**PROG 8450 – Big Data Storage and Integration**

**Machine learning algorithm in Spark**

**Autor – Islamiyat Osuolale**

**ID – 8773642**

**About COVID-19 Cases in Toronto**

Toronto Public Health is responding to an ongoing COVID-19 outbreak, in the context of an evolving global pandemic. This data set contains demographic, geographic, and severity information for all confirmed and probable cases reported to and managed by Toronto Public Health since the first case was reported in January 2020. This includes cases that are sporadic (occurring in the community) and outbreak-associated. The data are extracted from the provincial Case & Contact Management System (CCM).[1]

### **Limitations**

The data in this spreadsheet are subject to change as public health investigations into reported cases and continuous quality improvement initiatives are ongoing, and additional cases continue to be reported. \*The data will be completely refreshed and overwritten on a weekly basis, extracted at 8:30 AM on the Tuesday of a given week, and posted on the Wednesday. Please note that these numbers may differ from those posted elsewhere, as data are extracted at different times, and from different sources.[1]

DATA: <https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/64b54586-6180-4485-83eb-81e8fae3b8fe/resource/fff4ee65-3527-43be-9a8a-cb9401377dbc/download/COVID19%20cases.csv>

**Purpose of Report**

The objective of our report is to predict fatal situations of people that were hospitalised base on covid19 cases in Toronto. We will be making use of Machine Learning Analysis in SparkML, analysing, and predicting some variables outcome in the “COVID19\_cases.csv” dataset. (Link is provided in the previous paragraph)

**Data Exploration**

Graphical user interface

Description automatically generated

Fig. 1 Explore Data

Data Exploration helps to know the kind of data we are about to analyze and have an idea of what is going on.

We can see the data features below.

| **Column** | **Description** |
| --- | --- |
| \_id | Unique row identifier for Open Data database |
| 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. Please note, cases may be reclassified or removed from posted datasets over time. This can occur due to duplicate resolution, or other ongoing data cleaning initiatives. In such instances, the Assigned ID of that case will no longer appear on the list and will not get assigned to another case. |
| 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). |
| 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) |
| 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. Please note that client postal code information is mapped to the most up-to-date census tract (CT) and neighbourhood information available from the city. As a result, neighbourhood information is not available for those with missing postal code or when postal code could not be mapped/linked to a CT. |
| 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 neighbourhoods. |
| Source of Infection | The most likely way that cases acquired their COVID-19 infection is determined by examining several data fields including:   * A public health investigator's assessment of the most likely source of infection. * Being associated with a confirmed COVID-19 outbreak * Reported risk factors such as contact with a known case or travel   If the public health investigator's assessment is absent, then the other data fields are used to infer source of acquisition using the following hierarchy:  Cases with episode dates before April 1 2020:   * 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.   Descriptions:   * **Household contact**: Case who acquired infection from a household contact with a confirmed or probable COVID-19 case (e.g. family member, roommate). * **Close contact with a case**: Case who acquired infection from a close contact with a confirmed or probable COVID-19 case (e.g. co-worker). * **Outbreaks**: Cases linked to known confirmed COVID-19 outbreaks. These could include the index case who may have acquired the infection elsewhere. Outbreaks settings include: * *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. * **Travel**: Case that travelled outside of Ontario in the 14 days prior to their symptom onset or test date, whichever is the earliest. * **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. * **No information**: Cases with no information on the source of infection |
| 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.](http://www.health.gov.on.ca/en/pro/programs/publichealth/coronavirus/docs/2019_case_definition.pdf) |
| 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. |
| Reported Date | The date on which the case was reported to Toronto Public Health. |
| Client Gender | Self-reported gender. Gender is a system that operates in a social context and generally classifies people based on their assigned biological sex. |
| Outcome | **Fatal:** Any case that has died and has been marked as Outcome equals 'Fatal' and Type of Death does not equal 'Disease of Public Health Significance was unrelated to cause of death' in the provincial reporting system (CCM).  **Resolved:** Cases who have:   * A case outcome description in CCM of 'Recovered' OR * Case outcome description is equal to 'Fatal' AND Type of Death is equal to 'Disease of Public Health Significance was unrelated to cause of death' OR * Today's date is more than 14 days from episode date AND the case is not currently hospitalized/intubated/in ICU AND Case outcome description is not equal to 'Fatal' where Type of Death is not equal to 'Disease of Public Health Significance was unrelated to cause of death'.   **Active:** All other cases |
| Currently Hospitalized | Cases that are currently admitted to hospital (i.e., no discharge date reported). |
| Currently in ICU | Cases that are currently admitted to the intensive care unit (ICU) (i.e. no discharge date reported). |
| 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). |
| 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). |
| 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). |
| 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) |

