**COVID-19 Data Analysis – Predicting Patient Recovery**

**Arnela Salkić, Nermina Durmić**

*Department of Information Technologies, International Burch University,*

*Faculty of Engineering and Sciences, Sarajevo, Bosnia and Herzegovina*

*All Correspondences to:* **Arnela Salkić** *Department of Information Technologies, International Burch University,*

*Faculty of Engineering and Sciences, Sarajevo, Bosnia and Herzegovina*

**ABSTRACT**

***Aim of this paper is to give insight in Covid 19 data and to try to predict whether individual person will recover from this virus. Furthermore, this paper aims to give some answers how information like the country, the age, and the gender of the patient, the number of cases in their country and whether they're from or have visited Wuhan can be used to make that prediction.***

***Study uses Novel Corona Virus (COVID-19) epidemiological dataset. Logistic regression model and Random Forest algorithm are used in order to make prediction, and the Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. Paper reveals that recovery/survival is supposed to depend on the age of the patient, gender and country from which patient come. Information, whether the patient is from Wuhan or has visited Wuhan, does not affect recovery/survival of patient.***

**Keywords:** ***Covid 19; Pandemic; Logistic Regression; Random Forest; Chi-Square Test.***

## INTRODUCTION

Coronaviruses are frequent RNA viruses, from the Coronaviridae family, which are responsible for digestive and respiratory infections in humans and animals. The virus owes its name to the appearance of its viral particles, bearing growths which evoke a crown.[1] COVID-19 is the infectious disease caused by the last coronavirus that was discovered. This new virus –and disease- was unknown before the outbreak occurred in Wuhan, China in December 2019. COVID-19 is now pandemic and affects many countries around the world. The most common symptoms of COVID-19 are fever, dry cough, and fatigue. Other less common symptoms may also appear in some people, such as body aches and pains, nasal congestion, headache, conjunctivitis, sore throat, diarrhea, loss of taste or smell, rash, or discoloration of the fingers on the hand or foot. These symptoms are generally mild and appear gradually. Some people, although infected, have only very mild symptoms.[2]

Most patients (approximately 80%) recover without the need for hospitalization. About one in five people with the disease have severe symptoms, including difficulty breathing. Older people and those with other health conditions (high blood pressure, heart or lung problems, diabetes or cancer) are more likely to have serious symptoms. However, anyone can get COVID-19 and become seriously ill. People of any age who have a fever and or cough associated with difficulty breathing / shortness of breath, chest pain / pressure, or loss of speech or difficulty getting around should see a doctor immediately. It is recommended, if possible, to call the care provider or health facility in advance, so that the patient is referred to the appropriate service.[2]

Data science will play a main role in the global response to the Covid-19 pandemic by analyzing data which is collected daily and by searching patterns in those datasets which would help humanity to fight pandemic.

Aim of the study is to give insight into Covid-19 pandemic by bringing answers like whether information like the country, the age, and the gender of the patient, the number of cases in their country and whether they're from or have visited Wuhan can be used to predict whether a random patient whose data we have will recover from this virus. The purpose of this study consists in providing prediction model for future recovery trend which can be used to specify additional measures which would help fighting Covid 19 pandemic.

This paper is organized as follows: section 2. Where several studies about Covid 19 topic have been reviewed and where research question of this paper have been introduced. In section 3. Used datasets and details have been presented as well as methods which were used to get answers about Covid 19 which later lead to the final conclusion about research question. Results are presented in section 4. Using visualization for better understanding, while last two sections 5. And 6. Are discussion and conclusion about conducted study.

