Predicting the fatality of COVID-19 using machine learning models

Daniel Choi, Ngoc Bui, Yi Shan

University of British Columbia

Predicting the fatality of COVID-19 using machine learning models

Daniel Choi, Ngoc Bui, Yi Shan

Abstract

The sudden spread of COVID-19 has confounded many researchers and the determination of contraction trends with demographic predictors is a topic of interest. The aim of our study was to provide additional insight towards the outcome of patients that contract the virus. This information could contribute to the determination of suitable preventive measures and treatments for those who are predicted to be more susceptible. Previously there have been a number of different models derived from machine learning methods in order to predict the trends in the growth and contraction of the disease. However, less research has been done to predict fatality after contraction of COVID-19 depending on individual characteristics of the patients such as age and gender, which is what our study aims to develop. We used data from Alberta, Ontario, and Japan to predict the fatality of individuals who contracted COVID-19 and determined the accuracy of different prediction models. More specifically, we applied Ridge Regression, Lasso Regression, Logistic Regression, and Support Vector Machine (SVM) Regression to each dataset and subsequently validated the regression models through train test split with 40% of the data. The results showed that SVM performed the best for all three datasets with 92.0% accuracy for Alberta, 85.1% accuracy for Ontario, and 92.2% accuracy for Japan.

Introduction

Statistics and machine learning have been used for epidemiology and related applications. In the COVID-19 research field, previous studies have discovered trends in infection rates, fatality rates, and symptom severity with relation to age, sex, income, ethnicity, and the like [1,2,3,4,5]

Methods - Machine Learning

Ridge and Lasso Regression

Extensions of linear regression.
Determine whether a particular factor has a stronger effect on the outcome of the case

Logistic Regression

Logistic regression is a statistical model that is often used as a classification model that can classify two different possible outcomes using a logistic function. The logistic regression is actually an extension of linear regression but focuses more on classification problems rather than predicting probabilities of the outcomes.

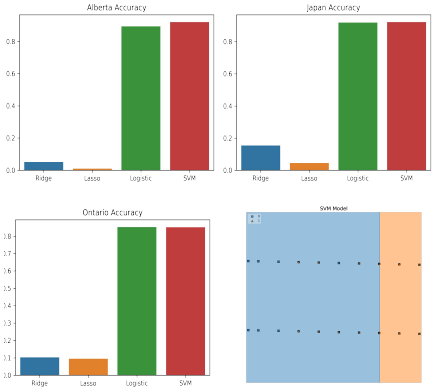
SVM

SVM is a supervised machine learning model that is optimally used for classification data and regression analysis. It does this by generating the most optimal hyperplane that separates the classified data using support vectors as the most critical decision factors

Results

The results that we found after running the models and validation was that both the Logistic and SVM models overall did very well and performed the best compared to the other models. SVM did slightly better with more accuracy as overall the accuracy was above 90% for most regions that were analyzed. Model coefficients also show that males seem to be marginally affecting fatality more than females as there are more weight on its coefficient when predicting the outcome of the patients.

Accuracy Scores Summary Graphs



Discussion

High survival and low fatality rates contribute to higher accuracy of the models

Other variables/factors also affect fatality (comorbidities)

SVM can be used to predict classifications (in our case, fatal or non-fatal cases) while Ridge/Lasso Regression can only predict numerical values and linear relationships between two variables

Acknowledgement

We would like to acknowledge and thank our mentor Nicole Zhang for making this project possible.

