Public health Infobase on Covid-19
Report 2: Task
CSCI -527 Big Data

Venkata Krishna Manne - 202103591,
Kedarnath Pamarthi - 202004457,
Sarvesh Gadgil -202005945

1 Abstract

A pneumonia outbreak of unknown origin was reported in Wuhan, Hubei Province, by Chinese health authorities toward the end of December 2019. A few days later, the scientific community received access to a novel coronavirus genome, which had previously been kept secret. The International Committee on Taxonomy of Viruses' Coronavirus Study Group originally designated this novel coronavirus 2019-nCoV, which is now known as SARS-CoV-2. The Betacoronavirus genus, subgenus, and family that SARS-CoV-2 belongs to is the Coronaviridae. Since its discovery, the virus has spread throughout the world, killing thousands of people and severely affecting our economies and health care systems. In this project, we'll use a variety of machine learning models to predict the outcome of the virus, including whether it will increase or decrease in the coming months.[3]

2 Introduction

The COVID-19 pandemic has caused a shocking global loss of life and poses an unprecedented threat to food systems, public health, and the workplace. The pandemic has wreaked havoc on the economy and society. Tens of millions of people face starvation, and the number of undernourished people, which is currently estimated to be close to 690 million, could rise by as many as 132 million by the end of the year. Figure 1 shows an overview of COVID-19 in Canada, while Figure 2 shows the Data Dictionary, which includes the number of cases and fatalities at the provincial and federal levels from January 2020 to the present. It also mentions how many tests were performed and how many patients recovered.

3 Dataset

The dataset used in this project is [8]. This dataset is based on the Covid-19 Dataset, which keeps track of daily statistics on the number of cases in each Canadian province. The dataset includes the following information. ID, Province name, Reporting date, Total Cases, Total Deaths, and Total Active Cases. To eliminate all unwanted data, we performed a data-cleaning process. These data consist of 10333 rows and 40 columns in total.

4 Evaluation

To analyze the dataset, we must first clean the existing dataset by removing ”empty” and ”Null” values or filling with a default value for column types such as integer or decimal. The ability to evaluate and validate the models that we build is a critical step in supervised machine learning. Using unbiased data is one way to create an effective and valid model. We can increase our confidence that our model will work well with new data by reducing bias in our model. To accomplish this, we must divide the provided dataset into train and test segments in a 70:30 ratio using the Scitkit Learn train test split. We can fit a model using the training portion of our dataset and validate it using the test portion of our dataset.

5 Feature Correlation Heat Map

A type of plot called a correlation heat map shows the strength of relationships between numerical variables. To determine that variables are related to one another and how strongly they are related, correlation plots are used. Typically, a correlation plot includes a number of numeric variables, each of which is represented by a column. The relationships between each pair of variables are shown by the rows. Positive values indicate a strong relationship, while negative values indicate a weak relationship. The values in the cells represent the strength of the relationship. You can use correlation heat maps to identify possible relationships between variables and to gauge how strong these relationships are. Furthermore, outliers can be found and linear and nonlinear relationships can be found using correlation plots. The cells' color-coding makes it simple to quickly spot any relationships between the variables. Finding both linear and nonlinear relationships between variables is possible with the aid of correlation heat maps.[6]

- **Figure 3.** As we can see from figure 3 in order to predict the number of deaths the features with high correlation are ”numtotal”, ”numconf”, ”numrecover” and other features which has values above 0.5.

6 Models

To ensure that the training data is consistent across training data, we are using a standard dataset. We are using KNN Nearest Neighbors, a random forest classifier, and a decision tree to analyze the dataset.

6.1 K - Nearest Neighbor Algorithm

K-nearest neighbours (KNN) is a type of supervised learning algorithm used for both regression and classification. By calculating the separation between the test data and all of the training points, KNN attempts to predict the appropriate class for the test data. Then choose the K points that are most similar to the test data. The KNN algorithm determines which classes of the ”K” training data the test data will belong to, and only the class with the highest probability will be chosen. The value in a regression scenario is the average of the ’K’ chosen training points.[7]

- **Figure 2.** we are using a K-NN classifier to train the dataset by feeding it the mapped train dataset. Once the model has been trained, we feed it the test dataset and ask it to predict the outcome based on the trained model.
  We obtained an accuracy of **59.96** after successfully running the model on test datasets.

