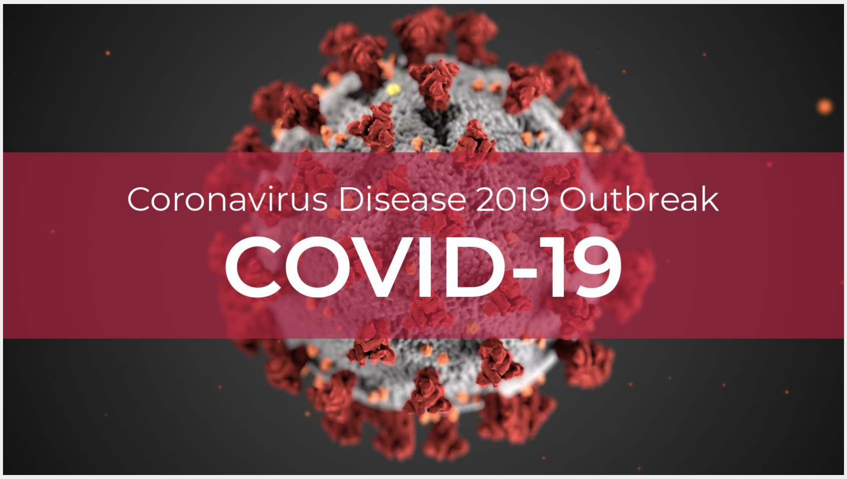


Final Project

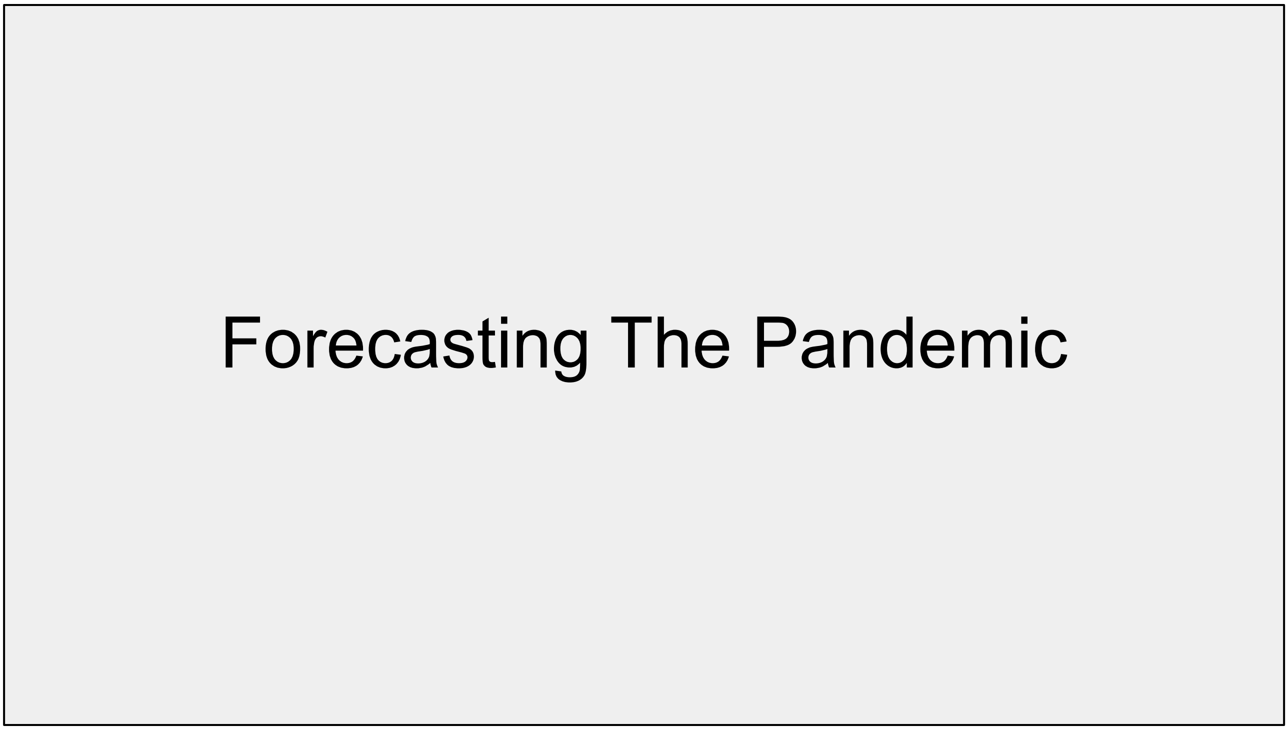
Business Case

Machine Learning:

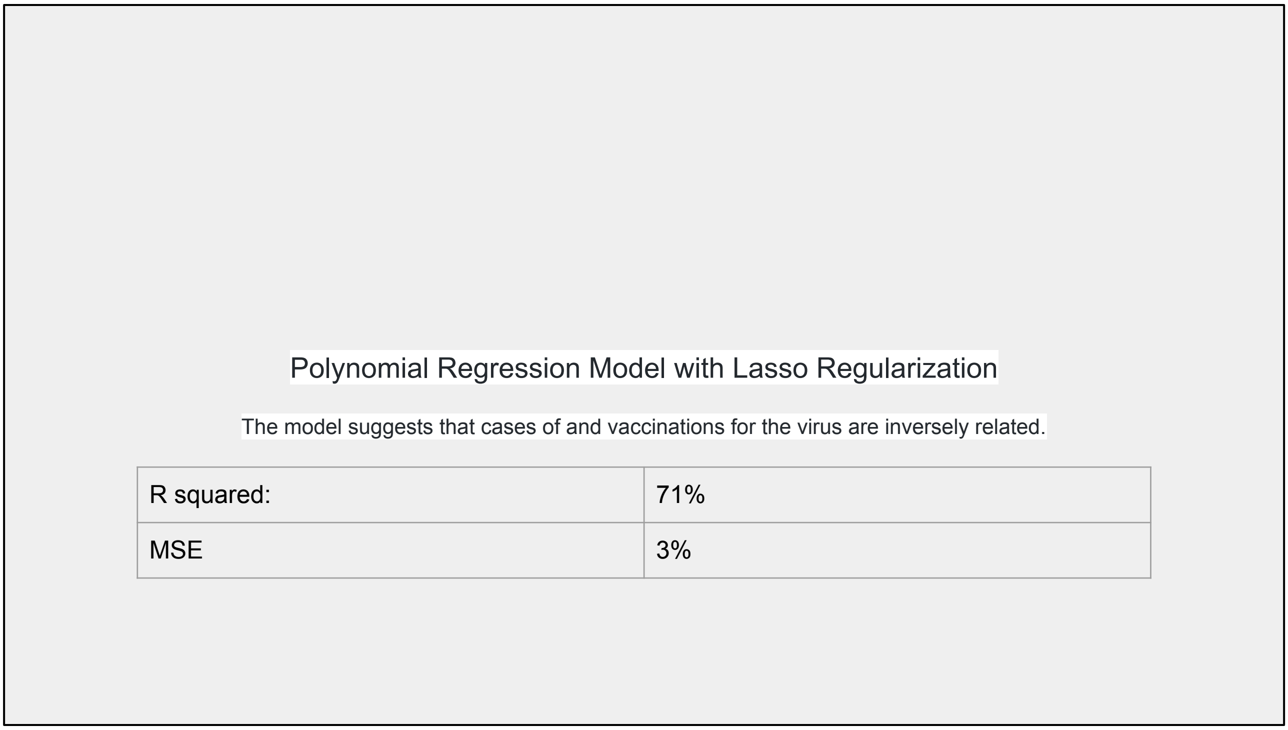
- Regression modeling:
  - A polynomial regression with lasso regularization model is used to determine the polynomial relationship between cases of and vaccinations for SARS-CoV-2.
  - A SARIMA model is used to determine the progression of cases of and vaccinations for SARS-CoV-2.
- Classification Modeling:
  - A logit model is used to determine the log odds relationship between government covid-19 regulations and cases of SARS-CoV-2.
  - An XGBoost model is used to determine the feature importance of government covid-19 regulations to cases of SARS-CoV-2.

