PROYECTO FINAL

Pontificia Universidad
JAVERIANA
Cali

INTELIGENCIA ARTIFICIAL PARA ANÁLISIS DE DATOS

Daniel Cano Salgado
William Chavez Gonzalez
Alexander Valencia Altamirano
Carlos Ortega López

Ejercicio 1: Regresión

Predicción de Precios de Vehículos Usados

El presente proyecto tuvo como objetivo desarrollar un modelo capaz de predecir el precio de vehículos usados a partir de sus características, utilizando técnicas de regresión y el conjunto de datos "*Automobile Dataset*". A lo largo del análisis, se llevaron a cabo las siguientes etapas:

1. Análisis Exploratorio de Datos (EDA): Se realizó una exploración inicial del conjunto de datos, identificando valores faltantes en la variable `normalized-losses` y la presencia de *outliers* en variables como `price`, `horsepower` y `peak-rpm`. Además, se observó que existe una alta correlación entre el precio y variables como `curb-weight`, `engine-size` y `horsepower`.
2. Preprocesamiento de Datos: Se imputaron los valores faltantes del conjunto utilizando la mediana y se codificaron las variables categóricas mediante la estrategia de *One-Hot Encoding*. Asimismo, se estandarizaron las variables numéricas para asegurar una escala común.
3. Modelado y Evaluación: Se entrenaron tres modelos de regresión: *Regresión Lineal*, *Random Forest* y *Gradient Boosting*. Se utilizó validación cruzada en el proceso para evaluar su desempeño y se optimizaron los hiper parámetros del modelo *Random Forest* mediante el uso de *Grid Search*. Los resultados en el conjunto de prueba mostraron un rendimiento superior del modelo *Gradient Boosting*, con un RMSE de 2486.53, un MAE de 1821.95 y un R^2 de 0.87.
4. Importancia de Características: Se analizó la importancia de las características en cada modelo. En el caso de la *Regresión Lineal*, se observaron coeficientes significativos para los valores de `curb-weight`, `engine-size` y `horsepower`, lo que indica su fuerte influencia en el precio. Por otro lado, en los modelos de árbol (Random Forest y *Gradient Boosting*), las características más importantes fueron `horsepower`, `curb-weight`, `engine-size` y `width`. Estos resultados son consistentes con las correlaciones observadas en el EDA.

Conclusiones:

- El modelo *Gradient Boosting* demostró ser el más efectivo para predecir el precio de vehículos usados en este conjunto de datos, superando a la *Regresión Linea*l y *Random Forest* en términos de RMSE, MAE y R^2.
- Las características más relevantes para la predicción del precio son `horsepower`, `curb-weight`, `engine-size` y `width`. Esto sugiere que factores como la potencia del motor, el peso del vehículo, el tamaño del motor y el ancho son determinantes en la formación del precio.
- El análisis exploratorio de datos y la correcta selección de características son fundamentales para obtener modelos de predicción precisos.

Posibles Mejoras Futuras:

- Recolección de más datos: Tener acceso a un conjunto de datos más amplio y diverso podría mejorar la capacidad de generalización del modelo.
- Ingeniería de características: Se podrían crear nuevas características a partir de las existentes, como la relación peso/potencia, para capturar información adicional relevante para la predicción del precio.
- Exploración de otros modelos: Se podrían probar modelos más complejos, como redes neuronales, para evaluar si logran un mejor rendimiento que los modelos de regresión utilizados en este proyecto.
- Análisis de outliers: Se podría profundizar en el análisis de *outliers* para determinar si su eliminación o tratamiento mejora el rendimiento del modelo.

Ejercicio 2: Clasificación

Diagnóstico de Diabetes

El objetivo principal de este ejercicio fue desarrollar un modelo de clasificación para predecir la presencia de diabetes en pacientes, utilizando el conjunto de datos "*Pima Indians Diabetes Dataset*" y técnicas de aprendizaje automático en Python con scikit-learn. Se siguieron los siguientes pasos:

1. Análisis Exploratorio de Datos (EDA):
    - Se realizó una inspección inicial de los datos, verificando la presencia de valores faltantes (no se encontraron) y examinando estadísticas descriptivas (media, desviación estándar, percentiles) de las características.
    - Se visualizaron histogramas y diagramas de caja para comprender la distribución de cada variable y detectar posibles valores atípicos.
    - Se analizó la matriz de correlación entre las características, revelando relaciones moderadas entre algunas de ellas y la variable objetivo (Outcome).
2. Preprocesamiento de Datos:
    - Se trataron los valores atípicos utilizando el método del rango intercuartílico (IQR).
    - Se estandarizaron las características numéricas para que tuvieran una escala común.
3. Modelado y Evaluación:
    - Se dividió el conjunto de datos en dos partes: 80% para entrenamiento y 20% para prueba.
    - Se entrenaron tres modelos de clasificación:
        - Regresión Logística
        - Decision Tree
        - Random Forest
        - Random Forest con hiperparámetros optimizados
        - Random Forest con validación cruzada
    - Se calcularon métricas como accuracy, precisión, recall para comparar los modelos.
4. Importancia de Características:
    - Tras obtener los resultados del heatmap de correlación entre las diferentes variables, encontramos que aquellas cuya correlación fuese mayor a 0.2 son aquellas más fuertes y que podrían facilitar el descarte de una u otra. Por ejemplo la Age con Pregnancies que tiene una correlación mucho mayor de 0.54.
    - También nos permite ver la correlación con el Outcome (target) entre las cuales están BMI, Age y Glucose que fueron previamente identificadas como importantes para el modelo.

- Tras evaluar los feature_importances de un modelo basado en **DecisionTree**, la **Glucose**, BMI, **DiabetesPedigreeFuncton** y **Age** fueron evaluados como las variables más importantes, siendo **Glucose** la más importante. Por ende, se utilizaron en los modelos posteriores.

Conclusiones:

- De acuerdo con el fi de la Regresión Logística se puede evidenciar que los valores que se encuentran por encima de 0.01 son aquellas variables que más ingerencia e importancia tendrán al momento de clasificar la variable objetivo Outcome: **Pregnancies**, **Glucose**, **BMI**, **DiabetesPedigreeFunction** y **Age**.
- De acuerdo con la **matriz de confusión** presentada en el modelo de Regresión Logística, se puede evidenciar que el modelo es fuertemente sensible a detectar Verdaderos Positivos y tiene una aceptable capacidad de identificar también Verdaderos Negativos.
- Tras utilizar el modelo de RandomForestClassifier sin realizar optimización estructurada por hiperparámetros, pudimos evaluar con el auc y el accuracy que el modelo clasifica perfectamente en entrenamiento aunque en test disminuye su capacidad con un accuracy del 78%.
- Al realizar un cuarto modelo de RandomForest con validación cruzada en malla con GridSearchCV y un scoring roc_auc de área bajo la curva, se puede identificar que el modelo sigue siendo más equilibrado ajustado a un auc del 0.866 para test y un accuracy del 79.2%

Finalmente los siguientes modelos fueron comparados utilizando las métricas de **precision** y **recall**:

1. Decision Tree
2. Random Forest
3. Random Forest con hiperparámetros optimizados
4. Random Forest con validación cruzada

Conclusión: el modelo que más nos pareció estable en cuanto a ambas métricas fue el de **RandomForest sin optimización** ya que con un **0.5 de recall *_alcanza casi un_ *95% de precisión**, lo cual es el mejor balance para determinar la clasificación de **Diabetes**.