Final Project Report

On

Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery

Submitted By

Nitin Sharma (202IT017)

Praful Kumar (202IT020)

Under Supervision of

Dr. Nagamma Patil

Department of Information Technology

National Institute of Technology, Surathkal

ACKNOWLEDGEMENT

We would like to express our deep gratitude to **Dr Nagamma Patil**, our research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to thank her for her advice and assistance in keeping our progress on schedule.

We would also like to extend our thanks to our batch mates for pitching in with their suggestions and finally we wish to thank our respective parents for their support and encouragement throughout our study.

Nitin Sharma(202IT017) **Praful Kumar(202IT020)**

DECLARATION

We, **Nitin Sharma (202IT017)** and **Praful Kumar (202IT020)**, enrolled in M.Tech program of **Information Technology** department in **NIT Surathkal** hereby declare that we own the full responsibility for the information and results provided in the project titled **"Predictive Data Mining Model for Novel Coronavirus Patient Recovery"** under guidance of **Dr. Nagamma Patil** for the partial fulfillment of Data Mining course.

We have taken care in all respect to honor the intellectual property right and have acknowledged the contribution of others for using them in academic purpose and further declare that in case of any violation of intellectual property right or copyright we as candidates, will be fully responsible for the same. My supervisor should not be held responsible for full or partial violation of copyright or intellectual property rights.

Nitin

Nitin Sharma (202IT017)

IT, M-Tech 2020-22

Praful

Praful Kumar (202IT020)

IT, M-Tech 2020-22

TABLE OF CONTENTS

Contents **PageNo.**

LIST OF FIGURES

Figures **PageNo.**

LIST OF TABLES

Tables **PageNo.**

INTRODUCTION

Recently, the whole world has been drastically affected by Coronavirus (COVID-19) which had brought the world to a stand still. Even though there is still time for COVID vaccines to come out, still we need to put efforts into better understanding the natures of this virus to decrease the fatality rate and data modelling of such diseases can be very helpful in understanding their impact in future.

Statistical models available at our disposal, and they are good at understanding the relation between variables. The major difference between machine learning and statistics is their purpose. Machine learning models are designed to make the most accurate predictions possible. Statistical models are designed for inference about the relationships between variables. Many statistical models can make predictions, but predictive accuracy is not their strength.

In our project we are going to develop data mining models for predicting recovery of COVID-19 affected patients of South Korea. We are going to apply K-Nearest Neighbor, Decision Tree, Random Forest, Naive Bayes, Logistic Regression, Support Vector Machine(SVM) and Artificial Neural Network(ANN) on the KCDC dataset that is available on Kaggle and we will be using python as our programming language and google colaboratory as out working environment.

The result of our work shows that Random forest prediction model is the best suited for our requirements and achieved an accuracy of 99.39%. Accuracies of rest of the model lied in between 97% to 99%.

Keywords : *KNN, Naive Bayes, Random Forest, ANN, SVM, COVID, Coronavirus, Patient, Recovery.*

LITERATURE SURVEY

2.1 Related Work

Covid-19 research has been a hot topic in recent times and in this section we look at some notable contribution and work related in this field.

- **[2016] : Al-Turaiki I, Alshahrani M, Almutairi T.** build predictive models for MERS-CoV infections using data mining techniques. Accuracy of the models was between 53.6% and 71.58%.
- **[2020] : L. J. Muhammad, Md. Milon Islam, Sani Sharif Usman, Safial Islam Ayon** build predictive models using Data Mining Techniques. Accuracy of 97% to 99% using different models.

Data mining has been widely used for the prog-nosis and diagnosis of many diseases. Some other contribution of people in medical field using Data Mining models are:

- **[2012] : Ferreiraet al** used data mining to improve the diagnosis of neonatal jaundice in newborns. They applied many data mining model and finally came up with the result that the best predictive models were obtained by using Naive Bayes, multilayer perceptron, and simple logistic and achieved highest accuracies of 89%.
- **[2017] : Prof. Mamta Sharma1,Farheen Khan2, Vishnupriya Ravichandran** did a comparative study on different data mining techniques for Heart Disease prediction and achieved maximum accuracy of 99.25% with Neural Network.

2.2 Motivation

Covid-19 has been hard on people around the globe. It bought economies down, made the whole world come to a stand still and forced people to hide behind their doors. With no vaccines out as a permanent solution to this, we can still strive for understanding this disease better. Taking this as a motivation, we, through our project work, are going to build some data mining driven models to help us understand the condition of people affected by covid and chances of them getting released from hospital, i.e. achieving full recovery.

