IE 7275: Data Mining in Engineering

USE CASE STUDY REPORT

ON

LIFE EXPECTANCY PREDICTION MODEL

Submitted to: Prof. Xuemin Jin

Group No.: Group 22
Student Names: Aviral Agrawal and Nidhi Agrawal

Executive Summary

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that effect of immunization and human development index was not taken into account in the past. To resolve this issue, we have considered data from year 2000 to 2015 for all the countries which include important immunization like Hepatitis B, Polio, and Diphtheria. It will help us to focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well.

This study is focused upon a prediction problem in which we predicted Life Expectancy of individuals based on the attributes such as Adult mortality, infant deaths, alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, diphtheria, HIV/AIDS, GDP, Population, thinness (1-19 years), thinness (5-9 years), Income composition of resources and schooling. The data-set related to life expectancy, health factors for 193 countries has been collected from www.kaggle.com (link) the open online database. We used regression and classification as our prediction algorithm by dividing data into a training set and validation set. We find the RMSE and correlation values to compare predictors responsible for getting the best result for our response variable which is life expectancy.

We have used Multiple Linear Regression, Multiple Linear Regression with Principal Component Analysis and Regression Tree and found that for the Regression Tree, the value of RMSE is 0.38944 which is lowest and the correlation value is 0.91924 that is the highest. Using classification, we have first done the sampling into training and validation data sets and then used K-Nearest Neighbor and Classification Tree where we got the accuracy as **80.10%** and **86.69%** respectively which is better for the Classification Tree.

I. Background and Introduction

Although there have been a lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition, and mortality rates. It was found that the effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on the data set of one year for all the countries. Hence, this gives the motivation to resolve both the factors stated previously by formulating a regression model based on the mixed-effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries.

(a) The problem query

The problem that we are going to tackle is to predict the life expectancy on the basis of Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness(1-19 years), thinness (5-9 years), Income composition of resources, schooling. It is expected that we will be able to predict the life expectancy of individuals using these attributes as accurately as possible which will help countries to identify areas to work on to improve overall life expectancy. Our aim is to come up with a model that can validate the whole data.

(b) The goal of our study

The goal of our study is to compare the model used which are KNN and Classification Tree on the basis of accuracy. Also, we will do regression analysis using Multiple Linear Regression, Multiple Linear Regression using Principal Component Analysis (PCA) and Random Forest Regression and then compare them using the RMSE and correlation values. Then predicting the life expectancy for the best algorithm.

(c) Possible solution

By doing the regression analysis and finding the RMSE and correlation values, we will be able to identify life expectancy based on the selected attributes and choose the model that best fits our data. We categorize the life expectancies as low, average and high and then use KNN and classification tree algorithms to find out the category in which the particular individual falls based on the given attributes. The most accurate model gives the most accurate result. This will help the person know what can be the possible life expectancy based on their attributes.

II. Data Exploration and Visualization

Data exploration starts from the Kaggle website where the data was collected by the WHO and the United Nations website. The data from 2000-2015 for 193 countries have been considered. It consists of 20 columns and 2938 rows which means 20 predicting variables. The major attributes in our data are the following:

Attribute	Description
Country	Country
Year	Year
Status	Developing or developed status
Life expectancy	Life expectancy in age
Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)

Infant deaths	Number of Infant Deaths per 1000 population
Alcohol	Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)
Percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita (%)
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	Measles - number of reported cases per 1000 population
BMI	Average Body Mass Index of the entire population
under-five deaths	Number of under-five deaths per 1000 population
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total Expenditure	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV/AIDS	Deaths per 1000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Population	The population of the country
Thinness (1-19 years)	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Thinness (5-9 years)	Prevalence of thinness among children for Age 5 to 9(%)
Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Number of years of Schooling(years)

Table 1: Attribute Description

When we first looked at the data, we eliminated the rows having null values to make the dataset efficient and predict more accurate results. Also, we have removed the columns such as country, year, status, GDP and population because we don't require them for our study for now. We plot the correlation for life expectancy to find out which attribute relates the most to life expectancy.

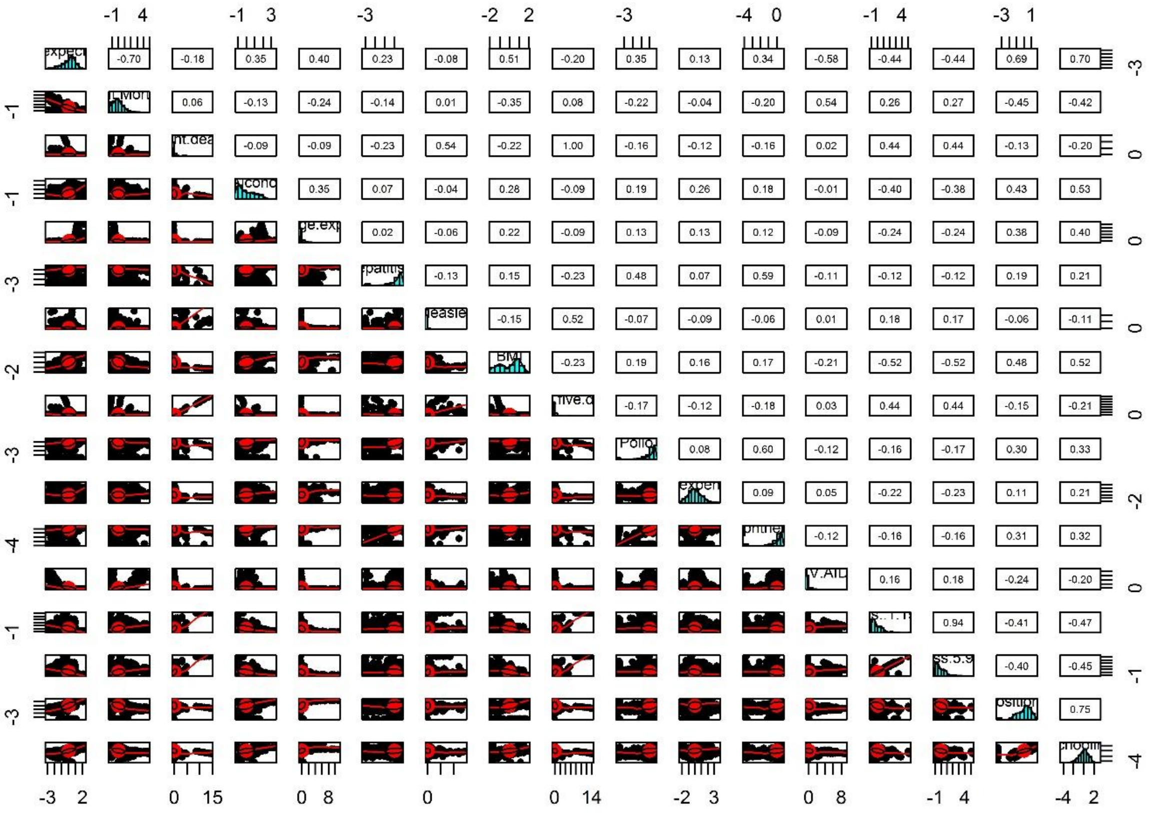


Figure 1: Correlation plot

From the correlation plot, we inferred that the most important attributes which affect the life expectancy the most are Adult Mortality, Alcohol, BMI, Total expenditure and schooling. Out of these attributes, Adult Mortality, Alcohol and Total expenditure have a negative correlation with life expectancy whereas BMI and schooling have a positive correlation with life expectancy.

