

**Universidad Tecnológica De Querétaro**

**2da Evaluación**

**Diego Ulises Mireles Marcial**

**viernes, 11 de agosto de 2023**

**UTEQ, Av. Pie de la Cuesta 2501, Nacional, 76148 Santiago de Querétaro, Qro.**

**Grupo: IDSG05**

Contenido

[SA 3](#_Toc142664319)

[Análisis supervisado (Beisbol) 3](#_Toc142664320)

[Justificación del algoritmo. 3](#_Toc142664321)

[Descripción del diseño del modelo 3](#_Toc142664322)

[Evaluación y optimización del modelo. 4](#_Toc142664323)

[Enlace hacia un repositorio que contenga el modelo obtenido. 5](#_Toc142664324)

[Gráfica personalizada e interpretación de resultados. 5](#_Toc142664325)

[Análisis supervisado (diabetes\_indiana) 6](#_Toc142664326)

[Justificación del algoritmo. 6](#_Toc142664327)

[Descripción del diseño del modelo. 6](#_Toc142664328)

[Evaluación y optimización del modelo. 8](#_Toc142664329)

[Enlace hacia un repositorio que contenga el modelo obtenido. 9](#_Toc142664330)

[Gráfica personalizada e interpretación de resultados. 9](#_Toc142664331)

[Análisis no supervisado (Samsung) 10](#_Toc142664332)

[Justificación del algoritmo. 10](#_Toc142664333)

[Descripción del diseño del modelo. 10](#_Toc142664334)

[Enlace hacia un repositorio que contenga el modelo obtenido. 12](#_Toc142664335)

[Gráfica personalizada e interpretación de resultados. 12](#_Toc142664336)

[Análisis no supervisado (comprar\_alquilar) 13](#_Toc142664337)

[Justificación del algoritmo. 13](#_Toc142664338)

[Descripción del diseño del modelo. 13](#_Toc142664339)

[Enlace hacia un repositorio que contenga el modelo obtenido. 14](#_Toc142664340)

[Gráfica personalizada e interpretación de resultados. 15](#_Toc142664341)

# SA

## Análisis supervisado (Beisbol)

1. Basado en el conjunto de datos "beisbol.csv", implemente el algoritmo de regresión de su preferencia y entregue un reporte que incluya:

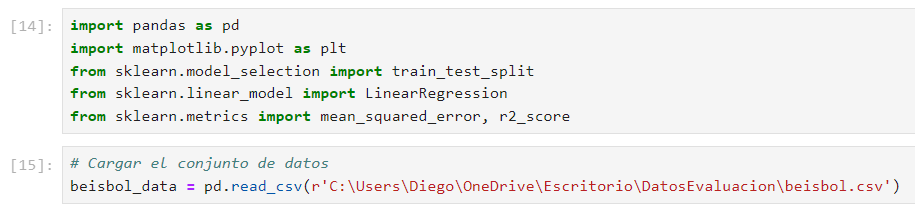
### Justificación del algoritmo.

En este caso, hemos utilizado el algoritmo de regresión lineal para resolver el problema. La regresión lineal es una técnica estadística que se utiliza para modelar la relación entre una variable independiente (en este caso, los datos de bateos y equipos) y una variable dependiente (los datos de carreras anotadas). La elección de la regresión lineal se justifica por la suposición de que existe una relación lineal entre las características y la variable objetivo en el conjunto de datos.

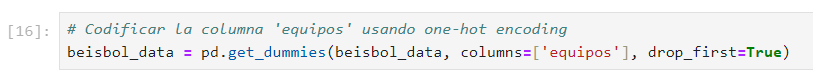
### Descripción del diseño del modelo

El diseño del modelo implica los siguientes pasos:

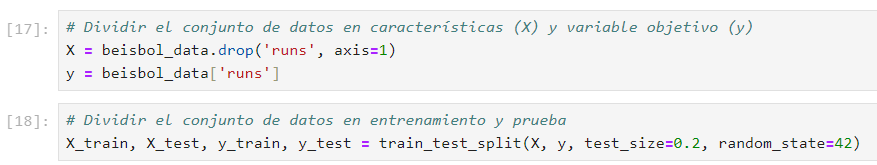
* Se cargan los datos del archivo CSV en un DataFrame utilizando la biblioteca Pandas.



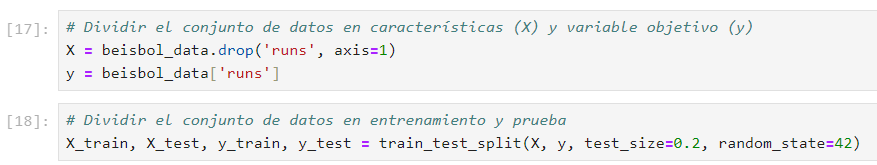
* Se realiza el preprocesamiento de datos: Se codifica la columna categórica 'equipos' utilizando one-hot encoding para convertirla en variables numéricas.



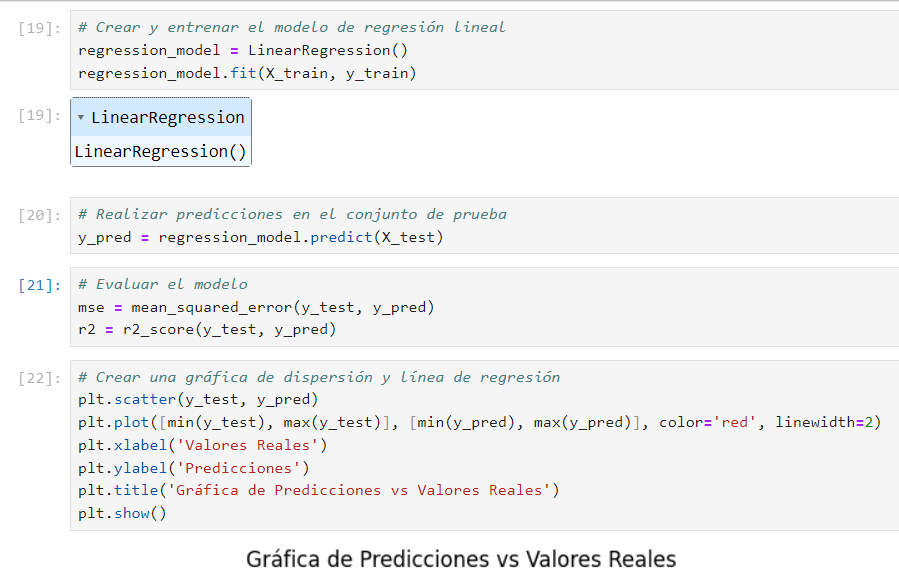
* Se dividen los datos en características (X) y variable objetivo (y).

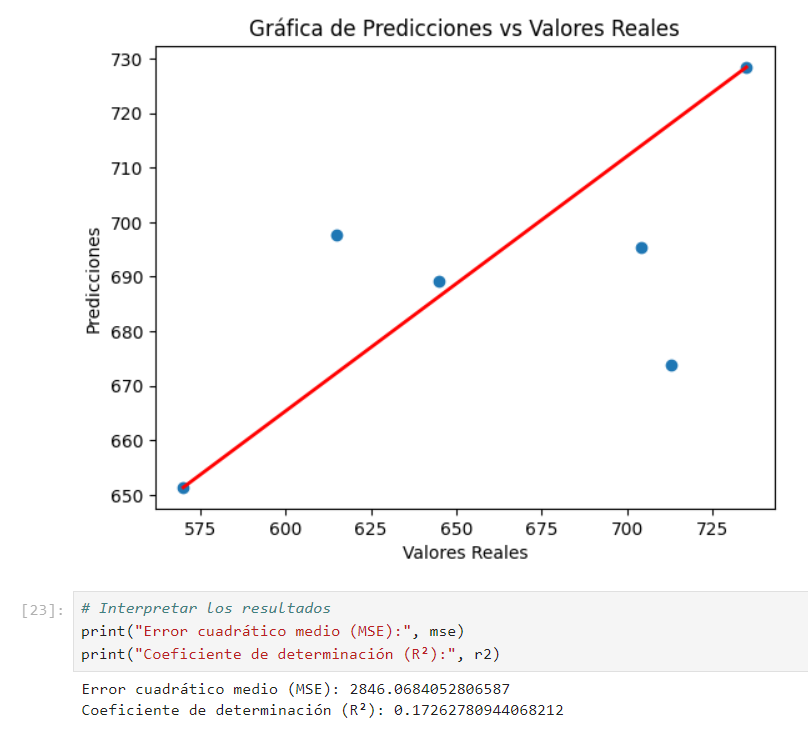


* Se divide el conjunto de datos en conjuntos de entrenamiento y prueba utilizando la función train\_test\_split de Scikit-learn.