- **Figure 4.** figure 4 shows the predicted number of deaths for each reporting day. The predicted cases were generated using KNN model.

6.2 Random Forest Classifier

An approach to supervised learning is the random forests algorithm. It can be applied to both classification and regression. Additionally, it is the most adaptable and user-friendly algorithm. Trees make up a forest. A forest is said to be more robust the more trees it has. On randomly chosen data samples, random forests generate decision trees, get predictions from each tree, and vote on the best answer. Additionally, it offers a fairly accurate indication of the feature's importance. There are many uses for random forests, including feature selection, image classification, and recommendation engines. It can be used to spot fraud and forecast illnesses.[4]

- **Figure 2.** Figure 2 displays the accuracy of the random forest classifier model which is **64.16**.This model has the highest accuracy.

- **Figure 5.**figure 4 shows the predicted number of deaths for each reporting day. The predicted cases were generated using the Random Forest Classifier.

6.3 Decision Tree

The decision Tree algorithm is a part of the supervised learning algorithm. The decision tree algorithm can be used to resolve classification and regression issues as well, unlike other supervised learning algorithms. Using a Decision Tree, the objective is to develop a training model that can be used to predict the class or value of the target variable by learning straightforward decision rules inferred from historical data (training data). In decision trees, we begin at the tree's base to predict the class label for a record. We contrast the root attribute's values with the record's attribute. We proceed to the next node by following the branch that corresponds to that value based on the comparison.[5]

- **Figure 2.**Figure 2 displays the accuracy of the Decision tree model which is **34.44**.

7 Results

Even after applying various features to the test and train data, as we can see in figure 2, the ”Decision tree” model is not very accurate when compared to the ”Random forest classifier” and ”K Nearest Neighbor model.” Our research shows that the ”Random forest classifier” Model performs better for this dataset. Figures 4 and 5 show how the model produces the anticipated number of death cases for each reporting date.

8 Conclusion

We all know that the COVID-19 pandemic has drastically reduced the number of human lives on the planet and that it presents an unmatched challenge to public health, food systems, and the employment sector. Since the pandemic has caused such severe economic and social disruption, tens of thousands of people are now in danger of experiencing extreme poverty. So that we can prepare for any potential pandemic, we need more research on this kind of contagious disease.

References

[1] https://plotly.com/python/time-series/. plotly.

[2] https://stackoverflow.com/questions/27037241/changing-the-rotation-of-tick-labels-in-seaborn-heatmap. Stack overflow.

[3] Minieri M Ciotti M, Angeletti S. https://pubmed.ncbi.nlm.nih.gov/32259829. pubmed.gov, october 2019.

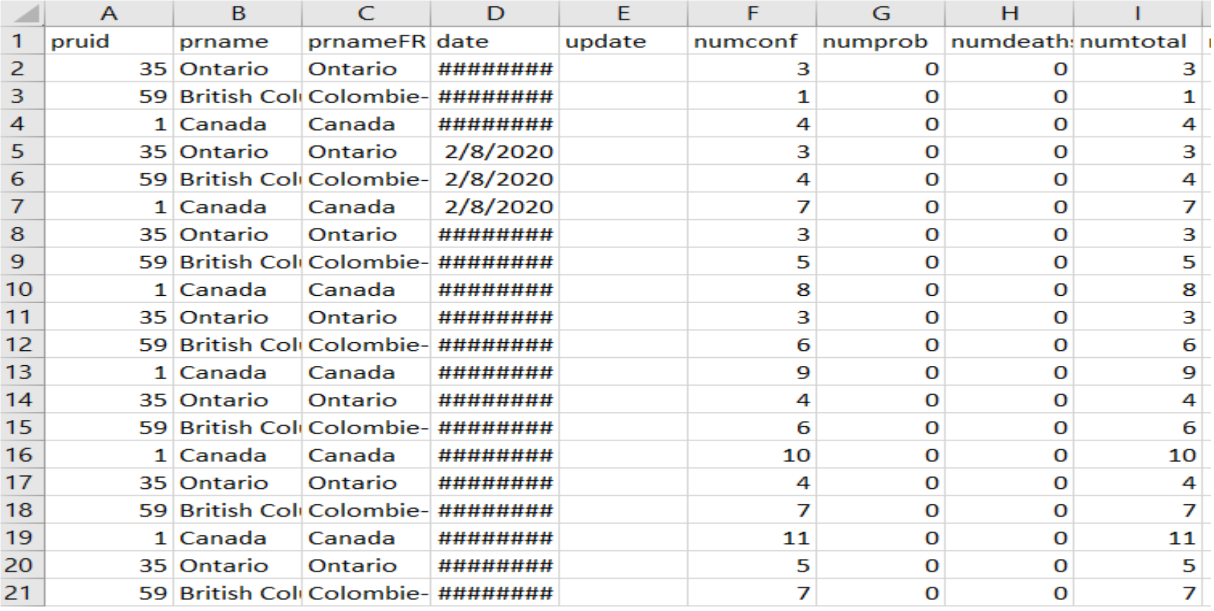
[4] Random forest Algorithm DataCamp. https://www.datacamp.com/tutorial/random-forests-classifier-python. DataCamp 2022, Jan 2022.

