**University of Ottawa**

**School of Electrical Engineering and Computer Science CSI4142 Fundamentals of Data Science**

**Project Phase 4: Data Mining**

**Group 12**

# Part A. Data summarization, data preprocessing and feature selections

1. Data Summarization at https://github.com/angemichaella/CSI4142\_project/blob/main/data\_mining/data\_summarization.ipynb

Deliverable Part A: Submit one page of notes to explain how you preprocessed the data. Your notes should detail any data transformation and data quality issues that you encountered.

For this part of the assignment, a number of changes were made to the dataset:

## Creating the initial dataframe

After determining which was going to be the target class (‘age\_group’), a dataframe that would contain all the data that was needed for the model training was created. Data from the fact table, the weather table, the phu table and the person table was combined into the new dataframe.

## Target Categorization

The target class’ values were all strings and had been transformed into integers as some of the learners, gradient boosting for example, require target class’ values to be numerical. Each one of the age ranges was replaced with a numeric value. For example, ‘80s’ was transformed into ‘80’’.

## Handling missing values

A first data transformation operation was performed, which consisted of removing all the missing values from the dataframe obtained in the previous step. Roughly 5% of the data was imputed.

## Handling categorical attributes

Going forward, categorical attributes were transformed into numeric values. The transformation was done using oneHotEncoder. For example, the municipality columns had values such as Ottawa, York, etc which were changed into new columns such as municipality\_ottawa, municipality\_york, etc.

## Normalization of numerical attributes

To ensure all attributes are of equal importance during the learning process, numerical values were normalized. Examples of some of the values that were normalized are mean\_temperature, which indicates the average temperature for a given day, and the number of resolved cases. MinMaxScaler was used to normalize numerical attributes.

## Feature selection

To remove any redundant attributes, a feature selection method was applied. For this assignment, RFECV was used, which is a cross-validation version of RFE. Using a minimum number of features to select equal to 5, roughly 6 out of 33 features were removed.

## Undersampling the majority class(es)

To balance class distribution, Near-Miss was used to remove some of the samples from the majority class. Before the transformation, some of the classes had up to 5000 samples, but since the minority class had only 1109 samples, all the classes were transformed in a way that each one of them would have only 1109 samples.

# Part B. Classification (Supervised Learning)

## Algorithm Results Comparison

|  |  |  |  |
| --- | --- | --- | --- |
|  | Decision Tree | Gradient Boosting | Random Forest |
| Accuracy | 17.5% | 14.3% | 14.7% |
| Precision | 20.0% | 13.8% | 11% |
| Recall | 17.6% | 14.4% | 11.2% |
| Construction Time | 0.5 sec | 0min 15.4sec | 18 sec |

## Classification Summary

1. (15 marks) Submit a 200 to 300 words summary explaining the actionable knowledge nuggets your team discovered. That is, you should explain what insights you obtained about the Covid-19 data, when investigating the models produced by the three algorithms.

Our models were not very accurate (approx. 17% for the decision tree) so that shows that the attributes we had were not very strong predictors of the age group (our chosen class label) of patients. We also used a Decision Tree Classifier neural network to select the optimal number of features, including the municipality, gender, case\_acquisition, temperature, precipitation, mobility trends, resolved, fatal, and outbreak attributes. Plotting the trees showed how attributes are being used to predict the age and provided some insights; for example the first attribute the tree split on was mobility trends of workplaces by checking whether it was <= 0.53. This split had a gini score of 0.89, showing high uncertainty between class classification. Classification for the other mobility trend columns were similar, showing that they alone are not strong predictors of age. However, this may be in part due to mobility trends and weather trends being linked to the fact table by reported date. This was something that should be changed in the data staging. As for the rest of the tree, all the gini scores for early nodes were high, so simple/useful insights could not be obtained from the tree. However, interesting information was found during the data summarization (such as a correlation between temperature and number of cases, a higher chance of fatality in the Toronto PHU, etc...).

# Part C. Detecting outliers

1. In the process of detecting outliers in our data, the first step (and the hardest one) was to find the columns where it will make sense to find outliers in them, so for that I choose to use the different columns from the mobility trends table and two columns (mean\_temperature and precipitation) from the weather table. On each of the chosen column I applied the one-class SVM to fit my data, after that I used the score\_sample method on my classifier to generate the score of each value (which determines if the value is abnormal or not), and plot the results on an only x-axis graph to be able to determine the presence of outliers on my column. Overall after checking the results not many outliers were found in the chosen column, in fact only three values that we can qualify as abnormal were found in the grocery\_and\_pharmacy column of the mobility table. The results showed more the presence of multiple clusters in the data used.