2.3 PROBLEM STATEMENT

GOAL is to build different Data Mining driven data models that help us predict whether a person diagnosed by covid is going to get fully recovered, and then perform a comparative study on these models to come up with the best suited model for our problem statement.

2.4 OBJECTIVES

A) Understanding the data-set first and then after performing pre-processing on the data, build classification models and compare their accuracies.

B) Develop following data mining models :

- K-Nearest Network (KNN)
- Decision Tree
- Logistic Regression
- Naive Bayes
- Support Vector Machine (SVM)
- Random Forest
- Artificial Neural Network (ANN)

C) Use Grid Search to perform hyper-parameter tuning and trial & error in the case of neural network to achieve as high accuracy as we can.

PROPOSED METHODOLOGY

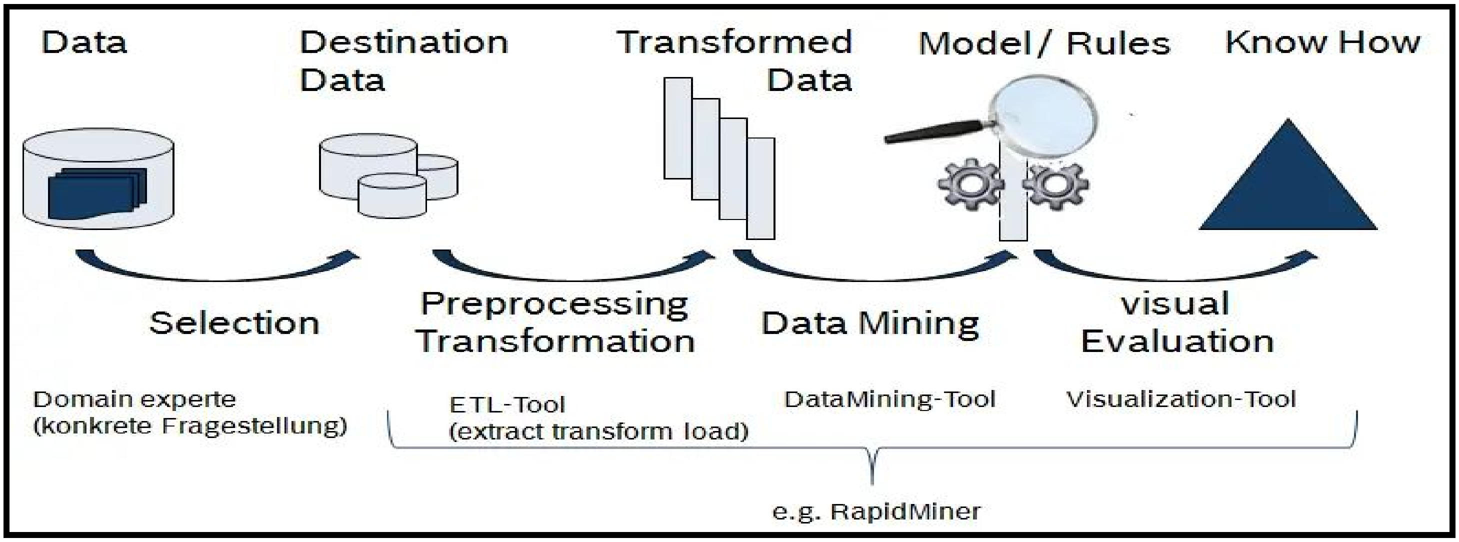


Figure 3.1

Dataset and Working Environment

The dataset has been obtained from Kaggle and made available by KCDC. It is an epidemiological dataset of COVID-19 patients in South Korea. This dataset consist of total 5164 records and 14 columns which include : patient_id, age, sex, country, infection_case, state (deceased/released/isolated), etc.

As for the working environment, we will be working on google collaboratory and using python as a programming language.

