



# UNIVERSIDAD DE SALAMANCA

GRADO EN ESTADÍSTICA

FACULTAD DE CIENCIAS

INSTITUTO DE INVESTIGACIÓN BIOMÉDICA

DE SALAMANCA

Trabajo de Fin de Grado

## **Predicción de eventos cardiovasculares y hemorrágicos en pacientes con doble antiagregación con modelos Machine Learning**

Prediction of cardiovascular and bleeding events in patients with  
double antiaggregation with machine learning models

Autor: **Pablo Pérez Sánchez**

Tutores: Dr. José Manuel Sánchez Santos

Dr. Pedro Luis Sánchez Fernández

Salamanca, 2020



**UNIVERSIDAD DE SALAMANCA**

GRADO EN ESTADÍSTICA

FACULTAD DE CIENCIAS

INSTITUTO DE INVESTIGACIÓN BIOMÉDICA  
DE SALAMANCA

Trabajo de Fin de Grado: Anexo

**Predicción de eventos cardiovasculares y hemorrágicos en  
pacientes con doble antiagregación con modelos Machine  
Learning**

Prediction of cardiovascular and bleeding events in patients with  
double antiaggregation with machine learning models

Tutores:

Dr. José Manuel Sánchez Santos

Dr. Pedro Luis Sánchez Fernández



Autor: Pablo Pérez Sánchez





Necesitamos enseñar que la duda no sea tímida, sino bienvenida y debatida.

No hay problema en decir: no lo sé.

Richard Feynman



En cumplimiento a la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (LOPD), no se pueden divulgar las bases de datos empleadas en el presente trabajo, debido a que estas pertenecen al Hospital Clínico de Salamanca.





# INDICE

|   |           |
|---|-----------|
| <b>1. Introducción</b>                                  | <b>3</b>  |
| 1.1. Cardiología  | 3         |
| 1.1.1. Trastornos cardíacos. El infarto                 | 3         |
| 1.2. Inteligencia artificial                            | 5         |
| 1.2.1. Aprendizaje automático                           | 6         |
| 1.2.2. IA en medicina                                   | 7         |
| 1.3. Motivación y objetivos                             | 8         |
| 1.4. Estado del arte                                    | 8         |
| <b>2. Material y métodos</b>                            | <b>10</b> |
| 2.1. Descripción de los datos                           | 10        |
| 2.2. Problemas de clasificación                         | 13        |
| 2.2.1. Matriz de confusión                              | 14        |
| 2.2.2. Curva ROC  | 15        |
| 2.2.3. Error de entrenamiento y error de generalización | 16        |
| 2.3. Algoritmos Machine Learning                        | 17        |
| 2.3.1. Clasificadores lineales                          | 17        |
| 2.3.2. Clasificadores combinados                        | 21        |
| 2.4. Evaluación   | 23        |
| 2.4.1. Cross validation                                 | 24        |
| 2.4.2. Validación externa                               | 25        |
| 2.5. Entorno de desarrollo                              | 25        |
| 2.6. Planteamiento de nuestro problema                  | 26        |
| 2.7. Flujo de datos                                     | 26        |
| <b>3. Resultados</b>                                    | <b>32</b> |
| <b>4. Discusión</b>                                     | <b>39</b> |
| <b>5. Conclusiones</b>                                  | <b>43</b> |
| <b>6. Bibliografía</b>                                  | <b>44</b> |
| <b>7. Summary</b>                                       | <b>46</b> |



# 1. Introducción

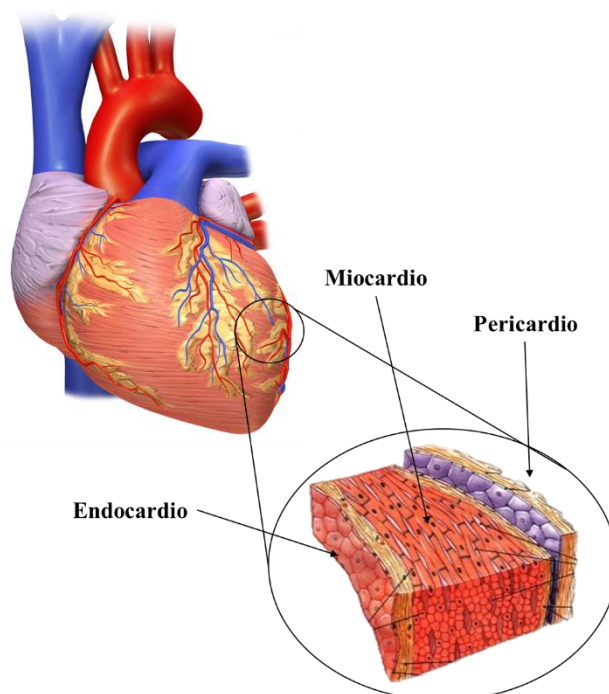
## 1.1. Cardiología

La cardiología es una rama de la medicina que se especializa en el tratamiento y diagnóstico de los trastornos del corazón, los vasos sanguíneos y el sistema circulatorio. Dentro de la cardiología encontramos subespecialidades como electrofisiología cardíaca, ecocardiografía, cardiología intervencionista y cardiología nuclear.

El funcionamiento básico del sistema cardiovascular implica el proceso del corazón para mover el oxígeno y los nutrientes a través de la sangre. Este proceso de transportar la sangre alrededor de todo el cuerpo se le denomina sistema circulatorio. El sistema cardiovascular es un circuito cerrado que funciona de manera unidireccional gracias al corazón, el cual es su órgano principal. Su cometido es bombear la sangre, facilitando el transporte del oxígeno y los nutrientes a las células del cuerpo. Está formado por 3 capas: pericardio, miocardio y endocardio.

- Pericardio: es una membrana fibrosa que cubre el corazón.
- Miocardio: es el tejido muscular del corazón, es el encargado de la contracción para bombear la sangre en el sistema circulatorio.
- Endocardio: es una membrana que recubre el interior de las cavidades del corazón.

Para el buen funcionamiento del corazón se tiene que mover la sangre por el sistema coronario, este sistema está compuesto por un conjunto de arterias y venas coronarias que riegan el miocardio.



Fuente: Wikimedia Commons y elaboración propia

**Ilustración 1.** El corazón y las capas por las que está formado.

### 1.1.1. Trastornos cardíacos. El infarto

Los trastornos del sistema cardiovascular son estudiados y tratados por la cardiología. Entre toda la variedad de trastornos, el más significativo es la cardiopatía isquémica.

La cardiopatía isquémica se produce cuando parte del miocardio recibe una cantidad insuficiente de sangre y oxígeno<sup>1</sup>. Aparece de manera específica cuando hay una descompensación entre el aporte de oxígeno y la necesidad del mismo. La causa más frecuente de isquemia del miocardio es el estrechamiento de una arteria coronaria, lo que disminuye la circulación sanguínea al miocardio y ocasiona un tránsito insuficiente por parte de la arteria coronaria afectada. La cardiopatía isquémica se puede presentar en forma de tres manifestaciones clínicas diferentes: angina de pecho estable, angina de pecho inestable e infarto agudo de miocardio. Los dos últimos también son también conocidos como síndrome coronario agudo (SCA).

El infarto agudo de miocardio (IAM) es una enfermedad que ocurre como consecuencia de la obstrucción de una arteria coronaria por un trombo, disminuyendo la cantidad de oxígeno que llega a las células. El IAM se diagnostica<sup>2</sup> cuando se constata un aumento de los valores de troponina cardíaca (cTn) por encima del límite superior de referencia (LSR) del percentil 99 y al menos una de las siguientes condiciones:

- Síntomas de isquemia miocárdica.
- Cambios isquémicos nuevos en el electrocardiograma (ECG).
- Aparición de ondas Q patológicas.
- Evidencia por imagen de pérdida de miocardio viable o anomalías regionales de la motilidad de la pared nuevas siguiendo un patrón compatible con una etiología isquémica.
- Identificación de un trombo coronario por angiografía o autopsia.

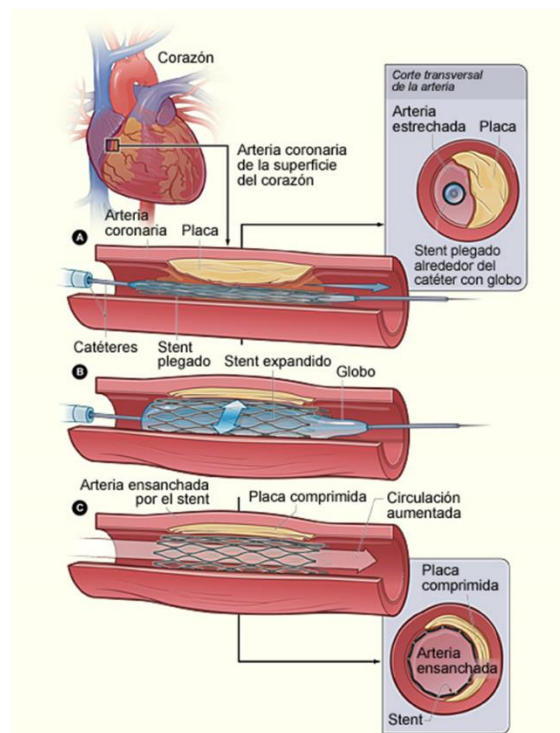
La troponina es una proteína globular que se vierte al torrente sanguíneo tras muerte celular, su determinación es sencilla a través del análisis de sangre una gota de sangre. Para la identificación de ondas Q, las cuales traduce necrosis miocárdica y por ende ausencia de repolarización, precisamos de la realización de un electrocardiograma.

La angina de pecho estable es un síndrome clínico aislado que se debe a una isquemia miocárdica transitoria. Es un dolor recurrente en el tórax, se suele desencadenar tras ejercicio físico y tras poco minutos de descanso se puede mitigar.

La angina de pecho inestable es un signo de alto riesgo de infarto de miocardio o muerte súbita. Se caracteriza por molestia en el tórax o los brazos. Surge durante el reposo, es intensa o va in crescendo (más intensa, duradera o frecuente).

El tratamiento principal ante un IAM es el intervencionismo percutáneo. Esta intervención consiste en realizar un punción femoral, radial o braquial percutánea. Se pasará un catéter guía a través de la arteria periférica y se avanzará hasta el trastorno coronario que se desea tratar. Al llegar a él realizan una abertura mecánica de la arteria. La abertura se produce mediante un balón expandible tal y como se puede observar en la ilustración 2, a su vez se implanta un stent para que no se vuelva a taponar ese mismo segmento arterial. Un stent es una estructura mecánica cilíndrica que se puede asemejar a un muelle.

Uno de los tratamientos seguidos tras la intervención percutánea es la doble antiagregación plaquetaria (DAP)<sup>3</sup>. La DAP consiste en administrar al paciente dos fármacos antiplaquetarios distintos con el objetivo de evitar nuevos sucesos isquémicos adversos. Uno de los dos fármacos antiplaquetarios siempre será la aspirina o ácido acetil salicílico, el otro medicamento será uno de los tres siguientes fármacos: clopidogrel, ticagrelor o prasugrel. La eficacia del tratamiento DAP ha confirmado que reduce el riesgo isquémico y con ello futuros sucesos cardiovasculares adversos. Aunque como efecto secundario aumenta el riesgo de sufrir una hemorragia.



Fuente: Wikimedia Commons

**Ilustración 2.** Intervención percutánea para la colocación de stent.

Cuando se habla del riesgo de sufrir una hemorragia<sup>4</sup>, se refiere al riesgo de sufrir una pérdida de sangre que se sucede desde el sistema vascular en una cavidad. Durante el presente trabajo se definió como hemorragia: la necesidad de transfusión o cirugía, taponamiento cardiaco, hematomas o, a nivel de parámetros de laboratorio, una disminución de la puntuación de hemoglobina y hematocrito.

## 1.2. Inteligencia artificial

La Inteligencia Artificial (IA) es una rama de la ciencia computacional, la cual busca comprender y construir entidades inteligentes. El término se acuñó formalmente en el año 1956 en una conferencia impartida en la ciudad de Dartmouth por John McCarthy. Sin embargo, el propósito de crear máquinas capaces de pensar por sí mismas se remonta a la Antigua Grecia. Aunque durante ese tiempo nacieron algunos ejemplos de IA, no fue hasta 1842 con la construcción de la primera computadora analítica creada por parte de Ada Lovelace cuando fue realmente posible crear un claro ejemplo de IA.

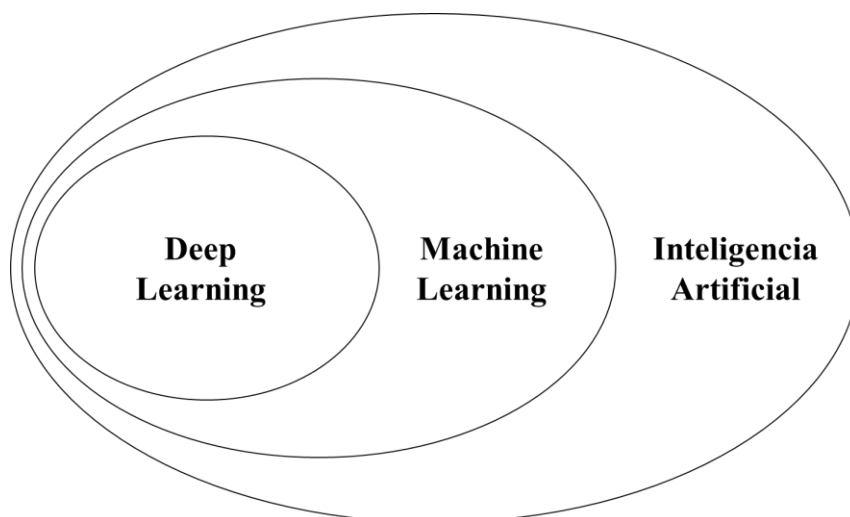
A lo largo de la historia, la definición de IA ha obtenido diferentes enfoques o perspectivas, sin embargo en la actualidad cuatro de ellos son los más utilizados: pensar como un humano, actuar como un humano, pensar racionalmente y actuar racionalmente<sup>5</sup>.

De estos cuatro enfoques, el más utilizado para definir la IA es el que hace referencia a pensar como un ser humano, aunque este simplemente es el enfoque basado en los conocimientos de la ciencia cognitiva, intentando imitar el cerebro humano. La ciencia cognitiva es el estudio científico de la mente y sus procesos.

Dentro del enfoque de actuar como un ser humano encontramos el test de Turing, una prueba de la capacidad de una computadora para exhibir un comportamiento inteligente que equivalga al de un humano. Para demostrar esta capacidad es necesario que un interrogador humano realice preguntas de forma escrita a un humano o a una maquina y leyendo las respuestas, discernir si es un humano o por el contrario no lo es.

El filósofo Aristóteles fue de los primeros en intentar codificar racionalmente el pensamiento a través de los llamados silogismos. Un ejemplo de uno de los silogismos más destacados es “Sócrates es un hombre. Todos los hombres son mortales. Sócrates es mortal”. De esta manera, comenzó el campo de la lógica donde a partir del siglo XIX se crearon programas que sigan silogismos y se basen en reglas lógicas para pensar racionalmente.

Otro de los enfoques citados anteriormente es el de actuar racionalmente, en la IA esto se relaciona con un agente o autómatas racional, el cual es aquel que intenta alcanzar el mejor resultado o el mejor resultado esperado en el caso de haber incertidumbre.



*Ilustración 3.* Diagrama Inteligencia Artificial, Machine Learning y Deep Learning.

Centrándonos en una perspectiva más práctica y moderna, en la actualidad la IA se utiliza principalmente para:

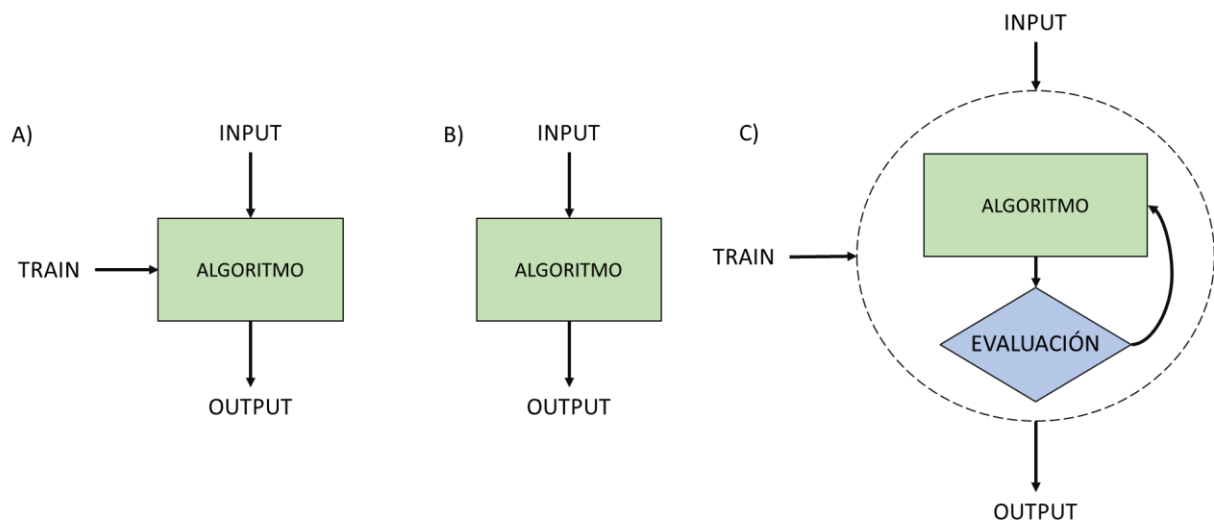
- i. Creación de programas que realizan tareas sin haber sido programados específicamente para ello, si no que han aprendido a partir de ejemplos de la realización de dicha tarea.
- ii. Detección de patrones en conjuntos de datos más allá de lo que una persona puede intentar estudiar.
- iii. Aplicar técnicas a masivas cantidades de datos que habitualmente requerirían de la labor de una persona humana durante un largo periodo de tiempo.
- iv. Implementación de redes neuronales que imitan el funcionamiento del sistema nervioso de los seres vivos

Gracias a todas las nuevas implementaciones que están surgiendo, en la actualidad la IA está cobrando una mayor importancia en nuestras vidas. Desde hablar con el asistente personal de nuestro móvil (Cortana, Siri, Google, ...) hasta la conducción de un vehículo de manera autónoma.

### 1.2.1. Aprendizaje automático

En la IA uno de los objetivos principales es que un agente tenga la capacidad de aprender a desempeñar tareas sin necesidad de ser entrenadas explícitamente para ello. De esto trata el aprendizaje automático o Machine Learning (ML). El aprendizaje automático es un conjunto de técnicas o algoritmos que otorgan a una maquina la capacidad para conseguir este fin.

Dentro del aprendizaje automático<sup>5</sup> existen diferentes formas: supervisado, no supervisado y por refuerzo.



**Ilustración 4.** Tipos de aprendizaje automático. A) Aprendizaje supervisado. B) Aprendizaje no supervisado. C) Aprendizaje por refuerzo.

#### ➤ Aprendizaje supervisado.

El aprendizaje supervisado es una técnica que trata de obtener una función desde unos datos de entrenamiento compuestos por vectores de entrada (*inputs*) con sus respectivos vectores objetivo (salida u *outputs*). Cuando este vector objetivo se trata de una variable cuantitativa hablamos de regresión, mientras que si es una variable categórica se habla de problema de clasificación.

➤ Aprendizaje no supervisado.

El aprendizaje no supervisado es una técnica que extrae propiedades o características internas de un conjunto de vectores de entrada sin necesidad de vectores objetivo. El uso más común es el análisis de conglomerados, es decir se utiliza a nivel exploratorio para encontrar patrones o agrupaciones ocultas en los datos.

➤ Aprendizaje por refuerzo.

En el aprendizaje por refuerzo, el programa aprende a través de una serie de premios o castigos. Dependiendo del resultado que obtenga en cada paso, el agente evaluará cómo actuar posteriormente. Es decir, el sistema obtendrá un resultado y en función de si es bueno o malo se le premiará o castigará, de esta forma en el siguiente paso actuará de forma diferente.

Dentro del aprendizaje automático encontramos una serie de algoritmos que se les conoce como redes neuronales, esta parte del ML se le denomina aprendizaje profundo o Deep Learning. Estas redes neuronales tratan de imitar al sistema nervioso del ser humano, están compuestas por capas de neuronas en la cual cada una aprende de manera individual. En general, las redes neuronales cuentan con gran número de capas, esto permite abordar tareas con mayor nivel de complejidad. Estos algoritmos se utilizan para reconocimiento de imágenes o el reconocimiento de voz, como para la detección de fraudes financieros.

### **1.2.2. IA en medicina**

En la actualidad la IA se está abriendo paso en el mundo de la medicina, desde facilitar la creación de nuevos fármacos, pasando por la predicción de enfermedades mediante análisis automático de imágenes médicas, hasta el estudio del genoma humano. La IA se ha instaurado en la medicina actual con la intención de ayudar en la elaboración y la rapidez del diagnóstico, también para acompañar y facilitar la labor al médico, nunca para sustituirlo. La IA también es una gran oportunidad para países en vías de desarrollo que no cuentan con personal cualificado en el ámbito médico.

