Course: CIND 820 - Big Data Analysis Project (F2021)  
Chang School of Continuing Studies, Ryerson University

Predictive Analysis on  
COVID-19 cases in Toronto

Submitted by-

Md Anwarul Choudhury

[anwarul.choudhury@ryerson.ca](mailto:anwarul.choudhury@ryerson.ca)

Student ID: 501138610

Table of Contents

[1. Abstract 3](#_Toc89434982)

[2. Dataset Overview 3](#_Toc89434983)

[3. Methodology 3](#_Toc89434984)

[4. Data Analysis, Review and Preparation 4](#_Toc89434985)

[4.1. Basic Findings 4](#_Toc89434986)

[4.2. Data Findings and Preparation (By Attributes) 5](#_Toc89434987)

[4.2.1. \_id: 5](#_Toc89434988)

[4.2.2. Assigned\_ID: 5](#_Toc89434989)

[4.2.3. Outbreak Associated: 5](#_Toc89434990)

[4.2.4. Age Group: 6](#_Toc89434991)

[4.2.5. Neighbourhood Name: 6](#_Toc89434992)

[4.2.6. FSA: 6](#_Toc89434993)

[4.2.7. Source of Infection: 6](#_Toc89434994)

[4.2.8. Classification: 7](#_Toc89434995)

[4.2.9. Episode Date: 7](#_Toc89434996)

[4.2.10. Reported Date: 8](#_Toc89434997)

[4.2.11. Gender: 8](#_Toc89434998)

[4.2.12. Outcome: 8](#_Toc89434999)

[4.2.13. Currently Hospitalized: 9](#_Toc89435000)

[4.2.14. Currently in ICU: 9](#_Toc89435001)

[4.2.15. Currently Intubated: 9](#_Toc89435002)

[4.2.16. Ever Hospitalized: 9](#_Toc89435003)

[4.2.17. Ever in ICU: 9](#_Toc89435004)

[4.2.18. Ever Intubated: 10](#_Toc89435005)

[5. Explorative Data Analysis 10](#_Toc89435006)

[5.1.1. Outbreak Associated: 10](#_Toc89435007)

[5.1.2. Age Group: 10](#_Toc89435008)

[5.1.3. Source of Infection: 11](#_Toc89435009)

[5.1.4. Reported Date: 12](#_Toc89435010)

[5.1.5. Gender: 12](#_Toc89435011)

[5.1.6. Outcome: 13](#_Toc89435012)

[5.1.7. Ever Hospitalized: 13](#_Toc89435013)

[5.1.8. Ever in ICU: 13](#_Toc89435014)

[5.1.9. Ever Intubated: 14](#_Toc89435015)

[6. Correlation Matrix 14](#_Toc89435016)

[7. Predictive Modelling 15](#_Toc89435017)

[7.1. Classification using Decision Tree: 15](#_Toc89435018)

[7.2. Classification using Naïve Bayes 16](#_Toc89435019)

[7.3. Classification using Random Forest 17](#_Toc89435020)

[7.4. Evaluation and Comparison 18](#_Toc89435021)

[8. Interpretations and Conclusion: 19](#_Toc89435022)

[9. Limitations/ Further Analysis 19](#_Toc89435023)

# Abstract

This study conducted some explorative analysis on Toronto COVID-19 data and to find if there is any significant relationship between the data attributes. This might help to understand the COVID-19 case patters in Toronto to help better fight the pandemic in Toronto area.

The study also developed classification models based on the available data to predict the outcome (patient survival or deceased) of a given patient(s).

# Dataset Overview

The dataset contains about 177000 reported COVID cases in Toronto and 18 attributes (City of Toronto’s Open Data Portal, Dataset: “About COVID-19 Cases in Toronto”). The dataset is being updated on a weekly basis by TPH (Toronto Public Health). Below is the description of the data features (City of Toronto’s Open Data Portal, Dataset: “About COVID-19 Cases in Toronto”).

