PROYECTO FINAL XAI

Máster en Inteligencia Artificial

Jimena Monteagudo Ruiz

Diciembre 2025

Introducción

En los últimos tiempos, ha aumentado el uso de modelos de machine learning y de inteligencia artificial para realizar la toma de decisiones en las empresas. Estos también se utilizan en muchos otros ámbitos como la medicina, las finanzas o incluso la política. Aunque muchas veces los resultados que se obtienen de estos modelos son buenos, en ocasiones no se puede entender cómo llegan a tomar sus decisiones. Esto puede llevar a desconfianza tanto de la persona que lo está usando como del cliente al que se le aplica, además de que no se llega a entender bien si el modelo tiene ciertos sesgos o si depende demasiado de algunas variables. Es por esto por lo que las técnicas de explicabilidad cobran una importancia especial.

Asimismo, para los stakeholders la explicabilidad es un tema que también resulta muy relevante. Las personas a las que afecta una predicción deben tener el derecho de entender cómo y por qué se ha tomado esa decisión. Por otro lado, las personas que crean el modelo deben conocer su modelo y comprender cómo funciona.

Mi motivación para este proyecto es construir un modelo predictivo y entender cómo ese modelo realiza las predicciones, ver si tiene algún sesgo y poder explicar cómo fuciona realmente. Además, he empleado un dataset de los pasajeros que viajaban en el Titanic, que es un tema que me ha interesado siempre, por lo que realizando este proyecto de explicabilidad podré ver qué variables influyen más o menos a la hora de predecir la supervivencia de un pasajero y por lo tanto ver qué factores influyeron en aquel momento a la hora de salvar a distintas personas.

Para llevar a cabo estos objetivos, se ha entrenado un modelo de Random Forest. A continuación, se han aplicado una serie de técnicas de explicabilidad tanto locales como globales y algunos sanity checks para evaluar la fiabilidad de las explicaciones. Todo ello se desarrollará a continuación.

Datos y Métodos

Dataset

Como ya he mencionado antes, para este proyecto se ha empleado un dataset del Titanic, disponible públicamente desde la librería de Seaborn y perfecto para realizar tareas de clasificación, como pretendíamos en este caso. El objetivo del problema es predecir si un pasajero sobrevivió o no al naufragio a partir de un conjunto de variables. Aunque el dataset no es excesivamente grande, he considerado que resulta adecuado para el análisis de explicabilidad, ya que se pueden interpretar las variables con facilidad y contrastar los resultados del modelo con lo que sucedió realmente en aquella época.

Antes de empezar, se han seleccionado una serie de variables para trabajar con ellas, de manera que todas fueran numéricas, ya que de esta manera podemos realizar un preprocesado de los datos más sencillo y realizar unas explicaciones mucho más claras. Entre ellas, destacan las variables acerca del género, la clase en la que viajaba el pasajero, el precio del billete o la presencia de otros familiares a bordo.

El conjunto de datos se ha dividido en subconjuntos de entrenamiento, validación y test, manteniendo la proporción de clases. Esta separación permite evaluar tanto el rendimiento del modelo como la estabilidad de las explicaciones en datos no vistos durante el entrenamiento.

Model

El modelo de predicción que se ha empleado es un Random Forest Classifier que es una técnica basada en un conjunto de árboles de decisión, estos capturan relaciones no lineales entre los datos e interacciones entre las variables de una forma relativamente sencilla. El modelo intenta predecir si un pasajero sobrevivió o no al Titanic en base a una serie de datos de los mismos.

El rendimiento del modelo se ha evaluado con la métrica ROC-AUC. Representa el área bajo la curva ROC, que relaciona la tasa de verdaderos positivos con la tasa de falsos positivos. Un valor de AUC cercano a 0,5 indica un comportamiento del modelo similar al azar. Con ello podemos ver la diferencia entre rendimiento en entrenamiento y validación para descartar un posible overfitting en el modelo.

Técnicas XAI

Para analizar el comportamiento del modelo se han aplicado varias técnicas de explicabilidad que nos dan explicaciones globales, locales y sanity checks que explicaremos a continuación.

En primer lugar, se han utilizado técnicas de explicabilidad global que han consistido en aplicar un SHAP al modelo. Se ha analizado la contribución media de cada variable a la predicción de supervivencia, así como la dirección de su efecto. Se han realizado representaciones con gráficos de bar plot y beeswarm que mostraré en los siguientes apartados, que nos permiten entender qué factores influyen más en el comportamiento general del modelo.

