EXPERIMENT ASSESSMENT

ACADEMIC YEAR 2025-26

Course: AIH Lab

Course code: CSDOL 7012

Year: BE SEM: VII

Experiment No. 3
Title: To perform EDA on healthcare data using Seaborne and perform data Modeling
Name: Ayush Gupta
Roll Number: 11
Date of Performance:28/07/25
Date of Submission:4/08/25

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Performance	5	
Understanding	5	
Journal work and timely submission.	10	
Total	**20**	

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Performance	5	3	2
Understanding	5	3	2
Journal work and timely submission.	10	8	4

Checked by

Name of Faculty : Mrs.Kranti Gule

Signature :

Date :

Aim: To perform EDA on healthcare data using Seaborne and perform data Modeling

Objective: The objective of this study is to conduct Exploratory Data Analysis (EDA) on healthcare data employing Seaborn for data visualization and subsequently perform data modeling. Through EDA, we aim to gain an in-depth understanding of the healthcare dataset, uncovering key insights, trends, and potential anomalies. Following EDA, the focus shifts to data modeling, where we aim to develop predictive or descriptive models that can provide actionable intelligence.

Theory:

Exploratory Data Analysis (EDA) is an essential phase in data analysis, particularly in healthcare, where the understanding of data intricacies is crucial. Seaborn, a Python data visualization library, serves as a valuable tool for this process. EDA involves data summarization, visualization, and pattern identification to reveal underlying structures within healthcare data.

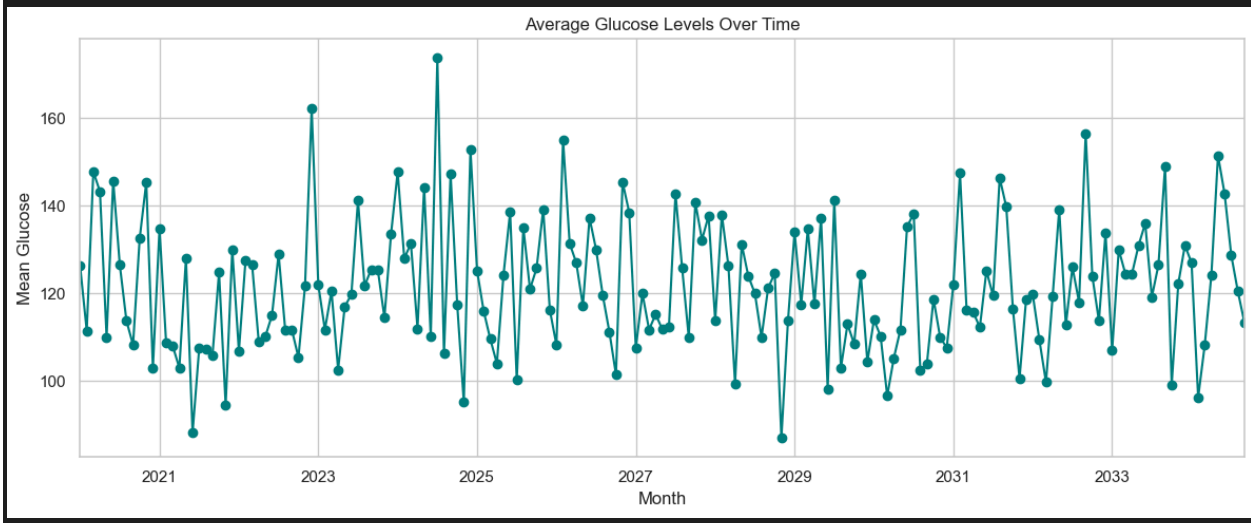
Seaborn's capabilities allow us to create informative plots such as histograms, scatter plots, and heatmaps, enabling us to explore relationships, distributions, and correlations in the healthcare dataset. Through EDA, we aim to identify trends, outliers, and potential areas for deeper investigation.