For this analysis 8 attributes were used and 10 attributes were excluded. Details of each data feature has been discussed in the Data Analysis, Review and Preparation section below.

# Methodology

To complete the project work and analysis below methodology (steps) were followed:

**Step 1: Data collection**

Dataset was available and downloaded as csv file from City of Toronto’s Open Data Portal. Dataset is being updated on a weekly basis. For this analysis, data was downloaded on 15-Sep-2021, any update on data after that date, was not reflected on this analysis.

**Step 2: Data Preparation and cleaning**

The raw data was not ready for actual analysis. Therefore, data preparation was performed mainly for data cleaning which included missing values, superfluous data, removing incorrect data, removing irrelevant attribute (from the context of this project) etc. Below were the outcomes of this step:

* Removed attributes: \_id, Assigned\_ID, Neighbourhood Name, FSA, Episode Date, Reported Date, Currently Hospitalized, Currently in ICU, Currently Intubated
* Removed all records with ‘Probable’ cases for the Classification attribute and then removed Classification attribute
* Removed all records with ‘Pending’ cases for the Source of Infection attribute
* Removed all records with missing values for Age Group attribute
* Modified (merge) the gender types as discussed in 4.2.11 section for Client Gender attribute
* Removed all records with ‘Active’ cases for Outcome attribute
* Numeric encoding performed for all rest 8 attributes
* Export the final data csv file as “Final\_COVID\_DataSet.csv”

**Step 3: Explorative data analysis (EDA)**

In this step, further analysis was conducted with the cleaned data to get some more data insight. This included general data statistics, frequency distribution, outlier analysis, correlation etc. The dataset had a class imbalance issue, therefore, Synthetic Minority Oversampling Technique (SMOTE) for categorical attributes was used to resolve it.

The details of the EDA outcome included in section 5.

|  |
| --- |
| Step 1: Data Collection  Step 2: Data Preparation and Cleaning  Step 3: Explorative Data Analysis  Step 4: Predictive Modelling  Step 5: Evaluation and Comparison  Step 6: Interpretation and Presentation |
| *Diagram: Project Methodology (steps)* |

**Step 4: Predictive modelling**

After EDA, the final dataset was used to develop classification models. For developing the classification models, Decision Tree, Naïve Bayes and Random Forest algorithms will be used. ‘Sklearn’ package for Python programming language was used to develop the models.

**Step 5: Evaluation and comparison**

This step conducted the evaluation and comparison of the predictive models built in the previous step. The outcome of the comparison discussed in ‘Predictive Modelling’ section of this document.

**Step 6: Interpreting the result and presentation**

Finally, in this step the analysis findings were discussed and presented.

# Data Analysis, Review and Preparation

## Basic Findings

* 16 qualitative data, 2 numeric data
* Some NULL or missing value were found for different attributes (discussed in the below table how missing value was handled)
* 176449 rows (records) in the initial dataset.

| Data Attribute | Data type |
| --- | --- |
| \_id | Numeric |
| Assigned\_ID | Numeric |
| Outbreak Associated | Nominal |
| Age Group | Ordinal |
| Neighbourhood Name | Nominal |
| FSA | Nominal |
| Source of Infection | Nominal |
| Classification | Nominal |
| Episode Date | Date |
| Reported Date | Date |
| Client Gender | Nominal |
| Outcome | Nominal |
| Currently Hospitalized | Nominal |
| Currently in ICU | Nominal |
| Currently Intubated | Nominal |
| Ever Hospitalized | Nominal |
| Ever in ICU | Nominal |
| Ever Intubated | Nominal |

## Data Findings and Preparation (By Attributes)

### \_id:

* Unique row identifier for the dataset
* Type: Numeric
* Since this attribute was only used to identify each record, it has been removed during analysis.

### Assigned\_ID:

* A unique ID assigned to cases by Toronto Public Health for the purposes of posting to Open Data, to allow for tracking of specific cases.
* Type: Numeric
* Since this attribute was only used to track each individual cases, it has been removed during analysis.