Deep Learning:

- Clustering:
  - An NLP model that uses K-means clustering to create class labels for word document importance from a CNBC covid-19 article and then classifies the word document importance using naive bayes machine learning.
- Neural Networks:
  - A CNN model is used to determine whether SARS-CoV-2 is present in CT-scans.
  - A bidirectional RNN model is used to predict the next nucleotide in the sequence of a SARS-CoV-2 genome.



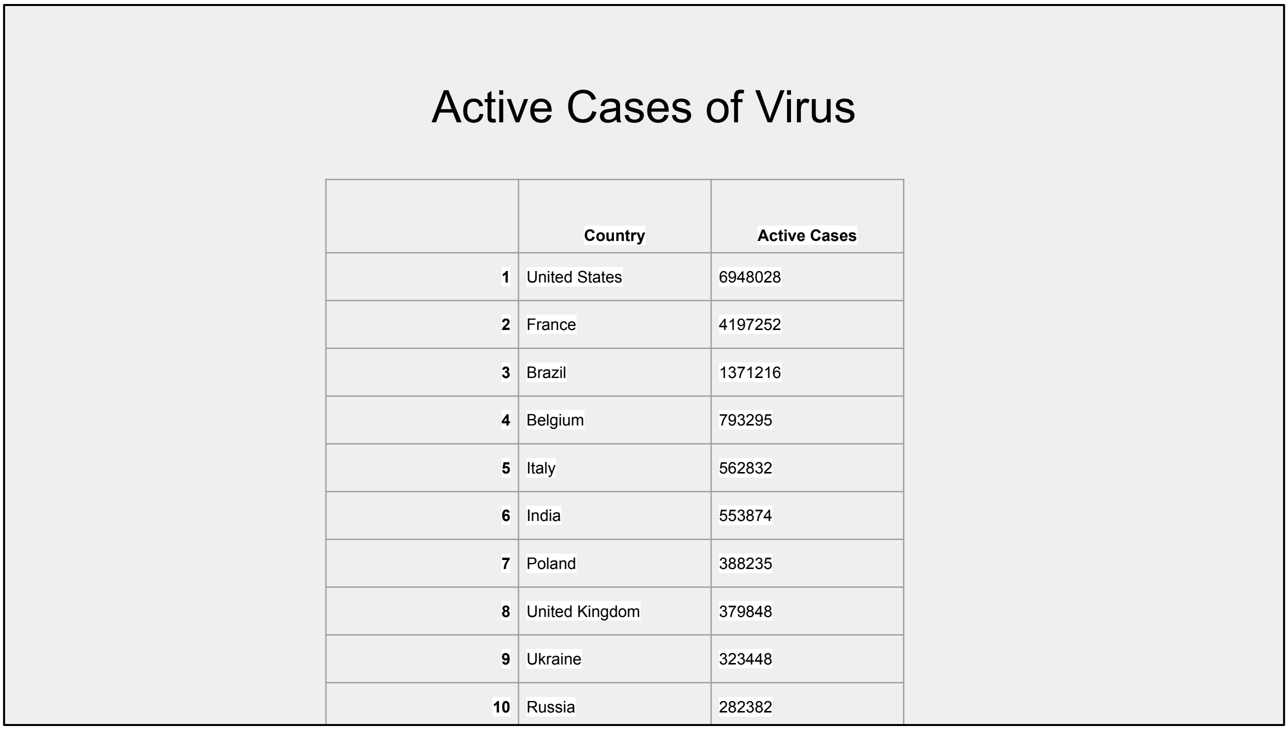
Forecasting The Pandemic



Polynomial Regression Model with Lasso Regularization

The model suggests that cases of and vaccinations for the virus are inversely related.

