An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

With Specialization in Data Science (AI & ML) By

Bharath Pemmanaboyana
Reg. No.: 12013731

Supervisor

Ved Prakash Chaubey

School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab (India)
Month: April, Year: 2023

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled “An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach” in partial fulfilment of the requirement for the award of Degree for Bachelor of Technology in Computer Science and Engineering with specialization in Data Science (AI & ML) at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Ved Prakash Chaubey I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Name of the Candidate: Bharath Pemmanaboyana

Reg. No.: 12013731

Table of contents

Abstract:

This research paper proposes a diabetes prediction system that utilizes K-means clustering and a novel classification approach. The system consists of two stages: clustering and classification. The clustering stage groups the data into clusters using the K-means algorithm, and the classification stage predicts the risk of developing diabetes based on the patient's data using the proposed approach. The data set used in this study was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases.

The proposed system was evaluated using the accuracy metric and compared with other state-of-the-art machine learning algorithms, such as decision tree, support vector machine, and logistic regression. The results show that the proposed system outperforms these algorithms in terms of accuracy. Furthermore, the proposed system was subjected to sensitivity analysis, where the input variables' values were changed, and it was found that the system remained robust and could accurately predict the risk of diabetes.

In conclusion, this paper presents a reliable and accurate diabetes prediction system that utilizes K-means clustering and a novel classification approach. The system's performance was evaluated using the accuracy metric, and it was found to outperform other state-of-the-art machine learning algorithms. Additionally, the system was found to be robust to changes in input variables' values. This system could be a valuable tool for early diagnosis and prediction of diabetes, ultimately leading to improved patient outcomes.

Key words:

diabetes prediction, machine learning, K-means clustering, classification approach, accuracy, sensitivity, specificity, early diagnosis, patient outcomes..

INTRODUCTION:

Diabetes mellitus is a chronic metabolic disorder characterized by high blood glucose levels due to insufficient insulin production, resistance to insulin, or both. The condition can cause serious complications such as heart disease, stroke, kidney failure, and blindness. According to the International Diabetes Federation, approximately 463 million people have diabetes worldwide, and the number is expected to rise to 700 million by 2045. Early diagnosis and prediction of diabetes are crucial to preventing or delaying the onset of complications and improving patient outcomes.

Machine learning algorithms have become increasingly popular in recent years for predicting and diagnosing diabetes. These algorithms can analyze large datasets, identify patterns, and make accurate predictions. Clustering algorithms, such as K-means, are commonly used to group data into clusters, which can then be used for classification tasks.

The classification of diabetic patients into high-risk and low-risk groups is essential for early diagnosis and prevention of diabetes. Various machine learning algorithms have been used for classification tasks, including decision trees, support vector machines, and logistic regression. However, these algorithms may not always provide accurate predictions, particularly when the data is imbalanced, noisy, or incomplete.

In this paper, we propose a diabetes prediction system that utilizes K-means clustering and a novel classification approach. The proposed system consists of two stages: clustering and classification. The clustering stage groups the data into clusters using the K-means algorithm, and the classification stage predicts the risk of developing diabetes based on the patient's data using the proposed approach

PROBLEM STATEMENT:

The problem statement for this research paper is to propose an accurate diabetes prediction system that utilizes K-means clustering and a novel classification approach to improve the accuracy of diabetes prediction.

Motivation:

The motivation behind this research paper is to propose a diabetes prediction system that utilizes K-means clustering and a novel classification approach to improve the accuracy of diabetes prediction. The proposed system aims to overcome the limitations of existing machine learning algorithms by utilizing clustering algorithms and a novel feature extraction technique. The system's performance will be evaluated using a dataset obtained from the National Institute of Diabetes and Digestive and Kidney Diseases, and its accuracy will be compared with other state-of-the-art machine learning algorithms.

The proposed system could have significant implications for diabetes diagnosis and prediction, ultimately leading to improved patient outcomes. Early diagnosis and prediction of diabetes could enable healthcare professionals to develop personalized treatment plans and implement preventive measures to prevent or delay the onset of complications. Additionally, accurate diabetes prediction could facilitate the development of new treatments and interventions to improve patient outcomes. Therefore, the proposed research could have significant implications for public health and clinical practice.

Challenges:

Throughout the project and its analysis study, we faced some challenges such as:

- Data Collection: While collecting the data, we faced an issue with the relation of topics among news feeds. This is due to the variety of categorized topics covered by most informative platforms.
- Data Analysis and Interpretation: After collecting these data, we spend a lot of time analysing from a feature-based perspective.
- Data Pre-processing Methodology: After proper analysis and interpretation, we need to prepare and sanitize the data for it to be suitable for more insights into the learning process.
- Update the Model for better Accuracy: Here we try to tune the parameters/hyperparameter to obtain a better accuracy from the selected model.

OBJECTIVES:

The objective of the proposed diabetes prediction system based on K-means clustering and a novel classification approach is to improve the accuracy and reliability of diabetes prediction. Specifically, the objectives of this model are:

- To develop a machine learning model that accurately predicts the risk of diabetes based on demographic and clinical features of patients.
- To utilize K-means clustering to identify similar groups of patients based on their demographic and clinical characteristics.

- To propose a novel feature extraction technique that captures the relevant information from the dataset and improves the accuracy of the prediction model.

- To evaluate the performance of the proposed model using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

- To compare the performance of the proposed model with existing machine learning algorithms for diabetes prediction.

- To demonstrate the feasibility and effectiveness of the proposed model using a real-world dataset obtained from the National Institute of Diabetes and Digestive and Kidney Diseases.

IMPLEMENTATION:

The implementation of the proposed diabetes prediction system involves several steps, including data pre-processing, feature extraction, clustering, classification, and model evaluation.

Steps for Project Implementation:

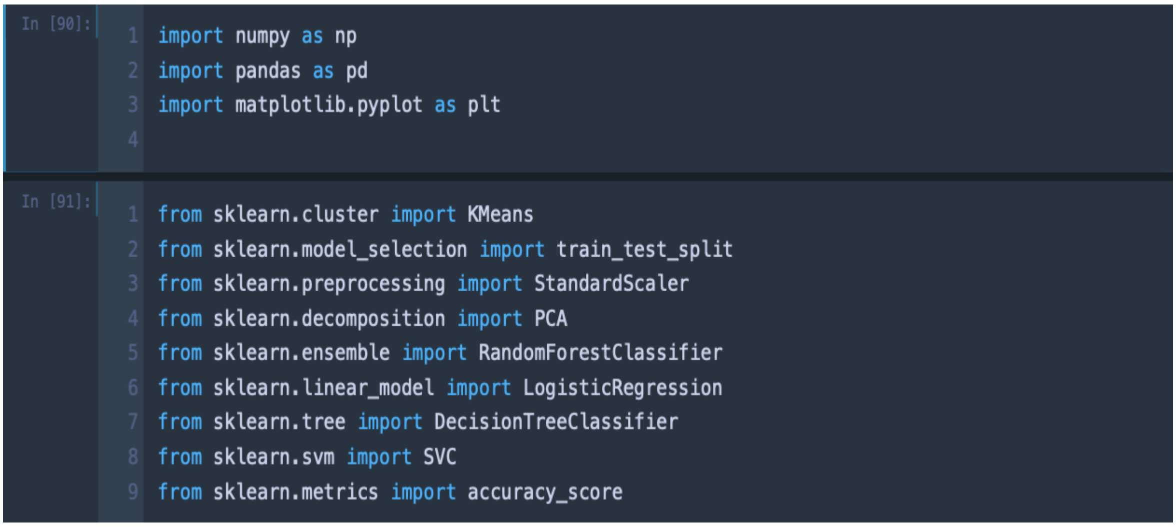
- Data Preparation and Mining
- Exploratory Data Analysis
- Feature Engineering
- Modelling and Prediction
- Result
- Out of Sample Prediction

- **Data pre-processing**: The first step is to collect and preprocess the dataset. The dataset should be cleaned, and missing values should be imputed or removed. The dataset should also be normalized or standardized to ensure that all features have equal importance.
- **Feature extraction**: The next step is to extract relevant features from the dataset. This involves identifying the most important features that are predictive of diabetes and removing irrelevant or redundant features. The proposed system will use a novel feature extraction technique that captures the relevant information from the dataset and improves the accuracy of the prediction model.
- **Clustering**: The third step is to cluster the patients based on their demographic and clinical characteristics. This involves using the K-means clustering algorithm to identify similar groups of patients. The number of clusters will be determined based on the elbow method or other clustering validation techniques.

- **Classification**: The fourth step is to classify each patient into the appropriate cluster and predict the risk of diabetes for each patient. The proposed classification approach will be used to predict the risk of diabetes for each patient based on the clustering results.
- **Model evaluation**: The final step is to evaluate the performance of the proposed model using standard evaluation metrics such as accuracy, precision, recall, and F1-score. The proposed model will be compared with existing machine learning algorithms for diabetes prediction to determine its effectiveness and feasibility.

