IES Pere Maria Orts

Sistemas de Aprendizaje Automático

Práctica 2: Clasificación de sistemas de aprendizaje automático

---

Autor:
Kenny Berrones

Profesor:
David Campoy Miñarro

iesperemariaorts

GENERALITAT
VALENCIANA

Índice

1. Introducción

Esta práctica se centra en explorar los distintos tipos de aprendizaje automático mediante la implementación de modelos específicos y el análisis de sus resultados. Los objetivos son entender y aplicar técnicas de:

Aprendizaje supervisado (regresión y clasificación), utilizando conjuntos de datos estructurados con etiquetas para que el modelo pueda aprender a predecir o clasificar valores desconocidos. Aprendizaje no supervisado (clustering), con el cual se agrupan datos no etiquetados en categorías o segmentos, según patrones subyacentes. Aprendizaje semi-supervisado, una combinación de los dos métodos anteriores, donde se utilizan datos parcialmente etiquetados para mejorar la precisión en tareas de clasificación.

Durante la práctica, se trabajará con datasets reales, como el de precios de viviendas en Boston, el de supervivencia en el Titanic y el de clientes de un centro comercial, y se evaluarán distintos parámetros y métricas para entender la eficacia de los modelos. Al final, se documentarán las pruebas, resultados y conclusiones obtenidas para cada modelo.

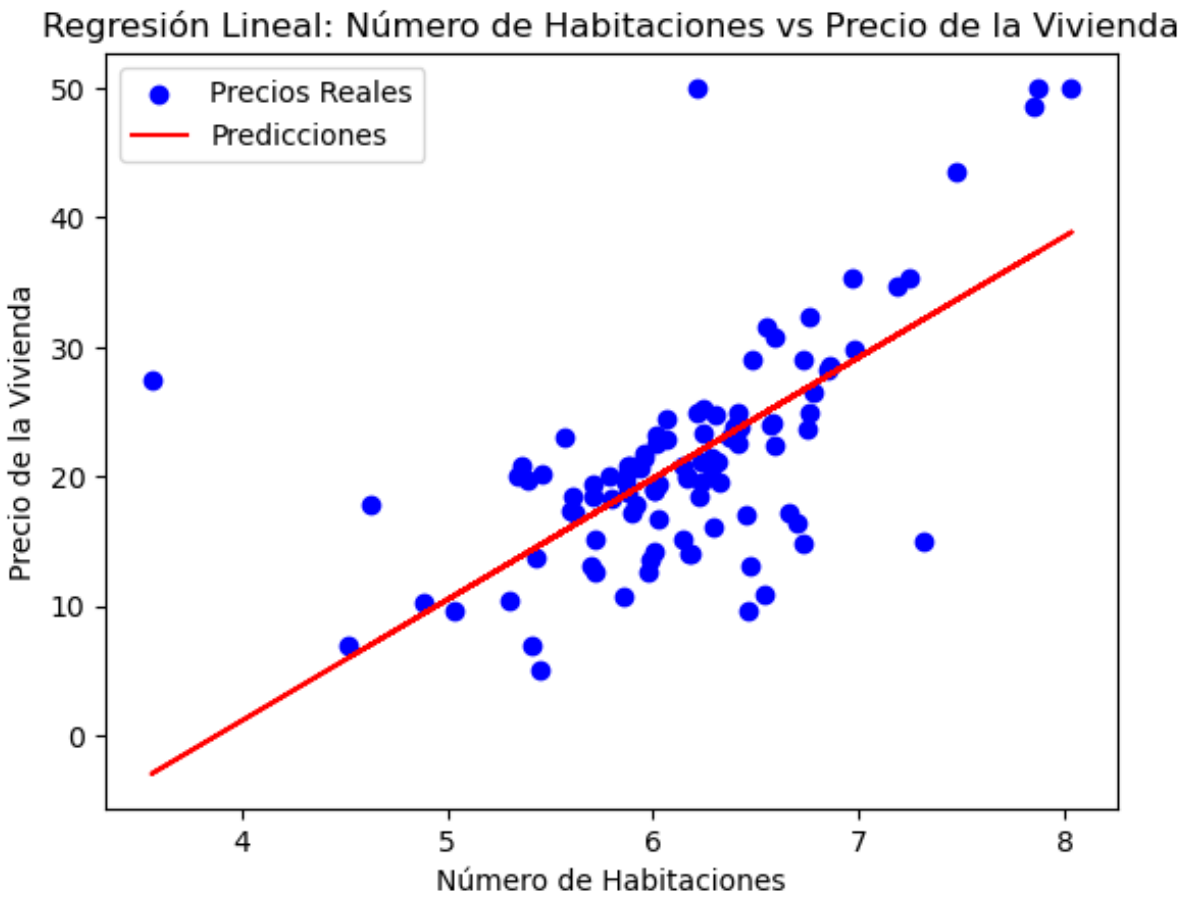
2. Aprendizaje automático supervisado: Regresión

La idea es que en base a unos datos ya existentes podamos predecir un valor para datos nuevos. En este caso vamos a usar el dataset Boston Houses Prices, este cuenta con distintos parámetros que influyen más o menos para el precio de una vivienda.