**LITERATURE REVIEW AND RESEARCH QUESTION**

Previous studies and data exploration have been made concerning this topic, however the goal of each study differs. For some data scientists on Kaggle, the model to develop was a regression model to predict the number of cases each day in each country. These studies findings suggested that China and Italy, who were the first to contract the virus, didn't record many cases during the month of January and then skyrocketed to the top of the coronavirus world meter list, which conveys that the virus spreads at a very high speed. The most affected countries were also in the northern hemisphere, hence the hope that the virus would disappear little by little as the weather around the globe gets calmer. The findings also included China's astonishing results after lockdown which encouraged many countries to also undergo a confinement mode. Models used comprised of linear regressors and learning algorithms such as XgBoost. [3][4]

Nadia AL-Rousan and Hazem AL Najjar in their work [5] are analyzing Covid-19 pandemic in South Korea based on recovered and death cases. Their study is based on statistically analysis the effect of factors such as region, sex, birth year, infection reasons and released or diseased date on the reported number of recovered and deceases cases. The X2 test is used to find the impact of the previous attributes on the number of recovered and deceases cases. Result of their work shows that confirmed date and infection reasons do not affect deceased cases, while on the other side confirmed date and region variables are significant with recovered cases. Besides, the results found that multinomial logistic

regression could give initial indicator about the possibility to survive or die based on the collected data. It is found that the results of multinomial logistic are in line with the results of the X2 test.

Dr. Anis Kouba in his work [6] aims to answer several questions about Covid-19 pandemic: How does COVID-19 spread around the world? What is its impact in terms of confirmed and death cases at the continent, region and country levels? How does its severity compare with other epidemic outbreaks, including Ebola 2014, MERS2012, and SARS 2003? Is there a correlation between the number of confirmed cases and death cases? Data analysis is based on Novel Coronavirus COVID-19 Data Repository provided by Johns Hopkins University. Tableau Professional software was used to analyze the collected data and to develop visualization dashboards about the Coronavirus disease. Methodology consists in creating descriptive models of the Coronavirus outbreak using statistical charts to understand the nature of the spread and its impact. Analysis was developed at three levels, namely, at the country-level, at region-level and continent-level. Each level provides different granularities towards understanding the distribution of the disease around the world. The descriptive model provides different types of statistical charts, including bar charts, geographic maps, heatmaps, box plot, and packed bubbles, to represent different features of the COVID-19 outbreak. Dr.Kouba also develop some predictive models using linear and polynomial regressions to predict the evolution of the outbreak, given the historical data.[6]

In [7], the authors investigated the impact of preventive measures, such as social distancing, lockdown in the containment of the virus outbreak. They developed prediction models that forecast how these measures can reduce the mortality impact of aged people. Mathematical model is performed on the spread of the Novel dataset coronavirus that considers both, the age and social contact structure. Authors conclusion is that the three-week lockdown will be insufficient in India. Their model suggests that sustained periods of lockdown with periodic relaxation will reduce the number of cases to levels where individualized social contact tracing and quarantine may become feasible.

The authors of [8] addressed the question about how the virus has spread from the epicenter of Wuhan city to the whole world. They have also analyzed the impact of preventive measures such as quarantine and city closure in mitigating the adverse impact of the spread. The authors have demonstrated visual graphs and developed a mathematical model of the disease transmission pattern.

Research question of this paper aims at studying the chances that a patient diagnosed with Covid-19 must recover from this illness. This research aimed to identify which information (variables) about patient are dependent with patient recovery. Questions which are stated

at the beginning of this study are: whether patient age is important in patient recovery/survival; whether patient gender is important in patient recovery/survival; whether information is patient from Wuhan or has visited Wuhan is important in patient recovery/survival; whether patient country from which comes from is important in patient recovery/survival. It is a classification problem (target variable is a response “Survived” 0 or 1) rather than a regression problem.

## METHODOLOGY

### A. Data Collection

With more than 10 million cases worldwide among with only half have recovered while the other half remains active, 500 000 deaths, the Coronavirus has become a pressing issue in every part of the world.

Because no country knows precisely the number of infected people among its citizens, the number of cases relies on the tests that the country performs and provides for the people.

In the data-oriented world that we live in today, all countries have been keeping track of all tests that they’ve done daily along with the results. Which fortunately enabled us to do this study.

