INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Machine Learning Assignment 1

Venkata Sai Suvvari 20CS10067
Shashank Goud Boorgu 20CS30013

Regression Tree

Question:

i Split TrainBTree into 70%-30% to form training and testing sets, respectively. Build Regression Tree. Train the classifier using sum of squared errors (no packages to be used for Regression Tree).

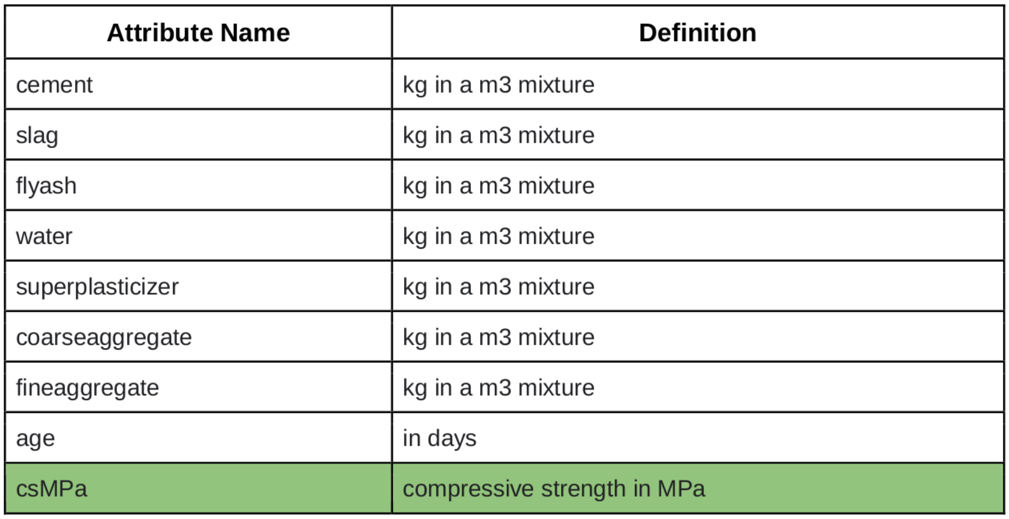
ii Repeat (1) for 10 random splits. Print the best test accuracy and the depth of that tree.

iii Perform rule post pruning operation over the tree obtained in (2). Plot the variation in test accuracy with varying depths. Print the depth for which the model over-fits. Print the pruned tree obtained in hierarchical fashion with the attributes clearly shown at each level.

iv Prepare a report including all your results.

Attribute Name and Definition:

Figure 1: Attributes and Definition



Attribute Name	Definition
cement	kg in a m3 mixture
slag	kg in a m3 mixture
flyash	kg in a m3 mixture
water	kg in a m3 mixture
superplasticizer	kg in a m3 mixture
coarseaggregate	kg in a m3 mixture
fineaggregate	kg in a m3 mixture
age	in days
csMPa	compressive strength in MPa

Algorithm:

a. The data should be split into 70%-%30 for training and testing separately for ten times and the best tree with lowest mean squared error should be taken.

b. The tree is built using training data by finding the threshold values possible and finding the threshold value at which the minimum value of error occurs.

c. Recursively the tree is built having 10 data points in each split.

d. The tree is allowed to over-fit.

e. The rule post pruning is applied to lower the over-fitting.

f. The criterion for comparing the error is **Variation** which is the mean of squared errors.

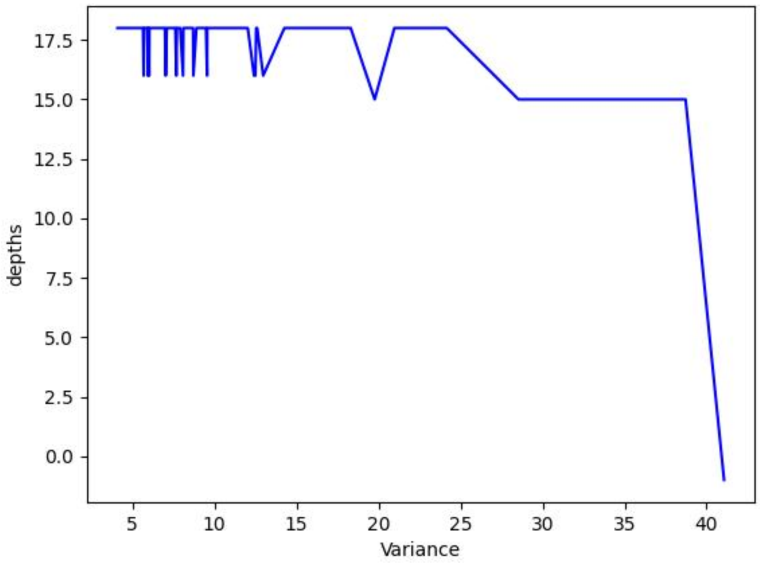
g. **Variation:** $Var = \sum(predictions[i] - Y[i])^2/(\text{length of dataset})$

h. The pruning shall be done until the error of the pruned tree is not worse than the original tree on the data set.

