2. COVID PREDICTION USING MACHINE LEARNING TECHNIQUES

Dr. K. Rohini
Professor, Department of Information Technology,
VISTAS, rrohini16@gmail.com

Dr. K. Kasturi
Associate Professor, Department of Information Technology,
VISTAS, Kastur.scs@velsuniv.ac.in

Hariharan.K
Student, Master of Science in Data Science and Business Analytics
Department of Information Technology, VISTAS

ABSTRACT:

The COVID-19 pandemic has had a profound impact on global public health and the economy. Machine learning algorithms have played a crucial role in predicting the spread of the virus and assessing its impact. This work aims to develop a predictive model for COVID-19 using machine learning techniques. We leverage a comprehensive dataset of COVID-19 cases, including demographic, geographic, and epidemiological information. The main goal of this

work is to develop a machine learning-based predictive model for COVID-19. This model aims to forecast the spread of the virus, including the number of cases, infection rates, and potential hotspots.

The results highlight the potential of machine learning algorithms in predicting COVID-19 outcomes with reasonable accuracy, demonstrating their value in aiding healthcare professionals in making informed decisions.

Key words: Machine learning algorithms; COVID-19; Predictive model ;Logistic Regression; Model; Accuracy; Data Collection; Preprocessing;

1. INTRODUCTION

The COVID-19 pandemic has had a profound impact on global public health, necessitating the development of predictive models to help monitor and manage the spread of the virus. Machine learning algorithms, such as Random Forest and Logistic Regression, have proven to be effective tools in this context. In this study, we explore the application of these algorithms to predict COVID-19 outcomes, with a focus on infection rates, hospitalizations, and mortality.

Our study presents the results of COVID-19 prediction using Logistic Regression. We discuss the model's accuracy and effectiveness in forecasting infection rates, hospitalization probabilities, and mortality risks in different regions.

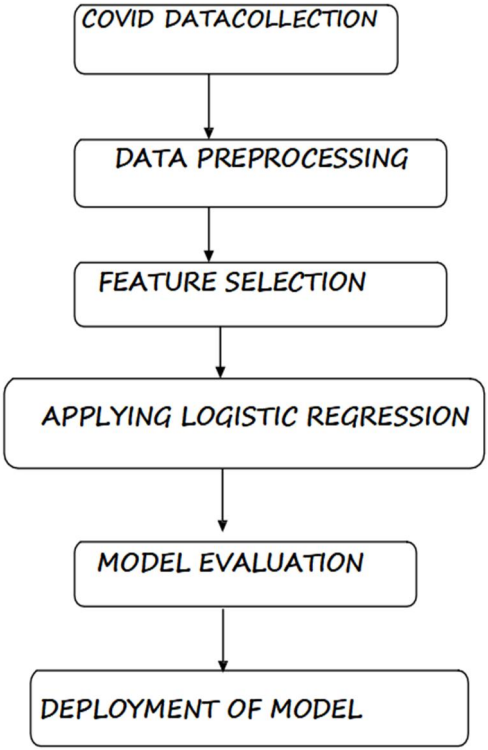
Predicting COVID-19 outcomes using machine learning algorithms like logistic regression and random forest involves using historical data to build models that can forecast disease progression or risk. Here's an overview of how these algorithms can be applied for COVID-19 prediction. Quantity of data, as well as the choice of features, greatly influence the model's predictive performance. Additionally, model evaluation should consider the trade-offs between false positives and false negatives, depending on the specific goals of the prediction task.

Finally, given the evolving nature of the COVID-19 pandemic, it's important to stay updated with the latest research and adapt models as new insights and data become available. This COVID-19 is spread from person to person, Artificial intelligence based electronic devices can play a pivotal role in preventing the spread of this virus. Asthe role of healthcare epidemiologists has expanded, the pervasiveness of electronichealth data has expanded too .The increasing availability of electronic health data presents a major opportunity in healthcare for both discoveries and practical applications to improve healthcare .This data can be used for training

machinelearning algorithms to improve its decision-making in terms of prediction.

2. Schematic Diagram:

The following schematic diagram shows the general workflow of Corona prediction using logistic regression and random forest models.



2.1 Data Collection:

Data Source:

Gather data from reputable sources such as government health agencies (e.g., CDC, WHO), public health organizations, and research institutions. Utilize open data repositories and APIs for COVID-19 data, which may include case counts, testing data, vaccination data, and hospitalization records.

