Reporte de Proyecto: Implementación de una Técnica de Aprendizaje Máquina sin Frameworks

By: Dafne Fernández Hernández A01369230

1. Introducción

El aprendizaje automático ha revolucionado el análisis de datos y la toma de decisiones en diversos campos, desde la medicina hasta la economía. En este proyecto, el enfoque fue implementar un modelo de regresión lineal desde cero utilizando Python, evitando el uso de bibliotecas de aprendizaje automático como scikit-learn o TensorFlow. Este ejercicio permite una comprensión más profunda de los fundamentos matemáticos y algoritmos que sustentan los modelos de aprendizaje automático.

El objetivo principal fue construir un modelo de regresión lineal simple para analizar y predecir la relación entre dos variables: delitos contra mujeres (AoW) y violencia doméstica (DV) en India, basándonos en datos obtenidos del dataset "Crimes Against Women in India (2001-2021)". Los resultados se evaluaron mediante el coeficiente de determinación ($\mathbf{R^2}$), proporcionando una medida de qué tan bien el modelo se ajusta a los datos observados.

---

2. Descripción del Dataset

2.1. Fuente de los Datos

El dataset utilizado en este proyecto proviene de Kaggle y contiene datos de crímenes cometidos contra mujeres en India durante el periodo 2001-2021. Este dataset es relevante para estudios sobre violencia de género y se emplea en este proyecto para investigar la relación entre dos tipos específicos de delitos reportados: delitos generales contra mujeres (AoW) y casos específicos de violencia doméstica (DV).

2.2. Descripción de las Variables

- AoW (Delitos contra Mujeres): Esta variable independiente (X) representa la cantidad de delitos reportados contra mujeres. Incluye diversos tipos de crímenes, como acoso sexual, abuso, y otros actos violentos que afectan a las mujeres.
- DV (Violencia Doméstica): La variable dependiente (y) representa los casos reportados de violencia doméstica, un subtipo específico de delitos contra mujeres que ocurre en el entorno del hogar.

El objetivo del análisis es determinar cómo los delitos generales contra mujeres influyen en la incidencia de violencia doméstica, y si es posible predecir los casos de violencia doméstica basándose en los datos de delitos reportados.

---

3. Implementación del Modelo

3.1. Regresión Lineal Simple

La regresión lineal es una técnica estadística fundamental utilizada para modelar la relación entre una variable

independiente $\boldsymbol{X}$ y una variable dependiente $\boldsymbol{y}$. La regresión lineal simple se basa en la suposición de que existe una relación lineal entre las dos variables, que puede ser descrita por la ecuación de una línea recta:

$$\boldsymbol{y = mx + b}$$

Donde:

- $\boldsymbol{m}$ representa la pendiente de la línea, indicando el cambio en $\boldsymbol{y}$ por unidad de cambio en $\boldsymbol{X}$.
- $\boldsymbol{b}$ es la intersección con el eje $\boldsymbol{y}$, indicando el valor de $\boldsymbol{y}$ cuando $\boldsymbol{X = 0}$

La pendiente $\boldsymbol{m}$ refleja la dirección y la magnitud de la relación entre las variables, mientras que la intersección $\boldsymbol{b}$ proporciona un punto de referencia de donde comienza la relación.

3.2. Método de Descenso de Gradiente

Para ajustar los parámetros $\boldsymbol{m}$ y $\boldsymbol{b}$, se utilizó el método de descenso de gradiente. Este es un algoritmo iterativo que minimiza la función de costo, en este caso el error cuadrático medio (MSE), que mide la diferencia promedio al cuadrado entre los valores observados y los valores predichos:

$$\boldsymbol{Error = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (mx_i + b)\right)^2}$$

El descenso de gradiente actualiza los valores de $\boldsymbol{m}$ y $\boldsymbol{b}$ en cada iteración, moviéndose en la dirección del gradiente negativo de la función de costo para encontrar el mínimo local:

$$\boldsymbol{m = m - \alpha \left(\frac{\partial Error}{\partial m}\right) b}$$
$$\boldsymbol{= b - \alpha \left(\frac{\partial Error}{\partial b}\right)}$$

Donde $\boldsymbol{\alpha}$ es la tasa de aprendizaje, un parámetro que controla el tamaño de los pasos dados hacia el mínimo de la función de costo.

3.3. Evaluación del Modelo: Coeficiente de Determinación (R^2)

El coeficiente de determinación $\boldsymbol{R^2}$ se utiliza para evaluar el rendimiento del modelo. Es una medida que indica la proporción de la varianza en la variable dependiente que es predecible a partir de la variable independiente:

$$\boldsymbol{R^2 = 1 - \frac{SSR}{SST}}$$

Donde:

$\boldsymbol{SSR}$ es la suma de los residuos al cuadrado:, $\boldsymbol{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, que representa la variabilidad de los errores de predicción.

