COVID DATA for MEXICO

Presented by Frida Oyucho

Business Understanding

This project aims to provide an in-depth analysis of the COVID-19 situation in Mexico and develop a machine learning-based model to predict new cases of the virus. By leveraging historical data on COVID-19 cases, the project seeks to identify trends, patterns, and potential correlations with various factors that may contribute to the spread of the virus. The insights gained from this analysis and the predictive model will help inform public health policies and strategies for better managing and mitigating the impact of the pandemic.

Data Understanding

The data was extracted from the Government of Mexico's ministry of Health national data repository. The data conatins 10,000 rows with 22 columns. The target variable for modelling is **['covid_res']** and predictor variables after preprocessing the data areas listed: **['sex', 'patient_type','month_name', 'intubed', 'pneumonia', 'pregnancy', 'diabetes', 'copd', 'asthma', 'inmsupr', 'hypertension', 'other_disease', 'cardiovascular', 'obesity', 'renal_chronic', 'tobacco', 'contact_other_covid', 'icu','dead','age_groups']**

COVID Background and Analysis Objective

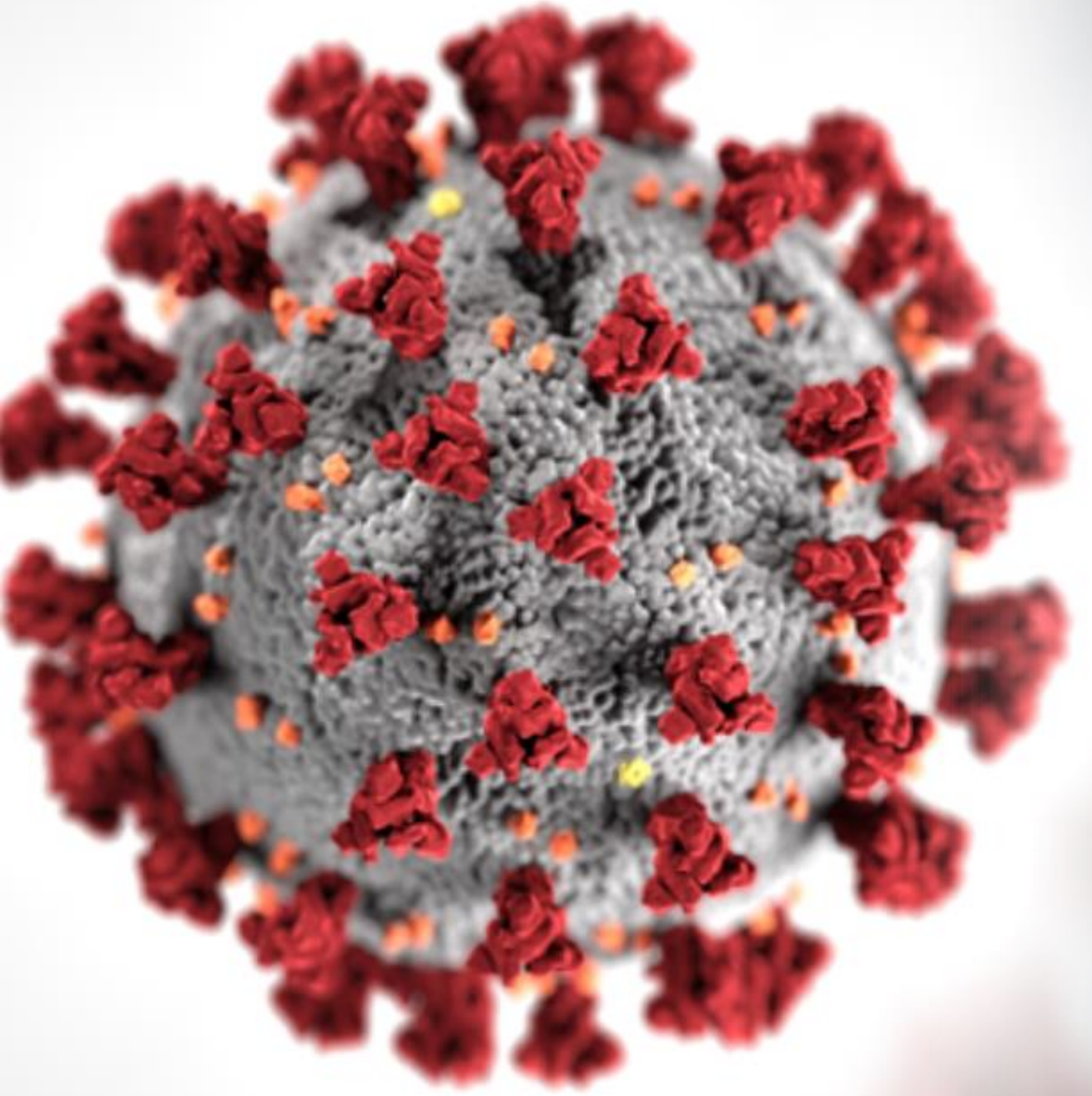
Background

Coronavirus disease 2019 (COVID-19) is an illness caused by a novel coronavirus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; formerly called 2019-nCoV), which was first identified amid an outbreak of respiratory illness cases in Wuhan City, Hubei Province, China

Understanding the underlying health conditions associated with COVID is crucial for health practitioners to help them stratify patients' risk for developing severe covid complications.

