Journal of Theoretical and Applied Information Technology
31st August 2020. Vol.98. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMIZING DECISION TREE CRITERIA TO IDENTIFY THE RELEASED FACTORS OF COVID-19 PATIENTS IN SOUTH KOREA

ISTON DWIJA UTAMA [1], IVAN DIRYANA SUDIRMAN [2]
[1] Entrepreneurship Department, BINUS Business School Undergraduate Program, Bina Nusantara University Jakarta, Indonesia
[2] Entrepreneurship Department, BINUS Business School Undergraduate Program, Bina Nusantara University Jakarta, Indonesia

ABSTRACT:

This study aims to identify the released factors for patients infected by SARS CoV 2 virus. By taking samples from South Korea Country, then we were using the data mining process for classifying problems by optimizing the decision tree model from RapidMiner Studio educational 9.6.00 edition. Based on our study showed that the accuracy of this model is 80.46% +/- 1.68% (micro average 80.46%). The result for this study showed that location or region is a dominant factor rather than age and sex, therefore the South Korea policy to take the drive-through, walk-through, tracing, and use digital maps to let society aware is effective to prevent the spreads of diseases. The Implication of this study reconfirms that stay away from an infected area or social distancing such as staying at home is the right decision to minimize and reduce pandemic spreads, and other countries can adopt or modify the strategy that already did in South Korea.

Keywords: *SARS CoV2, South Korea, Data Mining, Decision Tree Model*

1. INTRODUCTION

The spread of a new virus called SARS CoV 2 started in China in 2019. Then the virus outbreak has an outbreak beyond China. The spread of new viruses in China started to be suppressed as this work was published, but the spread of viruses outside of China continued to increase. South Korea, Iran, and Italy are the three non-China countries with most Covid-19 outbreaks, the name of the disease caused by the SARS CoV 2 virus.

There are 275,125 cases of Coronavirus, 11,376 deaths and 90,943 recovered, according to worldometers.info when this report was published. There are 80,967 cases in China, 8,652 cases in South Korea, 47,021 cases in Italy and 19,644 cases in Iran [1]. Although there are many cases across the world, the mortality rate is lower than SARS (Severe Acute Respiratory Syndrome) or MERS (Middle East Respiratory Syndrome), according to the WHO report [2].

While this virus death rate is comparatively low compared to SARS and MERS, it appears that the rate of spread of the SARS CoV 2 virus may impact the economies of affected countries. As per the report on 21st March 2020, There are 450 cases in Indonesia and the tourism sector begins to feel its impact. Occupancy rates of hotels fell by 40%. The retail industry has the potential to lose USD 48 million in turnover. The effect will be felt by more than 495 product forms or 13 percent of commodities for export to China. It is expected that as many as 299 imported products from China will decline or even vanish from the Indonesian market as a result of this virus[3].

SARS CoV 2 virus is a modern virus, and there is currently no effective medicine to treat virus-induced diseases. Much of this virus is still unclear, how it infects others, what factors are most important in causing death, whether or not the probability of death can be predicted. South Korea is one of the countries that has managed to control mortality from this virus. At 21st March 2020, South Korea have 8.799 cases with “only” 102 death, which is one of the lowest death rates among other Asian countries. Every country have a different approach to handle this situation. South Korea, in this case, their government prefer to take the massive rapid test for all citizen in a certain area which have confirmed Covid-19 status by the drive-through and walk-through diagnostic testing[4],

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

meanwhile in other regions of Asian countries such as India and Indonesia, they have a different policy to manage the spreads of Covid-19 case. In India, for example, the government prefers to take the lockdown policy, this policy may take multi-dimensional consequences in social, public health system, and economy [5]. In Indonesia itself, this diseases firstly occur in March 2020, then after this occurrence, government take action immediately by providing some regulation and policy such as reallocating budget, create a special task force, the prohibition of a group or mass gathering and activity in the public and local environment, tax incentive for tax holders which affected by Covid-19, and big scale social restrictions policy [6]

Recently, most of the businesses can gather a massive amount of data from everywhere, not only in business sectors, other sectors such as society, science, and technical engineering, medical, and politics also can take advantage of it. The computerization in every daily aspect of life and the fast technological development in terms of data collection and collection tools resulted in the growth of explosive data volume [7]. Data mining involves a standardized research approach to help translate business issues, propose appropriate data transformations and data mining techniques, and provide tools for assessing results effectiveness and documenting experience [8]. Several researchers have been using data mining techniques to extract insights from data. The improved ability of computers to process data and the Internet speed makes the use of data mining more widespread.[9]. Four types of patterns can be identified by data mining techniques, namely association, correlation, cluster, sequential. In this research, we will be performing a data mining type of classification, which is the most commonly used technique of data mining. Classification forms part of the family of machine learning and also employs supervised learning.