[1]

**Correlation Analysis**

Checking the relationships between the variables to determine the suitable ones for our analysis, we made use Tableau for visualization and R programming for computing the Corrgram graph just to be sure.

Chart, bubble chart

Description automatically generated

Fig. 2. Outcome of Covid-19 Cases in Toronto

According to the image above we can see that the youths have high numbers of covid cases from unknown sources in Toronto, which means they were probably careless and that is so reasonable. It makes sense because it is reality, we youths are a little bit nonchalant about the virus.

Moving on, we can also view that older people have the least cases but more fatal cases. Giving us an overview of how covid has been fatal and dangerous for old people, they probably have more count of covid case death in Toronto. It also shows that most people above 60 with fatal cases got infected from outbreaks of the virus.

Let us have a look at the corrgram graph, this will explain the correlations between each variable in the dataset.

Chart, scatter chart

Description automatically generated

Fig. 3. Corrgram Graph of Covid Cases in Toronto

In Fig. 3., we can see that outcome is highly correlated with most of the variables. This will make it easy for us to chose outcome as our dependent variable in our algorithm. We would like to know how much cases were resolved and the ones that are still in active including the ones that are fatal.

**Machine Learning Algorithm in Spark**

We built two models called training and test, making use of seed number 1234, this was right after reading in and modifying our dataset by choosing the column we would like to work on. These variables are Outcome, Age.Group, Ever.Hospitalized, Ever.Intubated, and Currently.in.ICU. Trying to know how many fatal cases we have in Age.Group and the remaining variables.

During the algorithm process, we can see that not all the choosing variables can be compared with outcome because we want to get the most accurate result for our prediction. So, we compared Outcome with Ever.Hospitalized, meaning what is the predicted accuracy of the outcome in people that were hospitalized due to covid-19.

*NOTE:* *We will take you through spark algorithm in the APPENDIX.*

Chart, pie chart

Description automatically generated

**Fig. 4. Prediction of Fatal Outcomes of People That Were Hospitalized**

Using python to compute the prediction result from our spark algorithm, we were able to predict that there were a lot of fatal situations for people that were hospitalized due to covid-19. It predicted about 75% were predicted to have fatal situations and the remaining were either resolved or active.

In conclusion, there are extremely fatal cases amongst hospitalized people in Toronto. We also have cases of people that are still active and the ones that are already resolved but not as much as the bad situation cases.

**APPENDIX**

Here the data was read in, viewing the data with “YourDF.show(10)” and schema with “YourDF.schema(10)”

Text

Description automatically generated

**Fig. 5. Reading in and viewing the data schema**

The next stage is to set your training and test data set, we slitted the data to 80, 20 respectively. Setting our seed to “1234” so we have the same random data in each set and created a new random forest object.

Likewise, we used pipeline to set our stages which are the string indexer, the vector assembler and the random forest classifier object.

Text

Description automatically generated

**Fig. 6. Setting Training Data, Test Data, Assembler, Random Forest, and Pipeline**

Lastly, we created the metric of accuracy percentage, comparing Outcome to the prediction column.

For regression we used RegressionEvaluator, we joined all with cross-validator, set the pipeline for estimator and Multiclassevaluator for evaluator.

Training model most times gives us the best model and test is used to compare or dependent variable “Outcome” with prediction and accuracy.

We have maxdept of entropy = 4 in the second fold.

Text

Description automatically generated

**Fig. 7. Prediction and Accuracy**

**References**

1. Toronto Open Data: COVID-19 CASES IN TORONTO https://open.toronto.ca/dataset/covid-19-cases-in-toronto/