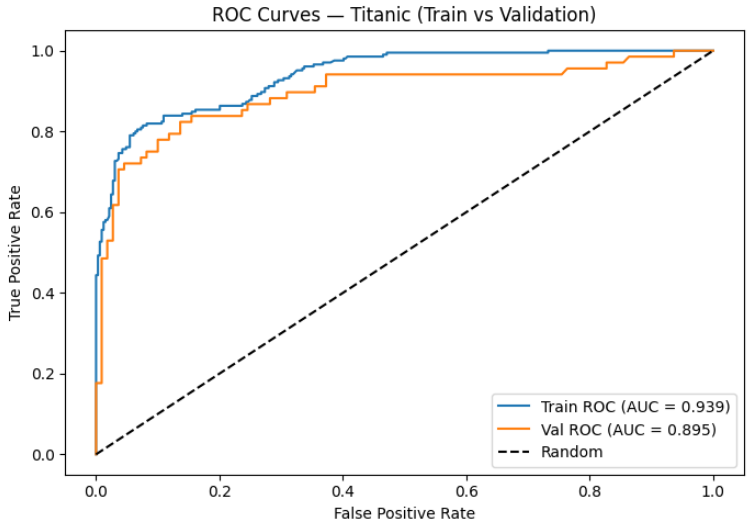
En segundo lugar, se han aplicado técnicas de explicabilidad local para ver cuánto han influido las variables a la hora de realizar la predicción para un pasajero en concreto. En este caso, he empleado dos técnicas para poder comparar sus resultados. La primera de ellas es LIME, nos permite explicar decisiones concretas del modelo mediante aproximaciones locales interpretables. La segunda técnica aplicada es SHAP una vez más, pero de manera local, que nos permite estimar de forma exacta la contribución de cada variable en una predicción específica. Estas explicaciones locales resultan especialmente útiles para analizar casos concretos y comparar decisiones entre pasajeros.

Finalmente, se han llevado a cabo distintos sanity checks con el objetivo de evaluar la fiabilidad de las explicaciones obtenidas. En particular, se entrenaron modelos con etiquetas aleatorias para comprobar que las explicaciones no dan buenos resultados al tener etiquetas falsas y se eliminaron algunas de las variables consideradas importantes para observar el impacto en el rendimiento del modelo. Estos experimentos permiten verificar que las explicaciones reflejan patrones reales aprendidos por el modelo.

Resultados

Model Performance

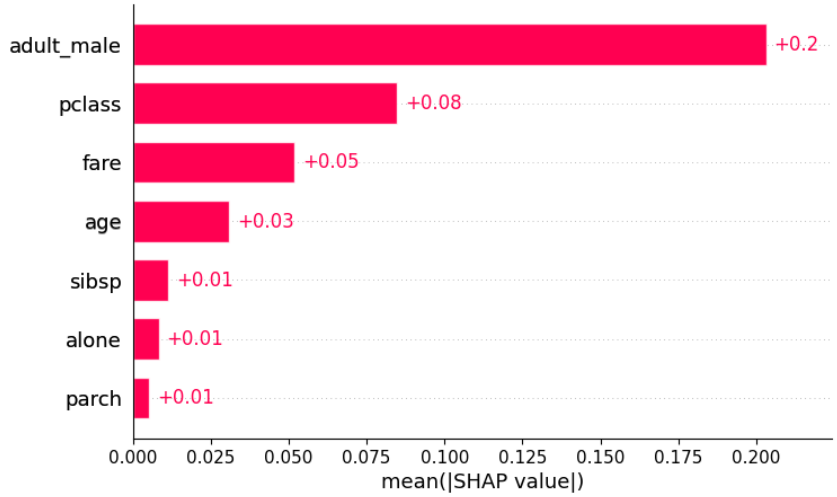
El modelo basado en Random Forest ha obtenido unos resultados satisfactorios y precisos en su tarea de clasificación. La métrica ROC-AUC obtenida fue alta tanto en el conjunto de train como en el de test, y no hay una diferencia demasiado grande entre ambos.



Como se puede ver en la gráfica, aunque el valor de train siempre está ligeramente por encima del de validación, la diferencia entre sus puntuaciones AUC no es demasiado grande, por lo que se puede afirmar que el modelo no tiene overfitting y por lo tanto el modelo no está simplemente aprendiendo patrones en los datos de train.