There are several classification techniques in data mining, some of which are decision tree analysis, statistical analysis, neural networks, and bayesian classifiers. Each technique offers its own advantages. Bayesian classifier model are popular technique that used by researchers in machine learning due to their simplicity in allowing each attribute variables to contribute toward the final decision independently and equally from other attribute variables. However, the conditional independence assumption in this model is strong and can make Bayesian classifiers unable to process two or more attribute variables of evidence together, however if this technique use in the appropriate domain of study or research, this technique offer quick data training, fast data analysis, and proper decision making as well as relevant interpretation of test result [10].

decision tree can cope with several combinations of terms and can create an impressive test result in many kind of domain study or research. This technique offer genuine simplicity of model interpretations and help their users to consider some important variables in dataset first by placing it into the top of tree [10].

Neural networks technique are a powerful technique for showing complex relationships between inputs and outputs. Neural networks are complicated and for certain domains, they can be massive because containing a large number of nodes and synapses. This technique has difficulties in understanding and time-consuming in the decision-making process. For a better classification performance, researcher have an option to combine the neural networks with support vector machines [10]

The previous study, which compares Bayesian classifiers, decision tree, and neural networks to analyze and predict the heart diseases case showed that neural network is the most accurate rather than decision tree and Bayesian classifiers [11], detail about result can be seen at table 1.

Table 1. Comparison of Bayesian Classifiers, Decision Trees, and Neural Networks

Classification techniques	Accuracy
Bayesian classifier	90.74%
Decision trees	99.62%
Neural networks	100%

Another study that conduct by [10] compares the performance measurement with several parameters for decision tree and bayesian classifier can be seen in the table 2.

Table 2. Comparison of Bayesian Classifier and Decision Tree Classifier

Classification techniques	Accuracy	Precision	Recall	F-Measure
Bayesian classifier	95.20%	99.37%	95.23 %	97.26%

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Decision tree classifier	94.85%	98.31%	95.90 %	97.09%

The result from table 1 dan 2 showed that all of the classifiers achieve an impressive result in all attributes of measurement.

Before researchers choose one from several classifier techniques, researchers try to consider some advantages and limitations of it as we mention on the table 3 [12]

Table 3. Advantages and Limitations of Classifier Techniques

Classifiers Techniques	Advantages	Limitations
Decision tree	1. Model relatively easy to be interpreted, 2. Easy to implement, 3. Can be used for discrete and continuous values 4. Deals with noise	1. A small dataset can lead to a different decision tree, 2. Need large numbers of the training dataset, 3. Overfitting
Bayesian classifiers	1. Simple to implement, 2. Strong computational efficiency and classification rate 3. Predicts the accurate result for most of problem classification and prediction	1. The algorithm precision can be decreased if the dataset is few, 2. To generating a good result, it requires a huge number of data records
Neural networks	1. Easy to use and can be adjust with few numbers of parameters, 2. It is no needed to learns and reprogramming, 3. Easy to implement, 4. Appropriate with the wide range of real life problems.	1. Needs high processing time when a neural network is large, 2. Hard to know how many layers and neurons are mandatory, 3. Learning process can be slow.

According to the explanation and comparison of three kinds of classifiers techniques above, in this study, we intend to find the process behind the data of patients who have recovered or released by using the decision tree analysis technique and will see how accurate the model is with the results.

Each country has different policies and regulations implementation, which is more effective and efficient against this pandemic. By study the pattern of released factors Covid-19 patients in South Korea as a sample, other countries may learn and can take an approach which appropriate and fit with the behavior of citizen in their countries. We also might offer several policy suggestions that might be effectively applied to manage and handle the spreads of the virus.

2. THEORY AND METHODOLOGY

Before exploring more about data mining, we should be aware of some terms about data mining. There is various type of data sets such as [7]:

a. Record, such as relational records, data matrix, data documentation, and data transaction

b. Graph and network, such as world wide web (www), social or network information, and molecular structures.

c. Ordered, such as data video, temporal data, sequential data, and genetic sequence data.

d. Spatial, image, and multimedia.

