**Data Mining**

**ISCG 7426**

**Semester 2**

**August 2020**

****

**Assignment-1**

**Case Analysis of Novel Corona Virus-COVID 19**

**(Reported during period 26 February 2020 – 1 August 2020)**

**Prepared By: Priya Payyadi Surendran**

**ID:1533656**

**Supervisor: Miss Natt Nuntalid**

**Declaration:” Wherever this assignment draws on the work of others, such sources are clearly acknowledged.”**

Table of Contents

[**1.** **Problem and Requirement Analysis** 2](#_Toc47982514)

[**a.** **What is the application background of the proposed data mining research?** 2](#_Toc47982515)

[**b.** **Who are the stakeholders of the research, and what are their requirements?** 2](#_Toc47982516)

[**c.** **What data mining analytics tasks can be formulated from the requirements?** 3](#_Toc47982517)

[**2.** **Data Discovery** 3](#_Toc47982518)

[**a.** **What operational procedures do you need to follow?** 3](#_Toc47982519)

[**b.** **What data are in demand?** 3](#_Toc47982520)

[**c.** **What data can we collect in practice for the research?** 4](#_Toc47982521)

[**d.** **What is the gap between demand and available?** 4](#_Toc47982522)

[**3.** **Data Collection and Validation** 4](#_Toc47982523)

[**a.** **What is the software and hardware setup of data collection?** 4](#_Toc47982524)

[**b.** **What are the requirements we should meet when collecting data?** 6](#_Toc47982525)

[**c.** **How to validate your collected data?** 6](#_Toc47982526)

[**4.** **Data Pre-processing** 6](#_Toc47982527)

[**a.** **How to integrate data from different sources?** 6](#_Toc47982528)

[**b.** **How to do data cleaning?** 7](#_Toc47982529)

[**c.** **How to do data transformation?** 9](#_Toc47982530)

[**5.** **Data Summarization** 10](#_Toc47982531)

[**a.** **Give a brief introduction of your dataset.** 10](#_Toc47982532)

[**b.** **What is the format and statistical description of your dataset?** 11](#_Toc47982533)

[**c.** **What are your suggestions for future data collection?** 11](#_Toc47982534)

[**References** 12](#_Toc47982535)

## **Problem and Requirement Analysis**

### **What is the application background of the proposed data mining research?**

The current pandemic COVID-19 poses prolonged and imperative coercions to the wellbeing and health of the people worldwide. It is causing us to face unrivalled level of uncertainty. People who are affected with this virus, doctors, hospitals and our economy need answers to few queries that can guide them to make proper decisions which will direct them to take suitable progression of action. The data mining topic chosen is the analysis of the cases of COVID-19 in New Zealand from 26th February 2020 to 1st August 2020 which can be applied to medical analysis to get a better insight on how to manage health crisis of this virus outbreak. As confirmed corona virus cases exceeds seventeen million globally, a fight against this battle has inspired many to implement innovative methods to identify the patterns in these cases of COVID -19. The growing importance on data mining in medical fields can deliver the precise environment for change and advancement (Durairaj & Ranjani, 2013).

### **Who are the stakeholders of the research, and what are their requirements?**

A myriad of stakeholders can be linked to the medical research topics. In order to address the critical public health crisis due to the COVID-19 pandemic, important role of the stakeholders in providing the health care delivery and management is crucial. To better understand the rapidly evolving nature of the pandemic, detailed enquiry with the impacted stakeholder group is inevitable. A vital step is stakeholder engagement to take robust decisions and timely course of action throughout.

The first and foremost stakeholder is the public scattered globally. Awareness about COVID-19 is mandatory among the public and they should follow the safety measures as a precaution against coronavirus. The utmost need is to know that these safety measures are effective or not .

The next major stakeholders are those belonging to the Health and medical fields like public and private health care providers, Health organizations, scientists and researchers. Health care providers like public and private hospitals, Pharmacies, clinics etc must have the resources and utilities to incorporate the inflow of patients. The frontline doctors must be educated with the specific need of healthcare providers used to manage COVID patients. Facilitating interaction with healthcare staffs on the availability and importance of specific drugs etc. Preparing for providing better facilities and services beforehand in order to incorporate the COVID affected patients for speedy recovery. Health organizations like WHO, healthcare professional organizations, emergency response organizations to expediate the information shared on the COVID-19 and regarding shortages in the supplies of the medical equipment and medicines, so they can take preventive measures to help the affected countries. Scientists and researchers can make use of the data collected to learn the behaviour of the virus so that they can discover a better drug to fight against the virus.

