***Data Mining and Predictive Modelling Assignment 6***

Name: Rushikesh Jyoti

Division: A

Roll no: 27

SRN: 201901139

*Question*

* Read the dataset **"prostrate.csv"** that is provided to you.
* Build a suitable decision tree predictive model to predict **Tumor Log Volume** based on predictor information.
* Plot the decision tree and develop some metrics to determine the accuracy of your model. (Compute various evaluation parameters of the tree model built).
* Cross validate and optimize the model using hold back V-fold technique.

Use method of pruning to avoid over-fitting and derive the best size of the decision tree.

*Code*

setwd("C:\\VS\_Workshop\\Sem 6\\Data Mining and Predictive Modelling\\Assignments\\Ass6") *# nolint*

library(datasets)

library(caTools)

library(caret)

library(Metrics)

library(party)

library(dplyr)

library(magrittr)

library(tree)

library(corrplot)

*df* *=* read.csv("./prostate.csv")

head(df)

corrplot(cor(df), *diag=*FALSE)

*# lbph doesnt have enough correlation with lcavol so drop it*

*df* *=* select(df, *-*lbph)

head(df)

*#Splitting dataset into 4:1 or 80:20 ratio for train and test data*

sample\_data *<-* sample.split(df, *SplitRatio* *=* 0.8)

train\_data *<-* subset(df, sample\_data *==* TRUE)

test\_data *<-* subset(df, sample\_data *==* FALSE)

*# Create the decision tree model using ctree and plot the model*

model *<-* tree(lcavol *~* ., train\_data, *mincut=*1)

*# The minimum number of observations*

*# to include in either child node = 1*

model

plot(model)

*# Pruning the tree*

prune.tree(model)

*cut\_model* *=* prune.tree(model, *k=*9)

cut\_model

plot(cut\_model)

*predictions* *=* predict(model, test\_data)

*# predictions = predict(cut\_model, test\_data)*

predictions

print(test\_data*$*lcavol)

print(predictions)

errors *=* *function*(pred) {

    mae *<-* MAE(test\_data*$*lcavol, pred)

    mse *<-* mse(test\_data*$*lcavol, pred)

    rmse *<-* RMSE(test\_data*$*lcavol, pred)

    r2 *<-* R2(test\_data*$*lcavol, pred)

    cat("\nMean Absolute Error:", mae, "\nMean Squared Error:", mse, "\nRMSE:", rmse, "\nR-squared:", r2, "\n")

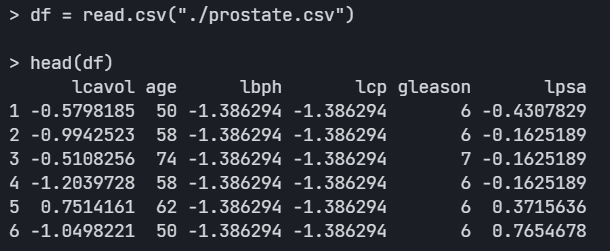
}

errors(predict(model, test\_data))

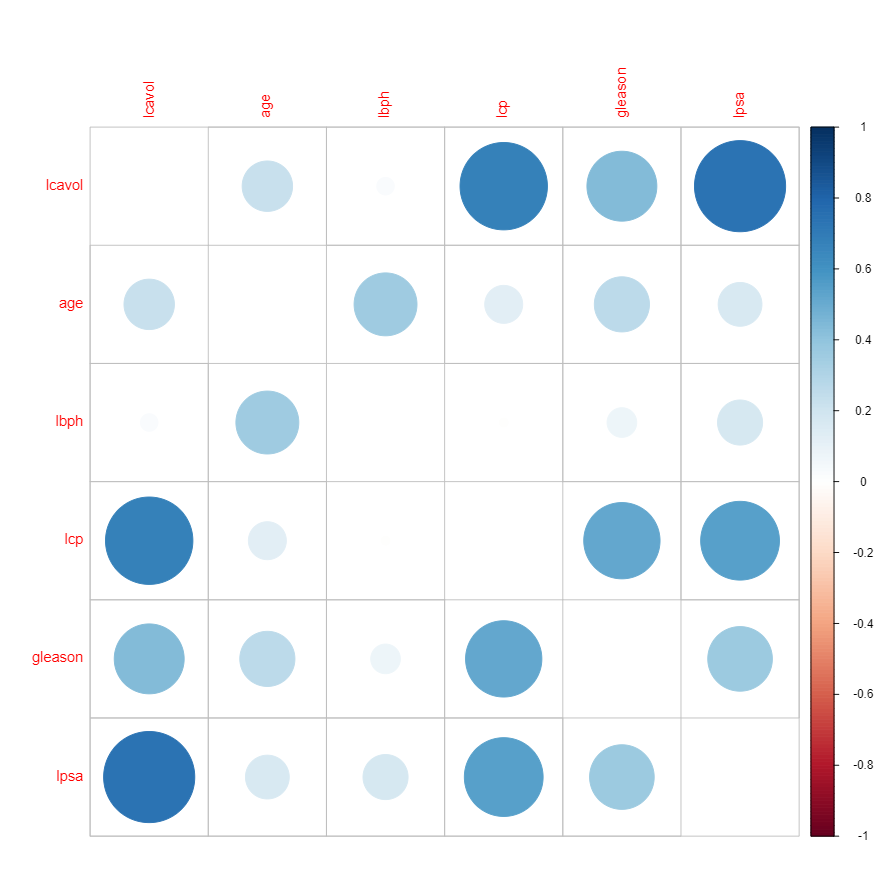
errors(predict(cut\_model, test\_data))