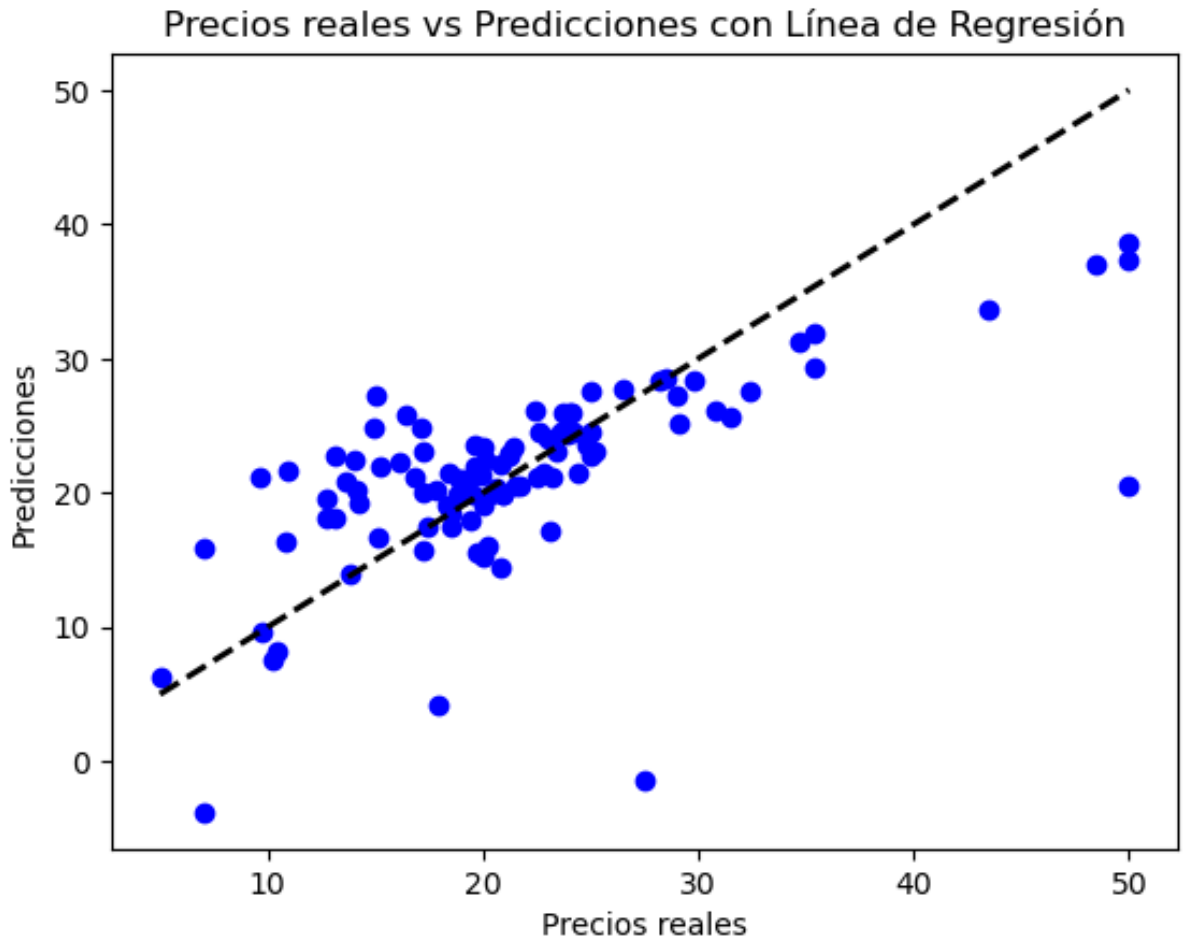
CRIM: Tasa de criminalidad per cápita por área.
ZN: Proporción de terreno residencial dividido en zonas para lotes mayores a 25,000 pies cuadrados.
INDUS: Proporción de acres de negocios no minoristas por ciudad.
CHAS: Variable ficticia de Charles River (= 1 si el tramo limita con el río; 0 en caso contrario).
NOX: Concentración de óxidos nítricos (partes por 10 millones).
RM: Número medio de habitaciones por vivienda.
AGE: Proporción de unidades ocupadas por sus propietarios construidas antes de 1940.
DIS: Distancias ponderadas a cinco centros de empleo de Boston.
RAD: Índice de accesibilidad a carreteras radiales.
TAX: Tasa de impuesto a la propiedad por $10,000.
PTRATIO: Proporción alumno-maestro por localidad.
B: Proporción de personas de ascendencia afroamericana por ciudad.
LSTAT: Porcentaje de población de estatus bajo.
MEDV: Valor medio de las viviendas ocupadas por sus propietarios en miles de dólares.

Figura 1: Características del Dataset de Boston

Tras ejecutar el primer código obtenemos el siguiente gráfico:



Tras esto vamos a ejecutar el segundo código proporcionado y obtenemos la relación entre los precios reales y las predicciones que ha hecho el modelo de regresión lineal, obtenemos lo siguiente:



En la gráfica anterior hemos relacionado el precio de las viviendas con las características de **tasa de criminalidad** y el **número de habitaciones**. Tras esto vamos a calcular el coeficiente de determinación, este parte del coeficiente de correlación de Pearson que se trata de un valor que oscila entre -1 o 1 y nos indica como las variables influyen en las demás una a una, por ejemplo, como influye la zona de la ciudad para el precio de una habitación de un hotel. Por su parte, el **coeficiente de determinación** nos indica el comportamiento de una variable en función de otras, como podría ser nuestro caso, el como influye el número de habitaciones en el precio de una casa en Boston.

