14/06/2024

TRABAJO PRACTICO

1 Cuatrimestre ISTEA
Materia APRENDIZAJE AUTOMATICO I

Federico Gauna
Moreno Priscila
Juan Fernando Torossi
María Victoria Barreto Rojas

INTRODUCCION

En este trabajo práctico, nos enfocaremos en la relación entre el nivel de analfabetismo y la mortalidad infantil en las distintas provincias de Argentina.
El objetivo principal de este trabajo es desarrollar un algoritmo en Python que permita estimar y analizar la relación entre estas dos variables. Para ello, se emplearán dos enfoques de modelado:

1- Modelo de regresión lineal: Este modelo permitirá representar la relación entre el nivel de analfabetismo (variable objetivo) y la mortalidad infantil (variable observable) mediante una línea recta. La regresión lineal es un método estadístico sencillo pero poderoso para identificar y cuantificar la relación entre dos variables.

2- Modelos de regresión polinómica: Para capturar posibles relaciones no lineales entre las variables, se utilizarán modelos de regresión polinómica de diferentes grados (2, 3, 4 y 5). Estos modelos permiten ajustar curvas más complejas y podrían ofrecer una mejor representación de la relación entre el analfabetismo y la mortalidad infantil si la relación subyacente es más complicada que una simple línea recta.

El desarrollo de estos modelos no solo permitirá evaluar la relación entre las variables de interés, sino que también proporcionará una herramienta analítica que puede ser utilizada para futuros estudios. A continuación, se detallarán los pasos específicos para la implementación y evaluación de los modelos mencionados, así como los resultados obtenidos de su aplicación a los datos de las provincias argentinas.

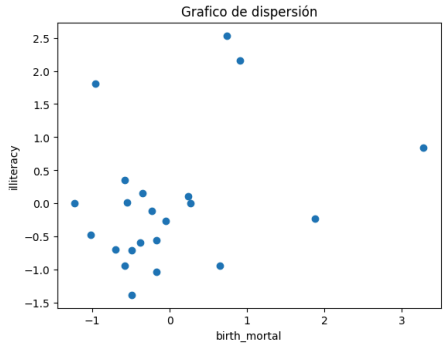
DESARROLLO

En la primera etapa para lograr el análisis de los datos y lograr realizar las conclusiones fue realizar la **“Carga y Preprocesamiento de Datos”**. Como primer paso, realizamos la exploración inicial de datos y detectamos datos faltantes (NaN). Para subsanar estos datos calculamos la media de las columnas numéricas y las rellenamos con la media. Luego utilizando “StandadScaler” logramos estandarizar las columnas “illiteracy” (variable objetivo) y “birth_mortal” (variable observable). Esto asegura que los valores tengan la misma escala.
La segunda etapa **“Creación del Modelo de Regresión”**, realizamos la partición del conjunto de datos, en conjuntos de entrenamiento y prueba. Se utiliza un tamaño de prueba del 40% y un 60% para el entrenamiento y una semilla aleatoria (random_state=1) para garantizar la reproducibilidad.

Definimos nuestras variables:

'x' representa la variable independiente (en este caso, “birth_mortal”).
'y' representa la variable dependiente (en este caso, “illiteracy”).

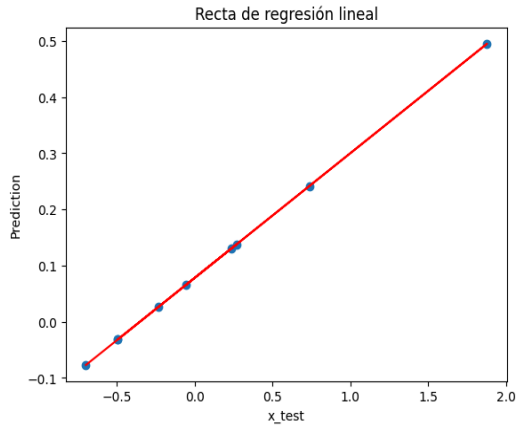
Generamos un primer gráfico de dispersión para ver la relación entre 'birth_mortal' e 'illiteracy'.



Al observar este gráfico de dispersión, no existe una tendencia definida entre las dos variables. Los puntos se encuentran dispersos sin forma o dirección específica. Hay cierta concentración de puntos en la parte inferior izquierda del gráfico y algunos puntos más dispersos en la parte superior derecha. Se puede concluir que no hay una relación lineal fuerte entre la tasa de mortalidad infantil y la tasa de analfabetismo en el conjunto de datos. Esto puede sugerir que el modelo de regresión lineal ***no es el adecuado*** *para lograr predecir la tasa de analfabetismo a partir de la tasa de mortalidad infantil.*

Modelo de Regresión Lineal

Creamos un gráfico que muestra los datos de prueba y la recta de regresión. Evaluamos el modelo calculando las métricas de error, Error Absoluto Promedio (**MAE**), el Error Cuadrático Promedio (**MSE**) y la Raíz Cuadrada del Error Cuadrático Medio (**RMSE**), así como el puntaje **R2**, para evaluar el desempeño del modelo.