Data Preprocessing

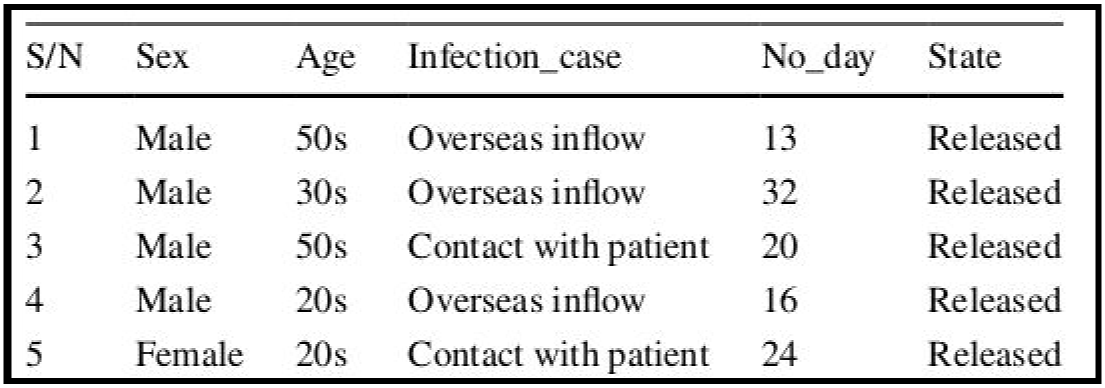
Original dataset from kaggle consists of 5165 tuples and 14 columns out of which number of relevant columns that are considered for the prediction purpose are 7. We have dropped records which had isolation in state as we are only considering released and deceased states of patient. We also dropped records with missing confirmed_date, released_date or deceased_date. Categorical encoding was performed on column ‘infection_case’.

For records with missing values in sex, age and infection_case we used last observation carried forward imputation technique to fill in the null values. We created a new column no_of_days by using following criteria :

- (released_date - confirm_date) in case of released patients

- (deceased_date - confirm_date) in case of deceased patients

At the end of pre-processing, the resultant dataset that we had consisted of 1646 tuples and 8 columns and this dataset was further divided into training and testing set with 80% training dataset and 20% going to testing. Sample dataset after pre processing :



S/N	Sex	Age	Infection_case	No_day	State
1	Male	50s	Overseas inflow	13	Released
2	Male	30s	Overseas inflow	32	Released
3	Male	50s	Contact with patient	20	Released
4	Male	20s	Overseas inflow	16	Released
5	Female	20s	Contact with patient	24	Released

Table 3.1

Data Distribution Via Parameters

- Gender wise distribution :

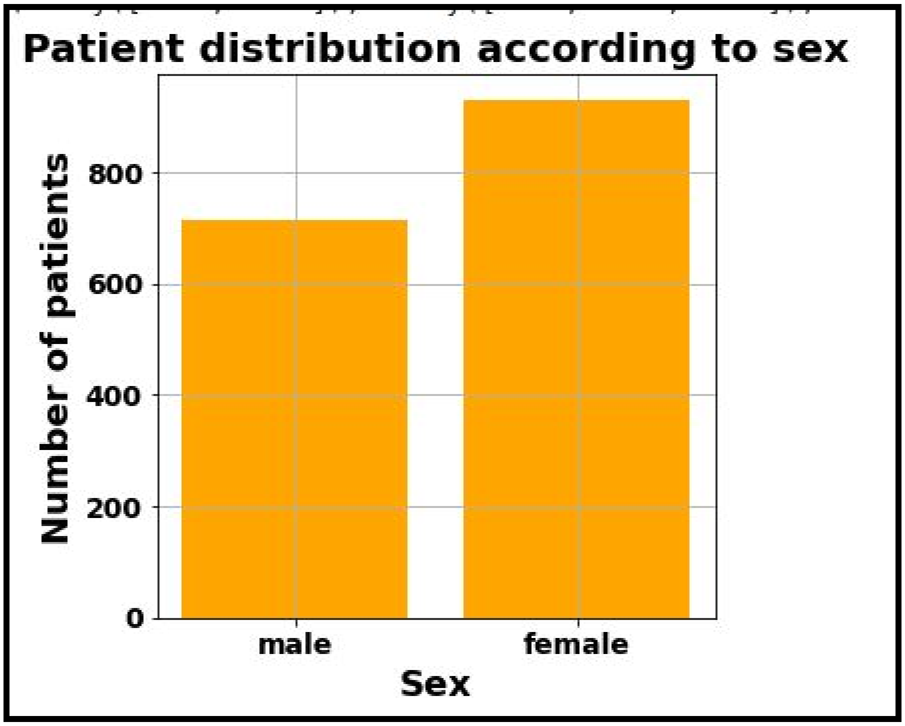


Figure 3.2

- Patient ‘state’ wise distribution of data :