Following EDA, the study shifts towards data modeling, where we leverage statistical and machine learning techniques to build predictive or descriptive models. These models can provide actionable intelligence, such as disease risk assessments, patient outcome predictions, or treatment optimization strategies. Data modeling plays a pivotal role in transforming raw healthcare data into actionable insights, enhancing decision-making in clinical practice, research, and healthcare management.

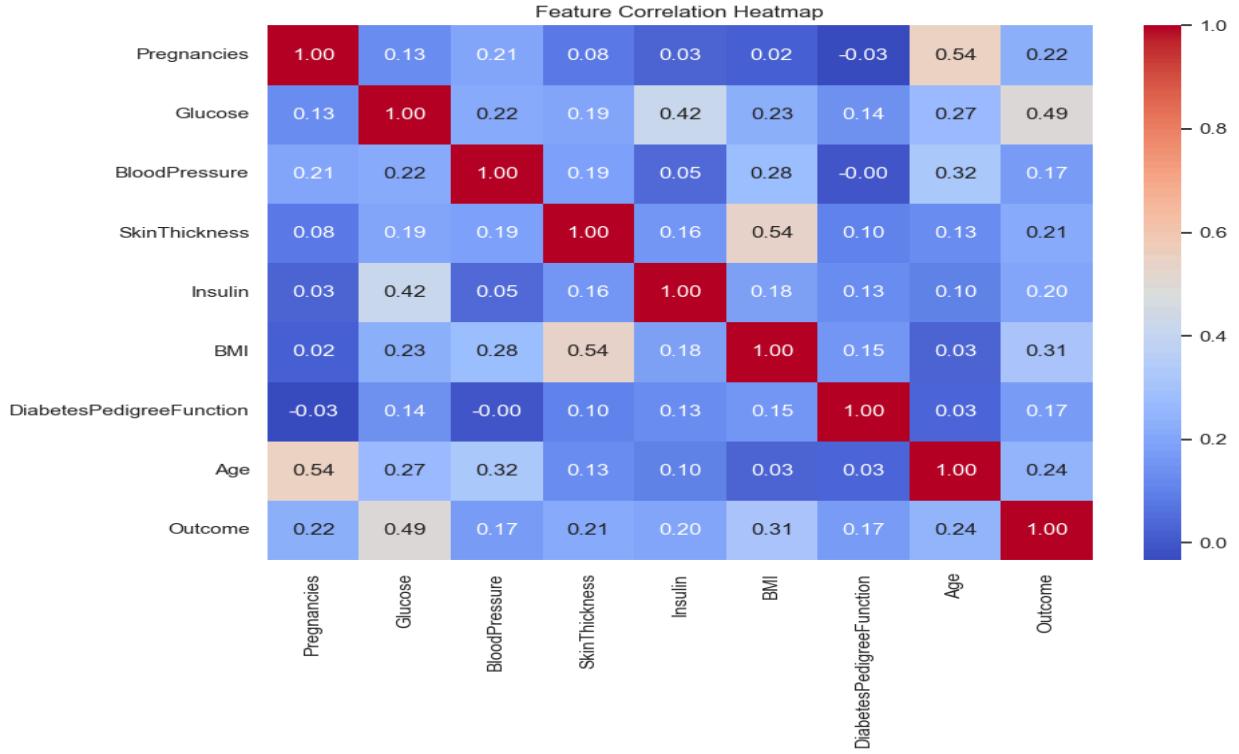
In essence, this study's theory underscores the importance of conducting thorough EDA with Seaborn to extract valuable insights from healthcare data and subsequently employing data modeling techniques to translate these insights into practical applications for improving healthcare outcomes and decision support.

