Predictive Modeling for COVID-19 Diagnosis Using Machine Learning

presented by Subeena k.k

TABLE OF CONTENT

Introduction

In the wake of the COVID-19 pandemic, the importance of early and accurate diagnosis has become critical for effective virus management and containment. While widespread testing has proven effective, the process can be resource-intensive and time-consuming. With the advancements in machine learning, we can harness the power of data to aid in the diagnostic process. By analyzing patient symptoms and demographic information, machine learning models can serve as a valuable tool for predicting COVID-19 cases, potentially easing the burden on healthcare systems. This project aims to explore the potential of machine learning in predicting COVID-19 diagnoses based on readily available information such as symptoms and demographic data.

Project Objective

The objective of this project is to develop a robust machine learning classification model for predicting COVID-19 diagnoses. The model will be built using a comprehensive dataset that includes key symptoms such as cough, fever, sore throat, shortness of breath, headache, and demographic information such as age, sex, and known contacts with infected individuals. The goal is to accurately classify whether a patient is likely to have COVID-19 based on these features, contributing to more timely identification and supporting healthcare professionals in making informed decisions.

Data Overview

The dataset used for this project consists of 278,848 records with 11 features. These features encompass both symptom-related data and demographic information that are crucial for building a predictive model for COVID-19 diagnosis.

The key columns include:

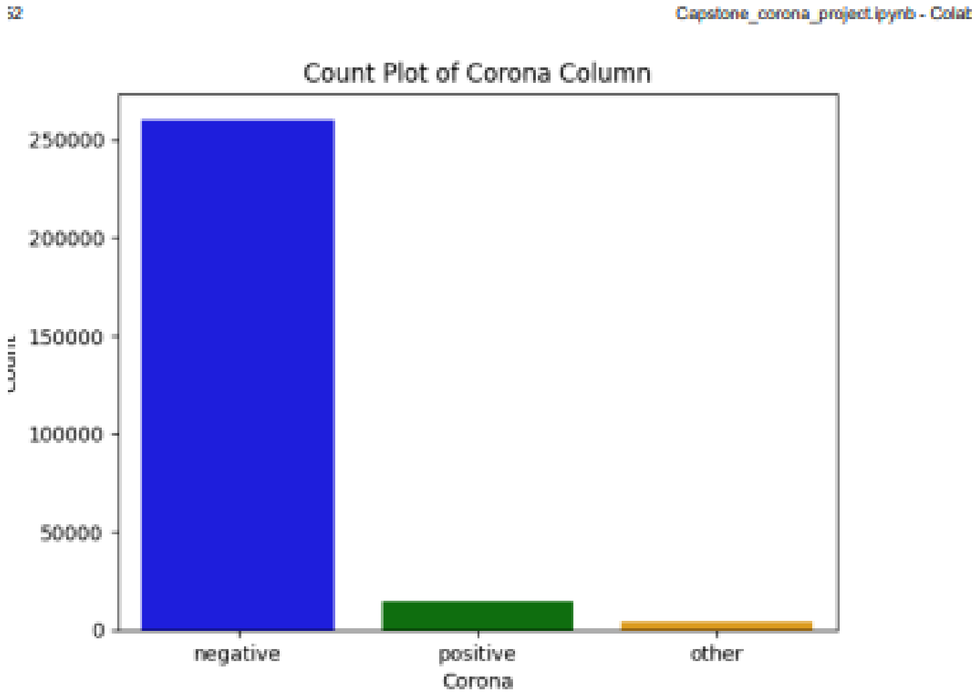
- Ind_ID: Unique identifier for each individual
- Cough_symptoms: Presence of cough symptoms (True/False)
- Fever: Presence of fever (True/False)
- Sore_throat: Presence of a sore throat (True/False)
- Shortness_of_breath: Presence of shortness of breath (True/False)
- Headache: Presence of a headache (True/False)
- Corona: COVID-19 test result (positive/negative)
- Age_60_above: Indicator of whether the individual is aged 60 or above
- Sex: Gender of the individual (male/female)
- Known_contact: Information on whether the individual had contact with a confirmed COVID-19 case
- Test_date: Date of the COVID-19 test

SQL-Based Analysis

To derive meaningful insights from the dataset, we performed several SQL queries to answer key analytical questions, including:

1. Number of COVID-positive patients who experienced shortness of breath.
2. Count of COVID-negative patients who had both fever and sore throat.
3. Monthly ranking of positive COVID-19 cases.
4. Number of female COVID-negative patients who experienced cough and headache.
5. Count of elderly COVID-positive patients who had breathing problems.
6. Three most common symptoms among COVID-positive patients.
7. Least common symptom among COVID-negative individuals.
8. Most common symptoms among COVID-positive males with a known contact abroad.

