Cmpt 459 Milestone 1 Report

Group members: Cheng Zhang 301364914, Zi Heng Wang 301303595, Shan Ying Yang 301305759

1.1 Cleaning messy outcome label

After simplifying the outcome label, we can find that hospitalized cases are dominating the majority of the data. More than 135 thousand patients are under protection from hospitals. In addition, it is clear to see that over 65 thousand patients recovered from COVID-19. The number of nonhospitalized and deceased cases are 779 and 4031 respectively, which are significantly smaller compared to the other two types of cases. In 1.1, the cleaning of outcome labels resolve the ambiguities and inconsistencies in the data, which benefits the further prediction on the outcome_group. The count for the “deceased” group is 4031, for “hospitalized” is 135726 for “nonhospitalized” is 779, for “recovered” is 65310.

1.2 Outcome label

The prediction of the outcome_group labels in the cases_2021_train.csv and cases_2021_test.csv should be included in the “predictive” type of data mining task.

1.3 Exploratory Data Analysis(EDA)

The exploratory data analysis can be found within eda.ipynb, where various data visualization techniques were performed to better understand the raw datasets. The data visualizations and generated missing value statistics are saved as figures in the plots folder.

1.4 Data Cleaning and Imputing Missing Values

The columns with missing data values for both of the training and testing datasets are age, sex, province, country, date_confirmation, additional_information, and source. All the entries with missing age values are removed because the amount of missing data is overly vast to impute. The following tables explain how missing values from different data files are processed during the data cleaning steps.

Table 1. Data Cleaning/Data Imputation for the Training and Testing Datasets

Missing Values	Data Cleaning Explanation
age	There are 182,793 and 90,013 rows of missing age entries in the training dataset and testing dataset, respectively. All age values are rounded to the nearest integer, and all the ranged age values are reformatted by taking the average of the given range. For example, the value of “20-29” is recomputed as (20+29)/2 that gives 24.5, and rounds to 24.
sex	The gender column also has a large amount of missing data. Since estimating the gender seems unreasonable to impute the missing values, the missing gender values are categorized with a new class of “unknown”.
province/country	For the missing provinces and states we used Reverse Geocoding to compute the value using the corresponding latitude and longitude coordinates. The filled in provinces can be useful data for analyzing the correlations between certain geographic areas and COVID cases. The only country that is missing for both of the training and testing files is “Taiwan”, so we manually filled in the missing country as “China”.
date_confirmation	The missing date_confirmation values are filled with the mode of the date_confirmation column, in which the most frequent confirmation date appearing in the column substitutes all of the missing dates. For dates values that are recorded as a range, the first appearing date is selected to replace its original value.
additional_information/ source	The missing additional and source values of both files are imputed by replacing the null values with an empty string.
outcome_group	The entire column of outcome_group values for the testing file has missing values, and is imputed by replacing the column values with an empty string.
latitude/longitude	The latitude and longitude for training and testing datasets do not have missing values.

Table 1. Data Cleaning/Data Imputation for the Location Datasets

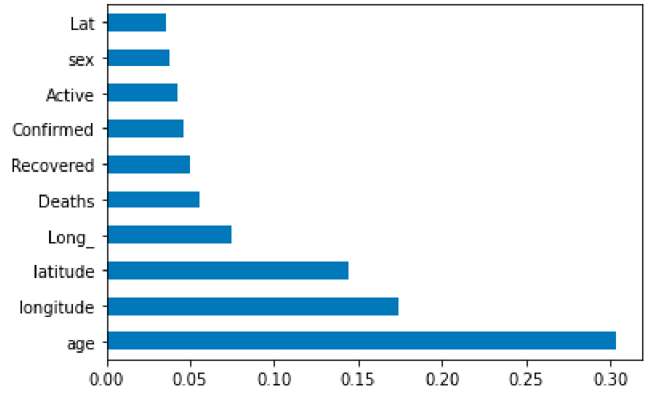
Missing values	Data cleaning explanation
Province_State	Similar to how provinces for the training/testing data are imputed, the Province_State for the location file is imputed by reverse geocoding with the given latitude and longitude. However, some of the latitude and longitude values are missing for the entries as well, so provinces cannot be found with the lack of information.
Lat/Long_	The Lat and Long_ columns have 89 null values. As a result, the entries with missing Lat/Long_ values are removed from the DataFrame, considering that they also lack the Incident_Rate attributes and Case_Fatality_Ratio. Since we removed all missing lat/long, entries with null Province_State also get removed.
Recovered/ Active/ Incident_Rate/ Case_Fatality_Ratio	The Recovered and Active columns both have 3275 null values, which is unreasonable to drop all of them considering the location file only has 4003 entries. The Incident_Rate and the Case_Fatality_Ratio have 90 and 48 null values correspondingly. For Milestone 1, the missing values for Recovered, Active, Incident_Rate, Case_Fatality_Ratio columns are imputed by replacing them with the mean value of the similar Country_Regions. For example, the entries that have null Recovered value for “Canada” will be replaced with the mean of “Canada”’s all non-null Recovered value.

1.5 Dealing with Outliers

For detecting the outlier section, we made two boxplots to visualize the existing outliers in the “location_2021_processed.csv”. The first boxplot covers these attributes: “Confirmed”, “Deaths”, “Active”, and “Recovered”. The second boxplot covers these attributes: “Case_Fatality_Ratio” and “Incident_Rate”.For the first boxplot, all attributes except “Deaths” had a vast amount of outliers. However, it is difficult to categorize these data as outliers considering different locations have different values for each attribute, and these data points that lie outside of the whisker cannot explain that distant data points are outliers for these categories. On the other hand, the second boxplot shows that the “Incident_Rate” has a cluster of data points that lie above the maximum of the whiskers. For similar reasons, the amount of these distant data are overly large so we cannot impute them as easily. Therefore, we decided to keep these data for future training prior to performing classification tasks.

1.6 Joining the Cases and Location Dataset

The joining of the Individual Cases and Location Datasets we chose to join based on “Country” and “Province” pairs. This was done in three phases. Firstly, l rename the irregular countries to the same form as in the cases_train. For instance, we change “US” to “United States” and “Korea, South” to “South Korea” so it’s compatible with the ‘country’ column. Secondly, we extracted the Individual Cases that had known Country and Province attributes and joined those cases with the corresponding data for those Country and Province pairs in the Locations dataset. This became our working data frame. Finally, to handle the individual cases which only had a known Country attribute (Province value is empty), we decided to aggregate entries in the Locations dataset to the Country level.



1.7 Feature Selection

For feature selection, l used feature importance to figure out the score for each feature of my data in 1.6. Feature importance is an inbuilt class that comes with Tree Based Classifiers, and l used Extra Tree Classifier for extracting the top 10 features for the datasets. In the plot, we find that age, latitude and longitude domains the majority of feature scores. In conclusion, we should discard “Combined Key”, “additional information”, “chronic_disease_binary”, “Last_Update” , while all 10 features mentioned in the plot should be selected.