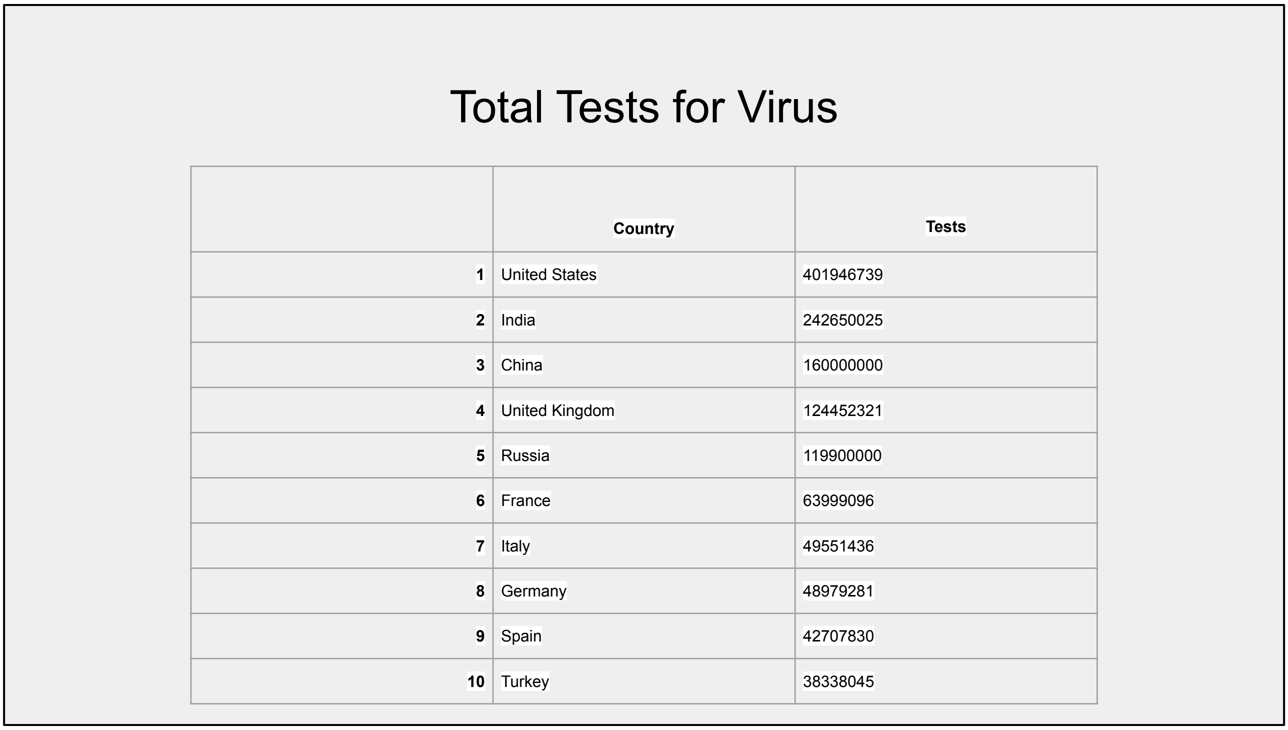
R squared:	71%
MSE	3%



Active Cases of Virus

	Country	Active Cases
1	United States	6948028
2	France	4197252
3	Brazil	1371216
4	Belgium	793295
5	Italy	562832
6	India	553874
7	Poland	388235
8	United Kingdom	379848
9	Ukraine	323448
10	Russia	282382

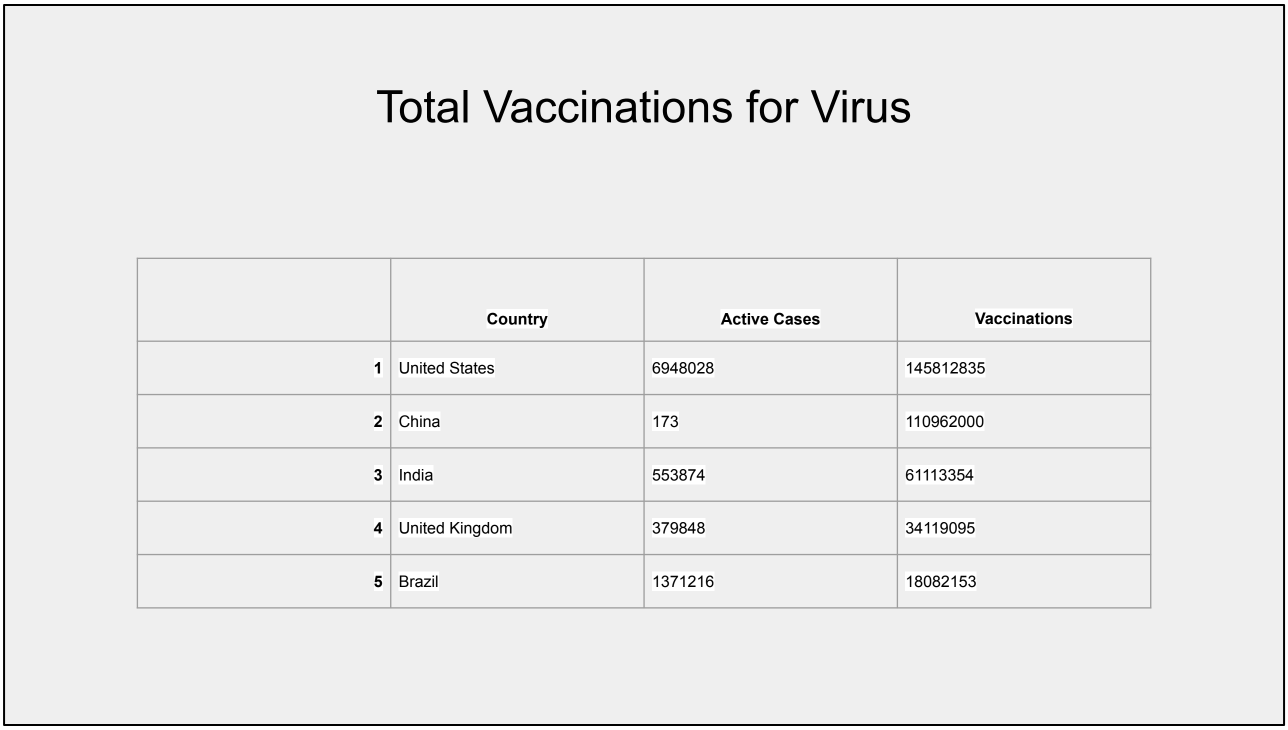
The United States has cumulatively had, and still has, the most cases of the virus.



Total Tests for Virus

	Country	Tests
1	United States	401946739
2	India	242650025
3	China	160000000
4	United Kingdom	124452321
5	Russia	119900000
6	France	63999096
7	Italy	49551436
8	Germany	48979281
9	Spain	42707830
10	Turkey	38338045