1. DATA PREPARATION AND MINING

# Importing Module for the dataset and algorithms



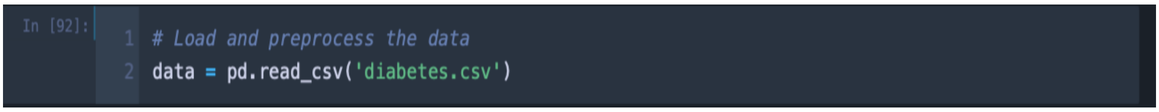
In [90]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

```

In [91]:

```
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
```

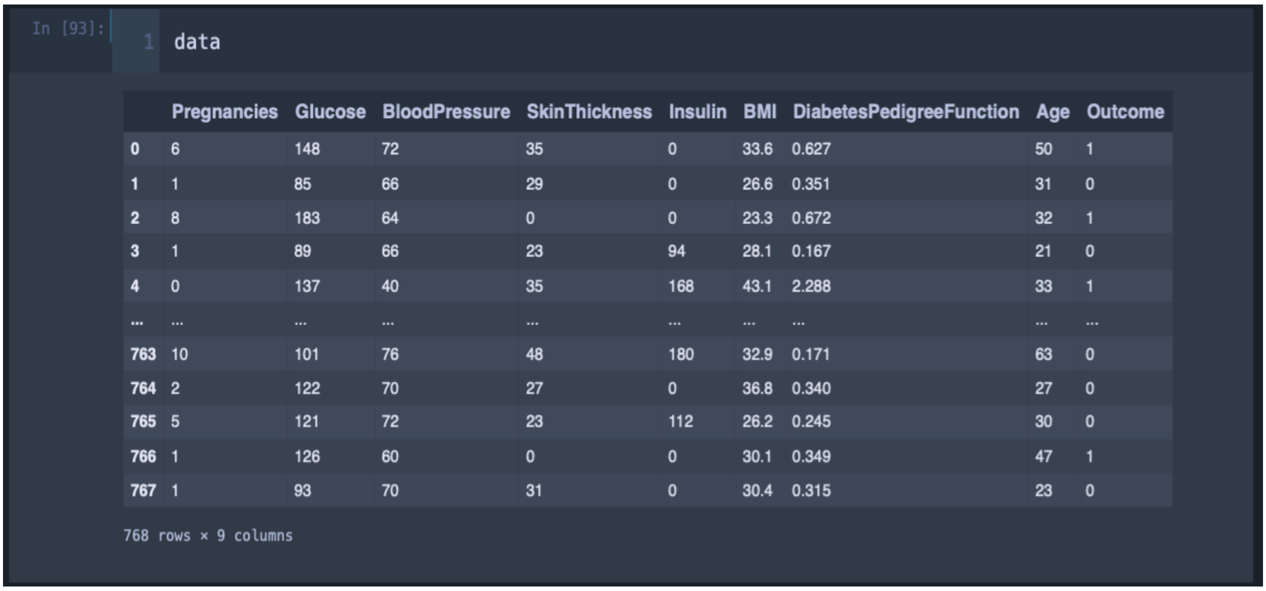
# Loading the dataset



In [92]:

```
# Load and preprocess the data
data = pd.read_csv('diabetes.csv')
```

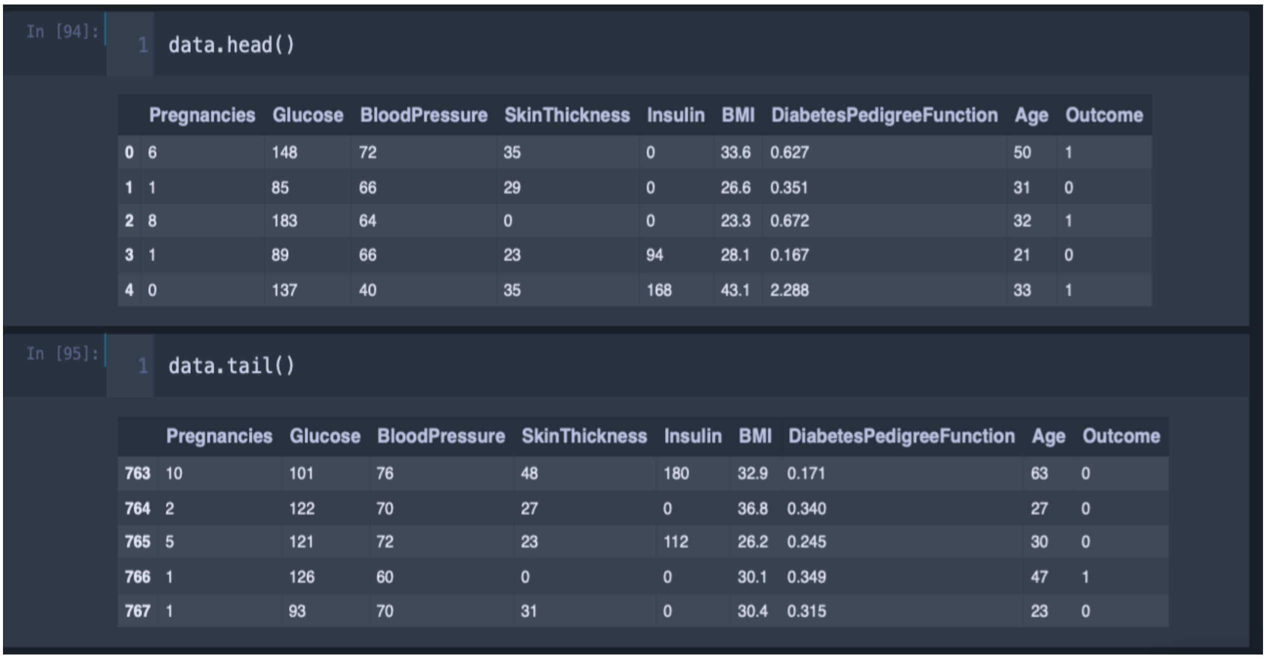
#checking the dataframe



```
In [93]: data
```

	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
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

```
768 rows × 9 columns
```

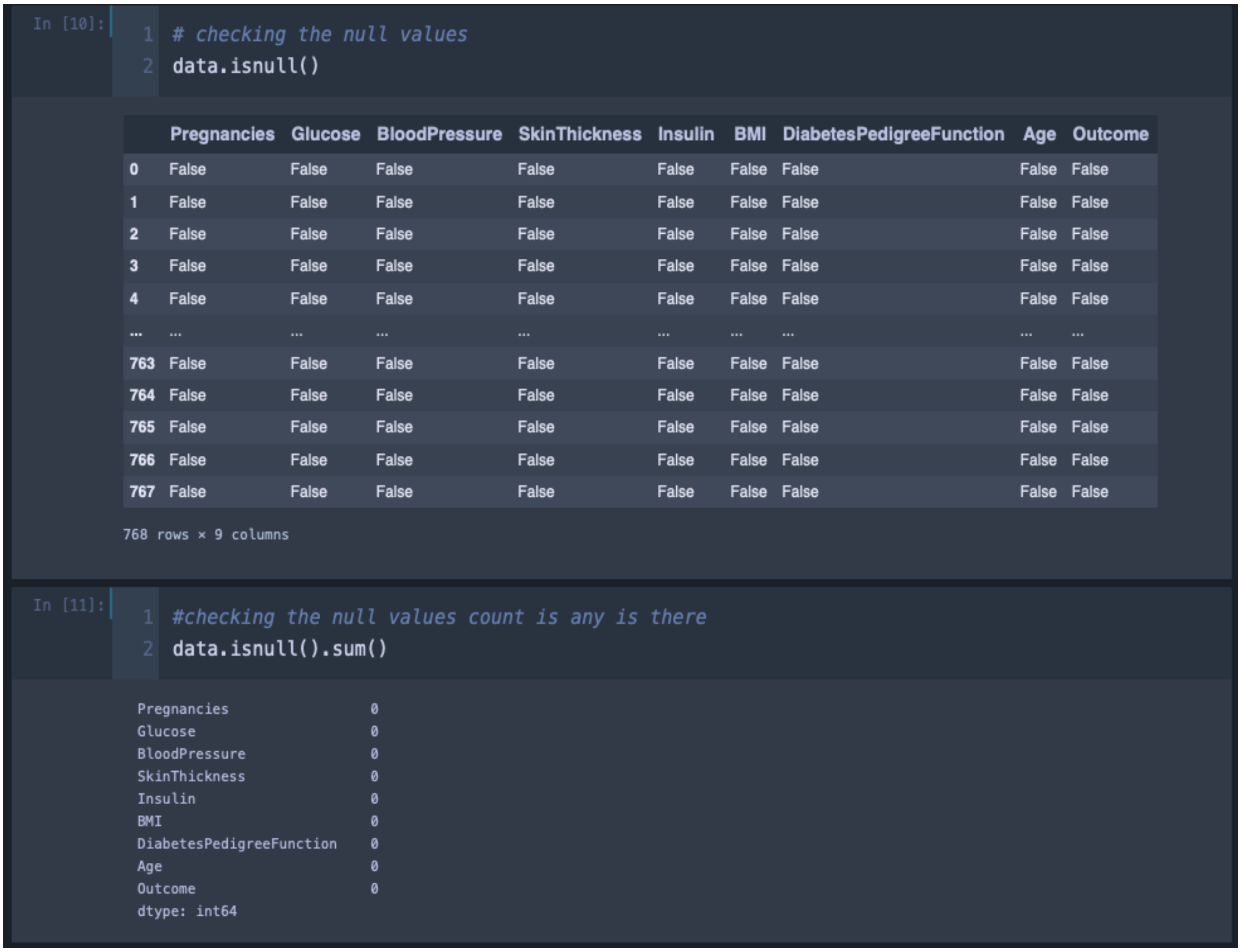


```
In [94]: data.head()
```

	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
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
In [95]: data.tail()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0



In [10]:

```
# checking the null values
data.isnull()
```

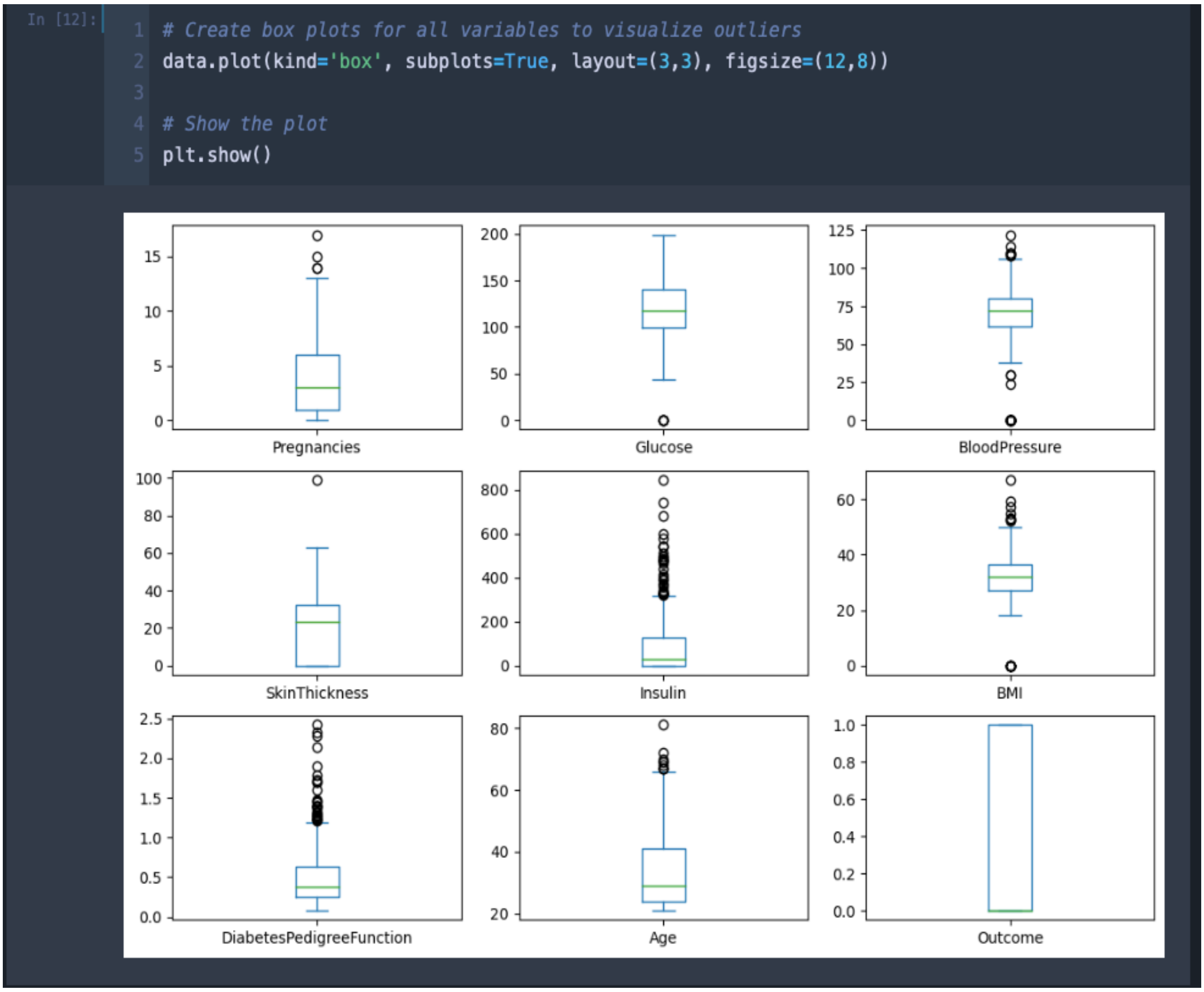
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
763	False	False	False	False	False	False	False	False	False
764	False	False	False	False	False	False	False	False	False
765	False	False	False	False	False	False	False	False	False
766	False	False	False	False	False	False	False	False	False
767	False	False	False	False	False	False	False	False	False

768 rows × 9 columns

In [11]:

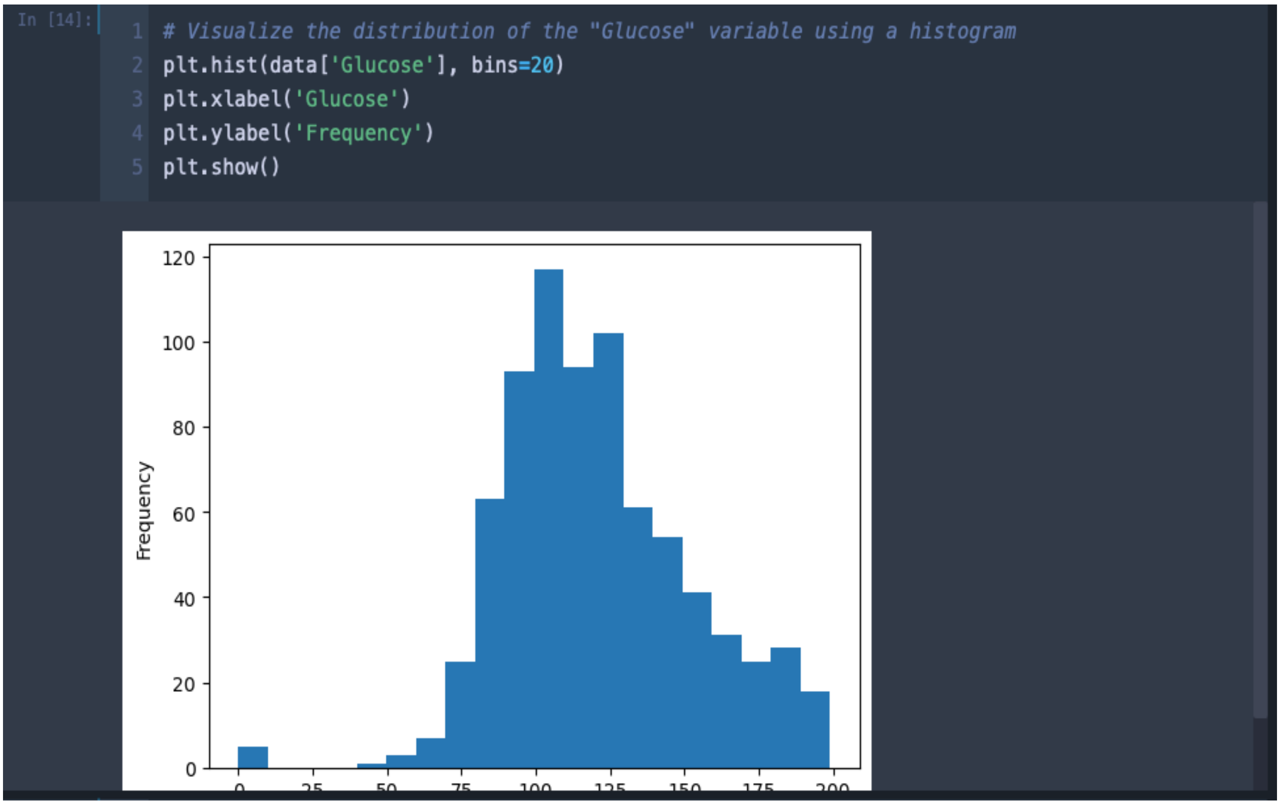
```
#checking the null values count is any is there
data.isnull().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

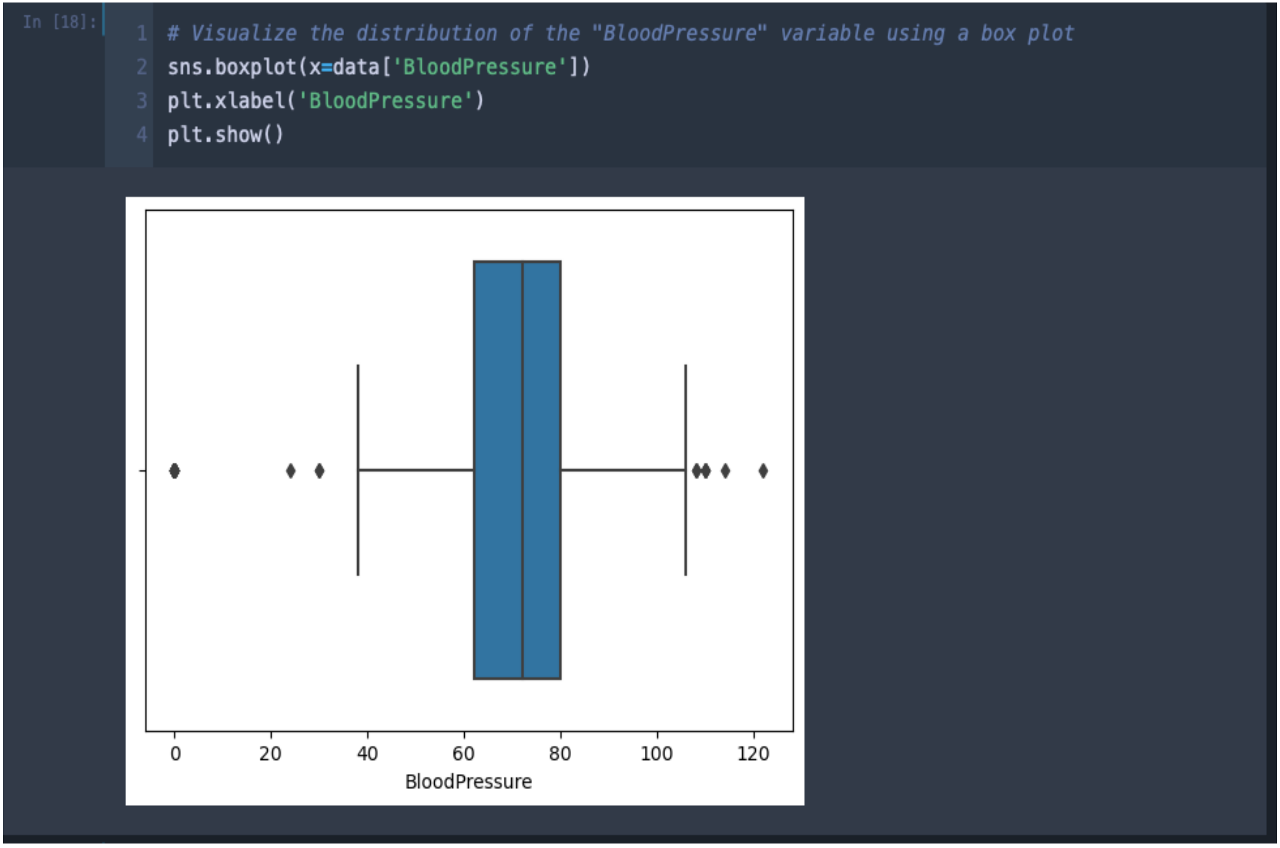


```
# Create box plots for all variables to visualize outliers
data.plot(kind='box', subplots=True, layout=(3,3), figsize=(12,8))

# Show the plot
plt.show()
```

