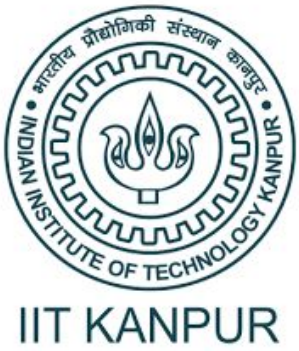


IIT KANPUR

IME 672A: DATA MINING AND KNOWLEDGE DISCOVERY

PROJECT TITLE: Find whether a person is covid positive or negative

GROUP #23

1.	Divyansh Tripathi	170257
2.	Hemraj Meena	170299
3.	Parimal Mukul	170460
4.	Parth Pandey	170462
5.	Piyush Sunil Samarth	170472

CONTENT:

- Problem description
- Data Understanding
- Data Preprocessing
- Modeling
- Results and Interpretation

PROBLEM DESCRIPTION

Based on the data set given we have to train a model to predict whether a person with a given set of attributes is covid positive or negative.

The training data obtained is from the Mexican government and hence, the analysis is valid for Mexico or maybe North America. The pandemic stats and behaviors are extremely different for Asian countries when compared to North American or European countries owing to a far lower case fatality rate for Asia.

DATA UNDERSTANDING

1. The data contains 566602 rows and 23 columns(attributes). Different patient attribute provided are {id, sex, patient_type, entry_date, date_symptoms, date_died, intubed, pneumonia, age, pregnancy, diabetes, copd, asthma, inmsupr, hypertension, other_disease, cardiovascular, obesity, renal_chronic, tobacco, contact_other_covid, covid_res, icu}.
2. These attributes present the medical history of the patients and statistics related to patient id, sex, the date on which the patient was admitted to the hospital, the date when covid symptoms manifested in the patient, the date on which the patient died covid contact, covid test result, etc.
3. These attributes are of type, non-ordered categorical and discrete attributes.

DATA PREPROCESSING

In data preprocessing, firstly we checked for different unique values taken by different attributes. The following unique values corresponding to different attributes were found.

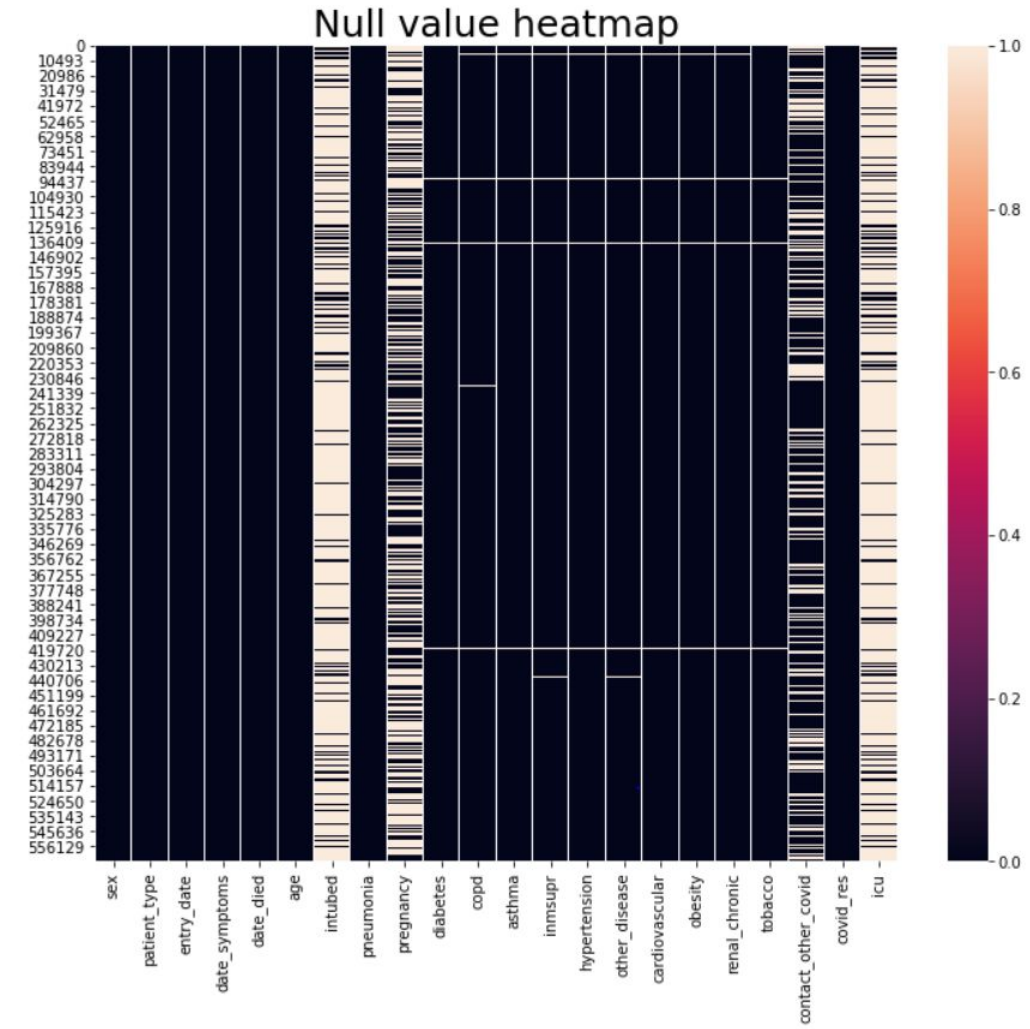
sex column - [2 1]	hypertension column - [2 1 98]
patient_type column - [1 2]	other_disease column - [2 1 98]
intubed column - [97 2 1 99]	cardiovascular column - [2 1 98]
pneumonia column - [2 1 99]	obesity column - [2 1 98]
pregnancy column - [97 2 1 98]	renal_chronic column - [2 1 98]
diabetes column - [2 1 98]	tobacco column - [2 1 98]
copd column - [2 1 98]	contact_other_covid column - [2 99 1]
asthma column - [2 1 98]	covid_res column - [1 2 3]
inmsupr column - [2 1 98]	icu column - [97 2 1 99]

