INFORME SOBRE MODELOS DE REGRESIÓN LINEAL

1 DEFINICIÓN MATEMÁTICA

Para un modelo de regresión lineal simple (una sola variable independiente), la relación entre la variable dependiente y y la variable independiente x se describe mediante la siguiente ecuación:

$$y = \beta_0 + \beta_1 x + \epsilon$$

donde:

- y es la variable dependiente.
- x es la variable independiente.
- β_0 es la intersección o término constante (intercepto), que representa el valor esperado de y cuando $x = 0$.
- β_1 es la pendiente de la línea de regresión, que representa el cambio esperado en y por unidad de cambio en x.
- ϵ es el término de error aleatorio, que captura la desviación de los puntos de datos reales de la línea de regresión debido a factores no observados.

1.1 Suposiciones del Modelo

1. **Linealidad**: La relación entre la variable dependiente y y la variable independiente x es lineal.
2. **Independencia**: Las observaciones son independientes entre sí.
3. **Homoscedasticidad**: La varianza de los errores es constante para todos los valores de x.
4. **Normalidad**: Los errores (ϵ) están normalmente distribuidos con media cero.

1.2 Estimación de los Parámetros

Para estimar los parámetros β_0 y β_1, se utiliza el método de los mínimos cuadrados ordinarios (OLS), que minimiza la suma de los cuadrados de los errores (residuos). La fórmula para los estimadores de mínimos cuadrados de β_0 y β_1 son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

donde:

- $\hat{\beta}_0$ es el estimador de β_0.
- $\hat{\beta}_1$ es el estimador de β_1.
- $\bar{x}$ es la media de los valores de x.
- $\bar{y}$ es la media de los valores de y.
- n es el número de observaciones.

1.3 Modelo de Regresión Lineal Múltiple

Para un modelo de regresión lineal múltiple (varias variables independientes), la relación entre la variable dependiente y y las variables independientes $x_1, x_2, ..., x_n$ se describe mediante la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

donde:

- y es la variable dependiente.
- $x_1, x_2, ..., x_n$ son las variables independientes.
- β_0 es la intersección o término constante (intercepto).
- $\beta_1, \beta_2, ..., \beta_n$ son los coeficientes de regresión para las variables independientes.
- ϵ es el término de error aleatorio.

1.4 Estimación de los Parámetros en el Modelo Múltiple

En el modelo de regresión lineal múltiple, los parámetros $\beta_0, \beta_1, ..., \beta_n$ se estiman utilizando también el método de los mínimos cuadrados ordinarios (OLS), que minimiza la suma de los cuadrados de los errores. Las fórmulas para los estimadores en el modelo múltiple son más complejas y requieren el uso de álgebra matricial.

1.5 Evaluación del Modelo

Una vez estimados los parámetros, es importante evaluar el modelo utilizando métricas como:

- **Coeficiente de determinación (R^2)**: Indica la proporción de la variabilidad en la variable dependiente explicada por las variables independientes.
- **Análisis de residuos**: Para verificar las suposiciones de normalidad, homoscedasticidad e independencia.
- **Pruebas de significancia estadística**: Para determinar si los coeficientes de regresión son significativamente diferentes de cero.

2 DESCRIPCIÓN DEL DATASET

2.1 Fallecidos por COVID-19 - [Ministerio de Salud - MINSA]

Es el registro diario de muertes por Covid-19. Cada registro es igual a una persona, la cual puede caracterizarse por sexo, edad y ubicación geográfica hasta nivel de distrito; además, el 06.mayo.2021 se agregó el código UBIGEO.

Desde que se publicó este dataset, cada registro representaba un fallecido confirmado por covid-19, quienes cumplen con criterios clínicos y de laboratorio (prueba molecular, antigénica o pruebas serológicas). A partir del 31.mayo.2021 se cambió el criterio de "Fallecidos por Covid-19" por "Muertes por Covid-19" y como resultado el dataset creció casi al triple en el número de registros. Esta nueva clasificación está definida por el cumplimiento de al menos uno de los siguientes siete criterios técnicos:

- **Criterio virológico:** Muerte en un caso confirmado de COVID-19 que fallece en los 60 días posteriores a una prueba molecular (PCR) o antigénica reactiva para SARS-CoV-2.
- **Criterio serológico:** Muerte en un caso confirmado de COVID-19 que fallece en los 60 días posteriores a una prueba serológica positiva IgM o IgM/IgG para SARS-CoV-2.