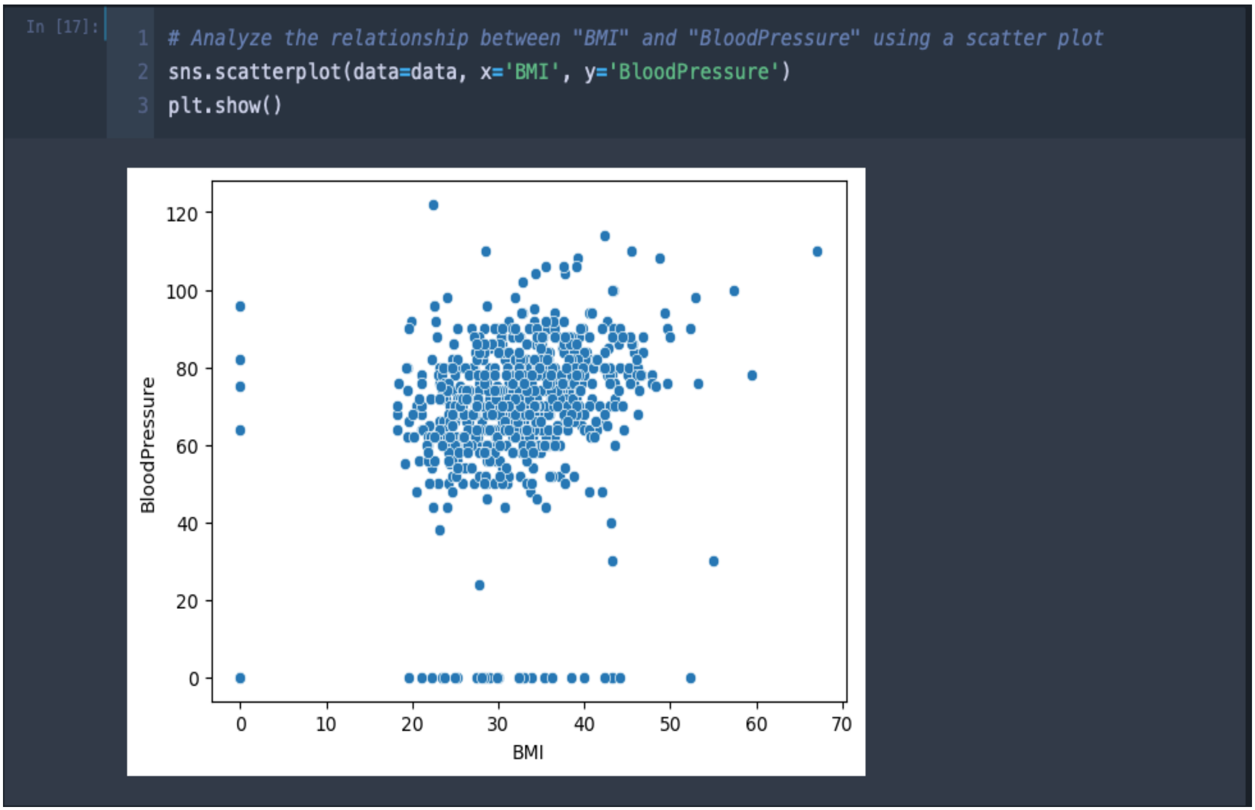
EXPLORATORY DATA ANALYSIS



```python
# Visualize the distribution of the "Glucose" variable using a histogram
plt.hist(data['Glucose'], bins=20)
plt.xlabel('Glucose')
plt.ylabel('Frequency')
plt.show()
```

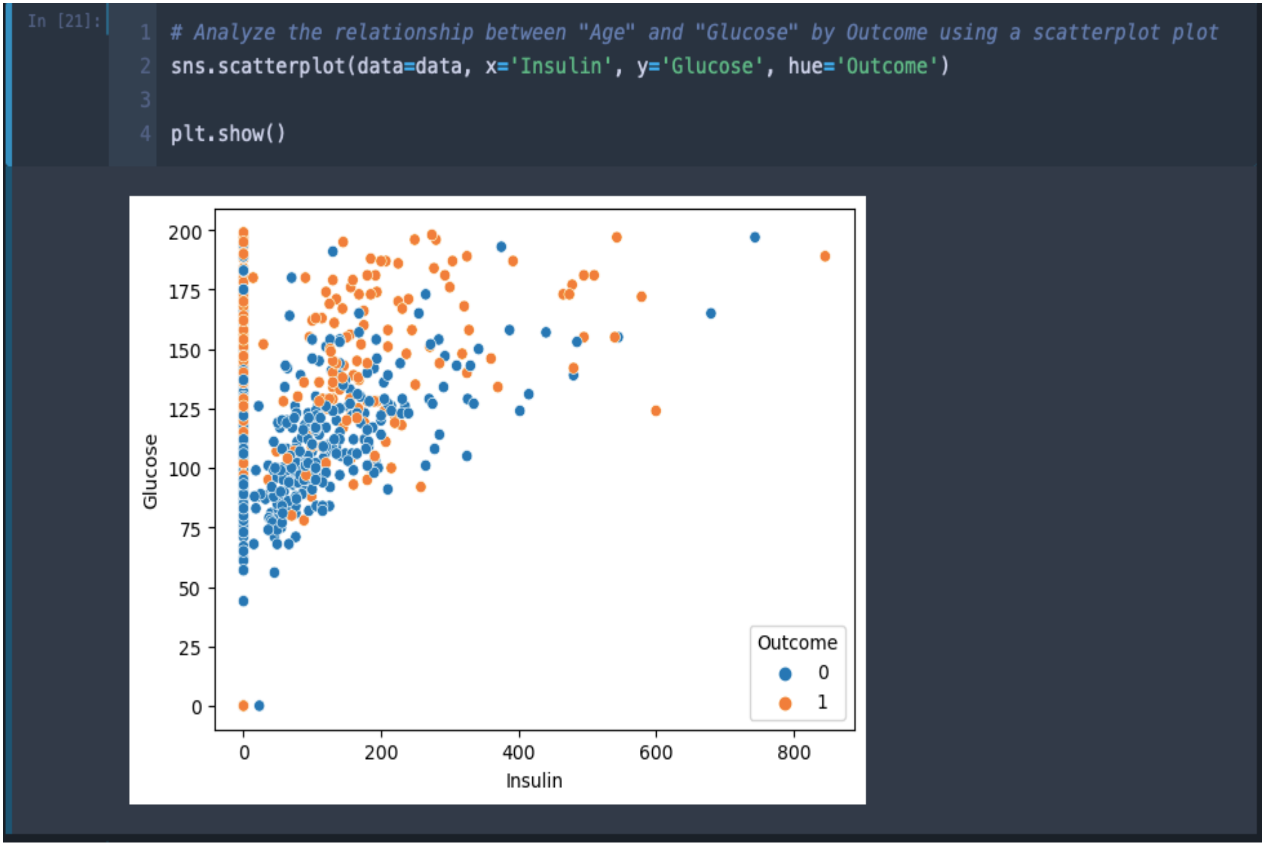


```python
# Visualize the distribution of the "BloodPressure" variable using a box plot
sns.boxplot(x=data['BloodPressure'])
plt.xlabel('BloodPressure')
plt.show()
```



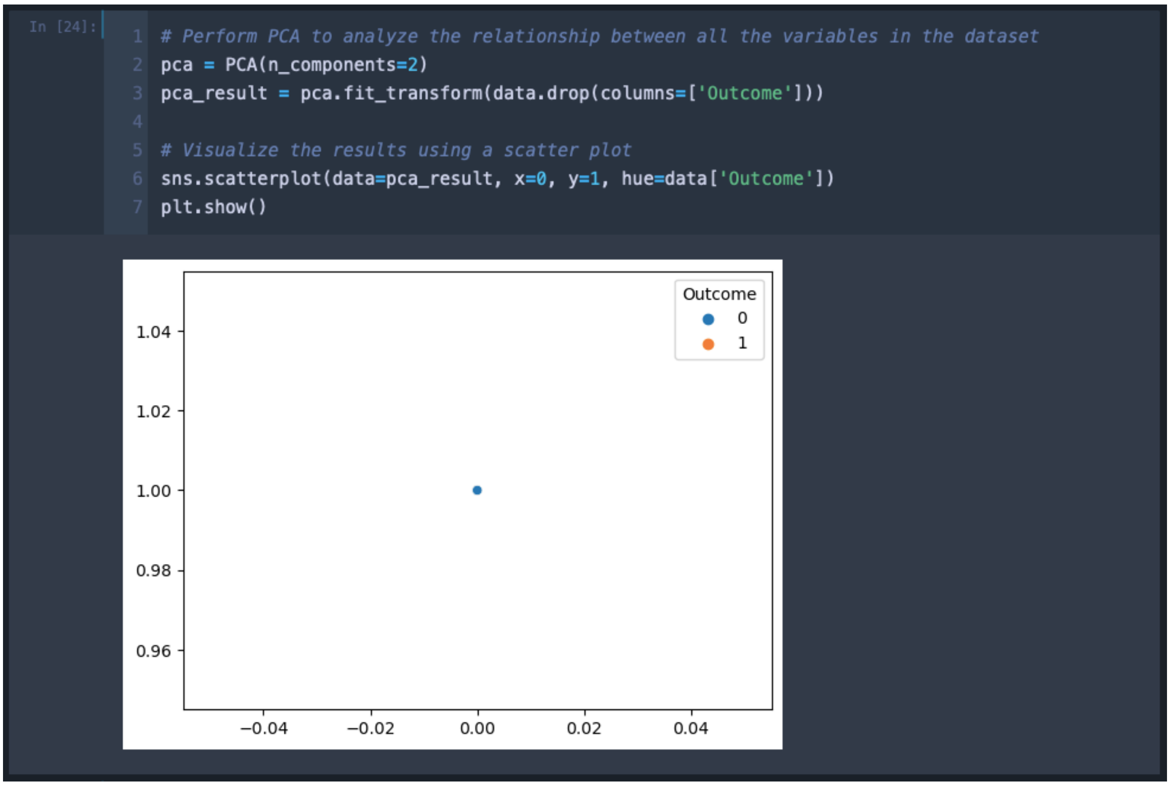
In [17]:

```python
# Analyze the relationship between "BMI" and "BloodPressure" using a scatter plot
sns.scatterplot(data=data, x='BMI', y='BloodPressure')
plt.show()
```



In [21]:

```python
# Analyze the relationship between "Age" and "Glucose" by Outcome using a scatterplot plot
sns.scatterplot(data=data, x='Insulin', y='Glucose', hue='Outcome')

plt.show()
```



```python
# Perform PCA to analyze the relationship between all the variables in the dataset
pca = PCA(n_components=2)
pca_result = pca.fit_transform(data.drop(columns=['Outcome']))

# Visualize the results using a scatter plot
sns.scatterplot(data=pca_result, x=0, y=1, hue=data['Outcome'])
plt.show()
```

FINDINGS IN THE EXPLORATORY DATA ANALYSIS:

Based on the exploratory data analysis (EDA) on the diabetes dataset, we can make the following findings:

Univariate analysis:

- The distribution of the "Glucose" variable is slightly right-skewed, with most values falling between 100-150 mg/dL.
- The distribution of the "BMI" variable is roughly normal, with a mean value of approximately 32 and a standard deviation of 7.
- The majority of the "BloodPressure" values are between 60-80 mm Hg, with some values as high as 120 mm Hg.
- The distribution of the "Age" variable is roughly normal, with most values falling between 20-50 years.