As presented in the above table, the values 97, 98, 99 indicate that the data is not available for these cells. In the further process, these values are replaced with NaN.
These attributes are categorical so we convert them into different categories as follows:
(a) Sex *{Female 1; Male 2}*
(b) Patient_type *{Outpatient 1; Inpatient 2}*
(c) Columns with preconditions like pneumonia, pregnancy, diabetes, copd, asthma, inmsupr, hypertension, other_disease, cardiovascular, obesity, renal_chronic, tobacco, contact_other_covid, icu. *{Yes 1; No 2}*
(d) Covid result *{Positive 1; Negative 2; Results Awaited 3}*

To check how skewed the data provided is, we find the following, the number of covid positive results = 220657, and the number of covid negative results = 279035, and the number of covid awaited results = 66910.

Based on this we can say that the data is not skewed, as the number of covid positive results and number of covid negative results are very close.

Next, we check for the NULL values in the data set and find that no NULL value is present in the data set.



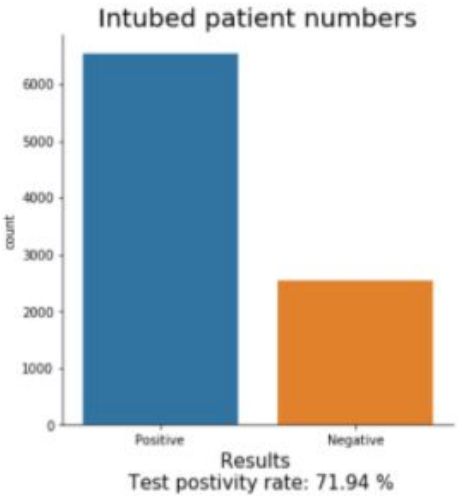
A large number of null values can be seen in intubed, pregnancy, contact_other_covid and icu attributes.

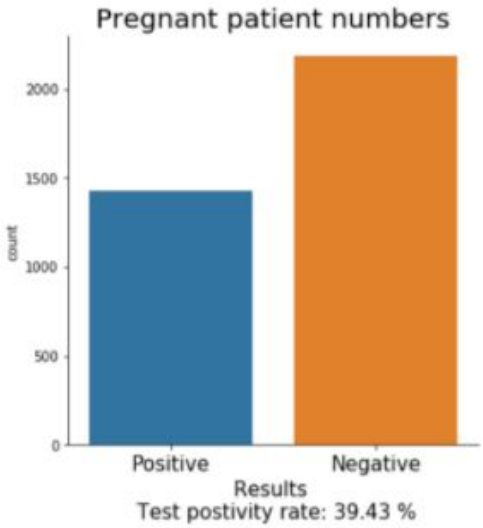
Hypothesis:

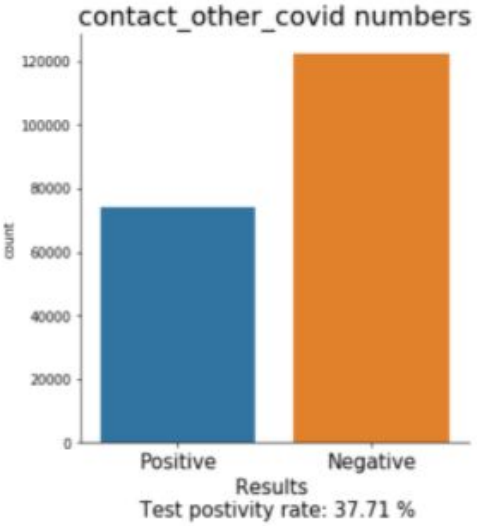
Here, we have to predict whether a person is Covid +ve or -ve based on the preconditions. So, we have proposed the following points in our hypothesis and attempt to verify them

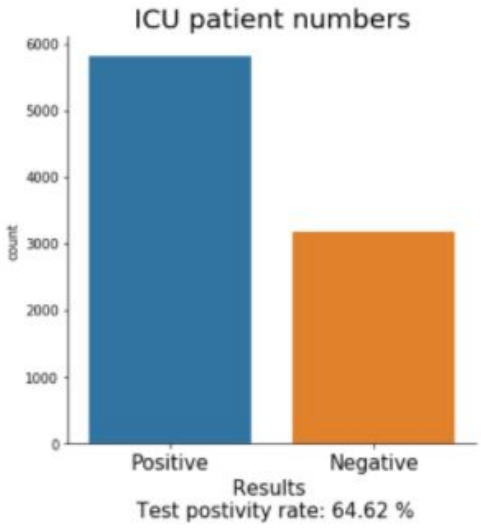
1. The fatality or the 'date_died' column has no role to play in determining Covid +ve or -ve. So we can drop that column.
2. We know that persons suffering from underlying health conditions especially respiratory infection have high chances of being positive. So, we check the preconditions like pneumonia and asthma and whether the person is intubated.
3. So, we attempt to check how these and different other features contribute to a person being positive or negative. Our main focus will be on the cases with positive or negative results, therefore we'll neglect awaiting results.

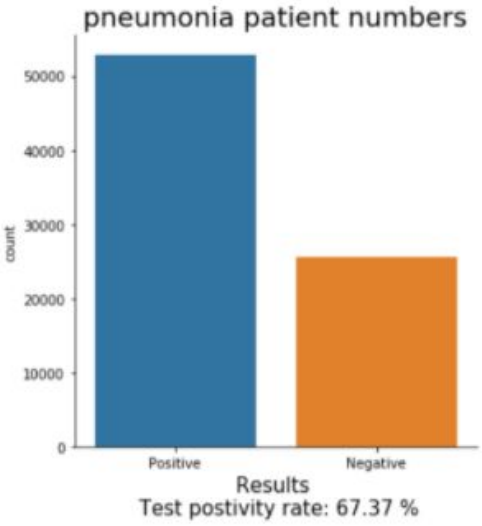
Now by applying pre-processing techniques, we observed the relationship of **Covid19 Positivity** with **all the categorical variables** as shown below. (Eg. Corresponding to attribute *'intubated'* from the set of covid positive results what is the count of intubated patients and not intubated patients are plotted below)

