Analyzing Long Covid-19 Data to Predict Long Covid-19 Cases

Katrina Dotzlaw, Ryan Dotzlaw, Sean Stzurm, Da Tan

Department of Computer Science

University of Manitoba

Winnipeg, MB, Canada

*Abstract*— Covid-19 data has been collected and analyzed since the start of the pandemic in 2019. Investigation of the post-Covid-19 condition (or Long Covid-19) began recently, and related data is being collected constantly. In this paper, we discover interesting associations in Long Covid-19 demographic data and cluster common symptoms. Using this information, we create a classifier that aims to predict the development of Long Covid-19 in patients. Our predictive model shows promise in identifying individuals at risk of developing Long Covid-19 and highlights demographic information that could indicate an increased risk of developing Long Covid-19.

Keywords—Covid-19, Long-Covid-19, data mining, prediction, association rules, clustering

# Introduction

Sean is writing

# Background and Related Works

## Long-Covid-19 Research

Sean is writing

## Related Works

Sean is writing

# Methodology

This section will describe the process of discovering association rules, examining the relationship between symptoms and demographics, and creating a classifier.

## Association Rule Mining

Our approach to association rule mining applies the Apriori algorithm to the demographic data provided by the US Census Bureau [1] and identifies frequent characteristics. We use the following python libraries in our algorithm:

* math
* numpy
* matplotlib.pyplot: for graph creation
* mlxtend.frequent\_patterns: for the Apriori function and association\_rules function
* pandas: for data structures used with mlxtend

The census data is preprocessed before being read into our mining function. Certain demographic information included in the census (like employment status) will not be indicative of developing Long Covid-19 and is removed. Once the data has been cleaned, we read it into an array-like data frame. Additional processing must be done to the data before frequent patterns can be mined. We turn the numeric patient ages into a range and store the ranges in a new column, named ‘*age range’.* Next, we identify columns that contain categorical data and split them into multiple columns using binary mapping. For example, the column ‘*birth gender’* is categorical, containing [‘M’, ‘F’]. After applying the binary map, the categorical birth gender data becomes [[1,0], [0,1]], where 1 indicates the presence of a feature and 0 indicates the absence of a feature.

After preprocessing, the following columns will be used in subsequent frequent pattern mining and association rule mining:

* age range: [18,36], (36, 54], (54, 72], 72+
* symptom severity: none, mild, moderate, severe
* race: White, Black, Hispanic, Asian, other
* birth gender: male, female
* current gender: male, female, transgender, other
* vaccinated: yes/ no
* long covid: yes/no
* impacted: yes/ no
* booster: yes/no
* number doses:
* treat oral: received oral antiviral medication yes/no
* treat mono: received monoclonal antibody medication yes/no
* current symptoms: yes/no

Before mining frequent patterns, we generate graphs of demographic information with respect to Covid-19 and Long Covid-19. The occurrences in each column are counted and normalized to a percentage that is then displayed in a graph. Comparisons between graphs will be discussed in later sections.

Minimum support of a frequent item is determined based on the following formula:

[SITE]

Essentially, the minimum support is , where x increases as the number of rows in the dataset increases. Using the Apriori function from mlxtend, the calculated minimum support, and the columns previously mentioned, we identify frequent item sets.

The frequent item sets are then used in conjunction with the association rules function from mlxtend to mine interesting association rules. We determine rule to be interesting if it meets the minimum confidence of 0.3, where confidence of every rule is calculated by: [2]. The association rules are then separated into two groups *long\_covid\_1* (where Long Covid-19 is the consequent) and *long\_covid\_0* (where **not** Long Covid-19 is the consequent). The former group of rules are sorted by confidence ascending.

## Demographic-Symptom Clustering

To create a classifier, we needed to explore the relationships between demographic groups and Long Covid-19 symptom combinations, focusing on symptoms that have significantly different frequencies among the demographic groups. Since this analysis required symptom data, we used the Kenya and Malawi Long Covid-19 survey, which consisted of 677 and 679 Long Covid-19 cases from Kenya and Malawi respectively [3]. To examine the differences in symptom occurrences in the demographic groups, we will calculate the frequencies of symptom in each group and perform a chi-square test to identify significant differences among the groups. The chi-square test will be done with a p-value threshold of 0.05.

Next, we will focus on the age-gender subgroups and compare the similarities between them in terms of the representative symptoms prevalent in each group. We cleaned the Kenya and Malawi datasets, eliminating social-state features[[1]](#footnote-1) that we deemed irrelevant to our study. Also, we reduced the number of symptoms by eliminating those with less than 5% of overall occurrence, since infrequent symptoms increase the dimensionality for the similarity comparison as well as the clustering algorithm.

