

Diabetes prediction

# AI/ML Assignment-1

**Lakshay Dahiya, 500121078, BTech CSE**

**Diabetes Prediction Using Machine Learning**

## This project aims to predict whether an individual has diabetes based on a set of health features using various machine learning models. The goal is to help healthcare professionals identify potential diabetes patients early, which can significantly improve diagnosis and treatment outcomes.

📊 **What’s Inside the Project?**

1. **Data Preprocessing:**

## The dataset has been cleaned and preprocessed by handling missing values, scaling features, and addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

1. **Model Building:**

## Random Forest: A robust model that works well for classification tasks.

## Logistic Regression: A simple but effective linear model for binary classification.

## Support Vector Machine (SVM): A powerful model that works well in high-dimensional spaces.

1. **Model Evaluation:**

## Accuracy, Precision, Recall, F1-score, and ROC-AUC are used to evaluate each model.

## ROC curves are plotted to visually compare the models’ performances.

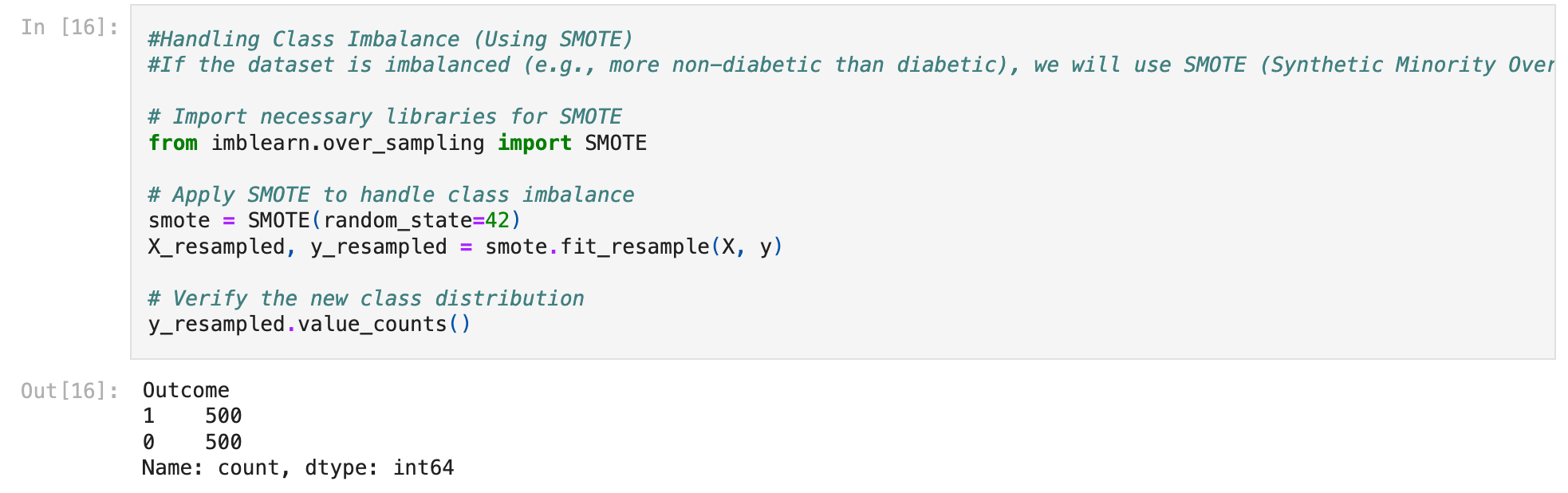
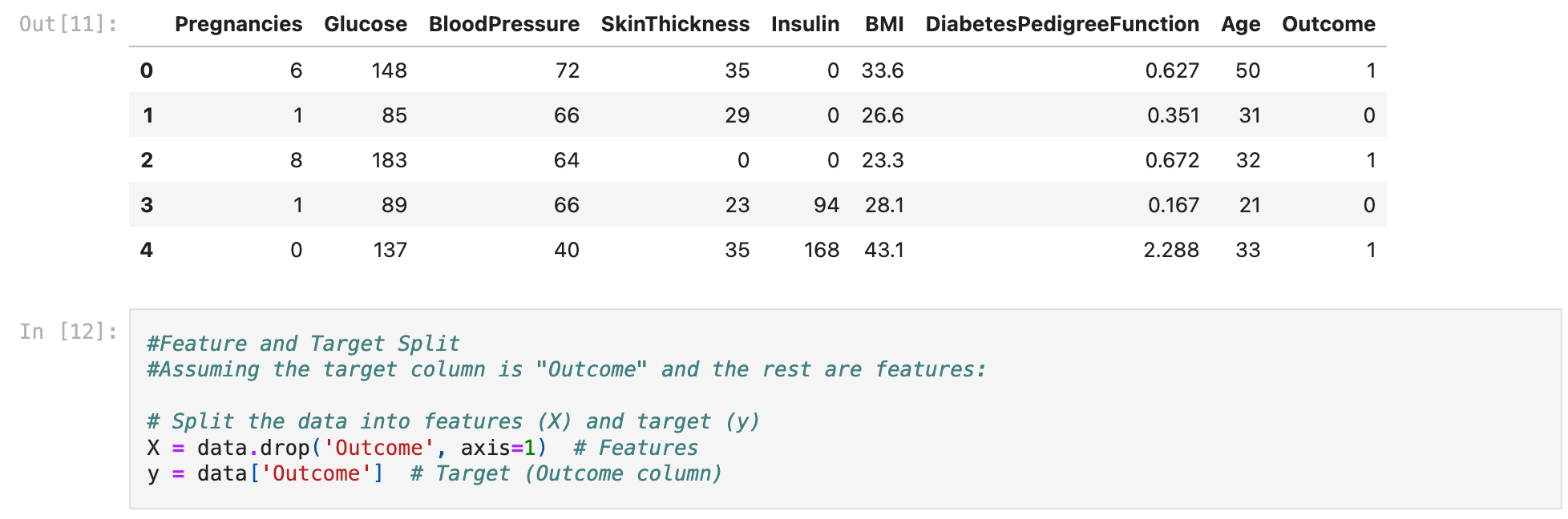
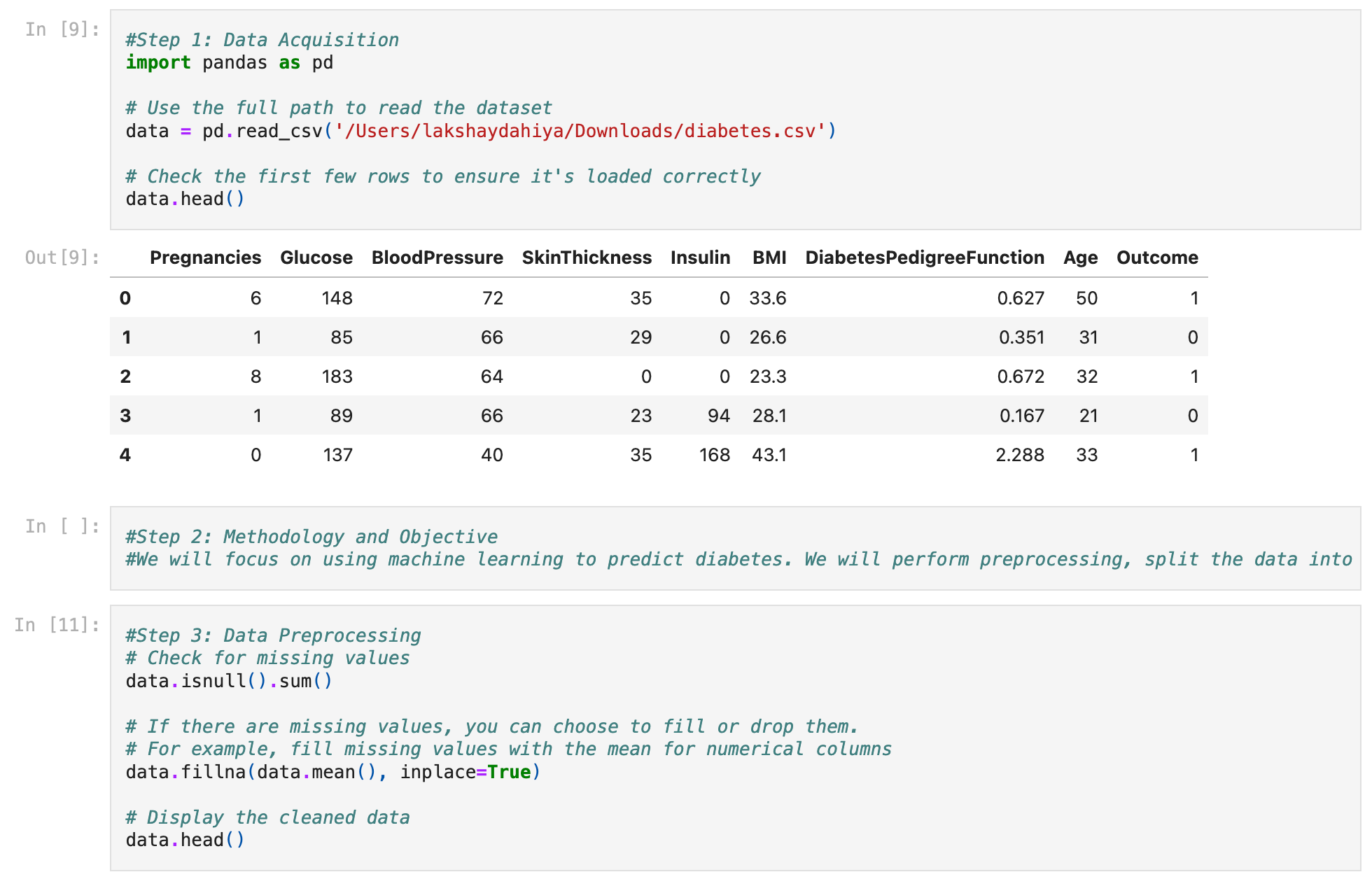
1. **K-Fold Cross-Validation:**

