BIS Assignment - Report
Index – 20020406 (2020/IS/040)

Github link - https://github.com/Dushanee/BIS_Assignment/tree/main

1.

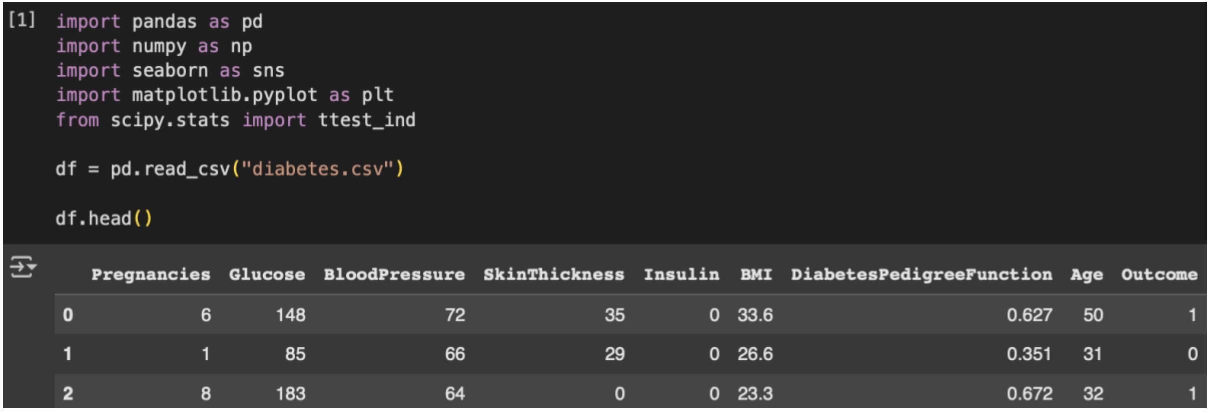
- Business Domain – **Healthcare - Predicting Diabetes Risk**
- Dataset - Pima Indians Diabetes Database
- Source - Kaggle (“uciml/pima-indians-diabetes-database")
- The healthcare industry relies on data analytics to predict diseases and improve preventive healthcare measures.
- Diabetes is a chronic condition that needs early detection to prevent complications.
- This dataset contains 768 patient records with 8 medical features.
- Target Variable - Outcome (1 = Diabetic, 0 = Non-Diabetic)
- Key Features - *Glucose, Blood Pressure, BMI, Age, Insulin, Skin Thickness, Diabetes Pedigree Function, and Pregnancies.*

2.

- Business Question - **"How can we identify high-risk individuals for diabetes based on key health indicators, and what preventive measures can be suggested?"**
- Analytics Process
  - Data Collection – Extract dataset from Kaggle.
  - Data Preprocessing – Handle missing values & clean the data.
  - Exploratory Data Analysis – Identify diabetes prevalence and feature impact.
  - Statistical & Machine Learning Analysis – Perform Linear Regression, Correlation Analysis, and Clustering.
  - Data Visualization – Graphically represent trends using Matplotlib & Seaborn.

3.

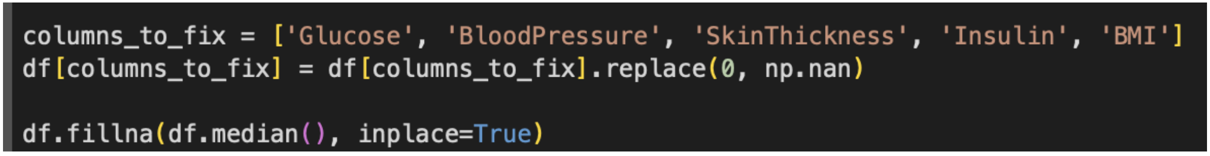
- Loading the data



```
[1]  import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
     from scipy.stats import ttest_ind

     df = pd.read_csv("diabetes.csv")

     df.head()
```

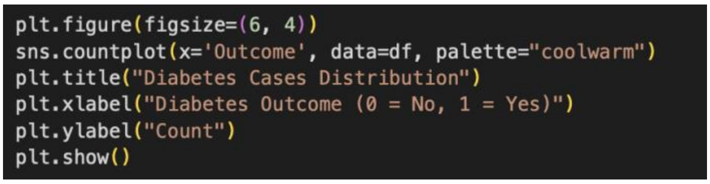
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1

