2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) | 979-8-3503-8944-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ACCAI61061.2024.10602005

Analysis and Prediction of COVID-19 using Machine Learning

Mr. M.Parthiban and Dr. Anna Alphy

[1]Research Scholar, Department of CSE, SRM Institute of Science and Technology , Delhi-NCR Campus, Modi Nagar, Ghaziabad – 201204
[2]Assistant professor, Department of CSE, SRM Institute of Science and Technology ,Delhi-NCR Campus, Modi Nagar, Ghaziabad – 201204

E-mail : pm9590@srmist.edu.in, annaa@srmist.edu.in

Abstract- In order to support the control and containment of the virus, accurate and efficient predictive models must be developed. The COVID-19 pandemic has presented hitherto unheard-of challenges for international health systems. This work offers a novel approach for COVID-19 infection prediction using binary classification models, specifically Random Forest and XGBoost algorithms. These models are well-known for their excellent accuracy and efficiency when handling complicated datasets. This research focuses on utilizing the advantages of ensemble learning techniques to improve prediction accuracy, in contrast to earlier approaches that depended on different machine learning techniques with differing degrees of success. The technique involves a comprehensive preprocessing of the data to generate a dataset containing COVID-19 symptoms and patient demographics. Next, the XGBoost and Random Forest models are used to classify individuals as COVID-19 either beneficial or detrimental. The aim of the research is to determine the predictive power of these models for COVID-19 infections through a comparative analysis of the models under various experimental conditions.

Keywords— COVID-19, XGBOOST, Random-forest, Binary Classification, Data Analysis, Predictive Modelling.

I. INTRODUCTION

The COVID-19 pandemic was started by the SARS- CoV-2 virus, which has had a the economy worldwide. Effective instruments for anticipating and controlling the virus's spread are becoming more and more important in light of this extraordinary disaster. Predictive modelling based on machine learning algorithms is a helpful tactic for aiding the general public, policymakers, and medical experts in understanding and responding to the pandemic. This paper addresses the binary classification problem of COVID-19 infection status prediction using two potent machine learning techniques: Random Forest and XGBoost. For timely and targeted public health interventions, a precise assessment of the COVID-19 infection status is necessary. By applying machine learning algorithms and relevant features, we want to develop robust models that are able to distinguish between persons who are positive or negative examples. The use of Random Forest and XGBoost is justified by their demonstrated effectiveness in handling complex datasets, detecting nonlinear relationships, and generating higher forecast accuracy. The algorithms have complementing strengths, and a comparative analysis will reveal whether one is better suited for the specific purpose of binary COVID-19classification.

II.RELATED WORKS

Many studies have looked into the application of ML methods for binary classification in COVID-19 prediction. This study usually aims to increase diagnostic capabilities and provide decision support to medical professionals. The majority of datasets used by current systems comprise test results, clinical data, demographic data, and symptoms. Public health organizations, healthcare databases, and hospitals are possible sources of these datasets. Many machine learning techniques are used for binary classification, including RF, SVM, XGBoost and Logistic Regression. Sensitivity and specificity are crucial for COVID-19 binary classification. Some systems discuss potential therapeutic applications of the prediction models, such as integration with electronic health records (EHRs) or telemedicine platforms. Real-time resource allocation and decision-making may be aided by integration. In currentsystems, issues and constraints pertaining to ethical considerations, generalization to varied populations, and data quality are frequently explored.

Based on computerized medical records from Tongji Hospital, L. Yan and H.-T. Zhang [1] created a prognostic model with the XGBoost machine learning algorithm to estimate mortality risk in COVID-19 patients, improving early intervention and healthcare resource allocation. In orderto facilitate early intervention and the most efficient use of medical resources, L. Sun, et al. [2] created an SVM prediction model that reliably distinguishes between mild and severe/critical COVID-19 cases based on 36 clinical indicators. This model performed exceptionally well in both training and testing datasets.

The severity can be predicted with over 81% accuracy using a machine learning model developed by H. Yao, et al.

