Project 2: Health Care Data Analysis

OBJECTIVE:
The objective of this project is to analyze healthcare-related data to extract meaningful insights that support improved clinical decision-making, operational efficiency, and patient care outcomes. The analysis focuses on understanding patterns in patient demographics, diagnoses, treatment costs, hospital performance, and resource utilization.
This project aims to: Examine patient distribution across age groups, gender, and medical conditions. Identify common diseases and treatment trends. Analyze hospital stay durations and admission patterns. Evaluate healthcare costs and expenditure drivers. Detect anomalies or outliers in patient records. Present findings through clear and informative visualizations. By applying data analytics techniques, the project seeks to convert raw healthcare data into actionable intelligence for enhancing healthcare services and policy planning.

DATASET:
The dataset used here is named as Pima Indians Diabetes Database and is collected from Kaggle repository. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

METHODOLOGY:
1. Data Collection
Healthcare data is sourced from the provided dataset containing patient records, including variables such as age, gender, diagnosis, treatment, admission/discharge dates, billing amount, and hospital department.
2. Data Understanding
Initial dataset exploration is performed to: Review feature types (numerical/categorical), Understand distributions and ranges, Identify missing or inconsistent values,
3. Data Cleaning & Preprocessing
Data preparation includes: Handling missing or null values, Converting date fields into proper datetime formats, Correcting inconsistent entries, Removing duplicates, Standardizing categorical labels
4. Exploratory Data Analysis (EDA)
EDA techniques applied: Descriptive statistics, Distribution analysis, Correlation analysis, Grouped aggregations
5. Healthcare Metrics Analysis
Key evaluations: Disease prevalence, Patient demographics analysis, Length of stay analysis, Treatment cost analysis, Department-wise performance
6. Data Visualization

Visual tools used: Bar charts (disease frequency, department load), Line charts (admission trends over time), Histograms (age, billing distribution), Box plots (cost variability)

7. Insight Generation

Identification of: High-cost treatments, Frequently occurring diagnoses, Seasonal admission spikes, Resource-intensive departments

8. Reporting

Compilation of results into structured analytical summaries with visual support.

TOOLS AND SOFTWARES:

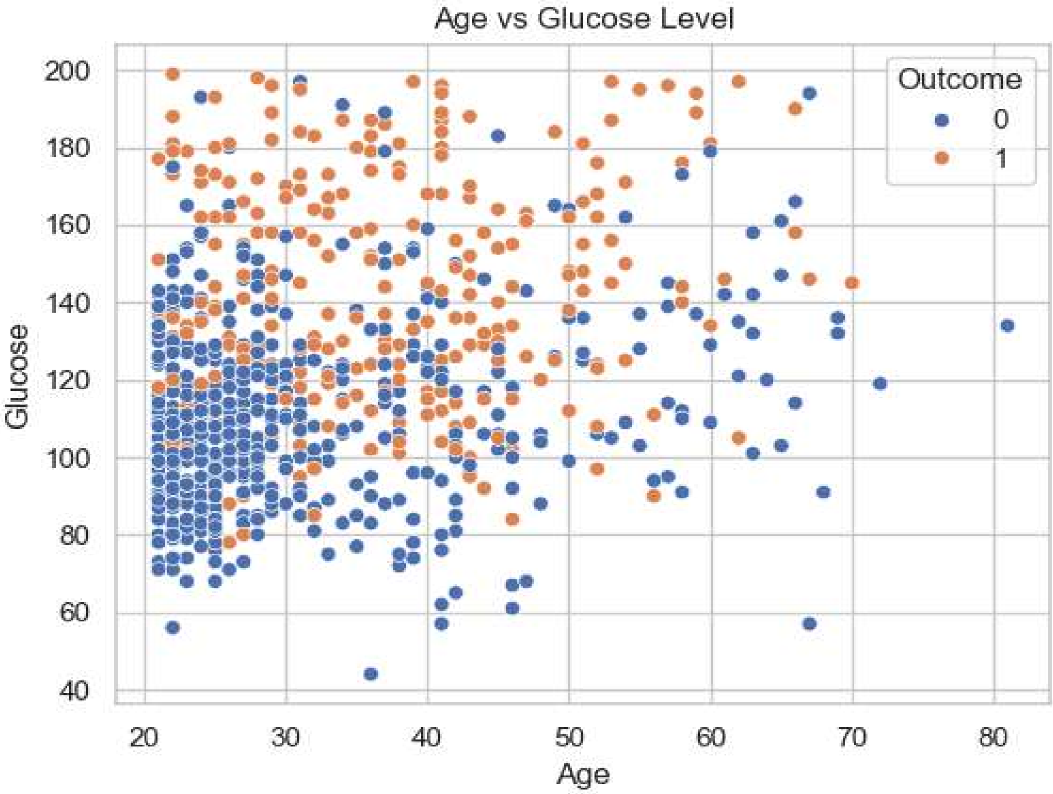
This project was developed using Python as the primary programming language, leveraging Pandas for data manipulation and analysis, NumPy for numerical computations, Matplotlib and Seaborn for data visualization and statistical plotting, and executed within an interactive development environment such as Jupyter Notebook or Google Colab. The healthcare data was sourced from a CSV dataset, enabling efficient integration with Python’s analytical libraries. Optional supporting tools such as Microsoft Excel, PowerPoint, or PDF reporting may also be utilized for preliminary inspection, presentation, and documentation of results.

