Introduction

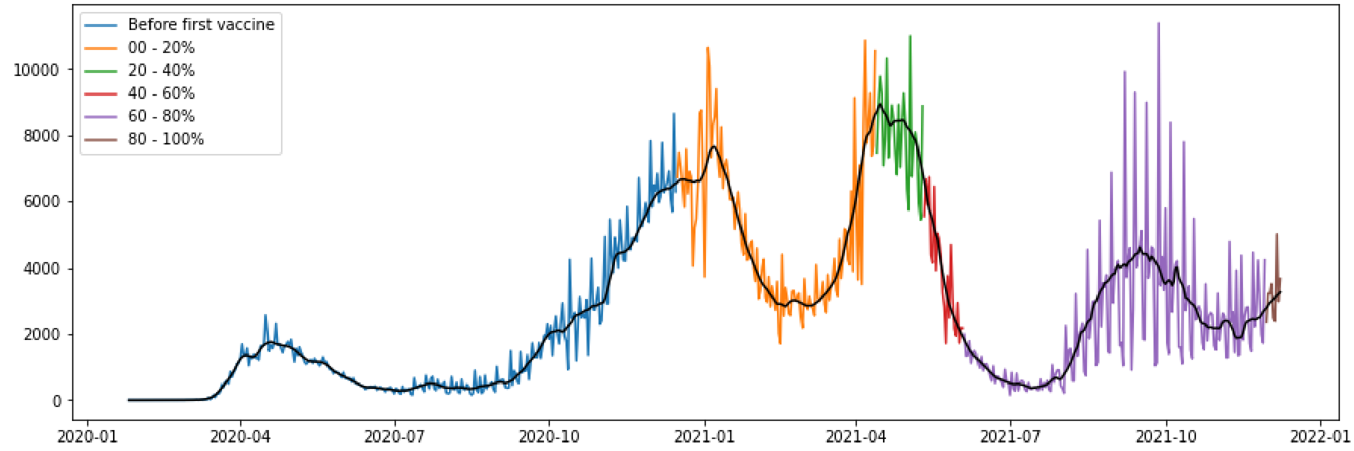
In this paper we will explore how Canadian, American, and International COVID-19 data can be analyzed to compare the impact of the virus in different countries as well as form predictions about future trends. The first question we are interested in answering is how vaccination has affected both the spread and the mortality rate of the COVID-19 virus. The second question we are interested in answering is how do the deaths resulting from COVID-19 differ between countries. The last question we will discuss is how accurately can future COVID-19 cases be predicted. To answer the first two questions we employed various statistical analysis tests and to answer the last question we utilized multiple machine learning models.

Data

The data we used in the analysis of COVID-19 data was pulled from the Our World in Data github repository which contains a large amount of historical COVID-19 data ranging from the start of the pandemic to the present day. This repository was downloaded and the file “owid-covid-data.csv” was copied to be cleaned and split for later use in statistical analysis and machine learning. This file contains daily information about the cases, vaccination, deaths, and more for each country. The data was processed by reading the file into a pandas dataframe and then selecting the columns with useful data and writing it to a cleaned CSV file containing international covid data to be used in further analysis. The columns of interest in the COVID-19 dataset were the country code, the total and daily case count, the total and daily death count, as well as vaccination information. Two additional subsets of the data were also made for the Canadian data and the American data. Apart from removing excess columns and making partitions, there was not much transformation that needed to be applied to the data as it was already in a well constructed format.

The following plot is meant to be a visual representation of what the Canadian COVID-19 data looks like. This figure is a plot of the daily Canadian case count which helps show the evolution of the pandemic. To highlight the main trend of the data, a LOESS line has been fit to the data.

Cases in Canada



Analysis Techniques

One technique that we used to compare datasets was the Mann-Whitney U-test. This is a nonparametric statistical test that indicates if samples from one group of data are larger or smaller than samples from a second set. The test assumes that the observations of both groups are independent and the values can be sorted.

