Title: Diabetes Prediction

**Course Title: Machine learning Course number: COMP 7745- 002**

**Submitted By:**

**Vikas Ravula**

### OBJECTIVES:

The objective of this project is to develop a machine learning model that can predict the likelihood of a patient having diabetes based on various medical and lifestyle factors. The model will be trained on a dataset of historical patient data and will be evaluated based on its ability to accurately predict the presence or absence of diabetes in new patients.

### ABSTRACT:

This project aims to develop a machine learning model for predicting the likelihood of diabetes in patients based on medical and lifestyle factors. The model will be trained on a dataset of 768 observations and 9 variables obtained from Kaggle. The dataset includes only female patients of Pima Indian ancestry who are at least 21 years old. The model will be evaluated based on its accuracy in predicting the presence or absence of diabetes in new patients. Five machine learning algorithms, including Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, and KNeighborsClassifier will be implemented for this purpose. Early detection and prevention of diabetes is critical for better patient outcomes, and machine learning can assist healthcare professionals in achieving this goal.

A confusion matrix is a useful tool for evaluating the performance of machine learning models in classification tasks. It presents a table of true positive, false positive, true negative, and false negative predictions made by the model. The true positive and true negative values represent the correct predictions made by the model, while the false positive and false negative values represent the incorrect predictions. From the confusion matrix, we can calculate various performance metrics such as accuracy, Mean squared error, roc\_curve accuracy to assess the effectiveness of the model. This paper will present the confusion matrices and performance metrics for several machine learning models used to predict the likelihood of diabetes in patients based on medical and lifestyle factors.

### INTRODUCTION:

* **What is diabetes?**

Diabetes is a chronic health condition characterized by the body's inability to produce or use insulin effectively, a hormone that regulates blood sugar levels. The three main types of diabetes are type 1, which involves a lack of insulin production; type 2 (the most prevalent type), which entails inadequate insulin production or utilization; and gestational diabetes, which occurs during pregnancy. Other less common types of diabetes also exist.

### How many Canadians live with diabetes? (Prevalence)

According to the Public Health Agency of Canada's latest data, approximately 3.0 million Canadians, equivalent to 8.1% of the population, were diagnosed with diabetes in 2013-2014. This prevalence translates to 1 in 300 children and youth (1-19 years old) and 1 in 10 adults (20

years and older). The incidence of diabetes typically rises with age, and males (8.7%) are more affected than females (7.6%) in most age groups. Type 1 diabetes accounts for 10% of diabetes cases in Canada, while type 2 diabetes constitutes 90% of all cases.

### Dataset - Attributes Information

The dataset utilized in this study is the Pima Indians Diabetes Database, which is maintained by the National Institute of Diabetes and Digestive and Kidney Diseases research center. The dataset comprises eight independent variables, also known as attributes, and one dependent variable or outcome, which is a binary classification of positive or negative indicated by 1 and 0. The attributes are defined as follows:

* Number of times pregnant
* Plasma glucose concentration 2 hours in an oral glucose tolerance test
* Blood pressure (mm Hg)
* Triceps skinfold thickness (mm)
* 2-Hour serum insulin (mu U/ml)
* Body mass index (weight in kg/(height in m)^2)
* Diabetes pedigree function
* Age (years)
* Outcome - Class variable (0 or 1)

### BACKGROUND:

* + The dataset was obtained from Kaggle :

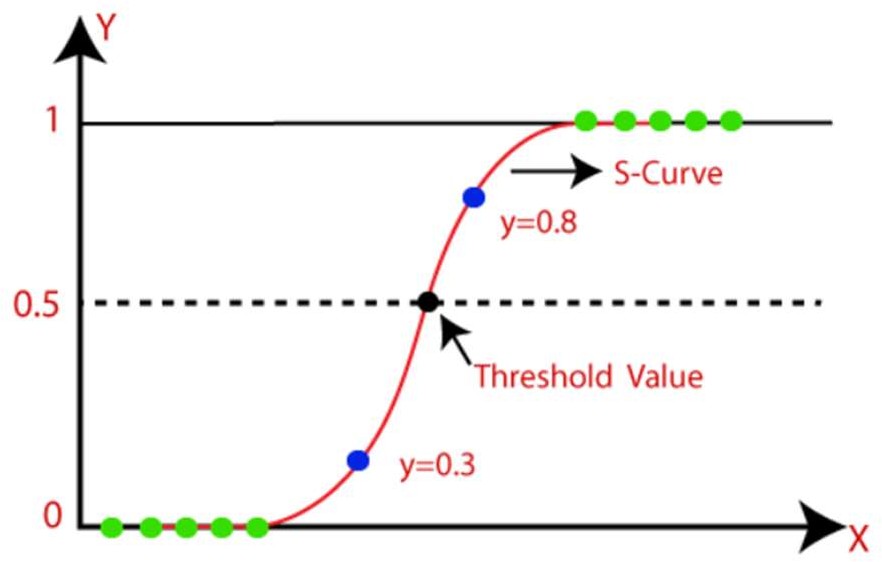
https://[www.kaggle.com/datasets/mathchi/diabetes-data-set](http://www.kaggle.com/datasets/mathchi/diabetes-data-set)

* + Data contains 768 observations and 9 variables.
  + The selection of instances from a larger database was subject to various restrictions, including the requirement that all patients be women of Pima Indian ancestry who are at least 21 years old.

### METHODOLOGY:

**Logistic regression:**

Logistic regression is a classification technique that categorizes observations into multiple classes. Unlike linear regression, which generates continuous numerical values, logistic regression produces a probability value that can be mapped to discrete classes using the logistic sigmoid function.

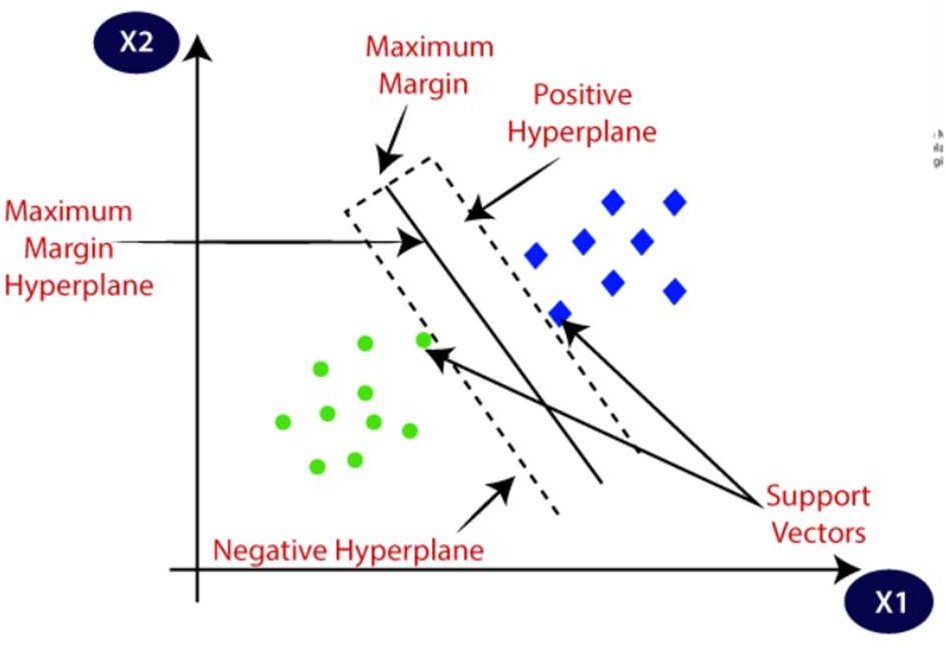


**K-nearest neighbors (KNN)**

K-nearest neighbors (KNN) is a simple supervised machine learning algorithm that can be used for both classification and regression problems. It is easy to set up and use, but its performance slows down considerably as the amount of data grows. KNN operates by computing the distances between a query and all instances in the dataset, selecting the number of instances (K) closest to the query, and then voting for the most common label (in classification) or averaging the labels (in regression).

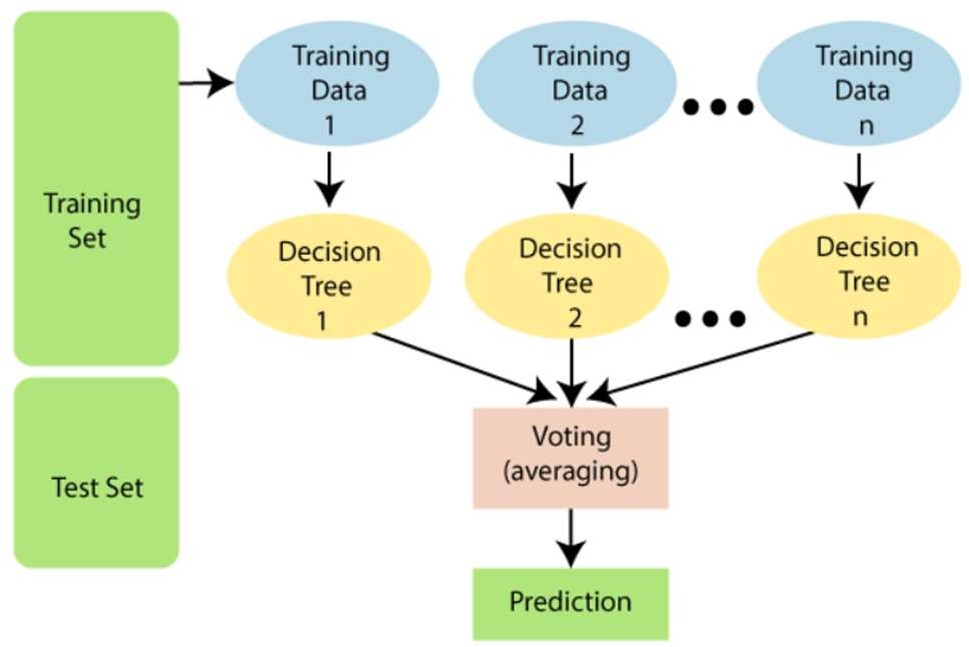
**Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression tasks, but it is primarily used for classification problems. In the SVM approach, each data point is represented as a point in n-dimensional space, where n corresponds to the number of features available, with each feature value being the value of a specific coordinate. The goal is to identify the hyperplane that best separates the two classes for classification.