- $\boldsymbol{SSR}$ es la suma total de cuadrados: $\boldsymbol{\sum_{i=1}^{n}(y_i - \bar{y})^2}$,, que representa la variabilidad total en los datos observados.

Un $\boldsymbol{R^2}$ cercano a 1 indica un modelo que se ajusta bien a los datos, mientras que un valor cercano a 0 sugiere un modelo pobre.

---

4. Resultados y Análisis

4.1. Parámetros del Modelo

Después de entrenar el modelo utilizando el método de descenso de

gradiente, se obtuvieron los siguientes parámetros:

- Pendiente (m): **624.89**
- Intersección (b): **642.21**
- Coeficiente de Determinación (R^2): **0.361**

4.2. Interpretación de los Resultados

- **Pendiente (m):** La pendiente indica que por cada incremento unitario en los delitos contra mujeres (AoW), se espera un aumento de aproximadamente 624.89 casos de violencia doméstica (DV). La intersección de 642.21 sugiere que, incluso en ausencia de delitos reportados, se esperarían alrededor de 642.21 casos de violencia doméstica. Esto puede reflejar una base de violencia doméstica inherente o factores no capturados en los datos.
- **Coeficiente de Determinación (R^2):** Un $\boldsymbol{R^2}$ de 0.361 indica que el modelo explica el 36.1% de la variabilidad en los casos de violencia doméstica. Esto sugiere que el modelo tiene un ajuste relativamente bajo, lo que significa que la relación entre los delitos contra mujeres y la violencia doméstica es menos fuerte en comparación con otros posibles modelos.

4.3. Predicciones del Modelo

- Predicciones para nuevos datos: [120.55, 120.85]

Las predicciones para nuevos datos muestran valores relativamente bajos, lo cual puede ser indicativo de un modelo que podría no capturar adecuadamente todas las complejidades del problema.

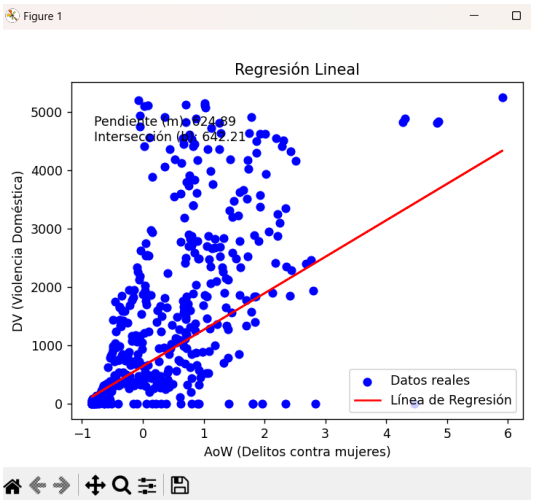
4.4. Errores de Epochs

Los errores de las épocas durante el entrenamiento muestran una disminución progresiva, indicando que el modelo está mejorando su ajuste a los datos. Sin embargo, los errores iniciales son elevados, lo que puede ser causado por varios factores:

- **Tasa de Aprendizaje:** Si la tasa de aprendizaje $\boldsymbol{\alpha}$ es demasiado alta, el modelo puede estar realizando grandes saltos en el espacio de parámetros, lo que puede llevar a una convergencia lenta y errores elevados en las primeras épocas. Ajustar la tasa de aprendizaje podría mejorar el proceso de ajuste.
- **Inicialización de Parámetros:** La elección de valores iniciales para los parámetros $\boldsymbol{m}$ y $\boldsymbol{b}$ puede influir en la rapidez con la que el modelo converge. Inicializaciones no óptimas pueden llevar a una mayor magnitud de errores al principio.
- **Complejidad del Modelo:** El modelo de regresión lineal simple puede no ser suficiente para capturar la relación compleja entre delitos contra mujeres y violencia doméstica, especialmente si hay factores no lineales o interacciones entre variables no consideradas.

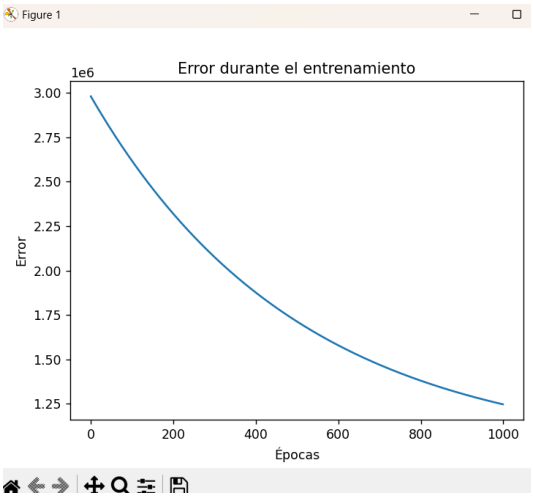
4.5. Interpretación de las Gráficas

- **Gráfica de Ajuste del Modelo:**



La gráfica muestra la línea de regresión ajustada en comparación con los datos reales. Aunque la línea sigue la tendencia general, la dispersión de los puntos alrededor de la línea indica que el modelo no captura completamente todas las variaciones en los datos. Esto es consistente con el bajo coeficiente de determinación (R^2) obtenido.

