torrecillas. (2021). Aplicación de ciencia de datos en la optimización de modelos predictivos para la estimación de mortalidad en ictus isquémico. Hospital Universitario Torrecárdenas. Almería.
5. Smith EE, Shobha N, Dai D, Olson DM, Reeves MJ, Saver JL, et al. (2013). A risk score for in-hospital death in patients admitted with ischemic or hemorrhagic stroke. Journal of the American Heart Association, 2(1), 1020–1029. DOI: 10.1161/JAHA.112.005207.
6. Torrecillas, Juan Manuel García. (2019). Memoria de registro del modelo a validar. Modelo predictivo para la mortalidad hospitalaria en el ictus isquémico no lisado. Almería.

EN PRODUCCIÓN. GALERADAS

EN PRODUCCIÓN. GALERADAS

8. PAPEL DE LOS MODELOS PREDICTIVOS DE REGRESIÓN EN LA ESTIMACIÓN DE LA MORTALIDAD: GENERALIDADES.

AUTORES: JUAN MANUEL GARCÍA TORRECILLAS. CARMEN ROSA GARRIDO. ADRIÁN APARICIO MOTA. LUIS TÉLLEZ RAMÍREZ. MARÍA DEL CARMEN LEA PEREIRA.

EN PRODUCCIÓN - GALERADAS

CONCEPTUALIZACIÓN. MODELOS PREDICTIVOS Y PROGNOSIS

Actualmente se abren paso una serie de técnicas denominadas "análisis predictivo" y "modelado predictivo" que buscan, a través de la estadística de modelización y la ciencia de datos, obtener predicciones basadas en los datos.

Aunque el modelado predictivo se está implantando en múltiples áreas, muy especialmente en el mundo de la empresa y negocios, es innegable el papel que viene desarrollando y su crecimiento exponencial en las disciplinas biomédicas. Una búsqueda por las palabras clave "predictive model" en Pubmed arrojó 12.765 resultados desde 1965 hasta la actualidad y, en los últimos 5 años, arroja 6.808 resultados, en una progresión claramente exponencial año a año.

El objetivo de los modelos predictivos (*MP*) radica, de un modo simplificado, en la estimación del riesgo de ocurrencia de una determinada variable de resultado a partir de un conjunto de variables predictoras, lo cual puede tener utilidad tanto pronóstica como diagnóstica.

La **prognosis o pronóstico** es realmente una de las bases de la medicina hipocrática. Definimos la prognosis como el conocimiento anticipado de algún suceso y concretamente en Medicina, sobre el curso de una enfermedad. Esta palabra viene del griego πρόγνωσις que se compone del prefijo "antes de" y vocablo "conocimiento" en griego. En nuestra área la prognosis es el juicio emitido sobre el resultado esperable o desenlace en una determi-

EN PRODUCCIÓN - GALERADAS

nada patología o enfermedad y, en el caso que nos ocupa, sobre el ictus isquémico y sus muy diferentes y posibles evoluciones.

Siguiendo estas especificaciones, en nuestro proyecto de ejemplo y que pretende ser el núcleo sobre el que se fundamenta esta monografía de investigación, disponemos de una serie de variables independientes, a las que también podemos llamar predictoras; estas variables predictoras, a través de determinados métodos estadísticos (regresión logística, regresión de Cox, Random Forest, árboles de clasificación, redes Neuronales, etc.) nos permitirán estimar el riesgo de que ocurra un evento determinado (variable resultado) que es la muerte del paciente en nuestro caso. En el caso que nos ocupa la finalidad es la estimación de un riesgo (meramente predictiva) aunque en otros casos la finalidad es fundamentalmente diagnóstica.

Este capítulo que aborda las generalidades sobre modelos predictivos no pretende ser más que una introducción a los conceptos básicos y los elementos mínimos que han de estudiarse para el modelado mediante métodos de regresión; otros métodos como los árboles de clasificación y el Random Forest son estudiados en otros capítulos de la monografía.

No se pretende pues que el lector tenga un conocimiento ambicioso del modelado predictivo y quedamos satisfechos si conseguimos despertar la curiosidad por este amplio conjunto de técnicas tan en boga.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - CALIFRADAS

MODELOS BASADOS EN REGRESIÓN SEGÚN LA VARIABLE DE RESULTADO

1. RESULTADO CUANTITATIVO CONTINUO

Cuando el resultado que se pretende estimar o determinar mediante el modelo es de carácter cuantitativo o continuo el procedimiento usual es el de regresión lineal (RL). En este caso la variable de resultado, respuesta o también llamada de criterio (sea Y) es predicha por una cierta combinación de las variables predictoras según la siguiente forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad [1]$$

Básicamente, una vez conocidas las variables y estimados los coeficientes ya es posible determinar el valor de Y que deseamos predecir a través de esta formulación lógico-matemática que lo hace posible. De la existencia de una variable explicativa o más de una dependerá que nos enfrentemos ante una RL simple o una RL múltiple respectivamente.

La regresión lineal no es el único método, pero sí el más frecuente a la hora de estimar variables dependientes de carácter cuantitativo. Una generalización de la RL simple es la RL múltiple, tal como hemos visto en la formulación anterior, ya contiene un número de variables predictoras superior a dos como se ha comentado previamente.

En el caso de la RL simple los parámetros β_0 y β_1 son conocidos constante y pendiente respectivamente. El conjunto de los Betas se suele denominar "coeficientes de regresión" sin más, y éstos serán dos en el caso de la RL simple (β_0 y β_1) o múltiples (hasta β_k) en el caso de la RL múltiple.

En el caso de la RL simple la respuesta nos viene dada por una combinación lineal de la variable explicativa. Esta variable explicativa se acompaña de un coeficiente que ha de interpretarse como el cambio promedio en la variable respuesta por cada unidad de aumento de la variable explicativa.

En el caso de la RL múltiple hay algunos cambios de interés. Por un lado el coeficiente de una de las variables explicativas se interpreta de igual modo (cambio promedio en las variables dependientes por cambio en una unidad de la explicativa).

Condiciones que han de darse para la RL múltiple:

a. Linealidad: la relación entre la variable dependiente y las variables independientes debe ser lineal, si esto no ocurre estamos ante un error de especificación. Para comprobar este supuesto se pueden utilizar gráficos de dispersión parcial entre la variable dependiente y cada una de las variables predictoras.

b. No colinealidad: En este tipo de modelos los predictores han de ser independientes y es importante que hayamos descartado la existencia de colinealidad entre ellos. La colinealidad consiste básicamente en que una variable predictora esté relacionada linealmente con uno o más de los restantes predictores, y esto acaba provocando un artefacto por el cual es casi imposible conocer adecuadamente la contribución de cada predictor individual al modelo.

c. Los residuos deben seguir una distribución Normal con media cero y desviación típica 1, $N(0,1)$. La forma de verificar que ocurre este supuesto suele discurrir la aplicación de un test de

EN PRODUCCIÓN - GALERADAS

normalidad, por ejemplo el test de Kolmogórov-Smirnov, y adicionalmente por una herramienta gráfica tipo histograma.

d. Homocedasticidad: Bajo este término nos referimos a la variabilidad constante de los residuos a lo largo de todo el rango de valores pronosticados. En el caso buscado, la varianza será constante, lo cual motivará que los residuos se distribuyan de modo aleatorio, es decir, sin un patrón específico (cónico, por ejemplo) y manteniendo un mismo grado de dispersión.

e. Independencia entre los residuos (no autocorrelación): los residuos obtenidos del modelo de predicción calculado han de ser independientes entre sí. Para detectarla puede utilizarse el test de Durbin-Watson, que toma valores entre 0 y 4. Toma el valor 2 cuando los residuos son totalmente independientes, valores inferiores a 2 indican autocorrelación positiva y valores superiores a 2 suponen autocorrelación negativa.

En algunas ocasiones, cuando no se cumple alguno de estos supuestos, puede ser útil realizar transformaciones sobre las variables para poder calcular un modelo de regresión lineal, la transformación más común es la logarítmica.

2. RESULTADO DICOTÓMICO

Cuando el resultado de la variable de respuesta adopta dos categorías (Sí/No, Muerto/Vivo, etc.) entonces estamos ante un caso en el que procede el uso de un modelo de regresión logística binaria (*RLB*).

La *RLB* fue desarrollada por Cox en 1.958 y consiste en un método de regresión para estimar la probabilidad de ocurrencia de una variable binaria en función de una serie de predictores que pueden ser de cualquier tipo (cuantitativos, cualitativos: dicotómicos o politómicos).

EN PRODUCCIÓN - GALERADAS

Este caso adopta la siguiente formulación desde el punto de vista formal:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad [2]$$

$$\text{Donde } p = P[Y = 1]$$

En este caso se denota el resultado como 1 para representar la presencia del evento de interés, por ejemplo, presencia de la enfermedad, y para 0 notar la ausencia evento de interés, por ejemplo, ausencia de enfermedad. Este modelo estima la probabilidad de que Y sea igual a 1 (probabilidad de que ocurra el evento de interés para la variable dicotómica).

En la RLB se obtiene una información adicional, y es que podemos conocer cuál es la probabilidad de que la variable dependiente pertenezca a cada una de las dos categorías en función de los diferentes valores que vaya adoptando la predictora o independiente.

Definición

Volviendo a la notación matemática previa y para definir adecuadamente la RLB consideremos una variable de resultado Y tal que pueda adoptar el valor 1 (ocurrencia del fenómeno) o de 0 (no ocurrencia del fenómeno).

La notación $log(\frac{p}{1-p})$ con la que iniciábamos la definición previa no es más que la probabilidad de ocurrencia del fenómeno (p) dividido por la probabilidad de no ocurrencia del mismo ($1-p$), y a esto se le denomina Odds (ventaja). Básicamente nos está indicando cuántas veces es más probable que ocurra un evento frente a la probabilidad de que no ocurra.

El problema del uso de las Odds en este caso es que puede adoptar valores entre 0 e infinito. Si las Odds ratio es inferior a 1 se considera que la variable predictora puede ser un factor de protección a presentar el evento de estudio, si toma valores supe-

EN PRODUCCIÓN - GALERADAS

riores a 1 indica que la variable predictora puede ser un factor de riesgo a presentar el evento de estudio, y si toma el valor 1 se considera que las dos categorías comparadas son iguales. Una forma de abordar este problema es realizar el logaritmo de la odds, en este caso el ya citado $log(\frac{p}{1-p})$, al cual pasamos a denominar ahora "$Logit(p)$".

Si manipulamos la expresión [2] algebraicamente podemos llegar a la siguiente formulación que es bastante cómoda de utilizar:

$$Logit(p_i) = ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i}$$

Y desde aquí es sencillo llegar a:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i})}}$$

En este momento hemos simplificado mucho el procedimiento mediante la manipulación algebraica al tiempo que ya, los valores de la probabilidad estimada se encuentran entre los valores 0 y 1, y además siguen una distribución de morfología sigmoide:

EN PRODUCCIÓN · GALERADAS

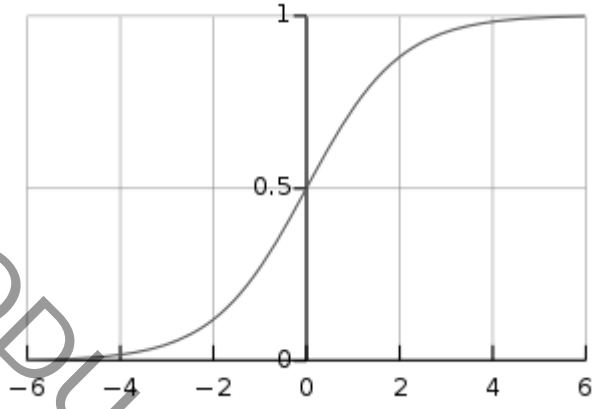


Fig. 1. Función sigmoide en Regresión Logística (Fuente CC Wikipedia)

En el caso que nos ocupa en el presente manual, los métodos de regresión logística han estado muy presentes en la elaboración de modelos y remitimos al lector al capítulo correspondiente para más detalles. No obstante, avanzamos el hecho de que se consigue estimar la probabilidad de que un paciente fallezca una vez ingresa por un ictus isquémico a partir de una función de carácter logístico que queda representada según la anterior formulación y esto se acompaña de una serie de medidas sobre la bondad del modelo obtenido que se comentarán a lo largo de este texto.

3. TIEMPO EN EL RESULTADO

En ocasiones es importante conocer si se ha producido un evento y el tiempo que ha transcurrido hasta que se produce este. El modelo de regresión de Cox tiene en cuenta ambos aspectos, si se presenta o no el evento de interés y cuánto tiempo ha tardado en aparecer. Por ejemplo, sería adecuado plantear un modelo de Cox para estimar si un paciente que se diagnostica un determi-

EN PRODUCCIÓN. GALERADAS

nado tumor va a fallecer o no, teniendo en cuenta el tiempo que lleva expuesto a la enfermedad.

Aunque no es el objeto de este manual, hay una serie de conceptos dentro del estudio o análisis de la supervivencia que debemos conocer:

1. Bases y concepto de censura: El modelo previamente establecido por Cox, realmente puede aplicarse para la predicción del evento teniendo en cuenta el tiempo transcurrido hasta el mismo. El evento no tiene por qué ser la mortalidad sino la ocurrencia de determinado fenómeno. La respuesta sería pues el tiempo desde el origen (cuando se produce un determinado diagnóstico, por ejemplo) hasta la ocurrencia de un determinado fenómeno que provoca el abandono del seguimiento del sujeto en estudio. Los fenómenos que provocan esta salida del estudio pueden ser haber presentado el evento de estudio u otros motivos como la invalidez, el reingreso, etc. Este es el llamado *analysis of time to event data*.

De otro lado, puede darse un fenómeno conocido como "censura". En ocasiones no todos los individuos se siguen hasta que ocurre un determinado desenlace, sino que salen del estudio antes, son los denominados datos "censurados". En esta situación se indica que el tiempo está censurado a la derecha (en la curva temporal) y como ejemplos de salida de los casos podríamos citar algunos:

- Finalización antes de lo previsto del estudio.
- El paciente fallece, pero no debido a la etiología en estudio sino a motivos no relacionados.
- El paciente se deslocaliza, no tenemos modo de contactar con él y no sabemos qué ha ocurrido, etc.

Por tanto, es evidente que los datos censurados hacen que el análisis de supervivencia varíe de modo sustancial, son datos pues

EN PRODUCCIÓN - GALERADAS

que han de mirarse con un muy especial detalle. Todos los tiempos censurados por ser retirados o perdidos suelen ser analizados del mismo modo pero se trata de casos muy distintos.

EN PRODUCCIÓN - GALERADAS

3. MÉTODOS ESTADÍSTICOS

En un análisis de supervivencia el objetivo inicial es el cálculo de la función de supervivencia o supervivencia sin más, que se denota por $S(t)$, siendo dicha función la probabilidad de que un individuo sobreviva desde la fecha de inicio del estudio hasta un momento t del estudio. Por el contrario, la función de riesgo es la probabilidad de que muera en este momento t , y se denota mediante $\lambda(t)$. Este análisis puede realizarse mediante técnicas paramétricas o no paramétricas según la distribución del tiempo sea o no conocida respectivamente.

a. Método de Kaplan-Meier (KM): Básicamente se utiliza para la estimación de la función de supervivencia mediante métodos no paramétricos. El método de KM calcula la proporción acumulada de supervivencia para el tiempo individual de cada paciente cada vez que ocurre un evento. Mediante este método se obtienen funciones escalonadas de la supervivencia. Es decir, la función de supervivencia se define como:

$$S(t) = \prod_{t_i \leq t} p_i$$

En el caso del ictus isquémico, mostramos un ejemplo de ello elaborado con métodos propios.

EN PRODUCCIÓN GALERADAS

La comparaciones de curvas de supervivencia, como las que muestra el ejemplo puede realizarse mediante el método KM (test de log-rank) que a continuación se detalla.

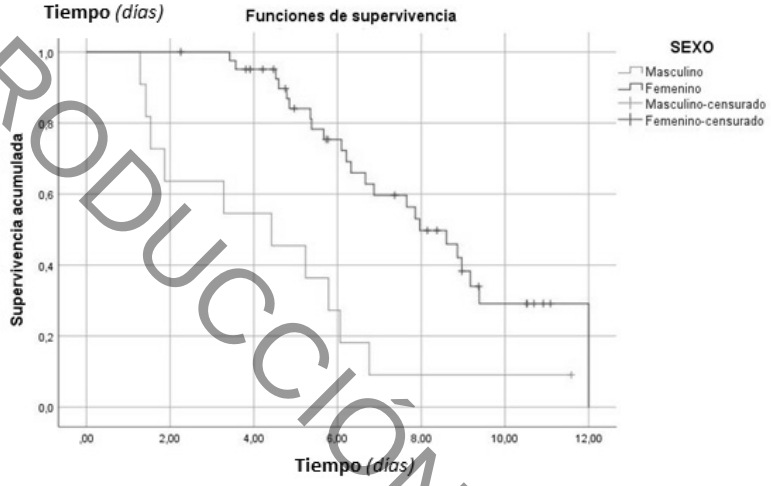


Fig.2. Disponible bajo licencia Creative Commons en https://estamatica.net/analisis-de-supervivencia-regresion-de-cox/

Dadas las funciones de supervivencia en el ingreso por ictus para diferentes valores de la estancia se trata de contrastar la igualdad de dos funciones de distribución de tiempos de supervivencia.

$$H_0: F_1 = F_2$$

$$H_1: F_1 \neq F_2$$

En el caso que nos ocupa, F_1 es la función de distribución de tiempos de supervivencia masculina y F_2 es la femenina.

Para llevar a cabo el contraste se aplica el Test de log-rank que determina en cada uno de los eventos el número observado y esperado en cada grupo, asumiendo que no hay diferencias signifi-

EN PRODUCCIÓN - GALERADAS

cativas entre ellos, es decir, bajo la premisa de que la hipótesis nula H_0 es cierta.

b. Modelo de regresión exponencial: Este modelo paramétrico basado en la distribución exponencial, expresa la tasa de riesgo en el instante t, $h(t)$, como combinación lineal de las k variables predictoras X. Si $Texp(\lambda)$ entonces podemos definir las función de supervivencia como:

$$S(t) = e^{-\lambda t}$$

Realizando el proceso antes mencionado, podemos expresar la tasa de riesgo de la siguiente forma:

$$h(t) = \lambda e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}$$

c. Modelo de Riesgos proporcionales de Cox: Aunque ya ha sido comentado, por ser el más utilizado merece la pena insistir en él. Este modelo de regresión es una técnica multivariante que nos permite evaluar y detectar la relación entre un conjunto de variables (independientes, predictoras) y la tasa de ocurrencia del evento en estudio (la denominada función de riesgo). El método también nos permite conocer el tiempo libre de evento (supervivencia) para cualquier individuo, una vez conocidas las variables explicativas, a partir de las cuales se hace la estimación.

Una regresión de Cox se interpreta de un modo similar a la regresión logística con la diferencia fundamental de que nos aporta las Hazard Ratio en lugar de las Odds Ratio de la regresión logística.

El término Hazard Ratio (HR) es la tasa de riesgos instantáneos, esta tasa se obtiene como el cociente entre el riesgo instantáneo de que se presente en el evento de estudio en un periodo temporal concreto, entre dos grupos de población, por ejemplo, grupo de expuestos y no expuestos, o grupo de intervención y control. Si la HR es igual a 1 indicaría que existe igual riesgo de

EN PRODUCCIÓN. GALERADAS

presentar el evento en ambos grupos, si HR es superior a 1, se correspondería a un mayor riesgo en el grupo de expuesto frente al de no expuestos, y si HR es inferior a 1, se revelaría que el riesgo de presentar el evento de estudio en el grupo de expuestos es menor que en el grupo de no expuestos.

En este supuesto consideramos que el logaritmo de la Hazard ratio (HR) del suceso "muerte" se relaciona con las variables explicativas o independientes según la siguiente formulación:

$$log(\frac{h(t)}{h_o(t)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \text{ [3]}$$

donde $h_o(t)$ es el denominado "riesgo base" y representa el riesgo de fallecer cuando todas las variables adoptan el valor 0 y $h(t)$ representa el riesgo en el instante t.

Dado este caso y la ecuación [3] podríamos calcular o predecir la probabilidad de muerte de un sujeto en un momento determinado, conociendo los valores de las diferentes variables independientes y sustituyendo en la ecuación a estas y al valor del tiempo para el que queremos hacer la estimación.

Al igual que hemos hecho previamente, aunque no es objeto del capítulo un estudio detallado sino sólo conceptual de estos, la ecuación del modelo para una regresión de Cox sería:

$$ln(\lambda t) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \text{ [4]}$$

En el modelo de Cox el cociente de riesgos instantáneos o HR de dos sujetos no depende del tiempo, sino que es constante y los riesgos además son proporcionales (de ahí el nombre de la técnica). El parámetro HR se define de la siguiente manera:

$$HR = e^{\beta(X_i - X_j)} \text{ [5]}$$

EN PRODUCCIÓN - GALERADAS

Supongamos que tenemos dos individuos que se diferencian en una unidad en términos de X. En este caso, el aumento de una unidad en la variable explicativa multiplica el riesgo por la exponencial del parámetro [5] (Cantalejo, 2015). Además, dependiendo del signo del parámetro se podrá observar si el riesgo aumenta o disminuye a medida que aumenta la predictora.

EVALUACIÓN DEL RENDIMIENTO EN CUANTO A LAS PREDICCIONES

Además de la clara utilidad del desarrollo y uso de un modelo predictivo, esto no sería de ningún interés si no fuera posible determinar "cuán bueno" es el modelo que hemos alcanzado. En muchas áreas como la Cardiología, las enfermedades infectivas, el cáncer y muy especialmente en el campo que nos ocupa, el ictus, los modelos han de poder aportarse junto a métricas que nos permitan conocer el rendimiento de los mismos. En las próximas líneas vamos a hacer mención de un modo muy básico a la discriminación y calibración de los modelos; cualquier intento de ir más allá requeriría un nuevo capítulo para detallar este apartado, excediendo por tanto las aspiraciones de este manual.

Medidas para evaluar la discriminación y calibración del modelo

La **discriminación** de un modelo es una medida que nos indicará la capacidad del mismo para predecir quién sufrirá un evento ictal (utilizando el tópico de esta monografía) y diferenciarlo de quién no lo sufrirá. Viene a ser pues la probabilidad de hacer una clasificación correcta para un par de pacientes dados de tal modo que uno tenga el resultado que queremos predecir frente a otro que no.

El Área Bajo la Curva (AUC) o siguiendo la terminología anglosajona, ROC ("receiving operator curve") constituye la métrica

EN PRODUCCIÓN. GALERADAS

habitual en la medición del grado de discriminación de un modelo. Veamos de una forma intuitiva la interpretación de un AUC del 84%: tendríamos pues una probabilidad de clasificar correctamente un par de pacientes (uno con el evento de interés, otro no) del 84%. Esta curva se conforma mediante la confrontación, para cada punto de corte, de la sensibilidad en el eye Y y del complementario de la especificidad en el eje X.

Por otro lado, debemos introducir el concepto de "**calibración**" del modelo. Trátese de un valor que nos muestra la diferencia entre la media de los resultados realmente obtenidos y los predichos. En consecuencia, la calibración expresa el grado de concordancia entre los resultados observables en la realidad frente a los que predice el modelo que estamos evaluando. Técnicamente (Núñez et al, 2011) la calibración "es la probabilidad de modelo de producir una estimación no sesgada de la probabilidad de ocurrencia de un evento". La calibración suele medirse mediante el test de Hosmer-Lemeshow (Lemeshow et al, 1989) o, aunque con menor frecuencia, en el caso de modelos de regresión de Cox, mediante el test de Gronnesby y Borgan. Otro método usado con frecuencia además de la calibración general es la pendiente de calibración.

Tanto en el caso de la discriminación como de la calibración, la técnica más fiable para su determinación debe estar basada en el remuestreo o bootstraping ya que al realizar su determinación basándonos en este proceder aseguramos la validez interna del modelo de un modo simultáneo.

Desde un punto de vista gráfico la discriminación, estimada como AUC o ROC puede observarse en la figura 3, al tiempo que la calibración suele mostrarse mediante la confrontación de 10 deciles de riesgo de los valores observados frente a los predichos (calibración global del modelo) o bien mediante la pendiente de calibración o mediante una curva *loess* que enfrenta los valores observados frente a los predichos (figura 4)

EN PRODUCCIÓN - GALERADAS

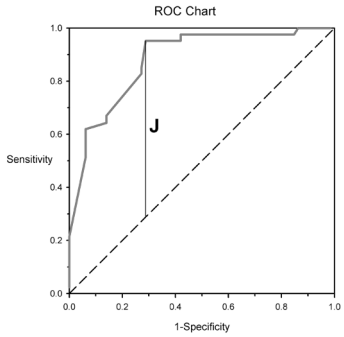


Fig. 3. Curva ROC con área bajo la curva. Disponible con licencia CCC en la URL https://commons.wikimedia.org/wiki/File:ROC_Curve_Youden_J.png

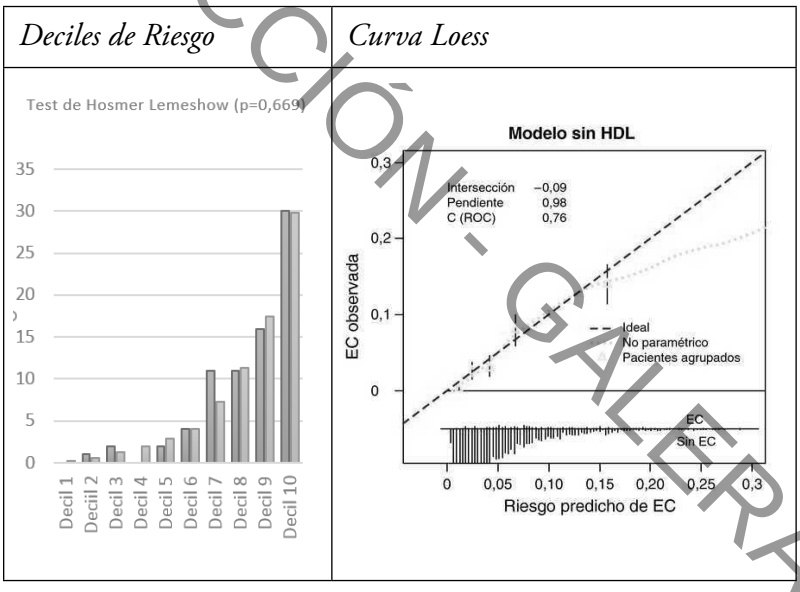


Figura 4. Gráficos utilizados en calibración

Deciles de riesgo, elaboración personal. Curva Loess (Steyerberg et al), disponible en la URL https://www.revespcardiol.org/es-medidas-del-rendimiento-modelos-prediccion-articulo-S0300893211004830

EN PRODUCCIÓN - GALERADAS

BIBLIOGRAFÍA

1. Sánchez-Cantalejo Ramírez y Sánchez-Cantalejo Castañeda. Análisis de supervivencia. Cuadernos metodológicos de la EASP. 2014.
2. Turkson A, Addor J, Ayiah-Mensah F. The Cox Proportional Hazard Regression Model Vis-à-Vis ITN-Factor Impact on Mortality Due to Malaria. Open Journal of Statistics. 2021;11:931-962.
3. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer International Publishing; 2019.
4. Kuhn M, Johnson K. Applied Predictive Modeling. Springer New York; 2013.

EN PRODUCCIÓN - GALERADAS

9. MODELO PREDICTIVO PARA LA ESTIMACIÓN DE MORTALIDAD EN EL ICTUS ISQUÉMICO NO LISADO MEDIANTE METODOLOGÍA DE REGRESIÓN LOGÍSTICA.

AUTORES: ADRIÁN APARICIO MOTA. MARÍA DEL CARMEN LEA PEREIRA. JUAN MANUEL GARCÍA TORRECILLAS.

Los modelos predictivos agrupan una variedad de técnicas de modelización y aprendizaje no supervisado. Estas técnicas pueden ser utilizadas para predecir la mortalidad de los pacientes que sufran alguna patología o calcular el riesgo de mortalidad de los pacientes según variables clínicas o fisiológicas. El objetivo de un modelo predictivo para la estimación de la mortalidad es evaluar la probabilidad de que un individuo de nuevo ingreso fallezca en base a una serie de variables independientes, ya sea durante el ingreso o en un número determinado de días, meses o años.

EN PRODUCCIÓN. GALERADAS

EN PRODUCCIÓN - GALERADAS

INTRODUCCIÓN

"Todos los modelos están equivocados, pero algunos son útiles" (George Edward Pelham Box).

Partimos de esta premisa con la idea de hacer ver que no hay un modelo que sea capaz de predecir correctamente al 100% (siempre que intervengan variables que sean estocásticas), sin embargo, mediante técnicas estadísticas, somos capaces de estimar con cierta confianza, el desenlace de un evento. En nuestro caso, la mortalidad de los pacientes es la variable estocástica a predecir.

Antes de la existencia de los modelos predictivos para la estimación de la mortalidad, el médico ya recurría a ciertas estimaciones de dicha mortalidad con el pronóstico (lo que hoy llamamos diagnóstico diferencial). El pronóstico era un juicio que el médico hacía según las características de los pacientes o durante el curso de una enfermedad. Dicho pronóstico intenta dar una probabilidad de supervivencia, ya que, esto ayudaba a la hora de actuar. Uno de los problemas de este pronóstico dado por el personal sanitario es que estaba sesgado por la visión subjetiva del médico y el número de pacientes limitados con características similares que había podido observar dicho médico.

Esta es una de las ventajas de un modelo predictivo, pese a no tener un juicio clínico, es capaz de dar una probabilidad en base a un número determinado de pacientes, analizarlo, y comparándolo con individuos similares, dar una probabilidad de su mortali-

EN PRODUCCIÓN - GAMERADAS

dad. Este modelo puede alcanzar a muchos más pacientes y, si se tiene cuidado con los sesgos, lo hará de manera más objetiva. Son por tanto, una herramienta más a tener en cuenta junto al juicio clínico de un profesional.