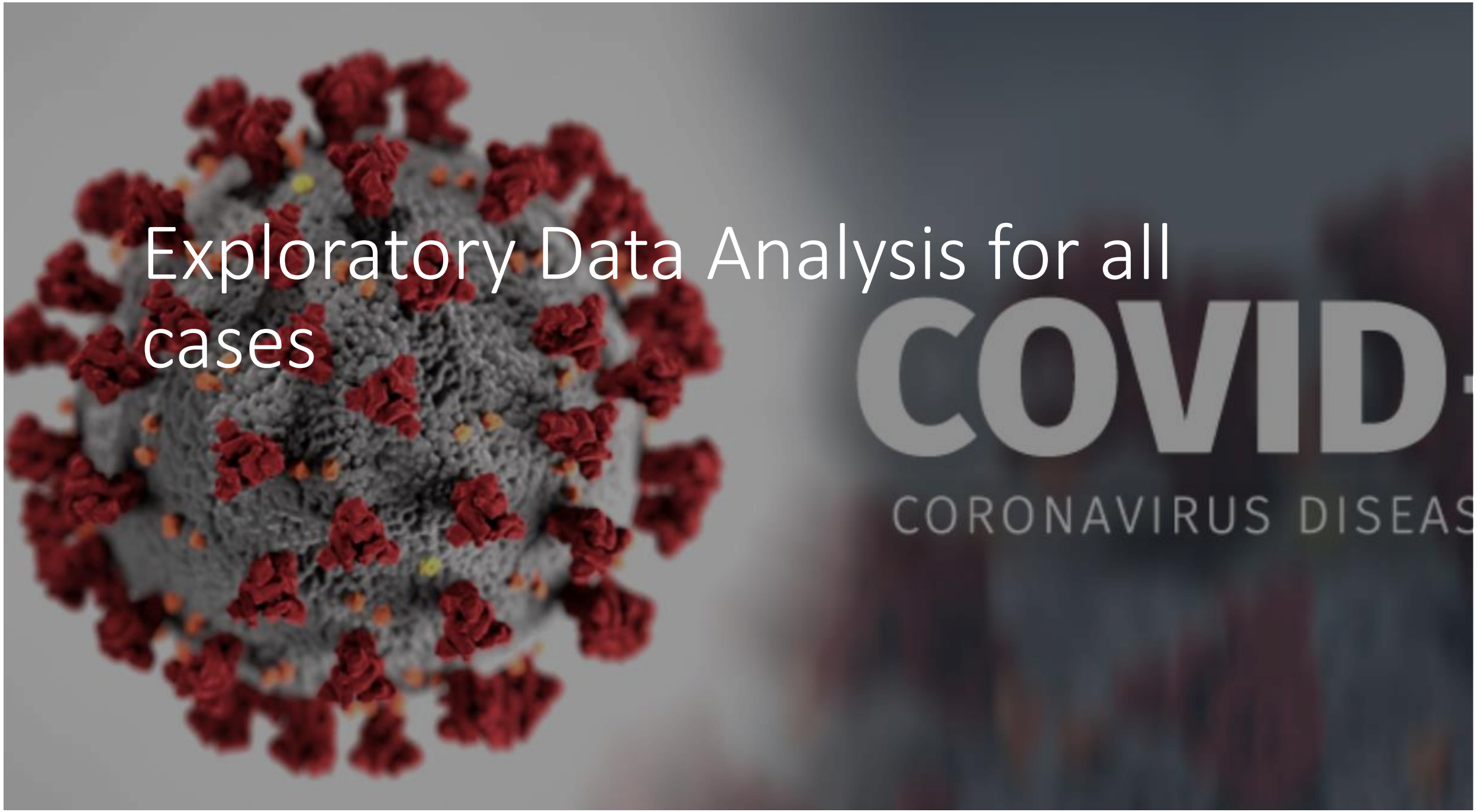
Objectives

- Conduct Exploratory Data Analysis: Analyze the COVID-19 situation in Mexico, identifying trends and patterns in the spread of the virus.
- Investigate potential correlations between different factors and the number of new COVID-19 cases.
- Evaluate and compare the performance of the developed Machine learning models .
- Provide actionable insights and recommendations to inform public health policies and strategies for controlling the spread of COVID-19 in Mexico.



Methodology

- The data analysis was based on clients tested for COVID clients who were admitted to care between January to December 2020
- Data preparation was done to align data types, detect and deal with: missingness, duplicate records and outliers
- Feature engineering was done to two variable: ‘dead’ , ‘age_groups’ and ‘month_names’
- Exploratory Data analysis was done on: Univariate, Bi-variate and Multivariates
- Data Preprocessing was done to prepare the data for modeling(Normalization using scalar & One hot enocoding)
- Data modelling was done using three models: Logistic Regression, Random Forest & XGboost

