Predicting Covid-19 Infected Patients Recovery

Classification models

Alexander Berg
IT-security and software testing
Dalarna University
Borlänge, Sweden
h17alebe@du.se

Josefin Paananen
IT-security and software testing
Dalarna University
Borlänge, Sweden
h17jospa@du.se

Abstract: **Covid-19 was confirmed a global pandemic in March 2020. A global social and economic disruption was caused by the pandemic, which includes the largest global recession since the Great Depression. This study aims to predict Covid-19 patients' recovery by using data mining models. With the help of Python, predictive classification models were created. The results revealed that the model developed with Random Forest algorithm is the most efficient predicting the recovery of infected patients with an accuracy of 95.99%. Ultimately older people and contact with a high number of other individuals proved to be major factors and at great risk of developing complications from Covid-19 which could lead to death.**

I. Introduction

Covid-19 is believed to be originated from Wuhan in the Hubei Province, China. The available evidence suggests that the virus has a zoonotic [1] source. Considering there is a limited close contact between humans and bats, it is more likely that the virus transmitted through intermediate animal hosts such as raccoons. However, this is yet to be identified. (World Health Organization, 2020) Other theories point at the live animal markets being the site of a superspreader event where one person infected many other people. (Letzter, 2020) What started as an epidemic mainly limited to China has turned into a global problem having claimed over 1.1 million lives and close to 44 million infected around the world as of the 27th October 2020. (Worldometer, 2020)

Covid-19 causes respiratory illnesses of varying severity. Currently there is no clinically proven vaccine and many already existing drugs such as Hydroxychloroquine (malaria medication) and antibiotics have been tested on patients without any clinically proven success. (World Health Organization, 2020) To protect each other the World Health Organization urges us to keep a safe distance to others, stay at home as much as possible, to wash your hands often and to cover when coughing/sneezing.

II. Problem definition

The study was conducted to get a better understanding what features determine whether a Covid-19 infected patient needs further medical attention or if they will succumb to the virus. Various classification techniques could eventually mitigate the load on the healthcare systems shoulders.

III. Hypothesis

By using classification models, it is possible to predict a COVID-19 patients' recovery.

IV. Litterature review

The World Health Organization stated that the group at risk of severe illness are people of high age, people with health conditions, and conditions that affect their immune system. (World Health Organization, 2020) A study suggests that it is possible to use decision-making technologies to handle the virus and help the healthcare organization with a proper suggestion in real-time to avoid spreading. (Vaishya, Javaid, Haleem Khan, & Haleem, 2020) As of now, many researchers are focused on using data science as an aid tool for the COVID-19 pandemic. However, the main topics are about forecasting the spreading therefore previous studies with the same main goal are sparse.

V. terminology

A. Classification techniques

Classification is a supervised machine learning task. Classes are also called targets/labels and is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). To understand how a given input variables relates to the class the classifier utilizes training data. (Sidath Asiri, 2018)

B. Decision Tree

A decision tree is a decision support tool that using a tree-like model of decisions and their feasible consequences, including

[1] Zoonotic disease is an infectious disease cause by a pathogen that has jumped from a non-human to a human. (Wikipedia, 2020)

chance events, cost of resources, and utility. The tree is built by partitioning the data into smaller units until each partition is pure. Noise in the data must be removed hence the decision tree is pruned to improve its performance. (Wikipedia, 2020)

C. Feedforward Neural Network

Commonly seen in its simplest form as a single layer perceptron is a type of ANN (Artificial Neural Network) often used in classification tasks. In this model the information flows forward and there are no feedback connections meaning that information moves only in one direction as opposed to multi-layer perceptron. At the input layer each input is weighted. The inputs are then added together to get a sum of weights. If the produced value is above a set threshold the neuron fires and takes the activated value which is typically 1. (DeepAI, u.d.)

