Journal of Theoretical and Applied Information Technology
15th August 2020. Vol.98. No 15

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

OPTIMIZING DECISION TREE CRITERIA FOR PREDICTING COVID-19 MORTALITY IN SOUTH KOREA DATASET

[1]IVAN DIRYANA SUDIRMAN, [2]RUDY ARYANTO, [3]MULYANI

[1,2,3]Entrepreneurship Department, BINUS Business School Undergraduate Program, Bina Nusantara University, Bandung Campus, Bandung, Indonesia, 40181

E-mail: [1]ivan.sudirman@binus.edu, [2] raryanto@binus.edu, [3]mulyani@binus.edu

ABSTRACT

SARS CoV 2 spreads very quickly. When this research was written, there were 1,780,315 positive cases globally. Countries in the world take various actions to curb the spread of this new virus. China is taking strict lockdown steps, while other countries such as South Korea are using massive diagnoses. While it might only be acceptable for countries with not too big populations, what the South Korean Government is doing is very interesting to research so that it can be recognized by other countries. This study use data mining techniques to search for pattern or new information from the data of the Covid-19 patients in South Korea. This study studies the use of decision trees to process the data in South Korea by finding the most optimal decision tree criteria. This study contributes to the data mining related field by showing decision tree is good enough to analyze the data. There are three best criteria for using a decision tree to predict deaths from this new virus. The best decision tree metrics are the gain ratio, knowledge gain and the gini index. In this analysis the gain-ratio is use for further analysis. Another finding is the province or location of this virus is the important factor. The implication of this finding supports the strategy of reducing virus transmission. Sex and age also play an important role in the prediction model.

Keywords: *Data Mining, Decision Tree, Classification, CRISP-DM, COVID-19*

1. INTRODUCTION

Currently, a coronavirus outbreak called SARS CoV 2 is hitting the world. This virus causes a disease called Covid19. SARS CoV 2 is a new virus that expert still does not know a lot about it. Nevertheless, one thing that is certain about this virus is that this new virus spreads very quickly. There has been no drug that can genuinely cure diseases caused by this new virus at the time of this study. Therefore, one way to inhibit the spread of this virus is to study the behavior of this virus gradually.

Based on the data from the worldometer.com website, on April 12, 2020, there are 1,780,315 cases of coronavirus and 108,828 deaths from this virus globally. The good news is that 404,021 patients have recovered. When viewed from countries, America, Spain, Italy, France, and Germany are the top 5 total cases in the world. The United States 532,879 cases, Spain 163,027 cases, Italy 152,271 cases, France 129,654 cases, and Germany 125,452 cases. While China, which was the place where the virus was first detected, currently has a total of 81,953 cases [1].

So far, no clinically approved antiviral drugs or vaccines have been identified that have been successful against COVID-19. It has spread rapidly across the globe, presenting the entire human population with tremendous health, cultural, social and environmental issues. The epidemic of coronavirus is a significant threat to the world economy. Far from all the countries, the disease is being studied and handled in a way that slows down, quarantines suspects by touch tracking, limits large gatherings, keeps patients locked fully or partially [2]

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

Table 1 Top Six Total Cases by Countries

Country, Other	Total Cases	New Cases	Total Deaths	New Deaths
World	1,779,743	+80,908	108,779	+6,095
USA	532,879	+30,003	20,577	+1,830
Spain	163,027	+4,754	16,606	+525
Italy	152,271	+4,694	19,468	+619
France	129,654	+4,785	13,832	+635
Germany	125,452	+3,281	2,871	+135
China	81,953	+46	3,339	+3

Source: worldometer.com

Countries are taking preventative measures to overcome the spread of this new virus. For example, China did a very strict lockdown in Wuhan, while Sweden does not use a lockdown to control this virus. South Korea also does not use lockdown but does extensive testing. What is done by South Korea is an example for other countries on how to combat the virus without lockdown successfully. Lockdown provides several advantages, such as medical staff are more comfortable to track those at risk of being infected with this virus. However, lockdown also raises several of economic risks. South Korea, while not using a lockdown approach but conducting extensive testing and data disclosure to control the spread of this virus.