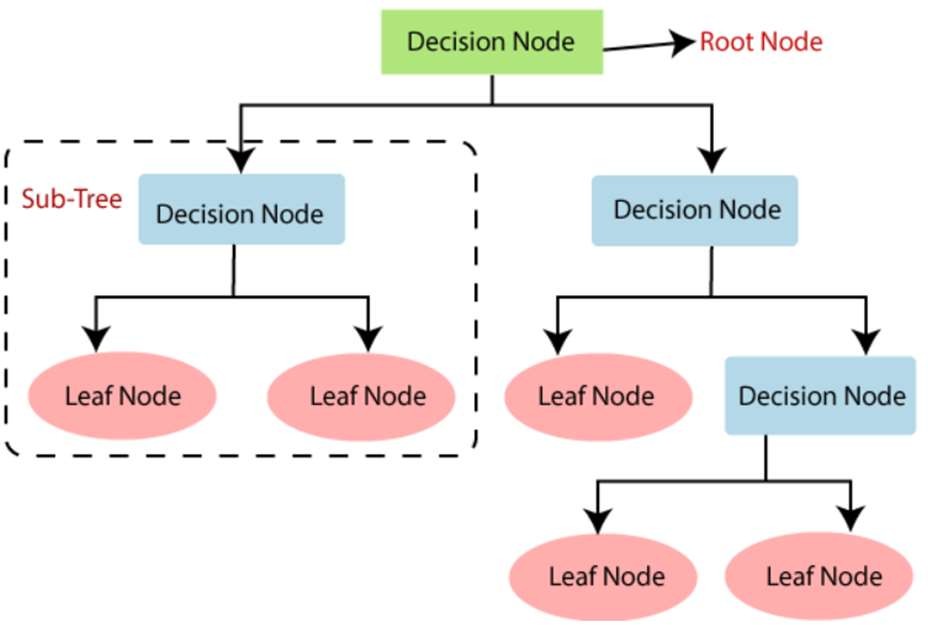
**Random Forest Classifier**

Random Forest Classifier is an ensemble learning technique used for classification, regression, and other problems. During training, it generates numerous decision trees. For classification problems, the output of random forests is the class chosen by the most trees. For regression tasks, it returns the average or mean prediction of individual trees. Random Forests address the tendency of decision trees to overfit their training set. In general, Random Forests outperform decision trees, but their accuracy is lower than that of gradient-boosted trees. However, their performance can be affected by the data features.



**Decision Tree Classifier**

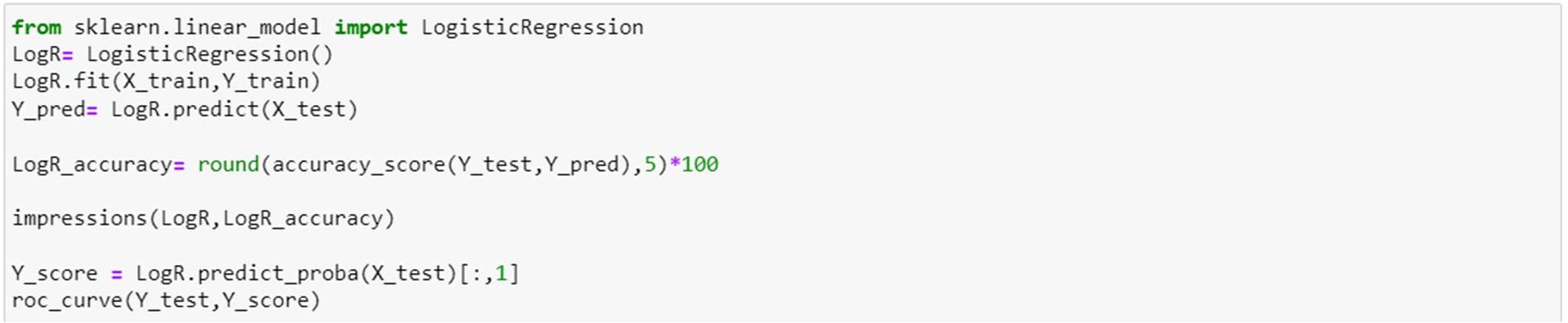
Decision Tree Classifier is a supervised learning technique that finds applications in machine learning, data mining, and statistics. It uses a classification or regression decision tree as a predictive model to draw conclusions about a set of observations.



# Model Description:

## Logistic Regression:

Logistic regression helps in assigning the observations to one of several classes. Below is the implementation for the logistic regression.



## K-Nearest Neighbour:

The k-nearest neighbours (KNN) technique is a straightforward supervised machine learning approach that may be applied to classification and regression issues.



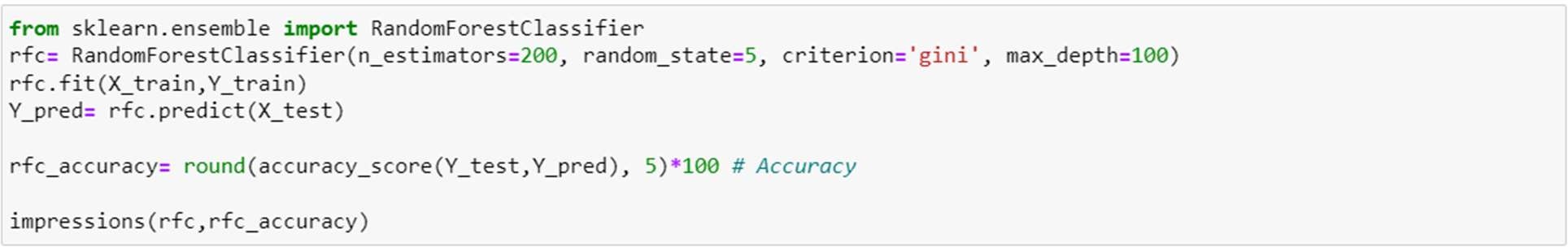
## Support Vector Machine:

SVM is a supervised machine learning technique that can be used for classification or regression tasks.



## Random Forest Classifier:

Random forest is an ensemble learning method for classification, regression, and other problems that works by generating a large number of decision trees during training.



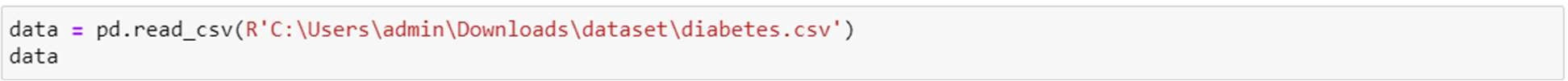
## Decision Tree Classifier:

Decision tree learning is a supervised learning method that has applications in statistics, data mining, and machine learning.