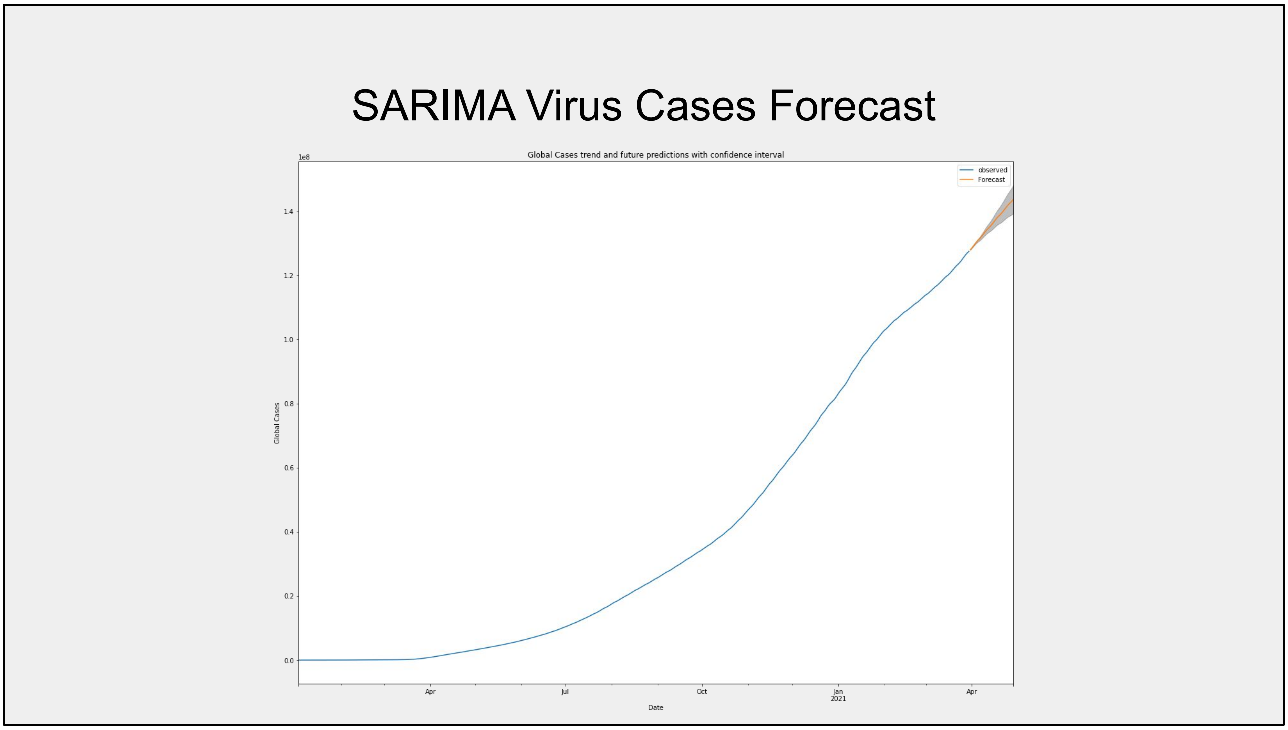
The United States likely has the most cases because it has tested the most.



Total Vaccinations for Virus

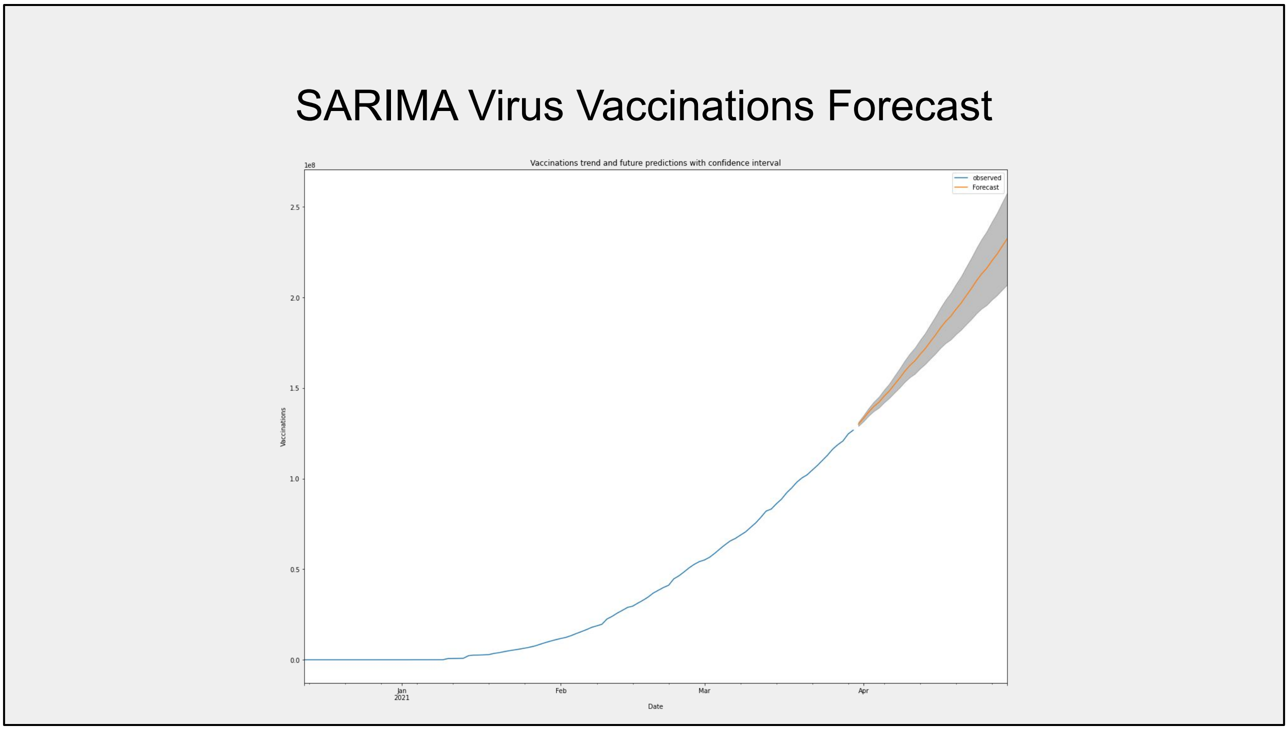
	Country	Active Cases	Vaccinations
1	United States	6948028	145812835
2	China	173	110962000
3	India	553874	61113354
4	United Kingdom	379848	34119095
5	Brazil	1371216	18082153

The United States leads in vaccinations.