- **Gráfica de Errores de Epochs:**



La gráfica de errores de épocas muestra la evolución del error cuadrático medio (MSE) durante el entrenamiento. La tendencia general es una disminución en el error, indicando que el modelo está aprendiendo y mejorando su ajuste con el tiempo. Sin embargo, los errores altos en las primeras épocas reflejan que el modelo está en proceso de ajuste y que las tasas de aprendizaje y inicialización pueden estar afectando el rendimiento.

---

5. Análisis Crítico y Mejoras Futuras

5.1. Limitaciones del Modelo

Aunque la implementación de la regresión lineal proporciona un enfoque inicial para modelar la relación entre delitos contra mujeres y violencia doméstica, hay varias limitaciones a considerar:

- **Simplificación de la Realidad:** La regresión lineal simple asume una relación lineal directa entre las variables, lo cual puede ser una simplificación excesiva de la realidad. La violencia doméstica es un fenómeno complejo influenciado por múltiples factores, y un modelo más sofisticado podría ser necesario.
- **Errores en Datos:** La calidad de los datos puede afectar la precisión del modelo. Los errores iniciales altos pueden reflejar problemas en la calidad o integridad de los datos.

5.2. Mejoras Propuestas

Para mejorar el modelo y la calidad de las predicciones, se podrían considerar las siguientes mejoras:

- **Modelos Multivariables:** Ampliar el modelo para incluir variables adicionales podría ayudar a capturar mejor la complejidad de la relación entre los delitos contra mujeres y la violencia doméstica.
- **Técnicas Avanzadas:**Considerar modelos más complejos como la regresión polinómica o técnicas de aprendizaje automático como los árboles de decisión y redes neuronales podría proporcionar un mejor ajuste a los datos.
- **Análisis de Series Temporales**: Si los datos están disponibles en forma de series temporales, se podría aplicar un análisis de series temporales para capturar tendencias, estacionalidades y fluctuaciones cíclicas que pueden influir en la relación entre delitos contra mujeres y violencia doméstica. Este tipo de análisis podría permitir un modelado más preciso y una mejor predicción futura.
- **Validación Cruzada:** Implementar técnicas de validación cruzada para evaluar el rendimiento del modelo de manera más robusta. La validación cruzada podría ayudar a evitar el sobreajuste y proporcionar una estimación más precisa de la capacidad del modelo para generalizar a nuevos datos.
- **Análisis de Outliers**: Realizar un análisis de outliers para identificar y manejar datos atípicos que podrían estar influyendo desproporcionadamente en los resultados del modelo. El tratamiento adecuado de los outliers podría mejorar la precisión del modelo.

- **6. Conclusión**
- Este proyecto ha permitido implementar y comprender en profundidad un modelo de regresión lineal simple sin la ayuda de frameworks o bibliotecas avanzadas, centrándose en los fundamentos matemáticos del aprendizaje automático. El modelo desarrollado proporciona una herramienta básica para analizar la relación entre delitos contra mujeres y violencia doméstica en India. Sin embargo, la moderada capacidad explicativa del modelo, como lo indica el coeficiente de determinación ($\boldsymbol{R^2 = 0.5866}$), sugiere que es necesario un análisis más profundo y modelos más complejos para capturar la verdadera naturaleza de esta relación.
- El proceso de desarrollo y análisis de este modelo ha destacado la importancia de comprender no solo la implementación técnica de un algoritmo, sino también las

implicaciones y limitaciones del modelo al aplicarlo a problemas del mundo real. Para futuros proyectos, se recomienda la incorporación de técnicas de modelado más avanzadas, un análisis más exhaustivo de los datos y la consideración de múltiples variables para mejorar la precisión y utilidad de las predicciones.

- **7. Referencias**

- BALAJI. (2021). *Crimes Against Women in India (2001-2021).* Kaggle.com. https://www.kaggle.com/datasets/balajivaraprasad/crimes-against-women-in-india-2001-2021
- Team, C. (2022, October 25). *NCRB Report 2021: Crime In India - ClearIAS*. ClearIAS. https://www.clearias.com/ncrb-report-2021/
- Crypto1. (2020, October 2). *Gradient Descent Algorithm: How Does it Work in Machine Learning?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/how-does-the-gradient-descent-algorithm-work-in-machine-learning/
- Bobbitt, Z. (2020, October 23). *How to Find Coefficient of Determination (R-Squared) in R*. Statology. https://www.statology.org/r-squared-in-r/