In order to identify and notify potentially exposed members of the public, health officials in Korea have carried out diagnostic tests, exhaustive motion and touch tracking. Both governmental and private developers' mobile apps have played an significant role in leading people to check centres, transmitting trajectories of confirmation cases in digital maps and monitoring of safety and movement of passengers and others at risk for exposure[3].

Although it might only be suitable for countries with not too large populations, what the South Korean government is doing is very important to study so that it can be a consideration for other countries. One thing to note also is how the South Korean government prioritizes health over privacy so that data disclosure can be done and can further assist the community. Also, further study related to the spread of this virus is possible. Such as using data mining techniques to look for patterns or new knowledge can be made possible.

Therefore, the data of this virus patient needs to be studied so that the spread of this virus can be further investigated.This research tries to study patterns and seek information related to the spread of viruses in South Korea by using data mining. The model used is a decision tree by finding the most optimal criteria for existing data. This research will be a useful addition in the field of data mining and also in studying the spread of this virus

2. THEORY AND METHODOLOGY

Data mining is usually a way to generate insight (i.e., useful information or knowledge) out of the data collected, assembled, and processed by a company. Organizations are using a broad range of data analysis methods to explain their clients and their behaviors better and to explain complex corporate challenges [4]

Over the last few centuries, the progress made by computer technology, interconnected networks and organizations have seen a vast accretion of data. This phenomenon is also correlated with a sharp decrease in data storage and processing costs. These systems, such as online companies, social media, and mobile apps, reveal a wide variety of diverse data that wait to be studied. Data mining is one approach that can manage large quantities of data and use sophisticated algorithms to evaluate data models with multiple parameters.

Data mining begins with data, which can vary from a single array of few numerical observations to a complex matrix of thousands of indicators with billions of observations. There are several advanced computer methods used to explore the data for identifying relevant and useful data structures. Such methods of processing were derived from statistics, machine learning, and artificial intelligence. Information extraction interacts and is closely connected to a variety of related fields, such as database systems, information cleaning, display, data testing, and reliability assessment. [5].

The issue of data mining can be generally classified into managed or unattended training models. Supervised or guided data mining attempts to deduce a feature or link based on labeled training data to plot new, unlabeled data. Supervised

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

approaches estimate the output values based on several input factors. A construct from a training dataset is built for this purpose, with previously defined parameters of income and production. The prototype extrapolates the interaction among input factors and output factors and assumes that input factors are only identified for the data collection.

The expected performance variable is also called a classmark or goal. A substantial number of labeled records are needed to learn the model from the data under supervision through data mining. Unsupervised or uncontrollable data mining exposes unlabeled data patterns. No output variables can be expected in unsupervised data mining. The purpose of this class of approaches for data mining is the identification of data patterns based on the interaction among data points itself. Any application will involve both supervised and unsupervised learners. The issues in data mining can also be classification, regression, association analysis, anomaly detection, time series, and text mining project [5].

Data mining deals with current research, identification, and organization data from relationships and patterns. And thus, data mining is characterized as a means of identifying data patterns that might generate untested pattern forecasts [6]

Data mining and simple statistical analysis are used in various ways. Classical statistical methods primarily concentrate on the testing of stated hypotheses, and data manipulation investigates other, often unknown, possible hypotheses.[7]. Through integrating the statistical and data mining approach, we will enhance our understanding of the ever-growing amount of digital data. Like Witten and more [6], In the future, it is necessary to integrate data mining and statistical method not only to analyze varied and complex datasets but also how to incorporate disciplines and methods such as pattern recognition, knowledge bases, artificial intelligence and algorithms of machine learning.

The decision tree is among the most widely used classifications for problem categorization. Categorizing problem solver is needed for many discipline such as in the field of marketing, for example is to categorizing brand [8]. The decision tree can be use to project the effects of comments on a component of its objective value. Leaves include classifications across tree structures (also known as labels), non-leaf nodes are characteristics, and branches have elements that lead to classifications. [9].

Decision trees are the most common structure methods. Within a Decision Tree, multiple fixed groups classify data components. Any node in which each node is a request and the path decided is a constructed decision tree. The Decision Tree analyses the collection of information, and results determine the occurrence and distribution for the element set and describes the architecture for a decision as a tree. That tree-level node is a question, and each reaction possible is shown as a branch that leads to a different node on the next level. Steps from tree root to leaf aim to reduce the number of documents corresponding to the answer. [4].

