CS767

Final Project

**Predicting COVID19 Deaths from pre-conditions**

Introduction

The COVID-19 pandemic identified an urgent need for data-driven insights to support clinical decision-making and public health responses. One of the most critical questions during the crisis was identifying which patients were most at risk of severe outcomes, including death. Many patients infected with the virus only experience mild symptoms of respiratory illness and can recover in a few days without any medical intervention. However, other patients developed more sever illness, many of them with pre-existing conditions that put them at more risk. Understanding how these underlying health conditions affect COVID-19 mortality helps health care professionals make decisions that will be best for the patients and prevent overloading of the hospitals. This project aims to contribute to that understanding using machine learning techniques.

For this project, I used a publicly available dataset provided by the Mexican government, which includes data on over 1 million patients. Each record contains anonymized patient data on age, sex, and the presence of various pre-existing medical conditions, as well as the whether the patient survived or not. My objective was to develop and evaluate machine learning models that can predict mortality using these data.

By focusing on pre-conditions alone, the project aims to assess how well we can identify high-risk patients at the point of testing or admission. Several classification models were tested, and their performance was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. This report presents the methodology, results, and analysis of the models used.

Methods

The creation of machine learning models from the dataset had a few key steps: data cleaning, class balancing, building models, and evaluating the models. For cleaning the data, I first removed all rows in which the patient was not confirmed to have COVID-19. I then dropped all columns which were not related to pre-existing information available to the hospital. After that, I filled in all missing values, since it is a safe assumption that if there is no data on a patient having a particular pre-condition, then they likely did not have it. I then balanced the classes using random under sampling of the majority class. This method was chosen since the dataset is already large so it wouldn’t have made sense to make it even bigger with an oversampling technique. So I found the proportion of majority class to minority class and then sampled from the majority class so it would have an equal number of rows and the minority class. Then I built the models testing 4 different machine learning models: logistic regression, random forest, gradient-boosted trees (GBT), and naïve bayes. These models were all built using the native spark functions and then were evaluated for accuracy, precision, recall, and F1-score.

Results

Logistic regression has an accuracy of 0.839, a precision of 0.843, a recall of 0.836, and an F1-score of 0.840. Random forest has an accuracy of 0.836, a precision of 0.819, a recall of 0.868, and an F1-score of 0.843. GBT has an accuracy of 0.842, a precision of 0.829, a recall of 0.866, and an F1-score of 0.847. Naïve bayes has an accuracy of 0.786, a precision of 0.776, a recall of 0.813, and an F1-score of 0.764.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score |
| Logistic Regression | 0.839 | 0.843 | 0.836 | 0.840 |
| Random Forest | 0.836 | 0.819 | 0.868 | 0.843 |
| Gradient-Boosted Tree | 0.842 | 0.829 | 0.866 | 0.847 |
| Naïve Bayes | 0.786 | 0.776 | 0.813 | 0.764 |

Discussion

The results of this project demonstrate that machine learning models can achieve reasonably high performance in predicting COVID-19 patient mortality based solely on pre-existing conditions. Among the models tested, GBT yielded the best overall performance, achieving an accuracy of 0.842 and the highest F1-score of 0.847. This suggests that GBT is particularly effective at balancing precision and recall in this classification task, likely due to its ability to model complex nonlinear relationships and interactions between features.

Logistic regression also performed strongly, with an accuracy of 0.839 and an F1-score of 0.840. Since it is a relatively simple model and easily interpretable, it is compelling for this usage so the decision can be explained to clinicians and patients. Although it slightly underperformed compared to GBT in terms of recall and F1-score, its consistency across all metrics makes it a reliable baseline model.

Random forest performed slightly better than Logistic Regression in terms of recall (0.868 vs. 0.836) but had slightly lower precision. This indicates that while it was more successful at identifying patients who eventually died, it also produced more false positives. In a clinical context, this trade-off might be preferable, since it will miss fewer high-risk cases.

Naïve Bayes, while the fastest and simplest model, showed the weakest performance across all metrics. Its lower F1-score of 0.764 and accuracy of 0.786 suggest that its strong independence assumptions do not hold well for this dataset, which likely contains correlated features.

Overall, logistic regression, GBT, and random forest all performed comparably and are all promising models for this usage. Perhaps more complex models such as neural networks would be able to better predict patient outcomes, at the tradeoff of being harder to interpret.

Conclusion

Overall, the models performed quite well despite relying solely on patient pre-conditions without access to any other patient data. This demonstrates the predictive value of pre-existing health data and the potential for early risk assessment tools. Overall, I was able to succeed at my objective of using machine learning to predict patient mortality from COVID-19 based on their pre-existing conditions at a decently high accuracy.