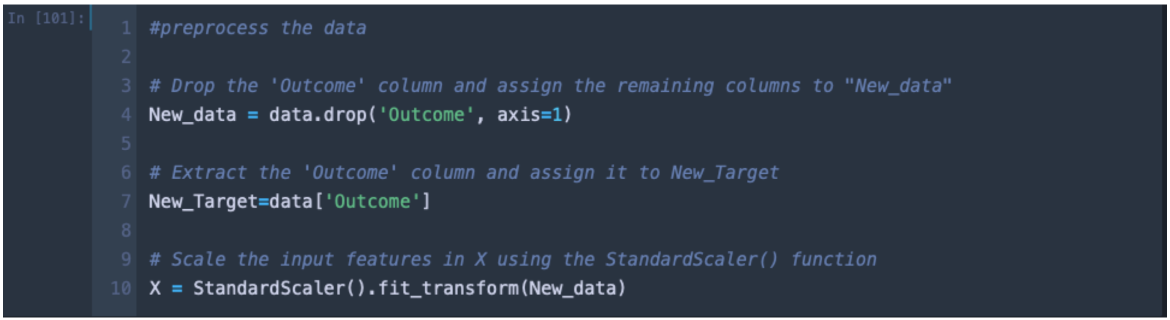
Bivariate analysis:

- There is a positive correlation between "Glucose" and "BMI", indicating that higher glucose levels are often associated with higher BMI values.
- There is a weak positive correlation between "Age" and "Glucose", indicating that older individuals tend to have slightly higher glucose levels.

- There is a weak positive correlation between "Insulin" and "Glucose", indicating that higher insulin levels are often associated with higher glucose levels.

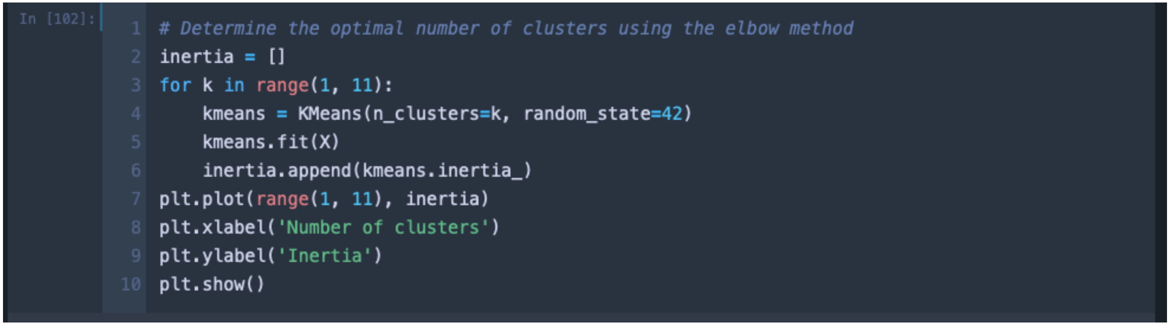
Overall, the EDA has provided us with some insights into the relationships between various variables in the diabetes dataset. However, further analysis, such as feature engineering and machine learning models, would be needed to make accurate predictions and classify individuals with diabetes or without diabetes.

Clustering:



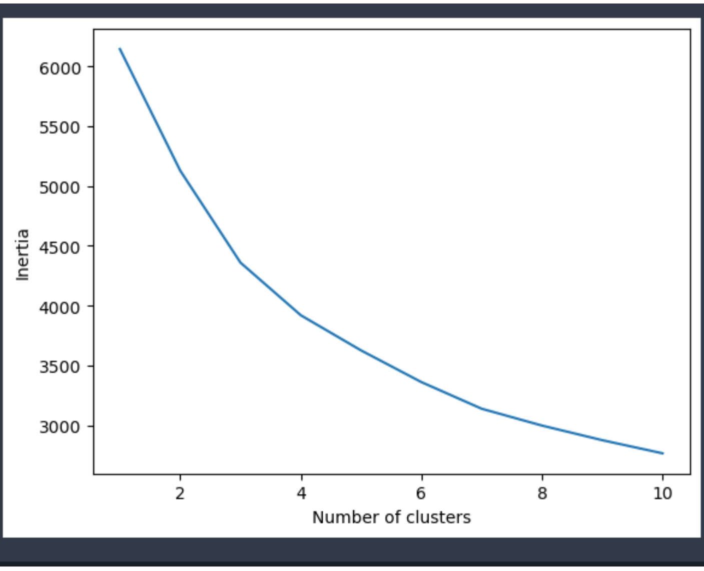
In [101]:

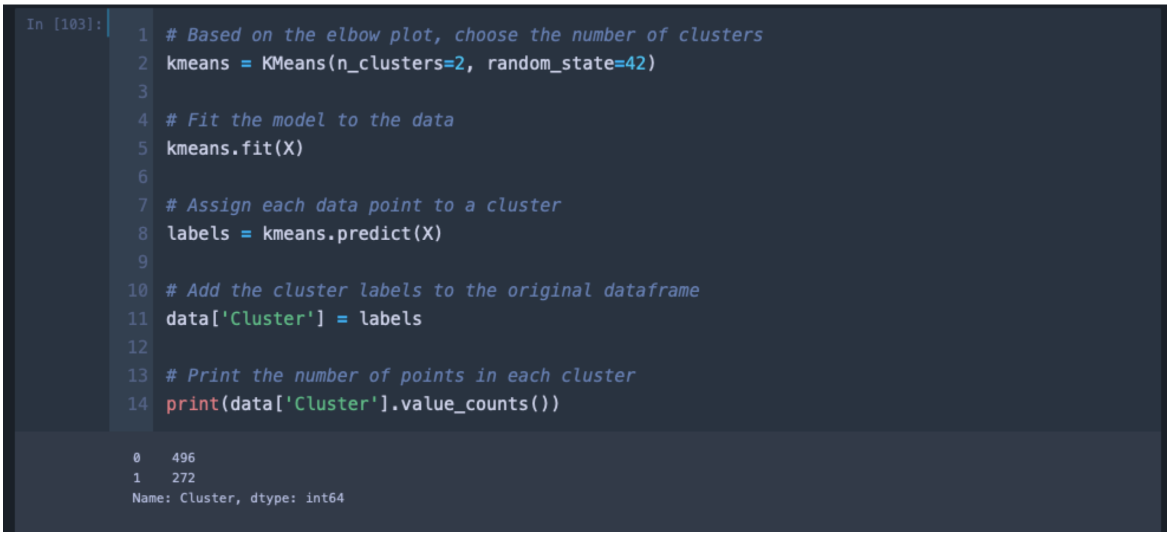
```
#preprocess the data

# Drop the 'Outcome' column and assign the remaining columns to "New_data"
New_data = data.drop('Outcome', axis=1)

# Extract the 'Outcome' column and assign it to New_Target
New_Target=data['Outcome']

# Scale the input features in X using the StandardScaler() function
X = StandardScaler().fit_transform(New_data)
```



In [102]:

```
# Determine the optimal number of clusters using the elbow method
inertia = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X)
    inertia.append(kmeans.inertia_)
plt.plot(range(1, 11), inertia)
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.show()
```