The federal, state and local government are aware of this virus and a prompt decision must be made to save the country. An accurate decision on time can save us all from the pandemic. An example for failed lockdown is India, where they started allowing people from other countries in without proper quarantine measures whereas country like New Zealand was successful in locking down the entire country and eradicate the virus. This shows the importance of the role of the government and allied organization in eradicating the pandemic by bending or tightening the laws and regulations. This also includes Educational Institutions and Tourism as they are the major income earner of a country like New Zealand. Due to the border restrictions, new student intake as well as tourist inflow especially during the peak time like winter had been stopped. This has affected the economy as a whole and especially students since many of them had to relocate and migrate to online learning. The ministry of education as well as tourism had to take apt decision to safeguard the students and tourists (New Zealand takes early and hard action to tackle COVID-19, 2020).

Then comes the employers as they need to know when their employee can return to office healthy and when they can resume their work to run the company successfully. Many employers opted for work from home to help their employees to protect themselves as well as to keep the business going.

The financial institutions and the insurance company are the next line stakeholders . With this pandemic many lost their jobs and as far as a bank is concerned, the loan repayment must be promptly paid else the account will turn in to Nonperforming asset. So, they need this information to restructure their loan policy to help the customers . The insurance companies needs to know how to carry the claims and the need to restructure the premiums of the policies due to COVID 19 (Healthcare Stakeholders, 2020).

### **What data mining analytics tasks can be formulated from the requirements?**

The concept used is association as this concept of data mining can be used to detect association between, sex, age and cases being reported. This can be used to analyse whether special preventive measures among males and females of various age group can be taken or if already taken will help them to recover or fight against the virus ( Han, Pei, & Kamber, 2011). This can be used to compare and arrive at the conclusions like probability of COVID-19 being completely eradicated or probability of increases in cases reported.

## **Data Discovery**

### **What operational procedures do you need to follow?**

First and foremost, a proper identification of a problem or an opportunity that can be analysed. Here, I have taken COVID-19 case analysis of New Zealand. The next procedure is to find in detail the association of the problem or what exactly should be analysed and why exactly it is happening. In this case, I chose to study the pattern of the virus COVID-19 in different age group and sex. The next step is based on the previous procedure and we decide what data to be collected, how to be collected and from where to be collected. After deciding on these, plan a method for data collection and a source for data collection. Here, I have collected the data from Ministry of Health New Zealand from the start of the outbreak of the virus on 26th February 2020 till 1st August 2020 of different age group and sex. Then finally make sure that the data collected and prepared is valid and useful for analysing the problem or opportunity.

### **What data are in demand?**

All information that gives the details of the pattern of the COVID-19 is in demand for the data analysis. The data collected contains date of case confirmed, age, sex, travel details, geographical location, recovery and death rates etc. The data in demand can be the way how they recovered after confirming the COVID-19 . More details like whether the patients were hospitalized, or they were carrying on their own home remedies? These are useful information which can be used by the health care providers to understand the situation correctly and take a prompt decision to facilitate the affected patients by providing them with proper medical facilities which may include some special requirements. The government can make use of the data to consider new laws regulating the safety of the population of the country as well as to provide awareness among the public to take the precautions against the virus.

### **What data can we collect in practice for the research?**

The Ministry of Health, New Zealand is the source of information gathered regarding the novel corona virus infected individuals and it consist of number of confirmed and probable cases reported daily and categorized by different District Health Board(DHB). The individual cases reported which are either confirmed or probable cases. The data collected can be categorized according to region of District Health Board, date of reporting or date of arrival in New Zealand as well as country and the date on which the flight departed from the country, they visited last, in case of international travellers. The age, sex, confirmed cases or probable cases and the deceased status of everyone is gathered.

### **What is the gap between demand and available?**

The information accessible from the Ministry of Health , New Zealand is date of case reported, age, sex, DHB, Country last visited, date of departure and arrival, case status and deceased status. The data in demand can be the way how they recovered after confirming the COVID-19 . More details like whether the patients were hospitalized, or they were carrying on their own home remedies? If hospitalized, then how long , what were the medicines given, whether they were in Intensive Care Unit or some other special care rooms? Whether all age group requires same time period for recovery? If home remedies were followed, then what are those? These data were not available as of now and this is a long gap between demand of data and its availability.

