Introduction to Artificial Intelligence
Project
COVID-19 Infection Analysis and Prediction

Deadline: 2023, June 1, 11:00 PM

Cases of a novel coronavirus were first reported in Wuhan, Hubei province, China, in December 2019 and have since spread across the world. Epidemiological studies have indicated human-to-human transmission in China and elsewhere. Epidemiological data is needed during emerging epidemics to best monitor and anticipate spread of infection.

The dataset has been made available publicly as of 20th January, 2020 containing different information about the patients: clinic, demographic and geographic.

There is also a Github repository available https://github.com/beoutbreakprepared/nCoV2019
The dataset is also available on kaggle platform (https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge), choose the file COVID19_line_list_data.csv

The data will be also available on moodle and Teams.

Note: To load a *.csv* file in Python, you can use the csv_read() function from Pandas library.

The epidemiological situation regarding the COVID-19 outbreak is continuously evolving. Each of the rows represents a single individual case and ID. A description of the fields in the database is shown in this paper: *Epidemiological data from the COVID-19 outbreak, real-time case information*

The goal of this project is to process this dataset using artificial intelligence methods in order to help the community to better understand the spread of the COVID-19 infection.

The project contains the following 4 parts:

1. ***Analysis of the dataset***
2. ***Bayes Nets***
3. ***Machine Learning***
4. ***Improving the results and Theoretical formalism***

1. Analysis of the dataset:

In order to analyse the dataset, you have to extract some statistical information from the given dataset, for example: the type of data, the missing values, outliers, the correlation between variables, etc. If there are missing values, you can replace them by the mean, median or mode of the concerning variable.

A. Compute the correlations between the variables. Which variables are most correlated with the target (**outcome**) ? Explain the results.

B. Plot the dataset using scatter and analyse the obtained result. Use the PCA (Principal Component Analysis) to project the dataset.

2. Bayes Nets:
In this part we will use Bayes model to compute some predictions from the data sets. In this context we want to answer to the following questions:

A. What is the probability for a person to have symptoms of *COVID-19* ***(symptom_onset=date)*** if this person visited Wuhan (**visiting Wuhan = 1**) ? Consider that ***(symptom_onset=N/A)*** means that the patient is asymptomatic.

B. What is the probability for a person to be a true patient if this person has symptoms of *COVID-19* ***(symptom_onset=date)*** *and* this person visited Wuhan?

C. What is the probability for a person to death if this person visited Wuhan?

D. Estimate the average recovery interval for a patient if this person visited Wuhan?

3. Machine Learning:

In this part we use a machine learning method in order to predict the **outcome**: patients outcome as either '**died'** or '**discharged'** from hospital. You can use the K-Nearest Neighbours (K-NN) or Bayes Classification.

A. The obtained results should be validated using some external indexes as prediction error (Confusion matrix and Accuracy) or others as Recall, F-Meseaure,... The obtained results should be analysed in the report and provide a solution to ameliorate the results.

B. Use the Regression to predict the **age** of persons based on other variables. You have the choice on these explanatory variables? How you choose these variables? Compute the quality of the prediction using MSE error (Mean Squared Error).

C. Apply a clustering method (K-means) on the dataset to segment the persons in different clusters. Use the Silhouette index to find out the best number of clusters. Plot the results using scatter to visually analyse the clustering structure.

4. Improving the results and Theoretical formalism

A. The data is unbalanced. You can balance it by reducing randomly the majority class. Assume that you extract randomly samples that are balanced. How the prediction results will change?

B. How you can better mange the missing values?

C. To find the best parameters for the models, the **Grid-search** algorithm can be used which is available in scikit-learn library. Explain the algorithm and use it for the learning models to find the best parameters.

D. Give the algorithmically (mathematical) formalism of the method which give the best results. Explain all the parameters of the used method and their impact on the results. Some comparison with public results should me made to conclude the project.