AI BASED DIABETICS PREDICITION SYSTEM

DEVELOPMENT PART 2

DATA CLEANING

Data cleaning is a crucial step in the data preprocessing process, especially when working with real-world data. It involves identifying and correcting errors, inconsistencies, and missing values in your dataset to ensure that it is accurate, reliable, and suitable for analysis or machine learning. Here are the key steps involved in data cleaning:

Data Inspection:

Begin by inspecting your dataset to get a sense of its structure, including the number of rows and columns, data types, and a summary of the data.

Handling Missing Values:

Identify and handle missing data. Missing values can be problematic for analysis or modeling. Options for handling missing data include:

Removing rows with missing values (if the number of missing values is small).

Imputing missing values using techniques such as mean, median, mode, or more advanced imputation methods based on the data distribution.

Using domain knowledge or other data sources to fill in missing values.

Data Transformation:

Convert data into a consistent format. This might include:

Standardizing data types (e.g., converting dates to a uniform format).

Encoding categorical variables as numerical values (e.g., one-hot encoding or label encoding).

Scaling or normalizing numerical features.

PYTHON SCRIPT

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore", category=UserWarning)

RED = "\033[91m"
GREEN = "\033[92m"
YELLOW = "\033[93m"
BLUE = "\033[94m"
RESET = "\033[0m"

df = pd.read_csv("/kaggle/input/diabetes-data-set/diabetes.csv")

# DATA CLEANING
print(BLUE + "\nDATA CLEANING" + RESET)
# --- Check for missing values
missing_values = df.isnull().sum()
print(GREEN + "Missing Values : " + RESET)
print(missing_values)
# --- Handle missing values
mean_fill = df.fillna(df.mean())
df.fillna(mean_fill, inplace=True)
# --- Check for duplicate values
duplicate_values = df.duplicated().sum()
print(GREEN + "Duplicate Values : " + RESET)
print(duplicate_values)
# --- Drop duplicate values
df.drop_duplicates(inplace=True)

# DATA ANALYSIS
print(BLUE + "\nDATA ANALYSIS" + RESET)
```

```
# --- Summary Statistics
summary_stats = df.describe()
print(GREEN + "Summary Statistics : " + RESET)
print(summary_stats)
# --- Class Distribution
class_distribution = df["Outcome"].value_counts()
print(GREEN + "Class Distribution : " + RESET)
print(class_distribution)


# DATA VISUALIZATION
print(BLUE + "\nDATA VISUALIZATION" + RESET)
# --- Pair Plot
print(GREEN + "PairPlot : " + RESET)
sns.pairplot(df, hue='Outcome',diag_kind='kde',palette = "Blues")
plt.title("Pairwise Relationships")
plt.show()
# --- Histogram for age distribution
print(GREEN + "Histogram : " + RESET)
sns.histplot(df["Age"], bins=10, kde=True,palette = "Blues")
plt.xlabel("Age")
plt.ylabel("Count")
plt.title("Age Distribution")
plt.show()
# --- Box plot to visualize glucose levels by outcome
print(GREEN + "BoxPlot : " + RESET)
sns.boxplot(x="Outcome", y="Glucose", data=df,palette = "Blues")
plt.xlabel("Diabetes Outcome (0: No, 1: Yes)")
plt.ylabel("Glucose Level")
plt.title("Glucose Levels by Diabetes Outcome")
plt.show()
# --- Correlation heatmap
print(GREEN + "Correlation Heatmap : " + RESET)
correlation_matrix = df.corr()
```

```
sns.heatmap(correlation_matrix, annot=True, cmap="Blues")

plt.title("Correlation Heatmap")

plt.show()


# SAVING THE FILE

df.to_csv("/kaggle/working/cleaned_diabetes.csv", index=False)

print(BLUE + "\nDATA SAVING" + RESET)

print(GREEN + "Data Cleaned and Saved !" + RESET)

print("\n")
```

OUTPUT

```
DATA CLEANING

Missing Values :

Pregnancies   0   Glucose   0

BloodPressure 0 SkinThickness

0    Insulin    0    BMI    0

DiabetesPedigreeFunction    0

Age 0 Outcome 0 dtype: int64

Duplicate Values :

