**MD2201 Data Science**

**Course Project**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S. No** | **Div.** | **Batch No** | **Group No** | **Roll No** | **Gr. No** | **Name of Student** |
| 1 | CS-D | 1 | 5 | 2 | 12110299 | Ripul Handoo |
| 2 | 3 | 12111511 | Rishi Purohit |
| 3 | 7 | 12111529 | Harshal Runwal |
| 4 | 14 | 12110300 | Sanchit Anand |
|  |  |  |  |

**Science**

**Date of performance:**

# Project Title: Mobile Price Classification

# Data Set Name: Mobile Features

# Data set Link: https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification

# Objective: The objective of this mobile price classification project is to develop an accurate and efficient machine learning model that can classify mobile phones into different price ranges based on various features such as hardware specifications, brand, and customer reviews. The aim is to provide valuable insights to businesses and consumers for making informed decisions about mobile phone purchases, and to improve the overall user experience in the mobile industry.

# Project Description: The mobile price classification project is aimed at developing a machine learning model that can classify mobile phones into different price ranges based on various features such as hardware specifications like RAM, Bluetooth, Wi-Fi, mobile weight, touch screen etc. The dataset used for this project had 2000 data points This project is particularly useful for businesses and consumers who are looking for a reliable and efficient way to evaluate the value of mobile phones before making a purchase.

# The project involves collecting data on various mobile phone models from different sources, such as online marketplaces and manufacturer websites. The data is then pre-processed to extract relevant features such as processor speed, screen size, RAM, camera quality, and customer reviews. The pre-processed data is then used to train a machine learning model that can accurately classify mobile phones into different price ranges.

# The machine learning model can be further improved by optimizing various parameters such as the type of algorithm used, the number of features selected, and the size of the training data set. The final model can be deployed as a web-based application or integrated into existing mobile phone review websites to provide consumers with a more comprehensive and accurate evaluation of the value of different mobile phone models.

# Overall, this project has the potential to significantly improve the user experience in the mobile phone industry by providing businesses and consumers with valuable insights that can inform their purchase decisions.

# CODE:

#datavisualization

library(dplyr)

library(tidyr)

library(ggplot2)

library(gridExtra)

library(ggcorrplot)

df <- read.csv("train.csv")

head(df)

summary(df)

#chceking the corelation

corr <- round(cor(df), 8)

print(ggcorrplot(corr))

df$blue <- as.factor(df$blue)

df$dual\_sim <- as.factor(df$dual\_sim)

df$four\_g <- as.factor(df$four\_g)

df$three\_g<- as.factor(df$three\_g)

df$touch\_screen <- as.factor(df$touch\_screen)

df$wifi <- as.factor(df$wifi)

df$price\_range <- as.factor(df$price\_range)

# Bar Chart Subplots

p1 <- ggplot(df, aes(x=blue, fill=blue)) + theme\_bw() + geom\_bar() + ylim(0, 1050) + labs(title = "Bluetooth") + scale\_x\_discrete(labels = c('Not Supported','Supported'))

p2 <- ggplot(df, aes(x=dual\_sim, fill=dual\_sim)) + theme\_bw() + geom\_bar() + ylim(0, 1050) + labs(title = "Dual Sim") + scale\_x\_discrete(labels = c('Not Supported','Supported'))

p3 <- ggplot(df, aes(x=four\_g, fill=four\_g)) + theme\_bw() + geom\_bar() + ylim(0, 1050) + labs(title = "4 G") + scale\_x\_discrete(labels = c('Not Supported','Supported'))

print(grid.arrange(p1, p2, p3, nrow = 1))

print(prop.table(table(df$blue))) # cell percentages

print(prop.table(table(df$dual\_sim))) # cell percentages

print(prop.table(table(df$four\_g))) # cell percentages

# Bar Chart Subplots

p1 <- ggplot(df, aes(x=price\_range, y = battery\_power, color=price\_range)) + geom\_boxplot(outlier.colour="red", outlier.shape=8, outlier.size=4) + labs(title = "Battery Power vs Price Range")

