Data Mining Coursework - 2

Thomas Tasioulis - 998273

April 2020

1 Introduction

Coronavirus, or COVID-19 as it is officially called was declared a pandemic disease by World Health Organization (WHO) on December 30 2019. By February 28, 2020 more than 82,000 cases were confirmed [6] in more than 26 different countries and this situation is still ongoing for many countries around the globe with severe implications to public health. At the current stage of COVID-19 we are still dealing with everyday new cases and deceases. Thus, there is an urgent need for active countermeasures [3, 4] for all countries in order to both contain the spread and secondly to reach "flattening the curve" [5] of health care systems in order to be more efficient and effective during this process; the appropriate "flattening" is crucial for the reduction of the consequences' magnitude and can make the virus more manageable by each country's healthcare system.

In this paper we have analyzed a dataset from the Korea Centers for Disease Control Prevention (KCDC) [2] in order to perform data mining techniques and to extract significant knowledge around patients, locations and places that the local governments should be excessively concerned. Hence, the aim of the research was to answer research questions as "Which activities emerge higher risk for subsequent spread of infection?", "Which provinces within Korea are the most popular to travel?" and "Which factors indicate higher probability of a perish, or will lead to isolation of them?".

2 Proposed Method

Dataset Description

The original dataset [2] is consisted of 12 files each one with different number of columns. For our analyis we have used only 3 of 12 files, "*Case*", "*Patient Info*" and "*Patient Route*". "*Case*" file is consisted of 8 columns, "*Patient Info*" of 18 columns and "*Patient Route*" of 8 columns while there is a direct connection between those files based on different columns among those files. Withing those files there is information such as Province, Country, Province destination, Infection Case (where did the infection took place), state of patient (released, isolated or deceased), age group and many others.

Network analysis

In order to identify locations and activities which led to higher spread we performed a network analysis to each infection case. More specifically, we isolated data from each province and we created a network for each one based on the confirmed cases for those locations. The separation between places of high interest and low interest we used Fruchterman and Reingold [1] (a physics-like simulation) algorithm such as strong connections are closer to the nodes and weak connections are far away from the nodes. Furthermore, we used different size of edges within our visualization, After a normalization in the confirmed cases for each activity we applied full-solid lines to places with more than 25% of the cases while in the other cases we have used dashed lines to visualize.

During the network analysis we performed normalization to the factors that were measured. At each case, the normalization has been done with formula of equation (1) in order to set values between [0-1] and thus, they could be compared more easily.

$$\hat{x} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Network for patient routes

We also performed a network analysis for the patient routes before their admission/confirmation of the case. At this case we measured the number of routes from each province to any other and we created 3 different scales for the routes; those with frequency less or equal to 10% of the total routes, those which are in between 10% and 30% of the total and those with frequency more than 30%. Then we created a network in order to find patterns and visualize the potential danger areas where the government should apply the restrictions.

3 Results and Discussion

3.1 Descriptive Visualizations

Firstly we initialized some basic visualizations in order to identify which age groups are mostly in danger from COVID-19 and how the pandemic was spread throughout all this period in South Korea.

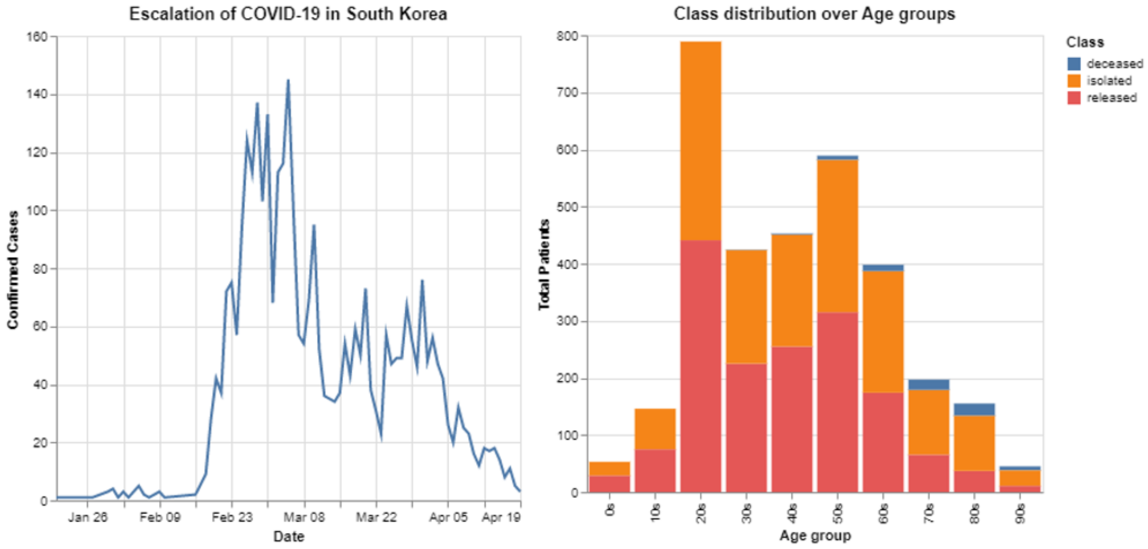

Figure 1: Left Image: Trendline of COVID-19 cases in South Korea, Right Image: Distribution of "State" between different age groups

