COSMOEM

Final Technical Report

Abdurrakif Ülgü	Ömer Yancı	Muhammed Faruk Yıldırım	Cem Ufuk Yılmaz
181180075	181180076	181180077	181180079

ABSTRACT

COVID-19 virus keeps spreading all around the world and affects our lives in different ways. Scientists are working to prevent the progression of the pandemic in different ways. Data science does not need a high amount of resource thus it is one of the best approaches for research on the pandemic. In three simple steps, collecting, organizing, and analyzing data, new findings can be obtained.

INTRODUCTION

Pandemic is growing day by day and COVID-19 cases keep increasing. Virus affects relatively high amount of people, and the number of hospitalization services may be insufficient. As a result, some cannot access healthcare services and unfortunately die.

To prevent this situation, we can identify people who need healthcare services more by using machine learning methods.

METHODOLOGY

Our main goal is predicting patients that have higher probability of mortality risk.

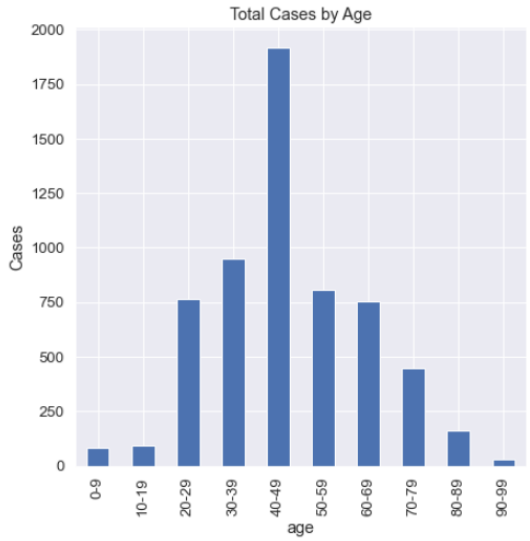

Figure 1 - Cases December 1, 2019 to February 5, 2020

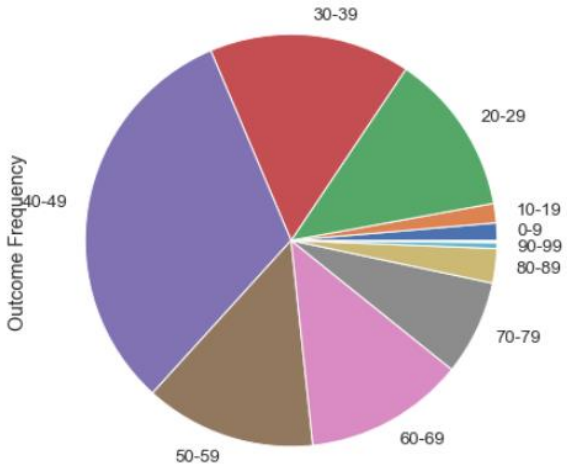

Figure 2 - Covid Mortality Frequency by Age

Hospitalized patients are asked for symptoms and the treatment process is started. These symptoms are stored as individual patient data so they can be used for AI research. Our dataset for individual patient data includes features that may affects mortality rate. Using correlation method, there are eight feature that affects mortality risk: Pneumonia, respiratory distress, septic shock, heart attack, kidney failure, dyspnea, gasp, and chest pain (fig. 3).

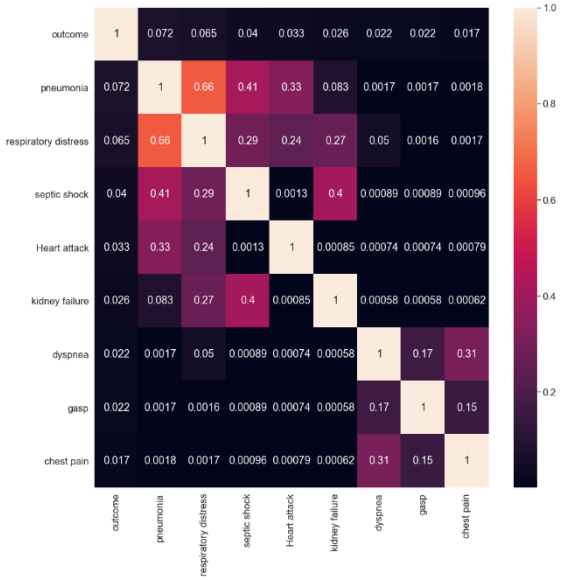

Figure 3 - Heatmap of Mentioned Features

Using these features and rest of them, we trained four ML approaches for predictions: Decision Tree, K-Nearest Neighbors, Logistic Regression and Random Forest Classifier.
First, we shuffled and split our dataset as 0.75 train and 0.25 test. Train part is used for training model and test part used for finding accuracy. We also used 10-fold cross validation for better understanding on models and ROC (Receiver Operating Characteristic) table (fig. 4) for visualizing and model selection. As the area under the curve goes higher, performance of the model is higher too.