[3] using SVM. Utilizing data from blood and urine tests, this model examines 28 distinct biomarkers, offering insights into their biological implications on disease progression and serving as a means for early risk evaluation. Hu C, Z. Liu, and Y. Jiang [4] constructed a logistic regression model based on four critical dimensions at hospital admission, providing a reliable method to assess mortality risk among severe COVID-19 patients with commendable accuracy.The COVIDC system was created by Wajid [5]. It is an automated system that uses chest CT scans and deep learning to quickly detect and analyze the severity of COVID-19. Its high accuracy outperforms previous approaches and works well in practical diagnostic situations. In order to optimize patient triage during resource shortages, C. An [6] created a ML model utilizing data from this disease patients in South Korea. They found that LASSO and linear SVM were highly effective in predicting death with over 90% sensitivity and specificity.

With a focus on identifying critical lesion markers such ground-glass opacity for improved diagnostic support, Y. Song and S. Zheng [7] created a DL based CT diagnosis system that achieves great performance in reliably.

In order to speed up quarantine and treatment procedures and mitigate the delays and false negatives associated with conventional pathogenic testing, S. Wang and B. Kang[8] suggested using artificial intelligence to analyze CT scans for early COVID-19 identification. Arun Sharma, Sheeba Rani, and Dinesh Gupta [9] improved upon previous approaches and advanced the use of AI-based categorization in biomedical imaging for COVID-19 diagnosis and management by creating effective DL models employing chest X-ray images for quick COVID-19 screening. A DL system for early COVID-19 screening from CT images was created by X. Xu et al. [10]. It achieved 86% accuracy by separating COVID-19 from Influenza-A pneumonia providing a promising additional diagnostic tool for doctors. In a study focusing on classification techniques rather than medical diagnostic accuracy, R Jain, et al. [11] employed deep learning CNN models to analyze chest X-ray pictures for COVID-19 detection and discovered that the Exception model outperformed others with a 97.97% accuracy rate. In order to support medical personnel in making decisions,Ahmet's [12] study on computer-aided detection of COVID-19 by ML algorithms on CT and X-ray images showed high accuracy rates, facilitating quick diagnosis and differentiating COVID-19 pneumonia from other forms.

III.PROPOSED METHODOLOGY

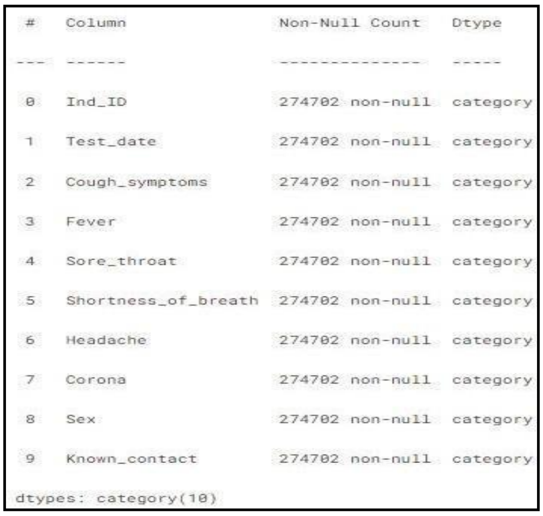
Data Source

This project's dataset was found on the reliable platform Kaggle, which facilitates the exchange and discovery of datasets. The dataset has a number of elements, such as symptoms, pertinent health indicators, and demographic data. Dataset Description: This dataset includes patient symptoms that are essential for diagnosing COVID-19 infections. The nature of the columns is categorical. The columns' specifics are:

- ID (Personal Identification)
- gender (men or women).
- Age 60 years or older (yes/no)
- Test date (date when tested for COVID)
- Cough (1/0).
- Fever (1/0).
- Sore throat (1/0).
- Breath Shortness (1/0).
- Headache (1/0).
- Known contact with an individual confirmed to have COVID-19 (1/0).
- Corona positive or negative.