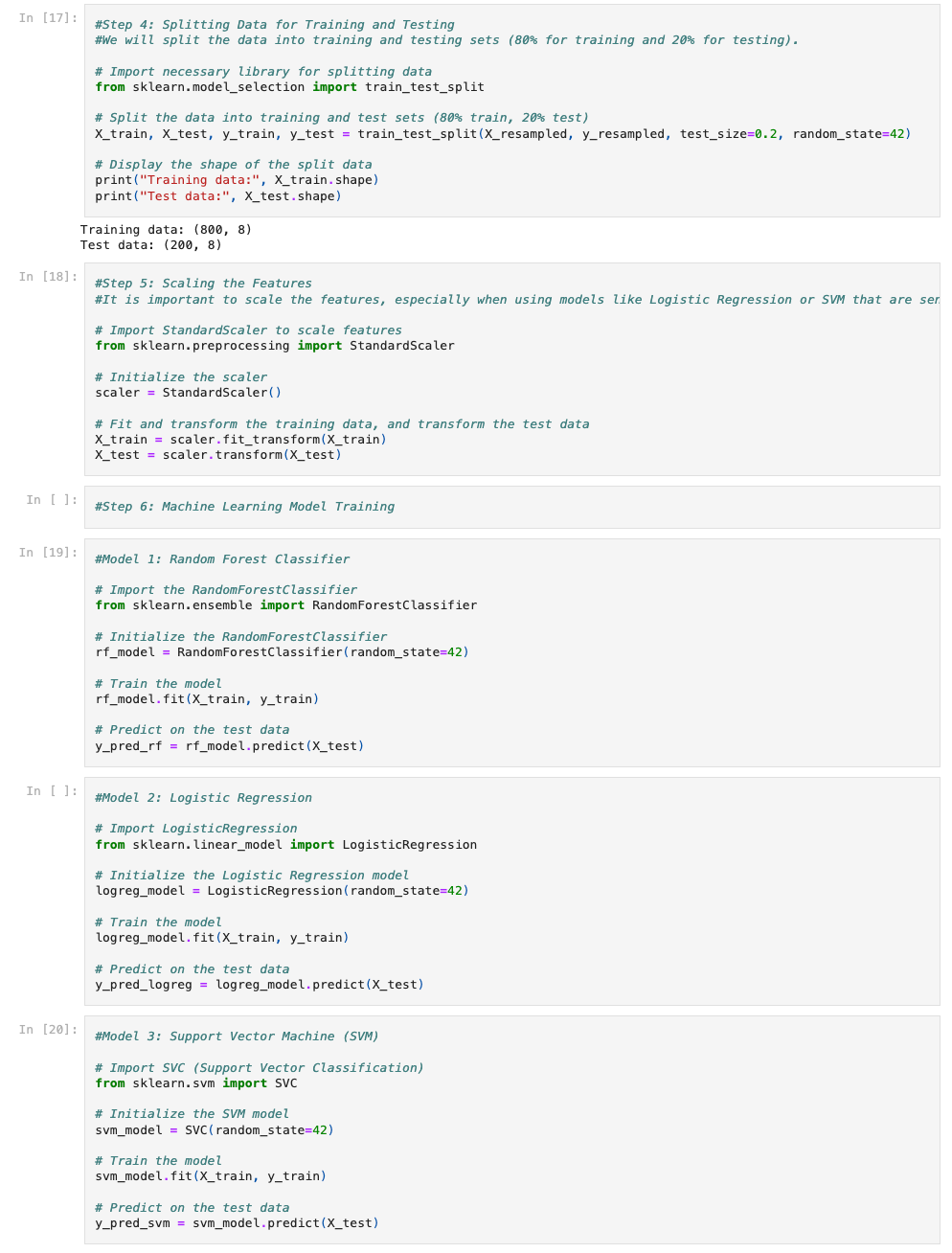
## Cross-validation (CV) is performed to assess how well the models generalize to unseen data. This step is essential for avoiding overfitting.

