Métodos de aprendizaje automático como ayuda a la toma de decisiones médicas

Alumno: Alejandro Simón Rodríguez

Tutora: Ana Carpio Rodríguez

Máster de Ingeniería Matemática 2022-2023

Facultad de Matemáticas UCM

UNIVERSIDAD
COMPLUTENSE
MADRID

Índice

Resumen: En el campo de la medicina se deben tomar decisiones complicadas basadas en las experiencias de profesionales, pero sin ser capaces de cuantificar los factores que influyen y las distintas opciones existentes. El aprendizaje automático permite desarrollar herramientas que utilizando datos de otros pacientes ayude a la toma de decisiones y cuantifique las distintas posibilidades. En este trabajo se desarrolla una herramienta que ayuda a determinar el éxito de la indometacina como prevención de la pancreatitis, una complicación habitual tras someterse a la CPRE. Además, se demuestra la utilización de métodos bayesianos para la elección de hiperparámetros y los valores Shapley como herramienta para interpretar los modelos de predicción.

Palabras clave: Aprendizaje Automático, Métodos Bayesianos, Selección de Hiperparámetros, Valores Shapley, Pancreatitis, Catboost, Random Forest, XGBoost, Indometacina

1. Introducción

Las herramientas de aprendizaje automático [1] pueden ser una herramienta de apoyo a la hora de tomar decisiones sobre procedimientos o tratamientos médicos. Es conocido que muchos medicamentos o intervenciones médicas pueden no tener efecto o incluso suponer un riesgo para el paciente.

Identificar los factores que tienen aquellos pacientes donde la intervención tiene una mayor probabilidad de éxito, así como aquellos factores que tienen los pacientes donde la probabilidad de fracaso es alta es un campo donde los modelos de aprendizaje automático pueden suponer un salto diferencial. Esto modelos no solo mejorarían la tasa de éxito de las diferentes técnicas médicas si no también reducirían el impacto en la salud del paciente, ya sea con secuelas temporales o crónicas.

La colangiopancreatografía retrógrada endoscópica (CPRE) es un procedimiento médico que permite diagnosticar y tratar condiciones del hígado, de la vesícula biliar, de los conductos biliares y del páncreas. Sin embargo, está técnica tiene una alta probabilidad de causar efectos secundarios (15-25%) llegando a derivar en pancreatitis, que puede llegar a ser mortal.

Hasta el descubrimiento de la indometacina ([2]), se había intentado evitar mediante más de 35 agentes farmacológicos esta complicación mediante prevención química. No obstante, la falta de pruebas clínicas de calidad y el hecho de no haber un conjunto de medidas generalizadas para evitar esta complicación dificultaba el descubrimiento de un medicamento que redujese las posibilidades de complicaciones.

Como todo fármaco tiene importantes efectos secundarios para el paciente como pueden ser malestar estomacal, úlceras, sangrado en el tracto gastrointestinal, mareos, dolor de cabeza y retención de líquidos, además de aumentar el riesgo de problemas cardiovasculares y renales.

El objetivo es de este proyecto es entrenar un modelo sobre el conjunto de pacientes al que se le administró la indometacina, de forma que el modelo aprenda a distinguir cuales son las características más importantes para determinar si surte efecto el medicamento.

Una vez identificadas esas características, se construye una herramienta personalizada que aconseja o rechaza el uso de la indometacina en función de cada paciente. De esta forma lograremos no administrar el medicamento en aquellos pacientes con baja probabilidad de efecto, reduciendo los efectos secundarios y la incomodidad del paciente al ser un medicamento administrado por vía rectal. En caso de aconsejar su uso, se está logrando reducir el riesgo de complicaciones.

2. Descripción de la investigación desarrollada

2.1. Descripción de la metodología del estudio

Se trata de un estudio [2] multicentro, aleatorio, controlado mediante placebo llevado a cabo en cuatro hospitales universitarios distintos: Michigan, Indiana, Kentucky y Ohio. El principal objetivo de la investigación es analizar el efecto de la indometacina rectal sobre pacientes con un elevado riesgo de desarrollar pancreatitis tras la técnica CPRE.

En la investigación participaron 602 pacientes, los cuales se dividieron en dos grupos: los pacientes que recibirían la indometacina (295) y el grupo placebo (307). El 9.2% de los pacientes que recibieron la indometacina desarrollaron pancreatitis aguda mientras que en el grupo placebo sucede en el 16.9% de los casos.

Un aspecto vital para la investigación es definir cuando un paciente sufre pancreatitis. Para ello debe cumplir tres condiciones:

- Dolor abdominal superior
- Aumento de las enzimas pancreáticas de al menos 3 veces el límite normal en un rango de 24 horas
- Estar hospitalizado tras someterse a la CPRE por un mínimo de dos noches.

Con respecto a la metodología del análisis, tras someterse el paciente a la CPRE, si el paciente cumple los criterios de inclusión, se administraba aleatoriamente dos supositorios de indometacina de 50 mg o dos supositorios idénticos como placebo. Estos supositorios son administrados en la misma habitación donde se desarrolla la CPRE y se elige la vía rectal porque las investigaciones hasta la fecha confirman que los medicamentos esteroidales son más eficaces por esa vía.

Finalmente, la indometacina utilizada pertenece a dos farmacéuticas distintas, aunque ambos medicamentos tienen las mismas propiedades comprobadas tras un análisis farmacológico y tras el estudio realizado.

2.2. ¿Qué es la CPRE?

El diagnóstico y el tratamiento de afecciones del páncreas, la vesícula biliar, los conductos biliares y el hígado se realiza mediante una colangiopancreatografía retrógrada endoscópica (CPRE) [3].

Un endoscopio especializado llamado duodenoscopio es introducido durante una CPRE a través de la boca, el esófago y el estómago hasta el duodeno (la primera parte del intestino delgado). Este dispositivo, que incluye una luz junto a una cámara en el extremo, permite al médico evaluar el estado interno del organismo.

Un pequeño catéter se desliza a través del endoscopio hasta llegar al duodeno. Una vez allí, se inyecta un tinte de contraste en estos conductos, lo que permite identificar su posición y observar cualquier anomalía u obstrucción.

Los gastroenterólogos, que son especialistas en el diagnóstico y tratamiento de afecciones gastrointestinales, suelen ser los encargados de realizar esta técnica. Al ser tan invasiva, este procedimiento generalmente se lleva a cabo después de que el paciente se haya sedado.

Sin embargo, debido a que este procedimiento conlleva riesgos como pancreatitis, sangrado, infección o perforación del tracto digestivo, la decisión de someterse a una CPRE debe basarse en las condiciones del paciente. Para ello se realiza un estudio donde se investigan las características y el perfil de riesgo del paciente, de forma que ayude a tomar una decisión al equipo médico.

2.3. ¿En qué consiste la pancreatitis?

La pancreatitis [4] es la inflamación del páncreas por definición. Esta inflamación puede tener distintos grados de seriedad, yendo de ser una inflamación leve para incluso llegar a ser mortal. Desde un punto de vista biológico, la inflamación es producida debido a que las enzimas pancreáticas se activan antes de tiempo y comienzan a digerir el tejido pancreático en lugar de hacerlo en el intestino.

Los motivos que pueden causar esta inflamación son muy diversos, entre los más habituales están: cálculos biliares que bloquean el conducto pancreático, el consumo excesivo de alcohol, el trauma abdominal, los altos niveles de triglicéridos y, por último, la presencia de anomalías congénitas en la estructura del páncreas.

Uno de los síntomas característicos de la pancreatitis aguda es el dolor abdominal intenso, generalmente en la parte superior, que puede extenderse hacia la espalda. Acompañando al dolor abdominal, la presencia de náuseas, vómitos, fiebre e hipersensibilidad al tacto en el abdomen son otros de los síntomas comunes. La pancreatitis aguda puede llegar a causar complicaciones como infecciones, abscesos, daño a órganos cercanos o insuficiencia orgánica en los casos más graves.

Para evaluar los síntomas clínicos, se realizan análisis de sangre para medir los niveles de enzimas pancreáticas y otros indicadores inflamatorios, así como la realización de pruebas de imagen como ecografías abdominales, tomografías computarizadas (TC) o resonancias magnéticas (RM) son los métodos más utilizados para realizar el diagnóstico de pancreatitis aguda.

El tratamiento de la pancreatitis aguda se enfoca en controlar el dolor, administrar líquidos por vía intravenosa con el objetivo de mantener al paciente hidratado y permitir que el páncreas descanse mediante un ayuno, además de monitorizar posibles complicaciones. En situaciones más graves, puede ser necesario ingresar al hospital y tomar medicamentos para controlar los síntomas. Si el estado del paciente es crítico se realiza una intervención quirúrgica como último recurso.

2.4. Indometacina y sus efectos secundarios

En este apartado vamos a explicar superficialmente que es la indometacina y cuáles son sus efectos secundarios para motivar el uso del medicamento únicamente en aquellos casos donde sea necesario.

Se trata de un medicamento que pertenece a la clase de los antiinflamatorios no esteroidales, conocidos también como AINEs. Su principal función es aliviar el dolor, la inflamación y la fiebre [5]. Para ello evita la formación de prostaglandinas en el organismo, enzimas que se producen de forma natural en respuesta a una lesión u otras enfermedades, provocando inflamación y dolor. Existen numerosas formas de aplicarla, desde administración tópica, oral o rectal.

Sin embargo, como todo medicamento posee efectos secundarios [6], por ejemplo, puede aumentar el riesgo de ataques al corazón o accidentes cerebrovasculares potencialmente mortales. También puede causar sangrado en el estómago o intestino, lo cuales pueden llegar a ser fatídicos para el paciente.

También es posible que origine signos de reacción alérgica desde ronchas, problemas para respirar o incluso reacciones en la piel como sarpullidos o quemazones en los ojos.

2.5. Descripción de los datos

El conjunto de datos [2] está compuesto por 602 pacientes y 33 variables. Para la realización del estudio se dividen los pacientes en dos grupos, aquellos pacientes que se les administró la indometacina y aquellos que solo recibieron un placebo.

