Tema 2 - Trabajo práctico

Inteligencia Artificial Explicable 2024/25

Juan Eizaguerri Serrano juan.eizaguerri@alumnos.upm.es
Daniel Hernández Martínez daniel.hernandezmar@alumnos.ump.es

Índice

1. Introducción

En esta primera versión de la práctica vamos a resolver un problema de clasificación. El objetivo del problema es clasificar los individuos según su riesgo en dos clase: bueno y malo. Para ello contamos con dos conjuntos de datos: un conjunto de entrenamiento y otro que será utilizado a la vez como validación y test. Para la versión final se contará con un conjunto de datos diferente para validación y test.

Para la realización de esta primera versión se comenzó realizando un pequeño análisis de los datos. Posteriormente se hizo un breve estudio de los hiperparámetros y se creó un árbol con los datos obtenidos. Sobre este árbol, y aprovechando la naturaleza inherentemente explicativa, se procedió a realizar explicaciones tanto globales como locales y ha realizar las diversas métricas de evaluación correspondiente, observando el patrón de alta explicabilidad y baja capacidad predictiva propio de los árboles de decisión. Finalmente se transformó el árbol en reglas para estudiar si se favorecía la explicabilidad.

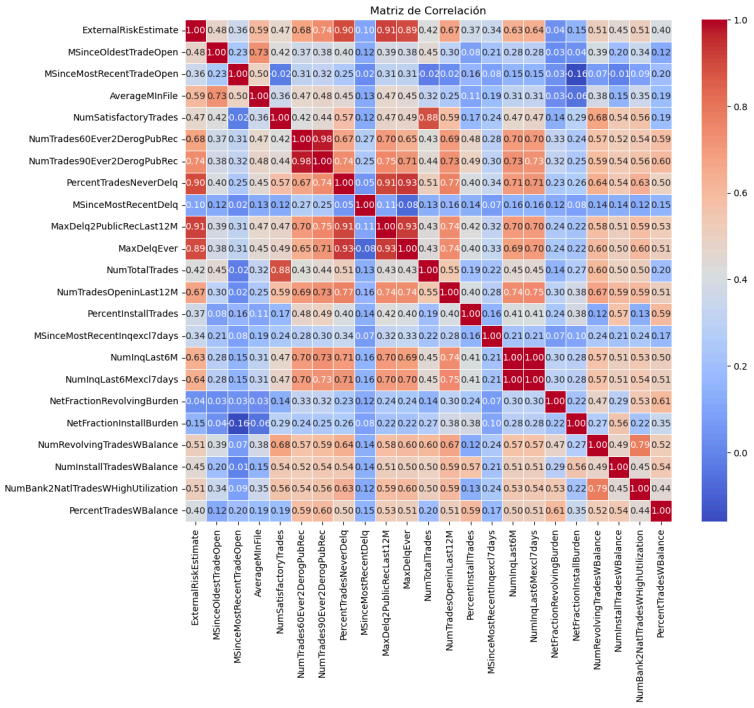
2. Metodología

El objetivo de este trabajo es el de desarrollar, entrenar y evaluar distintos clasificadores,teniendo en cuenta las métricas de confusión de los modelos frente a un conjunto de datos de validación, y discutiendo la interpretabilidad global y local de dichos modelos.

El conjunto de datos cuenta con un total de 24 variables, siendo la primera de ellas *RiskPerformance* la que se tratará de predecir. Esta variable toma los valores de cadena 'Good' o 'Bad', que pueden ser mapeados a los valores 1 y 0 respectivamente. El resto de variables toman valores enteros. Los datos están divididos en un conjunto de entrenamiento que cuenta con 6459 registros y otro de validación de 2000 registros. Observando la variable *RiskPerformance* se observa un balanceo de los datos aproximadamente equitativo en ambos conjuntos.