### Outbreak Associated:

* Outbreak associated cases are associated with outbreaks of COVID-19 in Toronto healthcare institutions and healthcare settings (e.g. long-term care homes, retirement homes, hospitals, etc.) and other Toronto congregate settings (such as homeless shelters).
* Type: Nominal
* There was no missing values for this data attribute

### Age Group:

* Age at time of illness. Age groups (in years): ≤19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+, unknown (blank).
* Type: Ordinal
* There were 109 missing values for this attribute. As the number of missing values are very nominal, the records with missing values have been removed from the analysis.

### Neighbourhood Name:

* Toronto is divided into 140 geographically distinct neighborhoods that were established to help government and community agencies with local planning by providing socio-economic data for a meaningful geographic area.
* Type: Nominal
* There were 2849 missing values for this attribute.
* This attribute was only used for some general statistical analysis. It was not used for classification analysis as it was not relevant. Therefore, the attribute was removed during classification analysis.

### FSA:

* Forward sortation area (i.e. first three characters of postal code) based on the case’s primary home address. FSA values are generated from client postal codes. One FSA can span multiple neighborhoods.
* Type: Nominal
* The focus of the analysis was based on the neighborhood. Therefore, this feature was removed during analysis.

### Source of Infection:

* The source of infection is represented in the following hierarchy:
  + - Travel > Outbreak (settings described below) > Household Contact > Close Contact > Community > No information
    - Cases with episode dates on or after April 1 2020:
      * Outbreak (settings described below) > Household Contact > Close Contact > Travel > Community > No information.
* Type: Nominal

|  |  |
| --- | --- |
| Value | Description |
| Close Contact | Case who acquired infection from a close contact with a confirmed or probable COVID-19 case (e.g. co-worker). |
| Community | Cases who did not travel outside of Ontario, did not identify being a close contact with a COVID-19 case, and were not part of a known confirmed COVID-19 outbreak. |
| Household Contact | Case who acquired infection from a household contact with a confirmed or probable COVID-19 case (e.g. family member, roommate) |
| No Information | Cases with no information on the source of infection |
| Outbreaks, Congregate Settings | confirmed outbreaks in Toronto in shelters, correctional facilities, group homes, or other congregate settings such as hostels or rooming houses. |
| Outbreaks, Healthcare Institutions | confirmed outbreaks in Toronto in long-term care homes, retirement homes, hospitals, chronic care hospitals, or other institutional settings. |
| Outbreaks, Other Settings | confirmed outbreaks in Toronto in workplaces, schools, day cares, or outbreaks outside of Toronto. We do not validate outbreaks that occur in other health units, as such these cases may not be linked to confirmed outbreaks. |
| Pending | Cases yet to be determined on the source of infection |
| Travel | Case that travelled outside of Ontario in the 14 days prior to their symptom onset or test date, whichever is the earliest. |

* There were 45,195 records (25% of the total data) in ‘No Information’ category. Therefore, the category was not removed, or data was not modified.
* There were only 79 records in ‘Pending’ category. As this is very nominal and the information was not determined, these records will be removed from the analysis.

### Classification:

* The application of the provincial case definition to categorize the cases as ‘Confirmed’ or ‘Probable’, according to standard criteria. Please refer to the Ontario Ministry of Health website for Ontario’s current provincial case definitions.
* Type: Ordinal
* There were no missing values for this attribute.
* There were 4941 records in ‘Probable’ category. As the analysis considers only ‘Confirmed’ cases, all ‘Probable’ cases were removed first and then the entire ‘Classification’ attribute was removed from analysis.

### Episode Date:

* The episode date is a derived variable that best estimates when the disease was acquired, and refers to the earliest available date from: symptom onset (the first day that COVID-19 symptoms occurred), laboratory specimen collection date, or reported date.
* Type: Date
* For the purpose of this analysis, the ‘Episode Date’ attribute was excluded from this analysis.