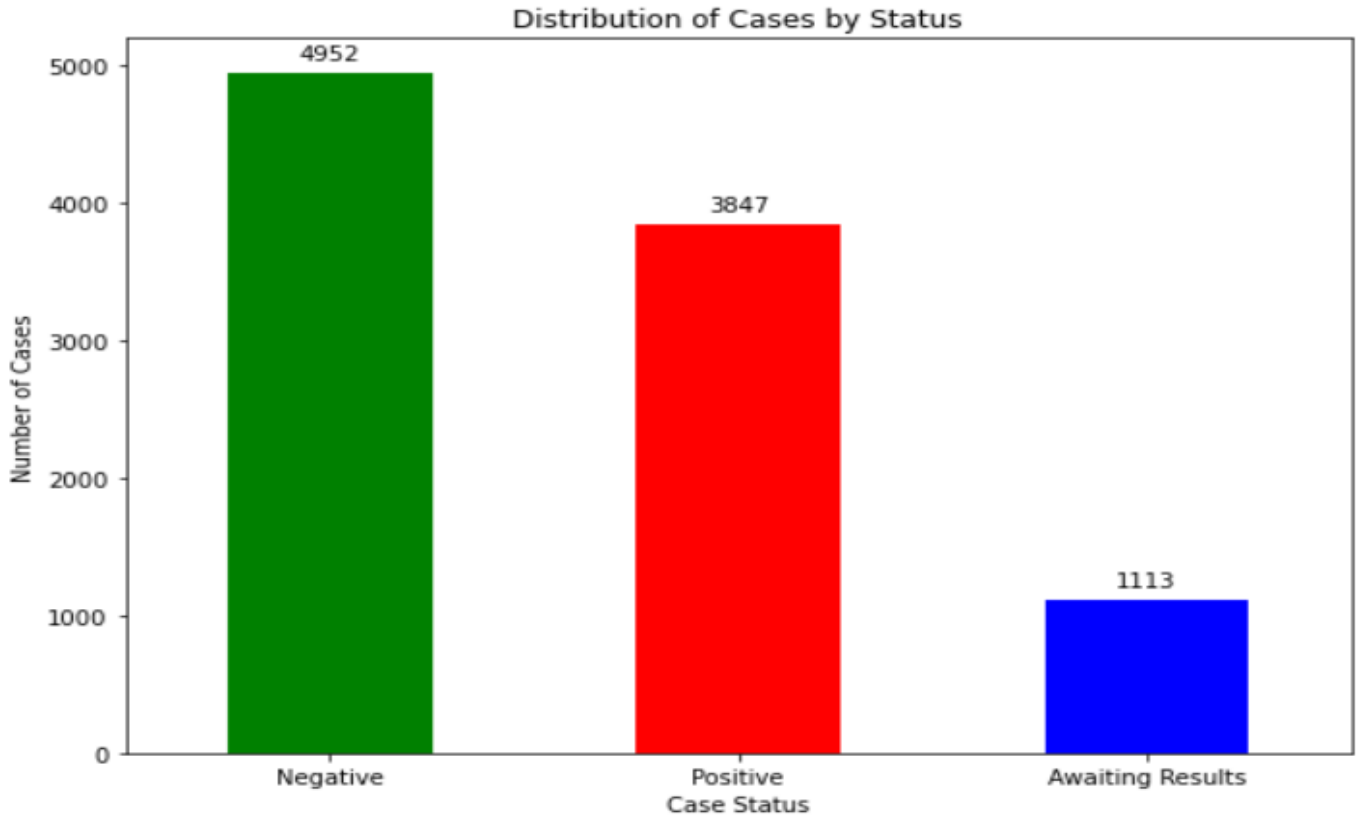


Exploratory Data Analysis for all cases

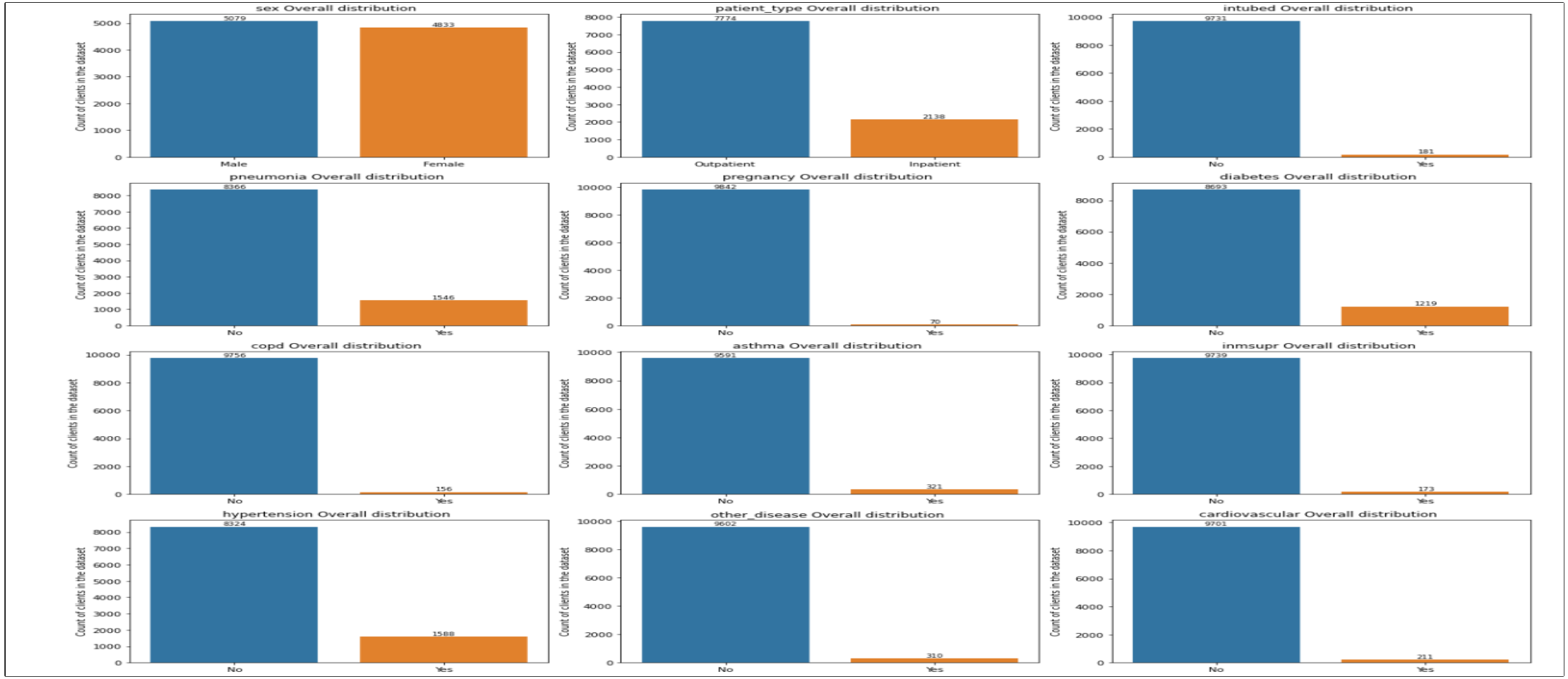
Distribution of cases by Status

