Matias Davila
Aprendizaje automático

UNIVERSITAT Carlemany

Tema 8 Datos del Titanic

Matias Davila
Aprendizaje automático

UNIVERSITAT Carlemany

1. Carga y exploración inicial de los datos

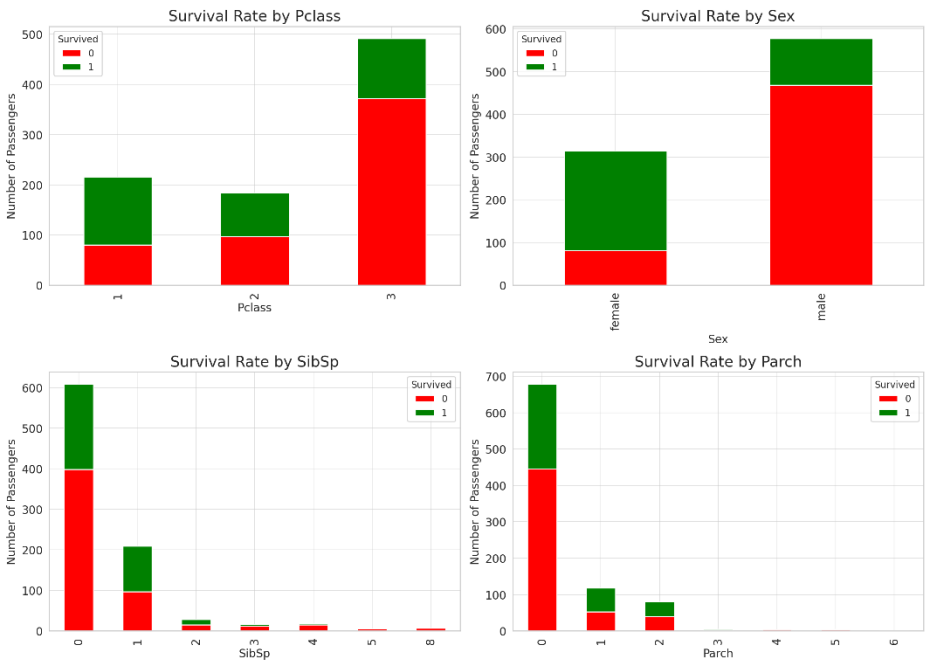
Comenzamos este trabajo cargando el conjunto de datos del Titanic desde un archivo csv. Una vez cargados, visualizamos las primeras filas con head () esto con el propósito de entender con qué tipo de datos estamos trabajando y la estructura que este conjunto de datos tiene. Los datos incluyen detalles sobre los pasajeros del Titanic, como su edad, sexo, clase de boleto, entre otros.

Resultados: El conjunto de datos tiene varias columnas, entre las cuales se encuentran: **PassengerId**, **Survived**, **Pclass**, **Name**, **Sex**, **Age**, **SibSp**, **Parch**, **Ticket**, **Fare**, **Cabin** y **Embarked**. Cada fila representa a un pasajero y sus detalles.

2. Análisis de las características

Luego se realizará un análisis exploratorio de los datos para poder generarnos una mejor idea de la relación entre las características de los pasajeros y su tasa de supervivencia. Utilizamos gráficos de barras ya que al visualizar la distribución de la supervivencia según diversas características es mucho más sencillo comprender.

Resultados: Observamos que las mujeres tenían una mayor probabilidad de supervivencia que los hombres esto puede ser debido a que fueron prioridad en botes salvavidas. Además, los pasajeros de primera y segunda clase tuvieron tasas de supervivencia más altas en comparación con la tercera clase, pudo haber sido por posición de sus recamaras y la cercanía a los botes o quizás se les dio preferencia simplemente. También notamos que el número de familiares a bordo (ya sean hermanos, cónyuges, padres o hijos) influyó en la tasa de supervivencia.