Further, for classification, when we classified our data based on the Life expectancy as low, average and high, we compared the results from the start year of our data collection that is 2000 and the end year of our data collection that is 2015. We plotted bar graphs for the category count in 2000 and category count in 2015 (shown below in *figure 2*).

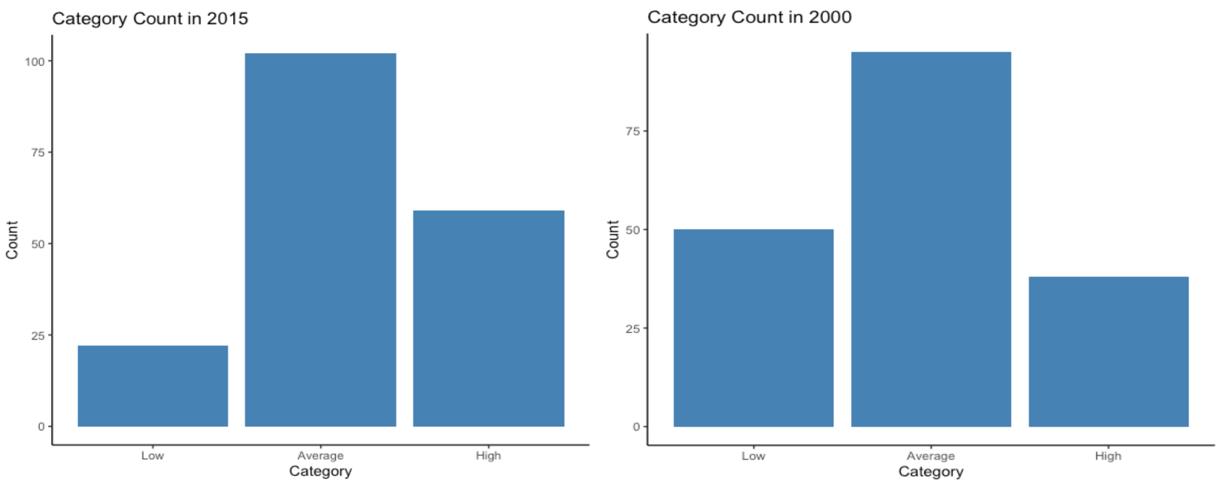


Figure 2: Bar graph for Life expectancy category count in 2000 and 2015

From *figure 2*, we saw that people having a low life expectancy have reduced as we see from year 2000 to 2015 and the count of people having an average as well as high life expectancy have increased.

III. Data Preparation and Preprocessing

The most important step is to prepare the data to get the best results from supervised machine learning. We prepared the data to get the most efficient output by removing any rows with null values in it and the columns that are not required.

	No. of rows	**No. of columns**
Before	2938	20
After	2083	17

Table 2: Data Cleaning

For doing the regression analysis, we standardize our data to get better accuracy with the regression models in the later stages.

For doing the classification analysis, we categorize the life expectancy into low, average and high and then use K-nearest neighbor and classification tree models to identify the category in which the particular individual falls based on the attributes he possesses.

The next important step for both regression as well as classification analysis is to divide the data into training and validation sets (60% training and 40% validation).

IV. Data Mining techniques and Implementation

The following problem is a prediction problem in which our response variable(y) is life expectancy and our predictors (x) are all the other variables available. The flowchart will describe the problem in the best way possible:

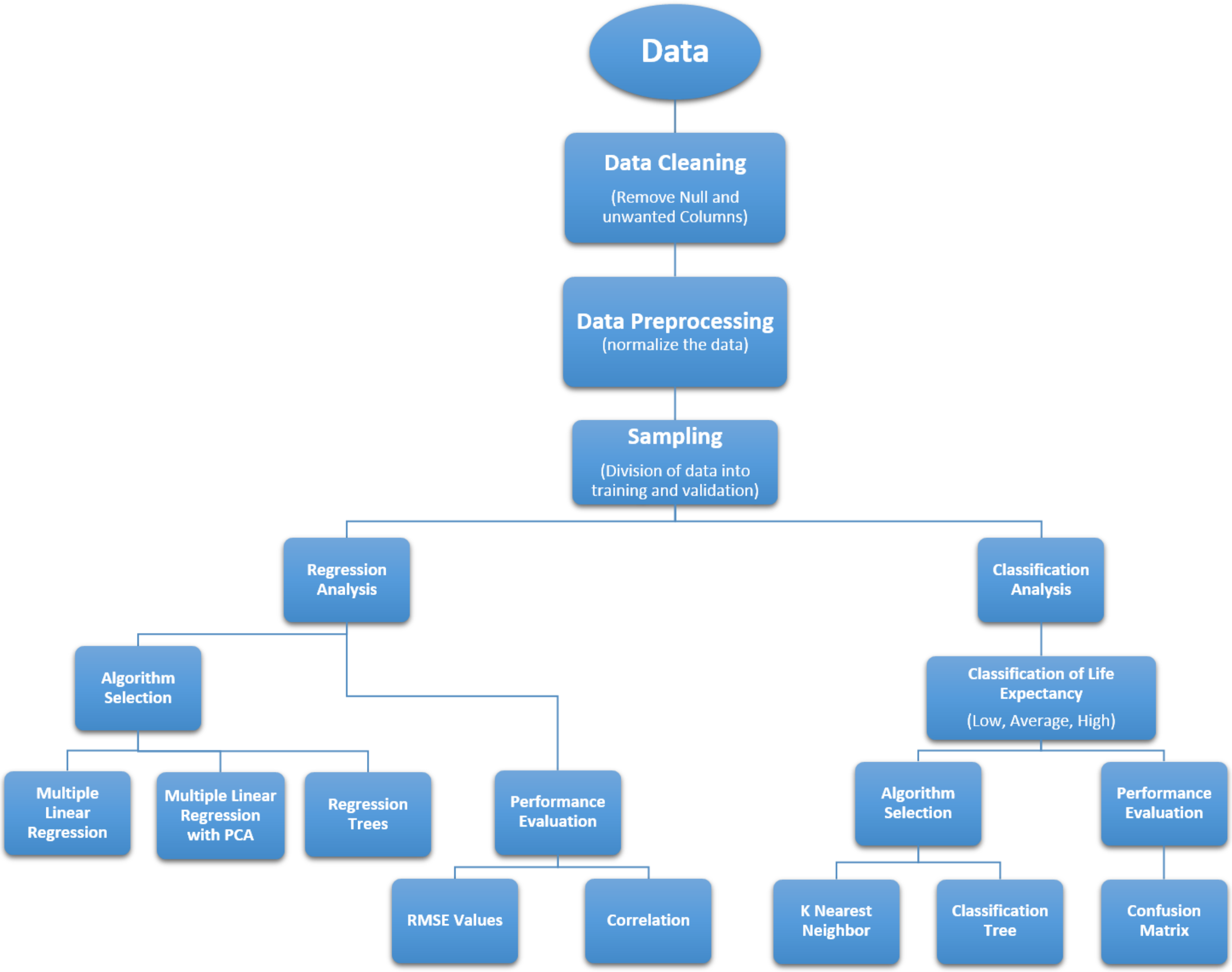


Figure 3: Flow Chart

The algorithms used are Regression analysis in which we have used Multiple Linear Regression, Multiple Linear Regression using Principal Component Analysis and Regression tree. Also, we have used the K-Nearest Neighbor algorithm and classification tree to determine which model is most accurate to determine the category in which a particular individual will fall based on the life expectancy.

