Nile University
جامعة النيل

CBIO 313

Predicting COVID-19 Status Using Machine Learning on Clinical and Laboratory Data

Name: Ashraqat Mohamed Abdelhamid

ID: 221000836

Under The supervision of: Dr. Muhammed Elsayeh

TA: Malak Abdel Monsef

Table Of Contents

Abstract

This study investigates the application of supervised machine learning techniques to predict COVID-19 status (Healthy or Diseased) based on a structured dataset comprising clinical and laboratory parameters. A Random Forest classifier, along with Logistic Regression and Support Vector Machine (SVM), was trained and evaluated to determine the most effective model. Feature engineering was employed to enhance predictive performance, and all models were assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Exploratory Data Analysis (EDA) was conducted to identify relevant patterns. The final and best-performing model was deployed using a Streamlit-based web application to allow real-time predictions. This work demonstrates how machine learning can support early diagnostic processes in healthcare.

1.0 Introduction

The COVID-19 pandemic has emphasized the need for rapid and reliable diagnostic tools, especially in resource-constrained environments. Traditional diagnostic methods, while accurate, may be time-consuming and dependent on laboratory infrastructure. Machine learning provides an opportunity to augment clinical decision-making using existing patient data. By leveraging historical and real-time health data, machine learning models can provide immediate insights and support physicians in identifying high-risk patients.

In the context of COVID-19, where the virus affects multiple organ systems and presents with a wide range of symptoms, relying solely on a single diagnostic modality can be limiting. Blood biomarkers and vital signs offer a rich source of information that, when analyzed using computational methods, can improve diagnostic accuracy. The ability to automate prediction based on these features can help healthcare systems scale their response, especially when facing patient surges.

The objective of this project is to develop and compare machine learning models to predict COVID-19 status based on clinical and laboratory features. The study investigates Logistic Regression, Support Vector Machine, and Random Forest algorithms to identify the most effective classifier. The selected model is then deployed via a web application to demonstrate practical applicability. Through this work, we aim to demonstrate how an end-to-end machine learning pipeline can be constructed, from data preprocessing and exploratory analysis to model training, evaluation, and deployment.

2.0 Materials & Methods

2.1 Dataset Description

The dataset used in this study consists of clinical and laboratory records containing sixteen attributes. These include vital signs such as age, temperature, heart rate, respiratory rate, and oxygen saturation. In addition, various blood parameters such as WBC count, RBC count, hemoglobin, platelet count, D-dimer, CRP, ferritin, LDH, and ALT are included. The target variable, COVID_status, is categorical and represents whether a patient is Healthy or Diseased.

2.2 Data Preprocessing and Feature Engineering

Initial exploration confirmed that there were no missing values. The categorical target variable was encoded using label encoding to convert 'Healthy' and 'Diseased' into binary numerical values. A new feature, referred to as inflammation_index, was engineered by computing the average of CRP, Ferritin, and LDH values. This index was designed to represent the overall inflammatory state of a patient. To standardize the input values, all numerical features were scaled using StandardScaler to ensure consistency in model input.

2.3 Exploratory Data Analysis (EDA)

Various statistical and visual analysis techniques were applied to better understand the distribution and relationships among features. Histograms and boxplots were used to analyze the distribution and variance of each feature relative to COVID status. A correlation heatmap revealed strong relationships among inflammatory markers and the target variable. Scatter plots of respiratory indicators further supported that oxygen saturation was generally lower among diseased individuals. EDA supported the hypothesis that certain features, particularly those related to inflammation and oxygen levels, are significant indicators of COVID-19 status.

2.4 Model Development

The task of predicting COVID-19 status was approached as a binary classification problem. The dataset was split into training and testing subsets in an 80:20 ratio. Three machine learning algorithms were selected for initial experimentation: Logistic Regression, Support Vector Machine (SVM), and Random Forest Classifier. Each model was trained on the scaled dataset, and their performances were evaluated using a consistent set of metrics, including

accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). The comparison of these models enabled us to determine which classifier was best suited for the prediction task.

2.5 Model Evaluation and Fine-Tuning

After evaluating the initial performance of all three models, Random Forest emerged as the best-performing algorithm. It consistently achieved higher values in precision, recall, and AUC-ROC across both training and test sets. Consequently, Random Forest was selected for further optimization. Hyperparameter tuning was conducted using GridSearchCV, where a range of values for the number of estimators and the maximum depth of the trees was systematically tested.