### Reported Date:

* The date on which the case was reported to Toronto Public Health.
* Type: Date
* For the purpose of this analysis, the ‘Reported Date’ attribute was excluded from this analysis.

### Gender:

* Self-reported gender. Gender is a system that operates in a social context and generally classifies people based on their assigned biological sex
* Type: Nominal
* Values and # of records:

|  |  |
| --- | --- |
| Value | # of records |
| FEMALE | 87070 |
| MALE | 86532 |
| NON-BINARY | 24 |
| NOT LISTED, PLEASE SPECIFY | 4 |
| OTHER | 24 |
| TRANS MAN | 3 |
| TRANS WOMAN | 2 |
| TRANSGENDER | 22 |
| UNKNOWN | 2868 |

* There are very small number of records in categories: ‘NON-BINARY’, ‘NOT LISTED, PLEASE SPECIFY’, ‘OTHER’, ‘TRANS MAN’, ‘TRANS WOMAN’ and ‘TRANSGENDER’
* Therefore, for the purpose of this analysis, all the above marked categories and ‘UNKNOWN’ category were modified as ‘OTHER’ category.

### Outcome:

* Describes the outcome for each case
  + - FATAL: Cases with a fatal outcome reported (patient deceased)
    - RESOLVED: Cases not reported as deceased and who are either reported as ‘recovered’ or where the report date is more than 14 days from symptom onset and the case is not currently hospitalized.
    - ACTIVE: All other cases
* Type: Nominal
* Values and #of records:

|  |  |
| --- | --- |
| Value | # of records |
| RESOLVED | 171295 |
| FATAL | 3638 |
| ACTIVE | 1616 |

* All cases marked as ‘Active’ were removed from analysis

### Currently Hospitalized:

* Cases that are currently admitted to hospital (i.e., no discharge date reported).
* Describe the current status of a case where the ‘Outcome’ attribute value is defined as ‘Active’. As all ‘Active’ cases were removed from this analysis, therefore, this data feature was removed from this analysis.

### Currently in ICU:

* Cases that are currently admitted to the intensive care unit (ICU) (i.e. no discharge date reported).
* Describe the current status of a case where the ‘Outcome’ attribute value is defined as ‘Active’. As all ‘Active’ cases were removed from this analysis, therefore, this data feature was removed from this analysis.

### Currently Intubated:

* Cases that were intubated related to their COVID-19 infection (includes cases that are currently intubated and those that have been discharged or deceased).
* Describe the current status of a case where the ‘Outcome’ attribute value is defined as ‘Active’. As all ‘Active’ cases were removed from this analysis, therefore, this data feature was removed from this analysis.

### Ever Hospitalized:

* Cases that were hospitalized related to their COVID-19 infection (includes cases that are currently hospitalized and those that have been discharged or are deceased).
* Type: Nominal (Yes/ No)
* Values and #of records:

|  |  |
| --- | --- |
| Value | # of records |
| YES | 11394 |
| NO | 165155 |

* No missing value for this attribute

### Ever in ICU:

* Cases that were admitted to the intensive care unit (ICU) related to their COVID-19 infection (includes cases that are currently in ICU and those that have been discharged or are deceased).
* Type: Nominal (Yes/ No)
* Values and #of records:

|  |  |
| --- | --- |
| Value | # of records |
| YES | 2272 |
| NO | 174277 |

* No missing value for this attribute

### Ever Intubated:

* Cases that were intubated related to their COVID-19 infection (includes cases that are currently intubated and those that have been discharged or deceased).
* Type: Nominal (Yes/ No)
* Values and #of records:

|  |  |
| --- | --- |
| Value | # of records |
| YES | 1358 |
| NO | 175191 |

* No missing value for this attribute

# Explorative Data Analysis

### Outbreak Associated:

* For the analysis and modelling purpose, the data was converted to numeric encoding as below:

|  |  |  |
| --- | --- | --- |
| Value | Numeric Encoding | # of records |
| Outbreak Associated | 1 | 21738 |
| Sporadic | 2 | 148248 |
|  | Total | 169986 |

* The correlation matrix indicated that there is strong correlation between ‘Outbreak Associated’ and ‘Source of Infection’ attributes.

