CMPT 459 D1 Group Project Milestone 1

Vishal Krishna Venkatakumar: vvenkata@sfu.ca

Manthan Desai: mda74@sfu.ca

Evan Coulter: ercoulte@sfu.ca

1.1

The number of cases per outcome group after cleaning messy outcome labels is as follows

deceased: 4031

hospitalized: 135726

nonhospitalized: 779

recovered: 65310

1.2

The data mining task of predicting the outcome group labels in the cases dataset is Classification.

1.4

After removing the entries with missing age values in the cases dataset we then reduced all the various age value formats to be the same. We found that most age values were already in the correct format of just a single integer. However there were some formats that didn't match up correctly, for example, "22-80", "22-", and "3.5". For the "22-80" case we took the average of the two values e.g. (80+22)/2 = 51. For the "22-" case we just extracted the integer value from this, e.g. "22-" = 22. Finally for the 3.5 case we just removed the extra decimal and fraction value, e.g. "3.5" = 3.

In addition to cleaning the age values, we imputed missing data for the province and country columns in the cases dataset. To do so we used Geopy, a 3rd party python package that allowed us to use Georeversal along with the latitude and longitude columns to impute the missing province or country data.

1.5

One attribute we found outliers in was the Case_Fatality_Ratio in the locations_2021 dataset. We initially manually inspected the dataset and found some fatality_ratios that were clearly outliers such as entries that had fatality ratios over 100% or exactly 0%. In addition, we removed entries with less than 2000 confirmed cases as we felt entries below this threshold seemed to show much more variance in their Case_Fatality_Ratio. Afterwards we used a boxplot to visualize the outliers remaining and removed those as their Case_Fatality_Ratio was too high and were outliers under a boxplot test. See /code/plots/task-1.3/Figure_5_boxplot_of_case_fatality_ratio_some_outliers_removed.png for the box plot visualization.

In addition, we also found outliers in the outcome_group_column, specifically when Chronic_disease_binary is true and outcome_group_column equaled anything but deceased. We found this through the pie chart visualization /code/plots/task-1.3/Figure_7_outcome_group_distribution_for_chronic_disease_cases.png where it can be seen that an overwhelming majority of cases where Chronic_disease_binary = true resulted in outcome_group_column = deceased.

Number of rows in joined datasets

cases_train: 23016

cases_test: 11366

1.7

The features we deem as important are age, chronic_disease_binary, country, and fatality_rate. The rest of the features are not particularly relevant in classifying the cases into outcome groups. We determined age to be an important feature based on the results of a series of pie charts depicted in /plots/task-1.3/Figure_6_pie_charts_outcome_group_change_with_age.png, where the dataset was divided into age groups and the outcome group ratios per age group are depicted. An increase in the mortality rate is clearly visible in the older age groups therefore age must be considered in the classification of the outcome_group class. Chronic_disease_binary was chosen because in Figure 7, the pie chart of outcome groups for cases where chronic_disease_binary is true depicts a very high mortality rate of the patients suggesting a correlation between the feature and the outcome_group class. As for country and fatality_rate we choose these features as they could be easily matched to our entries in the cases dataset joined on the country column and the fatality rate would help to provide more accurate classifications about whether or not a case entry's outcome_group results in a value of deceased.

Several features were discarded for this classification task, such as latitude, longitude, province, and sex. The first three of those features were discarded because they were made redundant due to selection of the country feature after the cases and location datasets were merged. Sex attribute was discarded after pie charts outlining the distribution of outcome groups based on each gender showed no noticeable differences. Several other attributes like date confirmed, source, and additional information were not considerable because they had no relevant information to the classification task.