# P8106 DS2 Final Report: Predictive Modeling of Severity of COVID-19 illness

Ze Li



## 1. Introduction

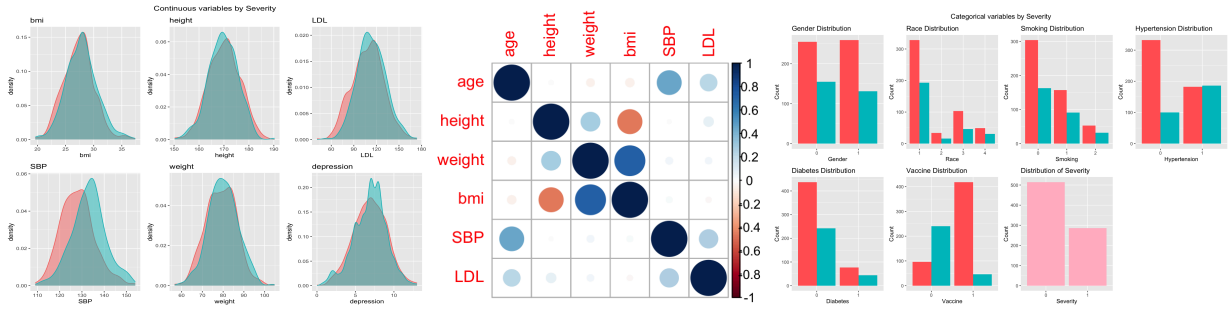
The study focuses on the severity levels of COVID-19 infections, analyzing factors from personal characteristics to health details such as gender, race, smoking status, body mass index (BMI), pre-existing health conditions, and COVID-19 infection details in order to predict severity outcomes and improve management and preparedness strategies.Our study using a dataset of 1000 participants from three cohort studies. These models could significantly contribute to patient management strategies, enhance public health responses, and inform future pandemic preparedness.

## 2. Exploratory Data Analysis

Our initial steps omit the null values and convert numerical codes to categorical labels for variables such as gender, race, smoking status, hypertension, diabetes, vaccination status, and severity of the COVID-19 infection. This process facilitated a more intuitive analysis and visualization. A thorough examination for missing values ensured the integrity of our dataset, enabling accurate and reliable insights.

a. Continuous Variables Analysis

Height, weight, BMI, SBP, LDL, and depression were analyzed using density plots by severity level0 and level1, revealing the distribution patterns among different severity levels. The distribution plots reveal distinct patterns across two COVID-19 severity levels for systolic blood pressure (SBP) and depression scores, with higher values correlating with increased severity. BMI, height, weight, and LDL cholesterol show similar distributions for both severity levels. The correlation analysis highlights strong positive associations between BMI and weight, and between weight and SBP, suggesting their potential as key predictors in modeling COVID-19 severity.




b. Categorical Variables Analysis

The bar charts depict the distribution of gender, race, smoking status, hypertension, diabetes, and vaccination status across two COVID-19 severity levels. Notable findings include higher proportions of severe cases among those with hypertension and among unvaccinated individuals.

## 3. Model Training

To develop a robust predictive model for severity of COVID-19 infection, we employed a multi-model approach. We chose to compare the performance of eight different regression techniques: logistic regression, MARS, and GAM, LDA, QDA, NB, classification trees, SVC, and SVM. Each of these models helps us to understand the relationship between the predictors and the recovery time. We used Receiver Operating Characteristic (ROC) to measure the performance of each regression model and compare the performance between different models.

**a. Data Preparation**

Prior to modeling, we ensured that categorical variables were correctly encoded as factors. To validate our models, the dataset is divided by using an 80-20 split for training and testing, respectively, ensuring that we have a representative sample for model validation. For each model, we used 10-fold cross-validation to evaluate model performance and tune the hyperparameters..

**1. Logistic Regression (GLM)**

Based on the response variable severity being binary, we consider that logistic regression, a generalized linear model (GLM) with a logistic function, is ideal for modeling the probability of occurrence of the two severity levels, facilitating the prediction of outcomes based on various predictor variables. The ROC for the linear model was 0.8961.

**2. Penalized Regression (GLMNET)**

We used the GLMNET method for penalized logistic regression. We tuned the models by experimenting with various values of lambda (the regularization penalty) from exponential value of-5 to 5 with length 50, and setting alpha between 0 and 1. The parameter of best tune is when alpha = 0.25 and lambda = 0.078. The ROC from the GLMNET model was 0.8961.

**3. Multivariate Adaptive Regression Splines (MARS)**

We assume the relationship between predictors and response is non-linear. We set parameters for a model in four-way interactions and the model training process should try retaining anywhere from 2 to 20 of these functions after the pruning process. The best model uses 4 basis functions (splines) and has the interactions of degree 1. From the variable importance plot, we can find that vaccine1 and SBP are the most important variables.

**4. Generalized Additive Model (GAM)**

Employing a Generalized Additive Model (GAM) allowed us to capture complex non-linear relationships between predictors and the response variable. GAM's flexibility in allowing both linear and non-linear terms gave us a nuanced understanding of the data, evident from the model's UBRE score is -0.2345662. The ROC of the GAM model was 0.8972.

**5. LDA**

The Linear Discriminant Analysis (LDA) model was trained using the train function with ROC as the metric for evaluation. The curve ascends sharply towards the upper left corner, indicating a high true positive rate (sensitivity) relative to the false positive rate (1 - specificity).