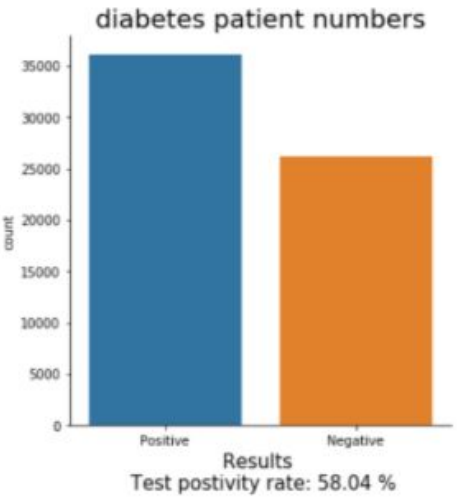


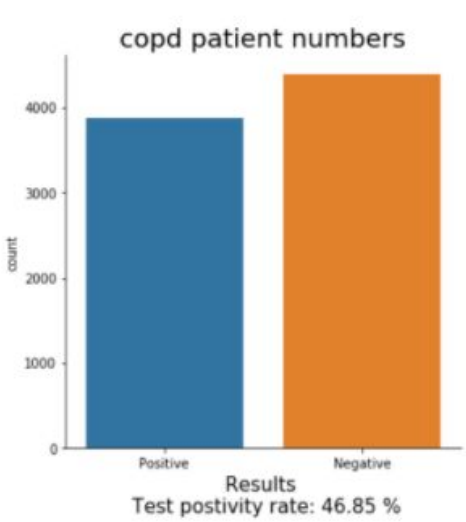


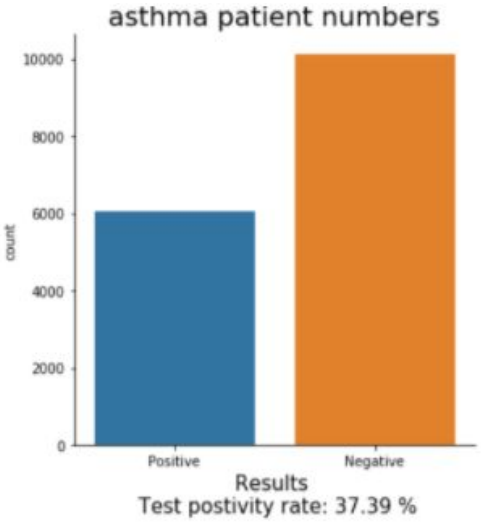


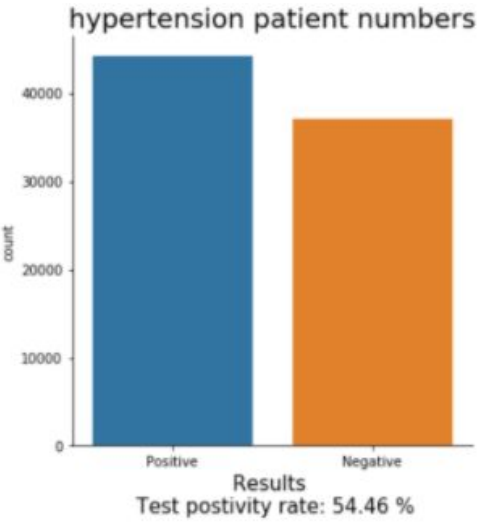


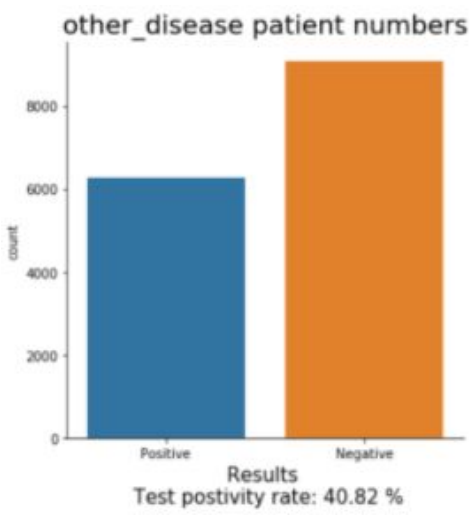


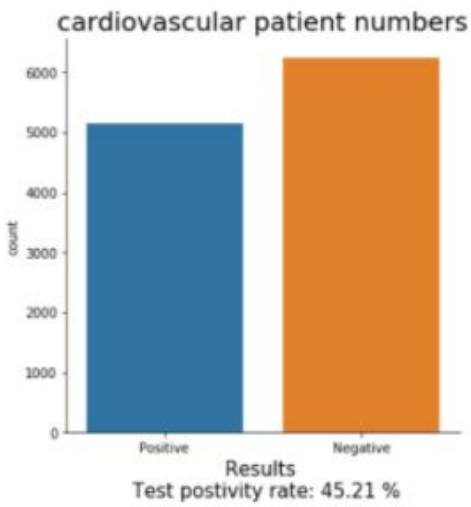


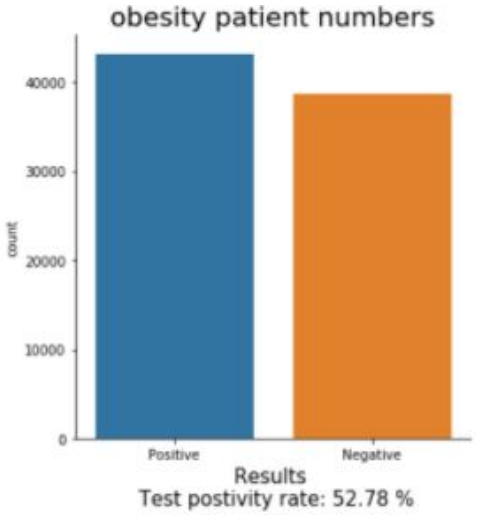


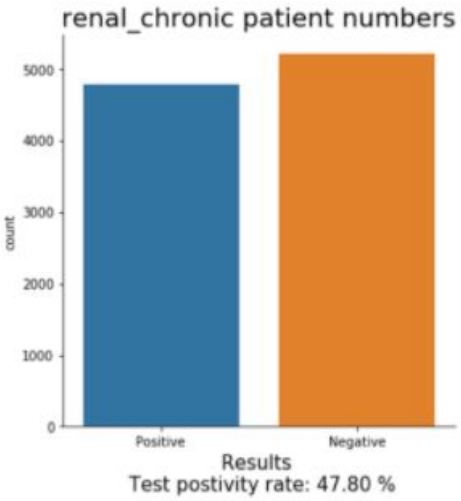