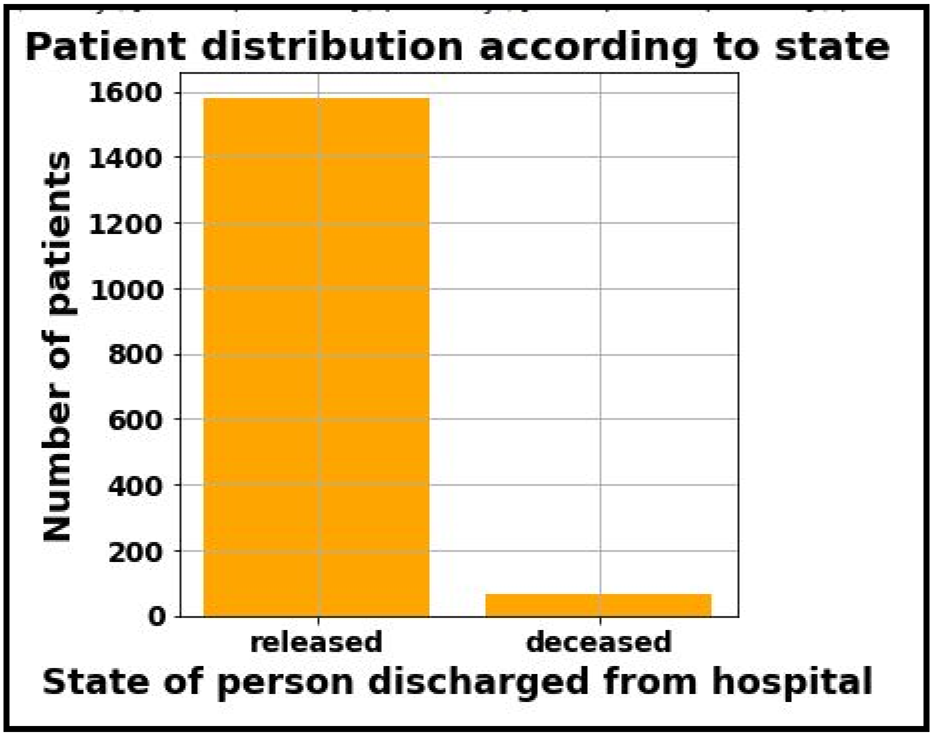


Figure 3.3

- ‘Age’ wise distribution of data

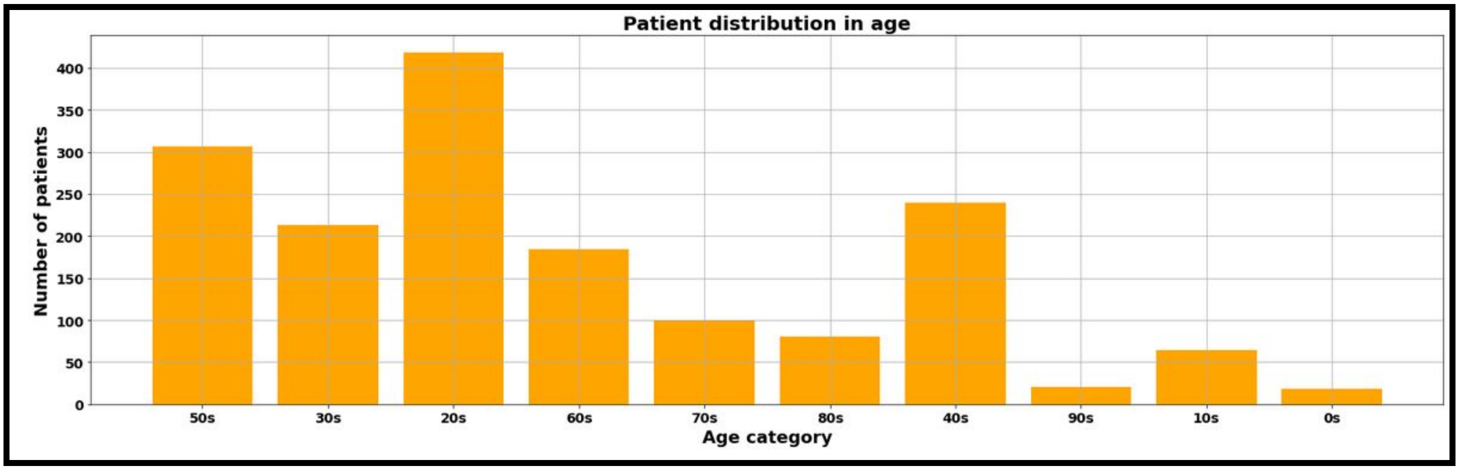


Figure 3.4

- ‘Infection_case’ wise distribution of data

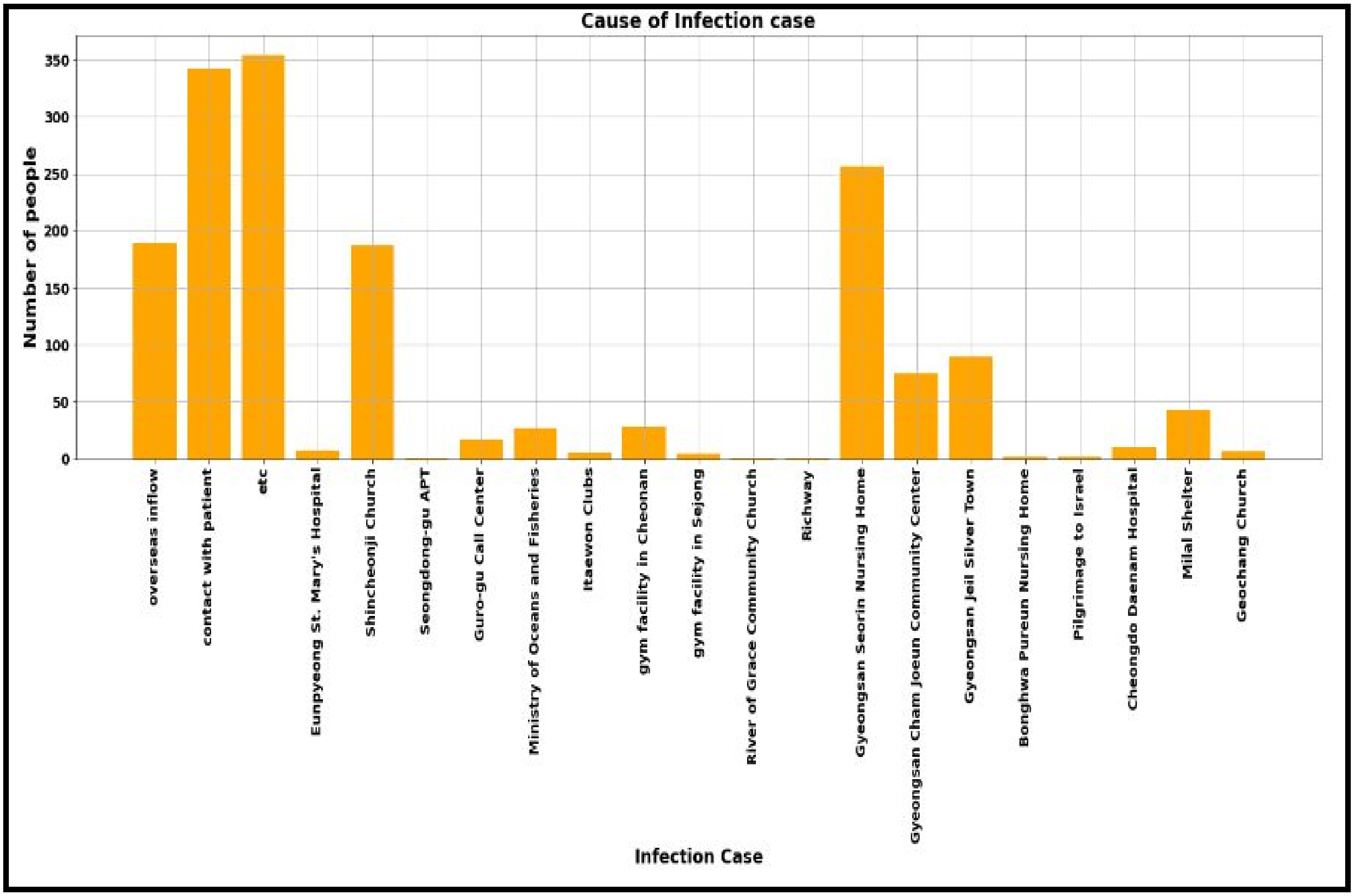


Figure 3.5

Different Techniques and Data Models

- **K-Nearest Network (KNN):** it is a simple, easy to implement supervised machine learning algorithm that can be used to solve both regression and classification problems. It uses some similarity measures like distance between the dataset points in order to solve the problems. The ideology behind this algorithm is that similar points lie closely to each other in data distribution space.
- **Decision Tree:** it is a machine learning algorithm that can be used for both classification as well as regression. It uses a tree-like model of decisions in order to represent decisions and decision making. For splitting of the tree we use 'gini' and 'entropy' criterion. It is also the base for Random forest classification algorithm as that algorithm uses many trees in the forest with the name as 'number of estimators'.
- **Logistic Regression:** it is an algorithm used to determine the association between the

categorical dependent variables against the independent variables. This is generally used when the dependent variable has binary values such that 0 or 1, True or False, Yes or No, etc.