In data sets, there are attributes (or dimensions, or variables, or features) that can be divided into several types such as [7]:

a. Nominal

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This type can be as a states, categories, or "name of things". Example: occupation, zip codes, marital status, and ID Numbers

b. Binary

This type of attribute has only two kinds of states (0 and 1). This type of attributes separates into two kinds, first is symmetric binary, which both of outcomes are equally important such as gender, and second is asymmetric binary, which both of outcomes are not equally important such as in medical result test (positive or negative of diseases)

c. Numeric

This type of attribute can be divided into two kinds, first is interval, which measures on a scale of equal-sized (standardized) units and there is no true zero-point such as temperature in Celsius or Fahrenheit, and second is ratio, which has inherent zero-points such as temperature in length, Celsius, counts, and monetary quantities. For example, we can notice that 10 meters are larger than 5 meters.

Data mining is an innovative method in which data are identified with real, new, potentially helpful and ultimately meaningful patterns in structured databases. [13]. Data mining is part of broad and complicated data sets for the collection of numerical and categorical data. The term also refers to more advanced methods, including text, web, and spatial data. [14].

Data mining is concerned with current data from the Associations and patterns with study, recognition, and establishment. And data mining is therefore defined as an identification method for patterns in data that can produce untested predictions of unknown patterns [15]. Data mining also can be defined as the process of finding trends and patterns that previously unknown in the database and using them to construct predictive models. For instance, it also can be described as data selection and discovery using a huge amount of data stores to reveal previously unknown patterns or knowledge [7]. Data Mining can be used in many areas of knowledge such as machine learning, database and data warehouse technology, information retrieval, statistics, neural networks, knowledge-based systems, pattern recognition, high-performance computing, image and signal processing, artificial intelligence, spatial or temporal data analysis, and data visualization [7]

There are different uses for data mining and basic statistical analysis. Classical statistical approaches mostly concentrate on testing specified hypotheses and data manipulation looks at other potential hypotheses, often unknown. [16]. The only way to increase knowledge and understanding of the ever-growing number of digital data would be by incorporating the statistical and data mining approach as Witten and others.[15] referred to, In the future, a combination of data mining and statistical approach would not only be appropriate for analyzing different and complex information but also to combine disciplines and techniques such as pattern recognition, bases of knowledge, artificial intelligence and machine learning algorithms.

One of the most commonly used classifications for classifying problems is the decision tree. We use the decision tree to map a finding of its objective value from comments on an element. Leaves contain classifications (also known as labels) within tree structures; non-leaf nodes are features and branches contain combinations of features leading to classifications [17]. Commonly, the procedure of classification will be made based on the following steps [18]

a. Define the classification classes,

b. Select the features,

c. Sampling of training data, this can be made by supervised learning, unsupervised learning, semi-supervised learning, and active learning,

d. Estimate the universe statistics, we have to create a proper decision rule in this step,

e. Classification, several classifications can be made such as Multi-level slice classifier, Decision tree classifier, Minimum distance classifier, Maximum likelihood classifier, Other classifiers techniques such as the expert system of fuzzy theory,

f. Verification of results.

Decision trees are the most sophisticated methods for separating object sets into groups. In a Decision Tree, a set of predefined groups classifies the data elements. A completed decision tree is a tree where every node represents a question and the way to decide the direction depends on the reaction. The decision tree analyzes the array of data and effects, determines the frequency and distribution of values in the variable set and creates the decision model as a tree. The nodes at each level of this tree are a query, and each answer possible is expressed as a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

branch pointing at the next level to another node. Increasing steps along the journey from tree root to leaves tend to decrease the number of records that suit the answers along the way [9]. Several criteria of performance measurement that the researcher needs to consider regarding implement the decision tree classifiers [19]:

a. Stability, higher value of stability value means better performance,

b. Simplicity, a decision tree can be both explanatory model and predictive model, so it is important to make decision tree models as simple as possible. The simplicity can be came from simplicity based on number of leaves and simplicity based on rule size,

c. Discriminatory power of the leaves, these criteria can be used to measure the usefulness of decision tree, a higher level of discriminatory power of leaves means higher performance.

Basically, each tree node contains a set of records that match the answers to the questions along the path to the given node. Each problem divides the view into two smaller segments and each route between the root node and every other node is unique. Every node in the tree is also a rule, and at each point in the tree the record set which conforms to this rule can be evaluated and the record set can be evaluated. The analyst uses the model to look for a desirable result when the decision-making process moves through the boom and ends when the crossing hits the branch. The "thinking process" used by a model is a straightforward feature of the decision tree model, and the analyst knows how the model has come to a specific conclusion.[9].