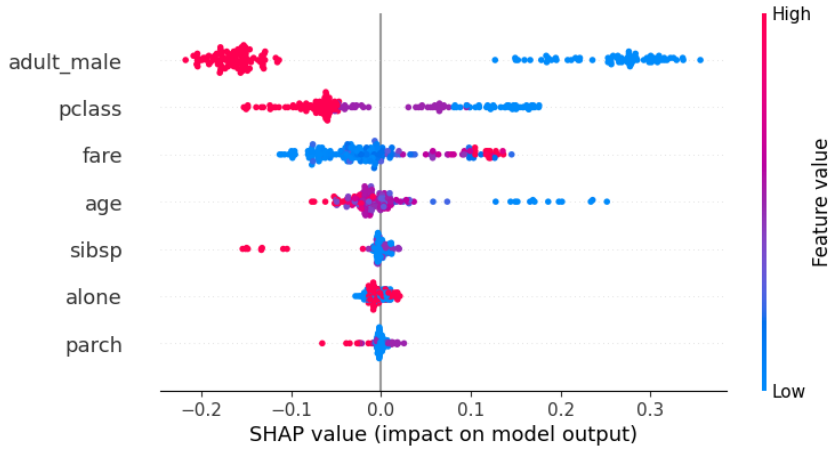
Explicaciones globales

Para analizar el comportamiento global del modelo se emplearon explicaciones basadas en SHAP. En primer lugar, se calculó la importancia media de cada variable mediante un gráfico de barras, que permite identificar qué características influyen más en la predicción de supervivencia.



Los resultados muestran que variables como adult_male, pclass y fare tienen un peso claramente mayor que el resto a la hora de tomar las decisiones. En concreto la variable adult_male es la que mayor peso tiene, por lo que ser un hombre adulto influyó mucho a la hora de saber si se iba a sobrevivir o no.

Pero para saber de qué manera influyó cada variable, es decir, en qué dirección empuja la predicción un valor u otro, se ha empleado un gráfico beeswarm.