*Output*

Taking a look at our dataset

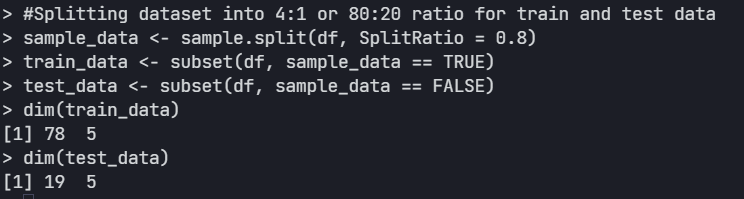


Correlation Plot of dataset

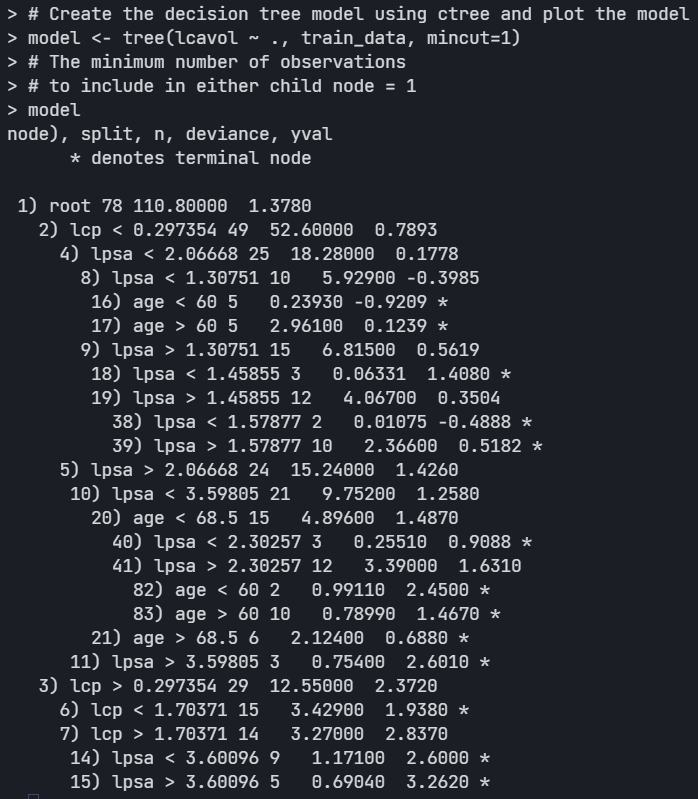


Since lbph is not correlated with lcavol we shall remove it

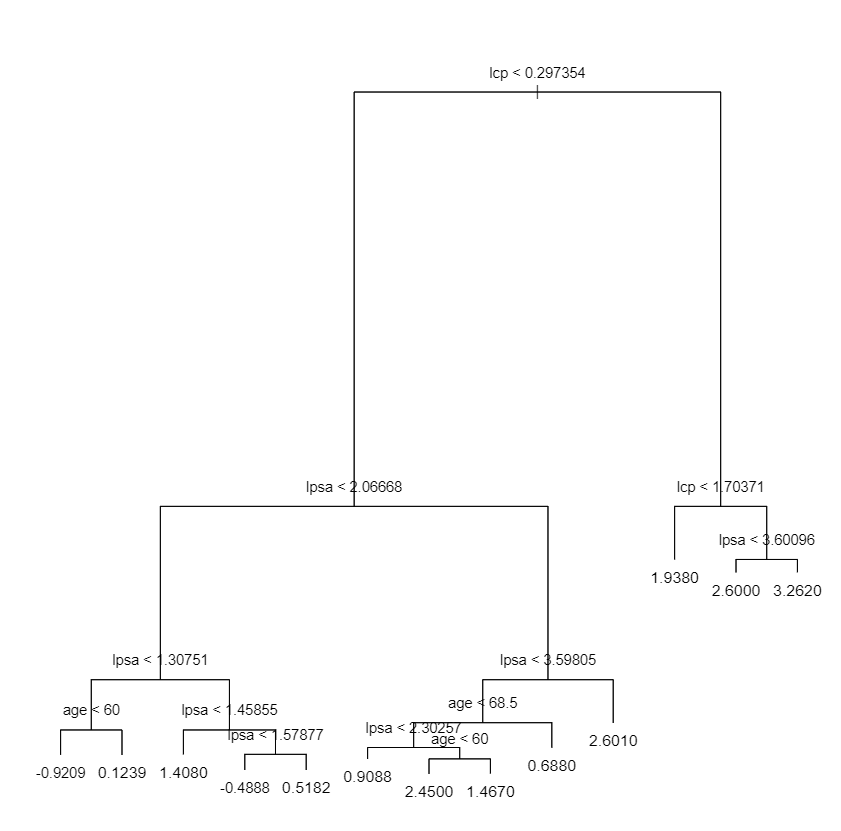
Split the data into 80:20 ratio



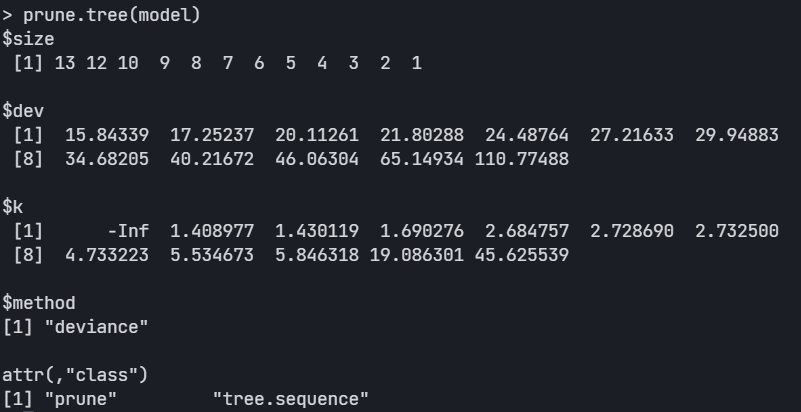
Creating the model



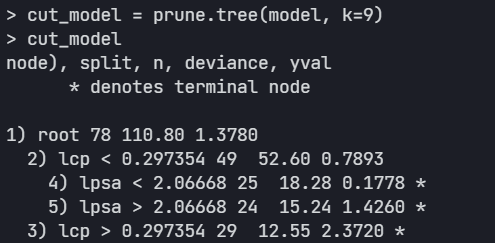