After preprocessing, we had 6 age-gender subgroups and we calculated the occurrence percentages and computed the similarities within each group. Specifically, for each pair of age-gender subgroups and , we computed the cosine similarity [4] between them:

The higher the value, the more similar the two age-gender groups.

To explore the relationship between the inner structure of the symptoms and the age-gender demographics, we performed K-Mode clustering on the Kenya and Malawi survey datasets. Using the 15 symptoms, we optimized the number of clusters according to the metric of within-cluster Sum of Squares and the Sum of Squares over the total clusters. We then mapped the clusters to the original age subgroups and analyzed the results.

## Creating a Supervised Classifier

To explore how well the features in the US Census dataset can predict the development of Long Covid-19, we created a predictive classifier. First, we removed the features *‘current symptoms’* and *‘impacted’* as they are only available when an individual develops Long Covid-19. Then we removed the cases that did not have Covid-19 and those whose ‘*symptom severity’* is NA[[2]](#footnote-2). We also removed the ‘*treat oral’* and *‘treat mono’* features because more than 90% of the values are NAs.

After preprocessing, we had 23349 cases with 7 predictive variables. Among the 23349 cases that had Covid-19, 6468 of them developed Long Covid-19 and 16881 didn’t. Since most of the classification algorithms are sensitive to unbalance datasets, we randomly sampled 70% (4528) of the positive cases and took them in the training dataset. We also randomly sampled an equal number of negative cases and put them into the training set. All the remaining cases became the independent test set, which comprised 1940 positive and 12353 negative cases.

Before training the model, we performed descriptive analysis and feature selection. To do descriptive analysis, we examined all of the candidate predictors to determine which distinguisher is the strongest. The ‘Boruta\_8.0.0’ package from R 4.2.0 was used to perform feature selection. Boruta selects features by wrapping the Random Forest algorithm inside, and then randomly shuffles, trains the data, and reports importance ranks for the features [5]. Any features rejected by Boruta were removed from the preprocessed data. Then we trained the classifiers on the preprocessed data and evaluated it using AUC[[3]](#footnote-3).

Using the R package ‘rpart’, we built a decision tree. The R package ‘caret’ was used to tune 2 parameters: ‘*cp’* and ‘*max\_depth’*, representing the complexity of the tree and maximum depth the tree can grow respectively. We then preformed 10-fold cross validation to further tune the parameters.

We also built a Random Forest model using the R package ‘randomForest’. We tuned the parameter ‘*mtry*’ (number of features for growing a random tree) using the ‘tuneRF’ function. We then used the optimized ‘*mtry’* value to train a Random Forest model and evaluated it on the independent test dataset using the AUC.

# Analysis

In this section, the association rules found are explained and the demographic information is analyzed. Additionally, the correctness of the classifier is examined, and predictive results are explained.

## Demographic Analysis – Week 46

The demographic information relating to Covid-19 and Long Covid-19 from the week 46 US Census Bureau [ref] was mined and graphed to compare distributions. The results of the demographic analysis will be compared to the results of demographic information collected and examined from a later census dataset.

We found that most Covid-19 patients are in the 30-60 age range. As shown in fig.1, the smallest percentage of individuals with Covid-19 are younger than 30 and older than 80.

Similarly to the age distribution of Covid-19 patients, most of the individuals experiencing Long Covid-19 are in the 30-60 age range (fig.2.). Individuals younger than 30 and older than 80 had the smallest incidence of Long Covid-19.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Fig. 1. Age Distribution of Covid-19 Patients Fig. 2. Age Distribution of Long Covid-19 Patients

In this dataset, 62.39% of individuals with Covid-19 were assigned female at birth, while 37.61% were assigned male at birth (fig.3). The amount of individuals assigned female at birth increases to 72.68% when examining individuals with Long Covid-19 (fig.4). The percentage of individuals assigned male at birth that have Long Covid-19 is 27.32%. This could indicate that individuals assigned female at birth are more likely to develop Long Covid-19, although this would require further analysis. It is unclear if this is an accurate representation of the birth gender of Covid-19 patients, or if there was just a higher instance of individuals assigned female at birth that responded to this survey.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 3. Birth Gender Ratio for Covid-19 Patients Fig. 4. Birth Gender Ratio for Long Covid-19 Patients