The Mann-Whitney U-test was used to compare the difference in deaths resulting from COVID-19 cases between both Canada and the USA to see if one country had a significantly lower rate of deaths per hundred cases. This was repeated for the difference in deaths between Canada and International data. The Mann-Whitney U-test was also used to see if there was a difference in deaths per hundred cases between dates with a low vaccination rate and dates with a higher vaccination rate which may indicate that the vaccine lowers the chance of death from a case. The final use of the test was to compare if there was a difference in daily case counts between a period of low vaccination and high vaccination in Canada.

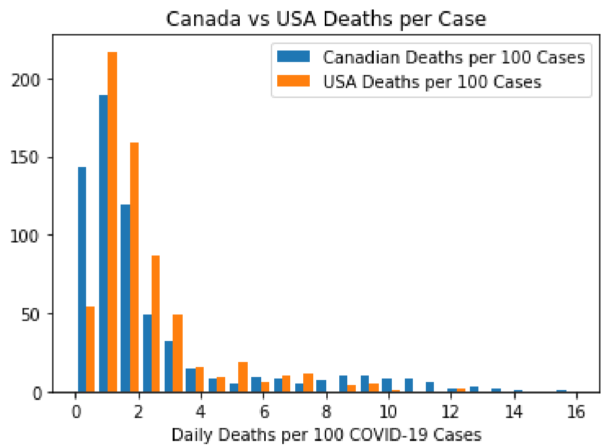
Another technique that we used was the ANOVA or Analysis of Variance test which is a statistical test that determines if any of the means in a set of groups differ. The test assumes that the samples are independent and identically distributed, the sets are normally distributed, and the sets have equal variance. The test was used to see if data with different vaccination rates had a noticeable difference in case count.

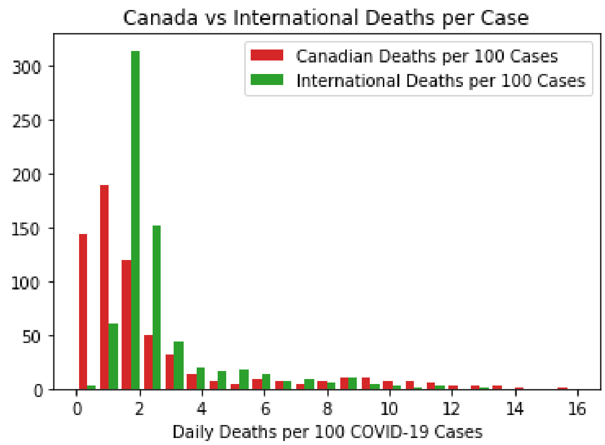
Machine Learning is the study of algorithms that make predictions on unseen data with the inputs we already have; it helps simulate real-world possibilities in a virtual environment. With the help of various models we built we estimated the number of new cases and deaths, which can prove useful for making forecasts about next waves or peaks.Though the models cannot predict what will happen, they can help us understand what might happen in certain scenarios, this can help us plan and achieve the best possible outcome.

