PROJECT REPORT

1. PSEUDO CODE:

a) LOGISTIC REGRESSION

```
library(caret)

library(RANN)

library(stats)

library(glmnet)

library(class)

ionosphere <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-
databases/ionosphere/ionosphere.data", header=FALSE)

ionosphere$V35 <- ifelse(ionosphere$V35 == "g", 1, ifelse(ionosphere$V35 == "b", 0, ""))

trainIndex <- sample(1:nrow(ionosphere), 0.8 * nrow(ionosphere))

train <- ionosphere[trainIndex, ]

test <- ionosphere[-trainIndex, ]

train_label <-train[1:280,35]

test_label <-test[1:71,35]

# used package to create 10 folds cross validation

ctrl <- trainControl(method="cv", number=10)

mod_fit <- train(V35~., data=train, method="glm", family="binomial",trControl = ctrl)

pred = predict(mod_fit, test)

confusionMatrix(pred, test$V35)

cm<-table(pred,test[,35])
```

b) K- NEAREST NEGHBOUR

```
library(class)

trainIndex <- sample(1:nrow(ionosphere), 0.8 * nrow(ionosphere))

train <- ionosphere[trainIndex, ]

test <- ionosphere[-trainIndex, ]

train_label <-train[1:280,35]

test_label <-test[1:71,35]

test_pred <- knn(train, test,cl =train_label, k=20)

test_pred

val<-table(test_pred,test_label)

y<-(diag(val))/(sum(val))

Z<-sum(y)

print(Z*100)

library(caret)

# used package to create 10 folds cross validation

ctrl <- trainControl(method="cv", number=10)

mod_fit <- train(V35~V34 + V33 + V32 + V31 + V30 + V29 + V28 + V27 + V26 + V25 + V24 + V23 +
V22 + V21 + V20 + V19 + V18 + V17 + V16 + V15 + V14 + V13 + V12 + V11 + V10+ V9 + V8 + V7 + V6,
data=train, method="glm", family="binomial",trControl = ctrl)

pred = predict(mod_fit, test)

confusionMatrix(data=pred, test_label)

cm<-table(pred,test[,35])

y<-(diag(cm))/(sum(cm))

Z<-sum(y)

Z

library(pROC)

f1 = roc(V35~V34, data=train)

f1

plot(f1, col="red")
```

c) RANDOM FOREST

```
library(class)
library(caret)
library(randomForest)
trainIndex <- sample(1:nrow(ionosphere), 0.8 * nrow(ionosphere))
train <- ionosphere[trainIndex, ]
test <- ionosphere[-trainIndex, ]
train_label <-train[1:280,35]
test_label <-test[1:71,35]
# used package to create 10 folds cross validation
rf_model<-train(V35~.,data=train, method="rf", trControl= trainControl (method="cv",
number=10 ), prox=TRUE, allowParallel=TRUE)
print(rf_model)
pred = predict(rf_model, test)
confusionMatrix(data=pred, test_label)
cm<-table(pred,test_label)
y<-(diag(cm))/(sum(cm))
Z<-sum(y)
print(Z*100)
library(pROC)
f1 = roc(V35~V34, data=train)
f1
plot(f1, col="red")
```

d) BOOSTING

```
## rpart,caret,adabag,ippred library should be loaded
library(rpart)
library(caret)
library(adabag)
library(ipred)

# Load the dataset
Ionosphere <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-
databases/ionosphere/ionosphere.data", header=FALSE)
dataset <- Ionosphere
dataset <- dataset[,-2]
dataset$V1 <- as.numeric(as.character(dataset$V1))
dataset$V35 <- ifelse(dataset$V35 == "g", 1, ifelse(dataset$V35 == "b", 0, ""))
# used package to create 10 folds cross validation
training_set <- trainControl(method="cv", number=10)
Grid <- expand.grid(maxdepth=25,mfinal=10, coeflearn="Zhu")
train_aboost <- train(V35~ ., data=dataset, method = "AdaBoost.M1",
trControl = training_set,tuneGrid=Grid)
ADABOOST_ACCURACY <- 100*(train_aboost$results$Accuracy)
```

e) BAGGING

```
## rpart library should be loaded
library(rpart)
library(caret)
library(adabag)
library(ipred)
# Load the dataset
Ionosphere <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-
databases/ionosphere/ionosphere.data", header=FALSE)
dataset <- Ionosphere
dataset <- dataset[,-2]
dataset$V1 <- as.numeric(as.character(dataset$V1))
dataset$V35 <- ifelse(dataset$V35 == "g", 1, ifelse(dataset$V35 == "b", 0, ""))
# used package to create 10 folds cross validation
training_set <- trainControl(method="cv", number=10)
train_bag <- train(V35~ ., data=dataset, method="treebag",
trControl=training_set)
BAG_ACCURACY <-  100*(train_bag$results$Accuracy)
BAG  ACCURACY
```

VALIDATION TECHNIQUES USED-

1. K-FOLD CROSS VALIDATION
2. AREA UNDER ROC

RESULTS TABLE:

1. Number of instances in dataset: 1
2. Number of attributes in dataset: 34
3. How many fold cross validation performed: Performed 5 cross validations for each classifier with no. of folds between (10 to 300)

CLASSIFIER	TECHNIQUE	ACCURACY	AREA UNDER ROC
Logistic Regression	10-fold Cross Validation	89.57%	0.6338
K-Nearest Neighbor	10-fold Cross Validation	82.28%	0.5414
Bagging	10-fold Cross Validation	91.76%	0.8452
Random Forest	10-fold Cross Validation	93.23%	0.8948
Boosting	10-fold Cross Validation	93.01%	0.8634

ANALYSIS:

From the above results, we can see that Random Forest give highest accuracy as compared to other classifier techniques. Also, boosting gives comparable accuracy estimates. A random forest is an ensemble bagging or averaging method that aims to reduce the variance of individual trees by randomly selecting many trees from the dataset, and averaging them.

Since error is composed from bias and variance, a too complex model has low bias but large variance, while a too simple model has low variance but large bias, both leading a high error but two different reasons. Random forest reduces variance of many "complex" models with low bias. On the other hand, Logistic regression is a pretty well-behaved classification algorithm that can be trained as long the features are roughly linear and the problem is linearly separable. Since our model is not perfectly linear, it does not give good accuracy. In boosting, one is learning from other which in turn boosts the learning. So, the accuracy improves after each learning step. K-nearest neighbor and Logistic Regression does not perform well on this dataset and hence can be classified as weak classifiers.

Also, from our experiments we conclude that number of folds in cross validation influence the accuracy of the models more than other parameters. From our experiments, we find that accuracy (confusion matrix) and cross validation are good evaluation metrics. However, Area under ROC also gives competitive results (between 0 and 1) and hence is classified as a good evaluation metric