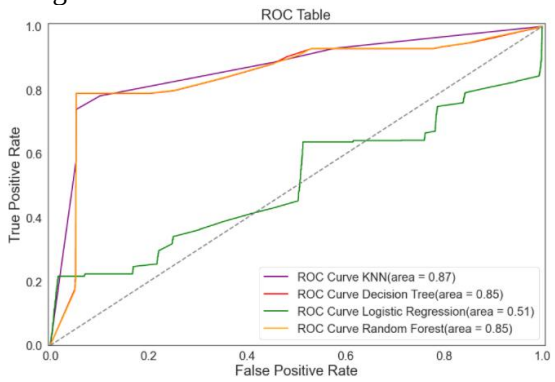

Figure 4 - Receiver Operating Characteristic Table

Decision Tree:

Decision trees are type of supervised learning. Tree keeps splitting continuously as the depth of the tree parameter increases. Our model trained using 9015 patient data with 54 attribute per patient. After training, calculated prediction accuracy on test set is 86.32 percent. Using 10-fold cross validation method, accuracy of model is 0.85 which is accurate enough to make good predictions.

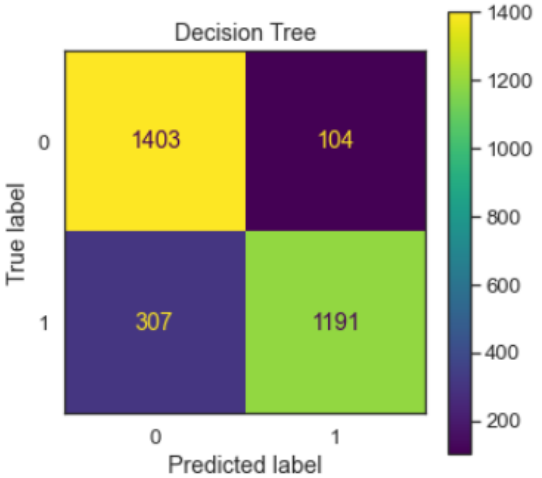

Figure 6 - Confusion Matrix of Decision Tree

K-Nearest Neighbor:

K-Nearest Neighbor, KNN for short, is very simple supervised learning algorithm for classification and regression. After training, calculated prediction accuracy on test set is 86.55 percent little bit higher than the decision tree model. Using 10-fold cross validation method, accuracy of model is 0.87 which is the most accurate method to predict cases.

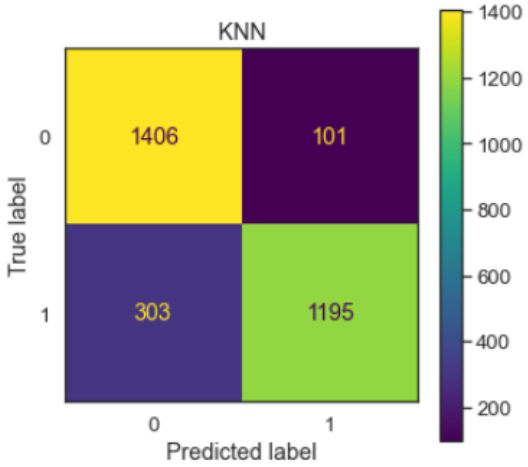

Figure 5 - Confusion Matrix of KNN

Logistic Regression:

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. After training, calculated prediction accuracy on test set is 54.24 which is the worst model. Using 10-fold cross validation method, accuracy of model is 0.51 which is the worst model too. Model performs very poor because we feed around 50 features to model which is not efficient for logistic regression.

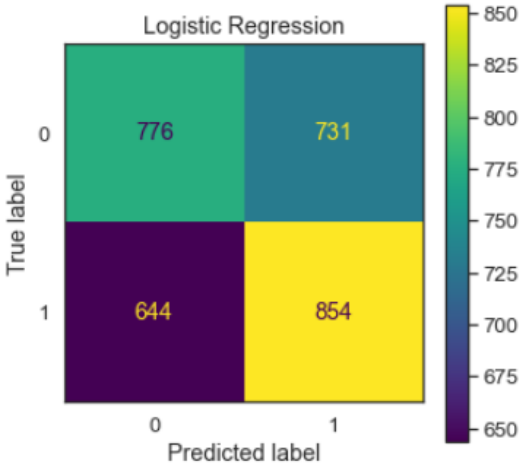

Figure 7 - Confusion Matrix of Logistic Regression

Random Forest

Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. As we do not have many target classes, random forest model performs very similar to decision tree model. Model performs with 86.42 percent accuracy. After testing with cross validation method, performance is 0.85 which is the same value of decision tree model.