Partición	n_{total}	$\#n_{Good}$ (%)	$\#n_{Bad}$ (%)
Train	6459	3109 (0.48)	3350 (0.52)
Validation	2000	967 (0.48)	1033 (0.52)

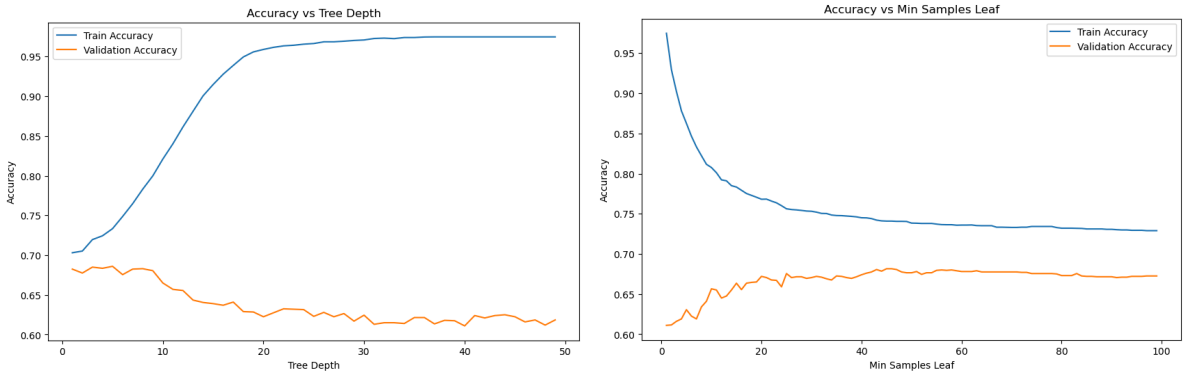
La matriz de correlación del conjunto de datos muestra un alto grado de correlación entre muchas de las variables, por lo que en muchos casos pueden aportar información similar a la hora de realizar predicciones.



Se entrenan distintos modelos de clasificadores utilizando los datos de entrenamiento y se evalúa su desempeño frente al conjunto de validación.

En primer lugar se utiliza un clasificador lógico, en concreto un árbol de decisión. Este clasificador tiene la ventaja de ser inherentemente explicable, además de requerir un grado de preparación de los datos muy bajo, pues permite el uso de datos categóricos y sin normalizar.

Con el objetivo de buscar el mejor clasificador de árbol de decisión posible, se calcula la precisión del modelo en función de los parámetros *max_depth*, que limita la profundidad máxima del árbol, y *min_samples_leaf*, que establece el número mínimo de muestras que pueden ser afectadas por una regla del árbol.



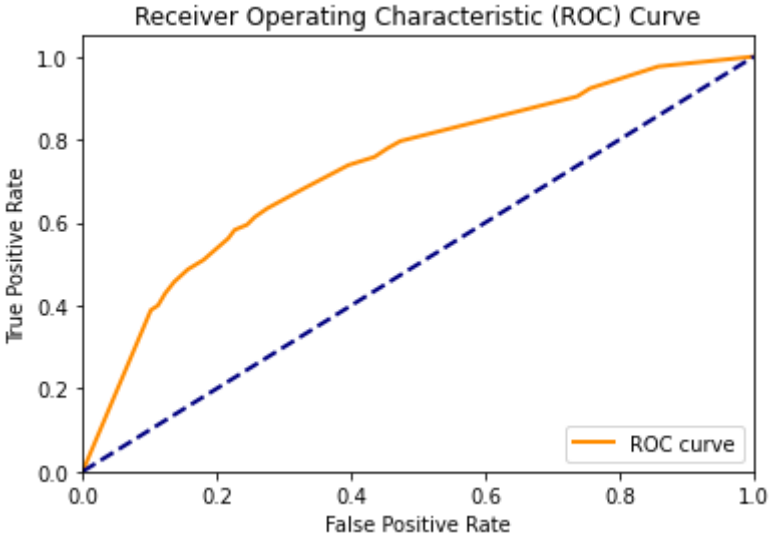
Se observa que a medida que aumenta la profundidad máxima del árbol puede producir un sobreajuste muy grande, ya que se crean reglas muy específicas haciendo que el modelo “memorice” los datos de entrenamiento, lo que tiene un efecto negativo a la hora de evaluar con datos que no había visto anteriormente. Por el contrario, el parámetro *min_samples_leaf* es una herramienta para combatir el sobreajuste, evitando las hojas ajustadas a datos específicos del conjunto de entrenamiento. El efecto de este parámetro deja de tener un efecto significativo alrededor de *min_samples_leaf*=40.