By leveraging SQL for these queries, we were able to extract critical insights and identify patterns within the data, contributing to our overall understanding of COVID-19 symptomatology.

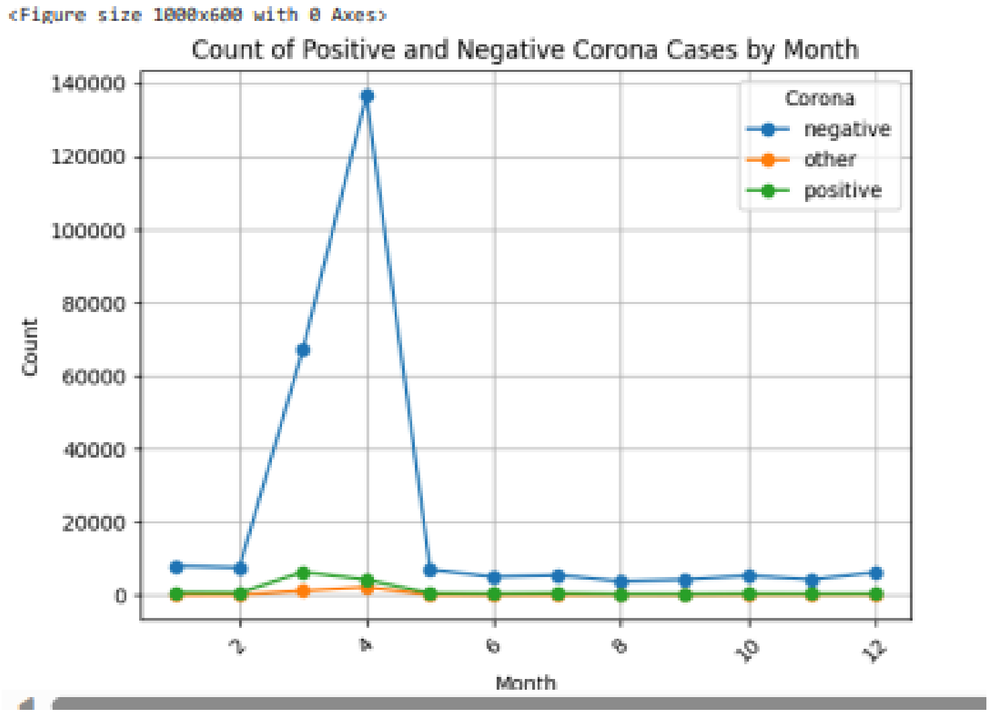


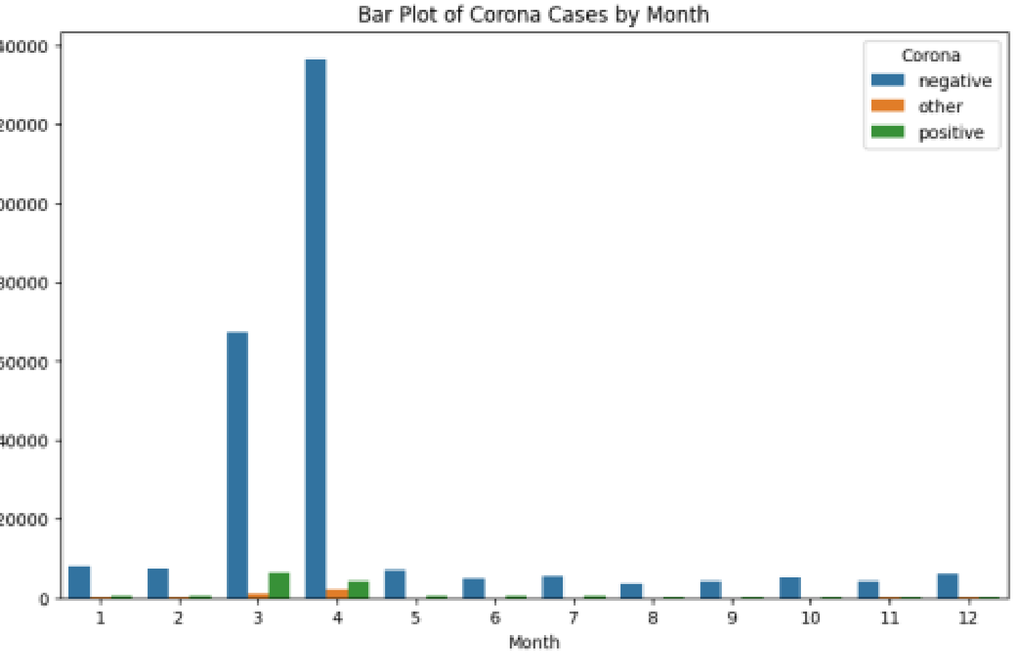
Count Plot for 'Corona' Column

- There is a higher count of negative cases compared to positive cases.
- This indicates that most individuals in the dataset tested negative for COVID-19.