A general strategy for carrying out data mining projects is generally pursued. In order to maximize the likelihood of success in undertaking data mining programs, data mining researchers and practitioners suggested a variety of procedures (workflows or basic step-by-step methods). A European group of companies indicated that the Cross-industry Standard Data Mining Method— CRISP-DM — was introduced in the mid-1990s to be a non-proper procedure for data mining [9], this is probably the most popular structured method. This is a six-step approach, beginning with a well-understood understanding of the market and the need for a data mining project (i.e., the application domain) and ending with the implementation of the solution this satisfies the relevant market requirement. Though these steps are sequential, there is typically a lot of backtracking. Data mining, since it is a method that is focused on the knowledge and testing, can be very iterative, depending on the problem situation and the know-how of the analyst, i.e., one can expect to go back and forth many times. Since further steps are focused on the results of previous measures, a special emphasis must be on earlier steps to ensure that the entire process is not done in the wrong way. The DM-CRISP model is considered as having three executive steps to monitor and evaluate the data. The following steps include [8]:

a. Business understanding, means that researchers should understand the main objective of their study then convert it into data mining problems and developing a project plan.

b. Data understanding, which consists of analyzing and examining data, this phase also includes data collecting, verify the quality of data, and data exploration

c. Data preparation, this phase includes the data selection, data cleaning, data formating and integration, then construct the final dataset for modeling

Many researchers have conduct studies both in the public sectors and private sectors, in many different type of industries, which already use the data mining approach with some reason and purposes to support the organizational or business objectives such as cost reducing, enhance research or information, increase sales volume, detect fraud and waste in operational organization, measure and improve the organizational performance or search trends in a huge amount of data in the organization. The size of the organization that implements the data mining also came from large and small-medium sized company to design a more cost-effective strategy by optimizing core valuables activities that support the organizational performance [20]. In the case study, this method also already implemented in the medical sector to analyze the heart diseases prediction by use 15 attributes and compare the result by using three classifiers which are naïve bayes, decision tree, and neural networks [11] and study the readiness level of unemployment in Ireland by using 139 attributes and compare the result by implementing the decision tree, weight by correlation, and weight by chi squared statistic [21].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

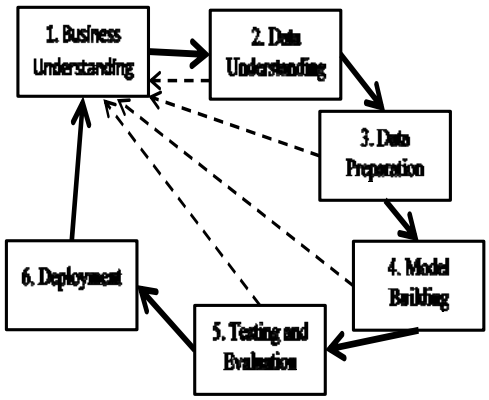

Figure 1. CRISP-DM

1. Business Understanding

The first step in CRISP-DM is Business Understanding.

The research in this study does not focus on business but the transmission of Covid-19 in South Korea, as is referred to here as Situation Understanding.

2. Data Understanding

A number of key points must be taken into account when defining and gathering data. In order to understand the most important details, the analyst should first of all be very clear and descriptive about the concept of data mining.

3. Data Preparation

In contrast with other steps of CRISP-DM data pre-processing requires more time and energy, many agree that this phase represents about 80 percent of the total amount of time spent on data mining. This huge effort is generally clarified by the fact that real-world data is unreliable (deficiencies in attributes of interest or data aggregates only), messy (containing anomalies or outliers), and ambiguous (containing inconsistencies in codes or names).

4. Model Building

Various modeling methods for an already designed data set are then selected and used to meet different business requirements. We use the Decision Tree approach in this research as we explore the mechanism behind the results.

5. Testing and Evaluation

The model is tested for consistency and generality. This stage checks to what extent this model is achieving its targets and to what extent.

6. Deployment

Depending on requirements, the implementation step is as simple as report development or as a continuous data mining process across the business. The user, not the data analyst, conducts implementation steps in several cases.

To discover the pattern of Covid-19, this study uses a patient dataset taken from Kaggle provide by datartist and 12 collaborators [22]. who has a structured dataset based on the report materials of KCDC (Korea Centers for Disease Control & Prevention) and local governments.

