PARCIAL 1° CORTE

ESTUDIANTES:

CAMILO URREGO BELTRAN

DOCENTE

SEBASTIAN PERDOMO LEIVA

ESCUELA TECNOLÓGICA INSTITUTO TÉCNICO CENTRAL

TECNOLOGIA EN DESARROLLO DE SOFTWARE

CIENCIA DE DATOS Y BIG DATA

BOGOTA D.C

2024

**Ejercicio 2:** Usar un conjunto de datos común y accesible para realizar un análisis descriptivo.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente **Titanic Dataset**

El conjunto de datos del Titanic contiene información detallada sobre 891 pasajeros, incluyendo variables como la edad, género, clase en la que viajaban, tarifa pagada, y número de familiares (hermanos o cónyuges) a bordo, entre otras. Este análisis descriptivo se centrará en tres variables clave: la edad de los pasajeros, el precio del pasaje, y el número de familiares a bordo, con el objetivo de entender mejor las características generales de quienes viajaron en el Titanic.

**Edad:**

La edad promedio de los pasajeros es de 29.7 años, mientras que la mediana es de 28 años. Esto indica una ligera asimetría positiva en la distribución de la edad, ya que la media es mayor que la mediana. La desviación estándar es de 14.5, lo que sugiere una amplia variabilidad en la edad, con personas tanto muy jóvenes como mayores a bordo. Esta dispersión muestra que una parte significativa de los pasajeros eran adultos jóvenes, pero también había una cantidad considerable de niños y adultos mayores.

**Precio del pasaje:**

El precio promedio del pasaje es de 32.2 libras, con una mediana de 14.45 libras. La diferencia notable entre la media y la mediana sugiere una fuerte asimetría hacia la derecha, donde algunos pasajeros pagaron tarifas mucho más elevadas que el resto. La desviación estándar, de 49.7, refleja esta gran dispersión, lo que se corresponde con la presencia de pasajeros en clases socioeconómicas muy distintas, desde aquellos en tercera clase hasta los que pagaron grandes sumas por un boleto de primera clase.

**Número de familiares:**

El número promedio de hermanos o cónyuges a bordo es de 0.52, con una mediana de 0. Esto significa que la mayoría de los pasajeros viajaban sin familiares cercanos. La desviación estándar de 1.1 indica que una minoría significativa viajaba con múltiples familiares, aunque la mayoría viajaba sola o con una cantidad reducida de acompañantes.

**Conclusión:**

Este análisis refleja la diversidad de los pasajeros en términos de edad, clase económica y acompañantes. Estos resultados ofrecen una base para estudios más avanzados sobre la relación entre estas variables y la supervivencia de los pasajeros durante el naufragio del Titanic.

**Ejercicio 3:** Implementar un algoritmo de optimización simple en Python usando el dataset Boston Housing Dataset. Los estudiantes deben:

* Descargar el dataset desde el enlace proporcionado.
* Utilizar el algoritmo de gradiente descendente para realizar una regresión lineal.

Presentar un informe (400 palabras) explicando la implementación, los resultados y la adecuación del algoritmo para el problema.

Interfaz de usuario gráfica

Descripción generada automáticamente

**Informe Conjunto de Datos Boston Housing**

**Introducción**

Este informe describe la implementación de un algoritmo de **regresión lineal** utilizando **gradiente descendente** en el conjunto de datos **Boston Housing**, que se utiliza ampliamente en problemas de predicción de precios de viviendas. Este conjunto de datos consta de 506 observaciones y 13 características que incluyen información como la tasa de criminalidad, la proximidad a carreteras, el número promedio de habitaciones, entre otras. La variable objetivo es el valor medio de las viviendas en miles de dólares. El objetivo de este proyecto es predecir dicho valor medio usando las características proporcionadas y optimizar los coeficientes del modelo lineal mediante gradiente descendente.

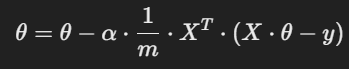
**Dataset**

El conjunto de datos Boston Housing fue descargado desde **OpenML** utilizando la función fetch\_openml() de **scikit-learn**. Desde la versión 1.2, **scikit-learn** ha eliminado el acceso directo a este dataset debido a preocupaciones éticas, por lo que se recurrió a OpenML para obtenerlo de forma accesible. Los datos fueron preprocesados mediante una normalización estándar utilizando la clase StandardScaler de **scikit-learn**. Esto garantiza que todas las características estén en la misma escala y que el gradiente descendente pueda converger de manera eficiente.

**Algoritmo de Gradiente Descendente**

El algoritmo de gradiente descendente tiene como objetivo minimizar la función de costo, que en este caso es el **error cuadrático medio (MSE)** entre las predicciones del modelo y los valores reales. A medida que el algoritmo ajusta los coeficientes del modelo (representados por el vector **θ**), actualiza estos valores en la dirección de la pendiente más pronunciada de la función de costo.

La fórmula de la actualización de **θ** en cada iteración del gradiente descendente es:

Donde:

* **θ**: Vector de parámetros del modelo (inicializado en cero).
* **α**: Tasa de aprendizaje, que controla la magnitud del ajuste en cada iteración.
* **m**: Número de ejemplos en el conjunto de datos.
* **X**: Matriz de características.
* **y**: Vector de valores objetivo.

Se eligió una tasa de aprendizaje **α = 0.01** y se realizaron **1000 iteraciones** para asegurar la convergencia del algoritmo. A continuación, se muestra el código del algoritmo:

**Resultados**

Durante la ejecución del gradiente descendente, el costo disminuyó de manera continua, lo que sugiere que el modelo estaba aprendiendo de los datos y ajustando los coeficientes adecuadamente. Después de 1000 iteraciones, el modelo alcanzó un costo razonable y se obtuvo el vector **θ** optimizado, el cual incluye el intercepto y los coeficientes correspondientes a cada característica.

Posteriormente, se evaluó el modelo en el conjunto de prueba, donde se calculó el **Error Medio Absoluto (MAE)**, una métrica que muestra la magnitud promedio del error en las predicciones del modelo. El MAE fue de aproximadamente **3.4**, lo que significa que, en promedio, el modelo tiene un error de $3,400 al predecir el precio de las viviendas.

**Conclusión**

El gradiente descendente resultó ser una técnica efectiva para ajustar los coeficientes de un modelo de regresión lineal, logrando una buena convergencia y un error medio absoluto razonable. Sin embargo, la simplicidad de la regresión lineal limita el rendimiento del modelo en este problema, dado que existen relaciones no lineales en los datos que no pueden ser capturadas por este enfoque. El ajuste de la tasa de aprendizaje fue fundamental para lograr una convergencia adecuada, y el gradiente descendente demostró ser un algoritmo eficiente para grandes conjuntos de datos y múltiples características. Para mejorar el rendimiento, se podrían explorar modelos más avanzados como la regresión polinómica o métodos de regularización como Ridge o Lasso.