3. Resultados

Clasificador	Mejores parámetros	ACC_{train}	ACC_{test}	AUC
Árbol de decisión	*max_depth=5* *min_samples_leaf*=20	0.73	0.69	0.73

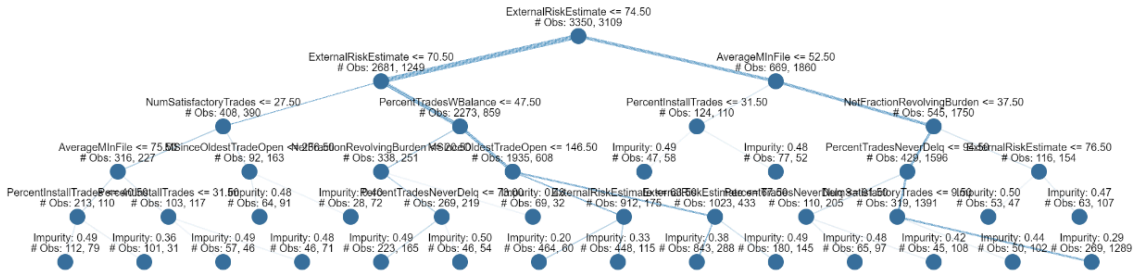
Se realiza una evaluación del rendimiento del modelo entrenado utilizando los datos de entrenamiento y validación. La tabla anterior muestra un resumen de los resultados obtenidos en estas pruebas.

Para el modelo de árbol de decisión se obtiene un *accuracy* del 73% en el conjunto de entrenamiento y 69% en el de test. Observando la curva ROC frente a datos de test se observa un desempeño significativamente mejor que el de un clasificador aleatorio.



4. Modelos

4.1. Interpretación global



Analizando el árbol de podemos observar como la variable más importante del conjunto de datos para nuestro árbol es el estimado del riesgo externo (ExternalRiskEstimate), ya que es la que se utiliza en el nodo raíz. Esto quiere decir que es la variable que sirve para dividir el conjunto de datos de forma más efectiva. Cabe recalcar que esto sucede para nuestro árbol en concreto y no tendría por qué ser así para todos los árboles que solucionen el problema. Podemos observar que al utilizar esta variable para dividir el conjunto de datos entre los individuos que presentan un estimado del riesgo mayor y menor a 74.50, se obtienen dos conjuntos bastante diferentes. Uno de ellos muestra un número mayor de individuos cuyo riesgo es bueno (clase 0) y el otro contiene un número mayor de individuos con riesgo malo (clase 1). Esa diferencia entre los individuos de ambos subconjuntos muestra como ExternalRiskEstimate es una variable capaz de discriminar de forma bastante efectiva entre ambas clases, lo que explica por qué es la primera que se utiliza para dividir el conjunto de datos. Además esta también es utilizada en la siguiente división realizada en el subárbol izquierdo, reforzando así la importancia de esta variable. Debido a esto, si queremos explicar las decisiones tomadas por nuestro árbol, podríamos afirmar de forma bastante segura que uno de los factores más importantes es el estimado del riesgo externo, pudiendo dar así una primera idea del por qué un individuo ha sido asignado a una clase.

Si analizamos los *features importances* de las variables, podemos observar que la variable que presenta un mayor valor es ExternalRiskEstimate, coincidiendo con la primera variable utilizada en el árbol. Aún así, cabe recalcar que el que una variable tenga un gran *feature importance* no implica que esta vaya a ser utilizada en todos los árboles. Puede que sea una de las más prioritarias en cierto árbol mientras que no sea utilizada en absoluto por otro árbol. En nuestro caso, todas las variables que presentan un *feature importance* positivo son utilizadas en el árbol, aunque esto no tiene por qué pasar en todos los árboles. Sin embargo, ninguna de las variables que presentan un valor de *feature importance* igual a 0 son utilizadas. Esta es una forma muy directa de poder mostrar qué variables no influyen en la decisión, pudiendo dar información sobre qué aspectos no se tienen en cuenta a la hora de decidir la clase de un individuo.

Si volvemos a la variable ExternalRiskEstimate, podemos observar que no solo es la primera que se utiliza, si no que además se usa otras 4 veces en nuestro árbol de profundidad 5. Esto explica por qué es la variable con mayor *feature importance*.

Si queremos seguir profundizando en qué variables son relevantes, podríamos destacar las siguientes utilizadas en el árbol, las cuales son AverageMinFile, NumSatisfactoryTrades, PercentTradesWBalance, PercentInstallTrades y NetFractionRevolvingBurden. Cuanto más bajemos en el árbol, menor influencia tendrán dichas variables en la decisión, por lo que observar las variables utilizadas en nodos muy inferiores no nos da información muy relevante a nivel global, aunque sí puede ser útil a nivel local, al tratar de explicar la decisión tomada para un individuo en concreto.