50% of clients tested for covid were negative. 38.8% were positive

N=10,000



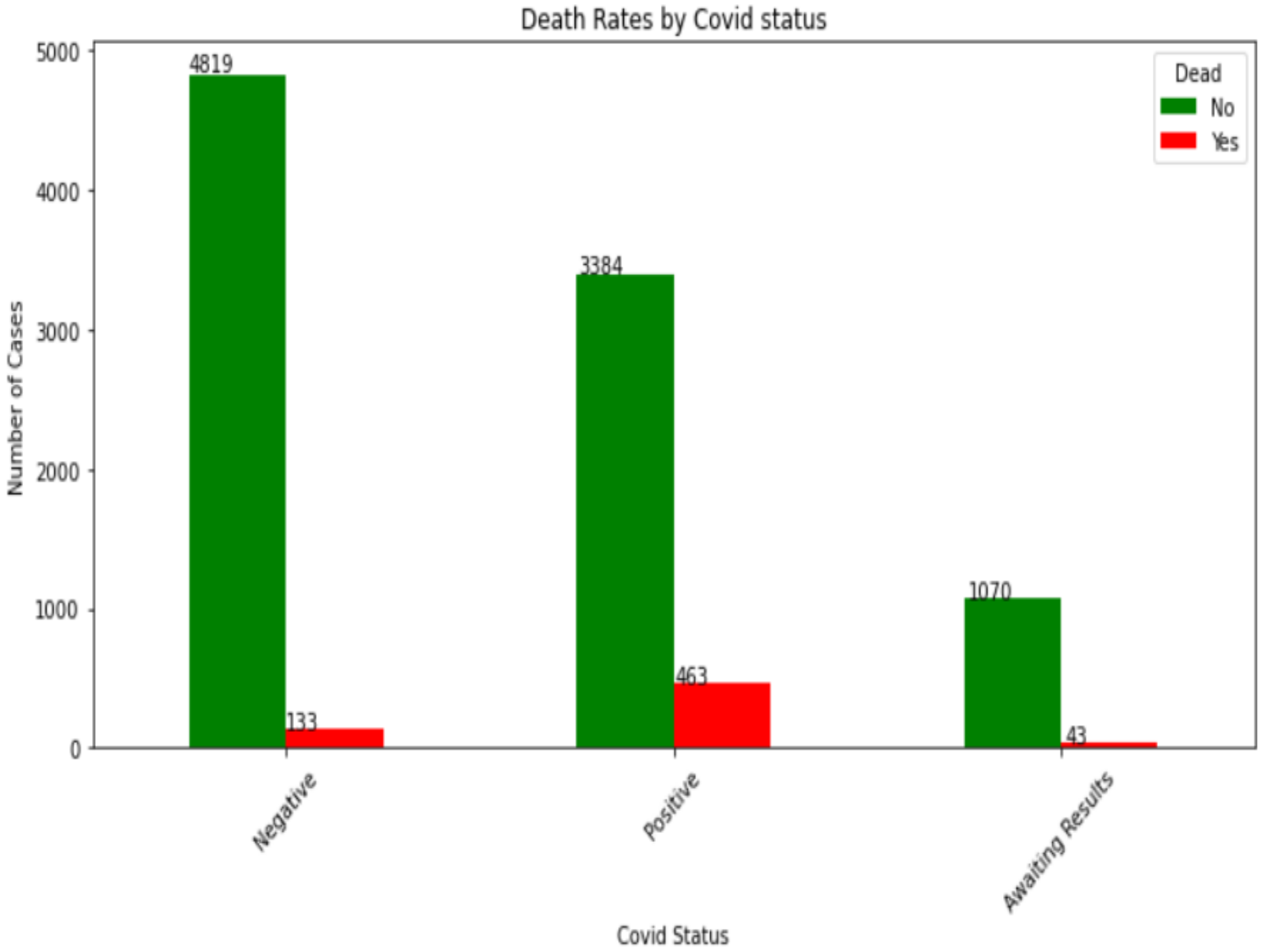
Distribution among Categorical variables



