Project: **Exploratory Data Analysis (EDA) on Diabetes Dataset**

Vivek Tejasbhai Jariwala

Email: [vj84491n@pace.edu](mailto:vj84491n@pace.edu)

**Abstract:**

Exploratory Data Analysis (EDA) is a critical step in data science that allows us to uncover patterns, detect anomalies, and get insights before applying machine learning algorithms. In this project, we are working on a diabetes progression dataset, investigating dependencies among variables, outliers detection, and feature importance. This project follows a step-by-step approach using NumPy, Pandas, and Matplotlib for a detailed study of the data.

**Dataset Overview:**

* The dataset is obtained from Stanford University's Machine Learning Repository and contains 442 records. The target variable "Y" represents the progression of diabetes. The dataset consists of various clinical features such as age, BMI, blood pressure, and serum measurements.
* Dataset Link: <https://hastie.su.domains/Papers/LARS/diabetes.data>

**Data Preparation:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

file\_path = "diabetes.data.txt"

df = pd.read\_csv(file\_path, sep=r'\s+', header=0)

pd.set\_option('display.max\_columns', None)

pd.set\_option('display.max\_rows', None)

**Data Overview & Checking for Missing Values:**

print(df.shape)

print(df.isnull().sum())

print(df.dtypes)

The dataset contains no missing values, ensuring a clean analysis.

**Ensuring Numeric Data:**

print(df.dtypes)

**Handling Outliers:**

Using boxplots to identify potential outliers in the dataset:

plt.figure(figsize=(12, 6))

df.boxplot()

plt.xticks(rotation=45)

plt.title("Boxplot to detect outliers")

plt.show()

A graph with lines and dots

AI-generated content may be incorrect.

Here,

* Each column in df is represented as a separate box in the plot.
* **The box represents the Interquartile Range (IQR)**:
  + The bottom of the box = **25th percentile (Q1)**
  + The top of the box = **75th percentile (Q3)**
  + The middle line inside the box = **Median (Q2)**
* **The "whiskers"** represent the range of data within **1.5 times the IQR**
* **Outliers** are shown as **small circles or dots** beyond the whiskers.
  + These values are significantly different from the majority of data points.
  + Identifies outliers based on the **Interquartile Range (IQR) method.**

# Compute Q1 (25th percentile) and Q3 (75th percentile)

Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1 # Interquartile Range

# Define outliers as values below Q1 - 1.5\*IQR or above Q3 + 1.5\*IQR

outliers = ((df < (Q1 - 1.5 \* IQR)) | (df > (Q3 + 1.5 \* IQR)))

# Print outlier values per column

for column in df.columns:

outlier\_values = df[column][outliers[column]]

if not outlier\_values.empty:

print(f"Outliers in {column}:")

print(outlier\_values.to\_list())

A screenshot of a computer

AI-generated content may be incorrect.

**Feature Correlation Analysis**

compute the correlation matrix:

corr\_matrix = df.corr()

print(corr\_matrix)

Now, visualize the correlation using a heatmap:

plt.figure(figsize=(10, 6))

plt.imshow(corr\_matrix, cmap="coolwarm", interpolation="none")

plt.colorbar()

plt.xticks(range(len(corr\_matrix.columns)), corr\_matrix.columns, rotation=45)

plt.yticks(range(len(corr\_matrix.columns)), corr\_matrix.columns)

plt.title("Feature Correlation Heatmap")

plt.show()

A heatmap chart with different colors

AI-generated content may be incorrect.

# Feature importance using correlation with target variable 'Y'

correlation\_with\_target = df.corr()["Y"].abs().sort\_values(ascending=False)

print("\nFeature Importance based on correlation with Y:")

print(correlation\_with\_target)

Feature Importance based on correlation with Y:

Y 1.000000

BMI 0.586450

S5 0.565883

BP 0.441482

S4 0.430453

S6 0.382483

S1 0.212022

AGE 0.187889

S2 0.174054

SEX 0.043062

S3 -0.394789

Name: Y, dtype: float64

# Feature importance using OLS regression

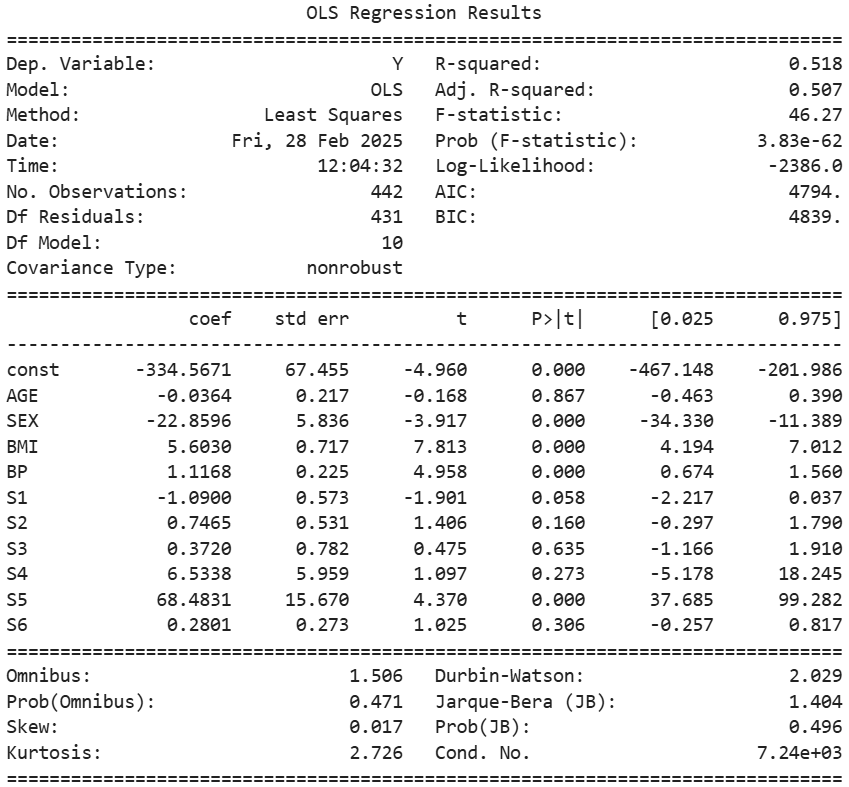
X = df.drop('Y', axis=1)

y = df['Y']

X = sm.add\_constant(X)

model = sm.OLS(y, X).fit()

print(model.summary())



**Model Performance**

* **R-squared**: The R-squared value is **0.518**, meaning that approximately **51.8%** of the variability in the target variable (Y) is explained by the independent variables in the model. This indicates a moderate fit.
* **Adjusted R-squared**: The adjusted R-squared is **0.507**, which accounts for the number of predictors in the model. It is slightly lower than the R-squared, indicating that some predictors may not contribute significantly to the model.
* **F-statistic**: The F-statistic is **46.27** with a very low p-value (**3.83e-62**), indicating that the model is statistically significant overall.

**Insights:**

**2. Significant Predictors**

The coefficients and their p-values help identify which predictors are statistically significant:

* **SEX**: The coefficient is **-22.8596** with a p-value of **0.000**, indicating that sex is a significant predictor of diabetes progression. Being female (assuming SEX=2) is associated with a **22.86-unit decrease** in diabetes progression compared to males (SEX=1).
* **BMI**: The coefficient is **5.6030** with a p-value of **0.000**, indicating that BMI is a significant predictor. A **1-unit increase in BMI** is associated with a **5.60-unit increase** in diabetes progression.
* **BP**: The coefficient is **1.1168** with a p-value of **0.000**, indicating that blood pressure is a significant predictor. A **1-unit increase in BP** is associated with a **1.12-unit increase** in diabetes progression.
* **S5**: The coefficient is **68.4831** with a p-value of **0.000**, indicating that S5 (a blood serum measurement) is a significant predictor. A **1-unit increase in S5** is associated with a **68.48-unit increase** in diabetes progression.

**3. Non-Significant Predictors**

Some predictors are not statistically significant (p-value > 0.05):

* **AGE**: The coefficient is **-0.0364** with a p-value of **0.867**, indicating that age does not significantly affect diabetes progression in this model.
* **S1**: The coefficient is **-1.0900** with a p-value of **0.058**, which is close to the significance threshold but not statistically significant at the 5% level.
* **S2**, **S3**, **S4**, **S6**: These predictors have p-values greater than 0.05, indicating they do not significantly contribute to the model.

**Possible Issues:**

* The dataset might have been collected from a **specific demographic or region**, leading to bias.
* If data is not diverse, insights may not **generalize well** to other populations.
* Example: If most patients in the dataset are from a specific age group or ethnicity, the model may not work well for different populations.
* Outliers can **skew statistical measures** like mean and standard deviation, leading to misleading insights.

**Diabetes Progression Insights Dashboard:**

[**https://public.tableau.com/views/DiabetesProgressionInsightsDashboard/DiabetesProgressionInsightsDashboard?:language=en-US&:sid=&:redirect=auth&:display\_count=n&:origin=viz\_share\_link**](https://public.tableau.com/views/DiabetesProgressionInsightsDashboard/DiabetesProgressionInsightsDashboard?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link)