Results

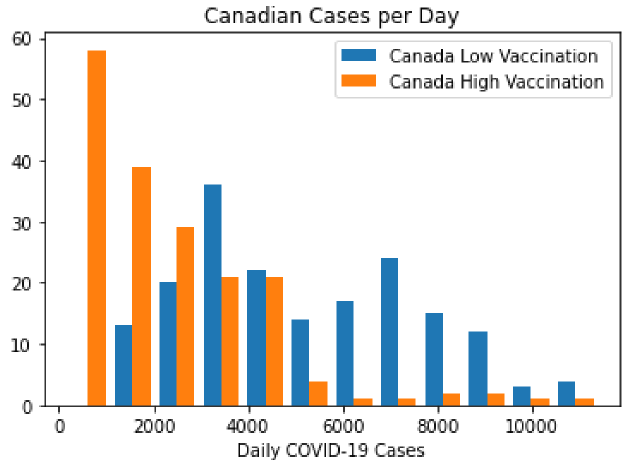
Mann-Whitney U-Test Results

The first comparison done using the Mann-Whitney U-Test was a comparison between Canada and the USA for the daily deaths per one hundred cases. The test returned a p-value of ~0.0001130 which allows us to safely assume that the samples in one of the sets of data are larger. As seen in the “Canada vs USA Deaths per Case” plot below, this test result is visualized as the USA deaths per 100 cases averages noticeably higher than the Canadian deaths per 100 cases. This same test was repeated to compare Canada to global data and provided similar results with a p-value of ~1.505e-24. In both cases it is observed that Canada has a definitively lower rate of death per case than both the USA and the global average.

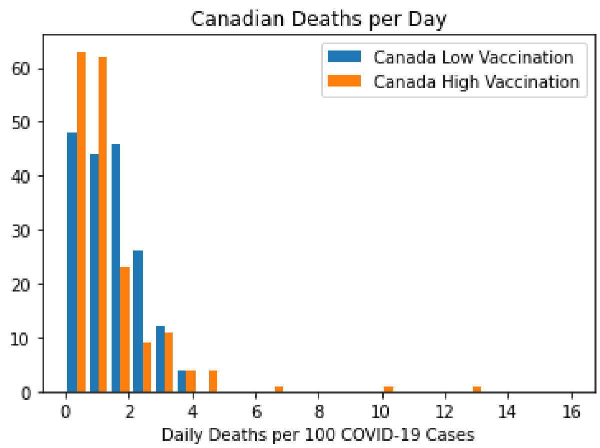


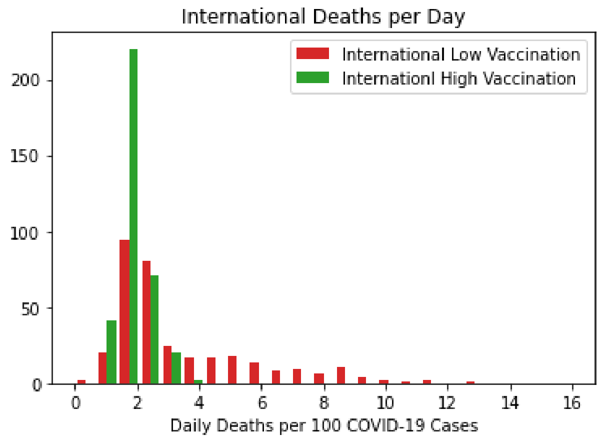


The set of Canadian COVID-19 data was split at the median vaccination percentage value and compared with a Mann-Whitney U-Test to determine if there was a difference in daily new cases between a period of low vaccination and a period of high vaccination. The test returned a p-value of ~5.153e-29 which allows us to safely assume that there is a difference between the samples. This difference becomes apparent when plotting as seen in the figure below. From this information it can be concluded that the period of high vaccination averages a much lower daily case count than the period of low vaccination.



Both the set of Canadian and international COVID-19 data were split at the median vaccination percentage value to compare the deaths per 100 cases during a period of lower vaccination and a period of higher vaccination. For this comparison between dates with low vaccination and high vaccination the Mann-Whitney U-Test returned a p-value of 0.0018 for the Canadian data and a p-value of 3.333e-33 for the international data. Both of these values indicate that there is a difference between the samples. These results can be seen in the histograms below as the deaths per 100 cases are noticeably higher when the vaccination rates are lower.





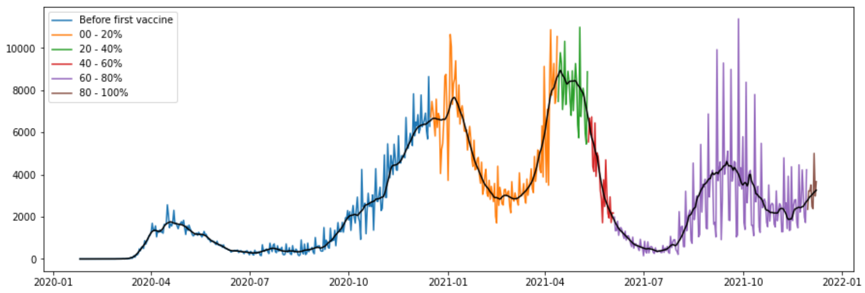
ANOVA Test Results

To compare the effectiveness of the vaccine in Canada and the United States, the data was put through an analysis of variance test. This test determines if any of the means of any of the groups differ. To accomplish this, the data was split into sections according to Canada's rollout of the vaccine. The data was split into sections consisting of:

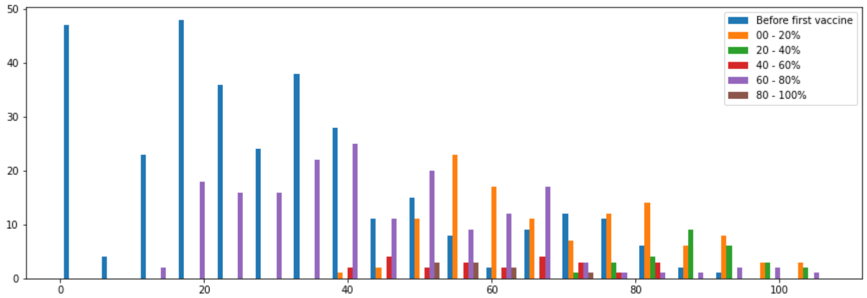
Event	Time-frame
No Canadians vaccinated	January 26, 2020 - December 15, 2020
0 - 20 % Canadians vaccinated	December 16, 2020 - April 12, 2021
20 - 40% Canadians vaccinated	April 13, 2021 - May 10, 2021
40 - 60% Canadians vaccinated	May 11, 2021 - June 3, 2021
60 - 80% Canadians vaccinated	June 4, 2021 - November 29, 2021
80 - 100% Canadians vaccinated	November 30, 2021 - December 8, 2021

By using the dates these events occurred, we can compare the same date ranges between Canada and the United States. If we were to compare the same percentages, we expect the graphs to look similar as the death rate decreases and the vaccine rate increases.

Cases in Canada

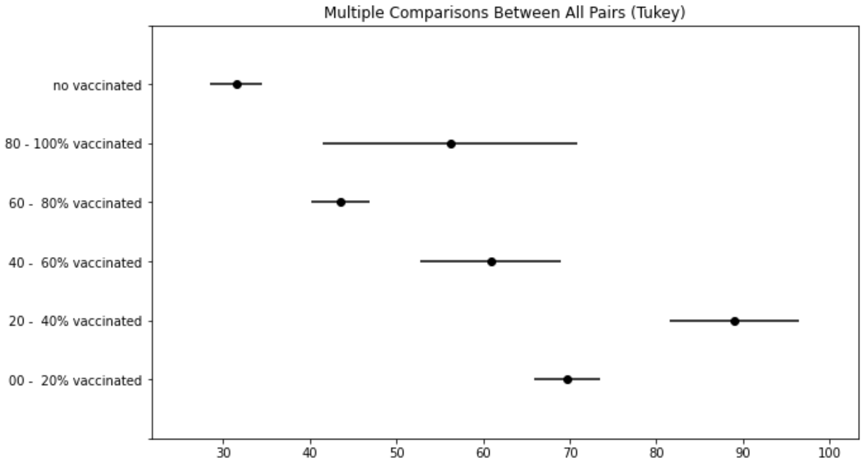


Starting with Canada, we can graph and visualize the data to get a grasp of what the data looks like. The data is quite noisy and is smoothed out to get a smooth curve. In order to run an ANOVA test, the data must be a normal distribution; or a nice bell curve. In order to accomplish that, the number of new cases was square rooted, which gave a more normal distribution.