Los campos del aprendizaje automático o ML y aprendizaje profundo o Deep Learning de la IA ofrecen unos resultados ventajosos en la predicción de enfermedades y los resultados ante un tratamiento<sup>6</sup>. Esto se debe al gran número de variables que se pueden involucrar en un algoritmo o red neuronal, estos métodos nos permiten tratar grandes bases de datos con heterogeneidad u homogeneidad entre sus variables. A partir de los datos clínicos de un paciente, se podría determinar cómo su organismo puede responder ante un tratamiento, lo que permitiría buscar el proceso que más se adapte a sus necesidades y que mayor eficacia pueda tener. Por tanto, esta tecnología ofrece grandes avances en la personalización de los tratamientos y en la prevención de enfermedades.

Algunos ejemplos habituales de IA en medicina son el tratamiento de imágenes para la predicción de una enfermedad, como por ejemplo el diagnóstico de cáncer de mama a partir de las imágenes tomadas de una mamografía<sup>7</sup>. O el tratamiento de grandes conjuntos de datos biológicos de los pacientes para la predicción de enfermedades<sup>8</sup>.

### 1.3. Motivación y objetivos

En el mundo se producen 32,4 millones de IAM al año, además en 2017 un 30% de las muertes en España fueron por enfermedades cardiovasculares. En 2015, según datos del INE, el IAM fue la mayor causa de muerte de nuestro país. La probabilidad (aproximada) de superar un IAM, varía dependiendo del lugar donde se produzca, dado que el tiempo de intervención es clave en estos casos, desde un 40% de mortalidad en los casos extrahospitalarios a un 10% de mortalidad en los casos intrahospitalarios<sup>9</sup>. Aunque el porcentaje de reincidencia de un IAM es bajo (menos de un 10%), la medicación y el estilo de vida seguido después de sufrir un IAM es clave en estas ocasiones. No obstante, el impacto de estas enfermedades no debe medirse solo en términos de mortalidad, puesto que estas enfermedades suelen llevar asociadas patologías con un gran impacto en la calidad de vida tanto de los pacientes como de sus familias.

Como hablamos anteriormente, una de las terapias utilizadas como tratamiento tras un IAM es la DAP. Por ello la elaboración de este trabajo gira entorno a pacientes que tras haber sufrido una cardiopatía isquémica han sido seguido este tratamiento, se estudió el riesgo isquémico (o riesgo de sufrir un reinfarto) y el riesgo hemorrágico o de sangrado.

La implicación que concierne a este trabajo no se cierra a la elaboración de un Trabajo de Fin de Grado, sino que puede servir para la ayuda al facultativo en la decisión de medicación tras un IAM. El uso de la IA con la creación de modelos que sean capaces de realizar trabajos monótonos que realizan los médicos en su día a día, hasta modelos capaces de ayudar a la predicción de enfermedades o el riesgo a padecerla, son imprescindibles para obtener una mejor sanidad, en la cual el médico esté más centrado en el paciente y se obtenga una medicina más personal.

El fin de este trabajo es conseguir una predicción del riesgo isquémico y hemorrágico. Haciendo uso de algoritmos de IA que se explicarán a lo largo del texto, se pretende obtener una predicción correcta de ambos eventos en pacientes que tomen la medicación DAP durante 12 meses siguientes a sufrir un IAM.

### 1.4. Estado del arte

Este Trabajo de Fin de Grado se apoya en los trabajos previos realizados por Sergio Raposeiras-Roubin y colaboradores, Cardiólogo clínico en el Hospital Universitario Álvaro Cunqueiro (Vigo, España), sobre los estudios RENAMI y BleeMACS y en la metodología realizada en Sampedro-Gómez Jesús (2020)<sup>10</sup>.

Respecto a los estudios RENAMI y BleeMACS se ha seguido tres artículos diferentes para estudiar y comprender mejor estas bases de datos:

- i. En Fabrizio D'Ascenzo (2020)<sup>11</sup> se realizó una comparación del riesgo diario promedio de isquemia y hemorragia durante el primer año tras una intervención coronaria percutánea por una IAM. Concluyen que ambos riesgos son equivalentes al año, que durante las dos primeras semanas el riesgo de sangrado es mayor, que en el primer año los pacientes sin revascularización completa tienen un mayor riesgo isquémico y los pacientes sin elevación del ST y con medicación ticagrelor obtienen un mayor riesgo de sangrado.
- ii. En Garay Alberto (2017)<sup>12</sup> crearon una nueva puntuación para la predicción de hemorragia en pacientes de edad avanzada. Evaluaron la incidencia y las características del sangrado. Se concluyó que los pacientes de más de 75 años son tratados con mayor frecuencia con clopidogrel



respecto a ticagrelor y prasugrel y que los pacientes de edad avanzada tuvieron una incidencia significativamente mayor de hemorragia posterior al alta.

- iii. En Raposeiras-Roubín Sergio (2019)<sup>13</sup>, su objetivo fue analizar la incidencia y los predictores de trombosis del stent tras sufrir un síndrome coronario agudo y recibir la medicación DAP. Concluyeron que la trombosis del stent es infrecuente y que las variables elevación del segmento ST, edad avanzada, antecedente de IAM y la creatinina se asocian con un mayor riesgo de trombosis del stent.

Respecto a Sampedro-Gómez Jesús (2020) referenciado anteriormente, se utilizó para seguir una metodología propia de investigación en IA en el ámbito de la medicina. En él desarrollan un modelo aplicable a la práctica clínica diaria para predecir la restenosis del stent. La restenosis es el estrechamiento de un vaso sanguíneo o arteria que había recibido un tratamiento quirúrgico anteriormente por ese mismo motivo. Hacen uso de algoritmos de ML y técnicas como la validación cruzada para obtener modelos más robustos. Finalmente obtienen un modelo con mejores resultados que los discriminadores actuales.

## 2. Material y métodos

### 2.1. Descripción de los datos

La población de estudio proviene de tres registros diferentes: “Bleeding complications in a Multicenter registry of patients discharged with diagnosis of Acute Coronary Syndrome” (BleeMACS), “REgistry of New Antiplatelets in patients with Myocardial Infarction” (RENAMI) y un registro propio del hospital universitario Álvaro Cunqueiro (Vigo, España) y del hospital universitario Virgen de la Arrixaca (Murcia, España). Contamos con un total de 26.076 pacientes con SCA sometidos a una intervención coronaria percutánea.

BleeMACS<sup>14</sup> es un registro multicéntrico internacional que cuenta con 15.364 pacientes de 16 centros hospitalarios procedentes de Europa (Alemania, Polonia, Países Bajos, España, Italia, Grecia, Macedonia), Asia (China y Japón), Norteamérica (Canadá) y Sudamérica (Brasil), entre los años 2003 y 2014. Los pacientes fueron elegidos mediante tres criterios de inclusión obligatorios: paciente dado de alta con diagnóstico de SCA con infarto del miocardio con elevación del segmento ST (SCACEST), con infarto del miocardio sin elevación del segmento ST (SCASEST) o con angina inestable. Este registro fue diseñado para predecir el sangrado en pacientes con SCA. El tratamiento DAP se realizó durante 12 meses, excepto en pacientes tratados con anticoagulantes orales, en estos pacientes el tiempo de tratamiento se basó en el juicio clínico del facultativo responsable. Los datos fueron recogidos durante el tiempo de tratamiento de los pacientes. Se estableció como punto final del seguimiento tener una hemorragia grave, la cual se definió como sangrado intracraneal o cualquier otro sangrado que conduce a hospitalización y/o transfusión de glóbulos rojos.

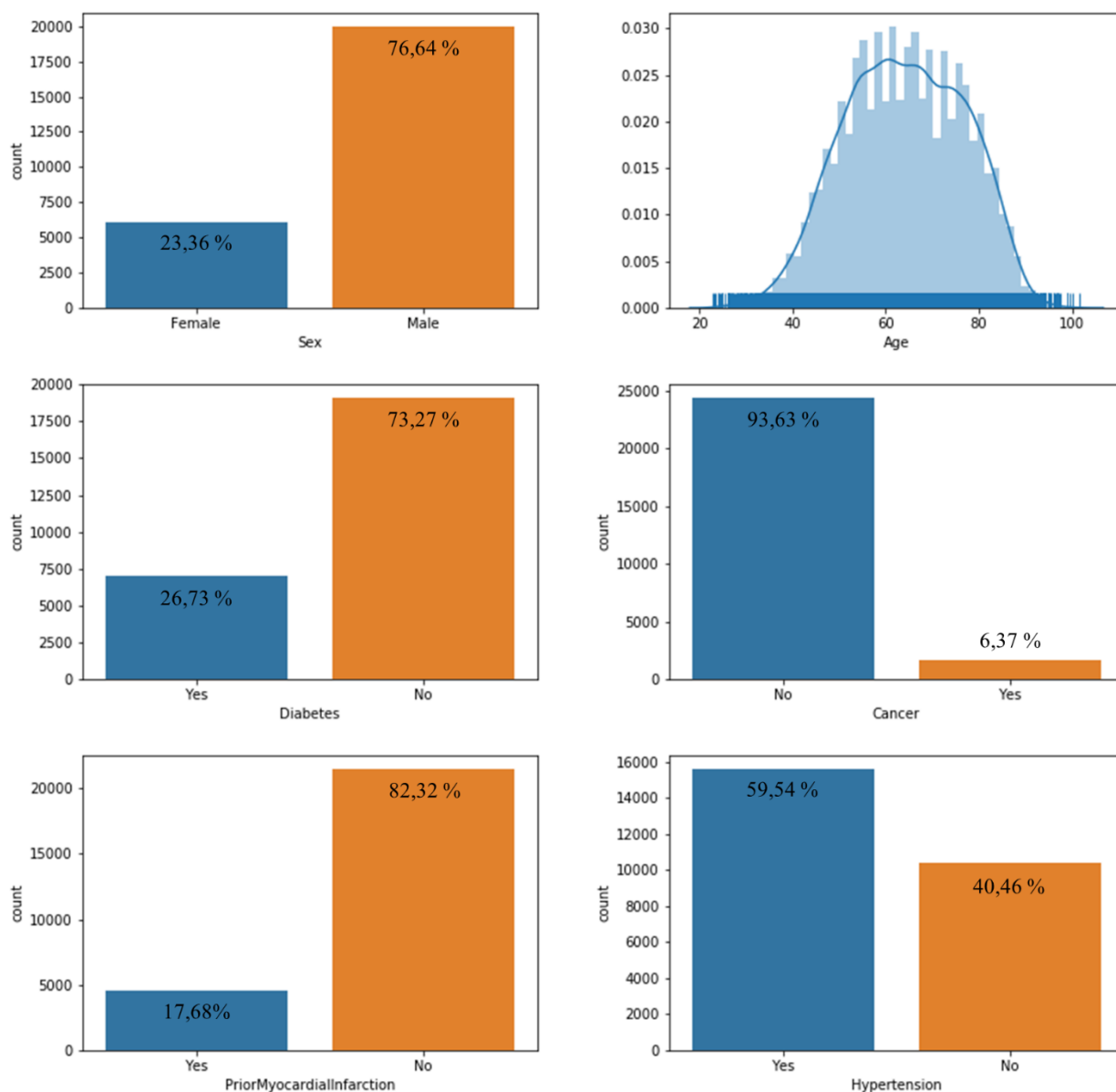
RENAMI<sup>13</sup> es un registro multicéntrico que contiene información sobre 4.442 pacientes de 11 centros europeos (España, Italia, Suiza, Grecia, Serbia y el Reino Unido) entre los años 2012 y 2016. Los criterios de inclusión del registro fueron los siguientes: paciente dado de alta con diagnóstico SCA, tener evidencia de estenosis coronarias angiográficamente significativas durante el ingreso por SCA, realización de intervención coronaria percutánea durante el ingreso, seguir el tratamiento DAP tras SCA. Este conjunto de datos fue elaborado para evaluar la incidencia y los predictores de trombosis del stent en pacientes con SCA que reciben DAP, con el uso de aspirina más prasugrel o ticagrelor<sup>3</sup>. El tratamiento DAP se realizó durante 12 meses. El seguimiento fue censurado en la aparición de ST, muerte, suspensión DAP o a los 12 meses de seguimiento.

El registro procedente de los hospitales universitarios de las localidades Vigo y Murcia, cuenta con un total de 6290 pacientes. Este no pertenece a un estudio, sino a la práctica habitual de los hospitales mencionados anteriormente. En el conjunto de datos se realiza un seguimiento a los pacientes que han sufrido SCA y han recibido tratamiento DAP en algún momento.

En el conjunto de datos se pueden observar 49 variables, de las cuales 38 son variables categóricas, 10 son variables numéricas y una es variable de identificación. Algunas de las variables de la base de datos son la edad de los pacientes, país al que pertenecen, el sexo, si ha sufrido anteriormente un IAM, el tiempo de la medicación o si ha sufrido muerte, infarto o sangrado; además también se tiene una variable tiempo en la cual se indica cuando sucedió el evento, de esta manera tenemos en total 4 variables de tiempo: medicación, reinfarcto, hemorragia y muerte. Con ellas y acompañadas de sus variables eventos, somos capaces de crear las variables respuesta que se usarán en este trabajo. En el ANEXO se puede ver una pequeña definición de las variables que se observan en la Tabla 1.

Nos encontramos con una población mayoritariamente masculina (76.6%) y con una media de 64 años de edad. El 26.7% de los pacientes presenta diabetes y el 6.4% cáncer. Aunque los pacientes de nuestro estudio han sufrido un infarto en el momento en el que se comenzó a recoger sus datos, esto no quiere decir que sea el primer infarto que han tenido, ya que el 17.7% de la población que vamos a estudiar ya ha sufrido un IAM con anterioridad. También podemos observar que el 59.9% de nuestro conjunto de pacientes padece hipertensión, reconocida como uno de los cinco factores de riesgo principales para sufrir un IAM<sup>2</sup>. El 52.8% de los pacientes sufrió un SCACEST. Toda esta información

se puede ver gráficamente en la ilustración 5 y en la Tabla 1 se puede observar ampliada todas las variables de estudio.



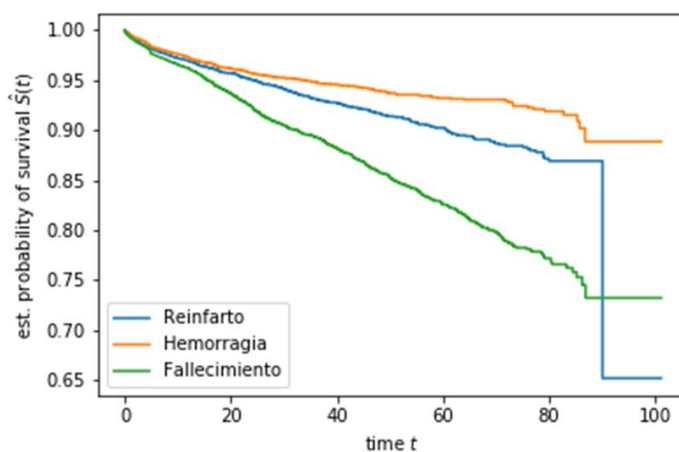
**Ilustración 5.** Gráfica de las variables “Sexo”, “Edad”, “Diabetes”, “Cáncer”, “IAM previo” e “Hipertensión”. En ellas se puede ver cómo está distribuida la población a estudiar.

**Tabla 1.** Características de la base de datos.

|  |         |
|--|---------|
| Age - Edad (AÑOS)  | 64 ± 13 |
| Sex - Sexo femenino (%)  | 23.36   |
| Diabetes - Diabetes mellitus (%)                               | 26.73   |
| Hypertension - Hipertensión (%)                                | 59.94   |
| Dyslipemia - Dislipemia (%)                                    | 56.02   |
| Peripheral Artery Disease - Enfermedad vascular periférica (%) | 6.14    |
| Prior Myocardial Infarction - IAM (anterior) (%)               | 17.68   |
| Prior PCI - Intervención percutánea coronaria (anterior) (%)   | 15.70   |
| Prior CABG - Bypass coronario (anterior) (%)                   | 3.25    |
| Prior Stroke - Accidente cerebrovascular (anterior) (%)        | 6.29    |
| Cancer - Cáncer (%)  | 6.37    |

|   |              |
|---|--------------|
| Prior Bleeding - Hemorragia (anterior) (%)                  | 4.24         |
| STEMI - SCACEST (%)   | 52.86        |
| Unstable Angina - Angina inestable (%)                      | 12.86        |
| Killip2orMore (%)   | 11.25        |
| Femoral Access - Acceso femoral (%)                         | 37.54        |
| Multivessel – Enfermedad multivazo (%)                      | 39.68        |
| Drug Eluting Stent - Stent farmacológico (%)                | 46.71        |
| No stent (%)  | 4.40         |
| Complete Revascularization - Revascularización completa (%) | 69.44        |
| ReAMI in hospital - Reinfarto en el hospital (%)            | 1.61         |
| Bleeding in hospital - Hemorragia en el hospital (%)        | 4.17         |
| IH transfusion - Transfusión IH (%)                         | 3.49         |
| Hemoglobin on Admission - Hemoglobina al ingreso (g/dl)     | 14.00 ± 1.67 |
| Creatinine on Admission - Creatinina al ingreso(mg/dl)      | 0.99 ± 0.53  |
| Clopidogrel (%)   | 78.07        |
| Ticagrelor (%)  | 12.86        |
| Prasugrel (%)   | 7.07         |
| Oral Anticoagulation - Anticuagulantes orales (%)           | 4.94         |
| Betablockers - Betabloqueantes (%)                          | 84.10        |
| ACEI_ARB (%)  | 77.96        |
| Statin - Estatina (%)                                       | 94.80        |
| ReInfarction - Reinfarto (%)                                | 3.12         |
| Bleeding - Hemorragia (%)                                   | 3.78         |
| Death - Fallecimiento (%)                                   | 5.80         |
| Transfusion - Transfusión (%)                               | 2.90         |

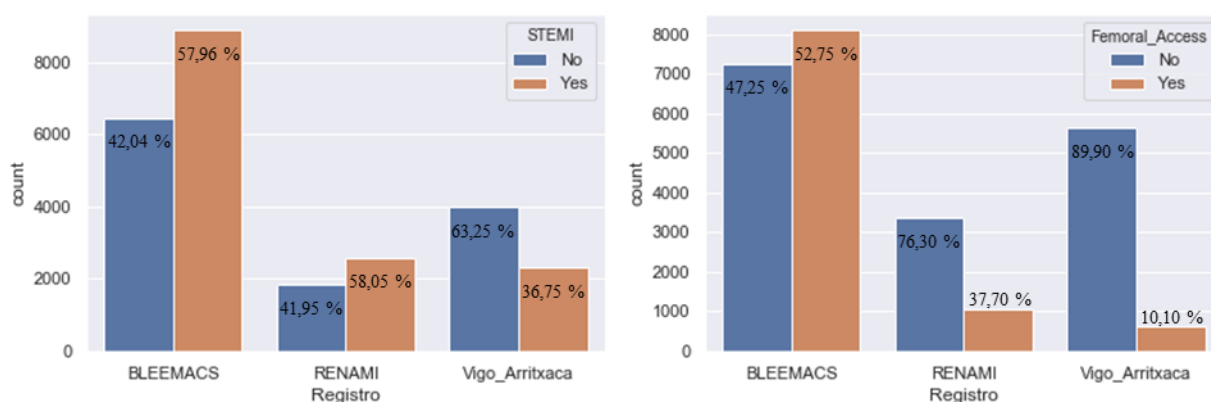
En la ilustración 6 se visualiza una gráfica de un análisis de supervivencia. Podemos observar que a medida que avanza el tiempo hay una caída pronunciada en los tres eventos que se analizan. Esta caída tan brusca se debe al bajo número de pacientes a los que todavía se les realiza seguimiento en ese tiempo tan avanzado. El registro BleemACS solo realiza el seguimiento durante 12 meses y la gran mayoría de los pacientes de RENAMI también se les realiza un seguimiento durante ese periodo de tiempo. Con lo cual, los únicos pacientes que se les realiza un seguimiento tan extenso son los procedentes de Vigo y Murcia. En el mes 90 de los 4 pacientes que siguen teniendo un seguimiento uno de ellos sufre un reinfarto, por lo que la probabilidad de sobrevivir en ese momento del tiempo baja un 25%.



**Ilustración 6.** Curva de supervivencia de reinfarto, hemorragia y fallecimiento. En el eje horizontal se plasma el tiempo en meses. En el eje vertical la probabilidad de sobrevivir a uno de esos eventos.