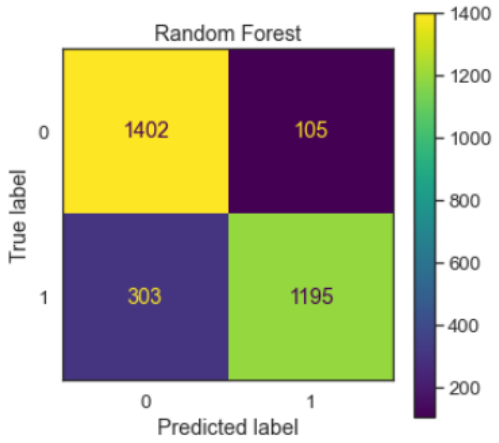

Figure 8 - Confusion Matrix of Random Forest

DATASETS

[1] https://doi.org/10.1038/s41597-020-0448-0

This data set is generally used on an individual basis. In our current data, our members are individuals, and they have other disease states. Thus, we were able to detect the effect of other diseases on Covid 19 with visualizations and machine learning.

[2] https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

Confirmed case numbers are kept in this dataset. Thus, we were able to examine the rate of increase in cases according to days.

[3] https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv

In this data set, the number of people who died due to covid 19 is kept. With this dataset, we were able to examine mortality between countries.

[4] https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv

In this data set, people who have recovered are kept.

[5] https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/08-22-2020.csv

We tried to reach up-to-date data with this data set.

[6] https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports_us/08-22-2020.csv

With this data set, we reached hospital data. This data such as test rate, hospitalization rate, mortality rate are kept.

[7] https://www.kaggle.com/reinoso/covid19-global-analysis/data

In this data set, country-based information is kept. There is information such as total vaccination, type of vaccination.

[8] https://www.kaggle.com/reinoso/covid19-global-analysis/data?select=worldometer_coronavirus_daily_data.csv

In this data set, there is information about the daily number of cases.

EXPERIMENTS

After working on mortality prediction, we focused on the more general topics: Spread of covid by country, vaccinated person by country etc.

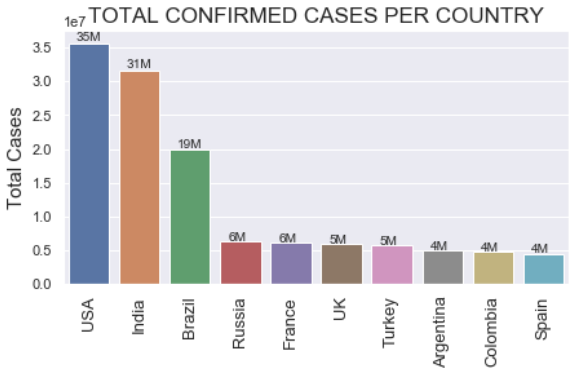

Figure 9 - Total Confirmed Cases per Country

When we look at the table (fig. 9), we see the number of cases in different countries. We realize that the increase in the number of populations significantly affects the number of cases.

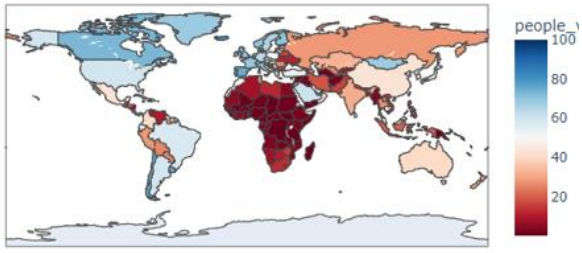

Figure 10 - Percentage of Vaccinated Population per Country

When we look at our image (fig. 10), the vaccination rate is very high in the American and European regions. However, we see that vaccination is low in a certain part of Africa and Asia. We guess that this situation is due to the economic, political, and military power of the countries.

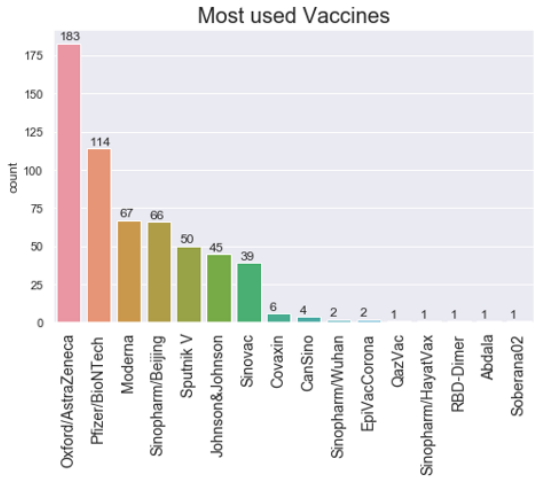

Figure 11 - Most Used Covid Vaccines

In this section, we see the most preferred companies in vaccines, which caused great concern when they were first released.