As Figure 1 indicates the disease started to have epidemic form between 9 to 23 February similarly to many other countries at that time. Moreover, we see two huge peaks between February 23 and March 10 when there were almost 140 new cases each day. However, the possible restrictions that government took seem to have an impact in the spread of the disease as it is observed a steep decrease in the cases after March 8 and eventually on April 19 the new cases reached less than 10 daily. Regarding to the class distribution it is observed that people who perish are statistically above the age of 50, even though the percentage of people who generally perish are quite small compared to the total population. Thus, it would be safe to say that Korean's healthcare system seem to work quite efficiently and the higher risk is on people who belong to be the middle age and above (50+) compared to the younger ones.

3.2 Network Analysis

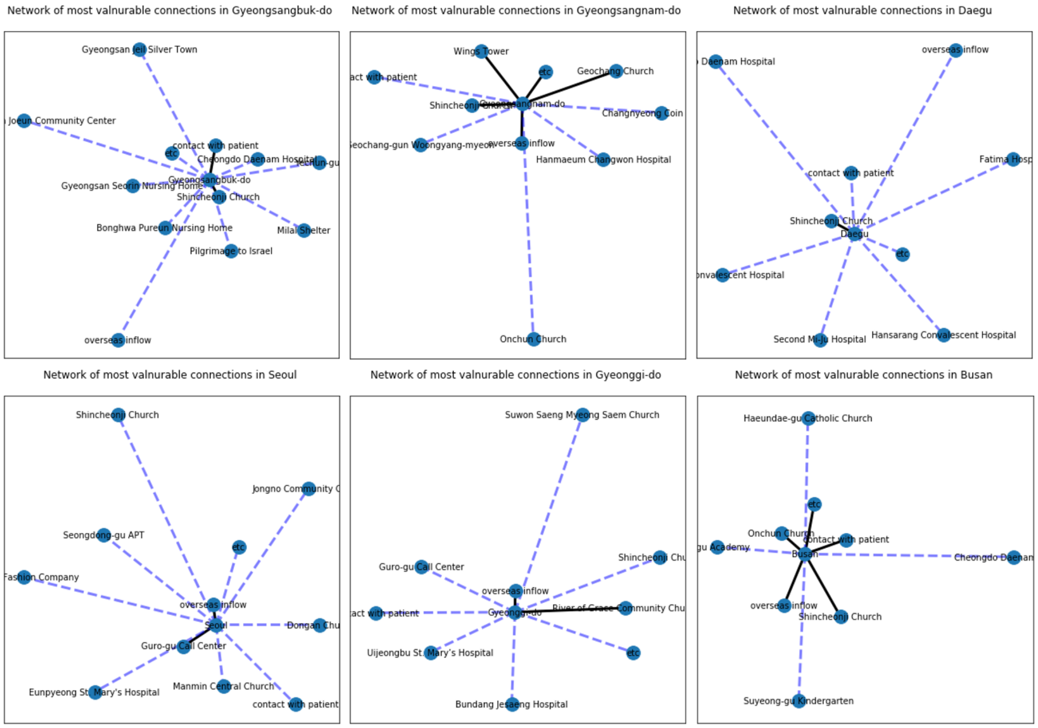

Figure 2: Network analysis for 6 provinces in South Korea regarding to activities that led to confirmed cases. Full width represents more than 25% of cases due to the respective activity while dashed lines less than 25%

From Figure 2 we can reasonably extract some key places that government should probably impose a lock down until COVID-19 fades away. As Figure 2 illustrates the common pattern of all six graphs are places of religious reasons (Church, Catholic Church etc.), overseas inflows and contact with patients (such as hospitals or medical clinics). Thus, if measures are going to be applied those figures can help to a more strategic choice of which places are considered the most dangerous. In the case of overseas inflows we have also observed that most of countries worldwide have imposed a full cancellation to flights from and to domestic land. Furthermore, we observe that Seoul which is the capital of South Korea has the least strong connections. However, this seems reasonable due to Seoul is one of the predominant places with cases and such it is quite difficult for any other place-activity to breach the 25% threshold. Lastly, there are additional networks for smaller provinces which are not displayed in this section and they are included in the demo code file.

