Jiayu Gu

Dr. Dr.Bari

Predictive Analytics

11.03.2023

Report

1. **1. Problem Statement:**

The central question I aim to address is whether it is possible to predict the presence of COVID-19 based on medical data.

1. **Data sources:**

Three primary data sources underpin this report. The first is the New York State Government Health Data site, from which I have extracted two datasets: the number of COVID-19 deaths in NYC and the daily positive test rate, calculated as the number of positive test results divided by the total number of tests conducted on that day. The second source is Google Trends data for the keywords "fever" and "cough" over the period from 2021 to 2022, analyzed to discern if these terms could serve as early indicators of COVID-19 outbreaks. The third source includes data on COVID-19 symptoms and corresponding test results.

1. **Feature engineering and preprocessing done**

I merged these datasets into two data matrices. The first includes the positive rate, fever, cough, and death data, consolidated from four tables based on the date to determine if the selected features can predict the presence of COVID-19. The data from Google Trends is indexed from 1 to 100. I expect the model to correctly predict the rate of positive cases. The second matrix is binary, indicating the presence (1) or absence (0) of symptoms, with the expected outcome being the model's ability to predict negative or positive test results. In RapidMiner, the initial step was to combine the data, followed by normalization and outlier processing for the first matrix. Since the second matrix lacked outliers, no normalization or outlier processing was required. As the feature set was small, dimensionality reduction through PCA was not utilized.

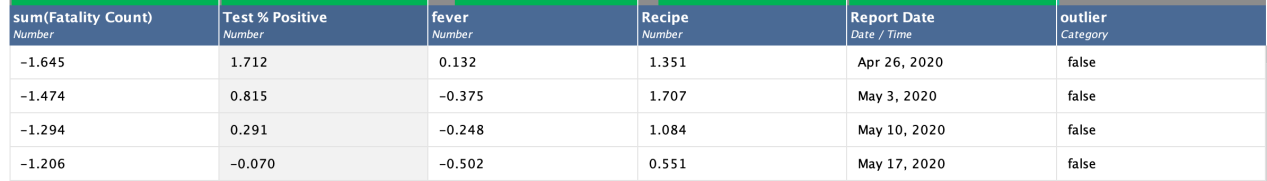


Figure 1 The feature of first matrix

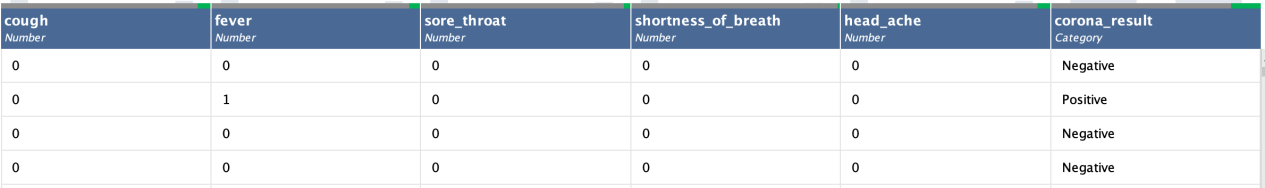


Figure 2 The feature of second matrix

1. Results from all the models in the form of a table

Figure 3 illustrates the relationship between the positive rate and other features. The positive rate data is numerical, thus the table displays the correlation with other features. The modeling results suggest that these features have a high correlation with the positive rate, making them suitable predictors for COVID-19.

For numerical outcomes, we applied a random forest model with the main hyperparameter set to least squares. The advantages of random forest include its efficacy with numerical data and high accuracy, while the drawbacks are the risk of overfitting and parameter complexity.

Figure 4 focuses on the symptoms used to predict test outcomes. The test result is categorical; hence, the table shows the accuracy between the test results and symptoms. The high predictive accuracy suggests these features are good indicators for COVID-19.

For categorical outcomes, we selected a decision tree model. The benefit of the decision tree is that it does not necessitate data normalization, but it can be computationally more complex than other algorithms.