Importación y tratamiento de los datos

La base de datos de estudio que se utilizará en este capítulo como ejemplo, recoge 117527 pacientes que han sufrido ictus isquémico y 82 variables, entre ellas, la variable dependiente "Exitus". El principal problema a la hora de crear un modelo predictivo para predecir la probabilidad de dicha variable dependiente, es que estamos trabajando con datos no balanceados, esto es, del total de pacientes, 110333 no fallecieron tras padecer ictus isquémico mientras que tan sólo 7194 pacientes fallecieron tras padecerlo.

El modelo que se genera es un clasificador trivial, es decir, clasificará todos los pacientes a un grupo para conseguir una mejor precisión global. No obstante, un examen cuidadoso de la tabla de clasificación nos permitirá conocer en qué momentos el modelo es útil y en cuáles no, como estudiaremos más adelante.

Para la creación del modelo de regresión logística múltiple se eliminan aquellas variables que tienen muchos "casos perdidos", las variables que clínicamente no influirán y por último aquellas variables que nos pueden generar colinealidad en el modelo o se calculan "a posteriori".

Finalmente, y siguiendo este ejemplo, retuvimos 11 variables independientes para crear un modelo de predicción de la variable dependiente Exitus.

- Variables cuantitativas: En nuestro caso, solamente se dispuso de la variable Edad que está definida como variable cuantitativa discreta.

EN PRODUCCIÓN - GALERADAS

- Variables cualitativas: Las demás variables dependientes del modelo fueron cualitativas dicotómicas, principalmente se trató de variables de presencia de otras enfermedades.
- Variable dependiente: Nuestra variable dependiente fue el Exitus, variable cualitativa dicotómica (Si/No) ocurre el suceso (fallecer).

Problemática de datos desbalanceados

Un conjunto de datos no balanceados suele ser un problema común en algunos ámbitos de la medicina. Este es uno de los principales inconvenientes en disciplinas como la minería de datos y el aprendizaje automático.

Un conjunto de datos se dice que está desbalanceado cuando en la variable dependiente una de sus clases (mayoritaria) está significativamente más representada que la otra. En nuestro caso la clase Vivos está significativamente más representada que la clase Exitus.

Por eso mismo, es adecuado abordar dicho desbalanceo con diferentes técnicas para reducir la clase mayoritaria, incrementar la clase minoritaria o ambas a la vez.

¿Qué podemos hacer?

En primer lugar, podríamos intentar conseguir más datos. Generalmente conseguir un mayor número de registros no es posible o no soluciona nada si nuestros registros ya son lo suficientemente grandes.

Para resolver el problema y elaborar un modelo para la predicción de la mortalidad aplicamos el método DownSampling. Éste consiste en reducir el número de registros de la categoría mayoritaria con el fin de obtener un conjunto balanceado. Este es un buen método para la muestra de este capítulo, ya que es lo suficientemente grande.

EN PRODUCCIÓN - GALERADAS

Análisis exploratorio de nuestra muestra

Las variables expuestas a continuación están recogidas como variables cualitativas dicotómicas, mediante las cuales se calculará la frecuencia relativa según el paciente sea exitus o no. Para el contraste de hipótesis de estas variables, diferenciando por Exitus = SI/NO, se utilizará el test exacto de Fisher.

	EXITUS NO	EXITUS SI	
Variables	Frecuencia relativa	Frecuencia relativa	P-valor
Sexo = Mujer	45.5%	61.3%	0.00*
Reingreso = Si	4.8%	9.8%	0.00*
Cardiopatía Isquémica = Si	3.1%	4.5%	0.00*
HTA=Si	60.5%	56.4%	0.00*
FA=Si	7.2%	11.3%	0.00*
Dislipemia=SI	40.8%	27.7%	0.00*
ICC=Si	1.4%	3.2%	0.00*
EPOC=Si	5.4%	6.4%	0.00*
DM=Si	3.3%	2.4%	0.00*
Infarto Basilar=Si	0.1%	0.0%	0.186

La variable de la edad está recogida como variable cuantitativa. En base a ello se calcularán medidas de tendencia central (medias e intervalo de confianza para la media). Se comprobará la normalidad de la variable cuantitativa con el test de Kolmogórov-Smirnov.

EN PRODUCCIÓN - GALERADAS

Para el contraste de hipótesis de igualdad de medidas de localización en los Exitus = SI/NO se utilizará el test de la t de Student para muestras independientes si estas se distribuyen según una distribución normal, en caso contrario se utilizará el test U de Mann-Whitney.

	EXITUS NO		EXITUS SI		
Variables	Media	IC para la media	Media	IC para la media	P-valor
Edad	73.48	[73.41, 73.56]	83.61	[83.40, 83.83]	0.00



Podemos observar en el gráfico cómo la variable Edad influye a la hora de fallecer por ictus isquémico. En los pacientes que fallecen la medida de localización (mediana) es mayor incluso que el tercer cuartil de los pacientes que no fallecen tras el ingreso por ictus isquémico.

EN PRODUCCIÓN - GALERADAS

Selección de un modelo para estimar la mortalidad en los pacientes con ictus.

La finalidad de un modelo es predecir la variable respuesta en observaciones futuras. El error que nos devuelve en primer lugar el modelo es el error del conjunto dónde hemos entrenado dicho modelo. Este error está sesgado ya que tiene en cuenta observaciones que ha "visto" anteriormente y forman parte de la elaboración de dicho modelo.

Existen diferentes estrategias para paliar este sesgo y validar un modelo en base a observaciones que no ha "visto" anteriormente. En nuestro caso utilizaremos el método K Fold Cross Validation en el subconjunto train (entrenamiento) y posteriormente, validaremos el modelo en el subconjunto test (o subconjunto de validación).

Este método también es iterativo, pero en este caso, dividiremos nuestro conjunto de forma aleatoria en k grupos de, aproximadamente, el mismo tamaño. En cada iteración se usa k-1 grupos para entrenar el modelo y el grupo excluido para su validación. Finalmente, el error del modelo será el promedio de las k-iteraciones realizadas por esta técnica en cada uno de los grupos de validación. Hay que tener en cuenta que una mejor estimación del error de predicción no implica un mejor criterio de selección del modelo.

Seleccionaremos el modelo en base al error mediante el método K-Fold Cross-Validation en nuestro conjunto de datos balanceados mediante downsampling.

Modelos de regresión logística para la estimación de mortalidad

Exponer un modelo de regresión logística como método para estudiar la asociación entre una variable dependiente dicotómica y varias variables predictoras suele ser una de las técnicas más utilizadas para la estimación de la mortalidad. En nuestro caso usaremos un modelo de regresión logística múltiple para la es-

EN PRODUCCIÓN - GALERADAS

timación de mortalidad en pacientes con primo ingreso de ictus isquémico.

El modelo logístico multivariante se expresa como:

$$log\frac{P}{1-P} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Una forma de expresar el mismo modelo es:

$$P = \frac{e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}$$

De esta manera el modelo permite calcular la probabilidad de que un individuo con determinados valores en las variables predictores presente la característica de interés, en nuestro caso, la mortalidad.

Procedemos a realizar lo expuesto anteriormente para estimar la probabilidad de fallecer tras padecer un ictus isquémico. Los datos recogidos están sacados de una base de datos del CMDB (Conjunto Mínimo Básico de Datos) de pacientes que padecieron ictus isquémicos de los años 2013 a 2015.

Modelo de regresión logística múltiple

Para la estimación de la mortalidad en pacientes con ictus isquémico, y teniendo en cuenta nuestros datos vamos a crear un modelo de regresión logística múltiple. De esta manera se intentará predecir la mortalidad mediante la variable "Exitus" recogida para cada paciente.

Previamente a decidir el método para entrenamiento y evaluación aplicamos un método para muestra de datos desbalanceadas. En la práctica fueron usados varios métodos, tanto de oversampling (Smote) como de undersampling.

EN PRODUCCIÓN - GALERADAS

Finalmente optamos por el método downsampling para equilibrar la muestra y proporcionar una mayor importancia a los pacientes fallecidos y poder detectarlos con mayor exactitud. El método que usaremos para entrenamiento y evaluación del modelo será el método K-Fold Cross Validation y Hold-Out. Este método consiste en dividir nuestro conjunto en train y test (70% y 30% de nuestra muestra respectivamente) de manera aleatoria, luego en el subconjunto train aplicaremos el método K-Fold Cross Validation, dividiendo la muestra en 5 submuestras y aplicando 5 modelos. El modelo validado final será un promedio de los 5 modelos, validando cada vez en el subconjunto con el que no se creó dicho modelo. Una vez elaborado el modelo lo evaluaremos con el subconjunto test, de esta manera evitamos el sesgo de evaluar el modelo con los mismos individuos con los que lo creamos.

Creamos el modelo de regresión logística. Observamos que 3 variables quedan fuera del modelo al no ser significativas (DM, Infarto Basilar y EPOC). Procedemos a eliminar aquellas variables no significativas y volvemos a crear el modelo excluyendo dichas variables para obtener un modelo más sencillo y seguir el principio de parsimonia.

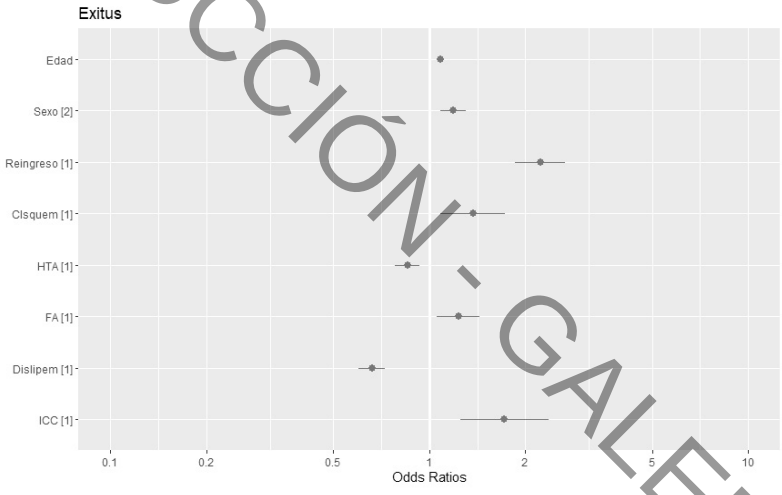
Variables	B	Error estándar	Wald	P-valor	OR
Edad	0,08	0	1067,72	0.00	1,08
Sexo	0,22	0,05	21,72	0.00	1,25
Reingreso	0,87	0,09	88,81	0.00	2,39
Cardiopatía Isquémica	0,44	0,12	13,78	0.00	1,55
HTA	-0,18	0,05	15,27	0.00	0,83
FA	0,22	0,08	7,92	0.01	1,25
Dislipemia	-0,42	0,05	73,37	0.00	0,66
ICC	0,5	0,16	9,76	0.00	1,65
Constante	-6,2	0,19	1062,79	0.00	0

EN PRODUCCIÓN - GALERADAS

Odds-Ratios

El odds ratio es una medida de asociación entre dos variables, en este caso es la asociación entre la variable estocástica a predecir con cada una de las variables predictoras (Correlación bivariada).

El odds ratio expresa si la probabilidad de que fallezca o no el paciente difiere en el aumento de la variable cuantitativa o en la ocurrencia o no de una de las variables categóricas. Es teóricamente la probabilidad de que fallezca por la probabilidad de que no fallezca teniendo en cuenta siempre la variable predictora para cada OR.



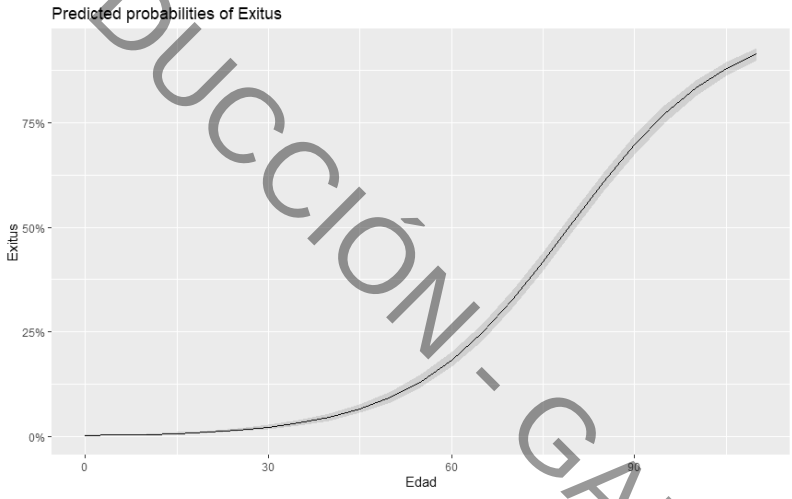
Se puede observar como las variables que influyen más a la hora de fallecer tras ingreso de ictus isquémico son las variables Reingreso e ICC.

Se observa como para las comorbilidades HTA y Dislipemia es un factor de protección en la mortalidad de ictus isquémico (aunque esto pueda parecer paradójico, tiene su claro significado clínico).

EN PRODUCCIÓN - GALERADAS

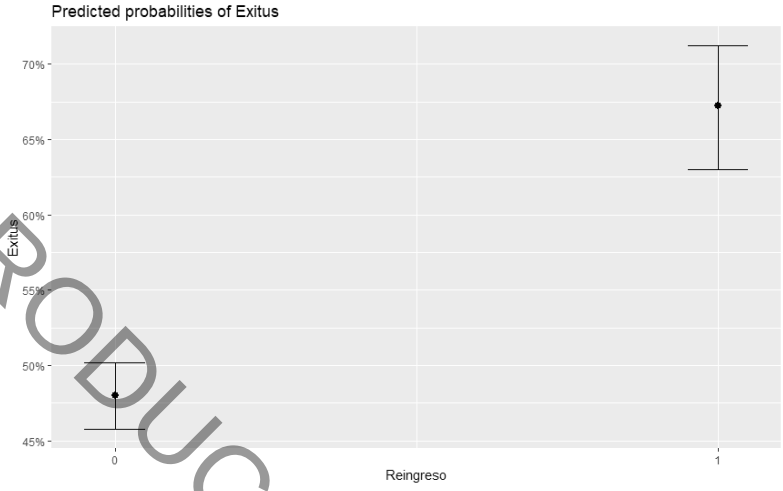
Respecto a la edad no podemos ver su importancia gráficamente comparándola con variables categóricas, ya que posiblemente tenga un mayor peso que otras variables si categorizamos dicha variable.

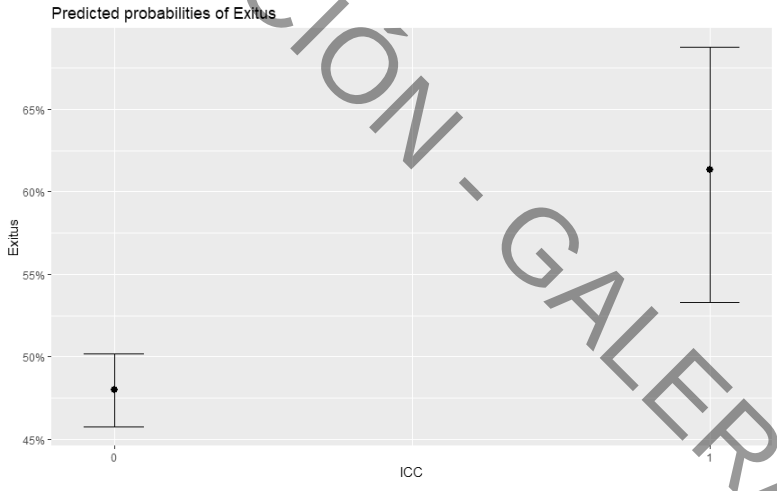
A continuación, se visualizará las probabilidades pronosticadas de fallecer por el modelo respecto a la edad de cada uno de los pacientes.



Si hacemos las mismas comparaciones para alguna de las variables categóricas más influyentes, se obtiene:

EN PRODUCCIÓN - GALERADAS





EN PRODUCCIÓN - GALERADAS

Cómo validar el modelo predictivo

1. *Bondad de ajuste en modelos de regresión logística.*

En los modelos de regresión logística existen varias medidas para comprobar cómo de bueno es el ajuste de las predicciones del modelo con los valores observados. Para nuestro modelo haremos uso de dos test para comprobar la bondad de ajuste; el test de Hosmer-Lemeshow y el test de subconjuntos de coeficientes.

Test subconjunto de coeficientes

Se utiliza el estadístico "deviance" para contrastar la nulidad de subconjunto de los coeficientes.

Se compara dos modelos uno con un número menor de regresores para contrastar la hipótesis nula:

$$\{H0: \quad B_r + 1 = \cdots, B_p = 0 \; H1: B_i \neq 0 \; \forall \, j \in \{r+1 \ldots p\} \qquad r < p$$

Se calcula la desviación del modelo para la hipótesis nula y se compara con el modelo con p regresores.

El p-valor obtenido en nuestro modelo en este test es de 0,00, por lo que lo consideramos un modelo adecuado.

Test de Hosmer Lemeshow

Las observaciones son clasificadas usualmente en 10 grupos, el primer grupo estará formada por las n/10 observaciones con los $\hat{\pi}$ más pequeños, el segundo por los n/10 observaciones con los siguientes $\hat{\pi}$ más pequeños y así sucesivamente.

$$HL = \sum_{i=1}^{g} \frac{(O_i - N_i \underline{\pi}_i)^2}{N_i \underline{\pi}_i (1 - \pi_i)}$$

Siendo:

O_i el número de fracasos.

$N_i \underline{\pi}_i$ la frecuencia esperada en cada grupo.

$\underline{N_i}$ número de observaciones en el grupo i.

EN PRODUCCIÓN - GALERADAS

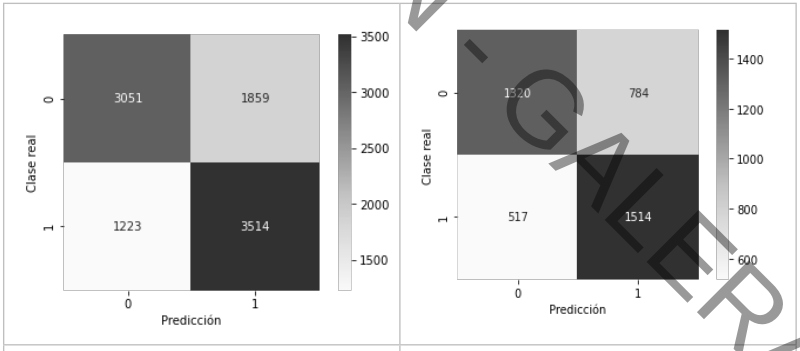
El p-valor obtenido en el test de Hosmer-Lemeshow es de 0,00, por lo que lo consideraríamos un modelo adecuado. Pero debido a la revisión de antecedentes, el test de Hosmer-Lemeshow no es una prueba adecuada cuándo queremos comprobar la fiabilidad de un modelo si el tamaño de la muestra es demasiado grande. (Allison, s.f.)

2. Métricas para la evaluación del modelo predictivo

Un modelo predictivo devuelve varias "métricas de evaluación" que podemos tener en cuenta todas ellas para evaluar nuestro modelo y comprobar su utilidad para la estimación de la mortalidad.

Matriz de confusión

Una de estas métricas de evaluación es la matriz de confusión sirve para mostrar de forma explícita cuando una clase es confundida con otra.



Conjunto train | *Conjunto test*

Podemos observar una precisión cercana al 70%, clasificando mejor los Exitus=1, siendo este el objetivo principal de este capítulo.

EN PRODUCCION - GALERADAS

Exactitud

Otra de las métricas a tener en cuenta es la exactitud o accuracy. La exactitud nos indica la habilidad de un modelo para predecir de forma correcta cada clase. Cuando hay datos no balanceados la exactitud no es una buena métrica.

$$exactitud(y, y\hat{}) = \sum_{i=0}^{n-1} 1(y_i\hat{} = y_i)$$

Precisión

Es la capacidad del clasificador de no asignar falsos positivos. En otras palabras, mide qué tan bueno es el modelo para encontrar verdaderos positivos.

$$Precisión = \frac{TP}{TP + FP}$$

TP → Verdaderos positivos.

FP → Falsos positivos.

La exactitud toma en cuenta todo tipo de error, mientras que la precisión da más importancia al error por falso positivo.

Recall (Sensibilidad)

Ésta nos indica la habilidad del modelo para detectar todos los verdaderos positivos.

$$Recall = \frac{TP}{TP + FN}$$

Puntuación F1

La puntuación F1 combina la precisión y el recall del modelo predictivo. Es la media armónica de la precisión y el recall.

EN PRODUCCIÓN - GALERADAS

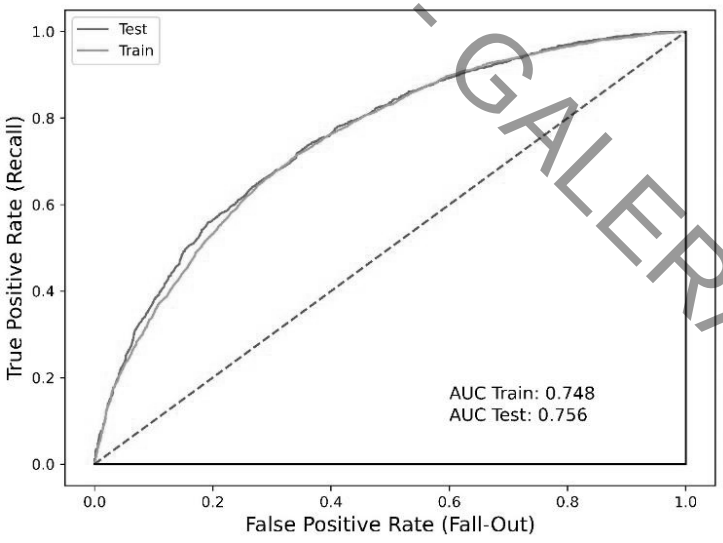
$$F1_{score} = 2 \cdot \frac{P \cdot R}{P + R}$$

La puntuación F1 es útil al ser el promedio de la precisión y recall, además, al utilizar la media armónica, es útil cuando se trabaja con clases desbalanceadas que suele ser un problema bastante frecuente para estimar la mortalidad.

Métricas asociadas a nuestro modelo:

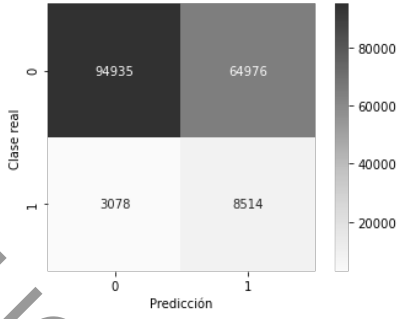
	Accuracy	*Balanced accuracy*	*F1-Score*	*AUROC*	*Precision*	*Recall*
Balanced Logistic Regression	**0,683676**	**0,68481**	**0,699195**	**0,75353**	**0,656211**	**0,748407**

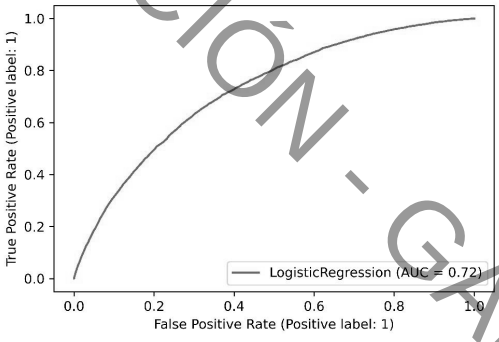
Curva ROC asociada al modelo



EN PRODUCCIÓN - GALERADAS

A continuación, vamos a validar nuestro modelo en una nueva base de datos de otros años 2008-2012. En este caso lo validamos en el subconjunto completo desbalanceado.





	Accuracy	*Balanced accuracy*	*F1-Score*	*AUROC*	*Precision*	*Recall*
Balanced Logistic Regression	**0,683676**	**0,68481**	**0,699195**	**0,75353**	**0,656211**	**0,748407**
	0,629493	**0,670936**	**0,207780**	**0,732073**	**0,121446**	**0,718857**

EN PRODUCCIÓN - GALERADAS

CONCLUSIÓN

El resultado del modelo es una herramienta útil que aporta información una vez realizado el diagnóstico de ictus isquémico (no tratado con fibrinolisis). Nuestro modelo es capaz de detectar la posibilidad de exitus de cada paciente y basándonos en él podemos crear escalas, *scores* o calculadoras que nos permitirán estimar la probabilidad de exitus de cada paciente al primo ingreso tras un ictus isquémico.

El modelo resultante tiene algunos inconvenientes entre los cuales destacamos la precisión del mismo cuando empleamos datos reales, y es que, en este contexto, no predice adecuadamente las personas que no van a fallecer.

Otro de los inconvenientes que hemos observado es la precisión global del modelo. Son necesarios nuevos estudios que incidan en el problema o en los que se valore la recogida de nuevas variables que sean de utilidad a la hora de predecir la probabilidad de exitus tras el primer ingreso.

Además el resultado de nuestro modelo también podría mejorar si la variable a predecir no fuera la mortalidad, ya que muchos pacientes sobreviven tras padecer un ictus, pero un gran porcentaje de ellos presentará secuelas irreversibles.

EN PRODUCCIÓN - GALERADAS

BIBLIOGRAFÍA

1. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken, NJ: Wiley; 2013. 528 p. ISBN: 978-0-470-58247-3.
2. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput. 2011;21:137-146.
3. Núñez E, Steyerberg EW, Núñez J. Regression Modeling Strategies. Rev Esp Cardiol (Engl Ed). 2011;64(6):501-507.
4. Allison P. Why I Don't Trust the Hosmer-Lemeshow Test for Logistic Regression. Statistical Horizon. 2013.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN GALERADAS

10. MODELO PREDICTIVO BASADO EN TÉCNICAS DE ÁRBOLES DE CLASIFICACIÓN PARA ESTIMAR LA MORTALIDAD POR ICTUS

AUTORES: FERNANDO RECHE LORITE. JUAN MANUEL GARCÍA TORRECILLAS.
LUIS TÉLLEZ RAMÍREZ.

EN PRODUCCIÓN - GALERADAS

1. INTRODUCCIÓN

La elaboración de buenos modelos predictivos basados en la información proporcionada por bases de datos es una tarea de suma importancia pues puede servirnos como herramienta de apoyo a la hora de tomar decisiones.

El ser humano tiene serios problemas a la hora de analizar la información proveniente de un número elevado de variables. Los ordenadores, con su enorme capacidad de cálculo, son capaces de manejar esa gran cantidad de información sin demasiados problemas. La cuestión fundamental es: ¿cómo podemos implementar un método eficaz para tal tarea?

Actualmente existen una gran cantidad de herramientas estadísticas con las que podemos implementar modelos predictivos, cada una de ellas con sus ventajas y sus inconvenientes. En el caso que nos concierne, nuestro objetivo principal consiste en presentar una estimación objetiva del riesgo de mortalidad que tiene una persona en el caso de sufrir un ictus. Así pues, la variable de interés que estamos considerando es de tipo binario pues solamente puede tomar dos valores: supervivencia o mortalidad.

A la hora de construir un modelo predictivo debemos plantearnos dos cuestiones básicas. En primer lugar, hemos de detectar cuales son las variables que influyen y en qué grado en la mortalidad del individuo que ha sufrido un ictus y, en segundo lugar,

EN PRODUCCIÓN - GALERADAS

hemos de encontrar la metodología adecuada que nos permita construir un modelo que pueda predecir la posibilidad que tiene el paciente de fallecer a partir de la observación de las variables del modelo.

La metodología de los árboles de clasificación permite realizar ambas tareas simultáneamente. Por un lado, selecciona las variables que más influyen en la variable objetivo y, a la vez, construye un modelo predictivo para estimar los valores de dicha variable a partir de los valores observados de las variables seleccionadas.

1.1. MÉTODOS DE CLASIFICACIÓN

Nuestro modelo predictivo, dado que la variable objetivo es de tipo categórico —en nuestro caso con dos niveles—, se basa en un método de clasificación. De forma general, los métodos de clasificación se pueden dividir en dos tipos: los métodos supervisados y los no supervisados.

En las metodologías supervisadas, se parte de una base de datos en la que se conoce el valor de la variable objetivo —en nuestro caso, la supervivencia del individuo al sufrir un ictus— y los valores del resto de las variables. A partir de esos datos se construye un modelo —proceso que se conoce como fase de entrenamiento— y, una vez validado, se utiliza para predecir los valores de la variable objetivo en un nuevo individuo al que se le han observado los valores del resto de las variables del modelo.

En los métodos no supervisados no se conoce a priori a qué grupo pertenece el individuo y, a partir de los datos observados, hemos de construir agrupaciones que segmenten nuestra muestra.

Para realizar clasificación supervisada podemos utilizar, por ejemplo, regresiones logísticas, análisis discriminante, redes neu-

EN PRODUCCIÓN - ALTERADAS

ronales, modelos gráficos probabilísticos o árboles de clasificación.

En este capítulo vamos a explicar los fundamentos básicos de los árboles de clasificación, metodología que hemos considerado interesante para construir un modelo predictivo para estimar la mortalidad en el caso de sufrir un ictus por los siguientes motivos:

- Selecciona variables a la vez que construye el modelo predictivo.
- Es una técnica no paramétrica, es decir, no presupone que se verifique ningún modelo.
- Maneja eficientemente relaciones no lineales.
- Permite trabajar con datos faltantes.
- Es robusto frente a datos anómalos.
- Trabaja simultáneamente con variables de todo tipo —numéricas y categóricas— sin necesidad de transformar los datos.
- Permite detectar las variables más importantes del modelo.
- Y, sobre todo, es un modelo de muy fácil interpretación.