3. RESULT AND DISCUSSION.

1. Situation Understanding.

The first step is to know what is going on with the info. On 31 December 2019, the Chinese WHO Country Office confirmed cases of pneumonia of unknown etiology to the Chinese government in Wuhan Town, Hubei Province of China. National authorities in China confirmed to the WHO from 31 December 2019 to 3 January 2020 a total of 44 patients with the unexplained etiological pneumonic disease. During this reporting period, the causative agent was not identified. Comprehensive information on the outbreak in the Wuhan region exposure field was collected from the WHO China National Health Commission on 11 and 12 January 2020.

The Chinese authorities discovered a new form of isolated Coronavirus on 7 January 2020. On 12 January 2020, China released the genetic code of the latest Coronavirus used in developing countries in single diagnostic kits. The Ministry of Public Health Thailand announced on 13 January 2020 in its first imported case of a new lab-confirmed coronavirus (2019-nCoV) from Wuhan Province of Hubei, China. On 15 January 2020, the Ministry of Health, Labor and Welfare of Japan announced the latest Coronavirus (2019-nCoV) developed by the laboratories in Wuhan, Hubei Province, China [23]. Since then, the virus that currently has the official name SARS CoV 2 and the resulting disease called Covid-19 has spread to many countries outside China.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. Data Understanding.

Data was a structured dataset based on the report materials of KCDC and local governments. We use PatientInfo.csv data which contains 17 columns or attributes. The attributes are:

a. patient_id,
b. global_num,
c. sex,
d. birth_year,
e. age,
f. country,
g. province,
h. city,
i. disease,
j. infection_case,
k. infected_by,
l. contact_number,
m. symptom_onset_date,
n. confimed_date,
o. released_date,
p. deceased_date,
q. state.

There is much missing value in the data and some of the columns are completely blank. Some attribute is not necessary for this research such as global_num, birth_year can be replaced by age and it is easier to understand. Infected_by is very interesting but has much missing value. After we analyze the data we move on to the next step.

3. Data Preparation.

First, we deleted the attributes that did not in accordance with the purpose of the study. Attributes like patient_id, global_num, birth_year, and more are not included in this research. The purpose of this study is to predict the patient's condition. Thus the label attribute is state attribute. But we have to reset the attribute a little because we only wanted to see which patient was released and which were isolated. To do this, we modify the data, state attribute was changed to released attribute. The released state was change to yes and isolated state were change to no. For the deceased records, as we do not need it, we took it out. Below are ten records of the dataset for example.

Table 4. Data Preparation

sex	age	country	province	infection_ca...	released
male	50s	Korea	Seoul	overseas infl...	yes
male	30s	Korea	Seoul	overseas infl...	yes
male	50s	Korea	Seoul	contact with p...	yes
male	20s	Korea	Seoul	overseas infl...	yes
female	20s	Korea	Seoul	contact with p...	yes
female	50s	Korea	Seoul	contact with p...	yes
male	20s	Korea	Seoul	contact with p...	yes
male	20s	Korea	Seoul	overseas infl...	yes
male	30s	Korea	Seoul	overseas infl...	yes
female	60s	Korea	Seoul	contact with p...	yes

4. Model Building.

We use RapidMiner Studio educational 9.6.00 for the model building phase. When we used Decision Tree modeling, we have to choose the optimal parameters. There are plenty of parameters we need to adjust. One of the important parameters is Decision Tree Criterion which has four parameters namely gain ratio, information gain, gini index, accuracy, and least square. In this research, we optimized the Decision Tree Criterion only using three out of four criteria because we can not use the least square criterion because the data type is not number. Below is the process we use in the RapidMiner Studio.

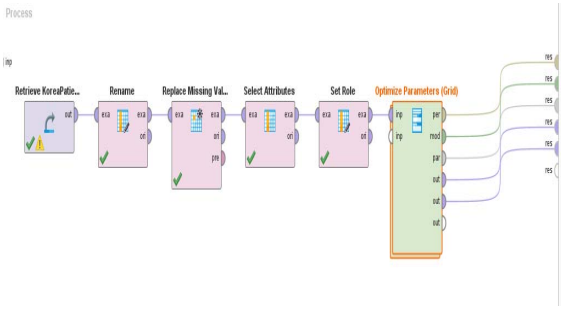

Figure 2. RapidMiner Process

We use rename parameters to rename the state attribute into released. Replace missing value is used again even though we already deal with missing value in the data preparation phase, we set the average data as a replacement for the missing value. After consideration, we did not include countries attribute and filtered it out on select

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

attributes parameters. The released attribute was set as the label in the set role parameters as we wanted to predict patients that had been released.

