**Big Data Solution Architecture, School of Applied Computer Science & Information Technology**

**PROG8420-23S-Sec1-Programming for Big Data**

**Assignment 10**

**Instructor: Jomis Varikayanickal John**

Task 1: Dataset Selection

1. Dataset URL:

<https://github.com/awesomedata/awesome-public-datasets/blob/master/Datasets/titanic.csv.zip>

1. Justification:

Titanic dataset contains information regarding the passengers on the Titanic, features of the passengers like age, gender, name, survival class. As the data set has numerical and categorical variables, classification tasks are best suitable. Here we are classifying all the numerical variables with the category gender. Dataset is preprocessed by identifying the missing and null values. Missing values for numerical variables are substituted by mean and null values for categorical variables are handled by dropping the columns. Feature Engineering is done by scaling the values from int data type to float. RandomForest and SupportVectorMachine models are created, trained, and tested with (80% training and 20% testing data). The model which provides the highest accuracy is the best suitable one for the classification.

Task 2: Data Exploration with Python

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Task 3: Data preprocessing with Python:

Task 4: Implement Machine Learning Models with Python

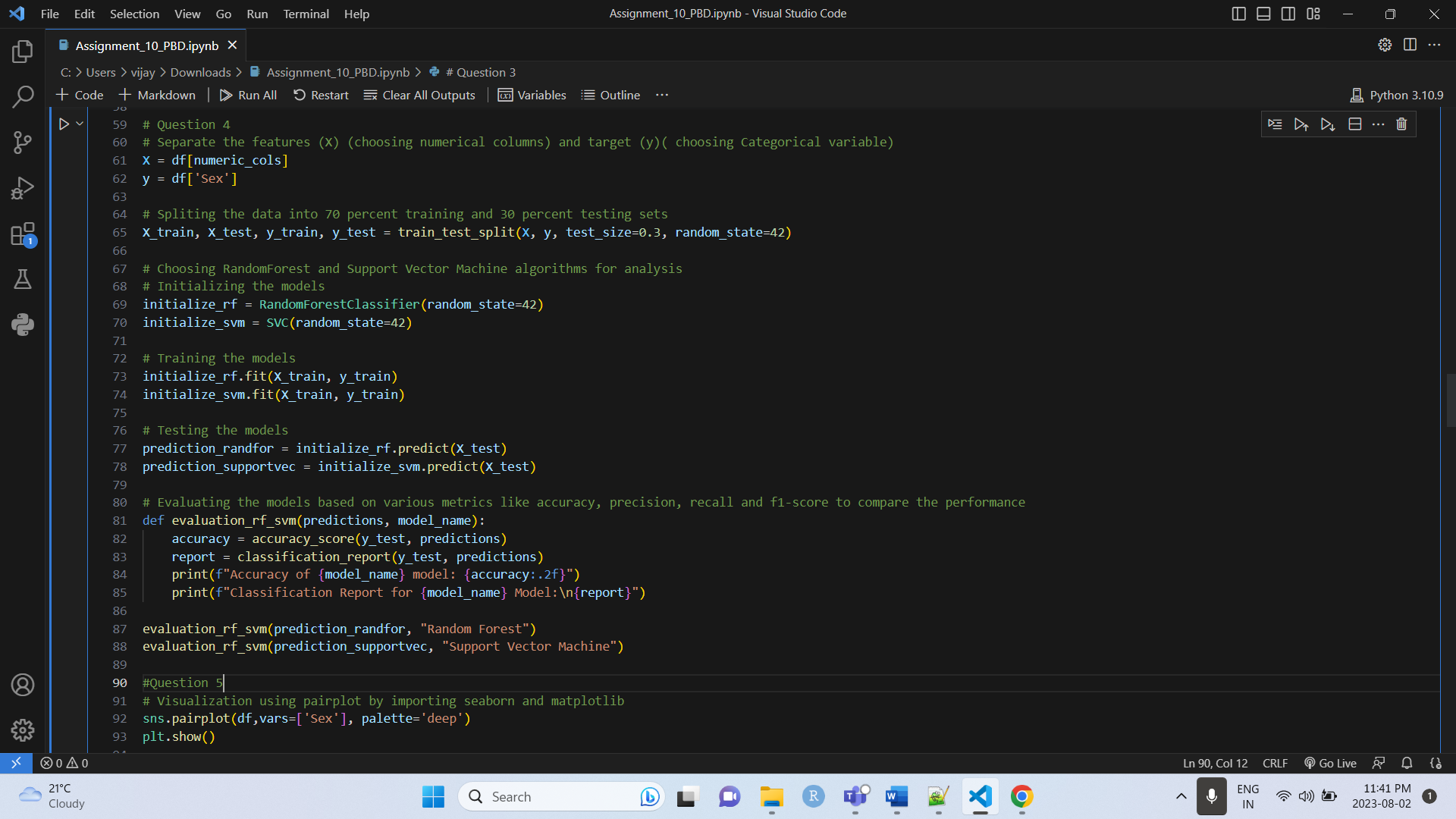
Task 5: Visualization with Python using

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



A computer screen shot of a black screen

Description automatically generated

A computer screen with a black screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A computer screen shot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated