Machine Learning Algorithm to Predict COVID-19 Deaths

Intro to Computer Science

CSC – 103 – 001

Early Birds

Providence College
The Mathematics & Computer Science, Providence, RI

Submitted to
Dr. Reza Sadeghi

Fall 2020

Final Project Report of Machine Learning Algorithm to Predict COVID-19 Deaths

Team Name

Early Birds

Team Members

1. Jared Quast (Team Leader)	jquast@friars.providence.edu
2. Julia Rose Sclafani	jsclafan@friars.providence.edu
3. Jacob Hefele	jhefele@friars.providence.edu
4. Brendan Kennedy	bkenned7@friars.providence.edu
5. Patrick Thompson	pthomps6@friars.providence.edu

Roles of Team Members

- **All team members worked on the report, and all helped with general project structure**

1. Jared Quast

- Data Cleaning

2. Julia Rose Sclafani

- Graphs

3. Jacob Hefele

- Train/Test Sets

4. Patrick Thompson

- Linear Regression Model

5. Brendan Kennedy

- Logistic Regression Model

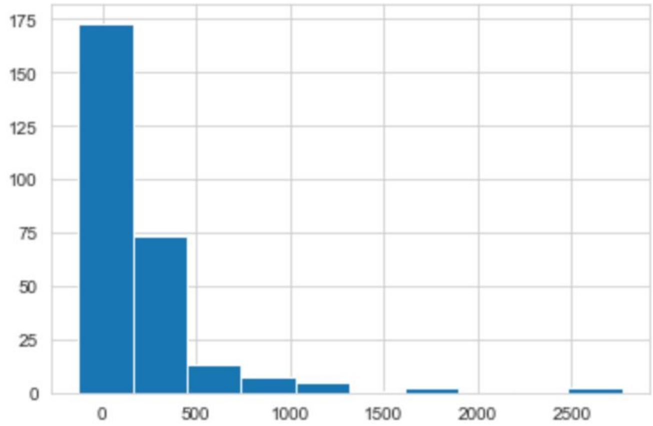
Table of Contents

Table of figures:

New Cases Histogram

X-axis: number of cases in a given day

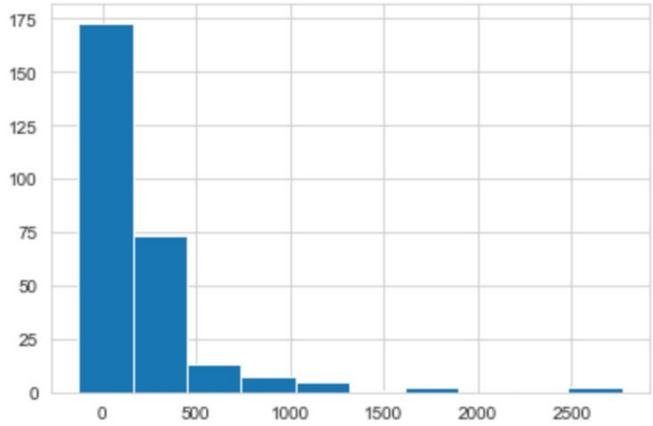
Y-axis: number of days in each range



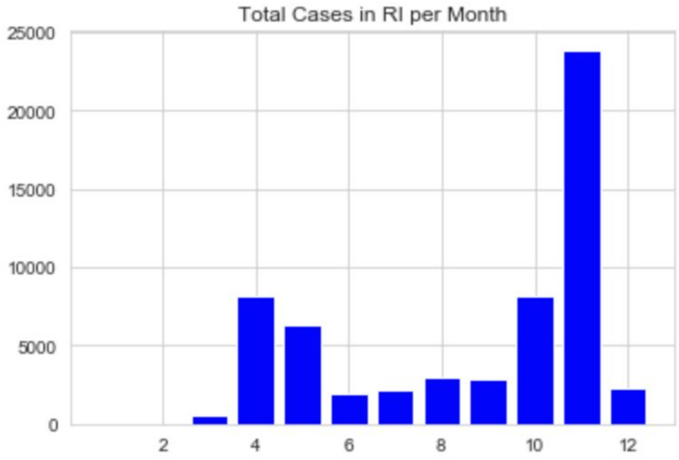
Deaths per Day Histogram

X-axis: number of deaths in a given day

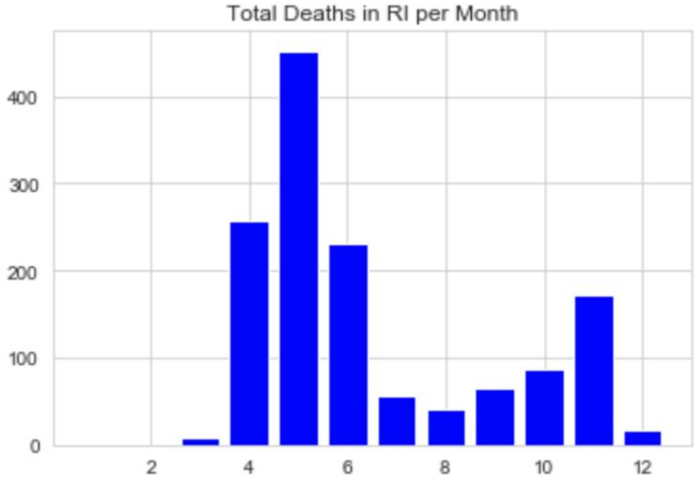
Y-axis: number of days in each range



Total Cases in RI per Month
X-axis: month
Y-axis: number of cases



Total Deaths in RI per Month
X-axis: month
Y-axis: number of deaths



Linear Regression Line: Total_Deaths_Estimate = 0.0071 * new_case + 3.43

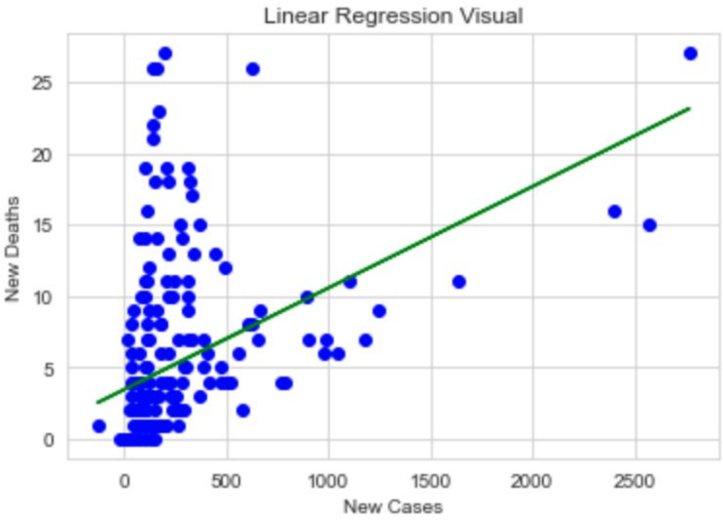
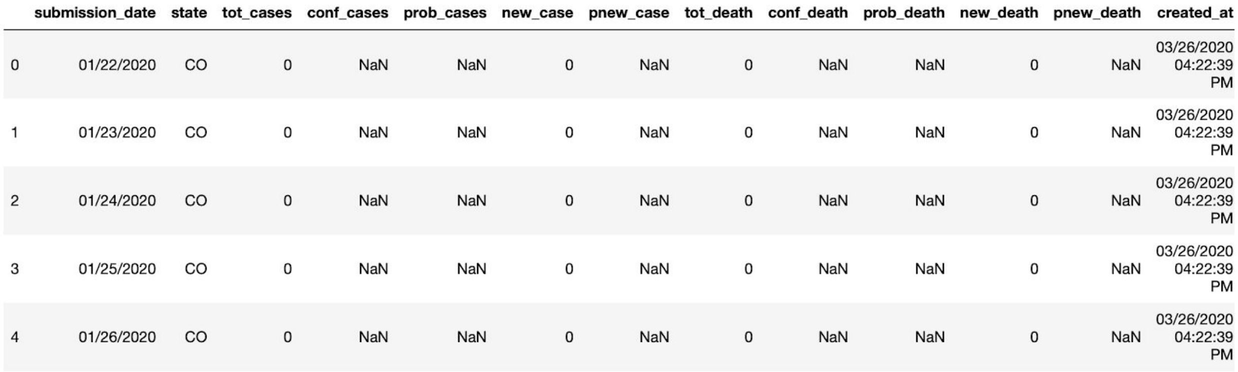