D. Random Forest

Random forest is a supervised learning algorithm that builds a "forest" which are a group of decision trees. Random forest is used for both classification and regression problems. Instead of searching for the most important feature while splitting the node, it searches for the best feature among a random subset of features. This usually results in a better performing model. (Donges, 2020)

E. Logistic Regression

Is a classification algorithm that is used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam. (Pant, 2019)

F. Cross-validation and K-Fold cross-validation

Cross validation techniques are used to test the effectiveness of machine learning models. Based on the model's performance on unseen data we can say whether the model is under-fitting/over-fitting/well generalized. The K-fold technique is generally less biased compared to other methods. To perform a K-Fold cross-validation the data is randomly split into K-folds, after which the model then is fit using the K minus 1 folds and data validated using the remaining K-fold. The process is repeated until every K-fold has been used for a test-set. Finally, the average of recorded scores will present the performance metric for the model. (M, 2018)

G. Feature importance

Is a technique that assign scores to input based on how useful they are at predicting a target variable. The features that are no useful are eliminated to reduce noise in the data which affects the model's accuracy. (Shaikh, 2018)

H. F1 Score

F1 score combines precision and recall to get an overall score that depicts how well the model is performing. The higher the F1 score is the better the predictive power of the classification procedure. A score of 1 means perfect classification whereas 0 is the lowest F1 score possible which points to bad classification. Precision is the number of correct positive results divided by the number of predicted positive results. The recall is the number of correct positive results divided by the number of actual positive results. (Sucky, 2020)

VI. Methodology

A. Dataset

The dataset was pulled from Kaggle[2] and it was collected from KCDC (Korea Centers for Disease Control & Prevention). It contains 5166 entries and 14 variables with information about infected patients. They include patient ID, sex, age, country, province, city, infection case, infected by (which ID the patient was infected by), contact number (number of contact with other people), symptom onset date, confirmed date, released date, deceased date and state (released, deceased, isolated). The timespan of the dataset is from 2020-01-20 to 2020-06-30.

B. Data Processing

Many of the variables had values missing, mainly deceased date, released date, contact number, symptom onset date, and infected by. Pre-processing included last observation forward fill as missing values reduces the prediction power. Isolation state was also excluded from the analysis and only released and deceased patients were considered. Following features were used in the research:

	sex	infection_case	contact_number	state	age_group
0	female	overseas inflow	3	0	50-59
1	male	Shincheonji Church	3	1	70-79
2	female	Shincheonji Church	3	1	60-69
3	female	Shincheonji Church	3	1	50-59
4	female	Shincheonji Church	3	1	50-59

Features used in the research.

C. Data Modeling

After the data processing steps the models were trained. All models were fit with the training data. The data was split into 70% training set (998 samples) and 30% testing set (429 samples).
A tenfold cross-validation was used to ensure the models robustness.
The default scores for each model:

- Decision Tree scored 97.49

[2] Kaggle is a website where data scientist and machine learning practitioners can find and publish data sets. (Kaggle, 2020)

- Random Forest scored 97.49
- Logistic Regression scored 96.49
- Feedforward Neural Network 96.49

The top models with the highest score were Decision Tree and Random Forest.

The next step was to perform a feature importance and for the features that had zero importance were dropped. Then another ten-fold cross-validation was done.

Cross-Validation Mean Score for each model:

- Random Forest 95.99%
- Decision Tree 94.40%

F1 scores for each model:

- Random Forest 97.72%
- Decision Tree 97.07%

VII. Analysis

A. *Data Visulatizaton of the dataset*

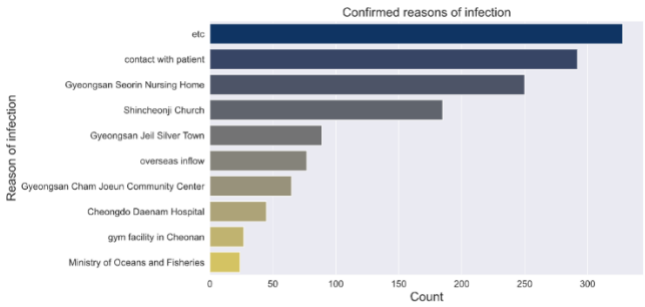

Figure 1 confirmed reasons of infection.

The most common reason of infection were other causes. (etc is other causes). The second most common infection reason is due to contact with other patients.

