PA 2: Classification - Decision Tree

Student Details (1 Point)

First Student Name and ID: ABC 1001XXXXXX

Second Student Name and ID: XYZ 1001XXXXXX

Notes: When submitting, fill your name and ID in this cell. [1 point]

Do not to forget to cite any external sources used by you.

Programming Assignment Submission Instructions (2 Points)

Step 1: Rename this submission file as 'yourLastName_Last4digitsofyourID_DT.ipynb' [1 point]

Step 2: Place this file inside the folder 'PA#2_Classification_yourLastName' [1 point]

Do not upload the database file [-20 points]

Programming Assignment Details (7 Points)

For this assignment use Jupyter notebook, Panda, and scikit.

1) Load Heart dataset from cardio_train.csv [1 points]

Features:

- Age | Objective Feature | age | int (days)
- Height | Objective Feature | height | int (cm) |
- Weight | Objective Feature | weight | float (kg) |
- Gender | Objective Feature | gender | categorical code |
- Systolic blood pressure | Examination Feature | ap_hi | int |
- Diastolic blood pressure | Examination Feature | ap_lo | int |
- Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
- Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
- Smoking | Subjective Feature | smoke | binary |
- Alcohol intake | Subjective Feature | alco | binary |
- Physical activity | Subjective Feature | active | binary |
- Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

All of the dataset values were collected at the moment of medical examination.

2) Create a dataframe and print the first and last five records of your dataset. [2 points]

3) Print the class labels. [2 points]

4) Split your dataset 70% for training, and 30% for testing the classifier. [2 points]

DecisionTree (10 Points)

1) Use gini and entropy to measure the quality of a split. [2 points]

2) Use comments to explain your code and variable names. [1 point]

3) Calculate and print the confusion matrix (use graphics instead showing a 2D array), and the classification Report (includes: precision, recall, f1-score, and support. [2 points]

4) Print the decision tree visualization. [5 points]

Naive Bayes (10 Points)

1) Use Naive bayes classifier (Gaussian) to predict the test data[5 point]

2) Use comments to explain your code and variable names[1 point]

3) Calculate and print the confusion matrix (use graphics instead showing a 2D array), and the classification Report (includes: precision, recall, f1-score, and support). [4 points]

Report (20 Points)

1) Describe the Decision Tree methods, and Naive Bayes classifier. Dont copy paste it from the internet. Write it on your own. [4 points]

2) Describe the datasets [3 points] like what do you understand from the dataset? and if you have done any pre-processing , and your code, please write down your observation. [2 points]

4) Visualization of the decision tree for gini and entropy.[4 points]

5) Interpret your results, compare gini and entropy [3 points]

6) Visualize the dataset, for the target variable - 2 graphs [4 points]

Do not to forget to cite your sources!

Please consult the TA before using any other packages apart from sklearn,numpy,pandas, matplotlib and seaborn.