```
install.packages("dplyr")
df_train <- read.csv("/usr/mounika/Desktop/CSP_Project/Dataset/
train.csv")
df_test <- read.csv("/usr/mounika/Desktop/CSP_Project/Dataset/train.csv")
df_val <- read.csv("/usr/mounika/Desktop/CSP_Project/Dataset/train.csv")

spark_write_parquet(df_train, path="/user/rstudio-user",
mode="overwrite")
spark_write_parquet(df_test, path="/user/rstudio-user", mode="overwrite")
spark_write_parquet(df_val, path="/user/rstudio-user", mode="overwrite")

dim(df_train)
dim(df_val)
dim(df_test)

df_train$Class <- factor(df_train$Class)

train  %>%
  select(Class) %>%
  group_by(Class) %>%
  summarise(count = n()) %>%
  glimpse

test %>%
  select(Class) %>%
  group_by(Class) %>%
  summarise(count = n()) %>%
  glimpse

#Logistic Regression model
Logistic_Model=glm(Class~.,test_data,family=binomial())
summary(Logistic_Model)

# build random forest model using every variable
rfModel <- randomForest(Class ~ . , data = train)
test$predicted <- predict(rfModel, test)

library(caret)
confusionMatrix(test$Class, test$predicted)

library(MLmetrics)
F1_all <- F1_Score(test$Class, test$predicted)
F1_all

options(repr.plot.width=5, repr.plot.height=4)
varImpPlot(rfModel,
           sort = T,
           n.var=10,
           main="Top 10 Most Important Variables")

rfModelTrim1 <- randomForest(Class ~  V17,
                             data = train)

test$predictedTrim1 <- predict(rfModelTrim1, test)

F1_1 <- F1_Score(test$Class, test$predictedTrim1)
F1_1
```

```
#trim2
rfModelTrim2 <- randomForest(Class ~  V17 + V12,
                             data = train)

test$predictedTrim2 <- predict(rfModelTrim2, test)

F1_2 <- F1_Score(test$Class, test$predictedTrim2)
F1_2

#trim3
rfModelTrim3 <- randomForest(Class ~  V17 + V12 + V14,
                             data = train)

test$predictedTrim3 <- predict(rfModelTrim3, test)

F1_3 <- F1_Score(test$Class, test$predictedTrim3)
F1_3

#trim4
# four variables
rfModelTrim4 <- randomForest(Class ~  V17 + V12 + V14 + V10,
                             data = train)

test$predictedTrim4 <- predict(rfModelTrim4, test)

F1_4 <- F1_Score(test$Class, test$predictedTrim4)
F1_4

# trim 5
rfModelTrim5 <- randomForest(Class ~  V17 + V12 + V14 + V10 + V16,
                             data = train)

test$predictedTrim5 <- predict(rfModelTrim5, test)

F1_5 <- F1_Score(test$Class, test$predictedTrim5)
F1_5

# ten variables
rfModelTrim10 <- randomForest(Class ~  V17 + V12 + V14 + V10 + V16
                              + V11 + V9 + V4 + V18 + V26,
                              data = train)

test$predictedTrim10 <- predict(rfModelTrim10, test)

F1_10 <- F1_Score(test$Class, test$predictedTrim10)
F1_10

# build dataframe of number of variables and scores
numVariables <- c(1,2,3,4,5,10,17)
F1_Score <- c(F1_1, F1_2, F1_3, F1_4, F1_5, F1_10, F1_all)
variablePerf <- data.frame(numVariables, F1_Score)


# plot score performance against number of variables
options(repr.plot.width=4, repr.plot.height=3)
ggplot(variablePerf, aes(numVariables, F1_Score)) + geom_point() + labs(x
= "Number of Variables", y = "F1 Score", title = "F1 Score Performance")
```

```
rf10 = randomForest(Class ~  V17 + V12 + V14 + V10 + V16
                    + V11 + V9 + V4 + V18 + V26,
                    ntree = 1000,
                    data = train)

options(repr.plot.width=6, repr.plot.height=4)
plot(rf10)
```