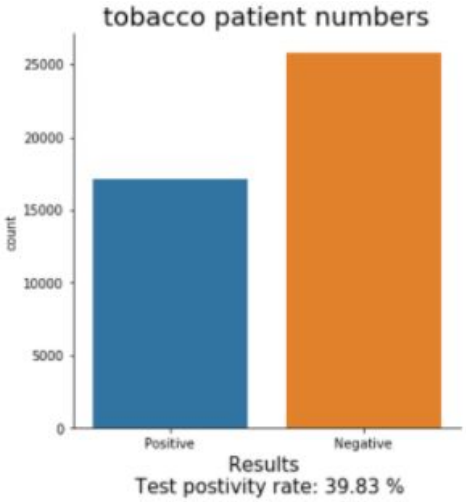


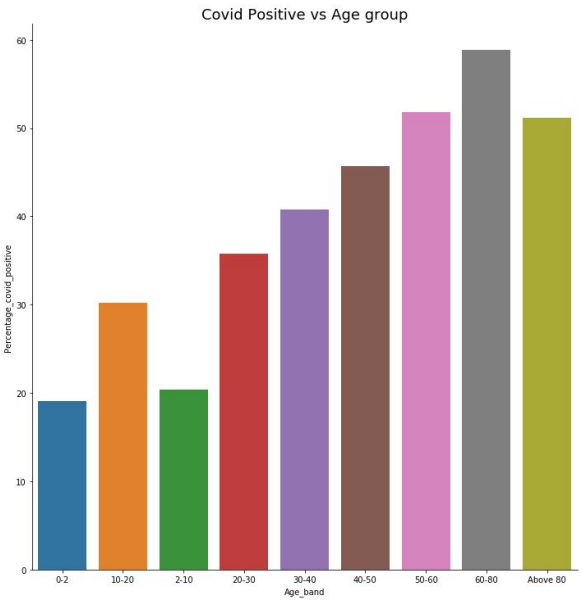

other_disease patient numbers
count
Positive
Negative
Results
Test postivity rate: 40.82 %

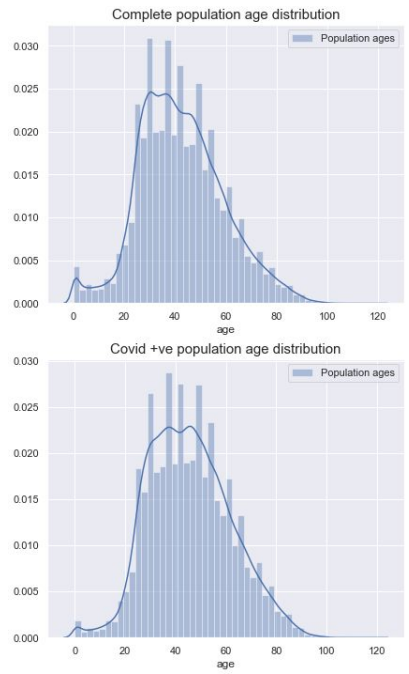

cardiovascular patient numbers
count
Positive
Negative
Results
Test postivity rate: 45.21 %