A la hora de estudiar la población diferenciando entre los tres registros que la componen no encontramos muchas diferencias entre las distribuciones muestrales que siguen las variables. La mayoría de las características de los pacientes siguen la misma distribución en los tres registros. Algunas de las variables en las que encontramos una diferencia en los registros son “Infarto agudo de miocardio con elevación del ST” o “acceso femoral”, como se puede observar visualmente en la ilustración 6. Observamos que “SCACEST” presenta un porcentaje parecido de presencia en los registros BleeMACS y RENAMI, en torno al 58%, mientras que en el registro de Vigo y Murcia presenta un número de casos menor, alrededor de un 37%. En la variable “acceso femoral” podemos advertir un cambio significativo a la hora de intervenir en el IAM. En el registro más antiguo (BleeMACS) se realizaba la intervención el 53% de las veces por esa vía, en cambio en RENAMI se llevaba a cabo en el 38% de los casos y en Vigo y Murcia el 10% de los casos. En la actualidad el acceso femoral es cada vez menos utilizado, dado que tiene un alto riesgo de provocar un sangrado<sup>15</sup>. Donde encontramos una mayor diferencia es en la medicación DAP que se le asigna al paciente, donde cada registro se recetó un antiagregante diferente a cada paciente, en la Tabla 2 se puede observar esta diferencia de medicación por registro. Se puede encontrar toda esta información ampliada a todas las variables en el ANEXO en las ilustraciones 1 y 2 y en la Tabla 1.

Las variables “peso” y “talla” no fueron recogidas en el registro BLEEMACS, por ello no fueron utilizadas a lo largo del trabajo. También se encuentra un fallo en la variable “peso” en el registro RENAMI, esta variable hace referencia al índice de masa corporal en vez de al peso del paciente.



**Ilustración 7.** Gráfica de las variables categóricas STEMI (IAM con elevación del segmento ST) y acceso femoral agrupadas por registro.

**Tabla 2.** Distribución del medicamento utilizado en el tratamiento DAP por registro.

|                      | Clopidrogel | Ticagrelor | Prasugrel |
|----------------------|-------------|------------|-----------|
| <b>BleeMACS</b>      | 91,6 %      | 4,1 %      | 4,3 %     |
| <b>RENAMI</b>        | 0 %         | 61,56 %    | 38,43 %   |
| <b>Vigo y Murcia</b> | 100 %       | 0 %        | 0 %       |

## 2.2. Problemas de clasificación

Podemos definir los problemas de clasificación como las aplicaciones de aprendizaje supervisado en las que se pretende obtener a partir de un vector de entrada, un número finito de categorías. Estas categorías pueden ser binarias (0 y 1) o multicategorías. En nuestro caso se trabajó el problema de

clasificación binario, dado que nuestras variables a predecir “Infarto 12 meses” o “Hemorragia 12 meses” cuentan con dos categorías.

En el ámbito de la IA existen diferentes herramientas que permiten la visualización del rendimiento de algoritmos empleados en métodos supervisados. Dos herramientas bastante usadas en los problemas de clasificación binarios son la matriz de confusión y la curva ROC.

### 2.2.1. Matriz de confusión

La matriz de confusión<sup>16</sup> resume el rendimiento de la clasificación de un algoritmo con respecto a un conjunto de datos. Es una matriz 2x2 y está indexada por la verdadera clase de un objeto y por la clase que asigna el clasificador.

**Tabla 3.** Matriz de confusión

|            |          | Clase predicción |          |
|------------|----------|------------------|----------|
|            |          | Positivo         | Negativo |
| Clase real | Positivo | VP               | FN       |
|            | Negativo | FP               | VN       |

- Negativo y Positivo es la condición que puede obtener la clase a predecir.
- VP es el número de predicciones correctas que se han hecho con la clase Positivo.
- FN es el número de predicciones incorrectas que se han hecho con la clase Positivo.
- FP es el número de predicciones incorrectas que se han hecho con la clase Negativo.
- VN es el número de predicciones correctas que se han hecho con la clase Negativo.

A partir de estos 4 resultados de la clasificación, se definen varias medidas del rendimiento de la clasificación:

- *Accuracy* o eficacia (AC) es la proporción total del número de predicciones correctas.

$$AC = \frac{VP + VN}{VP + VN + FP + FN}$$

- Sensibilidad o *recall* o ratio de verdaderos positivos (RVP) es la proporción de casos positivos que se han identificado correctamente.

$$Sensibilidad \text{ o } RVP = \frac{VP}{VP + FN}$$

- Especificidad o ratio de verdaderos negativos (RVN) es la proporción de casos negativos que se han identificado correctamente.

$$Especificidad \text{ o } RVN = \frac{VN}{VN + FP}$$

- Precisión o valor predictivo positivo (VPP) es la proporción del número de casos positivos bien predichos entre el total de los casos positivos predichos.

$$Precisión \text{ o } VPP = \frac{VP}{VP + FP}$$

- Resultado  $F_1$  es la media armónica de la precisión y la sensibilidad.

$$F_1 = 2 \cdot \frac{VPP \cdot RVP}{VPP + RVP}$$

- Ratio de falsos positivos (RFP) es la proporción de casos negativos que han sido identificados como positivos.

$$RFP = \frac{FP}{FP + VN}$$

- Ratio de falsos negativos (RFN) es la proporción de casos positivos que han sido identificados como negativos.

$$RFN = \frac{FN}{VP + FN}$$

- Valor predictivo negativo (VPN) es la proporción del número de casos negativos bien predichos entre el total de los casos negativos predichos.

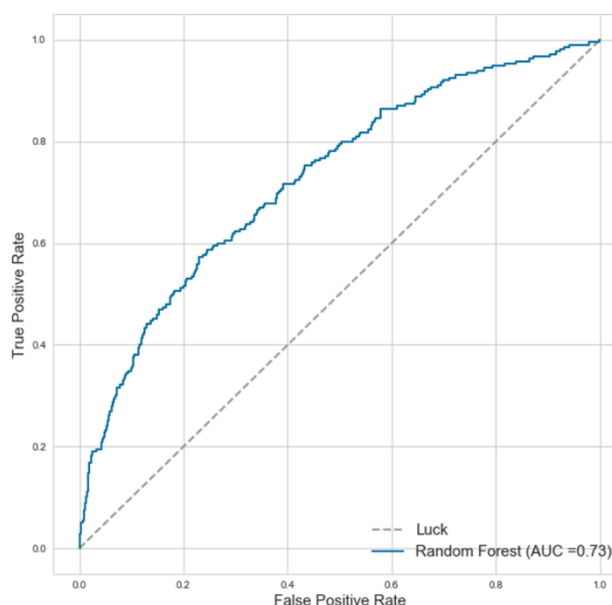
$$VPN = \frac{VN}{VN + FN}$$

Estas medidas posteriormente se usarán para evaluar los dos modelos. La métrica que se pretende optimizar depende del problema de clasificación que estemos tratando. Por ejemplo, en el caso de diagnosticar enfermedades que conlleven una prueba muy invasiva, la métrica que se optimizará será la especificidad, de tal manera que se pretenderá predecir como “enfermos” al mayor número de pacientes posibles que estemos muy seguros de que tengan la enfermedad, para evitar realizar la técnica diagnóstica invasiva a pacientes que no la necesitan. Otro ejemplo sería, diagnosticar una enfermedad que llevara un tratamiento poco invasivo. En ese caso, lo mejor sería optimizar la sensibilidad, de tal manera que no se nos “escape” ningún paciente con la enfermedad, aunque estemos también tratando a pacientes que no la tengan. En el siguiente punto se hablará de puntos de corte para establecer esa alta sensibilidad o especificidad.

## 2.2.2. Curva ROC

En el caso concreto de los algoritmos que hemos usado en este Trabajo de Fin de Grado para trabajar nuestro problema de clasificación binaria se asignará a cada individuo una probabilidad de pertenecer a cada una de las dos clases. Esta probabilidad da lugar a una clasificación continua que se puede representar de acuerdo a su sensibilidad y especificidad.

La curva ROC (Característica Operativa del Receptor)<sup>17</sup> es un gráfico que muestra el diagnóstico de un clasificador binario a medida que cambia el umbral de discriminación. Este gráfico enfrenta la especificidad contra la sensibilidad ajustadas a diferentes puntos de corte.



*Ilustración 8.* Curva ROC

En la curva ROC trabajaremos con el área bajo la curva (AUC). Este es el área que se encuentra por debajo de la curva y puede obtener un valor máximo de 1 y mínimo de 0. Se intentará optimizar este valor, de tal manera que, a un valor más alto, un mayor rendimiento del algoritmo, es decir, cuanto más cercana sea la curva ROC a la esquina superior izquierda, más alta será la exactitud de la prueba. El AUC está directamente ligado con el estadístico de Mann-Whitney.

$$AUC_1 = \frac{U_1}{n_1 n_2}$$

Donde  $U_1$  es el estadístico de Mann-Whitney<sup>17</sup>, también conocido como estadístico de la prueba de suma de rango de Wilcoxon y  $n_1$  y  $n_2$  los tamaños de las muestras. Esto quiere decir que si ordenamos nuestras instancias de acuerdo con la probabilidad que asigna un clasificador de pertenecer a la clase positiva, el AUC mide la probabilidad de que una instancia positiva elegida aleatoriamente esté por encima de una instancia negativa elegida aleatoriamente<sup>17</sup>.

Dada la probabilidad que se obtiene de los clasificadores, se nos permitirá poder variar el punto de corte para obtener una predicción más acorde a nuestros objetivos. Es decir, se podrá establecer el punto de corte dependiendo de la métrica que se prefiera optimizar. La curva ROC nos permite visualizar todos los puntos de corte, en términos de su sensibilidad y especificidad.

Tal como se habló en el punto 2.2.1, dependiendo del problema de clasificación al que se enfrente, se establecerá un punto de corte para optimizar la especificidad, sensibilidad o una combinación de ambas. Además, en el ámbito sanitario, hay que tener en cuenta la prevalencia de la enfermedad. Es posible que se estudien enfermedades con una baja prevalencia, como por ejemplo en este trabajo con el infarto/sangrado. Esto puede suponer un problema a la hora de clasificar o de optimizar las diferentes métricas. dado un clasificador, ante una población con una prevalencia baja, va a tener menor (o igual) valor predictivo positivo, y mayor (o igual) valor predictivo negativo que ante una población balanceada.

### 2.2.3. Error de entrenamiento y error de generalización

A la hora de medir las métricas no es lo mismo trabajar sobre un pequeño subconjunto, que sobre el total de pacientes del mundo. Es decir, nuestro algoritmo no trabajará igual para un subconjunto de la población que para la totalidad de ella.

Entrenar un algoritmo es variar los parámetros del modelo para mejorar las métricas que hemos definido previamente. Estas métricas no son aptas para entrenar con un ordenador, así que usamos una función de coste (o también llamada de pérdida) que sí sea apta y tenga una buena correlación con alguna de estas métricas. Dicho esto, nos referimos como error de entrenamiento a la diferencia que hay entre estas métricas y el punto ideal, usando el modelo en los datos de entrenamiento. Aludimos a error de generalización a esto mismo, pero aplicado a datos no vistos anteriormente.

Por ello la aclaración del primer párrafo, no hay que confundir el error de generalización con el de entrenamiento, dado que estos errores no tienen por qué ser iguales. Por ejemplo, al entrenar con un subconjunto de la población podríamos lograr obtener un error mínimo de entrenamiento, pero esto no garantizaría tener un error pequeño de generalización a la hora de trabajar con toda la población. Esta diferencia viene dada por un fenómeno llamado sobreajuste, en el punto 2.4 se hablará de él y de cómo estimar el error de generalización usando datos de entrenamiento.

Para disminuir el error de generalización en el entrenamiento también se pueden añadir términos a la función de coste que harán que el error de generalización sea más pequeño, pese a que aumente el error de entrenamiento. Estos se denominan términos de regularización.



## 2.3. Algoritmos Machine Learning

Un algoritmo de aprendizaje automático es un proceso, o conjunto de procesos, que ayuda a un modelo a ajustarse a los datos para lograr un objetivo. Un algoritmo de ML normalmente especifica la forma en que los datos se transforman de entrada a salida y como el modelo aprende una asignación adecuada de entrada.

Por lo tanto, podríamos dividir el algoritmo de ML en el modelo más el algoritmo de aprendizaje. El modelo especifica la función de mapeo y contiene los hiperparámetros, mientras que el algoritmo de aprendizaje actualiza los parámetros para ayudar al modelo a completar su objetivo. Los hiperparámetros se utilizan para parametrizar el proceso del modelo, no se utilizan para modelar los datos directamente, pero si influyen en la capacidad y características de aprendizaje.

Se utilizarán algoritmos de aprendizaje supervisado, estos algoritmos de ML funcionan a partir de datos de entrenamiento, es decir, pares de datos (entrada  $x$ , salida  $y$ ), por los cuales el algoritmo “aprende” por sí mismo. De esta forma al introducir un sujeto nuevo (entrada  $x$ ), el algoritmo realiza una predicción devolviendo una puntuación o probabilidad (salida  $y$ ).

En este trabajo se utilizaron dos tipos de clasificadores, clasificadores lineales y clasificadores combinados.

### 2.3.1. Clasificadores lineales

Se entiende por clasificador lineal a un algoritmo de clasificación que hace que su clasificación se base en una función lineal de predicción que combina un conjunto de pesos con el vector de entrada.

$$f(\vec{x}) = \vec{\omega} \cdot \vec{x} + b$$

Donde  $\vec{\omega}$  es un vector real de pesos,  $\vec{x}$  el vector de entrada,  $f$  la función que obtiene la salida y  $b$  una constante. Los pesos vendrán dados por el clasificador lineal en función de los datos que le proporcionemos.

Cuando  $f(\vec{x})$  sea mayor que 0, la clasificación será positiva. Mientras que cuando  $f(\vec{x})$  menor que 0, será una clasificación negativa. Cuando  $f(\vec{x}) = 0$  es un hiperplano que separa el espacio de vectores de entrada en dos.

#### 2.3.1.1. Regresión Logística

La regresión logística<sup>18</sup> es un modelo estadístico que utiliza la función logística (2.4.1.1.1) para modelar las probabilidades de pertenecer a una determinada clase a través de funciones lineales en  $x$ . El modelo tiene la siguiente forma:

$$h_{\vec{\omega}}(\vec{x}) = \text{Logistic}(\vec{\omega} \cdot \vec{x}) = \frac{1}{1 + e^{-\vec{\omega} \cdot \vec{x}}}$$

$$\text{donde la función logística es: } \text{Logistic}(z) = \frac{1}{1 + e^{-z}} \quad (2.4.1.1.1)$$

El proceso de ajustar los pesos de este modelo para minimizar el error en un conjunto de datos del que conocemos la verdadera clasificación, se llama regresión logística.

La función de coste de la regresión logística se crea a partir de la pérdida de la función  $h_{\vec{\omega}}(x)$  y una regularización. Para establecer la función de pérdida se buscará el  $\vec{\omega}$  que maximiza la verosimilitud global, lo que equivale a maximizar la suma de los log-verosimilitud individuales. Para un miembro positivo:

$$h_{\vec{\omega}}(\vec{x}) \cdot \frac{e^{\vec{\omega} \cdot \vec{x}}}{e^{\vec{\omega} \cdot \vec{x}} + 1} = \frac{e^{\vec{\omega} \cdot \vec{x}}}{e^{\vec{\omega} \cdot \vec{x}} + 1}$$

$$\log h_{\vec{\omega}}(\vec{x}) = \log \frac{e^{\vec{\omega} \cdot \vec{x}}}{e^{\vec{\omega} \cdot \vec{x}} + 1} = \vec{\omega} \cdot \vec{x} - \log(e^{\vec{\omega} \cdot \vec{x}} + 1)$$

La log-verosimilitud para un miembro de la clase negativa será:

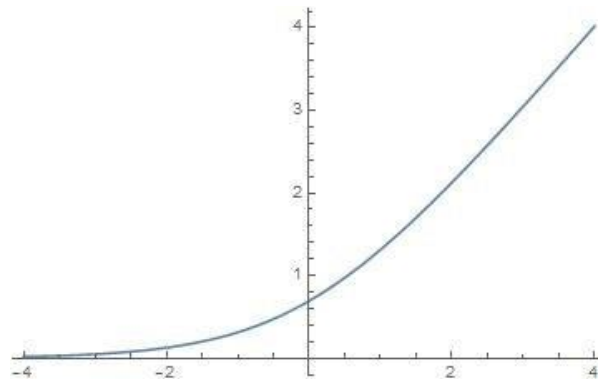
$$\log(1 - h_{\vec{\omega}}(\vec{x})) = \log\left(1 - \frac{1}{1 + e^{-\vec{\omega} \cdot \vec{x}}}\right) = \log(e^{\vec{\omega} \cdot \vec{x}} + 1) - \vec{\omega} \cdot \vec{x}$$

Si encontramos el valor de  $\vec{\omega}$  para el cual la suma de los log-verosimilitud es máxima, habríamos encontrado la función logística que más se ajusta a nuestros datos. Si multiplicamos la log-verosimilitud por -1, está se podrá interpretar como una función de pérdida que nos pone un coste si clasificamos mal una instancia, y la cual hay que minimizar.

$$\log(e^{\vec{\omega} \cdot \vec{x}} + 1) - \vec{\omega} \cdot \vec{x}$$

De esta manera tenemos la función de pérdida  $f(x)$  que usaremos para ajustar la regresión. Esta función  $f(x)$  tendrá dos posibilidades, dependiendo de si el valor de  $y$  tiene la etiqueta 0 o 1.

$$f_{\vec{\omega}}(x) = \begin{cases} \log(e^{-\vec{\omega} \cdot \vec{x}} + 1) + \vec{\omega} \cdot \vec{x}, & y = 0 \\ \log(e^{\vec{\omega} \cdot \vec{x}} + 1) - \vec{\omega} \cdot \vec{x}, & y = 1 \end{cases}$$



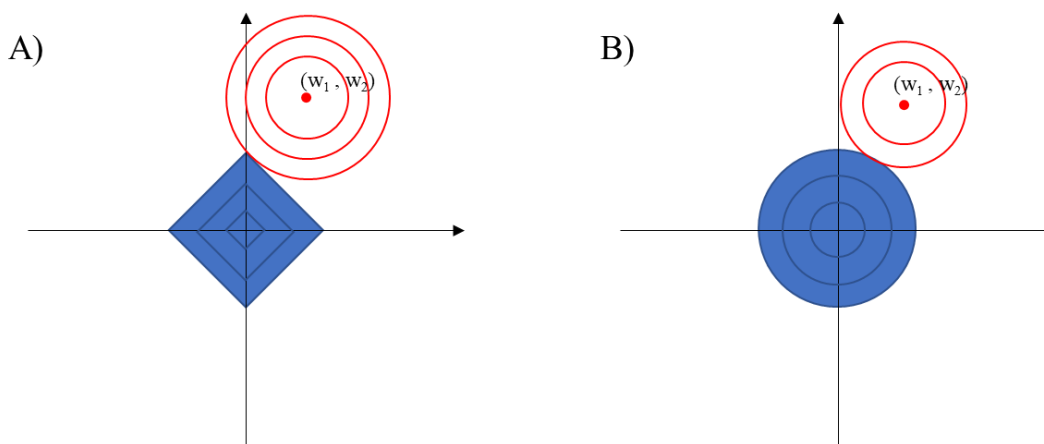
**Ilustración 9.** Gráfica de  $f(x)$  para la etiqueta 0.

La predicción de nuestro algoritmo será la siguiente, cuando el valor de  $\omega^*x$  sea mayor o igual a 0, se clasificará como 1. Mientras que, si el valor de  $x$  es menor a 0, se clasificara como 0. En la ilustración 9 se observa el gráfico de  $f(x)$  cuando su clasificación es 0, en el caso de que su clasificación sea 1 entonces la gráfica será su reflejo respecto del eje vertical. Esto viene explicado porque si  $\omega^*x$  es positiva cuando su clasificación real es 0, el algoritmo lo está clasificando mal y está generando mucha pérdida, mientras que si  $\omega^*x$  fuera negativa y su clasificación 0, no se generaría casi pérdida. Y sería lo mismo para la gráfica con respecto de la clasificación 1.

Por ello, se obtiene la siguiente función de pérdida en función de la clasificación de nuestra regresión logística, siendo  $y$  la clasificación verdadera de nuestros datos.

$$f_{\vec{w}}(\vec{x}, y) = y \cdot (\log(e^{\vec{w} \cdot \vec{x}} + 1) - \vec{w} \cdot \vec{x}) + (1 - y)(\log(e^{-\vec{w} \cdot \vec{x}} + 1) + \vec{w} \cdot \vec{x})$$

La función de coste de la regresión logística se puede formar a partir de la función de pérdida más una regularización. Dado que no es obligatorio añadir la regularización, la función de coste puede tener varias formas dependiendo de la regularización que se utilice. Podemos encontrar la regularización  $L_1$  o  $L_2$ .



**Ilustración 10.** Figuras explicativas de la regularización. A) Regularización  $L_1$ . B) Regularización  $L_2$ .

En la ilustración 10 podemos observar la regularización  $L_1$  y  $L_2$ , el punto  $(w_1, w_2)$  equivale al peso escogido sin usar la regularización, las curvas a su alrededor son la función de coste de la regresión logística. El punto que minimiza la suma de ambas funciones, regresión logística y regularización, tiene que ser tangente a una curva de nivel de cada función. A medida que se explican ambas regularizaciones, se irán explicando donde se produce la unión y la importancia que conlleva.

