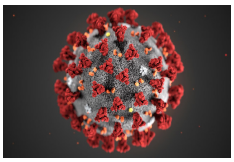
COVID-19 Data Analysis and Prediction: A Machine Learning Approach



Elaborated By :
Fatma BEN REJEB

Contents

Google Colab Link Data Source

1 Introduction

This project explores the application of machine learning models to predict COVID-19 diagnosis based on anonymized patient laboratory data from the Hospital Israelita Albert Einstein in São Paulo, Brazil. The analysis includes exploratory data analysis (EDA), preprocessing, and predictive modeling. The primary goal is to classify patients as COVID-19 positive or negative using their clinical and laboratory test results.

2 Problem Statement

Early and reliable detection of COVID-19 is crucial to reduce transmission and improve patient outcomes. RT-PCR tests, although the gold standard, may be time-consuming and resource-intensive. This project aims to investigate whether routine blood and laboratory test results can be leveraged by machine learning models to predict SARS-CoV-2 infection status efficiently.

3 Data Overview

The dataset contains information about 5,644 patients with 111 variables, including hospital admission variables, clinical outcomes, and laboratory results. Key features include:

- **SARS-CoV-2 exam result (Target)**: Indicates whether the patient tested positive or negative.
- **Patient age quantile**: Represents age groups divided into 20 quantiles for privacy (0 = youngest 5%, 19 = oldest 5%).
- **Patient admitted to intensive care unit (ICU)**: 1 if admitted, 0 otherwise.
- **Patient admitted to semi-intensive unit**: 1 if admitted, 0 otherwise.
- **Hemoglobin**: Oxygen-carrying protein level in blood.
- **Leukocytes**: Total white blood cell count, indicator of immune response.
- **Lymphocytes**: White blood cells critical for viral defense, often reduced in COVID-19 cases.
- **Platelets**: Key role in blood clotting and inflammation.
- **Respiratory Syncytial Virus**: Test result (positive/negative).

- **Influenza A**: Test result (positive/negative).

The dataset includes both categorical and numerical features, many of which contain missing values. Appropriate handling of missing data is a critical step in the preprocessing pipeline.

4 Methodology

4.1 Exploratory Data Analysis (EDA)

The dataset contains a large proportion of missing values across laboratory features. To ensure meaningful analysis, only variables with less than 90% missing values were retained. This reduced the dataset to two main feature groups:

- **Viral test variables:** These features exhibit between 76% and 89% missing values. Thus, viral test results are available for approximately 11–24% of patients.
- **Blood test variables:** These features exhibit more than 89% missing values. Consequently, blood test results are available for fewer than 11% of patients.

Overall, only about 10% of patients in the full dataset tested positive for SARS-CoV-2, which highlights the class imbalance problem for machine learning models.

4.1.1 Distribution and Missing Values

- A heatmap of missing values revealed that most viral and blood test variables had high sparsity.
- After filtering, only 1,350 rows remained for viral test variables (92% negative cases, 8% positive), and 600 rows for blood test variables (87% negative, 13% positive).

4.1.2 Viral Test Analysis

- Cross-tabulation heatmaps showed that co-infections with multiple viruses are rare.
- Most individuals who tested positive for another respiratory virus (e.g., Influenza A, RSV) tended to test negative for SARS-CoV-2, suggesting limited overlap between COVID-19 and other viral infections in this dataset.

4.1.3 Blood Test Analysis

- Distribution plots revealed differences between COVID-positive and COVID-negative patients in key blood parameters, notably:
  - **Platelets:** Lower platelet counts observed in positive cases.
  - **Leukocytes:** Reduced white blood cell counts in positive cases.
  - **Monocytes:** Differences observed across positive and negative groups.
- Correlation analysis showed that:
  - Red Blood Cells, Hematocrit, and Hemoglobin are strongly correlated.
  - Mean Corpuscular Hemoglobin (MCH) and Mean Corpuscular Volume (MCV) are also closely related.

4.1.4 Age Quantile Analysis

- A countplot of patient age quantiles showed that younger individuals appear less represented among COVID-19 positive cases in this dataset.
- However, since children can also be affected by the virus, this observation should be validated with larger, more balanced samples.

4.1.5 Hospital Admission Status

A new categorical feature `status` was engineered to capture hospitalization level:

- Supervision (regular ward admission)
- Semi-intensive care
- Intensive care
- Unknown

Distribution plots of blood parameters by hospitalization level suggest potential predictive power in determining the severity of illness.