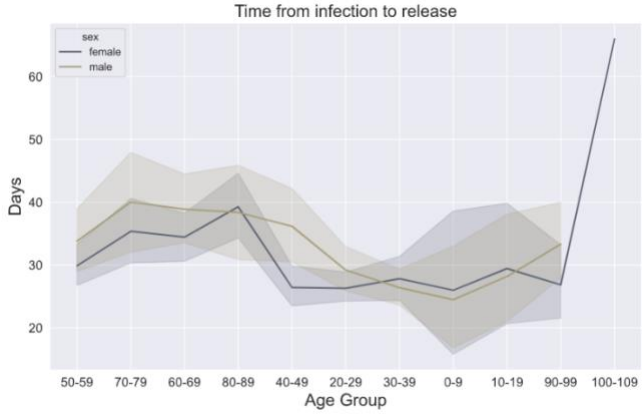

Figure 2 Time from infection to release.

Time from when confirmed infection until release from isolation. Shadows displays a 95% confidence interval.

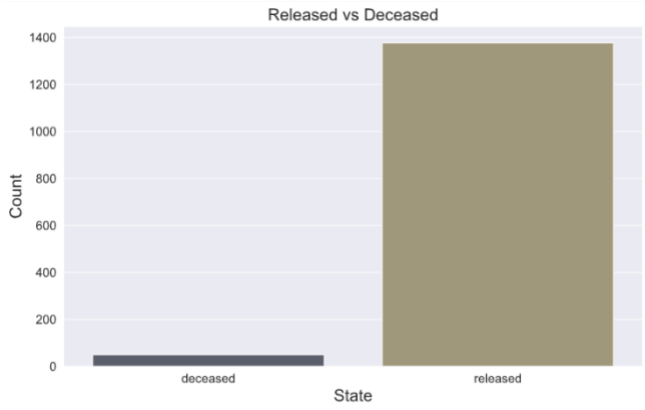

Figure 3 Released vs Deceased

A far greater number of people survived the infection.

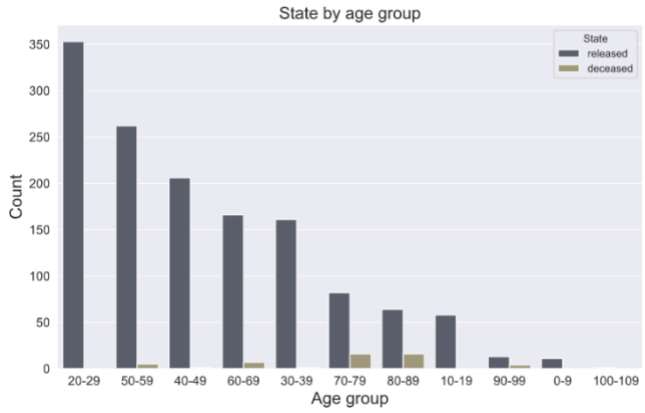

Figure 4 State by age group

Majority of released patients were between 20 and 29 years old while highest number of deaths was among individuals from 80 to 89 years old.

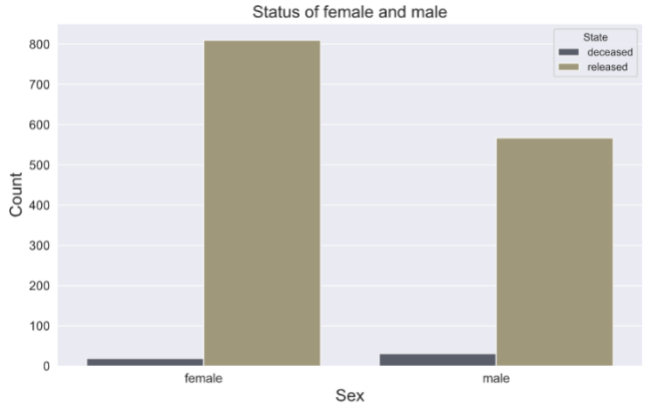

Figure 5 Status of female and male

A higher number of women were released compared to men, and there were more men than women that did not survive.

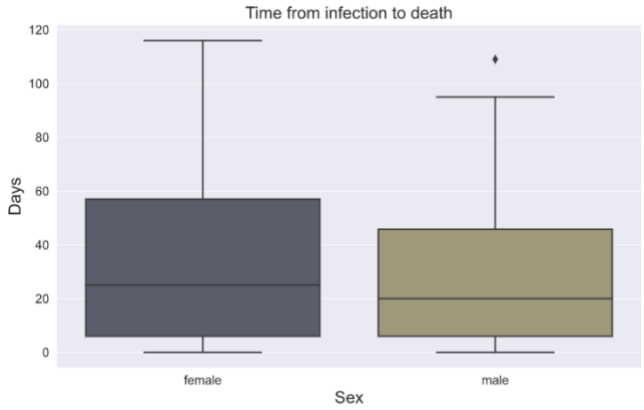

Figure 6 Boxplot Time from infection to death (only sex and days)

Figure reveals that there is a greater variability for females than males suggesting that males die sooner.