Data PreProcessing

In our project aimed at predicting COVID-19 infections using Random Forest and XGBoost algorithms, to guarantee ideal model performance, the dataset must be refined, which requires the use of the Data Preprocessing Module. This critical phase entails encoding categorical variables, such as symptoms and demographic information, to make them interpretable for machine learning algorithms, handling missing values to preserve data integrity, and eliminating duplicates to prevent bias.)



```
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   Ind_ID               274702 non-null  category
 1   Test_date            274702 non-null  category
 2   Cough_symptoms       274702 non-null  category
 3   Fever                274702 non-null  category
 4   Sore_throat          274702 non-null  category
 5   Shortness_of_breath  274702 non-null  category
 6   Headache             274702 non-null  category
 7   Corona               274702 non-null  category
 8   Sex                  274702 non-null  category
 9   Known_contact        274702 non-null  category
dtypes: category(10)
```

Fig. 1. Column Values

Fig.1. shows the dataset from Kaggle is categorical and contains a range of COVID-19-related datasets and outcomes, preprocessing involves carefully dividing the data into training sets and testing sets. This part is crucial for assessing the predictive power and generalizability of the models. By implementing these rigorous preprocessing steps,we hope to increase the precision and dependability of our prediction models and help with the early identification and treatment of COVID-19 infections.

Feature Selection and Engineering

In order to forecast COVID-19 infections, our work combines the state-of-the-art MLT RF and XGBoost. The Feature Selection and Engineering step is crucial. In order to determine which aspects in our dataset are most predictive, we first do a thorough domain knowledge study and apply exploratory data analysis (EDA). An extensive analysis is performed on the dataset. It includes a wide range of symptoms, including headache, fever, sore throat, cough, and breath shortness, along with detailed demographic information. Part of this is determining how relevant known interactions are to confirmed COVID-19 cases. To make sure that the algorithms can efficiently learn from the most pertinent data, the goal is to determine which features directlycontribute to the models' predicted accuracy. This procedure is essential because it has a direct impact on the model's capacity to extrapolate from the symptomatology anddemographic profiles found in the data to effectively identifypossible COVID-19 infections.

The methodology described for selecting features using the Chi-squared (χ^2) test in feature engineering, The formula for the Chi-squared statistic is:

$$\chi^2 = \underline{\sum ((Ot - Et_i)^2)}\ /\ Et$$

Ot = Observed frequency
Et = Expected frequency

Moreover, this phase's feature engineering component enables the creative development of additional variables that may reveal more complex correlations between the symptoms and COVID-19 infection rates. To add more complex prediction signals to the dataset, specific symptoms or demographic characteristics can be synthesized to identify greater risk clusters. Each feature's significance is carefully evaluated, and its impact on the model's performance isquantified, using tools included in Random Forest and XGBoost.These valuable insights enable us to prioritize essential variables as we refine our feature set. Employing a logical and empirical method in feature engineering andselection enhances the accuracy of COVID-19 predictions, thereby improving the model's overall predictive performance. Our primary goal is to build a highly accurate model that equips medical professionals with vital data to effectively manage and mitigate the COVID-19 crisis.

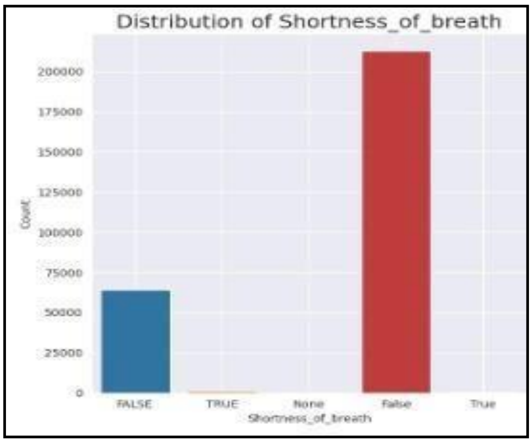


Fig. 2. Breath Shortness Distribution

Fig.2. distribution of breath shortness among COVID-19 patients depicts a significant trend in the manifestation of respiratory symptoms. Analysis of the data reveals that breath shortness is a prevalent symptom experienced by a considerable portion of individuals infected with the virus, spanning across various age groups and demographics. The graph illustrates a spectrum of severity and frequency of breath shortness, ranging from mild to severe cases.