### Age Group:

* For the analysis and modelling purpose, the data was converted to numeric encoding as below:

|  |  |  |
| --- | --- | --- |
| Value | Numeric Encoding | # of records |
| 19 and younger | 1 | 24161 |
| 20 to 29 Years | 2 | 35364 |
| 30 to 39 Years | 3 | 29699 |
| 40 to 49 Years | 4 | 24556 |
| 50 to 59 Years | 5 | 24341 |
| 60 to 69 Years | 6 | 15350 |
| 70 to 79 Years | 7 | 7592 |
| 80 to 89 Years | 8 | 5768 |
| 90 and older | 9 | 3155 |
|  | Total | 169986 |

* The correlation matrix indicated that there is significant correlation between ‘Age Group’ and ‘Outcome’ attributes.
* The correlation matrix indicated that there is significant correlation between ‘Age Group’ and ‘Ever Hospitalized’ attributes.
* The correlation matrix indicated that there is correlation between ‘Age Group’ and ‘Gender’ attributes.

### Source of Infection:

* For the analysis and modelling purpose, the data was converted to numeric encoding as below:

|  |  |  |
| --- | --- | --- |
| Value | Numeric Encoding | # of records |
| Close Contact | 1 | 11957 |
| Community | 2 | 58654 |
| Household Contact | 3 | 30661 |
| No Information | 4 | 43946 |
| Outbreaks, Congregate Settings | 5 | 2572 |
| Outbreaks, Healthcare Institutions | 6 | 11630 |
| Outbreaks, Other Settings | 7 | 8081 |
| Travel | 8 | 2485 |
|  | Total | 169986 |

* The correlation matrix indicated that there is significant correlation between ‘Source of Infection ’ and ‘Outcome’ attributes.
* The correlation matrix indicated that there is correlation between ‘Source of Infection’ and ‘Ever Hospitalized’ attributes.
* The correlation matrix indicated that there is correlation between ‘Source of Infection ’ and ‘Outcome’ attributes.

### Gender:

* For the analysis and modelling purpose, the data was converted to numeric encoding as below:

|  |  |  |
| --- | --- | --- |
| Value | Numeric Encoding | # of records |
| FEMALE | 1 | 83995 |
| MALE | 2 | 83566 |
| OTHER | 3 | 2425 |
|  | Total | 169986 |

* The correlation matrix indicated that there is correlation between ‘Gender’ and ‘Age Group’ attributes

### Outcome:

* For the analysis and modelling purpose, the data was converted to numeric encoding as below:

|  |  |  |
| --- | --- | --- |
| Value | Numeric Encoding | # of records |
| RESOLVED | 1 | 166382 |
| FATAL | 2 | 3604 |
|  | Total | 169986 |

* The correlation matrix indicated that there is significant correlation between ‘Outcome’ and ‘Age Group’ attributes
* The correlation matrix indicated that there is correlation between ‘Outcome’ and ‘Source of Infection’ attributes
* The correlation matrix indicated that there is strong correlation between ‘Outcome’ and ‘Ever Hospitalized’ attributes
* Class imbalance

### Ever Hospitalized:

* For the analysis and modelling purpose, the data was converted to numeric encoding as below:

|  |  |  |
| --- | --- | --- |
| Value | Numeric Encoding | # of records |
| NO | 0 | 158734 |
| YES | 1 | 11252 |
|  | Total | 169986 |

* The correlation matrix indicated that there is strong correlation between ‘Ever Hospitalized’ and ‘Age Group’ attributes
* The correlation matrix indicated that there is significant correlation between ‘Ever Hospitalized’ and ‘Source of Infection’ attributes
* The correlation matrix indicated that there is strong correlation between ‘Ever Hospitalized’ and ‘Outcome’ attributes

