ISSN 2278-3091

Volume 9, No.3, May - June 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse76932020.pdf

https://doi.org/10.30534/ijatcse/2020/76932020

# Analysis of Corona Virus spread uses the CRISP-DM as a Framework: Predictive Modelling

Muhammad Akbar Rivai [1], [2] Sfenrianto

[1, 2] Information System Management Department, BINUS Graduated Program – Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia, 11480. Email: [1]muhammad.rivai@binus.ac.id; [2]sfenrianto@binus.edu

## ABSTRACT

Corona Viruses (CoVs), which are indicated by sensory-positive RNA viruses, have a sign with a crown shape projected from the surface of the RNA genome to look very large, also have unique replication. Corona virus causes various diseases in mammals, birds ranging from enteritis in cattle, pigs, and chickens. Corona has problems such as respiratory diseases for infections in humans that cause death in infected R & D. In this study. We provide a brief introduction about Coronavirus that discusses the predictive results of spreading the Corona Virus to any country. We will also consider an outbreak of Coronavirus where the results we have found will find a solution for the future so that this virus does not spread widely to various countries.

**Key words:** Augmented reality, Big Data, Hadoop, mobile application, SME

## 1. INTRODUCTION

Coronaviruses (CoVs) are the largest group of viruses included in the order Nidovirales, which consists of the family Coronaviridae, Arteriviridae, Mesoniviridae, and Roniviridae. Koronavirinae consists of one of two subfamilies in the Coronavirida family. Another example is Torovirinae. Coronaviruses further divided into four genera, alpha, beta, gamma, and delta coronavirus [9].

The city of Wuhan in China is a global concern because of respiratory disease due to coronavirus2019-nCoV. In December 2019, an outbreak of pneumonia caused an unknown in Wuhan, Hubei province in China, by linking the epidemiologist to the Huanan Seafood Wholesale Market, where there are also live animal sales [8]. The WHO notification on December 31, 2019, by Chinese Health Authorities encourages health authorities in Hong Kong, Macao, and Taiwan to increase border surveillance, and incentivize and inhibit it can be a challenge in novels and health assistance for the community [13]. The Chinese health authorities have announced the public health measures, including intensive surveillance, epidemiology-investigation, and market closure on January 1, 2020.

Viruses spread so quickly from day one that they started to become a hot topic in the world. The study says half of all people who are infected with the new coronavirus start 40 to 59 years. The research presented only 10% of patients younger than 39 years. Therefore, children are not as exposed as adults because the epidemic begins during the Chinese New Year holiday - schools when closed.

Viruses have initially sorted into genera based on serology but now shared with phylogenetics. All viruses in the *Nidovirales* sequence are enveloped in non-segmented, sensory-positive RNA viruses. Some viruses have the largest identified RNA genome, containing up to 33.5 kilobases (kb) of the genome. Other standard features in the order of *Nidovirales* are:

1) Very permanent genome organization, with more significant genes than previous structural and accessory genes.

2) Non-structural genes by ribosomes. Frameshifting

3) One of unique or unusual enzymatic activity that is present in large polyprotein replicate-transcriptase.

4) Downstream gene expression by the synthesis of nested sub-genomic mRNAs. The big difference in the Nidovirus family is in the amount, type, and size of structural proteins. These differences cause significant changes in the structure and morphology of nucleo-capsid and virion [7]

### 1.1 CRISP-DM Framework

The CRISP-DM (Cross-Industry Standard Process for Data Mining) described in terms of; a hierarchical process model, which consists of four sets of abstractions (from generic to specific): phases, generic tasks, particular tasks, and sample processes (see figure 1 below). Methodology The CRISP-DM distinguishes between the Reference model and the User's Guide. Where in the Reference Model, represent a brief overview of the phases, tasks, and outputs produced [11].

The life cycle of a data mining project is broken down or determined into six stages, which explained in Figure 1 below. While each sequence presented in Figure 1; showed that the phases are tight. The arrows only change in the essential stages and phases that affect various dependencies between steps that take place, except for specific projects, it is necessary for the results of each aspect that have not yet run in the next process [1].

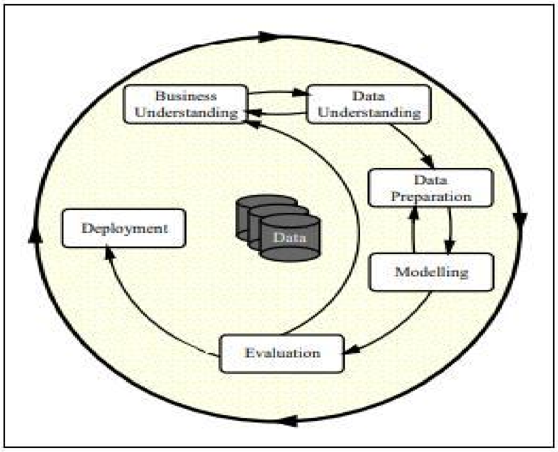
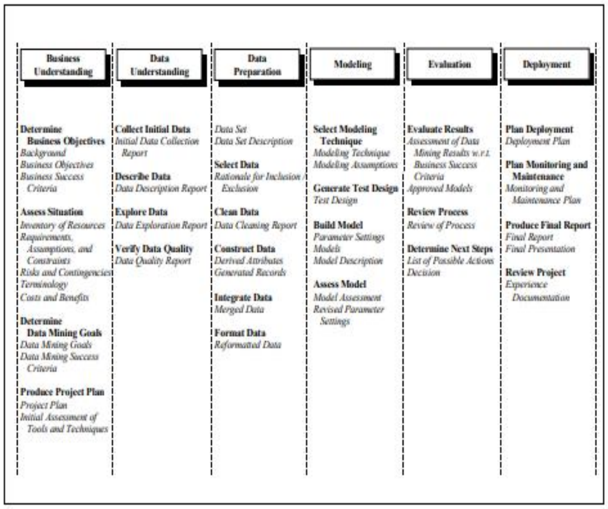


**Figure 1:** Four level Breakdown of the CRISP-DM Methodology for Data Mining

The outer circle in Figure 1 above, symbolizes the environmental circle of data mining. Data mining not completed when a solution achieved; Lessons learned during the process and from solutions deployed given new business questions that asked more focused. Further data on the mining process will benefit from previous experience.



| Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment |
|---|---|---|---|---|---|
| **Determine Business Objectives**<br>*Background*<br>*Business Objectives*<br>*Business Success Criteria*<br><br>**Assess Situation**<br>*Inventory of Resources*<br>*Requirements, Assumptions, and Constraints*<br>*Risks and Contingencies*<br>*Terminology*<br>*Costs and Benefits*<br><br>**Determine Data Mining Goals**<br>*Data Mining Goals*<br>*Data Mining Success Criteria*<br><br>**Produce Project Plan**<br>*Project Plan*<br>*Initial Assessment of Tools and Techniques* | **Collect Initial Data**<br>*Initial Data Collection Report*<br><br>**Describe Data**<br>*Data Description Report*<br><br>**Explore Data**<br>*Data Exploration Report*<br><br>**Verify Data Quality**<br>*Data Quality Report* | *Data Set*<br>*Data Set Description*<br><br>**Select Data**<br>*Rationale for Inclusion / Exclusion*<br><br>**Clean Data**<br>*Data Cleaning Report*<br><br>**Construct Data**<br>*Derived Attributes*<br>*Generated Records*<br><br>**Integrate Data**<br>*Merged Data*<br><br>**Format Data**<br>*Reformatted Data* | **Select Modeling Technique**<br>*Modeling Technique*<br>*Modeling Assumptions*<br><br>**Generate Test Design**<br>*Test Design*<br><br>**Build Model**<br>*Parameter Settings*<br>*Models*<br>*Model Description*<br><br>**Assess Model**<br>*Model Assessment*<br>*Revised Parameter Settings* | **Evaluate Results**<br>*Assessment of Data Mining Results w.r.t. Business Success Criteria*<br>*Approved Models*<br><br>**Review Process**<br>*Review of Process*<br><br>**Determine Next Steps**<br>*List of Possible Actions*<br>*Decision* | **Plan Deployment**<br>*Deployment Plan*<br><br>**Plan Monitoring and Maintenance**<br>*Monitoring and Maintenance Plan*<br><br>**Produce Final Report**<br>*Final Report*<br>*Final Presentation*<br><br>**Review Project**<br>*Experience Documentation* |

**Figure 2**: Overview of the CRISP-DM tasks

The picture (Figure 2) above explains the stages performed in data mining using the CRISP-DM method

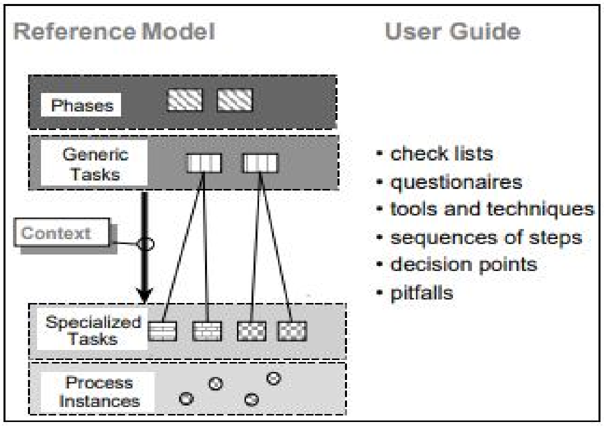
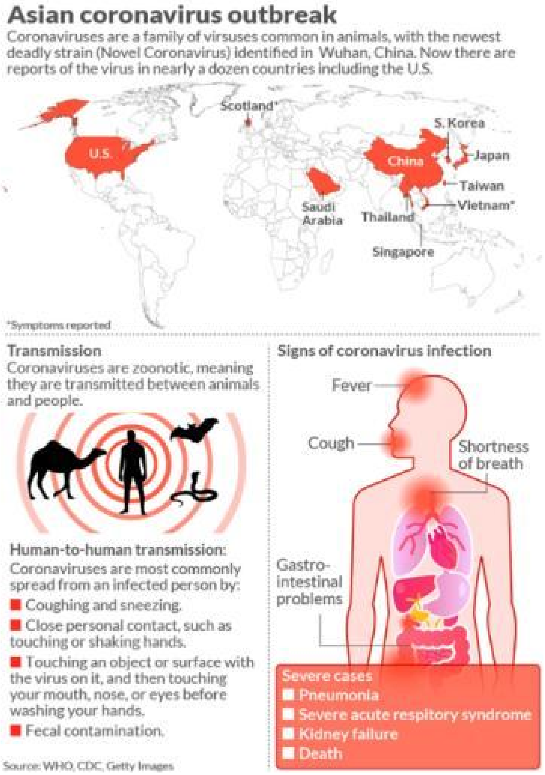


**Figure 3:** Phases of the Current CRISP-DM Process M**o**del

Above is the CRISP-DM Process Model Current Phase (figure 3), which seeks to find the context of the data



Asian coronavirus outbreak

Coronaviruses are a family of virsuses common in animals, with the newest deadly strain (Novel Coronavirus) identified in Wuhan, China. Now there are reports of the virus in nearly a dozen countries including the U.S.

*Symptoms reported

Transmission

Coronaviruses are zoonotic, meaning they are transmitted between animals and people.

Human-to-human transmission:

Coronaviruses are most commonly spread from an infected person by:

- Coughing and sneezing.
- Close personal contact, such as touching or shaking hands.
- Touching an object or surface with the virus on it, and then touching your mouth, nose, or eyes before washing your hands.
- Fecal contamination.

Signs of coronavirus infection

Severe cases
- Pneumonia
- Severe acute respitory syndrome
- Kidney failure
- Death

Source: WHO, CDC, Getty Images

**Figure 4:** Coronavirus Outbreak

The figure 4 above explains the symptoms and transmission of Coronavirus. Spread is very fast in the Corona Virus, causing it to happen everywhere. In this study we will try and discuss the results of research on countries that contaminated with the virus will be announced for the future will find out if there is a rare virus and prevent us from tackling the virus, so we can find solutions that allow strange illnesses to plague [18].