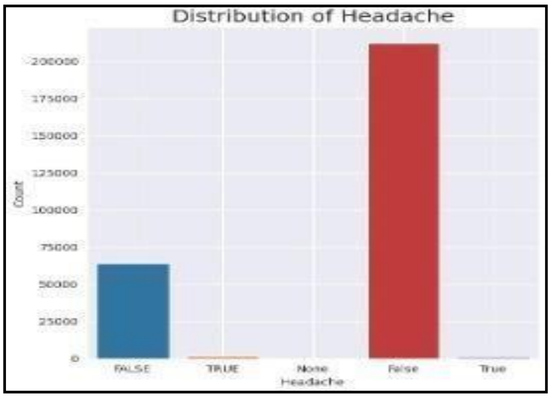


Fig. 3. Headache Distribution

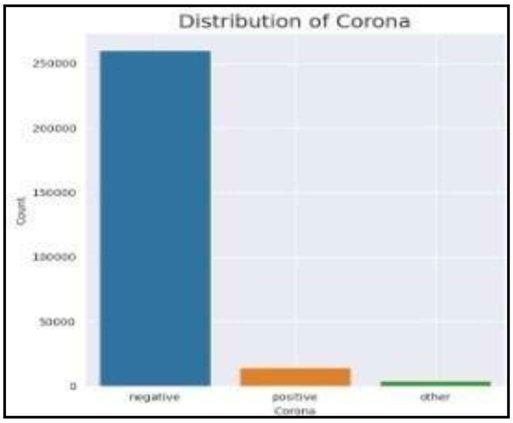


Fig. 4. Corona Distribution

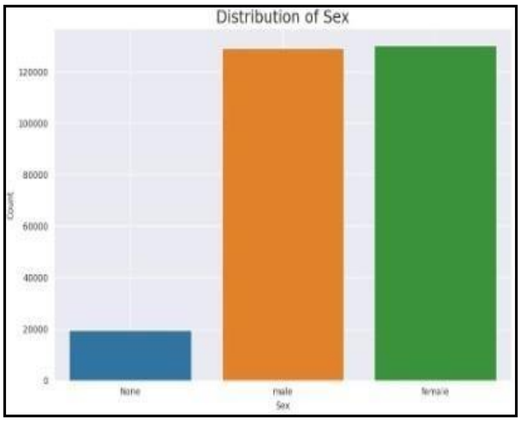


Fig. 5. Sex Distribution

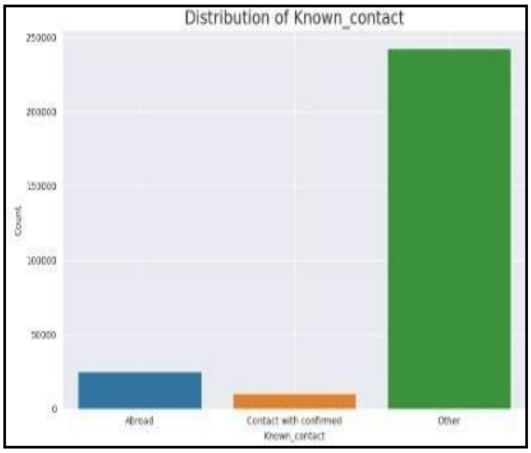


Fig. 6. Known Contact Distribution

Model Selection

Selecting COVID-19 prediction models involves a critical evaluation of the merits and drawbacks of Random Forest and XGBoost algorithms for binary classification tasks. Random Forest, an ensemble technique, constructs numerous decision trees during training and determines the class mode for classification.. It works remarkably well withcomplicated datasets that have multiple attributes, and it is resistant to overfitting. Random Forest's capacity to offer insights into feature relevance is very helpful in figuring out the major factors affecting COVID-19 prediction accuracy. Conversely, XGBoost, a gradient boosting technique, constructs an ensemble of weak learners in a stepwise manner by adding trees that repair the faults of the preceding ones and minimizing a loss function.