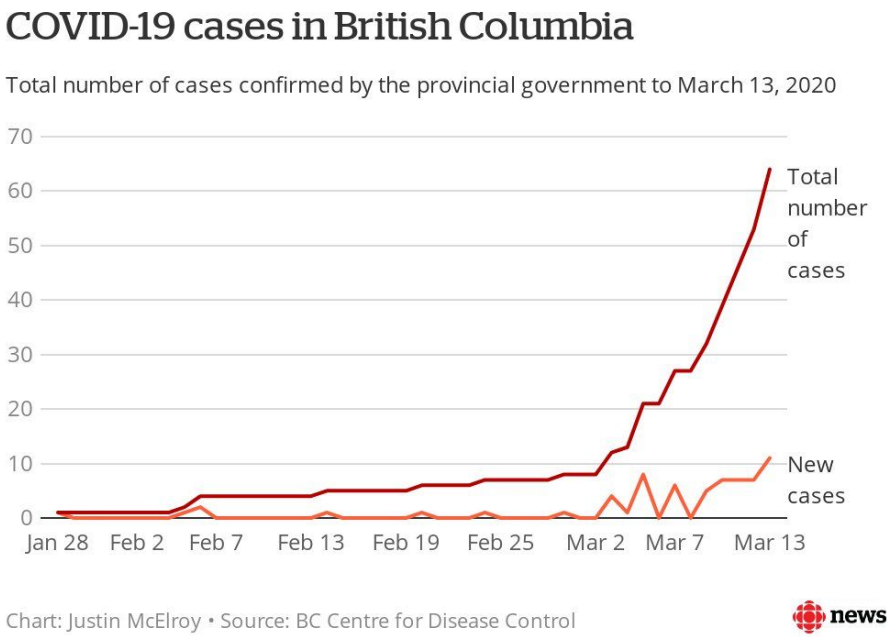
Conclusion

SVM was the best model overall on average compared to the other models. Some implications found in this study is that males are more susceptible than females when it comes to the fatality of COVID-19. This information can also contribute to determine suitable preventive measures and treatments for those who are predicted to be more susceptible. Moving forward from this study, it would be interesting to investigate other factors that could contribute such as the comorbidities of the patients, workplace environment, and possibly more. We can also try applying other machine learning models, focusing mainly on classification models such as random forest, and Naive bayes to find any better models that can predict the outcomes with better accuracy. Lastly, we can also look into statistical analysis on the significance of the risk factors that are contributing.

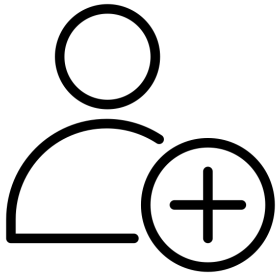
Reference / Bibliography

1. Hawkins, D. (2020). Differential occupational risk for COVID-19 and other infection exposure according to race and ethnicity. American Journal of Industrial Medicine, 63(9), 817-820. doi:10.1002/ajim.23145
2. Lien A, Edjoc R, Atchessi N, Abalos C, Gabrani-Juma I, Heisz M. (2020). COVID-19 and the increasing need for sex-disaggregated mortality data in Canada and worldwide. Canada Communicable disease report, 46(7/8), 231–5. doi:10.14745/ccdr.v46i78a03
3. Mehta, M., Julaiti, J., Griffin, P., & Kumara, S. (2020). Early Stage Machine Learning–Based Prediction of US County Vulnerability to the COVID-19 Pandemic: Machine Learning Approach (Preprint). JMIR Public Health Surveill, 6(3), e19446. doi:10.2196/19446
4. Pan, A., & Wu, T. (2020). Wuhan COVID-19 data – An example to show the importance of public health interventions to fight against the pandemic. Toxicology, 441, 152523. doi:10.1016/j.tox.2020.152523
5. Yadav, M., Perumal, M., & Srinivas, M. (2020). Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos, solitons, and fractals, 139, 110050. https://doi.org/10.1016/j.chaos.2020.110050
6. O'Brien, J., Du, K. Y., & Peng, C. (2020). Incidence, clinical features, and outcomes of COVID-19 in Canada: Impact of sex and age. Journal of Ovarian Research, 13(1). doi:10.1186/s13048-020-00734-4

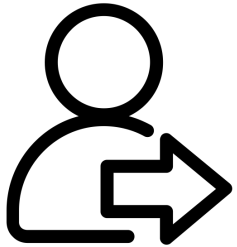
Introduction



Introduction



737 New Cases



5207 Cases

Introduction

Previous Studies:

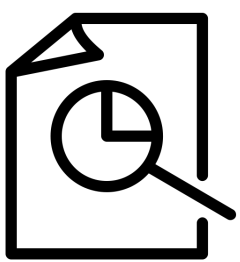
- Elderly more likely to experience severe symptoms (Pan et al., 2020)
- Females less susceptible to infection (Lien et al., 2020; Pan et. Al, 2020)
- Ethnic minority groups at higher risk of symptom severity and death (Hawkins, 2020)
- Machine learning effective for analyzing COVID-19 data (Mehta et. al., 2020; Yadav et. al., 2020)

Our Goal: Predict fatality rate with predictors of Sex and Age using ML

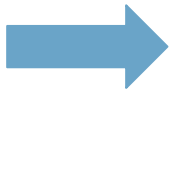
Introduction

- Learn to use Python for data processing
- Grasp a basic knowledge of Machine Learning models
- Test ML on a relevant topic of research

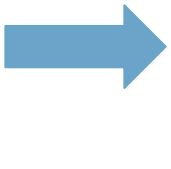
Methods

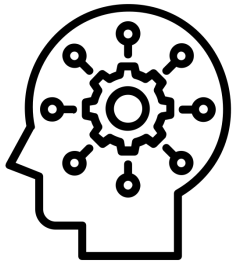


Data Collection



Data Cleaning and Reformatting





Machine Learning

Machine Learning Models

- Ridge Regression
- Lasso Regression
- Logistic Regression
- Support Vector Machine (SVM) Classification

