
Análisis y Regresión Lineal del Conjunto de Datos de Atención Médica 🏥

El informe se basa en un modelo de regresión lineal simple y múltiple, que busca encontrar una relación entre las variables independientes (características del paciente, como edad y género) y la variable dependiente, que en este caso es el resultado de la prueba (Test Results Num).

Preprocesamiento de los Datos

Antes de construir el modelo, se realizaron varias transformaciones en el conjunto de datos:

- **Codificación de Género:** La variable **Gender** (género) se convirtió de categorías de texto ("Male", "Female") a valores numéricos (0 para "Male" y 1 para "Female").
- **Codificación One-Hot:** Se aplicó la codificación **one-hot** a las variables categóricas **Medical Condition**, **Insurance Provider** y **Medication**. Este método crea nuevas columnas binarias (con valores **True** o **False**), una por cada categoría, lo que permite que el modelo de regresión las interprete como características numéricas.
- **Codificación de Resultados de Pruebas:** La variable objetivo **Test Results** (resultados de pruebas) se mapeó a valores numéricos, creando la columna **Test Results Num** con la siguiente correspondencia:
    - **"Normal"** -> **0**
    - **"Inconclusive"** -> **1**
    - **"Abnormal"** -> **2**
- **Selección de Características:** Se eliminaron las columnas que no eran relevantes o no eran numéricas para la regresión, como nombres, fechas, detalles del médico y el hospital, y las variables categóricas originales que ya fueron codificadas. Las características (X) y la variable objetivo (y) se prepararon para el modelo.

Construcción y Entrenamiento del Modelo

El conjunto de datos se dividió en dos partes:

- **Conjunto de Entrenamiento (70%):** Utilizado para **ajustar el modelo**, es decir, para que el algoritmo de regresión aprenda la relación entre las características y el resultado de la prueba.
- **Conjunto de Prueba (30%):** Utilizado para **evaluar el rendimiento del modelo** con datos que no ha visto antes, lo que proporciona una estimación más realista de su capacidad de predicción.

Se utilizó el modelo **LinearRegression** de la biblioteca **scikit-learn** para realizar el análisis. Al entrenar el modelo, se obtuvieron los coeficientes para cada característica, los cuales indican la relación y la importancia de cada una de ellas para predecir el resultado de la prueba.

Resultados y Evaluación del Modelo

Coeficientes y Estadísticas

Los coeficientes del modelo muestran el impacto de cada característica en la predicción. Sin embargo, en este caso, la mayoría de los coeficientes, a excepción de **Test Results Num**, son extremadamente cercanos a cero. Esto se debe a que la variable objetivo (Test Results Num) ya estaba incluida en el conjunto de características (X). Esto hace que el modelo de regresión sea perfecto, ya que predice la variable objetivo a partir de sí misma.

La regresión lineal es perfecta (R2=1.0), lo cual es un resultado irreal que indica que la variable objetivo (y) está contenida en el conjunto de características (X).

Métricas de Evaluación

- **R2 (R-cuadrado):** El valor de R2 es **1.0**. Un valor de 1.0 significa que el modelo explica el 100% de la variabilidad en los datos de la variable objetivo. Como se mencionó anteriormente, este es un resultado perfecto, pero incorrecto, que surge de un error en la selección de las variables.
- **MAE, MSE, RMSE:** Los valores de los errores **MAE** (Error Absoluto Medio), **MSE** (Error Cuadrático Medio) y **RMSE** (Raíz del Error Cuadrático Medio) son extremadamente cercanos a cero, lo que también es una consecuencia del error de preprocesamiento, indicando que las predicciones del modelo son casi idénticas a los valores reales.

Visualizaciones

Los gráficos de dispersión y de residuos confirman un resultado perfecto:

- **Gráfico de dispersión de y_test vs. predictions:** Los puntos forman una línea recta perfecta de 45 grados, mostrando que cada valor predicho coincide con el valor real.
- **Histograma de residuos:** La distribución de los errores muestra una única barra muy alta en cero, lo que significa que no hay errores o son insignificantes, lo que no es un resultado típico en la mayoría de los problemas de regresión.

Conclusión y Recomendación

El análisis de regresión lineal en este informe demuestra un **error crítico de preprocesamiento**: la variable objetivo **Test Results Num** se incluyó accidentalmente como una de las características (X). Por lo tanto, el modelo simplemente aprendió a predecir el valor de y directamente a partir de la misma variable, lo que resulta en un ajuste perfecto, pero completamente sin sentido.

Para que el análisis sea válido, se debe **eliminar la columna Test Results Num del conjunto de características X** antes de entrenar el modelo. Una vez corregido este paso, el modelo de regresión lineal podrá encontrar la verdadera relación entre las características del paciente (como edad, género, condición médica, etc.) y los resultados de las pruebas,

proporcionando coeficientes, métricas y visualizaciones que reflejen un rendimiento realista y útil.