For doing the regression analysis, we first define the RMSE function. After doing regression analysis using the different algorithms, we find RMSE values for each of them and the correlation values and compare the results.

For doing the classification, as we classify the life expectancy into 3 classes as low, average and high, we have applied the K-nearest Neighbor Algorithm as well as the Classification tree and then by using the confusion matrix we determine which is the most accurate out of the two.

V. Performance Evaluation

The performance evaluation or all the data mining techniques explored in our study is justified by calculating the RMSE values and correlation values for the three regression models and the confusion matrix for classification models. Also, we have used pruning for the regression tree as well as the classification tree where we have found that there is no change implying that our model is a best fit.

a) Regression Algorithms

***Multiple Linear Regression*:** As our data consists of several attributes, it helps us to determine the relative influence of all the attributes with the predictor variable that is the life expectancy in our case. The RMSE value that we got is 0.4474 and the correlation value as 0.8919.

***Multiple Linear Regression using PCA*:** Instead of considering all the attributes and finding relation with life expectancy, we find the principal components that are the attributes related the most to the life expectancy. As shown in *figure 4*, we got 4 principal components.
The RMSE value we got is 0.4816 and the correlation value is 0.8733.

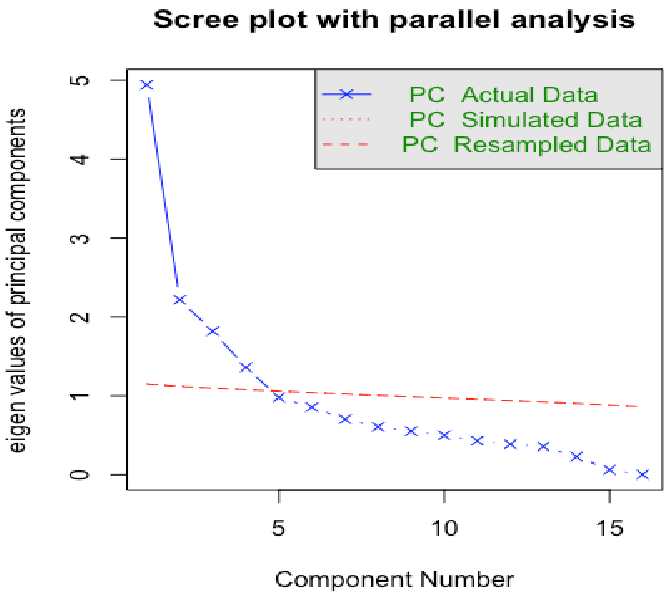


Figure 4: Principal Component Analysis

Regression Tree: The Random forest algorithm is one of the best algorithms to use for prediction problems and it saves time and it gives highest accuracy in most cases. The RMSE value we got is 0.3893 and correlation value is 0.9192 that is the best out of the 3 algorithms. *Figure 4* shows the regression tree.

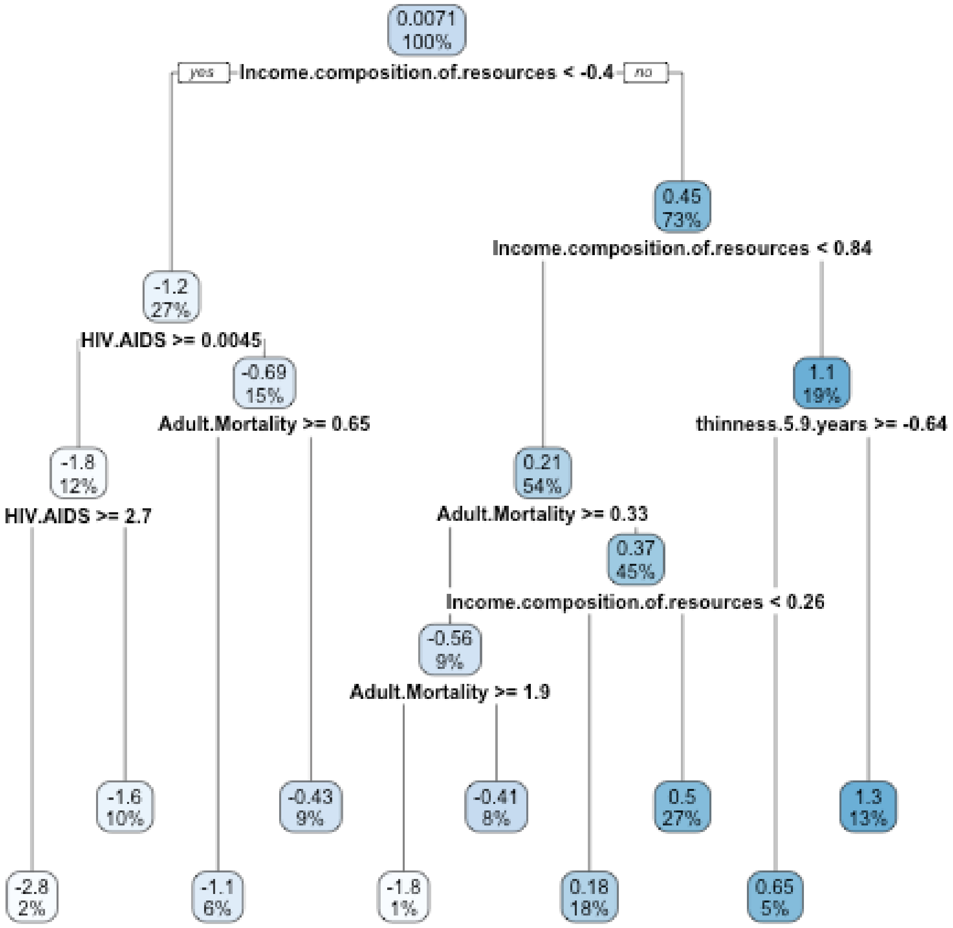


Figure 5: Regression tree

After pruning, there are no changes in the result which means that our model is the best fit.
For evaluating our error values, we have plotted box plots for our training and validation data (shown in *figure 6*).