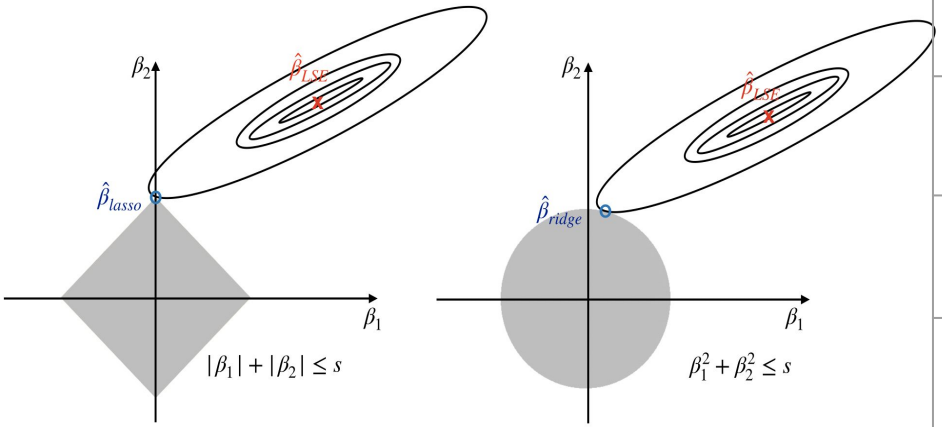
Validation Method

Train/Test Split

a. Splits data into two chunks of specific size
b. 1 chunk for train, 1 chunk for test
c. Trains data (regression) then tests for accuracy

Ridge and Lasso Regression

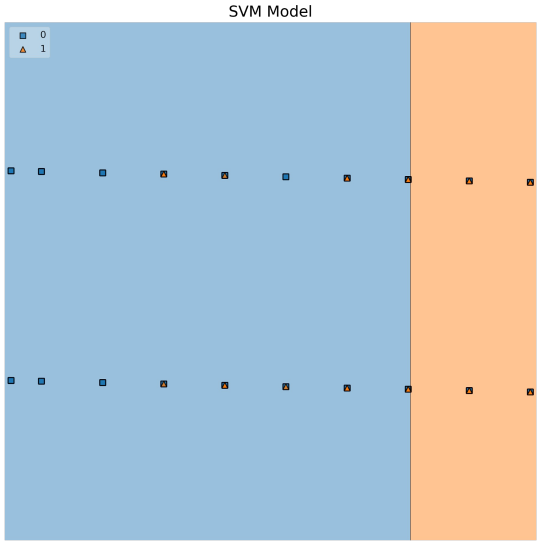
- Both Ridge Regression and Lasso Regression are extensions of linear regression that adds a penalty to the loss function during training.
- Used to determine if a factor plays an important role in the outcomes of the cases.
- Found that males are more susceptible to fatality than females



Accuracy Scores	Ridge	Lasso
Alberta	0.051	< 0.001
Ontario	0.103	0.095
Japan	0.154	0.045

Logistic Regression/SVM

- Classification models that can predict outcomes instead of probabilities of the outcomes
- Accuracy improved by almost 9 folds compared to Ridge and Lasso regression



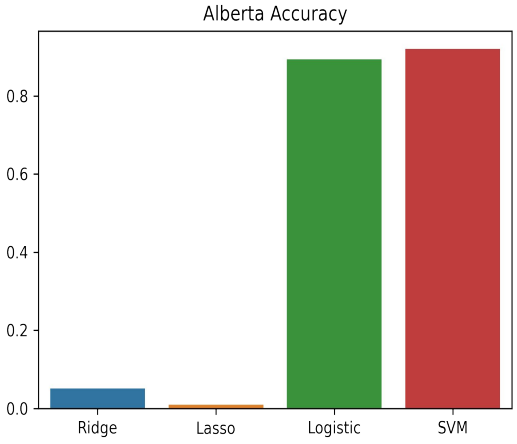
Accuracy Scores	Logistic	SVM
Alberta	0.893	0.920
Ontario	0.852	0.851
Japan	0.919	0.922