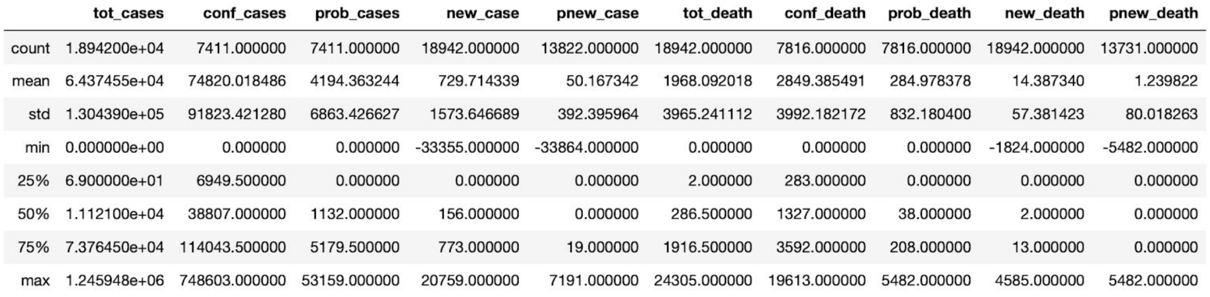
Table of Tables:

Raw Data



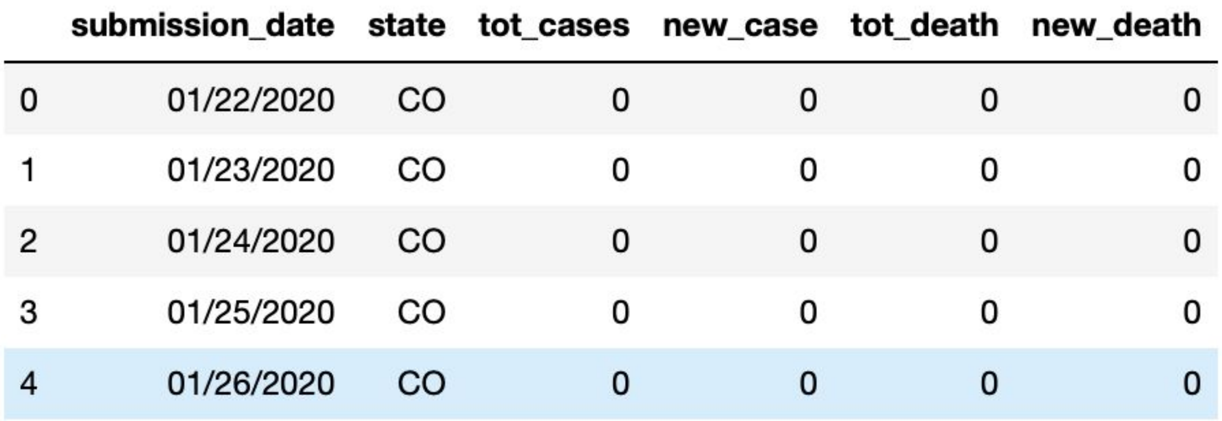
	submission_date	state	tot_cases	conf_cases	prob_cases	new_case	pnew_case	tot_death	conf_death	prob_death	new_death	pnew_death	created_at
0	01/22/2020	CO	0	NaN	NaN	0	NaN	0	NaN	NaN	0	NaN	03/26/2020 04:22:39 PM
1	01/23/2020	CO	0	NaN	NaN	0	NaN	0	NaN	NaN	0	NaN	03/26/2020 04:22:39 PM
2	01/24/2020	CO	0	NaN	NaN	0	NaN	0	NaN	NaN	0	NaN	03/26/2020 04:22:39 PM
3	01/25/2020	CO	0	NaN	NaN	0	NaN	0	NaN	NaN	0	NaN	03/26/2020 04:22:39 PM
4	01/26/2020	CO	0	NaN	NaN	0	NaN	0	NaN	NaN	0	NaN	03/26/2020 04:22:39 PM