Of the individuals with Covid-19 (fig.5), 60.91% identify as female while 36.79% identify as male. Only 0.41% identify as transgender and 1.89% identify as a different gender identity. The percentage of individuals who identify as female increases to 70.76% when examining individuals with Long Covid-19 (fig.6). The number of individuals identifying as male decreases to 26.63%. Transgender individuals and individuals that identify as a different gender identity make up 0.53% and 2.09% of the individuals with Long Covid-19. These results could indicate that more female-identifying individuals experience Long Covid-19.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 5. Gender Identity of Covid-19 Patients Fig. 6. Gender Identity of Long Covid-19 Patients

The distribution of ethnicities is generally the same for both Covid-19 patients (fig.7) and Long Covid-19 patients (fig.8). White individuals represented the majority of both Covid-19 patients and Long Covid-19 patients with 73.5% and 71.15% respectively. The percentage of Hispanic individuals with Covid-19 who developed Long Covid-19 increased slightly from 10.75% to 12.79%. The percentage of Black individuals with Covid-19 who had Long Covid-19 also increased from 7.1% to 8.11%. Only 4.16% of individuals with Covid-19 in this dataset were Asian, and 3.09% of individuals with Long Covid-19 were Asian. The percentage of mixed individuals remained similar with 4.48% having Covid-19 and 4.86% having Long Covid-19.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 7. Ethnicities of Covid-19 Patients Fig. 8. Ethnicities of Long Covid-19 Patients

The majority of Covid-19 patients in the week 46 dataset reported having either mild or moderate symptoms (fig.9), with 38.9% having mild symptoms and 41.33% having moderate symptoms. 14.48% of individuals with Covid-19 reported experiencing severe symptoms, while 5.3% reported having no symptoms. When examining symptom severity of Long Covid-19 patients, we found that 47.31% reported having moderate symptoms and 28.8% reported experiencing severe symptoms. This could indicate that Long Covid-19 symptoms are more likely to be moderate or severe, but further data analysis would need to be done. The number of individuals experiencing mild symptoms decreased to 22.4% when examining Long Covid-19 symptom severity. As seen in Fig.10., a small percentage of individuals reported experiencing no symptoms. These individuals could be asymptomatic and testing positive the amount of time required[[4]](#footnote-4) to receive a Long Covid-19 diagnosis, although it is more likely that they are part of the margin of error.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 9. Symptom Severity of Covid-19 Patients Fig. 10. Symptom Severity of Long Covid-19 Patients

85.75% of the individuals with Covid-19 (fig. 11) in this dataset received at least 1 vaccine while 14.25% were not vaccinated. When examining individuals with Long Covid-19 (fig.12), we found that 83.85% of individuals had at least 1 vaccine, while 16.15% were not vaccinated.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 11. Vaccination Rate of Covid-19 Patients Fig. 12. Vaccination Rate of Long Covid-19 Patients

## Explaining Interesting Association Rules – Week 46

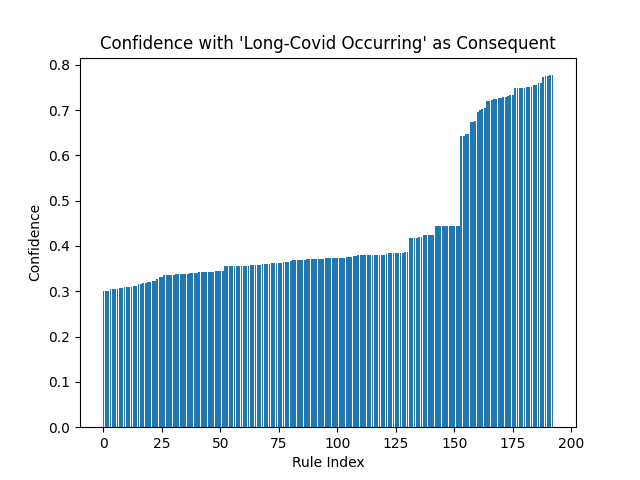


Fig. 13. Association Rules with ‘Long Covid Occurring’ as Consequent

Chart, histogram

Description automatically generated

Fig. 14. Association Rules with ‘Long Covid not Occurring’ as Consequent

As shown in fig. 14, we found over 800 rules with ‘*Long Covid Not Occurring’* in the consequent with the confidence ranging from approximately 0.5 to over 0.8. All of the rules are sorted by confidence ascending.

Our analysis found just under 200 association rules with ‘*Long Covid Occurring’* in the consequent (fig. 13). The confidence of these rules ranges from approximately 0.3 to approximately 0.8. Again, the rules are sorted by confidence ascending.

