Machine Learning Project Report

Data Analysis and Preprocessing

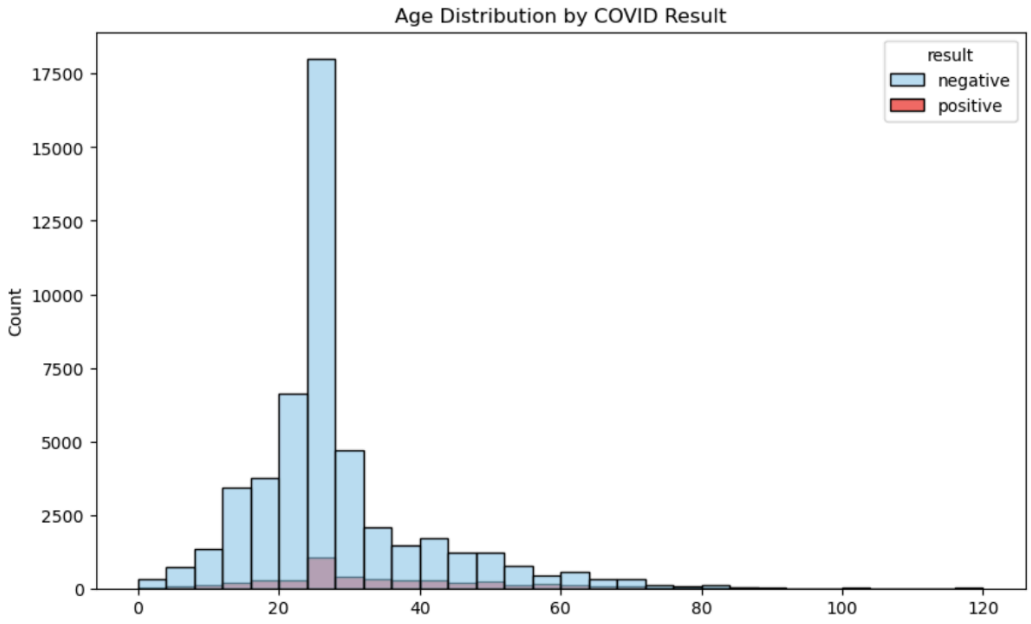
Initially, the dataset had 109,927 rows and 47 characteristics, which represented patient demographics, symptoms, and COVID-19 test results. Given that this was a classification task, we concentrated on ensuring that the target variable was properly stated, eliminating any unclear or invalid entries.Their target values were limited to positive and negative. Missing covid-19 symptoms were treated as negatives. We also assumed 2020 to be the collection year and such was used to compute the age. However, a data point reflected a negative value. We dealt with such value and missing data the same way by replacing them with the median age. The median age was chosen because the age distribution was skewed to the right.

Data Cleaning

- Target validation: Only test findings that were clearly labeled as positive or negative were kept.
- Handling missing values: Columns with more than 90% missing data were eliminated.
- Free-text fields, such as additional problems, were eliminated since they were unstandardized.
- Features without any variance in the dataset were dropped.These features were uncommon in literature as well.
- Chi-squared test was used to test the association of each symptom with the target feature.
- Feature without a statistically significance ($p < 0.05$) difference were eliminated
- These brought down the number of features to 16 initially
- However, features related to sore throat were found to be strongly correlated ($r=1$), one was dropped to make room for the other
- The final features for the final model training were 15 in all

Exploratory Data Analysis (EDA)

EDA revealed that the demographic skew towards younger persons, with about 75% of patients aged 34 years or younger. Although the diseases were not prevalent in the elderly population compared to their younger counterparts, they were very few and their behavioral qualities could be a factor. In other words, elderly people may have been meticulous about their health by adhering to safety protocols and avoiding social gatherings. Younger people on the other hand could have been very anxious about the lockdown and eager to go out or participate in social gatherings.



Sex Distribution

Of the 34,100 guys tested, around 6% were positive while of the 18,618 females tested, around 3% were positive. Male positivity stood at 8.9%, while female positivity was about 7.0%.

Other binary Features

Chi-squared test was used to test the association of each symptom with the target feature. Features without a statistically significant ($p < 0.05$) difference were eliminated. The final features were as seen in literature as common covid-19 symptoms, further validating the statistical basis.

Feature Engineering

Ages was engineered in our process, assuming the data collection date as year 2020. We also made an attempt to cluster the features into 3 specific groups: respiratory symptoms, gastrointestinal symptoms and neurological symptoms. This was considered in order to improve the model. However, it impacted negatively on the models. Therefore, they were later left out.

Model Development

We divided the dataset into training and test subgroups. To resolve the class imbalance, SMOTE was used to oversample the minority (positive) class. The following models were evaluated:

1. Random Forest
2. Logistic regression
3. Gradient Boosting
4. XGBoost
5. SGDClassifier
6. SVC (computational expensive) it was commented out because of elongated run time

Accuracy, recall, precision, F1 score, and confusion matrices were among the performance indicators used.

Choosing algorithms and feature selection

- Logistic Regression significantly beat others in recall (about 49%), detecting over half of true positive situations.
- Gradient Boosting and XGBoost produced equivalent recall, albeit at the expense of more false alarms or overfitting.
- Random Forest and SGDClassifier performed moderately but had lesser interpretability than Logistic Regression.
- The final choice prioritized Logistic Regression because of its balance of recall, interpretability, and stability.

Interpretation and Implementation of the Model

About 49% of COVID-positive patients were identified by logistic regression, and false positives (about 2,500 in the test set) were kept to a minimum.

This model ensures the interpretability of individual feature contributions while raising alerts for potential situations, making it a helpful early screening tool.

A Gradio web application was used to deliver the finished model, giving end users the ability to enter demographic and symptom information and obtain a real-time screening result.

Ethical Considerations

False negatives: About 50% of instances that tested positive were overlooked. This risk needs to be explained in detail since infections that go unnoticed could spread.

False positives: Unnecessary alarms may be sent to some healthy people, which could lead to worry or the need for further testing.

Clinical use: This instrument is not meant to be used as a diagnostic test, but rather as a screening tool. Laboratory testing for confirmation is still crucial.

Equity: The significance of regular fairness audits to make sure the model does not unfairly penalize subgroups is highlighted by gender variations in positive rates.

Privacy: Data management followed ethical guidelines as the data did not contain personal information that could be used to trace a patient.

Conclusion

We created a screening tool based on 15 clinically important features after methodically cleaning, engineering, and modeling an initial dataset of 109,927 records and 47 features. For deployment, the Logistic Regression model was chosen since it provided: ~73% accuracy, Recall: about 49% F1-score: about 0.22. As long as its limitations are recognized and utilized in conjunction with expert medical evaluation, this approach can aid in early detection and public health triage even if it is not a diagnostic tool.