[5] Decision tree algorithm KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. KDnuggets 2022, Jan 2022.

[6] Ajitesh Kumar. https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/: :text=with wordpress, april 2022.

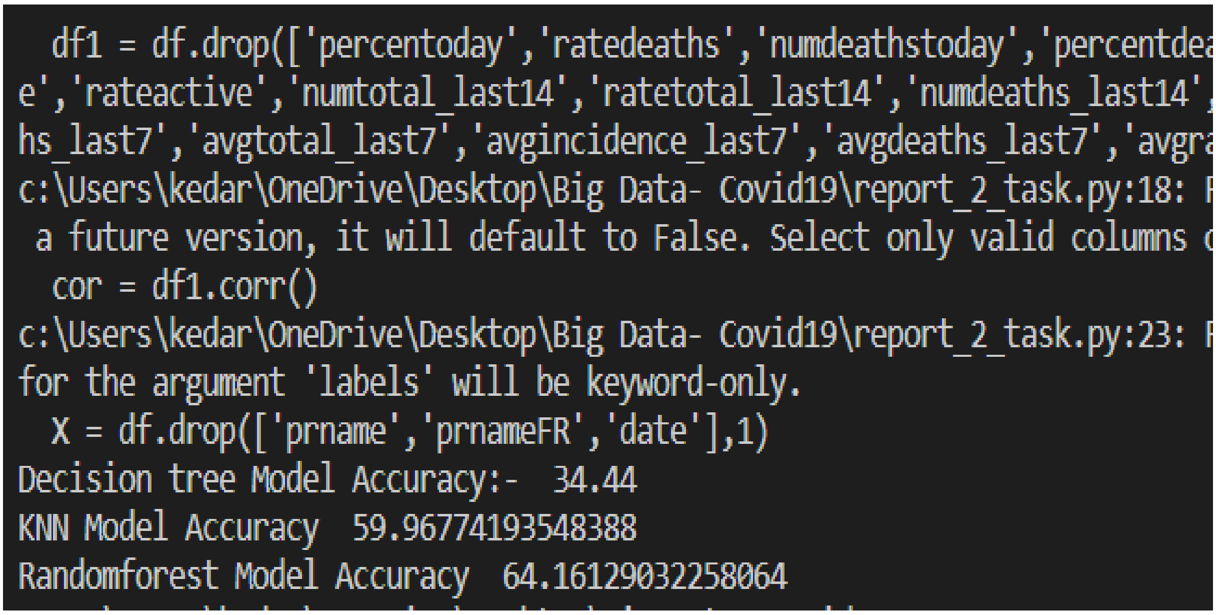
[7] KNN Algorithm Medium. https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4. DataCamp 2022, Jan 2022.

[8] Government of Canada Public Health Infobase. https://open.canada.ca/data/en/dataset/261c32ab-4cfd-4f81-9dea-7b64065690dc. Online, Government of Canada, Jan 2021.