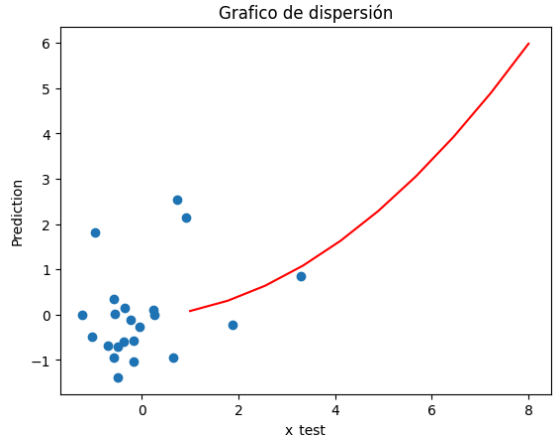


Los puntos azules representan los valores reales del conjunto de prueba ('x_test'). Cada punto muestra el valor de 'x_test' en el eje horizontal y la correspondiente predicción del modelo en el eje vertical. La recta de color rojo representa la relación lineal que el modelo ha aprendido de los datos de entrenamiento. Los puntos azules se encuentran relativamente cerca de la línea roja, lo que indica que el modelo de regresión lineal ha logrado un buen ajuste a los datos de prueba. La línea roja tiene pendiente positiva, lo que sugiere que existe una relación positiva entre la variable independiente y la variable objetivo. Las predicciones del modelo se encuentran relativamente cerca de los valores reales del conjunto de prueba, lo que indica que el modelo es capaz de predecir razonablemente bien los valores de la variable objetivo.

```
The mean absolte error (MAE) on test set: 0.70050
The mean squared error (MSE) on test set: 0.9558
The root mean squared error (RMSE) on test set: 0.9776
Variance score: 0.09
```

Con el resultado de estas métricas concluimos que el modelo no está funcionando bien, ya que el **R2** mide la cantidad de varianza en la variable objetivo que es explicada por el modelo. Un R2 de 1 indica que el modelo explica el 100% de la varianza, mientras que un R2 de 0 indica que el modelo no explica ninguna de la varianza. En este caso, el R2 es de 0.09, lo que significa que el modelo sólo explica el **9%** de la varianza en la **variable objetivo**.

Modelo de Regresión Polinómica de grado 2

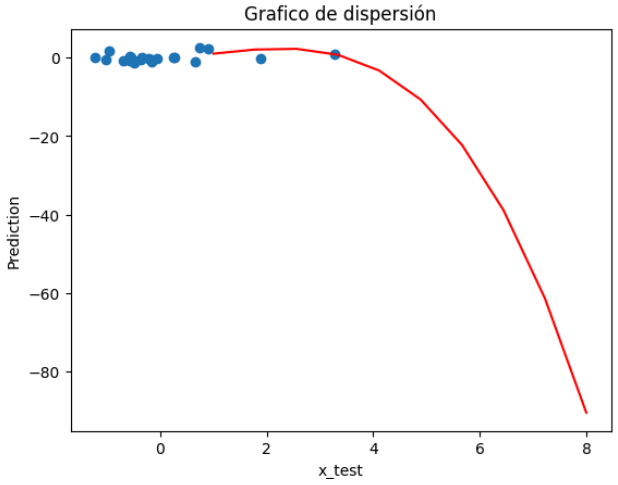


El grafico representa la relacion entre las dos variables, representadas por los puntos de color azul y la linea color rojo. x test representa los valores de la variable independiente/predictora en el conjunto de prueba. Y el eje y, representa los valores de la variable dependiente/respuesta predichos por el modelo para cada valor de "x test". Observamos que los puntos azules no siguen un patron lineal claro, ya que se encuentran dispersos, y la curva de color rojo trata de capturar la tendencia general de la relacion entre las variables.

```
The mean absolte error (MAE) on test set: 0.68530
The mean squared error (MSE) on test set: 1.0344
The root mean squared error (RMSE) on test set: 1.0170
Variance score: 0.01
```

Los resultados que arrojaron las métricas MAE, MSE y RMSE son altos. Esto nos indica que el modelo de grado 2 tiene un error alto cuando intenta hacer la predicción de la tasa de analfabetismo a partir de la tasa de mortalidad infantil. El valor de R2, conocido también como coeficiente de determinación, sólo explica el 1% de la variación en los datos de analfabetismo. Esto es muy bajo, y nos indica que no está capturando bien la relación entre la mortalidad infantil y el analfabetismo.