La regularización  $L_1$  o Lasso<sup>18</sup> se utiliza cuando se sospecha que varios atributos de entrada son irrelevantes. Con esta regularización se favorece que algunos de los coeficientes acaben valiendo 0, dado que es muy probable que la unión de ambas funciones se de en una de las esquinas de los rombos, esto implica un peso igual a 0, por ello esta regularización quita las variables irrelevantes. Esto hará que el modelo generalice mejor.

$$L_1(\vec{w}) = \lambda |\vec{w}|$$

El parámetro  $\lambda$  servirá para darle un peso a nuestra regularización.

Por lo que la función de coste o *Lasso regression* tendría la siguiente forma:

$$\sum_{i=0}^n f_{\vec{w}}(\vec{x}, y) + \sum_{j=0}^P L_1(\vec{w}_j)$$

Donde  $n$  será el número de muestras y  $P$  el número de variables.

La regularización  $L_2$  o Ridge<sup>19</sup> se emplea cuando se sospecha que varias variables de entrada están correlacionadas entre ellas. Ridge genera unos coeficientes más pequeños, esto se debe a que, a diferencia de la regularización  $L_1$ , no hay un punto más probable en el que se produzca la tangente, sino que es la misma alrededor de la regularización. Por ello se minimiza el efecto de la correlación entre las

variables de entrada y el modelo generaliza mejor<sup>19</sup>. Al contrario que Lasso, Ridge función mejor cuando la mayoría de las variables son relevantes.

$$L_2(\vec{\omega}) = \lambda \omega^2$$

Entonces la función de coste o *Ridge regression* tiene la siguiente forma:

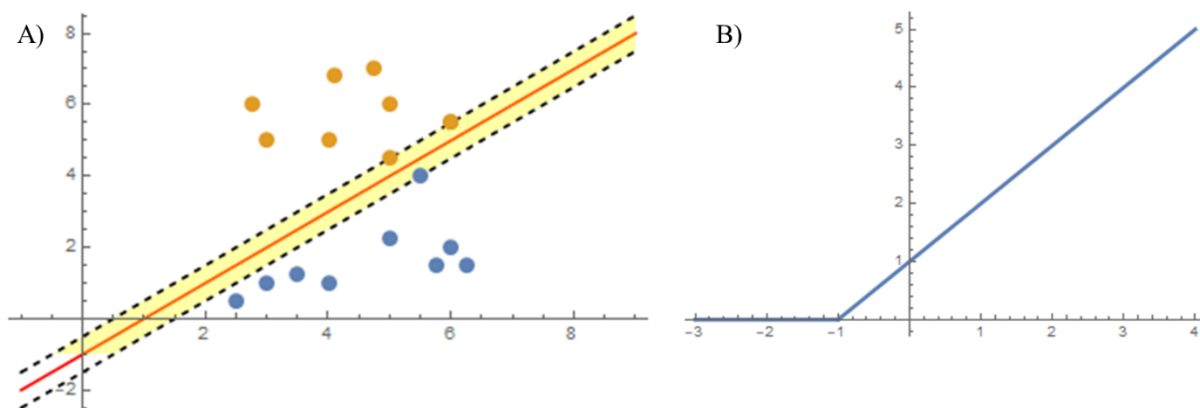
$$\sum_{i=0}^n f_{\vec{\omega}}(\vec{x}, \vec{y}) + \sum_{j=0}^P L_2(\vec{\omega}_j)$$

Existe una regularización combinada de ambos llamada ElasticNet, que se usará cuando se trabaje con un gran número de variables; donde algunas de estas variables tienen una gran correlación entre ellas y otras no son relevantes. Esta gráfica sería una unión entre las dos regularizaciones, la cual tendría forma de rombo abombado.

Aunque no existe una forma cerrada suficientemente simple para obtener el valor óptimo de  $\vec{\omega}$  con este modelo, se puede minimizar la función de coste utilizando métodos numéricos como el algoritmo Newton-Raphson<sup>18</sup> o el descenso de gradiente<sup>5</sup>.

### 2.3.1.2. Support Vector Machine.

Una máquina de vectores de soporte (SVM)<sup>20</sup> es un modelo de aprendizaje automático muy potente y versátil, es uno de los modelos más populares de ML. Las SVM son adecuadas para la clasificación de conjuntos de datos complejos.



**Ilustración 11.** A) Hiperplano de margen máximo y márgenes para un SVM entrenado con muestras de dos clases. B) Grafica de coste para valores igual a 0.

Las SVM en lugar de minimizar la pérdida empírica esperada en los datos de entrenamiento, intentan minimizar la pérdida de generalización esperada. No se sabe dónde pueden caer los nuevos puntos, pero bajo el supuesto probabilístico de que se extraen de la misma distribución que los datos de entrenamiento, se sugiere que se minimice la pérdida de generalización eligiendo el separador más alejado de los datos de entrenamiento. Llamamos a este separador, que se muestra en la Ilustración 11.A el separador de margen máximo. El margen es el ancho del área delimitada por líneas discontinuas en la figura, el doble de la distancia desde el separador hasta el punto de ejemplo más cercano.

La función de coste de SVM lo tiene la siguiente forma:

$$f_{\vec{\omega}}(\vec{x}, y) = \begin{cases} \max(0, 1 - \vec{\omega}\vec{x}), & y = 1 \\ \max(0, 1 + \vec{\omega}\vec{x}), & y = 0 \end{cases}$$

$$f_{\vec{\omega}}(\vec{x}, y) = y \cdot \max(0, 1 - \vec{\omega}\vec{x}) + (1 - y) \max(0, 1 + \vec{\omega}\vec{x})$$

En la Ilustración 11.B se observa el gráfico de  $f(x)$  cuando su clasificación es 0, si su clasificación fuera 1 entonces la gráfica sería su imagen simétrica respecto del eje vertical. Esto sucede porque si  $\omega \cdot x$  es positiva cuando su clasificación real es 0, el algoritmo lo está clasificando erróneamente y está generando mucha pérdida, mientras que si  $\omega \cdot x$  fuera negativa y su clasificación 0, no se generaría casi pérdida. Y sería lo mismo para la gráfica con clasificación igual a 1.

Añadiendo una de las regularizaciones explicadas en el punto 2.4.1.1, se establecería la función de coste para nuestro algoritmo. Por ejemplo, añadiendo la regularización  $L_2$ :

$$f_{\vec{\omega}}(\vec{x}, \vec{y}) = \sum_{i=0}^n y \cdot \max(0, 1 - \vec{\omega}\vec{x}) + (1 - y) \max(0, 1 + \vec{\omega}\vec{x}) + \sum_{j=0}^P L_2(\vec{\omega}_j)$$

La ilustración 11.B se observa el grafico de la función de coste cuando el valor de  $y$  es 0, para los valores de  $y$  igual a 1, la gráfica sería su reflejo respecto al eje vertical.

### 2.3.2. Clasificadores combinados.

Se entiende por clasificador combinados, aquellos clasificadores que utilicen múltiples algoritmos de aprendizaje. Estos clasificadores se utilizan para obtener un mejor rendimiento predictivo que clasificadores lineales. Su idea principal es contemplar múltiples hipótesis simultáneamente formando una hipótesis que se comporte mejor.

A causa del uso de varias hipótesis simultáneas, el coste computacional aumenta, por ello se suelen usar algoritmos rápidos, como por ejemplo los árboles de decisión.

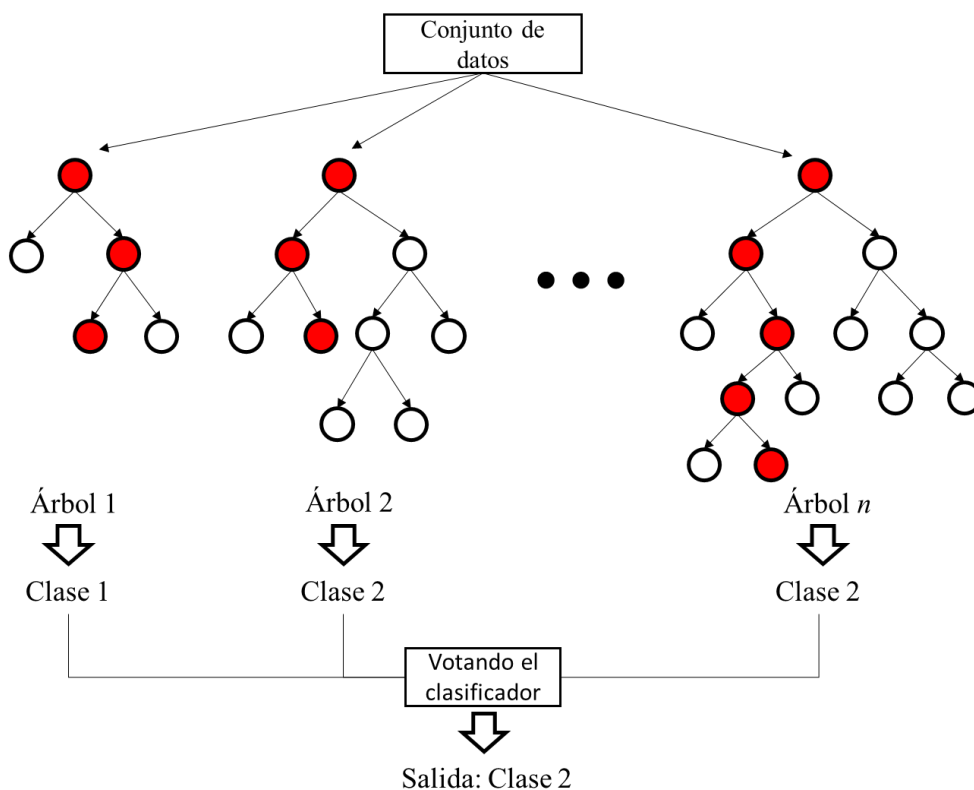


Ilustración 12. Diagrama del funcionamiento del algoritmo Random Forest.

### 2.3.2.1. Random Forest.

Los árboles de decisión son otro tipo de clasificadores, que tienen una apariencia similar a un árbol. Los árboles de decisión trabajan dividiendo el conjunto de datos repetidamente en base a una variable, de modo que cada división va a ser una rama que se volverá a subdividir hasta llegar a los nodos “hojas”. En cada división, el valor elegido para dividir a partir de la variable viene dado por un criterio de pureza que podemos seguir. El índice Gini es la media de varianza total entre las clases  $k$ :

$$G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$$

donde  $\hat{p}_{mk}$  representa la proporción de observación de entrenamiento en la  $m$ -ésima región que son de la clase  $k$ . Cuando los valores  $\hat{p}_{mk}$  son próximos a 0 o 1, el índice Gini adquiere un valor pequeño. El valor pequeño indica que un nodo contiene predominantemente observaciones de una clase, por esto Gini se considera una medida de pureza del nodo.

Una alternativa es la entropía:

$$D = - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$$

Como  $0 \leq \hat{p}_{mk} \leq 1$ , se puede deducir como  $0 \leq -\hat{p}_{mk} \log \hat{p}_{mk}$ . Se demuestra que la entropía toma un valor cercano a cero si los  $\hat{p}_{mk}$  están todos cercanos a 0 o 1. Por lo tanto, al igual que Gini, la entropía tendrá un valor pequeño cuando el nodo es puro.

A la hora de construir un árbol de decisión, el índice de Gini o la entropía se usan para medir la calidad de la división particular.

Los árboles de decisión predicen una respuesta cuantitativa. Al interpretar los resultados obtenidos, podemos observar las proporciones de clase entre los datos de entrenamiento de esa región.

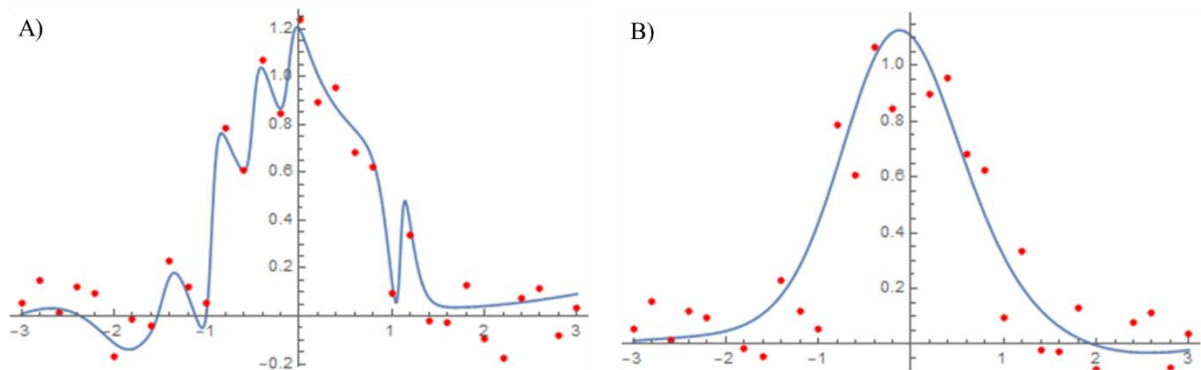
Random forest (RF) o bosques aleatorios, es un clasificador combinado generado a partir de árboles de decisión. RF va a ajustar múltiples árboles de decisión diferentes. A la hora de predecir un nuevo individuo, va a dejar que cada árbol vote cuál es la clasificación que considera correcta, estableciéndolo en una región del espacio de parámetros. Tras la votación de todos los árboles que componen el RF, al nuevo individuo se le asigna una probabilidad de pertenecer a una clase. Esta probabilidad viene dada por la suma de los resultados que obtiene cada árbol.

Para construir varios árboles de decisión diferentes, el algoritmo usa la técnica *bootstrapping*<sup>21</sup>. Esta técnica selecciona un subconjunto de los datos totales, para que cada árbol use datos diferentes. Cada vez que se considera una división en un árbol, se elige una muestra aleatoria de  $m$  predictores como candidatos divididos del conjunto completo de predictores  $p$ . La división solo puede usar uno de esos  $m$  predictores. Se toma una nueva muestra de  $m$  predictores en cada división, normalmente se utiliza  $m \approx \sqrt{p}$ .

Por ello al realizar un RF, en cada división del árbol, el algoritmo no considera la mayoría de predictores disponibles, sino que va variando cada vez el número de variables que utiliza para predecir. Por ejemplo, supongamos que hay un predictor muy fuerte en el conjunto de datos, junto con otros predictores moderadamente fuertes. Dado que están restringidas las variables por división, en las colecciones de árboles de decisión, la mayoría de los árboles tendrán este fuerte predictor en un punto diferente de su estructura. Por ello el RF es capaz de crear árboles diferentes para nuestro set de datos.

## 2.4. Evaluación.

Al realizar la evaluación de nuestros modelos, debemos tener en cuenta el sobreajuste de éstos (*overfitting*), que significa que el modelo funciona bien en los datos de entrenamiento, pero no generaliza bien.



**Ilustración 13.** A) En esta gráfica se puede observar un ejemplo de sobreajuste en la predicción de nuestros datos. B) En esta gráfica se puede observar la predicción de los mismos datos, esta vez sin producirse un sobreajuste.

El sobreajuste ocurre cuando el modelo es demasiado complejo en relación con la cantidad de ruido de los datos de entrenamiento. Las posibles soluciones son:

- Simplificar el modelo seleccionando otro con menos grados de libertad, reduciendo el número de atributos en los datos de entrenamiento o restringiendo el modelo.
- Recopilar más datos de entrenamiento.
- Reducir el ruido en los datos de entrenamiento.

Normalmente para entrenar un modelo de la mejor forma posible, éste no se entrena con todos los datos de los que se dispone, sino que este set de datos se divide en dos conjuntos diferentes: un conjunto de entrenamiento o *train* y otro de prueba o *test*. Medir las métricas usando los datos de entrenamiento no generalizaría bien para toda una población, y por eso es necesario un conjunto de datos de prueba. De esta manera entrenamos nuestro modelo con los datos de *train* para luego poder validarlo con los datos de *test*. Así se podrá observar si nuestro modelo predice correctamente o se ha producido *overfitting*. De esta manera, en el *test* calcularemos el AUC o  $F_1$ , entre otras métricas; y al pasarlas al conjunto de entrenamiento veremos si se ha producido un sobreajuste de nuestros datos. Hay algoritmos que por defecto sobreajustan los datos de entrenamiento, por ejemplo, el Random Forest. En el cual, si no se limitara la profundidad del árbol, se llegarían a nodos finales para cada individuo de los datos de entrenamiento, haciendo predicciones perfectas de estos datos.

En el caso que el conjunto de *train-test* provenga de una misma población, se pueden producir sesgos en el modelo. Ya sea por las características geográficas de esa población o por cómo se produce la recogida de los datos, entre otros muchos factores. Esto complica la generalización del modelo para otras poblaciones. Por ello, tener una población externa para usarla de *test*, es una clara ventaja ante una sola población, este factor le otorga más robustez al modelo.

En el caso clínico los sesgos que podemos encontrar son los siguientes:

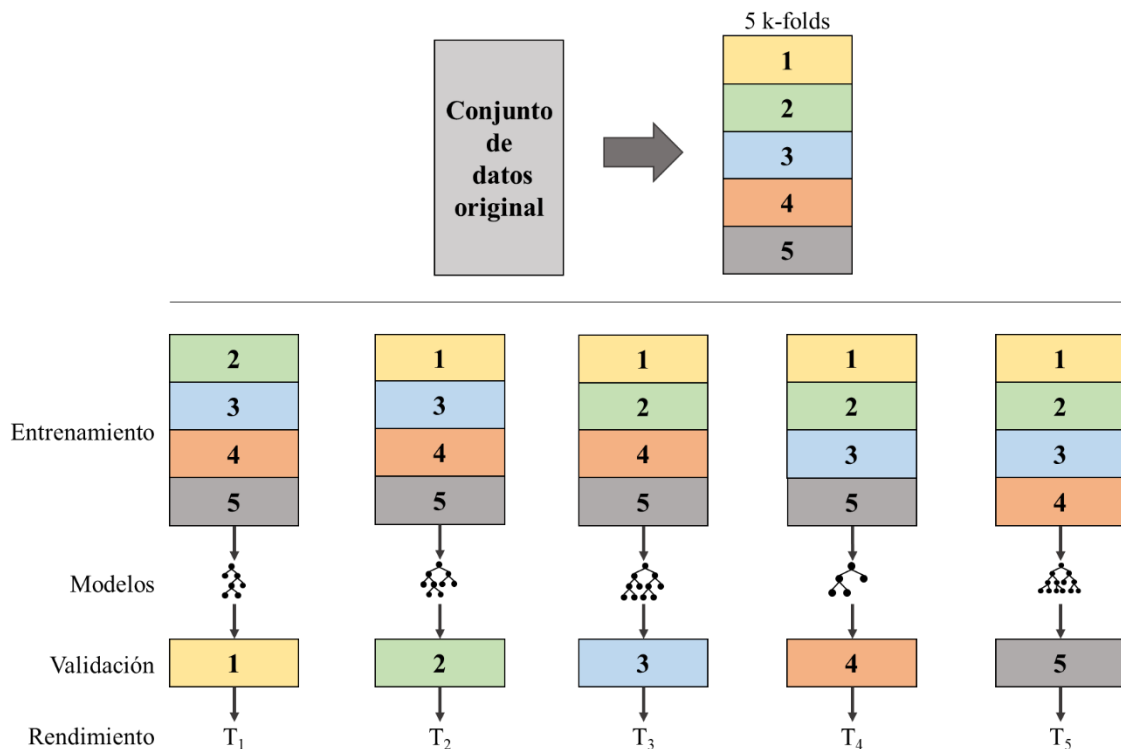
- Sesgos de selección.
  - Dentro de los sesgos de selección diferenciaremos entre dos:
    - Los producidos durante la selección de pacientes, hay que tener en cuenta edad de los pacientes, enfermedades anteriores, estado de salud, ...

- Los generados en el momento del análisis, puede haber problemas durante la realización del ensayo, que el paciente incumpla las normas, con el protocolo a seguir o con el programa de tratamiento que se había descrito.
- **Sesgos de medición.**  
El sesgo de medición se subdivide en dos sesgos más.
  - Los ocasionados durante la recopilación de datos, cuando los instrumentos o los sistemas para registrar los datos están defectuosos.
  - Los originados durante el análisis de los datos, nos podemos encontrar con subgrupos de pacientes que responden mejor a tratamientos, al usarlos para el análisis producirá un sesgo.
- **Sesgos de publicación.**  
En la práctica normal, las investigaciones publicadas suelen ser aquellas que obtienen buenos resultados o que son de gran importancia, dejando de lado aquellas investigaciones que no han obtenido buenos resultados. Al suceder esto, otros laboratorios o grupos de investigación pueden intentar realizar la misma investigación y llegar al mismo resultado fallido, gastando así recursos.

Para obtener un modelo más robusto usaremos dos técnicas, cross validation y validación externa.