Optimize Parameters (Grid) was used to find the best Decision Tree Criterion. Below we can see the inside process of the Optimize Parameters (Grid). The subprocess of the Optimize Parameters (Grid) is Cross-Validation as we wanted to find the best accuracy from the model's parameters.

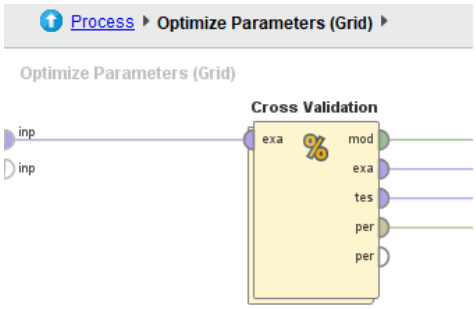

Figure 3. RapidMiner Optimize Parameters (Grid) Sub Process

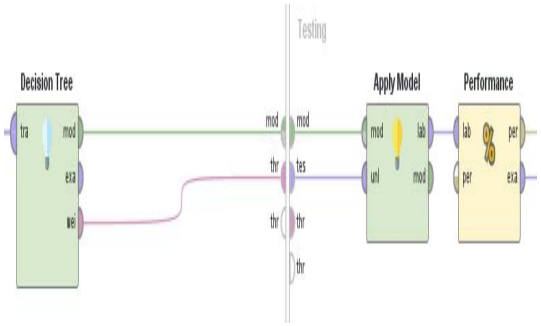

Figure 4. Training and Testing Decision Tree Model

We wanted to find the best accuracy for Decision Tree and then draw a conclusion from it. After all the parameters were set, we run the RapidMiner Studio. Below is the result of the optimized parameters for the Decision Tree Criterion.

Table 5. Optimized Parameters

iteration	Decision Tree.criterion	accuracy
1	gain_ratio	0.796
3	gini_index	0.798
4	accuracy	0.805
2	information_gain	0.794

From table 5 we can see that in this case, the accuracy criterion has the highest accuracy for Decision Tree. Thus we apply the criterion into the model and below is the confusion matrix.

Table 6. Confusion Matrix

	true yes	true no	class precision
pred. yes	31	39	44.29%
pred. no	156	772	83.19%
class recall	16.58%	95.19%	

We can see from table 6 above that there is 31 predicted for yes and correctly belong to yes group and there is 39 predicted as yes but actually, it belongs to no group. The same as for predicted no, 156 is true, yes and 772 is true no. The accuracy for this model is 80.46% +/- 1.68% (micro average 80.46%). Spearman rho is 0.185 +/- 0.066 (micro average: 1.851). root mean squared error: 0.380 +/- 0.013 (micro average: 0.380 +/- 0.000). Thus we confident enough that the model is good enough.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

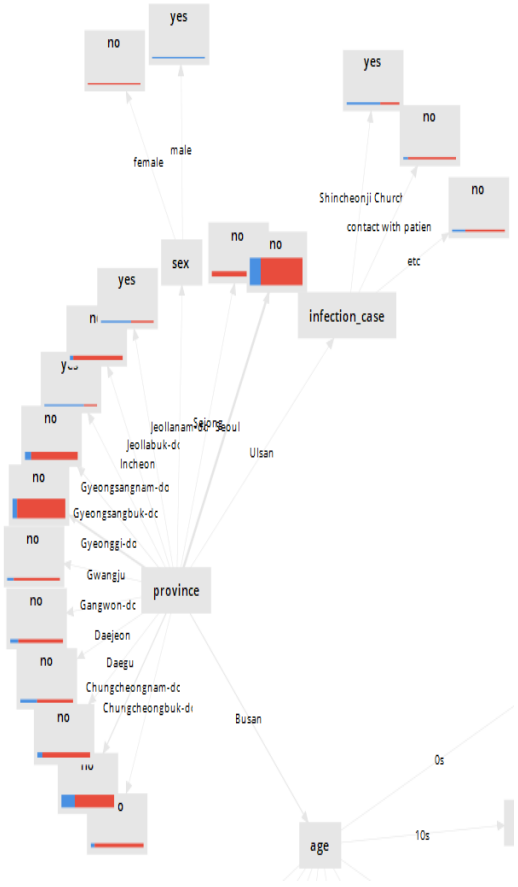

Figure 5. Decision Tree Part 1

Because the decision tree produced is quite large thus we use the radial presentation to have a better view. From figure 5 above it can be seen that the main factor for patients to be released is the province, not age. For Ulsan, because there was the Shincheonji Church incident, it branches into infection cases. From province also we can see that it branches in sex attributes with male have more chances to be released. For Busan Province, it branches again into age.

