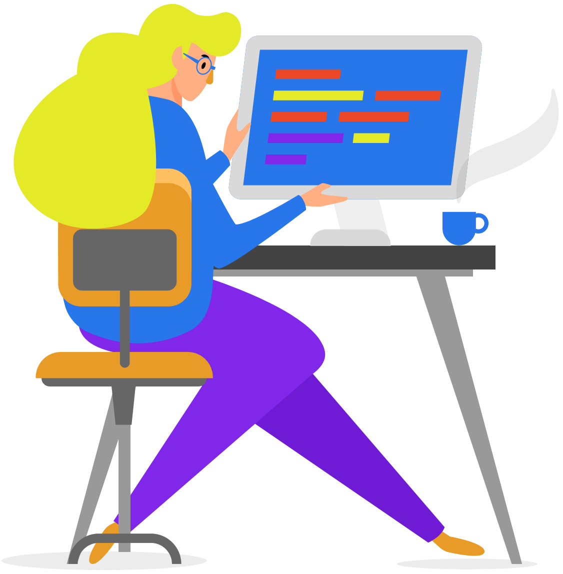
Predicting COVID-19 with Machine Learning

Data Processes - Second Assignment (2024-2025)

Group members:
Ádám Földvári
Álvaro Honrubia Genilloud
Jose Antonio Ruiz Heredia

Content



Introduction

Global Impact

COVID-19 has heavily impacted healthcare systems worldwide since 2020.They act according to environment

Project Focus

Analyzing synthetic data from two hospitals, specifically emergency department visits.

Patient Data

- Age
- Sex
- Body temperature
- Oxygen saturation
- Symptoms
- Comorbidities
- PCR results

…

Objective

Use machine learning to predict COVID-19 positivity based on these clinical features.

Goals and KPIs

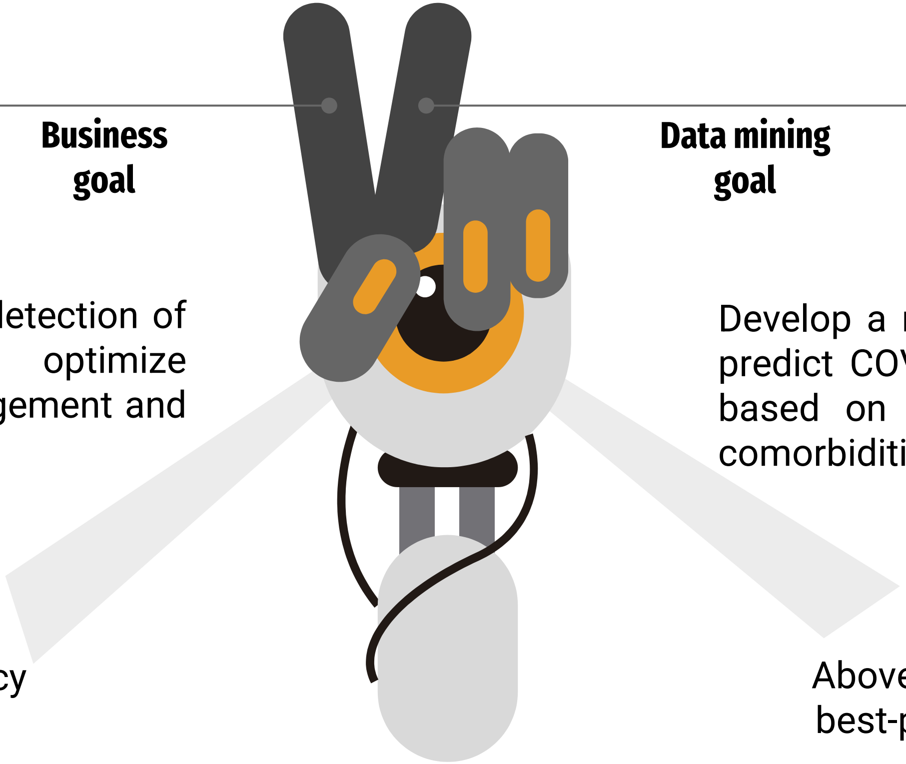
01

Business goal

Enhance early detection of COVID-19 to optimize resource management and reduce costs

Accuracy

Achieve 85% accuracy in + predictions



02

Data mining goal

Develop a robust model to predict COVID-19 positivity based on symptoms and comorbidities.