Program and Output:

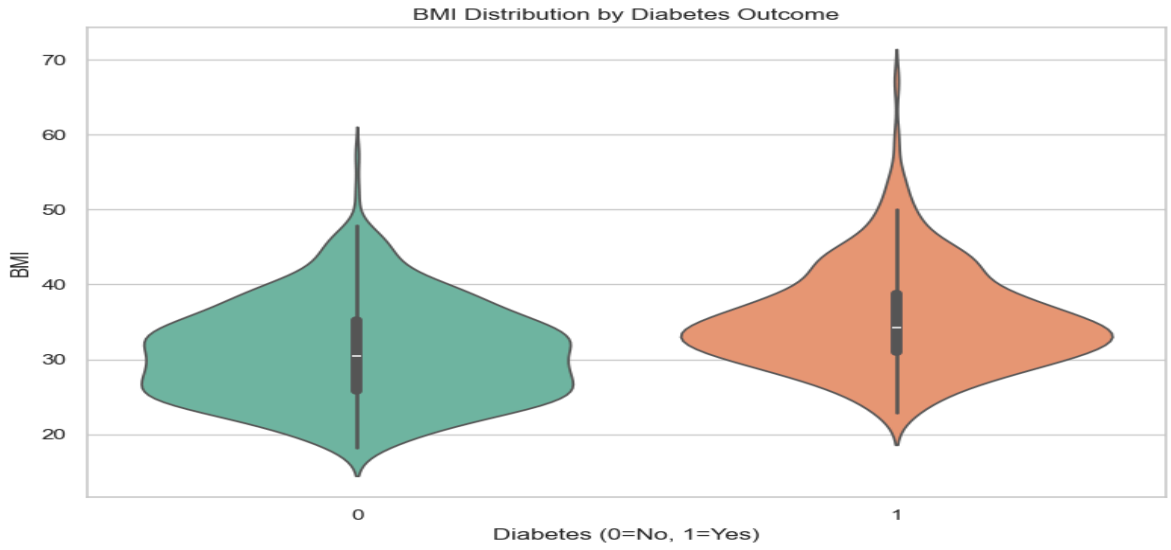
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style='whitegrid')
df = pd.read_csv("diabetes(cleaned).csv")
import numpy as np
df['VisitDate'] = pd.date_range(start='2020-01-01', periods=len(df), freq='W')
df.set_index('VisitDate', inplace=True)
plt.figure(figsize=(12, 5))
df['Glucose'].resample('M').mean().plot(marker='o', color='teal')
plt.title('Average Glucose Levels Over Time')
plt.xlabel('Month')
plt.ylabel('Mean Glucose')
plt.grid(True)
plt.tight_layout()
plt.show()
```



```
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot=True, fmt=".2f", cmap='coolwarm')
plt.title('Feature Correlation Heatmap')
plt.tight_layout()
plt.show()
```




```
plt.figure(figsize=(8, 6))
sns.violinplot(x='Outcome', y='BMI', data=df, palette='Set2')
plt.title('BMI Distribution by Diabetes Outcome')
plt.xlabel('Diabetes (0=No, 1=Yes)')
plt.ylabel('BMI')
plt.tight_layout()
plt.show()
```



Conclusion-

1.Was the healthcare data successfully collected for the selected disease? If yes, from which sources?

Yes, the healthcare data was successfully collected for diabetes, the selected disease. The dataset is commonly sourced from public health repositories like the UCI Machine Learning Repository or Kaggle, where real-world patient data related to diabetes indicators such as glucose, BMI, insulin, and age is made available for research and analysis**.**

2.What challenges were encountered during the data cleaning process? How were they Resolved?

During data cleaning, the main challenges included missing values and outliers in key health indicators like glucose and insulin levels. These were handled by using `fillna()` to fill missing values with appropriate methods (like forward fill or mean), and box plots were used to detect outliers, which were either removed or replaced with capped values to reduce distortion in analysis.

3.Difference between Matplot lib & Seaborn lib?

Matplotlib and Seaborn are both Python libraries used for data visualization, but they differ in functionality and ease of use. Matplotlib is a low-level library that gives you complete control over plot customization, making it highly flexible but often requiring more code to produce detailed visualizations. On the other hand, Seaborn is built on top of Matplotlib and provides a high-level interface that simplifies the creation of attractive and informative statistical plots. It comes with built-in themes and functions for visualizing complex relationships, making it more suitable for quick and elegant data exploration. In short, Matplotlib is ideal for custom visualizations, while Seaborn is preferred for cleaner, ready-to-use statistical graphics**.**