Una vez hemos entendido esto, hemos probado con distintas variables para calcular este coeficiente, hemos obtenido los siguientes resultados:

```
Características:
['CRIM', 'RM', 'LSTAT'], Coeficiente de determinación (R²): 0.56
['CRIM', 'RM', 'LSTAT', 'PTRATIO'], Coeficiente de determinación (R²): 0.62
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX'], Coeficiente de determinación (R²): 0.62
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS'], Coeficiente de determinación (R²): 0.64
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS', 'TAX'], Coeficiente de determinación (R²): 0.64
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS', 'TAX', 'RAD'], Coeficiente de determinación (R²): 0.67
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS', 'TAX', 'RAD', 'ZN'], Coeficiente de determinación (R²): 0.69
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS', 'TAX', 'RAD', 'ZN', 'INDUS'], Coeficiente de determinación (R²): 0.69
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS', 'TAX', 'RAD', 'ZN', 'INDUS', 'CHAS'], Coeficiente de determinación (R²): 0.69
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS', 'TAX', 'RAD', 'ZN', 'INDUS', 'CHAS', 'AGE'], Coeficiente de determinación (R²): 0.69
['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS', 'TAX', 'RAD', 'ZN', 'INDUS', 'CHAS', 'AGE', 'B'], Coeficiente de determinación (R²): 0.67
```

Figura 2: Coeficientes de determinación para distintas características

Como se puede apreciar en la imagen anterior, obtenemos distintos valores, vemos que

uno que da un coeficiente de 0.69 con el menor número de variables es la combinación de:

`['CRIM', 'RM', 'LSTAT', 'PTRATIO', 'NOX', 'DIS', 'TAX', 'RAD', 'ZN']`

La idea de este coeficiente es que sea lo más cercano a 1, que indicará que la capacidad del modelo para predecir los precios de las viviendas será mejor.

3. Aprendizaje automático supervisado Clasificación

A grandes rasgos, se entiende la clasificación como la separación en n clases objetivo. En nuestro caso vamos a utilizar el dataset de los fallecidos y de los supervivientes del Titanic, este dataset cuenta con distintas variables que influyen en si una persona hubiese sobrevivido o no. Tras ejecutar el código proporcionado obtenemos la siguiente matriz de confusión:

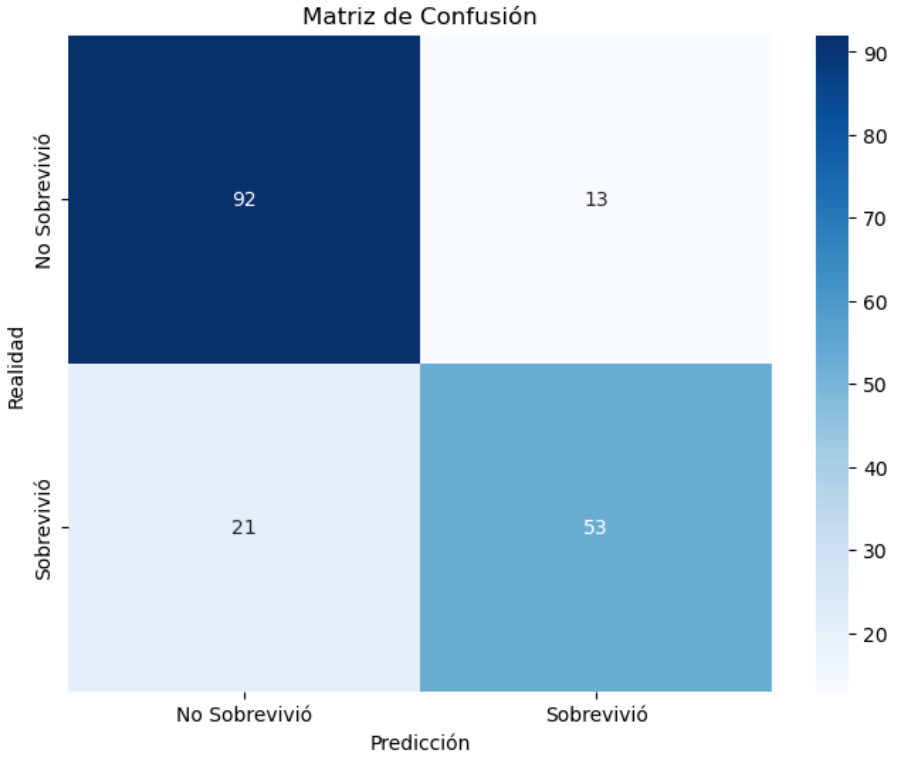
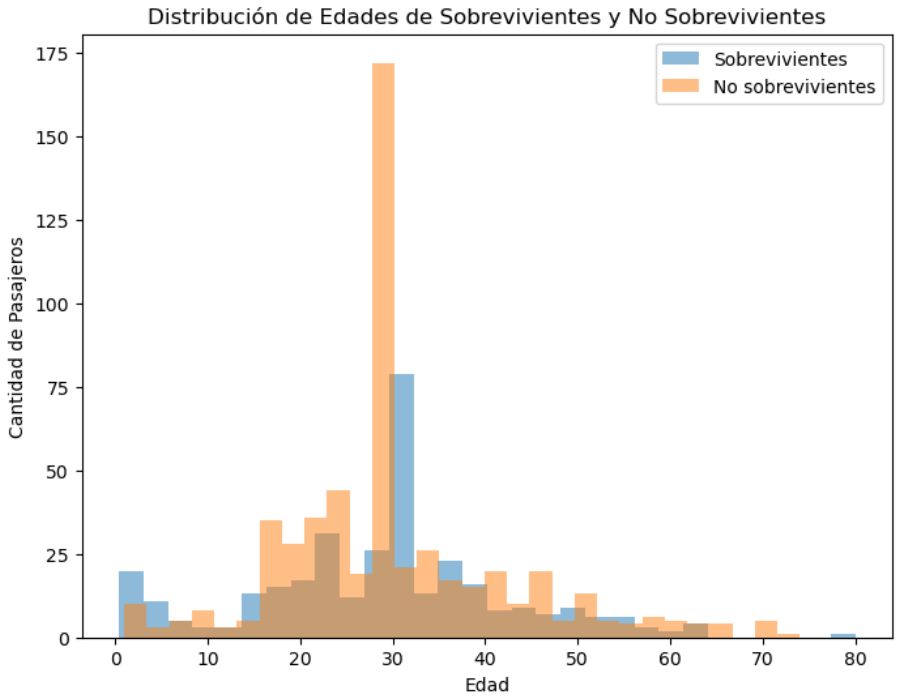


