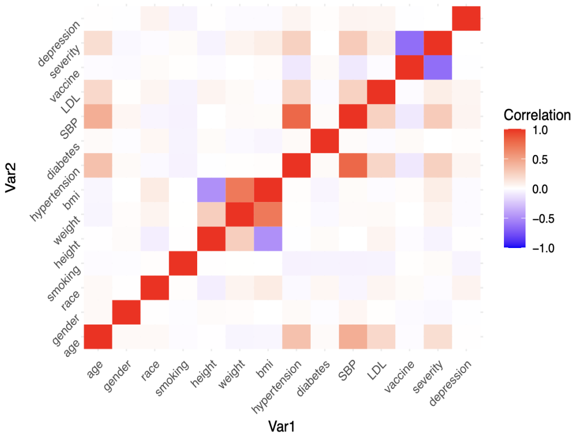
COVID-19 Severity Prediction Model

Exploratory Analysis and Data Visualization

The main goal of this research is to use the variables in the severity training dataset and the severity testing dataset to create models that could be used to identify the potential predictors for our main response variable, 'Severity' from COVID-19. The potential result could be to see what elements during the infection time worsen the patient's health condition.

This analysis began with exploring both datasets and then using visualization techniques to show the relationships between predictors. For the severity training datasets, there are 800 observations of the participants and 16 potential variables. For the severity testing dataset, there are 200 observations of the participants and also 16 potential variables. The first variable, "id", was deleted from both training and testing datasets for not being useful to the model training, leading to 15 remaining variables in both datasets. Figure 1 visually represents the relationships between various pairs of variables in our dataset. We observed that 'hypertension' has a strong correlation with 'SBP' and a moderate correlation with 'age'. Such high correlations can lead to multicollinearity, which undermines the accuracy and interpretability of our predictive model. Consequently, we decided to exclude 'hypertension' from our analysis. Additionally, since BMI can be derived from height and weight, we apply the principle of parsimony, maintaining the least number of predictors, to justify retaining 'BMI' and excluding 'height' and 'weight' for a simpler model.

Figure 1. Variables Correlation



Thus, the final variables retained in the severity training and testing datasets include "age," "gender," "race," "smoking," "BMI," "diabetes," "SBP," "LDL," "vaccine," "depression," and "severity," with severity serving as the response variable and the others as predictors.

Model training

We adopted a k-fold cross-validation approach with 10 folds for training the model, which involves dividing the dataset into 10 equal parts and running 10 cycles of training and testing. We used a fixed random seed of 1 to ensure consistency and reproducibility and utilized the caret package in R.

For building the model, various techniques were applied including Elastic Net, MARS, Linear Discriminant Analysis, Recursive Partitioning and Regression Trees, Random Forest, Adaptive Boosting, and Support Vector Machine Learning techniques for regression analysis. The final step involves comparing the root mean squared error of all models through a resampling technique to determine the most effective model. All of these models are running by the formula "severity ~ age + gender + race + smoking + bmi + diabetes + SBP + LDL + vaccine + depression".

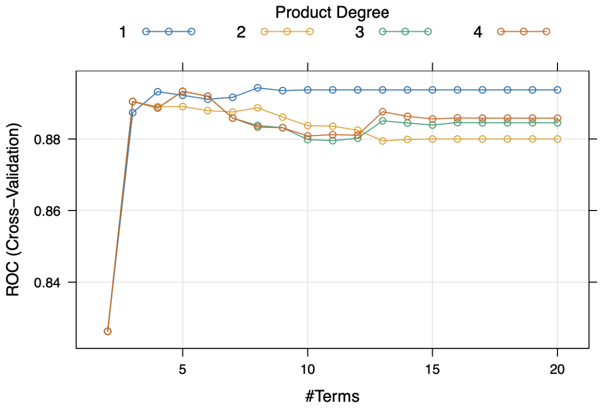
Elastic Net Model (Enet)

The Elastic Net training process sets the alpha degree for 0 to 1 with a length of 21, and lambda degree for 1 to -7 with a length of 100. The optimal parameters found are a grid setting of alpha equal to 0.2, and a lambda equal to 0.07162. Using the test dataset to assess the performance of this model, the confusion matrix gives an accuracy of 0.855 and a 95% confidence interval of (0.7984, 0.9007), sensitivity is 0.8889, and specificity is 0.7846.

Multivariate Adaptive Regression Splines (MARS)

The Multivariate Adaptive Regression Splines training process sets the degree parameter to range from 1 to 4 and the prune parameter from 2 to 20. The optimal parameters found are a prune setting of 8 and a degree of 1.(Figure 2)When the model's performance was assessed on the test data, it achieved an accuracy of 0.855 with a 95% confidence interval of (0.7984, 0.9007), a sensitivity of 0.8963, and a specificity of 0.7692, as determined from the confusion matrix.