En la siguiente tabla observamos información de algunas de estas variables y como se distribuyen en ambos grupos de pacientes:

Característica	Indometacina (N = 295)	Placebo (N = 307)
Edad -- años	44.4±13.5	46.0±13.1
Sexo femenino — no. (%)	229 (77.6)	247 (80.5)
Sospecha clínica de disfunción del esfínter de Oddi		
Ninguna	248 (84.1)	247 (80.5)
Tipo 1	38 (12.9)	43 (14.0)
Tipo 2	139 (47.1)	135 (44.0)
Tipo 3	71 (24.1)	69 (22.5)
Documentada en manometría	155 (52.5)	160 (52.1)
Historial de post-PCRE pancreatitis — no. (%)	47 (15.9)	49 (16.0)
Historial de pancreatitis recurrente — no. (%)	85 (28.8)	94 (30.6)
Canulación complicada (>8 intentos) — no. (%)	79 (26.8)	77 (25.1)
Esfinterotomía precortada — no. (%):j:	15 (5.1)	17 (5.5)
Pancreatografía		
Pacientes — no.{%)	249 (84.4)	260 (84.7)
Número mediano de inyecciones de agentes de	2	2
Esfinteroctomía pancreática terapeutica — no. (%)	172 (58.3)	170 (55.4)
Acinarización pancreática — no. (%)	15 (5.1)	12 (3.9)
Esfinteroctomía biliar terapeutica — no.{%)	172 (58.3)	171 (55.7)
Ampulectomía — no. (%)	9 (3.1)	9 (2.9)
Colocación de stent pancreático — no. (%)	246 (83.4)	250 (81.4)
Participación del entrenamiento en la CPRE — no. (%)	142 (48.1)	140 (45.6)

Observamos como se construyeron dos grupos similares en términos de edad, participantes y género. Otros factores que se repartieron en igual porcentaje son la disfunción del esfínter de Oddi, la recurrencia de la pancreatitis, la participación en el entrenamiento de la CPRE o la pancreatografía [7] entre otros.

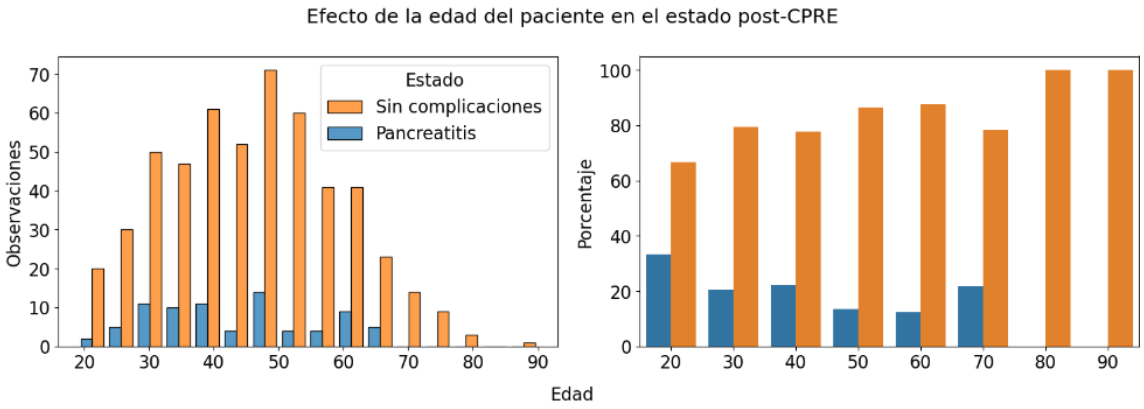
Otras variables de interés del estudio son:

- El riesgo asociado a cada paciente, siendo una puntuación del 1-5.5 otorgada por el equipo médico según las características de cada paciente.
- Si se realizó un cepillado del conducto pancreático para tomar muestras.
- Si hubo sangrado post-PCRE al perforar algún órgano gastrointestinal.
- Si el paciente ha tomado dosis de aspirina lo cual puede incrementar el riesgo de sangrado.

3. Análisis estadístico de los datos

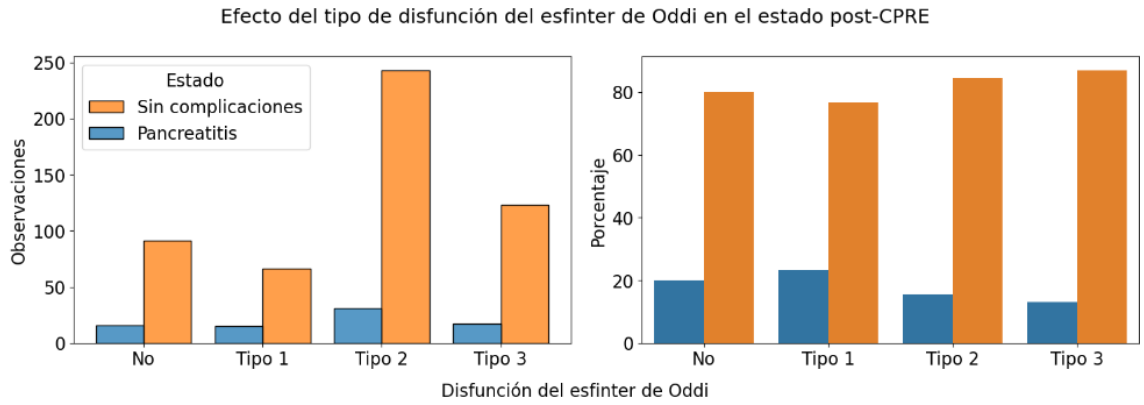
De las 33 variables del conjunto de datos, 30 son variables categóricas mientras que solo hay 3 variables numéricas: edad, riesgo e identificador.

Para entender mejor el conjunto de datos se ha realizado un análisis de ciertas variables sobre el grupo **placebo**, de forma que entendamos la relación con la posibilidad de desarrollar pancreatitis tras someterse a la CPRE. Por ejemplo:

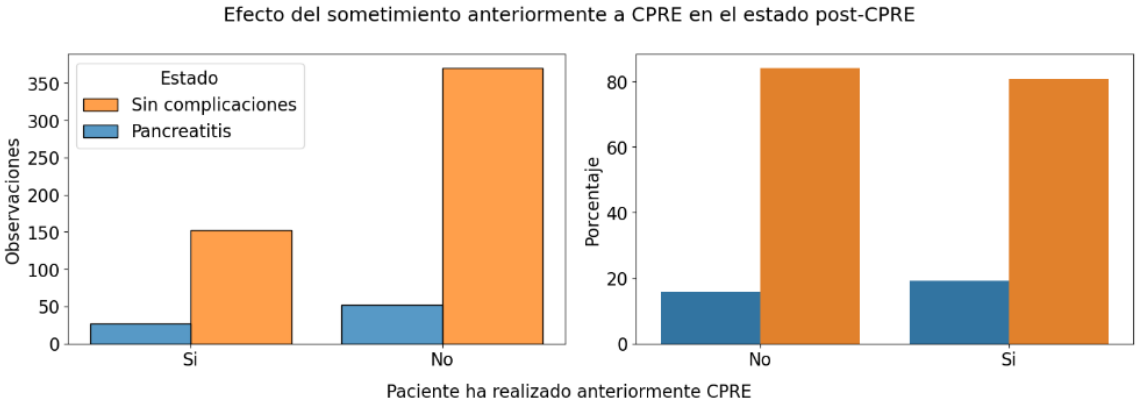


A la izquierda observamos la distribución de la edad de la muestra de pacientes y a la derecha el porcentaje de pacientes por grupo que desarrolla complicaciones post-CPRE. Aunque apenas tenemos observaciones a partir de los 75 años, tener una avanzada edad no parece ser un factor determinante en el desarrollo de pancreatitis post-CPRE.

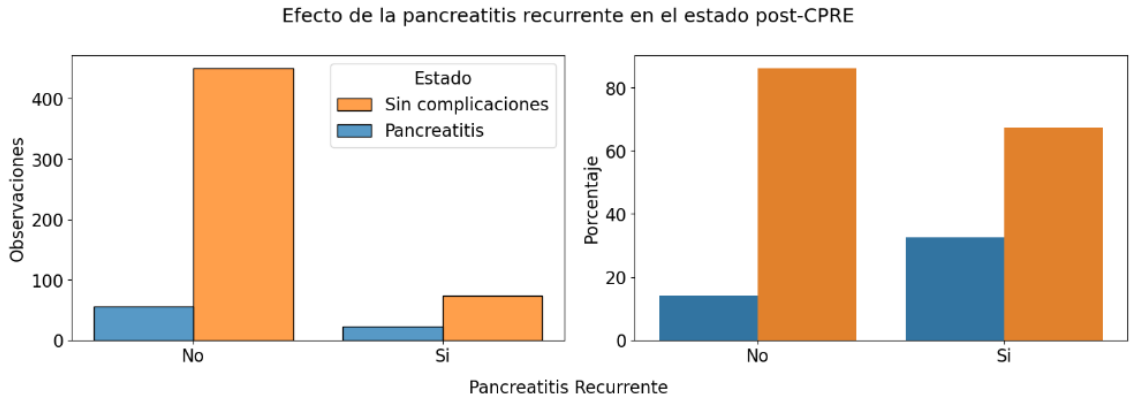
Otro factor que no parece ser determinante es la posibilidad de tener una disfunción en el esfínter de Oddi:



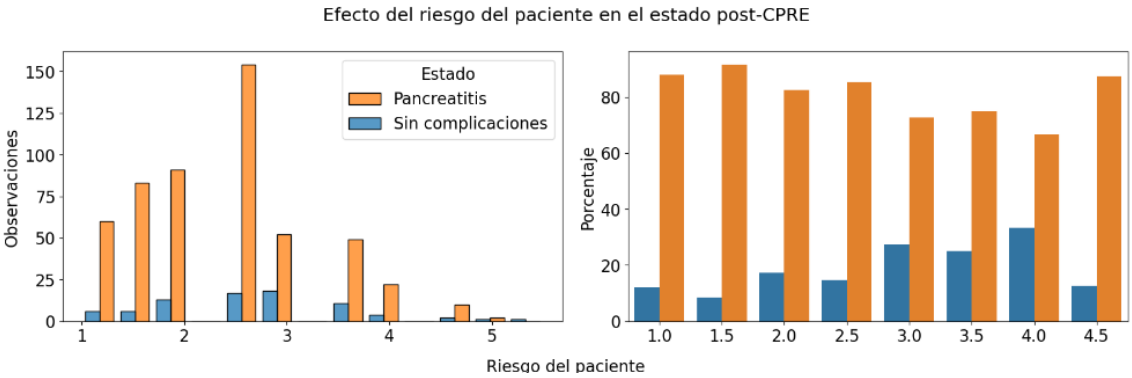
Si nos fijamos en el porcentaje de pancreatitis de aquellos que tienen disfunción (16%) versus aquellos que no tienen disfunción (20%), no parece haber un efecto determinante de la disfunción en el resultado de la CPRE. Sin embargo, hay una tendencia descendente cuanto mayor sea el tipo de disfunción, pudiendo ser un indicador de la infección.