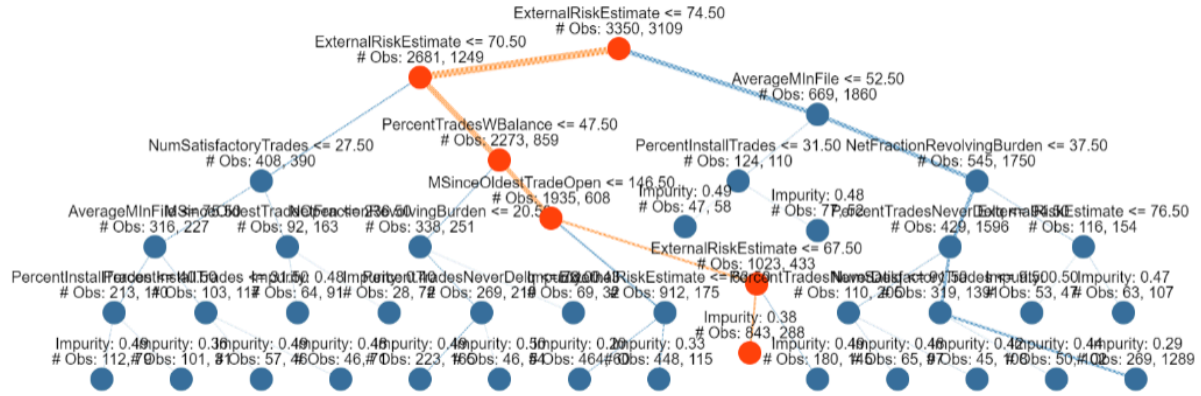
Si analizamos los nodos hoja, podemos observar que no existe ninguno que presente conjuntos completamente heterogéneos, dificultando así la fiabilidad del modelo. Aún así, si existen ciertos nodos que presentan una cantidad de individuos de una clase mucho mayor que de la otra, como es el caso del último nodo, con una proporción de individuos del 17% y 83%. Si algún individuo cae en ese nodo, podemos tener una gran seguridad de que será bien clasificado, mientras que si pertenece a un nodo con una proporción cercana al 50-50, como es el caso del sexto nodo hoja (46% y 54%), no podemos ofrecer una clasificación fiable.

Pese a que una profundidad de 5 pueda no parecer muy elevada, podemos observar como el árbol obtenido es bastante complejo, haciendo que sea difícil de interpretar. Debido a esto, una opción posible sería reducir la profundidad máxima. De esta forma podríamos perder capacidad predictora, pero favoreceríamos la interpretabilidad. Depende de cual sea el aspecto al que demos más importancia y la finalidad del modelo, que podemos considerar esta opción o descartarla completamente.

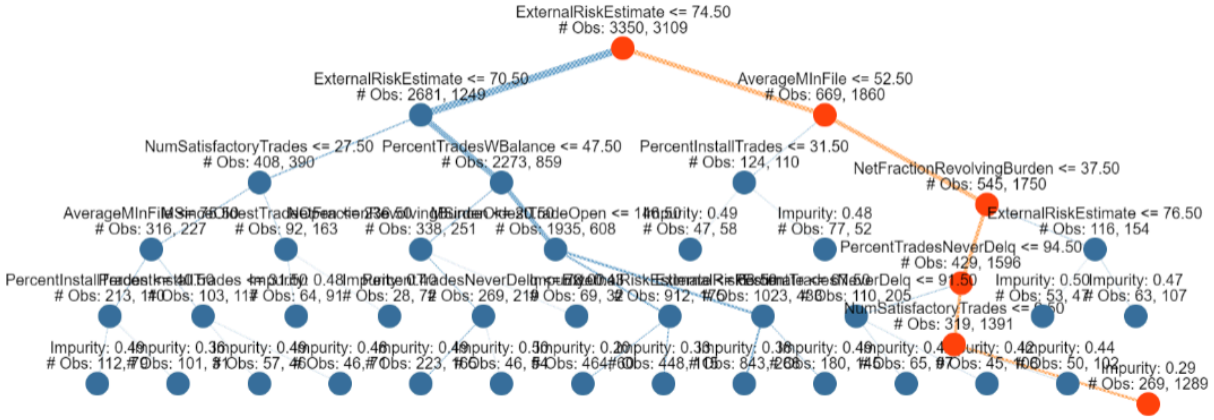
4.2. Interpretación local

En un árbol de decisión, es posible comprobar el razonamiento seguido por el modelo para llegar a la respuesta obtenida.

