THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Department of Computer Science and Engineering
COMP4331:Introduction to Data Mining
Fall 2020 Group Project
Due time and date: 11:59pm, Dec 21 (Mon), 2020.

IMPORTANT NOTES

- **Your grade will be based on the correctness and clarity.**
- **Late submission: 25 marks will be deducted for every 24 hours after the deadline.**
- **ZERO-Tolerance on Plagiarism: All involved parties will get zero mark.**

Dataset

You are given a dataset related to COVID-19: covid_train.csv [link] with 3864 records. Each record contains information about a certain country over a three-day period. There are a total of 90 features. The following is a brief description. Details can be found in data_description.pdf.

- *Countries Data*: General information about the countries. These features are static and do not change over the time periods.
- *Policies Data*: Policies implemented by the government. The values are binary or ordinal, and are averages taken over the three-day period.
- *Related Indexes*: Includes Stringency Index, Government Response Index, Containment Health Index, and Economic Support Index.
- *Survey Data*: Self-reporting data, which is weighted to adjust for survey biases, and the values are averages taken over the three-day period.
  - *Symptoms Data*: From pct_fever_weighted to pct_chills_weighted. These are the estimated percentages of people reporting that they have experienced the given symptoms for the past 24 hours.
  - *Activities Data*: From pct_ever_tested_weighted to pct_no_public_weighted. These describe the general activities of the people, such as traveling and wearing masks.
- *Mobility Data*: Location and transport data relative to a chosen baseline. The values are averages taken over the three-day period.
- prev_cases: The previous number of cases before the three-day period. Even though dates are not provided in this dataset, this serves as an indicator of the progress on the spreading of the disease for the countries.
- total_cases: The total number of cases recorded at the end of the three-day period. Please note that this attribute will **NOT** be included in the test set (covid_test.csv).
- new_cases_percentages: The number of newly confirmed cases over the last three days (i.e., total_cases - prev_cases) in percentage of the country's population. This is the target variable that your model has to predict, and is divided into four classes based on the severity of the spread of the disease.

Tasks

In the following, we describe the tasks that you **HAVE TO** perform. Besides this basic set of tasks, you can provide additional analysis to help understand the data.
To ensure that your results can be replicated, please set the random seed to 333 by inserting the following lines before your code:

```
>>> import numpy as np
>>> np.random.seed(333)
```

1. (Code and Report) This task uses ONLY the *Countries Data*. Drop all the duplicated records.

   (a) Data preprocessing: To deal with the missing data in this subset, train an imputer using scikit-learn's IterativeImputer to fit and transform the data. Next, standardize the data (*Countries Data*) such that the mean is 0 and standard deviation is 1. Remove all the records that are outliers.

   (b) Hierarchical clustering: Using the Euclidean distance to measure distance between samples, show the top five levels of the dendrograms obtained with (i) single-link distance, (ii) complete-link distance, and (iii) group average distance. From the dendrograms, extract the 3-cluster solutions. Visualize each clustering solution by projecting the data samples to 2D using t-distributed Stochastic Neighbor Embedding (t-SNE). Please use different colors for different clusters.

   (c) Clustering validity measures: There are a number of cluster validity measures that can be used to assess the clustering solution's quality. Here, you will use the Davies-Bouldin score, implemented in the function davies_bouldin_score, and the Silhouette score, implemented in the function silhouette_score. By using these scores, compare and discuss the 3-cluster results obtained in task 1(b).

   (d) Visualization: For each of the 3 clusters obtained by using the group average distance in task 1(b), show

      i. the names of the countries, and

      ii. summary statistics (including mean, standard deviation, and five-number summary) and boxplot for each attribute (use the values before standardization). Include two of these boxplots (i.e., corresponding to two attributes) in your report that best illustrate differences between the groups.

      iii. Create a scatter plot using the two attributes you chose in task 1(d)ii. Use different colors for the different clusters that each country belongs to.

2. (Code and Report) This task uses ONLY the *Policies Data* and new_cases_percentage. Based on new_cases_percentage, we first define the following classes

   - class 0: new_cases_percentage $\leq 0.000441\%$ of population;
   - class 1: new_cases_percentage $> 0.000441\%$ but $\leq 0.00221\%$ of population;
   - class 2: new_cases_percentage $> 0.00221\%$ but $\leq 0.00788\%$ of population;
   - class 3: new_cases_percentage $> 0.00788\%$ of population.