El hecho de haberse sometido con anterioridad a la CPRE no es un factor determinante a la hora de desarrollar pancreatitis tras someterse de nuevo a la CPRE. Sin embargo, haber desarrollado pancreatitis anteriormente si parece ser un factor que aumenta la posibilidad de desarrollar pancreatitis otra vez tras esta prueba:



Hay más del doble de probabilidades (14% vs 33%) de desarrollar pancreatitis tras la prueba. Además de haber tenido pancreatitis con anterioridad, el perfil de riesgo evaluado por el equipo médico parece ser una medida acertada:



Observamos que, a mayor riesgo, mayor es la probabilidad de desarrollar la infección. La falta de datos en pacientes de mayor riesgo puede ser el motivo de ese punto de inflexión en la tendencia.

3.1. Valores perdidos y valores atípicos.

Si realizamos una búsqueda de valores perdidos encontramos una única observación de la cual no se tienen datos para las columnas asa, asa81 y asa325, las cuales indica si el paciente ha tomado aspirina diariamente y así como la dosis. Al estar una única observación afectada se ha supuesto que dicho paciente no tomó aspirina.

No se han observado más valores perdidos o valores extraños en el conjunto de datos. Las variables que son constantes o casi constantes se han eliminado, pues solo aportan ruido. En este caso se han identificado tres variables casi constantes: *id*, *brush* y *pbmal*.

Con respecto a valores atípicos, en variables categóricas se suelen definir como una observación que tiene valores en categorías con baja frecuencia y por tanto se determina como atípica. Tras eliminar las variables casi-constantes no se aprecian observaciones que podamos identificar como valores atípicos.

4. Modelos propuestos

Una de las grandes complicaciones de este problema es el hecho de que casi todas las variables son categóricas, reduciendo enormemente los posibles modelos a utilizar. Si bien es cierto que se pueden utilizar técnicas como one hot encoding para transformar a numéricas estas variables, no todos los modelos se comportan bien ante esta situación.

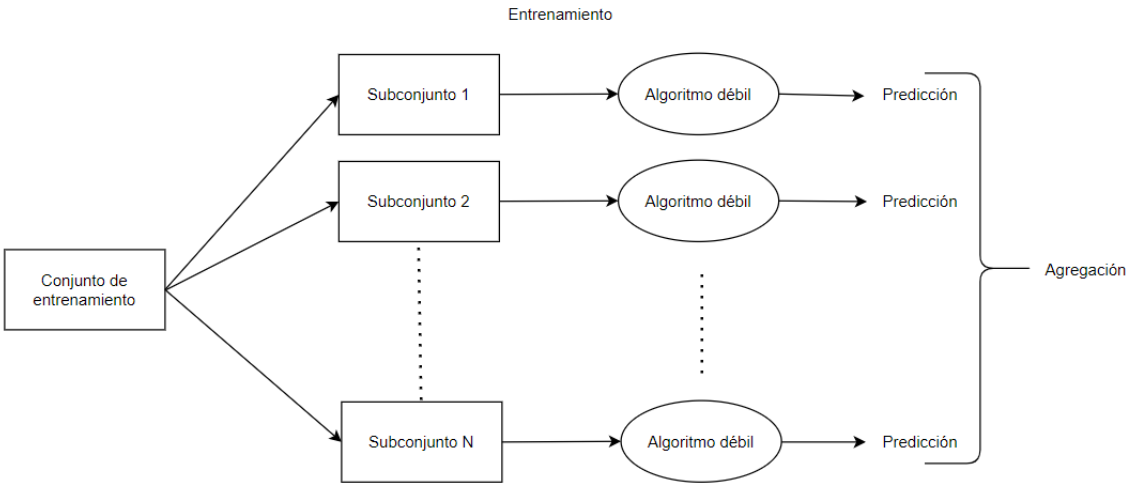
Además, el hecho de transformar las variables categóricas en numéricas tiene la consecuencia de complicar la interpretación del modelo. Se han ajustado 3 modelos: Random Forest, XGBoost y Catboost basados en árboles de decisión con diferentes técnicas de ensamblado.

La principal ventaja de los árboles de decisión es su interpretabilidad, por lo que permite no solo construir modelos capaces de predecir, si no también entender cuales son las variables que más influyen a la hora de predecir una clase u otra.

4.1. Random Forest

Es un modelo ensamblado que utiliza la técnica de bagging [8] tomando como modelos simples los árboles de decisión. El método de bagging consiste en combinar en paralelo cientos de algoritmos simples para obtener una predicción más estable (en términos de varianza del modelo) al ser la predicción del modelo ensamblado una agregación de las predicciones de los modelos simples. El método de agregación depende del objetivo del modelo, si se trata regresión se suele utilizar una media simple, mientras que si se trata de un problema de clasificación se realiza una votación, donde el peso depende de la proporción de elementos de esa clase en el conjunto, permitiendo lidiar con problemas desbalanceados.

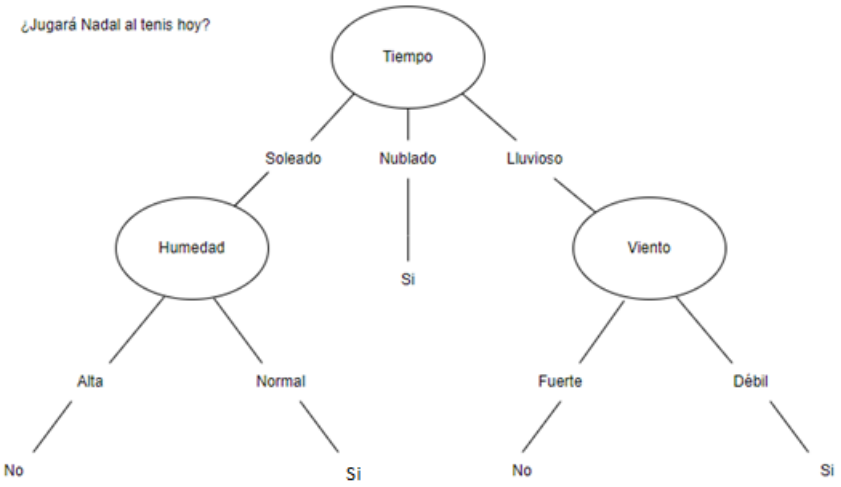
El siguiente ejemplo ilustra el método de bagging:



Para evitar obtener la misma predicción en cada algoritmo simple se toma una muestra aleatoria de observaciones y características para cada modelo, de forma que cada uno de los modelos simples es entrenado sobre un subconjunto distinto, teniendo como resultado un modelo complejo con una mayor capacidad de generalización. Posteriormente cada modelo simple realiza una predicción y la predicción final es la agregación de las predicciones simples. En el caso de Random Forest [9] se utiliza la media de las predicciones de los árboles de decisión.

Un árbol de decisión [10] es uno de los modelos más interpretables para problemas de clasificación o regresión en el ámbito del aprendizaje automático, por eso es altamente utilizado en el ámbito profesional. Se trata de una estructura jerárquica compuesta por el nodo raíz, las ramas, los nodos intermedios y los nodos hoja.

Los nodos son cada una de las variables que forman parte del modelo, donde el nodo raíz viene dado por la variable más destacada a la hora de realizar la predicción, mientras que las ramas son las distintas evaluaciones posibles de la variable.



En pocas palabras, dado un conjunto de variables y observaciones clasificadas en la clase objetivo, elegimos un criterio que nos permita seleccionar que variable utilizar para dividir el conjunto original de observaciones. En el ejemplo superior, el objetivo es responder a la pregunta sobre si Nadal juega al tenis hoy. Dado un histórico de días donde se conoce las condiciones temporales y si Nadal jugó o no, elegimos las características más relevantes según un criterio en cada nivel del árbol, y desarrollamos hasta que somos capaces de responder a la pregunta original.

Por ejemplo, si el tiempo está nublado, Nadal siempre va a jugar. Sin embargo, si es lluvioso y ventoso entonces el árbol de decisión apunta a que no jugará.