### Ever in ICU:

* For the analysis and modelling purpose, the data was converted to numeric encoding as below:

|  |  |  |
| --- | --- | --- |
| Value | Numeric Encoding | # of records |
| NO | 0 | 167758 |
| YES | 1 | 2228 |
|  | Total | 169986 |

* The correlation matrix indicated that there is strong correlation between ‘Ever in ICU’ and ‘Ever Intubated’ attributes

### Ever Intubated:

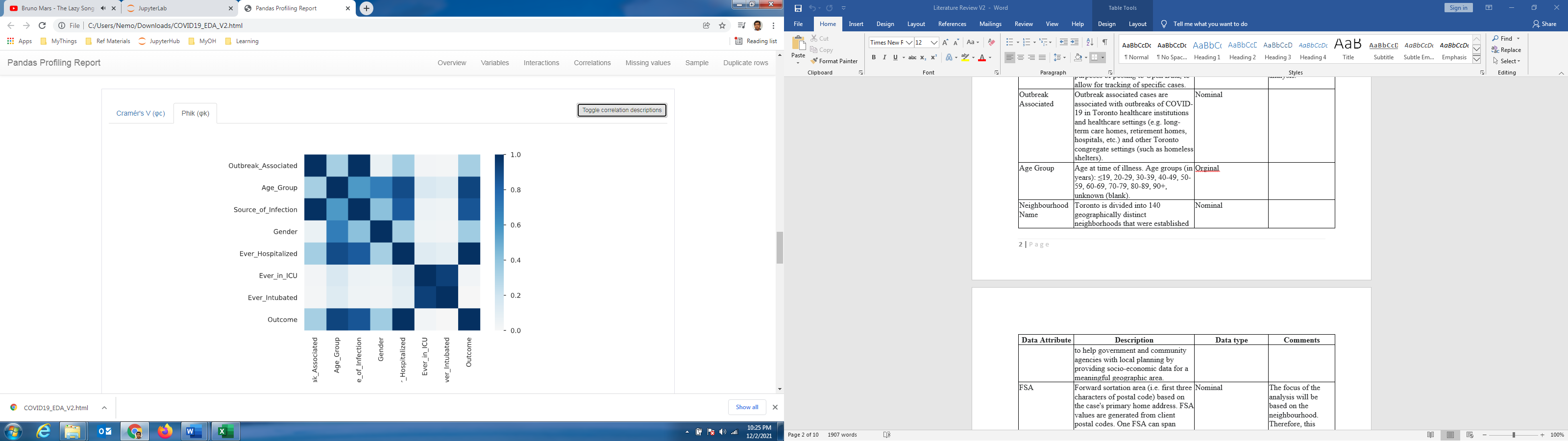
* For the analysis and modelling purpose, the data was converted to numeric encoding as below:

|  |  |  |
| --- | --- | --- |
| Value | Numeric Encoding | # of records |
| NO | 0 | 168649 |
| YES | 1 | 1337 |
|  | Total | 169986 |

* The correlation matrix indicated that there is strong correlation between ‘Ever Intubated’ and ‘Ever in ICU’ attributes

# Correlation Matrix

The below correlation matrix was developed using Phik Correlation Efficient (for categorical variables)



# Predictive Modelling

The following parameters were used to build a classification model using Decision Tree on the prepared data:

* Data were split into train and test set
* “Outcome” attributed was considered as the class name
* 70% data was used for training and 30% was used for test
* Max depth for the decision tree was set to 8.
* Python Sklearn package was used to develop the Decision Tree and Naïve Bayes

## Classification using Decision Tree:

Findings based on Confusion Matrix:

* The model correctly predicted 49708 out of 49980 ‘Resolved’ cases (99.4%)
* The model correctly predicted 49046 out of 49850 ‘Fatal’ cases (98.4%)
* The model incorrectly predicted 272 out of 49980 ‘Resolved’ cases as ‘Fatal’ (0.54%)
* The model incorrectly predicted 804 out of 49850 ‘Fatal’ cases as ‘Resolved’ (1.61%)

|  |
| --- |
| Confusion Matrix |
|  |
| |  |  |  |  | | --- | --- | --- | --- | |  |  | Predicted | | |  |  | Resolved | Fatal | | Actual | Resolved | 49708 | 272 | | Fatal | 804 | 49046 | |
| Assessment Matrix |
|  |

## Classification using Naïve Bayes

Another predictive modelling was performed using Naïve Bayes model. Below is the output for the model.

