**Project report**

* **Cs 677**
* **Junyu Mou**
* **April 13th, 2022**

Research scenario

This project is about research on covid-19 data set. Exploring the Covid-19 cases distributions and death rate. Dividing the countries into developed nations and developing nations to explore their performance.

Describe the data

The dataset has 10 features with 225 values each. The features are country names, countries’ other name, ISO code for county, population, continent name, total cases, total deaths, total cases per 1 million population, total death per 1 million population, death percentage.

Data preparation

In this part, I import data and modify the data for usage.

1. Import the data from .csv file

Application

Description automatically generated with medium confidence

1. Remove missing values

Text

Description automatically generated with medium confidence

1. Modify dataset



1. The standardizations were done with machine learning part

Exploratory data analysis

In this part I made data visualizations and plots to explore research questions.

Questions:

Text

Description automatically generated

There are three interactive data visualizations to represent the distributions for total cases, total death, and death rate. (Please see Jupiter Notebook)

A picture containing diagram

Description automatically generated

A picture containing graphical user interface

Description automatically generated

In the two graphs above, the United States, Indian and Brazil has most cases, and they have most deaths as well, you can see the details in cases and deaths in Notebook.

A picture containing diagram

Description automatically generated

Even though the United States, Indian and Brazil has most cases and deaths, the highest death rate is happened around middle east and north Africa.

Table

Description automatically generated

I get average total\_cases, average total\_deaths and average death\_percentage by NumPy by continents. The average is based on population.

Text

Description automatically generated

The data for continents are not balanced, for example there are only 5 countries in Northern America, and 58 countries in Africa. That’s why I got average based on population. If the average is based on country, the result will be misleading, continents with more countries the average cases and deaths will be low.

Chart, bar chart

Description automatically generated

This is histograms for number of countries in each continent.

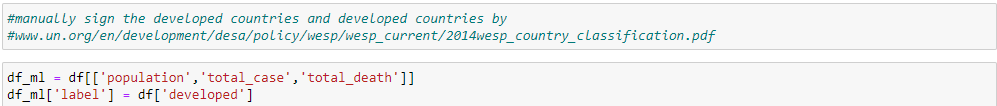
Chart, bar chart

Description automatically generated

This plot shows each continent total cases per 1 million people compared with total deaths per 1 million people.

Machine learning

For this part I will use 4 different classification algorithms to classify the dataset. I divided dataset into developing countries and developed countries, and use population, total cases, and total deaths as features.



Chart

Description automatically generated

The dataset is not balanced, so I oversampled the dataset:

Text

Description automatically generated

After oversampling, the dataset is balanced:

Chart, bar chart

Description automatically generated

Naïve Bayesian:

Chart

Description automatically generated

Logistic regression:

Graphical user interface, text, application

Description automatically generated

KNN:

Find best k:

Chart

Description automatically generated

KNN with best k:

Application

Description automatically generated with medium confidence

Decision Tree:

Chart

Description automatically generated with medium confidence

Chart, waterfall chart

Description automatically generated

As the result, Logistic regression has the best accuracy (0.83).

Prediction by best classifier

Text

Description automatically generated