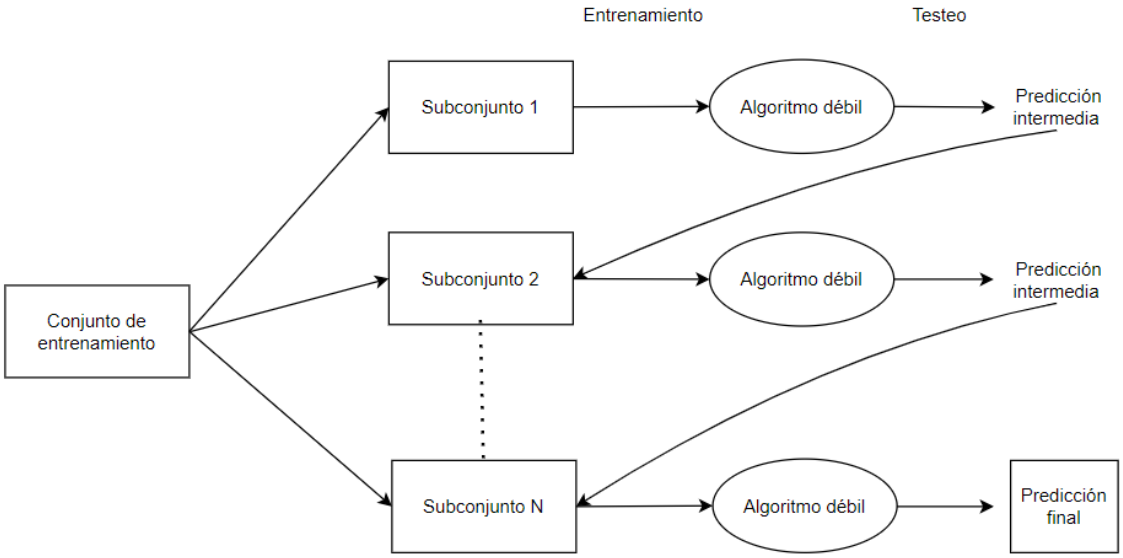
Los posibles criterios a seleccionar (entropía [11], gini index [12], etc…) priorizan aquellas variables capaces de crear conjuntos tan puros/homogéneos como sea posible, puesto que son aquellas que son capaces de dividir de mejor manera el conjunto y por tanto las que tiene un mayor poder predictivo. Una vez tengamos un subconjunto de observaciones donde todas pertenezcan a la misma clase tendremos un nodo hoja.

Una vez definida la estructura del árbol, clasificar es tan sencillo como navegar el árbol evaluando cada una de las variables correspondientes de la observación. Además, una variable tendrá mayor poder predictivo cuanto más cerca del nodo raíz se encuentre. Finalmente, estos modelos se conocen por su interpretabilidad ya que se analiza en lenguaje natural cuales son los factores que llevan a clasificar una observación en un determinado grupo.

Con respecto a las variables categóricas, este modelo requiere que las variables categóricas sean transformadas a valores numéricos previamente, para ello hemos utilizado la transformación *one hot encoding* [13] (10. Apéndice: Transformación de variables categóricas).

4.2. XGBoost

Es un modelo ensamblado que utiliza la técnica de boosting [14] tomando como modelos simples los árboles de decisión. El método de boosting consiste en combinar secuencialmente algoritmos simples donde cada modelo aprende de los errores del modelo anterior, asignándole un mayor peso en la función de pérdida a esas observaciones. En la siguiente imagen se ilustra esta técnica:



Este tipo de modelos ensamblados suele priorizar una mayor precisión debido al aprendizaje por refuerzo mencionado anteriormente, frente a la mayor estabilidad de los modelos ensamblados con la técnica de bagging. Esta mayor precisión tiene un coste, por un lado, suele tener un peor desempeño frente a conjuntos de datos con valores atípicos, y por otro tiene a resultar en modelos con alta precisión, pero alta varianza debido al aprendizaje por refuerzo. No obstante, presenta el inconveniente de que su tiempo de entrenamiento es mayor al ser secuencial mientras que los modelos basados en bagging pueden paralelizar el aprendizaje de sus modelos simples.

Tanto el algoritmo XGBoost [15] como Random Forest podan los árboles para aumentar la capacidad de generalización del modelo, sin embargo, la metodología difiere en ambos modelos. Mientras que Random Forest lo realiza después de haber entrenado a todos los árboles simples (podado del bosque) eliminando ramas innecesarias y fusionando nodos redundantes en todos los árboles, en XGBoost se realiza durante la etapa de *crecimiento* del árbol. Para ello existe un criterio de parada mediante el cual, si la ganancia a obtener no es suficiente, no se sigue creciendo esa rama.

Este modelo también requiere que las variables categóricas sean transformadas a variables numéricas por lo que aplicamos one hot encoding de nuevo.

4.3. Catboost

Catboost [16] es similar a Random Forest y XGBoost en el sentido de que todos son modelos ensamblados. También comparte la metodología boosting con el modelo XGBoost, sin embargo, la razón por la que lo he elegido es por su transformación de variables categóricas a valores numéricas.

La principal ventaja es que está especialmente diseñado para trabajar con variables categóricas como indica su nombre, utilizando una codificación especial interna conocida por *target encoding* [17] (10. Apéndice: Transformación de variables categóricas) de manera que puede aprovechar al máximo la información proporcionada por estas variables.

El beneficio de esta técnica radica en que es capaz de conservar información relevante entre las variables categóricas y la variable objetivo. Los modelos que la utilizan tienden a realizar sobreajuste, aunque Catboost cuenta con técnicas de regularización y validación cruzada para evitar este problema.

5. Elección del modelo

5.1. Elección de los hiperparámetros para cada modelo

Para la elección del modelo más adecuado para clasificar los pacientes se ha separado el conjunto en 3 subconjuntos: entrenamiento, validación y test. El conjunto test representa el 33% de las muestras mientras que el conjunto de entrenamiento más validación el 64%.

Se ha realizado *finetuning* sobre cada uno de los modelos buscando la mejor combinación de hiperparámetros para cada uno de ellos utilizando *grid search* junto con validación cruzada con 4 divisiones.

Al tratarse de un problema de clasificación binario donde el interés está en clasificar correctamente los casos positivos (cuando el paciente desarrollará pancreatitis) se ha tomado como medida de puntuación la *exhaustividad* (8. Apéndice: Métricas de rendimiento). Esta medida indica la ratio de pacientes con pancreatitis que son correctamente identificados por el modelo y el número total de pacientes que desarrollan la complicación. Los parámetros que se han tenido en cuenta en cada modelo son:

Modelo	Hiperparámetros
Random Forest	nº estimadores, nº variables, profundidad, criterio de división y peso de la clase positiva en la función pérdida
XGBoost	nº estimadores, ratio de aprendizaje, profundidad y peso de la clase positiva en la función pérdida
Catboost	nº iteraciones, ratio de aprendizaje, profundidad y peso de la clase positiva en la función pérdida

Usualmente se suele escoger aquel conjunto de hiperparámetros que tiene una mayor puntuación media donde la varianza de las puntuaciones no sea alta, ya que se preferible modelos estables a modelos inconsistentes con alta precisión de manera ocasional.

En esta ocasión en lugar de elegir el modelo con mayor puntuación media he querido estudiar si un modelo es significativamente mejor que otro desde un punto de vista estadístico. Para ello se suele realizar un test estadístico, pero los que se usan habitualmente asumen independencia de las muestras.

Normalmente cuando comparamos el rendimiento de dos modelos sobre un conjunto de particiones del conjunto de entrenamiento estás particiones muestran correlación en la

puntuación de los modelos ya que algunas son más fáciles de clasificar para los modelos mientras que otras son más difíciles, de manera que los modelos covarían y las puntaciones de los modelos no son independientes entre sí.

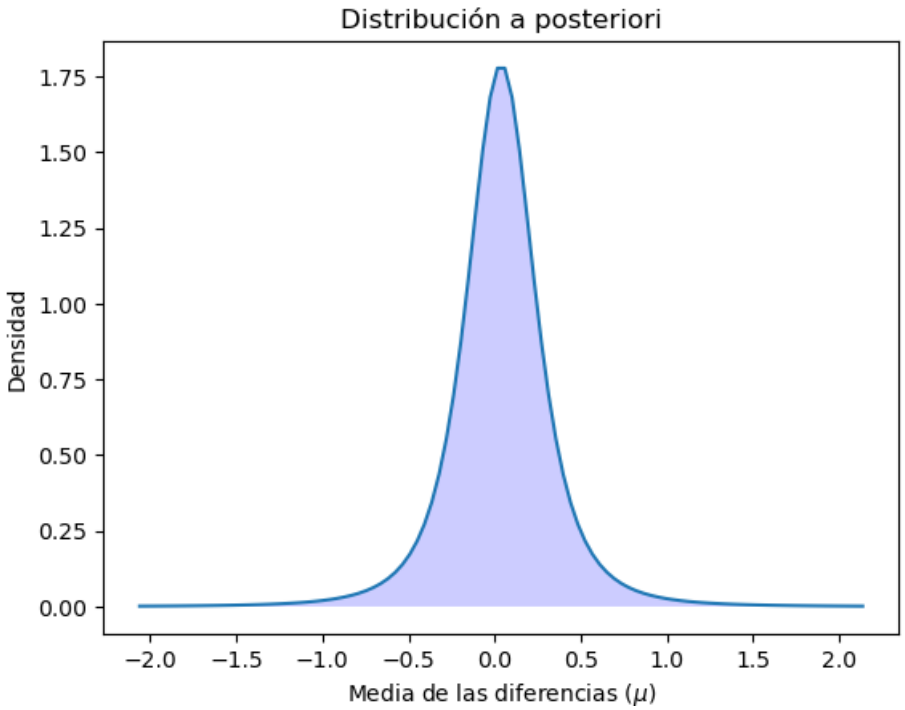
En este caso proponemos un enfoque bayesiano [18] que nos permita calcular la probabilidad de que un modelo sea mejor que otro en base a su rendimiento sobre las distintas particiones de los datos durante la validación cruzada.

Para realizar la estimación bayesiana definiremos la distribución a posteriori de forma que la distribución a posteriori tenga una forma cerrada, tomando una distribución a priori conjugada con la función de verosimilitud. En el artículo [18] Benavoli y sus compañeros sugieren que cuando queremos comparar el rendimiento de dos clasificadores podemos modelar la distribución a priori como una distribución Normal-Gamma (con media y varianza desconocidas) conjugada a la función de verosimilitud de la normal, expresando de esta manera la distribución a posterior como la distribución normal. Si marginalizamos la varianza de esta distribución a posteriori podemos definir la media como parámetro de una distribución t-Student, en particular:

$$St(\mu; n-1, \bar{x}, \left(\frac{1}{n} + \frac{n_{test}}{n_{train}}\right) \hat{\sigma}^2$$

donde n es el número total de muestras, $\bar{x}$ representa la diferencia de las medias en las puntuaciones sobre cada división, $n_{train}\ y\ n_{test}$ son el número de observaciones utilizadas para entrenamiento y testeo respectivamente, y $\hat{\sigma}^2$ representa la varianza de las diferencias observadas.