- A smaller proportion of individuals tested positive, reflecting a lower prevalence of COVID-19 within this sample.

Line chart over a month

- we can clearly see that The third and fourth months have some positive cases, with the third month having slightly more positive cases than the fourth month.
- Remaining months have very few positive cases compared to the third and fourth months.



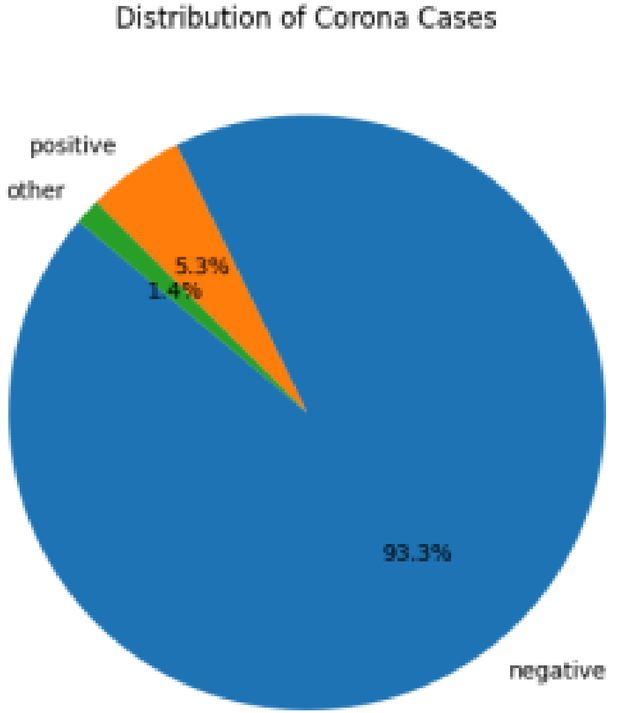


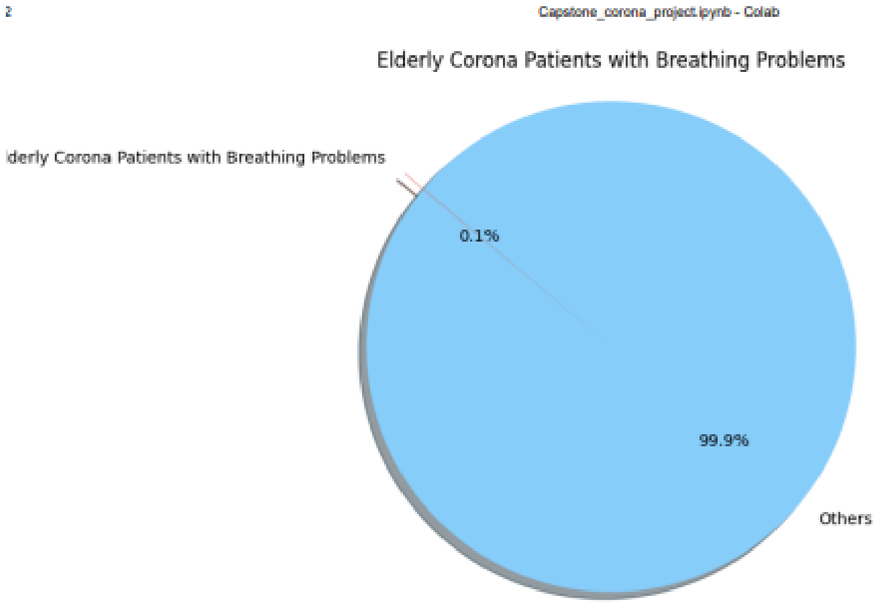
Bar plot for 'Corona' column against 'month'