SARIMA Virus Cases Forecast

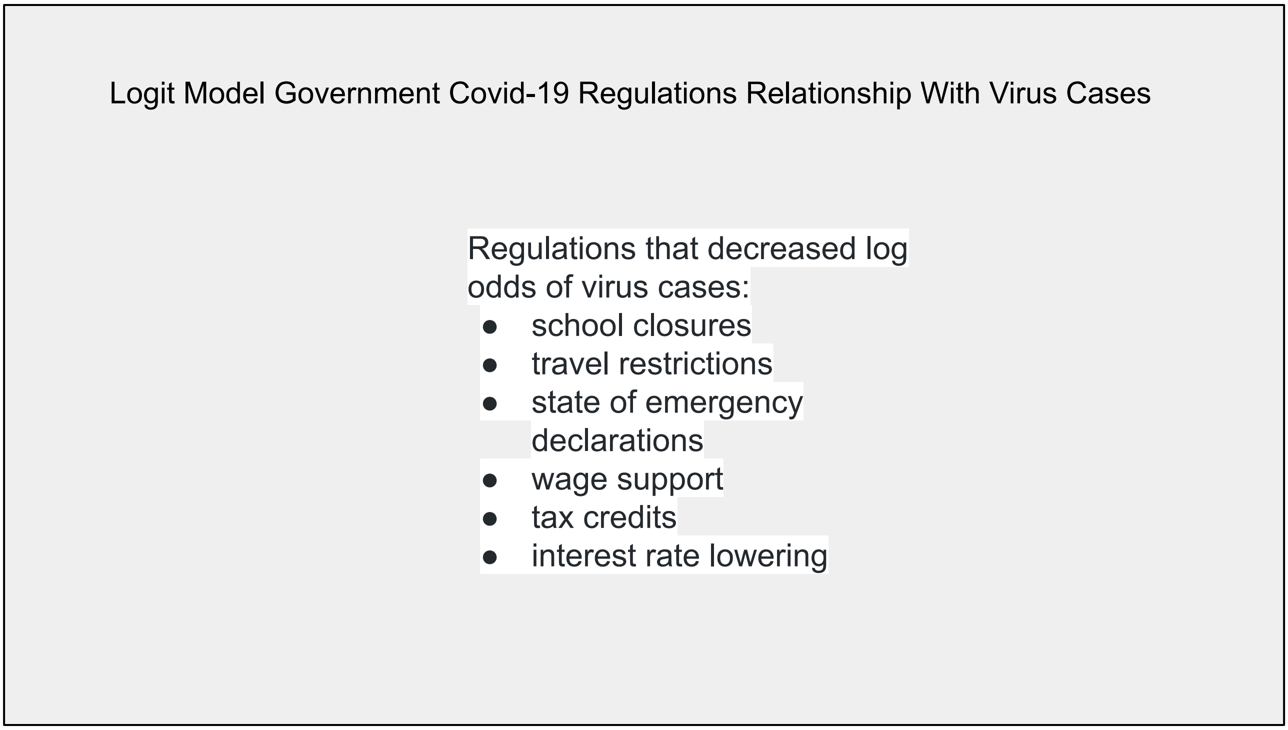
The cumulative cases of the virus are projected to increase by 12.8% over the next month.



SARIMA Virus Vaccinations Forecast

The cumulative virus vaccinations are projected to increase by 83.3% over the next month.

Government Regulations In Response To Covid-19



Logit Model Government Covid-19 Regulations Relationship With Virus Cases

Regulations that decreased log odds of virus cases:

- school closures
- travel restrictions
- state of emergency declarations
- wage support
- tax credits
- interest rate lowering

The data was divided into two classes: cases and no cases of the virus. Government regulations such as disallowing public gatherings and mandating wearing masks did not decrease the log odds of the presence of virus cases.

Logit Metrics

Precision Score: 0.8072260328601053

Recall Score: 0.8241521110703962

F1 Score: 0.8156012651852449

Accuracy Score: 0.817523923444976

Specificity Score: 0.8278185297400733

ROC AUC: 0.8176583048418117

XGBoost Model Government Covid-19 Regulations Relationship With Virus Cases

No statistically significant government covid-19 regulation had significantly greater feature importance than the others in predicting cases of the virus.

The data for this model was divided into three classes: small, medium, and large amount of virus cases.