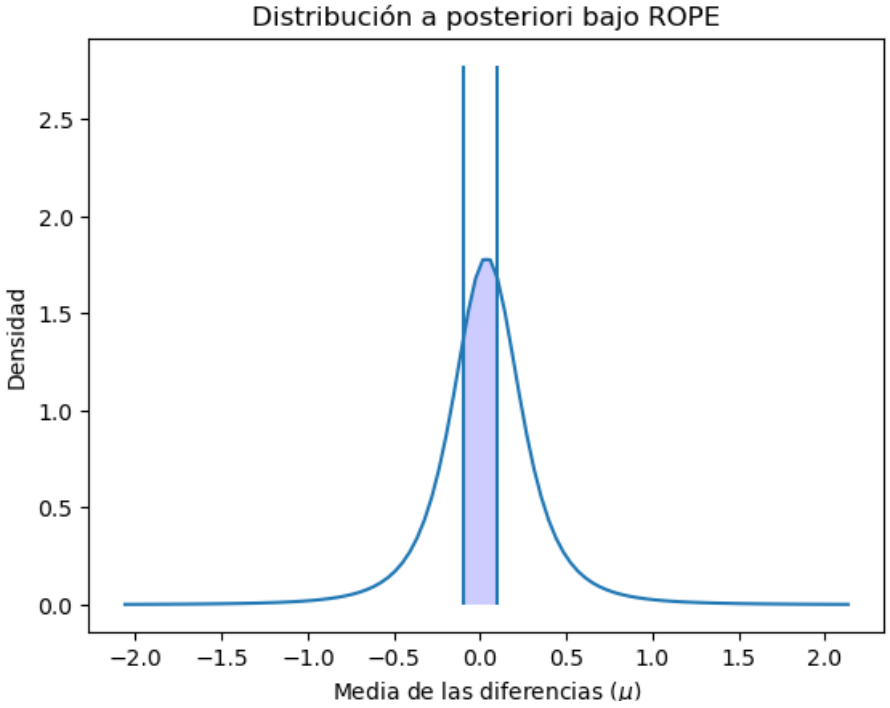
Podemos representar la distribución a posteriori usando el modelo definido anteriormente, representando las diferencias medias entre las puntuaciones de los modelos como:



De esta manera podemos calcular la probabilidad de que el primer modelo sea mejor que el segundo como el área bajo la curva de la distribución a posteriori desde cero a infinito y también en el caso contrario, calculando la probabilidad de que el segundo mejor sea mejor

que el primero como el área bajo la curva de la distribución a posteriori desde menos infinito hasta cero.

A veces un modelo no es mejor que otra de manera clara, por lo que tiene sentido definir regiones donde el rendimiento de un modelo es equivalente al otro. En este caso, definimos la región de equivalencia o ROPE como aquella donde la media de las diferencias se mueve en el intervalo [-0.01, 0.01], de manera que si difieren menos de un 1% en su rendimiento se consideran equivalente. Para ello, calculamos la probabilidad de que sean equivalente como el área bajo la curva de la distribución a posteriori en esa región como podemos ver en el siguiente gráfico:



Por cuestiones de visualización se ha representado la región [-0.1, 0.1] en lugar de la región original [-0.01, 0.01].

En resumen, este enfoque bayesiano nos permite calcular la probabilidad de que un modelo rinda mejor, peor o sean prácticamente equivalentes. Usando este enfoque hemos elegido la mejor combinación de hiperparámetros para cada modelo.

En el caso de Random Forest destacaron tres modelos, siendo el primero de ellos el más preciso y estable:

parameter_combination	split0_test_score	split1_test_score	split2_test_score	split3_test_score	mean_test_score	std_test_score
log_loss_5_log2_100_False_{0: 0.05, 1: 0.95}	0.2	0.0	0.25	0.2	0.1625	0.096014
log_loss_5_sqrt_100_False	0.4	0.0	0.00	0.2	0.1500	0.165831
gini_5_log2_100_True	0.0	0.0	0.25	0.2	0.1125	0.113880

Aun así, usamos el enfoque bayesiano para saber si es significativamente mejor que los otros dos:

	model_1	model_2	worse_prob	better_prob	rope_prob
0	log_loss_5_log2_100_False_{0: 0.05, 1: 0.95}	log_loss_5_sqrt_100_False	0.442	0.507	0.052
1	log_loss_5_log2_100_False_{0: 0.05, 1: 0.95}	gini_5_log2_100_True	0.245	0.682	0.074
2	log_loss_5_sqrt_100_False	gini_5_log2_100_True	0.416	0.549	0.035

La probabilidad de que el primer modelo sea mejor que el segundo en la primera comparación es del 50.7%, mientras que la probabilidad de que sea peor es del 44.2%. La probabilidad de que sean idénticos es del 5.2%. El primer modelo es claramente mejor que el tercero con un 68.2% de probabilidad. Por lo que, continuando el análisis de esta forma de las comparaciones, vemos que el tercer modelo es claramente peor que los dos primeros, y basándonos en la estimación bayesiana, la precisión media y la estabilidad del modelo, nos quedamos con la primera combinación de hiperparámetros.

En el caso de XGBoost todos los modelos tienen la misma puntuación sobre cada división, por tanto, no hay manera de elegir uno por encima de los demás usando el método descrito pues todos serían equivalentes:

parameter_combination	split0_test_score	split1_test_score	split2_test_score	split3_test_score	mean_test_score	std_test_score
0.1_15_100_0.95	0.2	0.0	0.25	0.2	0.1625	0.096014
1_5_100_0.95	0.2	0.0	0.25	0.2	0.1625	0.096014
0.1_10_100_0.95	0.2	0.0	0.25	0.2	0.1625	0.096014
1_10_100_0.95	0.2	0.0	0.25	0.2	0.1625	0.096014
0.1_5_100_0.95	0.2	0.0	0.25	0.2	0.1625	0.096014
1_15_100_0.95	0.2	0.0	0.25	0.2	0.1625	0.096014

En este caso, preferimos modelos simples a modelos complejos a nivel de profundidad y tras realizar varias pruebas sobre el conjunto test nos quedamos con el antepenúltimo modelo.

Por último, observamos el rendimiento de las mejores combinaciones de hiperparámetros para Catboost:

parameter_combination	split0_test_score	split1_test_score	split2_test_score	split3_test_score	mean_test_score	std_test_score
[0.05, 0.95]_5_100_0.001	1.0	1.0	0.75	0.6	0.8375	0.170935
[0.05, 0.95]_5_100_0.01	1.0	1.0	0.75	0.4	0.7875	0.245904
[0.05, 0.95]_15_100_0.001	0.8	1.0	0.75	0.4	0.7375	0.216145

El primer modelo es claramente el mejor al ser el de mayor precisión y estabilidad. Vemos los resultados obtenidos usando la estimación bayesiana:

	model_1	model_2	worse_prob	better_prob	rope_prob
0	[0.05, 0.95]_5_100_0.001	[0.05, 0.95]_5_100_0.01	0.245	0.682	0.074
1	[0.05, 0.95]_5_100_0.001	[0.05, 0.95]_15_100_0.001	0.150	0.809	0.041
2	[0.05, 0.95]_5_100_0.01	[0.05, 0.95]_15_100_0.001	0.245	0.682	0.074

Claramente el primer modelo es el mejor, en un 68.2% de probabilidad que el segundo y con un 80.9% que el tercero. Por tanto, nos quedamos con los hiperparámetros del primer modelo.

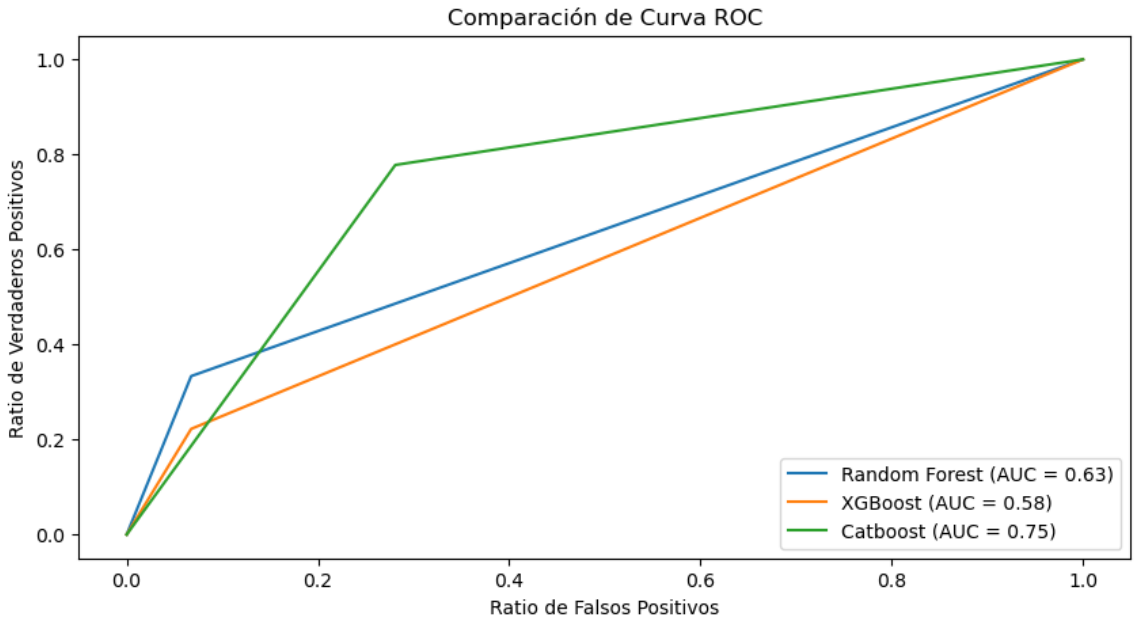
Los modelos finales elegidos con sus hiperparámetros son los siguientes:

Modelos	Nº Estimadores	Peso clase positiva	Ratio Aprendizaje	Profundidad	Criterio Separación	Max Nº Características
Random Forest	100	0,95		5	log loss	log2
XGBoost	100	0,95	0,1	5		
Catboost	100	0,95	0,001	5		

Si analizamos los hiperparámetros en común de los tres modelos, vemos como se prefieren árboles pequeños, con profundidad máxima 5 y como el número de estimadores deja de importar a partir de cierto punto. Además, es remarcable que aunque la proporción es 1/10 entre las clases binarias objetivo, los resultados son mejores si asignamos un peso de 0.95 a la clase positiva, en lugar de 0.9 como sugieren por defecto los modelos. Esto tiene sentido al utilizar como métrica de rendimiento la exhaustividad ya que premiamos más acertar correctamente pacientes que desarrollan pancreatitis.