Figure 7 Time from infection to death (days, age, and sex)

This figure shows us the difference in age group and sex from the time from infection to death. Shadows displays a 95% confidence interval.

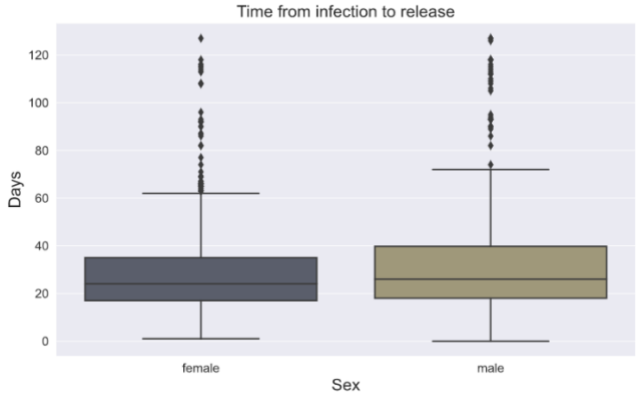

Figure 8 Boxplot of Time from infection to release (only sex and days)

There is a slightly greater variability for males than females however the duration from infection to release from isolation are similar.

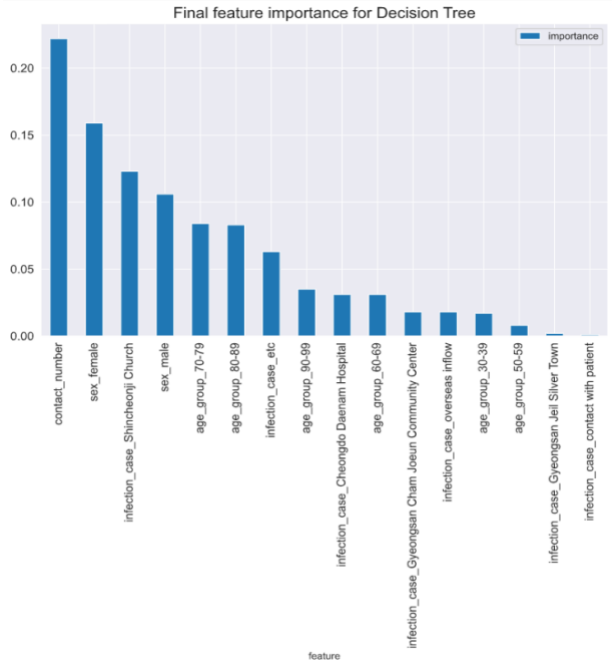

Figure 9 Important features for Decision Tree

Significant features are contact with other people, the sex and old age.

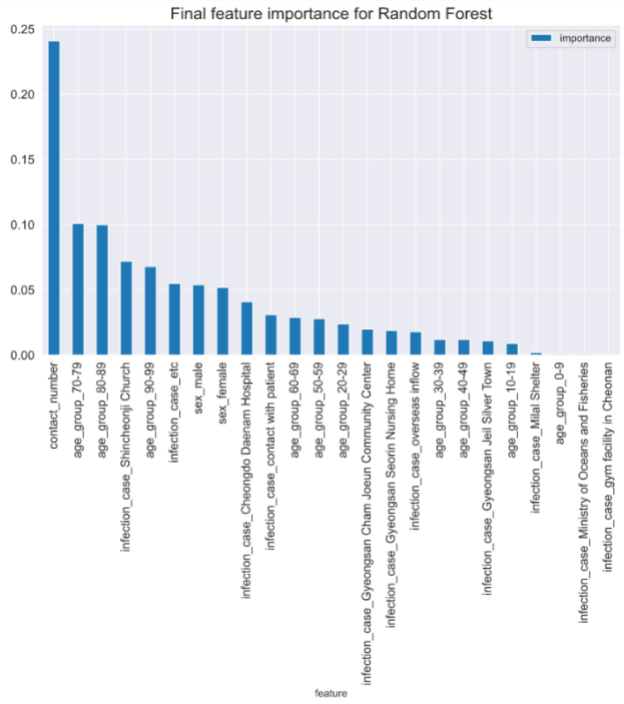

Figure 10 Important features for Random Forest

Significant features are contact with other people and old age.