XGBoost Metrics

Class:0

Precision Score: 0.8877333333333334

Recall Score: 0.8818543046357616

F1 Score: 0.8847840531561463

Accuracy Score: 0.9022162070715615

Specificity Score: 0.912831036841591

ROC AUC: 0.8995837368214652

Class:1

Precision Score: 0.8315385567520732

Recall Score: 0.8749674394373534

F1 Score: 0.8527003871295297

Accuracy Score: 0.8691140810917498

Specificity Score: 0.9005593536357986

ROC AUC: 0.8698059474660661

Class:2

Precision Score: 0.893686165273909

Recall Score: 0.7684630738522954

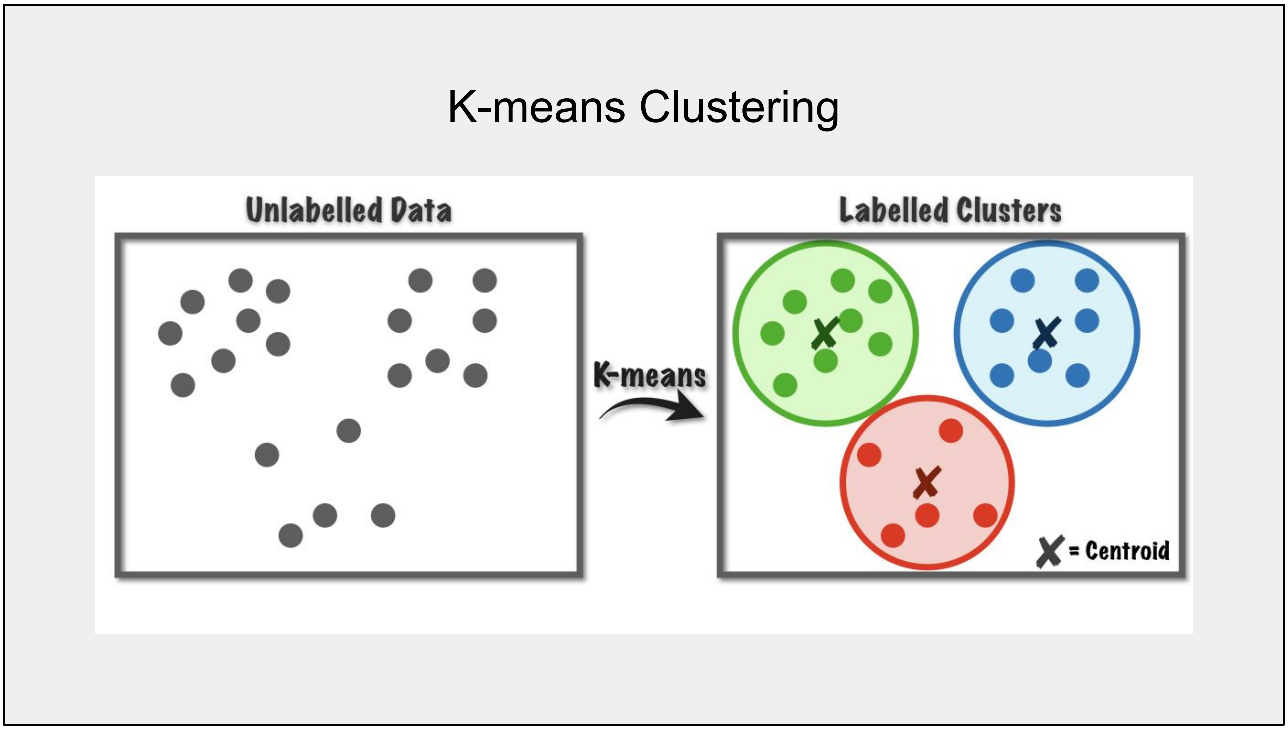
F1 Score: 0.8263575874651211

Accuracy Score: 0.9543788417075509

Specificity Score: 0.9627703960459593

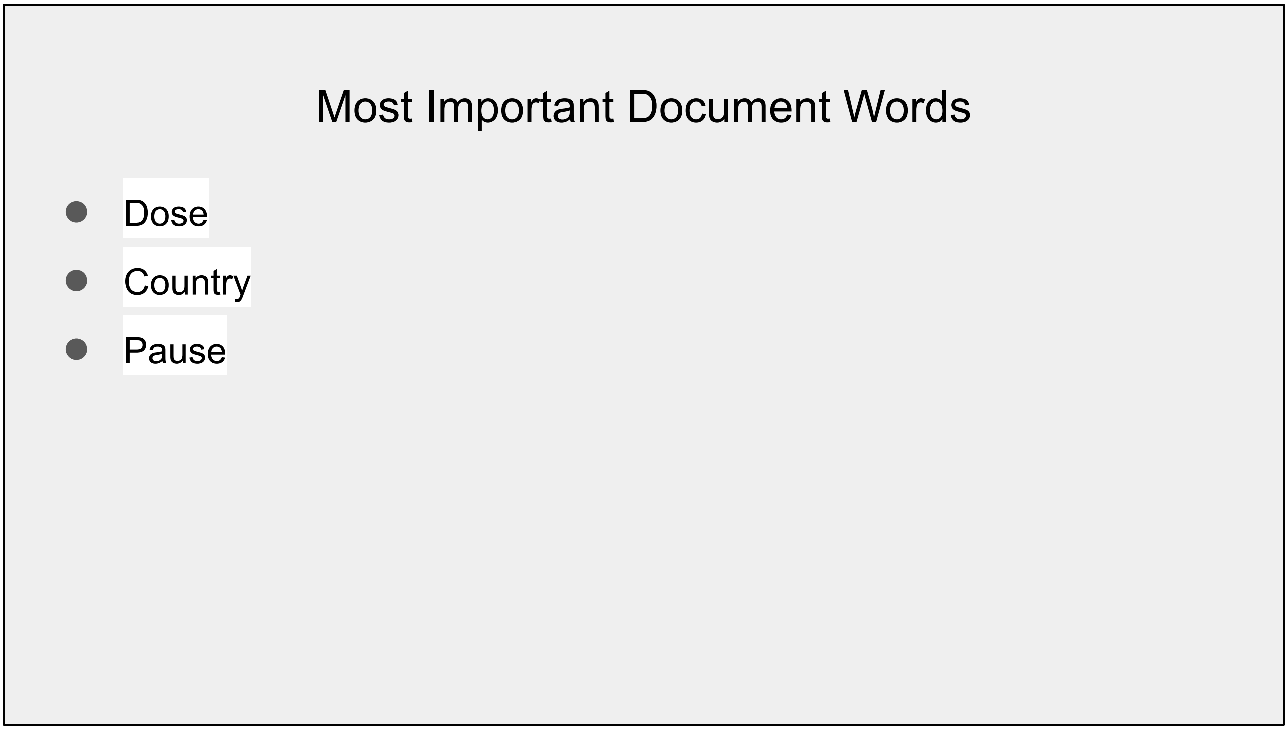
ROC AUC: 0.8767124930595861

Covid-19 Media



K-means Clustering

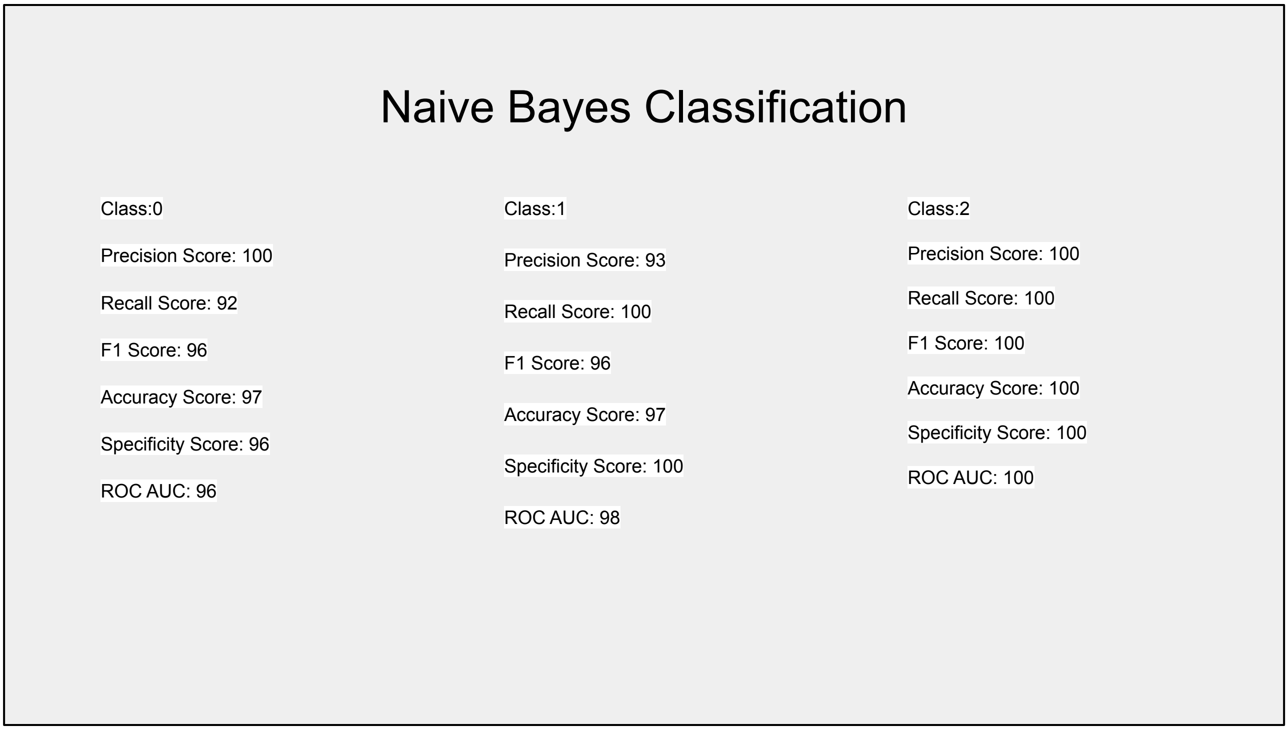
k-means clustering partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean, and is used to determine class labels for word document importance from the covid 19 article. .