- Preprocessing – handling missing values

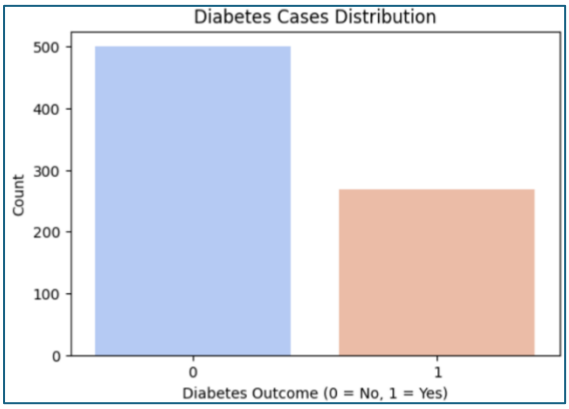


```
columns_to_fix = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
df[columns_to_fix] = df[columns_to_fix].replace(0, np.nan)

df.fillna(df.median(), inplace=True)
```

- Exploratory Data Analysis
    - Diabetes Outcome Distribution

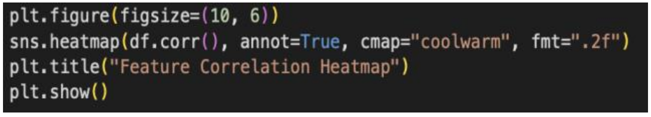


```
plt.figure(figsize=(6, 4))
sns.countplot(x='Outcome', data=df, palette="coolwarm")
plt.title("Diabetes Cases Distribution")
plt.xlabel("Diabetes Outcome (0 = No, 1 = Yes)")
plt.ylabel("Count")
plt.show()
```

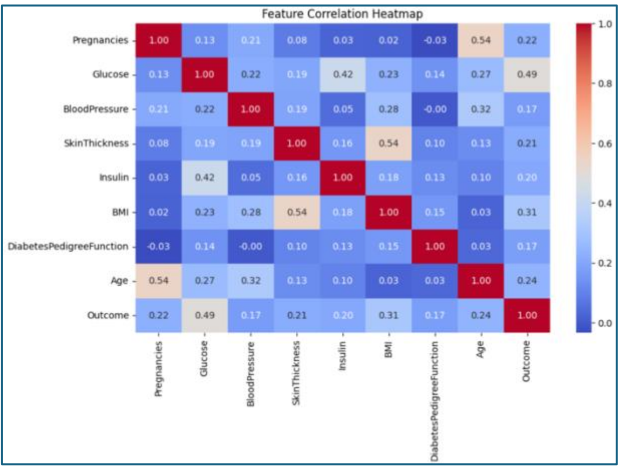


Interpretation - Approximately **35% of the dataset** Consists of diabetic patients. The dataset is slightly **imbalanced**, which is important when applying machine learning models.

- Feature Correlation Analysis (Heatmap)



```
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Feature Correlation Heatmap")
plt.show()
```



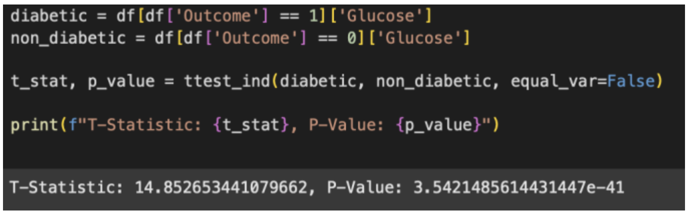
Interpretation

i. Glucose has the highest correlation (0.49) with diabetes.
ii. BMI (0.31) and Age (0.24) also contribute to diabetes risk.
iii. Blood Pressure & Insulin have weak correlations and may not be strong predictors

- Statistical Analysis - T-Test for Glucose Levels

Interpretation

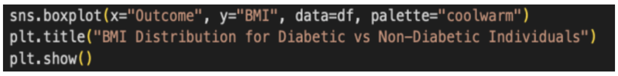
Glucose levels are significantly higher in diabetic individuals (p-value < 0.05). This confirms glucose is a key predictor for diabetes.