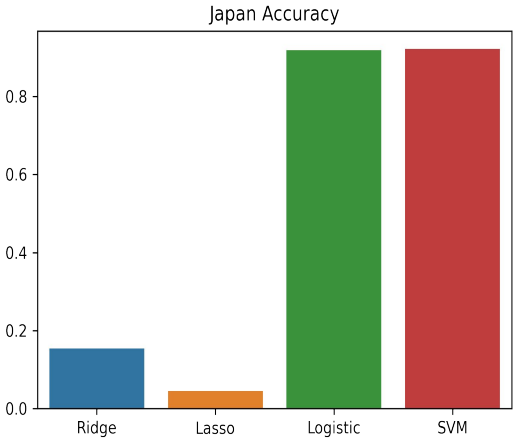
Results Summary

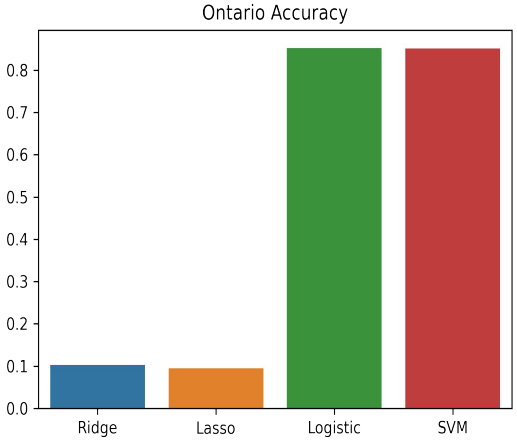
- SVM is the best model (out of SVM, Ridge Regression, Lasso Regression, Logistic Regression) in predicting fatality of COVID patients

Accuracy Scores	Ridge	Lasso	Logistic	SVM
Alberta	0.051	< 0.001	0.893	0.920
Ontario	0.103	0.095	0.852	0.851
Japan	0.154	0.045	0.919	0.922

Results Summary Graphs







Discussion

- Things to consider:
  - High survival and low fatality rates, which contributes to higher accuracy of the models
  - Other variables/factors also affect fatality (ex. comorbidities)
  - SVM can be used to predict classifications (in our case, fatal or non-fatal cases) while Ridge/Lasso Regression (the model that performs the worst) can only predict numerical values and linear relationships between two variables.
  - There is interest in this topic - a study in the journal of Ovarian Research (O'Brien, Du, & Peng, 2020) using statistical methods found that female patients have lower case fatality rates than male patients.

Challenges

- Limited access to applicable datasets
- Limited access to capable hardware
- Limited literature available on topic
- Beginners to python and machine learning

Conclusion

- SVM was the best model overall on average
- Implications:
    - Males are more susceptible than females
    - Applications of suitable preventive measures and treatment
- The next step:
    - More research using additional risk factors (comorbidities)
    - Use other machine learning models
        - Decision Trees
        - Random Forest
        - Naive Bayes
    - Statistical analysis
        - ANOVA

University of British Columbia

Predicting the fatality of COVID-19 using machine learning models

Daniel Choi, Ngoc Bui, Yi Shan

Abstract

The sudden spread of COVID-19 has confounded many researchers and the determination of contraction trends with demographic predictors is a topic of interest. The aim of our study was to provide additional insight towards the outcome of patients that contract the virus. This information could contribute to the determination of suitable preventive measures and treatments for those who are predicted to be more susceptible. Previously there have been a number of different models derived from machine learning methods in order to predict the trends in the growth and contraction of the disease. However, less research has been done to predict fatality after contraction of COVID-19 depending on individual characteristics of the patients such as age and gender, which is what our study aims to develop. We used data from Alberta, Ontario, and Japan to predict the fatality of individuals who contracted COVID-19 and determined the accuracy of different prediction models. More specifically, we applied Ridge Regression, Lasso Regression, Logistic Regression, and Support Vector Machine (SVM) Regression to each dataset and subsequently validated the regression models through train test split with 40% of the data. The results showed that SVM performed the best for all three datasets with 92.0% accuracy for Alberta, 85.1% accuracy for Ontario, and 92.2% accuracy for Japan.

Introduction

Statistics and machine learning have been used for epidemiology and related applications. In the COVID-19 research field, previous studies have discovered trends in infection rates, fatality rates, and symptom severity with relation to age, sex, income, ethnicity, and the like [1,2,3,4,5]

Methods - Machine Learning

Ridge and Lasso Regression

Extensions of linear regression.
Determine whether a particular factor has a stronger effect on the outcome of the case

Logistic Regression

Logistic regression is a statistical model that is often used as a classification model that can classify two different possible outcomes using a logistic function. The logistic regression is actually an extension of linear regression but focuses more on classification problems rather than predicting probabilities of the outcomes.