Most Important Document Words

- Dose
- Country
- Pause

K-means determined that 3 classes where optimal.



Naive Bayes Classification

Class:0

Precision Score: 100

Recall Score: 92

F1 Score: 96

Accuracy Score: 97

Specificity Score: 96

ROC AUC: 96

Class:1

Precision Score: 93

Recall Score: 100

F1 Score: 96

Accuracy Score: 97

Specificity Score: 100

ROC AUC: 98

Class:2

Precision Score: 100

Recall Score: 100

F1 Score: 100

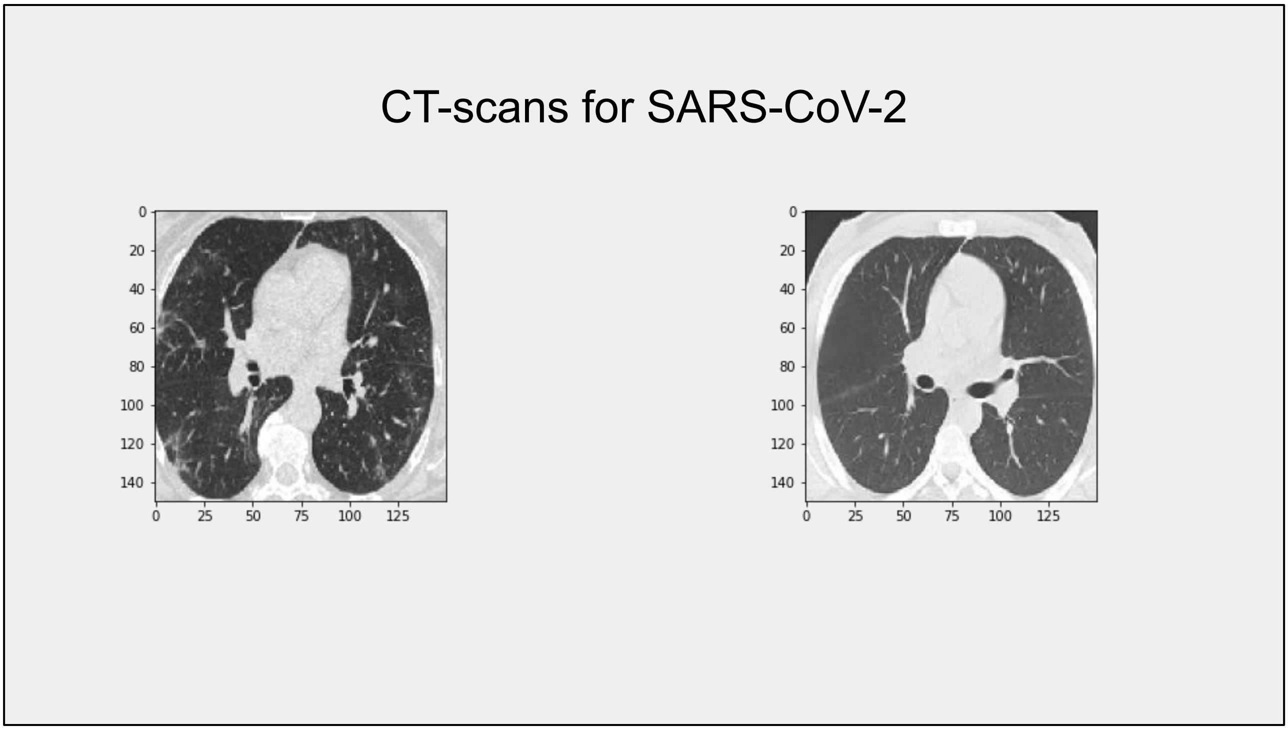
Accuracy Score: 100

Specificity Score: 100

ROC AUC: 100

Naive bayes is used to classify how important a word is to a document based on the posterior probability of it belonging to a class.