Figure 2. MARS Cross Validation Plot



.

Linear Discriminant Analysis (Lda)

The Linear Discriminant Analysis training process gives us a model performance with an accuracy of 0.84, a 95% confidence interval of (0.7817, 0.8879), a sensitivity of 0.8519, and a specificity of 0.8154 given by the confusion matrix.

Recursive Partitioning and Regression Trees (Rpart)

The Recursive Partitioning and Regression Tree training process gives us a model performance with an accuracy of 0.85, a 95% confidence interval of (0.7928, 0.8965), a sensitivity of 0.9259, and a specificity of 0.6923 given by the confusion matrix.

Random Forest

The grid we set for this training process was to sample the participants from a range of 1 to 8 of the decision tree range and the minimal node value varies from 2 to 16 with an increment of 2. Finally, we got the most optimal parameters for when minimal observations are 10 and mtry equals 1. (Figure 3)

The Random Forest confusion matrix gives us a model performance with an accuracy of 0.81, a 95% confidence interval of (0.7487, 0.8619), a sensitivity of 0.9704, and a specificity of 0.4769.

We also did a variable importance check for this process; the two most important variables are vaccine and SBP. (Figure 4)

Figure 3. Random Forest Cross Validation Plot

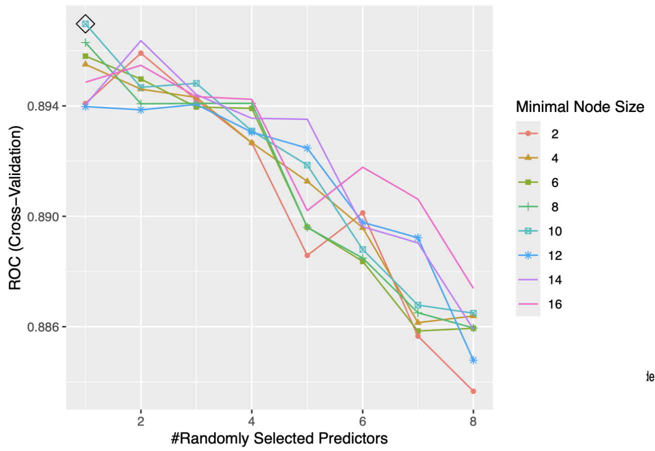
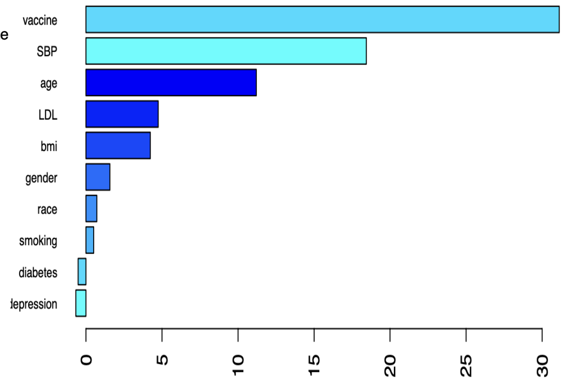


Figure 4. RF Importance Chart



Adaptive Boosting

This training process set the number of trees from 2000 to 5000, and 1 to 6 for the maximum depth of each tree in boosting. The final prediction values of the contribution are 0.001, 0.002, and 0.003 for each tree. The most optimal parameters for this process are when trees equal to 2000, maximum depth equals 4, and contribution equals 0.001.(Figure 5)

For the confusion matrix of the Adaptive Boosting training process, we obtain an Accuracy of 0.855, a 95% Confidence Interval of (0.7984, 0.9007), a sensitivity of 0.9185, and a specificity of 0.7231.

The most important variable of this training process is also the vaccine. (Figure 6)