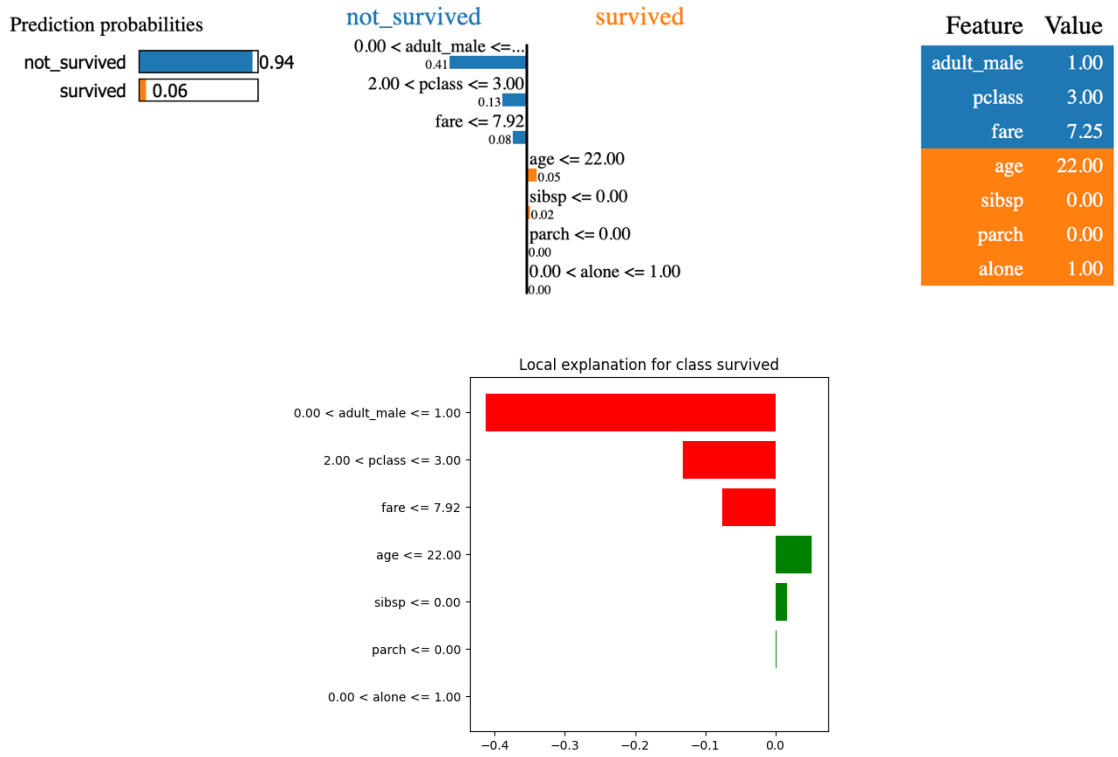


De esta manera podemos observar que efectivamente la variable adult_male es la que más influye a la hora de realizar las predicciones. Esta variable es binaria, con un 1 en el caso de ser un hombre adulto y con un 0 en el caso de no serlo. Podemos ver que si eras un hombre adulto tenías menos probabilidades de sobrevivir que si no lo eras, esto se puede deber a la política de priorizar a mujeres y niños a la hora de salvarse. También podemos ver que la variable de en qué clase social estabas afectaba a la hora de saber si ibas a sobrevivir, ya que aquellos pasajeros con peor clase (número de clase más alto) tenían menos probabilidades de salvarse que aquellos de mejores clases. El precio del billete también ha afectado, ya que aquellos billetes de menor coste tenían menos probabilidades de supervivencia que aquellos más caros.

Explicaciones locales

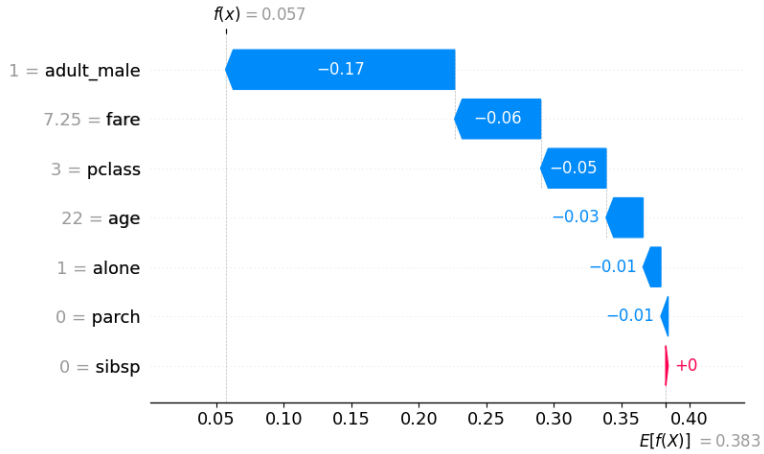
Además del análisis global, se han estudiado explicaciones locales para comprender decisiones individuales del modelo. En primer lugar, se ha aplicado la técnica LIME para explicar la predicción correspondiente a un pasajero concreto del conjunto de validación. LIME aproxima el comportamiento del modelo en el entorno local de ese dato y proporciona una explicación interpretable.

He elegido el siguiente pasajero, en la imagen podemos ver el valor de sus variables y su probabilidad de supervivencia.



Como podemos observar, este pasajero no sobreviviría y las principales variables que han influido en su caso concreto han sido el hecho de que sea un hombre adulto y que viajara en tercera clase. Esto coincide con lo que hemos visto en el SHAP aplicado al modelo globalmente.

Pero hagamos otra prueba. En este caso hemos aplicado SHAP localmente analizando los datos del mismo pasajero para ver si coinciden.



En este caso podemos ver que, para el mismo pasajero, las tres primeras variables que influyen a su no supervivencia obtenidas con LIME son las mismas que las obtenidas con SHAP, por lo que nos queda claro que son las más relevantes. Sin embargo, obtenemos algunos resultados distintos para el resto de variables, esto se debe a que ambos métodos calculan la importancia de las variables de manera diferente, por lo que para aquellas que no son muy significativas, los resultados pueden variar.

Sanity Checks

Finalmente, con el objetivo de evaluar la fiabilidad de las explicaciones, se realizaron varios sanity checks. En primer lugar, se entrenó el mismo modelo utilizando etiquetas aleatorias. Los valores obtenidos para el análisis de SHAP y de la curva ROC-AUC han sido los siguientes.

```
Random-labels AUC  train=0.824  val(real y)=0.379
ExactExplainer explainer: 179it [00:15,  4.24it/s]
Top SHAP (random labels):


fare          0.040865
alone         0.030491
adult_male    0.023269
age           0.019452
pclass        0.011283
parch         0.010110
sibsp         0.006803
dtype: float64
```

El modelo ha sido capaz de ajustarse a estas etiquetas aleatorias en el conjunto de train (AUC = 0.824), pero podemos ver que no tiene ninguna capacidad de predicción real a la hora de evaluarlo sobre el conjunto de validación (AUC = 0.379) como era de esperar. Además, al analizar las explicaciones SHAP correspondientes, las importancias de las variables se vuelven más planas y menos estructuradas, con valores similares y más pequeños para todas las variables. Esto demuestra que, al introducir etiquetas aleatorias, las explicaciones dejan de tener sentido y representan básicamente ruido. Por lo que podríamos decir que originalmente las explicaciones del modelo original se basan en patrones reales y explican el modelo.