Distribution Mortality rates

72% of deaths were from clients who tested positive

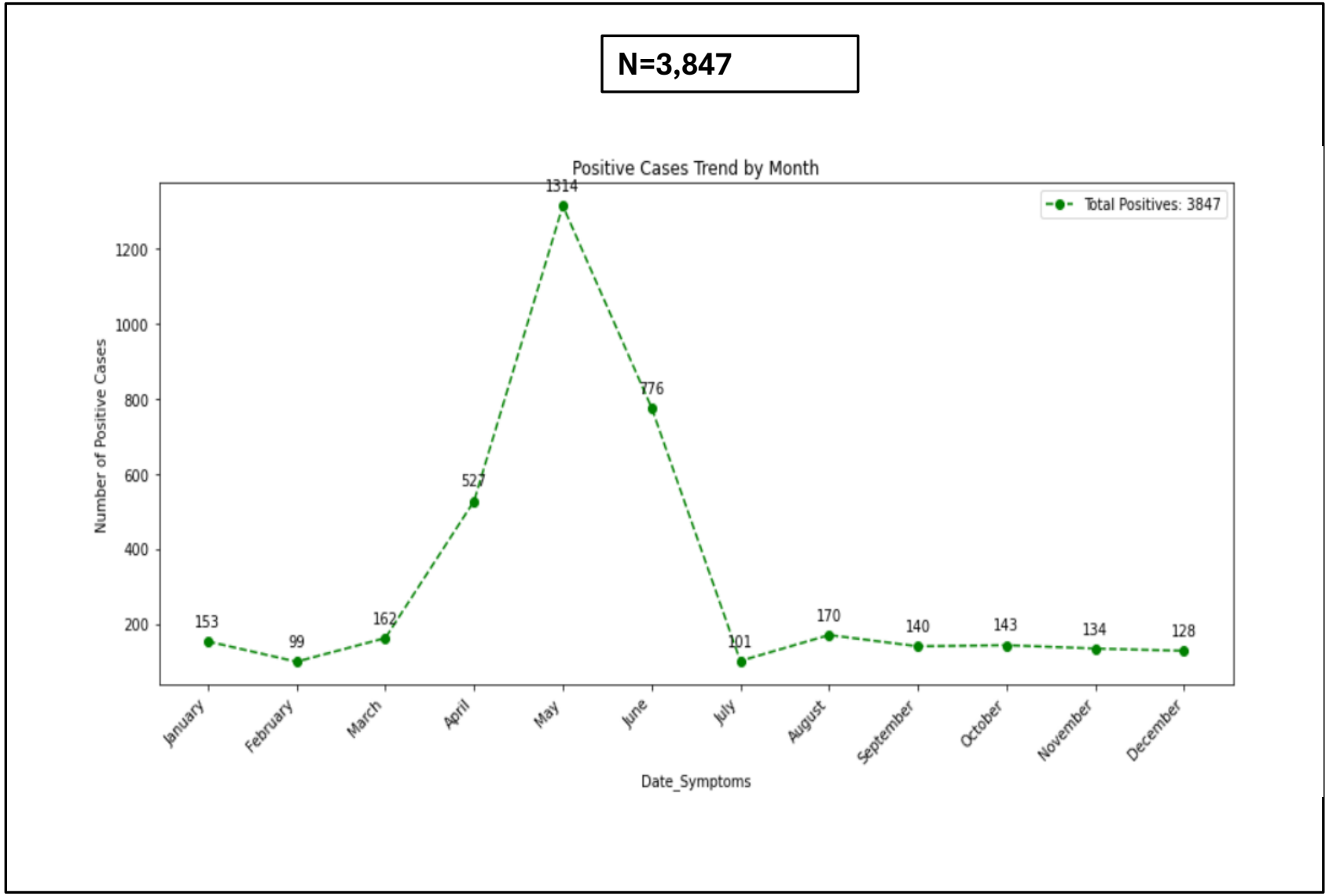
N=10,000



COVID Positive Clients

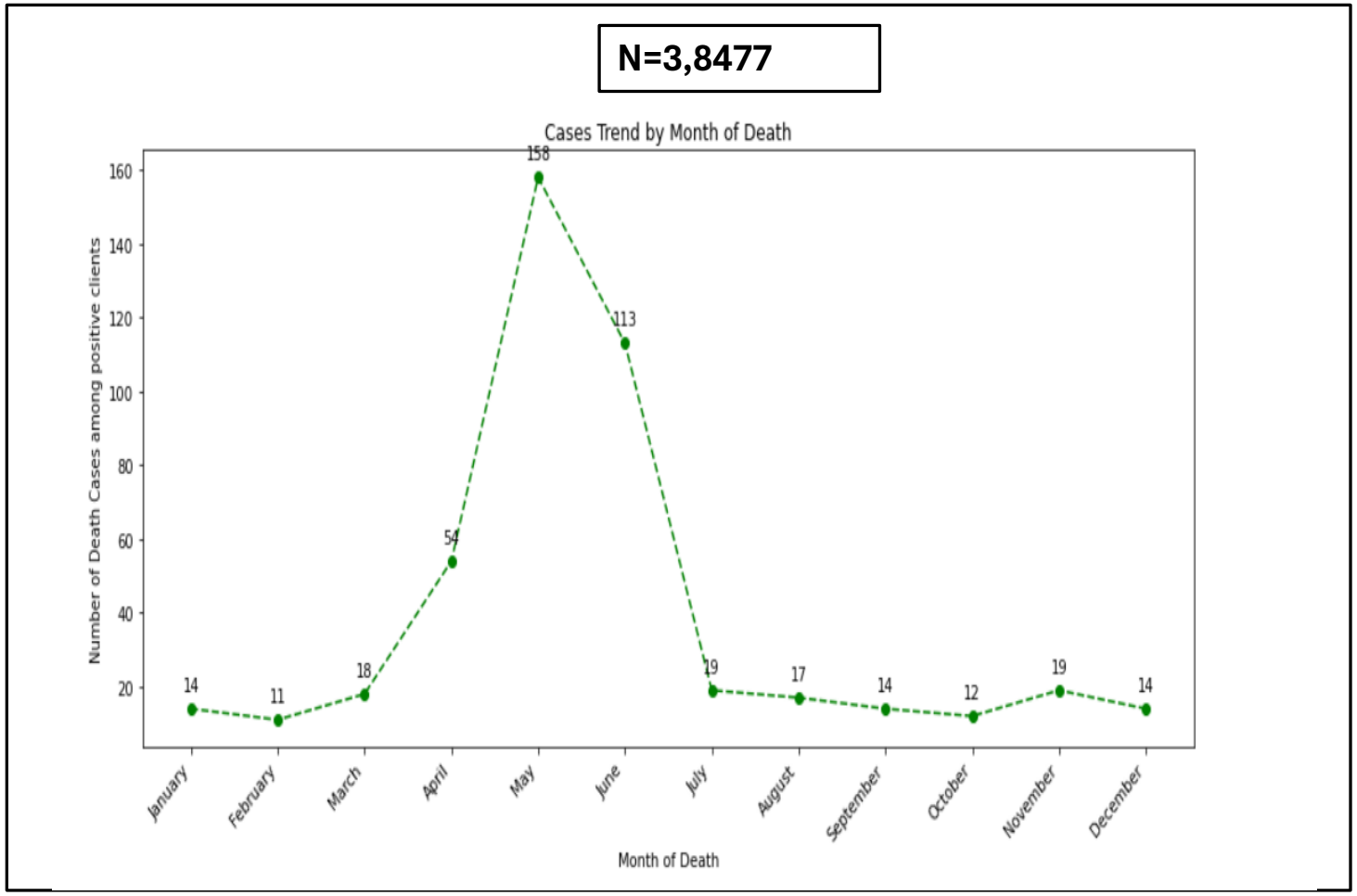
Trend Analysis on positive cases

Peak positivity rates was in the period of May 2020.



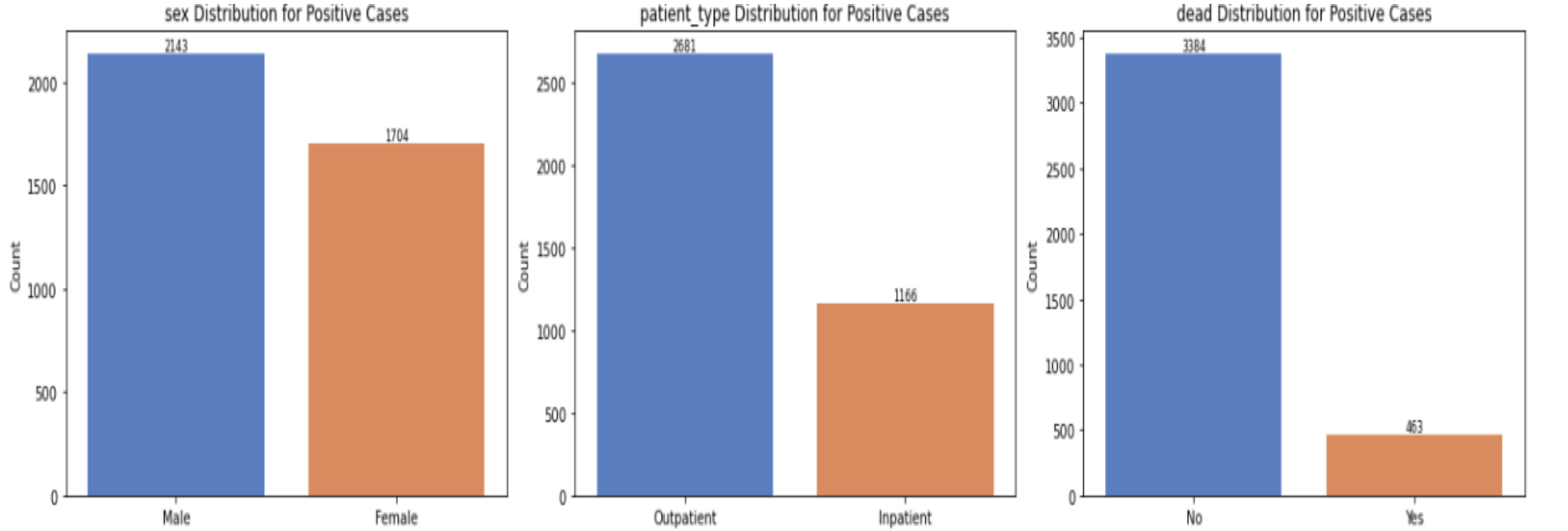
Trend Analysis on death rates for positive cases

Peak death rates was in the period of May 2020.