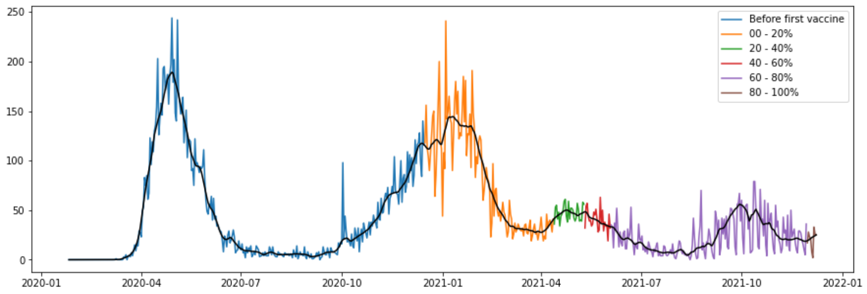
One thing that we can already see with the data is the number of cases actually went up on average when the vaccine started to roll out. This could be due to relaxed lockdown rules as the vaccine started to roll out.

Using the transformed data, we can run an ANOVA test, and find that there is a difference in the mean of the different groups. We can visualize this with a post hoc analysis using Tukey's Honest Significant Difference test. Running a test on these groups gives us the following graph: We can clearly see that there were a lower number of cases when there was no vaccination; however, that does not show the whole picture. As Canadians continued to get vaccinated, the number of cases started to decrease after 20 - 40% got vaccinated.

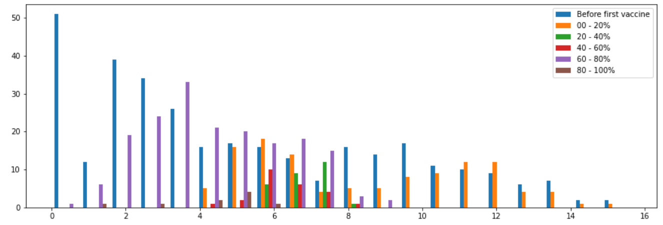


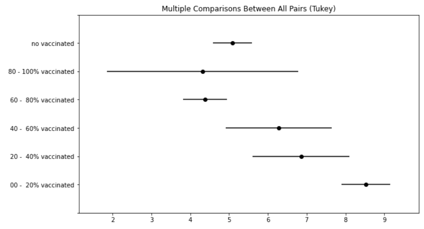
The vaccine seems to be effective in slowing the transmission of COVID in Canada, but does it decrease the death rate?

Death Rate in Canada



The number of deaths in Canada starts high at the start of the pandemic. The initial dip could be caused by lockdowns being implemented. We can see a general trend downwards from the start of the pandemic towards the end, with an initial spike in the middle as Canada started its rollout. Putting the data through the same process as above for the number of cases, we get the following graphs:





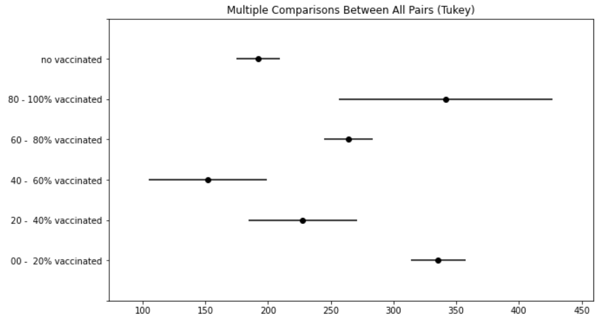
These visuals show that as the vaccine started to roll out, there was some significant decrease in the number of deaths.

For both new cases and deaths, there is a large margin for the mean of the 80 - 100% group. This is because Canada just recently hit 80% vaccination rate, so there is not as much data in the group.

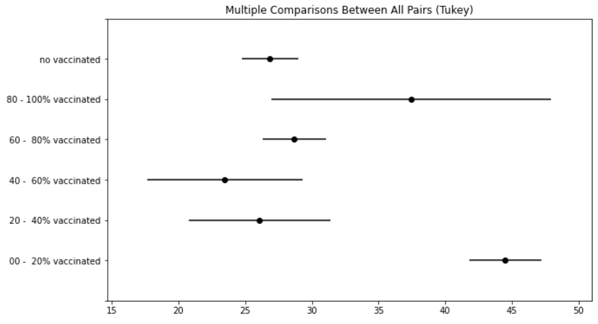
United States

Data from the United States tells a similar story with Canada. To look at the data from the United States, see the ANOVA notebook in the Github repository listed in the Appendix.