Notable for its regularization techniques that help minimize overfitting, XGBoost often outperforms other algorithms in terms of projected accuracy. Its adaptability in addressing class imbalances is crucial in the situation of COVID-19 binary classification, in which the percentage of positive cases may be substantially lower than that of negative cases. The choice to employ both Random Forest and XGBoost stems from their complimentary advantages: Random Forest offers superior accuracy and manages imbalanced data, while XGBoost offers robustness and interpretability. The combination of two models offers a synergistic method that improves prediction performance in the difficult COVID-19 binary classification job by utilizing the advantages of both algorithms.

Model Training

Training of the COVID-19 Prediction Model in Binary Classification using Random Forest and XGBoost, a rigorousprocess is employed to ensure that the algorithms learn from the provided data and adapt well to new situations. When employing Random Forest, bootstrapped sampling is used totrain each decision tree on a subset of the training set once a certain number of decision trees have been initialized. To encourage variation among the trees, random subsets of features are chosen at each node during the training phase to build each tree. The final categorization is then determined by adding up all of the predictions made by the trees using a voting method. An ensemble of decision trees is incorporatedinto the process, producing a reliable and accurate model.

Similar to this, XGBoost training begins with initializing an XGBoost classifier, after which a predetermined number of decision trees—also known as weak learners—are introduced to the ensemble one after the other. By addressing the mistakes committed by its predecessors, each succeeding tree maximizes a given loss function. To ensure that the model works well when applied to new data and to assist prevent overfitting, a regularization component is integrated into the XGBoost objective function. The learning rate, which controls each tree's contribution to the ensemble, is one important hyperparameter that has an impact on the training process.

Both Random Forest and XGBoost use cross-validation during training to assess their resilience and modify hyperparameters. The dataset is split up into multiple subsetsfor cross-validation purposes. The model is trained on one ofthe subsets, and its performance is valuated with remaining data. This is repeated multiple times to provide a more precise evaluation of functionality and ability to generalize. To optimize the prediction performance of the models, the cross- validation results are used to adjust the hyperparameters of both methods, which include the number of trees, maximum depth, and learning rate. For COVID-19 Prediction Binary Classification, Random Forest and XGBoost are both optimized by meticulous manipulation of training parameters to yield optimal classification results.

Model Evaluation

During the evaluation phase of our COVID-19 Prediction Binary Classification project, we meticulously assess the Random Forest and XGBoost models' performance using various metrics, including accuracy, precision, recall, F1 score, and AUC-ROC. This evaluation methodology enables us to comprehensively analyze each model's overall performance in accurately predicting COVID-19 cases. Accuracy assesses the models' ability to predict outcomes correctly overall, while precision focuses on their capability to correctly identify positive cases. The F1 score, which harmonizes recall and precision, provides a holistic perspective on the model's performance. Recall evaluates how effectively the models can capture all positive instances. The AUC-ROC statistic, which evaluates the models' ability to discern between plus and minus scenarios, highlights the models' discriminatory power.

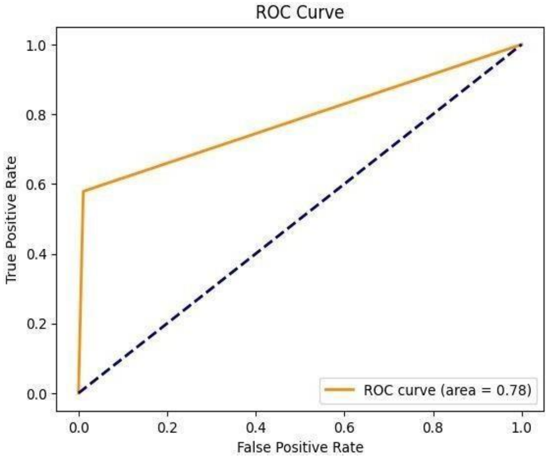


Fig.7. ROC curve

Fig.7. illustrates ROC curve over true and false positive rates. After the completion of training, predictions on the test set are generated using the syntax y_pred_rf = rfc_best.predict(x_test). The evaluation process calculates the accuracy of the model by dividing the number of correctly predicted instances by the total number of instances in the dataset.The formula for accuracy is

$$A = u(p) + u(n) / u(p) + u(n) + s(p) + s(n)$$

u(p) is true positives u(n) is true negatives s(p) is false positives s(n) is false negatives