	A	B	C	D	E	F	G	H	I
1	pruid	prname	prnameFR	date	update	numconf	numprob	numdeath:	numtotal
2	35	Ontario	Ontario	########		3	0	0	3
3	59	British Col	Colombie-	########		1	0	0	1
4	1	Canada	Canada	########		4	0	0	4
5	35	Ontario	Ontario	2/8/2020		3	0	0	3
6	59	British Col	Colombie-	2/8/2020		4	0	0	4
7	1	Canada	Canada	2/8/2020		7	0	0	7
8	35	Ontario	Ontario	########		3	0	0	3
9	59	British Col	Colombie-	########		5	0	0	5
10	1	Canada	Canada	########		8	0	0	8
11	35	Ontario	Ontario	########		3	0	0	3
12	59	British Col	Colombie-	########		6	0	0	6
13	1	Canada	Canada	########		9	0	0	9
14	35	Ontario	Ontario	########		4	0	0	4
15	59	British Col	Colombie-	########		6	0	0	6
16	1	Canada	Canada	########		10	0	0	10
17	35	Ontario	Ontario	########		4	0	0	4
18	59	British Col	Colombie-	########		7	0	0	7
19	1	Canada	Canada	########		11	0	0	11
20	35	Ontario	Ontario	########		5	0	0	5
21	59	British Col	Colombie-	########		7	0	0	7

Figure 1: Dataset of Covid-19



```
  df1 = df.drop(['percentoday','ratedeaths','numdeathstoday','percentdea
e','rateactive','numtotal_last14','ratetotal_last14','numdeaths_last14',
hs_last7','avgtotal_last7','avgincidence_last7','avgdeaths_last7','avgra
c:\Users\kedar\OneDrive\Desktop\Big Data- Covid19\report_2_task.py:18: F
 a future version, it will default to False. Select only valid columns c
  cor = df1.corr()
c:\Users\kedar\OneDrive\Desktop\Big Data- Covid19\report_2_task.py:23: F
for the argument 'labels' will be keyword-only.
  X = df.drop(['prname','prnameFR','date'],1)
Decision tree Model Accuracy:-  34.44
KNN Model Accuracy  59.96774193548388
Randomforest Model Accuracy  64.16129032258064
```

