Q3.

a) Preprocessing:
   i) Converted all the preprocessed files into csv file and then added the column or feature name to the dataset(csv file).
   ii) Appended all the four files, processed.cleveland, processed.hungarian, processed.switzerland and processed.va.
   iii) Dropped the samples having no data i.e. having null values.
   iv) Replacing output 2,3,4 to integer 1 from the “num” column which is the output label just because we have to predict where there is a presence of heart disease in the patient can be done just by two classes.
   v) Plotted the pairplot between them to check the range of values and different values.
   vi) Plotted the correlation heatmap just to check the features that are more relevant.
b) Splitted the dataset into a train and test set in the ratio 80:20.
c) Trained the decision tree model using both ‘entropy’ and ‘gini impurity’ as the criterion and found the accuracy as following:
   i) entropy_accuracy : 0.7833333333333333 (78.33%)
   ii) gini_accuracy : 0.8 (80%)
d) The best criterion was coming to be gini impurity, now after putting the parameters such as standard scaler, PCA, decision tree into the pipeline that will tune by GridSearchCV (scaling the data will help to range the data between same range so that there will not be any dominance of feature of data having higher values, applied PCA for reducing the dimension of the dataset). Selected few parameters such as gini criteria for decision tree classifier, list [2,4,6,8,10,12] for max_depth , min_sample_split list as[2,5,10, 20] (it will give the min data that must present at which the split can happen), and max_features as [None, ‘sqrt’, ‘log2’, 1, 5] and then put these hyperparameters into gridsearch and tuned it, and as a result we found :
   i) Best Criterion: gini
   ii) Best max_depth: 2
   iii) Best Number Of Components: 1
   iv) DecisionTreeClassifier(max_depth=2)
   v) Best min_samples_split: 2
   vi) Best max_features: None
e) After tuning the hyperparameters I was getting the accuracy of 86.67%
f) I have trained the model using RandomForestClassifier on the X_train dataframe and was getting an accuracy of 86.67%.
g) Then tuned the hyperparameter for random forest classifier by first defining the parameters such as pca__n_components, rf__n_estimators(means number of weak classifier), rf__max_depth, rf__min_samples_split, rf__max_features and then passed them into GridSearchCV as hyperparameters. And was getting results as follows:
   i) RandomForest - Best n_estimators: 100
   ii) RandomForest - Best max_depth: None
   iii) RandomForest - Best min_samples_split: 3
   iv) RandomForest - Best max_features: sqrt
   v) RandomForest - Best Number Of Components: 10
h) And accuracy was coming to be 83.33%.

Q2.

a) Loaded the thyroid csv file into csv file and then converted the categorical variables into categorical codes, dropped NaN, converted the dataframe having true, false as 1, 0 respectively.
b) Used gini index as criteria.
c) Functions:
    i) cost_function(): Calculates the impurity using the gini impurity.
    ii) make_split(): This function essentially divides the dataset into two subsets based on a specified feature and threshold.The split is done by evaluating which data points satisfy the condition (values in the selected feature less than the threshold) and which do not, creating a left child and a right child for further splitting.
    iii) max_depth(): checks if the depth of the tree is increased or not.
    iv) predict(): predicts the class of the input provided and returns the answer.
d) Calculated the effectiveness of the model using evaluation from the test set and the accuracy was coming to be 99.48%.