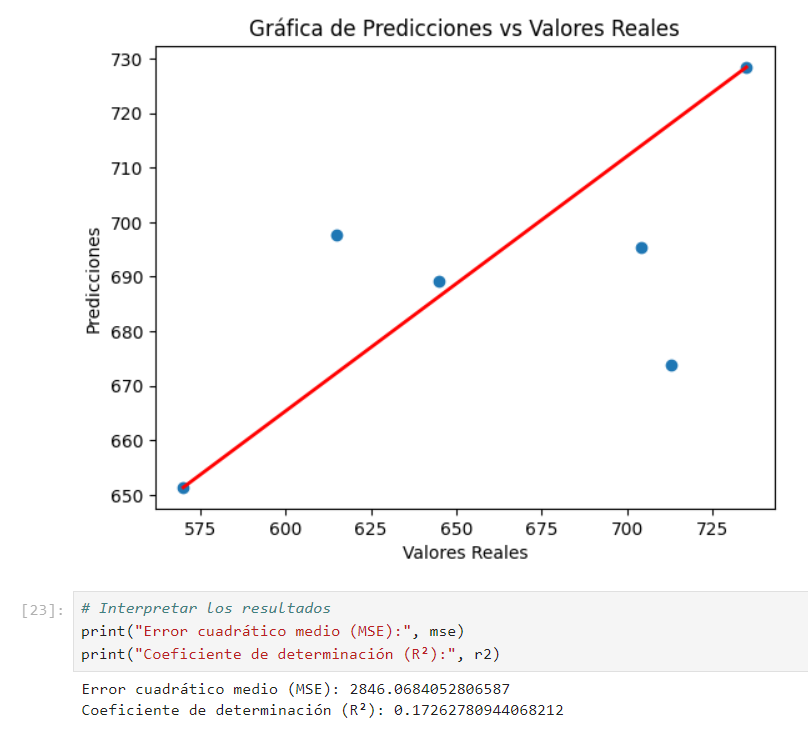
* Se crea una instancia del modelo de regresión lineal (LinearRegression) y se entrena en los datos de entrenamiento utilizando el método fit.





### Evaluación y optimización del modelo.

Después de entrenar el modelo, se realizan predicciones en el conjunto de prueba utilizando el método predict. Luego, se calcula el error cuadrático medio (MSE) y el coeficiente de determinación (R²) para evaluar el rendimiento del modelo. Estas métricas nos ayudan a entender qué tan bien se ajusta el modelo a los datos reales y cómo de bien puede predecir nuevas observaciones.

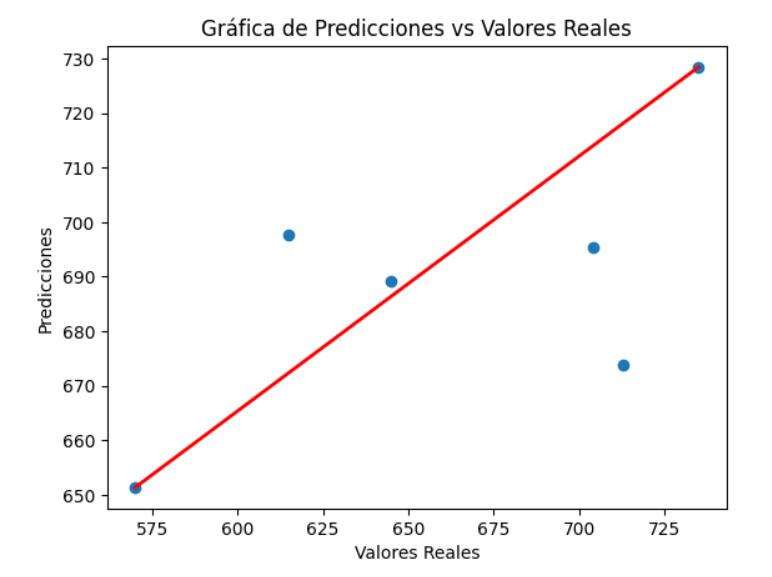


### Enlace hacia un repositorio que contenga el modelo obtenido.

Para poder acceder a los datos y resultados, visitar el siguiente repositorio:

<https://github.com/DiegoUlisesMM/Eva2BD>

### Gráfica personalizada e interpretación de resultados.



Representa visualmente la relación entre las puntuaciones predichas por el modelo de regresión y los valores reales de carreras anotadas en los partidos de béisbol. Cada punto en la gráfica representa un equipo y su respectiva puntuación real en el eje horizontal, mientras que en el eje vertical se encuentran las puntuaciones predichas por el modelo. Una dispersión cercana a la línea diagonal indica una predicción precisa, donde las puntuaciones predichas se alinean estrechamente con las puntuaciones reales.

se puede observar que en su mayoría las puntuaciones predichas se encuentran en proximidad a la línea diagonal, lo que sugiere que el modelo ha logrado capturar las tendencias generales en los datos. Sin embargo, también se pueden identificar algunas discrepancias notables entre las predicciones y los valores reales, especialmente en los extremos. Estas discrepancias podrían deberse a variaciones no consideradas por el modelo o a factores atípicos en los datos.

## Análisis supervisado (diabetes\_indiana)

2. Basado en el conjunto de datos "diabetes\_indiana.csv" implemente el algoritmo de clasificación de su preferencia y entregue un reporte que incluya:

### Justificación del algoritmo.

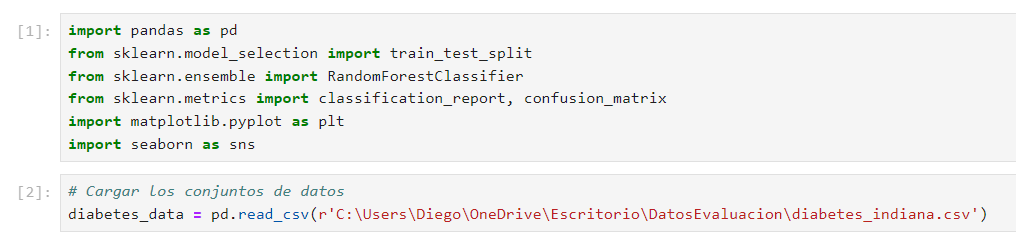
Se eligió utilizar el algoritmo RandomForestClassifier para abordar el problema de clasificación en el conjunto de datos de diabetes de Indiana. La justificación detrás de esta elección radica en las siguientes razones:

* Flexibilidad y Robustez: RandomForestClassifier es un método de aprendizaje automático versátil que puede manejar tanto problemas de clasificación como de regresión. Son resistentes al sobreajuste y funcionan bien en una variedad de situaciones.
* Manejo de Características: Es capaz de manejar múltiples características y son adecuados para conjuntos de datos con características numéricas y categóricas.
* Reducción de Variabilidad: Al promediar las predicciones de varios árboles de decisión, los bosques aleatorios tienden a reducir la variabilidad y el ruido en las predicciones.