## **Data Collection and Validation**

### **What is the software and hardware setup of data collection?**

The software used is SNOMED CT by the SNOMED International as a part of global effort to manage this pandemic. SNOMED CT is used to record, share, integrate and analyse COVID-19 data. SNOMED subsets used to code myriad of data elements related to COVID-19 since these data can be critical for frontline service delivery, retrospective data analysis and pandemic surveillance. SNOMED CT can be used to capture COVID-19 related data for data elements like Provider & Facility Details(healthcare profession, site of care etc), Patient Demographics(sex, age, occupation, ethnicity, living arrangements, care and support circumstances etc), Clinical Assessments(symptoms, severity, exposure events etc), Tests & Investigations (specimens, lab tests, results), Prevention, Treatment & Education (prevention, medicines, awareness, treatment equipment, administrative procedures etc). There are SNOMED codes for each data element, and these are stored in the patient’s record by health care providers and these are stored by each DHBs in a local database. Later, Ministry of Health collects these data and publishes it for general people after processing it (COVID-19 Data Coding using SNOMED CT, 2020).

A close up of a logo

Description automatically generated

Figure :Use of SNOMED CT

A screenshot of a computer

Description automatically generated

Figure : SNOMED CT Browser New Zealand Edition

The hardware used to collect COVID-19 data are swab collection kits and testing equipment of laboratories. Swab is like a cotton ear bud with longer stick and is used to collect samples from public , especially from suspected person. The swabs are mainly collected from upper respiratory parts like nose and throat and are called nasopharyngeal and oropharyngeal swabs. These swabs are inserted into the tubes and send to the labs for testing. Lower respiratory samples like sputum and bronchial fluids can also be collected. But collecting upper respiratory samples are easier and since RT-PCR(Reverse Transcription Polymerase Chain Reaction) technique is used to test these swab samples (COVID-19 testing, 2020) (COVID 19 Testing and Sample Requirements, 2020).

Figure : Swab collection kit Figure : RT PCR set up for COVID testing

### **What are the requirements we should meet when collecting data?**

The main requirement to encounter when gathering data is that the data collected is related to the research and is valid. The data collected should be the data in demand and there should not be a gap between data in demand and its availability. The data gathered should be useful for the stakeholders to analyse the problem or the opportunity under the research and it should answer all the questions. The data collected must provide a proper conclusion to the research and researchers must be able to derive it easily.

### **How to validate your collected data?**

To confirm the information collected, I need to make certain that the composed data helps the researchers to tackle the problem or opportunity under question. The data collected should be valid and clean. No quality data means no quality mining can be done. In order to convert the raw data, data transformation and data pre-processing like cleaning, integrate , transform and reduce the data according to the requirements of the research.

## **Data Pre-processing**

### **How to integrate data from different sources?**

Utmost need should be given to the careful integration of data from multiple sources before the data set is processed since it will aid in avoiding redundancies and inconsistencies which in turns improves data mining speed and quality. In order to integrating data into a coherent store , certain possible problems confronted are different names for different attributes, attributes may be derived from other tables, different representations and scales. These problems must be tackled correctly before integrating the datasets. Considering my dataset, I only used a single data set and as a result, integration was not required. If I’m to integrate another dataset with COVID-19 related data elements of New Zealand, then I would give more care to integrating the datasets when it comes to the date and age formats, names of attribute(country last visited and case status). The date format of my dataset is DD/MM/YYYY and whenever I integrate it with another source, I will make sure that the format is same. The age format is categorized into group in my dataset and if I’m integrating it, then I will make sure that the age falls under correct group. I will also make sure that the names of the country last visited is written completely and not in short form(E.g.: United States of America and not USA) and the case status is either confirmed or probable (instead of a yes or no/doubtful).

|  |  |  |
| --- | --- | --- |
| **Date notified of potential case** | **Sex** | **Age group** |
| 26/02/2020 | Female | 60 to 69 |
| 2/03/2020 | Female | 30 to 39 |
| 4/03/2020 | Male | 40 to 49 |
| 5/03/2020 | Male | 70+ |

|  |
| --- |
| **Case Status** |
| Probable |
| Confirmed |

|  |
| --- |
| **Last country before return** |
| United States of America |
| New Zealand |
| United Kingdom |
| United Arab Emirates |