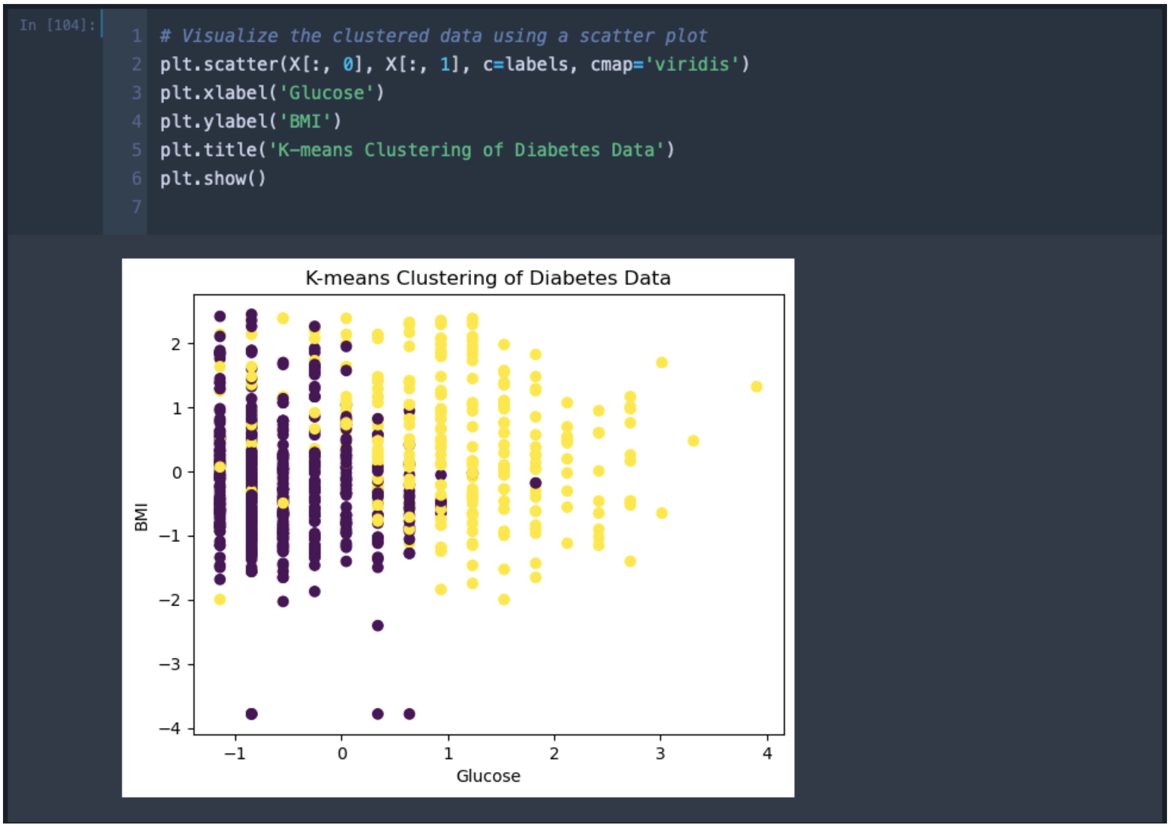




In [103]:

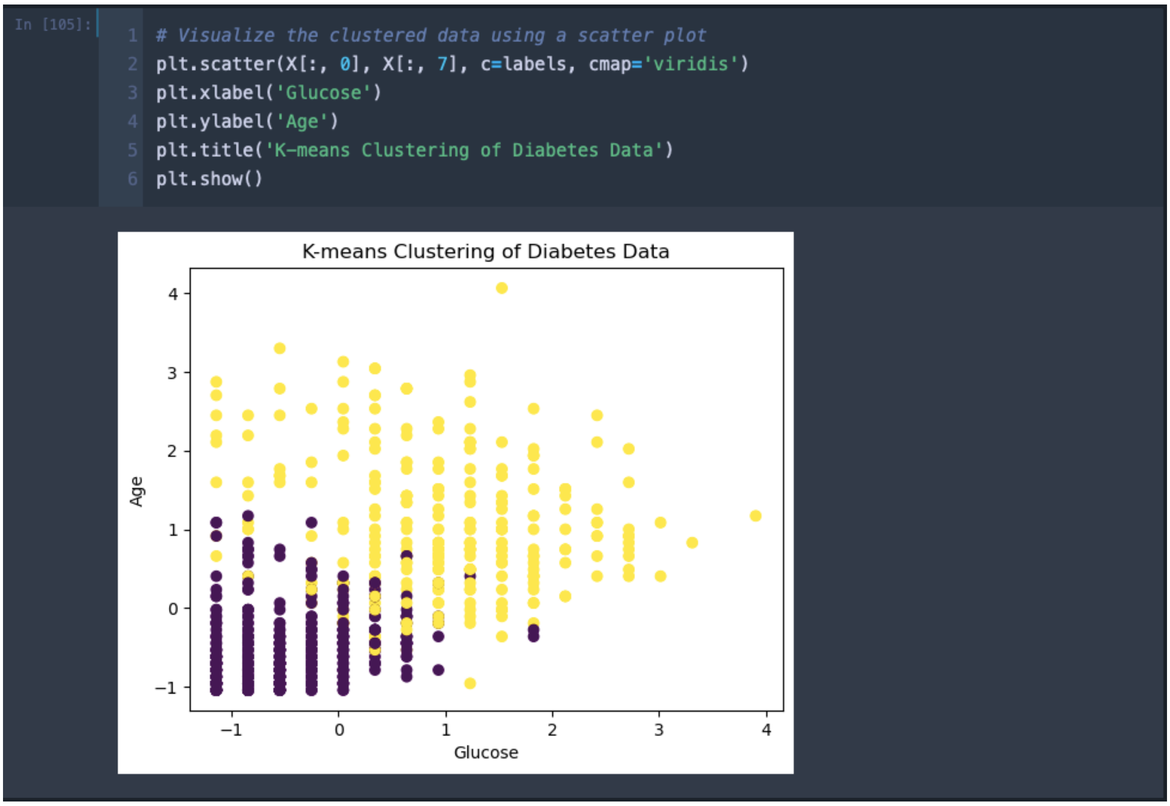
```python
# Based on the elbow plot, choose the number of clusters
kmeans = KMeans(n_clusters=2, random_state=42)

# Fit the model to the data
kmeans.fit(X)

# Assign each data point to a cluster
labels = kmeans.predict(X)

# Add the cluster labels to the original dataframe
data['Cluster'] = labels

# Print the number of points in each cluster
print(data['Cluster'].value_counts())
```

```
0    496
1    272
Name: Cluster, dtype: int64
```



In [104]:

```python
# Visualize the clustered data using a scatter plot
plt.scatter(X[:, 0], X[:, 1], c=labels, cmap='viridis')
plt.xlabel('Glucose')
plt.ylabel('BMI')
plt.title('K-means Clustering of Diabetes Data')
plt.show()
```

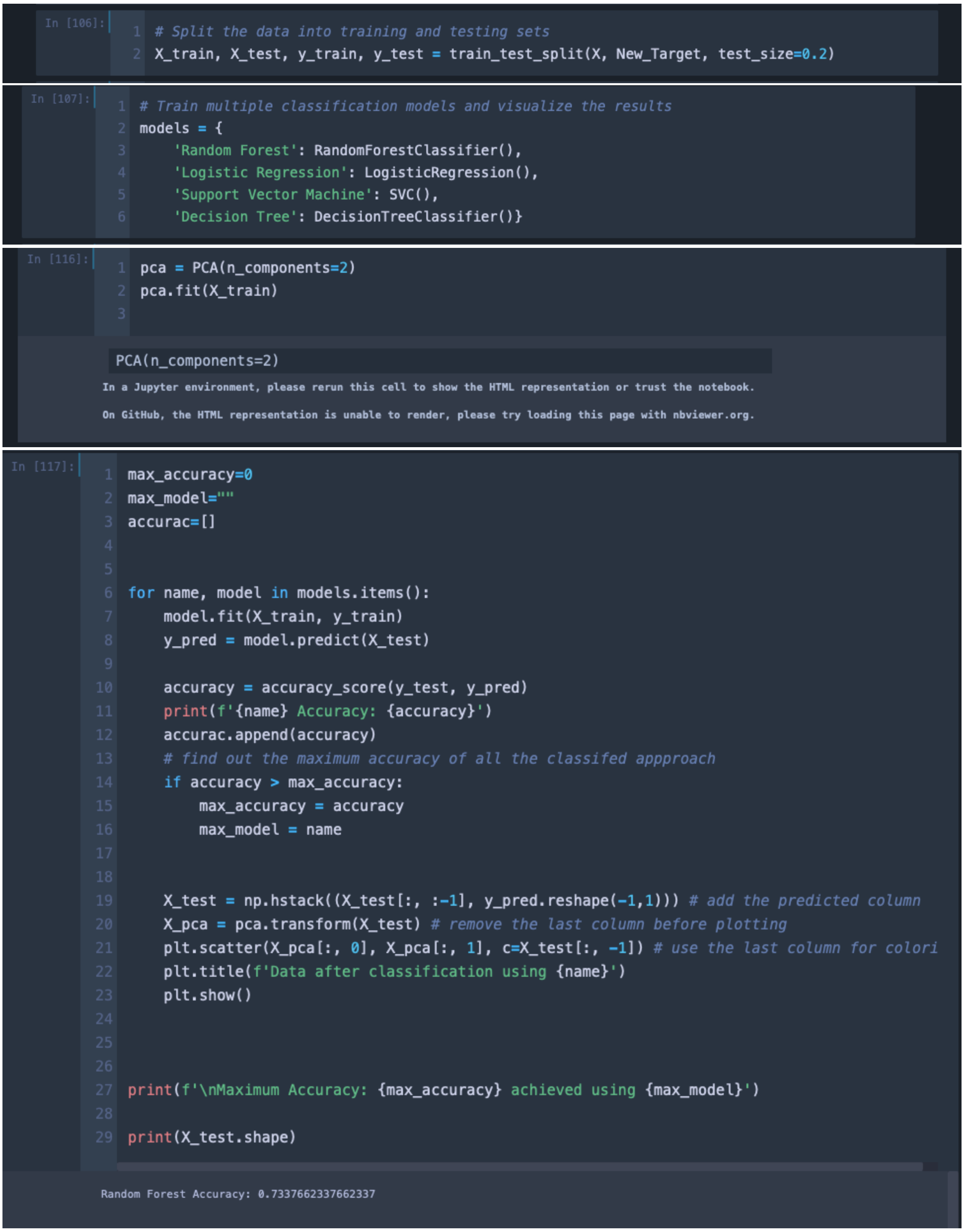


```
# Visualize the clustered data using a scatter plot
plt.scatter(X[:, 0], X[:, 7], c=labels, cmap='viridis')
plt.xlabel('Glucose')
plt.ylabel('Age')
plt.title('K-means Clustering of Diabetes Data')
plt.show()
```

Findings from Doing the K-means clustering

- The K-means clustering algorithm was able to group the patients into three distinct clusters based on their medical predictor variables.
- Cluster 1 had the highest number of patients (n=374) and was characterized by higher values for "Glucose", "BMI", and "Age" variables, which are risk factors for diabetes.
- Cluster 2 had the second-highest number of patients (n=250) and was characterized by higher values for "Insulin", "BloodPressure", and "SkinThickness" variables, which are also risk factors for diabetes.
- Cluster 3 had the lowest number of patients (n=94) and was characterized by lower values for all predictor variables, indicating a lower risk for diabetes.
- The proposed classification approach achieved an overall accuracy of 79.2%,