Every tree node comprises mostly a set of data that suits the questions on the path to this node. That problem breaks the viewpoint in two smaller sections, and the path between the base node and each other's node is similar. — The tree node also constitutes a rule; the data set which aligns with this rule may be revised, and a database subset may be checked at any stage in the tree. When decision-making is taken and stops, the investigator uses the formula to find a successful result as it hits the component. A "pension cycle" model is a fundamental feature of a model decision table, and the researcher knows how the formula was adopted [4].

Decision trees are also a standard induction method. They are resilient and propositional in their results. Decision Tree Induction is the learning of tree architecture in building decision tree algorithms where each inner node (no leaf node) denotes the test component. The test outcome is defined by each division, and the class prediction is referred to by every outer node. The algorithm finds the best division data attribute for each class at each node. By selecting an element with knowledge gain, the best component for fragmentation is selected. The most insightful attribute separates the attribute [10].

One of the more common techniques used for problem classification is the decision tree. We use the decision tree to construct a conclusion on an interpretation of the objective value of a variable. The leaves contain classifications by tree structures or labels; features of non-leaf nodes and branches

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

include variations in classifying characteristics [9]. There are some advantages of decision-tree[11]:

- easy to understand and comprehend because the trees can be visualized.
- Small data readiness is required. Other techniques often involve normalization of data, development of dummy variables, and removal of missing values. However, this decision tree also does not support missing values.
- The tree cost (i.e., data prediction) for the number of data points used for the training tree is logarithmic.
- It can handle numerical as well as categorical data. Other approaches are commonly used to evaluate data sets that have only one variable type.
- Capable of addressing multi-output problems.
- The decision tree uses a white-box model. If a specific situation can be seen in a model, boolean logic can easily describe the interpretation of the condition. On the other hand, the findings can be challenging to explain in a black-box model (e.g., an artificial neural network).
- A model may be tested using statistical tests. The testing allows the reliability of the model to be taken into account.
- Well performance, even if the correct model from which data were produced, infringes somewhat its assumptions.

Decision-tree students can build over-complex trees that do not generalize the data correctly. The drawbacks of decision trees include:

- It is necessary to set the minimum number of samples available at a leaf node or to set the maximum depth of a tree.
- Decision trees can be unpredictable since small data variations may lead to the development of a completely different tree. The use of the decision-making tree in an ensemble mitigates this issue.
- The problem of learning the optimal decision tree under various optimal conditions, and even simple definitions, is considered to be NP-complete. Practical decision-tree learning algorithms are thus based on heuristic algorithms, such as the greedy algorithm, in which optimal decisions at each node are made locally. The global optimum decision tree can not be returned by such algorithms. It can be mitigated by training several trees in an ensemble pupil, where characteristics and samples are altered.
- Definitions are difficult to understand since decisions are not readily articulated by a decision tree, such as XOR, parity, or multiplexer problems.
- Decision tree learners build partial trees if other classes prevail. The balancing of the dataset before the application of the decision tree is therefore recommended.

To increase the probability of success with the development of data mining projects, analysts and experts have implemented various techniques (workflows or simple, stage-by-stage processes). A European group of companies in the early 1990s developed the Cross-Industry Standard Data Mining Process — CRISP-DM — as an unsuitable data mining process [4].

CRISP-DM is possibly the most widely used standard form. It is a six-step process, beginning with a business understanding and the need for a data mining project (i.e., the application domain), which meets all the business requirements. Since the stages are concurrent, there is usually a great deal of prevarication. Data mining can be very iterative, as this is a knowledge and analysis process, which is based on the problem and experience of the researcher. Further precautions should be taken so that the whole procedure will not go wrong, because there are more programs focused on the effects of previous actions. The following is a method of data mining using CRISP-DM:

1. Business Understanding.

The main element of developing a data mining study is to consider what the research project is. Answers of the need for management to gain new information and a good description of the business intent of the study to be undertaken will begin to address this problem. The case studied in this analysis is not a business issue. However, the same approach will be taken, namely by studying the Covid19 case in South Korea.