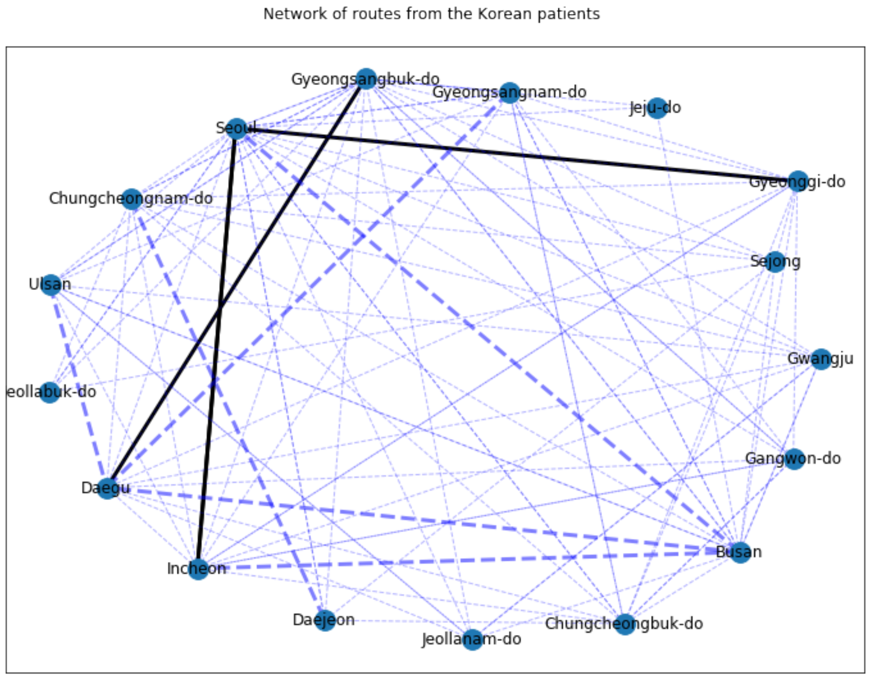

Figure 3: Network analysis of patients' routes. Full width lines represent more than 30% of total routes, semi dashed lines between 10% and 30% and small dashed lines lower than 10%

Figure 3 presents the network of patients' routes within the country of South Korea; by this network analysis we extract crucial information regarding to which places might be imposed into a lock down and maybe the ranking of them could mean how sooner or later this lock down should be imposed. In Figure 3 is shown that most patient trips became from and to Seoul, Daegu, Gyeongsangbuk-do, Gyeonggi-do and Incheon. Thus, those locations are crucial to be protected and stay isolated as the most patients have their routes from and to them. The secondary routes are from smaller cities compared to the major ones before. However, those trips might be as dangerous as the former due to they might become alternative destinations subsequently.

3.3 Factor Estimation

For the factor estimation we have used a supervised learning technique and more specifically Random Forest. Our model is consisted of maximum 5 nodes per tree and by 1000 trees. Our labeling system was the state of the patient as "Released, Isolated and Deceased" as they presented in

Figure 1. In order to measure each factor we split the dataset into train and test set by removing all rows that had "*Nan*" at any cell that we were interested in. The split was done under 2/3 train to 1/3 test rule and our key features were: "Province, Age, Infected by other, Existence of previous underlying disease, Existence of symptoms, Sex and Country". Thus, we trained our model and we measured the importance of factors on both train and test set. Results are shown in Figure 4.