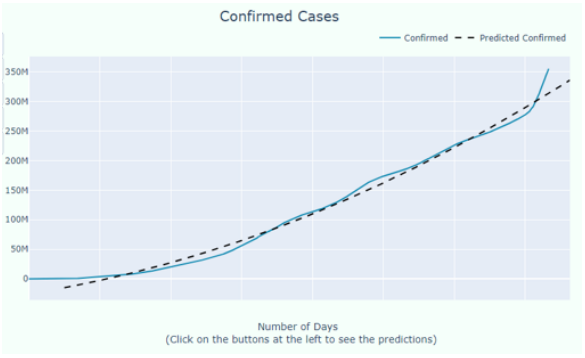

Figure 12 - Confirmed Cases and Prediction

In this part (fig. 12), we tried to make an estimation based on the currently confirmed cases. We used linear regression approach for prediction.

Figure 13 - Mortality Rate After Vaccination

Considering the decrease in mortality rate (fig. 13), we understand that vaccines have a significant impact on COVID-19.

RESULTS

We approached the problem in two stages. First, we found a dataset that includes individual patient data for mortality risk prediction. Then we tried to train a model and predict mortality risk by looking symptoms and disorders. So, patients that have relatively high mortality risk can get priority for medical treatment. We trained 4 models for this prediction and tested them with different approaches. By looking into results (fig. 4), we can say that KNN model has the best outcome. Main problem with the prediction and dataset is dataset includes patient data from early stage of pandemic. That means we do not have any vaccination information and new variants of COVID. Apparently, countries do not share these data publicly. Thus, we had to use this dataset, despite we obtain pretty accurate predictions in the end.

Second approach includes mostly visualizations of spread of covid and vaccinations. After our observations we assumed that vaccines and hygiene rules (social distancing, masks) have huge impact on mortality and infectiousness. We can see that in the countries which vaccination percentage is under 20 percent still has high mortality risk. Countries that have vaccination rate over 80 percent are gaining natural immunity, so some rules and restrictions are being reduced. We also observed that total confirmed cases increasing linearly. In the first stage of pandemic case count started increasing exponentially but as the vaccination started, we obtain more linear line which is promising.

CONCLUSION

After finished the project, we can say that we obtained pretty accurate results and good understanding on data science topics. During development stage, we followed the data science road map. At first, we framed the problem and figure out what we need to observe after our work is done. After that we found datasets and visualized them for better understanding. In this stage it was challenging to find fresh and wide datasets. Then we tried to extract features that we do not need to train model. Then we tried to find which model is best or we need to train model again with different features. By following these steps development process become much more workable in teams.

REFERENCES

[1] Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant applications of machine learning for COVID-19 pandemic. *Journal of Industrial Integration and Management*, *5*(04), 453-479.

[2] Gao, Y., Cai, G. Y., Fang, W., Li, H. Y., Wang, S. Y., Chen, L., ... & Gao, Q. L. (2020). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature communications*, *11*(1), 1-10.

[3] Dasgupta, A., Sun, Y. V, König, I. R., Bailey-Wilson, J. E., & Malley, J. D. (2011). Brief review of regression based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genetic Epidemiology, 35(S1), S5–S11. https://doi.org/10.1002/gepi.20642

[4] Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C. C. Y., Poon, R. W. S., Tsoi, H. W., Lo, S. K. F., Chan, K. H., Poon, V. K. M., Chan, W. M., Ip, J. D., Cai, J. P., Cheng, V. C. C., Chen, H., … Yuen, K. Y. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet, 395(10223), 514–523.

https://doi.org/10.1016/S0140-6736(20)30154-9

[5] Zoumpekas, T. (2020). Clustering World Countries affected by Coronavirus. COVID-19 Cluster Analysis. https://towardsdatascience.com/covid-19-cluster-analysis-405ebbd10049

[6] Leonel, J. (2019). Classification Methods in MachineLearning.

https://medium.com/@jorgesleonel/classification-methods-in-machine-learning-58ce63173db8

[7] Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., ... & Kraemer, M. U. (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific data*, *7*(1), 1-6.

[8] Novel Coronavirus (COVID-19) (healthmap.org)

[9] Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health*, *20*, 100178.

[10] Komorowski, M., Kraemer, M. U., & Brownstein, J. S. (2020). Sharing patient-level real-time COVID-19 data. *The Lancet Digital Health*, *2*(7), e345.