3. Construcción del modelo usando RandomForest

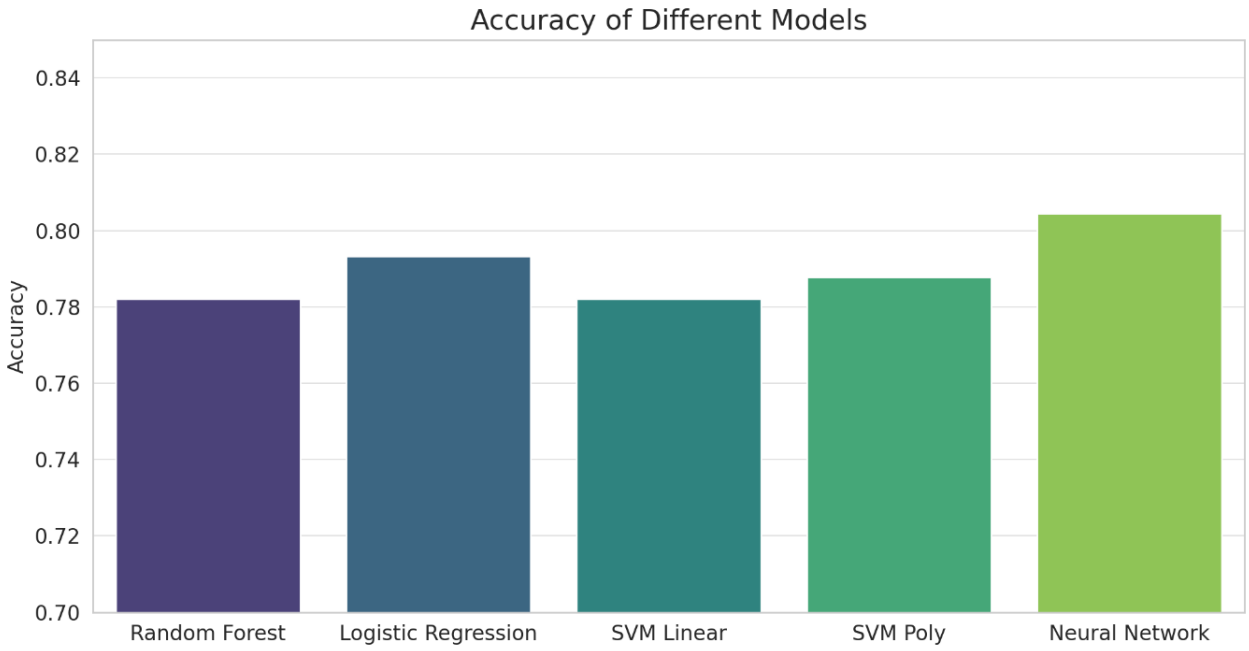
Usamos el algoritmo RandomForest, que es un ensamblado de árboles de decisión, para predecir si un pasajero sobreviviría o no basándonos en sus características. Antes de entrenar el modelo, preprocesamos los datos codificando características categóricas y dividimos el conjunto de datos en entrenamiento y prueba, tal y como lo hicimos en trabajos anteriores es importante esta separación de datos de prueba y entrenamiento para la precisión ya que vimos casos en los que si estos se modelan mal puede fallar precisión.

Resultados: El modelo de RandomForest logró una precisión de aproximadamente 78.21%.

4. Comparación de diferentes modelos

Además de RandomForest, entrenamos otros modelos, incluyendo Regresión Logística, Máquina de Vectores de Soporte (SVM) y Red Neuronal, para comparar su rendimiento en la precisión de la predicción de los datos según las características que elegimos.

Resultados: La Red Neuronal fue el modelo más preciso con una precisión de 80.45%, en segundo lugar, no encontramos que la Regresión Logística obtuvo un 79.33%. Los modelos SVM mostraron un rendimiento similar con precisiones en el rango del 78%.



5. Explicación de cada modelo

Regresión logística: Cuando comencé a explorar la regresión logística, me di cuenta de que es como un detective que intenta estimar las probabilidades. Este modelo intenta predecir la probabilidad de que algo pertenezca a una categoría específica. Es como preguntarse: "Dadas ciertas características, ¿cuál es la probabilidad de que este pasajero sobreviva?".

Máquina de vectores de soporte (SVM): SVM busca la mejor manera, o línea, para dividir y clasificar nuestros datos. Es como dibujar una línea en la arena y decir: "De este lado están los supervivientes y del otro lado los no supervivientes".

Máquina de vectores de soporte con un núcleo polinómico: Al profundizar en SVM, descubrí una versión más avanzada que utiliza lo que llaman un "truco de kernel". Puede sonar complicado, pero es básicamente una forma inteligente de transformar nuestros datos para que esa línea en la arena se adapte aún mejor. Imagina poder doblar y torcer la línea para que se ajuste perfectamente a los puntos.

Red neuronal: La red neuronal fue, para mí, como entrar en el mundo de la ciencia ficción. Inspirada en cómo funciona nuestro cerebro, esta técnica utiliza "neuronas" que procesan información en diferentes etapas. Es como si tuviera un mini cerebro en mi computadora, analizando y aprendiendo de los datos del Titanic para hacer sus predicciones.