CMPT459 Fall 2020
Data Mining
Martin Ester
TAs: Madana Krishnan Vadakandara Krishnan and Rhea Rodriguez

Milestone 1 of the Course Project

Deadline: October 15

Total marks: 100

Introduction

In our course project, you will work with two publicly available COVID-19 datasets.

- Case dataset
  The first file contains the data for individual cases, i.e. cases that tested positive for COVID-19.
- Location dataset
  The second file contains the number of cases based on location.

The two files were obtained from the following open-source repositories, which provide more information:

- https://github.com/beoutbreakprepared/nCoV2019
- https://github.com/CSSEGISandData/COVID-19

The datasets have been filtered by the TAs for use in our project. To ensure consistency, the datasets have been frozen on September 20th, 2020. You can download the datasets from

https://github.com/MadanKrishnan97/CMPT459CourseProject/tree/master/dataset

The data mining task will be to predict the outcome of a case. In this first milestone of the course project, you will do the data preprocessing as specified in the following.

Preprocessing Tasks

1.1 Exploratory Data Analysis **(20 marks)**
Perform exploratory data analysis to get an understanding of the datasets. Show visualizations and statistics for all attributes of both datasets. In particular, for every attribute print the number of missing values.

1.2 Data cleaning and Imputing missing values **(20 marks)**
Perform data cleaning steps, mainly on the age column. Reduce different formats (ex. 20-29, 25-, 13 months), to a standard format.
For all attributes with missing values, discuss why and how (if applicable) you impute missing values. Apply your imputation strategy to your datasets.

1.3 Dealing with outliers **(20 marks)**

Which attributes have outliers? How do you deal with them? Apply your strategy of dealing with outliers to your datasets.

1.4 Transformation **(10 marks)**

In the location dataset, aggregate the information for cases from the US from the country level, used in the location dataset, to the state level, used in the cases dataset. Explain your method of transformation.

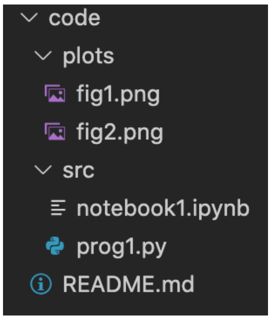
1.5 Joining the cases and location dataset **(30 marks)**

The two datasets can be joined using some shared features. You can use either 'province, country' or 'latitude, longitude'. Present your strategy for joining the datasets and motivate your design decisions. Apply your join strategy to create a dataset of cases with additional features inherited from their locations.

Submission (Code + Report)

2.1 Code

Submit a ‘code.zip’ file with the following contents. It should contain the code that you have written, figures obtained for Exploratory Data Analysis, and any important instructions for code execution. The structure can look like this.



```
code
  plots
    fig1.png
    fig2.png
  src
    notebook1.ipynb
    prog1.py
  README.md
```

2.2 Report

Briefly explain the approaches and steps followed in the five preprocessing tasks. Report should **NOT** be more than 2 pages. Submit a ‘report.pdf’ file.