0
```

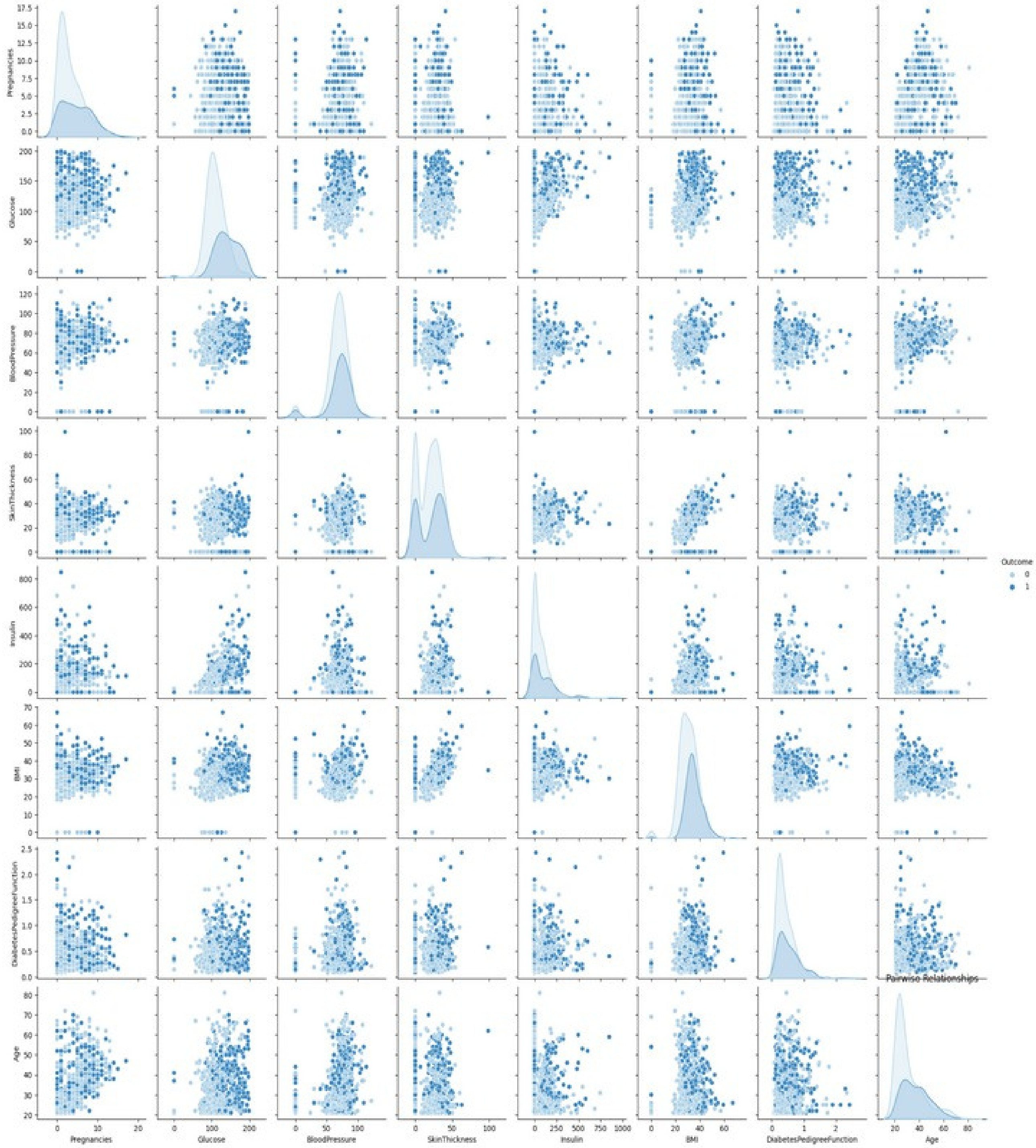
DATA ANALYSIS

```
Summary Statistics :