A nivel de clasificación supervisada, la metodología basada en redes neuronales está muy en boga. Es cierto que es un modelo que funciona muy bien a la hora de clasificar, pero tiene el inconveniente de ser una caja negra, es decir, no nos informa de cómo se realiza el proceso de clasificación ni cuales son las variables más importantes a la hora de realizar esta tarea, por lo que es necesario una selección inicial de variables para eliminar el posible ruido que haya en la base de datos.

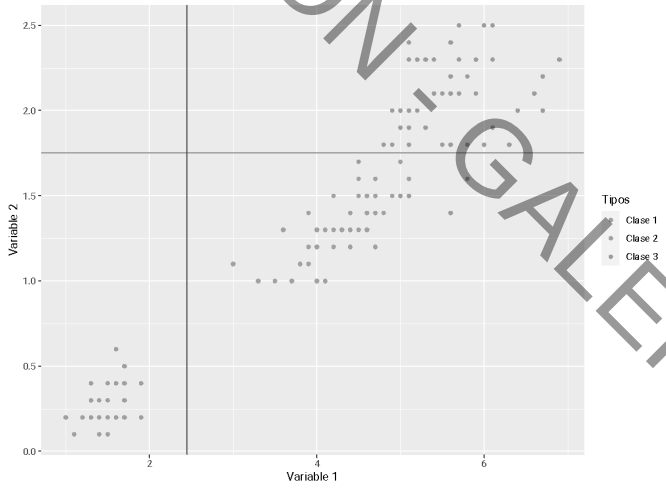
EN PRODUCCIÓN - GALERADAS

2. ÁRBOLES DE CLASIFICACIÓN

2.1. IDEAS BÁSICAS

¿Qué es un árbol de clasificación? Desde un punto de vista técnico, un árbol de clasificación es un modelo que pretende representar el conocimiento mediante una partición recursiva del espacio muestral.

Esto que parece una definición algo complicada lo podemos ver con claridad de forma gráfica con el siguiente ejemplo.



En la Figura 1 podemos ver un diagrama de puntos representando el comportamiento de dos variables a las que se le añade la información de la variable objetivo coloreando los puntos y que nos clasifica las observaciones en tres clases.

EN PRODUCCIÓN - GALERADAS

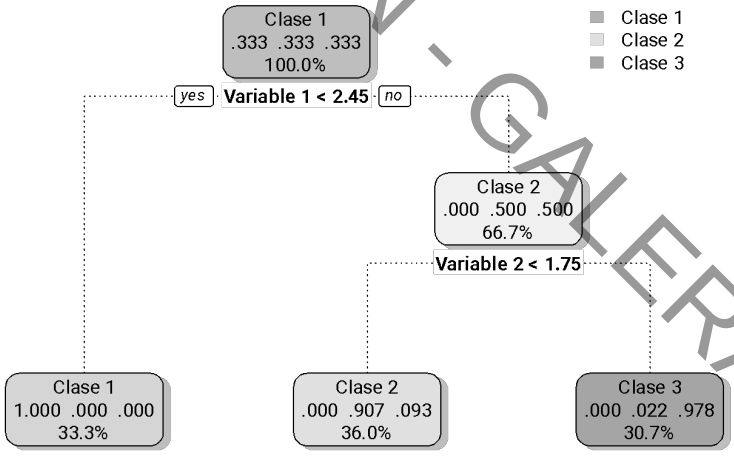
Con dos líneas —una azul y la otra roja— hemos realizado una partición del espacio muestral.

En este caso es sencillo pues solamente estamos presentando la información de dos variables.

En primer lugar, podemos ver que mediante la línea azul separamos claramente los individuos de la Clase 1 (en el gráfico aparecen en color rojo) solamente utilizando la Variable 1 seleccionando el valor 2,45.

En un segundo paso, considerando el valor de la Variable 2 en 1,75, separamos los individuos de la Clase 2 (en verde) de los de la Clase 3 (en azul). Como se puede observar, hay algunos individuos mal clasificados ya que algunos individuos azules están en el grupo de los verdes.

En un método de clasificación es muy complicado —y en muchos casos, no es deseable— una clasificación perfecta, por lo que la tasa de acierto puede ser un buen indicador —aunque no es el único posible— de la bondad de nuestro método.



Podemos ver en la Figura 2 la representación de esa información de forma gráfica mediante un árbol de clasificación.

EN PRODUCCIÓN - GALERADAS

Como podemos observar, en el nodo raíz (inicial) hay un tercio de los individuos en cada clase puesto que en la base de datos de entrenamiento hay 50 individuos de cada clase.

En el primer paso se ha realizado la pregunta de si los individuos tienen un valor inferior a 2,45 en la Variable 1.

Si la respuesta es Sí (yes), nos desplazamos por la rama de la izquierda y se etiquetan esos individuos como pertenecientes a la Clase 1 puesto que es la clase.

Como podemos observar en la Figura 2, esta pregunta hace que obtengamos una clasificación perfecta pues en el nodo resultante hay probabilidad 1 de que los individuos que cumplan esa condición pertenezcan a la Clase 1.

Si la respuesta es negativa (No), nos desplazamos por la rama de la derecha. En este caso, quedan un 50% de la Clase 2 y otro 50% de la Clase 3. En este camino hacemos el siguiente paso, que consiste en preguntar si los individuos tienen un valor inferior a 1.75 en la Variable 2.

El argumento es el mismo de la pregunta anterior. Si la respuesta es Sí (yes) las observaciones se etiquetan como de la Clase 2. En este caso la clasificación no es perfecta, pues hay un 90.7% de casos pertenecientes a la Clase 2 y un 9.3% de la Clase 3. Y así sucesivamente.

Como se puede observar, se trata de un proceso jerárquico ordenado en el que hay una sucesión de preguntas de tipo binario —si se cumple una condición o no— que nos lleva a un modelo de clasificación muy sencillo de interpretar.

La elección de las variables, el orden en el que se hacen las preguntas y el tipo de preguntas viene dada por la metodología de los árboles propuesta por Breiman et al. [1] que a continuación vamos a explicar con un cierto detalle.

EN PRODUCCIÓN - GALERADAS

2.2. GRAFOS

Un grafo es una representación gráfica de un conjunto de objetos, denominados nodos o vértices, unidos por unos enlaces y que se conocen como arcos o aristas.

Mediante un grafo se pueden representar multitud de situaciones reales, tales como planos del metro, estructura de una red informática, conexiones entre individuos en una red social o las relaciones entre diferentes variables.

En nuestro caso, esta representación gráfica la vamos a utilizar para establecer relaciones entre las variables involucradas en un problema concreto: la mortalidad acaecida después de sufrir un ictus.

Existen muchos tipos de grafos y hay una rama de la matemática intensamente dedicada a su estudio y comportamiento pues se trata de una herramienta extremadamente útil para modelar muchas situaciones reales. En nuestro caso, la estructura que vamos a utilizar es un grafo de *tipo árbol*.

Aunque la definición exacta es algo más exhaustiva, un árbol es un grafo en el que cualquier par de vértices están conectados exactamente por un camino. El árbol que nosotros vamos a utilizar es dirigido, es decir, los enlaces entre nodos tienen un sentido.

EN PRODUCCIÓN-GALERADAS

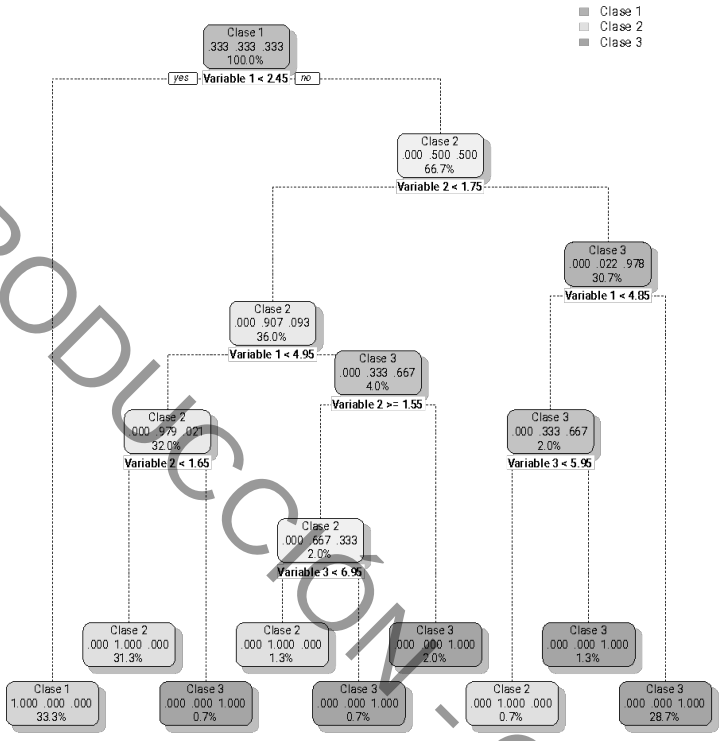


*Figura SEQ Figura * ARABIC 2: Árbol de clasificación asociado a la partición de la Figura 1*

Además, el tipo de árbol que se utiliza en el modelo de árboles de clasificación es binario, es decir, de cada nodo solo pueden salir dos enlaces.

En la Figura 3 podemos ver una estructura de árbol binario dirigido. El nodo superior se denomina *nodo raíz* y, como se puede observar, de él parten dos enlaces. Así sucesivamente se va construyendo todo el árbol hasta llegar a los nodos finales que se denominan *nodos hojas*.

EN PRODUCCIÓN - GALERADAS

2.3. EL PROCESO DE APRENDIZAJE

Desde un punto de vista conceptual, un árbol de clasificación es una forma de representar el conocimiento mediante una estructura resultante de una partición recursiva de un conjunto de observaciones.

Es decir, partimos de un conjunto de entrenamiento en el que conocemos los valores de las variables que intervienen en el problema y, además, a qué clase pertenece cada individuo.

El proceso recursivo consiste en que en cada paso se realiza una pregunta sobre una variable que se representa en un nodo interior. Esa pregunta es de tipo binario, tiene una respuesta afirmativa o negativa. En el caso de la Figura 3, en el nodo raíz preguntamos: ¿toma la Vatiable 1 un valor inferior a 2,45? Si una observación verifica esta condición la «mandaremos» al nodo de la izquierda y si no la verifica, a la derecha.

Posteriormente, decidiremos si seguimos partiendo o no. Cuando se decida parar el proceso, los nodos finales —hojas— nos dirán a qué clase pertenecen los individuos alojados en dicho nodo.

Algunas de las cuestiones que nos tenemos que plantear en este proceso son las siguientes:

- ¿Cómo se hacen las particiones?
- ¿Cómo se selecciona la mejor en cada momento?
- ¿Qué criterio se adopta para medir la homogeneidad de un nodo?
- ¿Cuándo paramos el proceso de división?
- ¿Cómo se asigna una etiqueta a un nodo?

El objetivo fundamental a la hora de construir un árbol de clasificación es que en cada paso se mejore la *homogeneidad* de los nodos resultantes. Por ejemplo, en la Figura 3 podemos observar que todas las observaciones que tienen un valor de la Variable 1

EN PRODUCCIÓN - GALERADAS

inferior a 2,45 pertenecen a la Clase 1, por lo que ese nodo es totalmente homogéneo y no necesitamos seguir partiendo. Sin embargo, las observaciones que no cumplen esta condición, la mitad son de la Clase 2 y la otra mitad de la Clase 3, por lo tanto, parece conveniente —como así se hace— continuar realizando particiones.

2.4. LAS PARTICIONES

Veamos ahora cómo se realizan las particiones. En primer lugar, establezcamos el tipo de preguntas que se realizan.

Si disponemos de una variable de tipo categórico X_i que toma los valores $X_i = \{x_1, x_2, \dots, x_k\}$ las preguntas son de tipo $\{X_i \in C\}$ siendo $C \subset \{x_1, x_2, \dots, x_k\}$.

Si la variable es de tipo continuo las preguntas son del tipo $\{X_i \leq k\}$, donde $k \in R$ que, aunque teóricamente puede ser un valor real arbitrario, se elige el punto medio de dos valores consecutivos de la variable X_i, hecho que simplifica los cálculos y aumenta la eficiencia del proceso.

En consecuencia, podemos manejar simultáneamente variables de tipo numérico y categórico sin necesidad de realizar ningún proceso adicional.

Si se verifica la condición el caso se adjudica al nodo de la izquierda y si no la verifica, se adjudica al nodo de la derecha.

Para decidir qué variables se seleccionan y qué preguntas se realizan nos basaremos en una *medida de impureza* para construir una medida de homogeneidad de una partición. Así el ordenador realizará todas las posibles preguntas de todas las variables y elegirá la variable con la pregunta que produzca la mayor homogeneidad en la partición resultante.

EN PRODUCCIÓN - GALERADAS

La metodología que vamos a utilizar para la construcción de los árboles de clasificación es la propuesta por Breiman et al [1] denominada CART (*Classification And Regression Trees*). Existen otros algoritmos (Chaid, C5.0, Conditional Trees, etc) que se diferencian esencialmente del CART en la forma de medir la homogeneidad y en la eficiencia del algoritmo.

2.5. MEDIDA DE IMPUREZA Y DE HOMOGENEIDAD DE LAS PARTICIONES

Una medida de impureza ϕ definida sobre un conjunto de valores $\{c_1, c_2, \dots, c_k\}$ que cumplen que:

- Son todos no negativos, es decir, $c_i \geq 0$ para todos los $i = 1,2, \dots, k$.
- Suman uno, $c_1 + c_2 + \cdots + c_k = 1$.

Además, tiene que verificar las siguientes tres propiedades:

1. ϕ tiene un único máximo en $\left(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k}\right)$.
2. ϕ alcanza su mínimo si $c_i = 1$ para un único i y $c_j = 0$ para cualquier $j \neq i$, y dicho valor mínimo es cero.
3. ϕ es una función simétrica de $c_1, c_2, \dots, c_k$.

Las medidas de impureza que se aplican en la metodología CART son la *medida de entropía* y el *índice de Gini*.

La medida de entropía de un nodo t se define como $i(t) = -\sum \quad p(t)\, log\, log\, (j|t)$ y el índice de Gini como

EN PRODUCCIÓN - GALERADAS

$i(t) = 1 - \sum \quad p(t)^2$, donde $p(t)$ es la probabilidad de que un caso del nodo t sea de la clase j.

Dada una medida de impureza, el siguiente paso consiste en construir la definición de una medida de la *bondad* de una partición *s* considerando la impureza el nodo raíz t y la de los nodos resultantes de la partición t_L y t_R. Esta medida de bondad se define de la siguiente forma:

$$\Phi(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Siendo *s* la partición que se realiza y *t* el nodo raíz. Entonces, se denota por p_L a la probabilidad de que un elemento forme parte del nodo de la izquierda t_L y por p_R a la probabilidad de que forme parte del nodo de la derecha t_R.

Veamos cómo se calcula la bondad de una partición con un ejemplo sencillo:

Partimos de un nodo t con 9 elementos, 6 de la clase *A* y 3 de la clase *B* y hemos hecho una partición s que nos ha proporcionado dos nodos, t_L con 4 elementos de los 9 y t_R con los 5 restantes con la composición de las clases resultantes que aparece en la Figura 4:

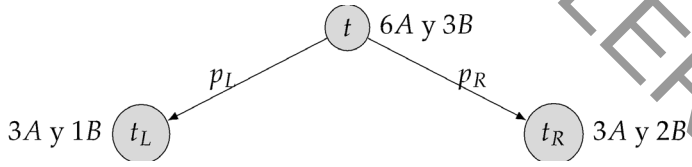


Figura 4: *Partición s del nodo t*

En este caso $p_L = \frac{4}{9}$ y $p_R = \frac{5}{9}$, ya que de los 9 elementos que hay en el nodo *t*, 4 se adjudican al nodo t_L y 5 al nodo p_R.

EN PRODUCCIÓN - GALERADAS

Las medidas de impureza de los nodos de la Figura 4, utilizando la medida de la entropía son:

- $i(t) = -(p(t) \cdot log\ log\ p(t) + p(t)p(t) = -\left(\frac{6}{9} \cdot log\ log\ \frac{6}{9} + \frac{3}{9} \cdot log\ log\ \frac{3}{9}\right) = 0.27$
- $i(t_L) = -(p(t_L) \cdot log\ log\ p(t_L) + p(t_L)p(t_L) = -\left(\frac{3}{4} \cdot log\ log\ \frac{3}{4} + \frac{1}{4} \cdot log\ log\ \frac{1}{4}\right) = 0.224$
- $i(t_R) = -(p(t_R) \cdot log\ log\ p(t_R) + p(t_R)p(t_R) = -\left(\frac{3}{5} \cdot log\ log\ \frac{3}{5} + \frac{2}{5} \cdot log\ log\ \frac{2}{5}\right) = 0.29$

Por lo tanto, la bondad de esta partición *s* cuyo nodo raíz es *t* será:

$$\Phi(s,t) = 0.27 - \frac{4}{9}0.244 - \frac{5}{9}0.29 = 0.014.$$

Ésta es la medida que el método CART utiliza para decidir cuál será la partición elegida, es decir, se realizan todas las posibles preguntas y se calcula la bondad de todas las particiones generadas por dichas preguntas y, en consecuencia, se elige la que tiene mayor índice de bondad.

2.6. EL PROBLEMA DEL SOBREAPRENDIZAJE

Una vez que hemos decidido cuál es la medida que nos determina la elección de las variables y de las preguntas a realizar la siguiente cuestión que nos planteamos es en qué momento debemos parar el proceso. ¿Por qué?

En teoría podríamos estar partiendo el espacio de las observaciones hasta que todos los nodos hoja fueran puros —con una clasificación perfecta—. Si seguimos esta estrategia indefinidamente, nos encontramos con una situación poco deseable: el *sobreaprendizaje* del modelo.

¿En qué consiste el sobreaprendizaje de un modelo?

EN PRODUCCIÓN - GALERADAS

Si dividimos hasta obtener nodos hoja puros le estamos creando un «traje a medida» al conjunto de entrenamiento, es decir, el modelo se ajusta muy bien a los datos con los que lo hemos construido, pero el modelo no es capaz de clasificar correctamente cuando aparecen datos nuevos. Por decirlo de otra forma, utilizando un símil cotidiano, un traje a medida se ajusta perfectamente a la persona que lo ha encargado, pero es muy difícil que le quede bien a nadie más.

Por lo tanto, hemos de establecer alguna estrategia para evitar el fenómeno del sobreaprendizaje en la construcción de nuestro modelo.

Para resolver este problema recurrimos en los árboles de clasificación a un proceso de *poda*. Para ello se construye una sucesión ordenada de árboles en las que se han ido eliminando «ramas» y se elige en el que se comete menos error de clasificación. Una forma de medir el error puede realizarse calculando la proporción de fallos de clasificación, es decir, la proporción de elementos mal clasificados.

Al hablar de error de clasificación hemos de ir con cuidado ya que cuanto más grande sea el árbol —más particiones se hayan realizado— menos error se cometerá, pues los nodos hoja serán cada vez más puros si se sigue dividiendo el árbol.

Para equilibrar error con complejidad del árbol se introduce un parámetro de *coste-complejidad* que penaliza el árbol proporcionalmente al número de hojas que éste tenga. Así si la disminución del error de clasificación no compensa la complejidad del árbol y la medida de error aumentará en lugar de disminuir.

Este parámetro de coste-complejidad puede ser fijado por el usuario en función de la observación de los valores de error de clasificación obtenidos en la secuencia de árboles construidos.

Pero ¿cómo se mide este error?

No parece razonable utilizar los mismos datos que hemos usado en el proceso de construcción del modelo para medir el error de clasificación ya que estaríamos subestimando el error real. Para esta tarea podemos utilizar dos estrategias:

EN PRODUCCIÓN - CALERADAS

- Utilizar como conjunto de prueba una parte de la muestra, es decir, se divide la muestra en dos partes. Con la primera se aprende el modelo y con el resto compruebo el error cometido. El problema principal de esta estrategia es que se necesitan bases de datos de un tamaño razonable para poder llevarlo a cabo.
- Otra estrategia consiste en utilizar una metodología de *validación cruzada*. En este caso se divide aleatoriamente la muestra en partes —por ejemplo, podríamos considerar 10 partes—. Se usan 9 de ellas para entrenar el modelo y la restante para calcular el error. Posteriormente se repite este proceso con todas las partes, es decir, todas pasan por el proceso de prueba. Este método evita sesgos y no se necesita tanta muestra.

2.7. LA PODA

Una vez que hemos establecido una estrategia para medir el error debemos elaborar un procedimiento para evitar el sobreaprendizaje del modelo.

Uno de estos procedimientos es el que se denomina —siguiendo con la terminología agrícola— el proceso de *poda* del árbol.

Esta metodología, se basa en la siguiente idea: a partir de un árbol maximal —con el mayor número de subdivisiones posibles— se va construyendo una sucesión de árboles podando las ramas que produzcan pequeños beneficios de bondad.

Desde un punto de vista operativo la secuencia de árboles utilizando los siguientes elementos:

- Denominemos por $|\widetilde{T}|$ el número de hojas que tiene el árbol T.
- Calculamos el error de clasificación $R(T)$.
- La medida de coste-complejidad viene dada, para cada α, por $R_\alpha(T) = R(T) + \alpha|\tilde{T}|$. Con esta medida penalizamos

EN PRODUCCIÓN - GALERADAS

árboles complejos ya que se incrementa el error si el árbol tiene muchas hojas.

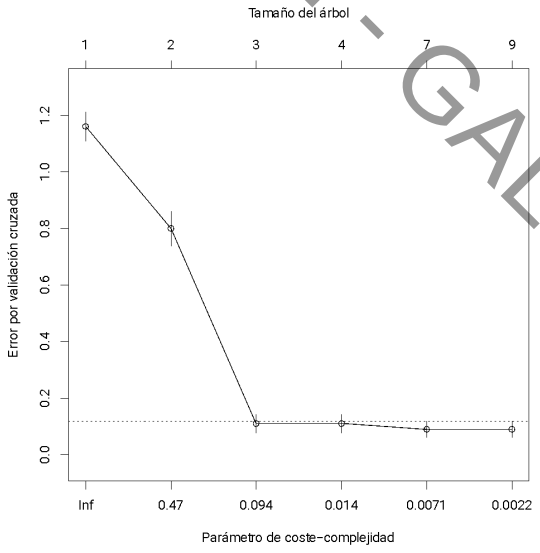
- Para cada α se busca el árbol T_α que minimiza la medida de coste complejidad R_α

Con esta estrategia se construye una secuencia de árboles variando el valor del parámetro de coste-complejidad. De esta forma se construye una secuencia ordenada de árboles

$$T_{max} \succ T_1 \succ \cdots \succ \{t_1\}.$$

De todos estos árboles elegiremos aquel que menos medida de coste-complejidad tenga o, en su defecto, cuando veamos que la ganancia de clasificación se obtiene no merezca la pena.

En la Figura 5 podemos ver de forma gráfica la evolución del error de clasificación según varía de mayor a menor el valor de parámetro de coste-complejidad α asociado al árbol de clasificación maximal que aparece en la Figura 3.

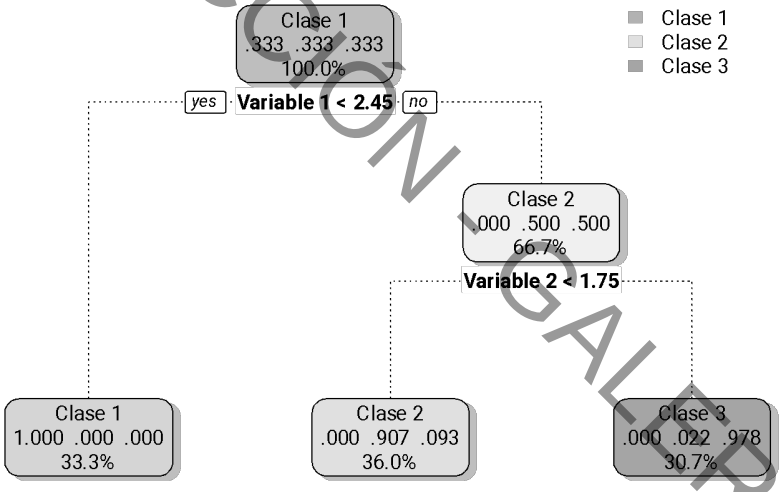


EN PRODUCCIÓN - GALERADAS

Cuando partimos de un valor muy grande tenemos solamente el nodo raíz. Al ir disminuyendo el valor vamos aumentando la complejidad del árbol y, por lo tanto, reduciendo el error cometido. Al estar este error penalizado por el número de hojas, hay un momento en el que dicho error se estabiliza y ya no merece la pena seguir realizando particiones.

Como se puede observar en la Figura 5, hay una fuerte caída inicial del error en los primeros árboles —de tamaño 1, 2 y 3— para, posteriormente estabilizarse, por lo que sería razonable elegir un parámetro de coste complejidad de 0.094.

En este caso el árbol podado resultante al ajustar el parámetro de coste-complejidad es el que aparece en la Figura 6.



Como podemos observar, este árbol tiene tres hojas, a diferencia del que aparecía en la Figura 3, que tenía 9.

En este caso los nodos hojas no son puros pero el error de clasificación cometido es muy bajo.

De esta forma se evita el sobreaprendizaje del modelo, con una estructura más sencilla y con una tasa de error aceptable.

EN PRODUCCIÓN - GALERADAS

3. APLICACIÓN DE LOS ÁRBOLES DE CLASIFICACIÓN A UN MODELO DE MORTALIDAD POR ICTUS

Una vez vista la metodología e interpretación de los árboles de clasificación para la construcción de modelos predictivos vamos a aplicarlo a un caso real.

Como cualquier caso real, en la construcción de un modelo predictivo surgen cuestiones que pueden perturbar la calidad del modelo. Por ello es necesario realizar previamente un estudio descriptivo completo para detectar posibles problemáticas que puedan afectar a nuestro modelo.

Sin entrar en muchos detalles, la base de datos utilizada para el estudio está formada por 171503 observaciones de pacientes que han sufrido un ictus durante el periodo 2008-2012. Una vez depurada dicha base de datos eliminando todo el ruido posible, se han considerado 13 variables, entre ellas la supervivencia del paciente que ha sufrido el ictus.

El primer problema con en el que nos encontramos —desde un punto de vista estadístico, no clínico— es el desequilibrio que hay en la variable de supervivencia, pues solamente fallecen 11592 individuos —el 6,76% de los pacientes que ingresan con un ictus— frente al 93,24% que sobreviven. Este hecho hace que en modelo predictivo la tasa de acierto sea muy grande, pero esencialmente debido a que el modelo se equivoca poco en al

EN PRODUCCIÓN - GALERADAS

predecir la supervivencia y hace un poco irrelevante los fallos de predicción en la clase minoritaria —la supervivencia—.

Este no es un problema menor que no tiene una solución sencilla. Para soslayar esta situación hemos seguido la siguiente estrategia a la hora de construir nuestro modelo predictivo. Se ha considerado a todos los pacientes que han fallecido y hemos seleccionado aleatoriamente la misma cantidad de pacientes entre el grupo de los que han sobrevivido. De esta manera obtenemos clases equilibradas. No es una solución óptima, pero nos va servir para construir un modelo predictivo razonable.

El proceso de construcción, una vez seleccionados los elementos de la muestra y que parte de 23184 observaciones —50% de fallecidos y 50% de supervivientes—, es el siguiente:

- Dividimos esta muestra en dos grupos: El 80% para entrenar el modelo y el 20% para ver la capacidad predictiva que éste tiene.
- Con la muestra de entrenamiento construimos un árbol de clasificación en el que se ha ajustado el parámetro de coste-complejidad para evitar el sobreaprendizaje.

El árbol aprendido, que podemos ver en la Figura 7, aparecen por orden de importancia las variables que mejor discriminan a los pacientes de la muestra.

EN PRODUCCIÓN - GALERADAS

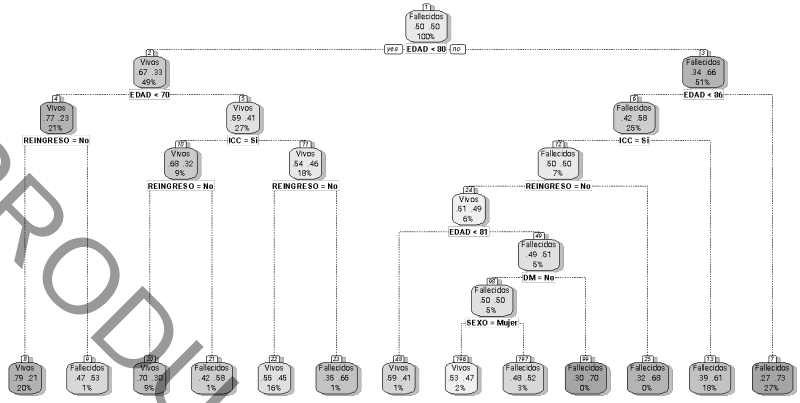


Figura 7: *Modelo predictivo basado en árboles de clasificación*

Como se puede observar, la primera variable que aparece en el modelo —nodo raíz— es la Edad —con un corte en 80 años—. Como se ha mencionado en el apartado de metodología, el árbol debemos interpretarlo en función de la pregunta que se hace en función de la variable que aparece en el nodo.

A modo de ejemplo, utilicemos la pregunta de este nodo: ¿tiene el paciente una edad inferior a 80 años?

Si la respuesta es Sí, nos desplazamos por la rama de la izquierda. En tal caso, se hace otra pregunta sobre la Edad: ¿es inferior a 70? Si es que Sí, se pregunta por el Reingreso: ¿No reingresa? Si la respuesta es que Sí, llegamos al nodo hoja final donde la probabilidad de sobrevivir es del 79%.

Resumiendo todo este razonamiento concluimos que para los pacientes de menos que 70 años que no reingresan, la probabilidad de supervivencia es de casi un 80%, mientras que si reingresan esta probabilidad se reduce hasta un 47%.

En los nodos hojas podemos ver las probabilidades de supervivencia que nos proporcionan las variables «padres» y, por lo tanto, podemos hacer perfiles de supervivencia.