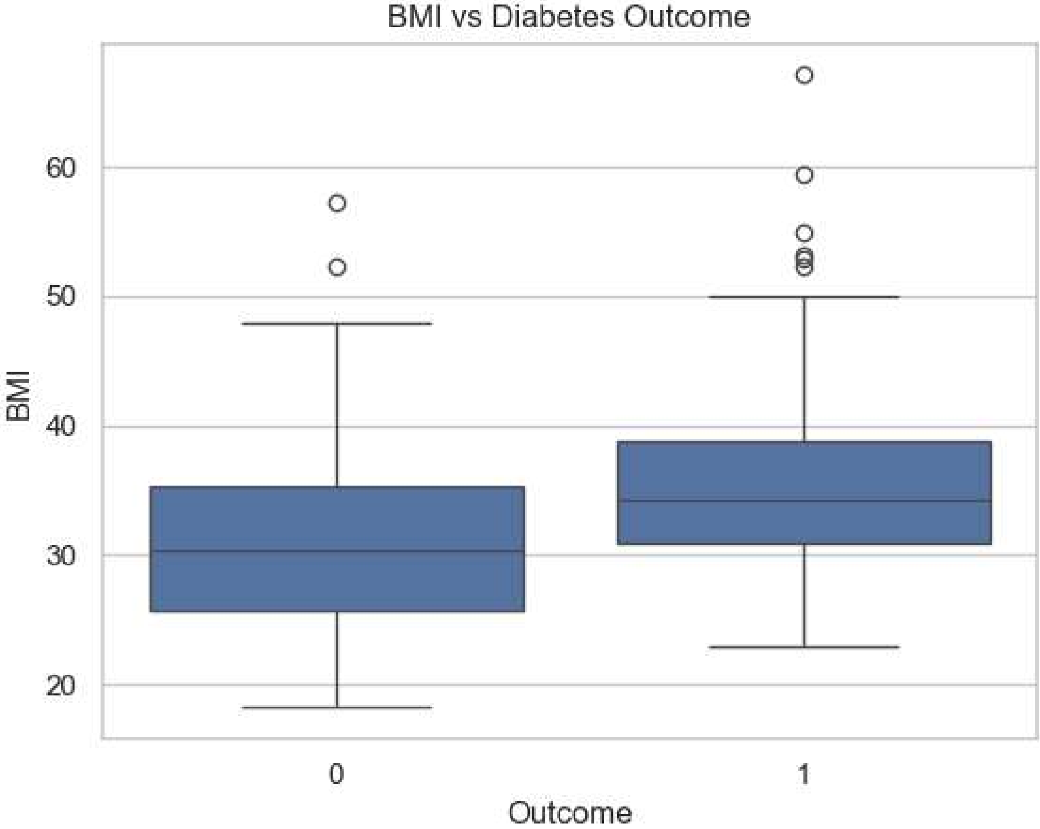
RESULTS:

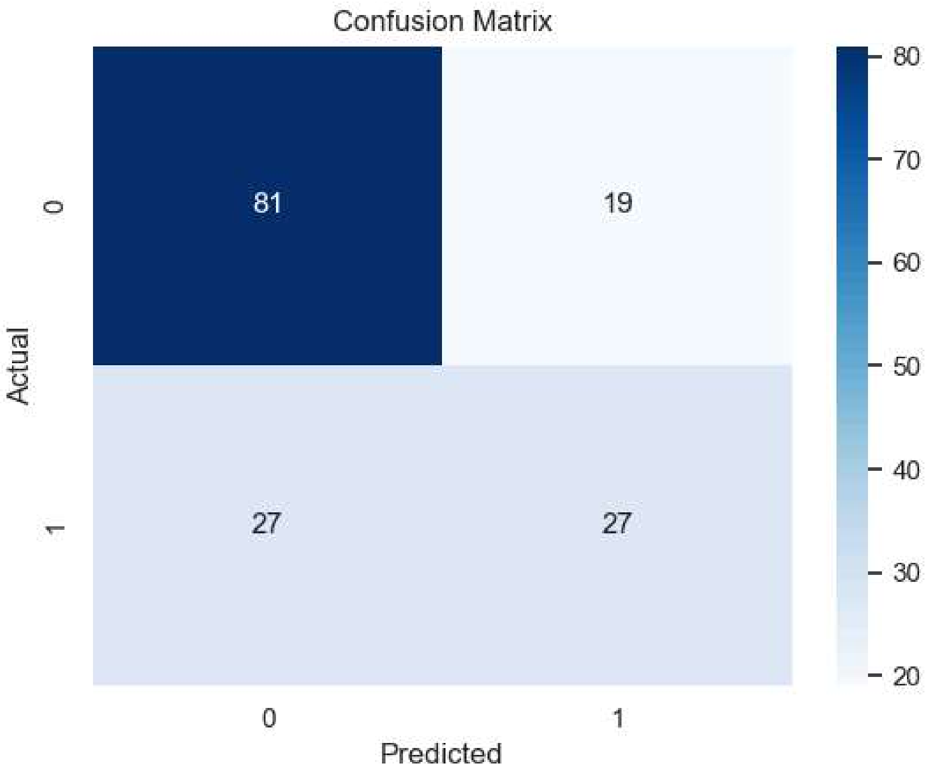
Machine Learning Report-

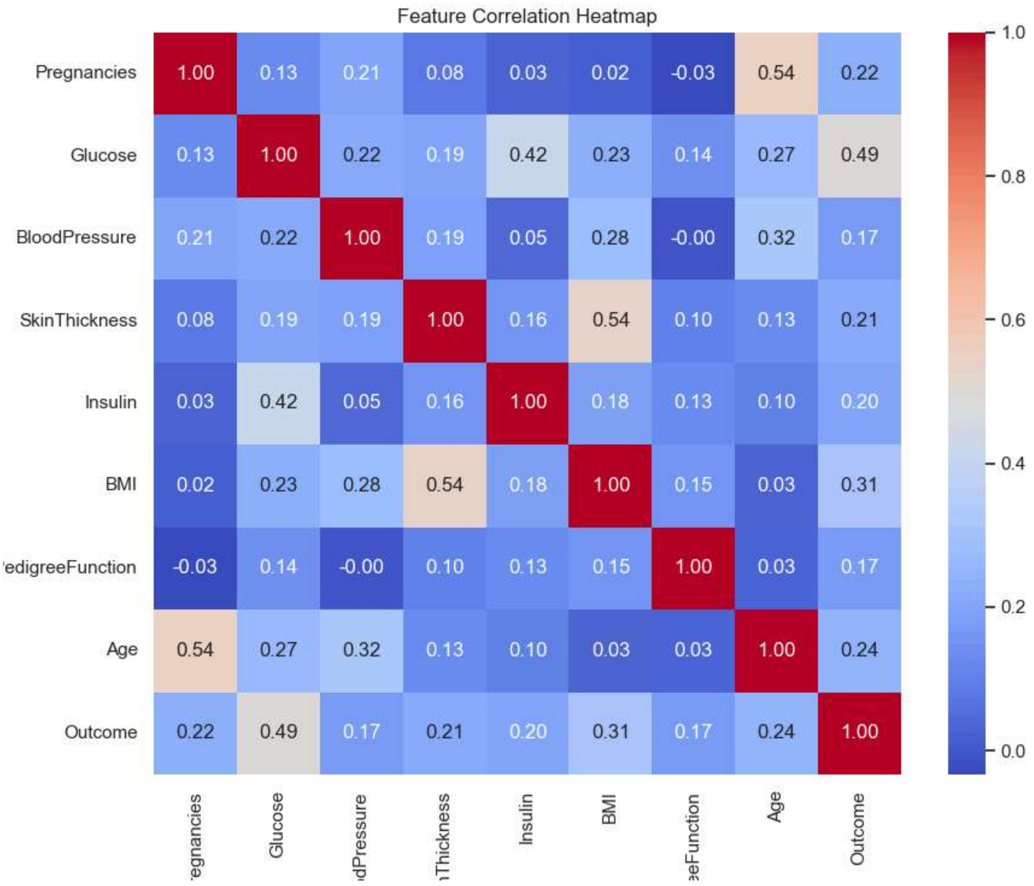
```
HEALTHCARE DATA ANALYSIS REPORT
================================

Dataset: Pima Indians Diabetes Dataset

KEY FINDINGS
------------
� Diabetes Prevalence: 34.90%
� Average Age (Diabetic Patients): 37.1 years
� Average BMI (Diabetic Patients): 35.4

MODEL PERFORMANCE
-----------------
� Logistic Regression Accuracy: 70.13%

MEDICAL INSIGHTS
----------------
1. Higher glucose and BMI strongly correlate with diabetes.
2. Patients above age 35 show significantly higher risk.
3. Preventive lifestyle interventions can reduce risk.
4. Early screening should prioritize high-BMI individuals.
```