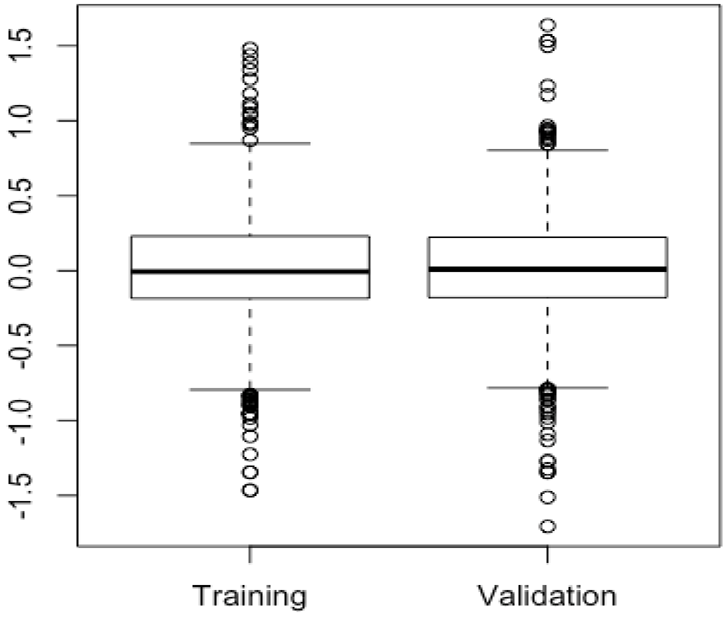
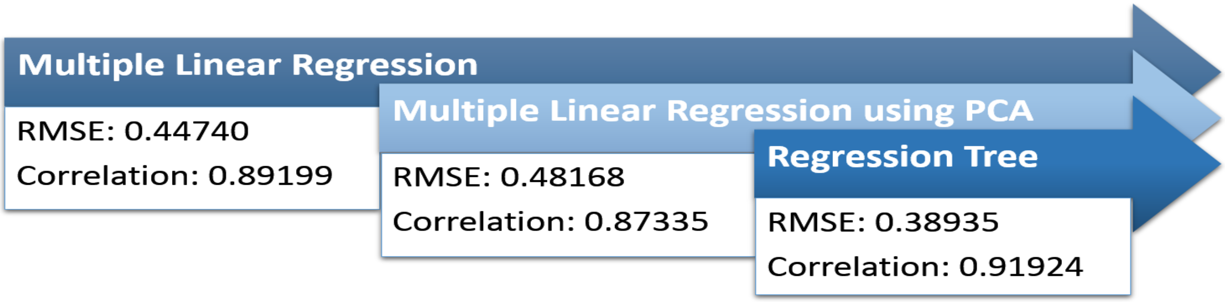


Figure 6: Boxplot for training and validation data

As we can infer from this, most of the error values lie close to zero. The outliers for the training data are less scattered that validates our RMSE values.



Multiple Linear Regression
RMSE: 0.44740
Correlation: 0.89199

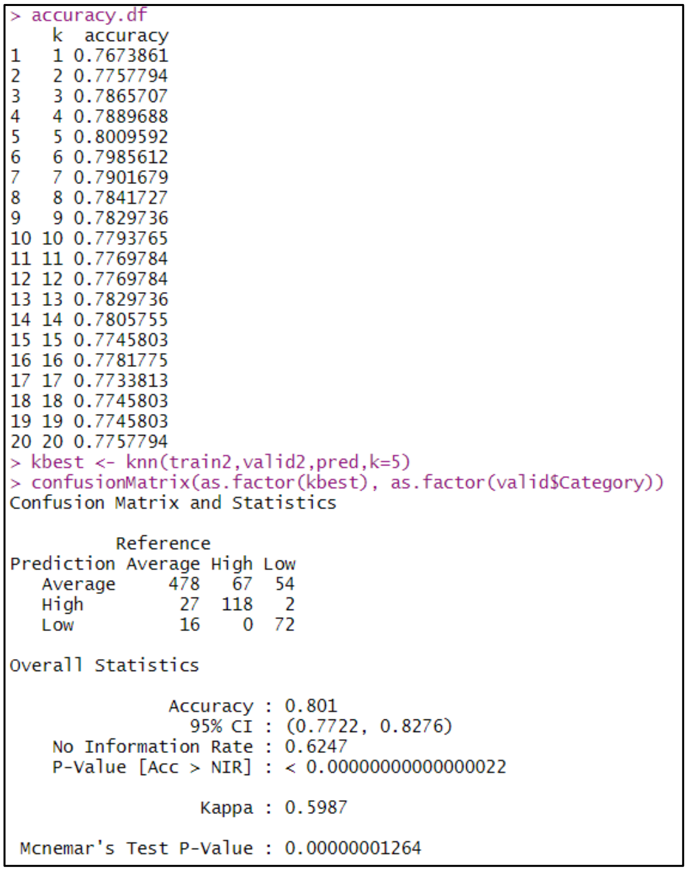
Multiple Linear Regression using PCA
RMSE: 0.48168
Correlation: 0.87335

Regression Tree
RMSE: 0.38935
Correlation: 0.91924

Figure 7: Performance Evaluation

b) Classification Algorithms

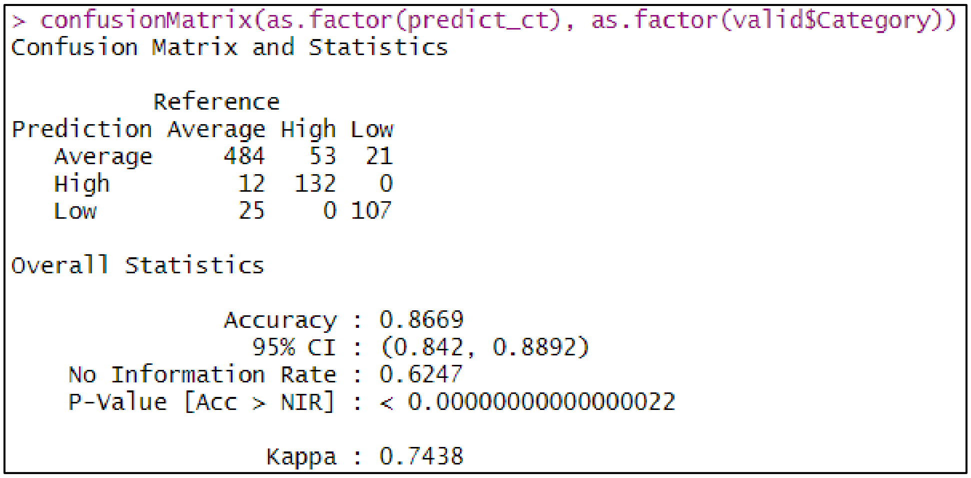
K-Nearest Neighbor: This is one of the easiest algorithms to apply in any case of classification problem. It classifies new cases on the basis of similarity measure (e.g.: distance functions). In k-NN classification, the output is a class membership.
For, k values, we have used values from 1 to 20 and determined the accuracy. A snapshot from R is shown in *figure 7* below which shows all the values for k=1 to 20 and the best value that is chosen that is accuracy=80.1% at k=5. Also, the confusion matrix is also shown.



```
> accuracy.df
    k  accuracy
1   1 0.7673861
2   2 0.7757794
3   3 0.7865707
4   4 0.7889688
5   5 0.8009592
6   6 0.7985612
7   7 0.7901679
8   8 0.7841727
9   9 0.7829736
10 10 0.7793765
11 11 0.7769784
12 12 0.7769784
13 13 0.7829736
14 14 0.7805755
15 15 0.7745803
16 16 0.7781775
17 17 0.7733813
18 18 0.7745803
19 19 0.7745803
20 20 0.7757794
> kbest <- knn(train2,valid2,pred,k=5)
> confusionMatrix(as.factor(kbest), as.factor(valid$Category))
Confusion Matrix and Statistics

          Reference
Prediction Average High Low
   Average     478   67  54
   High         27  118   2
   Low          16    0  72

Overall Statistics

               Accuracy : 0.801
                 95% CI : (0.7722, 0.8276)
    No Information Rate : 0.6247
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5987

 Mcnemar's Test P-Value : 0.00000001264
```