EN PRODUCCIÓN - GALERADAS

Como se puede observar de la estructura del árbol, solamente aparecen cinco variables: la Edad, el Reingreso, el ICC (insuficiencia cardíaca), DM y el Sexo. El resto de las variables de la base de datos son menos informativas que éstas a la hora de establecer la probabilidad de supervivencia de los pacientes.

Una forma de ver el comportamiento de este modelo predictivo consiste en predecir la supervivencia del 20% de los pacientes considerados como conjunto de prueba comparándolos con su situación real.

En nuestro caso los datos son los siguientes:

Predicción\ Observación	**Vivos**	**Fallecidos**	**Total**
Vivos	1541 (68.4%)	713 (31.6%)	2254 (100%)
Fallecidos	791(33.2%)	1592 (66.8%)	2383 (100%)
Total	2332 (50.3%)	2305 (49.7%)	4637 (100%)

Tabla 1: *Predicciones del modelo frente a observaciones reales*

A la vista de la Tabla 1 podemos ver que hay un 67,57% de acierto en las predicciones del modelo —este valor se calcula dividiendo en número de aciertos entre el total de observaciones, multiplicado por 100—, que podemos considerar mejorable pero que es una tasa de acierto razonable, máxime teniendo en cuenta el equilibrio de aciertos en ambas clases —supervivencia y fallecimiento—.

EN PRODUCCION - GALERADAS

4. CONCLUSIONES

En este trabajo hemos presentado brevemente la metodología de los árboles de clasificación aplicándola a un problema real.

Como método de clasificación, se trata de una metodología potente que, además, permite seleccionar las variables más importantes y establecer una jerarquía de importancia entre ellas a la hora de clasificar observaciones.

Los árboles de clasificación permiten manejar variables de diferentes tipologías sin necesidad de un tratamiento adicional y son muy robustos frente a los datos atípicos —a diferencia de otros métodos predictivos como son los modelos de regresión—. Además, la existencia de valores faltantes no es un problema para la construcción del modelo, siempre que el porcentaje de éstos en la muestra sea razonable.

La aplicación de estos modelos en el ámbito sanitario, complementados con otras metodologías, puede permitir construir herramientas que ayuden a los profesionales a facilitar el proceso de toma de decisiones.

Contribución de los autores: Fernando Reche Lorite y Luis Téllez Ramirez han desarrollado la parte metodológica y analítica. Juan Manuel García Torrecillas ha proporcionado la base de datos del estudio y se ha encargado de la parte clínica del estudio.

Agradecimientos: Fernando Reche Lorite quiere agradecer el apoyo del centro de investigación CDTIME de la Universidad de Almería por su apoyo en la realización de este trabajo.

EN PRODUCCIÓN - GALERADAS

5. BIBLIOGRAFÍA

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J Stone. *Classification and Regression Trees*. Wadsworth, Belmont, Ca, 1983.

L. A. Clark and D. Pregibon. Tree-based models. In J.M. Chambers and T.J. Hastie, editors, *Statistical Models in S*, chapter 9. Wadsworth and Brooks/Cole, Pacific Grove, Ca, 1992.

T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Second Edition. Springer, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2014. http://www.R-project.org.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

11. LA VALIDACIÓN DE LOS MODELOS PREDICTIVOS A LA LUZ DE LA ESTADÍSTICA Y LA CIENCIA DE DATOS.

AUTORES: LUIS TÉLLEZ RAMÍREZ. FERNANDO RECHE LORITE. CRISTINA ORTIZ PUERTAS. JUAN MANUEL GARCÍA TORRECILLAS.

Resumen del capítulo: Este capítulo tiene por objetivo hacer una síntesis de lo que implica validar un modelo para poder ponerlo en producción. Para ello, se verán primero algunas dificultades que pueden aparecer a la hora de entrenar un modelo. Se hablará de las técnicas más importantes para realizar dicha validación, y se propondrá una metodología propia para hacerlo. Por último, se estudiará la evaluación de clasificadores y se propondrán algunas claves del capítulo a modo de conclusión.

Para la lectura de este capítulo se sobreentiende que el lector está familiarizado con términos como modelo, aprendizaje supervisado, aprendizaje no supervisado, clasificación, regresión, y que más o menos controla algo relacionado a software estadísticos: SPSS, Stata, R, Python, Julia, etc. No obstante, este capítulo tiene un carácter técnico-divulgativo y lo que se busca es dar a conocer el proceso seguido y no tanto la aportación de material demasiado técnico (para ello, consultar bibliografía).

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

1. VALIDACIÓN DE MODELOS

Se entenderá por validación de modelos, un conjunto de procesos y técnicas cuyo objetivo es verificar que el modelo se comporta como esperamos. Esto es más importante de lo que a priori podría pensarse, ya que, si este paso no se realiza correctamente, el modelo podría ser más optimista o pesimista de lo que debiera ser (presenta sesgo u otros problemas que veremos más adelante). En definitiva, el modelo no se comporta como se espera que lo haga.

Se definen ahora dos conceptos que ayudarán a fijar los objetivos:

- **Generalización**: Es la habilidad de un modelo de mantener su capacidad de predicción en entrenamiento sobre nuevos datos.
- Un modelo de clasificación/regresión es **válido** en la medida en la que es capaz de generalizar a nuevos datos.
- Para este proceso, es clave tener presente el principio de parsimonia, que reza:
  "En igualdad de condiciones, la explicación más sencilla suele ser la más probable."

EN PRODUCCIÓN - GALERADAS

1.1. ¿POR QUÉ ES IMPORTANTE?

El objetivo de un modelo no es otro que entender la naturaleza de los datos y hacer predicciones sobre ellos. La validación determina si el modelo entrenado es fiable. Además, la validación de modelos permite reducir los costes, descubrir más errores, aumentar la escalabilidad, flexibilidad y calidad del modelo.

1.2. ENUNCIADO

Dado un conjunto de datos ***D*** y un algoritmo de aprendizaje A, si se aprende un modelo M:

1. ¿Qué garantía se tiene para ese modelo cuando se aplique a nuevos datos?
2. Dados dos algoritmos distintos (A_1, A_2) y un mismo conjunto de datos ***D,*** ¿Qué algoritmo es mejor?
3. Dado un algoritmo A con una serie de parámetros $p_1, \ldots, p_n$ ¿Cuál es la mejor parametrización dado un conjunto de datos ***D***? ¿Y para generalizar?

La validación de modelos viene a responder estas preguntas. Para ello, lo primero que se necesita es definir una medida (f) de rendimiento que nos diga cómo de bueno es un modelo respecto de los datos. Y lo más importante, es necesario estimar ese valor de forma honesta. El rendimiento de un modelo de aprendizaje supervisado refleja la diferencia entre las predicciones de éste para un conjunto de entradas **X**, y los valores **Y** esperados.

EN PRODUCCIÓN. GALERADAS

2. COMPLEJIDAD DEL MODELO

Se dirá que un modelo es más o menos complejo en función de la cantidad de parámetros e hiperparámetros que haya que estimar a la hora de entrenar y optimizar un modelo. La grandeza está en la sencillez, y muchas veces, por usar un modelo más complejo, no se obtienen resultados mejores, pero hay veces, que no queda otra que usar modelos complejos que tengan mayor expresividad.

Aumentar la complejidad del modelo:

- Puede disminuir el error sobre los datos de entrenamiento y nuevos datos.
- Depende del tamaño y naturaleza del conjunto de datos.

Un **aumento excesivo** de la complejidad del modelo:

- Puede disminuir el error sobre los datos de entrenamiento, pero aumentarlo sobre nuevos datos.
- A veces de manera dramática (lo que se conoce como *caer en sobreajuste*).

Sobreajuste: Se produce cuando el modelo se ajusta demasiado al conjunto de datos de entrenamiento, y por ello, no generaliza bien en la predicción de nuevos casos.

EN PRODUCCIÓN - GALERADAS

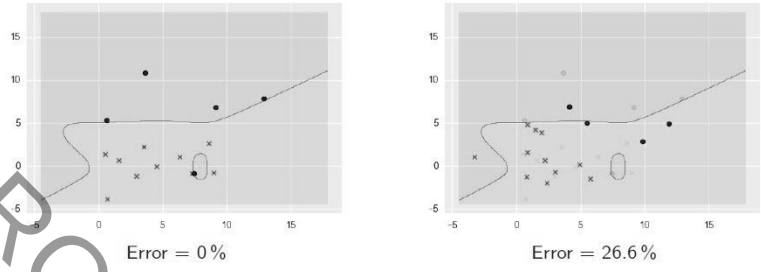


Figura 1. Sobreajuste de datos de entrenamiento.

2.2. SESGO O BIAS.

El sesgo o (bias) es el error que se produce debido a la incapacidad del modelo para ajustar los datos. Puede deberse a:

- Asunciones erróneas del algoritmo de aprendizaje.
- Falta de expresividad.

Un sesgo alto hace que el algoritmo no pueda capturar la relación entre las variables de entrada y las variables objetivo (**subajuste** o **underfitting**).

Ej. Uso de modelos de regresión lineal en la predicción de funciones no lineales.

2.3. VARIANZA

La varianza es el error que se produce debido a la sensibilidad ante pequeñas fluctuaciones o particularidades específicas en el conjunto de datos de entrenamiento.

- Esas particularidades no se presentan en los nuevos datos.

EN PRODUCCIÓN - GALERADAS

- Una varianza alta puede producir sobreajuste u overfitting.

Ej. Regresión logística con características polinómicas (grado 5). (Figura 1)



Figura 2. Sesgo - varianza

Además, podemos descomponer el error en términos del sesgo y la varianza de la siguiente forma:

$$Error = Sesgo + Varianza + Error\ Irreducible$$

EN PRODUCCIÓN - GALERADAS

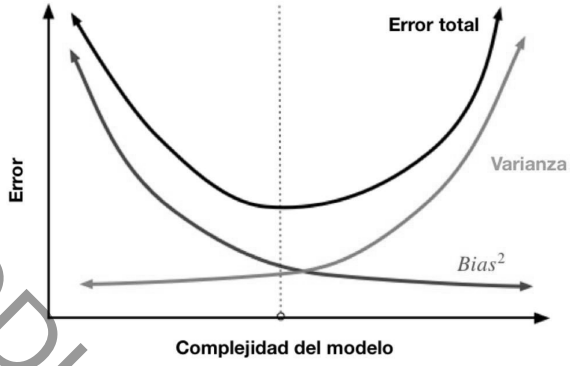


Figura 3. compensación sesgo-varianza

2.3. DEFINICIÓN DE RENDIMIENTO.

El rendimiento de un modelo de aprendizaje supervisado refleja la diferencia entre las predicciones de éste para un conjunto de entradas **X**, y los valores **Y** esperados.

Notas importantes:

- El rendimiento se corresponde al que presenta en la predicción de nuevos casos.
- La evaluación del modelo ha de hacerse sobre datos que no hayan sido utilizados en **ninguna fase** del proceso de entrenamiento. (Data Leakage).
- Es necesario desordenar aleatoriamente los ejemplos del conjunto original antes de hacer la partición (de cara a evitar sesgo de ordenación).
- A veces se trabaja con particiones estratificadas. Mantienen la proporción de casos de cada clase en entrenamiento y test. **Especialmente necesario cuando el conjunto de datos es pequeño o muy desbalanceado.**

EN PRODUCCIÓN - GALERADAS

2.4. PROBLEMAS DE REGRESIÓN.

El error cuadrático medio, entre una de las muchas métricas que existen para los problemas de regresión, se formula como:

$$ECM(h_\theta, X) = \frac{1}{m}\sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

2.5. PROBLEMAS DE CLASIFICACIÓN.

La tasa de acierto/error:

$$\text{ERROR}(c_\theta, X) = \frac{1}{m}\sum_{i=1}^{m} \text{error}\left(c_\theta\left(x^{(i)}\right), y^{(i)}\right)$$

$$\text{error}\left(c_\theta\left(x^{(i)}\right), y^{(i)}\right) = \begin{cases} 0 & \text{si } c_\theta\left(x^{(i)}\right) = y^{(i)} \\ 1 & \text{si } c_\theta\left(x^{(i)}\right) \neq y^{(i)} \end{cases}$$

$$ACIERTO(c_\theta, X) = 1 - \text{ERROR}(c_\theta, X)$$

Donde **X** representa un conjunto de datos de tamaño m; h_θ y c_θ, respectivamente, los modelos de regresión y clasificación; y θ representa los parámetros de los modelos.

Cuando realizamos el **entrenamiento** de un modelo sobre un conjunto de datos, se nos devuelve la **configuración de parámetros óptima**, es decir, aquella que **minimiza el coste con respecto a los datos utilizados en el aprendizaje**.

Evidentemente, el error del modelo aumentará en la predicción de nuevos casos. Puede no ocurrir en algunos casos, pero lo habitual es que aumente.

EN PRODUCCIÓN - GALERADAS

3. ASPECTOS A TENER EN CUENTA: ANÁLISIS DE LAS CURVAS DE APRENDIZAJE.

- El tamaño del conjunto de datos de entrenamiento es determinante en el rendimiento del algoritmo.
- El aumento en datos de entrenamiento disminuye la varianza.
- Un conjunto suficientemente grande de datos impide el sobreajuste.
- Una manera de analizar el comportamiento del algoritmo (si el error se debe al sesgo o a la varianza):
    - Dibujar los errores de entrenamiento y validación en función del número de ejemplos utilizados en el entrenamiento, y compararlos.

3.1. SUBAJUSTE.

- El aumento del tamaño del conjunto de datos no ayuda demasiado cuando el problema es que el modelo no tiene capacidad de ajustar (sesgo alto).
- En este caso, tanto el error de entrenamiento como el de validación/test se estancan en un valor alto.

EN PRODUCCIÓN - GALERADAS

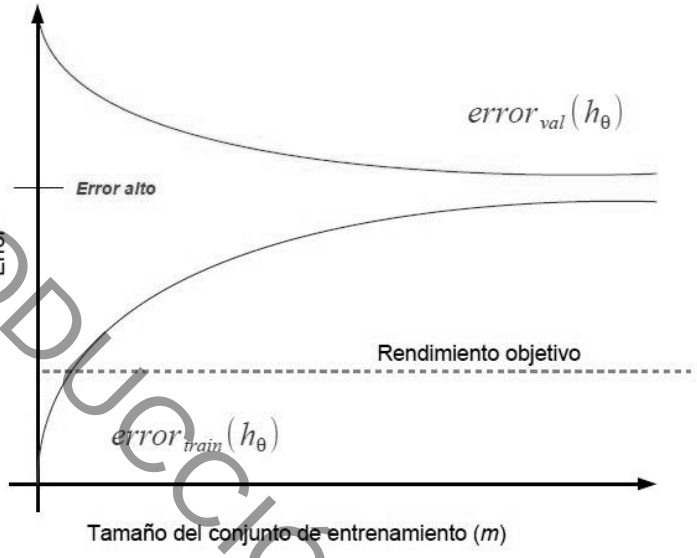


Figura 4 Subajuste del Modelo

3.2. SOBREAJUSTE.

- Cuando el modelo presenta sobreajuste, un aumento en el tamaño del conjunto de datos repercute en una mejor capacidad de generalización (se disminuye la varianza).
- El error de entrenamiento puede aumentar, pero el de validación disminuye, y ambos tienden a acercarse.

EN PRODUCCION - GALERADAS

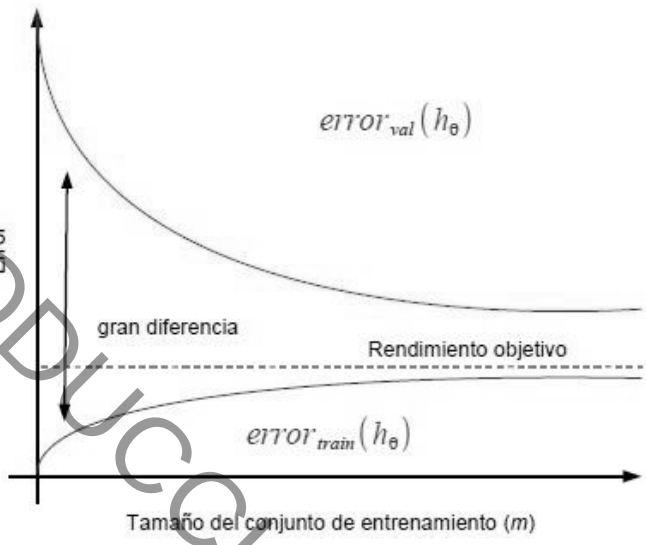


Figura 5 Sobreajuste

Cuando se aborda un problema de aprendizaje, no solamente hay que elegir el algoritmo adecuado. Es necesario establecer diferentes parámetros que tendrían un impacto en el rendimiento final. Por ejemplo:

- Variables (número) utilizadas.
- Creación de características polinomiales.
- Regularización en regresión lineal y logística.
- Topología y número de neuronas en redes neuronales.
- Limitar el número de dependencias en clasificadores bayesianos.
- Umbral de poda en árboles de clasificación, etc.

EN PRODUCCIÓN - GALERADAS

Cuando los modelos son muy expresivos pueden, dependiendo del volumen de datos, sobreajustar. Por eso se busca un equilibrio.

En la siguiente imagen, se puede apreciar cómo a medida que aumenta la capacidad de captar detalles, el error en Train se reduce, pero llega un punto en el que el modelo empieza a «memorizar» patrones no generales y empieza a aumentar el error en Test.

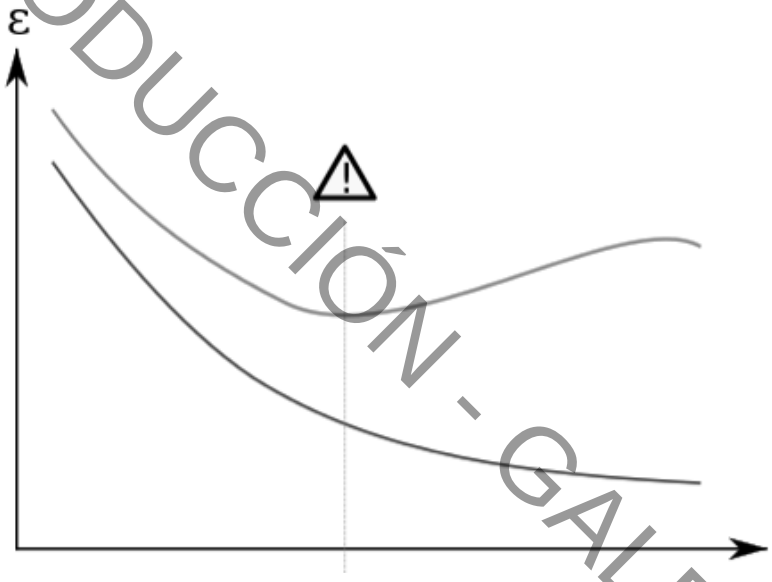


Figura 6 Curva de Sobreajuste

EN PRODUCCIÓN - GALERADAS

4. TÉCNICAS DE VALIDACIÓN.

- Train/test split
- k-Fold Cross-Validation
- Leave-one-out Cross-Validation
- Leave-one-group-out Cross-Validation
- Nested Cross-Validation
- Time-series Cross-Validation
- Wilcoxon signed-rank test
- McNemar's test
- 5x2CV paired t-test
- 5x2CV combined F test

Por no hacer esto eterno, se describirán únicamente las técnicas más utilizadas.

1. TRAIN – TEST SPLIT.

La técnica más básica de validación de modelos es realizar una división de entrenamiento/validación/prueba en los datos. Una proporción típica para esto podría ser 80/10/10 para asegurar suficientes datos de entrenamiento. Después de entrenar el modelo con el conjunto de entrenamiento, pasaremos a validar los resultados y a ajustar los hiperparámetros con el conjunto de validación hasta alcanzar una métrica de rendimiento satisfactoria. Una

EN PRODUCCIÓN - GALERADAS

vez completada esta etapa, pasaremos a probar el modelo con el conjunto de pruebas para predecir y evaluar el rendimiento.

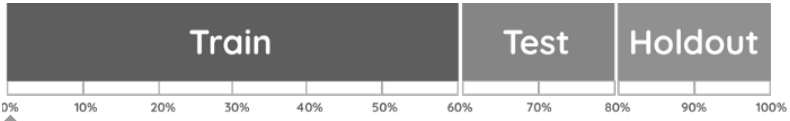


Figura 7 División en Train - Test

La evaluación entrenamiento/test no es adecuada cuando las bases de datos son pequeñas. Debido a la gran diferencia entre las dos particiones (depende de la semilla).

Puesto que seguramente todo problema, haya alguien que se lo ha planteado previamente, tenemos también soluciones:

- Hold-out repetido/iterado. Repetir *k* veces:

    1. Aleatorizar conjunto de datos. Dividir train/test.
    2. Aprender sobre el conjunto de entrenamiento y validar con el de test.
    3. Devolver como medida la media obtenida sobre los *k* conjuntos de test.

- Bootstraping. Repetir *k* veces:

    1. Obtener una muestra de tamaño *m* a partir del conjunto de datos mediante **muestreo con reemplazo**.
    2. Aprender el modelo a partir de la muestra obtenida.
    3. Validar el modelo sobre los datos OOB (out of Bag) no seleccionados en el muestreo.
    4. Devolver como medida la media obtenida sobre los *k* conjuntos de test.

EN PRODUCCIÓN - GALERADAS

¿Y si seguimos teniendo buena/mala suerte? ¿Se puede hacer más sistemáticamente? La respuesta es afirmativa y viene dada por la **Validación Cruzada**.

2. VALIDACIÓN CRUZADA K-FOLD CON UN CONJUNTO DE DATOS DE PRUEBA INDEPENDIENTE.

En el caso de que quisiéramos conservar la mayor cantidad de datos posible para la etapa de entrenamiento y no arriesgarnos a perder datos valiosos para el conjunto de validación. Esta técnica no requerirá que los datos de entrenamiento cedan ninguna porción para un conjunto de validación. En este caso, el conjunto de datos se divide en un número k de pliegues, uno de los cuales se utilizará como conjunto de prueba y el resto se utilizará como conjunto de datos de entrenamiento, y esto se repetirá n veces, según especifique el usuario. En una regresión, la media de los resultados se utilizará como resultado final. En una clasificación, se tomará como resultado final la media de los resultados (es decir, exactitud, tasa de verdaderos positivos, F1, etc.).

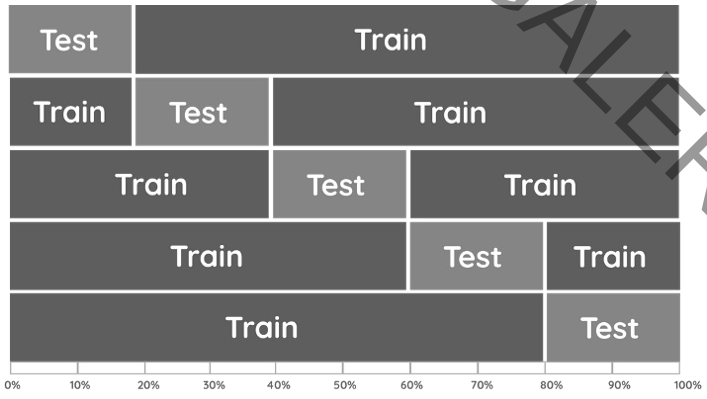


Figura 8 Validación cruzada 5 - fold

EN PRODUCCIÓN. GALERADAS

- En clasificación se suele hacer validación cruzada estratificada.
- **Leave One Out**: Cuando *k* es igual al número de registros de **X** (se verá más adelante).
- **Validación Cruzada Iterada**: r x k-CrossVal. Se realizan r iteraciones de una validación cruzada de k pliegues o divisiones y se devuelve la media.

3. LEAVE-ONE-OUT CROSS-VALIDATION WITH INDEPENDENT TEST DATA SET (VALIDACIÓN CRUZADA CON UN CONJUNTO DE DATOS DE PRUEBA INDEPENDIENTE).

La validación Leave-One-Out es similar a la validación cruzada k-fold. La iteración se lleva a cabo unas veces determinadas y el conjunto de datos se dividirá en **n-1** conjuntos de datos y el que se elimine será el dato de prueba.

Es de mencionar, que esta técnica es prohibitiva hablando en términos de coste computacional si la muestra es lo suficientemente grande.

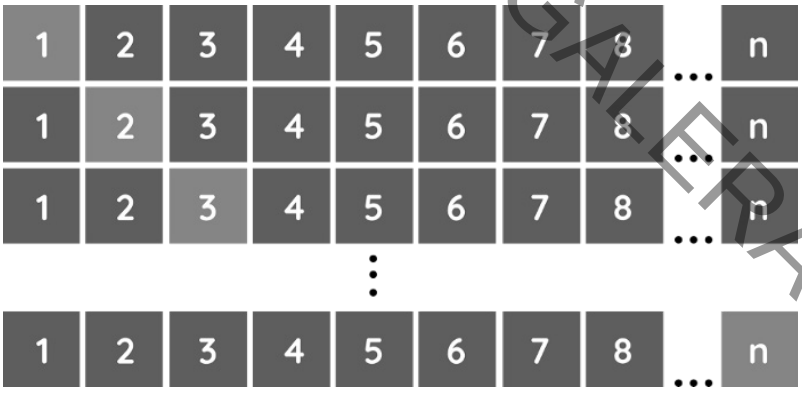


Figura 9 Leave One-Out

EN PRODUCCIÓN - GALERADAS

5. METODOLOGÍA PERSONAL PARA ENTRENAR Y VALIDAR A LA VEZ.

En esta sección, se detallará una metodología propia para entrenar modelos, ya que es buena, rápida y sencilla. Está basada en una combinación de las técnicas que han sido explicadas hasta ahora.

En primer lugar, se realiza una división en **Train** y **Test**, ya que es importante probar el modelo sobre datos que nunca ha visto, para comprobar el rendimiento. Además, procuro que la división en clases de ambos conjuntos sea estratificada (se conserve la proporción).

Una vez que ya tengo la división en conjunto de entrenamiento y test, olvidaremos completamente el conjunto de test durante un rato, y se procede al siguiente paso.

Ahora, centrándonos únicamente en el conjunto de entrenamiento, se realiza una **validación cruzada exhaustiva** (adjunto el enlace al objeto que utilizo de la librería **Scikit-Learn (Sklearn**), en **Python**.

Validación cruzada exhaustiva: Se entenderá por **Validación Cruzada Exhaustiva** una validación cruzada K-fold para cada combinación de hiperparámetros. Devolviendo el mejor modelo (con la mejor combinación de hiperparámetros, y validado con la validación cruzada). **La consecuencia** será la obtención de un modelo validado internamente, y por tanto a la hora de probarlo en el conjunto test, no debería variar mucho de los resultados obtenidos en la validación cruzada.

EN PRODUCCIÓN - GALERADAS

Esto suena un poco lioso al decirlo de esta manera, pero se puede ver de la siguiente manera:

Los **modelos** como bien sabemos se cargan como **objetos** (que tienen propiedades, parámetros, métodos, ...) (todo esto se presupone que el lector lo conoce), y al entrenarlos, obtenemos esos parámetros y entonces podemos usar sus métodos, como la realización de predicciones, etc. Por ejemplo, si consideramos el modelo Regresión Logística (LogisticRegression en Sklearn), tenemos los siguientes parámetros:

- penalty (penalización/ regularización) (l1, l2, elasticnet, ninguna), por defecto, l2.
- dual (opción para regularización) (booleano), por defecto, false.
- C (parámetro de regularización).
- Class Weight (peso de las clases) (balanceado, personalizado o ninguna), por defecto ninguna, etc.

Como se ha podido observar, cada modelo tiene distintos **parámetros** y cada parámetro tiene distintas **opciones disponibles**. Los parámetros más importantes de la regresión logística son: **penalty, C y Class Weight**.

class sklearn.linear_model**.**LogisticRegression(penalty='l2'**,** *, dual=**False,** tol=0.0001**,** C=1.0**,** fit_intercept=**True,** intercept_scaling=1**,** class_weight=**None,** random_state=**None,** solver='lbfgs'**,** max_iter=100**,** multi_class='auto'**,** verbose=0**,** warm_start=**False,** n_jobs=**None,** l1_ratio=**None)**

Esta es la forma que tiene un objeto de la clase LogisticRegression. Se puede apreciar la cantidad de parámetros que tienen estos modelos, y no es necesario tampoco conocer cómo funcio-

EN PRODUCCIÓN - GALERADAS

nan la mayoría de ellos, pero sí los más importantes. Ya que es vital conocer cómo funciona el modelo y sus parámetros (cómo influyen) para poder ajustarlo correctamente.

Imagínese que se pretende, por comodidad, obtener el modelo con los hiperparámetros **C** y **Class Weight** mejor ajustados a los datos. Entonces el objeto GridSearchCV, probará todas las combinaciones de parámetros posibles y hará una validación cruzada con cada una, devolviéndo la mejor. Es decir, si se tienen las siguientes listas de candidatos a hiperparámetros:

- **from** itertools **import** product
- C**,** class_weight**, = [**0.1**, **1**, **10**, **100**], [None,** 'balanced']
- **list(**product(C**,** class_weight**))**

Entonces, como le decimos que C pueda tomar 4 valores, y class_weight, 2 valores:

```
>>> (0.1, None)
>>> (0.1, 'balanced')
>>> (1, None)
>>> (1, 'balanced')
>>> (10, None)
>>> (10, 'balanced')
>>> (100, None)
>>> (100, 'balanced')
```

Esas son las 8 combinaciones posibles de parámetros a probar.

El problema está en que, para cada una de esas combinaciones, el modelo tendrá que realizar una validación cruzada sobre los datos de training. La validación cruzada tiene un coste computacional bastante alto, aquí, tendremos que hacer un mínimo

EN PRODUCCIÓN - GALERADAS

de 8 validaciones cruzadas. Pero se puede intuir el peligro que tienen estas combinaciones para un modelo más complejo.