Figura 3: Matriz de confusión para el aprendizaje supervisado: Clasificación

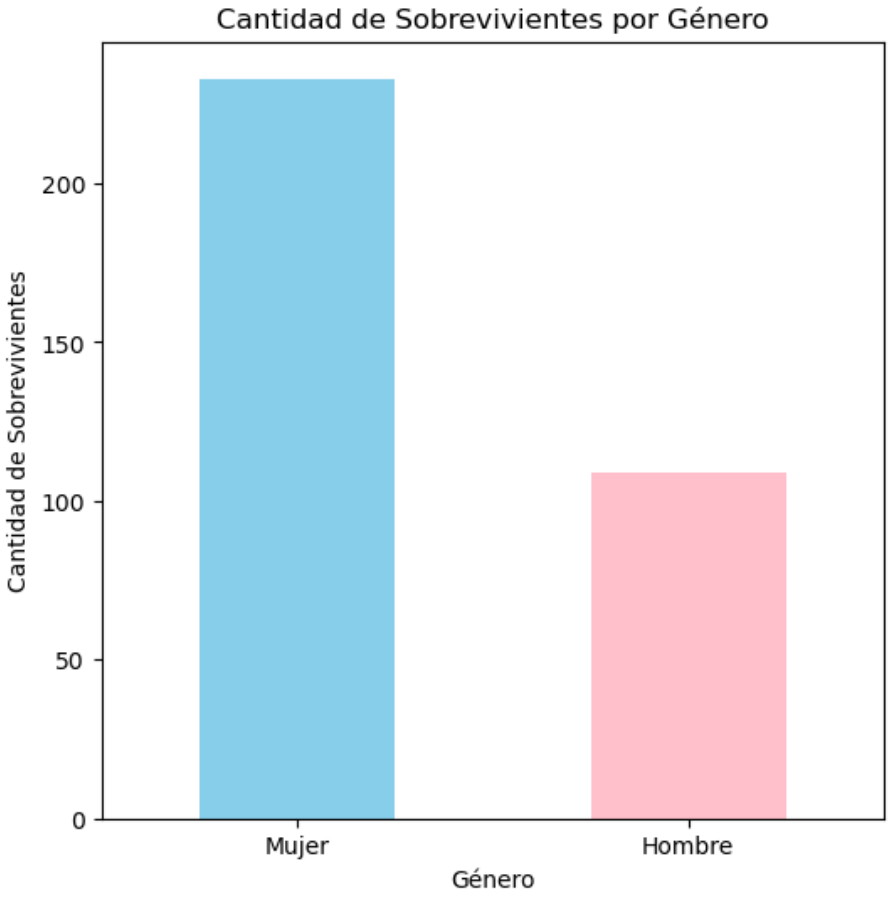
Si observamos la matriz de confusión, vemos que ha acertado en la mayoría de casos, concretamente en un 81 % de los casos.

Ahora vamos a ver como influye en la edad en la categorización, para ello tenemos la siguiente gráfica:



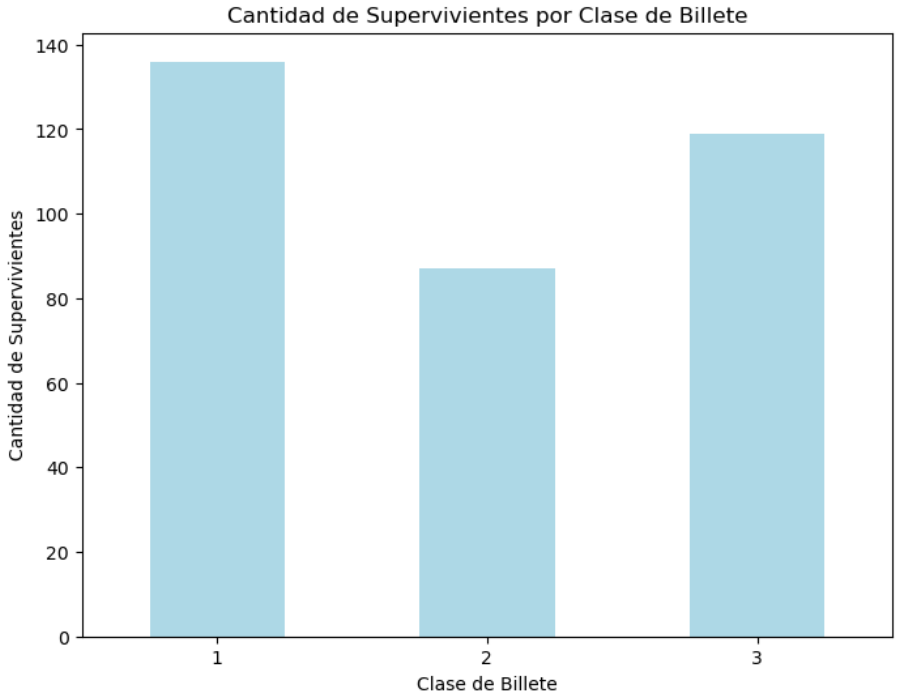
En la gráfica se puede observar que la edad sigue más o menos una distribución normal, teniendo un pico de tanto supervivientes como para lo no supervivientes en el rango de edad de 27 a 32 años.