Statistical Data



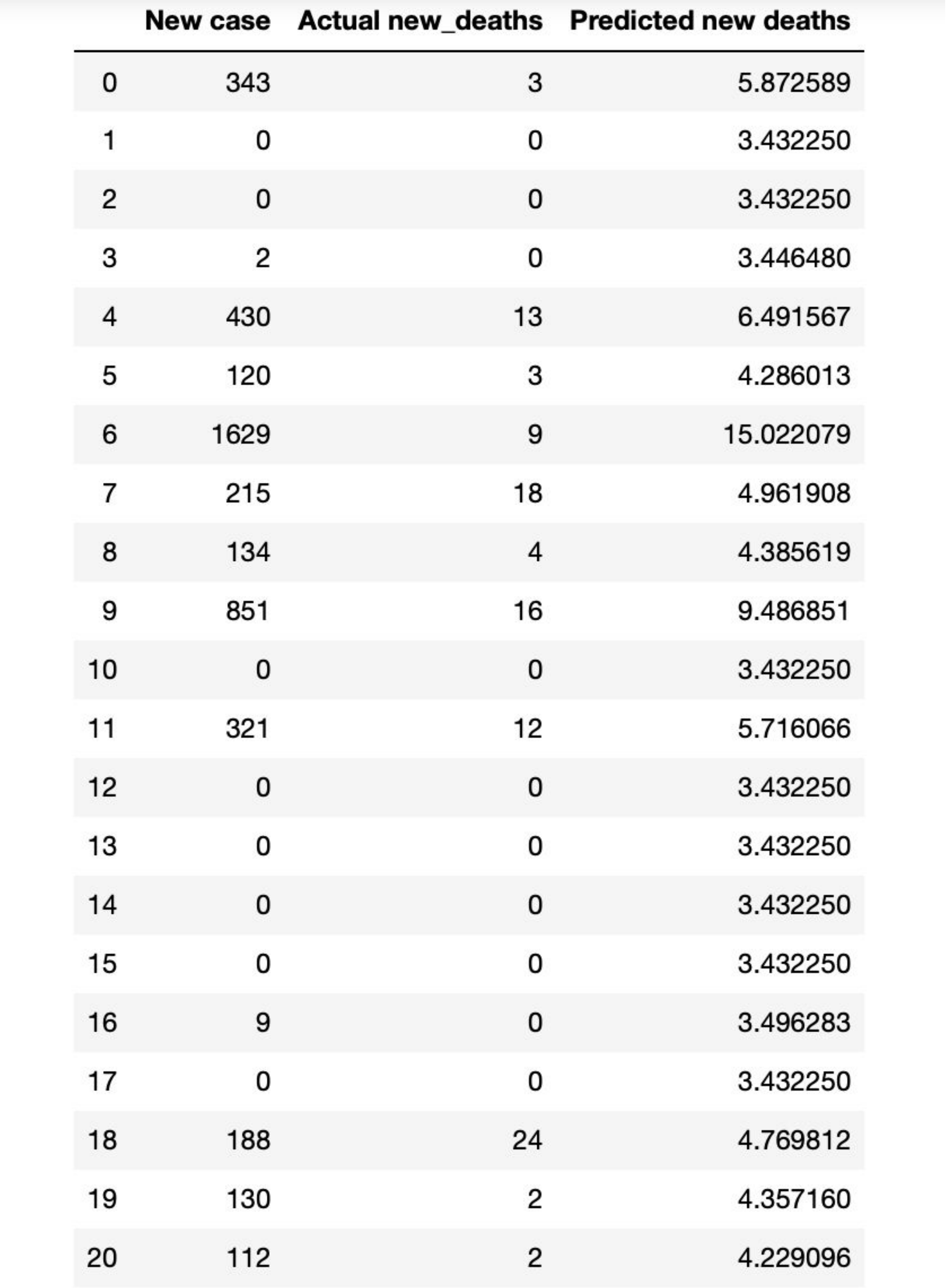
	tot_cases	conf_cases	prob_cases	new_case	pnew_case	tot_death	conf_death	prob_death	new_death	pnew_death
count	1.894200e+04	7411.000000	7411.000000	18942.000000	13822.000000	18942.000000	7816.000000	7816.000000	18942.000000	13731.000000
mean	6.437455e+04	74820.018486	4194.363244	729.714339	50.167342	1968.092018	2849.385491	284.978378	14.387340	1.239822
std	1.304390e+05	91823.421280	6863.426627	1573.646689	392.395964	3965.241112	3992.182172	832.180400	57.381423	80.018263
min	0.000000e+00	0.000000	0.000000	-33355.000000	-33864.000000	0.000000	0.000000	0.000000	-1824.000000	-5482.000000
25%	6.900000e+01	6949.500000	0.000000	0.000000	0.000000	2.000000	283.000000	0.000000	0.000000	0.000000
50%	1.112100e+04	38807.000000	1132.000000	156.000000	0.000000	286.500000	1327.000000	38.000000	2.000000	0.000000
75%	7.376450e+04	114043.500000	5179.500000	773.000000	19.000000	1916.500000	3592.000000	208.000000	13.000000	0.000000
max	1.245948e+06	748603.000000	53159.000000	20759.000000	7191.000000	24305.000000	19613.000000	5482.000000	4585.000000	5482.000000

