Read me:

I am using mac os and the IDE that i have used is VS studio code,if all the plugins are installed.

The project folder contain the 3 folders
1)code - contains the code and csv files
2)screenshots - contains all the screenshots added in the report
3)report - contains the report

To run the codes

1) Pre-processing, Data Mining, and Visualization:

a)What variables do you plan to use as the input features?

Visualise and explore the data set, python program to to run this is 1a.py
Python3 1a.py

I have done pandas profiling too to understand the dataset and program to run is
Python3 pandas_profiling.py

Ps: output of the panda_profiling is taken into html format, name of the file is pandas_profiling.html (html file), make sure you have installed pandas_profiling library

b)What pre-processing (if any) did you execute on the variables?
c)Which independent variables are strongly correlated (positively or negatively)?
d)How many significant signals exist in the independent variables?
e)What derived or alternative features might be useful for analysis (e.g. polynomial features)?

For all the above question, I have implemented in a single python file, that is 1b.py, to run this,
Python3 1b.py

2) Classification:

For this i have tried to implement Logistic Regression, Decision tree, Random Forest and all are implemented in a single python file and python program to run this is '2.py'

Python3 2.py

3) Evaluation:

compute the confusion matrix, the F1 value, the bias, and the variance for the default operating point. This is implemented in the python file 3a.py and it will show comparison results of all the 3 models.
Python3 3a.py

For at least one of your methods, generate a Receiver Operator Curve [ROC] by varying the operating point of the classifier.For this, I have implemented it in the '3b.py'.
Python3 3b.py

1. **Which classifier did the best?**
2. **What statistic are you using to decide which classifier is the best performer?**
3. **What does the bias and variance indicate as to what the best next steps to take would be to improve performance?**
4. **What would be a good operating point for the classifier for which you generated the ROC curve?**

Answers to all these questions are provided in the report.

4)Iteration:

I have taken the random forest model for iteration and I have done 3 iterations on it and compared the confusion matrix and accuracy. The python program to run this is '4.py'

Python3 4.py

I have done two extensions for this project

First Extensions:

I have Attempted an additional iterations to make your system better. I have used the same Random Forest model and done a fourth iteration by hyperparameter tuning and criterion to Entropy. The python program to run this first extensions is extensions_1.py

Python3 extensions_1.py

Second Extensions:

Using more ML methods, I have tried to use a support vector machine on the dataset, further comparing the accuracy and confusion matrix with other 3 previously implemented models. The python program to run this is extensions_2.py

Python3 extensions_2.py

PS: I have not used any time travel days for this project and the project is submitted before the due date.