4.1.6 Statistical Testing

To formally test differences between COVID-positive and negative groups, a **t-test** was applied:

- Null Hypothesis H_0: No significant difference in mean blood parameter values between groups.
- At $\alpha = 0.02$, the null hypothesis was rejected for Platelets, Monocytes, and Leukocytes.
- This confirms that these variables are significantly associated with COVID-19 status.

4.1.7 Plots

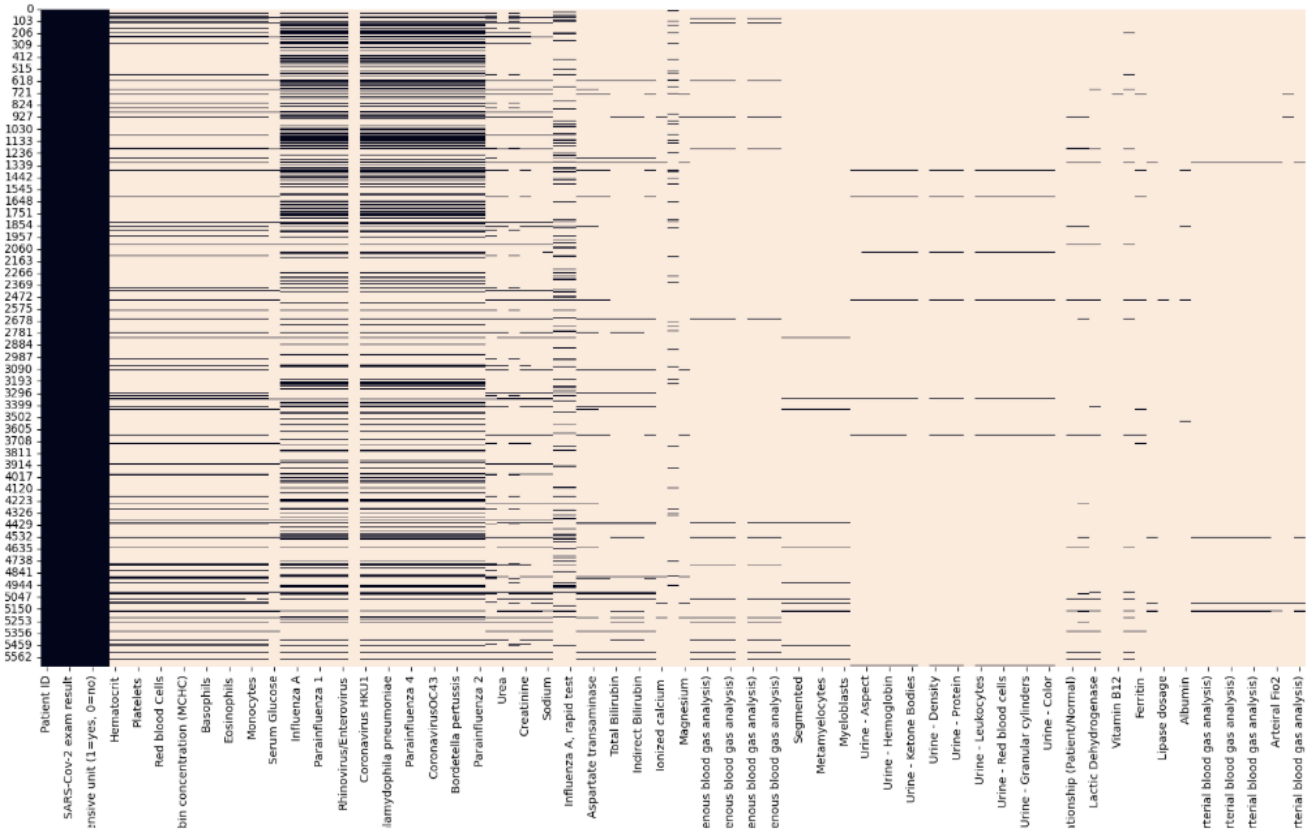


Figure 1: Missing values heatmap across features

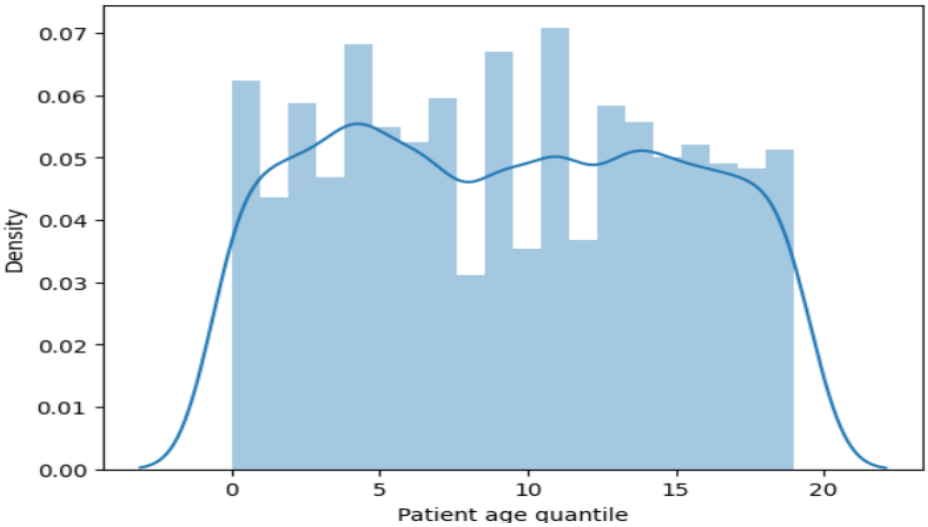


Figure 2: Distribution of Patient Ages

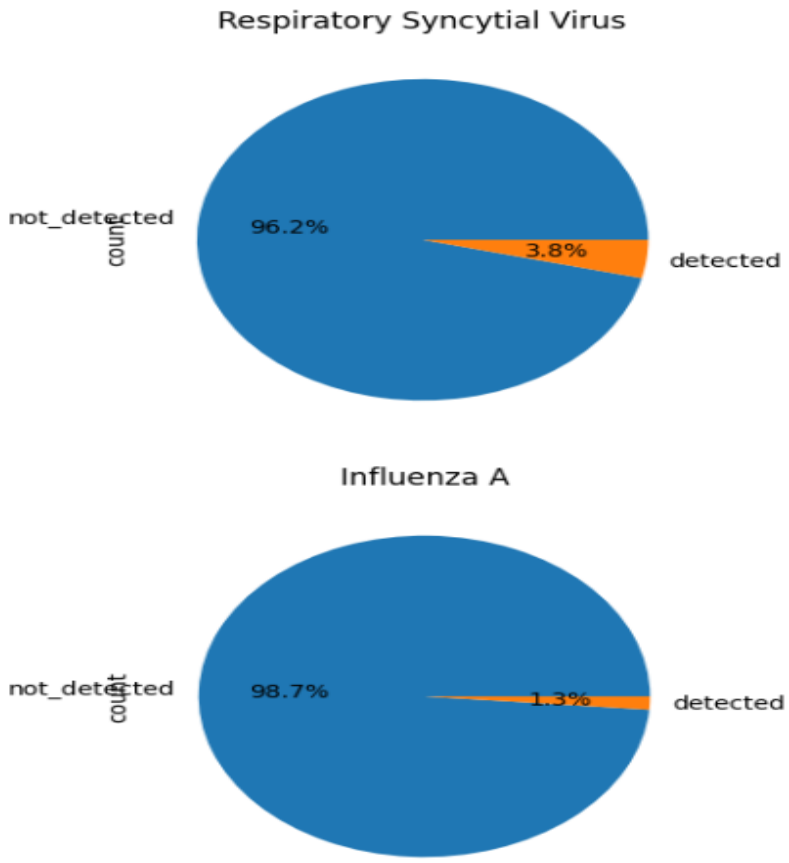


Figure 3: Class proportions of viral test results

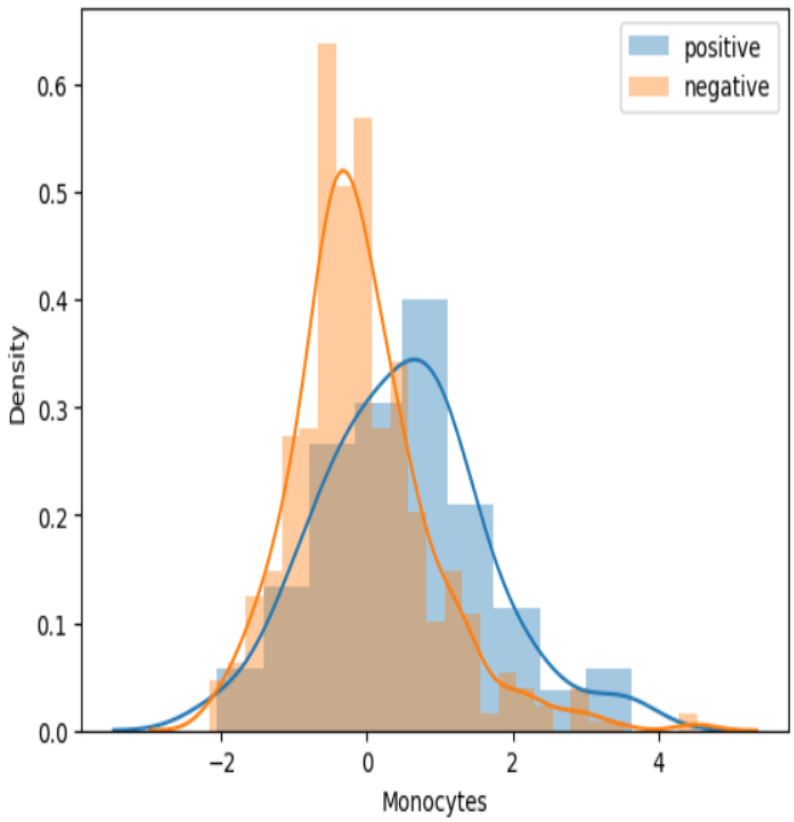


Figure 4: Monocyte levels in COVID-19 positive vs. negative patients

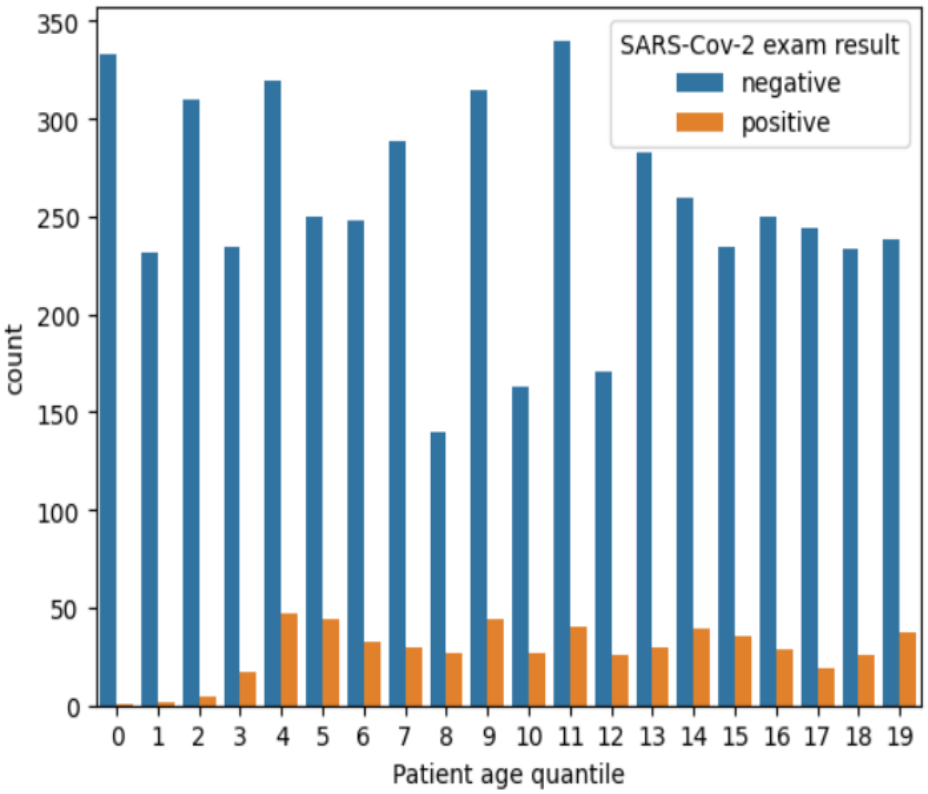


Figure 5: COVID-19 test outcomes across age quantiles