Lo bueno es que el modelo devuelto, aparte de ser óptimo en parámetros e hiperparámetros (entre los aportados), está validado internamente, y se puede comprobar ello en el conjunto test. Si fuese necesario, se puede repetir todo el entrenamiento, usando toda la muestra como si fuese entrenamiento, una vez se ha hecho este estudio.

Para no insistir más en este punto ni agregar demasiado código, se implementará en código y se subirá a un repositorio en GitHub para quien quiera acceder a ello.

	Accuracy	Recall	Precision	F1
Train	0.631444	0.718055	0.121932	0.208465
Test	0.626876	0.707878	0.119260	0.204129

Figura 10 Comparación train-test de un modelo validado con CV

Pero como se puede apreciar, es normal que los resultados en Train y test con este método sean muy parecidos.

EN PRODUCCIÓN - GALERADAS

6. EVALUACIÓN DE CLASIFICADORES.

Una pregunta que nos puede surgir cuando nos enfrentemos a problemas de clasificación, es la siguiente: ¿Es adecuado usar siempre la tasa de acierto? Lo cierto es que la evaluación es sensible a la distribución de las clases:

- A veces la distribución de las clases en los datos no está equilibrada (imbalanced data). Esto quiere decir que, si el tamaño del conjunto de datos es **m**, y el número de clases es **c**, no tiene por qué haber $\frac{m}{c}$ instancias de cada clase. Esto puede suponer un problema en la evaluación.

Se han estudiado 10000 pacientes, de los cuales 9900 están sanos y 100 tienen una enfermedad grave. Un clasificador C_1 predice que todos los pacientes están sanos. La tasa de acierto de dicho clasificador es de $\frac{9900}{10000} = 99\%$. ¿Es C_1 un buen clasificador?

Tenemos que echar mano de la matriz de confusión:

EN PRODUCCIÓN - GALERADAS

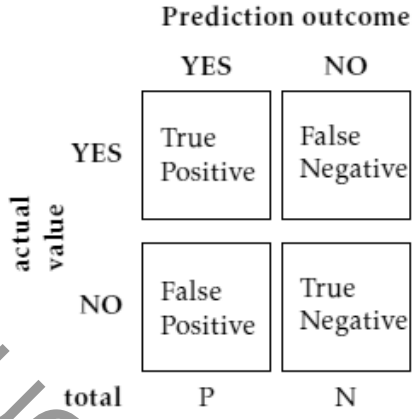


Figura 11 Matriz de confusión.

6.1. MÉTRICAS ASOCIADAS A LA MATRIZ DE SOBREAJUSTE.

- Tasa de acierto (Accuracy): $accuracy = \frac{TP+TN}{P+N}$
- Tasa de verdaderos positivos $TPR = \frac{TP}{TP+FN} = \frac{TP}{P}$
- Tasa de verdaderos negativos TNR (specificity): $TNR = \frac{TN}{TN+FP} = \frac{T}{N}$
- Tasa de falsos positivos FPR (negative error): $FPR = \frac{FP}{TN+FP} = \frac{FP}{N}$
- Valor positivo predictivo (precision): $PPV = \frac{TP}{TP+FP}$
- Macromedia: $MM = \frac{TPR+TNR}{2}$

EN PRODUCCIÓN - GALERADAS

Se cumple que: $FPR = (1 - specificity)$

Cuando una clase aparece con poca frecuencia, se suelen usar los índices *precision* y *recall* para resumir los resultados del modelo.

- **Precision** $\left(\frac{TP}{TP+FP}\right)$: De todos los casos clasificados como positivos, qué fracción de ellos lo son realmente. Qué fracción de los pacientes diagnosticados como enfermos lo están realmente.
- **Recall** $\left(\frac{TP}{TP+FN}\right)$: Qué fracción de casos positivos se ha detectado. Qué fracción de pacientes enfermos se han clasificado como tales.

6.2. BALANCE ENTRE PRECISION Y RECALL.

La regresión logística devuelve $h_\theta(x) \in [0,1]$, que es la probabilidad de que un caso sea positivo. Por defecto (salvo que se modifique), se clasifica como positivo el caso con probabilidad > 0.5, lo que se llama umbral = 0.5, pero se puede modificar.

Por tanto, si solamente se quiere predecir y = 1 cuando se está muy seguro:

- Umbral muy alto.
- Más precision.
- Menor recall.

Si lo que se busca es evitar falsos negativos:

- Se baja el umbral.
- El recall sería más alto.

EN PRODUCCIÓN. GALERADAS

- La precision sería más baja.

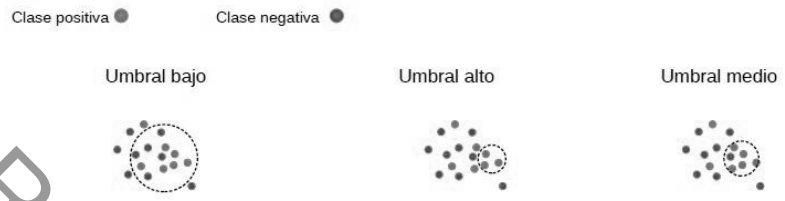


Figura 12 Umbrales de un modelo de clasificación

¿Cómo se comparan entonces los índices de precisión y recall? Véase qué ocurre en el siguiente ejemplo:

	Precision	Recall	Media	F_1 Score
Algoritmo 1	0.5	0.4	0.45	0.44
Algoritmo 2	0.7	0.1	0.4	0.175
Algoritmo 3	0.02	1.0	0.51	0.0392

En el ejemplo, la mejor media se da para un algoritmo con recall = 1 y precision ≈ 0, que no funciona bien porque predice casi todo positivo.

$$F_1 = 2\frac{P.R}{P+R}$$

F1-Score oscila entre 0 (si P = 0 o R = 0) y 1 (si P = 1 y R = 1).

6.3. ANÁLISIS ROC.

En el Análisis ROC (Receiver Operating Characteristics) se normaliza la matriz de confusión por columnas. El clasificador se representa como un punto en el espacio bidimensional dado por la tasa de verdaderos y falsos positivos. De la siguiente tabla:

EN PRODUCCIÓN - GALERADAS

	sí	no
sí	50	1000
no	50	8900
Total	100	9900

Si se normaliza, se llega a la siguiente:

	sí	no
sí	TPR = 0.5	FPR = 0.101
no	FNR = 0.5	TNR = 0.899
Total	1	1

Si para esta tabla el umbral es de 0.5, ya se ha obtenido el primer punto de la curva ROC, que se corresponde con:

$$(FPR, TPR) = (0.101, 0.5)$$

Si se hace variar el umbral entre todos los valores posibles y repetimos esta operación, se obtendrán todos los puntos de la curva, quedando lo siguiente:

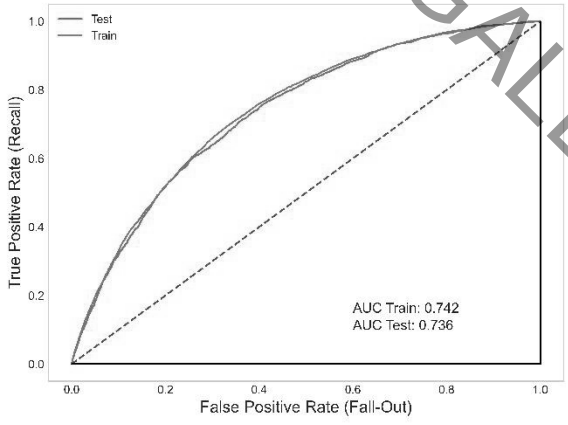


Figura 13 Curvas ROC

EN PRODUCCIÓN - GALERADAS

La curva ROC representa el rendimiento de clasificadores que utilizan score + umbral (Ej. regresión logística). Cada punto representa el rendimiento con un umbral. Se utilizan como umbrales los scores correspondientes a cada caso del conjunto de entrenamiento.

El área bajo la curva ROC (AUC) de un modelo de clasificación corresponde al área debajo de la curva ROC que va desde los puntos (0,0) a (1,1).

Puede interpretarse como la probabilidad de que el score que el modelo asigna a un ejemplo positivo sea superior al asignado a un ejemplo negativo.

Una vez llegado a este punto, una pregunta interesante podría ser: ¿Todos los errores son iguales?

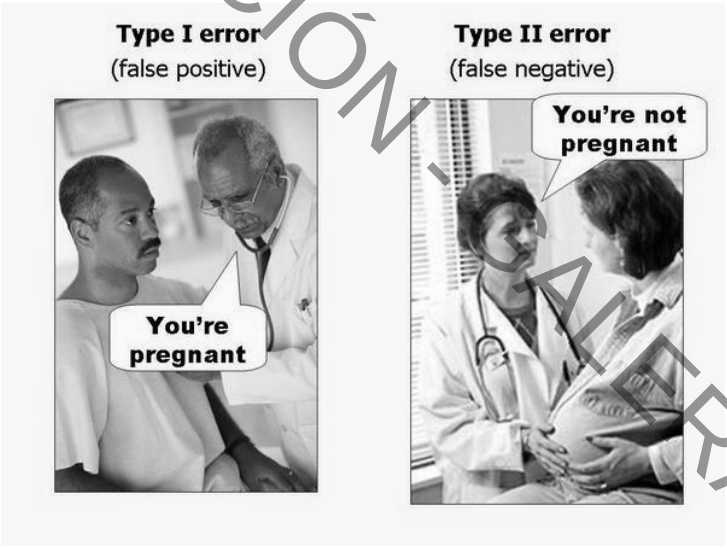


Figura Error tipo I y error tipo II

De los siguientes clasificadores, ¿Cuál es mejor? (Real (R) vs Predicción (P))

EN PRODUCCIÓN - GALERADAS

Array 1:

C_1	sí (R)	no (R)
sí (p)	0	0
no (p)	100	9900

Array 2:

C_2	sí (R)	no (R)
sí (p)	50	1000
no (p)	50	8900

Clasificador	C1	C2
Métrica	-	-
ACC	99%	89.5%
TPR	0%	50%
TNR	100%	89.5%
FPR	0%	10.1%
PPV	indefinido	9.52%
MM	50%	94.6%

En muchas ocasiones, los errores producidos no tienen la misma importancia, es más grave clasificar como sano a un paciente enfermo que clasificar como enfermo a un paciente sano. Cada uno de los tipos de errores (FP y FN) puede llevar un coste asociado:

Coste	sí	no
sí	0	100$
no	5000$	0

Aquí el objetivo será obtener el clasificador que minimice el coste.

EN PRODUCCIÓN - GALERADAS

7. CONCLUSIONES

- El verdadero rendimiento de un modelo de aprendizaje supervisado se evalúa sobre datos no utilizados sobre el aprendizaje.
- Si no es posible reservar datos para la evaluación, se puede utilizar validación cruzada.
- Un mayor número de datos para el entrenamiento disminuye el error.
- El error puede deberse a subajuste (demasiado sesgo) o sobreajuste (demasiada varianza).
- Un aumento de la complejidad del modelo puede repercutir positivamente (eliminación de sesgo/subajuste) o negativamente (aumento de la varianza/sobreajuste).
- Las curvas de aprendizaje y validación permiten determinar el origen del error y cómo mejorar el rendimiento del modelo.

EN PRODUCCIÓN - GALERADAS

8. BIBLIOGRAFÍA

1. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. New York: Springer; 2012.
2. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2008.
3. Cáceres RÁ. Estadística aplicada a las ciencias de la salud. Editorial Díaz de Santos, S.A.; 2007.
4. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
5. Smith EE, Shobha N, Dai D, Olson DM, Reeves MJ, Saver JL, et al. A risk score for in-hospital death in patients admitted with ischemic or hemorrhagic stroke. J Am Heart Assoc. 2013;2(1):e005207

EN PRODUCCIÓN - GALERADAS





Trabajo Fin de Máster
Máster en Ciencia de Datos e Ingeniería de Datos en la Nube

APLICACIÓN DE CIENCIA DE DATOS EN LA OPTIMIZACIÓN DE MODELOS PREDICTIVOS PARA LA ESTIMACIÓN DE MORTALIDAD EN ICTUS ISQUÉMICO

Autor: Luis Téllez Ramírez

Tutor: Jesús Martínez

Cotutor: Juan Manuel García Torrecillas



Convocatoria FPS 2020. Expdte: AP-0013-2020-C1-F1

Julio, 2021

EN PRODUCCIÓN - GALERADAS

RESUMEN

El propósito de este trabajo no es otro que la aplicación de algunas de las técnicas vistas a lo largo del máster de cara a trabajar en un ámbito donde puede haber una gran conexión, como es la medicina, y en particular, el desarrollo de modelos para estimar la gravedad de un paciente.

A su vez, se trabaja con una base de datos extraída del CMBD, completamente anonimizada, donde los pacientes hayan sufrido un ictus isquémico en alguno de los diagnósticos. Se pretende entonces desarrollar modelos alternativos tanto de Machine Learning como Deep Learning, aplicar algoritmos de aprendizaje supervisado y no supervisado, técnicas de resumen de información, de selección de variables y también de explicabilidad e interpretabilidad de modelos para comprobar si estos realizan predicciones de manera coherente con el criterio experto.

Finalmente, se pretende implementar estos modelos desarrollados en la plataforma AWS para poder utilizarlos y validarlos con nuevos datos externos de cara a su recalibración y mejora en un futuro.

EN PRODUCCIÓN - GALERADAS

ÍNDICE DE FIGURAS

- Figura 1.1. Esquema de uso de CMBD.
- Figura 2.1 Diferencia de Edad entre el grupo de fallecidos y no fallecidos.
- Figura 2.2 Bootstrapping para la diferencia de Edad.
- Figura 2.3 Fallecidos por estrato de Edad.
- Figura 2.4 Fallecidos – Sexo en el escenario global (tanto A como B).
- Figura 2.5 % Mujeres y Hombres en la muestra global (tanto A como B).
- Figura 2.6 Fallecidos por diversos factores.
- Figura 2.7 Prevalencia de los factores en la población.
- Figura 2.8 Anomalía en la variable FA.
- Figura 2.9 t-SNE Escenario A.
- Figura 2.10 t-SNE Escenario B.
- Figura 3.1 Oversampling - Cross Validation Erróneo.
- Figura 3.2 Oversampling - Cross Validation Correcto.
- Figura 3.3 Curva de calibrado perfecta.
- Figura 3.4 Algoritmo genético selección de predictores.
- Figura 4.1 Curva de calibrado de la Regresión Logística Balanceada.
- Figura 4.2 Modelo calibrado.
- Figura 4.3 Importancia de los predictores según Random-Forest.
- Figura 4.4 Test de importancia por permutaciones.
- Figura 4.5 Valores SHAP.

EN PRODUCCIÓN - GALERADAS

- Figura 4.6 Influencia de los predictores según valores SHAP.
- Figura 4.7 Distribución de probabilidad de falsos positivos.
- Figura 4.8 Distribución de probabilidades de modelo calibrado y no calibrado.
- Figura 4.9 Shap values de los falsos positivos.
- Figura 4.10 Comparación de Modelos por ROC.
- Figura 4.11 Poder ROC mediante eliminación recursiva de predictores.
- Figura 5.1 Importancia predictores escenario B.
- Figura 5.2 Importancia por permutación.
- Figura 5.3 Comparación modelos escenario B.
- Figura 5.4 Eliminación recursiva de predictores sin importancia.
- Figura 5.5 Eliminación recursiva de predictores importantes.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN GALERADAS

CAPÍTULO 1. INTRODUCCIÓN [0]

1. SITUACIÓN

En este capítulo se realiza una recopilación de información relacionada con el objetivo de situarnos. En este trabajo hay bastantes cosas ya hechas por parte del grupo investigador en el que actualmente me incluyo, y, por tanto, va a jugar un papel importante el hecho de contextualizar correctamente.

Este grupo está formado por: María del Carmen Lea Pereira (Medicina Interna), Patricia Martínez Sánchez (Neurología), María Isabel Álvarez Moreno (Medicina Interna), José Galván Espinosa (Gestión), María del Mar Iglesias Espinosa (Neurología), Juan José López Ramos (Informática), Juan Manuel García Torrecillas (Medicina de Familia y Epidemiología), Fernando Reche Lorite (Estadística e investigación operativa), María del Mar Rodríguez (Enfermería). Se trata de un proyecto de investigación e innovación que tiene lugar dentro del Sistema Sanitario Público de Andalucía, y tiene por título: "Optimización de un modelo predictivo para la mortalidad en ictus isquémico a través de una cohorte de "Real World Data".

Algunos de los objetivos del proyecto inicial, es la realización de una validación externa de un modelo (regresión multivariante) actualmente implementado. Este objetivo llevará un tiempo considerable (reunir casos para una muestra) y, mientras tanto, se ha visto oportuno hacer una revisión del trabajo realizado hasta ahora aportando una nueva metodología y técnicas innovadoras de

Ciencia de Datos. Muchas de estas técnicas han sido contempladas a lo largo del máster, y, aunque no sea el objetivo primordial del proyecto, el de este TFM sí será hacer un análisis exploratorio de los datos, tratar el preprocesamiento de estos, aplicar distintas técnicas de modelado predictivo, hablar de la selección de variables, interpretar/explicar los modelos utilizados, aplicar modelos más avanzados como el uso de redes neuronales, ensambles, y, en definitiva, tratar de poner en marcha este proceso en producción mediante el uso de servicios AWS.

2. FUNDAMENTOS

El ictus isquémico supone la segunda causa de mortalidad en nuestro país en la población general y la primera causa de mortalidad en la mujer [1]. A nivel mundial el ictus es la segunda causa de mortalidad y la tercera más común en los países industrializados [2]. Por tanto, nos enfrentamos a un auténtico problema de salud pública. La mortalidad hospitalaria según fuentes españolas procedentes de registros de carácter clínico se sitúa en torno al 12.9% [3]. Un porcentaje elevadísimo de la misma son ictus isquémicos no lisables, esto es, no subsidiarios de tratamiento fibrinolítico (un tipo de fármaco) por no cumplir los estrictos criterios para la aplicación de este tratamiento.

Hasta la fecha se han detectado una serie de factores de riesgo tanto para el desarrollo de un evento isquémico como para estimar la probabilidad de fallecer o presentar secuelas [3]. Los trabajos de Smith et al [4] permitieron la obtención de modelos predictivos para la mortalidad hospitalaria, tanto por ictus isquémico como hemorrágico, y con un escaso número de variables, alcanzando una capacidad discriminativa estimada mediante el C-Statistic de 0.85 y una excelente capacidad discriminativa. Existen otros trabajos de alto interés que han mostrado una me-

EN PRODUCCIÓN - GALERADAS

todología acertada en la elaboración de modelos predictivos para ictus [5].

Estos factores han sido obtenidos fundamentalmente a partir de registros hospitalarios y en algún caso desde el Conjunto Mínimo Básico de Datos, si bien no hemos encontrado en la literatura un modelo predictivo basado en CMBD* que nos permita una adecuada estimación de la probabilidad de fallecer durante la hospitalización por un ictus agudo no lisable (GRD 14).

3. JUSTIFICACIÓN

Se ha demostrado que la existencia de un plan integral de actuaciones que, desde la llegada del paciente al hospital maximice y optimice su atención repercute de manera beneficiosa en los pacientes que han padecido un ictus agudo, aumentando las probabilidades de recuperación [2]. A lo largo de los últimos veinte años no sólo el cambio en las actividades preventivas, sino la actuación precoz, reglada y acorde a los estándares de calidad más avanzados, ha demostrado conseguir una disminución importante tanto de la mortalidad como de las secuelas del ictus.

Una vez se produce la admisión del paciente con ictus, no sólo el tiempo y las medidas adoptadas, sino también la formación y el equipo que presta la atención ha mostrado ser un factor determinante en la reducción de la tasa de mortalidad intrahospitalaria.

El conocimiento de las variables subsidiarias de ser recogidas a pie de cama al ingreso o en los primeros días de estancia del paciente y que, integradas en un modelo predictivo nos permitan estimar la probabilidad de fallecer durante el ingreso, tiene importantes repercusiones en la atención al paciente:

I. Permitiría modular la intensidad y velocidad del esfuerzo diagnóstico terapéutico en aquellos pacientes con scores de riesgo de puntaje más elevado.

EN PRODUCCIÓN - GALERADAS

II. Facilitaría la integración en circuitos asistenciales específicos de los pacientes de alto riesgo
III. Sería un elemento clave para la optimización de la costo-eficiencia en la atención a pacientes con ictus no lisable.

CMBD: El conjunto mínimo básico de datos (CMBD) es una base clínico-administrativa de obligado cumplimiento para los hospitales de nuestro Sistema Nacional de Salud. A partir de los informes de alta que realizamos, las unidades de codificación de cada centro hospitalario se encargan de recoger un conjunto mínimo básico de datos.

Cada vez que un paciente llega al hospital, recibe una serie de diagnósticos, procedimientos, toma de datos Todo ello queda registrado en una instancia de la base de datos CMDB, es decir, un mismo paciente puede tener varias instancias de dicha base de datos asociadas a su ID, pero si el paciente vuelve al hospital por el mismo motivo (mismo diagnóstico) que se trató la vez anterior en menos de 30 días, entonces se continuará sobre la misma fila, y la variable REINGRESO (booleana) pasará a ser 1.

Es importante entender esto porque los datos con los que vamos a trabajar han sido obtenidos de esta base de datos. Además, si tenemos en cuenta que es una base de datos a nivel Nacional, no es difícil imaginar la cantidad de "ruido" que podemos encontrar en ella. Esta misma fuente de ruido será un "verificador" y a la vez un factor limitante en la calidad de los modelos (se verá más adelante).

EN PRODUCCIÓN - GALERADAS

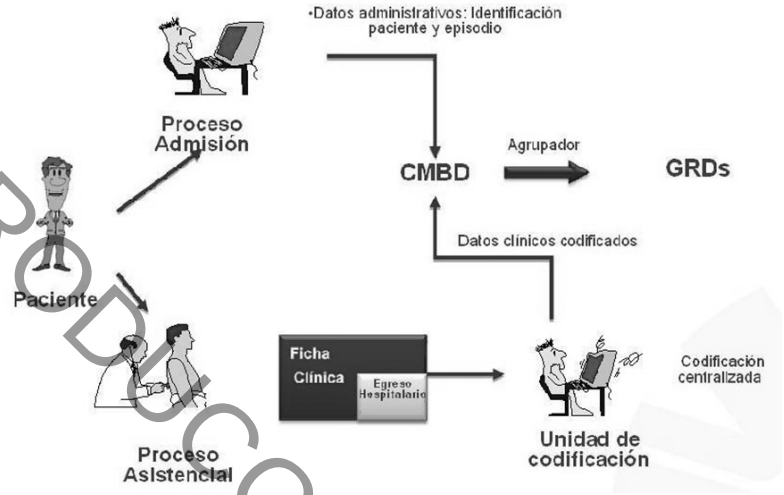


Figura 1.1. Esquema de uso de CMBD

El utilizar estas bases de datos para temas médicos tiene su inicio en EE. UU. en 1973, con "Uniform Hospital Discharge Data Set" (UHDDS), y tiene por objetivo la recogida de datos básicos que sirvan tanto para tema clínico como para tema administrativo. En 1982, la Comunidad Europea trata de imitarla, con "European Minimum Basic Data Set" (MDBS) con una finalidad de gestión, planificación, evaluación e investigación. Posteriormente se implantó en España el Conjunto Mínimo Básico de Datos en 1987, siendo de obligado cumplimiento desde 1992. Consta de un sistema de codificación propio GRD (Grupos relacionados por el diagnóstico).

4. HIPÓTESIS DEL TRABAJO

El estudio ya realizado parte de las siguientes hipótesis de trabajo:

EN PRODUCCIÓN - GALERADAS

I. Es posible identificar determinadas variables clínico-administrativas contenidas en el CMBD asociadas a la mortalidad intrahospitalaria de los pacientes con ictus no lisado.
II. Es posible elaborar un modelo predictivo de mortalidad intrahospitalaria basado en dichas variables que anticipe un mayor riesgo de mortalidad por ictus durante la hospitalización, facilitando el establecimiento de los recursos precisos para minimizar dicho desenlace.

La hipótesis del presente trabajo postula que, los pacientes ingresados con diagnóstico principal correspondiente a ictus isquémico agudo no lisado comparten características detectables, aislables y mensurables que permitirían la elaboración de un modelo logístico predictivo para la probabilidad de presentar mortalidad durante la hospitalización. Dicho modelo tendrá una capacidad discriminativa estimada mediante el área bajo la curva (estadístico C) no inferior a 0.70 y una calibración aceptable desde el punto de vista gráfico aún con valores de $\chi 2$ significativos para la prueba de Hosmer-Lemeshow.

5. DATOS UTILIZADOS

Para el desarrollo del modelo predictivo se estudiaron todos los episodios de hospitalización por el GRD 14 (ictus isquémico no lisado) a partir del Conjunto Mínimo Básico de Datos al alta (CMBD) de España desde 2008 hasta 2012. La información fue facilitada por el Instituto de Información Sanitaria del Ministerio de Sanidad y Consumo (IIS-MSC).

Para la codificación de los diagnósticos y procedimientos se ha utilizado la Clasificación Internacional de Enfermedades, 9ª edición Modificación Clínica, y para la agrupación previa de altas, el sistema de clasificación de pacientes de los Grupos Relacionados

EN PRODUCCIÓN - GALERADAS

por el Diagnóstico -AP-GRD-, versión 21.09. La información demográfica manejada procede del Instituto Nacional de Estadística (INE) [8].

6. DISEÑO

Se construyó una cohorte histórica a partir de todos los pacientes admitidos en el ingreso bajo el GRD 14, la cual fue desagregada posteriormente en dos grupos, de tal modo que se confrontaron los pacientes que no fallecieron durante el ingreso con aquellos que sí lo hicieron. Por tanto, se realizó un estudio de cohorte histórica con un grupo de comparación interna.

7. OBJETIVOS DEL PROYECTO

Principal: Validación externa mediante la determinación de la capacidad de discriminación y calibración de un modelo predictivo de mortalidad en ictus isquémico (modelo matemático presentado como patente y App como registro de propiedad intelectual) en una cohorte de pacientes reales hospitalizados.

I. **Específico 1: Recalibración** de los coeficientes del modelo teórico a partir de los datos de la cohorte real, determinando el punto de corte óptimo para la sensibilidad y especificidad en la predicción de mortalidad del modelo estadístico con los nuevos coeficientes.
II. **Específico 2: Trasladar y transferir** a todos los niveles asistenciales implicados en la atención al ictus una aplicación móvil y web recalibrada y optimizada

EN PRODUCCIÓN - GALERADAS

8. OBJETIVOS DEL TFM

Actualmente, el modelo de regresión logística se encuentra patentado y desplegado en una web, incluso su puesta en funcionamiento a nivel sanitario es inminente.

Sin embargo, para reunir datos con cierta calidad que sirvan para realizar una validación y calibración del modelo, hace falta tiempo. Durante ese "tiempo" lo que nos proponemos es:

- Hacer una revisión de la metodología utilizada. Implementar nuevas funciones que permitan ampliar el horizonte de opciones de cara a la investigación, esta vez de la mano de Ciencia de Datos.
- Traer nuevas técnicas utilizadas en Ciencia de Datos y proporcionadas por Librerías en Python como Sklearn, StatsModels, Tensorflow, Shap, XGBoost tratando de incorporar nuevas perspectivas en cuanto a la complejidad de modelos, explicabilidad de resultados, técnicas de resumen de información.
- Desarrollo de nuevos modelos de estimación de mortalidad con los que se espera igualar o mejorar los resultados, tanto para pre-ingreso del paciente, como para post-ingreso, lo que se corresponderá con los escenarios A y B.
- Una vez desarrollado y discutido el punto anterior, se propondrá a una puesta en funcionamiento en AWS mediante API Gateway y otros servicios, de manera que se puedan realizar estimaciones.

En definitiva, se pretende hacer un recorrido sobre muchos de los contenidos vistos en el máster de cara a ponerlos en funcionamiento en un caso real con posibilidad de realizar alguna publicación en este ámbito en un futuro.

EN PRODUCCION - GALERADAS

EN PRODUCCIÓN - GALERADAS

CAPÍTULO 2A. ANÁLISIS EXPLORATORIO

1. PREPROCESADO Y DESCRIPCIÓN DE LOS DATOS

La base de datos de la que disponemos es de SPSS (viene en formato .sav), debido a que los datos han formado parte de un proceso de estudio previo. Los datos gozan de cierta calidad, algunas variables han sido mapeadas para su uso, aunque al guardarse como archivo .sav pierden parte del procesamiento y hay que volver a realizarlo. Con pandas podemos leer archivos de este tipo con read_spss. Procederemos a detallar algunos de los pasos.

En dicha base de datos contamos con 186245 filas y 37 variables. Se trata de un problema en el que buscamos estimar una probabilidad, pero, en definitiva, es un problema de clasificación, donde la variable objetivo es 'EXITUS' (fallecidos según la jerga médica).

En la base de datos inicial hay más de 37 variables, pero estas se corresponden con temas administrativos y después de la decodificación y recodificación no tienen sentido.

De ahora en adelante, hablaremos de dos escenarios:

- **Escenario A:** Se desarrollarán todas las pruebas y modelos pertinentes, pero únicamente sobre variables que han sido seleccionadas por criterio experto. Estas variables son aquellas que pueden obtenerse de manera relativamente sencilla una vez ingresa el paciente. Se tendrían 10 variables predictoras y 1 variable objetivo.

- **Escenario B:** Puesto que disponemos de 37 variables, nos preguntamos si somos capaces de mejorar los resultados obtenidos en el Escenario A empleando únicamente técnicas de ciencia de datos.

Independientemente del escenario, se **normalizan los nombres de las columnas** y se ponen todos los nombres de las variables en mayúscula, evitando los acentos, caracteres especiales y cambio de 'ñ' por 'ny'.