When examining the results of our association rule mining, we can see that there are approximately 70 outliers that have a confidence greater than 0.5 (fig.13). The rules with the highest confidence (confidence >= 0.77) are as follows:

* *(Birth\_Gender\_F˄Current\_Symptoms\_Yes˄Treat\_Oral\_0) →Long Covid*
* *(Birth\_Gender\_F˄Current\_Symptoms\_Yes˄Current\_Gender\_F˄Treat\_Oral\_0) →Long Covid*
* *(Current\_Symptoms\_Yes˄Current\_Gender\_F˄Treat\_Oral\_0) →Long Covid*
* *(Birth\_Gender\_F˄Current\_Symptoms\_Yes˄Treat\_Mono\_0˄Treat\_Oral\_0)→Long Covid*
* *(Current\_Symptoms\_Yes ˄Treat\_Mono\_0˄Current\_Gender\_F˄Treat\_Oral) →Long Covid*

The rules listed above indicate that both individuals assigned female at birth and individuals that identify as female are more likely to develop Long Covid-19. However, our dataset contains a majority of individuals that are assigned female at birth and identify as female. Further analysis would need to be done to determine if this rule is widely applicable.

Additionally, our association rules show that individuals that have not received either antiviral medication or monoclonal antibody medication are more likely to develop Long Covid-19. This makes sense because not receiving treatment for an initial Covid-19 infection is more likely to cause symptoms to persist for the required amount of time to receive a Long Covid-19 diagnosis.

## Demographic Analysis – Week 49

The demographic information from the week 49 dataset from the US Census Bureau [ref] will be examined and compared to the results of the week 46 analysis.

The age distribution of Covid-19 patients (fig. 15) for this dataset is similar to the distribution of the week 46 dataset (fig.1), although there is a sharper peak around age 40. The ages of the Long Covid-19 patients also had a sharper peak at 40 (fig. 16) but had a larger cluster of data in the 30-60 age range than in the Covid-19 data. The distribution of Long Covid-19 ages is similar to that of the week 46 data(fig.2).

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Fig. 15. Age Distribution of Covid-19 Patients Fig. 16. Age Distribution of Long Covid-19 Patients

According to the week 49 Covid-19 data analyzed, 55.89% of individuals were assigned female at birth. 44.11% were assigned male at birth. Compared to the week 46 data analyzed, the percentage of individuals assigned female at birth decreased, and the number of individuals assigned male at birth increased.

For the week 49 data, there is a higher percentage of individuals assigned female at birth who have Covid-19 (fig.17). Fig.18. shows that 67.43% of individuals assigned female at birth reported experiencing Long Covid-19. 32.57% of individuals assigned male at birth reported experiencing Long Covid-19. Since both datasets have a high population of individuals assigned female at birth experiencing Long Covid-19, it is possible that those individuals are more likely to develop Long Covid-19.

Chart, pie chart

Description automatically generated



Fig. 17. Birth Gender Ratio of Covid-19 Patients Fig. 18. . Birth Gender Ratio of Long Covid-19 Patients

When examining the gender identity of individuals with Covid-19, we found that 55.08% of patients currently identify as female and 43.46% currently identify as male (fig.19). Individuals who currently identify as transgender make up 0.47% of patients in this data set, while 1.0% of patients reported identifying as a different gender identity. This ratio indicates that there is higher percentage of female-identifying Covid-19 patients than male-identifying. Compared to the results of the analysis done on the week 46 data, there is a lower percentage of individuals who identify as female present in this dataset.

In our analysis of Long Covid-19 (fig.20), we found that the number of female-identifying individuals increased to 66.07% and the number of male-identifying individuals decreased to 31.83%. The number of transgender individuals and individuals that have a different gender identity also increased to 0.65% and 1.45% respectively. Similar to the results of the week 46 analysis, the increase in individuals that identify as female experiencing Long Covid-19 could indicate that female-identifying individuals are more likely to experience Long Covid-19.

Fig.19. Ratio of Gender Identity in Covid-19 Patients Fig.20. Ratio of Gender Identity in Long Covid-19 Patients



Chart, pie chart

Description automatically generated

According to our analysis on week 49 data, 76.45% of the Covid-19 patients are White. Hispanic individuals made up 9.52% of Covid-19 patients. 5.77% of individuals in this dataset are Black, while 3.95% are Asian. 4.3% of individuals with Covid-19 are mixed. Fig.21. shows that patients who are White make up most of the Covid-19 patients included in this dataset. Compared to the week 46 dataset, there is an increased number of White, Black, Hispanic, Asian, and mixed individuals in this dataset that experienced Covid-19.