Ahora vamos a ver como influyó el género en los sobrevivientes, en la siguiente gráfica se observa:

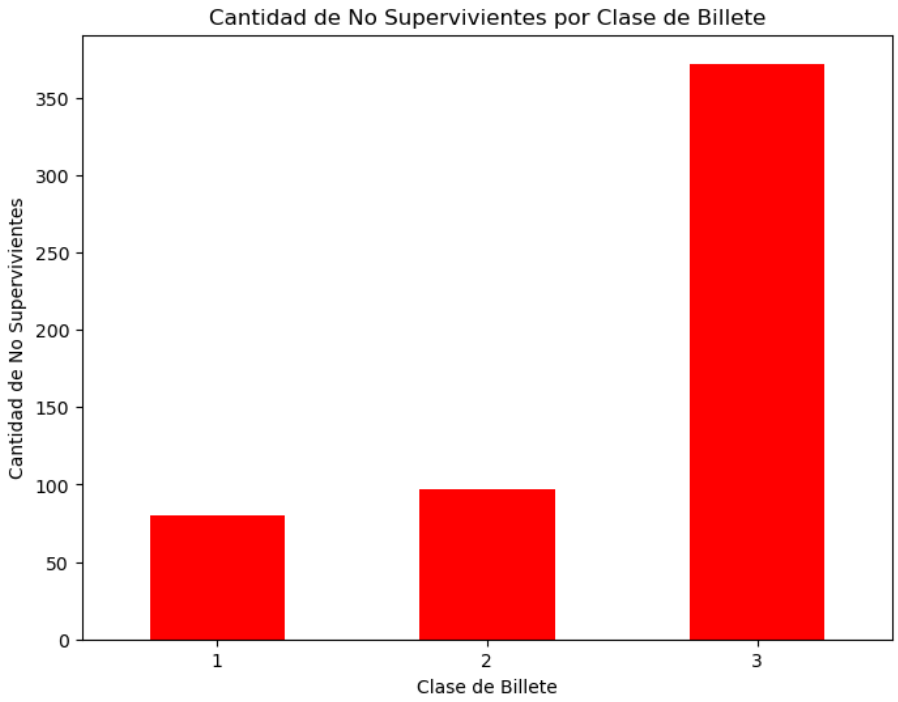


En la gráfica anterior se observa que la mayoría de supervivientes fueron mujeres, esto se debe a la "política" que se sigue a la hora de un accidente en alta mar, ya que primero van los niños, las mujeres, ancianos o personas con discapacidad.

Otra prueba que hemos realizado es el precio del billete, podríamos decir que esta variable refleja la clase socioeconómica, lo apreciamos en la siguiente gráfica:

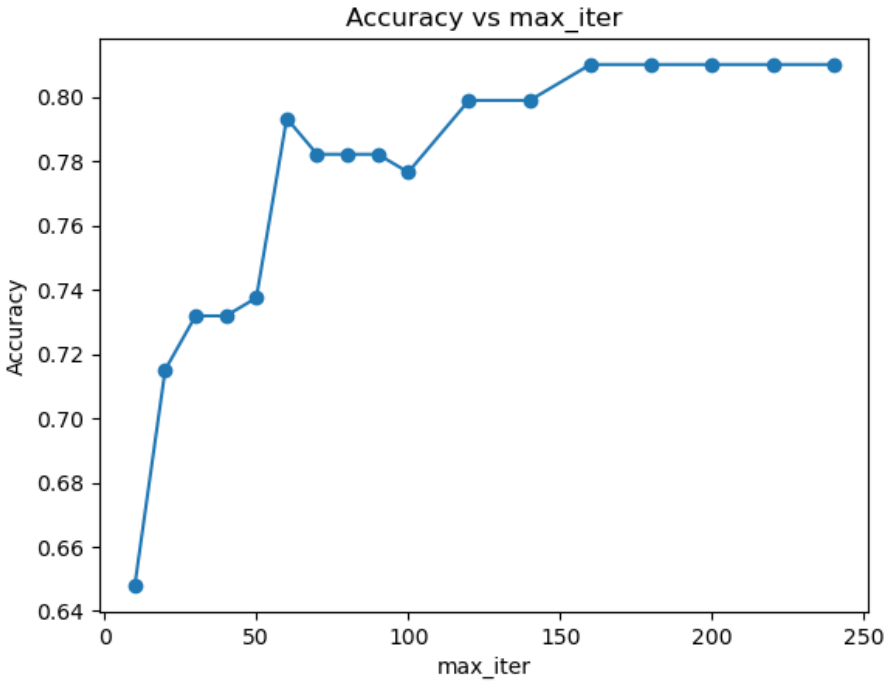


Aquí podemos apreciar que ha sobrevivido gente de todos las clases sociales más o menos en una proporción parecida, por lo que no podemos hacer un juicio de valor. Aunque vamos a observar la gráfica de los no supervivientes:



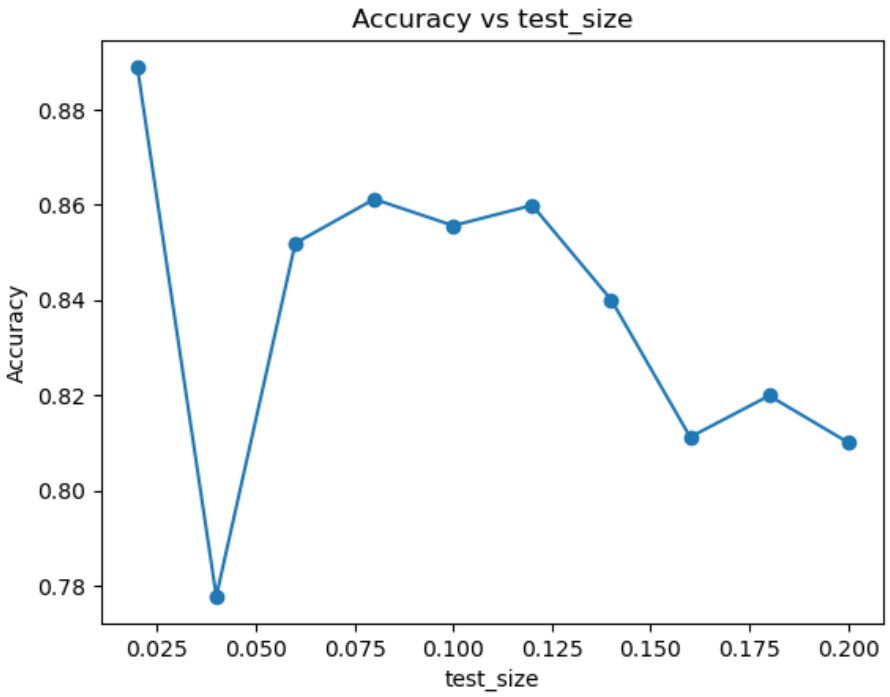
Si observamos la gráfica anterior podemos apreciar que si que han fallecido más personas de la tercera clase, luego de la segunda y finalmente de la primera.

Luego también hemos probado a cambiar el valor del parámetro max_iters para ver como evoluciona la precisión del modelo, hemos obtenido la siguiente gráfica:



Como se puede apreciar en la imagen anterior vemos que la precisión del modelo da 0.81 con un max_iter bastante más pequeño del que venía puesto por defecto, por lo que lo cambiaremos por un valor de 160 aproximadamente, en vez de 1000.

Por otro lado, también hemos probado a cambiar el test_size para ver que valores obtenemos para la precisión del modelo, hemos obtenido la siguiente gráfica:



Vemos que a menor valor de test_size obtenemos una mayor precisión, concretamente con un 2 % para el conjunto de pruebas obtenemos una precisión de 0.88, aunque creo que

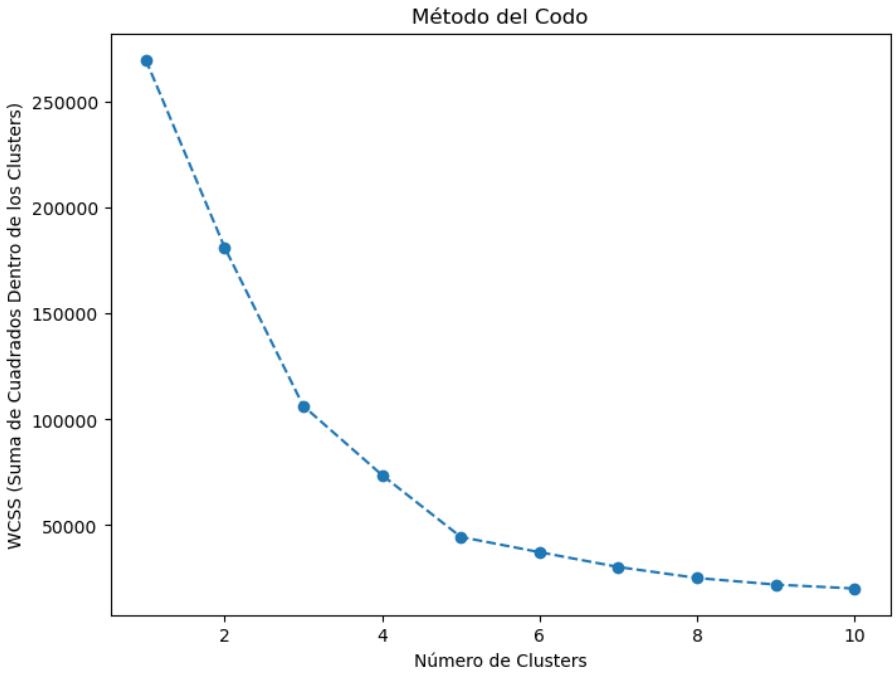
el valor ideal para este parámetro sería del 10 %, ya que tendremos suficientes imágenes para probar y también obtenemos una buena precisión. Cambiando este valor y el anterior, hemos podido mejorar la precisión del modelo a 0.85, lo podemos apreciar en la siguiente figura:

```
Precisión del modelo: 0.8555555555555555
Informe de clasificación:
              precision    recall  f1-score   support

           0       0.90      0.85      0.88        54
           1       0.79      0.86      0.83        36

    accuracy                           0.86        90
   macro avg       0.85      0.86      0.85        90
weighted avg       0.86      0.86      0.86        90
```