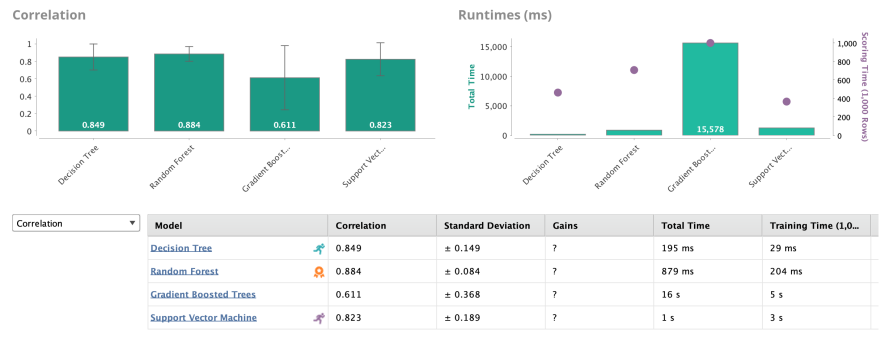


Figure 3 Three feature predict the rate of positive

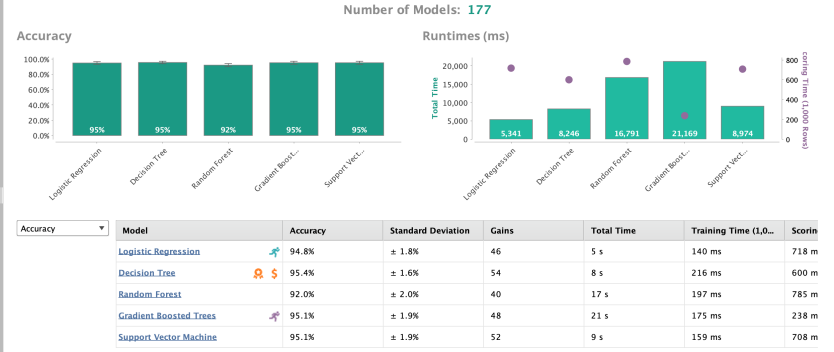


Figure 4 The Symptoms of Covid-19 to predict the resut of test

1. Results from the hyperparameter search

Matrix 1

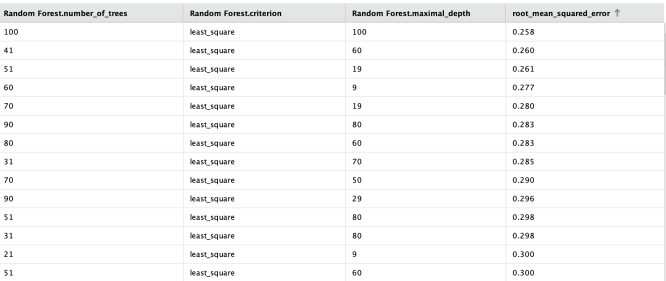


Figure 5 Random Forest optimize parameters(Grid) with criterion and maximal\_depth

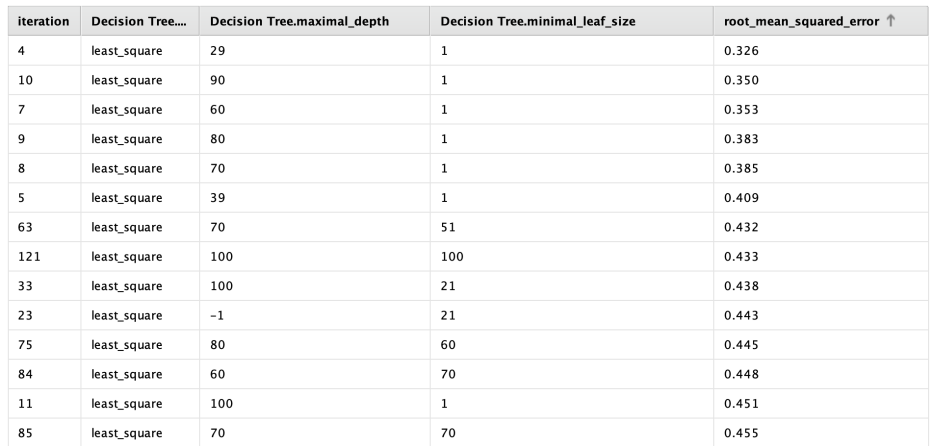


Figure 6 Decision tree optimize parameters(Grid) with maximal\_depth and minimal\_leaf\_size

Through the figure 5, the best hyperparameters set is the least\_square with 100 decision tree maximal depth. Through the figure 6, the best hyperparameters set is least\_square with 29 maximal depth and 1 minimal leaf size.