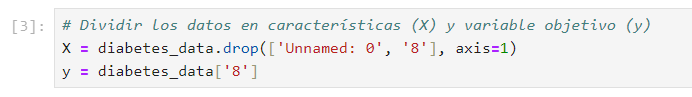
### Descripción del diseño del modelo.

El diseño del modelo se realizó siguiendo los siguientes pasos:

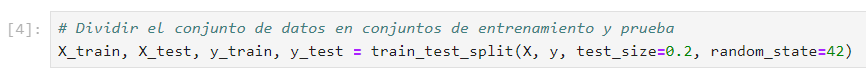
* Carga de Datos: Se cargó el conjunto de datos "diabetes\_indiana.csv" que contiene información sobre pacientes diabéticos.



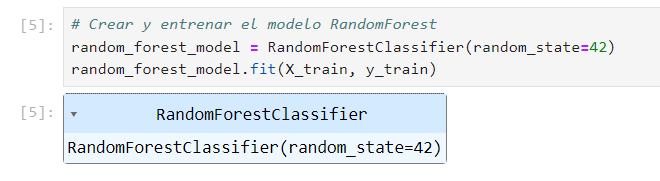
* Preprocesamiento de Datos: Se eliminó la columna "Unnamed: 0" que parece ser un índice no necesario. Luego, se dividieron los datos en características (variables independientes) y la variable objetivo (variable dependiente).



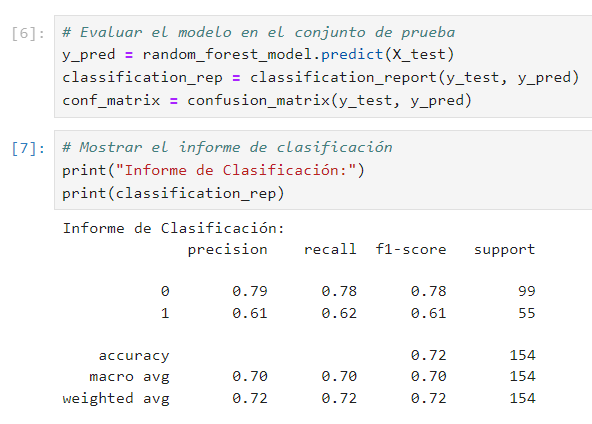
* División en Conjuntos de Entrenamiento y Prueba: Los datos se dividieron en conjuntos de entrenamiento (80%) y prueba (20%) utilizando la función train\_test\_split.

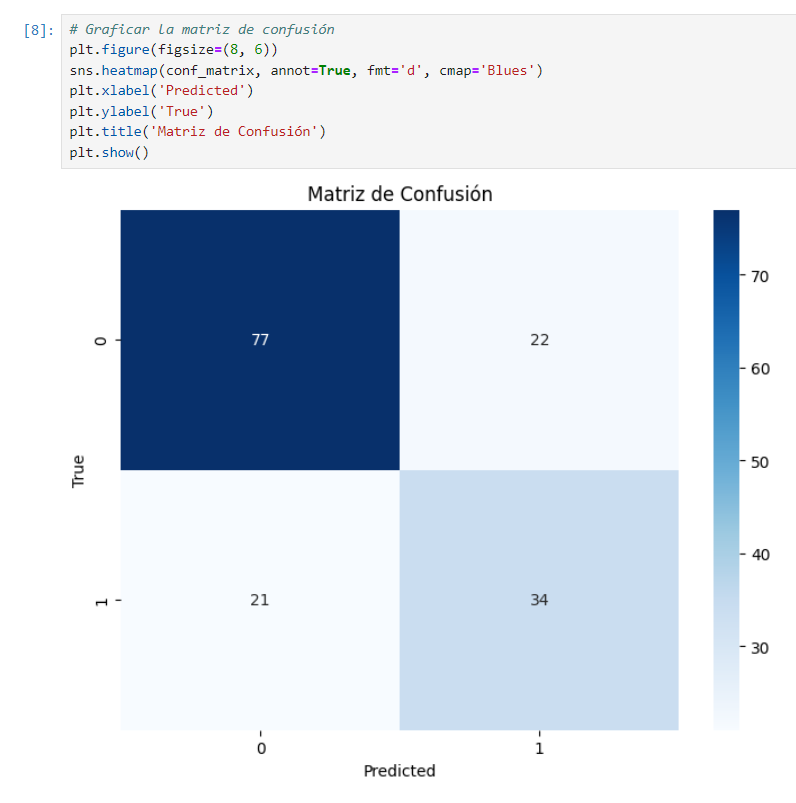


* Creación y Entrenamiento del Modelo: Se creó un modelo RandomForestClassifier con una semilla aleatoria de 42 y se entrenó utilizando el conjunto de entrenamiento.



* Evaluación del Modelo: Se realizaron predicciones en el conjunto de prueba y se generó un informe de clasificación que incluye métricas como precisión, recall, F1-score y soporte.



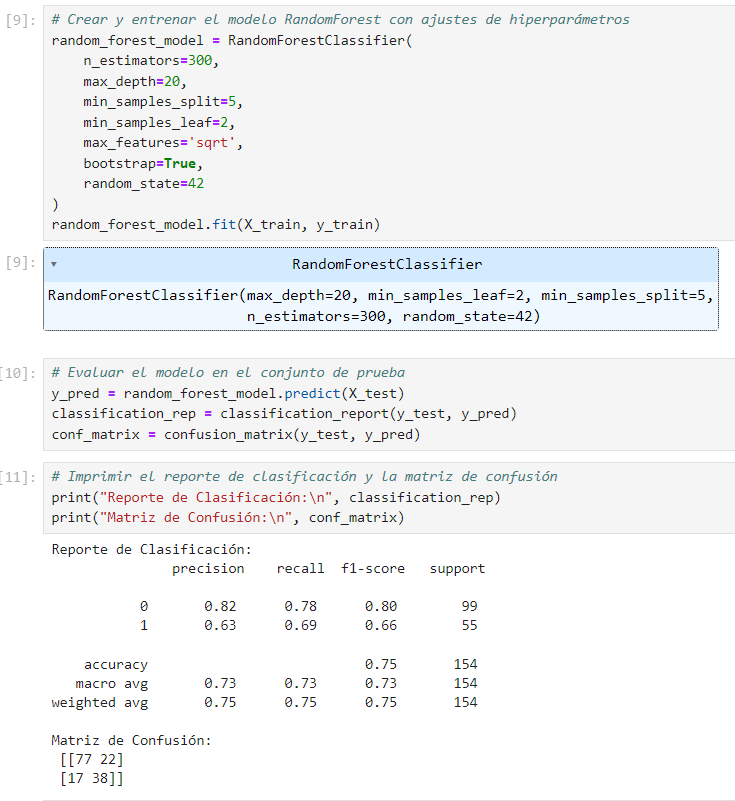


### Evaluación y optimización del modelo.

Se realiza una optimización adicional de los hiperparámetros del modelo RandomForest para mejorar su rendimiento. El código realiza una evaluación cuantitativa del modelo RandomForest utilizando métricas de evaluación específicas y realiza ajustes de los hiperparámetros con el objetivo de optimizar el rendimiento del modelo en la tarea de clasificación de la diabetes.

Los hiperparámetros que se ajustan incluyen el número de estimadores (n\_estimators), la profundidad máxima de los árboles (max\_depth), el número mínimo de muestras requeridas para dividir un nodo (min\_samples\_split) y el número mínimo de muestras requeridas en un nodo hoja (min\_samples\_leaf).

Estos hiperparámetros afectan la complejidad y la capacidad de generalización del modelo. Al ajustarlos, se busca encontrar un equilibrio entre el ajuste excesivo (overfitting) y la falta de ajuste (underfitting), mejorando así el rendimiento en datos no vistos.

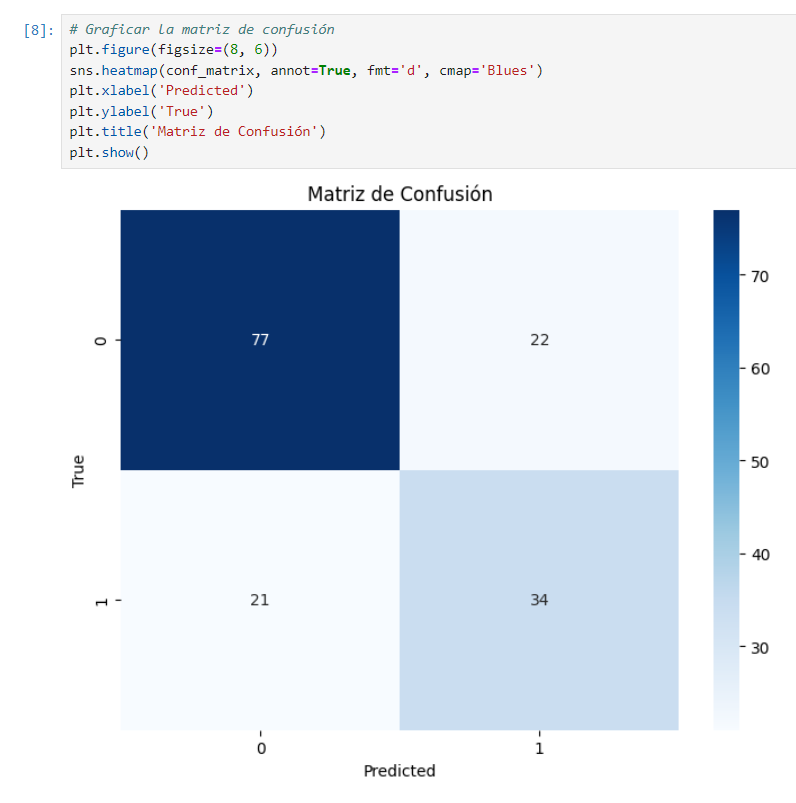


### Enlace hacia un repositorio que contenga el modelo obtenido.

Para poder acceder a los datos y resultados, visitar el siguiente repositorio:

<https://github.com/DiegoUlisesMM/Eva2BD>

### Gráfica personalizada e interpretación de resultados.



La gráfica de la matriz de confusión generada a partir de los archivos de diabetes de Indiana proporciona una representación visual de cómo el modelo de clasificación está realizando las predicciones en términos de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Cada una de las celdas de la matriz muestra la cantidad de instancias que caen en esa categoría.

En esta gráfica, si enfocamos en la diagonal principal (de la esquina superior izquierda a la inferior derecha), podemos observar que los verdaderos positivos y verdaderos negativos son mayores en comparación con los falsos positivos y falsos negativos. Esto indica que el modelo está acertando en la clasificación de las instancias en ambas clases (diabéticas y no diabéticas) en general.

La gráfica también puede ayudarnos a identificar en qué clase el modelo tiende a equivocarse más. Si los falsos positivos son significativamente más altos que los falsos negativos, podría indicar que el modelo está siendo más propenso a clasificar erróneamente instancias como positivas cuando en realidad son negativas. Del mismo modo, si los falsos negativos son mucho mayores que los falsos positivos, podría indicar que el modelo está perdiendo instancias que son positivas, pero se clasifican como negativas.

## Análisis no supervisado (Samsung)

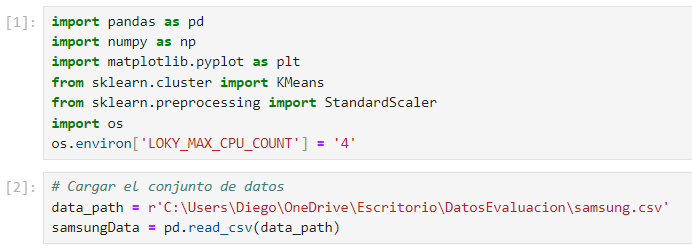
3. Basado en el conjunto de datos "samsung.csv" implemente el algoritmo de agrupación de su preferencia y entregue un reporte que incluya:

### Justificación del algoritmo.

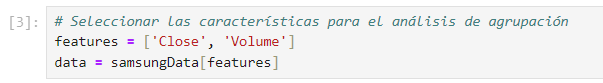
El algoritmo de agrupación K-Means ha sido seleccionado para este análisis debido a su simplicidad y eficacia en la agrupación de datos numéricos. K-Means es ampliamente utilizado en problemas de segmentación de datos y es capaz de agrupar observaciones similares en clusters basados en medidas de distancia. A pesar de que no está diseñado específicamente para series temporales, podemos adaptarlo para analizar cómo se agrupan los datos de precios de cierre y volúmenes de acciones de Samsung en un espacio bidimensional.

### Descripción del diseño del modelo.

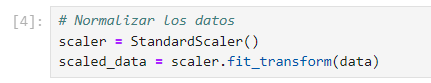
* Carga de Datos: Se cargaron los datos de precios de cierre y volúmenes de acciones de Samsung desde el archivo "samsung.csv".



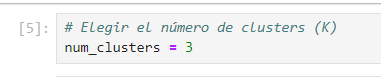
* Selección de Características: Se seleccionaron las características "Close" y "Volume" para el análisis de agrupación.



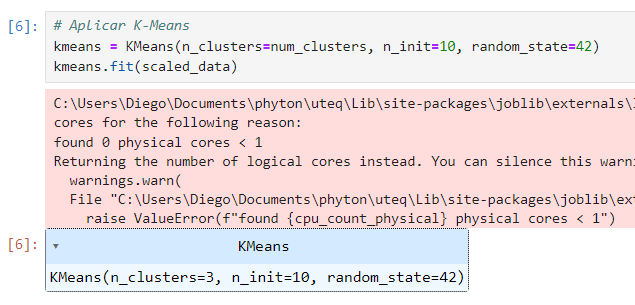
* Normalización de Datos: Los datos fueron normalizados utilizando la clase StandardScaler para garantizar que las características tengan la misma escala.



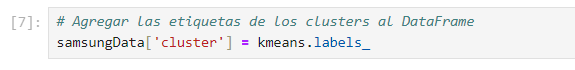
* Elección del Número de Clusters (K): Se eligió un número de clusters igual a 3, pero esta elección puede variar dependiendo de los objetivos del análisis.



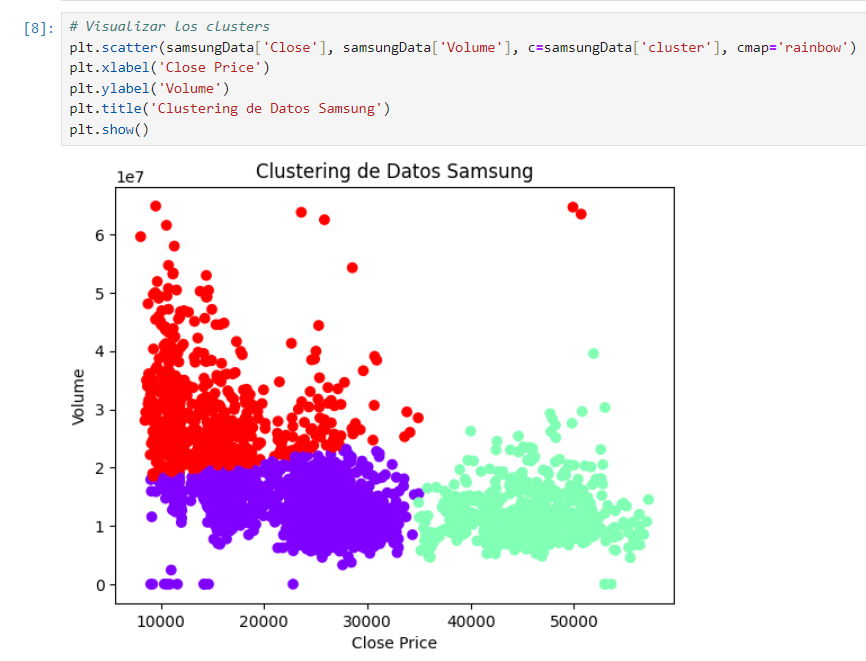
* Aplicación de K-Means: Se aplicó el algoritmo K-Means para agrupar los datos normalizados en clusters.