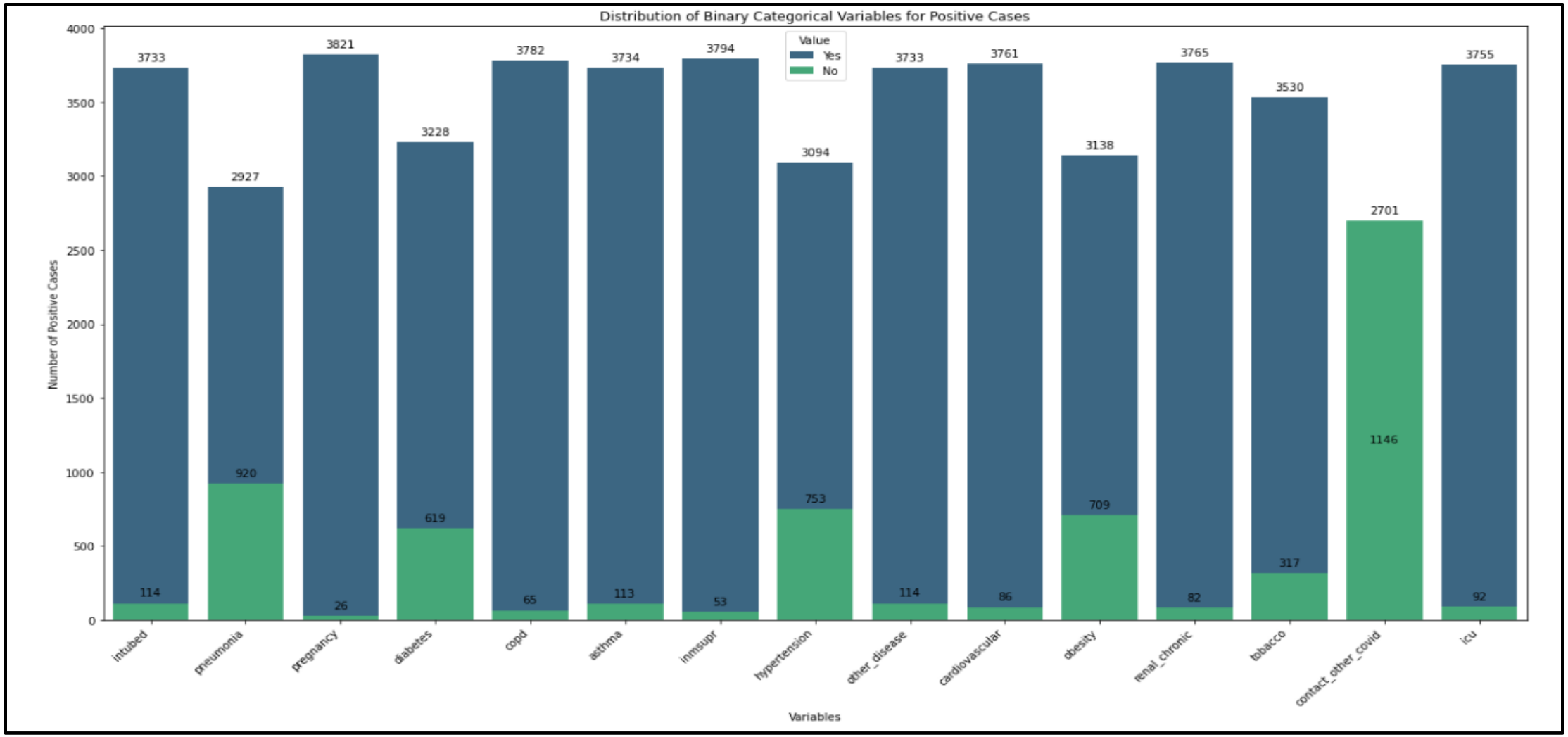


Distribution of variables among positive cases

N=3,847



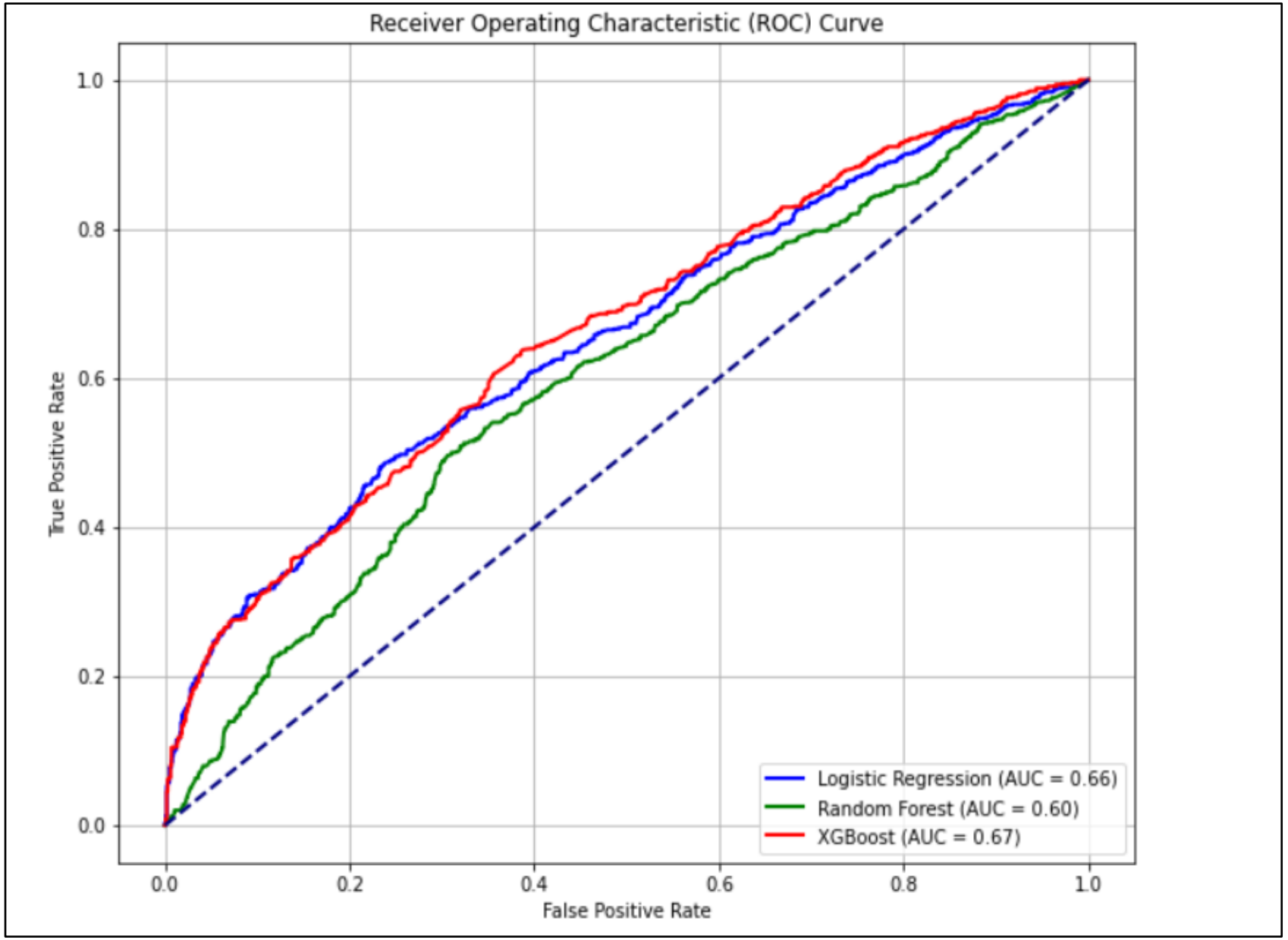
Distribution of diseases among confirmed cases



Visualizing Model Performance

ROC Curve visualizing the performance of the three models.