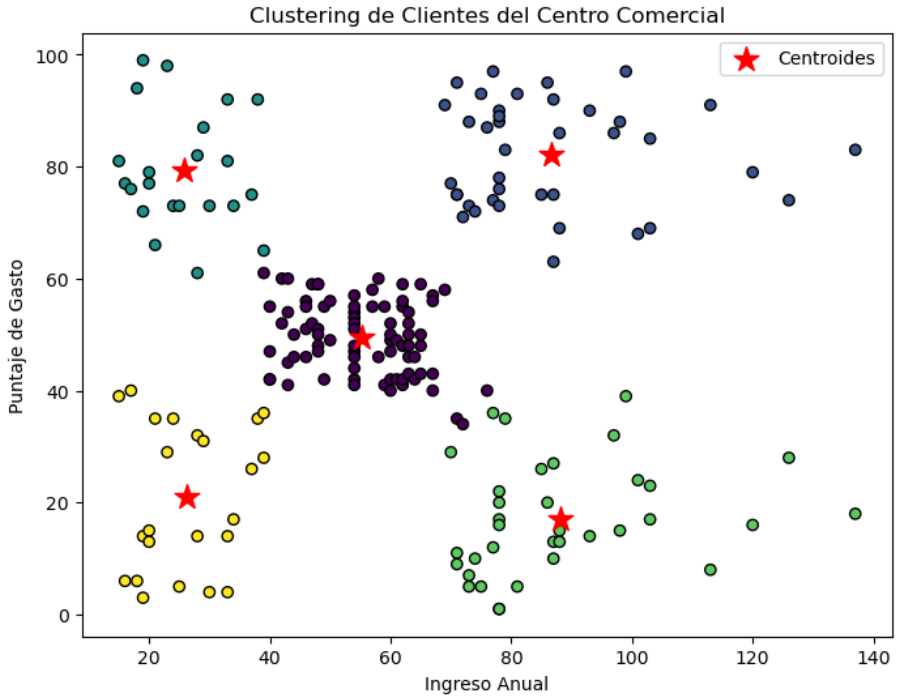
Figura 4: Resultado de cambiar test_size y max_iter

4. Aprendizaje automático no supervisado

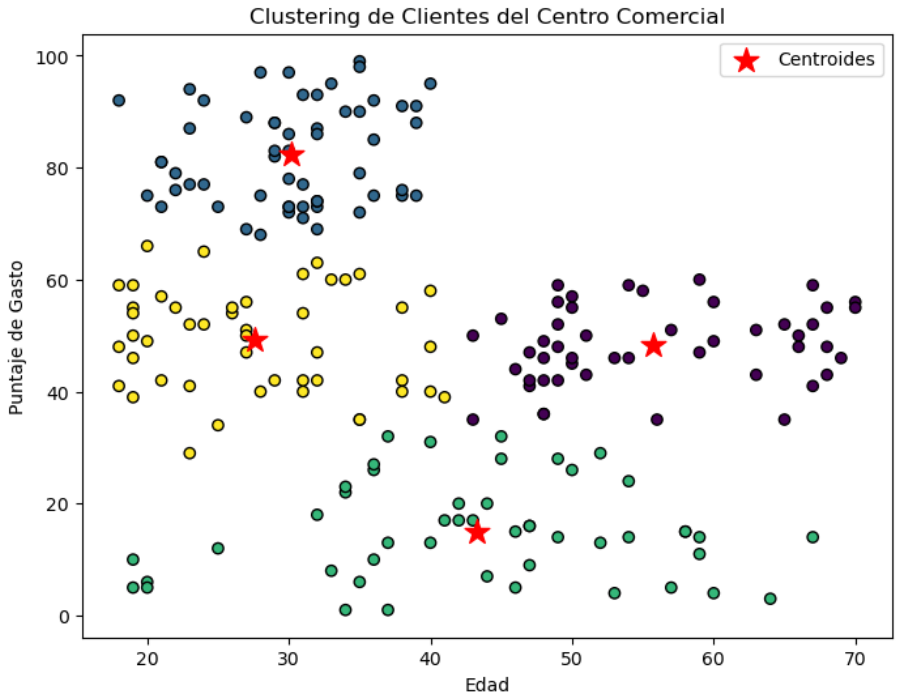
En esta parte vamos a usar el Clustering para agrupar según ciertas características del dataset, vamos a emplear un dataset sobre un centro comercial, tras ejecutar el código que se nos proporciona obtenemos las siguiente gráficas:



La gráfica anterior nos permite encontrar el valor "ideal" para el número de clusters, ya que calcula el valor donde la diferencia respecto a aumentar el número de clusters no es tan notorio. La otra gráfica que obtenemos es la que calcula el clustering de clientes por el ingreso anual y el puntaje de gasto.