5.2. Elección del mejor modelo

Para elegir el mejor modelo de los tres estudiados voy a evaluar su rendimiento sobre un conjunto no visto anteriormente, es decir, el conjunto de test. El objetivo es predecir correctamente tantos casos de pancreatitis como sea posible, por lo que la métrica más relevante para tomar la decisión sobre qué modelo escoger es la exhaustividad. Otras métricas han sido tenidas en cuenta como por ejemplo AUC o el área bajo la curva ROC:

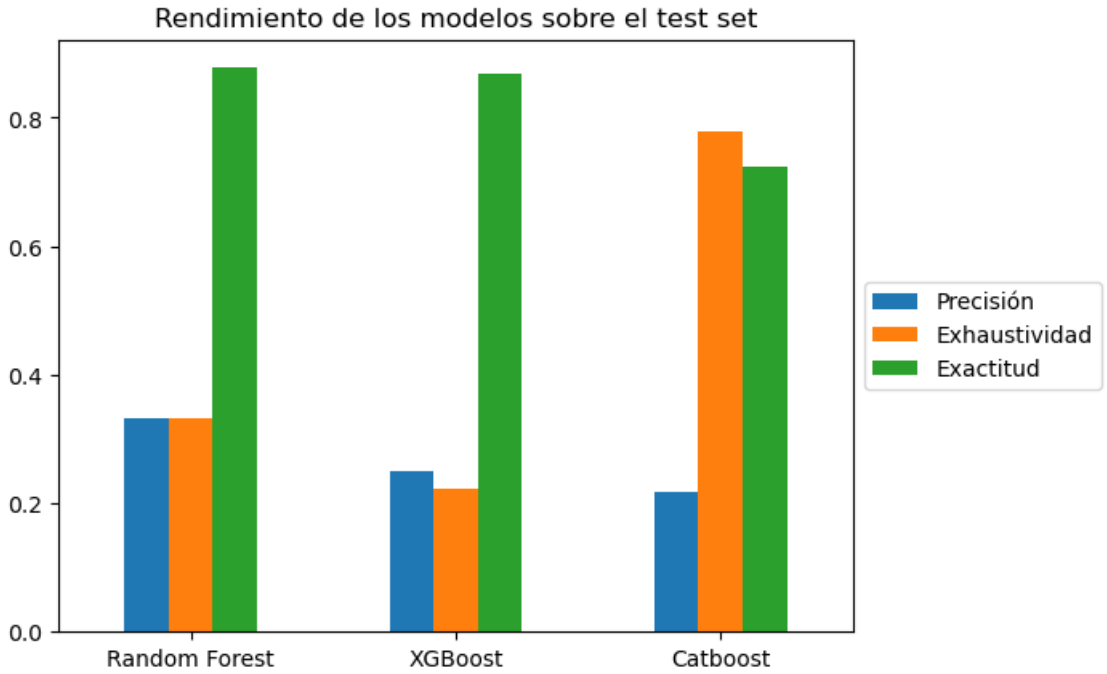


La curva ROC representa la ratio entre la sensibilidad (tasa de verdaderos positivos) y la especificidad (tasa de falsos positivos) a medida que se varía el umbral de clasificación del modelo. Este umbral representa a partir de limite de probabilidad se considera una observación como positiva. Por tanto, moviendo este umbral en el intervalo [0,1] se obtiene la curva ROC, calculando ambas ratios para cada valor del umbral.

El área bajo la curva ROC o AUC es una medida del rendimiento del modelo, donde un valor por debajo de 0.5 indica un modelo que no es mejor que un clasificador aleatorio y un valor igual a 1 sería el clasificador perfecto.

El AUC ordena claramente el rendimiento de los modelos, siendo el mejor el modelo Catboost seguido de Random Forest y por último XGBoost, con un rendimiento muy pobre.

Si analizamos otras métricas como precisión, exhaustividad y exactitud (8. Apéndice: Métricas de rendimiento) podemos evaluar el rendimiento de los modelos desde otros puntos de vista:

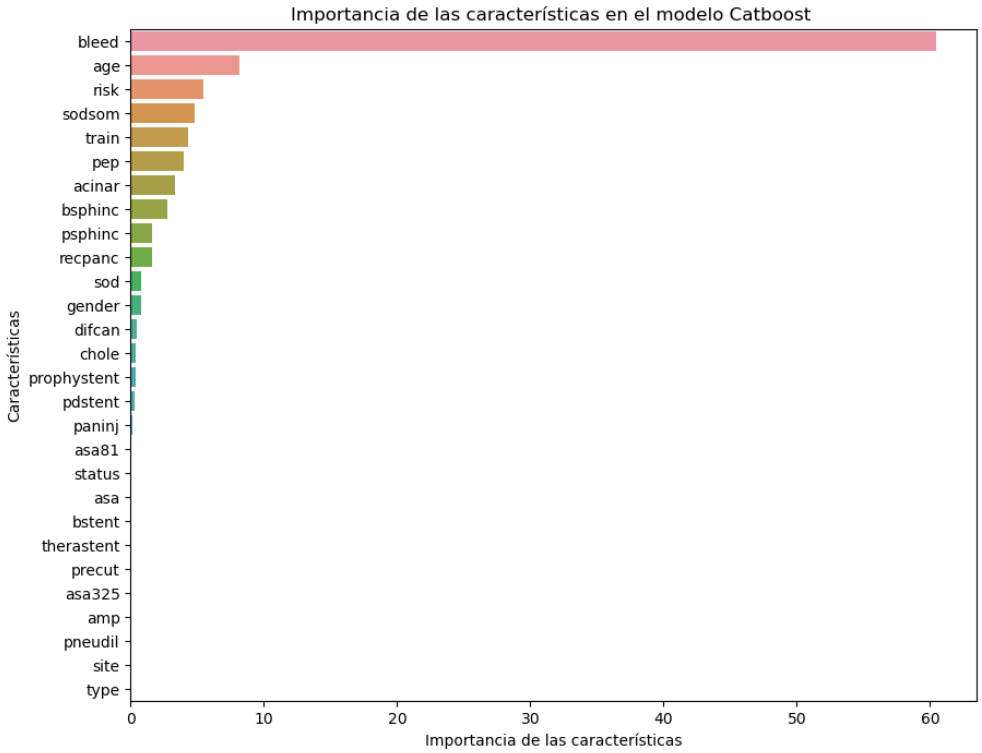


Analizando el gráfico superior vemos como los modelos Random Forest y XGBoost tienen mayor exactitud que Catboost a costa de una exhaustividad muy inferior al de este modelo. Esto se debe a que los dos primeros modelos predicen la mayor parte de los pacientes como que no desarrollaran complicaciones (clase mayoritaria), por lo que son incapaces de discriminar pacientes que desarrollaran pancreatitis (clase minoritaria). Sn embargo, el modelo Catboost sobre el conjunto test es capaz de identificar correctamente un 77% de los casos de pancreatitis, a costa de una menor precisión pues aumenta el número de falsos positivos.

Queda claro que dado el objetivo del problema y el rendimiento sobre un conjunto no entrenado anteriormente que el mejor modelo para realizar predicciones acerca de si el paciente desarrollara pancreatitis tras haberse tomado la indometacina es el modelo Catboost.

5.3. Interpretación del modelo seleccionado

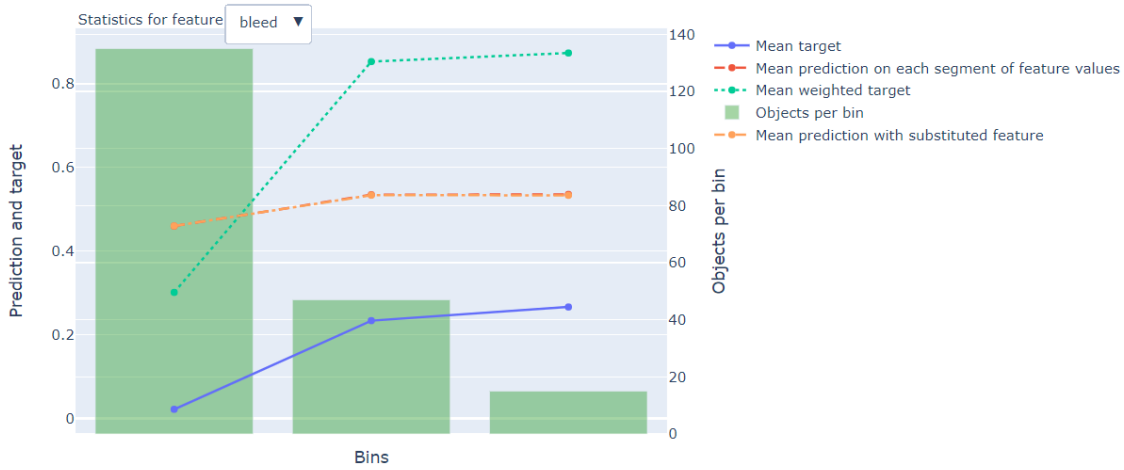
La principal ventaja de los modelos basados en árboles es su interpretabilidad. En particular, todos tienen por defecto una salida que calcula la importancia de una variable [19] para el modelo. Esta *importancia* se determina calculando el incremento en el error de predicción del modelo tras eliminar esa variable del modelo. Por tanto, una característica es importante si tras eliminarla del modelo el error de predicción aumenta y no es importante si no altera el error de predicción.