### 2.4.1. Cross validation.

Cross-validation o validación cruzada<sup>19</sup> es un método para estimar el error de predicción en un conjunto de datos limitado. Para definir una mejor evaluación de que tal han funcionado los modelos, se establece una hipótesis como por ejemplo el AUC. Aunque el AUC tenga un valor próximo a 1, eso no quiere decir que el modelo generalice bien, aquí es donde entraría la validación cruzada<sup>5</sup>. Para



**Ilustración 14.** Diagrama del funcionamiento de la validación cruzada.



entrenar cuando se tiene pocos datos, lo mejor sería usar todos los datos para el entrenamiento, pero a su vez tener algún modo de test para medir el AUC u otras métricas.

La cross-validation mediante *k-fold*, trata de dividir los datos en  $k$  subconjuntos iguales. Posteriormente realizamos  $k$  rondas de aprendizaje, en cada ronda  $1/k$  de los datos se representa como conjunto de prueba y los datos restantes como datos de entrenamiento. El valor final de la métrica que estemos midiendo viene promediado por los  $k$  valores que se han obtenido. Además, también se puede utilizar *stratified k-fold*, en este caso las divisiones se seleccionan de modo que el valor medio de respuesta sea aproximadamente igual en todas las particiones. En el caso de partición binaria, significa que aproximadamente cada partición contiene la misma proporción de etiquetas de clase<sup>18</sup>. Es habitual hacer este procedimiento con repeticiones, haciéndolo  $r$  veces eligiendo una partición diferente cada vez, de este modo  $r*k$  serán las mediciones diferentes que obtenemos y tenemos que promediarlas para obtener el valor de la métrica.

A la vez que podemos usar la media para estimar el AUC, podemos usar la desviación estándar de esas medidas para estimar el error que cometemos al dar la media, al hacer repeticiones en la cross-validation tenemos que corregir esta estimación. Normalmente el error estándar es conocido como  $\sqrt{\sigma^2/n}$  donde en nuestro caso  $n = k*r$ ,  $k$  será igual al número de *folders*,  $r$  el número de repeticiones y  $\sigma^2$  es la varianza de AUC en las  $n$  mediciones. Pero este valor está muy correlacionado, así que hemos de corregirlo, por lo que el error estándar será  $\sqrt{\left(\frac{1}{n} + \frac{n_1}{n_2}\right) \sigma^2}$  donde  $n_1$  y  $n_2$  son los tamaños de test y train<sup>22</sup>.

### 2.4.2. Validación externa.

Se entiende por validación externa<sup>23</sup>, cuando en el entrenamiento de un modelo se utiliza como datos de test un conjunto de datos de distinta procedencia al conjunto de datos con el que se entrenó el modelo. Al realizar esta validación, puedes observar si el modelo generaliza bien para otras poblaciones o si el modelo contiene sesgos.

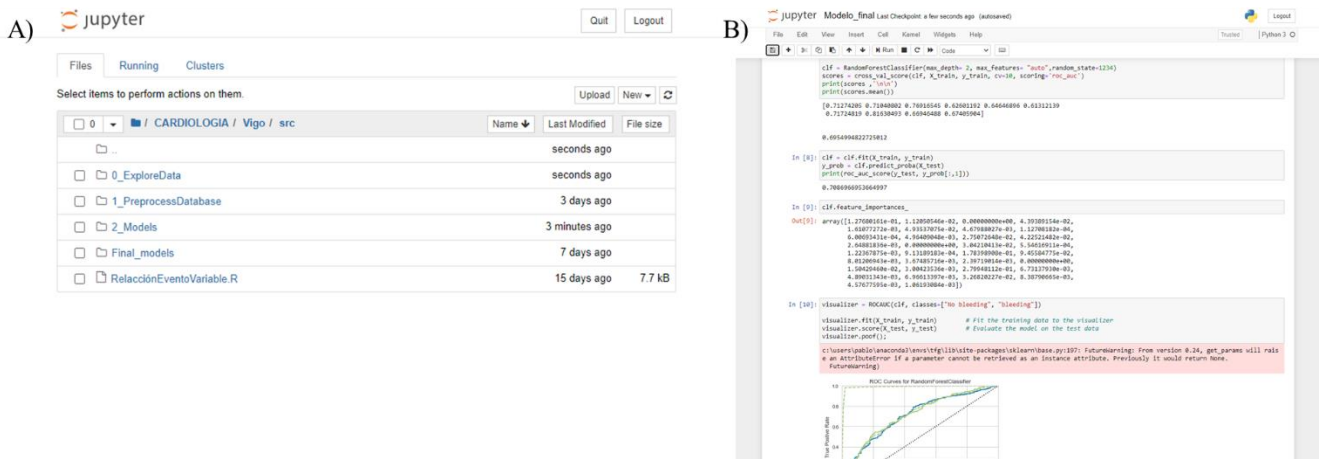
## 2.5. Entorno de desarrollo.

Para la realización de este trabajo de fin de grado se ha utilizado Python<sup>24</sup>. Python es un lenguaje de programación interpretado, dinámico y multiparadigma, ya que soporta orientación a objetos, programación imperativa y programación funcional. Su uso se ha realizado por medio de la interfaz Jupyter Notebook y la aplicación Anaconda Navigator.

Anaconda es un programa que permite instalar o tener acceso a diferentes interfaces de lenguajes de programación, como pueden ser Jupyter Notebook, Spyder o RStudio, además permite tener diferentes 'sesiones' en versiones diferentes del programa utilizado; y descargar y gestionar los paquetes usados durante la programación.

Jupyter Notebook es una interfaz que permite, en una aplicación web de código abierto, crear y compartir documentos que contengan código. Es una de las maneras más visuales que existe para editar y ejecutar código ya que permite mezclar código con formato texto.

Para comenzar a utilizar este software, deberá instalar Python en la versión 3.7 desde la página oficial y posteriormente instalar Anaconda Navigator, también desde la página oficial. El aspecto que tiene la interfaz Jupyter notebook se puede ver en la ilustración 15.A.



**Ilustración 15.** A) Captura de la interfaz de Jupyter Notebook. B) Captura de un notebook donde se escribe el código.

Dentro de Python se ha utilizado la Librería scikit-learn<sup>25</sup>, la cual ya tiene implementados los algoritmos de clasificación con los que se ha trabajado; y con matplotlib<sup>26</sup> y yellowbrick<sup>27</sup>, las cuales han servido para la elaboración de las gráficas durante el análisis de los datos o la visualización de la eficacia de los modelos. En la ilustración 15.B se puede observar un ejemplo de ajuste de un modelo y su predicción.

## 2.6. Planteamiento de nuestro problema.

Este trabajo se basó en la predicción del riesgo isquémico y hemorrágico de pacientes que han sufrido un infarto y han seguido la medicación DAP durante 12 meses, es decir, se va a predecir el riesgo del paciente de tener otro infarto o de sufrir una hemorragia en los primeros 12 meses. Este riesgo se va a utilizar para dar una recomendación al facultativo, a modo de ayuda con la decisión final del tratamiento y las revisiones que se va a someter al paciente.

## 2.7. Flujo de datos

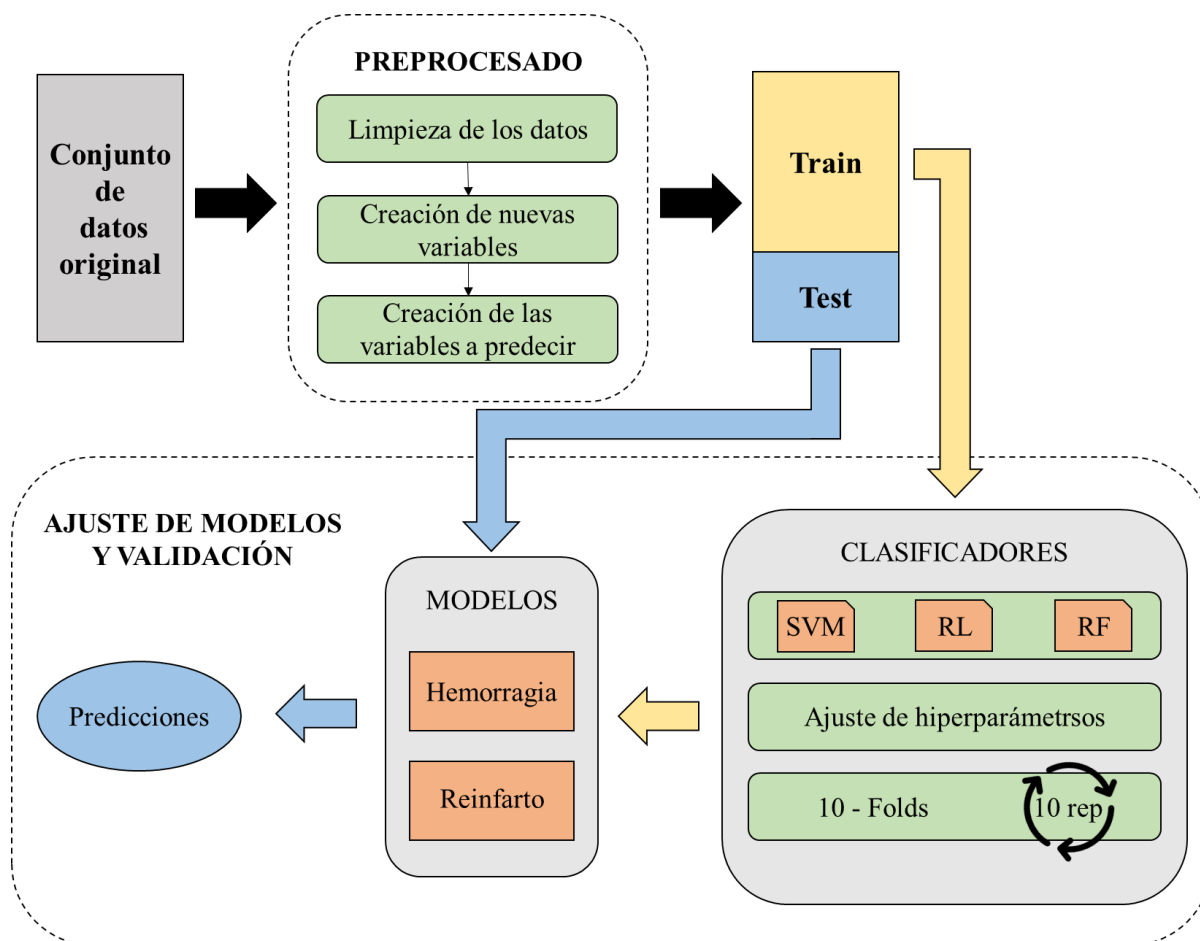
En la ilustración 16 se muestra el flujo de trabajo que se ha desarrollado durante este trabajo, se puede desglosar en dos pasos: (I) preprocesado de los datos y (II) ajuste de modelos y validación.

Todos los procedimientos explicados se pueden observar en la sección de “Código” de los ANEXOS. También se puede observar el código en mi cuenta de github<sup>28</sup>.

### I. Preprocesado de los datos

En este proceso de paso de variables categóricas a variables binarias, se crearon nuevas variables, se corrigieron los tiempos de los eventos y se generó las nuevas variables que se utilizarán para el estudio.

En “1\_PreprocessDatabase” del ANEXO se puede observar cómo se realizaron estos procedimientos en Python.



**Ilustración 16.** Flujo de trabajo llevado a cabo para la creación y validación de los modelos.

**a. Se pasaron las variables categóricas a variables binarias.**

Variables como sexo, diabetes, hipertensión, ..., se transformaron en variables binarias generando un diccionario y aplicando éste a las variables categóricas.

**b. Se creó una nueva variable IMC.**

Con la creación de esta variable fue diferente para los tres registros que componen nuestra población de estudio. Para la población “BleeMACS” no fue posible generar esta nueva variable dado que no contaba con las variables talla ni peso. Para el registro “RENAMI” esta variable ya existía bajo el nombre de Talla, así que se procedió a generar la nueva variable con los valores de la variable “talla”. Para los pacientes procedentes de “Vigo y Murcia” se generó esta variable a partir de las variables originales talla y peso, siguiendo la formula del índice de masa corporal, IMC:

$$IMC = \text{peso (Kg)} / \text{talla}^2(m)$$

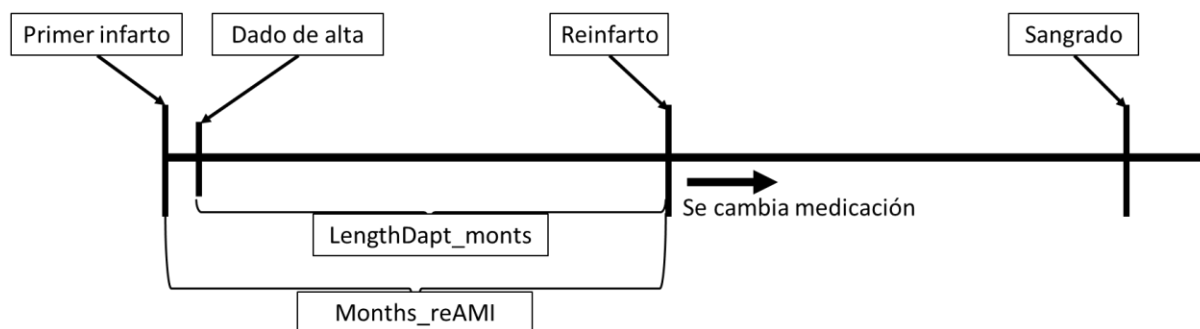
**c. Se crearon dos nuevas variables a partir de LVEFgroups.**

La variable “LVEFgroups” se refiere a la fracción de eyección del ventrículo izquierdo, en ella podemos distinguir tres grupos diferentes; LVEF menor que 40, entre 40 y 50, y mayor de

50. La variable referida a LVEF menor que 40 ya existe, por lo que se crearon otras dos variables binarias referidas a los otros dos grupos.

**d. Se establecieron bien los tiempos de sangrado y reinfarto en el registro en "Vigo y Murcia".**

En el registro "Vigo y Murcia" se encontró que de 441 pacientes que sufrían un reinfarto, 162 lo sufrían en las dos semanas siguientes a dejar el medicamento. Y de 286 pacientes que sufrieron un sangrado, 23 fueron en las dos semanas siguientes a dejar el tratamiento.



**Ilustración 17.** Workflow de un paciente en Vigo y Murcia. "LengthDapt\_months": tiempo de medicación, tiempo desde que se da el paciente de alta hasta el siguiente evento. "Months\_reAMI": tiempo de reinfarto, tiempo desde el primer infarto hasta el segundo. "Months\_Bleeding": tiempo de sangrado, tiempo desde el primer infarto hasta el sangrado.

Esto se debía a que el tiempo de medicación y el tiempo del evento no comenzaban en el mismo punto, pero sí terminaban cuando el paciente verdaderamente sufría ese evento. Este hecho viene explicado porque el tiempo de medicación se empezaba a contar en el momento del alta del paciente, mientras que el tiempo del evento se encontraba en el intervalo desde el infarto hasta el evento. Normalmente el tiempo que un paciente permanece hospitalizado tras un infarto del miocardio es 4-5 días.

Aparte, en este registro, tras el paciente sufrir un evento se le cambiaba la medicación, por ello el siguiente evento es inválido, dado que ya está alterado por la nueva medicación y no sigue las mismas condiciones que se están trabajando. Por ello, el segundo evento que se produce ha sido eliminado del estudio.

Para igualar los tiempos de medicación y el tiempo en el que se producía el evento, tras observar las ilustraciones 3 del ANEXO y tras hablar con facultativos del servicio de cardiología del Centro Asistencial Universitario de Salamanca, se acordó establecer sumar dos semanas al tiempo de los medicamentos para igualar los tiempos. Estas dos semanas vienen explicadas por la semana de ingreso del paciente más la semana que tardan en eliminarse completamente los efectos de los medicamentos en el paciente.

**e. Se crearon nuevas variables a predecir. El infarto/sangrado antes de los 12 meses durante la medicación.**

Se crearon cuatro nuevas variables teniendo en cuenta si el evento se produce durante o después del medicamento, además del tiempo en el que este se produce. Es decir, tendremos cuatro variables referentes a sangrado o infarto (dos variables a cada evento), en las cuales se

distinguió si el paciente sufrió el evento durante el tiempo que está tomando la medicación; o después de que dejará de tomar la medicación.

Aunque se crearon cuatro nuevas variables, se ha trabajado solo con dos de ellas. Las referentes al evento durante el tiempo que el paciente toma la medicación. A partir de estas dos variables se crearon otras dos nuevas, en la cual referenciaba si el evento se había producido antes de los 12 meses. Estas dos nuevas variables serán las que se usarán para la entrenar nuestro modelo y las cuales se intentará predecir.

## II. Ajuste de modelos y validación

En “2\_Model” del ANEXO se puede observar cómo se realizaron estos procedimientos en Python, además en mi cuenta de github<sup>28</sup> se encuentra el código utilizado en el sistema KoopaML mencionado a continuación.

Para la realización de este paso, lo primero que se llevó a cabo fue distinguir entre conjunto de entrenamiento y conjunto de prueba. Para generar el conjunto de entrenamiento se utilizaron los pacientes procedentes de los registros BleeMACS y RENAMI. Para el conjunto de validación se usaron los pacientes correspondientes a Vigo y Murcia. De esta manera podremos realizar una validación externa, dado que los datos proceden de conjuntos diferentes.

Para la obtención del riesgo de un nuevo paciente de sufrir un reinfarto o una hemorragia, se usaron los tres algoritmos explicados anteriormente en el punto 2.4. La elaboración del riesgo en cada evento se hizo independiente. Es decir, por un lado, se realizó la predicción de reinfarto. Por otro lado, se realizó el pronóstico de hemorragia.

Para cada uno de los eventos se usaron los algoritmos de ML: Regresión Logística, Support Vector Machine y Random Forest. El uso de estos algoritmos se realizó mediante la librería Scikit-learn<sup>25</sup>.

El conjunto de entrenamiento se utilizó para ajustar los 3 algoritmos de aprendizaje, donde se ajustaron los hiperparámetros de forma individual. Los hiperparámetros que se variaron en cada algoritmo se pueden observar en la tabla 4.

**Tabla 4.** Hiperparámetros utilizados para la creación de los modelos

|                            |              | <b>Hiperparámetros por modelo</b> |              |                      |                |
|----------------------------|--------------|-----------------------------------|--------------|----------------------|----------------|
| <b>Regresión Logística</b> |              | <b>Support Vector Machine</b>     |              | <b>Random Forest</b> |                |
| <b>C</b>                   | [0.1, 1, 10] | <b>C</b>                          | [0.1, 1, 10] | <b>max_depth</b>     | [1, 2, 5, nan] |
| <b>l1_ratio</b>            | [0, 0.5, 1]  | <b>Kernel</b>                     | linear       | <b>max_features</b>  | [auto, 1]      |
| <b>penalty</b>             | elasticnet   |                                   |              |                      |                |
| <b>Solver</b>              | saga         |                                   |              |                      |                |

Los hiperparámetros que se observan en la tabla 4 son aquellos que se modificaron a la hora de realizar los algoritmos con la Librería Scikit-learn, los demás hiperparámetros que consolidan los algoritmos no fueron modificados y se utilizaron sus valores por defecto. A continuación, se explica el significado de los hiperparámetros para la notación que se ha seguido durante el presente trabajo:

### i. Regresión logística.

El valor “C” equivale a la inversa del valor  $\lambda$  que multiplica a la regularización en la función de coste de RL (punto 2.3.1.1). Los valores de “l1\_ratio” van a la par que “penalty”, habiendo escogido una penalización en la regularización de elasticnet, al variar “l1\_ratio” se escogerá una regularización  $L_1$ , elasticnet o  $L_2$ , respectivamente según están escritos en la tabla 4. Como “solver”

se utilizó “saga” para el problema de optimización, “saga” realiza descenso del gradiente y los demás métodos no aceptaban la regularización  $L_1$  o elasticnet.

ii. Support Vector Machine.

El valor “C” adquiere el mismo significado en la RL, es la inversa de la constante  $\lambda$  que multiplica en la regularización de la función de coste. El núcleo que se ha usado para este problema ha sido lineal.

iii. Random Forest.

Los valores de “max:depth” hacen referencia a la profundidad máxima del árbol, cuando se establece “nan” el árbol crecerá hasta obtener nodos hoja puros. “max\_features” hace referencia al número de variables que tiene en cuenta en la división de cada árbol, el valor “auto” es la raíz del número de variables.

El ajuste de hiperparámetros en busca de la mejor combinación que nuestros modelos se ajusten a nuestros datos se denomina “hyperparameter tuning”. A la hora de entrenar el modelo con el conjunto de datos de entrenamiento, se realiza la elección de hiperparámetros probando todas las combinaciones y evaluándolas con una *cross-validation* de 10  *folds*.

En el momento de realizar la validación interna de la *cross-validation* anterior, no podemos usar la elección de hiperparámetros que hemos hecho anteriormente, porque hemos usado todos los datos de entrenamiento, y por lo tanto es una elección que hemos hecho con la información de todos los  *folds*.

Para solucionar este sesgo, cada vez que haya un  *fold* en el proceso de validación interna se volverá a elegir hiperparámetros usando solo la información proveniente de la información de los  *folds* de entrenamiento. Esto requerirá hacer una *cross-validation* de elección de hiperparámetros. Como la validación interna estamos haciendo 10 repeticiones y 10  *folds*, se requerirá hacer 100 elecciones de hiperparámetros, que a priori, podrían ser diferentes entre sí.

Puede parecer paradójico que la estimación del rendimiento de un modelo con unos hiperparámetros concretos combine los resultados de modelos con otras elecciones de hiperparámetros, pero de este modo evitaremos sobreestimar el rendimiento del algoritmo por filtrar información de los  *folds* de test a los  *folds* de training.

El proceso explicado en el párrafo anterior se realiza con cada combinación de hiperparámetros, esto adquiere el nombre de validación cruzada anidada. Como se puede observar el procedimiento es bastante repetitivo y requiere de un largo tiempo de computación. Por ellos, para automatizar el proceso, se utilizó un sistema llamado KoopaML<sup>29</sup>. Este sistema nos permite crear y automatizar diferentes flujos de trabajo gracias a la librería de Python “Luigi”. A la hora de ejecutarse, realizará los diferentes procesos de manera paralelizada.

Para apreciar el nivel de computación realizado simplemente durante la selección de los mejores hiperparámetros. Se realiza una *cross-validation* de 10  *folds* y 10 repeticiones, y a continuación se realiza otra *cross-validation* de 10  *folds*. Hasta este punto se habría entrenado 1000 veces nuestro modelo, pero esto sería para un solo conjunto de hiperparámetros, nosotros trabajamos con un total de 20 posibilidades de hiperparámetros distintas entre los tres algoritmos. Además, trabajamos con dos modelos finales, uno de hemorragia y otro de reinfarto. Con lo cual se entrenan los modelos 40.000 veces para hacer la validación interna, escoger los hiperparámetros, y entrenar un modelo definitivo.

Tras realizar el hyperparameter tuning se escogieron los mejores hiperparámetros de cada algoritmo. Se escogieron las siguientes combinaciones de hiperparámetros para los algoritmos:

A. Regresión logística.

Para la predicción de hemorragia y reinfarto se escogió la misma combinación de hiperparámetros. “C” = 0.1, “l1\_ratio” = 1, “penalty” = elastic y “solver” = saga.

B. Support Vector Machine.

En este caso, la predicción de ambos eventos también comparte los mismos hiperparámetros. “C” = 1 y “kernel” = linear.

C. Random Forest.

A diferencia de los dos algoritmos anteriores, en RF la predicción de los dos eventos presenta un conjunto de hiperparámetros diferentes. Para la predicción de hemorragia la combinación que se escogió fue “max\_depth” = 2 y “max\_features” = auto. Para la predicción de reinfarto se eligió “max\_depth” = 2 y “max\_features” = 1.

Con ello se generaron los respectivos modelos para cada algoritmo.

Por último, se realizó la validación externa donde se obtuvo el rendimiento de los modelos desarrollados sobre nuestro conjunto de validación externa.

### 3. Resultados

En la tabla 5 se muestran las frecuencias relativas de hemorragia o reinfarto durante los 12 primeros meses con medicación. Podemos observar la baja incidencia que tienen ambos eventos.

**Tabla 5.** Frecuencias de hemorragia o reinfarto.

|                                  |           | <b>Hemorragia</b> | <b>Reinfarto</b> |
|----------------------------------|-----------|-------------------|------------------|
| <b>Conjunto de entrenamiento</b> | <b>Sí</b> | 510 (2.58%)       | 542 (2.74%)      |
|                                  | <b>No</b> | 19276 (97.42%)    | 19244 (97.26%)   |
| <b>Conjunto de validación</b>    | <b>Sí</b> | 123 (1.96%)       | 141 (2.24%)      |
|                                  | <b>No</b> | 6167 (98.04%)     | 6149 (97.76%)    |

Se observó si había diferencias en las características de los pacientes respecto a si se presentaba el evento o no, para realizar esto se utilizó un test de Fisher para las variables categóricas y un test de Student para las variables continuas. Como hipótesis nula se estableció que los grupos eran iguales, como hipótesis alternativa que los grupos eran diferentes. Tras realizar la prueba, en los p-valores menores a 0.05 se rechazará la hipótesis nula y se aceptará la alternativa, por lo que podremos decir que existe diferencia entre los grupos dentro de la variable. Este proceso se realizó para nuestro conjunto de datos de entrenamiento y nuestro conjunto de validación, para observar si hay diferencias entre ellos. En las tablas 7 y 8 podemos observar los resultados obtenidos, la tabla 7 hace referencia a los datos utilizados para el entrenamiento de los modelos y la tabla 8 para los datos usados en la validación. Se observa que en algunas variables esta prueba no tiene un resultado. Esto viene dado porque en los conjuntos de datos siempre se elegía una misma variable, como puede ser el caso de ausencia de stent en los registros BleemACS y RENAMI (la presencia de stent era una de las condiciones de estudio) o en el caso del medicamento seguido en el tratamiento DAP (como se observa en la tabla 2 los medicamentos seguidos por cada registro). Llama la atención que dos tercios de las variables en los datos procedentes de Vigo y Murcia acepten la hipótesis nula y por tanto no haya diferencias entre los grupos (Tabla 8).

En la tabla 2 del ANEXO podemos observar los hiperparámetros evaluados y sus respectivos resultados en cada fold de la validación cruzada para la elección de hiperparámetros. Esta información la podemos ver resumida para el mejor conjunto de hiperparámetros en la tabla 6 y en la ilustración 18, donde se puede observar la capacidad discriminatoria de los 3 clasificadores de aprendizaje automático utilizados tras ajustar los modelos para hemorragia o reinfarto. Tal y como se contempla, el modelo que generaliza mejor fue RF. Aunque RF sea el modelo que mejor generaliza se decidió realizar un modelo para cada algoritmo para ver si hay diferencias a la hora de evaluar el rendimiento en un conjunto de datos externos.

**Tabla 6.** Resultados de los algoritmos utilizados para ajustar los modelos de clasificación en el conjunto de entrenamiento.

|                   |                               | <b>Resultados del entrenamiento (AUC)</b> |                   |
|-------------------|-------------------------------|---|-------------------|
| <b>Modelo</b>     |                               | <b>Media</b>                              | <b>Desviación</b> |
| <b>Hemorragia</b> | <b>Regresión logística</b>    | 0.72                                      | 0.05              |
|                   | <b>Support Vector Machine</b> | 0.56                                      | 0.07              |
|                   | <b>Random forest</b>          | <b>0.72</b>                               | 0.05              |
| <b>Reinfarto</b>  | <b>Regresión logística</b>    | 0.73                                      | 0.08              |



|                               |             |      |
|-------------------------------|-------------|------|
| <b>Support Vector Machine</b> | 0.55        | 0.05 |
| <b>Random forest</b>          | <b>0.73</b> | 0.05 |

**Tabla 7.** Resultados del test de Fisher y test de Student en las variables predictoras en el conjunto de datos uso para entrenamiento (Registro BleeMACS y RENAMI).

| Variable                         | Hemorragia                   |                              |         | Reinfarto                    |                              |         |
|----------------------------------|------------------------------|------------------------------|---------|------------------------------|------------------------------|---------|
|                                  | No<br>(N =19276)             | Sí<br>(N = 510)              | p-valor | No<br>(N =19244)             | Sí<br>(N = 542)              | p-valor |
|                                  | n(%)<br>$\bar{x} \pm \sigma$ | n(%)<br>$\bar{x} \pm \sigma$ |         | n(%)<br>$\bar{x} \pm \sigma$ | n(%)<br>$\bar{x} \pm \sigma$ |         |
| <b>Age</b>                       | 62.9±12.5                    | 69.4±11.6                    | < 0.001 | 62.9±12.4                    | 67.4±12.7                    | < 0.001 |
| <b>Sex</b>                       | 1.8±0.4                      | 1.7±0.5                      | < 0.001 | 1.8±0.4                      | 1.7±0.4                      | 0.005   |
| <b>Diabetes</b>                  | 4873<br>(25.3%)              | 138<br>(27.1%)               | 0.354   | 4816<br>(25.0%)              | 195<br>(36.0%)               | < 0.001 |
| <b>Hypertension</b>              | 11043<br>(57.3%)             | 361<br>(70.8%)               | < 0.001 | 11076<br>(57.6%)             | 328<br>(60.5%)               | 0.172   |
| <b>Dyslipemia</b>                | 10117<br>(52.5%)             | 275<br>(53.9%)               | 0.530   | 10146<br>(52.7%)             | 246<br>(45.4%)               | 0.001   |
| <b>PeripheralArteryDisease</b>   | 938 (4.9%)                   | 54 (10.6%)                   | < 0.001 | 917 (4.8%)                   | 75 (13.8%)                   | < 0.001 |
| <b>PriorMyocardialInfarction</b> | 2479<br>(12.9%)              | 84 (16.5%)                   | 0.019   | 2449<br>(12.7%)              | 114<br>(21.0%)               | < 0.001 |
| <b>PriorPCI</b>                  | 2640<br>(13.7%)              | 73 (14.3%)                   | 0.695   | 2606<br>(13.5%)              | 107<br>(19.7%)               | < 0.001 |
| <b>PriorCABG</b>                 | 528 (2.7%)                   | 18 (3.5%)                    | 0.272   | 516 (2.7%)                   | 30 (5.5%)                    | < 0.001 |
| <b>PriorStroke</b>               | 1084 (5.6%)                  | 44 (8.6%)                    | 0.006   | 1085 (5.6%)                  | 43 (7.9%)                    | 0.030   |
| <b>Cancer</b>                    | 1066 (5.5%)                  | 57 (11.2%)                   | < 0.001 | 1062 (5.5%)                  | 61 (11.3%)                   | < 0.001 |
| <b>Prior_Bleeding</b>            | 849 (4.4%)                   | 55 (10.8%)                   | < 0.001 | 876 (4.6%)                   | 28 (5.2%)                    | 0.466   |
| <b>STEMI</b>                     | 11186<br>(58.0%)             | 286<br>(56.1%)               | 0.388   | 11112<br>(57.7%)             | 360<br>(66.4%)               | < 0.001 |
| <b>UnstableAngina</b>            | 2397<br>(12.4%)              | 61 (12.0%)                   | 0.786   | 2423<br>(12.6%)              | 35 (6.5%)                    | < 0.001 |
| <b>Killip2orMore</b>             | 1767 (9.2%)                  | 87 (17.1%)                   | < 0.001 | 1743 (9.1%)                  | 111<br>(20.5%)               | < 0.001 |
| <b>LVEFless40</b>                | 2153<br>(11.2%)              | 72 (14.1%)                   | 0.046   | 2136<br>(11.1%)              | 89 (16.4%)                   | < 0.001 |
| <b>LVEFless50</b>                | 16426<br>(85.2%)             | 433<br>(84.9%)               | 0.850   | 16384<br>(85.1%)             | 475<br>(87.6%)               | 0.111   |
| <b>LVEFafter50</b>               | 4299<br>(22.3%)              | 127<br>(24.9%)               | 0.162   | 4292<br>(22.3%)              | 134<br>(24.7%)               | 0.191   |
| <b>Hemoglobin_Admission</b>      | 14.1±1.6                     | 13.0±2.0                     | < 0.001 | 14.0±1.6                     | 13.7±1.8                     | < 0.001 |
| <b>Creatinine_Admission</b>      | 1.0±0.5                      | 1.1±0.7                      | < 0.001 | 1.0±0.5                      | 1.1±0.7                      | < 0.001 |
| <b>Femoral_Access</b>            | 8893<br>(46.1%)              | 261<br>(51.2%)               | 0.025   | 8841<br>(45.9%)              | 313<br>(57.7%)               | < 0.001 |

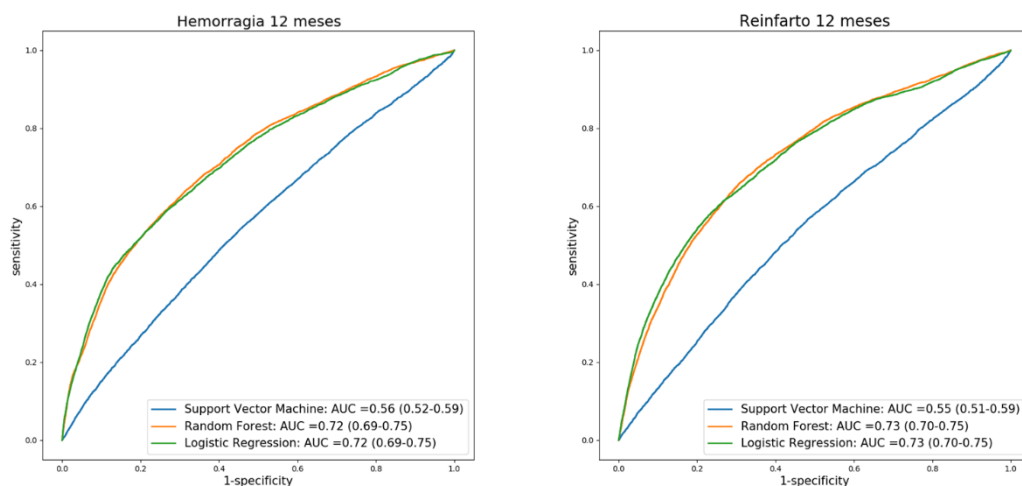
|                                   |                  |                |         |                  |                |         |
|-----------------------------------|------------------|----------------|---------|------------------|----------------|---------|
| <b>Multivessel</b>                | 7699<br>(39.9%)  | 228<br>(44.7%) | 0.031   | 7591<br>(39.4%)  | 336<br>(62.0%) | < 0.001 |
| <b>DrugElutingStent</b>           | 8745<br>(45.4%)  | 203<br>(39.8%) | 0.013   | 8773<br>(45.6%)  | 175<br>(32.3%) | < 0.001 |
| <b>NoStent</b>                    | 0.0±0.0          | 0.0±0.0        | ---     | 0.0±0.0          | 0.0±0.0        | ---     |
| <b>Revascularization_Complete</b> | 14147<br>(73.4%) | 337<br>(66.1%) | < 0.001 | 14182<br>(73.7%) | 302<br>(55.7%) | < 0.001 |
| <b>ReAMI_inhospital</b>           | 250 (1.3%)       | 4 (0.8%)       | 0.424   | 225 (1.2%)       | 29 (5.4%)      | < 0.001 |
| <b>Bleeding_inhospital</b>        | 862 (4.5%)       | 104<br>(20.4%) | < 0.001 | 909 (4.7%)       | 57 (10.5%)     | < 0.001 |
| <b>Clopidogrel</b>                | 13672<br>(70.9%) | 396<br>(77.6%) | 0.001   | 13622<br>(70.8%) | 446<br>(82.3%) | < 0.001 |
| <b>Ticagrelor</b>                 | 3277<br>(17.0%)  | 76 (14.9%)     | 0.232   | 3296<br>(17.1%)  | 57 (10.5%)     | < 0.001 |
| <b>Prasugrel</b>                  | 2327<br>(12.1%)  | 38 (7.5%)      | 0.001   | 2326<br>(12.1%)  | 39 (7.2%)      | < 0.001 |
| <b>OralAnticoagulation</b>        | 747 (3.9%)       | 52 (10.2%)     | < 0.001 | 758 (3.9%)       | 41 (7.6%)      | < 0.001 |
| <b>Betablockers</b>               | 16248<br>(84.3%) | 405<br>(79.4%) | 0.004   | 16208<br>(84.2%) | 445<br>(82.1%) | 0.189   |
| <b>ACEI_ARB</b>                   | 15267<br>(79.2%) | 414<br>(81.2%) | 0.293   | 15279<br>(79.4%) | 402<br>(74.2%) | 0.004   |
| <b>Statin</b>                     | 18209<br>(94.5%) | 480<br>(94.1%) | 0.695   | 18195<br>(94.5%) | 494<br>(91.1%) | 0.002   |

**Tabla 8.** Resultados del test de Fisher y test de Student en las variables predictoras en el conjunto de datos uso para validación externa (Registro Vigo y Murcia).

| Variable                         | Hemorragia           |                      |         | Reinfarto            |                      |         |
|----------------------------------|----------------------|----------------------|---------|----------------------|----------------------|---------|
|                                  | No<br>(N = 6167)     | Sí<br>(N = 123)      | p-valor | No<br>(N = 6149)     | Sí<br>(N = 141)      | p-valor |
|                                  | n(%)                 | n(%)                 |         | n(%)                 | n(%)                 |         |
|                                  | $\bar{x} \pm \sigma$ | $\bar{x} \pm \sigma$ |         | $\bar{x} \pm \sigma$ | $\bar{x} \pm \sigma$ |         |
| <b>Age</b>                       | 67.2±13.2            | 74.2±10.9            | < 0.001 | 67.3±13.2            | 69.1±13.6            | 0.109   |
| <b>Sex</b>                       | 1.7±0.4              | 1.7±0.5              | 0.013   | 1.7±0.4              | 1.7±0.4              | 0.522   |
| <b>Diabetes</b>                  | 1913 (31.0%)         | 45 (36.6%)           | 0.201   | 1899<br>(30.9%)      | 59 (41.8%)           | 0.007   |
| <b>Hypertension</b>              | 4134 (67.0%)         | 91 (74.0%)           | 0.120   | 4123<br>(67.1%)      | 102<br>(72.3%)       | 0.204   |
| <b>Dyslipemia</b>                | 4128 (66.9%)         | 87 (70.7%)           | 0.438   | 4114<br>(66.9%)      | 101<br>(71.6%)       | 0.277   |
| <b>PeripheralArteryDisease</b>   | 583 (9.5%)           | 25 (20.3%)           | < 0.001 | 590 (9.6%)           | 18 (12.8%)           | 0.196   |
| <b>PriorMyocardialInfarction</b> | 1995 (32.3%)         | 51 (41.5%)           | 0.041   | 1977<br>(32.2%)      | 69 (48.9%)           | < 0.001 |
| <b>PriorPCI</b>                  | 1348 (21.9%)         | 34 (27.6%)           | 0.125   | 1331<br>(21.6%)      | 51 (36.2%)           | < 0.001 |
| <b>PriorCABG</b>                 | 286 (4.6%)           | 15 (12.2%)           | 0.001   | 283 (4.6%)           | 18 (12.8%)           | < 0.001 |
| <b>PriorStroke</b>               | 501 (8.1%)           | 10 (8.1%)            | 1.000   | 504 (8.2%)           | 7 (5.0%)             | 0.210   |
| <b>Cancer</b>                    | 527 (8.5%)           | 11 (8.9%)            | 0.870   | 519 (8.4%)           | 19 (13.5%)           | 0.046   |
| <b>Prior_Bleeding</b>            | 192 (3.1%)           | 9 (7.3%)             | 0.017   | 198 (3.2%)           | 3 (2.1%)             | 0.630   |
| <b>STEMI</b>                     | 2273 (36.9%)         | 39 (31.7%)           | 0.258   | 2279<br>(37.1%)      | 33 (23.4%)           | 0.001   |