Figure : My dataset format

### **How to do data cleaning?**

As a part of data pre-processing, data cleansing is as important as data integration. Data in the real world is raw(dirty), noisy(filled with errors and outliers), inconsistent(contains discrepancies) and incomplete(missing attribute values). No quality data implies no quality data mining. Incomplete dataset can be managed by ignoring the instances containing the missing information(not possible if too many instances have missing values), using a constant like “unknown” to fill in the missing values or impute the missing values using attribute mean or decision tree to know what is the most probable value to fill in. Noisy data can be cleansed using clustering (where the errors are detected and outliers are removed), detecting the suspicious values by a combined inspection by human and computer or smoothening it out by fitting the data in to regression functions.

The dataset I collected contains many missing attribute values. The attributes ”Overseas Travel”, ”Last country before return”, ”Flight number”, ”Flight Departure date” and “Arrival Date” contains many missing values and it is not advised to delete the instances with missing values. I decided to replace the missing values with the constant value “unknown” which seems to be more apt for the research.

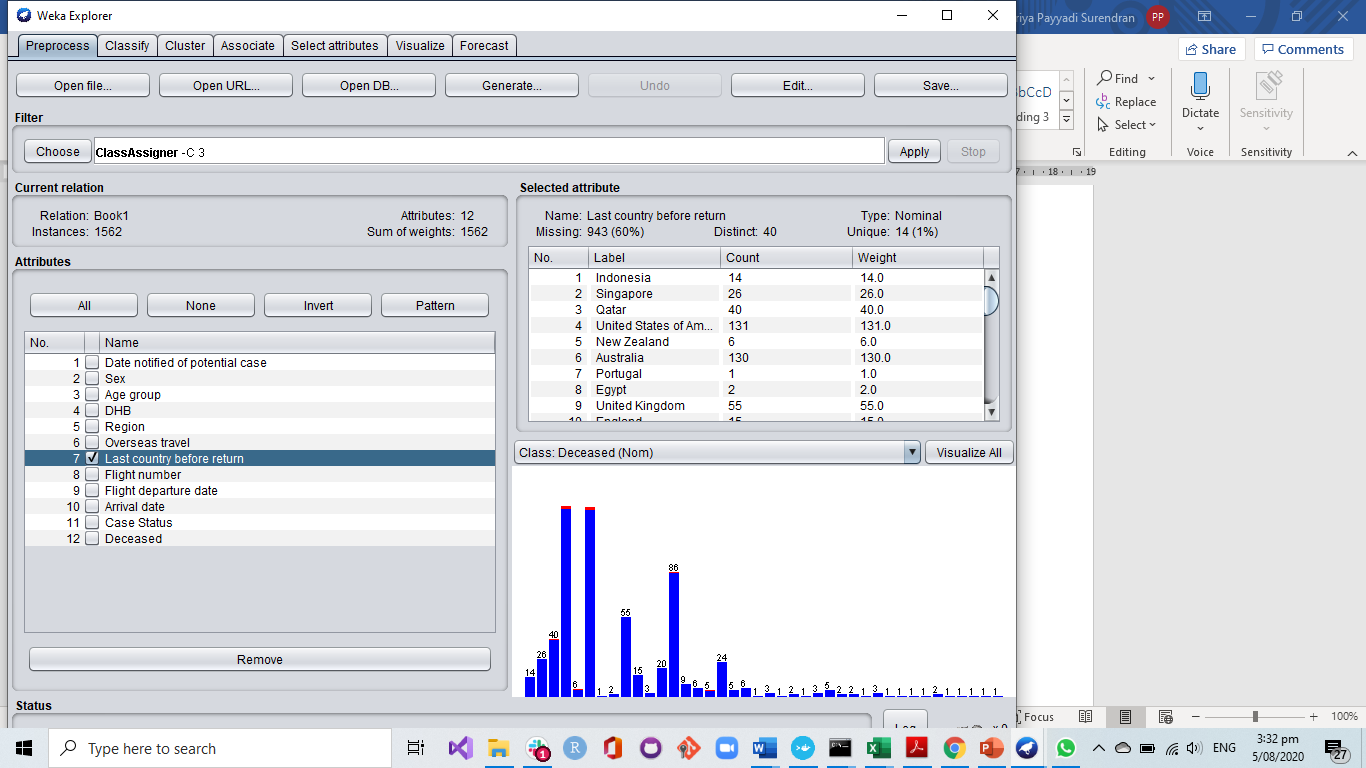


Figure :The attribute "Last country before return" with missing values-Weka

The attribute “Last country before return” to New Zealand has 943 missing values and I used the filter “ReplaceMissingValuesWithUserConstant” and substituted it with constant value ”Unknown”. The figure 7 shows the attribute after replacing the missing values, the unknown instances is 943 and 0% missing values.

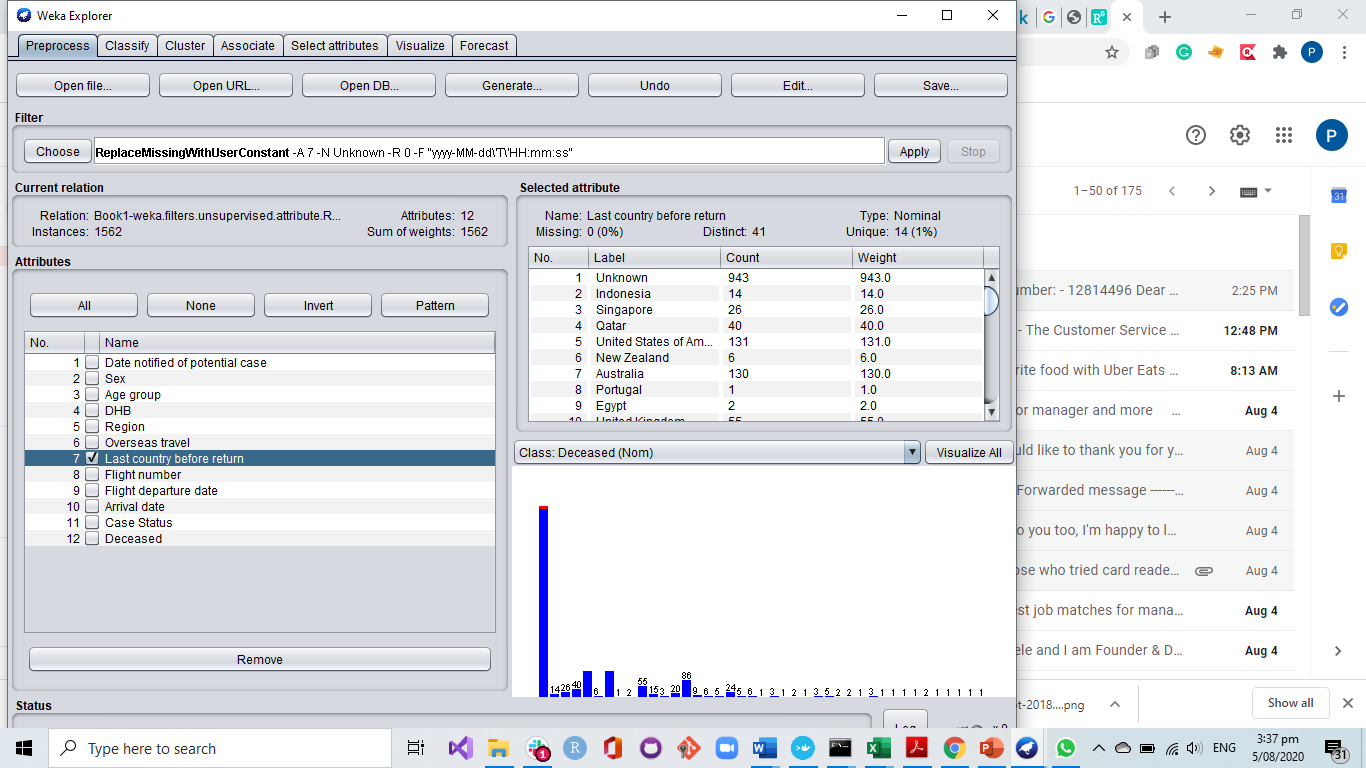


Figure :The attribute "Last country before return" after replacing missing values-Weka

### **How to do data transformation?**

The final part of data pre-processing is data transformation. Normalization, Aggregation, Discretisation, Generalisation, Attribute construction, and smoothing are the different methods to carry out data transformation. The attributes in the dataset collected is pretty much transformed. The age is already in categorical form and not in numerical form. The only attribute with most distinct values is DHB and the same can be clustered region wise to get a better understanding . I used generalisation to achieve it.