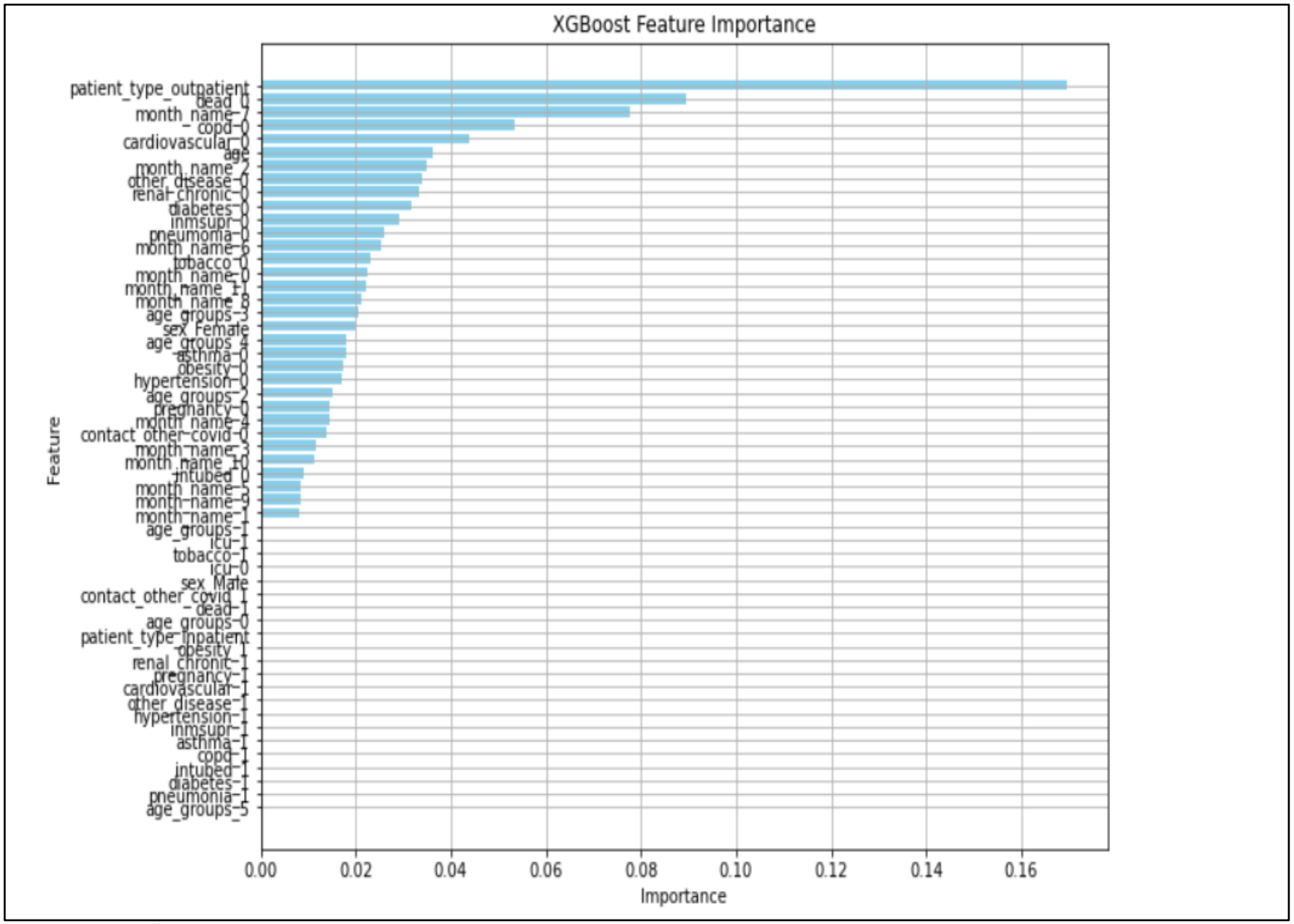
Hyperparameter tuning was done using grid search to all the models. The model that performed the best was XGBoost with an AUC of 67%