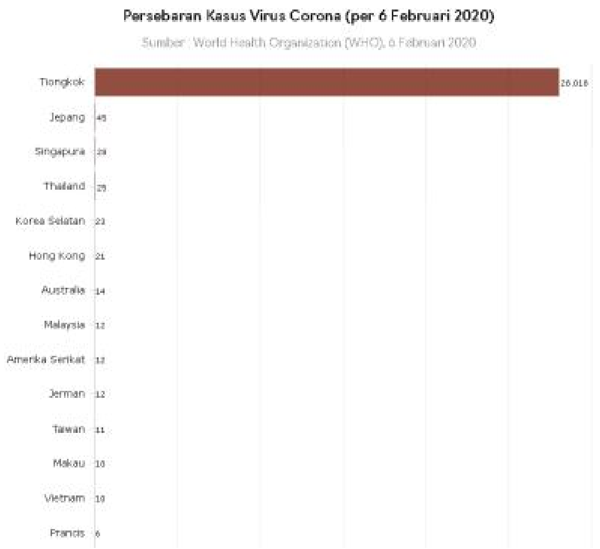


**Figure 5:** Coronavirus Spread-out Graphics

The picture (Figure 5) above is data that has been obtained about the spread of Coronavirus on 6 February 2020 which is so fast in China

## 2. BACKGROUND

The Coronavirus novel (2019-nCoV), better known as the Coronavirus, is a new type of Coronavirus transmitted to humans. This virus first discovered in the city of Wuhan, China, at the end of December 2019. This virus spread quickly and has spread to other regions in China and several countries[14].

Coronaviruses contain unsegmented RNA genomes. The genome comprises a structure of 5 caps along with three poly (A) tails, which allows it transplanted as mRNA for polyprotein replication translation. Gene replication that encodes nonstructural proteins (nsps) places two-thirds of the genome, around 20 kb, in contrast to structural and additional proteins, which only produce about 10 kb of the genomic virus. Tip 5 of the gene contains the leadership sequence and the untranslated region (UTR), which includes several structures. This virus can infect the respiratory system. In many cases, this virus only causes mild respiratory infections, such as flu [3].

This virus can also cause severe respiratory infections, such as pneumonia, Middle East Respiratory Syndrome (MERS), and Severe Acute Respiratory Syndrome (SARS). Coronavirus infection can cause sufferers to change flu, such as the runny and runny nose, headaches, coughing, sore throat, and fever, or severe infectious diseases, such as high fever, cough with phlegm, bleeding, breathing, and chest breathing.

The Chinese health authorities have announced the steps publicly to conduct intensive surveillance, epidemiological investigations, and market closure on January 1, 2020. Researchers can complete 2019-nCoV from patients in a short period on January 7, 2020, and carry out the 2019-nCoV genome sequencing. 2019-nCoV genetic sequence [2].

SARS is a zoonosis caused by SARS-CoV, which first appeared in China in 2002 before spreading to 29 countries in 2003 through a global travel-related outbreak with 8098 cases with a case fatality rate of 9.6%. SARS-CoV nosocomial transmission is also standard in bats and ferrets provided by ferrets on the market in China.

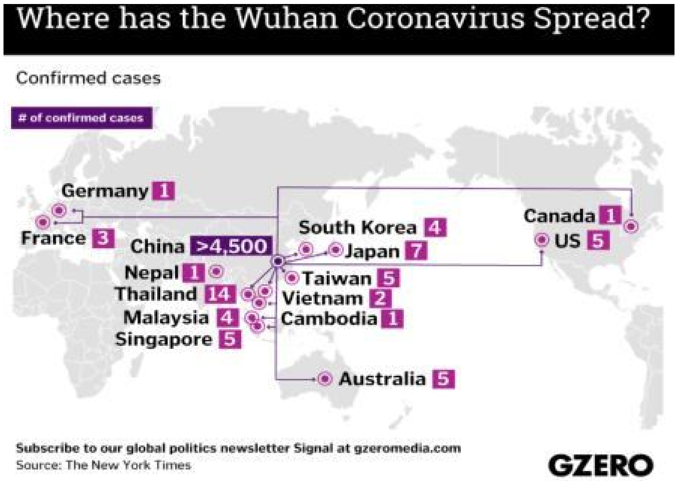


**Figure 6:** The  Wuhan coronavirus spreads

In this figure 6, it is a picture of the initial distribution of Coronavirus

MERS is a deadly novel of zoonotic human diseases endemic to the Middle East, caused by MERS-CoV. Humans considered having MERS-CoV, which cancels through contact with products or camels [13]. The recent outbreak of the pneumonia virus group due to 2019-nCoV in Wuhan Denpasar affects the significance of international health. It may be related to the sale of liar animals as a primary food in humans. On January 10, 2020, 41 patients were diagnosed with infection by indicated animals. The onset of the disease was 41 cases from December 8, 2019, to January 2, 2020. They are related to the condition of fever (> 90%), dry cough (80%), shortness of breath (20%), and respiratory distress (15%) [14].

Several fatal cases occurred in 61 years with stomach tumors and known cirrhosis in the hospital due to respiratory and severe pneumonia. The case has released in Wuhan since January 3, 2020. But the first case outside China was approved on January 13, 2020, in Chinese tourists [2].

## 3. METHODOLOGY

Two central problems affect the performance of K-means, the data discretization method has used and the type of clustering used. Reduction of Trimming Errors for

Enhancing Decision Tree Performance. Different approaches, multiple data collection methods, and different kinds of grouping methods, for example, are in studying the distribution process of a coronavirus spread. Various combinations of discretization methods and decision types and data collection are needed. For which combinations of combinations will give the regions of the country where the coronavirus spread occurs.

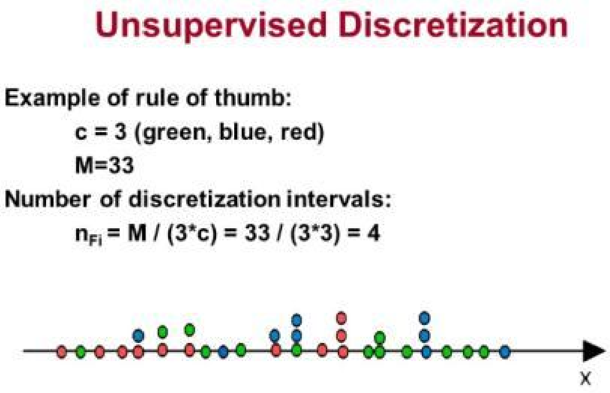
### 3.1 Data Dicretization

Discretization methods categorized as supervised or not supervised—the process of changing numerical data by mapping the data into intervals or concept labels. There are several techniques used, namely histogram analysis binning analysis decision tree analysis. Discretization methods that are not supported do not use information during the discretization process [6].