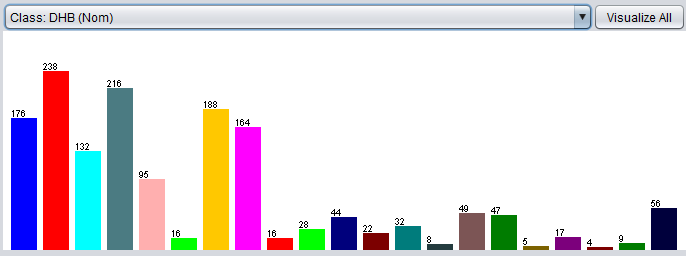


Figure : DHB cases-Weka

The DHB values reduced and a clear investigation in to finding the pattern of the virus spreading in New Zealand. I have replaced the” managed isolation and quarantine” instances of DHB with constant “Unknown” while transforming it in to DHB cluster region wise since they are not included in the DHB total.

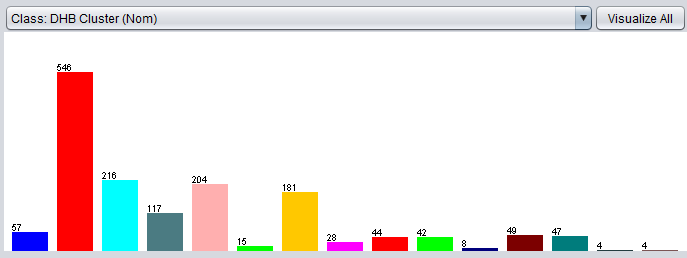


Figure :DHB Cluster cases-Weka

## **Data Summarization**

### **Give a brief introduction of your dataset.**

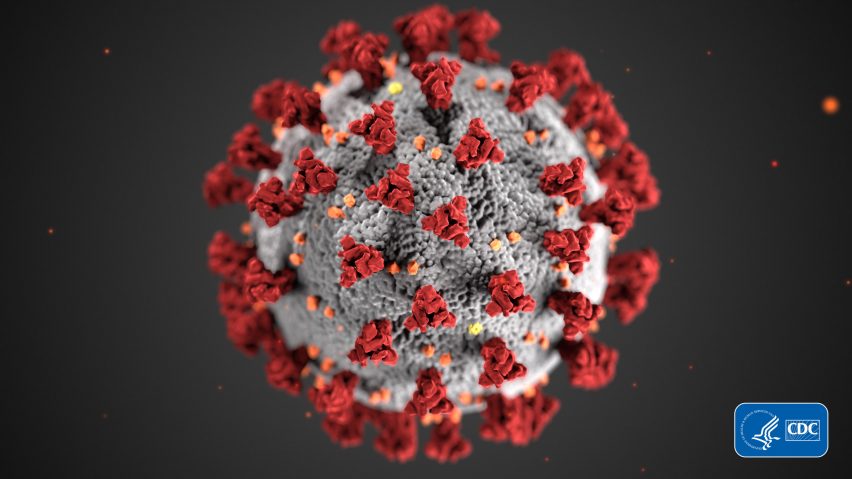


Figure :COVID-19

Corona Virus (CoVs) is a huge family of virus including Severe Acute Respiratory Syndrome (SARS-CoV). COVID-19 is a family of viruses that got the name from their spiky crown and the name was designated by World Health Organization on 02/11/2020. It was first reported in Wuhan, China and is also known as SARS-CoV-2 and is contagious respiratory virus. It originated in few animals like bats and later it was found that the disease was air borne. On 11/03/2020, WHO declared it as pandemic and as a part of preventing and controlling the spread, many countries adopted lockdown (Coronavirus disease (COVID-19) pandemic, 2020). The virus is still spreading irrepressibly, and many researchers and scientists are exploring on the possibilities to control the virus. The dataset I chose is the collection of data elements related to COVID-19 of New Zealand from the date of virus infection being first reported i.e. 26/02/2020 to 01/08/2020 till 9am by the Ministry of Health (COVID-19 (novel coronavirus), 2020). The dataset contains attributes of a person either confirmed with the virus or probable suspect like the date of travel and arrival, country last visited, age, sex etc. At present the spread is under control in New Zealand as we moved into level 1 with only border restrictions and with a few active cases being reported daily. Hence the dataset is updated daily .