Figure 5 Ada Boosting Cross Validation Plot

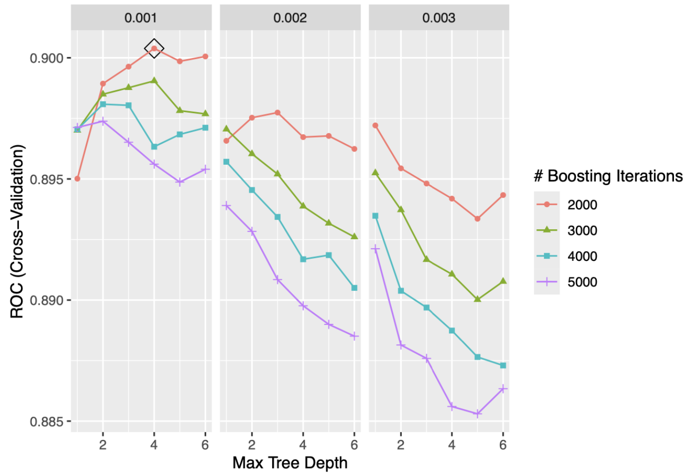
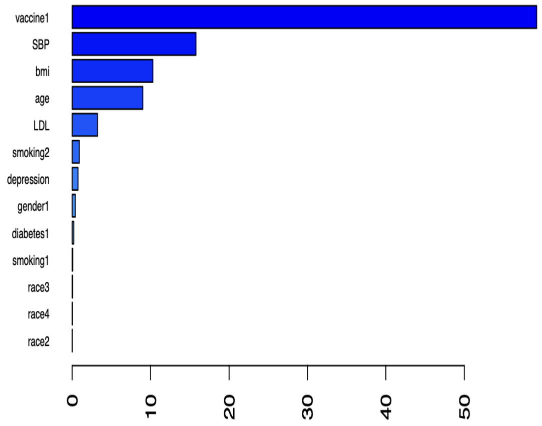


Figure 6 Ada Boosting Importance Plot

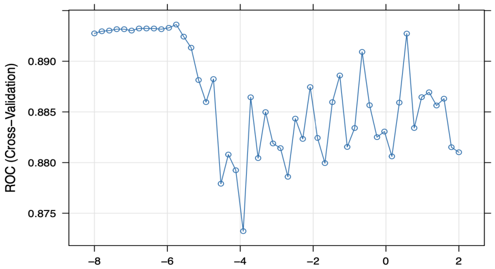


Support Vector Machine Learning (SVML)

This process specifies the exponential sequence of the regularization parameter ranging from -8 to 2. The most optimal parameter is when the parameter is equal to 0.00316658.(Figure 7)

For the confusion matrix of the Support Vector Machine Learning training process, we obtain an Accuracy of 0.84, a 95% Confidence Interval of (0.8519, 0.8879), a sensitivity of 0.8519, and a specificity of 0.8154.

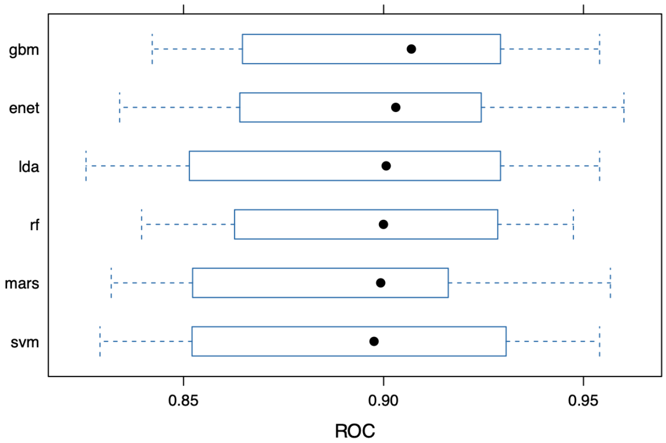
Figure 7. SVML Cross Validation Plot



Results

In the resampling process, by refitting all the models, except the rpart training process for using a different package, to samples from the training data, we obtained that the Adaptive Boosting is the best one with the highest ROC values and also the highest prediction accuracy. (Figure 8)

Figure 8. ROC comparison for All Models



Conclusions

To sum up, we applied Elastic Net, MARS, Linear Discriminant Analysis, Recursive Partitioning and Regression Trees, Random Forest, Adaptive Boosting, and Support Vector Machine Learning techniques to predict the severity of COVID-19. The best analysis process is Adaptive Boosting with the tune of 2000 trees and a maximum depth of 4 with a contribution of each tree equal to 0.001 to produce the highest Accuracy of 0.855 and an ROC of 0.9004. For this analysis procedure, as we checked for the variables' importance, the variable "vaccine" is the most important one contributing to the severity of COVID-19, participants without vaccination for COVID-19 had worse symptoms, so it is suggested that people should always have vaccination for repercussion. SBP is the second most important variable, suggesting that the value of a participant's systolic blood pressure is a good reflection of his health condition, along with the third "BMI", both two variables are great indicators of physical status. Therefore, getting more exercise, getting good rest, and managing pressure are good ways to get a not severe COVID-19 symptoms. "Age" is also an important variable, as someone ages, the probability of getting severe symptoms of COVID-19 rises, so people should take more care of the older ones in the family in such a dangerous time.