US New Cases relative to Canada Vaccination Rate



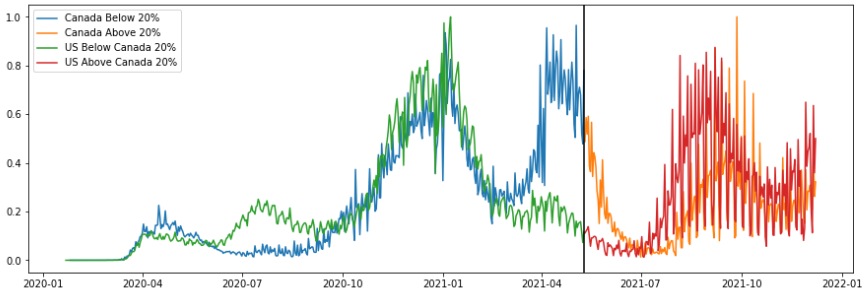
US New Deaths relative to Canada Vaccination Rate



We can see that there is a decrease in new cases and new deaths as the vaccine rolled out; however, there is a spike at the 80 - 100% group that was more intense than the Canada data showed.

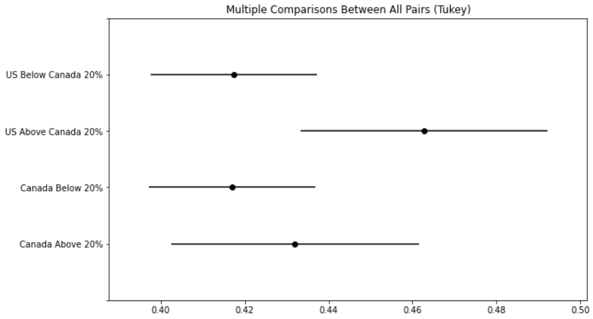
Canada against the United States

Finally the question we want to answer. How does Canada compare against the United States? To do this, instead of doing it in sections of 20%, the data was grouped into two groups for each country. The divide was when 40% of Canada was vaccinated. This divide was chosen since the first 3 groups (no vaccinated, 0 - 20% vaccinated, and 20 - 40% vaccinated) could be one group, and the remaining groups in the latter half. This gives us the following graph:

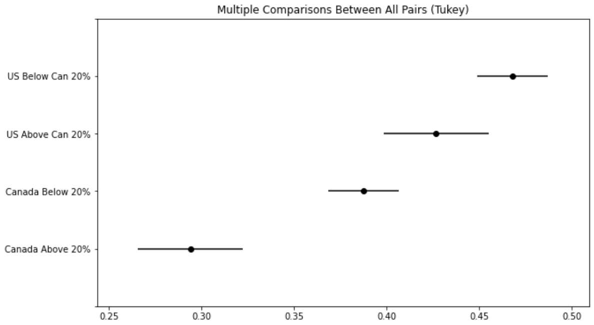


In order to get the graph above, the data had to be divided by the country max, so we could get a scale from 0.00-1.00. Checking both groups in a post hoc analysis shows us the following:

US and Canada New Cases relative to Canada Vaccination Rate



US and Canada New Deaths relative to Canada Vaccination Rate

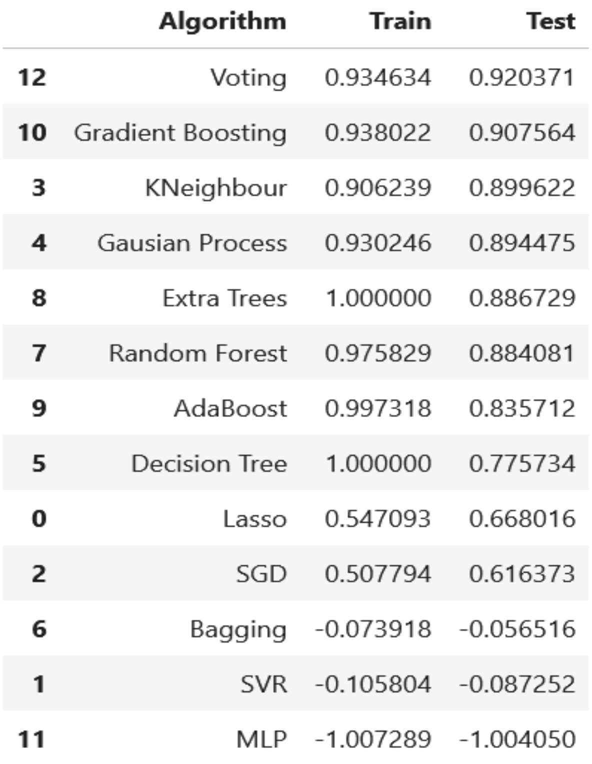


We can see that Canada and the United States had very similar case ratios before the vaccine. As the vaccine rolled out, the ratio of new cases in the United States increased much more than the ratio of new cases in Canada. We can also see that the vaccine decreased the number of new deaths in both countries. We also can see that the decrease in new deaths was much more intense in Canada compared to the effect in the United States.

At the time of writing, Canada was at 80% vaccinated, while the United States was at 71%. Future work could compare the two countries again when the United States gets to 80% to see if Canada still is in the lead of vaccinated, or if Canada has plateaued at that point.

Machine Learning Results

For the prediction of the number of cases we applied various regression models and selected the one with highest accuracy. Starting with the base models, we applied various techniques to make predictions more relevant, including but not limited to: scaling, feature selection (used people vaccinated, total cases and deaths/cases), encoding, splitting data into train and test, changing parameters (like max_depth, criterion and many more according to the model). The results for the models used in decreasing (test) accuracies are:



	Algorithm	Train	Test
12	Voting	0.934634	0.920371
10	Gradient Boosting	0.938022	0.907564
3	KNeighbour	0.906239	0.899622
4	Gausian Process	0.930246	0.894475
8	Extra Trees	1.000000	0.886729
7	Random Forest	0.975829	0.884081
9	AdaBoost	0.997318	0.835712
5	Decision Tree	1.000000	0.775734
0	Lasso	0.547093	0.668016
2	SGD	0.507794	0.616373
6	Bagging	-0.073918	-0.056516
1	SVR	-0.105804	-0.087252
11	MLP	-1.007289	-1.004050

We predicted new cases with an accuracy of ~92% with the help of Voting Regressor, a similar accuracy was achieved while predicting the number of deaths with the same model. With the help of the predicted number of cases and fatalities we will be better prepared to face the next wave.

Limitations

A possible limitation of the project was comparing Canada and the United States to the international whole. Comparing the international whole against each country will cause a skew towards lower GDP averages. A better comparison might be comparing countries with similar GDP or economic standings. Another possible limitation was a lack of access to an active case count number in our data which may have been helpful as a feature when training models. This number was not in the original dataset, likely because a definite value is impossible, but an approximation may have still been useful for our needs. If there was more time to work on this project it may be possible to gather an active case count number from a different source and include it in our analysis.

Project Experience Summary

Aidan: Applied ANOVA test and post hoc analysis to compare vaccine effectiveness in Canada and the United States.

Daniel: Prepared COVID-19 data for use in statistical analysis and performed Mann-Whitney U-Tests to find results about vaccination effectiveness.

Hiten: Built various machine learning models to predict the number of cases. This project improved my understanding of various algorithms for prediction and helped me learn how to improve the performance for better results, simultaneously helping me work on my soft skills like communicating problems with the group and coming up with creative solutions.

Appendix

- COVID-19 Data Source: https://github.com/owid/covid-19-data/tree/master/public
- Project Repository https://github.com/daniel-spooner/cmpt-353-final-project