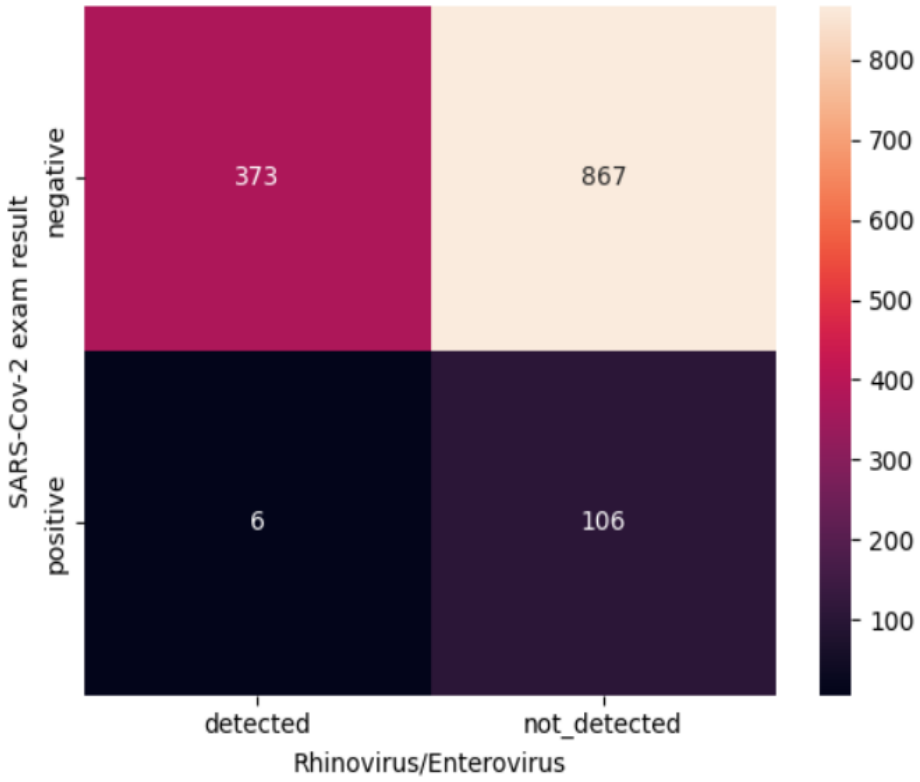


Figure 6: Co-occurrence heatmap of SARS-CoV-2 and Rhinovirus

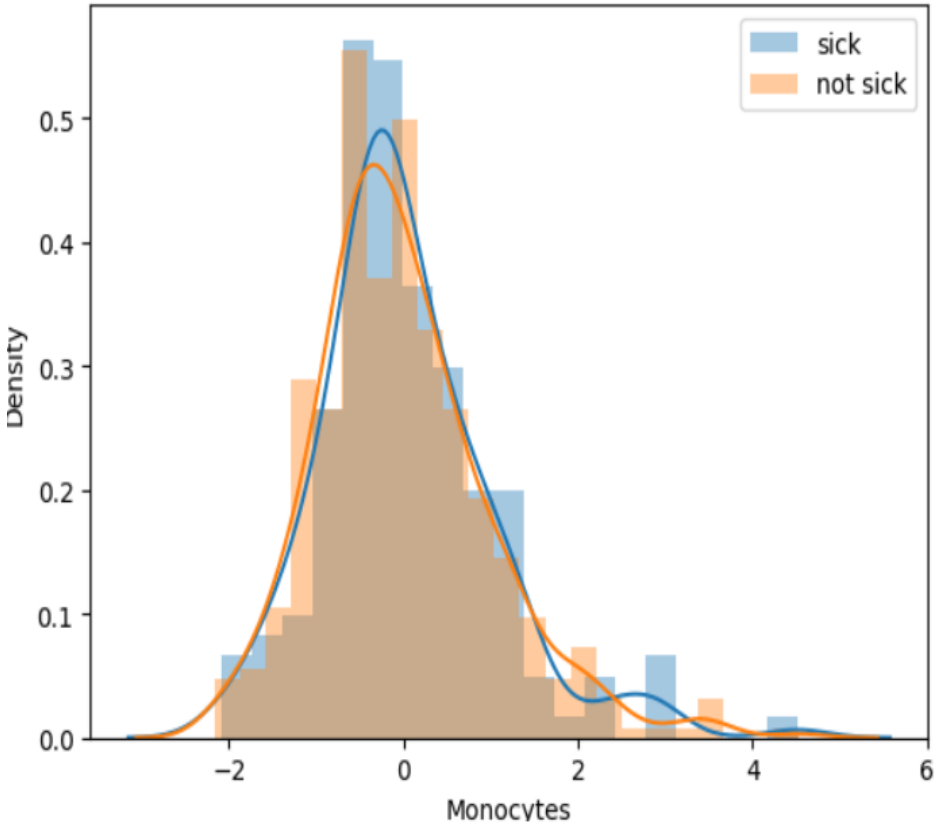


Figure 7: Monocytes distribution: other viral infections

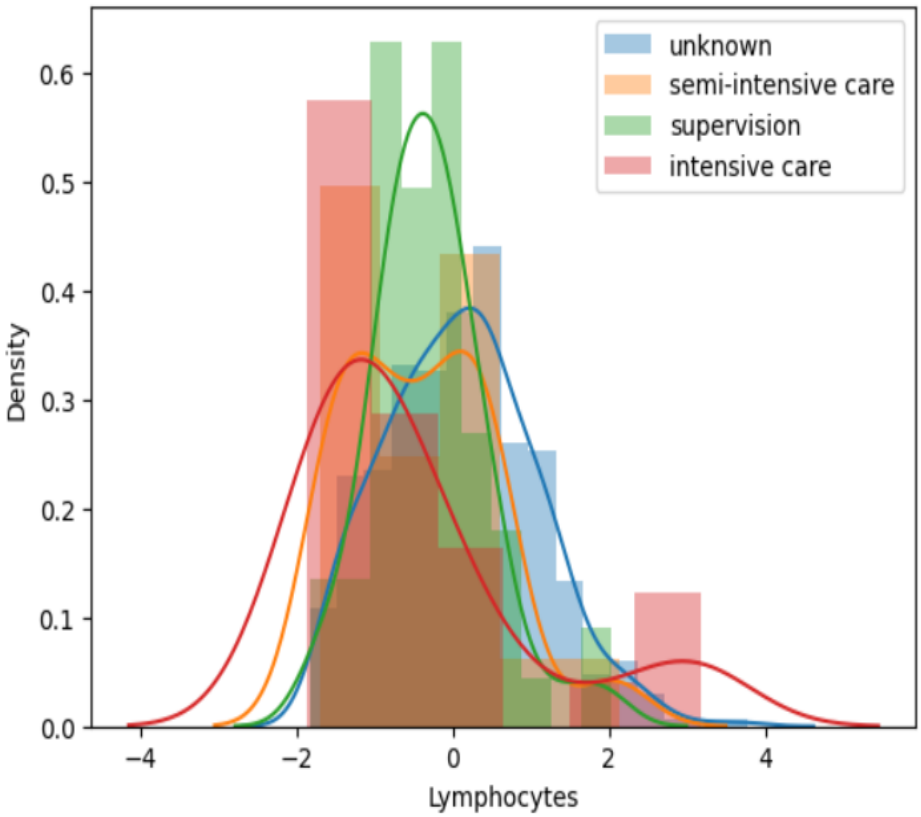


Figure 8: Lymphocytes distribution by hospitalization status

4.1.8 Insights

Key insights from EDA include:

- The dataset is highly sparse, requiring careful variable selection.
- SARS-CoV-2 positivity rate is around 10% in the full dataset, but increases sharply after filtering for patients with available viral or blood test results. This class imbalance must be considered when building predictive models.
- Viral co-infections are rare and generally mutually exclusive with SARS-CoV-2, suggesting that detection of other respiratory viruses may lower the likelihood of COVID-19 in this dataset.