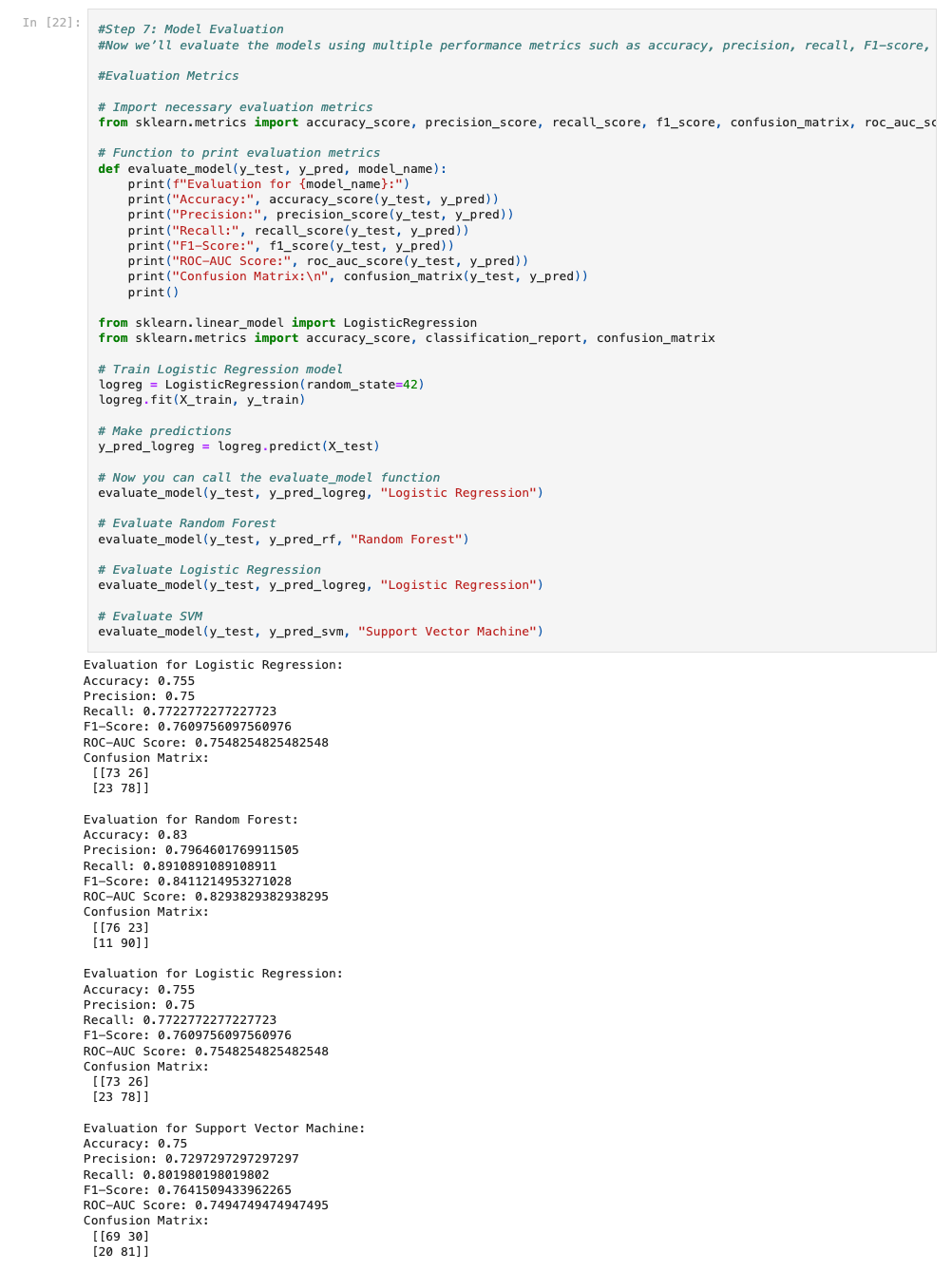
1. **Visualization:**