Figure 2: Accuracy scores of each model

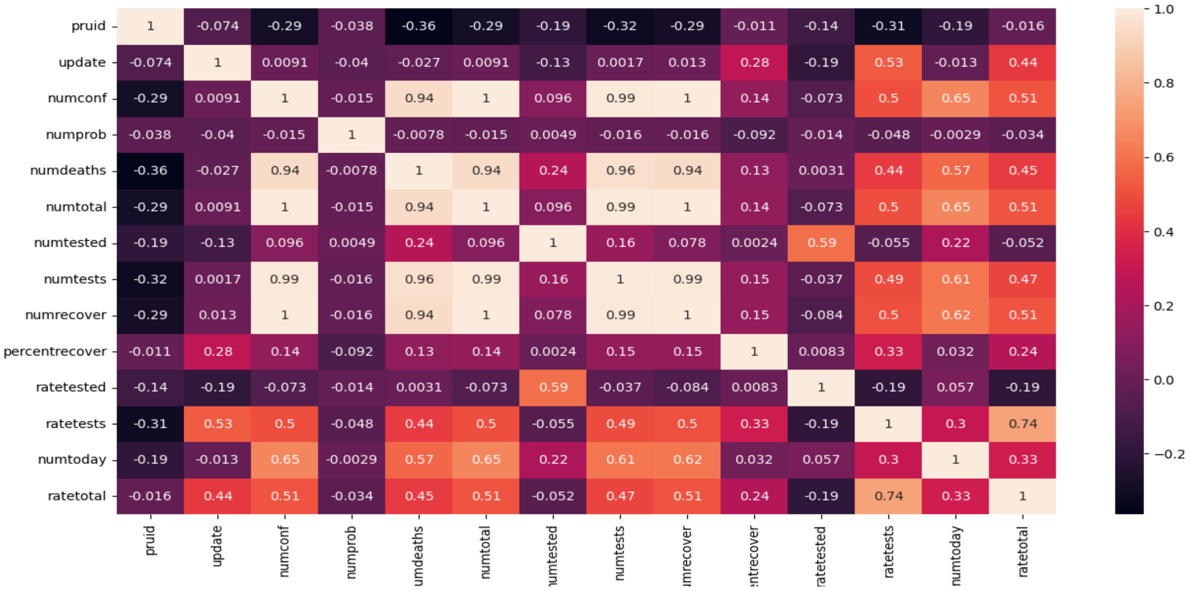


Figure 3: Feature Correlation using heat Map

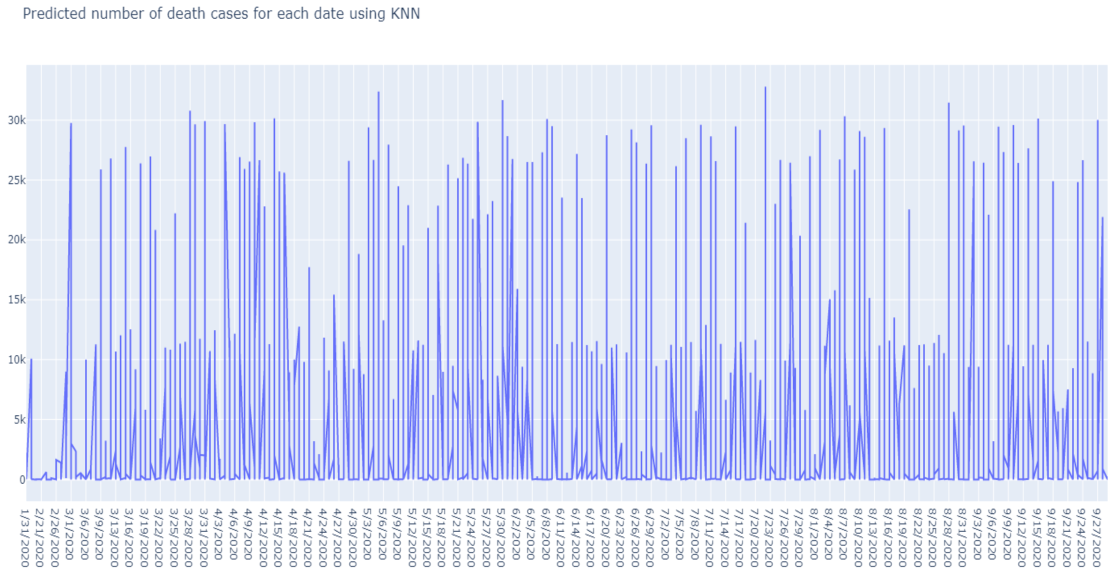


Figure 4: Predicted number of deaths for each day using KNN

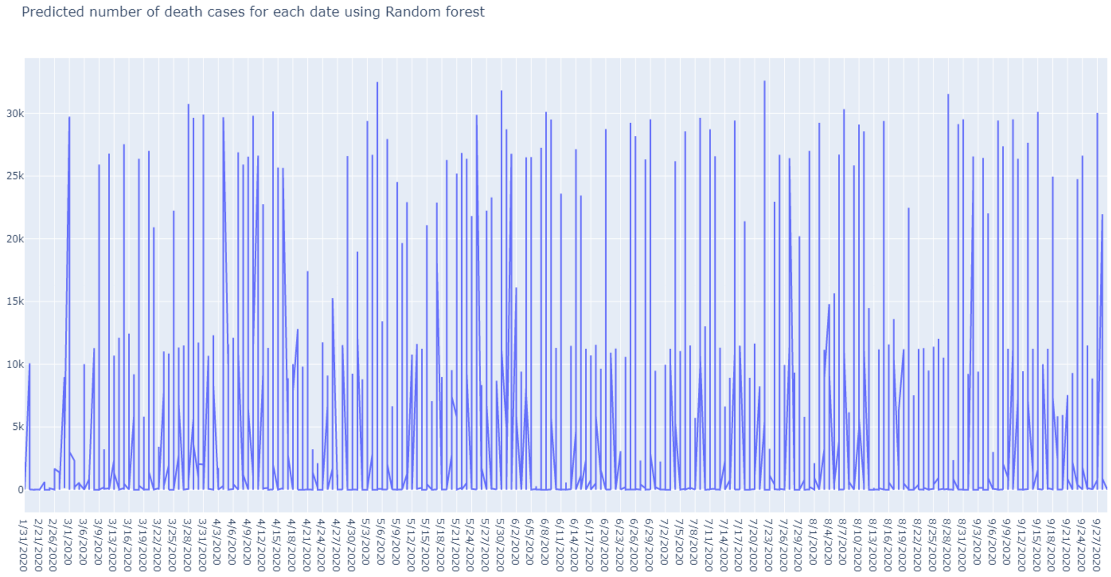


Figure 5: Predicted number of deaths for each day using RF