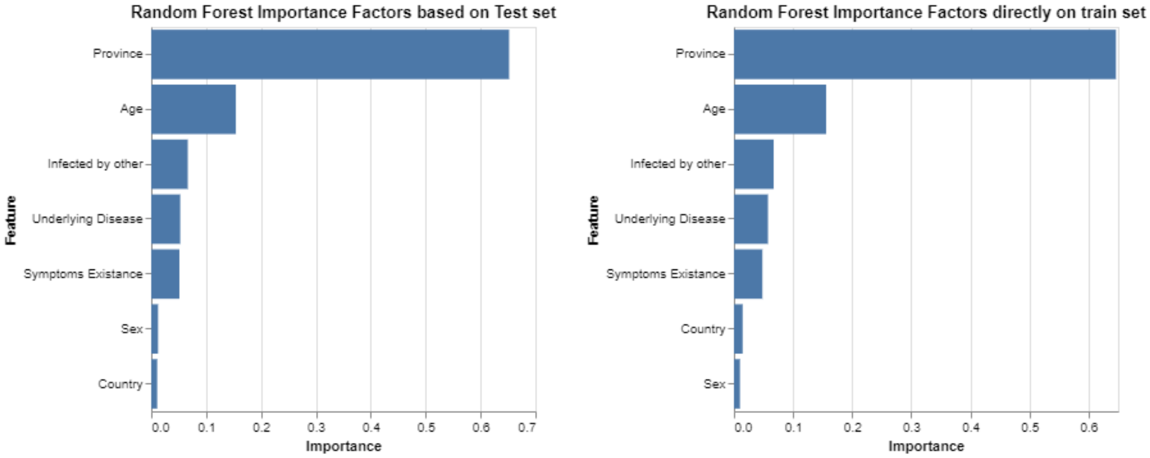

Figure 4: Factors estimation by using random forest on both Train and Test set

Hence the accuracy of the model was 72.40% on the test set and 74.20% on the train set, indicating a consistency of our model without overfitting the train set extensively. The estimations of the factors as they are shown in Figure 4 are quite the same between train and test set with a slight difference between Sex and Country on train and test, respectively; nevertheless their importance seems insignificant in comparison to the big picture. Therefore, we can observe that province location, age and infection by other member are the top 3 features while underlying disease existence seems to be significant as well. This interpretation makes sense when the problem is located within areas that we know that there is high spread of the infection; however there are limitation if it is going to be used widely as the model is not generic enough for this purpose.

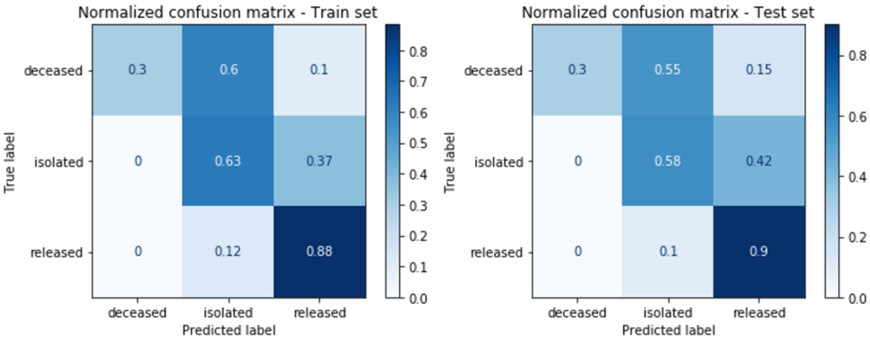

Figure 5: Confusion Matrix for Random Forest in Train and Test set, normalized by class

From the confusion matrix a lot of strenghts and weaknesses of the model can be derived. First and foremost, we can see that regarding to released patients the model has a decent capability to distinct them between released and isolated patients. At any case there is 100% accuracy between this and deceased which means that there will be no case which a released person might perish. However, in the isolated class there is a decent error rate (almost 40%) which means that there is 40% probability to release and a patient that should be isolated. Moreover, to that deceased class seems to have a large error rate as well (close to 70%). However, the actual error rate could be only 10% as in the worst scenario the patient would be isolated (60% probability) or classified as dead.

4 Conclusion

Overall, an explicit analysis has been done for a famous COVID-19 dataset related to South Korea [2]. The analysis has shown key places that the government should take strongly into consideration as potential infection places. Additionally, from the network of patients' routes there were patterns that emerged from major provinces within South Korea and probably should be the first places of countermeasures application. Moreover, regarding to patients analysis the extraction of the most significant parameters showed that besides Province, age and infection by other person seemed to be the most severe parameters which might determine the state of each patient and should be the first things under consideration if possible. Lastly, our model showed a decent accuracy rate and its error rate can be interpreted as the lesser evil of our model. Finally, in this pandemic crisis we are all living through, there is an urgent need of research among those issues that will contribute to existing knowledge around the issue and will help the prevention of a potential one.

References

[1] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[2] KCDC. Dataset from korea centers for disease control prevention (KCDC). https://www.kaggle.com/kimjihoo/coronavirusdataset/.

[3] Stephen M Kissler, Christine Tedijanto, Marc Lipsitch, and Yonatan Grad. Social distancing strategies for curbing the covid-19 epidemic. *medRxiv*, 2020.

[4] Joacim Rocklöv, Henrik Sjödin, and Annelies Wilder-Smith. Covid-19 outbreak on the diamond princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. *Journal of travel medicine*, 2020.

[5] Linda Thunstrom, Stephen Newbold, David Finnoff, Madison Ashworth, and Jason F Shogren. The benefits and costs of flattening the curve for covid-19. *Available at SSRN 3561934*, 2020.

[6] Annelies Wilder-Smith, Calvin J Chiew, and Vernon J Lee. Can we contain the covid-19 outbreak with the same measures as for sars? *The Lancet Infectious Diseases*, 2020.