- **Naive Bayes:** It is a classification technique based on the Bayes's theorem with an assumption that the presence of a particular feature in a class is unrelated to the presence of any other feature. This model is easy to build and very useful especially in the case of large datasets and also known to highly outperform the other sophisticated models.
- **Support Vector Machine (SVM):** this is an algorithm used for both the classification as well as regression tasks. This is based on the concept that the dataset points present can be easily separated with the help of a hyperplane such that both the classes of dataset points have their points at an equal distance from the hyperplane. This is highly useful especially due to its nature of giving high accuracy with lower computation power.
- **Random Forest:** this machine learning algorithm can be used for both classification and regression purposes. This algorithm uses decision trees as its building base as it contains a large number of decision trees and a large group of decision trees can act as an ensemble. Each decision tree predicts the class prediction and the class with the most number of votes is the output.
- **Artificial Neural Network (ANN):** This is a neural network that is inspired by the human brain and the way it learns new information and does prediction. This contains interconnected layers of neurons that help the model to train on a given dataset. Based on its style of learning it can be Feedback neural network or Feedforward neural network.

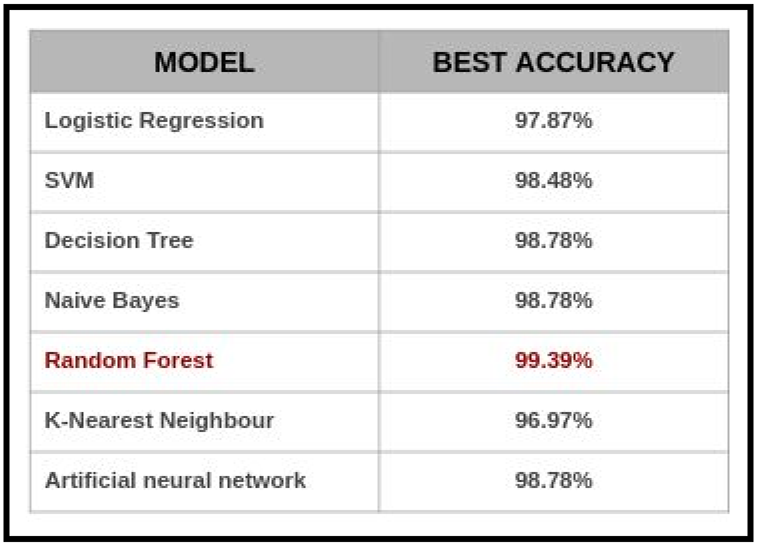
RESULT AND ANALYSIS

Performance Evaluation of Models

We used accuracies as a parameter to evaluate performance of each model. Accuracy determines the percentage of the dataset instances correctly classified for the model developed by the data mining algorithm. Expressed as:

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fn} + \text{fp}},$$

Some other performance evaluation techniques for the data mining model include specificity, sensitivity, but we will be only focussing on accuracy. Here is the accuracy obtained for each :



MODEL	BEST ACCURACY
Logistic Regression	97.87%
SVM	98.48%
Decision Tree	98.78%
Naive Bayes	98.78%
Random Forest	99.39%
K-Nearest Neighbour	96.97%
Artificial neural network	98.78%

Table 4.1

The result of our work shows that the Random **forest** prediction model is the best suited for our requirements and achieved an **accuracy of 99.39%**. Accuracies of the rest of the models lied in between 97% to 99%.