- Third and fourth months show a notable presence of positive cases:
  - The third month has a slightly higher number of positive cases than the fourth.
  - These months represent a peak period for COVID-19 positive cases.
- Remaining months have very few positive cases compared to the third and fourth months, indicating a decline or low incidence of COVID-19 during those periods.

Pie Chart for Corona Cases in %

- In this pie chart we can also see that 93.3% negative cases , 5.3 % positive cases and 1.4% other in the given dataset



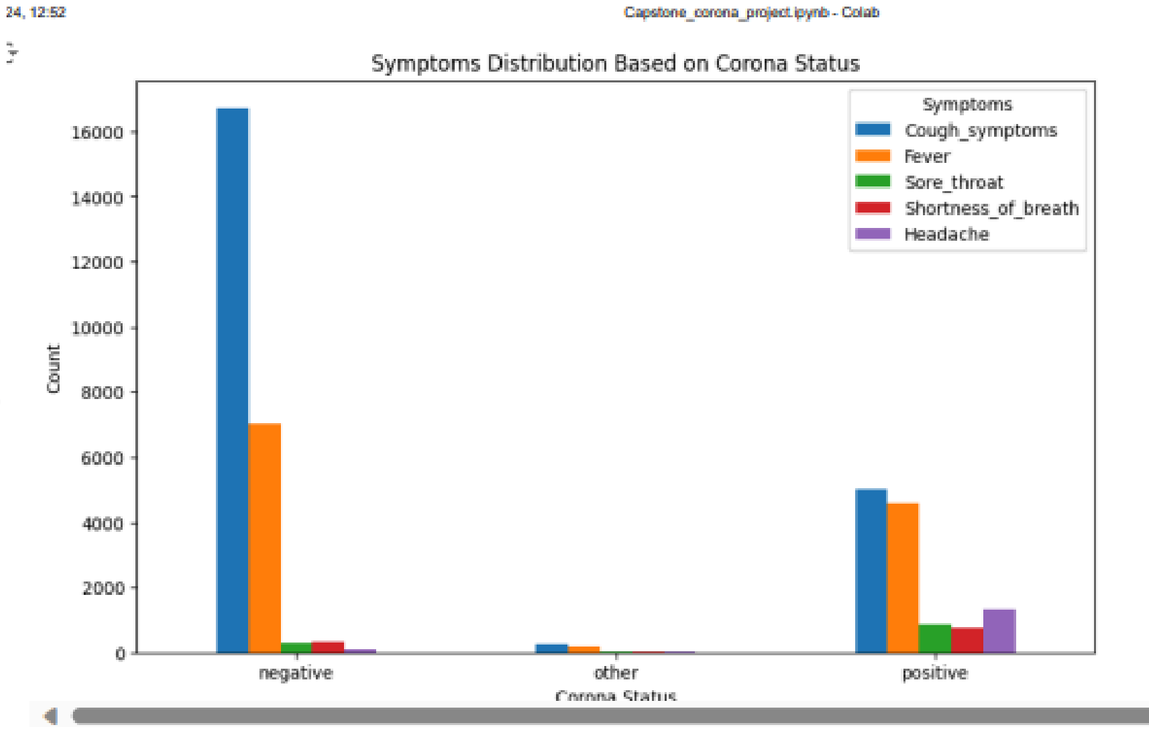


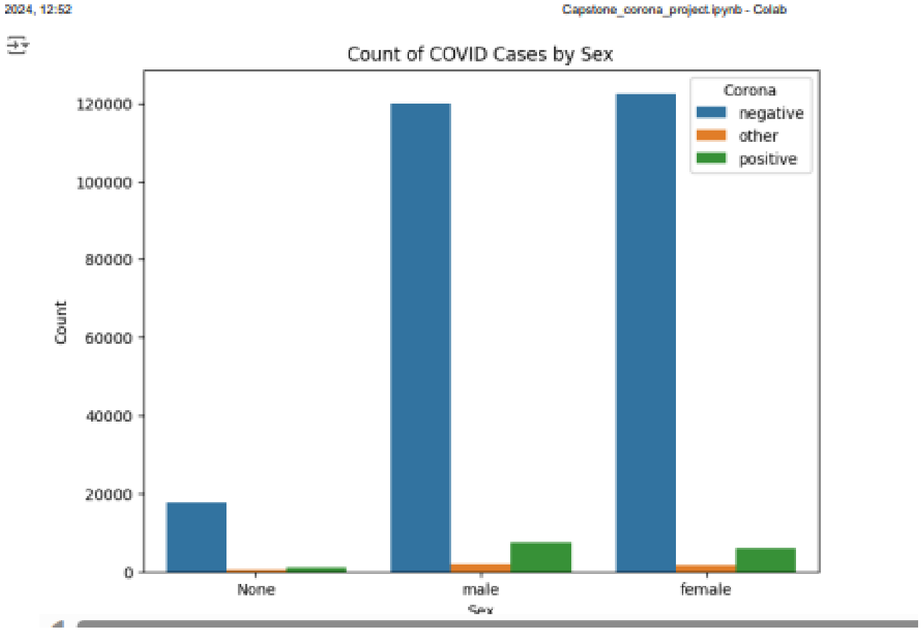
Pie chart -elderly corona patients who faced breathing problems