Dado este concepto de importancia de las características podemos ver en el gráfico las variables más importantes, donde destaca *bleed* o sangrado del paciente durante la CPRE por encima de todas. Otras variables con relevancia para el modelo son la edad, el riesgo asociado al paciente, si se había realizado una manometría del esfínter de Oddi, si un aprendiz participaba en la CPRE o si había padecido pancreatitis anteriormente tras someterse a otra prueba CPRE.

Así mismo podemos observar un conjunto de variables que apenas aportan información, como el tipo de disfunción de Oddi, el lugar del estudio, si había tomado aspirinas anteriormente o si se había realizado una ampulectomía por displasia o cáncer.

Catboost permite además profundizar en aquellas variables más relevantes, por ejemplo:



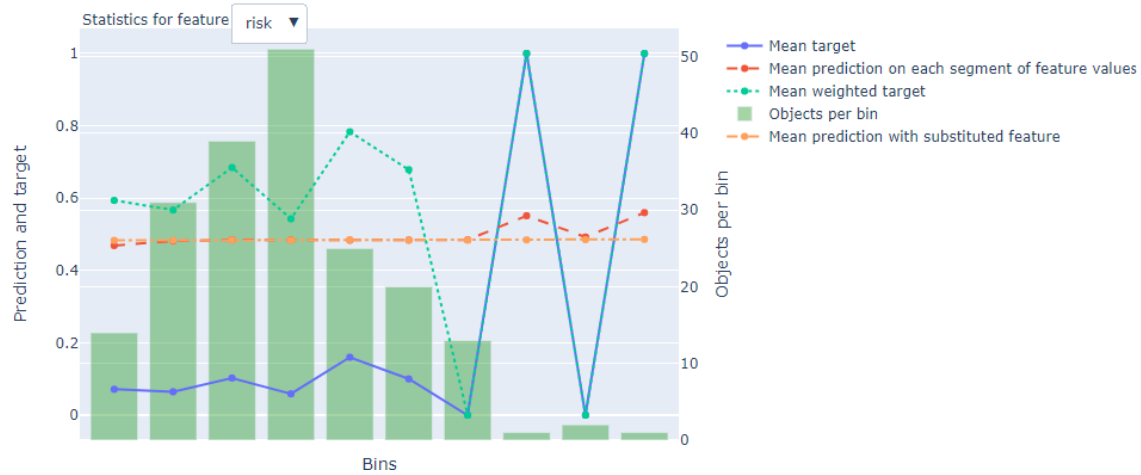
La variable *bleed* puede evaluarse de tres formas distintas: no sangrado, sangrado leve y sangrado grave. En lugar de fijarnos en la media de la variable objetivo, es mejor fijarnos en la media ponderada pues tiene en cuenta el desbalanceo entre clases. De esta forma vemos que en el grupo sin sangrado entorno al 30% tendría pancreatitis, mientras que en los otros grupos

es superior al 80%, por lo que sangrar durante la CPRE es un indicador claro de pancreatitis aun tomando indometacina. Los valores medios predichos para cada segmento coinciden con los valores medios predichos sustituyendo la variable. En poca palabra, el modelo en media predice que el paciente desarrolla pancreatitis si sangra, en caso contrario en media predice que no sufre la complicación post-CPRE.

Sin embargo, otras variables como la edad no permiten una interpretación clara que apunte a un grupo claro:

Como vemos en el gráfico superior no podemos apuntar al segmento de la población más envejecido, o a los jóvenes, o jóvenes adultos de manera clara.

Finalmente ahondaremos en la variable riesgo asignada previamente por los expertos ya que si permite una clara interpretación como podemos ver en el siguiente gráfico:



Aunque no existen números ejemplos de pacientes con alto riesgo, podemos observar como un alto riesgo es un claro indicador de que el modelo prediga complicaciones post-CPRE, mientras que pertenecer a un grupo de riesgo menor a 4 es un indicador para el modelo de que en media no se desarrollan complicaciones.

Si observamos los valores medios ponderados de la variable objetivo en los pacientes de mayor riesgo es fácil entender por qué el modelo realiza esas predicciones medias en esos casos.

6. Construcción de la herramienta personalizada.

El objetivo del proyecto se puede dividir en varias fases:

- Predecir qué pacientes desarrollarán pancreatitis a pesar de administrarles la indometacina.
- Predecir qué pacientes no desarrollarán pancreatitis y por tanto deben tomar la indometacina.
- Entender los factores mas determinantes que han llevado al modelo a tomar una predicción.

Actualmente contamos con un modelo entrenado sobre el conjunto de pacientes que recibió la indometacina para evitar desarrollar la pancreatitis tras el tratamiento CPRE. También tenemos un conjunto placebo donde conocemos si desarrollará o no la pancreatitis. Este conjunto puede utilizarse como el conjunto de prueba para nuestra herramienta personalizada.

El objetivo de esta herramienta es que, dado un paciente, realice una predicción sobre si la indometacina surte efecto y además provea de los factores más importantes para realizar esa predicción. Gracias a estos factores, el equipo médico y el paciente pueden tomar la mejor decisión en cada situación.

Para ello se han utilizado los valores Shapley [20], los cuales permiten entender mejor los resultados del modelo y como las distintas variables afectan a las predicciones realizadas por el modelo.

Estos valores se utilizan para calcular la importancia relativa de las variables en las predicciones realizadas por el modelo, explicando como cada característica contribuye al resultado final. Por esta razón son perfectos para la creación de herramientas personalizadas, pues permite evaluar cuales son los factores de riesgo del paciente en lenguaje coloquial, siendo fácilmente interpretables por el equipo médico y el paciente.

6.1. Ejemplos concretos con pacientes

Veamos con ejemplos concretos como se interpreta los valores de Shapley (9. Apéndice: Valores Shapley). El siguiente paciente se ha clasificado como que no sufrirá complicaciones post-CPRE. Si analizamos las variables más relevantes para esta clasificación tenemos:

La gráfica superior refleja las características en rojo que empujan al modelo a predecir que habrá complicaciones mientras que las variables en azul empujan al modelo a predecir que no habrá complicaciones. Las variables más relevantes para predecir que el paciente no desarrollará pancreatitis son:

- No se ha producido sangrado tras CPRE.
- Su perfil de riesgo es bajo, en concreto es 1.
- No se le ha realizado una esfinteroctomía en el páncreas.

El resto de las variables su peso es irrelevante. Las variables que apoyan a predecir pancreatitis son:

- El hecho de presentar coledocolitiasis, cálculos biliares bloqueando el conducto.
- Su edad de 24 años
- El hecho de haber desarrollado pancreatitis anteriormente.

Analicemos otro paciente, en este caso de un paciente que según el modelo padecerá pancreatitis a pesar de que tome indometacina:

Vemos como sangrado grave, el hecho de que un aprendiz participe en la CPRE y perfil de riesgo alto son las tres variables de mayor impacto para la predicción final. La única variable que apunta en otra dirección es el hecho de haber realizado una manometría.

De esta forma este modelo es una herramienta personalizada que permite ayudar al equipo médico a tomar decisiones sobre el uso de la indometacina, además de identificar a priori qué pacientes van a desarrollar pancreatitis y prepararse para esa consecuencia de la CPRE o probar otros tratamientos alternativos.

7. Conclusiones

La pancreatitis es una complicación que sucede entre el 15-20% de las veces tras someterse a la colangiopancreatografía retrógrada endoscópica (CPRE), un procedimiento médico que permite diagnosticar y tratar condiciones del hígado, de la vesícula biliar, de los conductos biliares y del páncreas.

A lo largo de este proyecto se ha probado como se pueden utilizar herramientas de aprendizaje automático (Random Forest, XGBoost y Catboost) para predecir enfermedades, en este caso con el aliciente de identificar aquellos pacientes donde el remedio(indometacina) frente a la complicación no surge efecto, a fin de evitar la administración de medicamentos que no tendrán efecto.

Otra investigación que se ha desarrollado en este proyecto es justificar cuales son los mejores hiperparámetros para cada modelo, no fijándonos únicamente en puntuaciones medias y estabilidad, si no desarrollando un método bayesiano que permite justificar con que probabilidad es mejor, peor o equivalente una combinación de hiperparámetros, con el objetivo de suavizar que ciertas divisiones del conjunto de datos sean especialmente favorables para ciertas combinaciones.

Además de identificar complicaciones adecuadamente, el hecho de utilizar modelos basados en árboles permite una mayor interpretabilidad de los resultados. Gracias a los valores Shapley podemos construir una herramienta personalizada que indique que pacientes sufrirán complicaciones o no y cuales son las diferentes características que llevan a esa conclusión.

8. Apéndice: Métricas de rendimiento

En este apéndice vamos a definir las métricas de rendimiento que han ayudado a elegir los hiperparámetros mas adecuados para cada modelo además del modelo mas conveniente para resolver el problema de clasificación desbalanceado.