Of the individuals reporting to have experienced Long Covid-19, 73.82% are White, 11.32% are Hispanic, 6.64% are Black, 5.73% are mixed, and 2.49% are Asian (Fig.22.). This could indicate that White individuals are more likely to experience Long Covid-19, however, recall that 76.45% of the Covid-19 patients in this data set are White and that demographic information can be skewed by inaccessible testing and treatment. Compared to the results from week 46, there is an increase in Black, Hispanic, Asian, and mixed individuals that develop Long Covid-19.





Fig.21. Ethnicity of Covid-19 Patients Fig.22. Ethnicity of Long Covid-19 Patients

The majority of Covid-19 patients reported experiencing mild or moderate symptoms, which is the same as the results of the week 46 analysis. As shown in Fig.23., 41.61% of individuals had mild symptoms and 41.77% had moderate symptoms. Severe symptoms – like hospitalization – were reported by 11.46% of Covid-19 patients in this data set. Only 5.16% of Covid-19 patients in this dataset reported experiencing no symptoms. Again, note that these results are from a self-reported survey. Symptom severity is relatively subjective and may not match the opinion of a health professional.

According to our analysis, symptom severity increased in individuals experiencing Long Covid-19, with the majority of individuals reporting moderate or severe symptoms. Since this matches the results from our previous analysis on week 46 data, it could be said that symptom severity increases when experiencing Long Covid-19. 49.46% of individuals reported experiencing moderate symptoms, while 25.45% of individuals reported experiencing severe symptoms. 23.47% of individuals with Long Covid-19 reported mild symptoms.





Fig.23. Symptom Severity in Covid-19 Patients Fig. 24. Symptom Severity of Long Covid-19 Patients

Among the Covid-19 patients in this dataset, 85.67% have received at least 1 vaccination, while 14.33% are unvaccinated (Fig. 25).

Similarly, 84.6% of individuals with Long Covid-19 reported to have received at least 1 vaccination. 15.4% of unvaccinated individuals reported experiencing Long Covid-19(Fig. 26.). Since these results follow the results of our previous analysis, it could indicate that Long Covid-19 is more likely to develop as a result of a breakthrough infection[[5]](#footnote-5), though this would need further research.





Fig.25. Vaccination Rate Among Covid-19 Patients Fig. 26. Vaccination Rate Among Long Covid-19 Patients

## Explaining Interesting Association Rules – Week 49

In our dataset, 27.56% of individuals reported experiencing Long Covid-19, while 72.44% reported not experiencing it. We are mainly interested in associations concerning Long Covid-19, although rules with ‘*long covid not occurring’* are also found.

An association rule with ‘*Long Covid Occurring’* as the consequent is of the form: .

An association rule with ‘*Long Covid Not Occurring’* as the consequent is of the form:

Comparing the graphs with ‘*Long Covid Occurring’* (fig. 27.)and ‘*Long Covid Not Occurring’* (fig. 28.) in the consequent, we can see that there are more rules in the latter group. Since 72.44% of patients in this dataset reported not experiencing Long Covid-19, it makes sense that we found more associations with ‘*Long Covid Not Occurring’* in the consequent*.*



Fig. 27. Confidence of Rules with ‘Long Covid Occurring’ in the Consequent

Chart, histogram

Description automatically generated

Fig. 28. Confidence of Rules with ‘Long Covid Not Occurring’ in the Consequent

Over 400 rules having ‘*Long Covid Not Occurring’* in the consequent were found, with confidence ranging from approximately 0.6 to greater than 0.8. The found rules were sorted in ascending order by confidence.

Our association rule mining discovered approximately 33 interesting associations with ‘*Long Covid Occurring’* in the consequent. The rules are sorted by confidence ascending. The confidence of all interesting rules ranges from approximately 0.3 to over 0.8.