- From this pie chart we can see that only 0.1% corona patients who faced breathing problems .also we calculate exact count is 169

Bar chart for corona column based on symptoms related columns

- In this pie chart we can also see that 93.3% negative cases , 5.3 % positive cases and 1.4% other in the given dataset



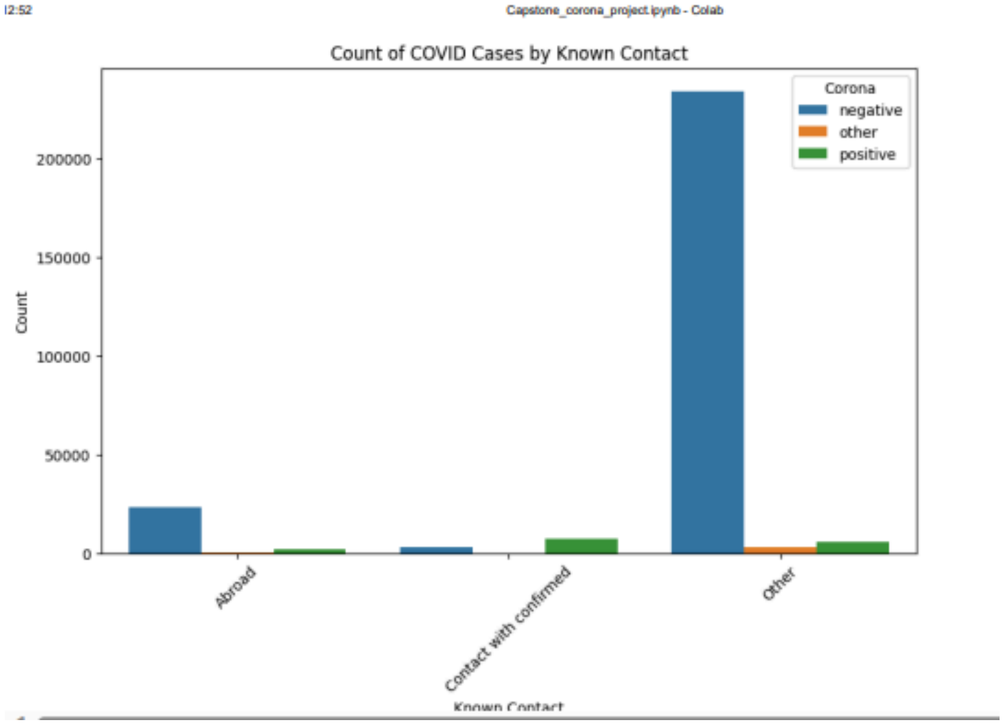


Count plot -Count of COVID Cases by Sex

- More positive cases are observed among males than females.
- The bars for 'positive' cases in the male category are significantly taller compared to those for females, indicating a higher incidence of COVID-19 among males in this dataset.