p2 <- ggplot(df, aes(x=price\_range, y = mobile\_wt, color=price\_range)) + geom\_boxplot(outlier.colour="red", outlier.shape=8, outlier.size=4) + labs(title = "Phone Weight vs Price Range")

p3 <- ggplot(df, aes(x=price\_range, y = ram, color=price\_range)) + geom\_boxplot(outlier.colour="red", outlier.shape=8, outlier.size=4) + labs(title = "RAM vs Price Range")

print(grid.arrange(p1, p2, p3, nrow = 1))

data = data.frame(MagaPixels = c(df$fc, df$pc), Camera = rep(c("Front Camera", "Primary Camera"), c(length(df$fc), length(df$pc))))

print(ggplot(data, aes(MagaPixels, fill = Camera)) + geom\_bar(position = 'identity', alpha = .5))

# Checking for Missing values

missing\_values <- df %>% summarize\_all(funs(sum(is.na(.))/n()))

missing\_values <- gather(missing\_values, key="feature", value="missing\_pct")

print(missing\_values %>% ggplot(aes(x=reorder(feature,-missing\_pct),y=missing\_pct)) + geom\_bar(stat="identity",fill="red")+coord\_flip()+theme\_bw())

f <- read.csv("train.csv")

#skewness

library(moments)

skewness(f)

hist(df$clock\_speed, col='red')

#logisticregression

#Importing Libraries And reading The file ----

library(tidyverse)

library(caret)

library(nnet)

library(caTools)

f<-read.csv("train.csv")

head(f)

f$blue <- as.factor(f$blue)

f$dual\_sim <- as.factor(f$dual\_sim)

f$four\_g <- as.factor(f$four\_g)

f$three\_g<- as.factor(f$three\_g)

f$touch\_screen <- as.factor(f$touch\_screen)

f$wifi <- as.factor(f$wifi)

f$price\_range <- as.factor(f$price\_range)

#spliting it into train and test data----

set.seed(123)

split<-sample.split(f,SplitRatio=0.8)

tr\_data<-subset(f,split==TRUE)

ts\_data<-subset(f,split==FALSE)

#Applying Multinomial Logestic Regression----

model <- nnet::multinom(price\_range ~., tr\_data)

#Summarize the model----

print(summary(model))

#prediction part----

predicted <- model %>% predict(ts\_data)

head(predicted)

#ts\_data <- cbind(ts\_data, predicted)

#final part of Confusion Matrix And Accuracy and all----

accu<-mean(predicted == ts\_data$price\_range) #accuracy of the multinomial logistic regression

t1<-table(ts\_data$price\_range, predicted)

confusionMatrix(t1, mode = "everything", positive="1")

#ROC----

predicted <- as.integer(predicted)

tr\_data$price\_range<- as.integer(tr\_data$price\_range)

s1 <- roc(ts\_data$price\_range, predicted)

s1

r <- ggroc(s1,colour="blue")

print(r)

#randomforest

library(randomForest)

library(caTools)

library(caret)

f<-read.csv("train.csv")

head(f)

f$blue <- as.factor(f$blue)

f$dual\_sim <- as.factor(f$dual\_sim)

f$four\_g <- as.factor(f$four\_g)

f$three\_g<- as.factor(f$three\_g)

f$touch\_screen <- as.factor(f$touch\_screen)

f$wifi <- as.factor(f$wifi)

f$price\_range <- as.factor(f$price\_range)

#spliting it into train and test data----

set.seed(123)

split<-sample.split(f,SplitRatio=0.8)

tr\_data<-subset(f,split==TRUE)

ts\_data<-subset(f,split==FALSE)

#Building Random Forest classifier

rf <- randomForest(price\_range ~ .,data=tr\_data, importance=TRUE)

importance(rf)

plot(rf)

#Predicting

pred = predict(rf, ts\_data)

#pred = as.numeric(pred)

cm = table(ts\_data$price\_range, pred)

confusionMatrix(cm, mode = "everything", positive = "1")