* The model correctly predicted 45957 out of 49980 ‘Resolved’ cases (92%)
* The model correctly predicted 49377 out of 49850 ‘Fatal’ cases (99.1%)
* The model incorrectly predicted 4023 out of 49980 ‘Resolved’ cases as ‘Fatal’ (8%)
* The model incorrectly predicted 473 out of 49850 ‘Fatal’ cases as ‘Resolved’ (0.9%)

|  |
| --- |
| Confusion Matrix |
|  |
| |  |  |  |  | | --- | --- | --- | --- | |  |  | Predicted | | |  |  | Resolved | Fatal | | Actual | Resolved | 45957 | 4023 | | Fatal | 473 | 49377 | |
| Assessment Matrix |
|  |

## Classification using Random Forest

Findings based on Confusion Matrix:

* The model correctly predicted 23198 out of 49890 ‘Resolved’ cases (46.5%)
* The model correctly predicted 26708 out of 49940 ‘Fatal’ cases (53.5%)
* The model incorrectly predicted 26692 out of 49890 ‘Resolved’ cases as ‘Fatal’ (53.5%)
* The model incorrectly predicted 23232 out of 49940 ‘Fatal’ cases as ‘Resolved’ (46.5%)

|  |
| --- |
| Confusion Matrix |
|  |
| |  |  |  |  | | --- | --- | --- | --- | |  |  | Predicted | | |  |  | Resolved | Fatal | | Actual | Resolved | 23198 | 26692 | | Fatal | 23232 | 26708 | |
| Assessment Matrix |
|  |

## Evaluation and Comparison

The below comparison table provides assessment information at a glance for all 3 models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assessment** | **Class** | **Decision Tree** | **Naïve Bayes** | **Random Forest** |
| Precision | Class 1 (Resolved) | 0.98 | 0.99 | 0.50 |
| Class 2 (Fatal) | 0.99 | 0.92 | 0.50 |
| Recall/ Sensitivity | Class 1 (Resolved) | 0.99 | 0.92 | 0.46 |
| Class 2 (Fatal) | 0.98 | 0.99 | 0.53 |
| f1-Scrore | Class 1 (Resolved) | 0.99 | 0.95 | 0.48 |
| Class 2 (Fatal) | 0.99 | 0.96 | 0.52 |
| Support | Class 1 (Resolved) | 49980 | 49980 | 49890 |
| Class 2 (Fatal) | 49850 | 49850 | 49940 |
| Accuracy | - | 0.99 | 0.95 | 0.50 |
|  |  |  |  |  |

# Interpretations and Conclusion:

* Decision Tree model is most accurate predicting cases based on Accuracy, Precision and f1-score comparing all 3 models.
* Decision Tree performed better than Naïve Bayes based on Sensitivity for ‘Resolved’ cases (0.99 in comparison to 0.92). On the other hand, Naïve Bayes performed better than Decision Tree based on Sensitivity for ‘Fatal’ cases (0.99 in comparison to 0.98). For this project, it is important to correctly predict ‘Fatal’ cases than ‘Resolved’ cases (1% indicates about 500 patients for this scenario).
* Random Forest is least accurate amongst all 3 models.
* For the above interpretation it is suggested to use Naïve Bayes for the Toronto COVID-19 dataset.

# Limitations/ Further Analysis

Further analysis could be done on this dataset, not performed due to time constraints. Some of them are listed below:

* Predictive Model to analyze ICU need
* Time series data analysis based on reporting date
* K-fold Cross Validation
* Detailed correlation analysis between attributes