Los mensajes de advertencia son principalmente relacionados con la configuración y ejecución del algoritmo K-Means. Sin embargo, estas advertencias no impiden que el código funcione.



* Visualización de Resultados: Se creó una gráfica bidimensional con los precios de cierre en el eje x y los volúmenes en el eje y. Los puntos fueron coloreados según el cluster al que pertenecen.

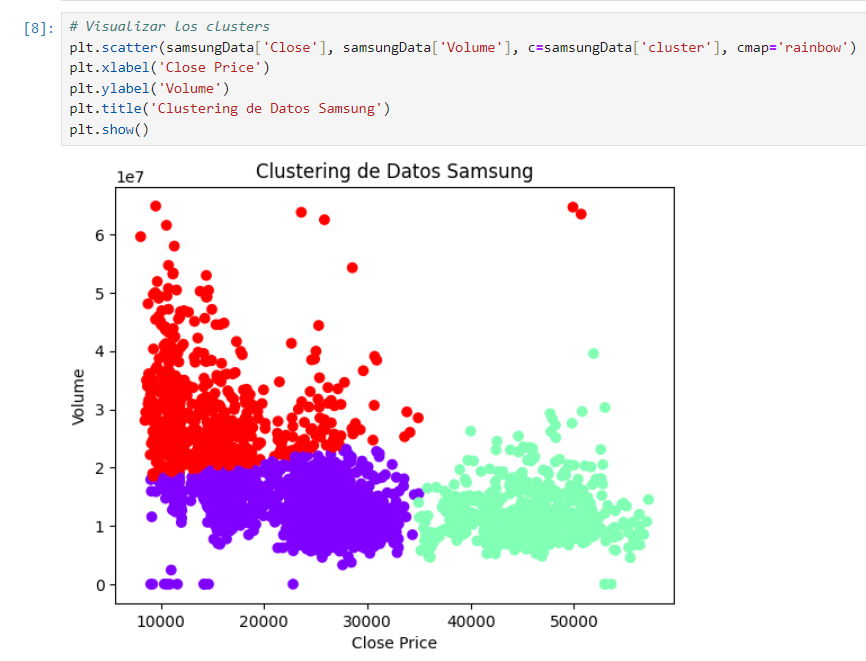


### Enlace hacia un repositorio que contenga el modelo obtenido.

Para poder acceder a los datos y resultados, visitar el siguiente repositorio:

<https://github.com/DiegoUlisesMM/Eva2BD>

### Gráfica personalizada e interpretación de resultados.



La gráfica de dispersión muestra la distribución de los datos de precios de cierre y volúmenes de acciones de Samsung en un espacio bidimensional. Cada punto en la gráfica representa una observación en el conjunto de datos. Los colores de los puntos indican a qué cluster ha sido asignada cada observación por el algoritmo K-Means.

Cluster 0 (Azul): Este cluster muestra observaciones con precios de cierre relativamente bajos y volúmenes moderados. Estas observaciones podrían representar períodos en los que el precio de las acciones de Samsung se mantuvo más bajo y el volumen de negociación fue constante.

Cluster 1 (Rojo): Este cluster agrupa observaciones con precios de cierre más altos y volúmenes bajos. Puede indicar períodos en los que el precio de las acciones aumentó significativamente, pero con un menor volumen de operaciones, lo que podría sugerir un interés selectivo por parte de los inversores.

Cluster 2 (Verde): Este cluster incluye observaciones con precios de cierre moderados y volúmenes altos. Esto podría indicar momentos en los que el precio se mantuvo en un rango intermedio mientras que el volumen de negociación fue considerablemente alto, posiblemente debido a noticias o eventos que generaron un mayor interés y actividad en el mercado.

En general, la agrupación proporciona una perspectiva interesante sobre cómo los precios de cierre y los volúmenes de negociación están relacionados en diferentes períodos de tiempo.

## Análisis no supervisado (comprar\_alquilar)

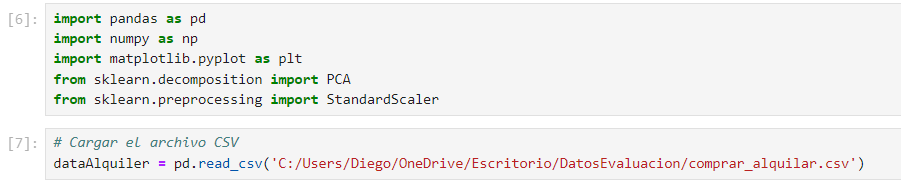
4. Basado en el conjunto de datos "comprar\_alquilar.csv" implemente el algoritmo de reducción de dimensionalidad de su preferencia y entregue un reporte que incluya:

### Justificación del algoritmo.

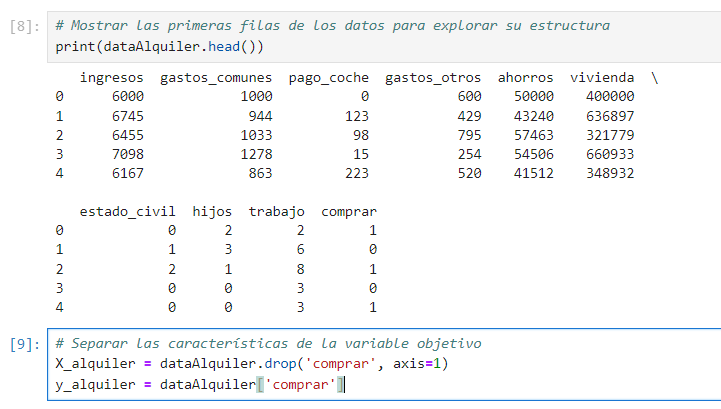
El Análisis de Componentes Principales (PCA) es un algoritmo de reducción de dimensionalidad que se utiliza para transformar un conjunto de datos en un nuevo espacio de características de menor dimensión, preservando la mayor cantidad posible de la varianza original. Esto es útil para simplificar los datos y eliminar la multicolinealidad entre características, lo que puede mejorar la eficiencia de los modelos de aprendizaje automático y facilitar la visualización de los datos.

### Descripción del diseño del modelo.

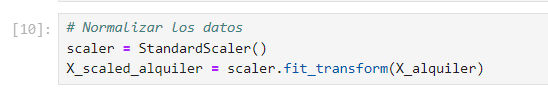
En este diseño, cargamos el archivo CSV "comprar\_alquilar.csv" que contiene información sobre características relacionadas con la decisión de comprar o alquilar una propiedad.