2.6 Deployment

After finalizing the best-performing model, it was serialized using joblib along with the associated scaler object. A user-friendly web application was developed using Streamlit to allow users to input patient data and receive instant predictions. The application automatically recalculates the engineered inflammation index and scales the inputs before making predictions. This deployment illustrates how machine learning can be effectively translated from development environments to practical tools.

3.0 Results & Discussion

3.1 Preprocessing and Feature Engineering

The data preprocessing phase confirmed that the dataset was clean, with no missing values or obvious inconsistencies. Label encoding was successfully applied to the categorical target variable, converting it into a numerical format suitable for modeling. The engineered inflammation_index, derived from CRP, Ferritin, and LDH values, aimed to capture the cumulative effect of inflammatory activity in the body. This feature was later confirmed to hold strong correlation with the target class and proved influential in the models' predictions. Standardization of features ensured that all variables contributed evenly to the machine learning models.

3.2 Exploratory Data Analysis

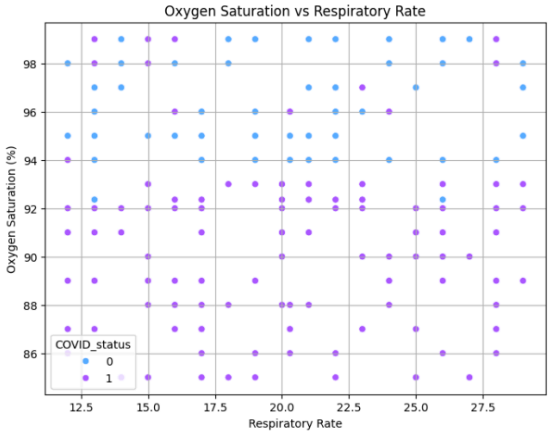

Figure 1: Scatter plot showing the relationship between respiratory rate and oxygen saturation colored by COVID-19 status.

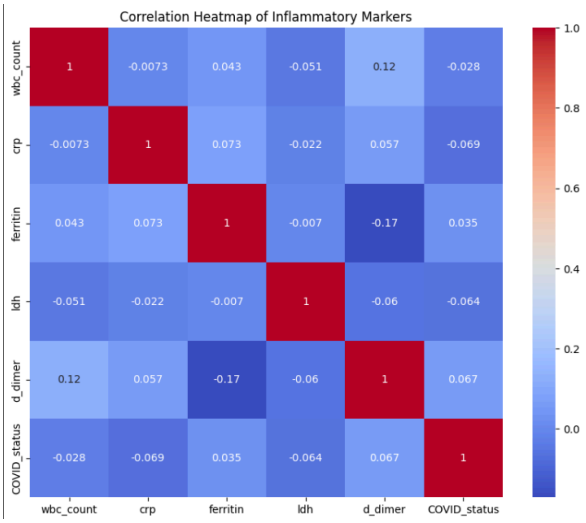

Figure 2: Correlation heatmap among inflammatory markers (CRP, Ferritin, LDH, etc.) and COVID status.

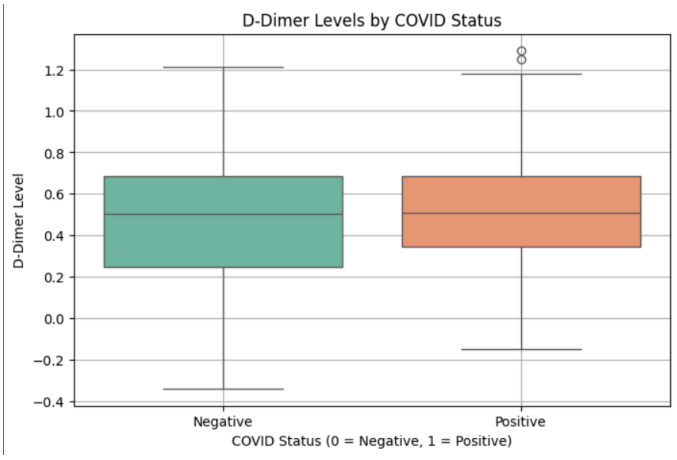

Figure 3: Boxplot comparing D-Dimer levels between Healthy (0) and Diseased (1) patients.

EDA revealed valuable insights into the dataset. As shown in Figure 1, oxygen saturation values were generally lower in diseased individuals, while respiratory rate was often higher. This inverse pattern suggests their combined diagnostic relevance. Figure 2 highlights a correlation heatmap of inflammatory markers. Although no extremely strong correlations were observed, modest relationships between CRP, Ferritin, LDH, and the COVID status indicated that these features contribute independently to model performance. Figure 3 shows the distribution of D-Dimer levels across classes. Diseased patients tended to exhibit higher levels and broader ranges, reinforcing the significance of D-Dimer as a predictive biomarker.