The discretization method started using class labels. To carry out discretization processes, such as the quadratic-based method and the entropy-based method. All discretization methods used as a step to pre-planning for a continuous attribute in the attribute data set to be discrete. The number of intervals used by the discretization technique is five; each method used for the pre-process of setting data benchmarks for trials in the decision tree [16].

### 3.2 Unsupervised Discretization

By discretizing without supervision, the same width interval and the same frequency will be used equally in extensive testing. The discretization algorithm is the same as the width interval. This algorithm determined the minimum and maximum values obtained from the discretized attribute. By dividing the range of the number of discrete intervals, the user has specified, as like in this figure 7 below.



Unsupervised Discretization

Example of rule of thumb:

c = 3 (green, blue, red)

M=33

Number of discretization intervals:

$n_{Fi}$ = M / (3*c) = 33 / (3*3) = 4

**Figure 7 :** Rule of thumb

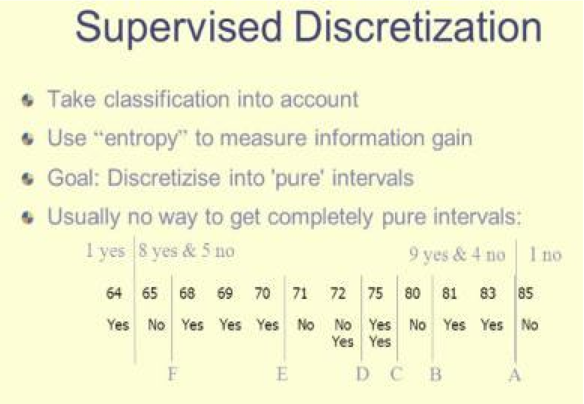
The algorithm with the same and exact frequency can determine the minimum and maximum values of the discrete attribute. By sorting all costs in ascending order and dividing the range into a specified number of users so that each interval containing the same amount of values is then will sort [12].

### 3.3 Supervised Discretization

In discretization, that begins with chi and the most famous discretization methods. The entropy discretization used in this model developed by Fayyad and Keki, Entropy is a measure of uncertainty contained in a training set. The process for selecting the selection method based on the unentropic way to choose the boundary to discretization. The samples sorted in ascending order and then entropy for each candidacy point to be counted. Positions taken approved for approval. In this model, the criteria stop at up to five attribute intervals.

### 3.4 Data Multiple Classifier

Multi classifier data retrieval makes training data into smaller data subsets and builds Decision Tree classifiers for each data subset. Applying a collection data algorithm for classification proves an increase in the success in the ranking of classifiers. This research has taken from the data collection section, and one of the rule steps was depicted in figure 8, where data collected between three and eleven parts. For each discretization method for each type of decision tree, which made the most successful, it would inform [4].



Supervised Discretization

- Take classification into account
- Use "entropy" to measure information gain
- Goal: Discretizise into 'pure' intervals
- Usually no way to get completely pure intervals:

| 64 | 65 | 68 | 69 | 70 | 71 | 72 | 75 | 80 | 81 | 83 | 85 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Yes | No | Yes | Yes | Yes | No | No<br>Yes | Yes<br>Yes | No | Yes | Yes | No |

**Figure 8:** Supervised discretization rules

There are many types of Decision trees. The difference is the mathematical model used to select the attributes added in expanding the decision tree rule structure. In this study, three types of testing most commonly used, namely, Information Strengthening, Gini Index, and Ratio Tree Strengthening, each of which discussed below.

### 3.5 Entropy Approach

Choose Entropy by agreeing on information conceptualized by selecting the slit attribute that replaces the entropy value, thereby maximizing Information Gain. To get the amount of Information Gain and Rain Gain, we must first calculate the cost of Entropy. Entropy use (Figure 9) to measure the inequality of objects on each branch based on an attribute [5].

**Figure 9 :** Entropy

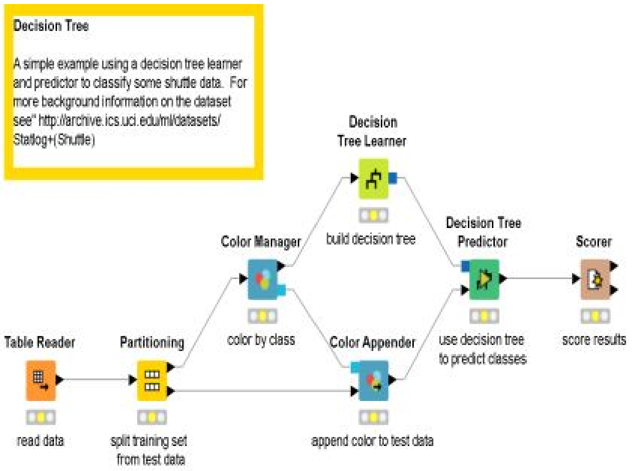
Identifying the placement attributes of the Decision Tree, one must calculate information for each quality and then choose attributes that additional information. Information reinforcement for each variety computed using the formula as shown below

$$E = \sum_{i=1}^{k} Pi \ \log_2 Pi$$

### 3.6 Trimming Data

To reduce the bias results from the use of Gain Information, a variant known as Gain Ratio introduced by Australian academic Ross Quinlan. Measurement of Information Strengthening is a test that is biased towards many results. That is, preferring to choose attributes that have large values. Gain Ratios adjust Gain Information for each attribute to free the breadth and uniformity of attribute values.

After extracting the decision tree rules, the pruning error rule uses to trim the derived decision rule. Proven error pruning is one of the fastest and most workable pruning methods. Applying pruning errors provides a more compact decision and reduces the number of extracted rules[15].


Decision Tree

A simple example using a decision tree learner and predictor to classify some shuttle data. For more background information on the dataset see" http://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle)

**Figure 10:** Decision tree step uses Knime analytic

As we can see in Figure 10, we need to calculate each combination, sensitivity, specificity, and Accuracy calculated, and sensitivity is a real comparison questioned positive posed is the ratio of sick people who try as ill. Uniqueness is the proportion of adverse events that are approved correctly. The negative request is the ratio of authorized health people, and the Accuracy is an example of a contribution passed successfully [10].

