Universidad Católica San Pablo

Laboratorio: Usando Sklearn

Computer Science — Tópicos en Inteligencia Artificial

Harold Alejandro Villanueva Borda

---

1. Introducción

La regresión lineal es un método de inteligencia artificial que predice valores continuos al encontrar una relación lineal entre variables. Se usa para modelar y predecir tendencias en datos, como precios o comportamientos futuros. Este informe se implementa una solución para el análisis del dataset escogido para realizar la práctica de laboratorio, en donde se aplican diferentes conceptos de regresión lineal.

El objetivo es determinar el mejor modelo predictivo basado en el Error Cuadrático Medio (MSE) para la mejor configuración del experimento. Además, se incluye un análisis completo de los datos, el preprocesamiento y los resultados obtenidos. En este informe se hace uso de la biblioteca Sklearn que ofrece Python con el fin de comparar los trabajos anteriores en donde solo se usó Numpy y Pandas, se espera que el margen de difrencias en los resultados sean m+inimas.

2. Dataset

El conjunto de datos utilizado es *Credit Card Fraud Detection Dataset 2023*, extraído del repositorio Kaggle [1]. Este conjunto de datos contiene transacciones con tarjeta de crédito realizadas por titulares de tarjetas europeos en el año 2023. Comprende más de 550.000 registros.

2.1. Características Clave

- id: Identificador único para cada transacción
- V1-V28: Características anónimas que representan varios atributos de transacción (por ejemplo, hora, ubicación, etc.)
- Cantidad: El monto de la transacción
- Clase: Etiqueta binaria que indica si la transacción es fraudulenta (1) o no (0)

3. Evaluación de los datos

- **Datos categóricos:** La columna `Class` es categórica (binaria), indicando si la transacción es fraudulenta (1) o no (0). El resto de las columnas son numéricas.
- **Valores NaN:** No se encontraron valores NaN en el dataset. Esto facilita el preprocesamiento, evitando la necesidad de imputar o eliminar datos faltantes.
- **Normalización:** Se aplicó la normalización Min-Max a todas las columnas numéricas, excepto `id`, `Class` y `Amount`. La normalización es necesaria para que todas las características tengan la misma escala y no haya sesgos en el modelo debido a diferencias en las magnitudes de las variables.

3.1. Matriz de Correlación

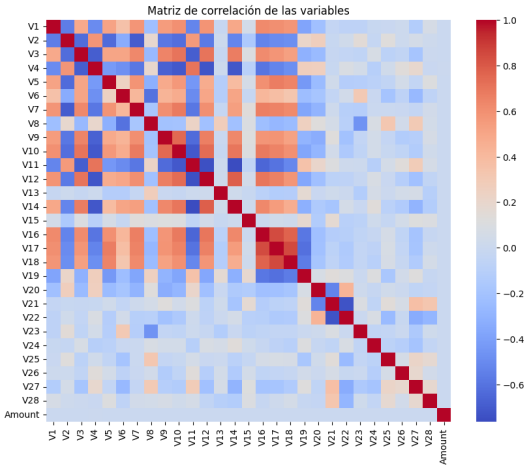

Figura 1: Matriz de correlación del dataset

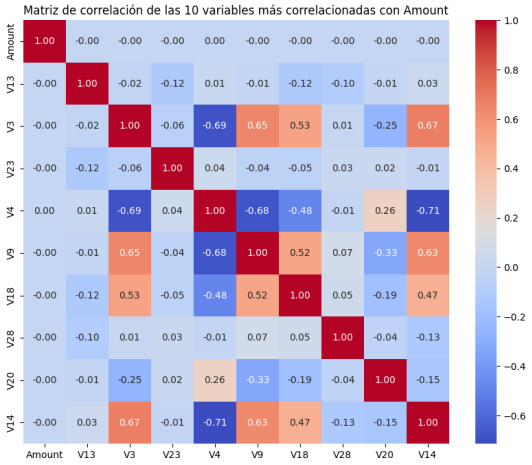

Figura 2: Matriz de correlación de las 10 variables más correlacionadas con Amount