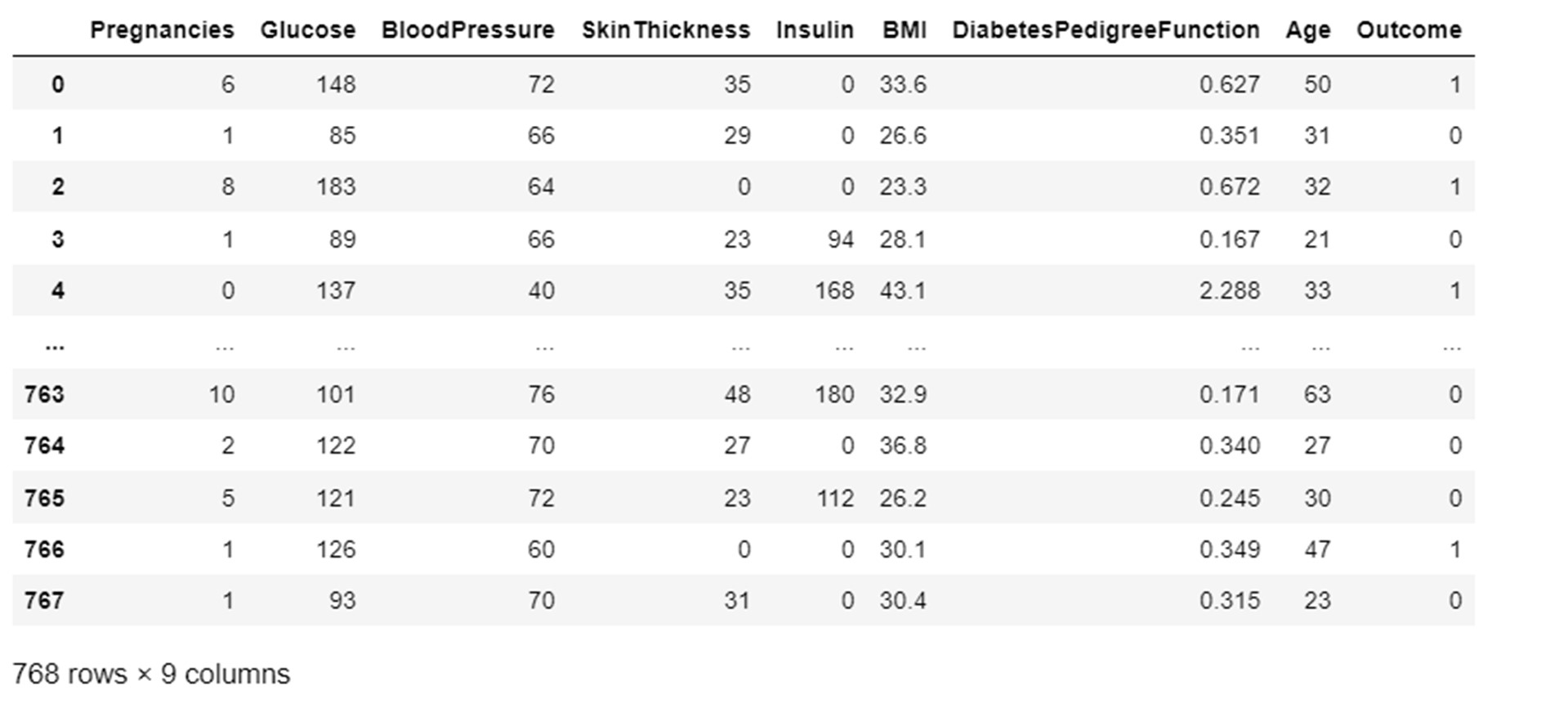


# Experimental results and Analysis:

## Loading the dataset:



**Data:**

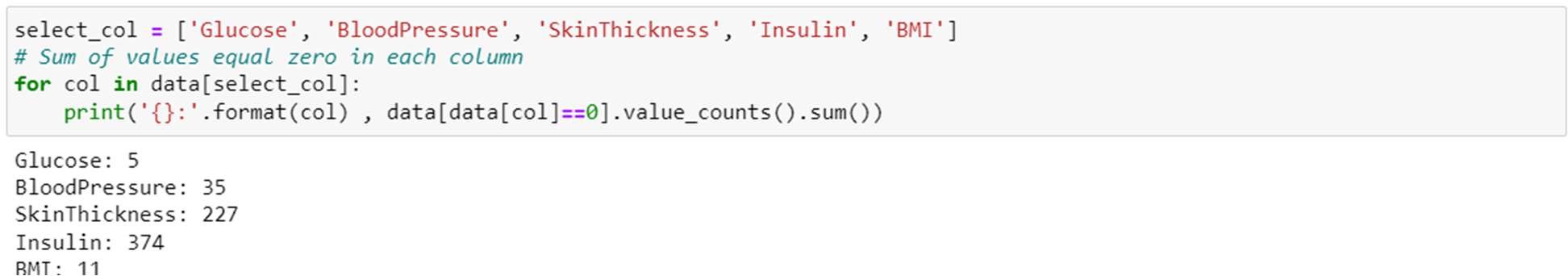


# Training and Testing Logs:

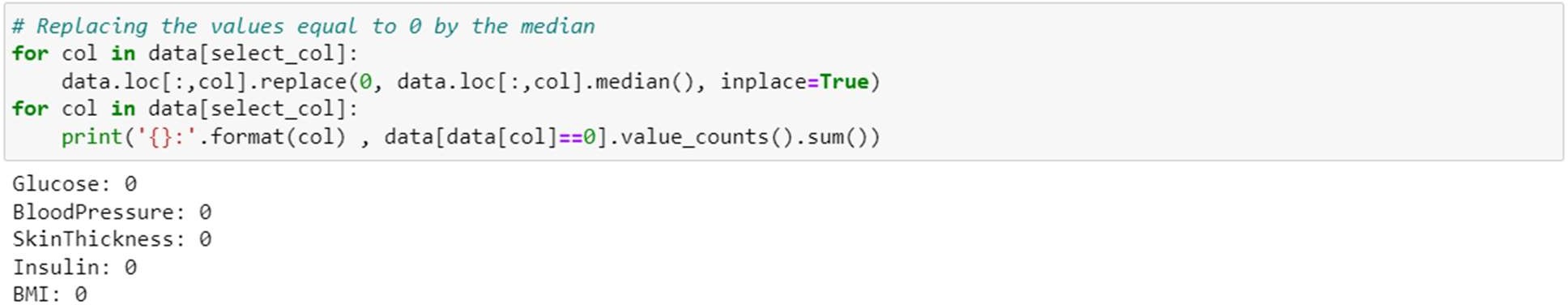
**Data Processing:**

### Checking for null values:

* **Sum of values equal zero in each column:**

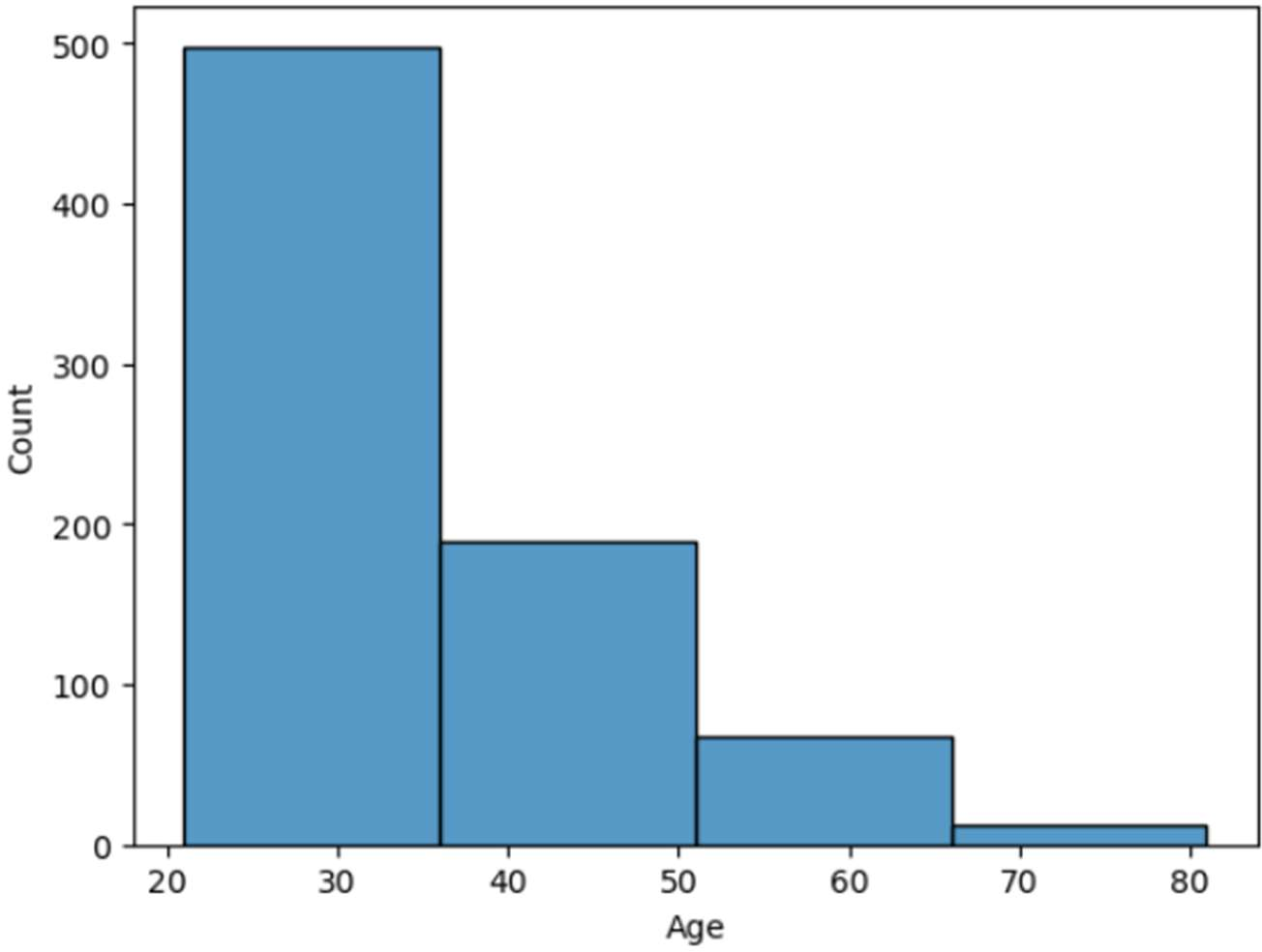


* **Replacing the values equal to 0 by the median**

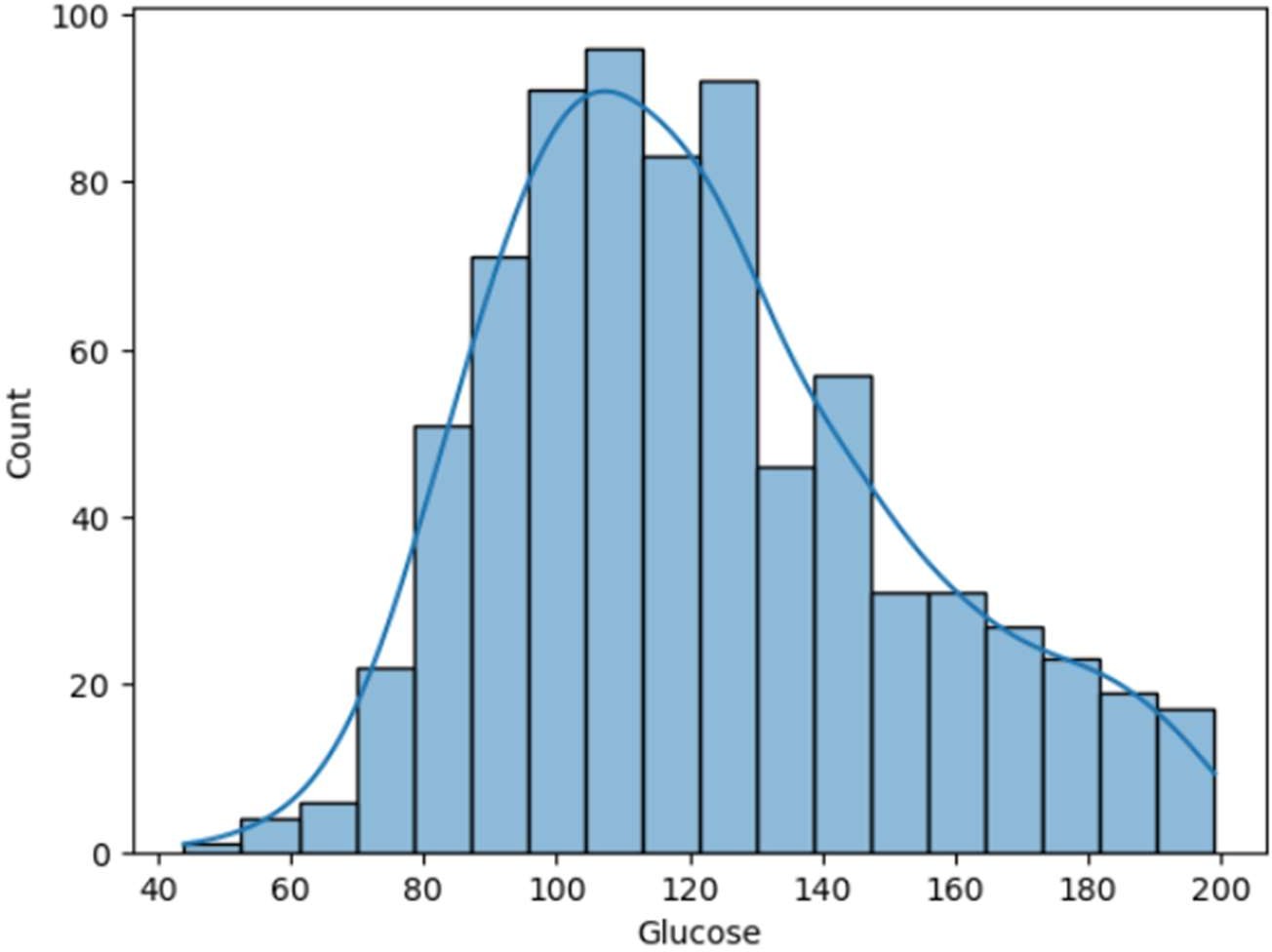


# DATA VIZUALIZATION:

### Age:

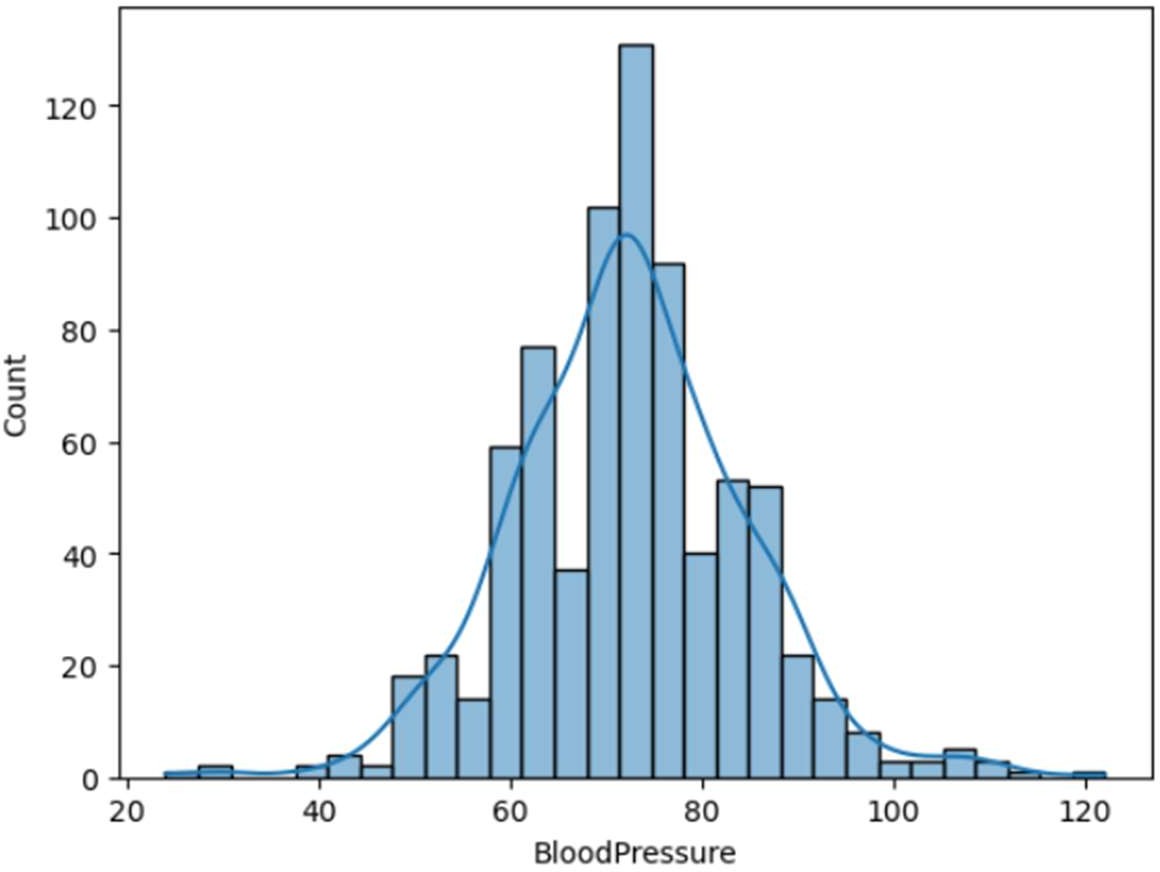


