#### **Machine Learning for Data Mining**

**Individual Assignment Report**

**Name: Khettri Bhishma**

Contents

[1. Loading Dataset & Checking Information 3](#_Toc87820442)

[2. Cleaning Dataset 3](#_Toc87820443)

[3. Inspecting Correlation using Heat Map 6](#_Toc87820444)

[4. Exploratory Data Analysis (EDA) 6](#_Toc87820445)

[5. Data Processing 10](#_Toc87820446)

[6. Train Test Split and Model Fitting 10](#_Toc87820447)

[7. All 7 Model Scores using different metrics 11](#_Toc87820448)

[i) Logistic regression 11](#_Toc87820449)

[ii) K nearest neighbors (KNN) 12](#_Toc87820450)

[iii) Naïve Bayes 13](#_Toc87820451)

[iv) Decision Tree 14](#_Toc87820452)

[v) Random Forest 15](#_Toc87820453)

[vi) SVM (Support-Vector Machines) 16](#_Toc87820454)

[vii) Neural Network (using multilayer perceptron) 17](#_Toc87820455)

[8. Final Score evaluation from above fitted model 20](#_Toc87820456)

# Loading Dataset & Checking Information

The dataset used on this assignment was mexico\_covid19.csv. It consists of 11 columns which are 'Age', 'Sex', 'Pneumonia', 'Diabetes', 'Ashma', 'Hypertension', 'CVDs', 'Obesity', 'CKDs', 'Tabacco', 'Result'.

While checking the dataset info in the below figure, all the columns’ data type were float 64 except Pneumonia which d types was object.

Graphical user interface, text

Description automatically generated

# 2. Cleaning Dataset

In the dataset, removed the all the empty row by dropping nan and change the columns data into Integers into Int 64 using the following screenshot codes.

Graphical user interface, text, table

Description automatically generated with medium confidenceGraphical user interface, text, application

Description automatically generated

The dataset shape, columns description and dataset details.

Table

Description automatically generated

While checking all the values of columns, there was extra values in some rows like 99, and 98. From the assignment description the value is only supposed to be 0 and 1 on each column except age so by using the below screenshot code by removing all the unnecessary value.

Graphical user interface, text

Description automatically generated with medium confidence

Checking result description, shape of data frame and if any null value

Graphical user interface, text, application, email

Description automatically generated

3. Inspecting Correlation using Heat Map

From the Heat map, the columns with largest Correlations were 'Result', 'Pneumonia', 'Age' and 'Diabetes'.

Chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

# 4. Exploratory Data Analysis (EDA)

The total patient was 261,409.

Chart, bar chart

Description automatically generated

From analysing the column result, the total patient with negative result were 60.71% and positive result were 38.69.



From analysing the column sex, the females had more negative result than male.

Chart, bar chart

Description automatically generated

From analysing the column Pneumonia, the patient with pneumonia had more positive result than negative result.

Chart, box and whisker chart

Description automatically generated

From analysing the column Diabetes, the patient with diabetes had more positive result than negative result.

Chart, box and whisker chart

Description automatically generated

From analysing the column Ashma, the patient with ashma had more negative result than positive result.

Chart

Description automatically generated

From analysing the column Hypertension, the patient with hypertension had more positive result than negative result.

Chart, box and whisker chart

Description automatically generated

From analysing the column CVDs, the patient with CVDs had more negative result than positive result.

Chart

Description automatically generated

From analysing the column Obesity, the patient with obesity had more positive result than negative result.

Chart, bar chart

Description automatically generated

From analysing the column CKDs, the patient with Pneumonia had more positive result than negative result.

Chart, box and whisker chart

Description automatically generated

From analysing the column Tabacco, the patient with tabacco had more negative result than positive result.

Chart, box and whisker chart

Description automatically generated

# 5. Data Processing

In the data processing the best columns according to correlations were age, pneumonia, diabetes, and results so those columns were only selected.

Graphical user interface, text, application

Description automatically generated

A standard scaler and dummy dataset were created to get the best accuracy score in the model.

A picture containing table

Description automatically generated

# 6. Train Test Split and Model Fitting

The dataset was divided into x and y. Imported train\_test\_split to create a separate train and test sets and the test size was 0.3 and random state was 0.

Graphical user interface, text, application, Word

Description automatically generated

Standard scaler was applied on X\_train and X\_test to improve the accuracy score.



# 7. All 7 Model Scores using different metrics

## Logistic regression

The accuracy score for logistic regression was 64.06%. which is a moderate score.



The other like metrics were also used such as classification report, precision, recall, f-1 score, roc curve and roc\_auc\_score to check the prediction score. All the scores had moderate prediction score.

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## K nearest neighbors (KNN)

The accuracy score for KNN was 58.19% which is a moderate score.



The other like metrics were also used such as classification report, precision, recall, f-1 score, roc curve and roc\_auc\_score to check the prediction score. All the scores had moderate prediction score.

A picture containing table

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, application

Description automatically generated

## Naïve Bayes

The accuracy score for naïve bayes was 63.1% which is a moderate score.



The other like metrics were also used such as classification report, precision, recall, f-1 score, roc curve and roc\_auc\_score to check the prediction score. All the scores had moderate prediction score.

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, application

Description automatically generated

## Decision Tree

The accuracy score for decision tree was 64.35% which is a moderate score.



The other like metrics were also used such as classification report, precision, recall, f-1 score, roc curve and roc\_auc\_score to check the prediction score. All the scores had moderate prediction score.

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, application

Description automatically generated

## Random Forest

The accuracy score for random forest was 64.36% which is a moderate score.



The other like metrics were also used such as classification report, precision, recall, f-1 score, roc curve and roc\_auc\_score to check the prediction score. All the scores had moderate prediction score.

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

A picture containing graphical user interface

Description automatically generated

## SVM (Support-Vector Machines)

The accuracy score for SVM was 64.1% which is a moderate score.



The other like metrics were also used such as classification report, precision, recall, f-1 score, roc curve and roc\_auc\_score to check the prediction score. All the scores had moderate prediction score.

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

A picture containing graphical user interface

Description automatically generated

## Neural Network (using multilayer perceptron)

For the neural network, the first, second, and the output layer was utilized under sequential.

Table

Description automatically generated

10 epochs were used to train the model

Table

Description automatically generated

The accuracy score for SVM was 64.1% which is a moderate score.



The confusion matrix predicted score on figure.

Chart, treemap chart

Description automatically generated

The other like metrics were also used such as classification report, precision, recall, f-1 score, roc curve and roc\_auc\_score to check the prediction score. All the scores had moderate prediction score.

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, application

Description automatically generated

# 8. Final Score evaluation from above fitted model

These are the list of every model algorithm accuracy score:

Text

Description automatically generated

According to the bar chart below, the highest prediction score is from Random Forest which score was 64.36% and the lowest prediction score is from K-Nearest Neighbors which score was 58.19%.

Chart, bar chart

Description automatically generated