Both datasets that have been worked with have been retrieved from Kaggle. The data set is the “Novel Corona Virus (COVID-19) epidemiological dataset”. The data is compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from various sources including the World Health Organization (WHO), DXY.cn, BNO News, National Health Commission of the People’s Republic of China (NHC), China CDC (CCDC), Hong Kong Department of Health, Macau Government, Taiwan CDC, US CDC, Government of Canada, Australia Government Department of Health, European Centre for Disease Prevention and Control (ECDC), Ministry of Health Singapore (MOH), and others. JHU CCSE maintains the data on the 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository on Github.[9]

This dataset contains 2 files, the first file “time_series_covid_19_recovered.csv” which provides us with daily information on the affected people in about 200 countries from the 22nd of January 2020 to the 16th of April 2020, and the second data set “COVID19 linelist data.csv” that contains observations of patients around the world where authorities have kept track of confirmed cases, recoveries and deaths.

### B. Data Analysis Method

This study consists of exploring the two datasets thoroughly, understanding the variables and seeing their distributions, visualizing the evolution of numerical variables (time series

of the numbers of cases in each country) and relationships between categorical ones (statistical testing). Afterwards, a simple logistic regression was performed to predict whether a patient recovers from Covid19 or not and some clustering techniques have been used to try to understand better recovery rates and how it is linked to the age of the patient. Logistic regression model and Random Forest algorithm are used, and the Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables.

Author of the paper would have opted for Text analysis of the columns “Summary” and “Symptoms” but were unable to do so due to lack of data (portion of Nas in the columns). However, other analysis methods have been used such as descriptive data analysis for the time serial evolution of the number of cases and inferential data analysis when performing statistical tests to reveal the dependence between some variables.

To conclude with predictive analysis with “Survived” as a target variable based on a percentage that conveys how much a patient is likely to be treated and survive the virus.

**C. Data Cleaning and Preprocessing**

Some countries are divided into provinces/district in both datasets. At first, it has been thought that each district has to be treated individually in order to join the two datasets. But different regions exist in the two data frames, most importantly some regions in the patients list do not exist in the time series data frame and would therefore stay empty if this data frame was filled with the number of cases by specific regions of a country. Which is why was decided to eliminate regions and work with total numbers of cases in a country.

Taking care of Nas is a very important part of every data exploration process. Using the ggpubr library in R, the number of Nas was visualized in each variable of the dataset and decide to get rid of all variables that contain more than 250 Nas (a little less than the 1/4th of the data). Following variables were selected to continue this study: age, gender, number of cases in country, from Wuhan, visiting Wuhan, death, recovery and country.

**D. Visualization**

* Plotting time series(number of cases by country)

In the initial time series data frame, each day is a variable and that is not a proper structure for visualizing time series. A series of steps is required to obtain a data frame that can be worked with:

- The data frame was transformed to have a column where each line is a day

- The header was removed and some variables that will not be used for this visualization
- All variables were converted to numeric type
- The created column containing dates was set as date-time so that it is recognizable as such by R when dates on the x-axis were plotted.

The number of cases per country were chosen to be visualized continent by continent. Those graphs will enable to follow a country's covid-19 outbreak and perceive at what moment the exponential spread has started. * Plotting survival against other categorical data

Using the ggplot2 library in R, data set has been grouped by category of the qualitative variable against which Survival using the "group_by" method was plotted.

The result is a histogram illustrating the counts of Survival values (0 or 1) in each category.

**E. Chi-Square Test of Independence**

This test is used to assess the existence or not of a relationship between two characteristics within a population, when these characteristics are qualitative or when one characteristic is quantitative and the other qualitative, or even when the two characteristics are quantitative but that the values have been combined. Note that this test makes it possible to check the existence of a dependence but in no case the direction of this dependence. Like in any statistical test, there is a null hypothesis H0 that has to be tested – determine whether there is enough evidence to "reject" this hypothesis. For this study test there is:

H0 : the two variables are independent

H1 : the two variables are not independent

á = 0.05 : tolerated risk

The mathematical steps were hardcoded into R to perform this test. The statistic that need to be calculated to perform this test is the following:

$$X^2 = \sum \frac{O-E}{E} \qquad (1)$$

Where O is the observed score and E is the expected score. [10]

The observed scores are in a table of the values that already exist in our dataset and are already in our disposition. The expected scores table is one of the same dimensions as the

observed scored. It is obtained by multiplying each value of the observed values with the sum of its row and dividing by the total number of observations.