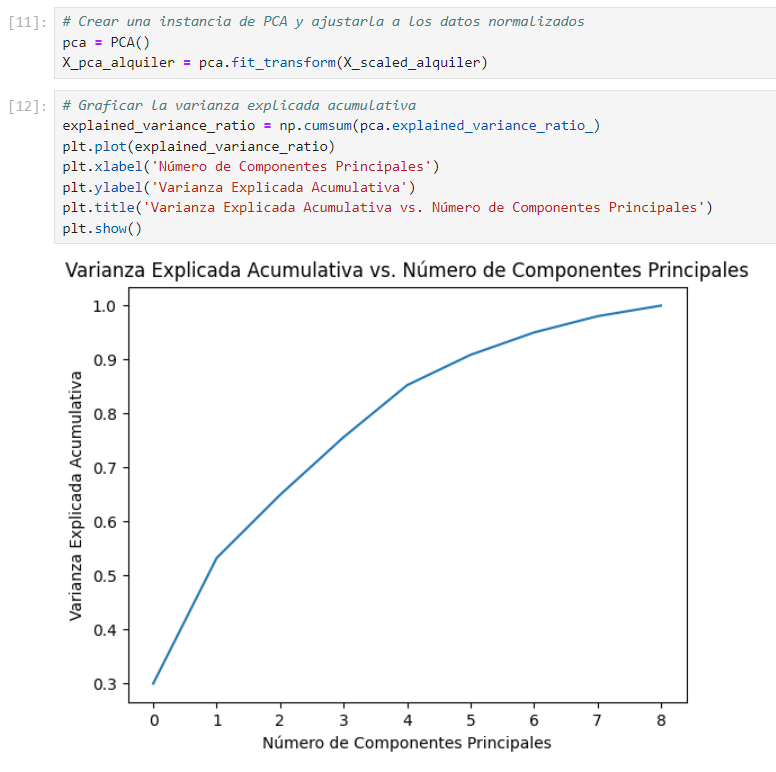
Eliminamos la columna de la variable objetivo "comprar" para realizar la reducción de dimensionalidad solo en las características.



Normalizamos los datos para asegurarnos de que todas las características tengan el mismo peso en el proceso de reducción.



Luego, utilizamos PCA para transformar los datos normalizados en un nuevo espacio de características. Esto nos dará nuevas variables llamadas "componentes principales" que son combinaciones lineales de las características originales. Estos componentes están ordenados de manera que el primero captura la mayor varianza en los datos, el segundo el siguiente mayor y así sucesivamente.

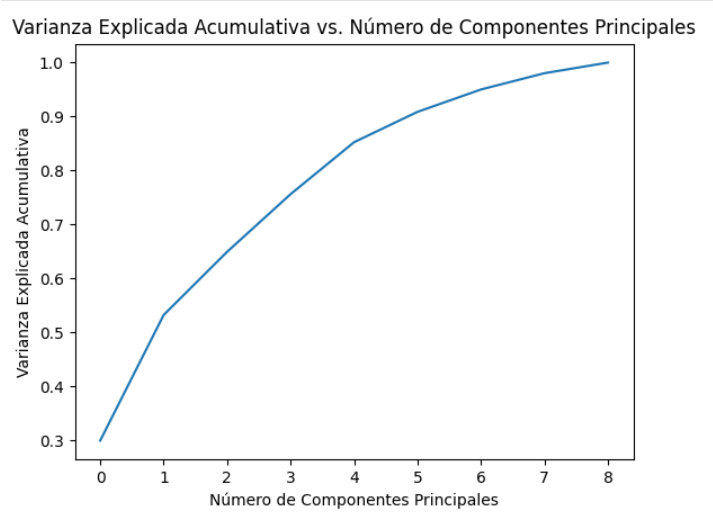


### Enlace hacia un repositorio que contenga el modelo obtenido.

Para poder acceder a los datos y resultados, visitar el siguiente repositorio:

<https://github.com/DiegoUlisesMM/Eva2BD>

### Gráfica personalizada e interpretación de resultados.



La gráfica de "Varianza Explicada Acumulativa vs. Número de Componentes Principales" proporciona información sobre cuánta varianza de los datos originales se conserva al considerar diferentes cantidades de componentes principales después de aplicar la reducción de dimensionalidad mediante PCA.

En la gráfica, el eje x representa el número de componentes principales utilizados, mientras que el eje y muestra la varianza explicada acumulativa hasta ese punto. La varianza explicada acumulativa indica cuánta información de los datos originales se mantiene al proyectarlos en un espacio de menor dimensión.

La interpretación de la gráfica implica buscar el punto donde la curva comienza a aplanarse o alcanzar una meseta. A medida que se agregan más componentes principales, la varianza explicada acumulativa aumenta. Sin embargo, en algún momento, agregar más componentes puede no aportar significativamente más información y puede conducir a una mayor complejidad sin beneficios significativos.

# DE

Basado en el conjunto de datos "diabetes\_indiana.csv" implemente un algoritmo de clasificación alterno al realizado en SA y entregue un reporte que incluya:

### Justificación del algoritmo.

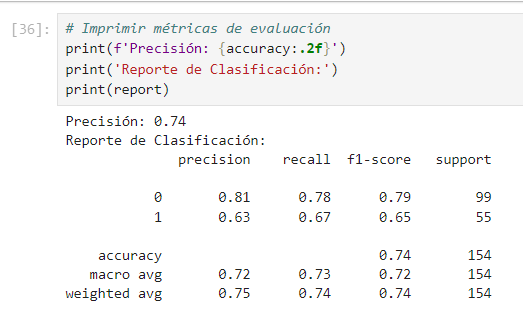
Gradient Boosting para este conjunto de datos debido a su capacidad para manejar relaciones no lineales entre características y etiquetas. Además, Gradient Boosting es robusto y puede trabajar bien incluso con datos ruidosos o con valores atípicos.

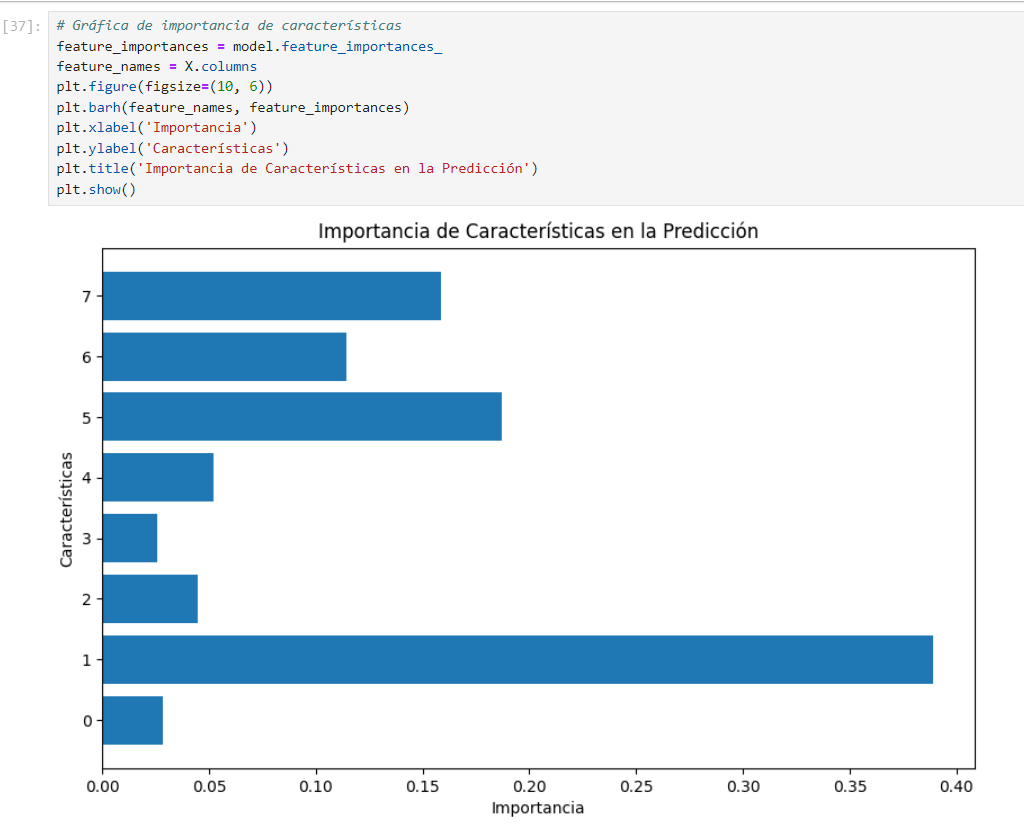
### Descripción del diseño del modelo.