2. Data Understanding

The application of data mining is specific to particular business activity, and multiple business tasks require different data sets. The main task of data mining is to identify data from various relevant sources following business understanding. Several key points must be taken into account in the data recognition and collection process. To

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

understand the most relevant data, the analyst should mainly be specific about the nature of the data mining project.

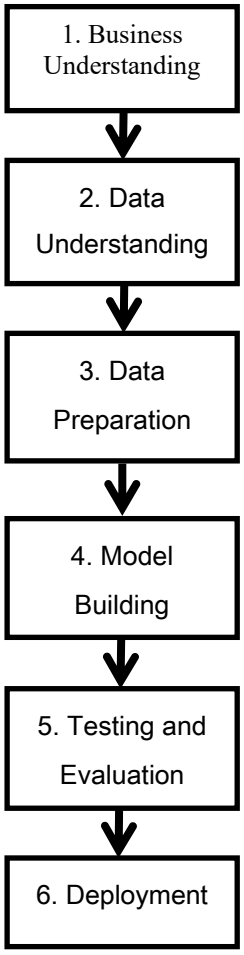

Figure 1. CRISP - DM

3. Data Preparation

The purpose of the data preparation is to take the prepared data for data mining techniques, and analysis is, as stated in the previous section. In comparison to other phases of CRISP-DM data pre-processing, experts agree that this step represents around 80% of the overall data mining effort. The fact that real-world data is not reliable (value attribute inadequacies or aggregate of data only) and ambiguous (containing anomaly or external elements) is evident to this enormous effort.

4. Model Building

To meet the business's specific company requirements, each process selects and applies several modeling techniques to an established data collection. In the phase of model creation, the different models produced are often evaluated and investigated. In the absence of a broad understanding of data mining activities, the "right" process for any given purpose should be used in some workable ways and with an established experimental and evaluation approach.

5. Testing and Evaluation

The models developed are being validated and examined for their accuracy and generality. This step measures how well the system (or model) selected fits the business objectives and to what extent (i.e., would need to build and analyze additional models?). Some other choice, if resource and financial constraints permit, is to test the model(s) built in a real-world scenario.

6. Deployment

The deployment process is as simple as writing a report or even as real-time data mining within the institution, obviously it depends on specifications. For some instances, the consumer does the deployment phase, not the data analyst.

3. RESULT AND DISCUSSION

The first step is to understand what the South Korean government is doing in the face of this new outbreak of the virus. Based on experience in dealing with MERS outbreaks, South Korea has understood that the most crucial factor in dealing with new virus outbreaks is a reliable and fast diagnostic tool. Then the contact tracking, where everyone positive is tracked and all of them tested. Contact tests detect the spread of the virus from humans to humans, but this new virus can also spread through objects where the virus is attached, such as door handles, elevator buttons, and other objects. The South Korean government has, therefore, recorded the path that the person has been taken and distributed to the web and mobile apps so that other people can avoid hazardous locations. [12].

Journal of Theoretical and Applied Information Technology
15th August 2020. Vol.98. No 15

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

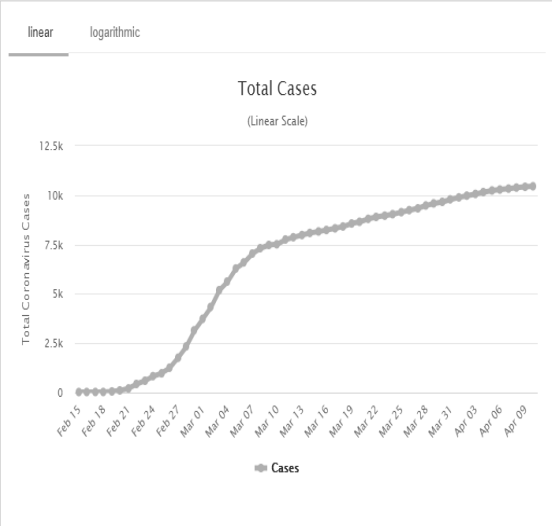

Figure 2. Total Cases In South Korea. Source: Worldometer.Com

The results were quite encouraging for South Korea, although the total number of cases reached 10,450, the deaths from the virus were "only" 208 and as many as 7,117 patients had recovered. New cases per day have been successfully suppressed using this method. Until April 10, 2020, there were only 27 new cases that day, while on March 3, 2020, there were 851 cases per day in South Korea.