3.3 Evaluation of Initial Models

Upon Evaluating the accuracies and performance of the models; Logistic Regression, Random Forest and Support Vector Machine, we saw the following results:

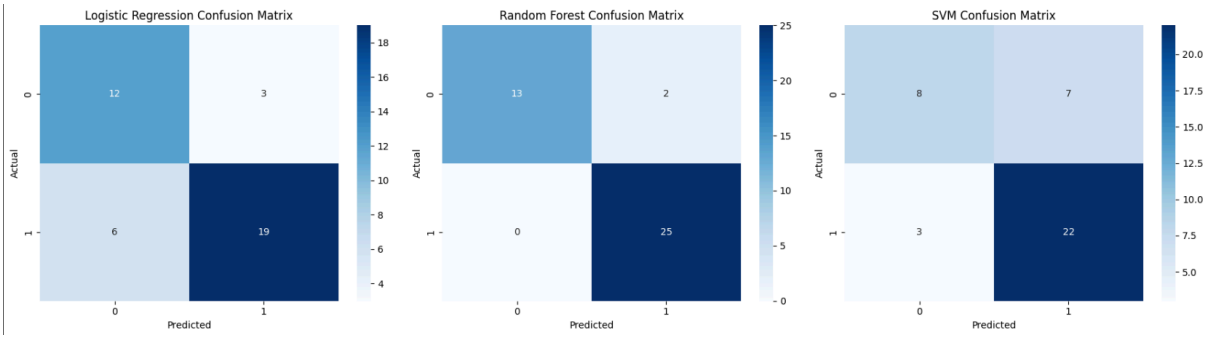

Figure 4: Confusion Matrix for the 3 models

```
Model: Logistic Regression
              precision    recall  f1-score   support

           0       0.67      0.80      0.73        15
           1       0.86      0.76      0.81        25

    accuracy                           0.78        40
   macro avg       0.77      0.78      0.77        40
weighted avg       0.79      0.78      0.78        40


Model: Random Forest
              precision    recall  f1-score   support

           0       1.00      0.87      0.93        15
           1       0.93      1.00      0.96        25

    accuracy                           0.95        40
   macro avg       0.96      0.93      0.95        40
weighted avg       0.95      0.95      0.95        40


Model: SVM
              precision    recall  f1-score   support

           0       0.73      0.53      0.62        15
           1       0.76      0.88      0.81        25

    accuracy                           0.75        40
   macro avg       0.74      0.71      0.72        40
weighted avg       0.75      0.75      0.74        40
```

Figure 5: Classification report for the 3 models

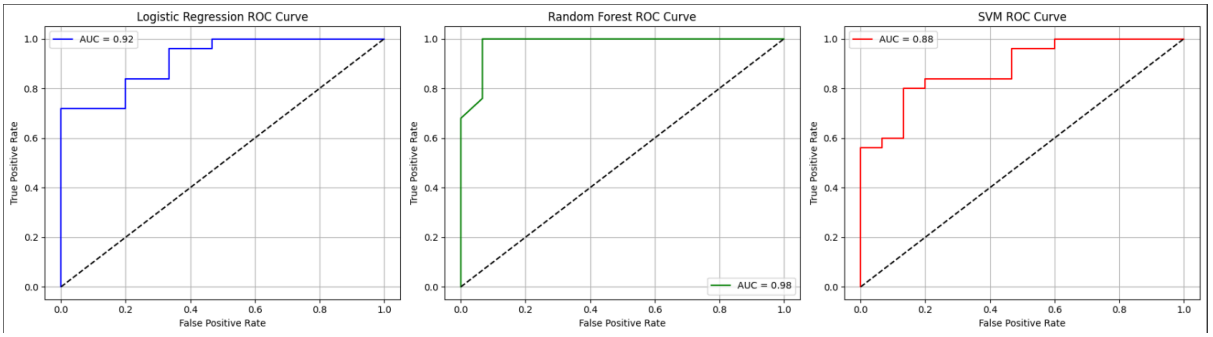

Figure 6: AUC-ROC plots

Logistic Regression achieved an accuracy of 0.78 and an AUC-ROC score of 0.92. SVM, while showing stronger recall, resulted in a lower overall accuracy of 0.75 and an AUC-ROC of 0.88. Random Forest outperformed both with an accuracy of 0.95 and an AUC-ROC of 0.98. It also provided superior F1-scores, reflecting its ability to balance precision and recall effectively. These findings guided the selection of Random Forest as the model for further tuning.