8.1. Exhaustividad

También conocida como *recall*, expresa la cantidad de observaciones positivas que somos correctamente capaces de clasificar sin tener en cuenta el número de falsos positivos. Se suele utilizar a la hora de resolver problemas donde el coste de cometer un falso positivo es mucho menor al coste de cometer un falso negativo, como por ejemplo la detección del cáncer. Si no detectamos a un paciente con cáncer, corre el riesgo de morir sin ser tratado adecuadamente. El cálculo de la exhaustividad se define como:

$$exhaustividad = \frac{verdaderos\ positivos}{verdaderos\ positivos\ + falsos\ negativos}$$

8.2. Precisión

Esta medida permite expresar la calidad del modelo en tareas de clasificación binaria. A diferencia de la exhaustividad tiene en cuenta el número de falsos positivos. Se calcula de la siguiente forma:

$$precision = \frac{verdaderos\ positivos}{verdaderos\ positivos\ + falsos\ positivos}$$

8.3. Exactitud

La exactitud mide el porcentaje de observaciones acertadas por el modelo en un problema de clasificación binario, tanto las observaciones positivas como las observaciones negativas. Sin embargo, puede llevar fácilmente al engaño en problemas desbalanceados como al que nos enfrentamos, ya que con clasificar todos los pacientes como que no desarrollarán pancreatitis, el modelo tendría una exactitud del 90%. El modelo sería inútil al ser incapaz de identificar pacientes que desarrollaran pancreatitis. Esta métrica se calcula como:

$$exactitud = \frac{verdaderos\ positivos + verdaderos\ negativos}{nº\ de\ observaciones}$$

9. Apéndice: Valores Shapley

Se trata de una métrica que permite medir la importancia relativa de las variables en las predicciones del modelo. Estos valores están basados en la teoría de juegos cooperativos y se utilizan para evaluar la contribución marginal de cada jugador en juegos de coaliciones.

Son valores justos y consistentes gracias a los axiomas de simetría y aditividad. El axioma de simetría establece que dos jugadores que aportan lo mismo a la coalición deben tener los mismos valores Shapley. Por otro lado, el axioma de aditividad establece que la suma de todos los valores Shapley es igual al resultado total del juego. De esta forma se proporciona una

manera justa de asignar un valor a la contribución de cada jugador al juego de coalición, teniendo en cuenta todas las combinaciones de jugadores y su impacto en los resultados del juego.

En un juego de coalición, los jugadores se alían para formar coaliciones y lograr un objetivo común. En este caso, para conectar la teoría de juegos con la teoría de modelos de aprendizaje automático debemos conectar ambas teorías. Por ello, veremos las variables de un modelo con los jugadores del juego, que colaboran para recibir un ganancia. La ganancia del juego se define como el valor real menos el valor medio para todas las observaciones. Finalmente, el juego es la función de predicción para una única instancia del conjunto de datos.

Los valores Shapley se calculan como la contribución marginal de cada valor de una variable a lo largo de todas las coaliciones posibles. Veámoslo mejor con un ejemplo.

Supongamos que queremos calcular el precio de un apartamento de 100 metros cuadrados en el Barrio Salamanca con el parque del Retiro a menos de 5 minutos. El precio predicho por el modelo es de 600000€ y queremos conocer la contribución del valor de cada variable al precio final. Los valores de las variables serían:

- Parque a menos de 5 minutos.
- Barrio Salamanca.
- 100 metros cuadrados de área.

Adicionalmente supondremos que la predicción media es de 650000€. Por tanto, el objetivo es explicar la diferencia entre ambas de 50000€ y como contribuyen los valores de las variables a ello.

Evaluemos cual es la contribución del hecho de que se encuentre a menos de 5 minutos. Eliminamos este valor y suponemos que se encuentra a 1 minutos del parque del Retiro y predecimos el precio final del modelo, en este caso es de 700000€, por tanto, la contribución de estar a 5 minutos es de 650000-700000, es decir, de -50000€. Si repetimos este proceso para todos los valores de la variable cercanía al Retiro obtendremos un valor mas preciso de la contribución de este valor de la variable a la predicción.

10. Apéndice: Transformación de variables categóricas

El problema planteado durante el proyecto tiene dos particularidades, la presencia de variables categóricas y el desbalanceo en la clase objetivo. Los modelos no son capaces de entender las variables cualitativas, por ello deben ser transformadas a variables numéricas. En las siguientes subsecciones se presentan los dos métodos utilizados para realizar esta transformación.

10.1. One hot encoding

Se trata de un método de transformación de variables categóricas a variables numéricas. Por cada categoría distinta de la variable categórica se crea una nueva variable booleana, que será verdadera en aquellas instancias donde haya presencia de dicha categoría y falsa en el caso contrario. Es mas sencillo de entender con un ejemplo.

Supongamos que tenemos la variable color compuesta por tres colores: azul, rojo y amarillo. Nuestro conjunto de datos está compuesto 10 observaciones y dos variables: un identificador y la variable color. Tendríamos, por tanto:

id	color
1	azul
2	azul
3	azul
4	amarillo
5	amarillo
6	rojo
7	azul
8	rojo
9	amarillo
10	rojo

Por cada valor de la variable color creamos una variable booleana y transformamos la variable categórica color en 3 variables numéricas:

id	color	es azul	es amarillo	es rojo
1	azul	VERDADERO	FALSO	FALSO
2	azul	VERDADERO	FALSO	FALSO
3	azul	VERDADERO	FALSO	FALSO
4	amarillo	FALSO	VERDADERO	FALSO
5	amarillo	FALSO	VERDADERO	FALSO
6	rojo	FALSO	FALSO	VERDADERO
7	azul	VERDADERO	FALSO	FALSO
8	rojo	FALSO	FALSO	VERDADERO
9	amarillo	FALSO	VERDADERO	FALSO
10	rojo	FALSO	FALSO	VERDADERO

La desventaja de este método es la creación de muchas variables en caso de variables categóricas con muchas categorías, donde además estas serían nulas la mayor parte de las veces, un comportamiento complicado de interpretar para algunos modelos.

10.2. Target encoding

Esta técnica de codificación consiste en transformar las variables categóricas en estadísticas resumen relacionadas con la variable objetivo. En problemas binarios, a cada valor de cada variable categórica se le asigna la proporción de ejemplos positivos con ese valor, siendo esta proporción la estadística resumen.

Por ejemplo, la variable *site* representa el centro donde el estudio del paciente se ha llevado a cabo, habiendo cuatro posibles destinos: Universidad de Michigan, Universidad de Indiana,

Universidad de Kentucky y Case Western. Podemos resumir esta variable tomando como estadística la media con respecto a la variable objetivo de la siguiente forma:

Id	Site	Objetivo	Site Codificado
1	UM	Falso	0
2	IU	Falso	0
3	UK	Falso	1/2
4	Case	Verdadero	1/3
5	UM	Falso	0
6	UM	Falso	0
7	IU	Falso	0
8	UK	Verdadero	1/2
9	Case	Falso	1/3
10	Case	Falso	1/3

Al no haber ningún paciente clasificado como positivo cuyo site es UM o IU, se codifican como cero. En el caso de site UK, contamos como uno de los dos pacientes es clasificado como positivo, por tanto, asignamos 0,5. Finalmente en el caso de Case Western, hay un caso positivo de tres casos totales, por tanto, se codifica como 0,3.

A diferencia de one hot encoding no crea tantas variables como categorías aumentando la dimensionalidad del problema considerablemente. Además, relaciona la variable categórica con la variable objetivo, otorgando mas información al modelo sobre la relación entre ambas.

Bibliografía

[1] M. C. W. B. P. M. L. S. M. L. Azan Zahir Virji, «Patients Like You: How Machine Learning Can Be Userd as a Shared Decision-Making Tool to Improve Care,» *NEJM Catalyst,* vol. DOI: 10.1056/CAT.21.004, p. 8, 2021.

[2] M. J. M. S. M. G. A. L. M. B. Joseph Elmunzer, «A Randomized Trial of Rectal Indomethacin,» *The new england journal of medicine,* p. 9, 12 April 2012.

[3] «Stanford Medicine,» [En línea]. Available: https://www.stanfordchildrens.org/es/topic/default?id=endoscopic-retrograde-cholangiopancreatography-ercp-92-P09228.

[4] «Stanford Medicine,» [En línea]. Available: https://stanfordhealthcare.org/medical-conditions/digestion-and-metabolic-health/acute-pancreatitis.html.

[5] «Clínica Universidad de Navarra,» [En línea]. Available: https://www.cun.es/enfermedades-tratamientos/medicamentos/indometacina.

[6] «Cigna Healthcare,» [En línea]. Available: https://www.cigna.com/es-us/knowledge-center/hw/medicamentos/indomethacin.

[7] «Clínica Universidad de Navarra,» [En línea]. Available: https://www.cun.es/diccionario-medico/terminos/pancreatografia.

[8] L. Breiman, «Bagging Predictors,» de *Machine Learning*, https://doi.org/10.1023/A:1018054314350, August 1996, pp. 123-140.

[9] L. Breiman, «Random Forests,» de *Machine Learning 45*, https://doi.org/10.1023/A:1010933404324, 2001, pp. 5-32.

[10] L. R. &. O. Maimon, «Decision Trees,» de *Data Mining and Knowledge Discovery Handbook*.

[11] C. Shannon, «A Mathematical Theory of Communication,» *The Bell System Technical Journal,,* vol. 27, pp. 379-423, 1948 October.

[12] F. A. Farris, «The Gini Index and Measures of Inequality,» *The American Mathematical,* vol. 117, pp. 851-864, December 2010.

[13] «One Hot Encoding,» [En línea]. Available: https://interactivechaos.com/es/manual/tutorial-de-machine-learning/one-hot-encoding.

[14] R. E. Schapire, «The Boosting Approach to Machine Learning,» December 2001.

[15] C. G. Tianqi Chen, «XGBoost: A Scalable Tree Boosting System,» June, 2016.

[16] G. G. A. V. A. V. D. A. G. Liudmila Prokhorenkova, «CatBoost: unbiased boosting with categorical features,» June 2017.

[17] F. P. F. T. J. e. a. Pargent, «Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features,» *Computational Stat,* vol. 37, pp. 2671-2692, February 2022.

[18] G. C. J. D. M. Z. Alessio Benavoli, «Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis,» *Journal of Machine Learning Research,* vol. 18, pp. 1-36, 2017.

[19] L. Breiman, «Random Forest,» de *Machine Learning Springer*, 2001.

[20] S.-I. L. Scott Lundberg, «A Unified Approach to Interpreting Model Predictions,» 2017.