Plotting our model



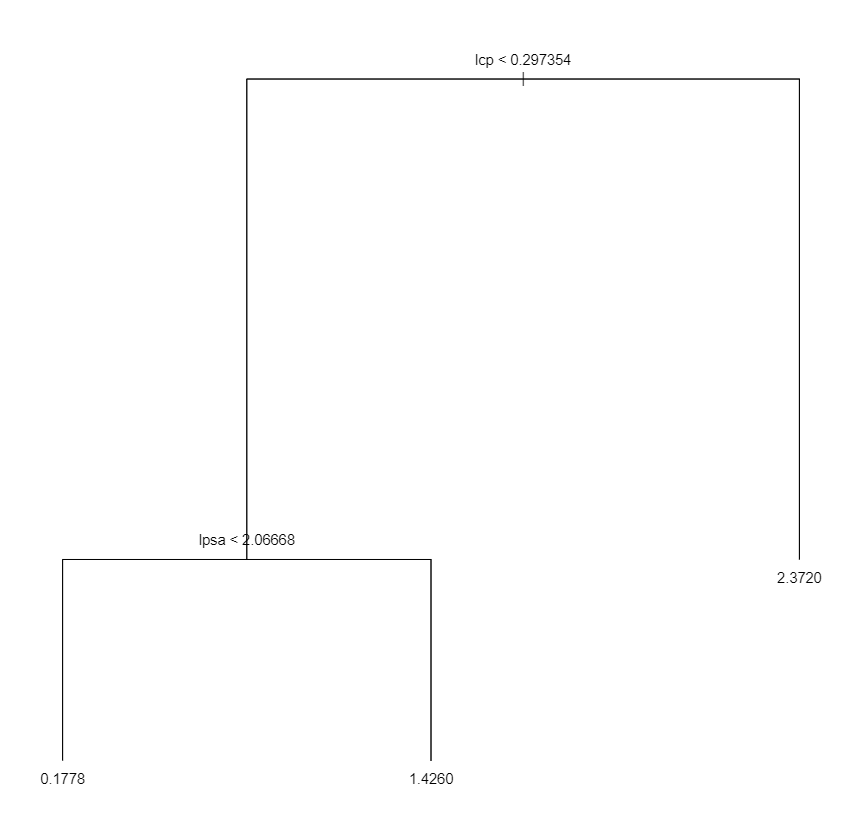
Let’s check some pruning status of our tree model



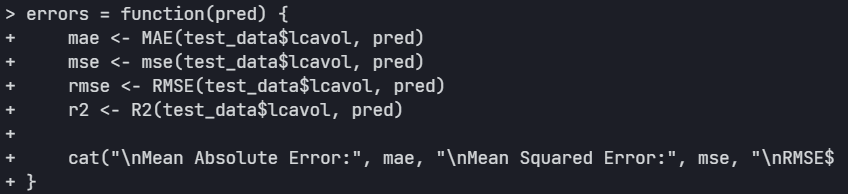
Let’s try pruning the tree with k=9



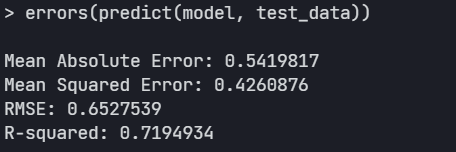
And we plot the pruned tree



Let’s use this custom function for errors (MAE, MSE, RMSE, R2)



Metrics of our original model



Metrics of pruned tree