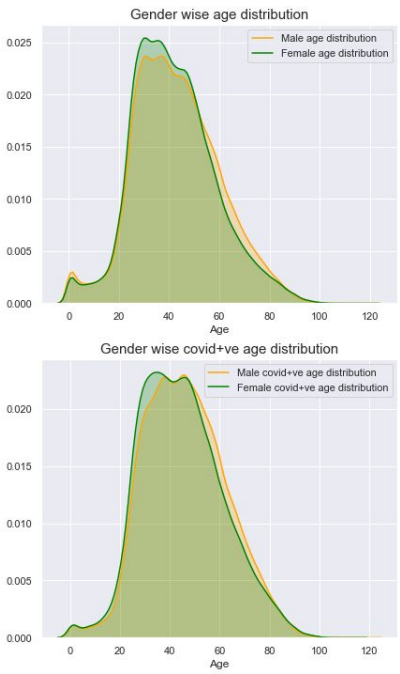

obesity patient numbers
count
Positive
Negative
Results
Test postivity rate: 52.78 %


renal_chronic patient numbers
count
Positive
Negative
Results
Test postivity rate: 47.80 %

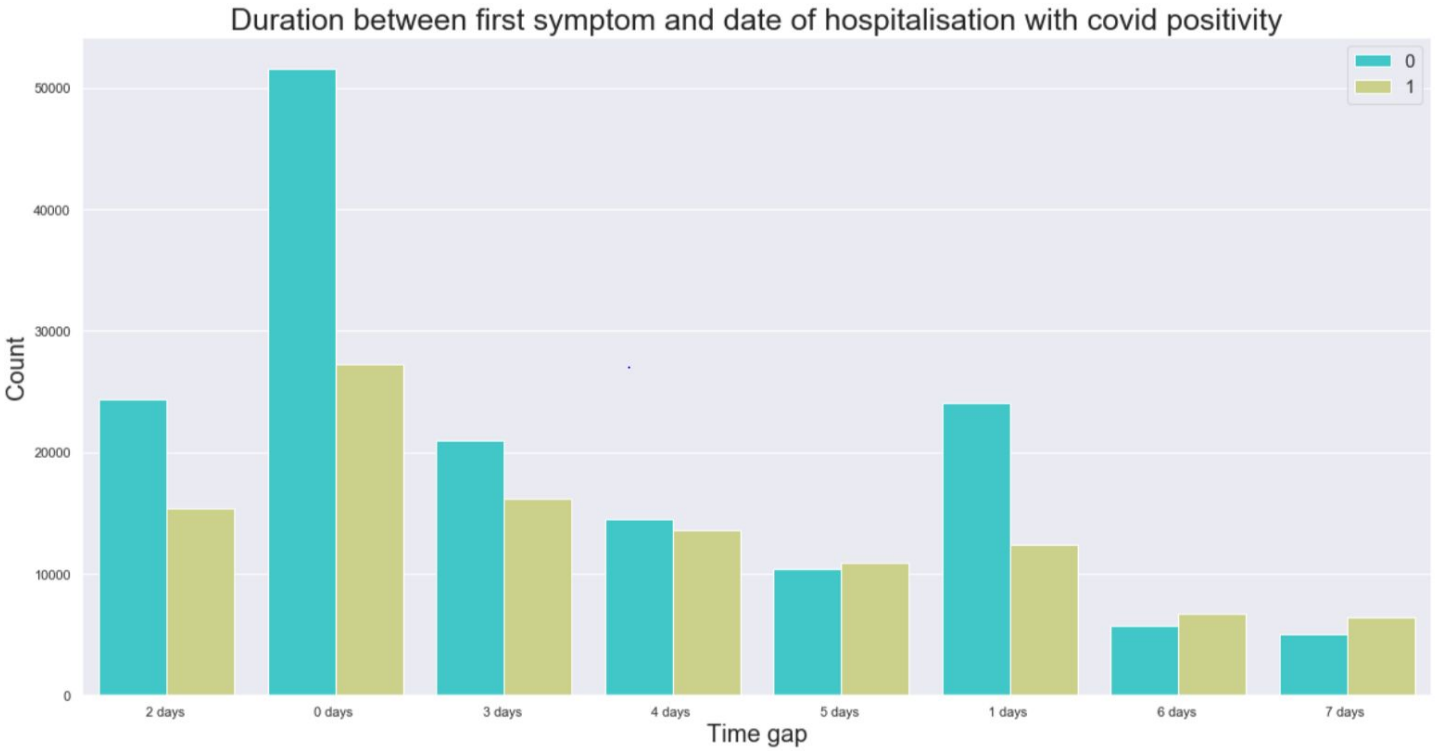

tobacco patient numbers
count
Positive
Negative
Results
Test postivity rate: 39.83 %


Covid Positive vs Age group
Percentage_covid_positive
0-2
10-20
2-10
20-30
30-40
40-50
50-60
60-80
Above 80
Age_band


Complete population age distribution
Population ages
age
Covid +ve population age distribution
Population ages
age


Gender wise age distribution
Male age distribution
Female age distribution
Age
Gender wise covid+ve age distribution
Male covid+ve age distribution
Female covid+ve age distribution
Age

There is also a significant relationship between the date of symptoms observed and the date of hospitalization. The more the gap the more will be the chance of being positive.



Chi-Square Test:

The Chi-square test is performed to find the correlation of various attributes with Covid_result. Since the attributes are categorical we have used the Chi-Square test and not Pearson correlation.
Hypothesis (H0) - If A and B are independent.

There is a **relationship** between the attribute *{covid_res}*, and the attributes *{sex, patient_type, inttubed, pneumonia, diabetes, copd, asthma, inmsupr, hypertension, other_diseases, cardiovascular, obesity, renal_chronic, tobacoo, contac_other_covid, icu}*