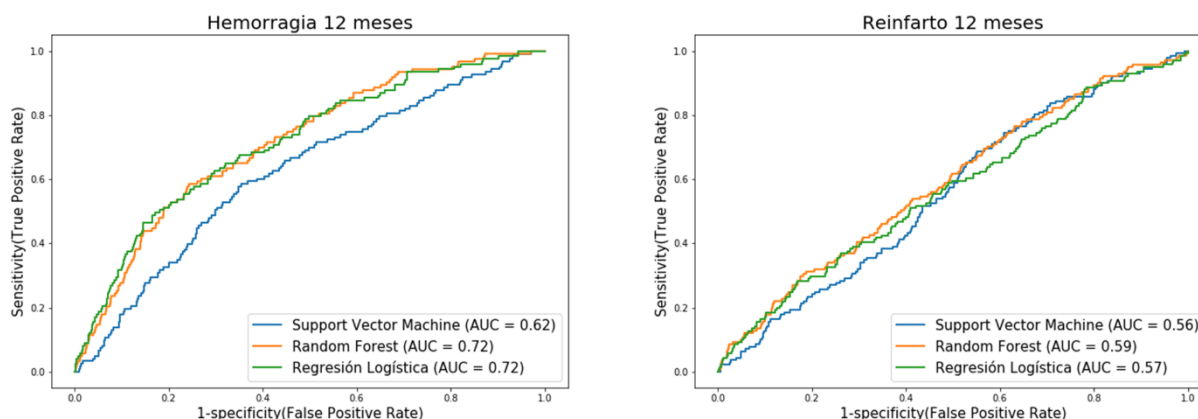
|                                   |              |             |         |              |             |       |
|-----------------------------------|--------------|-------------|---------|--------------|-------------|-------|
| <b>UnstableAngina</b>             | 871 (14.1%)  | 25 (20.3%)  | 0.067   | 875 (14.2%)  | 21 (14.9%)  | 0.807 |
| <b>Killip2orMore</b>              | 1046 (17.0%) | 34 (27.6%)  | 0.003   | 1056 (17.2%) | 24 (17.0%)  | 1.000 |
| <b>LVEFless40</b>                 | 949 (15.4%)  | 19 (15.4%)  | 1.000   | 944 (15.4%)  | 24 (17.0%)  | 0.556 |
| <b>LVEFless50</b>                 | 5201 (84.3%) | 107 (87.0%) | 0.530   | 5195 (84.5%) | 113 (80.1%) | 0.160 |
| <b>LVEFAfter50</b>                | 1620 (26.3%) | 31 (25.2%)  | 0.837   | 1608 (26.2%) | 43 (30.5%)  | 0.246 |
| <b>Hemoglobin_Admission</b>       | 13.9±1.8     | 12.8±2.1    | < 0.001 | 13.9±1.9     | 13.8±1.9    | 0.309 |
| <b>Creatinine_Admission</b>       | 1.1±0.7      | 1.2±0.6     | 0.107   | 1.1±0.6      | 1.1±0.9     | 0.228 |
| <b>Femoral_Access</b>             | 620 (10.1%)  | 15 (12.2%)  | 0.448   | 616 (10.0%)  | 19 (13.5%)  | 0.201 |
| <b>Multivessel</b>                | 2362 (38.3%) | 57 (46.3%)  | 0.075   | 2354 (38.3%) | 65 (46.1%)  | 0.066 |
| <b>DrugElutingStent</b>           | 3161 (51.3%) | 61 (49.6%)  | 0.717   | 3141 (51.1%) | 81 (57.4%)  | 0.148 |
| <b>NoStent</b>                    | 1116 (18.1%) | 32 (26.0%)  | 0.033   | 1129 (18.4%) | 19 (13.5%)  | 0.152 |
| <b>Revascularization_Complete</b> | 3562 (57.8%) | 60 (48.8%)  | 0.053   | 3552 (57.8%) | 70 (49.6%)  | 0.058 |
| <b>ReAMI_inhospital</b>           | 163 (2.6%)   | 3 (2.4%)    | 1.000   | 162 (2.6%)   | 4 (2.8%)    | 0.789 |
| <b>Bleeding_inhospital</b>        | 116 (1.9%)   | 6 (4.9%)    | 0.032   | 119 (1.9%)   | 3 (2.1%)    | 0.755 |
| <b>Clopidogrel</b>                | 1.0±0.0      | 1.0±0.0     | ---     | 1.0±0.0      | 1.0±0.0     | ---   |
| <b>Ticagrelor</b>                 | 0.0±0.0      | 0.0±0.0     | ---     | 0.0±0.0      | 0.0±0.0     | ---   |
| <b>Prasugrel</b>                  | 0.0±0.0      | 0.0±0.0     | ---     | 0.0±0.0      | 0.0±0.0     | ---   |
| <b>OralAnticoagulation</b>        | 461 (7.5%)   | 27 (22.0%)  | < 0.001 | 484 (7.9%)   | 4 (2.8%)    | 0.024 |
| <b>Betablockers</b>               | 5183 (84.0%) | 95 (77.2%)  | 0.047   | 5162 (83.9%) | 116 (82.3%) | 0.563 |
| <b>ACEI_ARB</b>                   | 4550 (73.8%) | 97 (78.9%)  | 0.215   | 4554 (74.1%) | 93 (66.0%)  | 0.033 |
| <b>Statin</b>                     | 5919 (96.0%) | 112 (91.1%) | 0.018   | 5898 (95.9%) | 133 (94.3%) | 0.385 |

Tras la construcción de cada modelo con su respectivo algoritmo de clasificación, se evaluó el rendimiento de estos mediante validación externa. Para ello se utilizó el conjunto de datos que datos que habíamos separado previamente procedente de Vigo y Murcia.



**Ilustración 18.** Curva ROC validación interna.

En la ilustración 19 podemos observar los resultados obtenidos tras realizar la validación externa. Podemos advertir que nuestros modelos generalizan mejor en el evento hemorragia, respecto al evento reinfarto.



**Ilustración 19.** Curva ROC validación externa.

En el evento hemorragia podemos observar cómo RF y RL tienen rendimiento parecido (AUC = 0.72), mientras que anteriormente en el entrenamiento del modelo RF generalizaba mejor que RL. Finalmente nos quedaremos con el algoritmo RF, dado que en el entrenamiento generaliza mejor y tiene el mayor rendimiento de los tres algoritmos.

En el evento referente a reinfarto se observa que los modelos no tienen un rendimiento muy alto. El mejor modelo es el que contiene el algoritmo RF con un AUC de 0,59. Los modelos no tienen un resultado de rendimiento tan alto respecto a cómo generalizaban en su entrenamiento.

Los modelos de aprendizaje automático tienen la capacidad de ajustarse dependiendo de si queremos una mayor sensibilidad o especificidad, esto sucede variando el punto de corte, donde podremos predecir correctamente un mayor porcentaje de evento del paciente o un mayor porcentaje de no evento. En la tabla 9 podemos observar la eficacia de nuestro modelo RF diferenciando entre optimizar dos métricas diferentes:  $F_1$  o suma de sensibilidad y especificidad. Podemos observar en ambos cómo optimizando el  $F_1$  obtenemos un mayor porcentaje de eficacia que optimizando la suma de sensibilidad y especificidad, esto se debe al desbalanceo de la base de datos respecto a la existencia del evento (Tabla 5); observamos que al optimizar  $F_1$  el porcentaje de verdaderos positivos o la sensibilidad tienen valores bajos, mientras que la especificidad del modelo o los verdaderos negativos obtienen un valor alto, utilizando ese punto de corte predeciremos el no evento mejor. A la hora de optimizar la suma de sensibilidad y especificidad en los eventos encontramos diferencias, mientras que en el evento hemorragia nos encontramos una sensibilidad y especificidad más balanceada; en el evento reinfarto nos encontramos sensibilidad alta frente a una baja especificidad, esto genera que nuestro modelo tienda más a predecir el evento reinfarto. Aunque se ha de destacar que el valor que se pretende optimizar no es la eficacia, dado que por ejemplo un algoritmo que predijera todos los pacientes como no hemorrágicos o no isquémicos obtendría un valor alto de eficacia, pero su utilización sería ineficaz.

Los algoritmos de aprendizaje automático establecen un peso a cada variable predictora en función de la mayor importancia que ejercen en la predicción. Este peso lo obtenemos mediante una función de scikit-learn para interpretar el funcionamiento de los algoritmos. También se quiso observar si había relación entre los eventos y las variables predictoras. Para realizar esta prueba se utilizó un modelo lineal generalizado.

**Tabla 9.** Rendimiento del modelo final según el punto de corte escogido. VP = verdadero positivo. FP = falso positivo. VN = verdadero negativo. FN = falso negativo. RVP = sensibilidad. RVN = especificidad. VPP = precisión. VPN = valor predictivo negativo.

| Modelo                | VP   | FP    | VN    | FN   | RVP   | RVN   | VPP  | VPN   | EFICACIA |
|-----------------------|------|-------|-------|------|-------|-------|------|-------|----------|
| <b>Hemorragia</b>     |      |       |       |      |       |       |      |       |          |
| <b>F1 óptimo</b>      | 0.7% | 10.6% | 87.5% | 1.3% | 35.8% | 89.2% | 6.2% | 98.6% | 88.2%    |
| <b>RVP+RVN óptimo</b> | 1.3% | 31.4% | 66.7% | 0.7% | 65%   | 68%   | 3.9% | 99%   | 68%      |
| <b>Reinfarto</b>      |      |       |       |      |       |       |      |       |          |
| <b>F1 óptimo</b>      | 0.2% | 2.7%  | 95.1% | 2.1% | 8.5%  | 97.3% | 6.7% | 97.9% | 95.3%    |
| <b>RVP+RVN óptimo</b> | 1.7% | 62%   | 35.7% | 0.5% | 76.6% | 36.6% | 2.7% | 98.6% | 17.4%    |

Los modelos lineales generalizados son una generalización de los modelos de regularización. Permiten relacionar la variable respuesta que puede no seguir una distribución Normal con las variables predictoras lineales por medio de una función de enlace<sup>30</sup>. Como hipótesis nula se estableció la variable predictor y variable respuesta son independientes, como hipótesis alternativa que si existía dependencia. Tras realizar la prueba, en los p-valores menores a 0.05 se rechazará la hipótesis nula y se aceptará la alternativa, por lo que podremos decir que existe dependencia entre la variable predictor y la variable respuesta. Esta información la podemos hallar en la tabla 10 donde encontraremos las tres variables con mayor peso de cada modelo en negrita.

**Tabla 10.** Importancia de las variables en el algoritmo RF y relación con el evento. Peso RF = el peso otorgado por el algoritmo Random Forest en los modelos. P-valor = resultado obtenido tras realizar el modelo lineal generalizado de la variable respuesta respecto a la variable predictiva.

| Variables                         | Hemorragia     |           | Reinfarto      |           |
|-----------------------------------|----------------|-----------|----------------|-----------|
|                                   | Peso RF        | p-valor   | Peso RF        | p-valor   |
| <b>Bleeding_inhospital</b>        | <b>0.27995</b> | < 0.001 * | 0.02736        | < 0.001 * |
| <b>Hemoglobin_Admission</b>       | <b>0.17840</b> | < 0.001 * | 0.04621        | < 0.001 * |
| <b>Age</b>                        | <b>0.12768</b> | < 0.001 * | 0.03068        | < 0.001 * |
| <b>Creatinine_Admission</b>       | 0.09456        | < 0.001 * | 0.03136        | < 0.001 * |
| <b>PeripheralArteryDisease</b>    | 0.04935        | < 0.001 * | 0.04364        | < 0.001 * |
| <b>Hypertension</b>               | 0.04394        | < 0.001 * | 0.01785        | 0.103     |
| <b>Prior_Bleeding</b>             | 0.04225        | < 0.001 * | 0.00552        | 0.692     |
| <b>OralAnticoagulation</b>        | 0.03268        | < 0.001 * | 0.01112        | 0.044 *   |
| <b>Killip2orMore</b>              | 0.03042        | < 0.001 * | 0.05657        | < 0.001 * |
| <b>Cancer</b>                     | 0.02751        | < 0.001 * | 0.02519        | < 0.001 * |
| <b>Revascularization_Complete</b> | 0.01504        | < 0.001 * | <b>0.08173</b> | < 0.001 * |
| <b>Sex</b>                        | 0.01121        | < 0.001 * | 0.00265        | 0.007 *   |
| <b>Betablockers</b>               | 0.00839        | < 0.001 * | 0.00297        | 0.155     |
| <b>Femoral_Access</b>             | 0.00801        | 0.001 *   | 0.02522        | < 0.001 * |
| <b>Prasugrel</b>                  | 0.00697        | 0.007 *   | 0.01226        | 0.002 *   |
| <b>Clopidogrel</b>                | 0.00673        | 0.016 *   | 0.03281        | < 0.001 * |
| <b>PriorStroke</b>                | 0.00496        | 0.019 *   | 0.01031        | 0.259     |
| <b>Ticagrelor</b>                 | 0.00489        | 0.517     | 0.02769        | < 0.001 * |
| <b>PriorMyocardialInfarction</b>  | 0.00468        | 0.015 *   | 0.03908        | < 0.001 * |
| <b>ACEI_ARB</b>                   | 0.00458        | 0.089     | 0.00388        | < 0.001 * |
| <b>Multivessel</b>                | 0.00367        | 0.005 *   | <b>0.07131</b> | < 0.001 * |
| <b>ReAMI_inhospital</b>           | 0.00300        | 0.31      | 0.04867        | < 0.001 * |
| <b>STEMI</b>                      | 0.00265        | 0.439     | 0.02748        | 0.013 *   |
| <b>DrugElutingStent</b>           | 0.0024         | 0.011 *   | <b>0.06751</b> | < 0.001 * |
| <b>Dyslipemia</b>                 | 0.00161        | 0.548     | 0.03236        | 0.006 *   |
| <b>LVEFless50</b>                 | 0.00122        | 0.831     | 0.00879        | 0.422     |

|                       |         |         |         |           |
|-----------------------|---------|---------|---------|-----------|
| <b>Statin</b>         | 0.00106 | 0.144   | 0.03783 | < 0.001 * |
| <b>LVEFafter50</b>    | 0.00091 | 0.319   | 0.00207 | 0.102     |
| <b>PriorCABG</b>      | 0.0006  | 0.005 * | 0.03195 | < 0.001 * |
| <b>LVEFless40</b>     | 0.00055 | 0.098   | 0.03644 | 0.001 *   |
| <b>PriorPCI</b>       | 0.00011 | 0.401   | 0.03581 | < 0.001 * |
| <b>Diabetes</b>       | 0       | 0.209   | 0.04318 | < 0.001 * |
| <b>UnstableAngina</b> | 0       | 0.582   | 0.02251 | < 0.001 * |
| <b>NoStent</b>        | 0       | 0.418   | 0       | 0.038 *   |

## 4. Discusión

### ❖ Ventajas de la IA en este tipo de problemas

El uso de IA en este tipo de problemas se está volviendo cada vez más habitual. Proporciona una ayuda al diagnóstico y al tratamiento del paciente. El uso de esta tecnología que permite enseñar a un modelo es de gran utilidad para el diagnóstico clínico. De esta manera se podrá ayudar al facultativo en la toma de decisiones sobre el tiempo del tratamiento DAP en el paciente tras haber sufrido un IAM.

Una de las principales ventajas del uso de la IA es la capacidad para trabajar con mucha información a la vez. Haciendo uso de esta capacidad ha sido posible trabajar con una base de datos de 26.076 pacientes y 49 variables, pudiendo comprobar si existen relaciones entre las variables y las salidas más allá de un análisis tradicional, al utilizar técnicas no lineales.

### ❖ Diferencia en las variables a la hora de predecir los eventos

A la hora de la predicción nuestros algoritmos generan un peso para cada variable, dependiendo de la mayor o menor importancia que tengan para la predicción final. En la tabla 10 se puede observar el peso obtenido por cada variable en el modelo. Además, en esa misma tabla podemos observar si existe relación entre la variable evento y la variable predictora, en el caso de que el p-valor obtenido por el modelo lineal generalizado sea menor a 0.05 existirá relación entre ambas variables.

En el modelo predictor de hemorragia a los 12 meses, observamos que la variable que obtiene una mayor importancia es el sangrado en el hospital, variable que además tiene una fuerte relación con el evento; seguida de la cantidad de hemoglobina que se le administra en sangre y de la edad del paciente, estas dos variables también tienen una fuerte relación con el evento. El hecho de sufrir una hemorragia durante el tiempo que has estado ingresado en el hospital tras la intervención percutánea es una de las variables con un mayor peso predictor<sup>31</sup>. Varios estudios han identificado la edad como un predictor importante de este evento<sup>31-33</sup>. La cantidad de hemoglobina que se le administra al paciente viene dada por su baja presencia en la sangre. La baja presencia de hemoglobina en sangre se puede sufrir por una hemorragia o por anemia. Encontramos artículos que también detectan esta variable con mucho peso a la hora de predecir el evento<sup>32</sup>. También advertimos que variables como ausencia de stent, diabetes o angina inestable obtienen un peso del clasificador igual a 0, es decir, nuestro algoritmo RF no las considera importantes para la clasificación. A su vez podemos ver que la relación independiente de las tres variables con el evento hemorragia no es significativa, dado que su p-valor es mayor a 0.05.

A la hora de predecir el reinfarto a los 12 meses de medicación, advertimos que a diferencia del modelo hemorrágico no hay ninguna variable que se le establezca un peso mucho mayor respecto a las demás. Aun así, las variables que tienen un peso mayor son revascularización completa, enfermedad multivaso y stent farmacológico, las tres variables tienen una alta relación con el evento. La enfermedad multivaso es un factor de riesgo, se produce en el paciente cuando tiene varias cardiopatías en diferentes arterias coronarias, esto acarrea un mayor riesgo respecto a solo tener una arteria afectada; una revascularización completa es un factor protector, este procedimiento se lleva a cabo cuando se arreglan los problemas de todas las arterias coronarias que sufren una afección, es decir, que el paciente presenta trastornos en varias arterias y son arregladas todas ellas. Respecto al stent farmacológico reduce el riesgo de sufrir un reinfarto, este tipo de stent es una evolución moderna de los stent convencionales para administrar medicamentos justo en el lugar de la lesión. Esta información la podemos ver avalada por otros artículos<sup>10,34</sup>. En este modelo solo la variable ausencia de stent no obtiene ningún peso, aunque es significativa con un p-valor de 0.038.

## ❖ Tratamiento de los puntos de corte

Siguiendo el contexto clínico en que se basa este trabajo, en el momento en que se vaya a utilizar nuestro algoritmo, lo mejor será intentar una mayor sensibilidad en la predicción de hemorragia y una mayor especificidad en la predicción de reinfarto. El modelo trabaja razonablemente bien para la predicción de hemorragia, dado que el riesgo de fallecimiento tras hemorragia es importante (alrededor de un 15%, tabla 11), por ello se busca un punto de corte más sensible para evitar hemorragias en aquellas personas con más riesgo. El rendimiento para predecir infarto de los algoritmos es pobre, no podemos plantear un cambio en el tratamiento a no ser que tengamos bastante seguridad de que el paciente va a beneficiarse con el cambio, por ello buscamos una alta precisión o una alta especificidad. De esta manera el médico tendrá más información sobre cómo ajustar el tratamiento DAP al paciente.

El uso de estos modelos sería una prueba de concepto, haría falta un estudio más riguroso de los beneficios y perjuicios de modificar la terapia de doble antiagregación antes de plantear esta aplicación.

**Tabla 11.** Tabla representativa del total de pacientes de los tres registros que han sufrido un evento. Los porcentajes referidos a hemorragia y reinfarto están hechos en base a toda la población. Los porcentajes de fallecimiento por hemorragia o reinfarto se realizaron en base al número de personas que habían fallecido de nuestra población.

| <b>Hemorragia</b> | <b>Reinfarto</b> | <b>Fallecimiento por hemorragia</b> | <b>Fallecimiento por reinfarto</b> |
|-------------------|------------------|-------------------------------------|------------------------------------|
| 633 (2.43%)       | 683 (2.62%)      | 94 (14.85%)                         | 154 (22.5%)                        |

Para la explicación de los puntos de corte, que se han comentado anteriormente, hay que tener en cuenta que el tratamiento DAP aumenta la probabilidad de sufrir una hemorragia a la vez que disminuye la probabilidad de sufrir un reinfarto. Observando además los porcentajes de los pacientes que sufren los eventos en nuestra base de datos, se puede apreciar en la tabla 11 que los eventos tienen una baja prevalencia.

Teniendo esta información lo mejor será dar la medicación estándar a la mayoría de los pacientes. Observando los cuatro posibles casos de pacientes durante la predicción de los riesgos de hemorragia y de isquemia, se daría las siguientes sugerencias:

- Paciente con riesgo de hemorragia y sin riesgo de isquemia.  
Se reducirá el tiempo de la medicación doble antiagregación.
- Paciente con riesgo de isquemia y sin riesgo de hemorragia.  
Se aumentará el tiempo de doble antiagregación.
- Paciente con riesgo de isquemia y hemorragia.  
Se mantendrá la terapia en 12 meses, tal como se suele programar, pero se harán revisiones más frecuentes a este paciente.
- Paciente sin riesgo de isquemia y hemorragia.  
Se mantendrá la terapia durante 12 meses.

## ❖ Sesgos de la base de datos

La población de estudio estaba formada por tres registros diferentes. Los registros BleeMACS y RENAMI son registros creados a partir de estudios clínicos, mientras que, los datos procedentes del hospital universitario de Vigo y del hospital universitario Arrixaca son datos obtenidos de la práctica



habitual. Es decir, los datos de BleeMACS y RENAMI están sometidos a criterios de inclusión sobre las patologías que debe de presentar y la práctica que se realizó para poder aceptar al paciente en el estudio. Mientras que los datos procedentes Vigo y Murcia se deben a la práctica habitual de ambos hospitales en el cual recogieron las mismas variables de los registros anteriores. La principal diferencia encontrada es el tiempo de medicación de los diferentes registros, mientras que BleeMACS y RENAMI se administra DAP durante 12 meses, a no ser que se produzca un evento. En Vigo y Murcia nos encontramos pacientes con un tiempo de medicación más irregular con un pico en 12 meses (ilustración 2 del ANEXO).

A su vez, es probable que BleeMACS y RENAMI, al ser estudios con criterios de inclusión, no sean una muestra completamente realista de los pacientes habituales de un hospital. Por otro lado, los datos están recogidos de manera sistemática. Vigo y Murcia suponen una muestra más realista, pero los datos es probable que no estén recogidos de manera tan sistemática. Los médicos, aunque se adhieren a las guías de práctica clínica, también hacen uso de su criterio subjetivo para determinar el mejor tratamiento para cada paciente de una manera que no está estandarizada (como lo estaría en un estudio) y/o adecuadamente recogida en las variables que disponemos.