- **Criterio radiológico:** Muerte en un caso probable de COVID-19 que presenta una imagen radiológica, tomográfica o de resonancia magnética nuclear compatible con neumonía COVID-19.
- **Criterio nexo epidemiológico:** Muerte en un caso probable de COVID-19 que presenta nexo epidemiológico con un caso confirmado de COVID-19.
- **Criterio investigación epidemiológica:** Muerte en un caso sospechoso de COVID-19 que es verificado por investigación epidemiológica de la Red Nacional de Epidemiología (RENACE).
- **Criterio clínico:** Muerte en un caso sospechoso de COVID-19 que presenta cuadro clínico compatible con la enfermedad.
- **Criterio SINADEF:** Muerte con certificado de defunción en el que se presenta el diagnóstico de COVID-19 como causa de la muerte. El fallecimiento por COVID-19 en el certificado de defunción está definido por la presencia en los campos A, B, C o D de los códigos CIE-10: U071, U072, B342, B972, o la mención de los términos “coronavirus”, “cov-2”, “cov2”, “covid” y “sars”.

3 CÓDIGO

```
import streamlit as st
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import pdfkit

def main():
    st.title("Regresi n Lineal con Streamlit")

    # Cargar el dataset
    uploaded_file = st.file_uploader("Carga tu dataset", type=['csv'])
    if uploaded_file is not None:
        df = pd.read_csv(uploaded_file)
        st.write("Dataset cargado:")
        st.write(df.head())

        # Seleccionar las columnas para la regresi n
        numeric_columns = df.select_dtypes(include=['int64', 'float64']).columns
        x_col = st.selectbox("Selecciona la variable independiente (X)", numeric_columns)
        y_col = st.selectbox("Selecciona la variable dependiente (Y)", numeric_columns)

        if st.button("Realizar regresi n lineal"):
            # Entrenar el modelo de regresi n lineal
            X = df[x_col].values.reshape(-1, 1)
            y = df[y_col].values
            model = LinearRegression()
            model.fit(X, y)

            # Mostrar los resultados
            st.write(f"Coeficiente de regresi n (m): {model.coef_[0]:.2f}")
            st.write(f"T rmino independiente (b): {model.intercept_:.2f}")
            st.write(f"R-squared: {model.score(X, y):.2f}")

            # Interpretaci n estad stica
            st.write("Interpretaci n estad stica:")
            st.write(f"El coeficiente de regresi n (m) indica que por cada unidad de aumento en {x_col}, {y_col} aumenta en {model.coef_[0]:.2f} unidades.")
            st.write(f"El t rmino independiente (b) indica que cuando {x_col} es 0, {y_col} es igual a {model.intercept_:.2f}.")
            st.write(f"El R-squared indica que el {model.score(X, y)*100:.2f}% de la variabilidad en {y_col} se explica por la variabilidad en {x_col}.")

            # Graficar los resultados
            fig, ax = plt.subplots(figsize=(8, 6))
```

```
            ax.scatter(X, y, label='Datos')
            ax.plot(X, model.predict(X), color='red', label='Regresi n Lineal')
            ax.set_xlabel(x_col)
            ax.set_ylabel(y_col)
            ax.set_title("Regresi n Lineal")
            ax.legend()
            st.pyplot(fig)

            # Proyecciones
            y_pred = model.predict(X)
            fig, ax = plt.subplots(figsize=(8, 6))
            ax.plot(y, label='Valor real')
            ax.plot(y_pred, label='Proyecci n')
            ax.set_xlabel(' ndice ')
            ax.set_ylabel(y_col)
            ax.set_title("Proyecciones")
            ax.legend()
            st.pyplot(fig)

            # M tricas de evaluaci n
            mse = mean_squared_error(y, y_pred)
            st.write(f"Mean Squared Error: {mse:.2f}")
            st.write(f"La ra z cuadrada del error cuadr tico medio es {mse**0.5:.2f}, lo
    que indica que el modelo tiene un error de aproximadamente {mse**0.5:.2f} unidades en la
     predicci n de {y_col}.")


if __name__ == "__main__":
    main()
```