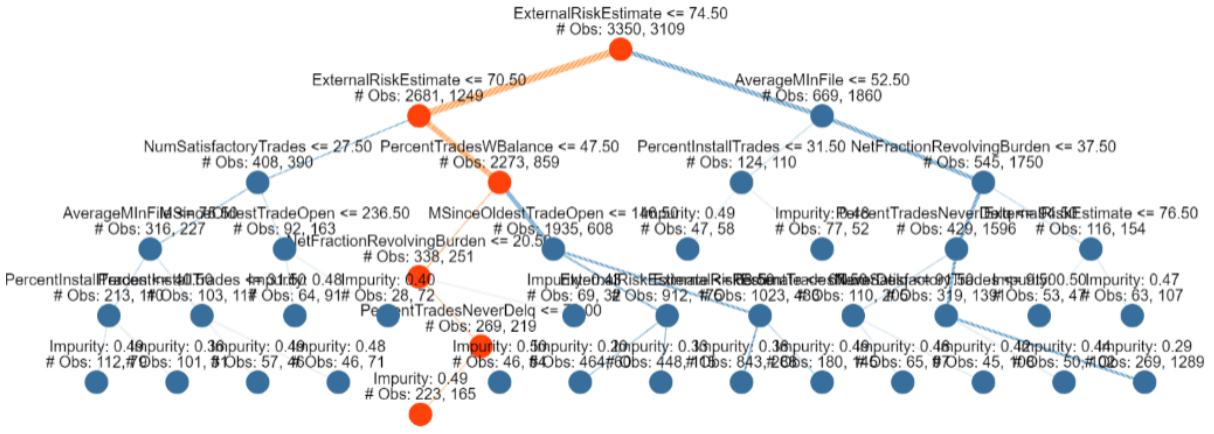
En la siguiente figura se muestra el camino del árbol seguido ante un ejemplo que ha sido clasificado correctamente como ‘Bad’, con una confianza de 0.745. Se observa que todos los nodos seguidos tienen un buen grado de acierto frente a los datos de entrenamiento.

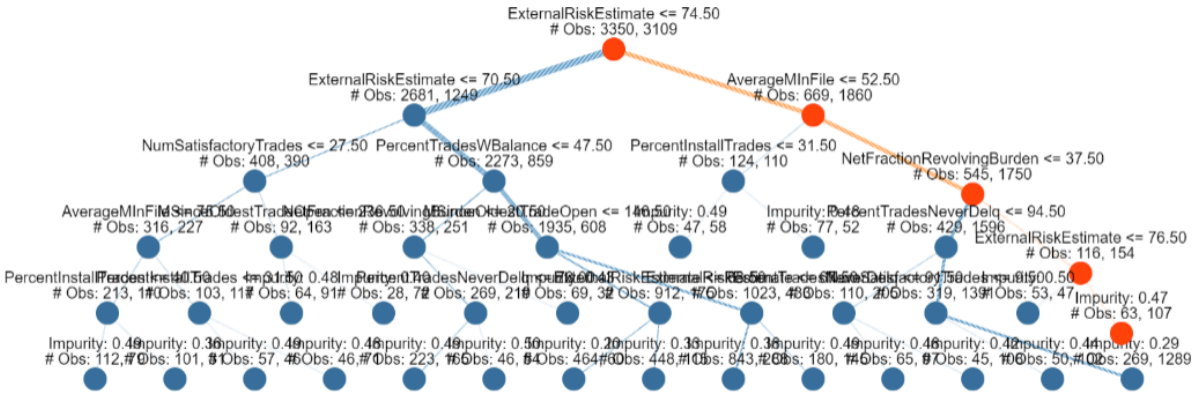


De forma similar, en la siguiente figura se muestra una clasificación correcta del valor ‘Good’ con una confianza todavía mayor, de 0.827.



Este método para interpretar los resultados es especialmente útil a la hora de comprobar por qué se ha equivocado el modelo en sus predicciones. A modo de ejemplo, a continuación se muestra un caso de falso positivo y otro de falso negativo.





En primer lugar, cabe destacar que la confianza del modelo para estas predicciones es significativamente menor a la de las predicciones verdaderas (PrScore=0.575 para el primer ejemplo y PrScore=0.629 para el segundo). Además, ambas predicciones son incorrectas por motivos similares: La existencia de al menos una regla que no es suficientemente discriminante.