En este caso, utilizaremos la implementación de Gradient Boosting de la biblioteca scikit-learn para crear el modelo. Dividiremos el conjunto de datos en características (X) y etiquetas (y), y luego dividiremos estos datos en conjuntos de entrenamiento y prueba. A continuación, crearemos un modelo GradientBoostingClassifier, lo entrenaremos en el conjunto de entrenamiento y evaluaremos su rendimiento en el conjunto de prueba.



Evaluaremos el modelo utilizando métricas de evaluación, como precisión, recall y F1-score. También se ajustarán los hiperparámetros del modelo para mejorar su rendimiento



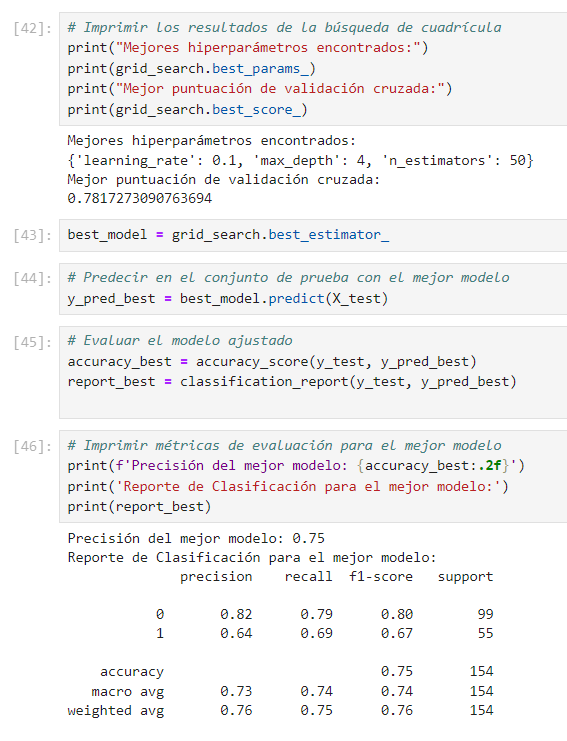


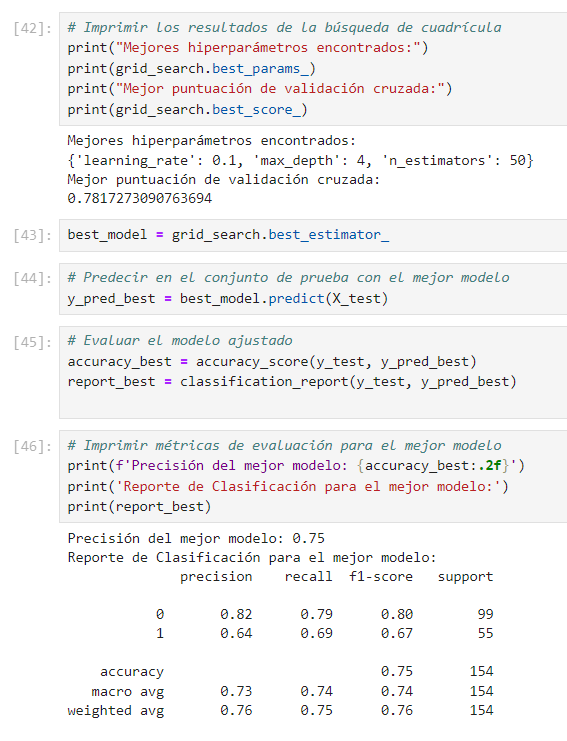
### Evaluación y optimización del modelo.

Ajustar los hiperparámetros del modelo puede mejorar su rendimiento. Uno de los enfoques comunes es realizar una búsqueda de cuadrícula (grid search) sobre diferentes combinaciones de hiperparámetros para encontrar la combinación óptima. A continuación, te proporciono los comandos para realizar una búsqueda de cuadrícula en el modelo Gradient Boosting









n\_estimators: Este hiperparámetro se refiere al número de árboles de decisión que se van a construir en el proceso de boosting. Un valor más alto de n\_estimators generalmente permite que el modelo sea más complejo y tenga un mayor poder predictivo. Sin embargo, un valor demasiado alto puede llevar a un sobreajuste. La búsqueda en la cuadrícula nos ayuda a encontrar un equilibrio adecuado para este valor, lo que puede conducir a un mejor rendimiento.

learning\_rate: El learning rate (tasa de aprendizaje) controla la contribución de cada árbol al modelo general. Valores más bajos de learning rate hacen que el modelo aprenda más lentamente y puede requerir más árboles para ajustarse bien a los datos. Valores más altos pueden llevar a un ajuste excesivo. La búsqueda de cuadrícula nos permite encontrar el learning rate que resulta en un buen equilibrio entre convergencia rápida y evitando el sobreajuste.

max\_depth: Este hiperparámetro controla la profundidad máxima de los árboles de decisión en el proceso de boosting. Árboles más profundos pueden capturar relaciones más complejas en los datos, pero también pueden llevar a un sobreajuste. Limitar la profundidad puede ayudar a prevenir el sobreajuste. La búsqueda en la cuadrícula nos ayuda a determinar la profundidad óptima que equilibra el poder predictivo con la generalización.

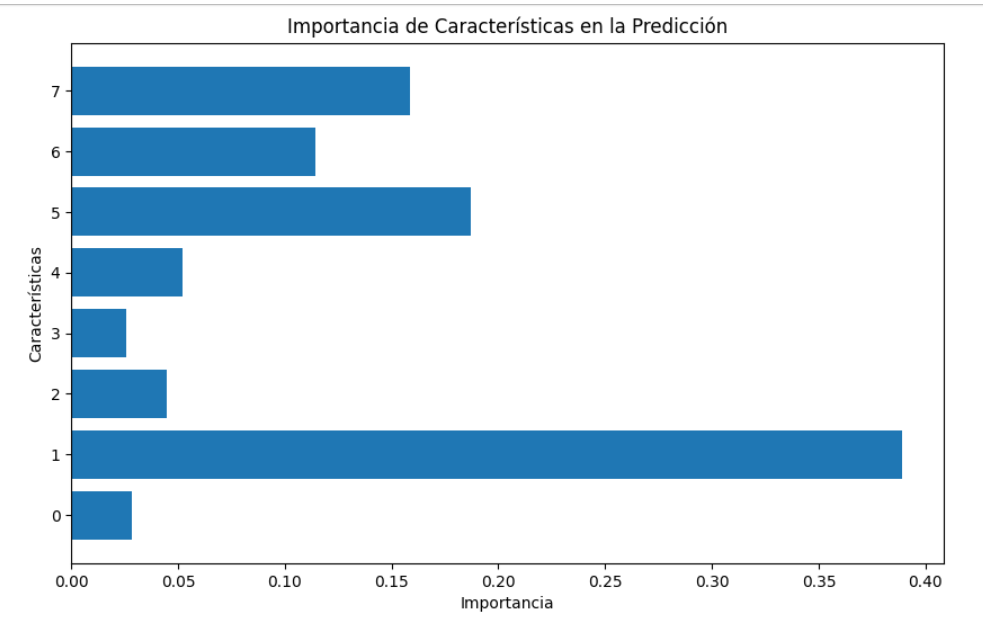
En resumen, al ajustar estos hiperparámetros, estamos controlando la complejidad y la velocidad de aprendizaje del modelo Gradient Boosting. La búsqueda de cuadrícula nos permite explorar diferentes combinaciones de estos hiperparámetros y encontrar el conjunto que optimiza el rendimiento del modelo en términos de precisión y capacidad de generalización en el conjunto de prueba.