As shown in fig. 27, there are approximately 5 outliers which indicates that those rules have a significantly higher confidence than the other rules found. The most significant rules all have a confidence of over 0.8 and are defined as follows:

* *(Current\_Symptoms\_Yes ˄Number\_Doses\_1.0˄ Vaccinated\_1) →Long Covid*
* *(Current\_Symptoms\_Yes ˄Number\_Doses\_1.0) → Long Covid*
* *(Current\_Symptoms\_Yes˄Race\_White)→Long Covid*
* *(Current\_Symptoms\_Yes ˄Birth\_Gender\_F)→Long Covid*
* *(Current\_Symptoms\_Yes)→Long Covid*

The above association rules indicate that vaccinated individuals are more likely to develop Long Covid-19, which supports the theory that Long Covid-19 could occur as the result of a breakthrough infection. Additionally, the rules we found show that White individuals are more likely to develop Long Covid-19. However, the majority of the individuals in this dataset are White so this rule might not be significant in practice. Also, we discovered that individuals assigned female at birth are more likely to develop Long Covid-19. The same caveat applies here as with ethnicity; the majority of our dataset was individuals assigned female at birth.

## Demographic-Symptom Clustering

As mentioned in *methodology,* we analyzed the prevalence of symptoms within demographic groups. For all the demographic groups, the frequencies of the symptoms were reported (fig. 29.) with headache, cough, and fatigue being the most prevalent Long Covid-19 symptoms.

Chart, bar chart

Description automatically generated

Fig. 29. Rank of Long Covid-19 Symptom Occurrences

We also discovered that there are significant differences in symptom occurrences among the demographic subgroups (fig. 30). Our analysis showed that Long Covid-19 patients of <=50 years old experienced more symptoms such as non-communicable diseases, pre-existing conditions, fatigue, and joint pain. This is reasonable since older people tend to have such conditions, although it is hard to link this to Long Covid-19 since there is a lack of control datasets that are Long Covid-19 negative.

We also noticed that in this dataset, more women experienced symptoms such as shortness of breath, chest pain, etc (fig. 30).

Chart, bar chart, waterfall chart

Description automatically generated

Fig. 30. Symptom Occurrences among Different Demographic Groups

Moreover, the prevalence of many symptoms such as cough, running nose, recurrent fever, etc. was significantly different between the two countries Kenya and Malawi. We are not able to trace the reason for this bias, which may be due to the different Covid-19 subtypes in these two countries, or the time distance of conducting the two surveys. In addition, we noticed that there was a statistical difference in symptom prevalence between the employed and unemployed groups (fig.30). The unemployed group seems to have a higher chance of developing symptoms such as fatigue, cough, etc. This might be because unemployed people on average tend to be elderly or have worse health conditions than the employed population. This is supported by the fact that pre-existing conditions and non-communicable diseases are much more prevalent in the unemployed population (Fig.30). There are other demographic differences in features of living type and the number of people living with them, but we still lack evidence to trace the cause of such deviations.

When comparing the symptoms of age-gender subgroups, we found a significant dissimilarity between the group ‘*age > 50’* and *‘age <= 50’* (fig. 31.)*.* We did not find any obvious differences between the gender groups or between the younger age groups. Table 1 shows the age-gender subgroups along with the associated cosine similarity.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 18-33\_Male | 34-50\_Female | 34-50\_Male | 18-33\_Female | >50\_Male | >50\_Female |
| 18-33\_Male | 1 | 0.976 | 0.976 | 0.985 | 0.915 | 0.887 |
| 34-50\_Female | 0.976 | 1 | 0.991 | 0.966 | 0.964 | 0.96 |
| 34-50\_Male | 0.976 | 0.991 | 1 | 0.96 | 0.962 | 0.948 |
| 18-33\_Female | 0.985 | 0.966 | 0.96 | 1 | 0.889 | 0.871 |
| >50\_Male | 0.915 | 0.964 | 0.962 | 0.889 | 1 | 0.975 |
| >50\_Female | 0.887 | 0.96 | 0.948 | 0.871 | 0.975 | 1 |

Table 1. Cosine Similarities among the Age-Gender Groups

Square

Description automatically generated with low confidence

Fig. 31. Heatmap Showing the Similarities among the Age-Gender Groups

Using the unsupervised learning methods described in *methodology,* we found that the optimal number of clusters is 11 (fig.32). Usually in K-Mode clustering the elbow point indicates an optimal K value, however choosing K=11 at the elbow point resulted in too many clusters. Table 2 demonstrates the 11 modes representing the 11 clusters.

Chart, line chart

Description automatically generated

Fig. 32. The Optimization Path for the Number of Clusters

A screenshot of a game

Description automatically generated with medium confidence

Table 2. The 11 Modes for the K-Mode Clustering

The results of mapping each mode to the original age group is shown in table 3. No obvious overlaps were found between the clustering results and the age groups. This might be because most of the symptoms occurred in a low frequency and that there are few symptoms (fig. 30.) that occur with significantly different frequencies between the age groups. However, we can see high concentrations of the modes ‘*non-communicable diseases’* and *‘pre-existing conditions’* for the >50 group in cluster 3. As previously mentioned, these modes tend to be more prevalent in individuals >50, so these results make sense.