Modelo de Regresión Polinómica de grado 3



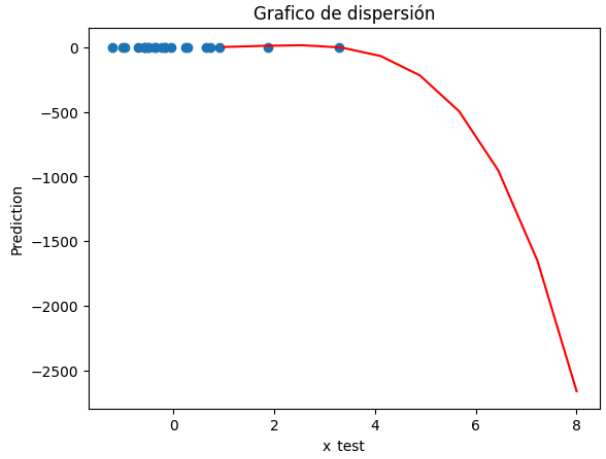
Observamos en el gráfico que la curva de color rojo alcanza un punto maximo aproximadamente en x=2, la variable Prediction (variable dependiente) aumenta inicialmente con "x test" (variable independiente), alcanza un maximo y luego disminuye. La mayoría de los puntos azules se concentran en un rango de x test bajo entre -1 y 2 aproximadamente, y hay pocos puntos para valores mayores. Los puntos no siguen la curva roja de manera precisa, mostrando una dispersión alrededor de ella.

```
The mean absolte error (MAE) on test set: 0.73772
The mean squared error (MSE) on test set: 1.1960
The root mean squared error (RMSE) on test set: 1.0936
Variance score: -0.14
```

El resultado de las métricas y el gráfico sugieren que el modelo de regresión polinómica de tercer grado no es adecuado para los datos. Ya que MAE, indica que en promedio, el modelo se equivoca en aproximadamente 74 % al predecir. MSE, tiene un valor alto por lo tanto falla en la predicción y RMSE indica un error promedio considerable, R2 arroja un

resultado negativo indica que el modelo no está capturando la variabilidad en los datos, dando como resultado un absurdo.

Modelo de Regresión Polinómica de grado 4



En este gráfico la mayoría de los puntos se concentran en un rango estrecho de valores de mortalidad infantil (x-test), mostrando poca variación. La curva roja representando la ecuación del polinomio intenta ajustarse a los datos, no logra capturar la tendencia general.

```
The mean absolte error (MAE) on test set: 2.51618
The mean squared error (MSE) on test set: 19.3481
The root mean squared error (RMSE) on test set: 4.3986
Variance score: -17.51
```

En base a las métricas, concluimos que el modelo no tiene un óptimo rendimiento. Si miramos los errores, en el MAE el modelo se equivocó al predecir el analfabetismo, MSE tiene un valor alto donde indica la presencia de errores en la predicción del modelo, RMSE muestra también un promedio alto de predicción del analfabetismo. Y por último R2, tiene un valor negativo, es un valor absurdo.

CONCLUSIÓN FINAL

En este trabajo práctico, nos propusimos explorar la relación entre el nivel de analfabetismo y la mortalidad infantil en las distintas provincias de Argentina utilizando modelos de regresión lineal y polinómica. A través de la implementación de estos modelos, buscamos no solo estimar dicha relación, sino también analizar la calidad del ajuste y la capacidad predictiva de los modelos desarrollados.

El modelo de regresión lineal mostró un valor de R^2 de 0.09, indicando que sólo explica el 9% de la variación en la tasa de analfabetismo. Este bajo valor sugiere que la relación entre la mortalidad infantil y el analfabetismo no es bien capturada por una línea recta. Los valores de MAE, MSE y RMSE fueron altos, señalando que el modelo tiene errores significativos en sus predicciones.

El Modelo de Regresión Polinomial Grado 2 con un valor de R 2 de apenas 0.01, este modelo tampoco logra capturar de manera adecuada la relación entre las variables. Solo

explica el 1% de la variación en los datos de analfabetismo. MAE, MSE y RMSE fueron elevados, indicando un error considerable en las predicciones.
Los polinomios de grado 3 y 4 el rendimiento de R2 fueron negativos, indicando que el modelo tampoco es el adecuado y las métricas continuaron siendo altas, reflejando errores en la predicción.
Este trabajo nos proporcionó una base inicial para entender la relación entre las dos variables estudiadas y, utilizando todos estos datos recolectados de las métricas de los distintos modelos, se puede concluir en que si tuviéramos que elegir uno modelo sería el de regresión lineal, ya que es el que mejor resultados dió. Aun así, debido a la poca cantidad de datos que se tienen para trabajar, el uso de estos modelos de regresión lineal y polinómica no llegan a capturar adecuadamente la variabilidad de los datos, mostrando en cada uno de los casos una capacidad predictiva baja.