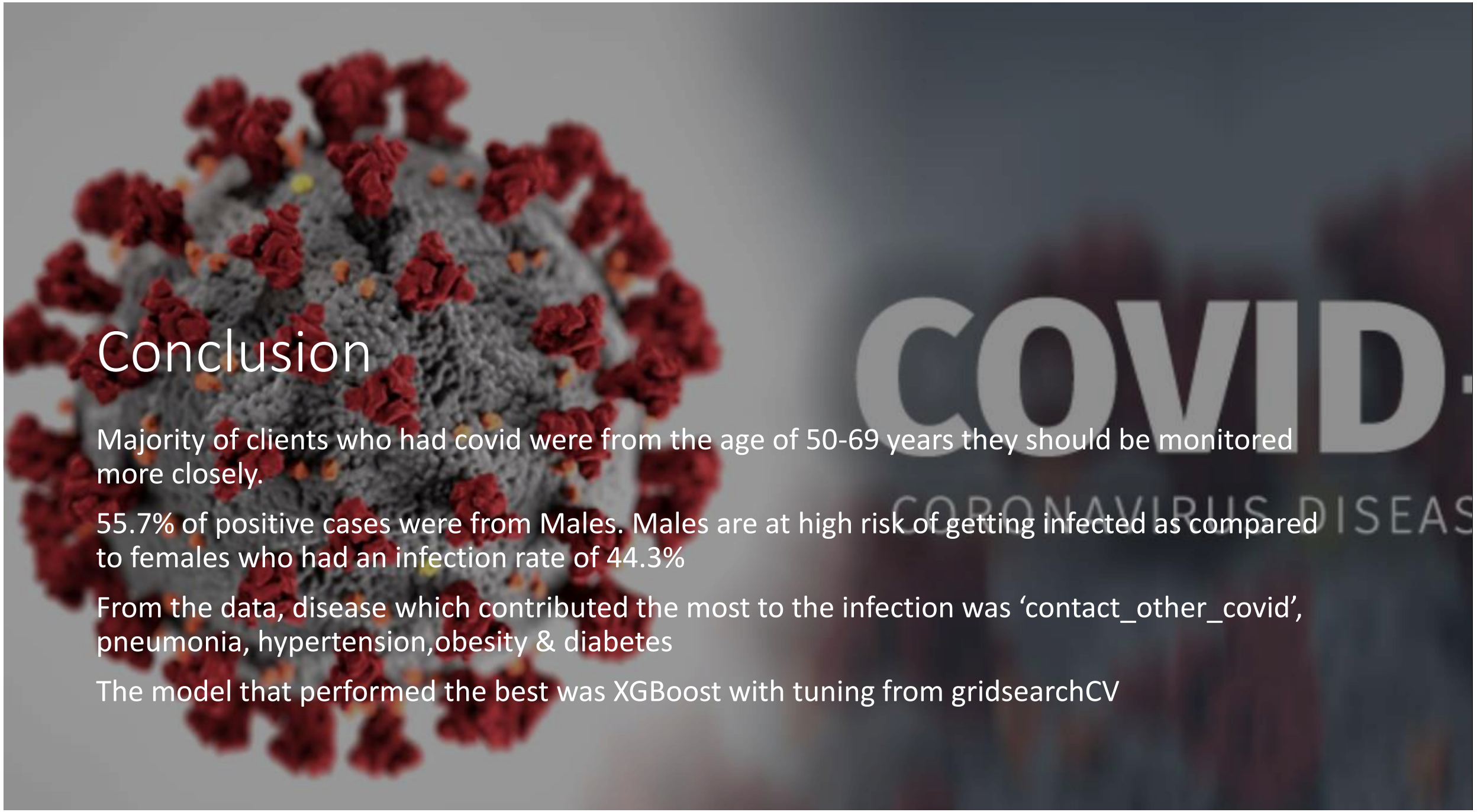


Feature Importance

From the model that performed the best, the top 5 features of importance were:

1. Patient_type_outpatient
2. died_0
3. date_symptoms
4. copd clients
5. cardiovascular clients





Conclusion

Majority of clients who had covid were from the age of 50-69 years they should be monitored more closely.

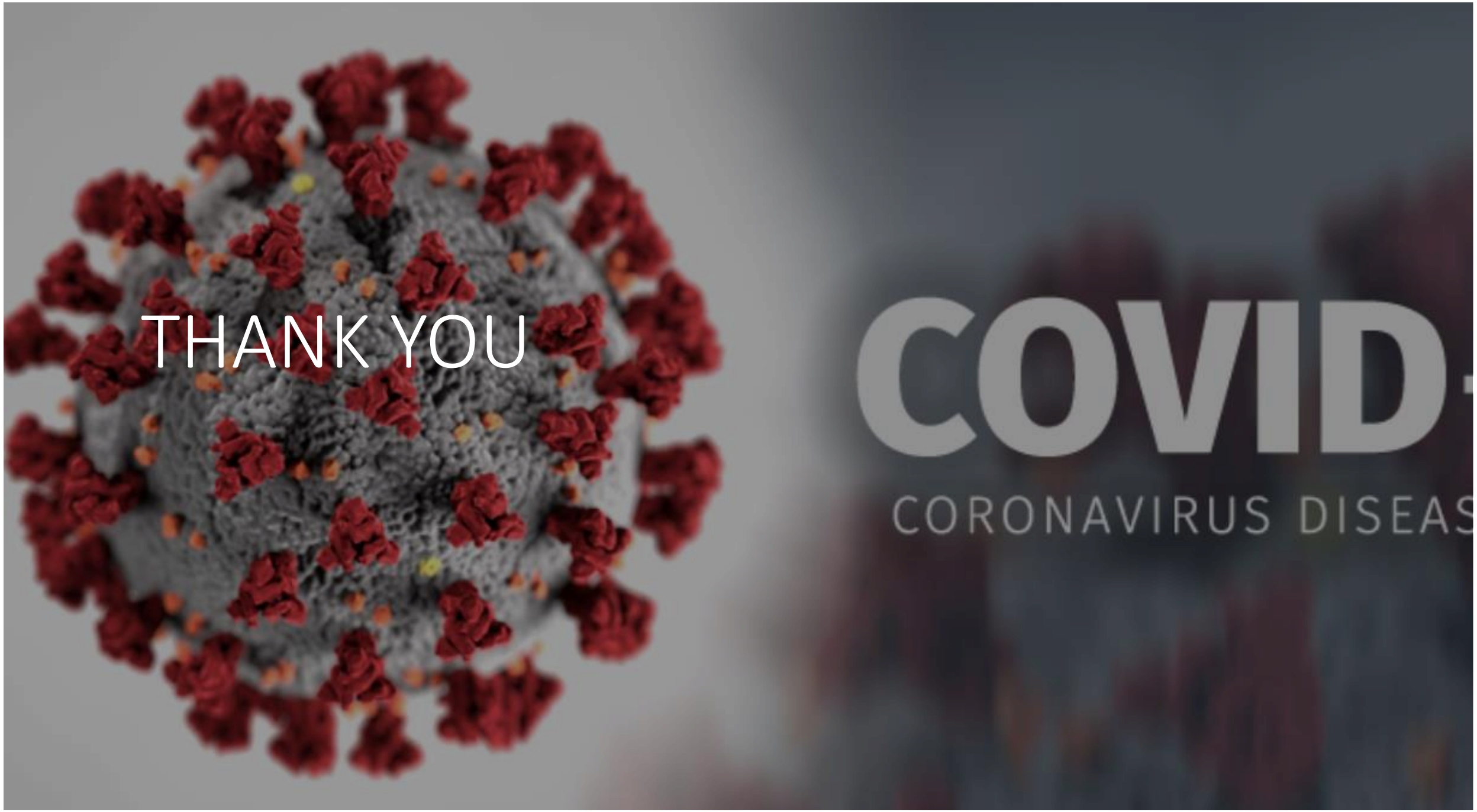
55.7% of positive cases were from Males. Males are at high risk of getting infected as compared to females who had an infection rate of 44.3%

From the data, disease which contributed the most to the infection was ‘contact_other_covid’, pneumonia, hypertension,obesity & diabetes

The model that performed the best was XGBoost with tuning from gridsearchCV

Recommendation

Date Symptoms : Implementing aggressive testing and contact tracing for individuals reporting early onset of symptoms can swiftly identify and isolate positive cases, curbing the spread of the virus.


THANK YOU
COVID
CORONAVIRUS DISEAS