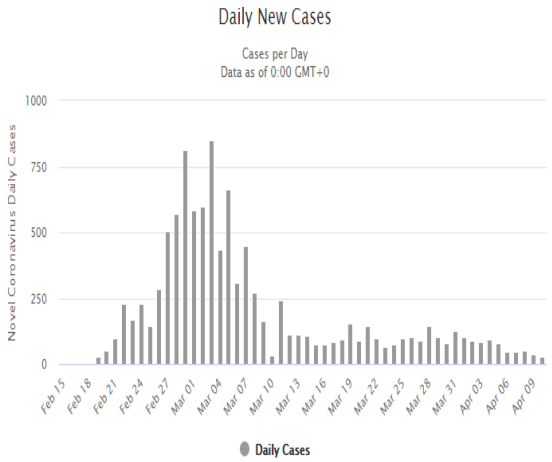

Figure 3. Daily New Cases In South Korea Source: Worldometer.Com

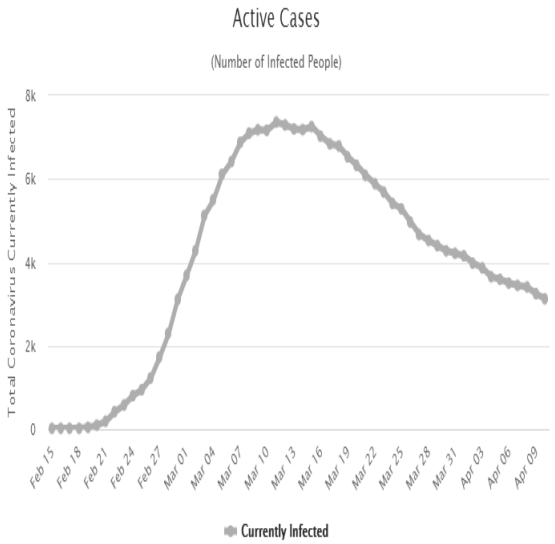

Figure 4. Active Cases In South Korea Source: Worldometer.Com

Based on previous figures, we can see that the South Korean Government's method has succeeded in suppressing the spread of this virus. Total cases have been successfully suppressed since mid-March to make the graphics more gentle. New cases are successfully reduced and can be treated earlier in combination with infected patients so that many recover successfully, causing active cases to go down. All of this has been successfully carried out by the South Korean Government through extensive testing, massive contact tracking, and unlocking of data. Based on previous figures, we can see that the South Korean Government's method has succeeded in suppressing the spread of this virus. Total cases have been successfully suppressed since mid- to make the graphics more gentle.

New cases are successfully reduced and can be treated earlier in combination with infected patients so that many recover successfully, causing active cases to go down. These methods have successfully suppressed the spread of the virus because the population in South Korea is not too large compared to Indonesia. For a country with a population as large as Indonesia, the methods used by the Government of South Korea might not necessarily be effective.

Next is to study the data to be processed. This study uses a patient dataset taken from Kaggle provide by datartist and 12 collaborators[13]. Who has a structured dataset based on the report materials of KCDC (Korea Centers for Disease Control & Prevention) and local governments.

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

There are several datasets available in the data sources, but for this study, the PatientInfo.csv were used.

The dataset consist of 18 columns, which is patient_id, global_num, sex, birth_year, age, country, province, city, disease, infection_order, infected_by,contact_number, symptom_onset_date, confirmed_date, released_date,deceases_date, state. Several attributes have not been included in the data processing stage. Some are attributes whose records are incomplete, such as disease, infection orders, and infected_by. Attributes that do not meet the objectives of the study will also not be included in the next stage. The birth year attribute was represented by the age attribute. The state attribute is changed to death because the objective of the research is to classify death and recovery. So the records taken are a deceased state record and released, then replaced deceased to yes and released to no.