- La primera matriz 1 muestra una baja correlación entre la mayoría de las variables, con pocas relaciones significativas. Las áreas rojas indican correlaciones positivas y las azules, negativas, pero en general, las relaciones son débiles.
- La segunda matriz 2, centrada en las 10 variables más correlacionadas con Amount, muestra que ninguna variable tiene una correlación fuerte con el monto. V13, V3 y V9 tienen correlaciones muy débiles, lo que indica que Amount no está fuertemente relacionado con ninguna variable individual en el dataset.

4. Experimentos y Resultados

En esta sección se evalúa el mejor modelo obtenido en los experimentos anteriores, que fue el experimento 3 con la configuración 80 %-20 %, tanto en la regresión lineal univariada como en la multivariada. Además, se verán las diferencias entre los experimentos en donde no se usó Sklearn y en los experimentos donde sí se usó.

4.1. Regresión Lineal Univaraiada

A continuación se mostrará los gráficos obetnidos del análisis de la regresión lineal univariada.

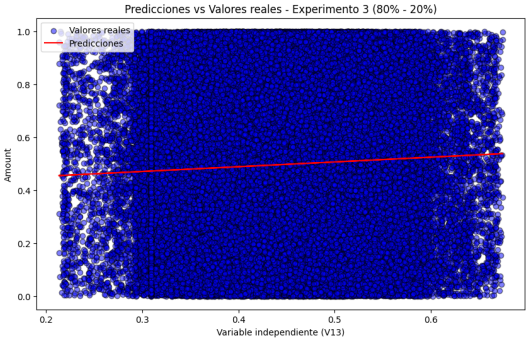

Figura 3: Gráfico obtenido sin el uso de Sklearn

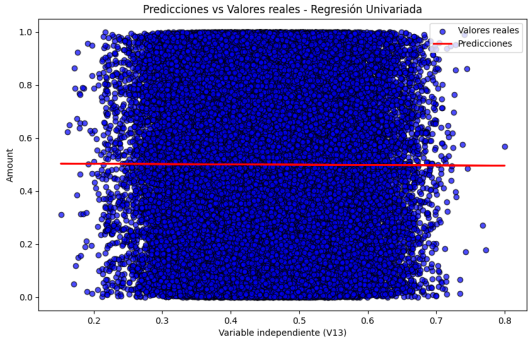

Figura 4: Gráfico obtenido haciendo uso de Sklearn

Sin usar Sklearn correspondientes al gráfico 3:

- ECM Experimento 3 (80 % - 20 %): 0.08306371826820705
- ECM en el conjunto de prueba: 0.08320944577546657

Usando Sklearn correspondientes al gráfico 4:

- ECM en entrenamiento: 0.08321529682275353
- ECM en prueba: 0.0831222626864957

Análisis comparativo

- El ECM del conjunto de entrenamiento es ligeramente menor que el del conjunto de prueba. Esto indica que el modelo está funcionando bien, ya que no muestra signos de sobreajuste (overfitting). La diferencia entre los errores es pequeña, lo que sugiere que el modelo generaliza de manera adecuada en el conjunto de prueba. La diferencia es pequeña, pero destaca que el modelo generado con Scikit-learn logra una ligera mejora en la precisión en el conjunto de prueba.

- Los ECM en los conjuntos de entrenamiento y prueba son muy similares. Esto indica que Scikit-learn está gestionando correctamente la regularización del modelo, lo que ayuda a evitar tanto el sobreajuste como el subajuste (underfitting). La diferencia mínima entre el ECM en entrenamiento y prueba sugiere que el modelo está bien equilibrado en términos de desempeño.

4.2. Regresión Lineal Multivaraiada

A continuación se mostrará los gráficos obetnidos del análisis de la regresión lineal multivariada.

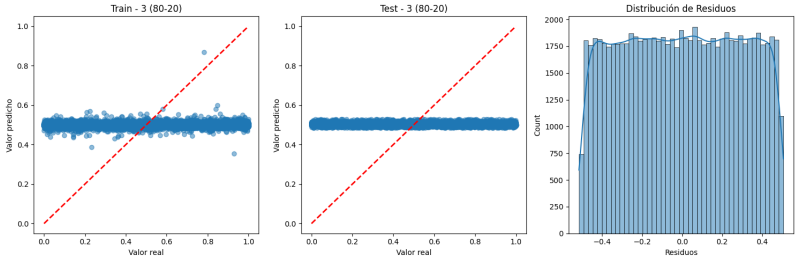

Figura 5: Gráfico obtenido sin el uso de Sklearn

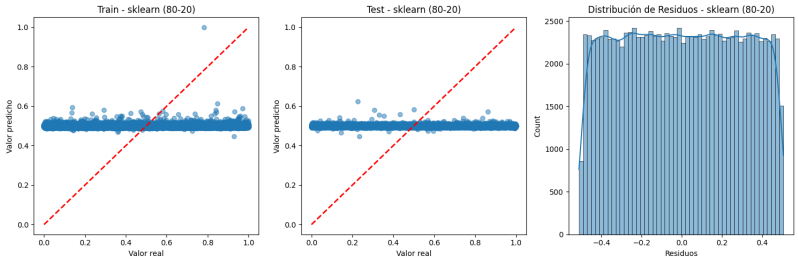

Figura 6: Gráfico obtenido haciendo uso de Sklearn

Sin usar Sklearn correspondientes al gráfico 5:

- MSE Train: 0.08336685409305604
- MSE Test: 0.08287106345311296
- Diferencia: 0.0004957906399430806

Usando Sklearn correspondientes al gráfico 6:

- MSE Train: 0.08320908581681044
- MSE Test: 0.08313167307966676
- Diferencia: 0.00007741273714368

Análisis comparativo

- En este caso, el ECM del conjunto de prueba es menor que el ECM del conjunto de entrenamiento. Esto podría indicar que el modelo sin Scikit-learn está generalizando bien, e incluso está funcionando ligeramente mejor en los datos de prueba que en los de entrenamiento. Sin embargo, una diferencia pequeña entre ambos ECMs puede ser una señal de suerte estadística, en lugar de sobreajuste (overfitting) o subajuste (underfitting).

- Al usar Scikit-learn, el MSE en el conjunto de entrenamiento es ligeramente menor que en el conjunto de prueba, lo cual es lo esperado en un buen modelo que no está sobreajustado. La diferencia entre ambos ECM es pequeña, lo que sugiere que Scikit-learn también está manejando correctamente el equilibrio entre el ajuste y la capacidad de generalización del modelo.

5. Conlusión

El análisis comparativo de la regresión lineal univariada y multivariada, tanto con como sin el uso de Scikit-learn, muestra resultados consistentes en términos de Error Cuadrático Medio (ECM). En ambos enfoques, los ECM de entrenamiento y prueba son muy similares, lo que indica que los modelos no presentan signos de sobreajuste ni subajuste.

- Sin usar Scikit-learn, los resultados indican una ligera ventaja en el conjunto de prueba, especialmente en el experimento multivariado, aunque esta diferencia podría atribuirse a variaciones estadísticas. El modelo parece generalizar bien sin mostrar un ajuste excesivo.

- Usando Scikit-learn, los ECM en entrenamiento y prueba son casi idénticos, lo que sugiere una mejor regularización del modelo. La implementación con Scikit-learn demuestra una mayor estabilidad y precisión en la generalización de los datos, minimizando la diferencia entre los conjuntos de entrenamiento y prueba.

6. Repositorio

La implementaci+on se encuentra en el siguiente respositorio: GitHub.

Referencias

[1] Nelgiriya Yewithana. *Credit Card Fraud Detection Dataset 2023*. Accessed: 2024-08-30. 2023. URL: https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023.