Collect data on COVID-19 cases, recoveries, deaths, and other relevant epidemiological information.Collecting data for COVID-19 prediction using machine learning is a crucial step to build effective models.

Collecting data for COVID-19 prediction using machine learning algorithms is a critical step in building accurate and reliable models. Here are the key considerations for data collection.Gather data from diverse and reliable sources. This can include government health departments, hospitals, clinics, research institutions, and public databases.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
Data = pd.read_csv("/content/drive/MyDrive/corona_tested_006 (2).csv.csv")
Data.head(10)
```

2.2 Data Preprocessing:

Data preprocessing is a critical step in preparing the COVID-19 data for machine learning (ML) algorithms. Proper preprocessing ensures that the data is clean, consistent, and suitable for training predictive models. Clean and preprocess the data to handle missing values, outliers, and inconsistencies. Ensure data quality by performing data validation and verification.Respect privacy and ethical considerations. Anonymize and de-identify personal information to protect individuals' privacy. Adhere to relevant data protection regulations and obtain necessary permissions for data usage.Data collection is an ongoing process, and maintaining the quality and relevance of your dataset is crucial for the success of

COVID-19 prediction models using machine learning algorithms. Regularly review and update your data collection procedures to adapt to changing circumstances and emerging research findings.Collaborate with public health authorities, epidemiologists, and domain experts to ensure that your data collection strategy aligns with the latest research and guidelines.

Dataset may be of different formats for different purposes, such as, if we want to create a machine learning model for business purpose, then dataset will be different with the dataset required for a liver patient. So each dataset is different from 21 another dataset. To use the dataset in our code, we usually put it into a CSV file. However, sometimes, we may also need to use an HTML or xlsx file

2.3 Feature Selection:

Normalize or standardize the numerical features to bring them to a similar scale. Feature scaling is an essential step in the data preprocessing pipeline for many machine learning algorithms, particularly those that rely on distance metrics or gradient-based optimization methods. It ensures that all features have a similar scale, preventing certain features from dominating the learning process due to their larger magnitude.

2.4 Model Training and Evaluation:

Train the Logistic Regression model on the training data. The model will learn the relationship between the features and the binary outcome. Train a logistic regression model on the training set using features and the target variable. The model is trained to maximize the likelihood of the observed outcomes given the input features and the model parameters. Evaluate the model's performance using the test data. Common evaluation metrics for binary classification tasks include accuracy, precision, recall, F1 score, and the ROC curve. Once the model is trained, evaluate its performance on the testing set.Evaluate the model on the test set using metrics such as accuracy, precision, recall, and F1 score.

3. PROPOSED WORK

Machine learning (ML) is a field of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to perform tasks without explicit programming. The primary goal of machine learning is to enable computers to learn from data and make predictions or decisions based on that learning. Machine learning (ML) is a field of artificial

intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to perform tasks without explicit programming. The primary goal of machine learning is to enable computers to learn from data and make predictions or decisions based on that learning. Machine learning algorithms rely on data to learn patterns, relationships, and insights. This data can be structured (e.g., databases) or unstructured (e.g., text, images, audio).Algorithms are trained on a dataset, which is typically divided into two subsets: the training data, used to teach the model, and the testing data, used to evaluate its performance.Models are the mathematical representations that capture patterns in the data. These models can range from simple linear equations to complex neural networks.

3.1 Correlation Analysis:

Calculate and visualize the correlation between features and the target variable (e.g., COVID-19 status, severity). Identify features with high positive or negative correlations.

The Implementation of Logistic Regression Technique

Logistic Regression is a statistical method used for binary classification tasks. Regression analysis can be used to infer causal

relationships between the independent and dependent variables. In the context of COVID-19 prediction, you can use Logistic Regression to predict binary outcomes, such as whether a patient is likely to be hospitalized or whether an individual is at high risk of contracting COVID-19.Logistic regression is a statistical method commonly used for binary classification tasks, making it suitable for COVID-19 prediction. Here's an overview of how logistic regression can be applied to predict COVID-19 outcomes logistic regression to the training data. In this binary classification context, the model will predict one of two outcomes, such as COVID-19 positive (1) or negative (0).logistic regression model estimates the probability of a person being COVID-19 positive based on the selected features.

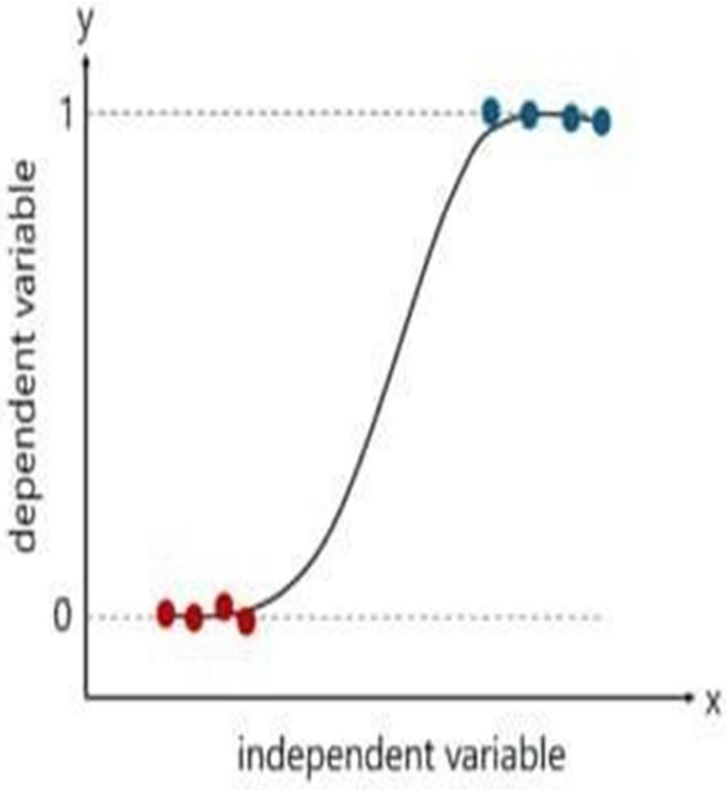
The model learns the coefficients for each feature during the training process.Logistic regression provides interpretable coefficients, which can help understand the impact of each feature on the prediction. Logistic Regression is a statistical method used for binary classification problems, where the outcome variable is categorical and has two classes[1]. Despite its name, logistic regression is a classification algorithm rather than a regression algorithm.