En segundo lugar, se ha eliminado la variable adult_male, ya que es la que SHAP ha considerado más importante globalmente para ver qué impacto genera en el modelo.

```
BASE Val AUC: 0.895
DROP 'adult_male' Val AUC: 0.781
Δ Val AUC: -0.114
```

Como podemos ver, al eliminar la variable, el modelo ha empeorado notablemente en los datos de validación, lo que nos demuestra que esta variable por sí misma jugaba un papel importante a la hora de realizar la predicción. Por lo que las explicaciones realizadas anteriormente en las que hemos obtenido adult_male como la variable más significativa, estaban en lo cierto.

Acciones e Insights

En cuanto a la mejora del modelo, se ha demostrado a lo largo de este análisis una fuerte dependencia de la variable adult_male, ya que es la variable más importante en todas las explicaciones. Aunque este resultado es coherente con el contexto histórico del hundimiento del Titanic, también muestra el riesgo de que el modelo dependa demasiado de una única variable. Es por esto que como futura mejora del modelo se podría considerar un entrenamiento de otros modelos alternativos que no dependan tanto de una única variable. Cuando he realizado el sanity check de eliminar adult_male, la fiabilidad del modelo ha descendido notablemente, pero todavía podría funcionar, por lo que reuniendo otro tipo de datos sobre cada pasajero y combinándolos con las otras variables relevantes que ya teníamos podríamos obtener un modelo que no sea tan dependiente de un solo dato.

Desde el punto de vista del dominio, las explicaciones confirman patrones que ya se conocían acerca de lo que sucedió en el Titanic. Como ya se sabía históricamente, en la supervivencia o no de una persona influyeron factores como su estatus social o la vulnerabilidad. Esto se ha podido comprobar a través de las explicaciones que el precio del billete, la clase en la que se viajaba y el hecho de ser un hombre adulto eran los factores que más influían a la hora de salvarse. Las explicaciones locales obtenidas con SHAP y LIME permiten además analizar casos individuales, mostrando cómo distintas combinaciones de características influyen en decisiones concretas del modelo. Todo ello resulta útil desde un enfoque educativo y analítico, ya que hace que el modelo sea una herramienta que nos permita analizar algunas dinámicas que ocurrían en la historia según las costumbres, o los movimientos sociales de la época.

Por todo esto podemos ver que mediante las técnicas de explicabilidad podemos ver el comportamiento del modelo, ver las dependencias entre variables e incluso emplear un modelo de manera educativa para explicar por qué han sucedido distintos eventos históricos.

Discussion

A pesar de los resultados obtenidos, este trabajo presenta varias limitaciones que conviene tener en cuenta a la hora de interpretar las conclusiones. En primer lugar, el dataset

utilizado es relativamente pequeño y está condicionado por un contexto histórico específico. Aunque el conjunto de datos del Titanic es utilizado con fines educativos, no es un escenario realista a día de hoy y es difícilmente aplicable de manera que se pueda obtener un beneficio económico de ello.

Otra limitación importante es el tipo de variables que se emplean. Algunas de ellas combinan información demográfica y de género, por lo que se debe tener en cuenta a la hora de introducir cuestiones de equidad y de sesgo. Aunque el análisis de explicabilidad ha permitido identificar esta dependencia y evaluar su impacto mediante sanity checks, el modelo sigue teniendo sesgos que ya estaban en los datos originales. Por lo que podemos seguir teniendo un problema ético, ya que aunque se pueda explicar una decisión, no tiene por qué ser éticamente correcta.

Por otro lado, las técnicas de explicabilidad que hemos empleado (SHAP y LIME) también presentan limitaciones. SHAP ofrece explicaciones consistentes y teóricamente fundamentadas, pero puede ser costoso computacionalmente y sensible a la elección del conjunto de referencia. Por su parte, LIME depende de aproximaciones locales y de decisiones como el muestreo o la discretización de variables, lo que puede dar lugar a explicaciones distintas para un mismo ejemplo. Además, las diferencias que hemos visto entre ambas técnicas muestran la importancia de no confiar en un único método de explicabilidad.

Como líneas de trabajo futuro, sería interesante ampliar el análisis incorporando métricas explícitas de equidad y comparando el comportamiento del modelo entre distintos subgrupos de la población. Además, podrían estudiarse otros modelos sin variables sensibles y evaluar la relación entre la precisión y la interpretabilidad. Finalmente, aplicar estas mismas técnicas de explicabilidad y sanity checks a un problema con una aplicabilidad más real permitiría ver cómo se puede utilizar la explicabilidad en casos de la vida cotidiana.