VIII. Discussion

The predictive models suggest that the most contributing factor for a patients' recovery is how many people they have been in contact with. Following most important factor is their age.

After doing a performance evaluation model, Random Forest had the highest accuracy. But the other models have a high enough accuracy as well. This research proves that it is possible to use classification models to predict a patients' recovery.

The dataset in use for the models is biased due to the last observation forward fill. However, if the missing values would've been removed the models' algorithm would have had nothing to work with, and with that, the classification models would have had a lower accuracy.

Suggestions for future work would be to use a dataset with fewer missing values. To bring this study further, a time series analysis in combination with a geographic clustering would be useful for the government and healthcare since they would be able to manage their resources accordingly.

The World Health Organization urged us to keep a safe distance, which this study can confirm since close contact with many people had the highest infection rate.

IX. Conclusions

In this study, four different data mining models have been developed to predict Covid-19 infected patients' recovery using an epidemiological dataset of Covid-19 patients of South Korea. The classification models used are Feedforward Neural Network, Logistic Regression, Decision Tree, and Random Forest. The model developed with Random Forest was the most accurate model to predict patients' recovery with an f1 score of 97.72% and Cross-Validation Mean Score of 95.99%. Decision Tree had an f1 score of 97.07% and a Cross-Validation Mean Score of 94.40%. Initially, all models performed well with high accuracy and it can be concluded that all four models are efficiently capable of predicting the possibility of recovery of patients infected with Covid-19. This confirms the reports hypothesis to be true. Ultimately important factors are older people and contact with a high number of other individuals are at great risk of developing complications from Covid-19 which may result in death.

X. References

DeepAI. (n.d.). *What is a Feed Forward Neural Network?* Retrieved from DeepAI: https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network

Donges, N. (2020, September 3). *A complete guide to the random forest algorithm*. Retrieved from Builtin: https://builtin.com/data-science/random-forest-algorithm

Kaggle. (2020, 10 26). *Kaggle*. Retrieved from https://www.kaggle.com/

Letzter, R. (2020, May 28). *LiveScience*. Retrieved from https://www.livescience.com/covid-19-did-not-start-at-wuhan-wet-market.html

M, S. (2018, November 13). *Why and how to Cross Validate a Model?* Retrieved from Towards Data Science: https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

Pant, A. (2019, January 22). *Introduction to Logistic Regression*. Retrieved from Towards Data Science: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

Shaikh, R. (2018, October 28). *Feature Selection Techniques in Machine Learning with Python*. Retrieved from Towards Data Science: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Sidath Asiri. (2018, June 11). *Machine Learning Classifiers*. Retrieved from Towards Data Science: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

Sucky, R. N. (2020, September 12). *A Complete Understanding of Precision, Recall, and F Score Concepts*. Retrieved from Towards Data Science: https://towardsdatascience.com/a-complete-understanding-of-precision-recall-and-f-score-concepts-23dc44defef6

Vaishya, R., Javaid, M., Haleem Khan, I., & Haleem, A. (2020). *Artificial Intelligence (AI) applications for COVID-19 pandemic.* New Delhi: Elsevier Ltd.

Wikipedia. (2020, October 13). *Decision Tree*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Decision_tree

Wikipedia. (2020, October 20). *Support Vector Machine*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Support_vector_machine

Wikipedia. (2020, October 19). *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Zoonosis

World Health Organization. (2020). *Coronavirus disease 2019 (COVID-19) - Situation report - 94.* Genève: World Health Organization.

World Health Organization. (2020, October 28). *COVID-19: vulnerable and high risk groups*. Retrieved from World Health Organization: https://www.who.int/westernpacific/emergencies/covid-19/information/high-risk-groups/

World Health Organization. (2020, October 12). *Q&As on COVID-19 and related health topics*. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub

Worldometer. (2020, October 26). Retrieved from https://www.worldometers.info/coronavirus/