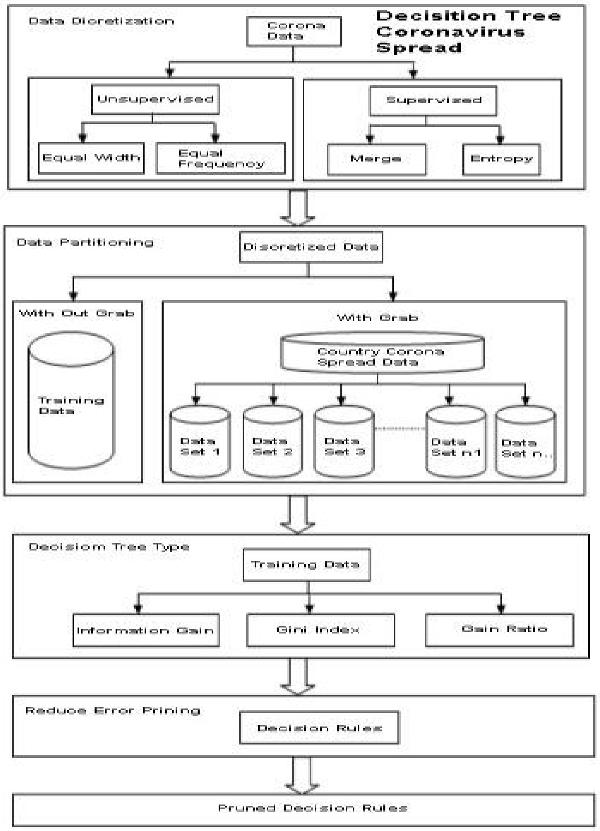


**Figure 11:** Decision tree corona virus spreads

In figure 11, that's flow steps in a decision tree to analyze the coronavirus data spreads

### 3.7 Summary

The research process involves discretization of data, data partitioning, selection of Decision Tree types, and the application of error pruning that results in trimmed Decision Trees. Data discretization divided into protected and not supervised methods. Methods initiated to involve the same width and the same frequency of compilation of the discretization method that begins involving both joints and entropy.

Decision trees are also useful for discussing data, finding hidden relationships between input variable estimates and target variables. Decision tree combines data tracking and modeling, so it is terrific as a first step in the modeling process so that it is complicated used as the final model of several other techniques.

Data partitions took with and without data retrieval. Three types of Decree Trees agreed, namely, reinforcing information, Gini index, and reinforcement ratio. Finally, we were decide pruning errors applied to everything extracted from the training data. The actual test is carried out executing each variant of each element in combination against all data combinations. Each option is then issued by itself and through various data sharing allocations.

The results of each variant through each data collection partition have approved for reporting errors that are applied. Overall, the Decision Tree is executed on a set of data to compile the findings presented here, related to the number of advantages, and it does not mean this method has no shortcomings. This decision tree can overlap; most of the criteria used are large and complicated in each predetermined input process. This result, of course, can increase the loading time by the amount of memory needed [17].

## 4. DATA

The data used in this study is the data set available on recap data from various portals and some inspired data from several web sites accessible at https: //db.cngb.org. Data sets have raw attributes. However, all experiments published are only experiments on that data. To facilitate comparisons with the literature, a revised test for this same attribute, which can see in Table 1. The data set contains 368 rows, but there are several values added to the data set that are the primary research material to find the various contradictions.

The machine must know which data set to look for to find a solution that helps it and which data set can be used to achieve the appropriate goal. This collection of data for approval is the Test-Set, while the data set for attaining it is called the Training-Set.This concept set training uses to create a machine learning model, while the Test-Set will be used to improve performance and correctness. This concept will increase the retrieval time according to the amount of memory needed.

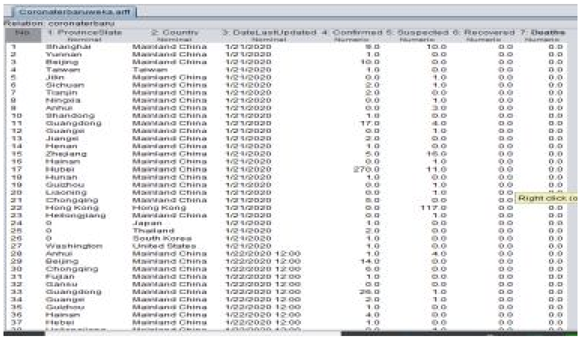
**Table 1:** Selected Data Set Attributed

| Name | Type | Description |
|---|---|---|
| Province | Continuous | Provinces in the country that spread the Corona virus |
| Country | Continuous | Country that spread the Corona virus |
| Date | Continuous | Last updated data of people affected by Corona Virus |
| Confirmed | Continuous | Continuous data on how many people have been confirmed to be affected by the corona virus |
| Suspected | Discrete | Suspect of a person affected by the corona virus<br>0 = No<br>1 = Yes |
| Recovered | Discrete | People who recover after being exposed to the corona virus<br>0 = No<br>1 = Yes |
| Deaths | Discrete | People who died after being exposed to the corona virus<br>0 = No<br>1 = Yes |

### 4.1. Dataset Description and Pre-Processing

We got a dataset about coronavirus outbreak from the Kaggle.com website. We use Coronavirus infection data published from  December 2019 to February 2020. This data has been published and categorized into several categories. For the overall category, the information provided: country, city, last updated, confirmed, suspected, recovered, and dead. Information that becomes specific information:

- Last updated: symptoms, current status, and patients who have had previous contacts with Coronavirus confirmed
- Recovery and death: patients who previously suffered from Coronavirus