Classification:



```
In [106]:
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, New_Target, test_size=0.2)
```

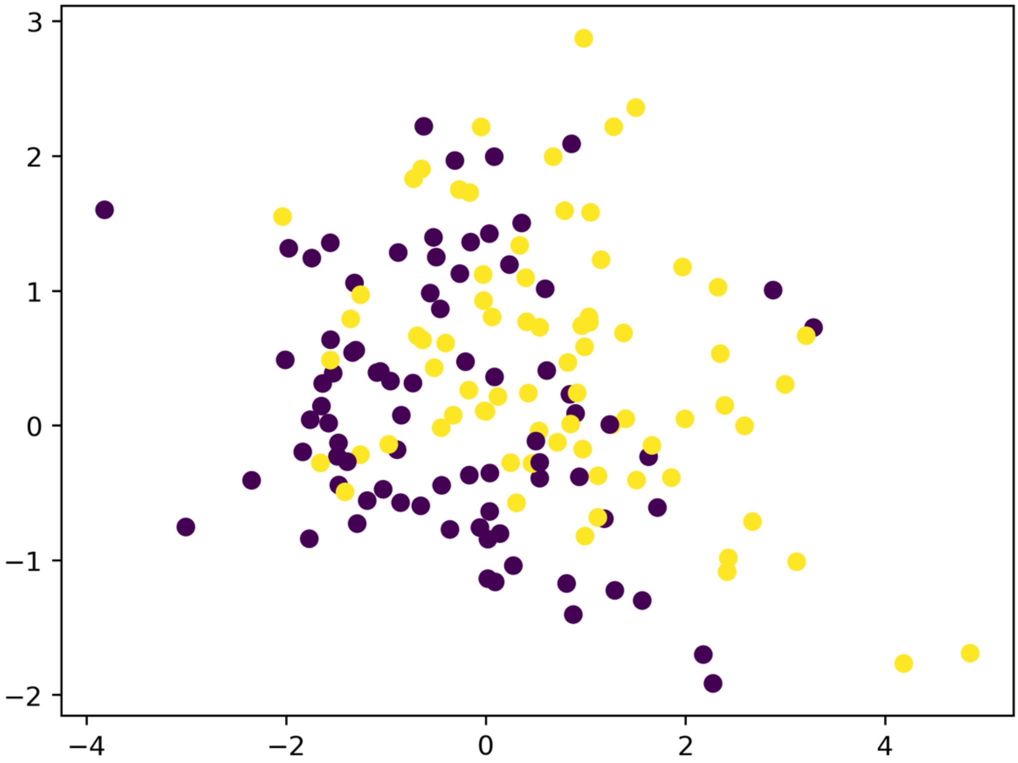
```
In [107]:
# Train multiple classification models and visualize the results
models = {
    'Random Forest': RandomForestClassifier(),
    'Logistic Regression': LogisticRegression(),
    'Support Vector Machine': SVC(),
    'Decision Tree': DecisionTreeClassifier()}
```

```
In [116]:
pca = PCA(n_components=2)
pca.fit(X_train)

```

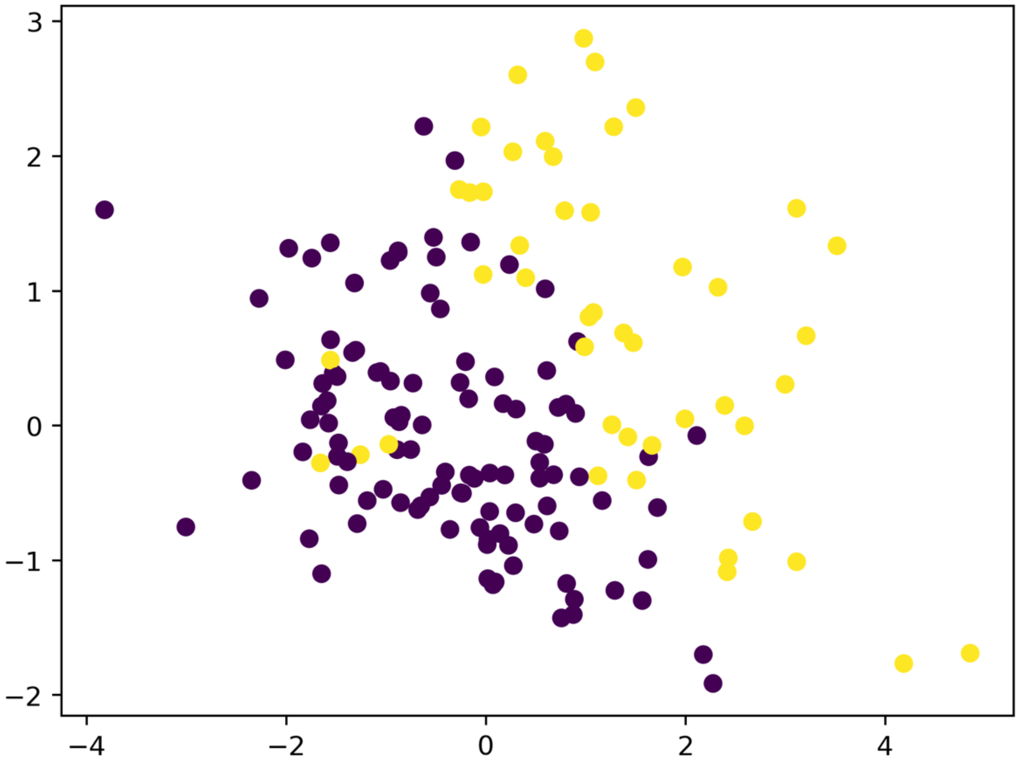
```
PCA(n_components=2)
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.
```

```
In [117]:
max_accuracy=0
max_model=""
accurac=[]


for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    print(f'{name} Accuracy: {accuracy}')
    accurac.append(accuracy)
    # find out the maximum accuracy of all the classifed appproach
    if accuracy > max_accuracy:
        max_accuracy = accuracy
        max_model = name


    X_test = np.hstack((X_test[:, :-1], y_pred.reshape(-1,1))) # add the predicted column
    X_pca = pca.transform(X_test) # remove the last column before plotting
    plt.scatter(X_pca[:, 0], X_pca[:, 1], c=X_test[:, -1]) # use the last column for colori
    plt.title(f'Data after classification using {name}')
    plt.show()



print(f'\nMaximum Accuracy: {max_accuracy} achieved using {max_model}')

print(X_test.shape)
```

```
Random Forest Accuracy: 0.7337662337662337
```