AUC

Above 0.72 AUC for the best-performing model

Common goal: Identify the top 5 features influencing predictions.

$$Cost\ Savings = \frac{(Initial\ Costs - Reduced\ Costs)}{Initial\ Costs} \times 100$$

Exploratory Data Analysis (EDA)

I — Data overview

Hospital 1 (H1): 14 712 records, 54 columns
Hospital 2 (H2): 12 736 records, 54 columns
Differences in column names and data structure

II — Observations

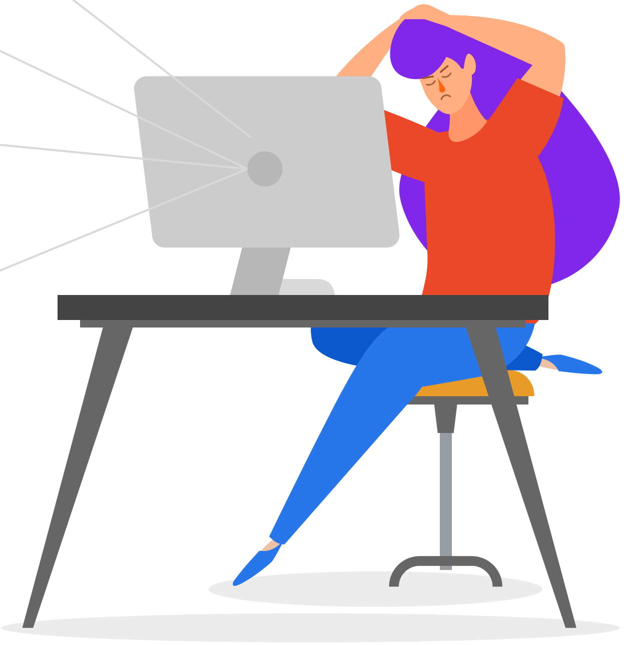
-Significant gaps in fever temperature and PCR results
-Adress missing values during preprocessing

III — Trends & patterns

-Similar seasonal trends in admissions in H1,H2 (2022 Jan)
-Age distribution is varying greatly between (H1,H2)
-A lot of positive cases among 25-34 age group

IV — Symptoms & correlations

-H1: strong symptom correlations
-H2: minimal correlations
-Most common symptoms/commorbidties: fever,hypertension



Data preprocessing

I

Reason:
Ensures data integrity by removing incomplete records that could skew analysis.

II

Reason:
Aligns datasets for seamless merging and consistent data handling.

III

Reason:
Converts categorical variables into numerical format for compatibility with machine learning algorithms.

Both datasets: H1, H2

IV

Reason:
Standardizes dates and numerical values to ensure uniformity across datasets.

V

Reason:
Imputes missing data to maintain dataset completeness and improve model accuracy.

VI

Reason:
Confirms that all preprocessing actions are correctly applied, ensuring data quality.

Merging two datasets

Data preprocessing

VII

Reason:
Removes columns with encoding issues or excessive unique values to streamline analysis.

VIII

Reason:
Derives features from date columns to capture temporal patterns.

Merged dataset

IX

Reason:
Represents time-based features cyclically to enhance model understanding of temporal trends.

X

Reason:
Confirms that after merge all preprocessing actions were correctly applied, ensuring data quality.

Merged dataset ready for training

Techniques, Tools, and Algorithms

Techniques

Confusion Matrix:
Accuracy
Precision and Recall

- Assess accuracy and balance.

F1-score

- Balance between precision and recall.

Correlation Analysis

- Identify relationships between variables

Tools

Python Libraries:
Data processing:

- Pandas,
- Numpy
- Scikit-learn

Data visualization:

- Matplotlib, Seaborn

Modelling:

- Scikit-learn

Algorithms

Random Forest:

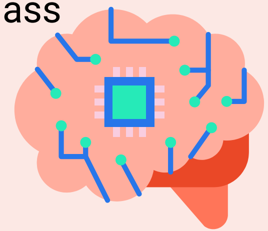
- Robust and interpretable

Decision Tree

- Provides feature insights

SMOTE

- Handles class imbalance



Modelling results

Class breakdown

Negative (0) PCR result:

- Precision : 73%
- Recall: Only 52%
- F1-score: 0.61

Moderate performance for this class.

Observation: The recall is low, indicating that many negative cases are missed. This suggests a need for better feature selection or model tuning to improve detection of negative cases.

Positive (1) PCR result:

- Precision: 92%
- Recall: 96%
- F1-score: 0.94

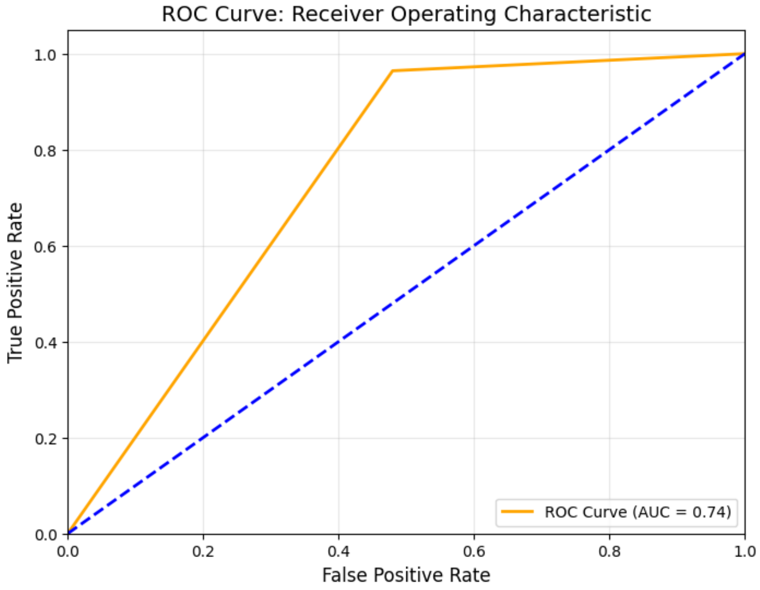
Excellent performance for this class. High precision and recall indicate reliable predictions.