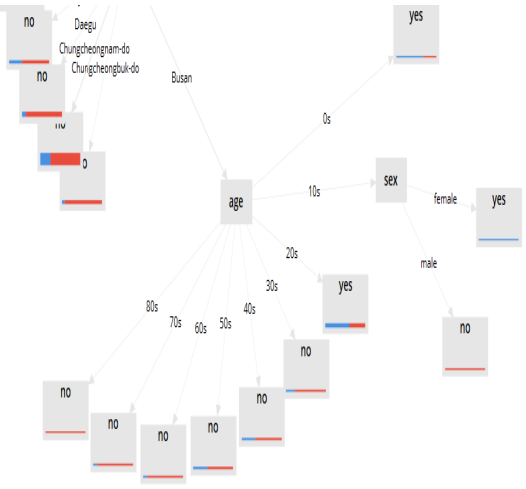

Figure 6. Decision Tree Part 2.

The age then split again into the age category from 0s to 80s. Then at the age of 10s, it split again into male or female, this time female has more chances to be released.

From the explanation given above, we can learn and understand that South Korea has a different spreads pattern with other countries. By using decision tree criteria, our findings showed that Province (area) had the significant variable that impacts the spreads of Covid-19 virus rather than age, gender, or hereditary disease. Therefore the government policy that conducts the massive Covid-19 test by walk-through, drive-through, contact tracing, and digital maps are effective to prevent the spreads, educate and give awareness for society. This findings also support the previous study that showed the effectiveness of drive-through and walk-trough in the area that have a high level of Covid-19 spreads in South Korea [4]

4. CONCLUSION.

Many countries have been affected by the rapid spread of the SARS CoV 2 virus. While the WHO estimates that the death rate from this virus is 3.4%, there is hysteria in the world. State economies were unstable, things were halted and even the Italian League, series A must be put off. This study shows that Decision Tree can accurately describe the data set to an accuracy of 80.46% and Spearman rho is 0.185, root mean squared error is 0.380 +/- 0.013.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

With the optimal criterion for Decision Tree is the accuracy criterion.

From the Decision Tree model, we can see the thinking process for a patient with a released state. The most important variable to split on is the province or area. From the result can gives us a clue that place or region is the most crucial determinant as a person will get infected or not. Therefore this study reconfirms that stay away from an infected area or staying at home is the right decision.

After the province variable, then the next variables to split on is sex, Ulsan, and Busan. Ulsan is the province where Shincheonji Church incident happened thus it split again into infection cases. But the Shincheonji Church branches to yes, this perhaps because the incident happened a while ago and most of the people who attended the activity are recovered by now. For infection cases that got the virus from contact with patients mostly are not released yet.

For Busan, age is the decision variable ranged from 0s to 80s. Age played an important role here, as the 30s to 80s have more chance of being treated longer compared to 20s, 10s, and 0s. This is a piece of good news for young Korean and also a confirmation that age has an important role in fighting the viruses. And children around 10 years of age, sex seem to determine the chances of recovery or to be released.

By study the Covid-19 case in South Korea, we can learn that the major factor of virus spreads that happen in this country is the area (province), then gender, and last age, there is no significant finding regarding to the hereditary disease in this case, therefore the case in South Korea slightly quite different with other countries such happens in China, Italy, India, or the USA. In South Korea, the government prefers not to choose lockdown strategy, they prefer to test covid-19 by conduct the walk-through, drive-through, tracing the suspect, and digital maps, and this strategy prove effectively. By considering the findings in South Korea, other countries also can be modified or adopt the same strategy that might be effective by announcing several policies to manage the spreads of Covid-19 such as massive Covid-19 test, tracing suspect using technology, and partially lockdown for certain area or district that have high level of Covid-19 cases. In Indonesia, after study from several Covid-19 cases in many countries, the government prefer to modify the policy by implement big scale of social restrictions in several districts of provinces to prohibit the activities of the society.

REFERENCES

[1] "Coronavirus Update (Live): 110,099 Cases and 3,831 Deaths from COVID-19 Wuhan China Virus Outbreak - Worldometer." https://www.worldometers.info/coronavirus/ (accessed Mar. 09, 2020).