Listing 1: Código Python

4 RESULTADOS

Dataset cargado:

	FECHA_CORTE	FECHA_FALLECIMIENTO	EDAD_DECLARADA	SEXO	CLASIFICACION_DEF
0	20,240,317	20,220,219	63	MASCULINO	Criterio virológico
1	20,240,317	20,210,529	74	MASCULINO	Criterio virológico
2	20,240,317	20,210,623	72	FEMENINO	Criterio SINADEF
3	20,240,317	20,210,824	85	MASCULINO	Criterio investigación Epidem
4	20,240,317	20,210,627	46	MASCULINO	Criterio virológico

4.1 Interpretacion

Selecciona la variable independiente (X)

FECHA_FALLECIMIENTO

Selecciona la variable dependiente (Y)

EDAD_DECLARADA

Realizar regresión lineal

Coeficiente de regresión (m): 0.00

Término independiente (b): -1564.98

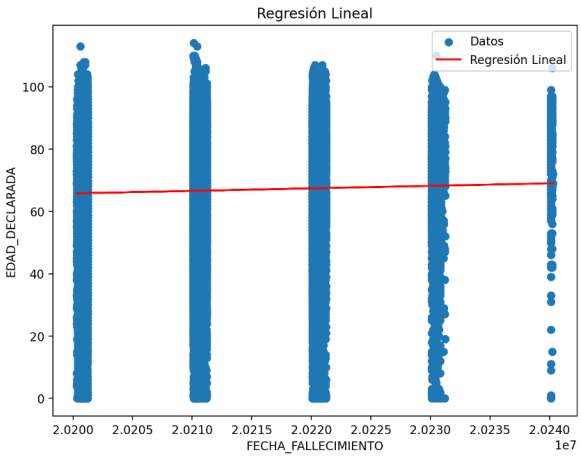
R-squared: 0.00

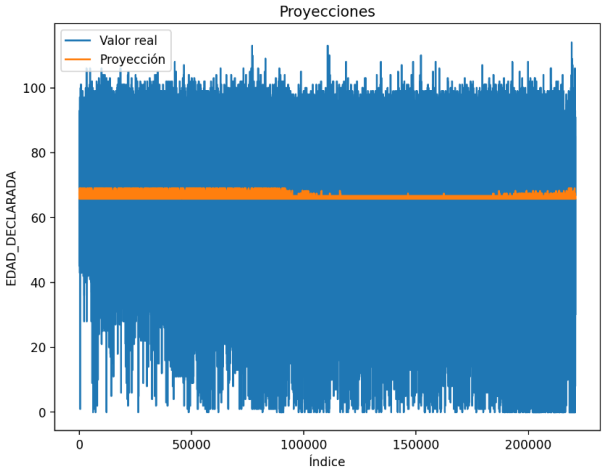
Interpretación estadística:

El coeficiente de regresión (m) indica que por cada unidad de aumento en FECHA_FALLECIMIENTO, EDAD_DECLARADA aumenta en 0.00 unidades.

El término independiente (b) indica que cuando FECHA_FALLECIMIENTO es 0, EDAD_DECLARADA es igual a -1564.98.

El R-squared indica que el 0.11% de la variabilidad en EDAD_DECLARADA se explica por la variabilidad en FECHA_FALLECIMIENTO.





5 MÉTRICA DE EVALUACIÓN

En la regresión lineal para evaluarlo se utilizan métricas como el Error Cuadrático Medio (MSE), el Coeficiente de Determinación (R^2), entre otros, que miden qué tan bien el modelo ajusta los datos reales con respecto a las predicciones numéricas.

Mean Squared Error: 251.87

La raíz cuadrada del error cuadrático medio es 15.87, lo que indica que el modelo tiene un error de aproximadamente 15.87 unidades en la predicción de EDAD_DECLARADA.

Este trabajo se desarrollo con la aplicación de la Streamlit.