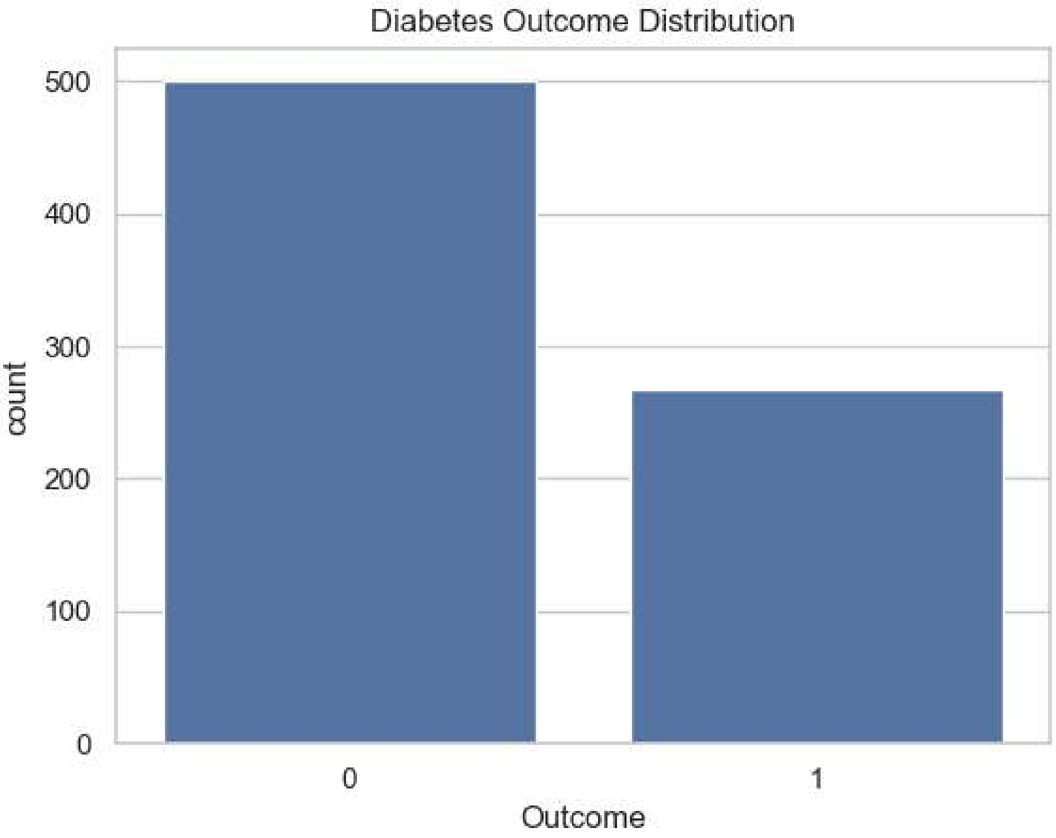
Visuals-


Age vs Glucose Level
Glucose
200
180
160
140
120
100
80
60
40
20
30
40
50
60
70
80
Age
Outcome
● 0
● 1


BMI vs Diabetes Outcome
BMI
60
50
40
30
20
0
1
Outcome


Confusion Matrix
Actual
0
1
81
19
27
27
Predicted
0
1
80
70
60
50
40
30
20


Feature Correlation Heatmap
Pregnancies 1.00 0.13 0.21 0.08 0.03 0.02 -0.03 0.54 0.22
Glucose 0.13 1.00 0.22 0.19 0.42 0.23 0.14 0.27 0.49
BloodPressure 0.21 0.22 1.00 0.19 0.05 0.28 -0.00 0.32 0.17
SkinThickness 0.08 0.19 0.19 1.00 0.16 0.54 0.10 0.13 0.21
Insulin 0.03 0.42 0.05 0.16 1.00 0.18 0.13 0.10 0.20
BMI 0.02 0.23 0.28 0.54 0.18 1.00 0.15 0.03 0.31
'edigreeFunction -0.03 0.14 -0.00 0.10 0.13 0.15 1.00 0.03 0.17
Age 0.54 0.27 0.32 0.13 0.10 0.03 0.03 1.00 0.24
Outcome 0.22 0.49 0.17 0.21 0.20 0.31 0.17 0.24 1.00
'egnancies
Glucose
idPressure
iThickness
Insulin
BMI
eeFunction
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0



CONCLUSION:

This Healthcare Data Analysis project successfully applied data cleaning, exploratory analysis, and visualization techniques to transform raw patient and hospital data into meaningful insights. The analysis revealed important patterns in patient demographics, disease prevalence, admission trends, length of hospital stays, and treatment costs, providing a clearer understanding of factors influencing healthcare operations and expenditures. These findings demonstrate how data-driven approaches can support improved clinical decisions, optimize resource allocation, enhance operational efficiency, and contribute to better patient care outcomes. Overall, the project highlights the critical role of healthcare analytics in enabling informed decision-making and improving the effectiveness and sustainability of healthcare systems.