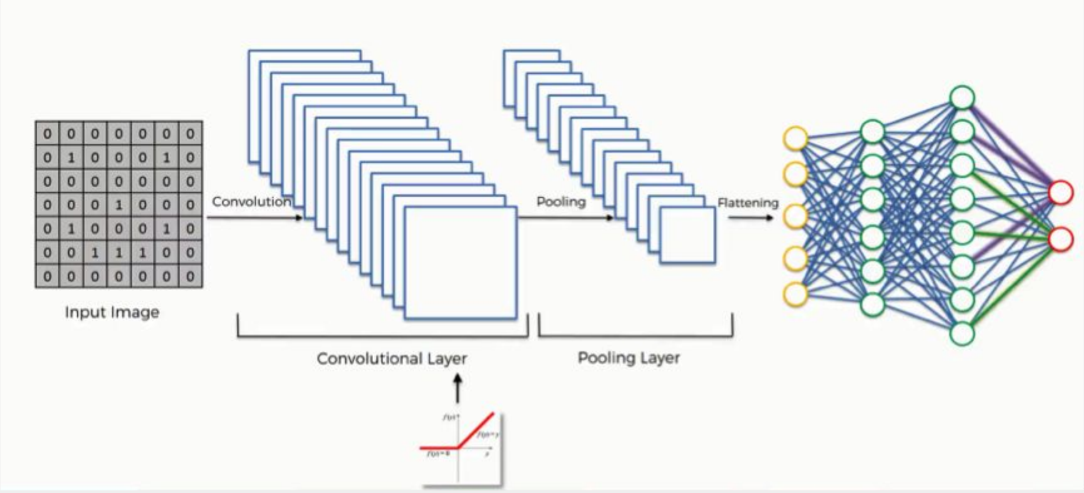
Sars-CoV-2 Biology



CT-scans for SARS-CoV-2

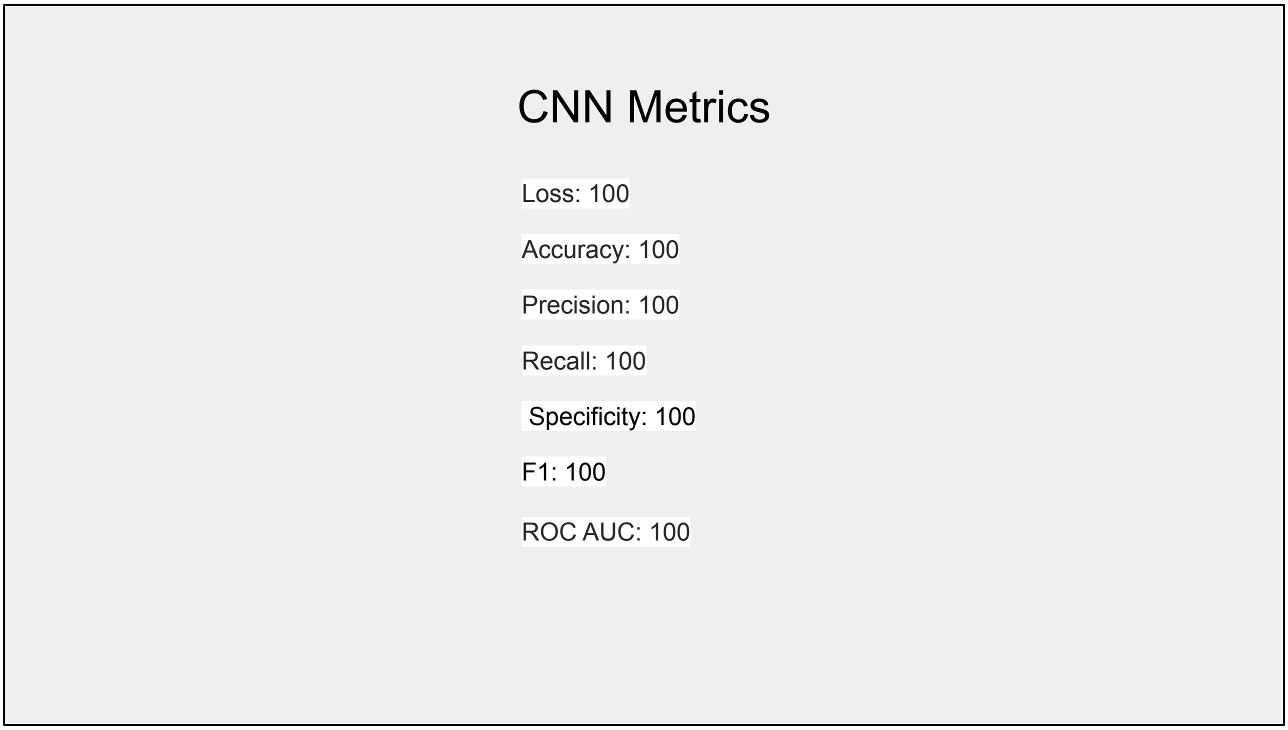
The CT-scan to the left is a negative case and the one to the right a positive case of SARS-CoV-2.

Convolutional Neural Network



0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

A convolutional neural network takes an image represented by a matrix of pixel values, convolutes the image with a filter matrix to find patterns and pools the result to reduce dimensionality, and then flattens the result to be classified.



CNN Metrics

Loss: 100

Accuracy: 100

Precision: 100

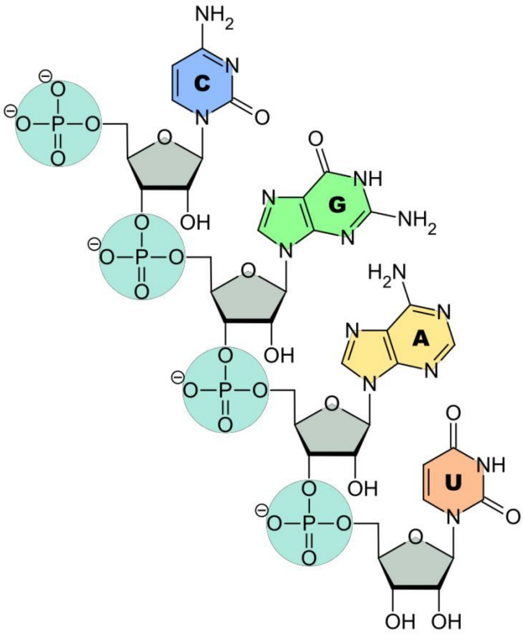
Recall: 100

Specificity: 100

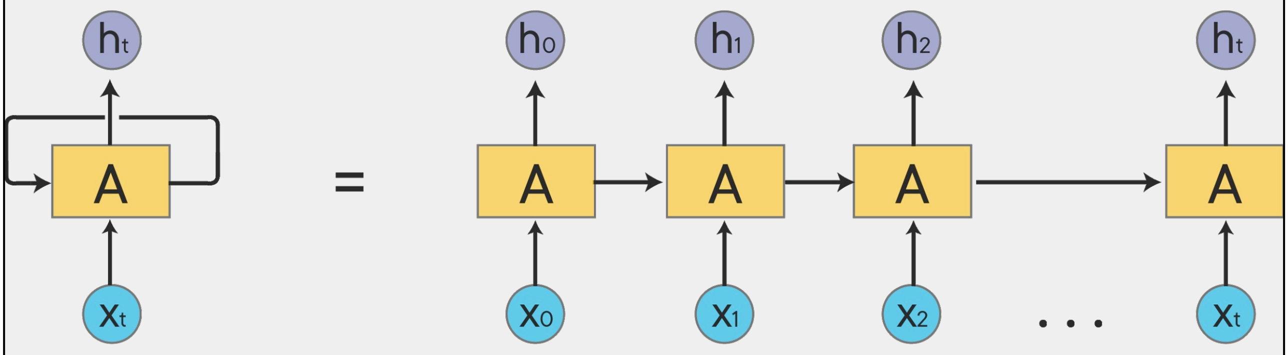
F1: 100

ROC AUC: 100

RNA



Recurrent Neural Network



Recurrent Neural Network is a neural network that passes its output, h, from a layer back into itself as input for the next layer. An RNN is used to predict the next nucleotide in the genome.

RNN Metrics

Loss: 40

Accuracy: 82

Precision: 94

Recall: 74

Specificity: 98

F1: 83

ROC AUC:

- Class 0: 85
- Class 1: 87
- Class 2: 85
- Class 3: 86

Conclusion

- Classifying whether a CT-scan has SARS-CoV-2 in it will increase clinical efficiency

- Predicting the next nucleotide in a genome will be useful in reconstructing fragmented genomes. .

- The virus spread is slowing and will decrease more as more people get vaccinated due because cases of and vaccinations for the virus have an inverse relationship.

- Instead of wearing masks and preventing gatherings, carrying handkerchiefs in which people could sneeze or cough and sanitizing areas were people gather would be sufficient in preventing the spread of SARS-CoV-2.

Future Work

Using neural networks to predict sequences of nucleotides or amino acids.

Thank You

Sources

https://www.ncbi.nlm.nih.gov/

https://datarepository.wolframcloud.com/

https://www.analyticsvidhya.com/blog/2020/02/mathematics-behind-convolutional-neural-network/

https://colah.github.io/posts/2015-08-Understanding-LSTMs/

https://stanford.edu/

https://machinelearningmastery.com/

https://towardsdatascience.com/

https://medium.com/

https://ruder.io/

https://stackoverflow.com/