Once these two tables are set, it can be proceeded to calculating the chi-squared statistic. The pvalue is calculated then. This is done by calculating the chi statistic for multiple random samples of our data, the pvalue is the proportion of the X2 statistic that are higher than the real X2 statistic of the sample.

If pvalue < 0.05 the null hypothesis is rejected, meaning that the two qualitative variables are independent.

If pvalue . 0.05 the null hypothesis is accepted, meaning that the two qualitative variables are dependent.

**RESULTS**

After cleaning and preprocessing the two datasets here are the resulting data frames that have been used for this study. There are 85 countries in the first data frame and the number of cases that appear each day in that said country.

| Time | Timor.Leste | Togo | Trinidad.and.Tobago | Tunisia | Turkey | Uganda | Ukraine | United.Arab.Emirates | Uruguay | US | Uzbekistan |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0020-01-22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0020-01-23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0020-01-24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0020-01-25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0020-01-26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0020-01-27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 1**

After cleaning the observations file and removing the variables that contained too many NAs, the following variables were decided to be kept: age, number of cases in country, gender, from Wuhan, visited Wuhan and the country of the patient. The two datasets were also joined by summing the number of cases during each month that was available in the first time series dataset for each country.

| age | case_in_country | gender | from.Wuhan | visiting.Wuhan | death | recovered | country | JanCases | FebCases | MarCases | AprCases |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 65 | 36 | unknown | 0 | 0 | 0 | 0 | France | 0 | 102 | 52910 | 352696 |
| 35 | 37 | female | 0 | 0 | 0 | 0 | France | 0 | 102 | 52910 | 352696 |
| -99 | 38 | male | 0 | 0 | 0 | 0 | France | 0 | 102 | 52910 | 352696 |
| 23 | 39 | female | 0 | 0 | 0 | 0 | France | 0 | 102 | 52910 | 352696 |
| 35 | 1 | male | 0 | 1 | 0 | 1 | Japan | 6 | 342 | 5635 | 10453 |
| 45 | 2 | male | 1 | 0 | 0 | 0 | Japan | 6 | 342 | 5635 | 10453 |
| 35 | 3 | female | 1 | 0 | 0 | 0 | Japan | 6 | 342 | 5635 | 10453 |
| 45 | 4 | male | 1 | 0 | 0 | 0 | Japan | 6 | 342 | 5635 | 10453 |

**Table 2**

By joining two datasets, plotting the evolution of cases in each country during these 4 months was also possible. Figure1 shows the results for the most popular countries during this pandemic (those that have been hit the most and most spoken of in the news).