Figure 8: k-NN for values of k from 1 to 20 and confusion matrix

Classification Tree: Classification trees are used to predict the membership of cases or objects into classes of a categorical dependent variable from their measurements on the predictor variable(s). Here, our categorical dependent variable is life expectancy which has been divided into the categories low, medium and large.



```
> confusionMatrix(as.factor(predict_ct), as.factor(valid$Category))
Confusion Matrix and Statistics

          Reference
Prediction Average High Low
   Average     484   53  21
   High         12  132   0
   Low          25    0 107

Overall Statistics

               Accuracy : 0.8669
                 95% CI : (0.842, 0.8892)
    No Information Rate : 0.6247
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7438
```

Figure 9: Confusion matrix for Classification tree

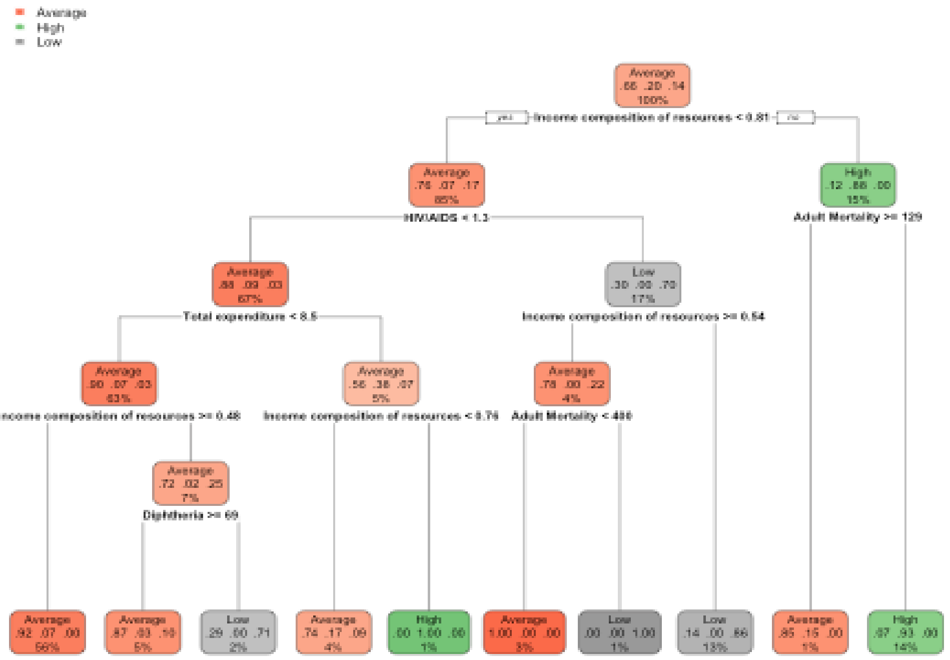
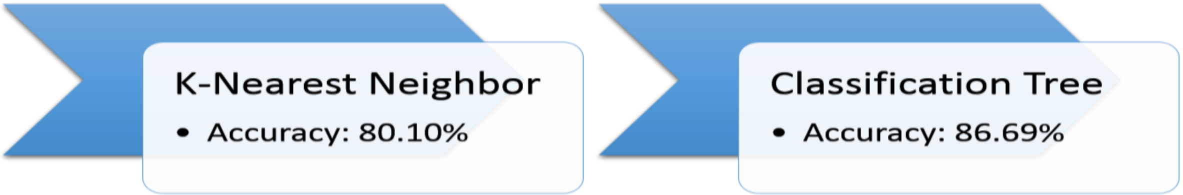


Figure 10: classification tree

The accuracy that we got from the classification tree is 86.69%. *Figure 8* shown above shows the classification tree.

After pruning, there is no change in the classification tree implying that our model obtained is the best fit.



K-Nearest Neighbor

- Accuracy: 80.10%

Classification Tree

- Accuracy: 86.69%

Figure 11: Performance Evaluation

VI. Discussion and Recommendation

The overall approach for the project is to predict the Life Expectancy from data given from years 2000 to 2015 which is properly justified by the approach that we had in the project. The scatter plot plotted to determine the most related attributes, selection of algorithms, refining the data in order to get the best accuracy, plotting the box plots for evaluation of training data and validation data, determining the RMSE and correlation values and accuracy for classification models.