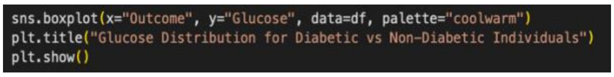
```
diabetic = df[df['Outcome'] == 1]['Glucose']
non_diabetic = df[df['Outcome'] == 0]['Glucose']

t_stat, p_value = ttest_ind(diabetic, non_diabetic, equal_var=False)

print(f"T-Statistic: {t_stat}, P-Value: {p_value}")
```

```
T-Statistic: 14.852653441079662, P-Value: 3.5421485614431447e-41
```

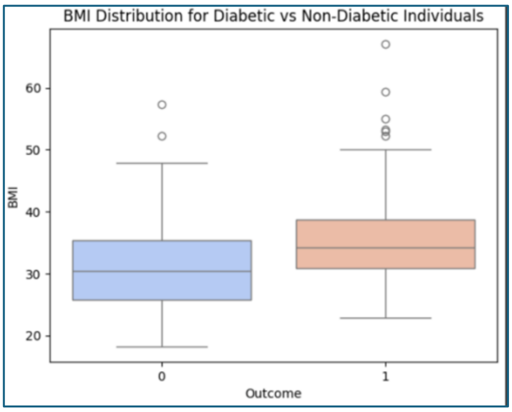
- Data Visualization

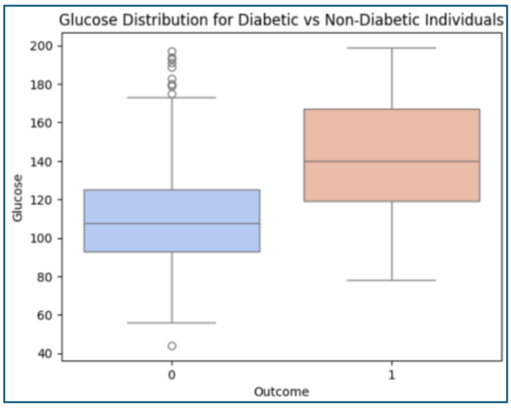


```
sns.boxplot(x="Outcome", y="BMI", data=df, palette="coolwarm")
plt.title("BMI Distribution for Diabetic vs Non-Diabetic Individuals")
plt.show()
```



```
sns.boxplot(x="Outcome", y="Glucose", data=df, palette="coolwarm")
plt.title("Glucose Distribution for Diabetic vs Non-Diabetic Individuals")
plt.show()
```

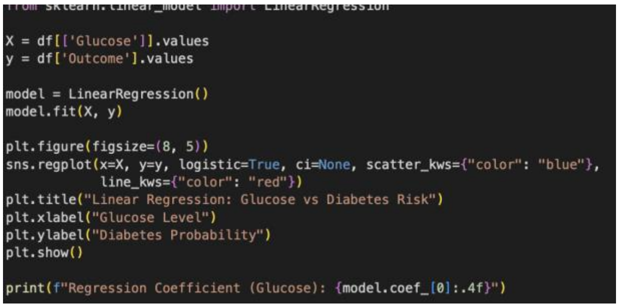




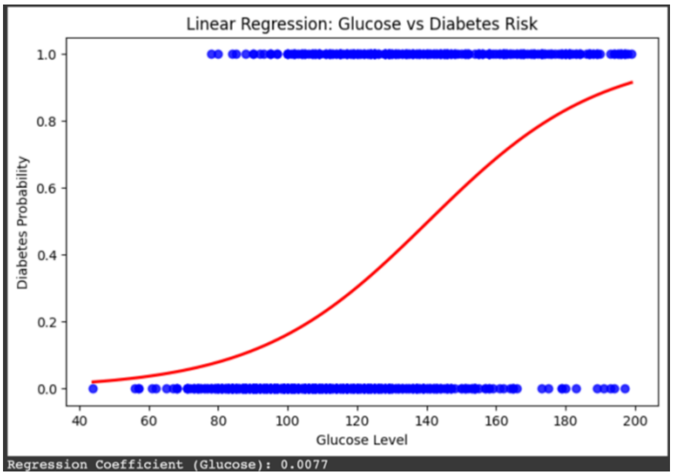
Interpretation

i. Diabetic patients have higher BMI & Glucose levels than non-diabetic patients.
ii. Preventive strategies should focus on reducing BMI & glucose levels.

- Regression Analysis (Linear Regression) - Does Glucose Predict Diabetes Risk?



```
from sklearn.linear_model import LinearRegression

X = df[['Glucose']].values
y = df['Outcome'].values

model = LinearRegression()
model.fit(X, y)

plt.figure(figsize=(8, 5))
sns.regplot(x=X, y=y, logistic=True, ci=None, scatter_kws={"color": "blue"},
            line_kws={"color": "red"})
plt.title("Linear Regression: Glucose vs Diabetes Risk")
plt.xlabel("Glucose Level")
plt.ylabel("Diabetes Probability")
plt.show()

print(f"Regression Coefficient (Glucose): {model.coef_[0]:.4f}")
```

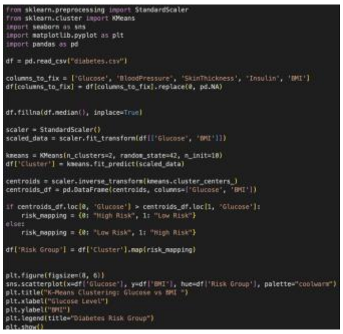


```
Regression Coefficient (Glucose): 0.0077
```

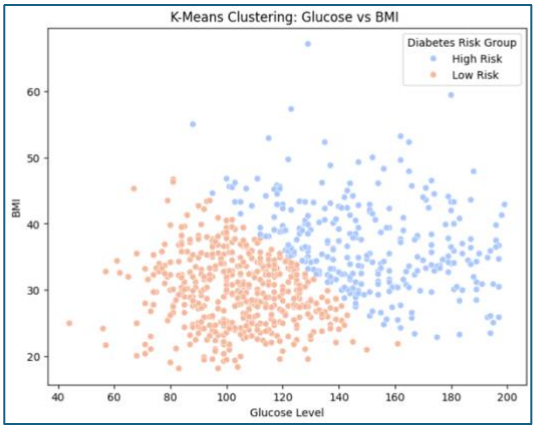
Interpretation

i. Higher glucose levels increase diabetes probability.
ii. Glucose has a positive regression coefficient, confirming its impact on diabetes.

- Clustering Analysis (K-Means) - Identifying Diabetes Risk Groups



```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv("diabetes.csv")

columns_to_fix = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
df[columns_to_fix] = df[columns_to_fix].replace(0, pd.NA)


df.fillna(df.median(), inplace=True)

scaler = StandardScaler()
scaled_data = scaler.fit_transform(df[['Glucose', 'BMI']])

kmeans = KMeans(n_clusters=2, random_state=42, n_init=10)
df['Cluster'] = kmeans.fit_predict(scaled_data)

centroids = scaler.inverse_transform(kmeans.cluster_centers_)
centroids_df = pd.DataFrame(centroids, columns=['Glucose', 'BMI'])

if centroids_df.loc[0, 'Glucose'] > centroids_df.loc[1, 'Glucose']:
    risk_mapping = {0: "High Risk", 1: "Low Risk"}
else:
    risk_mapping = {0: "Low Risk", 1: "High Risk"}

df['Risk Group'] = df['Cluster'].map(risk_mapping)


plt.figure(figsize=(8, 6))
sns.scatterplot(x=df['Glucose'], y=df['BMI'], hue=df['Risk Group'], palette="coolwarm")
plt.title("K-Means Clustering: Glucose vs BMI ")
plt.xlabel("Glucose Level")
plt.ylabel("BMI")
plt.legend(title="Diabetes Risk Group")
plt.show()
```



Interpretation

i. K-Means clustering divides patients into high-risk (red) and low-risk (blue) diabetes groups.
ii. Higher glucose & BMI indicate higher diabetes risk.
iii. This clustering helps in identifying high-risk individuals for early intervention

4. Findings & Business Implications

- Glucose is the strongest predictor of diabetes.
- BMI and Age also significantly impacting diabetes risk.
- Diabetes prevalence is higher in older individuals (>40 years).
- For Healthcare Providers
  - Target high-risk groups (High BMI & Glucose) for early screening.
  - Personalized treatment plans for weight loss & lifestyle changes.
- Insurance Companies can adjust premiums based on diabetes risk factors.
- Government & Health Agencies can spread awareness on diet & exercise.