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a 42 categorical dependent variable.

Logistic Regression is a statistical method used for binary classification problems, where the outcome variable is categorical and has two classes. It is named "logistic" because it models the probability of a certain class using the logistic function. Logistic Regression is used when the dependent variable is binary, meaning it has only two possible outcomes (0 or 1, True or False, Yes or No). The logistic regression model predicts the probability that an instance belongs to a particular class. The logistic function (also called the sigmoid function) is used to map any real-valued number to the range [2].

The parameters of a logistic regression are most commonly estimated by maximum-likelihood estimation (MLE). The Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression[3].

Many other medical scales used to assess severity of a patient have been developed using logistic regression.[4] Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease,kidney stones), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.).



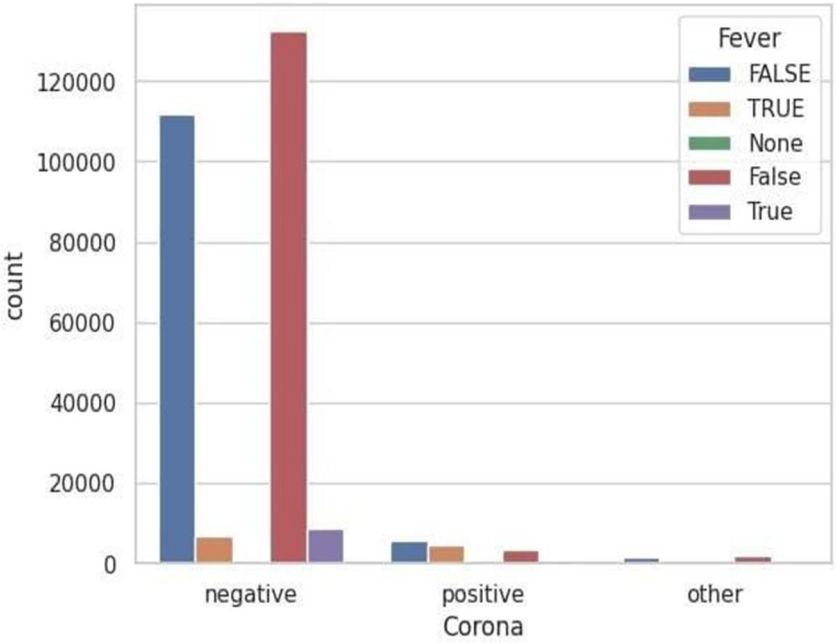
Logistic Regression is suitable for binary classification tasks, such as predicting.

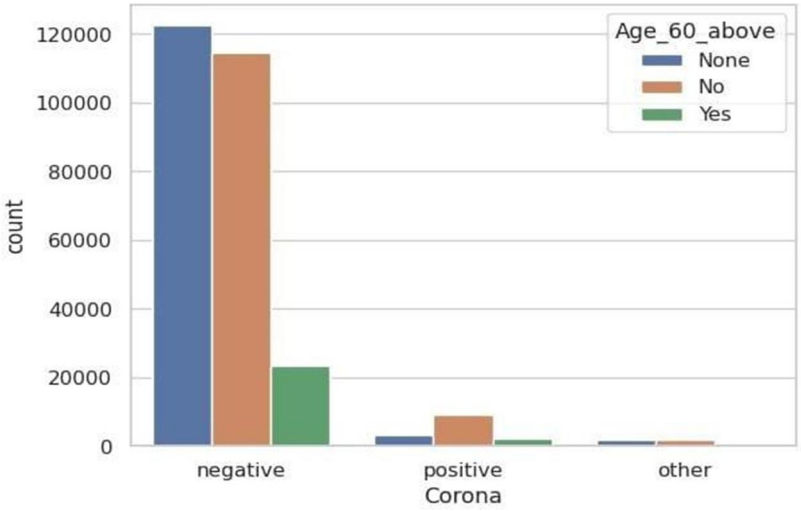
Model: Logistic Regression Accuracy: 0.93

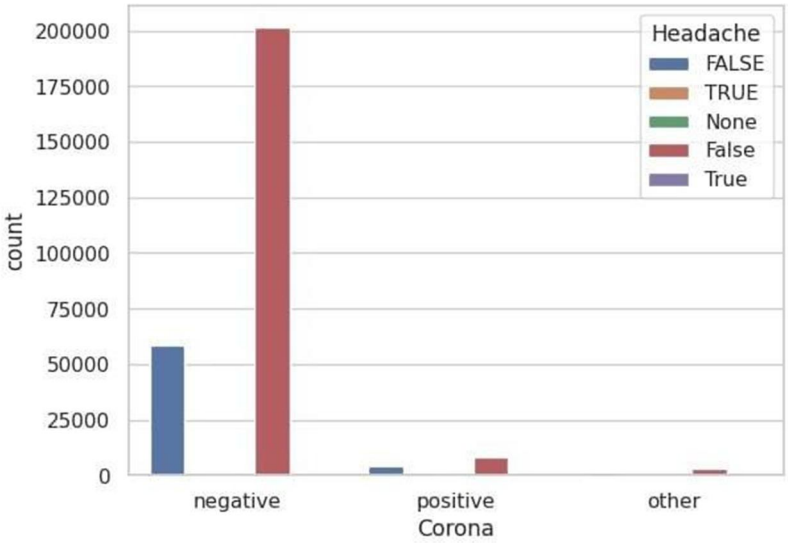
Precision: 0.87

Recall: 0.93

F1 Score: 0.9




Age_60_above
None
No
Yes
120000
100000
80000
60000
40000
20000
0
count
negative
positive
other
Corona


Headache
FALSE
TRUE
None
False
True
200000
175000
150000
125000
100000
75000
50000
25000
0
count
negative
positive
other
Corona

4. CONCLUSION AND FUTURE ENHANCEMENT

In conclusion, machine learning algorithms, including logistic regression and random forest, have demonstrated their usefulness in COVID-19 prediction tasks. Their application extends to predicting outcomes such as infection risk, disease severity, and mortality, contributing to more informed decision-making and improved healthcare responses during the pandemic. As our understanding of COVID-19 evolves and more data becomes available, the refinement and development of machine learning models for COVID-19 prediction will continue to be a valuable and ongoing area of research and application. In future this work can be extended in more number of algorithms with different datasets.

5. BIBLIOGRAPHY

[1] Authors: Imran Razzak, Saeid Anwar, and Paolo Perona Published in: IEEE Access, 2020

[2] Authors: Harmit S. Ranhotra, Mohammad H. Rahbar, and Shouq A. Alawad Published in: IEEE Journal of Biomedical and Health Informatics, 2021

[3] Authors: Rebeca Moraes, Tayana Conte, and Laís CastroPublished in: IEEE Access, 2021

[4] Authors: Alessandro M. Zani, Daniela Aloini, And Francesco Marcelloni Published In: IEEE Access, 2020

◆ ◆ ◆