This test shows that there is **no relationship** between *covid_res* and *pregnancy*, which is generally expected.

MODELING

Decision Tree Regressor:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.
● Complex trees are hard to interpret which pertains to our dataset of 23 different attributes. Therefore the results from this algorithm have a large error.
● Random Forests (RF) is the algorithm that builds many individual trees, pooling their predictions, these are simply better as our data is too complex to use decision tree regression.

Random Forest:

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. This broad technique of using multiple models to obtain better predictive performance is called **model ensembling**.

● In random forests, all the base models are constructed independently using a different subsample of the data.
● The random forest model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, random forests are able to capture non-linear interaction between the features and the target.

Neural Network:

A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function (here we are using the sigmoid activation function) controls the output**.**

Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable (Covid positive or negative), although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

RESULTS & INTERPRETATIONS

Performance Comparison: comparing the accuracy, recall, precision, and F1 score of different models.

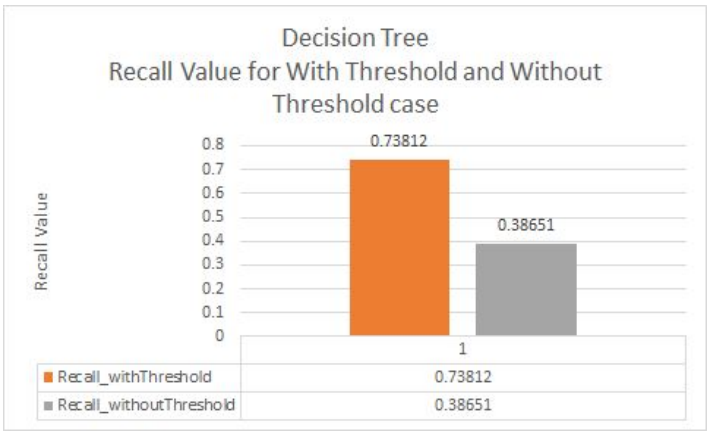
A) With Threshold Probability Function:-

