MD2201: Data Science

Name of the student: Rajeev Tapadia **Roll No.: 67**

Div: C **Batch: 3**

Date of performance:

Experiment No.6

Title: Classifier Performance

Aim: To measure the different performance parameters of a classifier.

Software used: Programming language R.

Data Set: Wisconsin Breast Cancer data set

Code Statement:

Apply KNN to the Wisconsin Breast Cancer data set. Split the data into training and testing samples. Scale the data and find the following

1. Accuracy
2. **Sensitivity**
3. Specificity
4. **Precision**

Code:

```
library(caret)

my_df<- read.csv("assignment 6\\breast cancer data.csv")
my_df$diagnosis <- factor (my_df$diagnosis, levels = c("B", "M"), labels = c(0,
1))

set.seed(123)

trainIndex <- createDataPartition(my_df$diagnosis, p= .7,
                                  times = 1,
                                  list = FALSE)

data_train <- my_df [ trainIndex,]
data_test <- my_df [-trainIndex,]

preproc_train <- preProcess(data_train[, -1], method = c("scale", "center"))
data_train_scaled <- predict(preproc_train, data_train[, -1])
```

```
data_test_scaled <- predict(preproc_train, data_test[, -1])

k<-5

model <- train (diagnosis ~ ., data = data_train_scaled,
                method = "knn", tuneGrid = data.frame(k = k))

predictions <- predict(model, data_test_scaled)

confusion_matrix <- confusionMatrix(predictions, data_test$diagnosis)
confusion_matrix

confusion_matrix$overall['Accuracy']
confusion_matrix$byClass['Sensitivity']
confusion_matrix$byClass['Specificity']
confusion_matrix$byClass['Precision']
```

Results:

```
> confusion_matrix$overall['Accuracy']
 Accuracy
0.9470588
> confusion_matrix$byClass['Sensitivity']
Sensitivity
  0.9813084
> confusion_matrix$byClass['Specificity']
Specificity
  0.8888889
> confusion_matrix$byClass['Precision']
Precision
   0.9375
>
```

Conclusion:

In this experiment we used the `caret` package to conduct classification on Wisconsin Breast Cancer dataset. Before training the KNN model, the training data is scaled and centered using z-score normalization. The KNN model is trained with k=5 using the training data. Subsequently, predictions are made on the scaled testing data. The performance of the model is evaluated using a confusion matrix, from which metrics like accuracy, sensitivity, specificity, and precision are derived. The reported accuracy of the model on the test

set is followed by metrics such as sensitivity, specificity, and precision, providing insights into the model's ability to correctly classify benign and malignant tumors. Overall, the code successfully demonstrates the application of KNN classification on the breast cancer dataset, with the obtained performance metrics serving as indicators of the model's effectiveness in distinguishing between benign and malignant tumors.