REPORT
MILESTONE 1
CMPT 459 – DATA MINING

SUBMITTED BY:

SAAYAN MAHESANIA

NAVJOT KAUR

1.1 Cleaning Messy Outcome Labels

The outcome of training data is cleaned here into new column outcome_group. where the cases for each outcome_group label are:

```
1.1 Number of cases:
outcome_group
deceased              1802
hospitalized         22084
nonhospitalized        312
recovered             7253
dtype: int64
```

1.2 Outcome Labels

The data mining task for prediction of outcome group labels is **Feature extraction** since an additional feature is derived, i.e., outcome_group from outcome.

1.3 Exploratory Data Analysis (EDA)

In this part, various plots were constructed for all three given datasets to visualize the data which can be found in the plots folder. Also, the count for missing values in the datasets was computed:

```
MISSING VALUES
CASES_TRAIN_2021_____________________________
age: 182793
sex: 180362
province: 604
country: 1
latitude: 0
longitude: 0
date_confirmation: 51
additional_information: 193444
source: 18394
chronic_disease_binary: 0
outcome: 0
outcome_group: 205846
```

```
CASES_TEST_2021_____________________________
age: 90013
sex: 88765
province: 296
country: 1
latitude: 0
longitude: 0
date_confirmation: 23
additional_information: 95206
source: 9160
chronic_disease_binary: 0
outcome_group: 101387
```

```
LOCATION_2021_____________________________
Province_State: 174
Country_Region: 0
Last_Update: 0
Lat: 89
Long_: 89
Confirmed: 0
Deaths: 0
Recovered: 3275
Active: 3275
Combined_Key: 0
Incident_Rate: 90
Case_Fatality_Ratio: 48
```

Plots for cases_train and cases_test and location:

1. Country (train_country, test_country, location_country) with count for all countries.
2. Province (train_province, test_province, location_province) where cases plot is for India and location plot is for US because the count for these countries was the highest.
3. Lat_long (train_lat_long, test_lat_long, location_lat_long). It is a scatterplot.
4. Date (train_date and test_date) by plotting months with the count of cases.

1.4 Data Cleaning and Imputing Missing Values

Here, all the missing values were imputed or removed. Also, the data is cleaned in this step.

For cases_train and cases_test dataset:

1. For the age, all the missing values were removed. It was seen that there were some float values and some range intervals where we rounded the former values and took the mean for the latter values.
2. For province, the latitude, longitude, and country data are used, and the values are predicted using K-Nearest Neighbors (KNN) algorithm because nearer the point to the cluster of longitude and latitude the high chances that it is that cluster province.
3. For sex, the missing values were replaced by ‘unknown’.

For location dataset:

1. The missing values for province were marked as unknown because the data is not used.
2. The missing values for latitude and longitude were removed because the data is not used.

3. The missing values for case_fatality_ratio is removed because on visualization many values for Confirmed, Deaths, Recovered, and Active for that rows were 0 and there was no way to calculate the ratio.

1.5 Dealing with Outliers

- The attributes that have outliers:
  For cases_train datasets: age (19), latitude (167), longitude (43).
  For cases_test dataset: age (8), latitude (91), longitude (18).
  For location dataset: latitude (106), longitude (145), confirmed (45), deaths (35), recovered (13), active (2), incident_rate (15), case_fatality_ratio (58).
- The outliers are detected by calculating z-scores of the quantitative components.
- For cases data, every outlier was removed except for age because mostly the outliers contained ages from 80-101. These ages seem reasonable in our opinion, so we did not remove them.
- Nothing was removed from location dataset because it did not seem viable as removing any outlier would shrink our joined dataset drastically suggesting those points were not outliers.

1.6 Joining the Cases and Location dataset

Here the cases and location dataset are **inner** joined on [‘province’, ‘country’]. The location dataset was cleaned before joining. The Korea South, US and Taiwan* were replaced with South Korea, United States, and Taiwan respectively. Since the column names for location dataset were different so it would have been redundant to keep country and Country_Region in the joined dataset, hence the column names were changed for Country_Region and Province_State to country and province respectively. The Lat and Long_ columns were removed because similar data is already present in the cases datasets. After the cleaning, the location dataset is joined with cases_train and cases_test.
The number of rows in cases_train dataset: 16037
The number of rows in cases_test dataset: 8946

1.7 Feature Selection

The feature selection of quantitative data for the joined datasets is done using Feature Importance, ExtraTreeClassifier method. We select the features with respect to the importance score of each feature. **Larger the score, the more important the feature**. The scores that we got were [0.3369 0.1457 0.2281 0.0099 0.0621 0.0369 0.0341 0.0364 0.0382 0.0713] for the respective features [age, latitude, longitude, chronic_disease_binary, Confirmed, Deaths, Recovered, Active, incident_rate, case_fatality_ratio]. By looking at the scores we selected the top 6 features which were [age, latitude, longitude, case_fatality_ratio] and discarded the rest, i.e., [chronic_disease_binary, Confirmed, Deaths, Recovered, Active, incident rate]. For the categorical data we kept [sex, province, country, date_confirmation] and removed the [additional_information, source]. Same features were selected for test and train datasets.
So, selected = [age, sex, province, country, latitude, longitude, date_confirmation, outcome_group, Confirmed, Death, case_fatality_ratio]