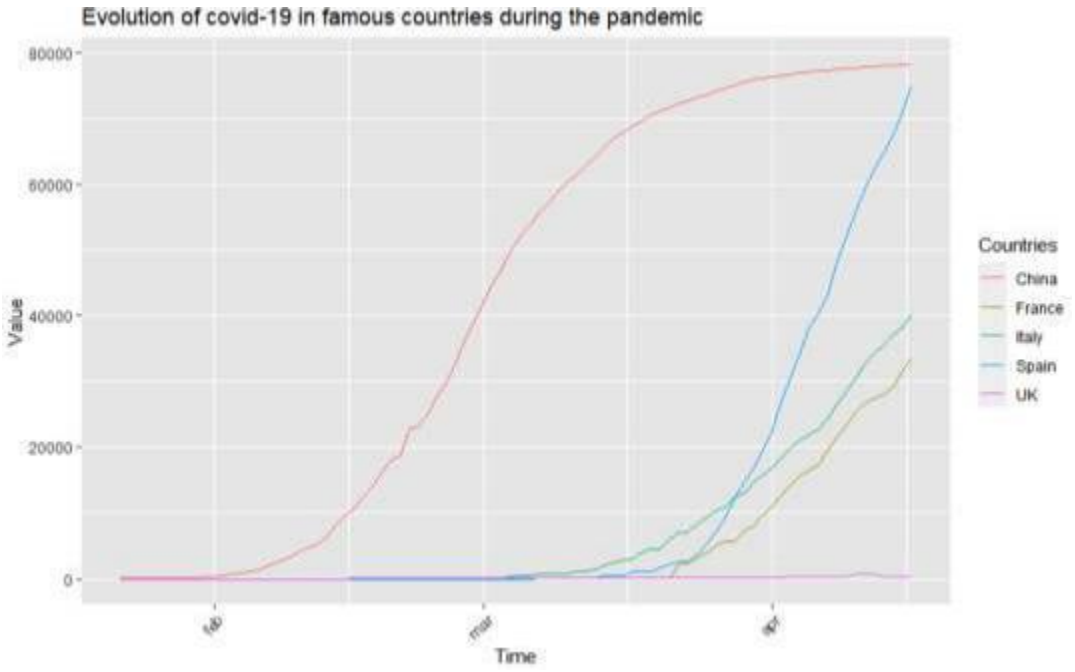


Fig 1:- Countries affected the most with Covid-19

As it can be noticed, China, the epicenter of this pandemic, was the first to witness an exponential rise in the number of cases early February. The other countries followed suite during the month of March.

To extract some insights from the data concerning the relationship between the survival of the patient and their gender and age was also possible. As it might have been expected before, gender is independent of the survival of the patient, but age does play a role in whether a patient will recover or not.

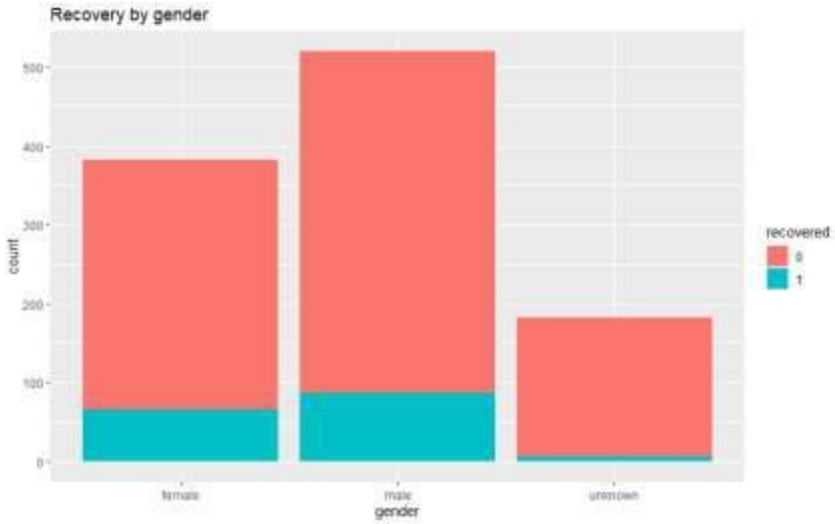


Fig 2:- Recovery by gender

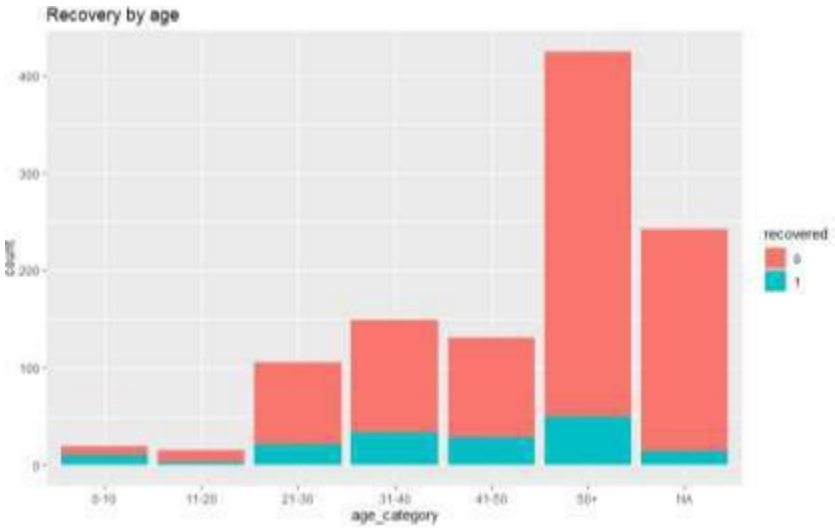


Fig 3:- Recovery by age

Another thing that might have been suspected is that a person would be less likely to recover if they're coming from