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster/ age group | 18-33 | 34-50 | >50 |
| 1 | 90 | 80 | 11 |
| 2 | 1 | 24 | 10 |
| 3 | 8 | 34 | 21 |
| 4 | 26 | 23 | 0 |
| 5 | 82 | 44 | 7 |
| 6 | 17 | 50 | 9 |
| 7 | 157 | 129 | 14 |
| 8 | 67 | 47 | 8 |
| 9 | 186 | 131 | 16 |
| 10 | 25 | 24 | 1 |
| 11 | 3 | 8 | 3 |

Table 3. Mapping of the 11 Clusters to the 3 Age Groups

## Classifier Results

Before we built the classifier, we performed descriptive analysis and feature selection on our dataset. We found 7 candidate predictors with only 1 numerical feature, *‘age’.* The other 6 features are categorical. As shown in fig. 33, we noticed that only ‘*symptom severity’* seemed to be a strong distinguisher.

Chart, bar chart

Description automatically generated

Fig. 33. Results of the Descriptive Analysis on the US Census Dataset

During the feature selection process, the Boruta algorithm (table.4) rejected ‘*vaccinated’* as a valid predictor and labeled ‘*age’* as tentative, while confirming the other predictors. Among the confirmed predictors, ‘*symptom severity’* was reported to have the highest importance rank, which agrees with our descriptive analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | meanImp | medianImp | minImp | maxImp | normHits | decision |
| AGE | 2.731 | 2.886 | -3.977 | 7.797 | 0.424 | Tentative |
| RACE | 7.793 | 7.724 | 3.293 | 13.994 | 0.96 | Confirmed |
| BIRTH\_GENDER | 24.65 | 24.551 | 16.532 | 30.367 | 1 | Confirmed |
| VACCINATED | 0.013 | -0.449 | -4.021 | 3.907 | 0.01 | Rejected |
| NUMBER\_DOSES | 7.922 | 7.951 | 3.177 | 13.942 | 0.96 | Confirmed |
| BOOSTER | 14.005 | 14.124 | 5.014 | 21.78 | 1 | Confirmed |
| SYMPTOM\_SEVERITY | 111.986 | 112.556 | 96.694 | 123.554 | 1 | Confirmed |

Table 4. Feature Selection Results from the Boruta Algorithm

After removing the rejected features, we built a decision tree (fig. 35) using the optimized hyperparameters, ‘*cp*=0.003’ and ‘*maxdepth*=3’ (fig. 34). The model was then tested on the test dataset and the AUC was calculated to be 0.706 (±0.011).

Next, we built a Random Forest model using the optimized hyperparameter, *‘mtry=2’.* The model was trained using our training data and evaluated using the independent test data (fig. 34). The AUC for this model is 0.721 (±0.011). The feature importance rank returned by the Random Forest model was recorded in table 5.

Chart

Description automatically generated with low confidence

Fig. 34. The optimization paths for parameters of the decision tree. The tuning metric is AUC, and the parameters tuned are complexity and maximal depth. The subgraphs at the second line indicate the performances of the two models Decision tree and Random Forest, in terms of AUC.

Timeline

Description automatically generated

Fig. 35. The Decision Tree Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | no | yes | MeanDecreaseAccuracy | MeanDecreaseGini |
| AGE | -0.002 | 0.003 | 0.001 | 198.178 |
| RACE | 0.004 | 0.004 | 0.004 | 75.789 |
| BIRTH\_GENDER | 0.008 | 0.023 | 0.015 | 79.842 |
| NUMBER\_DOSES | 0.009 | -0.002 | 0.004 | 38.779 |
| BOOSTER | 0.015 | 0.001 | 0.008 | 36.166 |
| SYMPTOM\_SEVERITY | 0.091 | 0.1 | 0.095 | 472.718 |

Table 5. The Feature Importance Rank Returned by Random Forest

The feature importance rank (table 5.) shows that the most important feature for the Random Forest classifier is ‘*symptom severity’* followed by *‘age’,* which agrees with our descriptive analysis (fig. 33).