Pictorial Representation using Confusion Matrix

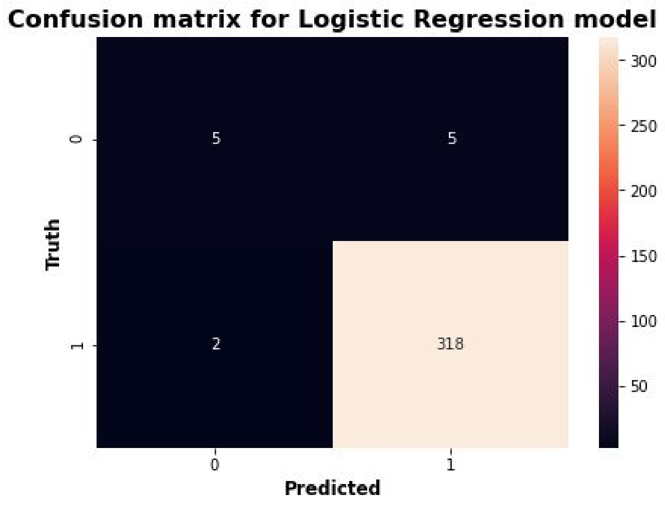


Figure 4.1

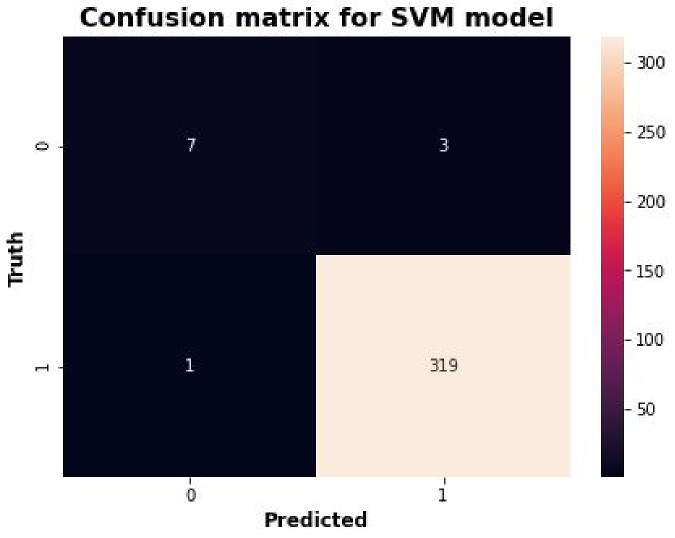


Figure 4.2

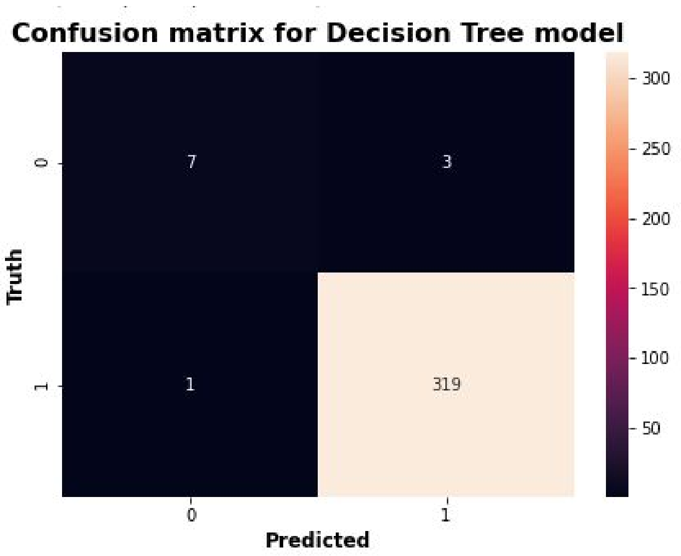


Figure 4.3

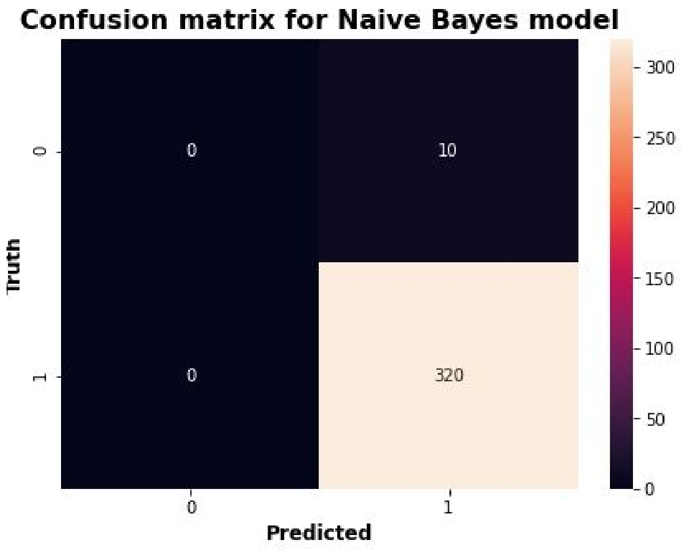


Figure 4.4

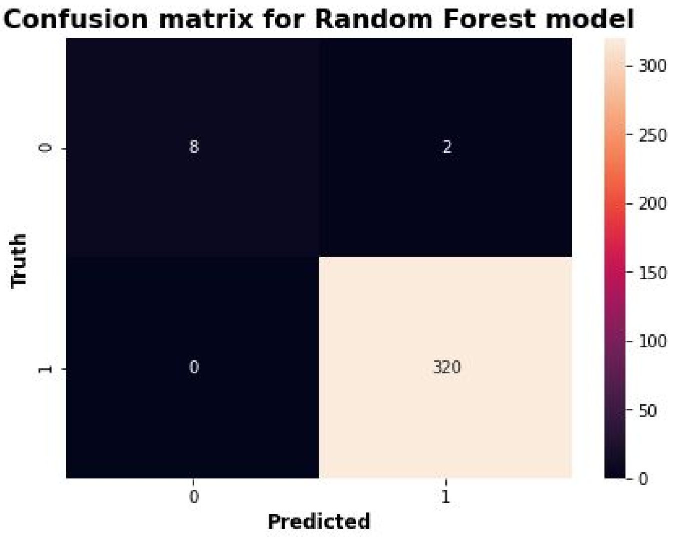


Figure 4.5

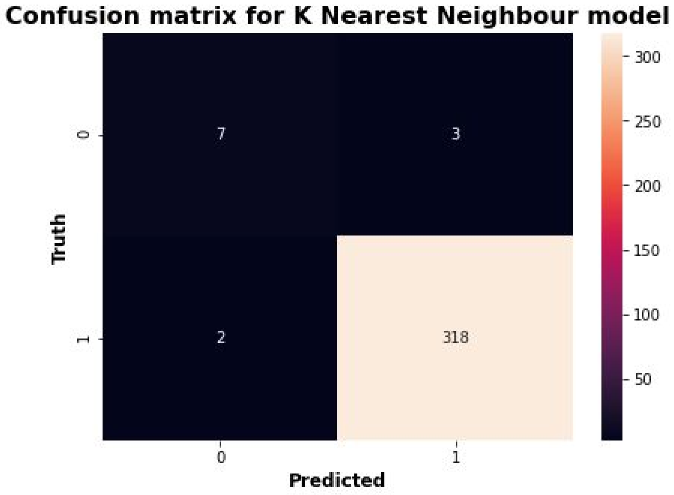


Figure 4.6

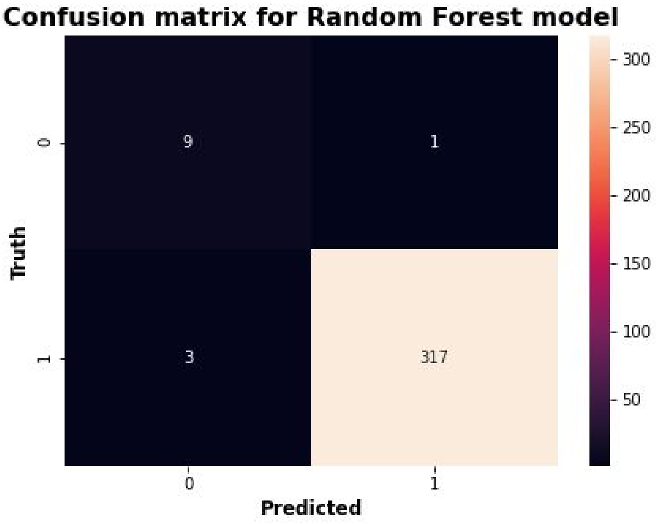
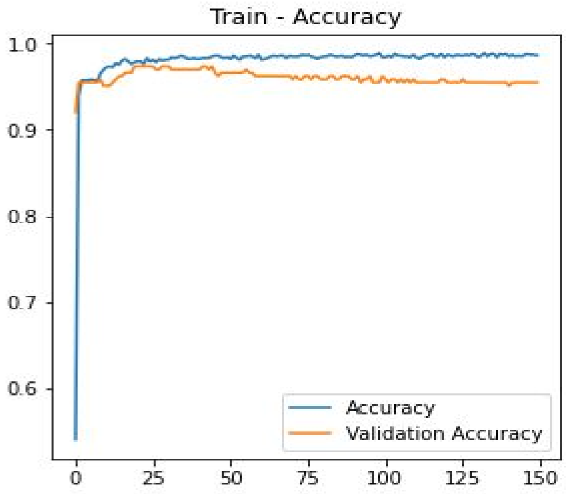


Figure 4.7



FOR NEURAL NETWORK

Figure 4.8

CONCLUSION AND FUTURE WORK

We can conclude that for the available dataset best test accuracy was achieved in the case of Random forest method which was 99.39% and worst accuracy was achieved in the case of K-nearest neighbour and it was 96.97%. For future work we can include the X-Ray image dataset of covid-19 patients and predict better results by using the deep learning techniques like CNN. Also, more care can be taken while noting down the data at the base level(here hospital staff) so that data pre-processing becomes more efficient and we have to discard lesser tuples.

REFERENCES

- Al-Turaiki I, Alshahrani M, Almutairi T paper on MERS Cov : https://pubmed.ncbi.nlm.nih.gov/27641481/

- L. J. Muhammad, Md. Milon Islam, Sani Sharif Usman, Safial Islam Ayon'spredictive models using Data Mining Techniques : https://link.springer.com/article/10.1007/s42979-020-00216-w

- Coronavirus dataset of Korea Centers for Disease Control & Prevention (KCDC) : https://www.kaggle.com/kimjihoo/coronavirusdataset/data