- Blood test analysis highlights that Platelets, Leukocytes, and Monocytes differ significantly between COVID-19 positive and negative patients, confirmed by statistical testing.
- Younger patients appear underrepresented among positive cases in this dataset, but further validation with larger balanced samples is necessary before drawing conclusions.
- Hospital admission level (regular ward, semi-intensive, ICU) is linked to abnormal blood test results, indicating that laboratory values may also serve as predictors of disease severity.

4.2 Data Preprocessing

To prepare the dataset for modeling, several preprocessing steps were applied. These steps aimed to address the high proportion of missing values, encode categorical variables, and engineer features relevant to the prediction task.

- **Feature selection:** Only variables with less than 90% missing values were retained. Based on the EDA findings, two sets of laboratory features were selected:
  - *Blood test variables:* Features with 88–90% missing values.
  - *Viral test variables:* Features with 75–80% missing values.

  Together with the key variables `Patient age quantile` and `SARS-Cov-2 exam result`, these formed the working dataset.

- **Train-test split:** The data was randomly split into training (80%) and testing (20%) subsets for model development and evaluation.
- **Encoding categorical variables:** Viral test results recorded as text labels (`positive`/`negative`, `detected`/`not_detected`) were mapped to binary values (1 = positive/detected, 0 = negative/not_detected).
- **Feature engineering:** A new binary feature, `Is sick`, was created to indicate whether a patient tested positive for at least one viral infection (excluding SARS-CoV-2). The original viral test columns were then dropped to reduce dimensionality.
- **Handling missing values:** Rows containing missing values in the selected features were removed to ensure data consistency.
- **Final dataset:** After preprocessing, the features (X) consisted of blood test variables, age quantile, and the engineered `Is sick` indicator. The target variable (y) was the SARS-CoV-2 test result.

4.3 Modeling

To predict whether a patient tested positive for SARS-CoV-2, multiple classification algorithms were developed and evaluated. Both baseline models and more complex pipelines were tested to capture non-linear relationships and select the most relevant features.

4.3.1 Baseline Models

The first set of models included:

- **Decision Tree Classifier:** A simple interpretable model used as a baseline.
- **Random Forest Classifier:** An ensemble of decision trees providing improved accuracy and robustness compared to a single tree.