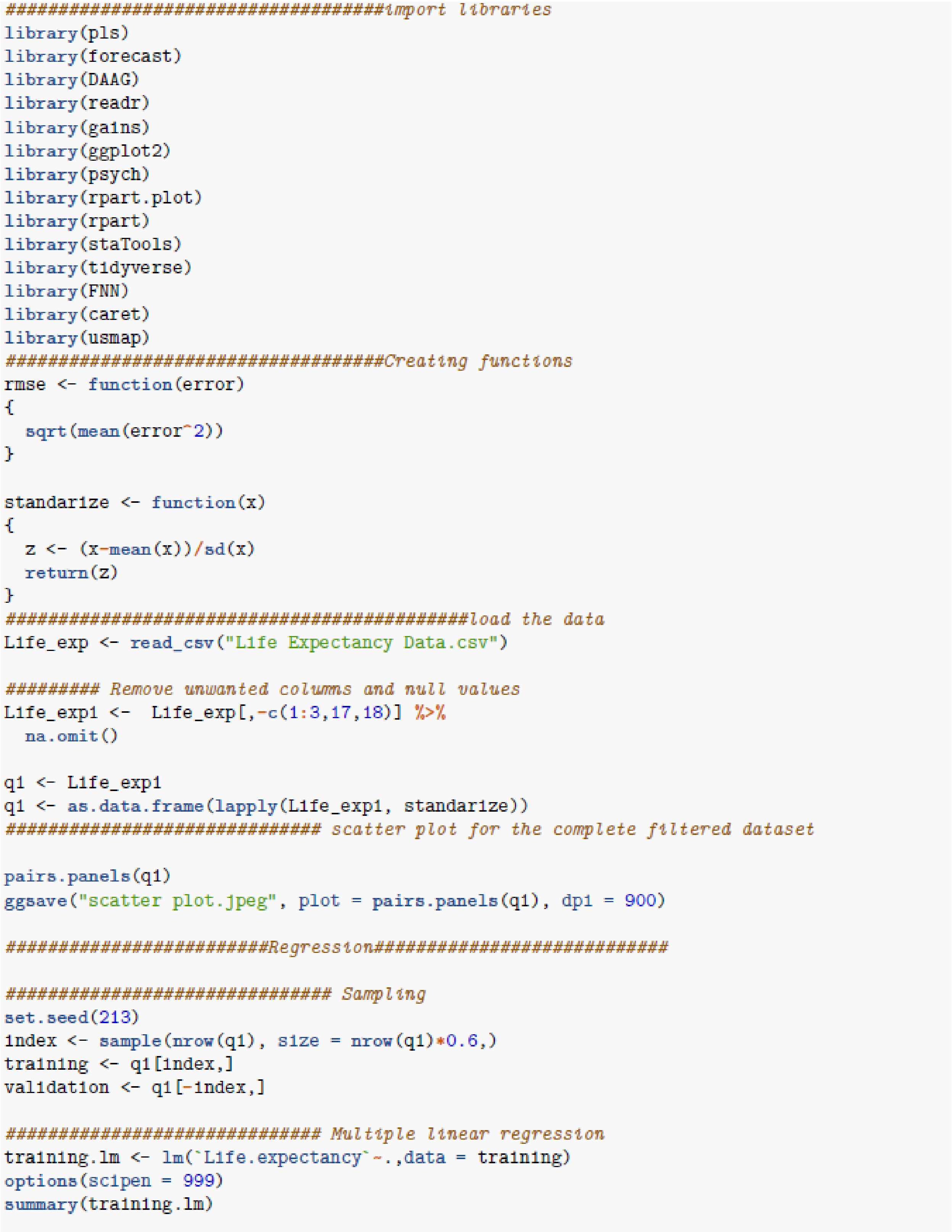
The recommendation can be that the accuracy from k-Nearest Neighbor Algorithm can be further improved by calculating the distance eight using a combination of coal mean based k-Nearest Neighbor and distance weight k-Nearest Neighbor.

VII. Summary

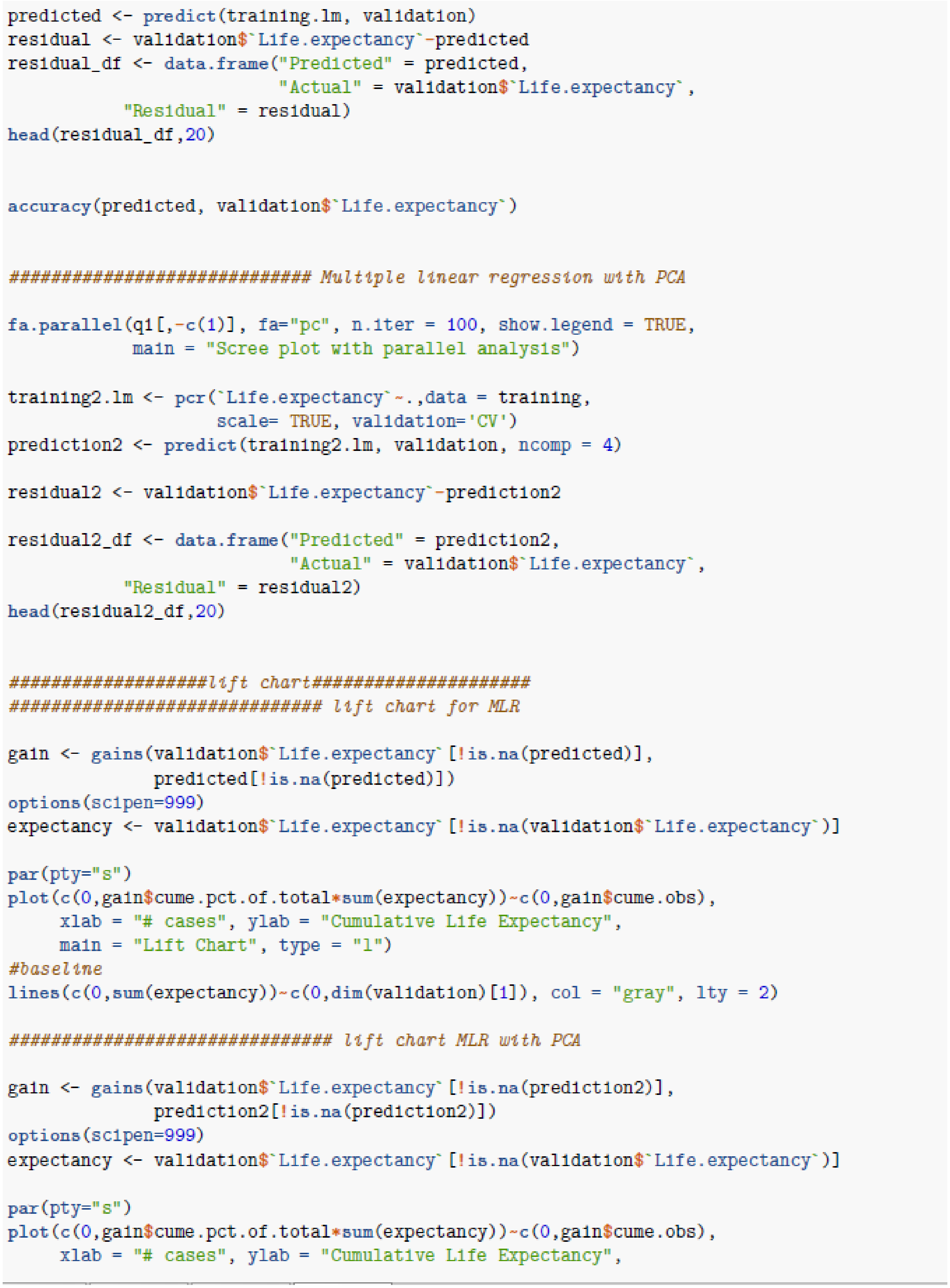
Starting with a prediction problem in which we predicted the life expectancy of an individual based on a number of attributes, we have used three different regression algorithms - Multiple Linear Regression, Multiple Linear Regression using Principal Component Analysis and Regression Tree. We created a scatter plot, compared the RMSE values and correlation values. The conclusion of regression analysis was that the Life Expectancy (our response variable) is predicted best by using the regression tree.

After converting the problem to classification problem by categorizing the life expectancy as low, medium and high we have used the K-Nearest Neighbor algorithm and classification tree. We created bar graphs for category count in 2000 and category count in 2015. The comparison between k-NN and classification tree is done on the basis of accuracy and using confusion matrix. K-Nearest Neighbor algorithm is a better classification algorithm as it gives a better accuracy of 86.69%.