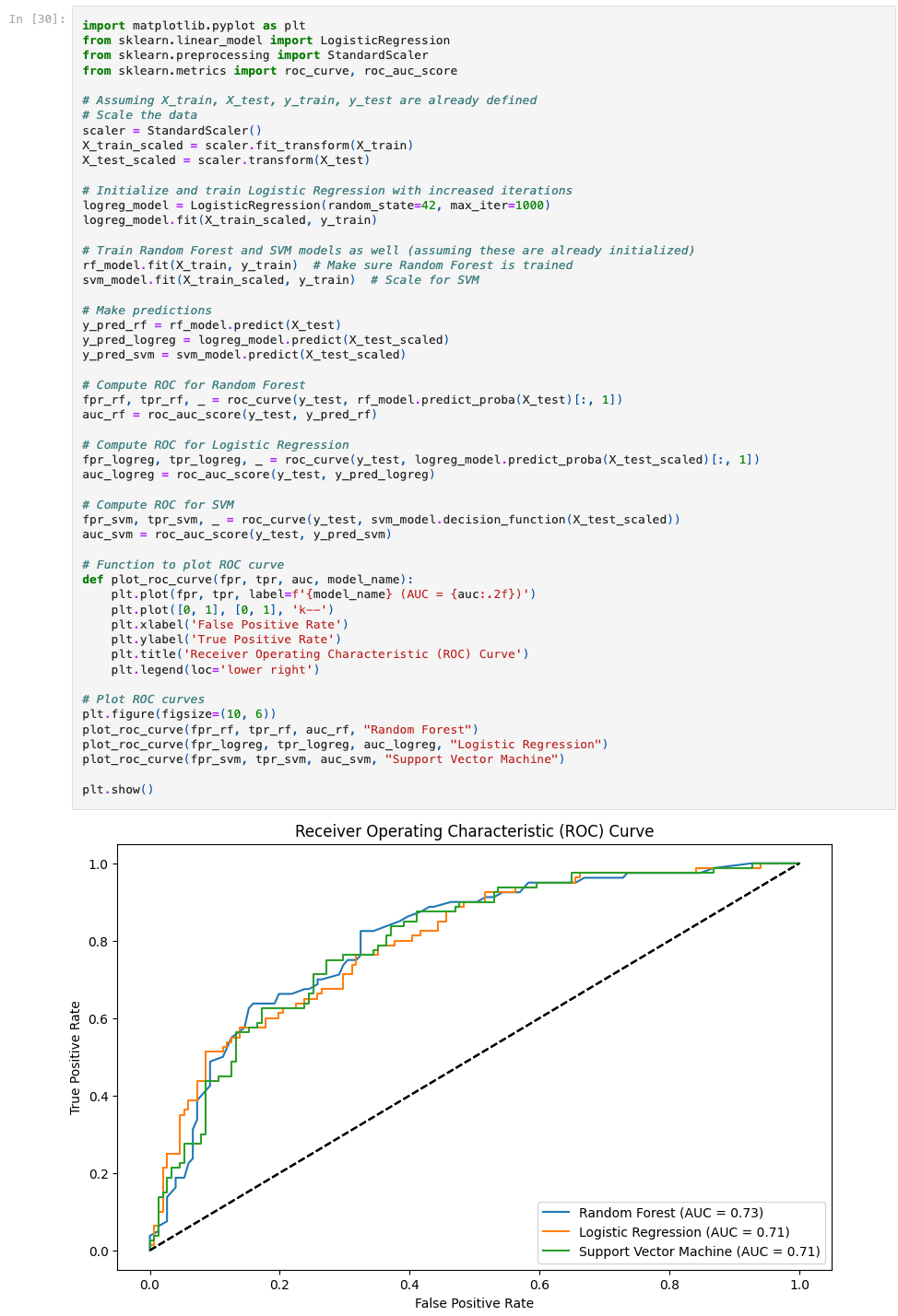
## ROC Curves: Visual comparison of model performance.