También observamos al realizar el análisis univariante del conjunto de entrenamiento y el conjunto de validación una diferencia entre ambos, mientras que en los registros de BleeMACS y RENAMI encontramos diferencia entre que se produzca el evento y que no se produzca en la mayoría de las variables, pero en el registro procedente de Vigo y Murcia la mayoría se acepta nuestra hipótesis de que no existen diferencias entre ellos. Toda esta información se puede observar en la tabla 4 del ANEXO.

#### ❖ **Adecuación de la metodología seguida en el trabajo**

Al trabajar con conjuntos de datos desbalanceados lo aconsejable es trabajar con técnicas como Under Sampling<sup>35</sup> (US). Esta técnica consiste en realizar un balance del conjunto de datos, eliminando algunos de los pacientes que forman parte del evento con un mayor número. A partir de esta técnica se puede dar lugar a clasificadores, como puede ser el Balance Bagging Classifier<sup>35</sup> (BBC). El uso de este clasificador otorga un mayor peso a la posibilidad del evento que tiene una menor probabilidad. En nuestro caso, con una baja prevalencia del evento, otorga más peso a tener la hemorragia o el reinfarto.

Tanto el clasificador BBC como la técnica US acompañada de otros algoritmos de clasificación (RL, SVM o RF) fueron utilizados para realizar pruebas. La utilización de estos algoritmos no supuso una mejora significativa en comparación con solo utilizar los algoritmos descritos en este trabajo. Por ello, se decidió no utilizarlos en la metodología.

Además de utilizar estas técnicas, se realizaron pruebas usando solo los datos procedentes de RENAMI y de Vigo y Murcia, dado que los pacientes de BleeMACS no contenían las variables “peso” y “talla”. Se generaron modelos añadiendo el IMC a las variables predictivas. Pero los resultados obtenidos en estos modelos tuvieron un resultado parecido a los modelos con todos los registros, por ello se prefirió usar una mayor cantidad de datos.

Por estas razones consideramos que la metodología que se ha utilizado durante el presente trabajo ha sido adecuada, dado que se llevaron a cabo distintas pruebas con metodología para datos desbalanceados y el uso de los algoritmos más conocidos de la IA. Finalmente, dado los resultados obtenidos preliminarmente al grueso del estudio, se estableció realizar los tres algoritmos RL, RF y SVM para la creación de los modelos de predicción.

Asimismo, el uso de validación cruzada y validación externa generan un modelo más robusto. Aunque observamos una clara diferencia en los resultados obtenidos en el modelo de predicción de reinfarto a la hora de generalizar el modelo con respecto al resultado de validación. Esto puede venir dado por el posible sesgo entre los registros utilizados. Aunque a priori la mayoría de las variables presentan la misma distribución, dentro de las tres variables con el mayor peso observamos como Vigo

y Murcia tiene un mayor porcentaje de stent farmacológico que los otros dos registros; o que en RENAMI se realizaban un 16% más de revascularizaciones completas respecto a BleeMACS y un 28% más respecto a Vigo y Murcia.

#### ❖ Rendimiento de los algoritmos

El rendimiento obtenido por los algoritmos no ha sido en general muy bueno. Funciona mejor en la predicción de hemorragia ya que la curva ROC tiene un valor de 0.72, y, aunque no podemos predecir de manera concreta quien va a tener hemorragia o no, puede servir para clasificar los pacientes por riesgo y ser capaces de tomar decisiones en los casos más claros. En el caso de reinfarto es un rendimiento probablemente insuficiente para llevarlo a la práctica clínica, 0.59 de AUC, en la cross-validation se obtenían mejores resultados, habría que realizar un análisis más exhaustivo de las diferencias entre los registros para poder usarlo.

#### ❖ Planes de futuro

Este proyecto no finaliza con la entrega de este trabajo. El proyecto fue premiado por la Fundación General de la Universidad de Salamanca con el premio de Transferencia de Conocimiento Universidad Empresa, con el objetivo de realizar un prototipo de este proyecto a nivel empresarial.

El objetivo final del proyecto es generar el riesgo isquémico y hemorrágico de pacientes con el tratamiento DAP a 6, 12 y 24 meses. Por ello el proyecto no acaba aquí, si no que será necesario la creación de “modelos en cascada” indicando el riesgo en estos meses. Además, la finalización del proyecto culminará con la creación de una aplicación en la cual se podrán introducir los datos del paciente, obteniendo entonces los riesgos que se predicen.

Hay que aclarar que esta aplicación no se utilizará con pacientes reales a menos que se obtengan unos resultados mejores y recalcar que es un resultado orientativo para el médico, nunca una sustitución de éste.

## 5. Conclusiones

En este Trabajo de Fin de Grado hemos desarrollado modelos en IA para la predicción del riesgo isquémico y hemorrágico de un paciente que ha sufrido un IAM y ha sido tratado con DAP. Durante su realización se han seguido los siguientes pasos:

- Este trabajo se realizó con una base de datos creada a partir de tres registros diferentes. BleeMACS, RENAMI y datos procedentes de los hospitales de Vigo y Murcia.
- Se realizó un estudio estadístico de todas las variables en función de los diferentes registros, donde se concluyó que la mayoría de las variables seguían la misma distribución.
- Se llevo a cabo un preprocesado de los datos. Crear nuevas variables, pasar variables existentes a binarias o establecer bien el tiempo de medicación.
- Se hizo uso de IA mediante modelos de aprendizaje automático para predecir la hemorragia o el reinfarto de un paciente a los 12 meses de medicación. Estos modelos se compusieron a partir de tres algoritmos diferentes: RL, SVM y RF.
- Se realizó el entrenamiento de los modelos con los datos procedentes de los registros BleeMACS y RENAMI. Se ajustaron los hiperparámetros a cada algoritmo, a su vez se realizó *cross-validation* de 10 repeticiones y 10 *k-folds*. Se obtuvieron los mejores hiperparámetros para cada modelo. El modelo que mejor generalizaba era RF con un AUC de 0.72 para hemorragia y 0.73 para reinfarto.
- Se utilizaron los datos procedentes de Vigo y Murcia como conjunto de validación para observar que tal generalizan los modelos. Tras realizar la validación externa de los modelos se volvieron a obtener los mejores resultados de RF, con un AUC de 0.72 para hemorragia y de 0.59 para reinfarto.
- Para la aplicación de estos modelos se ha propuesto un punto de corte de hemorragia sensible y un punto de corte de reinfarto más específico. Para el modelo hemorrágico con ese punto de corte tiene un 65% de sensibilidad y un 68% de especificidad. El modelo de reinfarto posee una sensibilidad del 8.5% y una especificidad del 97.3%.

Este trabajo ha ahondado en las posibilidades que tiene la IA y hemos querido demostrar como los datos recogidos durante la práctica clínica habitual o para estudios se pueden utilizar para personalizar los tratamientos y mejorar la calidad de vida de los pacientes.

## 6. Bibliografía

1. Jameson JL. *Harrison, principios de medicina interna, 20e*. McGraw-Hill Education; 2019.
2. Consenso ESC 2018 sobre la cuarta definición universal del infarto. *Revista Española de Cardiología*. 2019;72(1):72.e71 - 72.e27.
3. Valgimigli M, Bueno H, Byrne RA, et al. 2017 ESC focused update on dual antiplatelet therapy in coronary artery disease developed in collaboration with EACTS. *European Journal of Cardio-Thoracic Surgery*. 2017;53(1):34-78.
4. Mehran R, Rao SV, Bhatt DL, et al. Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium. *Circulation*. 2011;123(23):2736-2747.
5. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. 3 ed: Prentice Hall; 2010.
6. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, Inc.; 2019.
7. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Network Open*. 2020;3(3):e200265-e200265.
8. Uddin M, Wang Y, Woodbury-Smith M. Artificial intelligence for precision medicine in neurodevelopmental disorders. *npj Digital Medicine*. 2019;2(1):112.
9. Andrés E, Cordero A, Magán P, et al. Mortalidad a largo plazo y reingreso hospitalario tras infarto agudo de miocardio: un estudio de seguimiento de ocho años. *Revista Española de Cardiología*. 2012;65(5):414-420.
10. Sampedro-Gómez J, Dorado-Díaz PI, Víctor Vicente-Palacios E, et al. Machine Learning To Predict Stent Restenosis Based On Daily Demographic, Clinical And Angiographic Characteristics. *Canadian Journal of Cardiology*. 2020.
11. Average daily ischemic versus bleeding risk in patients with ACS undergoing PCI: Insights from the BleeMACS and RENAMI registries. *American Heart Journal*. 2020;220:108 - 115.
12. Garay A, Ariza-Solé A, Formiga F, et al. Prediction of Post-Discharge Bleeding in Elderly Patients with Acute Coronary Syndromes: Insights from the BleeMACS Registry. *Thromb Haemost*. 2018;118(05):929-938.
13. Raposeiras-Roubín S, Abu-Assi E, D'Ascenzo F, et al. Annual Incidence of Confirmed Stent Thrombosis and Clinical Predictors in Patients With ACS Treated With Ticagrelor or Prasugrel. *Rev Esp Cardiol (Engl Ed)*. 2019;72(4):298-304.
14. D'Ascenzo F, Abu-Assi E, Raposeiras-Roubín S, et al. BleeMACS: rationale and design of the study. *Journal of Cardiovascular Medicine*. 2016;17(10):744-749.
15. Nathan S, Rao SV. Radial Versus Femoral Access for Percutaneous Coronary Intervention: Implications for Vascular Complications and Bleeding. *Current Cardiology Reports*. 2012;14(4):502-509.
16. Ting KM. Confusion Matrix. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer US; 2017:260-260.
17. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006;27(8):861 - 874.
18. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2009.
19. Shalev-Shwartz S, Ben-David S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press; 2014.
20. Géron A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media; 2019.
21. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: with Applications in R*. Springer New York; 2014.
22. Bouckaert RR, Frank E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. 2004; Berlin, Heidelberg.

23. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.
24. Pilgrim M, Willison S. *Dive Into Python 3*. Vol 2: Springer; 2009.
25. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011;12:2825-2830.
26. Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering.* 2007;9(3):90-95.
27. Bengfort B, Bilbro R. Yellowbrick: Visualizing the Scikit-Learn Model Selection Process. *The Journal of Open Source Software.* 2019;4.
28. Código TFG. <https://github.com/PabloPerSa/TFG>. Accessed.
29. KoopaML. <https://github.com/IA-Cardiologia-husa/KoopaML>. Accessed.
30. Dunn PK, Smyth GK. *Generalized linear models with examples in R*. Springer; 2018.
31. Moscucci M, Fox KAA, Cannon CP, et al. Predictors of major bleeding in acute coronary syndromes: the Global Registry of Acute Coronary Events (GRACE). *European Heart Journal.* 2003;24(20):1815-1823.
32. Généreux P, Giustino G, Witzenbichler B, et al. Incidence, Predictors, and Impact of Post-Discharge Bleeding After Percutaneous Coronary Intervention. *Journal of the American College of Cardiology.* 2015;66(9):1036-1045.
33. Numasawa Y, Kohsaka S, Ueda I, et al. Incidence and predictors of bleeding complications after percutaneous coronary intervention. *Journal of Cardiology.* 2017;69(1):272-279.
34. Stone GW, Ellis SG, Cox DA, et al. A Polymer-Based, Paclitaxel-Eluting Stent in Patients with Coronary Artery Disease. *New England Journal of Medicine.* 2004;350(3):221-231.
35. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. *Learning from Imbalanced Data Sets*. Springer International Publishing; 2018.

## 7. Summary

### Introduction

The goal of this project is finding the ischemic and bleeding risk of patients who have been treated with Dual Antiplatelet Therapy after suffering an acute coronary syndrome with Artificial Intelligence. Acute coronary syndrome is a group of conditions that reduce blood flow to the heart, these conditions are Acute Myocardial Infarction and Unstable Angina.

Acute Myocardial Infarction is a heart disorder which occurs as a result of a coronary artery being blocked by a thrombus, decreasing the amount of oxygen reaching the cells. The main treatment after an acute myocardial infarction is a percutaneous intervention to remove the obstruction from the artery.

Unstable Angina is sign of a high risk for myocardial infarction or sudden death. It is characterized by discomfort in the chest or arms. It arises during rest, is intense or in crescendo (more intense, lasting or frequent). It occurs when the heart does not get enough blood flow and oxygen.

One of the treatments followed after the percutaneous intervention is Dual Antiplatelet Therapy, which consists of two antiplatelet drugs. One is aspirin and the other is one of these three drugs: clopidogrel, ticagrelor or prasugrel. This therapy reduces the risk of having a heart attack again but increases the risk of bleeding.

Artificial Intelligence belongs to the branch of computer science, which seeks to understand and build intelligent entities. Machine Learning is the branch of Artificial Intelligence whose objective is developing an agent able to learn how to perform tasks without the need of being explicitly trained to do so by using a set of techniques or algorithms that gives a machine the ability to achieve this task.

Within Machine Learning one of the techniques is supervised learning, which tries to obtain a function from training data composed of input vectors (inputs) with their respective target vectors (outputs).

Which automatic learning, which we find a series of algorithms known as neural networks, this part of the Machine Learning is called Deep Learning. These neural networks try to imitate the human nervous system. They are composed of layers of neurons in which each one learns individually. Neural networks with a large number of layers are frequently found.

Use of artificial intelligence is a great advance in medicine because it helps doctors in the diagnosis and its data processing capacity may be useful when creating new drugs or in research. Fields such as Machine Learning or Deep Learning offer us these advantages.

In 2017, 30% of the deaths in Spain were caused by cardiovascular diseases; and in 2015, acute myocardial infarction was the leading cause of death. The percentage of relapse for heart attack is less than 10% but medication and lifestyle play a very important role in this case. Therefore, this work revolves around one of the treatments followed by patients after an acute myocardial infarction, the double antiplatelet aggregation, predicting the risk of the patients suffering another ischemic or a bleeding episode.

### Material and methods

The population of study comes from three different records: BleeMACS, RENAMI and a separate register from the hospitals of Vigo and Murcia. All the patients in our data set have suffered from acute

coronary syndrome, i.e. they have suffered from acute myocardial infarction or unstable angina; and after their discharge from hospital they have followed the Dual Antiplatelet Therapy.

The BleeMACS registration is at international level and the RENAMI registration is at European level, while the data of the patients from Vigo and Murcia are recollected from the clinical practice of the University Hospitals of Vigo and Murcia.

In our dataset we found variables referring to the clinical data of the patient at the time of suffering the acute myocardial infarction, in addition to this, we also found variables which store how long the patient was observed and whether he or she suffered from a heart attack, bleeding, transfusion or died during the treatment. These time variables were used to generate the response variables of our study.

The dataset was composed of a predominantly male population with an average age of 64 years. 26.7% of patients have diabetes and 6.4% have cancer. 17.7% of the population has had an AMI before. We can also observe that 59.9% of patients suffered from hypertension and 52.8% of patients suffered from STEMI.

Machine Learning algorithms were used to find the ischemic and bleeding risk in patients with an acute coronary syndrome and Dual Antiplatelet Therapy. Three different algorithms and a set of techniques were used to create our predictive models to obtain a better model.

The metrics used in the evaluation of the models were the confusion matrix and the ROC curve. The confusion matrix in a 2x2 matrix and is indexed by the true class of an object and by the class that the classifier assigns. With it, different metrics can be generated to measure the performance of our classifier, such as accuracy, sensitivity, or specificity among others. The ROC curve is a graph which shows the diagnosis of a binary classifier as the discrimination threshold changes. This graph plots specificity against sensitivity adjusted at different cut-off points. The area under the curve is below the curve and can obtain a maximum value of 1 and a minimum value of 0.

The machine learning algorithms used were Logistic Regression, Support Vector Machine and Random Forest.

- Logistic Regression is a statistical model which uses the logistic function to model the probabilities of belonging to a certain class through linear functions.
- The Support Vector Machine is a very powerful and versatile self-learning model, it is one of the most popular models of Machine Learning. Support Vector Machines are suitable for the classification of complex data sets.
- Random Forest is a combined classifier generated from decision trees. Random Forest will fit multiple different decision trees. When predicting a new individual, it will let each tree vote on which classification it considers correct, setting it to a region of the parameter space. After this vote, the individual will be assigned a probability of belonging to a class.

The Logistic regression and the Support Vector Machine are linear classifiers. Both algorithms have a cost function, which we will try to decrease in order to obtain a better prediction. The cost function is generated from the loss function of the algorithm plus a regularization. The regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error, but not its training error.

Random Forest is a combination of sorters since it uses multiple sorting algorithms. Its main idea is to contemplate multiple hypotheses simultaneously forming a better behaving hypothesis.

The techniques used in the evaluation of the model to generate more robust algorithms and avoid overfitting were cross-validation and external validation. Overfitting occurs when the model is too complex in relation to the amount of noise in the training data. When generating prediction models, we may be suffering from biases, which may be generated in our model or may be inherited from the data used for its creation. The biases produced when generating the data cannot be avoided, but it is possible to do so with the ones produced during the training of the model can be. An example of these biases would be overfitting.

Cross-validation is a method to estimate the prediction error in a limited data set. It involves dividing the data into  $k$  equal subsets. Then we perform  $k$  learning rounds, in each round  $1/k$  of the data is represented as a test set and the remaining data as training data. The final value of the metric we are measuring is averaged by the  $k$  values that have been obtained.

The external validation uses a different set of data than the one used to train our model as a test. This way it can be observed whether the model we created generalizes well for other populations.

These two methods are useful to obtain a better and more robust predictive model.

The development of the project could be divided in two parts: pre-processing of the data and model fitting and validation. We used the Python programming language with the scikit-learn, matplotlib, seaborn and yellowbrick libraries.

- Pre-processing of the data:

In this part, changes were made to the type of variables, turning as many as possible into the binary form. Errors found in the times of the events were also corrected. The response variables of our models were generated from the variables of whether or not each event occurred and the time of the event.

- Model fitting and validation:

In model fitting and validation, a distinction was made between train and test data sets. In addition, two different models were made, one for reinfarction prediction and the other for bleeding prediction. The BleeMACS and RENAMI records were used as the train set. As a test set, data from patients in Vigo and Murcia were used to perform an external validation.

While training the model with the algorithm, different hyperparameters were used to choose the combination that fits the best our data. During the training of the model a 10 k-folds cross validation with 10 repetitions was performed.

After this procedure, the best combination of hyperparameters was chosen and the external validation of our data was carried out.

## Results

The result obtained from the training in the prediction of bleeding is 0.69 of area under the curve and from the prediction of reinfarction it is 0.70. In both models the algorithm with the best result was the Random forest.



After the training of the models, they were validated with our external validation set. The results obtained were 0.72 area under the curve in bleeding and 0.59 in reinfarction.

In addition, we can observe the variables that obtained a greater weight in the predictive models, the relationship of each variable with the event and the result to a Fisher test in which it was sought to find differences in the groups formed by the appearance of reinfarction or bleeding with the predictive variables.

## **Discussion and conclusions**

The use of artificial intelligence in this type of problem is advantageous because of the ability of being able to work with a lot of information at once. It has allowed us to work with a database with around 26,000 patients and 49 variables, checking whether there are relationships between the variables and the outputs beyond a traditional analysis or not, using non-linear techniques.

At the time of the prediction our algorithms generate a weight for each variable, depending on the greater or lesser importance they have for the final prediction. In the prediction of bleeding the variables that obtain a greater weight are “bleeding in hospital”, “admission haemoglobin” and “age”. In the prediction of reinfarction are “complete revascularization”, “multivessel” and “pharmacological stenting”.

Another advantage of using artificial intelligence algorithms is being able to vary the cut-off points of our algorithms to obtain a prediction adjusted to our clinical context. In our case, a cut-off point with greater sensitivity will be used for the prediction of bleeding and a cut-off point with greater specificity for the prediction of reinfarction.

For the bleeding model with that cut-off point it has a 65% sensitivity and a 68% specificity. The reinfarction model has a sensitivity of 8.5% and a specificity of 97.3%.

The differences between the results, training and validation, may have its explanation due to the difference in the origin of the data. The BleeMACS and RENAMI records are records created from clinical studies, while the data from the University Hospital of Vigo and the University Hospital of Murcia are data obtained from normal practice.

Different methodologies were tested at the beginning of the study, but the results were similar between them. Therefore, the methodology used was appropriate for the data set, performing techniques to obtain more robust models.

The final performance of the models was not very encouraging. It works better in predicting bleeding since the ROC curve has a value of 0.72, and, although we cannot predict concretely who will or will not have bleeding, it can be useful for classifying patients by risk and be able to make decisions in the clearest cases. In the case of reinfarction we got a regular performance, 0.59 of ROC curve, in the cross-validation better results were obtained, but a more comprehensive analysis of the differences between the records should be made in order to be able to use it.

Finally, this project has delved into the possibilities of the artificial intelligence and we wanted to demonstrate how data collected during regular clinical practice can be used to personalize treatments and improve the quality of life of patients.