**glucose**: The oral glucose tolerance test involves an overnight fast, followed by measuring the fasting blood sugar level. Next, you consume a sugary drink, and your blood sugar levels are periodically measured over the next two hours.

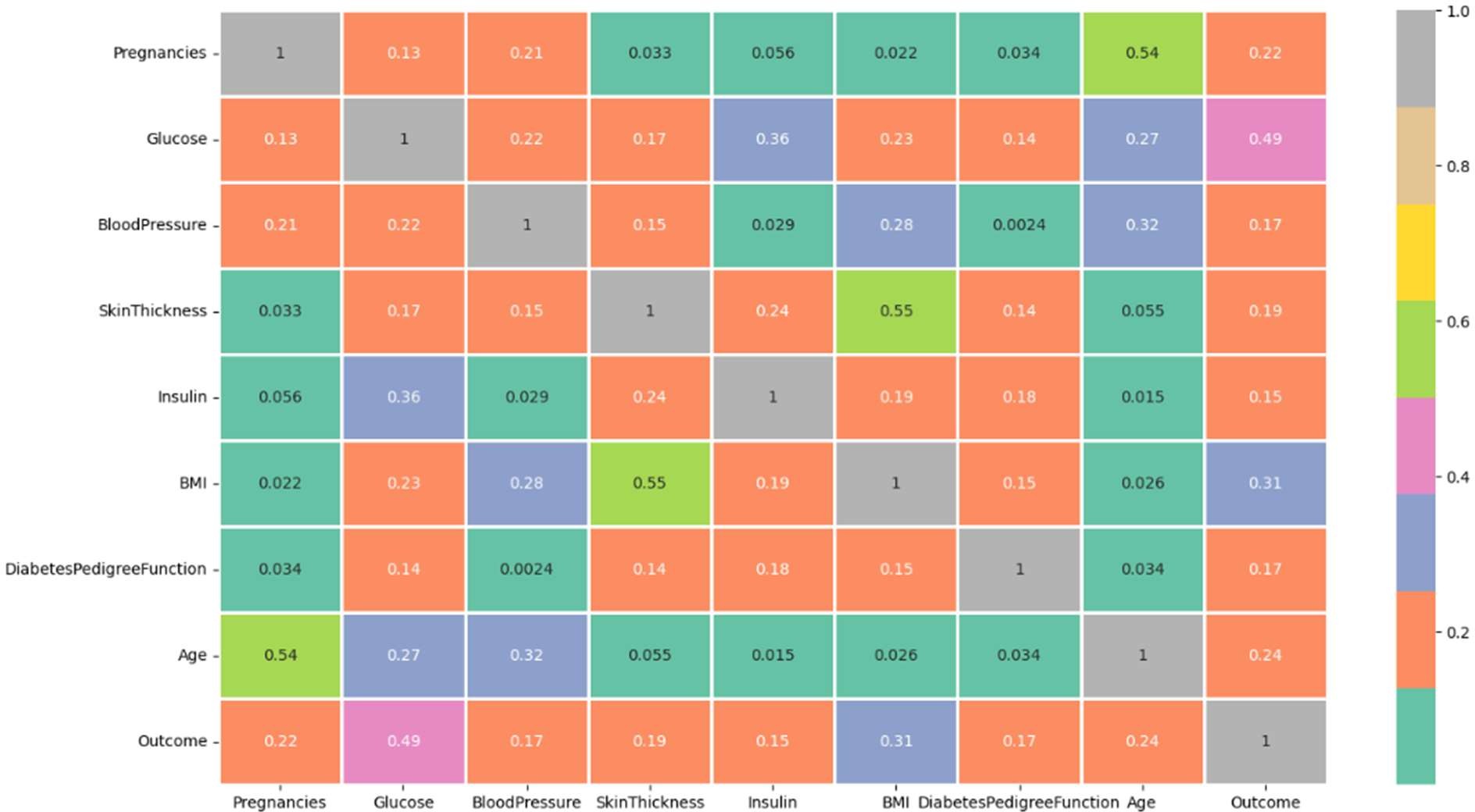


**Blood pressure:** The diastolic reading refers to the pressure in the arteries during the heart's resting period between beats. During this time, the heart is filling with blood and receiving oxygen.

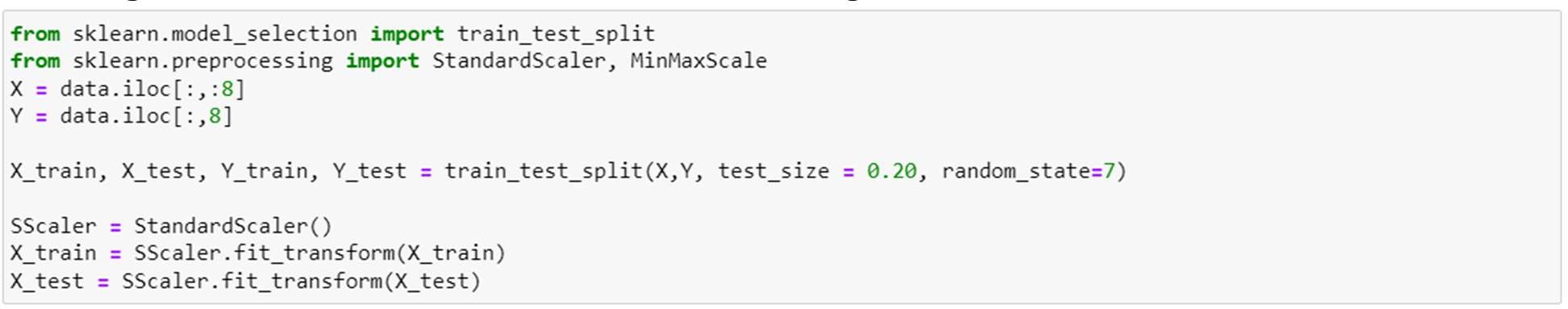
* Normal: Lower than 80 mmHg
* Stage 1 hypertension: 80-89 mmHg
* Stage 2 hypertension: 90 mmHg or more



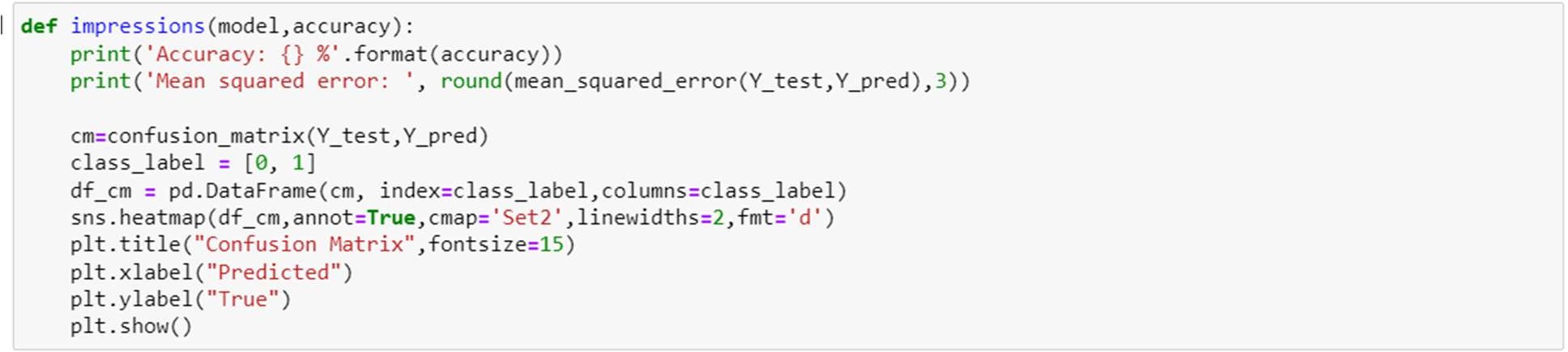
### Correlation matrix:



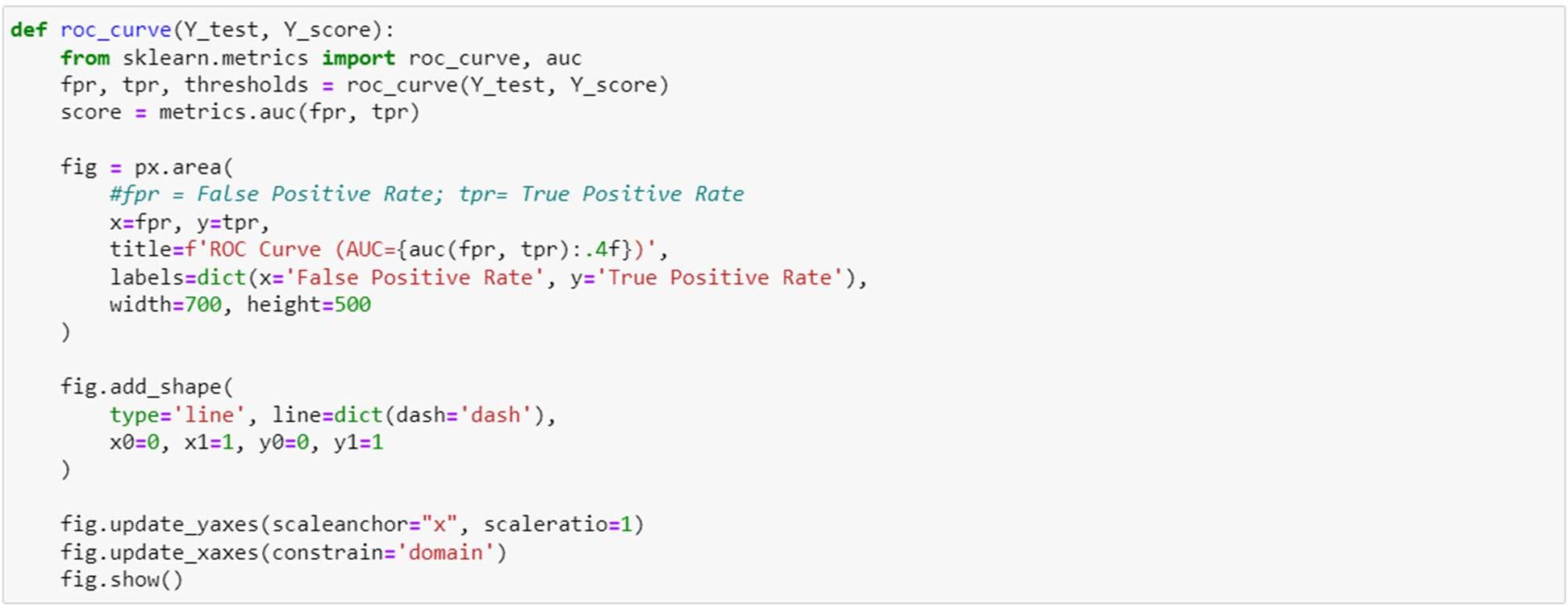
**Dividing Data into Train and Test Parts and Doing the StandardScaler, MinMaxScale:**



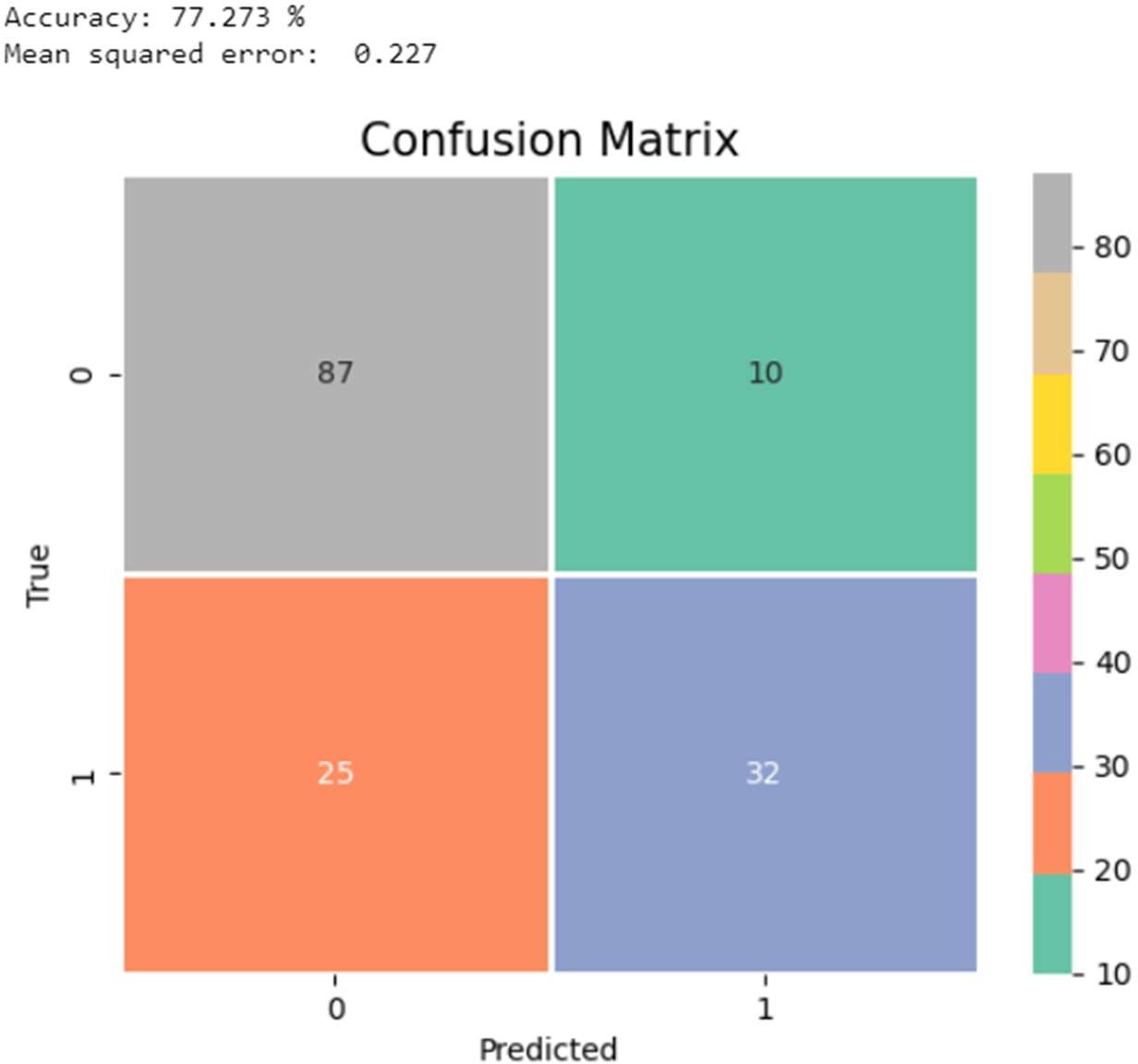
### Model building:

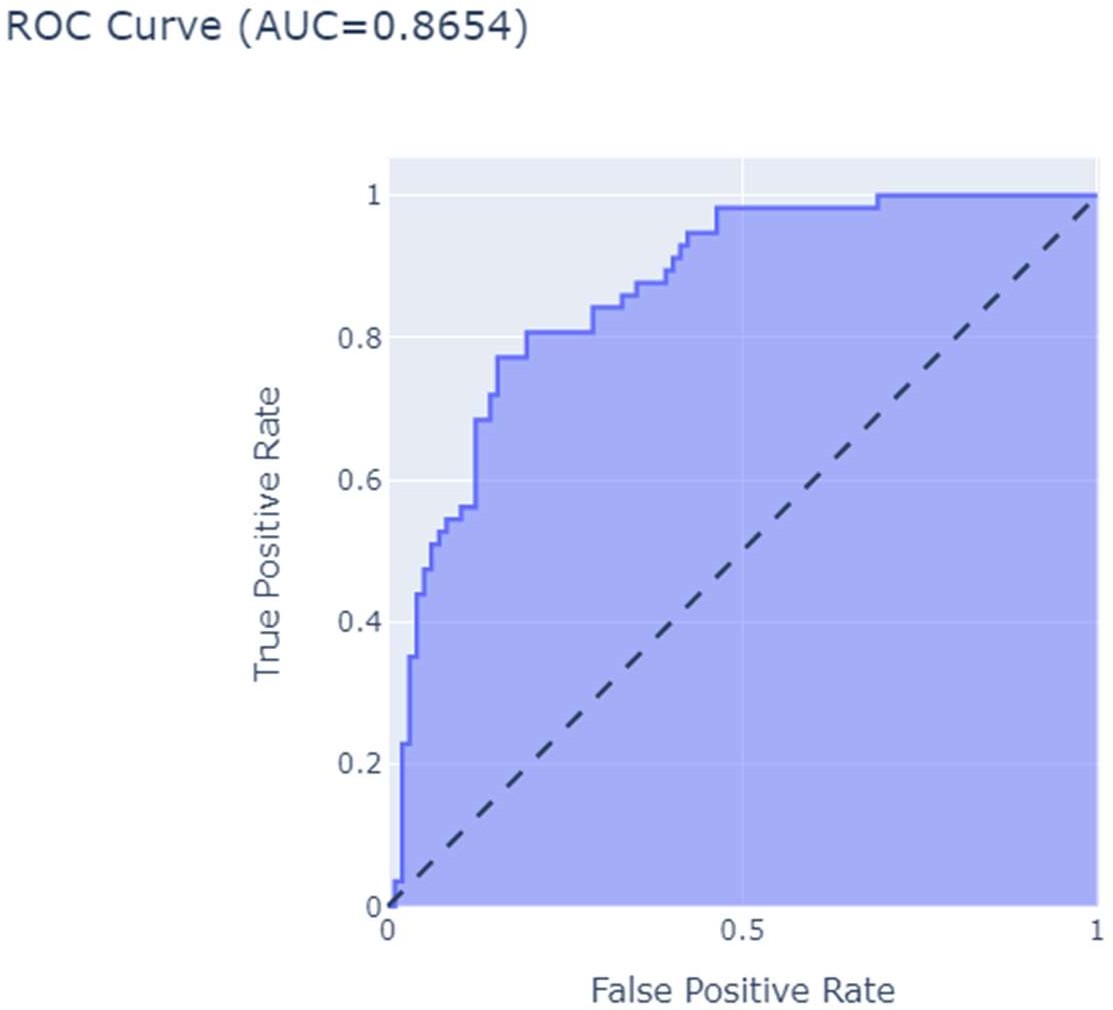


**Building Roc curve:**

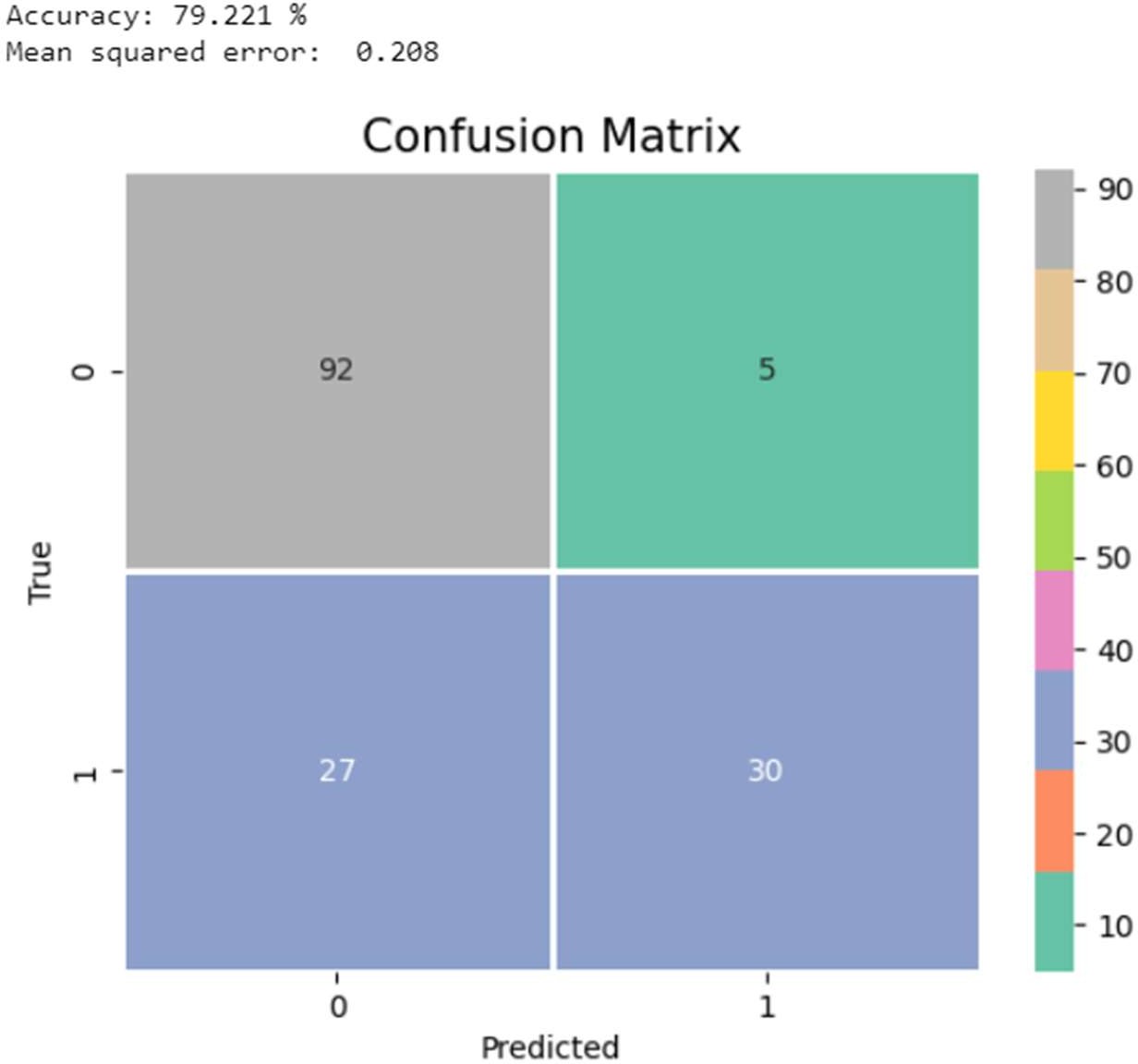


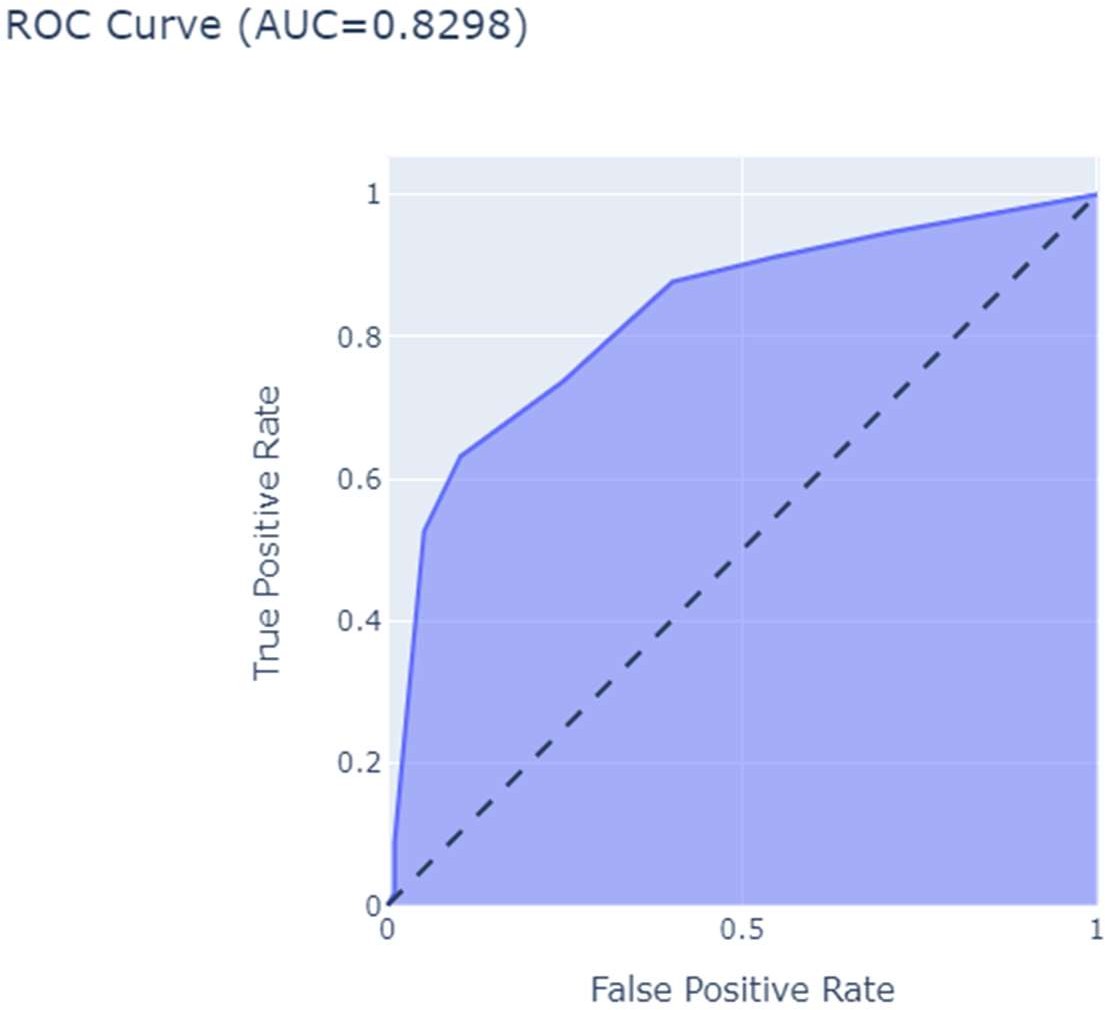
**Logistic Regression:**



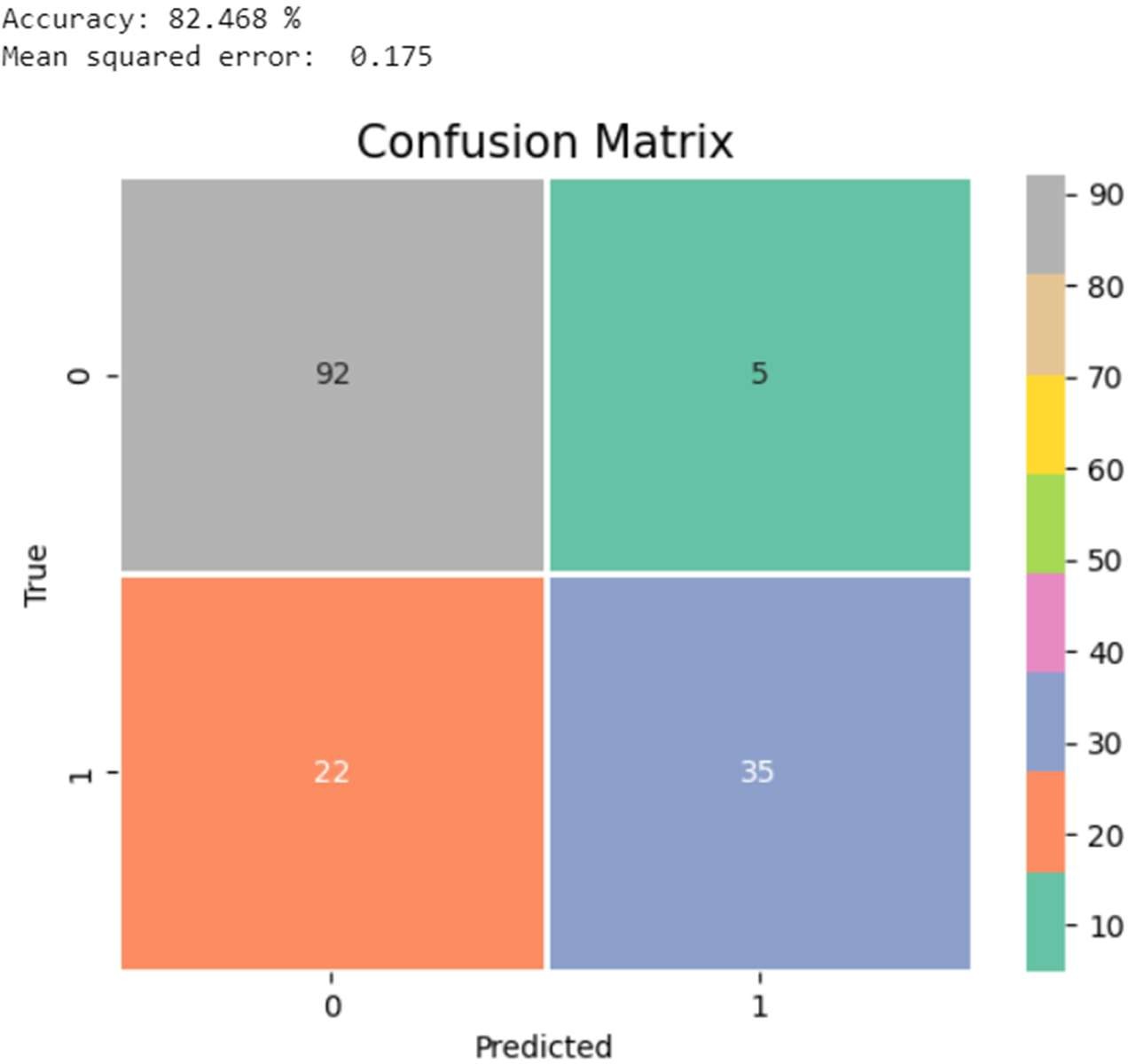