🟦 🟧 ⬜

Our selection of models for implementation is based onthe information gathered from this comprehensive evaluation. We identify the model that best satisfies the project's objectives of reliably and precisely predicting COVID-19 infections by comparing the performance of the Random Forest and XGBoost models with respect to these critical characteristics. This systematic and data-driven approach ensures that the chosen model is appropriate for real-world application and can significantly support public health campaigns by enabling the timely and accurate identification of potential COVID-19 cases.

V.Performance Analysis

Context of the ongoing COVID-19, the development and evaluation of predictive models for infection status classification are paramount to enhancing public health responses and management strategies. Among the contenders, the Random Forest and XGBoost models stand out for their robust performance in binary classification tasks,distinguishing individuals based on their COVID-19 infection status. Thoroughly assessing these models involves utilizing a range of metrics including accuracy, precision, recall, F1 score, and AUC-ROC, offering a comprehensive understanding of their predictive prowess. Accuracy is crucial for gauging overall effectiveness. TheF1 score, which balances precision and recall, provides a consolidated measure of the models' performance, while the AUC-ROC curve evaluates their ability to distinguish between positive and negative cases across different threshold settings.

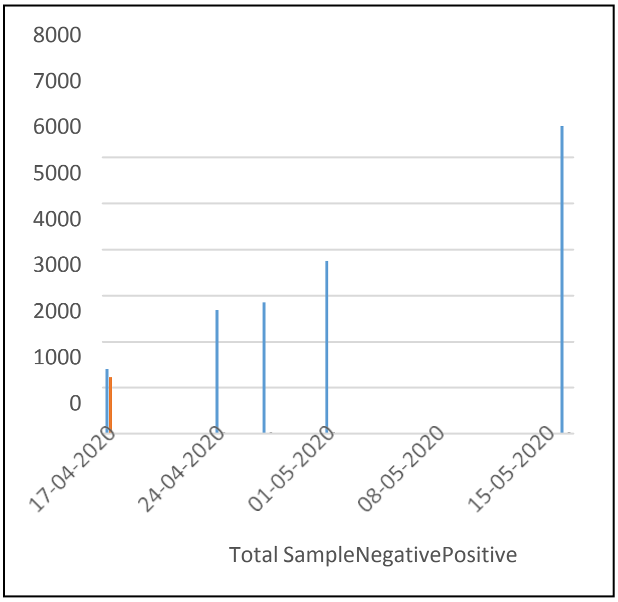


Fig. 8. Statewise Analysis

Fig.8. shows the analysis of the samples over days. The comparative analysis between the Random Forest and XGBoost models reveals a slight edge for XGBoost in termsof accuracy and efficiency. Boasting an accuracy rate reaching 98.81%, XGBoost not only showcases exceptional predictive prowess but also boasts optimized training times, rendering it a pragmatic choice in swift public health environments where timely decisions are paramount. This nuanced evaluation highlights the models' potential to notably aid in early COVID-19 detection and management, facilitating more precise and efficient containment and treatment approaches.