Matrix 2

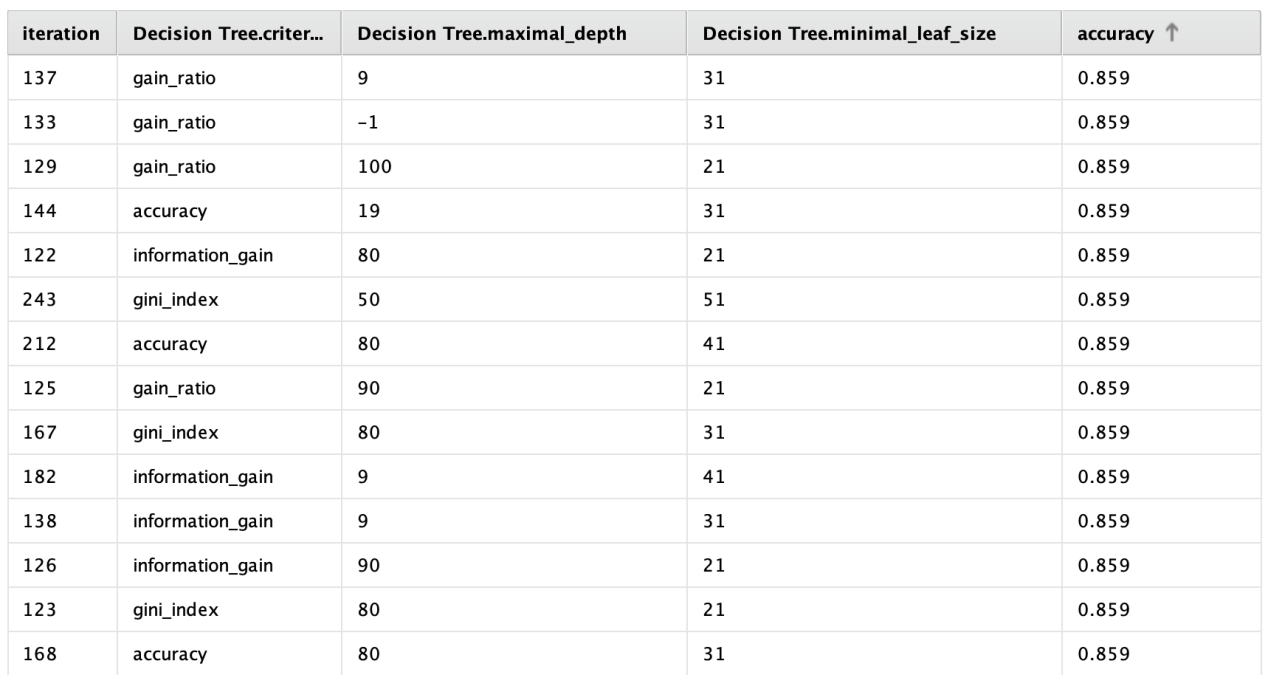


Figure 6 Decision tree optimize parameters(Grid) with criterion, maximal\_depth and minimal\_leaf\_size