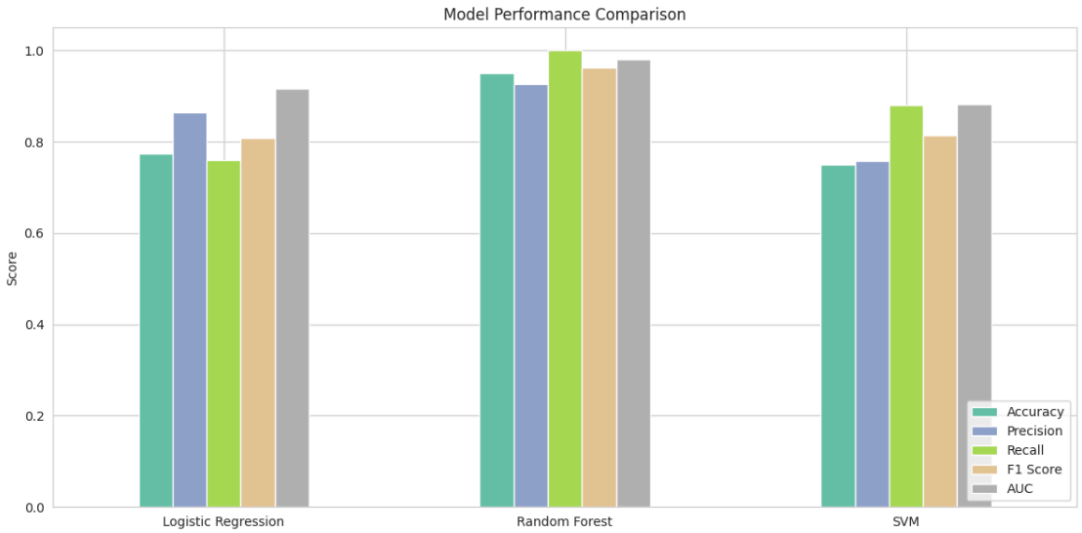

Figure 7: Bar graph showing the difference in performance between the 3 models, this graph shows the superiority of Random Forest across all evaluation metrics.

3.4 Hyperparameter Tuning Results

Using GridSearchCV, a grid of parameter combinations was evaluated, focusing primarily on the number of trees (estimators) and the depth of each decision tree.

```
Classification Report (Tuned Random Forest):
              precision    recall  f1-score   support

           0       1.00      0.87      0.93        15
           1       0.93      1.00      0.96        25

    accuracy                           0.95        40
   macro avg       0.96      0.93      0.95        40
weighted avg       0.95      0.95      0.95        40
```

Figure 8: Classification report for the tuned Random Forest Model

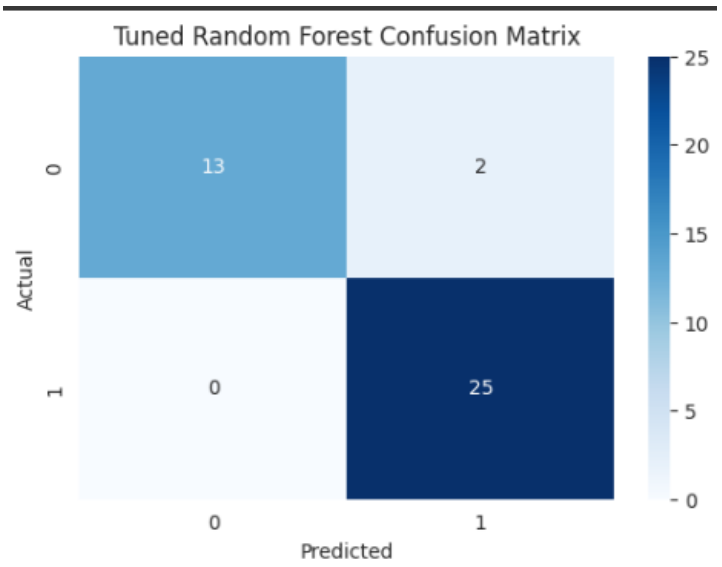

Figure 9: Confusion Matrix describing the classification power of the tuned model.