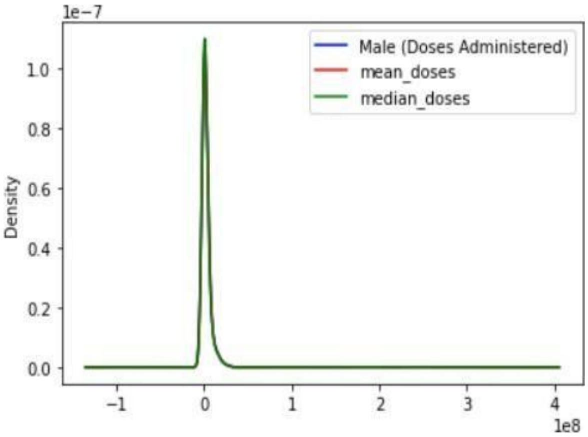


Fig. 9. Doses Administered

Fig.9. shows the doses administrated over the density. However, the evaluation also brings to light the inherent challenges and limitations associated with deploying machine learning models in the realm of infectious disease prediction. Potential biases within the training data, stemming from uneven sampling or demographic disparities, can skew the models' predictions, necessitating careful consideration and adjustment. Furthermore, the dynamic nature of the COVID-19 virus, with evolving strains and variable host responses, adds layers of complexity to the predictive modeling effort. Recognizing these challenges, the analysis underscores the importance of continuous model refinement and validation against emerging data. By doing so, the predictive models can remain relevant and accurate, offering valuable tools against this pandemic and future public health crises.

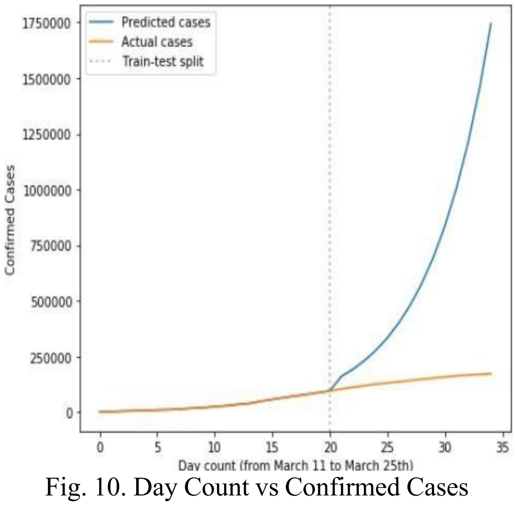


Fig. 10. Day Count vs Confirmed Cases

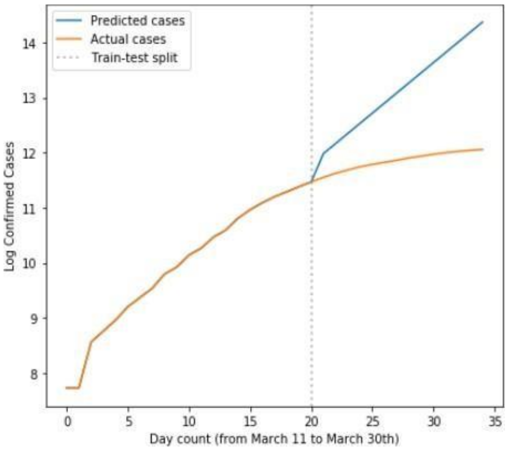


Fig. 11. Day Count vs Log Confirmed Cases

Fig.11. illustrates the day count and the confirmed cases of COVID-19 over a month which increases.

VI.FUTURE ENHANCEMENT

Improvements to the predictive power of models used for COVID-19 prediction, specifically when employing Random Forest and XGBoost algorithms for binary classification, are on the horizon with potential enhancements through sophisticated machine learning techniques. One pivotal area of advancement is in hyperparameter tuning, where the application of Bayesian optimization could significantly refine the optimization of model parameters, leading to enhanced performance. Additionally, an expanded approach to ensemble learning could offer substantial benefits. By integrating predictions from a broader array of machine learning algorithms beyond the current use of Random Forestand XGBoost, it's possible to leverage synergies between different models, potentially elevating the overall accuracy of COVID-19 predictions. This method of combining diverse algorithms can create a more robust model capable of capturing nuances in the data that single models might miss. Feature engineering presents a possible path for improving the efficacy of COVID-19 predictive models. Models can obtain a deeper, more nuanced picture of the factors impacting COVID-19 infection outcomes by adding new data sources and using state-of-the-art approaches to more efficiently extract and utilise information from existing data variables. This method solves one of the main problems in machine learning: making sure the model has access to relevant, high-quality data. It also improves the models' capacity to recognise patterns and anticipate outcomes. With enhancements including sophisticated hyperparameter tuning, extended ensemble techniques, and creative feature engineering, COVID-19 prediction employing RandomForest and XGBoost models appears to have a bright future with room for major gains in efficacy and accuracy.