Wuhan or if they've visited Wuhan. Here are the illustrations from the data:

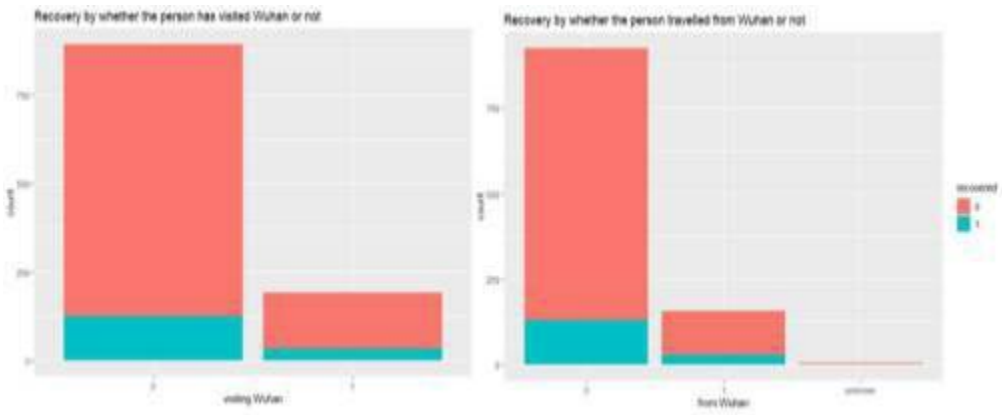


Fig 4:- Recovery dependency if patient visited Wuhan or not

To confirm these hypotheses, and because these are categorical variables, it has been decided to do a Chi-square test of independence to see whether survival was dependent on any of those variables. Results are summarized in the following table:

| Variable 1 | Variable 2 | Dependence |
|---|---|---|
| Survival | gender | yes |
| Survival | age | yes |
| Survival | country | yes |
| Survival | from.Wuhan | no |
| Survival | visited.Wuhan | no |

Table 3:- Categorical variables dependency

The final step of the study was creating a classification model with the data that have been prepared and try to predict whether a person would survive the virus or not. The logistic regression model gave an accuracy of 84% and random forest one of 91%. Random forest gave a better result than logistic regression because the model weighs certain features as more important than others (feature selection), the absence of assumption of a linear relationship like regression models do, and because random forest takes random samples from the data set, forms many decision trees, and then averages out the leaf nodes to get a clearer model (ensemble learning).

**DISCUSSION**

The analysis confirms that it is indeed possible to predict the recovery -or no- of a covid19 patient given some information about his situation such as age, gender, and the country it comes from. It is to author grand surprise that the statistical tests indicate that the Survived variable is actually dependent on the gender of the patient. Figure2 shows that males are more likely to be infected than females, based on data in our dataset. On the other hand, whether the patient is from Wuhan or has visited Wuhan does not affect in any way the outcome of the observation – author expectations were otherwise.

In line with this research paper question testing, Survival is supposed to depend on the age of the patient as all of us have witnessed during the previous months. The death rates were particularly high for the elderly while adults, teenagers and kids seemed to recover way more successfully from the illness. The country is also supposed to affect the outcome of the treatment since the latter depends on the country's means and dispositions to treat the patient. This study can contribute to predict future pandemic recovery trend by giving global picture of how many patients could be recovered from this virus and based on that information define further measures which can be overtaken. Future studies should take into account some variables that were in the "COVID19_line_list_data.csv" dataset but that couldn't have been used because of the number of missing data in those columns, these columns include : "cases_in_country", "reporting_date", "symptoms_on_set", "host_visit_date", "exposure_start" and "exposure_end". This information would certainly have been valuable to predict the survival of the patient.