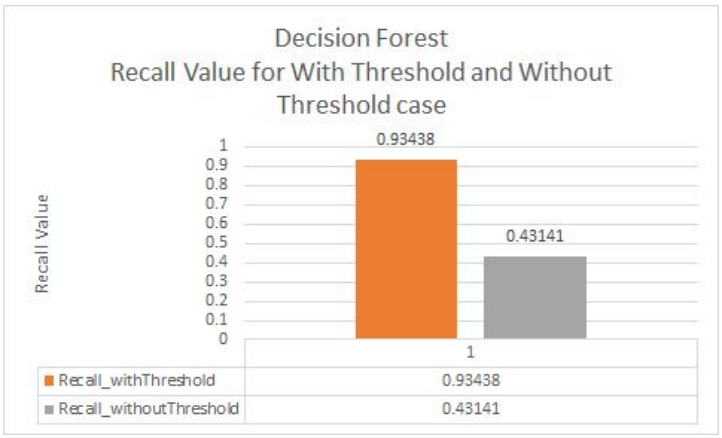
	Test Accuracy	**Training Accuracy**	**Recall**	**Precision**	**F1 Score**
Decision Tree	0.403432	0.262605	0.73812	0.40373	0.52196
Decision Forest	0.42459	0.382218	0.93438	0.430038	0.588997
Neural Network	0.51230	0.51334	0.912325	0.472728	0.622766
Logistic Regression	0.441252	0.440988	1.0	0.441252	0.612318

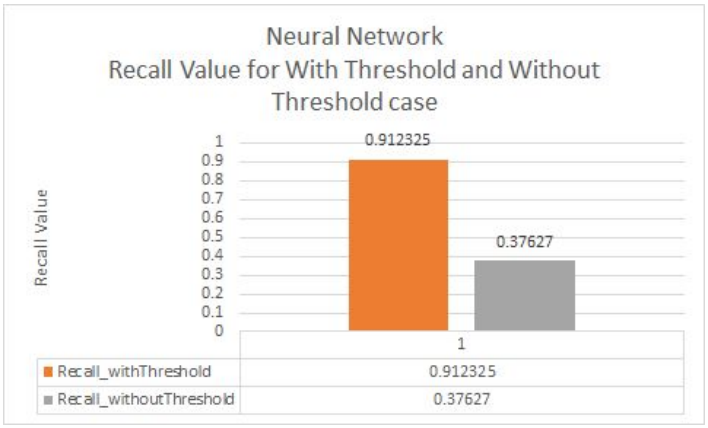
B) Without Threshold Probability Function:-

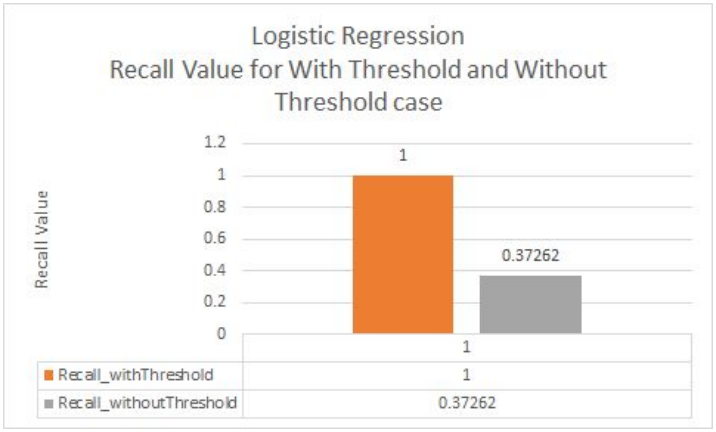
	Test Accuracy	**Training Accuracy**	**Recall**	**Precision**	**F1 Score**
Decision Tree	0.60147	0.77416	0.38651	0.57160	0.46118
Decision Forest	0.60839	0.77413	0.43141	0.57498	0.49295
Neural Network	0.63802	0.63878	0.37627	0.65681	0.47845
Logistic Regression	0.63281	0.63250	0.37262	0.64535	0.47245

The graphs below show the **effect** of the **Threshold probability function** on the **Recall Value** of different models. As can be seen in the graphs, the Recall value increases for the case when the Threshold probability function is used.









Significance of Threshold Probability function: We have used a threshold tuning for prediction as we wanted to maximize recall because we want to avoid False Negatives. This can be a very serious problem, especially in such medical cases. So we increase the recall by minimizing False Negatives and also trying to keep a decent accuracy. There will be a trade-off between these two. So, we used a new variable *recall*accuracy* and tried to maximize it.

Based on the above comparison, we can infer that the best model, if we are inclined towards increasing recall, is **Neural Networks**.

On the other hand, if we only want good accuracy (which is not preferred here) then the best model would be **Decision Tree or Decision Forest.**

We can see that the accuracy is not too good but satisfactory. This could be the case due to the type of features which we have and their role in telling/predicting whether or not a person is Covid +ve.