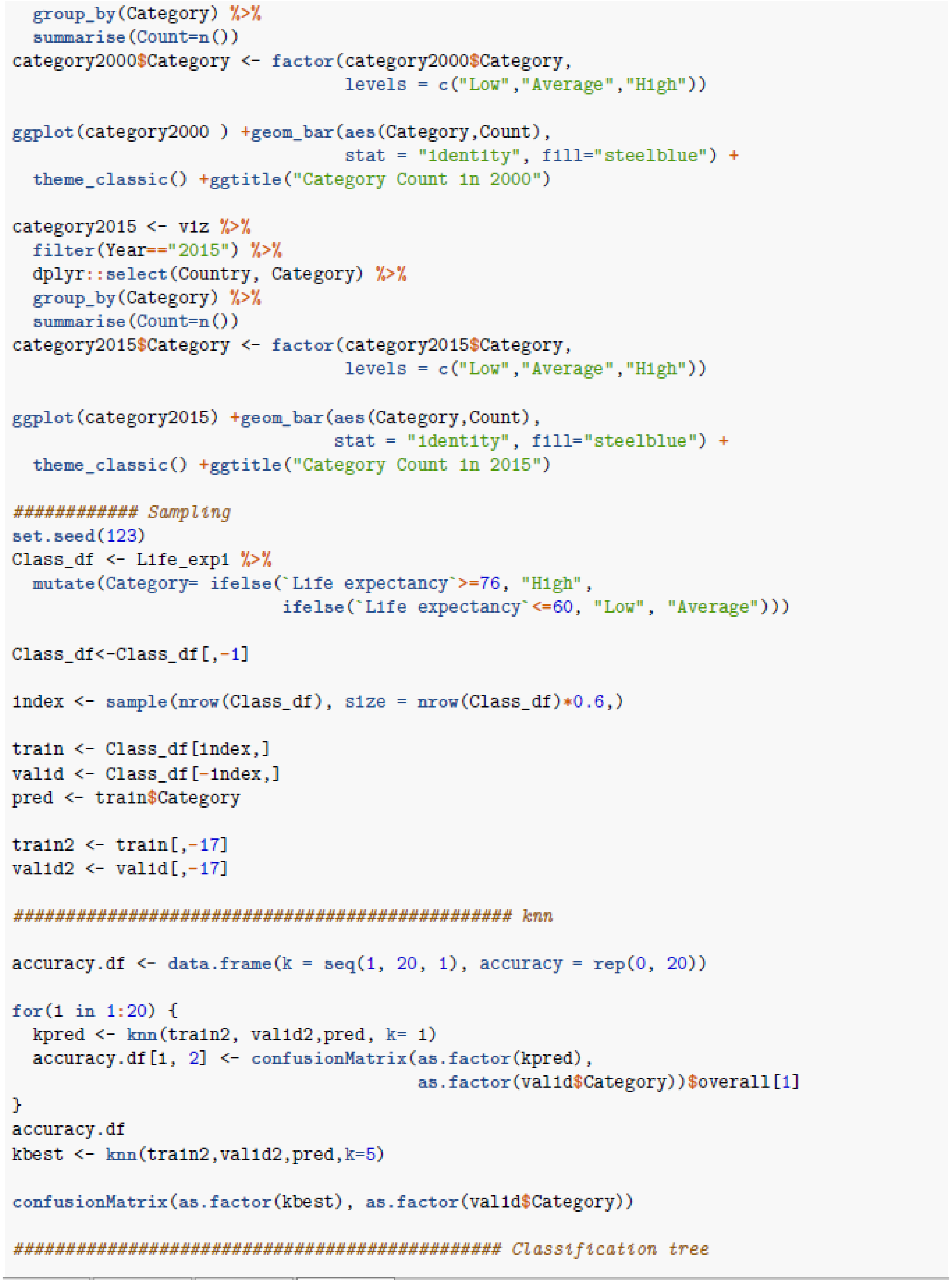
Appendix: R Code for Use Case Study



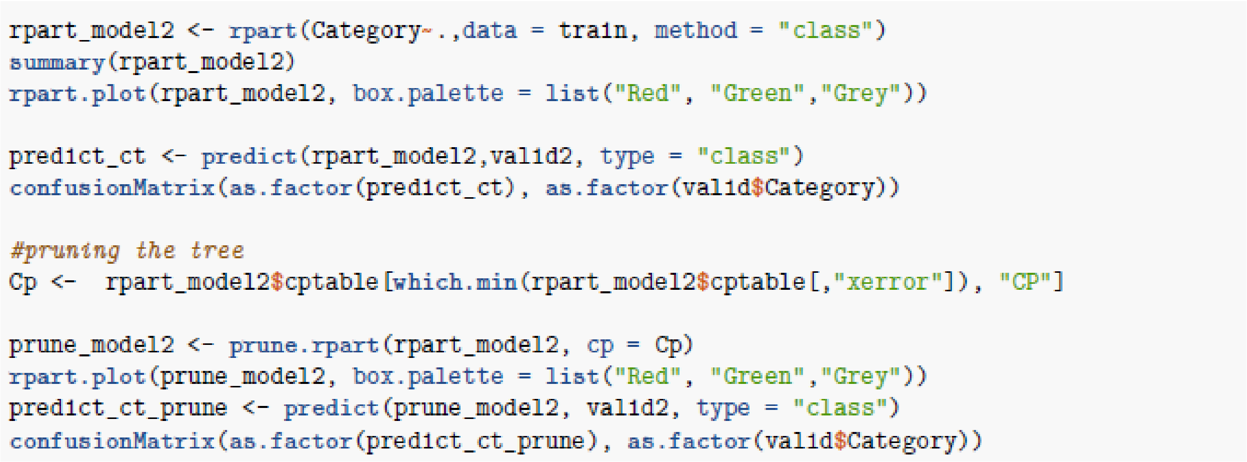
```r
####################################import libraries
library(pls)
library(forecast)
library(DAAG)
library(readr)
library(gains)
library(ggplot2)
library(psych)
library(rpart.plot)
library(rpart)
library(staTools)
library(tidyverse)
library(FNN)
library(caret)
library(usmap)
####################################Creating functions
rmse <- function(error)
{
  sqrt(mean(error^2))
}

standarize <- function(x)
{
  z <- (x-mean(x))/sd(x)
  return(z)
}
############################################load the data
Life_exp <- read_csv("Life Expectancy Data.csv")

######### Remove unwanted columns and null values
Life_exp1 <-  Life_exp[,-c(1:3,17,18)] %>%
  na.omit()

q1 <- Life_exp1
q1 <- as.data.frame(lapply(Life_exp1, standarize))
############################## scatter plot for the complete filtered dataset

pairs.panels(q1)
ggsave("scatter plot.jpeg", plot = pairs.panels(q1), dpi = 900)

#########################Regression############################

############################### Sampling
set.seed(213)
index <- sample(nrow(q1), size = nrow(q1)*0.6,)
training <- q1[index,]
validation <- q1[-index,]

############################## Multiple linear regression
training.lm <- lm(`Life.expectancy`~.,data = training)
options(scipen = 999)
summary(training.lm)
```



```r
predicted <- predict(training.lm, validation)
residual <- validation$`Life.expectancy`-predicted
residual_df <- data.frame("Predicted" = predicted,
                          "Actual" = validation$`Life.expectancy`,
           "Residual" = residual)
head(residual_df,20)


accuracy(predicted, validation$`Life.expectancy`)


############################## Multiple linear regression with PCA

fa.parallel(q1[,-c(1)], fa="pc", n.iter = 100, show.legend = TRUE,
            main = "Scree plot with parallel analysis")

training2.lm <- pcr(`Life.expectancy`~.,data = training,
                    scale= TRUE, validation='CV')
prediction2 <- predict(training2.lm, validation, ncomp = 4)

residual2 <- validation$`Life.expectancy`-prediction2

residual2_df <- data.frame("Predicted" = prediction2,
                           "Actual" = validation$`Life.expectancy`,
           "Residual" = residual2)
head(residual2_df,20)


###################lift chart####################
############################### lift chart for MLR

gain <- gains(validation$`Life.expectancy`[!is.na(predicted)],
              predicted[!is.na(predicted)])
options(scipen=999)
expectancy <- validation$`Life.expectancy`[!is.na(validation$`Life.expectancy`)]

par(pty="s")
plot(c(0,gain$cume.pct.of.total*sum(expectancy))~c(0,gain$cume.obs),
     xlab = "# cases", ylab = "Cumulative Life Expectancy",
     main = "Lift Chart", type = "l")
#baseline
lines(c(0,sum(expectancy))~c(0,dim(validation)[1]), col = "gray", lty = 2)

################################ lift chart MLR with PCA

gain <- gains(validation$`Life.expectancy`[!is.na(prediction2)],
              prediction2[!is.na(prediction2)])
options(scipen=999)
expectancy <- validation$`Life.expectancy`[!is.na(validation$`Life.expectancy`)]

par(pty="s")
plot(c(0,gain$cume.pct.of.total*sum(expectancy))~c(0,gain$cume.obs),
     xlab = "# cases", ylab = "Cumulative Life Expectancy",
     main = "Lift Chart", type = "l")
#baseline
lines(c(0,sum(expectancy))~c(0,dim(validation)[1]), col = "gray", lty = 2)


############################## Random forest RT

rpart_model <- rpart(`Life.expectancy`~., data = training, method = "anova")
rpart.plot(rpart_model) #to plot regression tree
summary(rpart_model) #to identify variable of importance

###boxplot of error values in training and validatoin dataset
pred_train <- predict(rpart_model,training)
pred_valid <- predict(rpart_model,validation)