**CONCLUSION**

This research aimed to identify which variables -pieces of information- about a patient is important to predict whether a person diagnosed with covid19 would survive the virus or not. Results conveyed that age, gender and country are the variables on which the target variable Survived depends.

Further research is needed to determine the reliability of these results since the dataset that have been worked with contains data until April 2020 only, while the pandemic is persisting and will still persist in the future.

**REFERENCES**


- Wikipedia. (July 2020). Coronavirus disease 2019 (Online). A v a i l a b l e : h t t p s : / / e n . w i k i p e d i a . o r g / w i k i / Coronavirus_disease_2019
- World health organization. (July 2020). Coronavirus disease ( C O V I D – 1 9 ) p a n d e m i c [ O n l i n e ] . Av a i l a b l e : https://www. who . int/emergencies/diseases/novel – c o r o n a v i r u s – 2 0 1 9 ? g c l i d = E A I a I Q o b C h M I n L X X _b2G6wIViqkYCh21BgipEAAYASAAEgIAdvD_BwE
- Kaggle competition.(July 2020). COVID19 Predictions u s i n g X G B O O S T [ O n l i n e ] . A v a i l a b l e : https://www.kaggle.com/anshuls235/covid19-explained-through-visualizations
- Kaggle competition.(July 2020). Nischay Dhankhar. Covid19 Week5 (Visuals+RandomForestRegressor) [ O n l i n e ] . Av a i l a b l e : h t t p s : / / w w w. k a g g l e . c o m /nischaydnk/covid19-week5-visuals-randomforestregressor

- N. AL-Rousan, H. AL-Najjar, April 2020, “Data analysis of coronavirus COVID-19 epidemic in South Korea based on recovered and death cases”, Journal of Medical Virology [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/jmv.25850
- Koubaa, Anis. (March, 29th 2020). “Understanding the COVID19 Outbreak: A Comparative Data Analytics and Study” [Online]. Available: https://www.researchgate.net/publication/340331977_Understanding_the_COVID19_Outbreak_A_Comparative_Data_Analytics_and_Study
- Singh, Rajesh & Adhikari, Ronojoy. (2020). “Age-structured impact of social distancing on the COVID-19 epidemic in India” [Online] . Available: https://github . com/ rajeshrinet/pyross
- B. Chen, M. Shi, X. Ni, L. Ruan, H. Jiang, H. Yao, M. Wang, Z. Song,Q. Zhou, and T. Ge, “Visual data analysis and simulation prediction for covid-19,” 2020, Cornel University, Physics, Medical Physics [arXiv:2002.07096]
- Kaggle, July 2020, Novel Corona Virus (COVID-19) epidemiological dataset [Online] . Available: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset#time_series_covid_19_ confirmed_US.csv
- Statistics Solutions, July 2020, Chi-Square Test of Independence [Online], Available: https://www.statisti cssolutions.com/non-parametric-analysis-chi-square/
- Kaagle, July 2020, COVID19- Inside Story of each Countries [Online]. Available: https://www.kaggle.com/pradeepmuniasamy/covid19-inside-story-of-each-countries
- Latif, Siddique & Usman, Muhammad & Manzoor, Sanaullah & Iqbal, Waleed & Qadir, Junaid & Tyson, Gareth & Castro, Ignacio & Razi, Adeel & Kamel Boulos, Maged & Crowcroft, Jon. (2020). “Leveraging Data Science To Combat COVID-19: A Comprehensive Review”. Research gate,10 . 13140/RG . 2 . 2 . 12685 . 28644/4 . Available: https://www.researchgate.net/publication/340687152_Leveraging_Data_Science_To_Combat_COVID-19_A_Comprehensive_Review#pf2
- Medium March 2020, R Tutorial: Analyzing COVID-19 Data, Introduction to using R in the real world [Online]. Available: https://towardsdatascience.com/r-tutorial-analyzing-covid-19-data-12670cd664d6

- Opuszko, Marek. (2020). Analytics of the COVID-19 (Corona) Spread using R [Online] . Available: https://www.researchgate.net/publication/340412951_Analytics_of_the_COVID-19_Corona_Spread_using_R.