DS 3001 Project Results

Kaitlyn Chou [1] **Mackenzie Chen** [2] **Mrunal Kute** [3]

Abstract

This document discusses our preliminary results based on our model. We ended up changing the scope of our research question, but we still used the original dataset. We decided to examine how people's race, underlying medical condition, and/or age group predict how likely they are to be hospitalized, committed to the intensive care unit (ICU), and/or pass away due to the coronavirus or COVID-19. Currently, we have performed no cross validations, tests for accuracy, or tests for robustness. However, we plan to do so in the future.

1. Decision Trees

Our first decision tree, found in Figure 1, is a classification tree that predicts if a case from our dataset will fall into one of two categories: "Hospitalization" or "Not Hospitalized". The orange/beige color represents no hospitalization, the blue color represents hospitalization, and the white color represents a more mixed or impure distribution. The elements interpreted at each node are "samples", which are the number of observations reaching each node, "Gini", which represents impurity, and "values", which represent the counts by class.

The white node in the top shows the variable, "underlying medical condition", as the strongest predictor of hospitalization. Moving to the right, if the case has an underlying medical condition, they are more likely to be hospitalized. Venturing towards the left, if the case has no medical condition, they are less likely to be hospitalized.

If there is no medical condition, the next best predictor is age. If a case is not in the 60-69 year age group or older, they are not likely to be hospitalized. However, if they are in the 60-69 year age group or older, they are more likely to be hospitalized. Even if a case has an underlying medical condition, age still matters but has less impact.

The next, less important predictor, is race/ethnicity. This predictor appears much lower in the tree, indicating its lower relevance in the prediction model. The race predictor actually refines the risk found in groups with similar ages/underlying medical conditions. For example, a racial group might have higher hospitalization rates than other groups in the same age group.

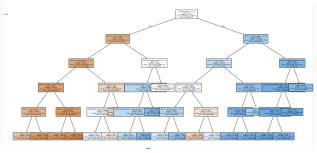

Figure 1. Decision tree that predicts for two categories: "Hospitalization" or "No Hospitalization"

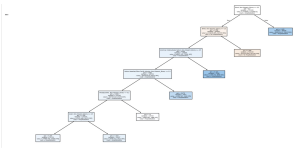

Figure 2. Decision tree that predicts for two categories: "Hospitalization" or "No Hospitalization" based on race

Our second decision tree predicts if a case from out dataset falls into either one of two categories: "Death" or "No Death". This tree has similar patterns to our first decision tree. The orange nodes predict "no death", the blue nodes predict "death", and the white node presents a mixed distribution. Once again, the most important predictors of death were the presence of underlying medical conditions and older age groups, while the less important predictor of death was race/ethnicity.

Our third decision tree, found in Figure 2, was much smaller, predicting hospitalization status based solely on race/ethnicity. The blue color represents hospitalization while the orange/beige color represents no hospitalization. For this tree, the strongest predictor of hospitalization was whether someone was Black/non-Hispanic. In fact, non-White minority groups were more likely to be hospitalized than White groups.

2. Statistics

To calculate some statistical measures based on our model, we decided to utilize a confusion matrix, found in Figure 3, to test for the accuracy level of our decision tree to make decisions. We decided to predict a binary outcome (0 vs.

Predicted	0	1
Actual		
0	162032	298
1	25221	284

Figure 3. Confusion matrix predicting the accuracy of our model

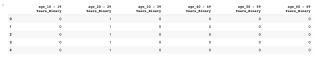

Figure 4. This table showcases the binary or dummy encoded age groups of our one-hot encoding.

1). ”0 to 0” is a true negative, which means that the model correctly predicted ”no hospitalization”). ”1 to 0” is a false negative, which means the model failed to detect a positive case. ”0 to 1” is a false positive, which represented none of our cases. ”1 to 1” is a true positive, which also represented none of our cases.

Based on the confusion matrix, we can see that the dataset is heavily imbalanced, with more dominant predictions of ”0” or ”non-hospitalization” cases. Therefore, we weighted classes in the decision tree in order to account for this major imbalance.

This same principle applied to other confusion matrices that examined ICU cases and death cases, so we weighted the classes in order to account for this imbalance.

3. Tables

We are in the process of developing tables to effectively represent our dataset. However, so far, we have been transforming the data to help us better generate decision trees. Specifically, we noticed that some of our variables are categorical, which is not conducive to successful and effective data analysis. Therefore, we had to transform some of those variables, such as underlying medical condition, hospitalization status, death status, and ICU status, to binary in order to produce better results.

For example, we used one-hot encoding to turn our categorical age group variable and our race variable into binary or dummy encoded age groups. The age group variable can be found in Figure 4. If a case has a ”1” in one of the age groups, that means that case is a member of that age group. If a case has a ”0” in one of the age groups, that means that case is not a member of that age group.

We continued to use one-hot encoding for the rest of the variables we wanted to examine that were categorical.

4. Main Results

Based on our confusion matrix assessing the accuracy of our dataset, our one-hot encoded tables, and our decision trees, we believe that the presence of an underlying medical condition is the most important predictor in determining whether someone will be hospitalized, committed to the ICU, or pass away from COVID-19. Specifically, if a person has an underlying medical condition, they are more likely to be hospitalized due to COVID-19 or worse.

The second most important predictor is someone’s membership in a certain age group. We discovered that people in the age group of 60-69 years and older are more likely to be hospitalized or pass away due to COVID-19. A less relevant predictor was race/ethnicity. We found that it refined risk found within similar health and age groups; for example, a subgroup within one age group could have higher rates of hospitalization because they were a different race.

Our analysis and identification of important predictors can help inform hospital triage decisions. For example, when admitting patients to the hospital to get treated for COVID-19, hospital workers can place more precedence on patients with underlying medical conditions because they are more vulnerable to death than other patients without underlying medical conditions. Additionally, patients who are 60-69 years or older can also be prioritized because they are also more vulnerable to COVID-19. Race/ethnicity is not as relevant, but it can still be used in decision-making.

5. Confidence in our Main Results

From our preliminary results, we firmly believe that the model could use some further improvement or tweaking to make it more accurate. Further cross validation and robustness tests are needed in order to further refine the decision-making ability of our model.

However, we do believe that our preliminary results are a promising start, and we are making good headway in terms of developing our model and ensuring that it is operating at the highest capacity. First, our dataset is highly imbalanced, which affects the interpretation of and stability of our decision trees. Although we used class weighting, additional techniques such as SMOTE oversampling, or stratified sampling may improve the model performance.

Further, our current analysis relies primarily on decision trees, which are interpretable but can be sensitive to noise. In the next phase, we plan to compare the decision-tree results with more robust models such as Random Forests, Gradient Boosted Trees, and Logistic Regression with regularization. These steps will help us better evaluate our predictors and strengthen the reliability of our conclusions.