Data with Dropped Columns



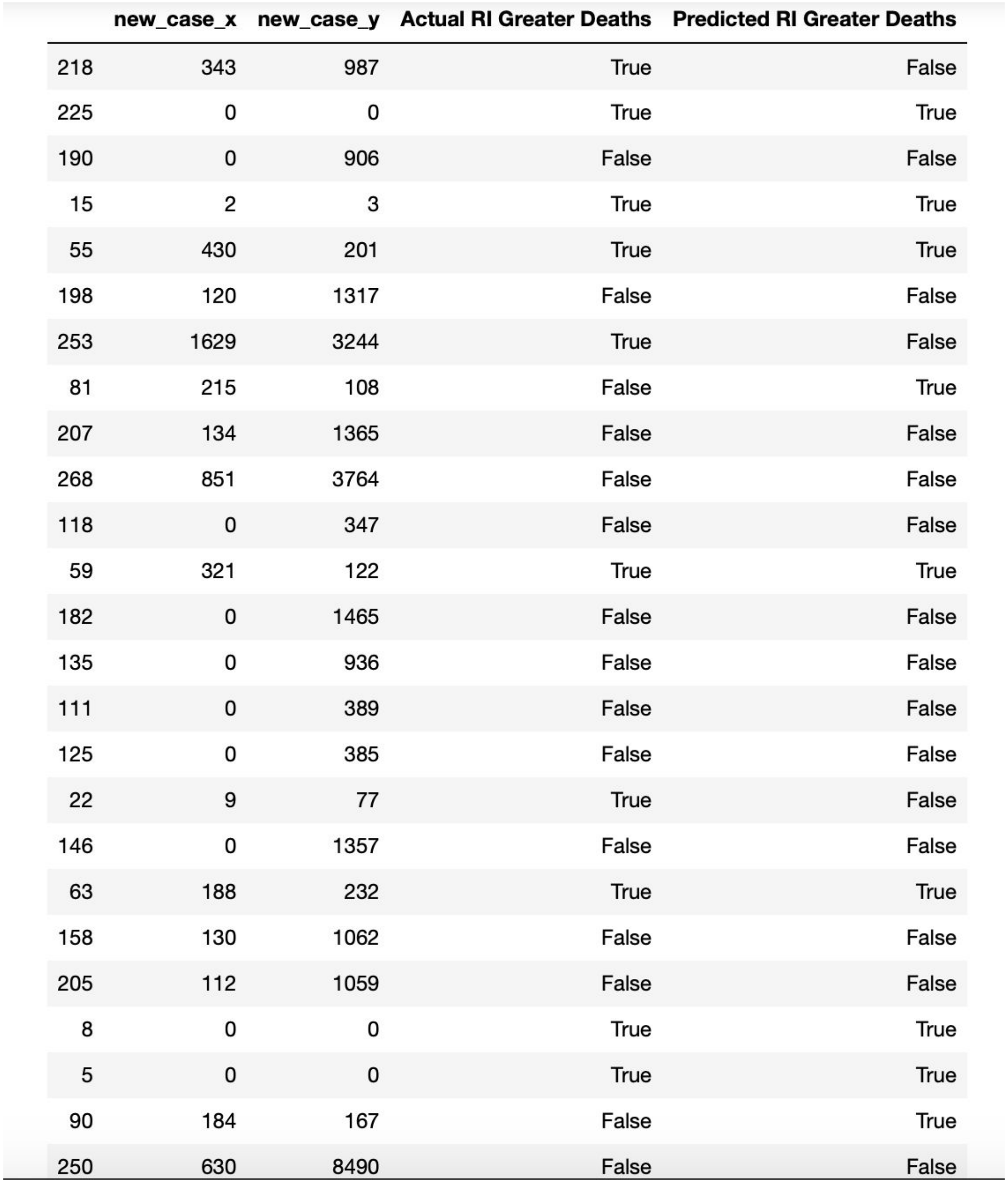
	submission_date	state	tot_cases	new_case	tot_death	new_death
0	01/22/2020	CO	0	0	0	0
1	01/23/2020	CO	0	0	0	0
2	01/24/2020	CO	0	0	0	0
3	01/25/2020	CO	0	0	0	0
4	01/26/2020	CO	0	0	0	0