### Enlace hacia un repositorio que contenga el modelo obtenido.

Para poder acceder a los datos y resultados, visitar el siguiente repositorio:

<https://github.com/DiegoUlisesMM/Eva2BD>

### Gráfica personalizada e interpretación de resultados.



La gráfica de importancia de características muestra la contribución relativa de cada característica en el proceso de toma de decisiones del modelo Gradient Boosting para predecir si una persona tiene diabetes o no. Las características con barras más largas son las que tuvieron una mayor influencia en la clasificación.

En el contexto de este conjunto de datos, la interpretación sería la siguiente:

Glucose (Glucosa): La concentración de glucosa en sangre es la característica más importante para determinar si una persona tiene diabetes. Valores más altos de glucosa tienden a aumentar la probabilidad de diabetes.

BMI (Índice de Masa Corporal): El índice de masa corporal también es un predictor significativo. Valores altos de BMI están correlacionados positivamente con la diabetes.

Age (Edad): La edad de la persona también tiene un impacto en la clasificación. Es posible que las personas mayores tengan una mayor probabilidad de desarrollar diabetes.

BloodPressure (Presión Arterial): Aunque menos influyente que las características anteriores, la presión arterial todavía contribuye en cierta medida a la clasificación.

SkinThickness (Espesor del Pliegue Cutáneo): Esta característica también influye en la clasificación, pero en menor medida en comparación con las anteriores.

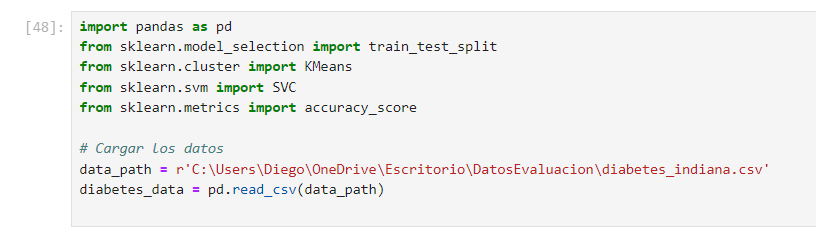
El resto de las características parecen tener una influencia más baja en la clasificación según este modelo en particular.

# AU

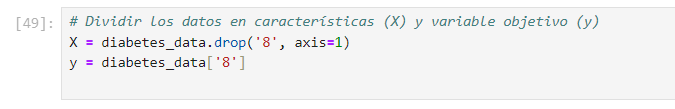
1. Elabora un documento donde implemente el aprendizaje por refuerzo (combinación de análisis supervisado y no supervisado)

Ejemplo básico que combina análisis no supervisado (K-Means) para la segmentación inicial de datos y análisis supervisado (SVM) para entrenar modelos en subconjuntos de datos segmentados. Luego, se combinan las decisiones de los modelos SVM entrenados en ambos clusters para obtener una estrategia de aprendizaje por refuerzo simple.

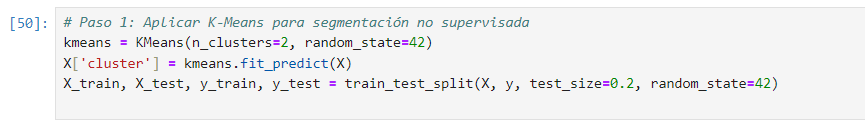
1. Carga de Datos: Se cargan los datos de diabetes\_indiana.csv y se divide en características (X) y variable objetivo (y).



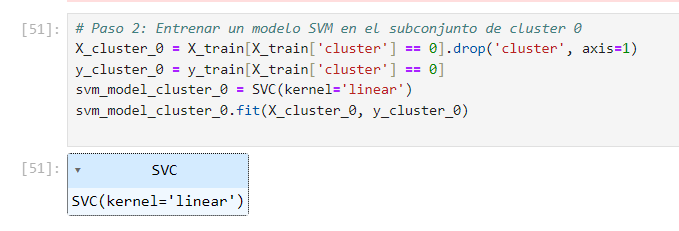
1. Segmentación no Supervisada: Se aplica K-Means para dividir los datos en dos clusters, permitiendo la identificación de patrones no evidentes.



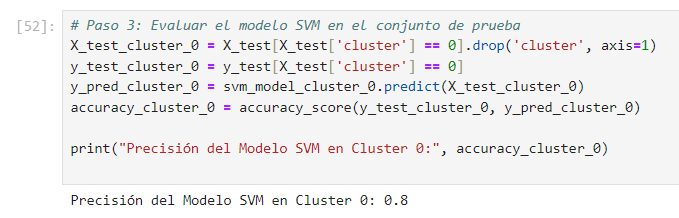
1. Entrenamiento del Modelo SVM en Cluster 0: Se entrena un modelo SVM con kernel lineal en el subconjunto de datos pertenecientes al cluster 0.



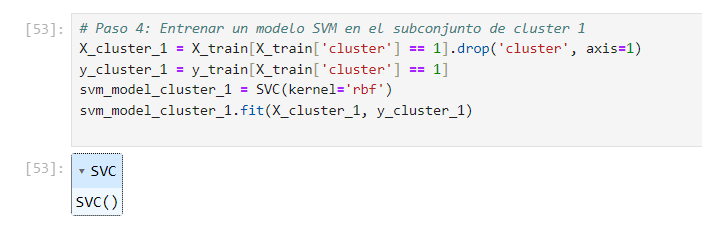
1. Evaluación del Modelo SVM en Cluster 0: Se evalúa el modelo SVM en el conjunto de prueba perteneciente al cluster 0 y se calcula su precisión.



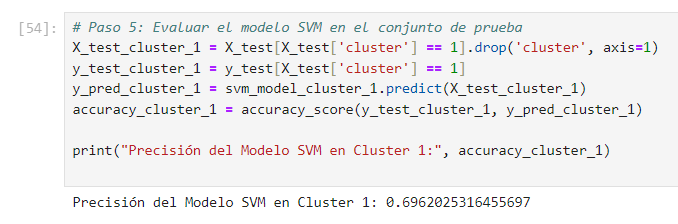
1. Entrenamiento del Modelo SVM en Cluster 1: Se entrena un modelo SVM con kernel RBF en el subconjunto de datos pertenecientes al cluster 1.



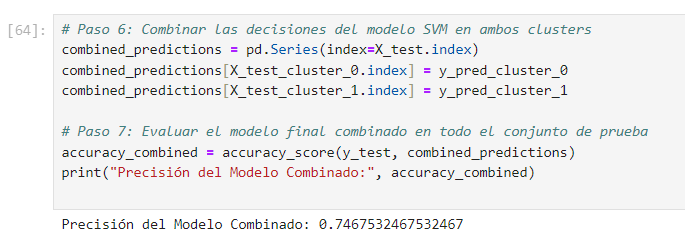
1. Evaluación del Modelo SVM en Cluster 1: Se evalúa el modelo SVM en el conjunto de prueba perteneciente al cluster 1 y se calcula su precisión.



1. Combinación de Predicciones: Se combinan las predicciones de ambos modelos SVM basados en los clusters para obtener un modelo final.



1. Evaluación del Modelo Combinado: Se evalúa el modelo final combinado en todo el conjunto de prueba y se calcula su precisión.



1. Visualización de Resultados: Se crea un gráfico de barras que muestra las predicciones combinadas y los valores reales en la muestra.