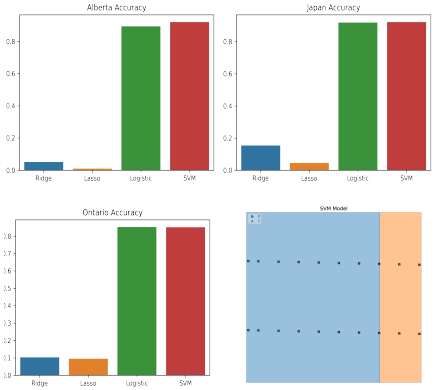
SVM

SVM is a supervised machine learning model that is optimally used for classification data and regression analysis. It does this by generating the most optimal hyperplane that separates the classified data using support vectors as the most critical decision factors

Results

The results that we found after running the models and validation was that both the Logistic and SVM models overall did very well and performed the best compared to the other models. SVM did slightly better with more accuracy as overall the accuracy was above 90% for most regions that were analyzed. Model coefficients also show that males seem to be marginally affecting fatality more than females as there are more weight on its coefficient when predicting the outcome of the patients.

Accuracy Scores Summary Graphs



Discussion

High survival and low fatality rates contribute to higher accuracy of the models

Other variables/factors also affect fatality (comorbidities)

SVM can be used to predict classifications (in our case, fatal or non-fatal cases) while Ridge/Lasso Regression can only predict numerical values and linear relationships between two variables

Acknowledgement

We would like to acknowledge and thank our mentor Nicole Zhang for making this project possible.

Conclusion

SVM was the best model overall on average compared to the other models. Some implications found in this study is that males are more susceptible than females when it comes to the fatality of COVID-19. This information can also contribute to determine suitable preventive measures and treatments for those who are predicted to be more susceptible. Moving forward from this study, it would be interesting to investigate other factors that could contribute such as the comorbidities of the patients, workplace environment, and possibly more. We can also try applying other machine learning models, focusing mainly on classification models such as random forest, and Naive bayes to find any better models that can predict the outcomes with better accuracy. Lastly, we can also look into statistical analysis on the significance of the risk factors that are contributing.

Reference / Bibliography

1. Hawkins, D. (2020). Differential occupational risk for COVID-19 and other infection exposure according to race and ethnicity. American Journal of Industrial Medicine, 63(9), 817-820. doi:10.1002/ajim.23145
2. Lien A, Edjoc R, Atchessi N, Abalos C, Gabrani-Juma I, Heisz M. (2020). COVID-19 and the increasing need for sex-disaggregated mortality data in Canada and worldwide. Canada Communicable disease report, 46(7/8), 231–5. doi:10.14745/ccdr.v46i78a03
3. Mehta, M., Julaiti, J., Griffin, P., & Kumara, S. (2020). Early Stage Machine Learning–Based Prediction of US County Vulnerability to the COVID-19 Pandemic: Machine Learning Approach (Preprint). JMIR Public Health Surveill, 6(3), e19446. doi:10.2196/19446
4. Pan, A., & Wu, T. (2020). Wuhan COVID-19 data – An example to show the importance of public health interventions to fight against the pandemic. Toxicology, 441, 152523. doi:10.1016/j.tox.2020.152523
5. Yadav, M., Perumal, M., & Srinivas, M. (2020). Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos, solitons, and fractals, 139, 110050. https://doi.org/10.1016/j.chaos.2020.110050
6. O'Brien, J., Du, K. Y., & Peng, C. (2020). Incidence, clinical features, and outcomes of COVID-19 in Canada: Impact of sex and age. Journal of Ovarian Research, 13(1). doi:10.1186/s13048-020-00734-4

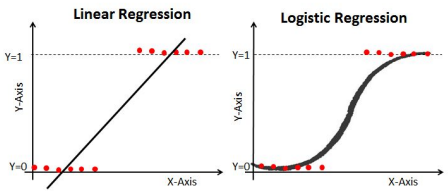
Appendix

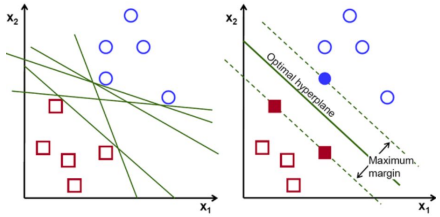
Ridge and Lasso Regression

- The Ridge Regression adds a penalty based on the sum of the squared coefficient values (L2) while the penalty of Lasso Regression model is based on the absolute value of the magnitude of the coefficients (L1).

Logistic:

- Extension of linear regression but uses a logistic function and focuses on classification problems between two possible outcomes
- Clearly separates the possible outcomes that is labelled as either 0 or 1





SVM

- A classification model that is very powerful with very little data.
- Can also be used in higher dimensions with multiple factors.