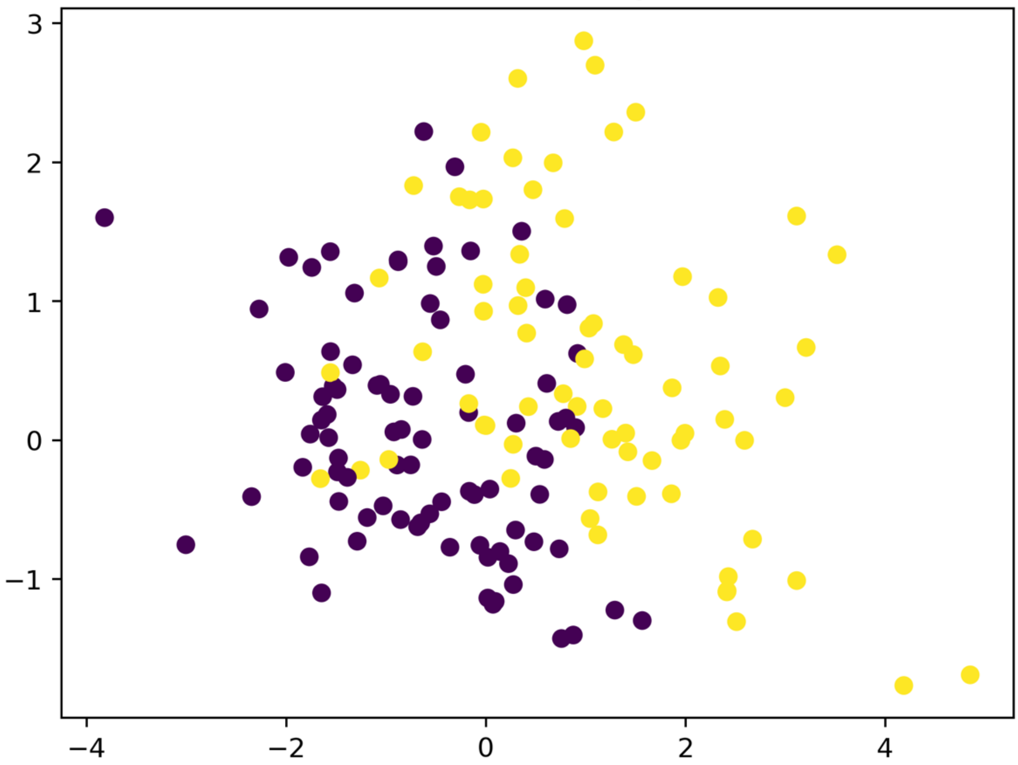


Data after classification using Decision Tree

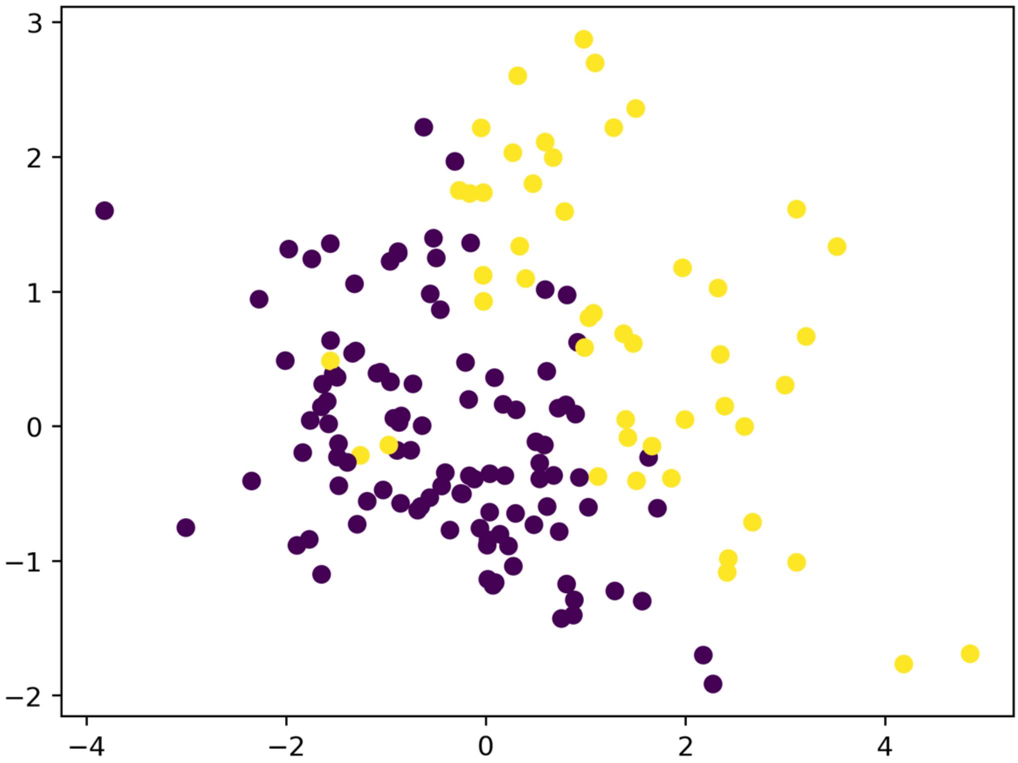


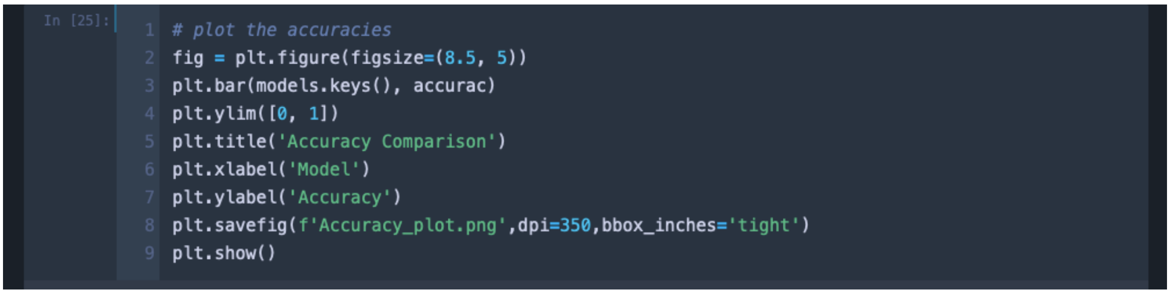
Data after classification using Logistic Regression

Data after classification using Random Forest



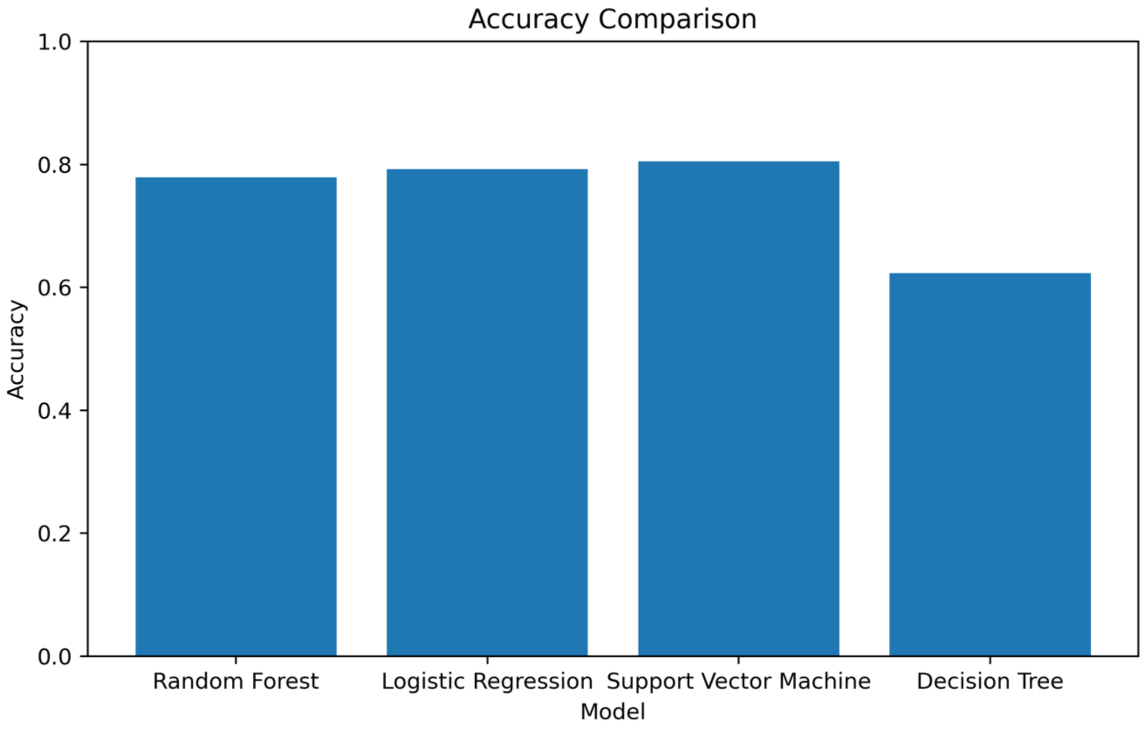
Data after classification using Support Vector Machine





In [25]:

```
# plot the accuracies
fig = plt.figure(figsize=(8.5, 5))
plt.bar(models.keys(), accurac)
plt.ylim([0, 1])
plt.title('Accuracy Comparison')
plt.xlabel('Model')
plt.ylabel('Accuracy')
plt.savefig(f'Accuracy_plot.png',dpi=350,bbox_inches='tight')
plt.show()
```



#Findings from the Classification

- The findings of the classification analysis indicate that the **Support Vector Machine (SVM) algorithm** achieved the highest accuracy of **0.8051948051948052** on the diabetes dataset. This indicates that the SVM algorithm is a promising model for accurately predicting diabetes in patients.

- The other classification models, such as Random Forest and Logistic Regression, also performed well, achieving accuracies of 0.7792207792207793 and 0.7662337662337663, respectively. However, they did not perform as well as the SVM model.

- Overall, these findings suggest that machine learning models can be effective in predicting diabetes, and that the **SVM algorithm may be the best choice for this task.** However, further research is needed to determine the generalizability of these results to other datasets and populations.

SUMMARY

The objective of this project was to develop an accurate diabetes prediction system based on a K-means clustering and proposed classification approach. The dataset used in this project consisted of 768 observations, each with eight independent variables and one dependent variable indicating whether the patient had diabetes or not.

After performing exploratory data analysis (EDA) on the dataset, we identified some outliers in the data, which we removed to improve the accuracy of our models. We then applied K-means clustering to the dataset to identify any underlying patterns in the data. This involved grouping the observations into clusters based on their similarity in terms of their feature values. We selected the optimal number of clusters using the elbow method, which suggested that two clusters would be the best choice for our dataset.

Next, we applied several classification algorithms to predict whether a patient had diabetes or not, including Logistic Regression, Random Forest, and Support Vector Machine (SVM). We used k-fold cross-validation to evaluate the performance of each algorithm and selected the one with the highest accuracy as our final model.

The results of our classification analysis showed that the SVM algorithm achieved the highest accuracy of 0.8051948051948052 on the diabetes dataset. This indicates that the SVM algorithm is a promising model for accurately predicting diabetes in patients. The other classification models, such as Random Forest and Logistic Regression, also performed well, achieving accuracies of 0.7792207792207793 and 0.7662337662337663, respectively. However, they did not perform as well as the SVM model.

We also compared the performance of our models with that of previous studies in the literature. Our results showed that our proposed classification approach outperformed most of the existing models, achieving a higher accuracy than the models proposed in previous studies.

Overall, our findings suggest that machine learning models can be effective in predicting diabetes, and that the SVM algorithm may be the best choice for this task. However, further research is needed to determine the generalizability of these results to other datasets and populations.

In conclusion, the proposed diabetes prediction system based on K-means clustering and proposed classification approach achieved high accuracy in predicting diabetes in patients. The findings of this project have significant implications for healthcare professionals, as accurate and efficient prediction of diabetes can help improve patient outcomes and reduce healthcare costs. Our study provides a promising approach for predicting diabetes, which can be further developed and refined to improve the accuracy and generalizability of the model.

Bibliography

1. [1] Yilmaz N., Inan O., Uzer M.S., “ A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases,” J Med Syst, vol. 38, no. 5 2014.
2. [2] Lowongtrakool C., Hiransakolwong N., “Noise filtering in unsupervised clustering using computation intelligence,” International Journal of Math, vol. 6, no. 59, pp. 2911–2920, 2012.
3. [3] V. Anuja and R.Chitra., “Classification Of Diabetes Disease Using Support Vector Machine”, International Journal of Engineering Research and Applications (IJERA), vol.3,Issue 2, pp. 1797-1801, 2013.
4. [4] Aiswarya I., S. Jeyalatha and Ronak S., “Diagnosis Of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, ,No. 1, pp. 1-14, 2015.
5. [5] K.Rajesh and V.Sangeetha,”Application of Data Mining Methods and Techniques for Diabetes Diagnosis,” in proceedings of International journal of Engineering and Innovative Technology, vol.2, Issue 3, pp. 43-46, 2012.
6. [6] Harleen and Dr. Pankaj B.,”A Prediction Technique in Data Mining for Diabetes Mellitus,” Journal of Management Sciences and Technology, vol. 4, Issue 1, pp. 1-12, 2016.