VII.CONCLUSION

To sum up, the COVID-19 Prediction Binary Classification project has shown encouraging results in correctly classifying people according to their chance of having COVID-19 infection. This has been achieved by utilizing the Random Forest and XGBoost algorithms. Strong performance metrics demonstrated the effectiveness of both models in binary categorization. The models' interpretability, especially when combined with XGBoost, enhanced the transparency of decision-making processes by offering insightful information about the variables affecting predictions. This result indicates that both Random Forest and XGBoost models demonstrate exceptional performance, with XGBoost slightly outperforming Random Forest in terms of accuracy and training time, achieving an accuracy rate of up to 98.81%. Even if the project's performance affects decision-making, resource allocation, and public health, it's important to recognize its limits, including potential biases inthe data or model limitations. Further studies should look into incorporating fresh datasets, utilizing advanced ensemble techniques, and adjusting the models to take into consideration fresh variations.

REFERENCES

[1] L. Yan, H.-T. Zhang, Y. Xiao et al., "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan", medRxiv, 2020.

[2] L. Sun, F. Song, N. Shi et al., "Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID- 19", Journal of Clinical Virology, vol. 128, pp. 104431, 2020.

[3] H. Yao, N. Zhang, R. Zhang et al., "Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests", Frontiers in Cell and Developmental Biology, vol. 8, pp. 1-10, 2020..

[4] Wajid Arshad Abbasi and Syed Ali Abbas, "COVIDC: An expert system to diagnose COVID-19 and predict its severity using chest CT scans: Application in radiology", Informatics in Medicine Unlocked, vol. 23, 2021, ISSN 2352-9148.

[5] An C, D.-W. Kim, J. H. Chang, Lim .H Y. J. Choi and S. W. Kim, "Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study", Scientific Reports, vol. 10, pp. 18716, 2020.

[6] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, et al., Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT imagesmedRxiv, 2020.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[7] Wang S, J. Ma, Zeng X, Kang B, M. Xiao, J. Guo, et al., "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)", medRxiv, 2020.

[8] Arun Sharma, Sheeba Rani and Dinesh Gupta, "Artificial Intelligence- Based Classification of Chest X-Ray Images into COVID-19 and Other Infectious Diseases", International Journal of Biomedical Imaging, vol. 2020, pp. 10, 2020.

[9] R Jain, M Gupta, S. Taneja et al., "Deep learning based detection & analysis of COVID-19 on chest X-ray images", Appl Intell, vol. 51, pp. 1690-1700, 2021.

[10] Ahmet "A new approach for computer-aided detection of coronavirus (COVID-19) from CT and X-ray images using machine learning methods", Applied Soft Computing, vol. 105, pp. 107323, 2021, ISSN 1568-4946.

[11] Harris E. CDC assesses risk from B.A2.86, highly mutated COVID-19 variant.JAMA.2023;330:1029

[12] orter AK, Kleinschmidt SE, Andres KL, Reusch CN, Krisko RM, Taiwo OA, et al. Antibody response to COVID-19 vaccines among workers with a wide range of exposure to per- and polyfluoroalkyl substances. Environ Int. 2022;169: 107537.

[13] Vengatesan K , A Kumar , M Parthiban "Analysis of Miral botnet malware issues and its prediction methods in internet of things"Conference on computer networks , 2020

[14] A Kumar , K Vengatesan , R Rajesh , M Parthiban ,"Review of gene subset selection using modified k-nearest neighbor clustering algorithm." Conference on Smart Systems and Inventive technology , 2018.

[15] Dillon MJ, Eleftheriou D, Brogan PA. Medium-size-vessel vasculitis. Pediatr Nephrol. 2010;25:1641–52.