   (a) We are interested in policy combinations that can reduce the number of new cases to $\leq 0.00221\%$ of the population in a three-day period. For each cluster in the clustering solution of Task 1(d), extract all policy combinations that are used in at least 30% of the records and lead to either class 0 or class 1 in at least 60% of the data.

   (b) Let S be the set of policies occurring in the combinations obtained from part (a). Draw a histogram for the number of policy combinations (obtained in part (a)) each policy in S occurs in.

3. (Code and Report) This task uses only the *Symptoms Data* and total_cases. For each attribute, if the record's original value is less than the attribute's median value, set it to 0 (i.e., corresponding symptom does not exist); otherwise, set it to 1 (i.e., corresponding symptom exists);

   (a) Extract all symptom combinations that appear in at least 20% of all the records.

   (b) Define the total number of cases in percentage of population as:

   $$\frac{\text{total\_cases}}{\text{pop\_total}} \times 100.$$

   This is considered high if it is larger than or equal to the corresponding median. Among all the symptom combinations obtained in part (a) above, find those that appear in at least 60% of all records and have high total number of cases in percentage of population.

   (c) Let S' be the set of symptoms occurring the combinations obtained from part (b) above. Draw a histogram for the number of symptom combinations each symptom in S' occurs in.

4. (Code and Report) In this task, use only the *Mobility Data*, pop_density and new_cases_percentages.

   (a) Create a scatter plot for each attribute (y-axis) against pop_density (x-axis, in log scale). Use different colors for the different classes in new_cases_percentages.

   (b) Divide the *Mobility Data* into five equal-sized parts based on the percentiles of pop_density. For each part, report the top 5 attributes that are most correlated with new_cases_percentages. Be sure to take both positive and negative correlated attributes into account.

5. (Code and Report) Perform additional analysis and discuss your findings. Present a set of suggestions (a minimum of 3) to stop the spread of the pandemic and support them with your findings from any previous tasks. Please direct your suggestions to specific countries or country groups, and be sure to include at least one suggestion targeting Hong Kong specifically.

6. (Code and Report) Predict new_cases_percentages based on all the data available provided in covid_train.csv. The prediction model should be mainly based on at least one of the models covered in the lectures (i.e., decision trees, neural networks, clustering and association analysis methods). On top of this backbone model, you can use additional data mining models/techniques to enhance its performance. Your data mining model will be evaluated by its error on the test set (which is hidden from you). Please be reminded that you are **NOT** allowed to use the attribute total_cases when training the model, as this attribute will not be present in the test set.

   Please write all your code in a Python notebook named group-project.ipynb, and use the library PyTorch to create the model. Save your trained data mining model as model.pt. At the end of your code, create a section for testing. It should read the csv file covid_test.csv, go through your data preprocessing procedure (be sure to take into account of missing data), and output the predictions of your model.

   In your report, provide details for your whole procedure including data preprocessing, model creation, training, predictions, etc. Keep in mind that this classification task is very difficult. With four classes, the accuracy for random guess is 25% on the test set, while a good model is able to achieve 80% on the test set.

Grading

Your grade will be evaluated by the following criteria:

1. (Report) Completeness and correctness for the tasks.

2. (Report) The plots are clear and communicate effectively.
3. (Code) The code reflects the report's description and executes as intended.
4. (Code and Report) The depths of your analysis (including the additional analysis you performed). Your work should demonstrate your understanding in the data and the suggestions provided are convincing and supported with data.
5. (Model) The performance of your data mining model on our test set.

Submission Guidelines

Please submit a report (report.pdf), a Python notebook (group-project.ipynb) for your code, and the model (model.pt). Zip all the files into either [group number]_group-project.zip or [group number]_group-project.tar.gz. Please submit the project by uploading the compressed file to Canvas. Note that the report and code should be clearly legible, otherwise you may lose some points if the report and code are difficult to read. Plagiarism will lead to zero point on this assignment.