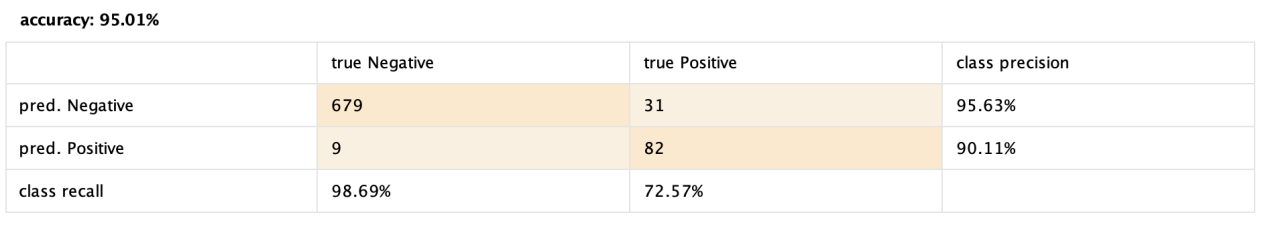


Figure 6 Decision tree performanceVector

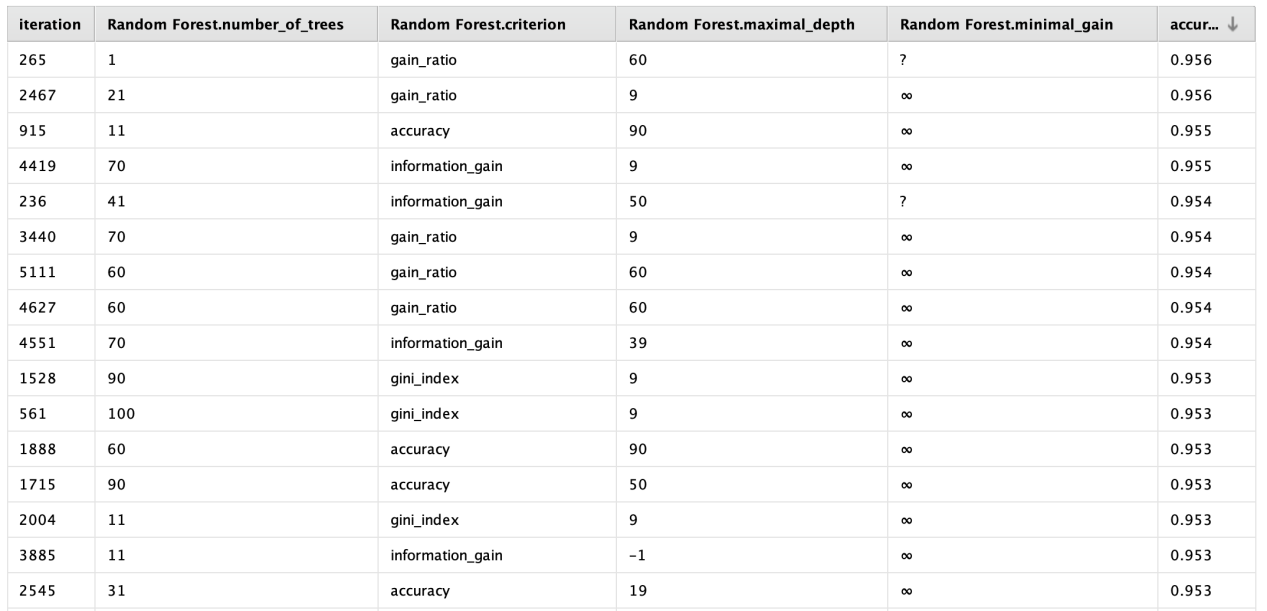


Figure 7 Random forest optimize parameters(Grid) with citerion, maximal\_dept and minimal\_gain

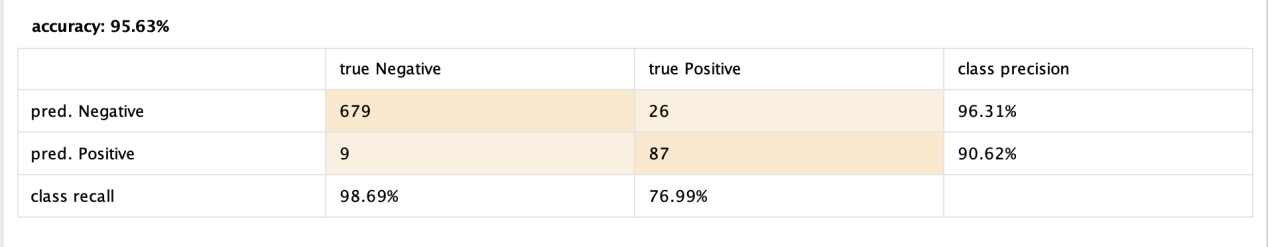


Figure 8 Random forest optimize parameters(Grid)

Through the figure 6, the best hyperparameters set is the gain\_ratio with 29 maximal deapth and 31 minimial leave. Through the figure 7, the best hyperparameters set is the gain\_ratio with 69 maximal depth.

1. Conclusions

The models demonstrate high accuracy in predicting COVID-19 based on the selected medical data. The optimal model for numerical predictions is the random forest with a maximum depth of 100. For categorical outcomes, the decision tree with a gain ratio, a maximum depth of 29, and a minimum leaf size of 31 is most effective. Consequently, the data supports the hypothesis that the presence of COVID-19 can indeed be predicted using medical data.

1. Link to the GitHub repo