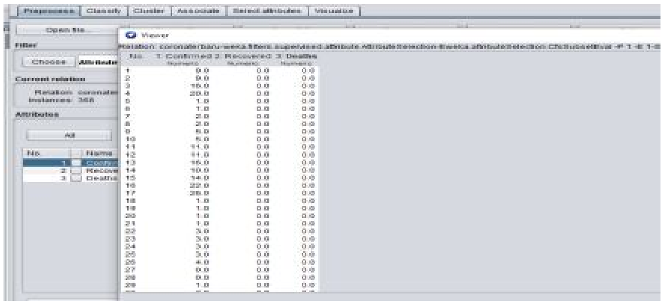


Relation: coronaterbaru

| No. | 1: ProvinceState<br>Nominal | 2: Country<br>Nominal | 3: DateLastUpdated<br>Nominal | 4: Confirmed<br>Numeric | 5: Suspected<br>Numeric | 6: Recovered<br>Numeric | 7: Deaths<br>Numeric |
|---|---|---|---|---|---|---|---|
| 1 | Shanghai | Mainland China | 1/21/2020 | 9.0 | 10.0 | 0.0 | 0.0 |
| 2 | Yunnan | Mainland China | 1/21/2020 | 1.0 | 0.0 | 0.0 | 0.0 |
| 3 | Beijing | Mainland China | 1/21/2020 | 10.0 | 0.0 | 0.0 | 0.0 |
| 4 | Taiwan | Taiwan | 1/21/2020 | 1.0 | 0.0 | 0.0 | 0.0 |
| 5 | Jilin | Mainland China | 1/21/2020 | 0.0 | 1.0 | 0.0 | 0.0 |
| 6 | Sichuan | Mainland China | 1/21/2020 | 2.0 | 1.0 | 0.0 | 0.0 |
| 7 | Tianjin | Mainland China | 1/21/2020 | 2.0 | 0.0 | 0.0 | 0.0 |
| 8 | Ningxia | Mainland China | 1/21/2020 | 0.0 | 1.0 | 0.0 | 0.0 |
| 9 | Anhui | Mainland China | 1/21/2020 | 0.0 | 3.0 | 0.0 | 0.0 |
| 10 | Shandong | Mainland China | 1/21/2020 | 1.0 | 0.0 | 0.0 | 0.0 |
| 11 | Guangdong | Mainland China | 1/21/2020 | 17.0 | 4.0 | 0.0 | 0.0 |
| 12 | Guangxi | Mainland China | 1/21/2020 | 0.0 | 1.0 | 0.0 | 0.0 |
| 13 | Jiangxi | Mainland China | 1/21/2020 | 2.0 | 0.0 | 0.0 | 0.0 |
| 14 | Henan | Mainland China | 1/21/2020 | 1.0 | 0.0 | 0.0 | 0.0 |
| 15 | Zhejiang | Mainland China | 1/21/2020 | 5.0 | 16.0 | 0.0 | 0.0 |
| 16 | Hainan | Mainland China | 1/21/2020 | 0.0 | 1.0 | 0.0 | 0.0 |
| 17 | Hubei | Mainland China | 1/21/2020 | 270.0 | 11.0 | 0.0 | 0.0 |
| 18 | Hunan | Mainland China | 1/21/2020 | 1.0 | 0.0 | 0.0 | 0.0 |
| 19 | Guizhou | Mainland China | 1/21/2020 | 0.0 | 1.0 | 0.0 | 0.0 |
| 20 | Liaoning | Mainland China | 1/21/2020 | 0.0 | 1.0 | 0.0 |  |
| 21 | Chongqing | Mainland China | 1/21/2020 | 5.0 | 0.0 | 0.0 |  |
| 22 | Hong Kong | Hong Kong | 1/21/2020 | 0.0 | 117.0 | 0.0 | 0.0 |
| 23 | Heilongjiang | Mainland China | 1/21/2020 | 0.0 | 1.0 | 0.0 | 0.0 |
| 24 | 0 | Japan | 1/21/2020 | 1.0 | 0.0 | 0.0 | 0.0 |
| 25 | 0 | Thailand | 1/21/2020 | 2.0 | 0.0 | 0.0 | 0.0 |
| 26 | 0 | South Korea | 1/21/2020 | 1.0 | 0.0 | 0.0 | 0.0 |
| 27 | Washington | United States | 1/21/2020 | 1.0 | 0.0 | 0.0 | 0.0 |
| 28 | Anhui | Mainland China | 1/22/2020 12:00 | 1.0 | 4.0 | 0.0 | 0.0 |
| 29 | Beijing | Mainland China | 1/22/2020 12:00 | 14.0 | 0.0 | 0.0 | 0.0 |
| 30 | Chongqing | Mainland China | 1/22/2020 12:00 | 6.0 | 0.0 | 0.0 | 0.0 |
| 31 | Fujian | Mainland China | 1/22/2020 12:00 | 1.0 | 0.0 | 0.0 | 0.0 |
| 32 | Gansu | Mainland China | 1/22/2020 12:00 | 0.0 | 0.0 | 0.0 | 0.0 |
| 33 | Guangdong | Mainland China | 1/22/2020 12:00 | 26.0 | 1.0 | 0.0 | 0.0 |
| 34 | Guangxi | Mainland China | 1/22/2020 12:00 | 2.0 | 1.0 | 0.0 | 0.0 |
| 35 | Guizhou | Mainland China | 1/22/2020 12:00 | 1.0 | 0.0 | 0.0 | 0.0 |
| 36 | Hainan | Mainland China | 1/22/2020 12:00 | 4.0 | 0.0 | 0.0 | 0.0 |
| 37 | Hebei | Mainland China | 1/22/2020 12:00 | 1.0 | 0.0 | 0.0 | 0.0 |

**Figure 12:** Number of Coronavirus cases in several countries

By selecting data (data selected from Figure 12), we used the cleaning feature on Weka by selecting a filter in the preprocess installation, which is the attribute selection filter. With the hope of getting the data obtained in the three most highlighted attributes; Death, Confirmed, Recovered. With the table display as follows:



Relation: coronater
Instances: 368

| No. | Name |
|---|---|
| 1 | Confir |
| 2 | Recove |
| 3 | Deaths |

Relation: coronaterbaru-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.CfsSubsetEval -P 1 -E 1-