Linear Regression Data



	New case	Actual new_deaths	Predicted new deaths
0	343	3	5.872589
1	0	0	3.432250
2	0	0	3.432250
3	2	0	3.446480
4	430	13	6.491567
5	120	3	4.286013
6	1629	9	15.022079
7	215	18	4.961908
8	134	4	4.385619
9	851	16	9.486851
10	0	0	3.432250
11	321	12	5.716066
12	0	0	3.432250
13	0	0	3.432250
14	0	0	3.432250
15	0	0	3.432250
16	9	0	3.496283
17	0	0	3.432250
18	188	24	4.769812
19	130	2	4.357160
20	112	2	4.229096

Logistic Regression Data



	new_case_x	new_case_y	Actual RI Greater Deaths	Predicted RI Greater Deaths
218	343	987	True	False
225	0	0	True	True
190	0	906	False	False
15	2	3	True	True
55	430	201	True	True
198	120	1317	False	False
253	1629	3244	True	False
81	215	108	False	True
207	134	1365	False	False
268	851	3764	False	False
118	0	347	False	False
59	321	122	True	True
182	0	1465	False	False
135	0	936	False	False
111	0	389	False	False
125	0	385	False	False
22	9	77	True	False
146	0	1357	False	False
63	188	232	True	True
158	130	1062	False	False
205	112	1059	False	False
8	0	0	True	True
5	0	0	True	True
90	184	167	False	True
250	630	8490	False	False

Introduction:

During March 2020, our entire world changed due to COVID-19. Our country shut down with businesses closing and families required to stay home. The rate of COVID cases continuously increased and hospitalization rates spiked. Our research question, Can a machine learning algorithm predict COVID-19 deaths in Rhode Island based on data of COVID-19 cases from March 1, 2020 to present day, is very relevant to help predict COVID deaths within the new year. It will be interesting to see if the rate will decrease as more people have already contracted the virus or with the potential of a vaccination. Our machine learning algorithm should be able to continue to predict COVID deaths with updated data.

As for teamwork, we all collaborated well together. We continuously met via Zoom and worked on the coding together. No one was afraid to ask a question or for help. It was beneficial working together as we could combine our knowledge of coding and figure out how to clean the data, create graphs, and perform our regression models.

Research Question:

Can a machine learning algorithm predict COVID-19 deaths in Rhode Island based on data of COVID-19 cases from March 1, 2020 to present day?