4.3.2 Feature Selection and Polynomial Pipelines

To reduce dimensionality and capture interactions between features, two extended pipelines were introduced:

- **Pipeline 1:** SelectKBest with ANOVA F-score (top 5 features) + Random Forest Classifier.

- **Pipeline 2:** Polynomial feature expansion (degree 2) + SelectKBest (top 10 features) + Random Forest Classifier.

Evaluation results showed that while all models tended to overfit (perfect training performance but weaker validation results), the pipelines with feature selection and polynomial expansion achieved better validation performance, reducing overfitting effects.

4.3.3 Extended Model Comparison

To further benchmark performance, additional classifiers were implemented with a common preprocessing pipeline (polynomial feature expansion and feature selection):

- **Random Forest Classifier**
- **AdaBoost Classifier**
- **Support Vector Machine (SVM)**
- **k-Nearest Neighbors (KNN)**

Each model was trained and evaluated using accuracy, precision, recall, and F1-score. Learning curves were also plotted to visualize training and validation performance across increasing dataset sizes. Feature importance plots were generated for Random Forest to assess the relative influence of blood test variables and engineered features.

4.3.4 Hyperparameter Tuning

A Randomized Search with 4-fold cross-validation was applied to optimize hyperparameters, focusing particularly on SVM:

- Kernel parameters: C and γ
- Polynomial feature degree
- Number of selected features (k in SelectKBest)

The best-performing configuration was then retrained and evaluated on the test set. The classification report indicated improved recall, which is particularly important in a medical context where minimizing false negatives is critical.

4.3.5 Threshold Adjustment

To further control the trade-off between precision and recall, the decision threshold of the final SVM model was adjusted based on the precision-recall curve. This allowed exploration of settings that favored higher recall, ensuring that more positive cases were correctly identified even at the expense of precision.

4.3.6 Model Performance Visualization

```
[[91  4]
 [ 8  8]]
              precision    recall  f1-score   support

           0       0.92      0.96      0.94        95
           1       0.67      0.50      0.57        16

    accuracy                           0.89       111
   macro avg       0.79      0.73      0.75       111
weighted avg       0.88      0.89      0.89       111
```