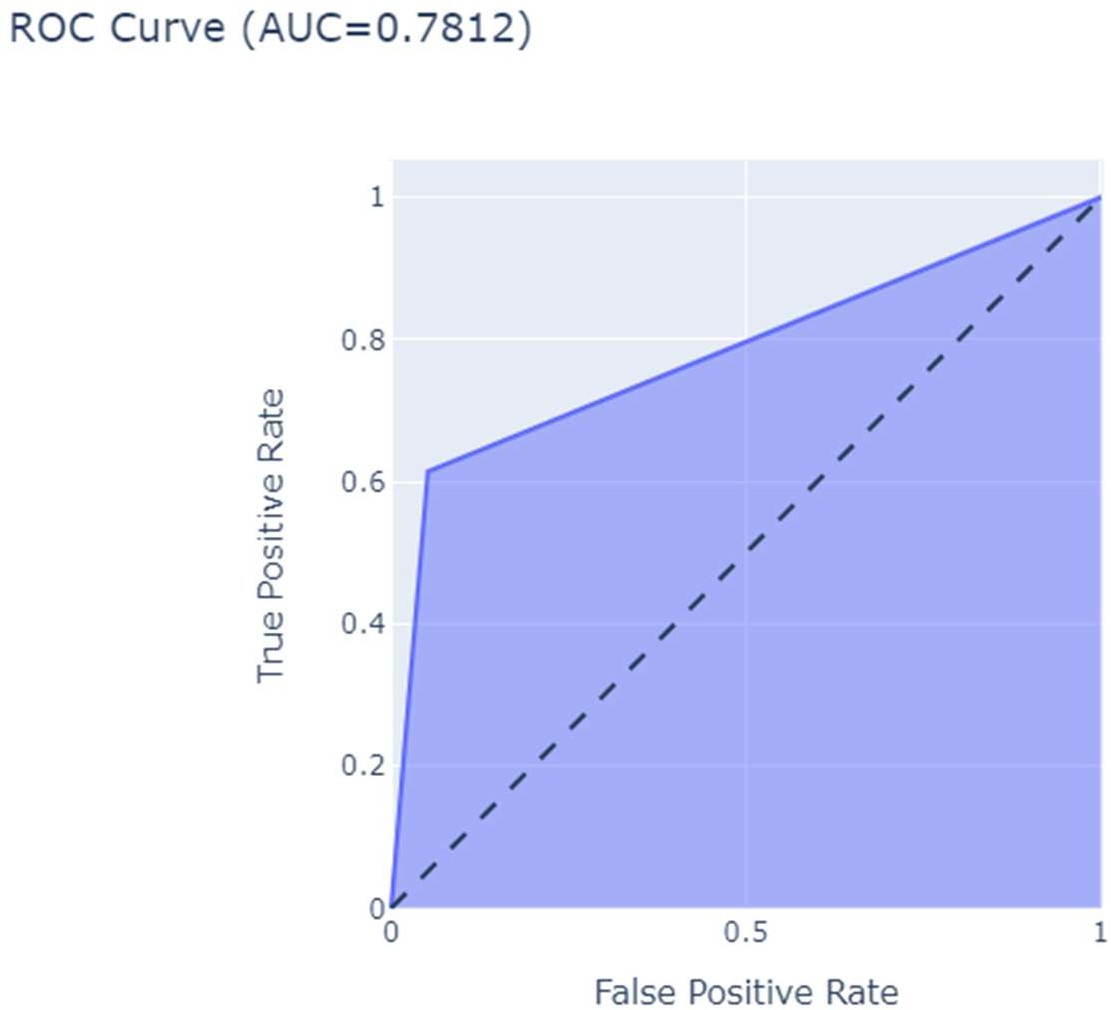
**K-Nearest Neighbour:**



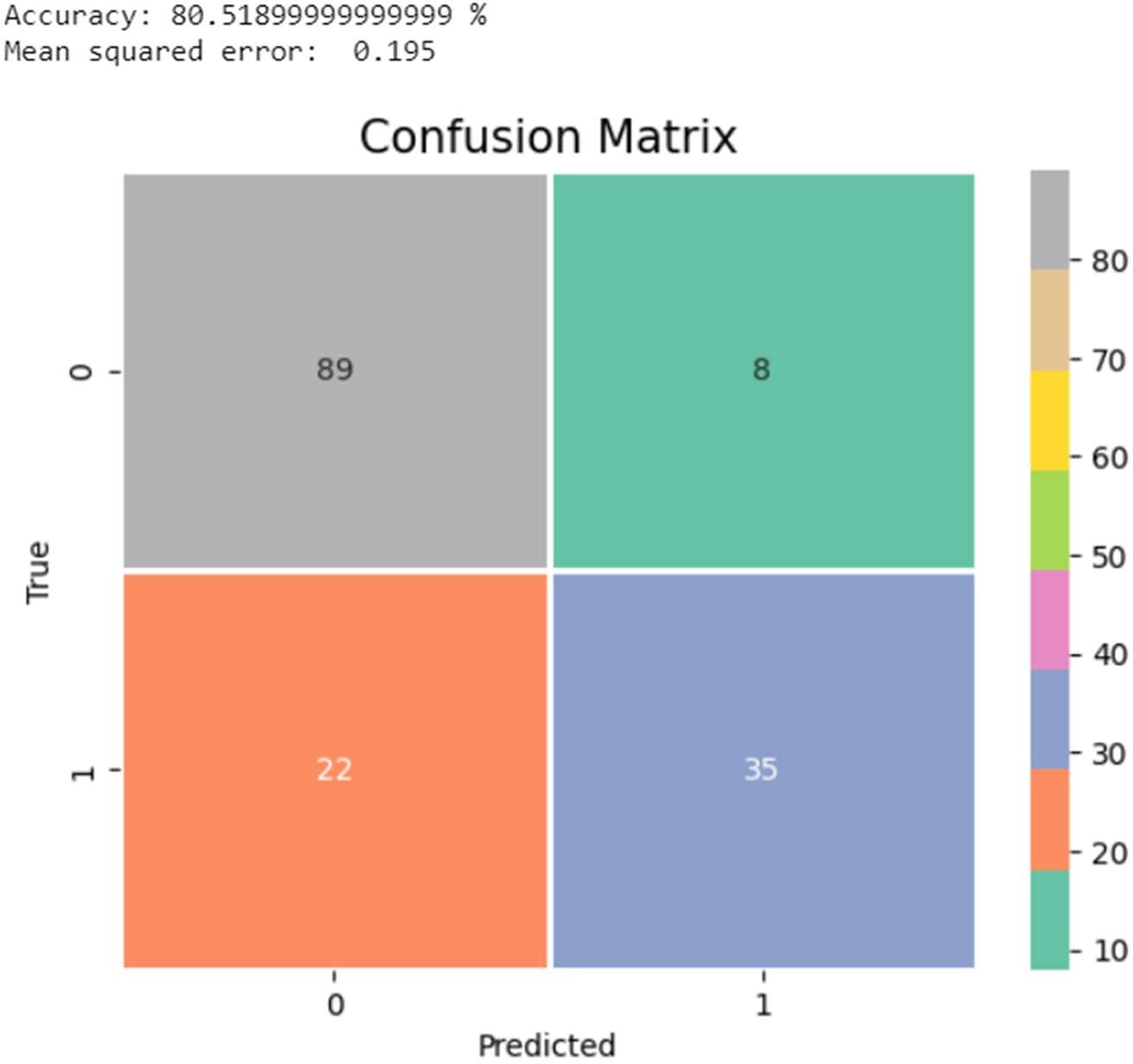


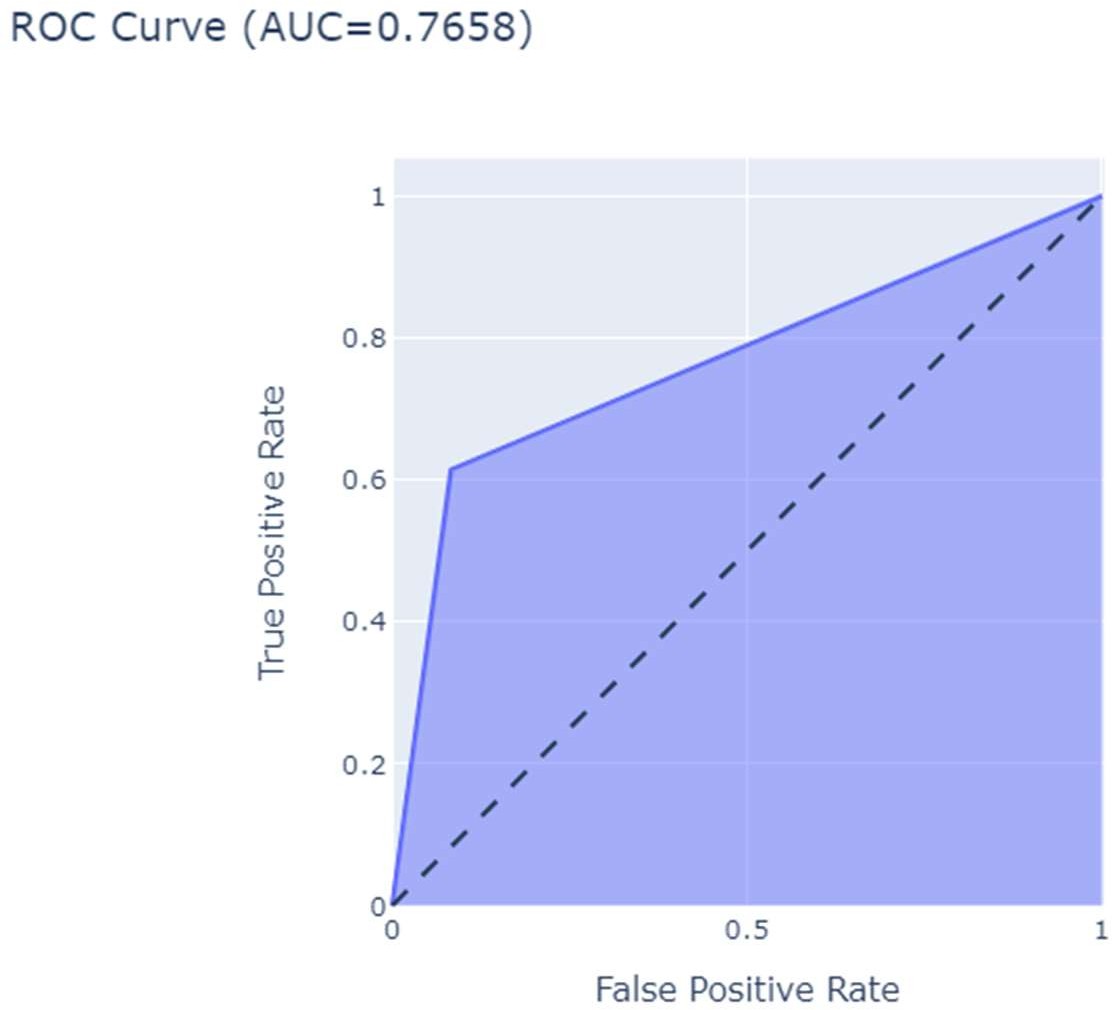
## Support Vector Machine:



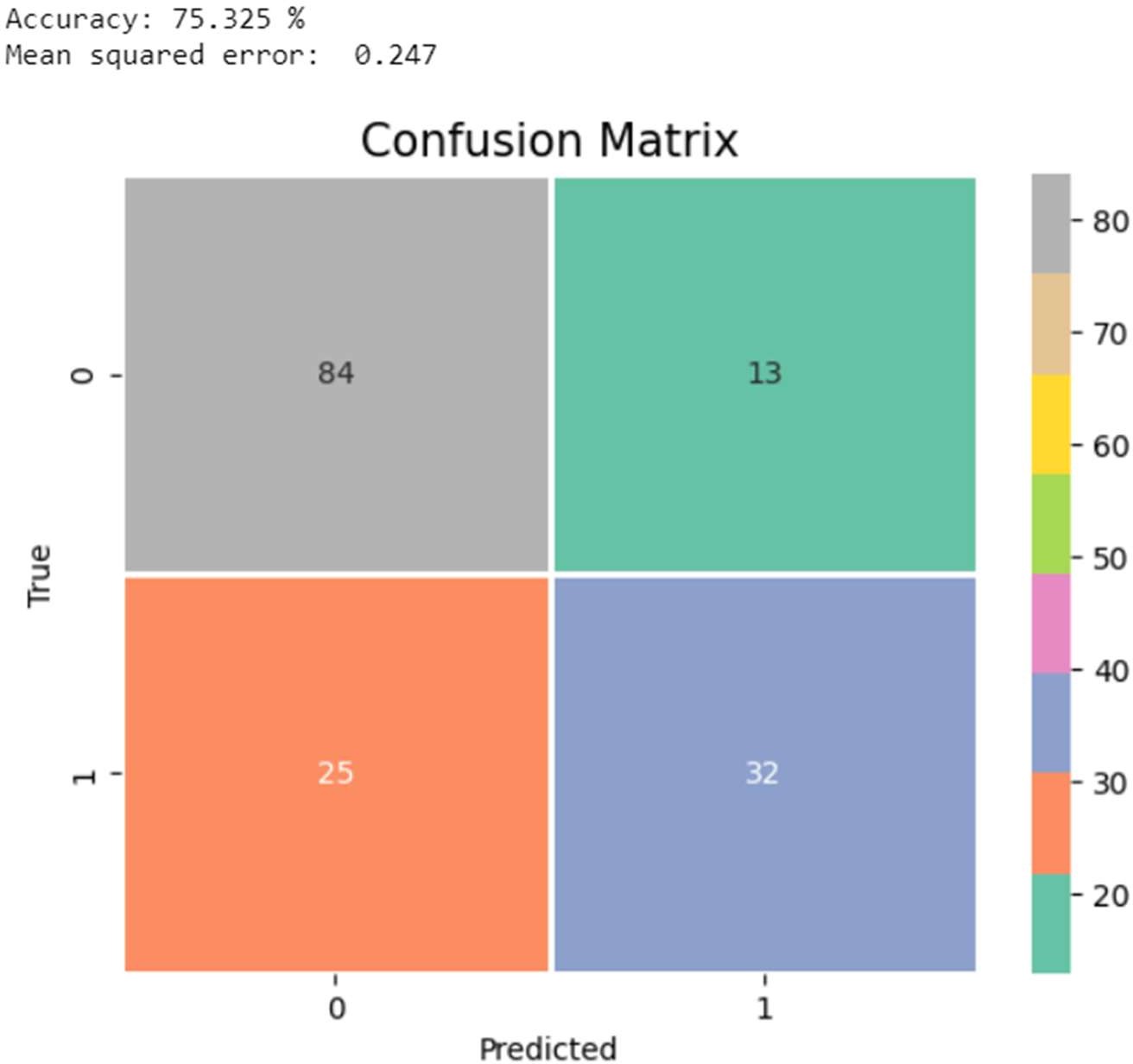


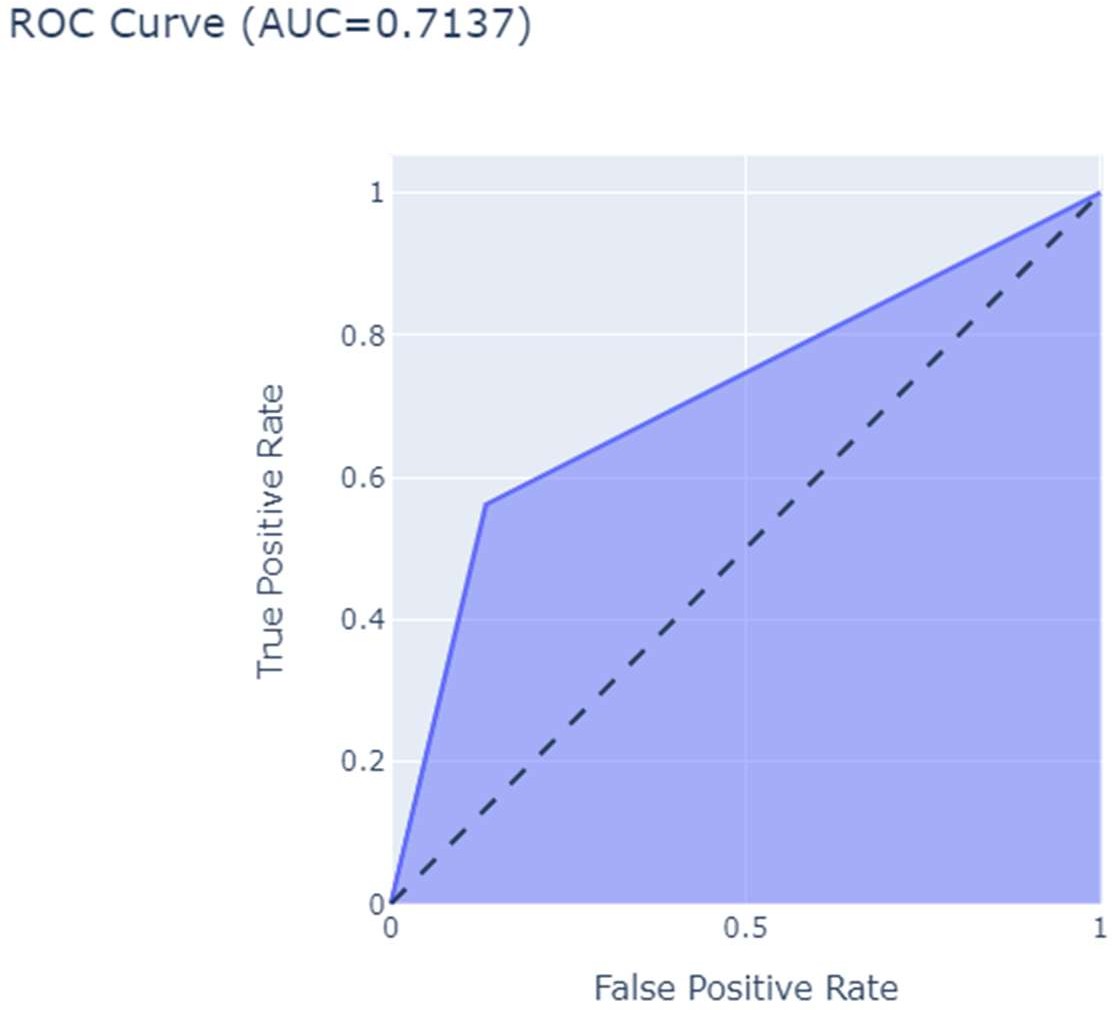
**Random Forest Classifier:**





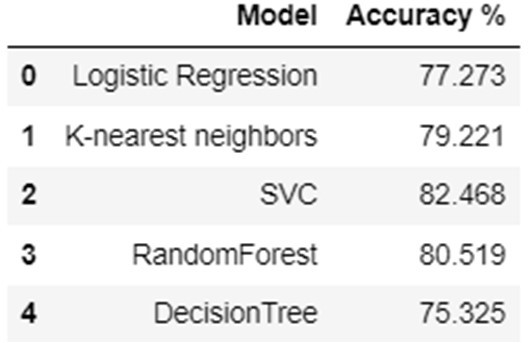
**Decision Tree Classifier:**





# Discussion and Model Comparison:

The below is the model comparison between the five machine learning models that are used in our project.



Based on the training accuracy percentages presented, it appears that the Support Vector Classifier (SVC) achieved the highest accuracy of 82.468%, followed by K-nearest neighbors with 79.221%, and Random Forest with 80.519%. Logistic Regression and Decision Tree Classifier have relatively lower accuracies at 77.273% and 75.325%, respectively

# Conclusion:

Based on the performance metrics and model comparison, the SVM model had the highest accuracy of 82.468% on the test dataset, followed by the K-nearest neighbors model with 79.221%. The other models, including Logistic Regression, Random Forest, and Decision Tree, also showed reasonable performance but were slightly less accurate than SVM and KNN.

According to the Diabetes dataset study results, Glucose is the stronger contributor for predicting diabetes.

The techniques employed in this study can be utilized to train a model that predicts the likelihood of diabetes in an individual based on input features such as Glucose, BMI, Age, and Number of Pregnancies

# Future Goals:

To conduct a more comprehensive study in the future for predicting the type of diabetes (Type 1 or Type 2), the company could consider providing additional medical information related to both types of diabetes, such as insulin levels, C-peptide levels, and autoimmune markers, in the dataset. This would enable the development of more precise and accurate predictive models for the different types of diabetes.

# References:

1. Public Health Agency of Canada. [Online] 2017.

https://[www.canada.ca/en/publichealth/services/chronic-diseases/diabetes.html.](http://www.canada.ca/en/publichealth/services/chronic-diseases/diabetes.html)

1. Introduction to Decision Trees . [Online] https://medium.com/greyatom/decision-trees-asimple-way- to-visualize-a-decision-dc506a403aeb.
2. Linear Models in R: Diagnosing Our Regression Model. [Online] https://[www.theanalysisfactor.com/linear-models-r-diagnosing-regression-model/.](http://www.theanalysisfactor.com/linear-models-r-diagnosing-regression-model/)
3. Residual Plots Part 3— Scale-Location Plot. [Online] https://medium.com/datadistilled/residual-plots- part-3-scale-location-plot-113e469b99c.