#ROC

pred <- as.integer(pred)

tr\_data$price\_range<- as.integer(tr\_data$price\_range)

s4 <- roc(ts\_data$price\_range, pred)

s4

#For Adding Two ROC Curves

#For Adding Two ROC Curves

plot(s1,col="red", main="ROC Curve")

plot(s4,col="black",add=TRUE)

legend(0.3,0.65,c("MLR","RF"),lty=1,col=c("red","black"),title="Graph type", cex = 0.6)

#svm

library(e1071)

f<-read.csv("train.csv")

#print(head(f))

f$blue <- as.factor(f$blue)

f$dual\_sim <- as.factor(f$dual\_sim)

f$four\_g <- as.factor(f$four\_g)

f$three\_g<- as.factor(f$three\_g)

f$touch\_screen <- as.factor(f$touch\_screen)

f$wifi <- as.factor(f$wifi)

f$price\_range <- as.factor(f$price\_range)

#splitting the data into train and test----

set.seed(123)

split<-sample.split(f, SplitRatio == 0.8)

tr\_data<-subset(f,split=TRUE)

ts\_data<-subset(f,split=FALSE)

svmfit <- svm(price\_range~ ., data = tr\_data, kernel = "linear", cost=10, scale = FALSE)

print(svmfit)

pred <- predict(svmfit, ts\_data)

t1<- table(ts\_data$price\_range, pred)

confusionMatrix(t1, mode="everything", positive ="1")

#Roc

pred <- as.integer(pred)

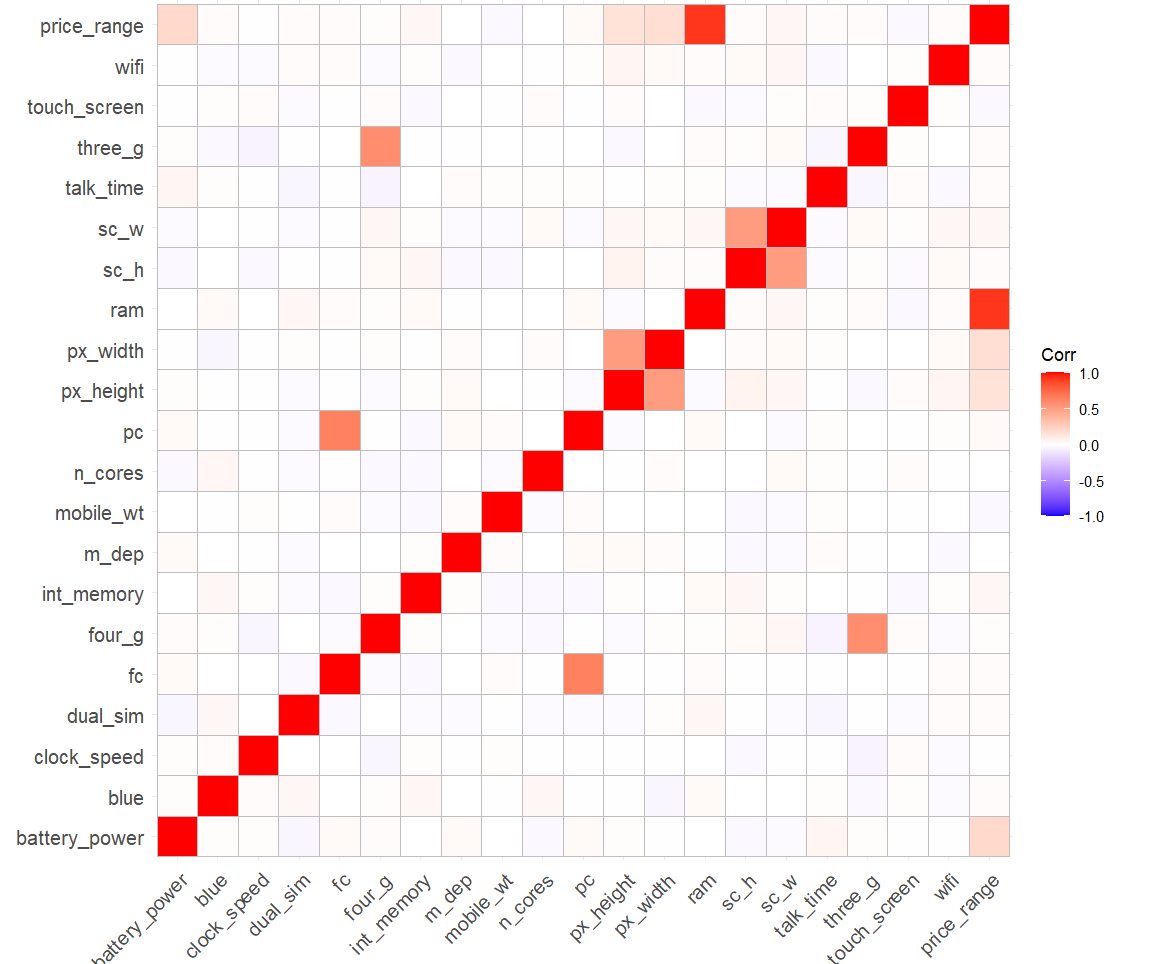
tr\_data$price\_range<- as.integer(tr\_data$price\_range)