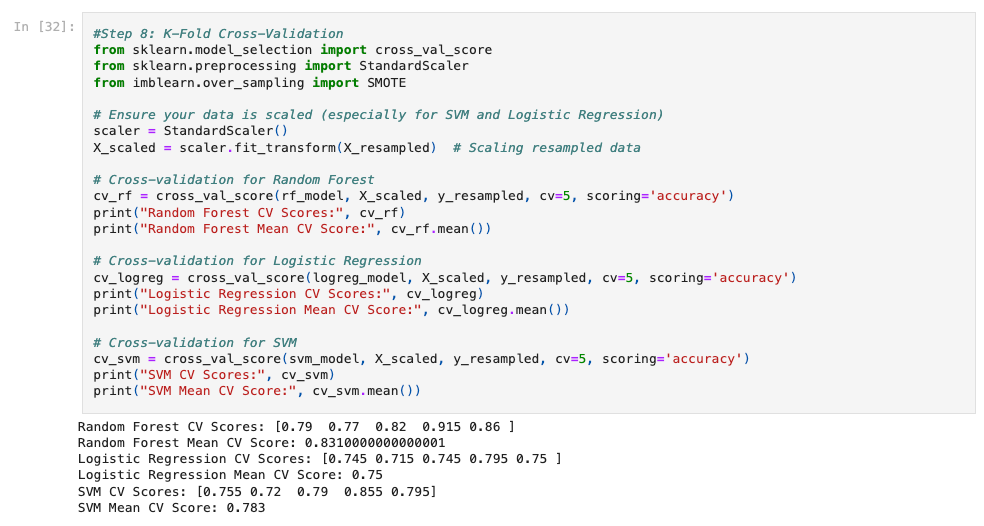
## AUC Scores: Quantitative measure to understand model effectiveness.











🤖 **How I Built This**

**1. Data Collection and Preprocessing:**

## I used the Pima Indians Diabetes Dataset, which contains several health features and a binary outcome indicating whether a person is diabetic. I made sure to handle missing values, normalize the data, and perform feature engineering where necessary.

**2. Model Selection:**

## I experimented with three machine learning models:

## Random Forest: Chosen for its robustness and ability to handle imbalanced data.

## Logistic Regression: A good baseline model for binary classification.

## Support Vector Machine (SVM): Selected for its capability in handling complex decision boundaries.

**3. Evaluation and Cross-Validation:**

## I used a combination of accuracy, precision, recall, F1-score, and ROC-AUC to evaluate the models. To make sure that the models weren’t overfitting, I also performed K-fold cross-validation.