Project description:

Modules and their uses:

- pandas: data analysis and manipulation
- numpy: computation
- seaborn: data visualization
- matplotlib: data visualization
- scikit-learn: predictive data analysis

Models:

- Linear Regression: A line of best fit, taking into account both cases and deaths, which predicts number of deaths on a given day in Rhode Island based on number of cases.
- Logistic Regression: The regression compared data detailing new cases and new deaths in the states of Rhode Island and Montana (Montana was chosen as a comparative state, as their number of total cases per day was closest to Rhode Island's number of total cases per day). The regression then compares the number of new cases in Rhode Island to the number of new cases in Montana on a given day. If the number of new cases in Rhode Island is greater than the new cases in Montana on that day, the logistic model predicts that Rhode Island will have more COVID-19 deaths on that day. If the number of new cases in Montana is greater than the new cases in Rhode Island on that day, the logistic regression model predicts that Montana will have more COVID-19 deaths on that day. The model then considers the actual number of COVID-19 deaths in Rhode Island and Montana and returns a true/false response, detailing whether the logistic model correctly predicted which state would have more deaths.

List of key variables:
-RI COVID-19 Cases
-RI COVID-19 Deaths
-COVID-19 Cases from Comparison State (MT)
-COVID-19 Deaths from Comparison State (MT)

List of functions:

Admin functions:

- Importing Modules
- Reading .csv file
- Data Cleaning
- Histograms

Guess functions:

- Linear Regression
- Logistic Regression

Business Problem

Problem Context

Hospitalizations from COVID-19 are on the rise as we move towards the winter months. With the holiday season underway, people may gather in larger groups, causing an increase in COVID-19 cases. This is concerning for both healthcare workers and citizens alike, as an increase in COVID-19 cases, will likely lead to an increase in COVID-19 deaths. The dataset will be utilized to project COVID-19 deaths based on COVID-19 cases from March 1, 2020 to present day.

Content

This dataset contains COVID-19 total cases, new cases, total deaths, and new deaths from the end of January 2020 to the beginning of December 2020, per day, per state. There are a total of 19,002 records recorded in this dataset.

Features:

- Date
- State
- Total Cases
- New Cases
- Total Deaths
- New Deaths

Mapping business problem to ML problem

The data set containing the features displayed above summarizes the nationwide effects of COVID-19 throughout 2020. This machine learning algorithm particularly focuses on the state of Rhode Island from March 1, 2020 to December 1, 2020. The algorithm is able to give the expected prediction of COVID-19 responsible deaths given the number of positive cases in a specific time period using linear regression. In addition, the algorithm also incorporates logistic regression as a way to predict whether or not Rhode Island will have more new deaths than the state of Montana based on each states' daily cases.

Report on Related Works

Several sources have reported predictions for the total number of deaths in the future, as a result of COVID-19 cases, including the CDC (Centers for Disease Control and Prevention) and the IHME (Institute for Health Metrics and Evaluation). Both of these sources have graphed the recorded number of COVID-19 deaths over the past months and included predictions for COVID-19 deaths. The CDC includes a wide array of predictions from reputable institutions including Columbia, Notre Dame, UCLA, and UMASS among others. The IHME includes a current prediction (likely based on several reputable sources) and three additional models based on different events occurring, including the presence of universal masks, a rapid vaccine rollout, and the easing of mandates. The consensus surrounding many of the models provided by the CDC and the IHME is that the death toll will continue to rise over the coming weeks, but will increase steadily. However, some models provided by the CDC are predicting there will be a rapid increase in the number of deaths from COVID-19 in the coming weeks, while other models predict there will be a dropoff in COVID-19 deaths over this same period. Ultimately, based on the majority of models provided by the CDC and IHME, it appears that COVID-19 deaths will continue rising at a steady rate over the coming weeks, but due to the fluidity of the situation at hand, it is difficult to make a prediction with complete accuracy.

References:

https://healthdata.gov/dataset/united-states-covid-19-cases-and-deaths-state-over-time

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

https://github.com/RezaSadeghiWSU/Liver-disease-prediction/blob/master/Liver_Disease_prediction.ipynb

https://pandas.pydata.org/

https://numpy.org/

https://seaborn.pydata.org/#:~:text=Seaborn%20is%20a%20Python%20data,attractive%20and%20informative%20statistical%20graphics.

https://matplotlib.org/

https://scikit-learn.org/stable/

https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

https://covid19.healthdata.org/united-states-of-america?view=total-deaths&tab=trend

https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a