Para en el **Escenario A** encontramos las siguientes variables (que serán comunes al B):

- **Edad**: Se considera como variable cuantitativa continua expresada en años completos.
- **Sexo**: Cualitativa booleana. Categorías Mujer vs Varón.
- **Exitus**: Cualitativa booleana. Indica la muerte del paciente.
- **Reingreso**: Cualitativa booleana. Indica si el paciente ha vuelto a ingresar por el mismo motivo en los 30 días posteriores al primer diagnóstico.
- **HTA**: Cualitativa booleana. Indica la presencia de hipertensión. DM: Cualitativa booleana. Indica la presencia de Diabetes mellitus. ARR: Cualitativa booleana. Indica la presencia de Arritmias.
- **DISLIPEM**: Cualitativa booleana. Indica la presencia de Dislipemia.
- **ICC**: Cualitativa booleana. Indica la presencia de insuficiencia cardíaca crónica.
- **ESTBASILAR**: Cualitativa booleana. Indica la presencia de estenosis de la arteria basilar.

EN PRODUCCIÓN - GALERADAS

Notas:

1. Estas son las variables con las que primero se desarrolló el modelo patentado con una regresión logística multivariante.
2. La idea inicial era usar la variable FA (fibrosis auricular en lugar de ARR), pero la calidad de la variable FA es deficiente, y, por tanto, se sustituye por ARR (que engloba a la variable FA), ya que está mejor codificada.
3. En el análisis que tendrá lugar en los siguientes puntos, se realizará una descripción global de la muestra.

Si estudiamos las variables que tienen más valores faltantes, obtenemos lo siguiente:

Variable	Valores faltantes
ARR	4740
VALV	2285
DM	2010
TNO_GI	1934
CISQ	1096
DISLIPEM	1043
FA	882
HTA	788
ICC	667
IRA	322
ANEMIA	291
TNO_HDROELEC	195

Tabla 1. Valores faltantes por variable.

EN PRODUCCIÓN - GALERADAS

Pero si atendemos únicamente a aquellas observaciones que tienen al menos un valor faltante en una de las 37 variables, obtenemos un total de 14742 registros, lo que corresponde a un 7.92% del total.

De esta manera, quedan 171503 registros donde los fallecidos ocupan el 6.76% de la muestra, por ello, no tiene sentido utilizar métricas como accuracy. Abordaremos esta cuestión más adelante.

Consideramos que al ser un tema delicado como es una enfermedad, realizar una técnica de imputación de valores faltantes puede ser arriesgado, y procedemos a la eliminación de esos registros.

Se transforman las variables binarias haciendo los siguientes cambios:

1. No -> 0
2. Sí -> 1
3. Vivo -> 0
4. Fallecido -> 1
5. Hombre -> 0
6. Mujer -> 1

Como hemos visto, salvo la edad, el resto son todas variables booleanas. A primera vista, podríamos pensar que esto es algo bueno, ya que si las clases **fallecer padeciendo Ictus y no fallecer padeciendo ictus** estuviesen "bien separadas" (esto es, si existiera una clara diferencia entre las características necesarias para pertenecer a una o a otra), se podría definir una frontera de decisión que separase las clases.

Pero ahora veremos que esto no es así exactamente, y que, aunque los predictores elegidos ayudan en las predicciones, no son suficientes como para discriminar correctamente.

EN PRODUCCIÓN - GALERADAS

Algunas variables más que tendremos en cuenta en el **Escenario B**:

EDAD_R: Edad estratificada transformada a cualitativa: '24-34':1, '35-44':2, '85 o más':7

NDA: Número de diagnósticos del paciente, variable categórica donde se ha hecho la siguiente transformación:
'0-3 diag':1, '4-7 diag':2, '>=12 diag':4

NPA: Número de procedimientos del paciente, variable categórica con la siguiente transformación: '0-3 proc':1, '4-7 proc':2, . '>= 16 proc':5

Tipo de Ingreso: Nos dice si el ingreso estaba programado o se debe a una urgencia: 'Urgente':1, 'Programado':2

ESTANCIA: Tiempo transcurrido en días completos desde fecha de ingreso hasta fecha de alta.

El resto de las variables se corresponden con otras enfermedades como arritmias, congénitas, obesidad siendo todas estas variables booleanas que indican la presencia.

2. EDAD EN LOS GRUPOS

El predictor Edad va a ser el más importante como veremos más adelante, y es que hay una clara diferencia entre los grupos de fallecidos y los que no fallecen.

Diferencia de medias observada: 8.7311641818239

EN PRODUCCIÓN - GALERADAS

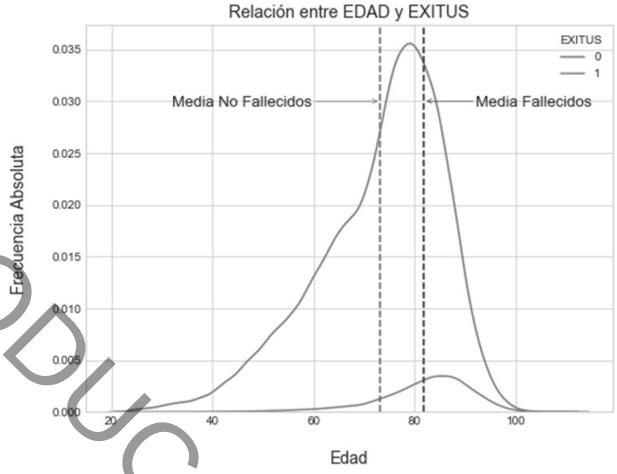


Figura 2.1 Diferencia de Edad entre el grupo de fallecidos y no fallecidos.

Podemos hacer un contraste de hipótesis mediante distribuciones Bootstraping, obteniendo lo siguiente:

- Intervalo al 95% de confianza: [-8.91621768, -8.54262454]

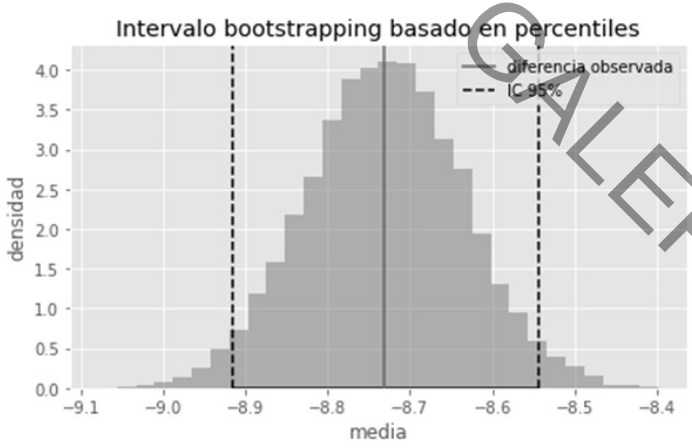
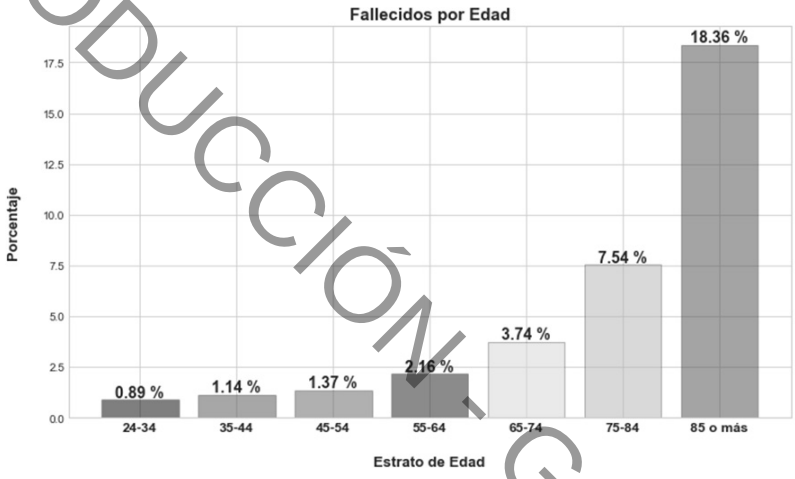


Figura 2.2 Bootstrapping para la diferencia de Edad.

EN PRODUCCIÓN - GALERADAS

Por tanto podríamos afirmar que sí existen evidencias para afirmar que la edad entre los dos grupos es distinta. El intervalo de confianza del 95% para la diferencia de medias obtenido por bootstrapping indica que, en promedio, la edad del grupo de fallecidos está entre 8.54 y 8.91 años más.

Una vez más, vemos como la Edad es un predictor clave, ya que aumenta el número de fallecidos de manera considerable con edades avanzadas.



3. SEXO

Como hemos comentado en el primer apartado, el fallecimiento por Ictus isquémico es más prevalente en las mujeres que en los hombres. Veámoslo en nuestros datos:

EN PRODUCCIÓN - GALERADAS

Figura 2.4 *Fallecidos – Sexo en el escenario global (tanto A como B).*

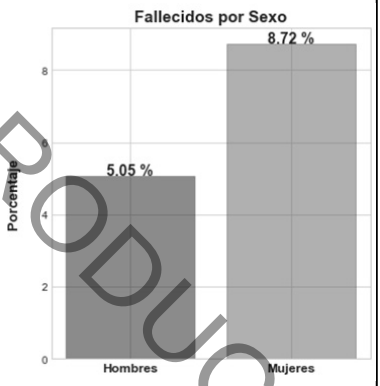
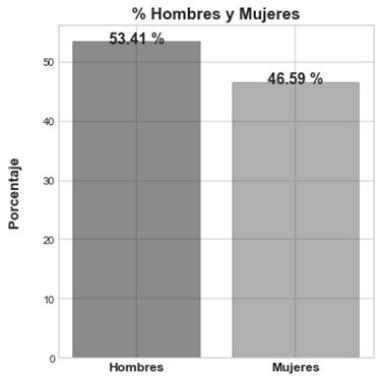


Figura 2.5 % *Mujeres y Hombres en la muestra global (tanto A como B).*



Claramente, fallecen más las mujeres que los hombres, y será determinante en la estimación de la probabilidad.

4. FALLECIDOS POR EL RESTO DE PREDICTORES

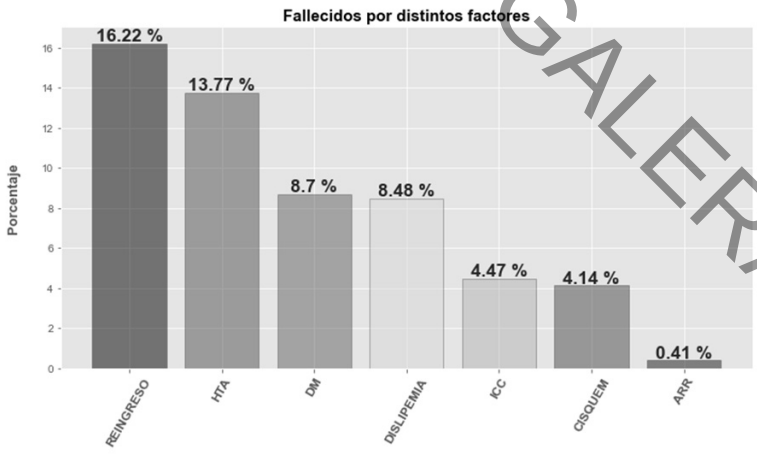


Figura 2.6 Fallecidos por diversos factores

EN PRODUCCIÓN - GALERADAS

Si además echamos un vistazo a la prevalencia de los factores en la población:

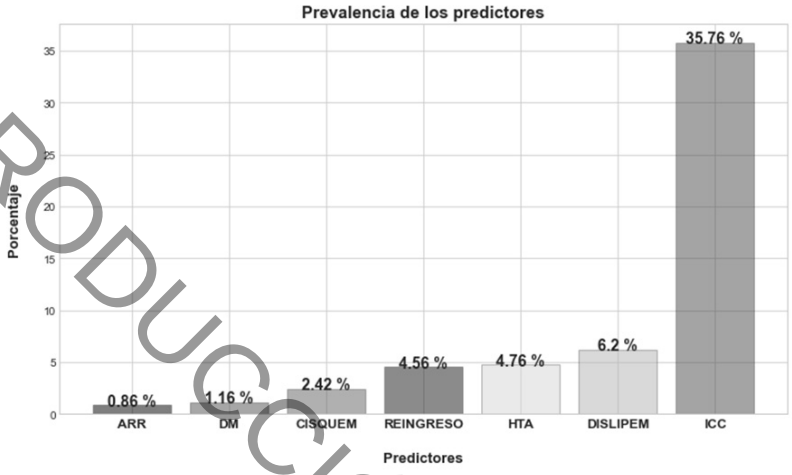


Figura 2.7 Prevalencia de los factores en la población.

Cuando se estaban visualizando estas variables, ha salido una anomalía en cuanto a la codificación de la variable FA.

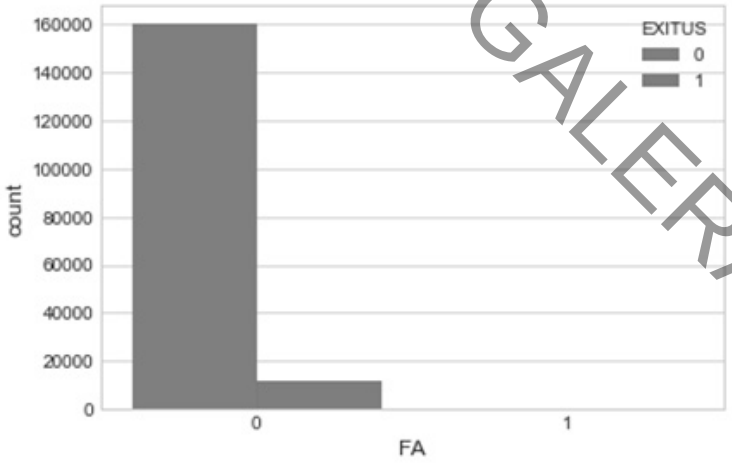


Figura 2.8 Anomalía en la variable FA.

EN PRODUCCIÓN - GALERADAS

Todas observaciones están codificadas como No, salvo un único caso que fallece. Es una variable que no nos proporciona ninguna información y que vamos a cambiar por la variable ARR, que, en particular, generaliza a la FA (existe multicolinealidad), y está mejor codificada.

5. DISCUSIÓN DE LA MÉTRICA

Como se ha comentado previamente, no tiene sentido utilizar la métrica **accuracy** por estar ante un problema desbalanceado.

Por un lado, queremos evitar los falsos negativos a toda costa, ya que es bastante grave decirle a un paciente que no tiene nada si este se encuentra muy grave. Por ello, **recall** podría ser una buena métrica.

Sin embargo, también es importante la **precisión**, que mide la habilidad de clasificar un verdadero positivo como positivo. Si maximizamos el recall pero muchos casos se nos van a falsos positivos, entendemos que el clasificador no discrimina correctamente.

Una buena métrica podría ser el **área bajo la curva precisión-recall** (AUPRC), métrica recomendada para casos de datos desbalanceados.

F1-Score será la métrica base que utilizaremos junto con precisión y recall a la hora de estudiar nuestros resultados. También tendremos en cuenta el comportamiento de la curva **ROC** (que mide la habilidad de distinguir entre las clases) y la métrica **Balanced- Accuracy.**

$$\textbf{\textit{Balanced Acc}} = \frac{TPR + TNR}{2}$$

$$\textbf{\textit{TPR (True positive rate)}} = \textbf{\textit{Recall}} = \frac{\#\,True\ positive}{\#\,Positives} = \frac{TP}{TP + FN}$$

$$\textbf{\textit{TNR (True negative rate)}} = \frac{\#True\ Negative}{\#Negatives} = \frac{TN}{TN + FP}$$

EN PRODUCCIÓN - GALERADAS

Como clasificador base para estudiar qué está ocurriendo, utilizaremos el modelo RandomForest, aunque entrenaremos también una regresión logística y probaremos algún otro modelo del estilo K-NN, Redes Neuronales, etc.

6. DIVISIÓN EN TRAIN-TEST

Quitando la variable FA, nos quedan 36 variables con las que procedemos a realizar una división en train (85%) / test (15%). Para asegurarnos que la muestra es siempre la misma, establecemos una semilla (random_state = 44). Dividiremos la muestra con 36 variables y posteriormente filtraremos las variables para crear la división en el Escenario A.

Como método de entrenamiento y evaluación se utilizará la validación cruzada + Holdout, ofrecida por el paquete Scikit-learn en el objeto GridSearchCV, el cual hace una búsqueda exhaustiva probando todas las combinaciones posibles de parámetros y quedándose con la mejor, a la vez que realiza una validación cruzada para validar el modelo internamente, no haciendo falta dividir a su vez el conjunto de entrenamiento en entrenamiento / validación.

Para mostrar los resultados de los distintos modelos, mostraremos las métricas comentadas en el apartado (2.5) de dos maneras distintas, train vs train y train vs test.

7. ALGORITMO DE REDUCCIÓN DE DIMENSIONALIDAD, T-SNE

Procedemos a utilizar un algoritmo de visualización de datos en alta dimensionalidad, como es el algoritmo t - SNE, el cual tampoco hay que interpretar de manera literal en cuanto a la interpretación de separación de clases, ya que es un algoritmo de

EN PRODUCCIÓN - GALERADAS

tipo “Embedding”, pero puede ayudarnos a interpretar cómo de separadas se encuentran las clases.

En él podemos ver que hay dos clases, pero que no están claramente diferenciadas, algunas instancias rojas sí, incluso algunas azules, pero no podemos asegurar la buena separación de las clases a priori.

Podemos ver como en el escenario A, los estratos de edad se ven marcados más claramente, donde va destacando el color azul conforme aumenta la edad. Esto lo vemos por la leve separación de zonas en la figura 2.9. Mientras que en el escenario B, a pesar de existir dos zonas con mayor densidad de observaciones de la misma clase, no parece existir una frontera de decisión clara.

Esto puede interpretarse como que no tenemos suficientes predictores como para separar las dos clases de una mejor manera.

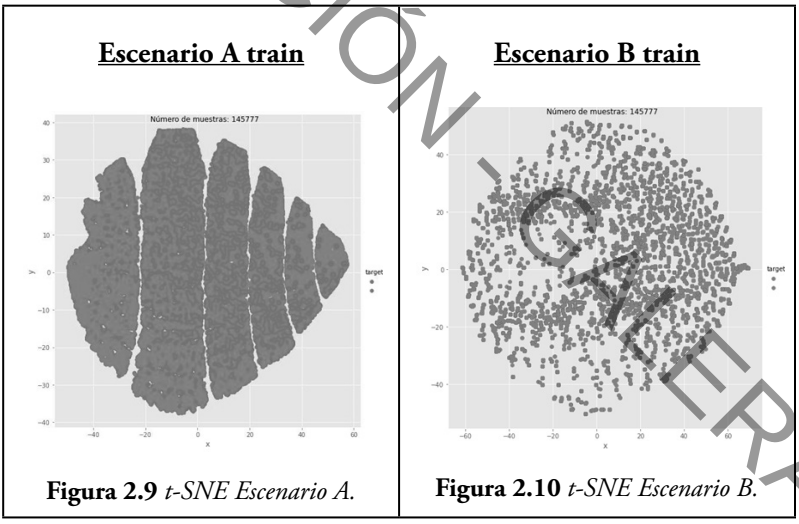


Figura 2.9 *t-SNE Escenario A.*

Figura 2.10 *t-SNE Escenario B.*

EN PRODUCCIÓN - CYTERADAS

CAPÍTULO 3. METODOLOGÍA

1. INTRODUCCIÓN

Hasta ahora, hemos explorado los datos y hemos dejado marcadas algunas pautas del camino a seguir, en cuanto a la métrica que utilizar para poder comparar los modelos. Sin embargo, aún no hemos explorado las opciones que tenemos con un conjunto de datos desbalanceado a la hora de entrenar un modelo, y es lo que se procede a discutir.

2. DATOS DESBALANCEADOS

Cuando nos enfrentamos a un problema desbalanceado, principalmente tenemos tres estrategias de abordaje:

- Ignorar el problema.
- Submuestrear la clase mayoritaria.
- Sobremuestrear la clase minoritaria.

Ninguna de estas formas de abordar el problema es perfecta, todas y cada una de ellas tiene ventajas e inconvenientes. Comentaremos brevemente cada una de estas estrategias y más adelante las terminaremos de reseñar al aplicarlas de diversas maneras usando el paquete **imb-learn**.

En primer lugar, **ignorar el problema**. Si construimos un clasificador usando los datos tal y como se presentan, la predicción

EN PRODUCCIÓN - GALERADAS

del modelo siempre devolverá la clase mayoritaria, es decir, acabará volviéndose un clasificador sesgado.

Submuestrear la clase mayoritaria. Esta es una de las estrategias más sencillas para tratar los datos desequilibrados. Existen diversas técnicas de submuestreo, pero no se ha probado que dichas técnicas aporten una mejora notable respecto a una simple selección aleatoria [27].

Sobremuestrear la clase minoritaria. Puede llevar a problemas de sobreajuste si no se hace de la manera correcta, por ejemplo, si duplicamos alguna de las entradas de la clase minoritaria y procedemos a hacer la validación cruzada, estaremos incorporando el mismo registro a parte de la muestra test y a parte de la muestra de ajuste, por tanto, acabaremos sobreestimando el funcionamiento de nuestro modelo.

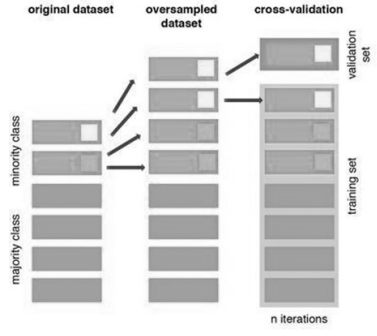


Figura 3.1 *Oversampling - Cross Validation Erróneo.*

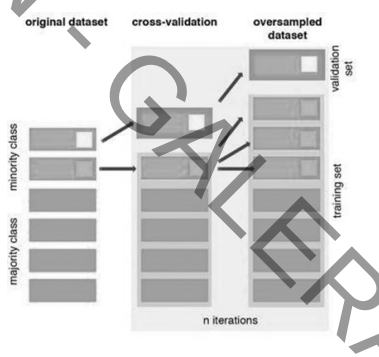


Figura 3.2 *Oversampling - Cross Validation Correcto.*

EN PRODUCCIÓN - GALERADAS

Otras técnicas que mejoran muchos de los resultados del sobremuestreo pueden ser las técnicas de SMOTE, que se basan en interpolaciones de la clase minoritaria.

Los modelos que entrenemos se guardarán en formato pickle (.pkl) formado por el string 'nombre_modelo' + '.pkl'.

Finalmente, nosotros vamos a utilizar únicamente las 2 primeras, por un lado, ignoraremos el problema y vemos qué ocurre con los modelos. Para después hacer un pipeline formado por un submuestreo aleatorio y un modelo.

3. CALIBRACIÓN DEL MODELO [22]

Recordemos que el objetivo del trabajo era estimar la probabilidad que tiene una persona que llega a un centro médico diagnosticado por ictus isquémico de fallecer (en primera instancia).

Estos modelos, tanto en el **escenario A** como en el **B** son capaces de predecir si una persona va a sobrevivir o fallecer teniendo en cuenta que tiene un ictus isquémico, atendiendo a una serie de variables.

Pero la manera en la que lo hace no es otra que estimar la probabilidad de que la observación correspondiente pertenezca a una de las clases y comprobar si sobrepasa un cierto umbral.

Sin embargo, para este problema, no nos interesa tanto si una persona sobrevive o muere (si la probabilidad estimada sobrepasa o no el umbral), sino si la estimación de probabilidad realizada por el modelo se ajusta a la realidad.

¿Hasta qué punto debemos confiar en las predicciones del modelo? Aquí es donde entra el proceso de calibración del modelo.

"Un modelo calibrado es aquel en el que, el valor estimado de probabilidad puede interpretarse directamente como la confianza que se tiene de que la clasificación predicha es correcta. Por ejemplo, si para un modelo de clasificación binaria (perfectamente

EN PRODUCCIÓN - GALERADAS

calibrado) se seleccionan las predicciones cuya probabilidad estimada es de 0.8, en torno al 80% estarán bien clasificadas."

"Todo lo que sabe un modelo es lo que ha podido aprender de los datos de entrenamiento y, por lo tanto, tiene una "visión" limitada." [21]

Definición: "La calibración de un modelo de clasificación consiste en reajustar las probabilidades predichas para que correspondan con la proporción de casos reales observados. En otras palabras, corregir las probabilidades predichas por un modelo cuando este las subestima o sobreestima."

Un modelo está perfectamente calibrado cuando, para cualquier valor p, la clasificación predicha con una confianza (probabilidad) de p, es correcta el p * 100 por ciento de las veces.

$$P(Y = Y|\hat{P} = p) = p, \qquad p \in [0,1]$$

Por ejemplo, si se seleccionan las observaciones cuya probabilidad predicha es $\hat{p}$ = 0.8, es de esperar que el porcentaje de esas observaciones bien clasificadas sea del 80%.

Esto debería ser una curva de calibrado perfecta:

EN PRODUCCIÓN - GALERADAS

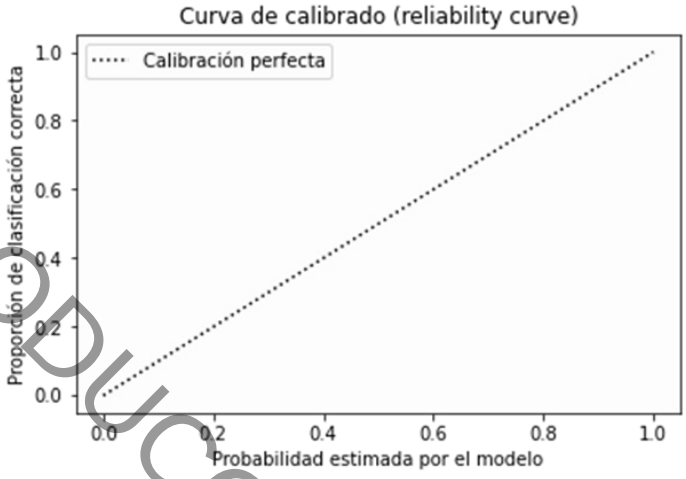


Figura 3.3 Curva de calibrado perfecta.

"El proceso de calibración es importante cuando se quieren emplear las probabilidades asociadas a las predicciones. Si únicamente son de interés las clasificaciones finales, la calibración no aporta valor. Esto último es importante tenerlo en cuenta ya que, calibrar un modelo, puede implicar reducir su porcentaje de clasificaciones correctas (*accuracy*)."

"Cuanto mejor calibrado esté el modelo, más próximos serán los valores de proporción empírica y de confianza, es decir, más se aproxima la curva obtenida a la diagonal. La curva de calibrado queda por encima de la diagonal si el modelo tiende a infravalorar las probabilidades y por debajo si las sobrevalora."

Aunque las curvas de calibración aportan información detallada, es interesante disponer de una métrica que permita cuantificar con un único valor la calidad de calibración del modelo. Brier score es la diferencia cuadrática media (mean squared difference) entre la probabilidad estimada por el modelo y la probabilidad real (1 para la clase positiva y 0 para la negativa). Cuanto menor

EN PRODUCCIÓN - GALERADAS

es su valor, mejor calibrado está el modelo. Esta métrica es adecuada solo para clasificaciones binarias.

Para nosotros, va a ser fundamental no solo el desarrollo del modelo, sino también su proceso de calibrado, con el que veremos cómo se comporta realmente.

4. ALGORITMO GENÉTICO DE SELECCIÓN DE VARIABLES

Recordemos que la base de datos inicial tiene 37 variables. Los primeros modelos, desarrollados sobre el **escenario A**, están basados en variables seleccionadas únicamente por criterio experto y además son variables pueden tomarse de forma directa cuando el paciente llegue al centro de salud.

Pero ¿Sería posible mejorar dichos resultados obtenidos por estos modelos usando únicamente técnicas de Ciencia de Datos? Introduciendo otros predictores que requieran el ingreso del paciente, algunas pruebas específicas (teniendo también cuidado de no introducir multicolinealidad en el modelo).

Lo que haremos será aplicar un algoritmo para selección genética de predictores propuesto por Joaquín Amat [20] y ver cómo se comportan los modelos propuestos hasta ahora. Los algoritmos genéticos son solo una de las muchas estrategias que existen para seleccionar los predictores más relevantes, y no tiene por qué ser la más adecuada en todos los escenarios. Existen estrategias iterativas Stepwise selection, modelos como Random Forest, Boosting y Lasso capaces de excluir automáticamente predictores, y técnicas de reducción de dimensión como PCA y t-SNE.

"Los algoritmos genéticos son métodos de optimización heurística que, entre otras aplicaciones, pueden emplearse para encontrar la combinación de variables que consigue maximizar la capacidad predictiva de un modelo. Su funcionamiento está inspirado en la **teoría evolutiva de selección natural** propuesta por

EN PRODUCCIÓN. GALERADAS

Darwin y Alfred Russel: los individuos de una población se reproducen generando nuevos descendientes, cuyas características, son combinación de las características de los progenitores (más ciertas mutaciones). De todos ellos, únicamente los mejores individuos sobreviven y pueden reproducirse de nuevo, transmitiendo así sus características a las siguientes generaciones."

Objetivo: Encontrar la combinación de predictores que da lugar al mejor modelo. **Configuración**: En nuestro caso, estamos ante un problema de clasificación, optimizaremos el modelo randomforest mediante la métrica f1, usando como máximo 50 generaciones, probabilidad de mutación de 0.1, método de selección "ruleta" (la probabilidad de que un individuo sea seleccionado es proporcional a su fitness relativo) y método de cruce "uniforme" (el valor que toma cada posición del nuevo individuo se obtiene de uno de los dos parentales), con una validación cruzada interna.

Funcionamiento del algoritmo:

1) En primer lugar, se crea una población inicial aleatoria de p individuos. En este caso, cada individuo representa una combinación de predictores.
2) Calcular el fitness de cada individuo de la población (calcular con una métrica de calidad del modelo). Mayor fitness cuanto mayor sea la métrica.
3) Crear una población vacía y repetir los pasos hasta que se hayan creado p nuevos individuos.