s3 <- roc(ts\_data$price\_range, pred)

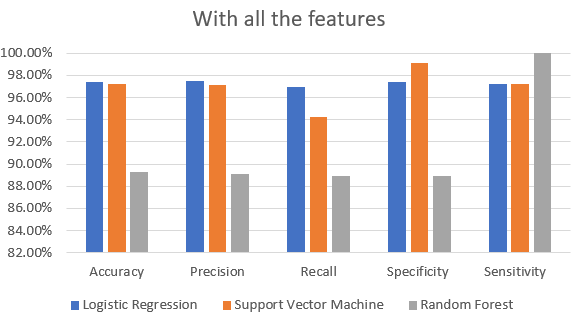
s3

# Results: Quantitative findings and Plots.

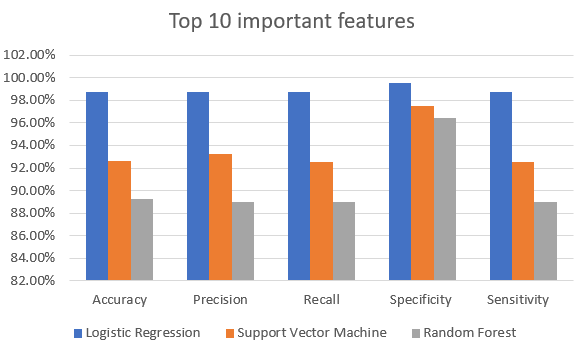
**Plot for relation between the features**

****

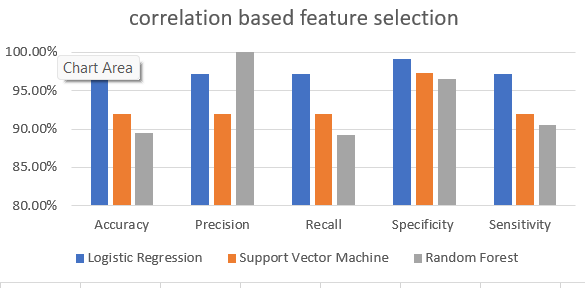
**Plot for comparison of algorithms with all features**

****

**Plot for comparison of algorithms with top 10 features**

****

**Plot for comparison of algorithms correlation-based features**

****

# Conclusions: In conclusion, the mobile price classification project is a valuable application of machine learning techniques in the mobile phone industry. By accurately classifying mobile phones into different price ranges based on various features, the project provides businesses and consumers with a reliable and efficient way to evaluate the value of different mobile phone models. This can improve the overall user experience in the mobile phone industry by ensuring that consumers make informed purchase decisions based on their specific needs and preferences. The project has several potential applications, such as being integrated into existing mobile phone review websites, providing valuable insights to businesses for product development and marketing strategies, and guiding consumers towards more cost-effective and suitable mobile phone purchases. Future work can focus on improving the accuracy and efficiency of the machine learning model, expanding the feature set, and incorporating additional data sources. Overall, this project has the potential to positively impact the mobile phone industry by improving the quality and accessibility of information available to businesses and consumers.