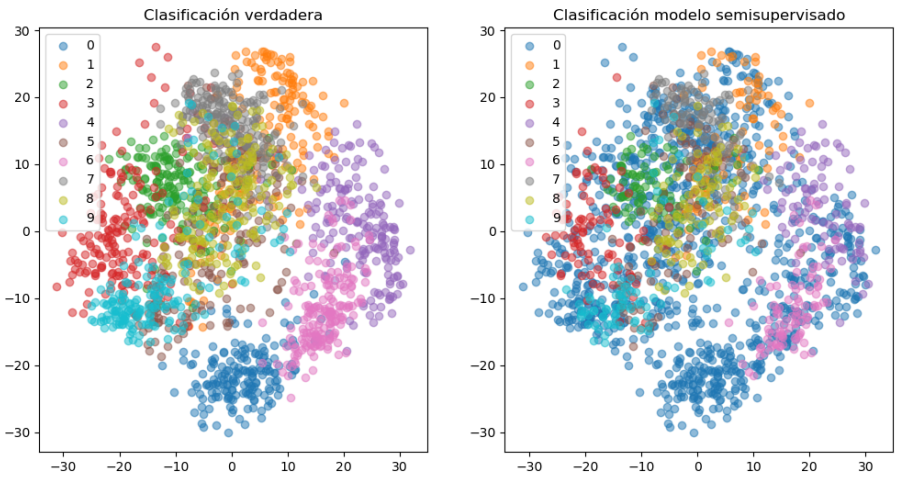


Vamos a realizar una prueba para la edad y el puntaje de gasto, hemos obtenido esta gráfica:

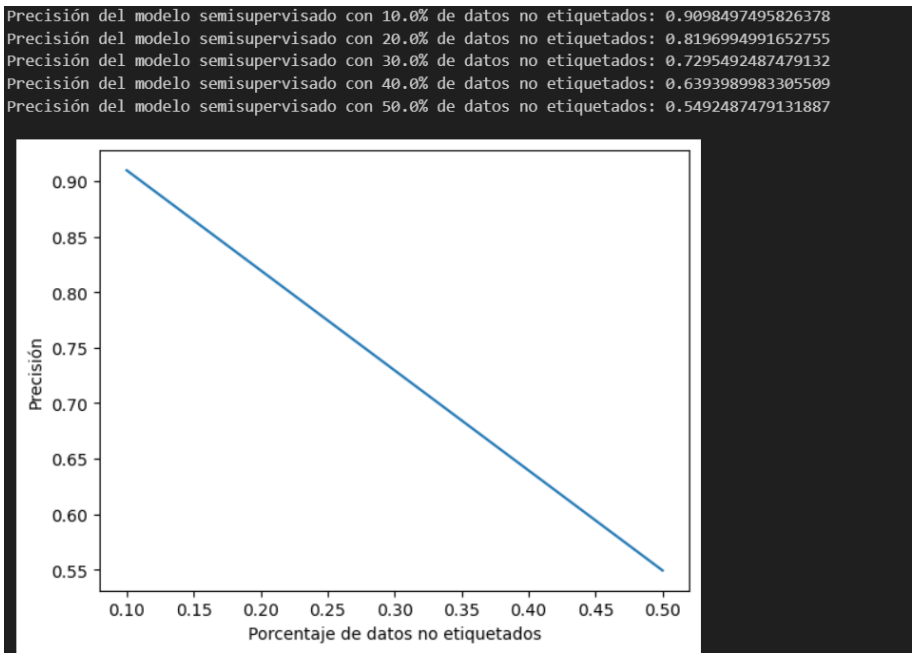


5. Aprendizaje semisupervisado

Finalmente, vamos a realizar el aprendizaje semisupervisado, aquí comprobaremos como este se comporta. Tras ejecutar el código obtenemos los siguientes resultados:



Además, el acierto que se obtiene con el modelo supervisado es de 0.54, sin lugar a duda una tasa pobre de acierto. He probado a modificar el parámetro porcentaje_no_etiquetados, a medida que este valor disminuye aumenta la precisión del modelo, esto tiene sentido, ya que estaremos entrenando con más datos etiquetados, esto lo podemos apreciar en la siguiente gráfica:



Como se puede apreciar, vemos que este valor influye mucho a la hora de mejorar la precisión del modelo.

6. Conclusiones

En esta práctica, se han explorado y evaluado distintas técnicas de aprendizaje automático, incluyendo modelos supervisados (regresión y clasificación), no supervisados (clustering) y semisupervisados. A través del análisis de datasets reales, como el de precios de viviendas en Boston y el de supervivientes del Titanic, se pudo observar cómo el ajuste

de parámetros clave (como el coeficiente de determinación en regresión y la matriz de confusión en clasificación) afecta el rendimiento de los modelos. Los resultados muestran que la precisión de los modelos supervisados puede mejorar significativamente ajustando parámetros como el tamaño del conjunto de prueba y el número de iteraciones. En el caso del aprendizaje semisupervisado, se evidenció que aumentar los datos etiquetados incrementa la precisión del modelo. En conjunto, estos experimentos subrayan la importancia de ajustar los parámetros y seleccionar el tipo adecuado de aprendizaje para optimizar el rendimiento en tareas específicas de clasificación y predicción.