    A. Seleccionar dos individuos de la población existente, donde la probabilidad de selección es proporcional al fitness de los individuos.
    B. Cruzar los dos individuos seleccionados para generar un nuevo descendiente (crossover).

EN PRODUCCIÓN - GALERADAS

C. Aplicar un proceso de mutación aleatorio sobre el nuevo individuo.
D. Añadir el nuevo individuo a la nueva población.

4) Reemplazar la antigua población de la nueva.
5) Si no se cumple el criterio de parada, volver al paso 2.

El término individuo hace referencia a una de las posibles soluciones del problema que se quiere optimizar, en nuestro caso, la selección de los mejores predictores.

Tras generar cada nuevo individuo de la descendencia, este se somete a un proceso de mutación en el que, cada una de sus posiciones, puede verse modificada con una probabilidad p.

Este paso es importante para añadir diversidad al proceso y evitar que el algoritmo caiga en mínimos locales porque todos los individuos sean demasiado parecidos de una generación a otra.

Resultados:

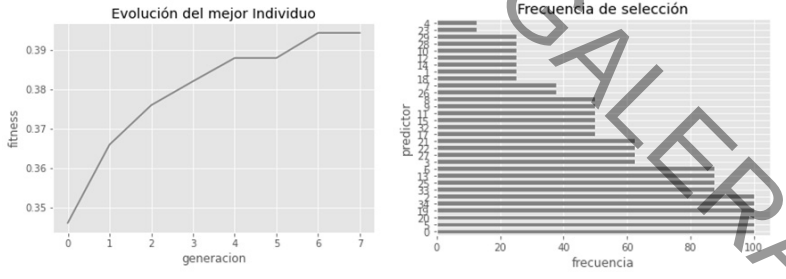


Figura 3.4 *Algoritmo genético selección de predictores.*

EN PRODUCCIÓN - GALERADAS

```
   Algoritmo detenido en la generación 7 por falta
cambio absoluto mínimo de 0.01 durante 5 genera-
ciones consecutivas.
   ----------------------------------------
   Optimización finalizada 2021-06-29 10:30:40
   ----------------------------------------
   Duración optimización: 2513.0435683727264 Núme-
ro de generaciones: 8
   Predictores óptimos: [ 0 1 2 5 6 9 11 12 13 14
15 17 18 19 20 22 25 27 28 33 34]
   Valor métrica óptimo: 0.39430238978439613
```

5. CONCLUSIÓN

En el escenario B, partiremos directamente del uso de estos 20 predictores que son óptimos para el randomforest entrenado con la métrica f1, en lugar de usar las 37 variables de las que partimos.

Además, probaremos hacer filtrados según el criterio de importancia asignado por el modelo randomforest y también una evaluación de modelo a la vez que se hace una eliminación recursiva de predictores.

Tanto en el escenario A como en el escenario B, probaremos un randomforest y una regresión logística, tanto balanceados en primer lugar, como en un pipeline con un submuestreo aleatorio previo. Para entrenar y ver cómo se comportan los modelos usaremos las métricas que han sido discutidas.

EN PRODUCCIÓN · GALERADAS

EN PRODUCCIÓN - GALERADAS

CAPÍTULO 4. ESCENARIO A

1. INTRODUCCIÓN

Este es el primer escenario que vamos a desarrollar, principalmente porque es más sencillo, y segundo porque el otro modelo englobará estas variables y parte del razonamiento aplicado en este caso podrá extrapolarse al escenario B.

Procedemos entonces a evaluar los modelos, a compararlos, calibrarlos, y tratar de estudiar la explicabilidad / interpretabilidad de estos.

Nota:

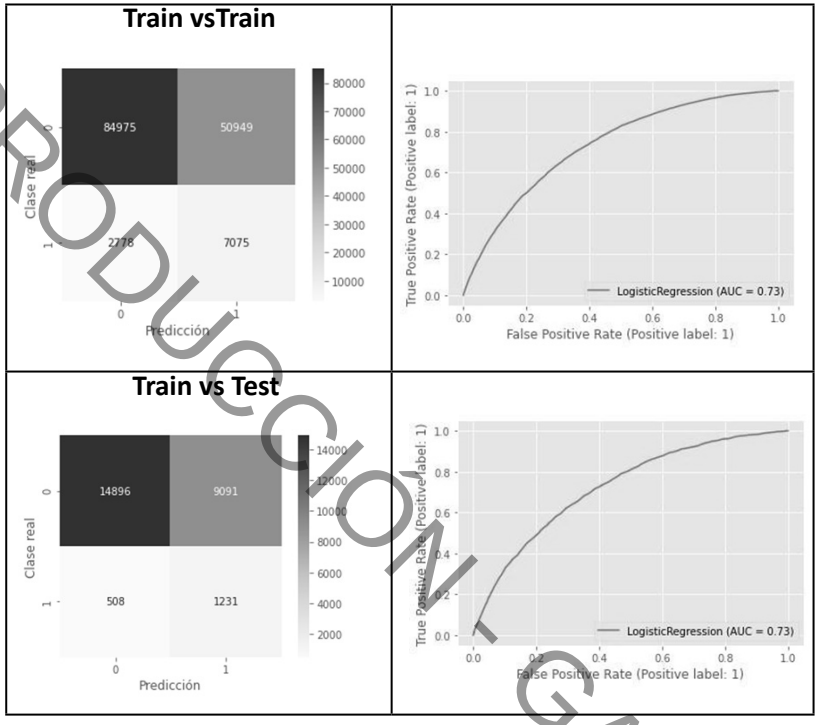
Recordemos que todos los modelos se van a entrenar usando la misma metodología, validación cruzada interna con una búsqueda exhaustiva del espacio de hiperparámetros, tratando de optimizar la métrica F1-Score.

Y, aunque pueda variar un poco la muestra de training (debido a que algún modelo estará compuesto por un pipeline, donde la primera parte sea hacer un submuestreo de la clase mayoritaria para entrenar con un 50-50), todos los modelos usarán la misma muestra test para poder comparar los resultados.

2. REGRESIÓN LOGÍSTICA

a. Regresión logística balanceada

Se ha utilizado una penalización inversamente proporcional al número de casos de la clase minoritaria.



EN PRODUCCIÓN - GALERADAS

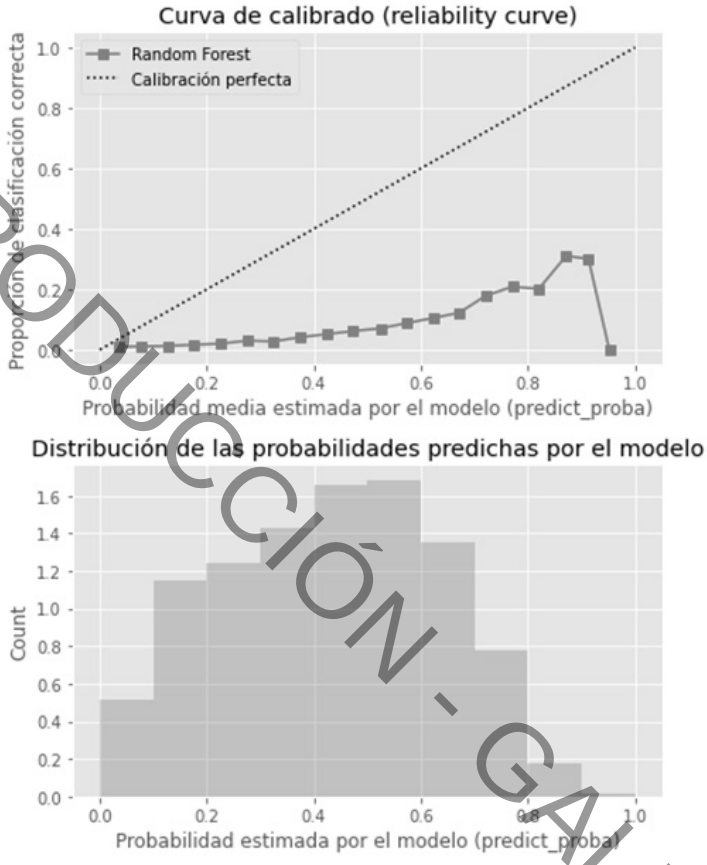


Figura 4.1 Curva de calibrado de la Regresión Logística Balanceada.

Por lo estudiado en la sección de calibración, sabemos que el modelo está sobreestimando las probabilidades, de ahí la baja precisión que estamos obteniendo y el gran número de falsos positivos. Recordemos que esto es preferible a la clasificación de falso negativo, pero básicamente es una cuestión de umbral, y lo que nos interesa es cuánto se ajusta dicha curva a la "perfecta".

```
Brier score = 0.21518750972103237
```

EN PRODUCCIÓN - GALERADAS

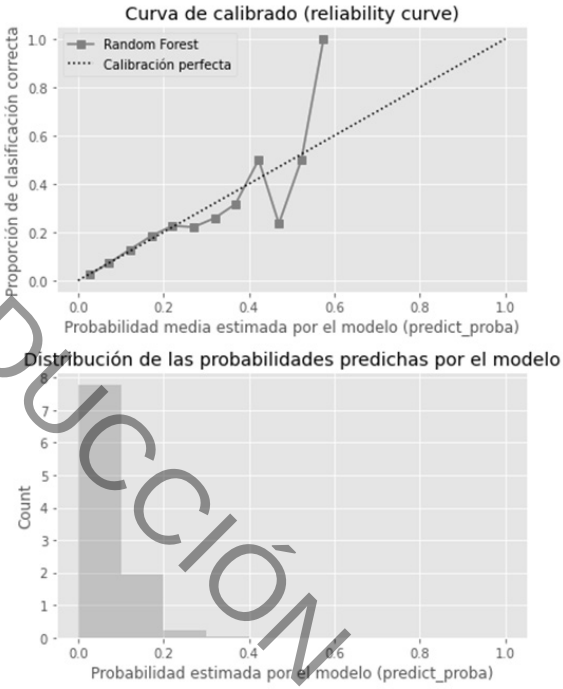


Figura 4.2 Modelo calibrado.

```
Brier score = 0.06001531521008679
```

Aunque la métrica de Brier sea menor, podemos observar que la calibración de este modelo es bastante deficiente debido a la gran cantidad de casos negativos del modelo, y acaba siendo un clasificador trivial.

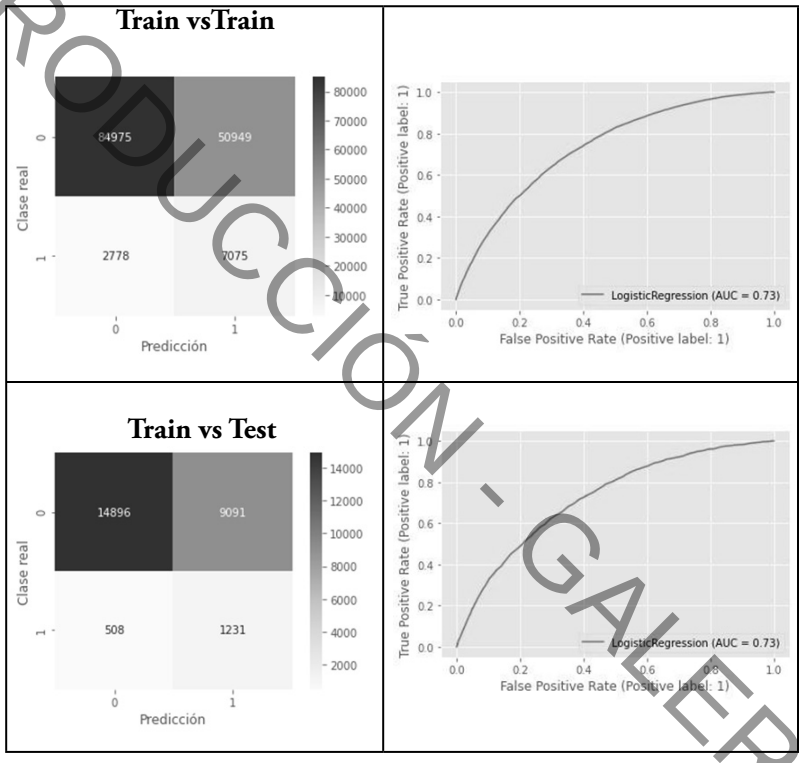
b. Pipeline Submuestrear + Regresión Logística

A pesar de haber realizado un submuestreo aleatorio sobre la clase mayoritaria, el modelo no consigue aprender a discriminar mejor que el visto en el punto anterior.

EN PRODUCCIÓN - GALERADAS

Vemos entonces que algo extraño está ocurriendo, probemos con un random forest usando las mismas técnicas comentadas y estudiemos la explicabilidad e interpretabilidad.

3. RANDOM FOREST BALANCEADO



EN PRODUCCIÓN - GALERADAS

	Accuracy	**Balanced accuracy**	**F1-Score**	AUROC	Precisión	**Recall**
Balanced Logistic Regression	0.633756	0.671532	0.208851	0.732837	0.122282	0.715214
Random Undersample + Logistic Regression	0.637426	0.671994	0.209760	0.732947	0.123003	0.711966
Balanced random forest	0.651886	0.673818	0.213537	0.733811	0.126013	0.699178
Under-sampling + Random forest	0.639648	0.661089	0.204715	0.712233	0.120337	0.685882

Tabla 2. Resultados de los modelos.

Como podemos observar, tenemos unos resultados similares tanto en el random forest como en la regresión logística, a pesar ser un modelo bastante más complejo que el otro. Tratemos de echar un vistazo a lo que está ocurriendo.

5. EXPLICABILIDAD E INTERPRETABILIDAD

Vamos a intentar echar un vistazo a algunos casos donde el modelo clasifique como fallecido, cuando realmente no fallece, tratando de entender por qué el modelo está clasificando de esa manera.

En primer lugar, veamos la asignación de importancia a los predictores por parte de nuestro modelo:

EN PRODUCCIÓN - GALERADAS

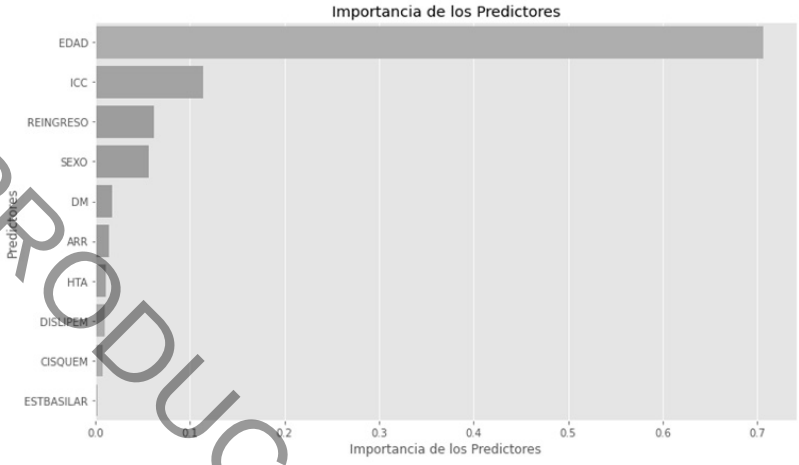


Figura 4.3 Importancia de los predictores según RandomForest.

Vemos que existe una gran diferencia entre el predictor Edad y el resto, y probablemente debido a ello, la gran cantidad de falsos positivos. Es sabido que la importancia asignada por el modelo RandomForest tiende a sobreestimar las variables cuantitativas, y a modo de corrección se suele aplicar la importancia por permutaciones:

EN PRODUCCIÓN - GALERADAS

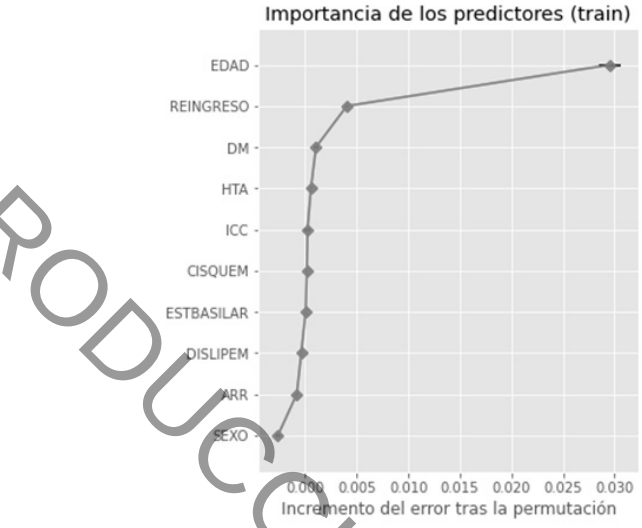
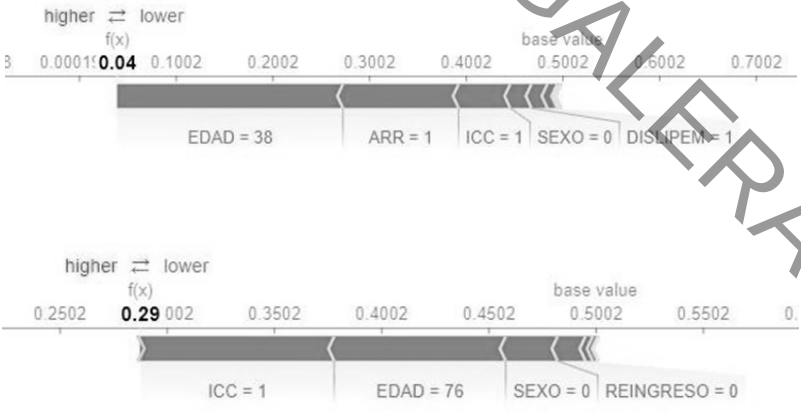


Figura 4.4 Test de importancia por permutaciones.

Si echamos un vistazo a algunas de las probabilidades estimadas por el modelo y nos vamos a los percentiles [0, 25, 50, 75, 100] podemos ver algunas diferencias:



EN PRODUCCIÓN. GALERADAS

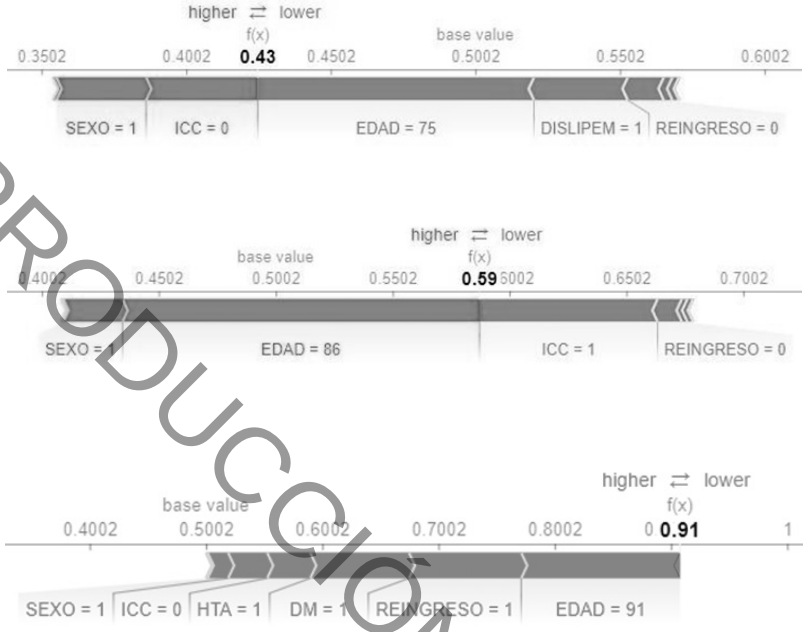


Figura 4.5 Valores SHAP.

Se ha demostrado que la HTA (hipertensión) es un factor protector frente al ictus isquémico, pero el modelo no parece detectarla, al menos en estos ejemplos, o si la detecta la tasa negativamente (puede deberse a la edad avanzada). Vemos también como la variable ICC la toma como positiva, disminuyendo la probabilidad. Además, el sexo también es un factor bastante marcado junto a la Edad.

EN PRODUCCIÓN - GALERADAS

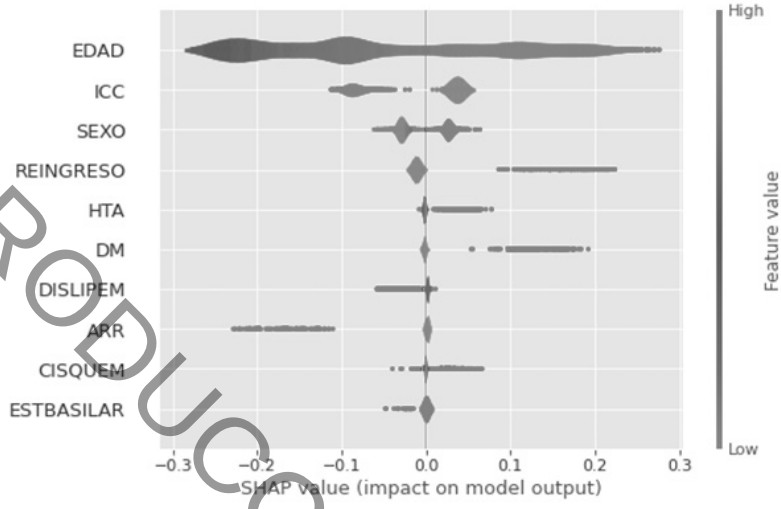


Figura 4.6 Influencia de los predictores según valores SHAP.

Aquí vemos cómo se comporta el modelo respecto a las observaciones de la muestra test y a los valores SHAP. En esta imagen vemos el azul como la disminución de la probabilidad de fallecer (aporte negativo) y la parte roja como un aumento de la probabilidad (aporte positivo). Así vemos cómo la Edad es mayoritariamente para negativo, en ICC hay cierto balance, aunque cuando afecta de manera negativa lo hace de manera mayor a cuando lo hace de manera positiva, el reingreso mayoritariamente afecta si es de manera negativa…

Si vemos qué ocurre con los falsos positivos:

EN PRODUCCIÓN - GALERADAS

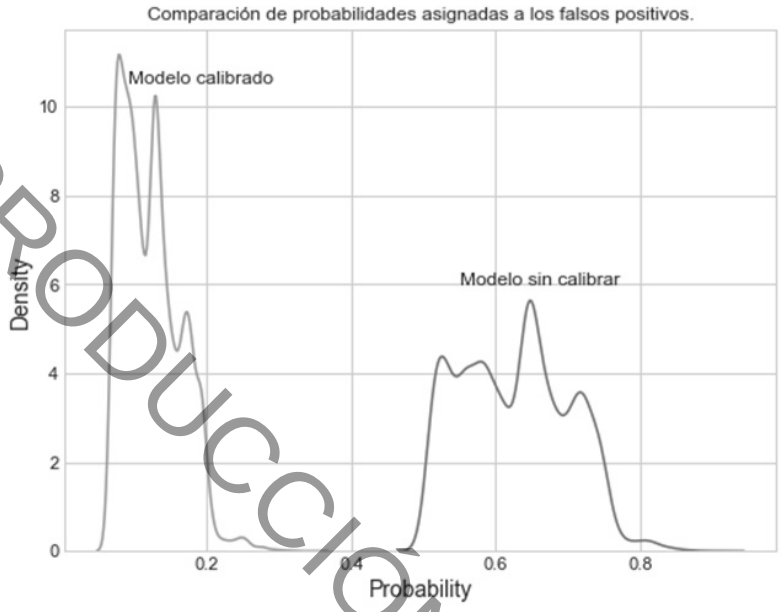


Figura 4.7 Distribución de probabilidad de falsos positivos.

Aquí vemos que, si miramos únicamente a los falsos positivos, el modelo calibrado es obvio que consigue neutralizar esos casos. Sin embargo, esta imagen está sesgada, ya que únicamente se está mostrando la probabilidad sobre los falsos positivos, si vemos cómo se comportan estos modelos frente a la muestra completa:

EN PRODUCCIÓN - GALERADAS

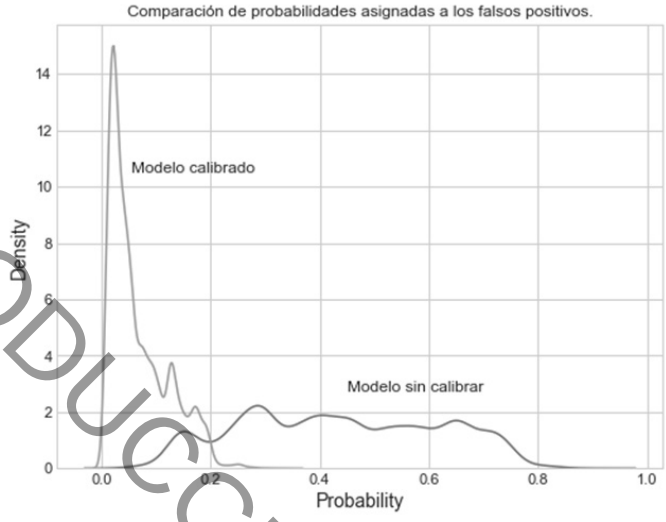


Figura 4.8 Distribución de probabilidades de modelo calibrado y no calibrado.

Podemos observar entonces que el modelo calibrado está sesgado de alguna manera por la muestra desbalanceada, y acaba siendo de alguna manera un clasificador trivial, el modelo sin calibrar ha captado un poco mejor la sensibilidad.

Es interesante ver cómo la mayoría de estos casos presentan una probabilidad entre 0.5 y 0.76, por tanto, se podría corregir subiendo un poco el umbral de clasificación. Aunque lo que nos incumbe es si la probabilidad asignada tiene sentido, veamos algún ejemplo.



EN PRODUCCIÓN - GALERADAS

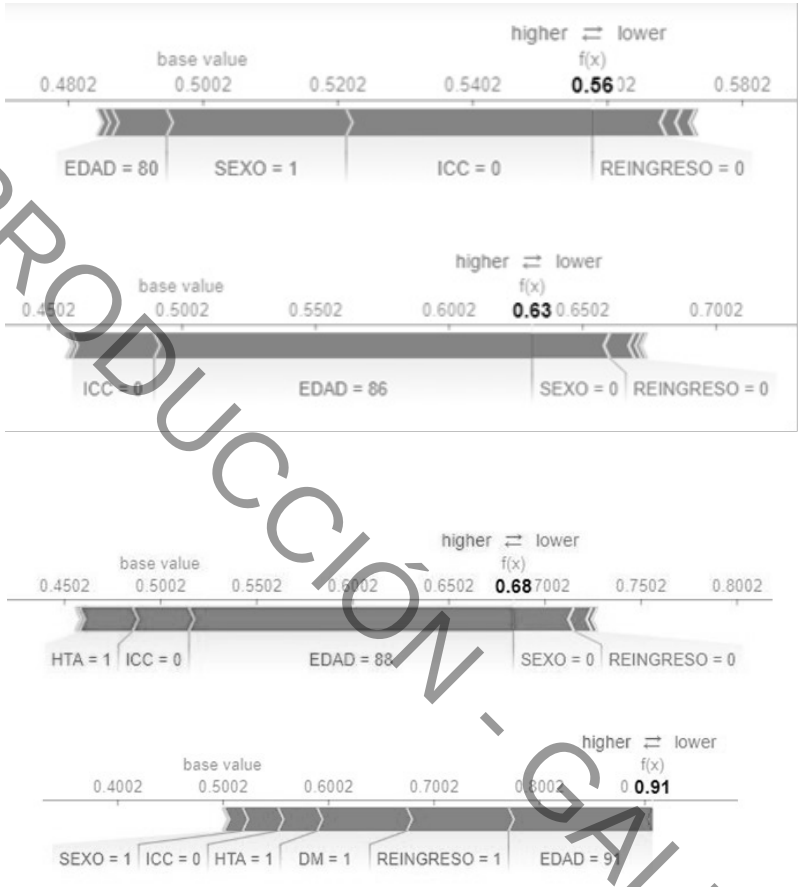


Figura 4.9 *Shap values de los falsos positivos.*

A primera vista no parece que estas probabilidades estén sobreestimadas, lo que ocurre es que tenemos muchos casos, que, aun siendo pacientes donde muchos con esas mismas características han fallecido, pueden no fallecer, y eso causa que aumente la cantidad de falsos positivos. Mayormente estimadas de esta manera por la Edad.

EN PRODUCCIÓN - GALERADAS

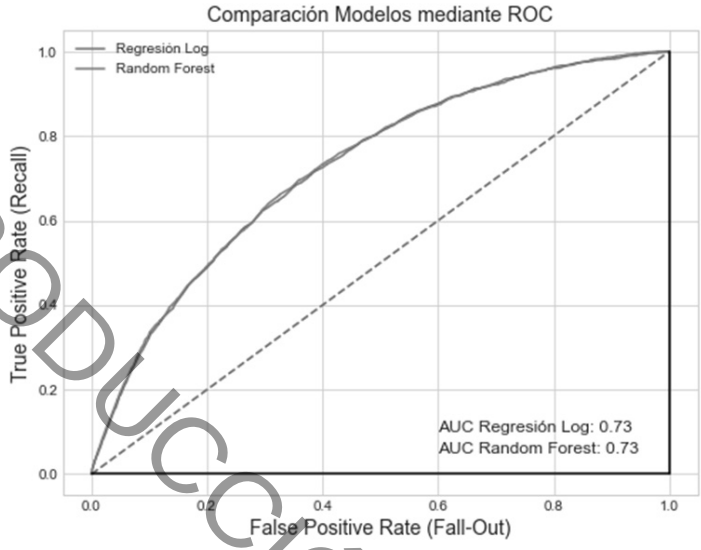


Figura 4.10 *Comparación de Modelos por ROC*

6. SELECCIÓN RECURSIVA DE PREDICTORES

Una pregunta que podríamos hacernos es la siguiente; A lo largo de esta sección hemos visto cómo los primeros predictores, EDAD, ICC, REINGRESO, SEXO tienen una importancia muy por encima del resto de predictores. ¿Cómo se vería afectado el modelo si empezamos a eliminar algunos de los predictores que tienen menor importancia?