Observation: The high precision and recall of the model validate its efficiency of distinguishing positive cases, which is vital for early identification.

Modelling results

Overall metrics

- Accuracy: The model accurately predict **90%** of the entire set of test samples
- Macro average: When all classes are treated equally, the figures for precision (0.82) and recall (0.74) indicate good performance, though there remains an opportunity to enhance recall for negative cases.
- Weighted average: The preservation of imbalance in the class distribution improves the results of Precision (0.89) and Recall (0.90).



Model Optimization (SMOTE)

Class breakdown

Negative (0) PCR result:

- Precision : 62%
- Recall: Only 64% (but showing better balance compared to the previous one)
- F1-score: 0.63

Moderate performance for this class.

Observation: Recalling negative cases was improved from 52 % to 64 %. Precision decreased slightly, indicating a trade-off between recall and precision.

Positive (1) PCR result:

- Precision: 93%
- Recall: 93%
- F1-score: 0.93

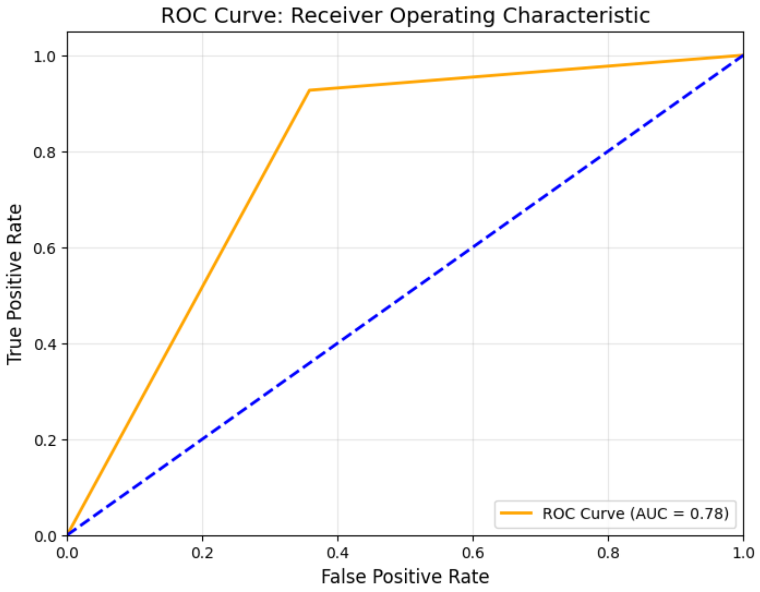
Excellent performance for this class.

Observation: Performance remains strong, maintaining high precision and recall.

Model Optimization (SMOTE)

Overall metrics

- Accuracy: The model accurately predict **88%** of the entire set of test samples
- Macro average: When all classes are treated equally, the figures for precision (0,78) and recall (0.78) indicate good performance, though there remains an opportunity to enhance recall for negative cases.
- Weighted average: The preservation of imbalance in the class distribution improves the results of Precision (0.88) and Recall (0.88).



Conclusions

	Top 5 features identified	
01	Age	Highlighting that older individuals are at a higher risk of severe COVID-19 symptoms and complications, making age a key factor in predicting PCR results.
02	Fever Temperature	Highlighting that elevated body temperature is a key indicator of a positive PCR test outcome. This is because fever is a common and prominent symptom in individuals suffering from COVID-19.
03	Oxygen saturation	Denotes that severe complications of COVID-19 are closely linked to oxygen saturation.
04	Fatigue/Malaise	Indicating that general feelings of tiredness and discomfort are significant predictors of a positive PCR test outcome. This aligns with fatigue being a common early symptom of COVID-19.
05	History of fever	The presence of a history of fever seems to be another noteworthy predictor underlining the importance of previous symptoms in the process of diagnosis.

Conclusions

Our Model

Strengths

- Successfully detects true positive cases
- Targets high-risk patients in the first place
- Precision and recall for positive PCR results are particularly strong
- Feature importance analysis

Weaknesses

- Intermediate performance for Class 0
- Enhancement to be made in separation between positive and negative cases

(The previous issues related to missing PCR results have been resolved by removing those entries entirely, eliminating the need for placeholder values. This enhanced the dataset's integrity.)

Model can be adopted in hospital settings to promote early detection of COVID-19

Thank you for your attention!