En el caso del falso positivo, la hoja a la que se llega ha tenido 223 aciertos frente a 165 fallos durante el entrenamiento, una proporción muy mejorable. Similarmente, en el falso negativo se llega a una hoja en la que se han observado 107 aciertos frente a 63 fallos.

Estos errores pueden ser solucionados aumentando la profundidad del árbol y con ello el número de reglas, para que el modelo se adapte mejor al conjunto de datos, pero como se ha visto en el apartado de metodología, esto no es recomendable ya que aumentaría significativamente el sobreajuste del modelo.

5. Discusión

Debido a la interpretabilidad inherente a los árboles de decisión, ha sido bastante sencillo entender las decisiones que toma el modelo, tanto a nivel global como local. Aún así, esto trae consigo un aspecto negativo, el cual es la baja capacidad predictora de los árboles. Normalmente, la explicabilidad y capacidad predictiva de un modelo suelen ser inversamente proporcionales. Debido a ello, un punto muy importante a la hora de decidir qué modelo utilizar para resolver nuestro problema es saber el equilibrio que deseamos obtener entre estos dos aspectos. Como hemos indicado en el apartado de resultados, este modelo, sin realizar una búsqueda exhaustiva de hiperparámetros, ha obtenido una *accuracy* del 73% en el conjunto de entrenamiento y del 69% en el conjunto de prueba. Estos valores no son muy altos, contrastando completamente con los altos niveles de explicabilidad del modelo. Debido a ello, consideramos que sólo se debería plantear usar árboles de decisión para resolver nuestro problema si la principal prioridad es entender por qué se toman las decisiones y no tanto el resultado del modelo. Esto puede concordar con el problema que abordamos, ya que se basa en decidir si una persona tiene un alto o bajo riesgo. Si la decisión que tomamos puede influenciar la vida de una persona, es importante poder explicar por qué la hemos tomado. Por el contrario, tampoco sería correcto basarnos en los resultados producidos por un modelo con baja capacidad predictora. Es por esto por lo que, en la versión futura de la práctica, abordaremos distintos algoritmos de caja blanca para analizar si estos ofrecen un mejor equilibrio entre explicabilidad y resultados.

Como sabemos, los árboles de decisión presentan ciertos problemas como la dificultad para encontrar relaciones aditivas entre las variables o que pequeños cambios en las variables pueden provocar grandes variaciones en las predicciones. Actualmente, no sabemos si estos problemas tienen un gran impacto en el modelo que hemos creado. Implementar distintos modelos que no presenten estos inconvenientes pueden ayudarnos a mejorar los resultados en caso de que estos tengan un gran impacto en el modelo final.

El árbol obtenido no es óptimo, esto puede influenciar el rendimiento del modelo. La utilización de un árbol óptimo puede llegar a igualar los resultados obtenidos por conjuntos de árboles, sin perder la explicabilidad que estos últimos no ofrecen. Debido a ello, podría ser interesante estudiar la implementación de un árbol óptimo para observar si mejora el rendimiento del modelo.

Si transformamos el árbol en reglas se podría realizar alguna simplificación. Un ejemplo es el primer subárbol de la izquierda. Esto es debido a que tanto en el nodo raíz como en el primer nodo del subárbol se evalúa la misma variable: ExternalRiskEstimate. Esto implica que juntando ambas reglas podríamos realizar una simplificación.
Para mejorar la capacidad predictora del modelo se podría realizar un estudio exhaustivo de los hiperparámetros. En este punto de la práctica se ha realizado una versión simplificada utilizando distintos modelos entrenados con el conjunto de entrenamiento y evaluados con un conjunto que actuaba a la vez como test y validación, obtenido del archivo “validation.csv”. Para la versión final se realizará este procedimiento utilizando tres conjuntos independientes (train, validation y test). Los hiperparámetros a estudiar serán profundidad, número mínimo de individuos en las hojas y criterio de separación (gini, entropy, log_loss). Mediante este estudio exhaustivo se espera obtener una mejora

significativa en la capacidad predictora.

6. Conclusiones

En este trabajo se ha utilizado un clasificador de árbol de decisiones para tratar un problema de clasificación binaria. La evaluación del modelo muestra un amplio rango de mejora en las métricas de confusión, sin embargo, este no deja de ser un modelo atractivo por su facilidad de uso, robustez ante datos categóricos y no normalizados, y sobre todo, su interpretabilidad inherente, que permite observar fácilmente tanto las reglas que conforman el modelo como el razonamiento seguido para cada una de las predicciones realizadas.