i. The pruned tree is later run over a new data split and the results are taken.

j. The plot is taken and the results are printed into a text file "q1_results.txt"

Results:



Major Functions:

a. **build_tree:** Builds the tree for each 70-30 split.

b. **get_optimal_split:** Traverses through all the features and finds the best feature to make a split to get optimal tree.

c. **print_tree:** Prints the tree in hierarchical fashion.

d. **post_prune_tree:** Recursively prunes the tree to until the error is not worse than original optimal tree.

e. **train:** Trains the tree during pruning to fill the values of leaf nodes of pruned tree.

f. **predict:** Finds the predictions of the dataset given.

g. **make_predictions:** This recursively goes to leaf node to get the predictions.

Bayesian Tree

Question:

a. Randomly divide the Train_B_Bayesian into 70% for training and 30% for testing. Encode categorical variables using appropriate encoding method (in-built function allowed).

b. A feature value is considered as an outlier if its value is greater than 2 x mean + 5 x standard deviation ($2 \times \mu + 5 \times \sigma$). A sample having maximum such outlier features must be dropped. Print the final set of features formed. Normalise the features as required.

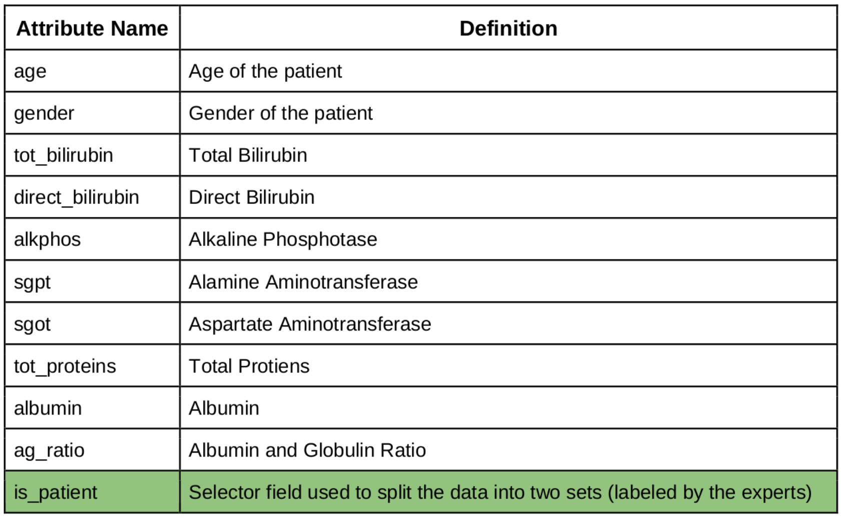
c. Train the Naïve Bayes Classifier using 5-fold cross validation (no packages to be used for Naïve Bayes Classifier). Print the final accuracy.

d. Train the Naïve Bayes Classifier using Laplace correction on the same train and test split. Print the final accuracy.

e. Prepare a report including all your results.

Attribute Name and Definition:

Figure 2: Attributes and Definition



Attribute Name	Definition
age	Age of the patient
gender	Gender of the patient
tot_bilirubin	Total Bilirubin
direct_bilirubin	Direct Bilirubin
alkphos	Alkaline Phosphotase
sgpt	Alamine Aminotransferase
sgot	Aspartate Aminotransferase
tot_proteins	Total Protiens
albumin	Albumin
ag_ratio	Albumin and Globulin Ratio
is_patient	Selector field used to split the data into two sets (labeled by the experts)

Algorithm:

a. We are given a data set that contains 583 liver patient records. 'is_patient' is a class label used to divide into groups(liver patient or not). We are using Gaussian Naive Bayes Classifier. The reason of using this is because we have continous data like sgot, albumin etc.

b. We have read the data using pandas read_csv function.

c. We have encoded the data using convertData function which takes arguments as data and changes the column corresponding to the feature 'gender' and we encoded 'Female' as 0 and 'Male' as 1.

d. We then found the outliers in each column by computing the value 2*mu + 5*sigma and removed those rows and normalized the data using (xi-xmin)/(xmax-xmin) in the function data_normalisation.

e. Then we divided the data into two parts of 70% training and 30% testing using the function sample_split.

f. We then performed 5-fold cross validation by dividing the training data set into 5 equal parts. Now we do 5 iterations and for each iteration we take one part(different part each time) as a testing set and the rest as a training set. Now we find the one with best accuracy and then test the original 30% training data set.