6. **QDA**

The QDA predictions provide a set of probabilities for the test instances, displaying a more flexible decision boundary than LDA due to different covariance structures assumed for each class. The ROC curve for the QDA model similarly shows a robust performance, with the curve approaching the upper left corner, indicating high sensitivity and specificity.

7. **Naive Bayers (NB)**

The Naive Bayes model explored here evaluates the performance of Gaussian and nonparametric kernel estimation methods by adjusting the bandwidth in the model. The ROC cross-validation score for the nonparametric approach shows a consistent improvement, indicating better model performance with a smoother probability density estimation.

8. **Rpart Classification Tree**

Key splitting variables include age, weight, smoking status, BMI, systolic blood pressure (SBP), and hypertension status to predict severity of COVID-19 infection. The optimal tuning parameter of rpart tree is approximately 0.0007526433 suggests that this is the threshold where adding any more splits does not significantly improve the model's accuracy on the validation set, considering the cost of complexity.

9. **Ctree Classification Tree**

The decision tree diagram reveals key predictors influencing the severity of COVID-19. The optimal parameter mincriterion = 0.7147945 from the ctree output suggests that the tree uses this threshold to maximize information gain and minimize impurity in the classification of severity, ensuring that each split provides substantial predictive value.

10. **Random Forest (RF)**

The ROC cross-validation performance of a Random Forest model shows that the optimal number of predictors (mtry) is 5, with a minimal node size of 12, optimizeing prediction accuracy while maintaining generalizability in predicting COVID-19 severity.

11. **Support Vector Machine Linear (SVML)**

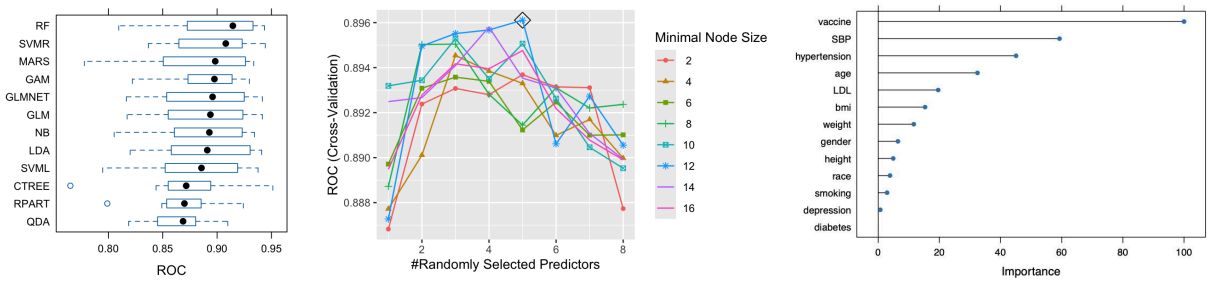
The ROC curve plot for the SVM model indicates a variation in performance across different cost values, with a notable peak in ROC cross-validation scores at lower cost levels. The optimal cost parameter, determined as approximately 0.0067, suggests that a very low penalty for misclassification leads to the best model performance in terms of ROC, which may imply that the data has a complex boundary that benefits from a softer margin classifier.

12. **Support Vector Machine Regression (SVMR)**

SVMR may offer better predictive performance for predicting the severity of COVID-19 illness compared to SVML. We set parameters for a model of cost between seq(1, 7) with length 50 and the sigma retaining from seq(-10, -2) with length 20 after the pruning process. The optimal parameter of sigma is 0.0005678 and cost is 526.0026.

## 4. Results

Upon examining the boxplots based on ROC values, the RF and SVMR model is distinguished by its notably higher ROC, as well as a relatively tight interquartile range. The compactness of the RF and SVMR boxplot signifies consistent performance across different folds of cross-validation, highlighting the model's robustness and reliability. Moreover, the absence of extended whiskers or outliers in the two plots further reinforces the stability of the model.




The random forest plot indicates that increasing the number of predictors generally improves model performance up to a certain point, after which the performance stabilizes or slightly declines, suggesting a balance between model complexity and overfitting. Further analyzing the variable importance of SVMR, we find the vaccine is the most important variable, suggesting it has a significant impact on predicting the severity of COVID-19variable.

## 5. Conclusion

In this study, we aimed to predict the severity of COVID-19 illness utilizing a dataset encompassing various personal characteristics, health details, and COVID-19 infection specifics. Through comprehensive exploratory data analysis and model training employing diverse regression techniques, we gained valuable insights into the predictors influencing COVID-19 severity and developed robust predictive models. It identified vaccination as the most important factor in predicting COVID-19 severity. This highlights the important role of vaccination in minimizing the pandemic's impact, as well as the significance of vaccination campaigns and policies in public health responses.

Furthermore, our multi-model method enabled us to assess the performance of various regression approaches, with Random Forest (RF) and Support Vector Machine Regression (SVMR) outperforming them significantly in terms of ROC values and resilience during cross-validation folds. These models present potential opportunities for precisely and reliably forecasting COVID-19 severity, allowing for more informed decision-making in patient care and public health actions.

Overall, our research contributes to a better knowledge of the characteristics that influence COVID-19 severity and offers significant insights for improving patient treatment techniques, public health initiatives, and future pandemic preparedness efforts. Using predictive modeling tools, we may better anticipate and handle the varied degrees of COVID-19 disease severity, ultimately leading to more effective pandemic mitigation and management strategies.