                  Pregnancies Glucose BloodPressure SkinThickness Insulin \

count 768.000000 768.000000 768.000000 768.000000 768.000000

mean 3.845052 120.894531 69.105469 20.536458 79.799479

std 3.369578 31.972618 19.355807 15.952218 115.244002
```

```
min 0.000000 0.000000 0.000000 0.000000 0.000000

25% 1.000000 99.000000 62.000000 0.000000 0.000000

50% 3.000000 117.000000 72.000000 23.000000 30.500000

75% 6.000000 140.250000 80.000000 32.000000 127.250000

max 17.000000 199.000000 122.000000 99.000000 846.000000


 BMI DiabetesPedigreeFunction Age Outcome

count 768.000000 768.000000 768.000000 768.000000

mean 31.992578 0.471876 33.240885 0.348958

std 7.884160 0.331329 11.760232 0.476951

min 0.000000 0.078000 21.000000 0.000000

25% 27.300000 0.243750 24.000000 0.000000

50% 32.000000 0.372500 29.000000 0.000000

75% 36.600000 0.626250 41.000000 1.000000

max 67.100000 2.420000 81.000000 1.000000

Class Distribution :

Outcome

0 500

1 268

Name: count, dtype: int64
```

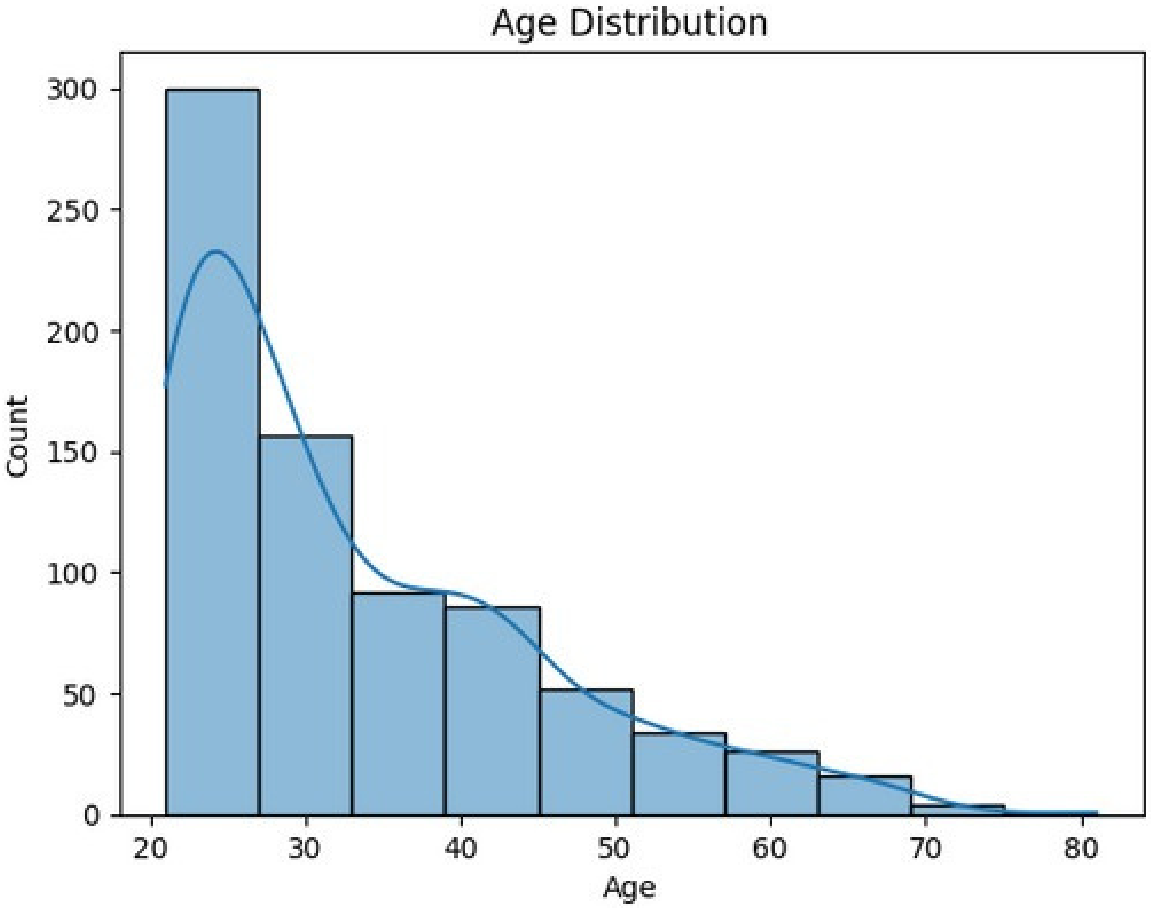
DATA VISUALISATION

Data visualization is the graphical representation of data to help people understand, analyze, and draw insights from data more effectively. It is an essential tool for data analysis, exploration, and communication.

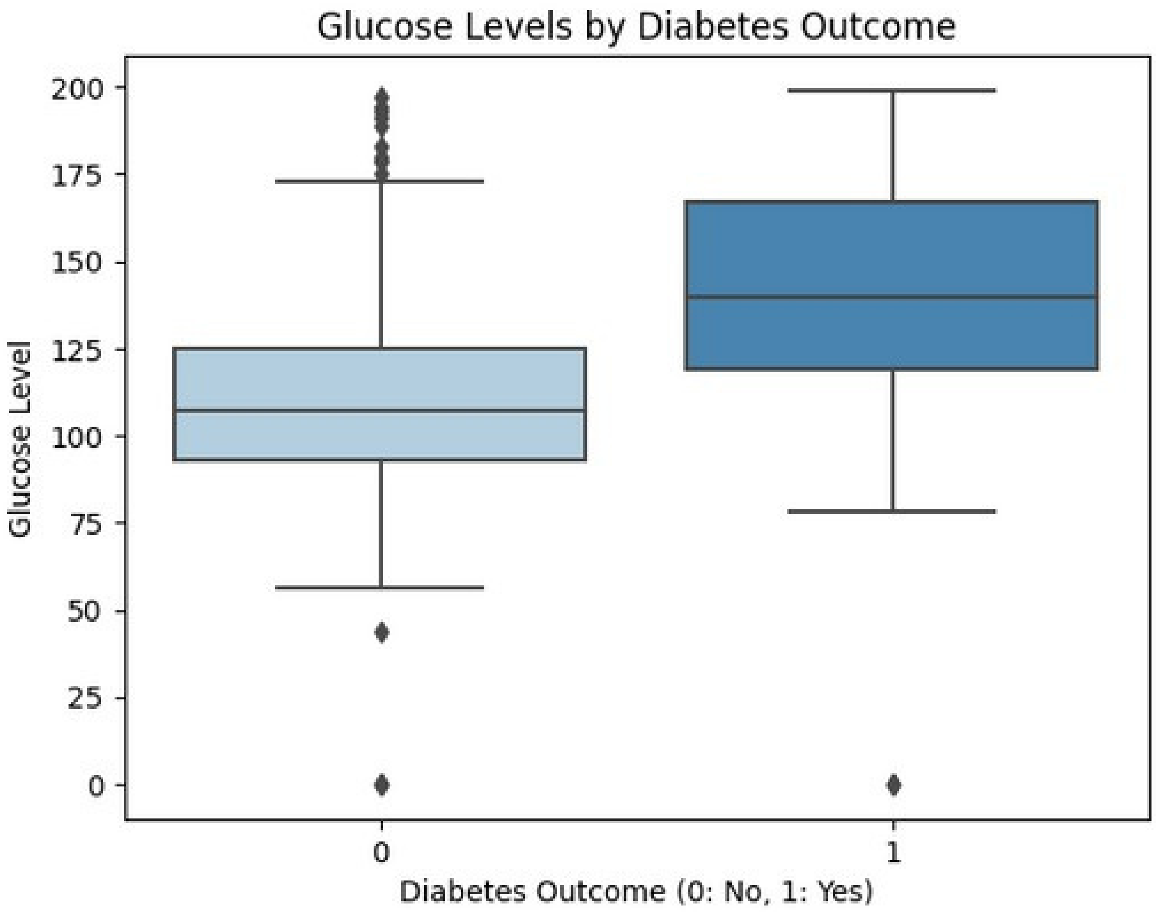


HISTOGRAM

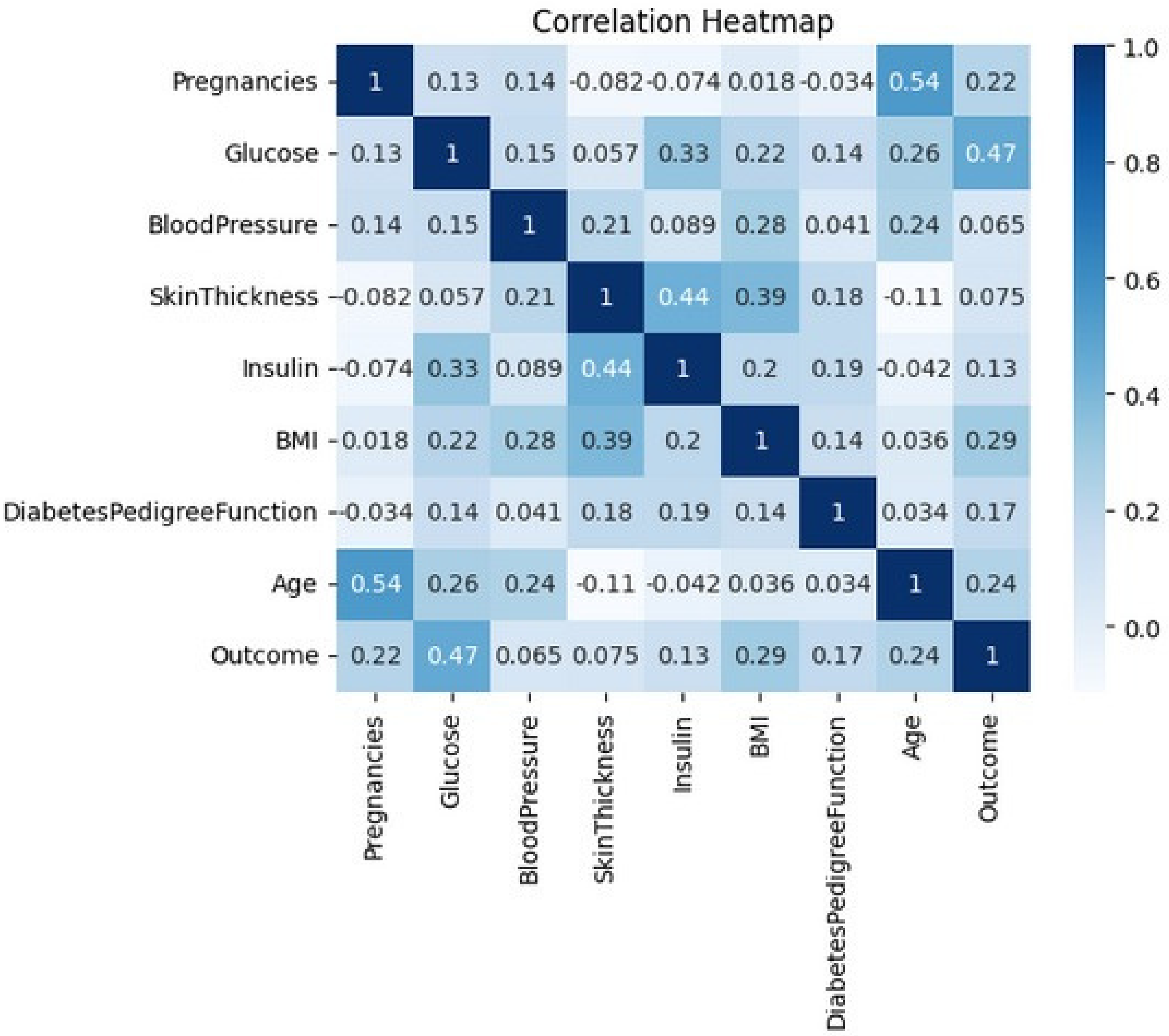
A histogram is a graphical representation of the distribution of a dataset. It's a way to visualize the frequency or count of data points in various numerical or categorical intervals or "bins." Histograms are particularly useful for understanding the shape, central tendency, and spread of a dataset.


Age Distribution
300
250
200
150
100
50
0
Count
20
30
40
50
60
70
80
Age

BOXPLOT



CORRELATION HEATMAP



CONCLUSION

The AI-based diabetes prediction system holds great promise in the fight against diabetes, a condition affecting millions of individuals worldwide. By leveraging the power of artificial intelligence and data-driven insights, we aim to improve public health, reduce the burden of the disease, and empower individuals to take control of their health. This project represents a step forward in the ongoing journey to harness the potential of AI for the betterment of society.

Name: Srileka V

Dept: CSE III year

Reg no.:810621104030