Confirmed cases are people with positive lab test and probable cases are people with exposure history and clinical symptoms and are treated like a confirmed case. This dataset contains provisional information which is taken at 9am daily from local data base EpiSurv (ESR).

The main contribution of the research are the exploration of COVID 19, a clear understanding of its enhancement and how research can predict the spread of the virus and how to curb it to eradicate the entire virus. The prediction of prevalence and incidence of this disease throughout the world is very critical in helping the healthcare providers to make key decision of the disease.

### **What is the format and statistical description of your dataset?**

The dataset format is Excel and was converted to CSV format for ease of use. There are 1562 instances and 11 attributes . The attribute ” DHB” which shows under which District Health Board the infected person is gives a distinct value and hence another attribute called “DHB Cluster” has been created for a better understanding of the region of infected person in New Zealand. All the attributes are nominal values. The data set covers the details of a person either confirmed with COVID-19 or a probable victim of the virus. The details includes the Date notified of potential case, Sex, Age Group, DHB, from Overseas or not, Last country before return, Flight number, Flight departure and arrival date, case status (confirmed/ probable), deceased or not . The date notified of potential case starts from the first case reported in New Zealand (26/02/2020) till 01/08/2020 9am. The cases reported are categorized by age, sex, DHB, and the cluster to which the DHB belongs. This categorization will help to study the pattern of virus and can be used to control and eradicate it.

### **What are your suggestions for future data collection?**

The data in demand can be the way how the patients recovered after confirming the COVID-19 . More details like whether the patients were hospitalized, or they were carrying on their own home remedies? If hospitalized, then how long , what were the medicines given, whether they were in Intensive Care Unit or some other special care rooms? Whether all age group requires same time period for recovery? If home remedies were followed, then what are those? These data were not available as of now and this is a long gap between the data in demand and data that are available. The data studied was confined to New Zealand alone, may be a larger dataset either globally or continent wise would help an in-depth study of the pattern of virus spreading which in turn help the doctors, authorities etc to curb the disease globally. The dataset was already tidy and hence could not apply much of the data integration and data pre-processing. Even though, it contained many missing values like, flight number, departure and arrival date. May be as a part of future data collection, a huge data set with more attributes that are in demand integrated from different sources can lead to a better data mining and a better research conclusion.

# 

# References

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining Concepts and Techniques.* Elsevier.

*Coronavirus disease (COVID-19) pandemic*. (2020, August 1). Retrieved from World Health Organization: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

*COVID 19 Testing and Sample Requirements*. (2020, August 4). Retrieved from ESR Science for communities: https://www.esr.cri.nz/assets/Test-Forms/COVID-19-sample-and-test-requirements-FEB.pdf

*COVID-19 (novel coronavirus)*. (2020, August 1). Retrieved from Ministry of Health-New Zealand: https://www.health.govt.nz/

*COVID-19 Data Coding using SNOMED CT*. (2020, August 4). Retrieved from SNOMED International Covid Guide: https://confluence.ihtsdotools.org/display/DOCCV19

*COVID-19 testing*. (2020, August 4). Retrieved from Unite Against Covid-19: https://covid19.govt.nz/covid-19/test-and-trace/covid-19-testing/#current-testing

Durairaj, M., & Ranjani, V. (2013). Data Mining Applications In Healthcare Sector: A Study. *Engineering International Journal of Scientific & Technology Research*.

*Healthcare Stakeholders*. (2020, August 1). Retrieved from Mckinsey & Company: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/helping-us-healthcare-stakeholders-understand-the-human-side-of-the-covid-19-crisis

*New Zealand takes early and hard action to tackle COVID-19*. (2020, August 2). Retrieved from World Health Organization: https://www.who.int/westernpacific/news/feature-stories/detail/new-zealand-takes-early-and-hard-action-to-tackle-covid-19

*Our response to COVID-19*. (2020, August 4). Retrieved from Unite Against COVID-19: https://covid19.govt.nz/