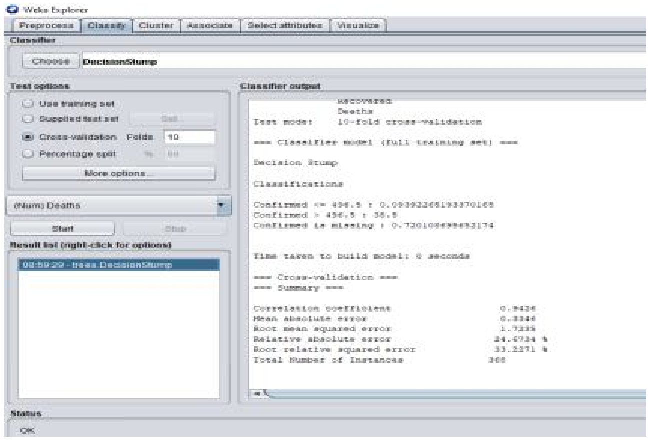
| No. | 1: Confirmed<br>Numeric | 2: Recovered<br>Numeric | 3: Deaths<br>Numeric |
|---|---|---|---|
| 1 | 9.0 | 0.0 | 0.0 |
| 2 | 9.0 | 0.0 | 0.0 |
| 3 | 16.0 | 0.0 | 0.0 |
| 4 | 20.0 | 0.0 | 0.0 |
| 5 | 1.0 | 0.0 | 0.0 |
| 6 | 1.0 | 0.0 | 0.0 |
| 7 | 2.0 | 0.0 | 0.0 |
| 8 | 2.0 | 0.0 | 0.0 |
| 9 | 6.0 | 0.0 | 0.0 |
| 10 | 5.0 | 0.0 | 0.0 |
| 11 | 11.0 | 0.0 | 0.0 |
| 12 | 11.0 | 0.0 | 0.0 |
| 13 | 16.0 | 0.0 | 0.0 |
| 14 | 10.0 | 0.0 | 0.0 |
| 15 | 14.0 | 0.0 | 0.0 |
| 16 | 22.0 | 0.0 | 0.0 |
| 17 | 26.0 | 0.0 | 0.0 |
| 18 | 1.0 | 0.0 | 0.0 |
| 19 | 1.0 | 0.0 | 0.0 |
| 20 | 1.0 | 0.0 | 0.0 |
| 21 | 1.0 | 0.0 | 0.0 |
| 22 | 3.0 | 0.0 | 0.0 |
| 23 | 3.0 | 0.0 | 0.0 |
| 24 | 3.0 | 0.0 | 0.0 |
| 25 | 3.0 | 0.0 | 0.0 |
| 26 | 4.0 | 0.0 | 0.0 |
| 27 | 0.0 | 0.0 | 0.0 |
| 28 | 0.0 | 0.0 | 0.0 |
| 29 | 1.0 | 0.0 | 0.0 |

**Figure 13:** Pre-process step uses Weka

The figure 3 above is a display of data processed in preprocess with Weka.

### 4.2. Analysis using the Stump Decision Algorithm

The Decision Stump operator is used to produce a decision tree with only one single separation. The resulting tree uses to classify examples that are not visible. This operator can be very efficient when upgraded with operators such as the AdaBoost operator. The Example of the Example given has several attributes, and each instance belongs to the class (like yes or no). The leaf node of the decision tree contains the name of the course, while the non-leaf node is the decision node. A decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute (please take a look in figure 14).



Classifier: DecisionStump

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) Deaths

08:59:29 - trees.DecisionStump

Classifier output

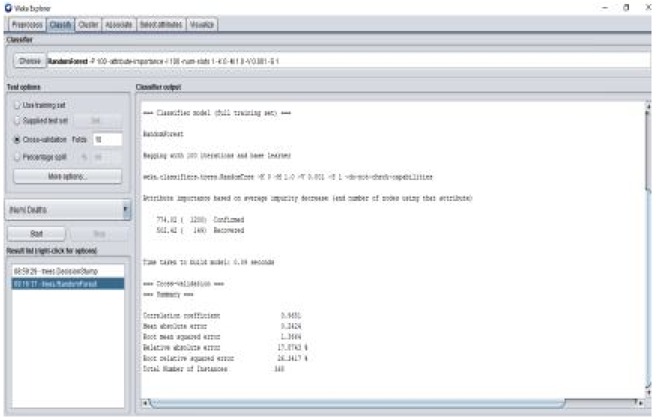
```
                Recovered
                Deaths
Test mode:      10-fold cross-validation

=== Classifier model (full training set) ===

Decision Stump

Classifications

Confirmed <= 496.5 : 0.09392265193370165
Confirmed > 496.5 : 38.5
Confirmed is missing : 0.7201086956521 74


Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9426
Mean absolute error                      0.3346
Root mean squared error                  1.7235
Relative absolute error                 24.6734 %
Root relative squared error             33.2271 %
Total Number of Instances              368
```

Status: OK

**Figure 14:** Decision stamp uses Weka

### 4.3. Analysis using Random Forest

The Random Forest Algorithm as in figure 15, is a supervised learning algorithm. The "forest" built was a decision tree ensemble, usually trained by the "pocketing" method. The general idea of the bagging method is that a combination of learning models improves overall results



Classifier: RandomForest -P 100 -attribute-importance -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) Deaths

08:59:29 - trees.DecisionStump
[illegible] - trees.RandomForest

Classifier output

```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Attribute importance based on average impurity decrease (and number of nodes using that attribute)

    774.02 (  1200)  Confirmed
    501.42 (   148)  Recovered


Time taken to build model: 0.[illegible] seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                 0.9651
Mean absolute error                     0.2424
Root mean squared error                 1.3984
Relative absolute error                17.8743 %
Root relative squared error            26.3417 %
Total Number of Instances             368
```

**Figure 15:** Random forest uses Weka

### 4.4. Knowledge Flow uses the KNIME Analytics Platform

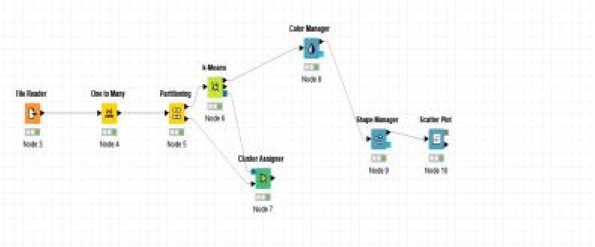


**Figure 16:** The example of knowledge flow uses the Knime analytics

In this analysis, use KNIME to find out the flow (figure 16) of knowledge created using the K-means algorithm. This node produces cluster centers for a predetermined number of clusters (no amount of dynamic clusters). K-means perform sharp clustering that delivers vector data to exactly one group. The algorithm ends when the cluster assignment does not change again. The clustering algorithm uses Euclidean distances on selected attributes [4].