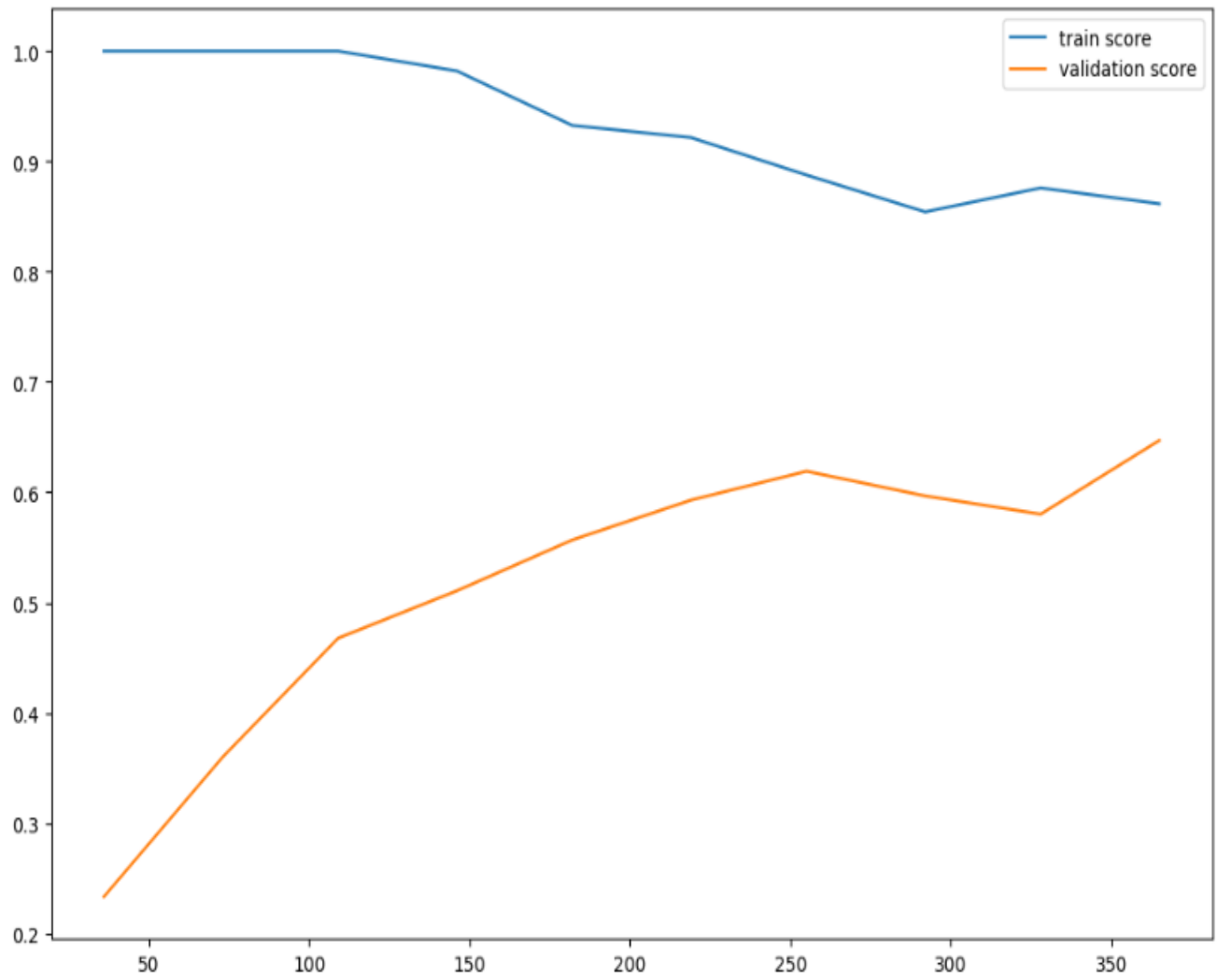


Figure 9: classification Report and Learning Curve of tuned SVM

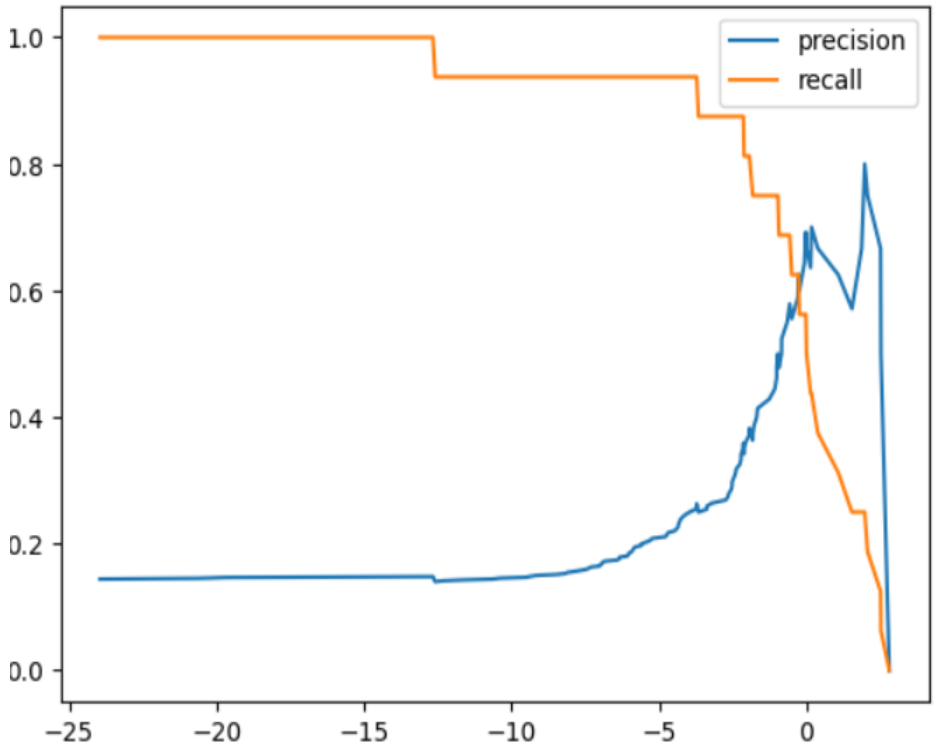


Figure 10: Precision Recall Curve

4.3.7 Insights

- Simple models such as Decision Trees showed strong overfitting and limited generalization.
- Ensemble methods (Random Forest, AdaBoost) and SVM performed better, with SVM showing the strongest results after hyperparameter tuning.
- Feature selection (SelectKBest) and polynomial expansion improved validation performance by focusing on relevant predictors.
- Adjusting the decision threshold provided a practical way to prioritize recall, an important consideration in clinical diagnostics.

5 Conclusion

This project shows that machine learning can leverage anonymized laboratory data to predict COVID-19 outcomes. While not a replacement for RT-PCR, such models could serve as complementary screening tools, especially in resource-constrained settings. Future work may include an application of explainability techniques (e.g., SHAP, LIME) to enhance clinical interpretability.