Count of COVID Cases by Known Contact

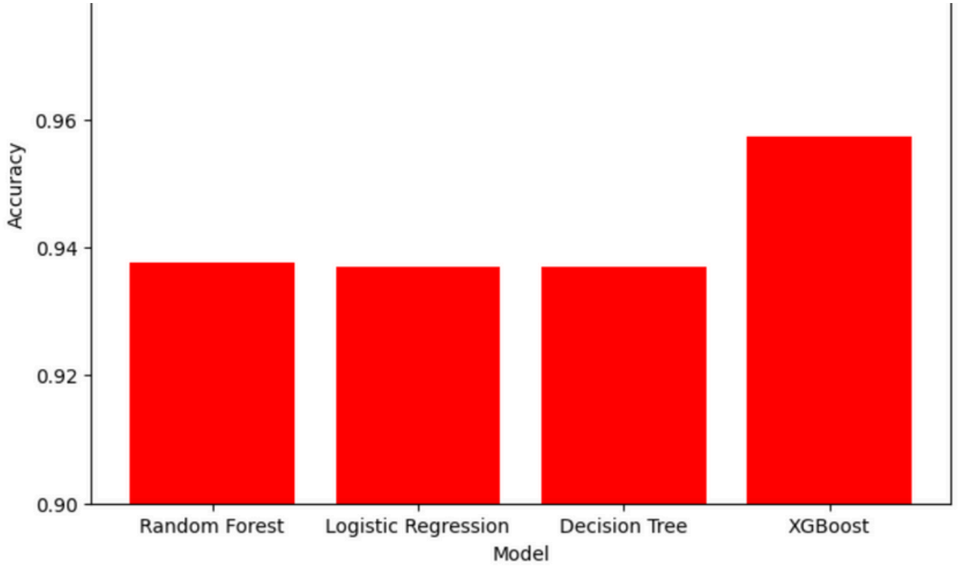
- From the count plot of COVID cases by known contact, it is evident that a larger number of positive cases are associated with individuals who have had contact with confirmed COVID-19 cases compared to those who have had contact abroad.



Model Building and Accuracy Evaluation

Models Developed:

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **XGBoost**



Cross-Validation Results:

- Logistic Regression: 93.69% accuracy
- Decision Tree: 93.69% accuracy
- Random Forest: 93.75% accuracy
- XGBoost: 95.72% accuracy

Key Insight:

- XGBoost achieved the highest accuracy, outperforming all other models.
- XGBoost’s superior performance is attributed to its advanced gradient boosting algorithm, which effectively handles complex data and enhances predictive accuracy.
-

Conclusion

Based on the cross-validated accuracy scores, XGBoost emerged as the most effective model for predicting COVID-19 status in our dataset. Its ability to handle intricate relationships between features and achieve high accuracy through advanced boosting techniques made it the optimal choice. As a result, we confidently accept our second null hypothesis, affirming that XGBoost is the most suitable model for this predictive task. This highlights its potential for enhancing COVID-19 diagnosis and supporting informed healthcare decision-making.

Recommendations

- **Focus on Testing During Peak Periods:**

Allocate more resources for testing and monitoring during peak periods, particularly the third and fourth months, which showed the highest number of positive cases. Early identification during these periods could help in controlling the spread of the virus.

- **Target Male Demographic for Preventive Measure:**

Given the higher incidence of COVID-19 among males compared to females, targeted preventive measures and awareness campaigns should be directed toward the male demographic to reduce their infection rate.

- **Emphasize Testing for Known Contacts:**

Prioritize testing for individuals who have had contact with confirmed COVID-19 cases, as they represent a larger proportion of positive cases. This strategy could help detect and isolate positive cases more effectively.

- **Strengthen Support for Individuals with Breathing Problems:**

Even though only 0.1% of COVID-19 patients faced breathing problems, specific medical support should be allocated for these individuals due to the severity of this symptom. Ensuring timely medical intervention can be critical for these cases.

THANK YOU