The K-means algorithm process generates the following table

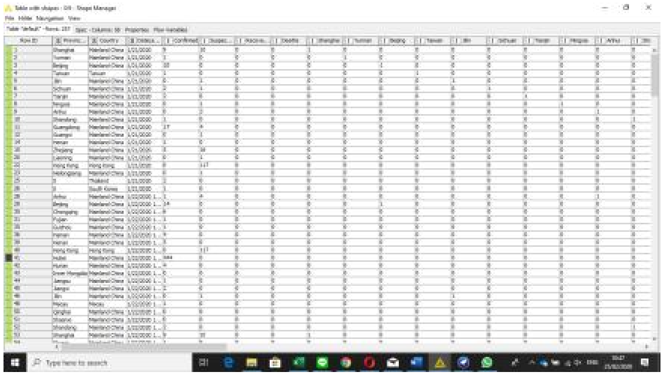


**Figure 17:** The K-means algorithm The K-means algorithm Scatter plot visualization Related to the final results of data mining using Coronavirus

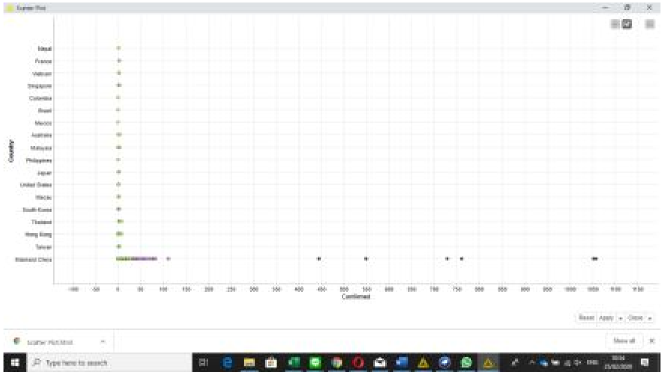


**Figure 18:** Scatterplot visualization results

The figure above has shown the results of K-means algorithm process in the scatterplot

## 5. CONCLUSION

The sensitivity, accuracy, and specificity results in determining the number of countries indicated using the same width, same frequency, chi merger, and entropy discretization with Information Gain, Gini Index, and Decision Making Ratios of Three Trees and correcting pruning. Counting methods for calculating two different requirements, counting the number cannot be combined and cannot compare between two groups.

Data review carried out to review the elements of data formation and estimates make sense in calculations and calculate the number of targets to be obtained based on the indicated country data, and the coronavirus has distributed. If an error is possible, it can be corrected, where the correction might not have data deleted.

## REFERENCES

[1] Anwar Lashari, S., Ibrahim, R., Senan, N., & Taujuddin, N. S. A. M. (2018). Application of Data Mining Techniques for Medical Data Classification: A Review. *MATEC Web of Conferences*, *150*, 1–6. https://doi.org/10.1051/matecconf/201815006003

[2] Backer Jantien, A., Don, K., & Jacco, W. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China. *Euro Surveill*, *25*(5), 20–28. https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062

[3] Carlos, W. G., Dela Cruz, C. S., Cao, B., Pasnick, S., & Jamil, S. (2020). Novel Wuhan (2019-nCoV) Coronavirus. *American Journal of Respiratory and Critical Care Medicine*. https://doi.org/10.1164/rccm.2014p7

[4] Dedić, N., & Stanier, C. (2017). Measuring the success of changes to Business Intelligence solutions to improve Business Intelligence reporting. *Journal of Management Analytics*, *4*(2), 130–144. https://doi.org/10.1080/23270012.2017.1299048

[5] Foley, É., & Guillemette, M. G. (2011). What is Business Intelligence? *International Journal of Business Intelligence Research*, *1*(4), 1–28. https://doi.org/10.4018/jbir.2010100101

[6] Hoptroff, R., & Kudyba, S. (2013). Data Mining and Business Intelligence. In *Data Mining and Business Intelligence*. https://doi.org/10.4018/978-1-930708-03-7

[7] Hui, D. S., I Azhar, E., Madani, T. A., Ntoumi, F., Kock, R., Dar, O., … Petersen, E. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases*, *91*, 264–266. https://doi.org/10.1016/j.ijid.2020.01.009

[8] Jiang, S., Xia, S., Ying, T., & Lu, L. (2020). A novel coronavirus (2019-nCoV) causing pneumonia-associated respiratory syndrome. *Cellular & Molecular Immunology*, *2001316*(February), 2001316. https://doi.org/10.1038/s41423-020-0372-4

[9] Kim, J. Y., Choe, P. G., Oh, Y., Oh, K. J., Kim, J., Park, S. J., … Oh, M. D. (2020). The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures. *Journal of Korean Medical Science*, *35*(5), e61. https://doi.org/10.3346/jkms.2020.35.e61

[10] Kolowitz, B. J., & Medical, P. (2011). Enabling Business Intelligence, Knowledge Management and Clinical Workflow With Singleview. *Issues in Information Systems*, *12*(1), 70–77.

[11] Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, *36*(5), 700–710. https://doi.org/10.1016/j.ijinfomgt.2016.04.013

[12] Lemahieu, W., vanden Broucke, S., & Baesens, B. (2019). Data Warehousing and Business Intelligence. In *Principles of Database Management* (pp. 551–589). https://doi.org/10.1017/9781316888773.019

[13] Lippi, G., & Plebani, M. (2020). The novel coronavirus (2019-nCoV) outbreak: think the unthinkable and be prepared to face the challenge. *Diagnosis*, *0*(0). https://doi.org/10.1515/dx-2020-0015

[14] Liu, T., Hu, J., Kang, M., Lin, L., Zhong, H., Xiao, J., … Ma, W. (2020). Transmission dynamics of 2019 novel coronavirus (2019-nCoV). *BioRxiv*, 2020.01.25.919787. https://doi.org/10.1101/2020.01.25.919787

[15] Olszak, C. M., & Batko, K. (2012). The use of business intelligence systems in healthcare organizations in Poland. *2012 Federated Conference on Computer Science and Information Systems, FedCSIS 2012*, 969–976.

[16] Ranjan, J. (2007). Need for real time Business Intelligence. *International Journal of Agile Systems and Management*, *2*(4), 425–443. https://doi.org/10.1504/IJASM.2007.015841

[17] Schermann, M., Hemsen, H., Buchmüller, C., Bitter, T., Krcmar, H., Markl, V., & Hoeren, T. (2014). An interdisciplinary opportunity for information systems research. *Business and Information Systems Engineering*, *6*(5), 261–266. https://doi.org/10.1007/s12599-014-0345-1

[18] Schieppati, A., Henter, J. I., Daina, E., & Aperia, A. (2008). Why rare diseases are an important medical and social issue. *The Lancet*, *371*(9629), 2039–2041. https://doi.org/10.1016/S0140-6736(08)60872-7