Final project is for 25 points and you are expected to work on an original idea that is substantial and covers at least 7 Python elements from the following list.

1. Use any data structure like list, dictionary, set or tuple (Used for Data Cleaning & Exploration)
2. List comprehension (Used for Data Munging to create Labels)
3. Dictionary comprehension (Used to select best Model based on Accuracy)
4. Functions (Used to show number of null, missing columns during Data Exploration)
5. Classes (Used to create instances of Classifier models)
6. User created iterators
7. Importing external modules (Used for Loading data, plotting, saving graphs, Machine Learning Models and measuring Accuracy)
8. Error checks using try-except
9. File input and output (Used to save Plots/Graphs to pdf)
10. Regular expression
11. Itertools (Used iteritems to access Dictionary list of models for selecting best model)
12. Decorators

Project report details

Please format the final project report according to the instruction below.

1. Introduction - This section, you should describe the problem that you are solving, any background information that will help the instructors to understand the program

INTRODUCTION : This project is Analysis of Census Income dataset downloaded from UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/census+income

Purpose of the project is being able to predict whether Income will be more than $50K or less than $50K based on Inputs as given in Census Income Dataset.

2. Requirements - List all the Python modules that need to be installed. If some of these modules need a specific version, please indicate so. You can also list any other conditions that are needed to run the program.

Python libraries used for the Census Income analysis are:

- Pandas :
  For analysis loading the data from Adult.csv as dataframe
  import pandas as pd
- Matplotlib :
  For plotting graphs, used pyplot module
  import matplotlib.pyplot as plt
  For saving plots/graphs to Pdf file, used backends.backend_pdf module
  from matplotlib.backends.backend_pdf import PdfPages

- Sklearn :
  For generating Training and Test data, used cross_validation module
  from sklearn.cross_validation import train_test_split
  Assumption for this analysis is splitting Data into 2 parts – 80% as Training, 20% as Test
  Decision Tree Classifier to fit (Training+Test) features and predict labels
  from sklearn import tree
  Creating an instance of the Decision Tree Classifier Class for our analysis
  cdt = tree.DecisionTreeClassifier()
  K Neighbors Classifier to fit (Training+Test) features and predict labels
  from sklearn.neighbors import KNeighborsClassifier
  Creating an instance of the K Neighbors Classifier Class for our analysis
  knc = KNeighborsClassifier()
  Support Vector Machine Classifier to fit (Training+Test) features and predict labels
  from sklearn.svm import SVC
  Creating an instance of the SVM Classifier Class for our analysis
  svc = SVC()
  Linear Logistic Regression Classifier to fit (Training+Test) features and predict labels
  from sklearn.linear_model import LogisticRegression
  Creating an instance of the Logistic Regression Class for our analysis
  lr= LogisticRegression()
  Random Forest Classifier to fit (Training+Test) features and predict labels
  from sklearn.ensemble import RandomForestClassifier
  Creating an instance of the Random Forest Class for our analysis
  lr= LogisticRegression()
  To measure accuracy of Prediction of models, used metrics module
  from sklearn.metrics import accuracy_score
  To choose Best Classifier model used model_Selection module
  from sklearn import model_selection

3. Description of the Python program . You need to describe the programs that you wrote.

Steps to figure out the best model for the Census Income dataset are as follows. Detailed description is in Ipython notebook Income_Analysis_Srabasti.py uploaded in zip file.

i. Load Data
ii. Analyze Data
iii. Feature Engineering
iv. Modeling
v. Choosing the Best Model

4. Screenshots of the program output - If you are using a specific hardware and cannot obtain screenshot, please enclose appropriate photographs

Attached Income_Analysis.pdf with Graphs for analysis and Final Model selection for Algorithm.

5. Conclusion - Describe in brief the problem you solved, the program you wrote and obtained output.

For Census Income dataset, after analyzing data, dropped data containing missing values. Feature engineering for parameters like marital status and creating labels has been done. For the purpose of the analysis Data is split into 2 parts – 80% as Training, 20% as Test.
Ran different Classifiers – Decision Tree, K Neighbors, SVM, Random Forest and Logistic Regression along with Accuracy reports for each of the Classifiers.
Finally, analyzing accuracy of each of the models, found the best classifier to be Logistic Regression and SVM.
Personally, I think using Logistic Regression model would be best since it is less time consuming as compared to SVC.

6. Python program - If the program is one file, please add it as one of the pages in the report.

7. Please make sure that the final report is in pdf format.

8. One zip folder - Add the files to one folder and zip it and upload the zip file. I only prefer zip files.