[2] B. L. J. Higgins-Dunn Noah, "WHO says coronavirus death rate is 3.4% globally, higher than previously thought," *CNBC*, Mar. 03, 2020. https://www.cnbc.com/2020/03/03/who-says-coronavirus-death-rate-is-3point4percent-globally-higher-than-previously-thought.html (accessed Mar. 09, 2020).

[3] "Efek Domino Virus Corona ke Industri Penunjang Pariwisata - Katadata.co.id," Mar. 02, 2020. https://katadata.co.id/berita/2020/03/02/efek-domino-virus-corona-ke-industri-penunjang-pariwisata (accessed Mar. 09, 2020).

[4] D. Lee and J. Lee, "Testing on the move: South Korea's rapid response to the COVID-19 pandemic," *Transportation Research Interdisciplinary Perspectives*, vol. 5, p. 100111, May 2020, doi: 10.1016/j.trip.2020.100111.

[5] A. Kumar, K. Rajasekharan Nayar, and S. F. Koya, "COVID-19: Challenges and its consequences for rural health care in India," *Public Health in Practice*, vol. 1, p. 100009, Nov. 2020, doi: 10.1016/j.puhip.2020.100009.

[6] R. Djalante *et al.*, "Review and analysis of current responses to COVID-19 in Indonesia: Period of January to March 2020," *Progress in Disaster Science*, vol. 6, p. 100091, Apr. 2020, doi: 10.1016/j.pdisas.2020.100091.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concept and techniques*, 3rd ed. Elsevier, 2012.

[8] M. J. Nodeh, M. H. Calp, and İ. Şahin, "Analyzing and Processing of Supplier Database Based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) Algorithm," in *Artificial Intelligence and Applied Mathematics in Engineering Problems*, vol. 43, D. J. Hemanth and U. Kose,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Eds. Cham: Springer International Publishing, 2020, pp. 544–558.

[9] R. Sharda, D. Delen, and E. Turban, *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, 4 edition. New York, NY: Pearson, 2017.

[10] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," vol. 4, no. 1, pp. 16–23, 2009.

[11] N. Bhatla and K. Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques," *International Journal of Engineering Research*, vol. 1, no. 8, p. 4, 2012.

[12] S. S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms," *Oriental Journal of Computer Science & Technology*, vol. 8, no. 1, pp. 13–19, Apr. 2015.

[13] U. Fayyad, "From Data Mining to Knowledge Discovery in Databases," p. 18.

[14] A. Lausch, A. Schmidt, and L. Tischendorf, "Data mining and linked open data – New perspectives for data analysis in environmental research," *Ecological Modelling*, vol. 295, pp. 5–17, Jan. 2015, doi: 10.1016/j.ecolmodel.2014.09.018.

[15] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3 edition. Burlington, MA: Morgan Kaufmann, 2011.

[16] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4 edition. Amsterdam: Morgan Kaufmann, 2016.

[17] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," *SIGMOD Rec.*, vol. 31, no. 1, pp. 76–77, Mar. 2002, doi: 10.1145/507338.507355.

[18] N. Jagtap, P. P. Shevatekar, and N. Mustary, "A Comparative Study of Classification Techniques in Data Mining Algorithms," *International Journal of Modern Trends in Engineering and Research*, vol. 4, no. 7, pp. 58–63, 2017, doi: 10.21884/IJMTER.2017.4211.VXAYK.

[19] K.-M. Osei-Bryson, "Evaluation of decision trees: a multi-criteria approach," *Computers & Operations Research*, vol. 31, no. 11, pp. 1933–1945, Sep. 2004, doi: 10.1016/S0305-0548(03)00156-4.

[20] Z. Bosnjak, O. Grljevic, and S. Bosnjak, "CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data," in *2009 5th International Symposium on Applied Computational Intelligence and Informatics*, Timisoara, Romania, May 2009, pp. 509–514, doi: 10.1109/SACI.2009.5136302.

[21] I. Wowczko, "A Case Study of Evaluating Job Readiness with Data Mining Tools and CRISP-DM Methodology," *IJI*, vol. 8, no. 3, pp. 1066–1070, Sep. 2015, doi: 10.20533/iji.1742.4712.2015.0126.

[22] "Data Science for COVID-19 (DS4C)." https://kaggle.com/kimjihoo/coronavirusdataset (accessed Mar. 21, 2020).

[23] "Novel Coronavirus (2019-nCoV) situation reports." https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed Mar. 09, 2020).