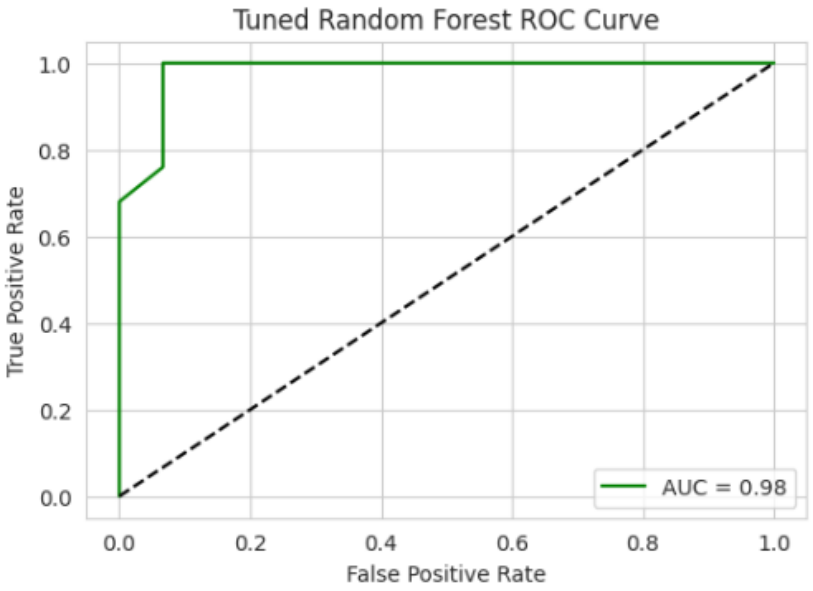

Figure 10: AUC-ROC plot for the tuned model

Using GridSearchCV, a grid of parameter combinations was evaluated, focusing primarily on the number of trees (estimators) and the depth of each decision tree. Although the process did not lead to measurable improvements in accuracy, precision, recall, or AUC-ROC, it reinforced the model's reliability and stability. The lack of significant change suggests that the model was already operating under near-optimal hyperparameter settings, and further tuning only validated its robustness.

These consistent scores highlight the robustness of the model and affirm that the initial configuration was highly effective for the dataset used. Visual outputs such as the confusion matrix and ROC curve underscore the reliability of this final model.

3.5 Deployment and Application Use

To translate the trained model into a practical diagnostic tool, the final Random Forest classifier was deployed using a Streamlit web application. The app allows users to input key patient features, including vital signs (such as age, temperature, respiratory rate, and oxygen saturation) and laboratory measurements (including CRP, D-dimer, ferritin, LDH, and others). Upon submission, the app automatically computes the inflammation index, applies the same scaling used during training, and returns a prediction indicating whether the patient is likely COVID-positive or negative. It also displays the model's confidence score. This interface provides an accessible and interpretable way for clinicians or users to obtain real-time predictions based on clinical data. The deployed model maintains an accuracy of 95%, demonstrating its robustness and reliability when integrated into a user-facing system.

4.0 Limitations & Future Work

While the model exhibits strong performance on the available dataset, it is important to recognize the limitations inherent in this study. The dataset is relatively small and may not reflect the full variability present in diverse populations. Additionally, the dataset does not include information about symptoms, comorbidities (the presence of two medical conditions), or demographic variables such as gender and ethnicity, which could further improve model performance.

In future work, larger and more diverse datasets should be incorporated. Further improvements can include the use of ensemble models and interpretability tools such as SHAP to enhance transparency. Integration with electronic health records (EHRs) and testing in real-world clinical settings would also increase the model's applicability.

5.0 Conclusion

The analysis presented in this project highlights the practical value of applying machine learning techniques to clinical and laboratory data for the prediction of COVID-19 status.

Through methodical data preprocessing, insightful feature engineering, and comprehensive evaluation of multiple classification models, the study identified Random Forest as the most effective model for the task. Despite attempts at fine-tuning, the original configuration already delivered robust results, underlining the model's suitability.

Furthermore, the deployment of the model using Streamlit illustrates the practical feasibility of turning academic insights into user-accessible tools. This application allows healthcare professionals or researchers to quickly obtain predictions based on patient data, showcasing a promising direction for integrating ML models into clinical workflows.

This study reaffirms the broader potential of data analysis and machine learning in supporting evidence-based decision-making in healthcare. As datasets grow in size and complexity, the ability to process and interpret them through ML becomes increasingly vital. These tools not only enhance diagnostic speed and precision but also offer scalability and cost-efficiency, which are critical during public health crises like the COVID-19 pandemic.

Appendices

GitHub Repository

Streamlit APP