EN PRODUCCIÓN GALERADAS

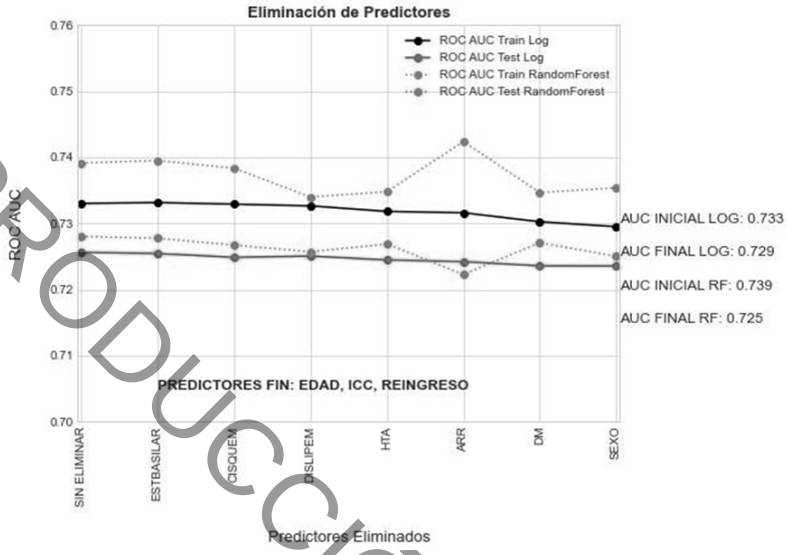


Figura 4.11 Poder ROC mediante eliminación recursiva de predictores.

7. CONCLUSIONES DEL CAPÍTULO

Como hemos visto hasta ahora, con los dos modelos iniciales, conseguimos un poder ROC medianamente considerable, pero qué ocurre con los resultados:

- La medicina no es una ciencia exacta, y según hemos visto, los falsos positivos tienen bastante sentido en cuanto a la asignación de probabilidad estimada, por ejemplo, en el último caso, persona de 91 años, mujer, con DM, REINGRESO, HTA, y, sin embargo, no fallece. Eso puede ser debido a que en el momento en que se tomaron los datos no había fallecido aún o que se salvan casos incluso en las situaciones más complicadas debido a los avances científi-

EN PRODUCCIÓN - GALERADAS

cos. Por eso tenemos tal grado de "error" y en parte no es malo que el modelo sobreestime un poco las probabilidades, que hemos visto que tienen cierto sentido, pero que, para calibrarlo, tendríamos que reducir de alguna manera las probabilidades estimadas por el modelo.

- Con estos dos primeros modelos, la tasa de falsos positivos es muy alta, debido a que es preferible este caso al falso negativo en este problema concreto.
- Se han probado otros modelos como VotingClassifiers, k-NN, redes neuronales y XGBoost, y no se obtienen mejores resultados. De hecho, la cantidad de falsos negativos aumenta considerablemente.
- La manera de mejorar estos clasificadores que suelen ser más potentes en cuanto a complejidad y coste computacional suele ser balancear la muestra hasta un 50-50, y aunque mejoran los resultados en cuanto a la clasificación, el poder ROC no acaba de subir por encima de 0.75.

Nota:

Tampoco esperamos que la ROC suba mucho más de 0.75 debido a la cantidad de "ruido" que tiene la base de datos, pero veremos si podemos mejorar este primer resultado, sin hacer uso del clasificador "perfecto".

También usaremos toda la base de datos en la cual aplicaremos un algoritmo genético para selección de predictores y desarrollo de un modelo post-ingreso, donde las variables que incluiremos no serán escogidas por criterio experto.

Además, podríamos plantearnos eliminar algunas de las variables que tienen menor importancia, y el modelo no se vería afectado en cuanto a capacidad de discriminación.

EN PRODUCCIÓN - CALERADAS

EN PRODUCCIÓN - GALERADAS

CAPÍTULO 5. ESCENARIO B

1. INTRODUCCIÓN

Al final del capítulo 3, obtuvimos los resultados del lanzamiento del algoritmo genético de selección de variables. Obtuvimos un individuo óptimo formado por 20 variables que optimiza los resultados de un randomforest entrenado con F1-Score.

Para empezar, nos quedaremos con dichas variables y procederemos a lanzar los modelos del escenario A.

2. Resultado de los modelos

	Accuracy	**Balanced accuracy**	**F1-Score**	AUROC	Precisión	**Recall**
Balanced Logistic Regression	0.680313	0.703465	0.235936	0.770848	0.140699	0.730236
Random Undersample + Logistic Regression	0.683414	0.703810	0.236999	0.770645	0.141567	0.727394
Balanced random forest	0.688579	0.716228	0.245163	0.788986	0.146609	0.7482008
Under-sampling + Random forest	0.671052	0.679860	0.220933	0.737112	0.131527	0.690045

Importancia de predictores por RandomForest:

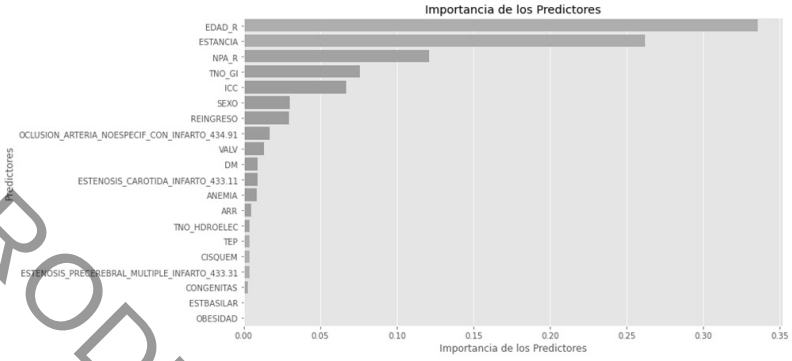


Figura 5.1 Importancia predictores escenario B

La importancia de la permutación de la característica se define como la disminución de la puntuación de un modelo cuando se baraja aleatoriamente un único valor de la característica. Este procedimiento rompe la relación entre la característica y el objetivo, por lo que la disminución de la puntuación del modelo es indicativa de cuánto depende el modelo de la característica. Esta técnica se beneficia de ser agnóstica al modelo y puede calcularse muchas veces con diferentes permutaciones de la característica [28].

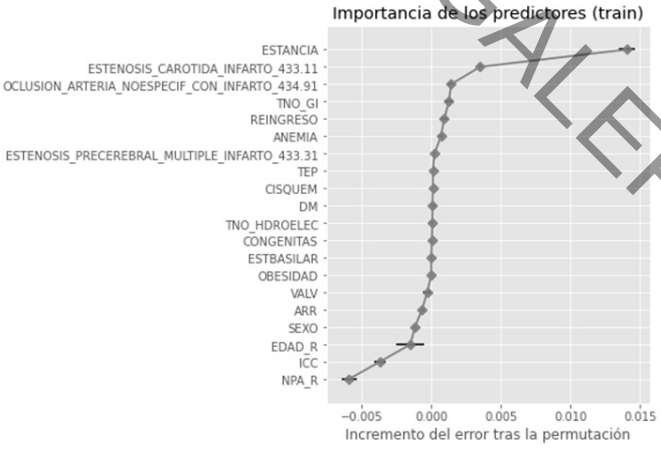


Figura 5.2 Importancia por permutación.

EN PRODUCCIÓN - GALERADAS

Si comparamos entonces los modelos mediante la curva ROC en el conjunto de Test:

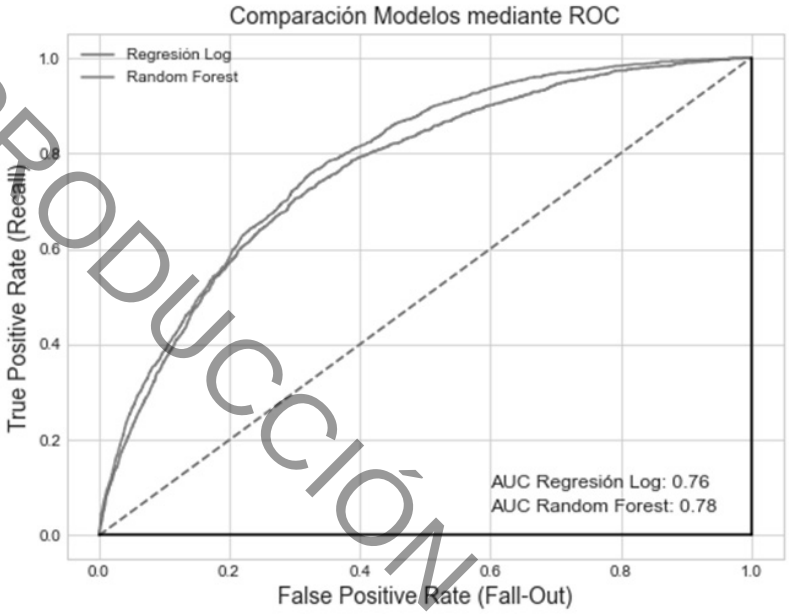


Figura 5.3 Comparación modelos escenario B.

Hemos obtenido un modelo post-ingreso (muchas de las variables requieren del ingreso del paciente para poder ser obtenidas) que es capaz de discriminar mejor que el que teníamos de partida.

Además, es posible que se puedan reducir la cantidad de predictores del modelo sin que ello conlleve un empeoramiento de los resultados Figura 5.4.

EN PRODUCCIÓN - GALERADAS

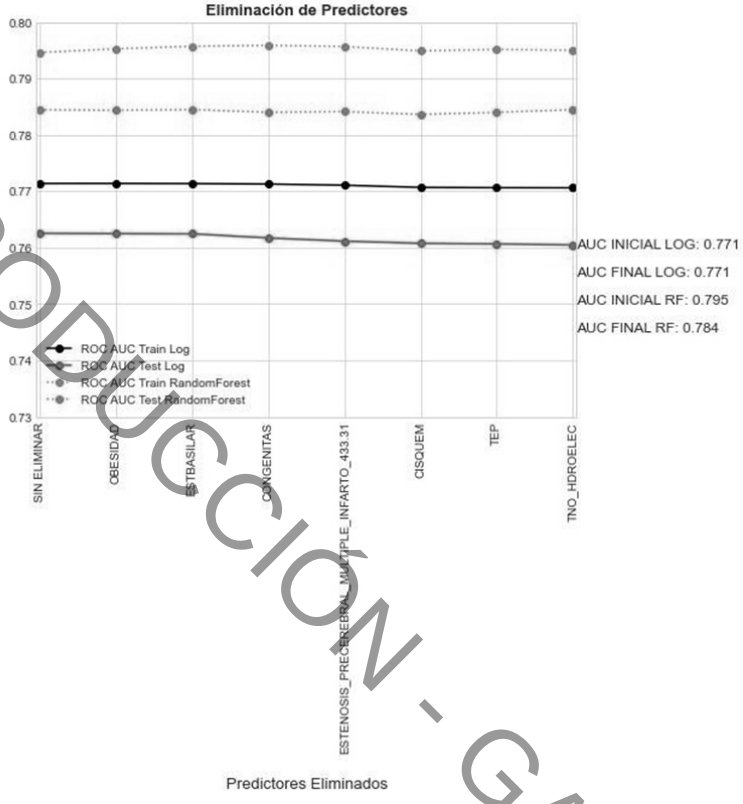


Figura 5.4 Eliminación recursiva de predictores sin importancia.

Si ahora relanzamos el modelo partiendo de selección de predictores cuya importancia es superior a 0.01, se obtiene la siguiente gráfica Figura 5.5.

EN PRODUCCIÓN - GALERADAS

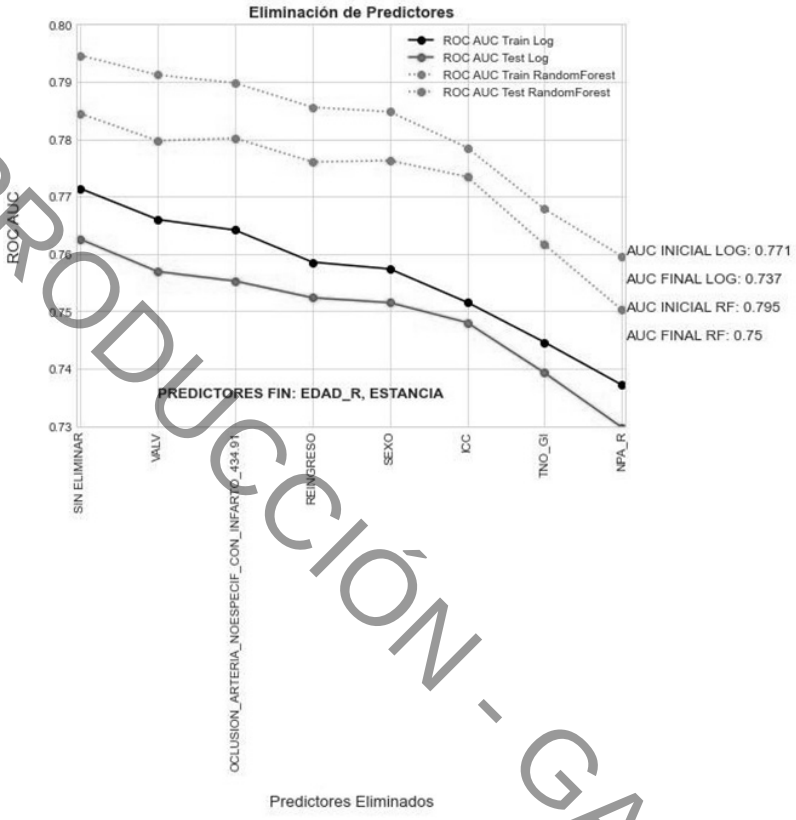


Figura 5.5 Eliminación recursiva de predictores importantes.

Vemos cómo ahora sí, el modelo empieza a sufrir mucho la eliminación de variables, por tanto, podría ser un buen criterio el seleccionar predictores cuya importancia asignada por el randomforest superase cierto umbral.

EN PRODUCCIÓN - GALERADAS

3. CONCLUSIONES DEL CAPÍTULO

A lo largo de este capítulo, se ha abordado la base de datos en su completo, y aplicando únicamente técnicas de Ciencia de datos (al contrario que en el otro escenario donde los predictores han sido escogidas por criterio experto), se han obtenido unos resultados que mejoran lo que habíamos logrado usando únicamente las variables del escenario A.

Posteriormente, se ha utilizado una eliminación recursiva de predictores basándose en la importancia asignada por el RandomForest, y hemos visto que se pueden eliminar aquellas que tienen importancia menor a 0.01, sin que esto afecte demasiado al modelo. Por tanto, podemos obtener un modelo en el escenario B, que teniendo aproximadamente la misma cantidad de variables (pero diferentes), consigue una mejor discriminación de las dos clases.

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

CAPÍTULO 6. CONCLUSIONES Y TRABAJO FUTURO

1. INTRODUCCIÓN

A lo largo de este capítulo se entenderá por aplicación, la arquitectura integrada por un sistema de peticiones (API), un modelo que haga inferencia sobre dichas peticiones, y una base de datos que acabe registrando todo el proceso.

El desarrollo futuro de este proyecto pasa por la puesta en producción, monitorización y recalibrado de los modelos. Para ello, en primer lugar, necesitaremos montar una estructura de soporte de peticiones, donde a su vez pueda desplegarse el modelo. Una vez montada dicha estructura (aplicación), será de vital importancia la colaboración por parte de los sanitarios que utilicen dicha aplicación para recoger nuevas observaciones, usarla para producir inferencia sobre éstas, y en un futuro, contrastar las estimaciones realizadas con las observadas.

2. ¿QUÉ MODELO PONER EN PRODUCCIÓN?

A lo largo de los 2 capítulos anteriores hemos visto las diferencias entre los distintos modelos (regresión logística y random forest), pero además, hemos observado también sus distintas versiones (sin calibrar y calibrado).

Por un lado, hemos observado que los clasificadores calibrados tienden a acertar más, pero no era el accuracy la métrica por la que estábamos escogiendo los modelos.

Ya que, si atendemos a las estimaciones realizadas por estos modelos, obtendremos muchos casos negativos, tiende a estandarizar más las probabilidades según la verdadera distribución de clases. Pero tengamos en cuenta que la distribución está claramente desbalanceada. Por tanto, si usamos este tipo de modelos, tendremos más "acierto" en cuanto a accuracy se refiere, pero disminuiremos la asignación de riesgo a los diversos pacientes. Lo que hemos visto, que sí que es asignado de una manera más correcta con el random forest sin calibrar.

Una asignación de riesgo considerable a un paciente recaerá en una atención más exhaustiva y precoz, que finalmente se inclinará por un falso positivo, es decir, se le habrá salvado la vida al paciente.

¿Con ello queremos decir que a todos los pacientes haya que tratarlos sobreestimando el riesgo? Pues claramente no, recordemos que con esto se prevé optimizar costo-eficiencia, por tanto, podría ser interesante hacer alguna ligera ponderación de las probabilidades estimadas por el modelo sin calibrar.

3. SERVICIOS AWS

Todo el trabajo se ha realizado usando un entorno interactivo como es Jupyter Notebooks, y en un futuro, el objetivo sería trasladarlo a los servicios de AWS.

Podríamos montar la aplicación, en primer lugar, usando SageMaker, ya que facilita mucho el proceso de puesta en producción del modelo.

Por tanto, únicamente necesitaríamos hacer uso de un bucket de S3 donde poder almacenar los datos con los que entrenaremos el modelo, el modelo, y las inferencias realizadas por

EN PRODUCCIÓN - GALERADAS

éste. Aunque la idea es usar alguna base de datos, como pueda ser DynamoDB de cara a una mayor escala de la puesta en producción del modelo. Ya que permite desarrollar, entrenar y desplegar bajo un control de versiones y un hosting gestionado y autoescalable.

Desde otro punto, se podrían implementar Lambdas para que, cada cierto tiempo, se fuese reentrenando el modelo con datos nuevos, montando una arquitectura que no dependiera de SageMaker. Haciendo únicamente uso de S3, API Gateway, Lambdas y DynamoDB o RDS.

4. CONCLUSIONES DEL TFM

A lo largo de este TFM, se ha hecho un recorrido por distintas técnicas de Ciencia de Datos para abordar un problema de estas características. Es la primera vez que realizo un trabajo de estas características de manera un poco más profesional, que se espera que tenga cierta aplicación y, además sobre un problema que está abierto.

Con ello me refiero a que no existe aún ningún modelo que de forma completamente determinista sea capaz de estimar un riesgo de mortalidad de manera perfecta.

La idea también ha sido acercar a la Medicina algunas técnicas de Ciencia de Datos que no son tan comunes aún, como es el uso de modelos más complejos que los modelos lineales, el estudio de explicabilidad/interpretabilidad de estos, la selección de predictores usando diversas técnicas

De cara al futuro esperamos reunir nuevos datos que hayan sido tomados de manera más rigurosa (datos obtenidos por médicos que colaboran en el proyecto), y poder realizar un estudio similar al realizado con esos nuevos datos. También que sirva como una validación externa de los modelos propuestos hasta ahora.

EN PRODUCCION - GALERADAS

BIBLIOGRAFÍA

1. Memoria de Registro del Modelo a Validar, Modelo predictivo para la mortalidad hospitalaria en el ictus isquémico no lisado. Juan Manuel García Torrecillas.
2. Ministerio de Sanidad y Política Social. Estrategia en Ictus del Sistema Nacional de Salud; 2009. 163 p.
3. Feigin VL, Lawes CM, Bennett DA, Anderson CS. Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. The Lancet Neurology. 2003;2(1):43-53.
4. Arboix A, García-Eroles L, Comes E, Oliveres M, Targa C, Balcells M, et al. Importancia del perfil cardiovascular en la mortalidad hospitalaria de los infartos cerebrales. Revista espanola de cardiologia. 2008;61(10):1020-9.
5. Smith EE, Shobha N, Dai D, Olson DM, Reeves MJ, Saver JL, et al. A risk score for in-hospital death in patients admitted with ischemic or hemorrhagic stroke. Journal of the American Heart Association. 2013;2(1): e005207.
6. Whisnant JP. Modeling of risk factors for ischemic stroke. The Willis Lecture. Stroke; a journal of cerebral circulation. 1997;28(9):1840-4.
7. Anon. n.d. "El Conjunto Mínimo Básico de Datos (CMBD) Como Fuente Para La Investigación En Medicina Interna. - Por Una Medicina Interna de Alto Valor."
8. López González, Ramiro A. n.d. XXIX CONGRESO NACIONAL DE LA SEMI. A CORUÑA, CMBD.
9. "Tablas de mortalidad - Instituto Nacional de Estadística. (National ..." 2015. 26 Jun. 2016.
10. Van Der Maaten, Laurens, and Geoffrey Hinton. 2008. Visualizing Data Using T-SNE. Vol.9.
11. Anon. n.d. "An Illustrated Introduction to the T-SNE Algorithm – O'Reilly." Retrieved June 23, 2021.
12. Anon. n.d. "T-SNE – Laurens van Der Maaten." Retrieved June 23, 2021.

EN PRODUCCIÓN - GALERADAS

13. Wattenberg, Martin, Fernanda Viégas, and Ian Johnson. 2017. "How to Use T-SNE Effectively.
14. Altmann, André, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. "Permutation Importance: A Corrected Feature Importance Measure." Bioinformatics 26(10):1340– 47.
15. Anon. n.d. "Permutation Importance vs Random Forest Feature Importance (MDI) — Scikit-Learn 0.24.2 Documentation." Retrieved June 25, 2021.
16. Breiman, Leo. 2001. "Random Forests." Machine Learning 45(1):5–32.
17. Anon. n.d. "3.2. Tuning the Hyper-Parameters of an Estimator — Scikit-Learn 0.24.2 Documentation."Retrieved June 25, 2021.
18. Anon. n.d. "Sklearn.Model_selection.GridSearchCV — Scikit-Learn 0.24.2 Documentation." Retrieved June 25, 2021.
19. Anon. n.d. "Welcome to the SHAP Documentation — SHAP Latest Documentation." Retrieved June 28, 2021.
20. Anon. n.d. "Seleccion de Predictores Mediante Algoritmo Genetico." Retrieved June 29, 2021.
21. "Calibrar Modelos de Machine Learning."
22. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers, B. Zadrozny & C. Elkan, ICML 2001.
23. Transforming Classifier Scores into Accurate Multiclass Probability Estimates, B. Zadrozny & C. Elkan, (KDD 2002).
24. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, J. Platt, (1999).
25. Predicting Good Probabilities with Supervised Learning, A. Niculescu-Mizil & R. Caruana, ICML 2005.
26. Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 200 International Conference on Artificial Intelligence (IC- AI'2000): Special Track on Inductive Learning Las Vegas, Nevada.

EN PRODUCCIÓN, GALERADAS

ÍNDICE

Agradecimientos 9
Autores de la obra 11
Prólogo 13
Funding 17
A modo de introducción… 19
Un libro que marca la diferencia 19

1. Neuroepidemiología del ictus avanzada en nuestro país. Condicionantes y medios. 21
1. Introducción 23
2. Metodología de la epidemiología del ictus. 25
2.1 ¿Por qué medir la incidencia del ictus? 25
2.2 Requisitos para una correcta evaluación de la incidencia 26
2.3. Estudios epidemiológicos de base poblacional 29
2.4. Epidemiología basada en estimaciones poblacionales. 36
3. Epidemiología del ictus en España 42
4. Discusión 47
5. Conclusiones 49
6. Bibliografía 50

2. Mortalidad del ictus antes y después de la era del tratamiento reperfusor (fibrinolisis/trombectomía) 55
Resumen 57
1. Introducción 58

EN PRODUCCIÓN - GALERADAS

2. La mortalidad del ictus en la era prefibrinolisis.
El rol de las Unidades de Ictus: el primer paso 61
3. La influencia de la fibrinolisis en la mortalidad del ictus 64
4. La revolución de la trombectomía mecánica en el ictus..... 67
Bibliografía 70

3. Cuidados post ictus en Atención Primaria de Salud 71
1. Introducción 73
2. Valoración en el entorno comunitario 75
3. Intervenciones post ictus en Atención Primaria 79
4. Conclusiones..... 82
Bibliografía 83

4. Análisis de la captación de financiación de proyectos sobre el ictus en Andalucía en el período 2019-2021. 85
Introducción 87
Bibliografía 97
Referencias adicionales: 97

5. Manejo del ictus de modo protocolizado en los Servicios de Urgencias 99
Necesidad de un protocolo
de actuación común 102
Código ictus en urgencias..... 103
Cadena asistencial del Ictus 105
¿Cuándo activar un código ictus? 107
Servicio de Urgencias Hospitalarias 108
Manejo inicial del paciente con ictus. 109
Diagnóstico diferencial en la sospecha de ictus 113
¿Se ha activado el código ictus?..... 114
Bibliografía 118

EN PRODUCCIÓN - GALERADAS

6. Aplicaciones móviles como medio de obtener estimaciones rápidas y fiables a partir de modelos complejos en ictus 121
1.- Introducción 123
2.- Desarrollo de una app móvil para determinar el riesgo de fallecimiento de un paciente por ictus isquémico 125
2.1.- Primer contacto 125
2.2.- Planificación 126
2.3.-Desarrollo 127
3.- Bibliografía 146

7. El papel de la ciencia de datos en el modelado predictivo: Aplicación al ictus isquémico y mortalidad. 147
Resumen 149
1. Introducción 150
1.1. Introducción al problema 153
1.2. Preprocesamiento de los datos 155
2. Análisis Exploratorio de Datos 157
2.1 Discusión de la métrica a utilizar 160
2.2 División en train-test 162
3. Modelos a utilizar 163
3. Resultados 165
3. SHAP values (SHapley Additive exPlanations) 167
4. Discusión 173
5. Conclusión 174
Bibliografía 176

8. Papel de los Modelos predictivos de regresión en la estimación de la mortalidad: generalidades 177
Conceptualización. Modelos predictivos y prognosis 179
Modelos basados en regresión según la variable de resultado ... 181
1. Resultado Cuantitativo Continuo 181

EN PRODUCCION - GALERADAS

2. Resultado dicotómico 183
3. Tiempo en el resultado 186
3. Métodos estadísticos 189
Evaluación del rendimiento en cuanto a las predicciones 193
Bibliografía 196

9. Modelo predictivo para la estimación de mortalidad en el ictus isquémico no lisado mediante metodología de regresión logística 197
Introducción 199
Conclusión 215
Bibliografía 216

10. Modelo predictivo basado en técnicas de árboles de clasificación para estimar la mortalidad por ictus 217
1. Introducción 219
1.1. Métodos de clasificación 220
2. Árboles de clasificación 222
2.1. Ideas básicas 222
2.2. Grafos 225
2.3. El proceso de aprendizaje 227
2.4. Las particiones 228
2.5. Medida de impureza y de homogeneidad de las particiones 229
2.6. El problema del sobreaprendizaje 231
2.7. La poda 233
3. Aplicación de los árboles de clasificación a un modelo de mortalidad por ictus 236
4. Conclusiones 240
Bibliografía 241

11. La validación de los modelos predictivos a la luz de la estadística y la ciencia de datos. 243

EN PRODUCCIÓN - GALERADAS

1. Validación de modelos..... 245
1.1. ¿Por qué es importante?..... 246
1.2. Enunciado 246
2. Complejidad del Modelo..... 247
2.2. Sesgo o bias..... 248
2.3. Varianza 248
2.3. Definición de rendimiento. 250
2.4. Problemas de regresión..... 251
2.5. Problemas de clasificación. 251
3. Aspectos a tener en cuenta: Análisis de las curvas de aprendizaje. 252
3.1. Subajuste..... 252
3.2. Sobreajuste..... 253
4. Técnicas de validación. 256
1. Train – Test Split. 256
2. Validación cruzada k-fold con un conjunto de datos de prueba independiente. 258
3. Leave-one-out cross-validation with independent test data set (Validación cruzada con un conjunto de datos de prueba independiente)..... 259
5. Metodología personal para entrenar y validar a la vez. 260
6. Evaluación de Clasificadores..... 264
6.1. Métricas asociadas a la matriz de sobreajuste. 265
6.2. Balance entre precision y recall..... 266
6.3. Análisis ROC. 267
7. Conclusiones..... 271
8. Bibliografía 272

TFM: Aplicación de Ciencia de Datos en la Optimización de modelos predictivos para la estimación de mortalidad en Ictus Isquémico 273

Resumen 274

EN PRODUCCIÓN - GALERADAS

Índice de Figuras **275**

Capítulo 1. Introducción [0] **277**
1. Situación 277
2. Fundamentos 278
3. Justificación 279
4. Hipótesis del trabajo 281
5. Datos Utilizados 282
6. Diseño 283
7. Objetivos del Proyecto 283
8. Objetivos del TFM 284

Capítulo 2A. Análisis Exploratorio **285**
1. Preprocesado y descripción de los Datos 285
2. Edad en los grupos 289
3. Sexo 291
4. Fallecidos por el resto de predictores 292
5. Discusión de la métrica 294
6. División en train-test 295
7. Algoritmo de reducción de dimensionalidad, t-SNE 295

Capítulo 3. Metodología **297**
1. Introducción 297
2. Datos desbalanceados 297
3. Calibración del modelo [22] 299
4. Algoritmo genético de selección de variables 302
5. Conclusión 305

Capítulo 4. Escenario A **306**
1. Introducción 306
2. Regresión logística 306
3. Random Forest balanceado 310
5. Explicabilidad e Interpretabilidad 311
6. Selección recursiva de predictores 319
7. Conclusiones del capítulo 320

EN PRODUCCIÓN - GALERADAS

Capítulo 5. Escenario B 322
1. Introducción 322
3. Conclusiones del capítulo 327

Capítulo 6. Conclusiones y trabajo futuro 328
1. Introducción 328
2. ¿Qué modelo poner en producción? 328
3. Servicios AWS 329
4. Conclusiones del TFM 330

Bibliografía 331

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS

EN PRODUCCIÓN - GALERADAS