error_train <- pred_train-training$`Life.expectancy`
error_valid <- pred_valid-validation$`Life.expectancy`
boxplot(error_train, error_valid,names = c("Training", "Validation"))


Cp <-  rpart_model$cptable[which.min(rpart_model$cptable[,"xerror"]), "CP"]

#pruning the tree
prune_model <- prune.rpart(rpart_model, cp = Cp)
rpart.plot(prune_model)

#############################comparison of regression models
####RMSE values
RMSE1 <- rmse(residual) #for MLR
RMSE2 <- rmse(residual2) #for MLR with PCA
RMSE3 <- RMSE(pred_valid, validation$`Life.expectancy`) #For regression tree
RMSE <- c(RMSE1, RMSE2,RMSE3)

###correlation values
cor_rt <- cor(pred_valid, validation$`Life.expectancy`) #for MLR
cor_mlr <- cor(validation$`Life.expectancy`, predicted) # for MLR with PCA
cor_pca <- cor(validation$`Life.expectancy`,prediction2) # For regression tree
Cor <- c(cor_mlr, cor_pca, cor_rt)
comparison <- as.data.frame(cbind(RMSE,Cor),
                            row.names = c("MLR", "MLR with PCA", "Regression Tree"))
colnames(comparison) <- c("RMSE", "Correlation_Value")
comparison

################################ Classification #############################

#############visualization
#barplot representing change in category count in year 2000 vs 2015

viz <- Life_exp %>%
  mutate(Category= ifelse(`Life expectancy`>=76, "High",
                          ifelse(`Life expectancy`<=60, "Low", "Average")))
category2000 <- viz %>%
  filter(Year=="2000") %>%
  dplyr::select(Country, Category) %>%
```



```
  group_by(Category) %>%
  summarise(Count=n())
category2000$Category <- factor(category2000$Category,
                                levels = c("Low","Average","High"))

ggplot(category2000 ) +geom_bar(aes(Category,Count),
                                stat = "identity", fill="steelblue") +
  theme_classic() +ggtitle("Category Count in 2000")

category2015 <- viz %>%
  filter(Year=="2015") %>%
  dplyr::select(Country, Category) %>%
  group_by(Category) %>%
  summarise(Count=n())
category2015$Category <- factor(category2015$Category,
                                levels = c("Low","Average","High"))

ggplot(category2015) +geom_bar(aes(Category,Count),
                               stat = "identity", fill="steelblue") +
  theme_classic() +ggtitle("Category Count in 2015")

############ Sampling
set.seed(123)
Class_df <- Life_exp1 %>%
  mutate(Category= ifelse(`Life expectancy`>=76, "High",
                          ifelse(`Life expectancy`<=60, "Low", "Average")))

Class_df<-Class_df[,-1]

index <- sample(nrow(Class_df), size = nrow(Class_df)*0.6,)

train <- Class_df[index,]
valid <- Class_df[-index,]
pred <- train$Category

train2 <- train[,-17]
valid2 <- valid[,-17]

################################################## knn

accuracy.df <- data.frame(k = seq(1, 20, 1), accuracy = rep(0, 20))

for(i in 1:20) {
  kpred <- knn(train2, valid2,pred, k= i)
  accuracy.df[i, 2] <- confusionMatrix(as.factor(kpred),
                                       as.factor(valid$Category))$overall[1]
}
accuracy.df
kbest <- knn(train2,valid2,pred,k=5)

confusionMatrix(as.factor(kbest), as.factor(valid$Category))

################################################# Classification tree
```



```r
rpart_model2 <- rpart(Category~.,data = train, method = "class")
summary(rpart_model2)
rpart.plot(rpart_model2, box.palette = list("Red", "Green","Grey"))

predict_ct <- predict(rpart_model2,valid2, type = "class")
confusionMatrix(as.factor(predict_ct), as.factor(valid$Category))

#pruning the tree
Cp <-  rpart_model2$cptable[which.min(rpart_model2$cptable[,"xerror"]), "CP"]

prune_model2 <- prune.rpart(rpart_model2, cp = Cp)
rpart.plot(prune_model2, box.palette = list("Red", "Green","Grey"))
predict_ct_prune <- predict(prune_model2, valid2, type = "class")
confusionMatrix(as.factor(predict_ct_prune), as.factor(valid$Category))
```