Although our descriptive analysis shows that most of the predictors are not good intuitively, the performances of the 2 classifiers are acceptable with AUCs of 0.706 and 0.721 for the decision tree and Random Forest classifier respectively. When comparing the AUCs, we can see that the Random Forest classifier performed better. In addition to AUC, we also evaluated our model using accuracy, F1, and other metrics using a default threshold of 0.5 (table 6.), although AUC remains our primary evaluation metric.

|  |  |  |
| --- | --- | --- |
| metric/model | Decision tree | Random Forest |
| AUC | 0.706 | **0.721** |
| Accuracy | **0.711** | 0.676 |
| Sensitivity | **0.723** | 0.678 |
| Specificity | 0.607 | **0.668** |
| Precision | 0.922 | **0.929** |
| Recall | **0.727** | 0.678 |
| F1 | **0.813** | 0.784 |

Table 6. Performance Metrics of the Decision Tree and Random Forest Models

# Conclusions

## Conclusions

## Limitations

Since Long Covid-19 is an active area of research, it was difficult to find open-source data to analyze. Most of the data sets we found required credentials from a reputable institution. Additionally, some of the data sets we originally planned to analyze were removed, presumably for private use by the CDC or governments. Due to the removal of data sets, we were unable to examine connections between Covid-19 variants and Long Covid-19 diagnosis.

Almost all the data sets we analyzed were self-reported surveys, which can often be biased and exclusionary. Individuals experiencing Covid-19 and Long Covid-19 that do not have internet access would likely be excluded from such surveys. Individuals that do not have access to safe and reliable health care might report symptoms inaccurately.

We were unable to find a publicly accessible data set containing Covid-19 symptoms and Long Covid-19 symptoms to develop a symptom-based predictive model. Again, this is likely because Long Covid-19 is an active area of research.

## Future Work

Ideally, Long Covid-19 data will become publicly available as research into the virus progresses. Finding supervised data that contains both Covid-19 and Long Covid-19 data would allow the creation of a symptom-based classifier.

Additionally, Covid-19 variant data could be used to find associations between different variants and the development of Long Covid-19. Our predictive model could be expanded to predict the development of Long Covid-19 based on variant diagnosis.

Further investigation could be done into the relationship between vaccination status and the development of Long Covid-19, as our demographic analysis indicated that Long Covid-19 could arise as a result of a breakthrough infection.

##### Acknowledgment

We would like to acknowledge the individual contributions of group members to this project. We would like to thank Sean Stzurm for his work cleaning the datasets, researching Long Covid-19, creating and narrating the project presentation, and writing the introduction and background information sections of this paper. We would like to thank Ryan Dotzlaw for his work preprocessing datasets, performing association rule mining, creating demographic graphs, and editing this paper. We would like to thank Da Tan for his work for clustering demographic and symptom data, creating a classifier, writing the methodology section of this paper, writing subsections of the analysis section (E, F), and editing the paper. We would like to thank Katrina Dotzlaw for her work researching Long Covid-19, analyzing the demographic information and association rules, writing the methodology section of this paper, writing parts (A, B, C, D) of the analysis section of this paper, editing and formatting subsections of the analysis section (E, F), and writing all sections in conclusions.

# References

|  |  |
| --- | --- |
| [1] | United States Census Bureau, "Hosehold Pulse Survey Public Use File (PUF)," 14-28; 1-13 September; June 2022. [Online]. Available: https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html. [Accessed September 2022]. |
| [2] | J. Hipp, U. Guntzer and G. Nakhaeizadeh, "Algorithms for Association Rule Mining - A General Survey and Comparison," *ACM SIGKDD Explorations Newsletter,* vol. 2, no. 1, pp. 58-64, 2000. |
| [3] | Humanitarian Data Exchange, "Kenya, Malawi, Long Covid-19 Effects Survey Dataset," OCHA, 20 January 2022. [Online]. Available: https://data.humdata.org/dataset/long-covidresearchagenda. [Accessed September 2022]. |
| [4] | K. Park, J. S. Hong and W. Kim, "A Methodology Combining Cosine Similarity with Classifer for Text Classification," *Applied Artificial Intelligence,* vol. 34, no. 5, pp. 369-411, 2020. |
| [5] | M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta package," *Journal of Statistical Software,* vol. 36, no. 11, pp. 1-13, 2010. |

1. Social-state features would include employment status, country, living in a rural or urban area, and the amount of people present in a household. [↑](#footnote-ref-1)
2. Less than 1% [↑](#footnote-ref-2)
3. Where AUC is the area under the curve. [↑](#footnote-ref-3)
4. The required amount of time symptoms persist after initial Covid-19 diagnosis is debated among health organizations. [↑](#footnote-ref-4)
5. An infection of vaccinated individuals. [↑](#footnote-ref-5)