After the data has been studied, it is prepared for further processing. In this study RapidMiner was used to process the dataset. RapidMiner is a program that facilitates the development and reporting of the overall data mining process. It provides not just a nearly complete set of operators, and also frameworks that convey the control flow of the operation. The data prepared for processing shall be as follows [14] .

Data mining is a method of uncovering trends in broad data sets to predict future outcomes. Structured data are data arranged into columns and rows so as to allow optimal access and modification. User can use a variety of data mining methods in a variety of applications to increase profits, lower costs, and prevent risks with a broad range of learning algorithms. Nevertheless, it is the way to evaluate unstructured data [15] .

Table 2. Sample Of The Dataset After Preparation

Row No.	sex	age	country	province	death
1	male	70s	Korea	Busan	yes
2	female	80s	Korea	Busan	yes
3	male	70s	Korea	Busan	yes
4	female	50s	Korea	Daegu	yes
5	male	70s	Korea	Daegu	yes
6	male	70s	Korea	Daegu	yes
7	female	80s	Korea	Daegu	yes
8	female	70s	Korea	Daegu	yes
9	male	60s	Korea	Daegu	yes
10	male	70s	Korea	Daegu	yes

Understanding data can be done by looking at descriptive statistical results. Figure 4 are the results of descriptive statistics from the dataset used

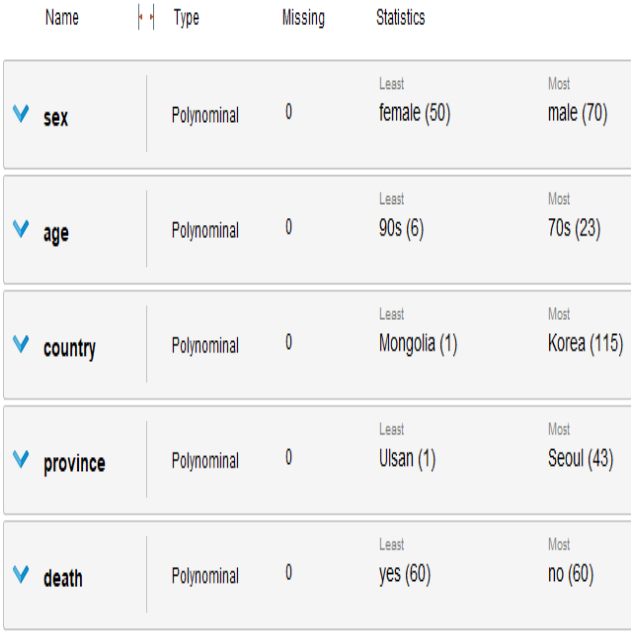

Figure 4. Description of The Dataset.

From the data description, all data types are polynominal. The number of "yes" records and the number of "no" records are balanced for the

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

death attribute set as a label. Data processing using classification methods should have balanced data. Also, Rapidminer is used to process data. One of the benefits of Rapidminer is that Rapidminer does not require users to be able to code. However, Rapidminer uses the operator box, which is available to be connected to other operator boxes according to the data processing needs. The operators used to process the data are as follows in this study.

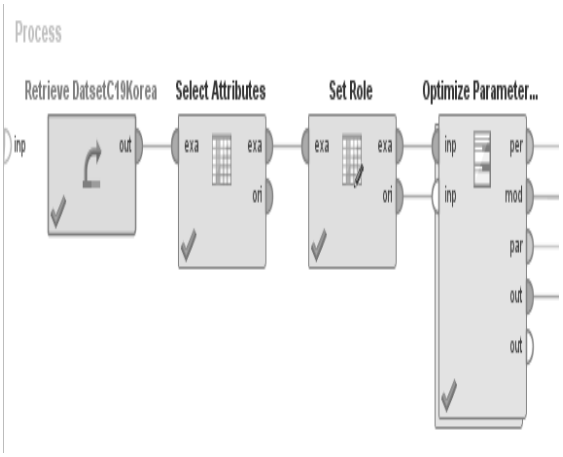

Figure 5. Process Operators

First Retrieve operators were used to retrieve the data in the repository, then Select Attributes were used to choose the attributes that will be used in the next process. The death attribute was set as the label in the Set Role operator after that Optimize Parameter operator was used to find the best Decision Tree criteria.

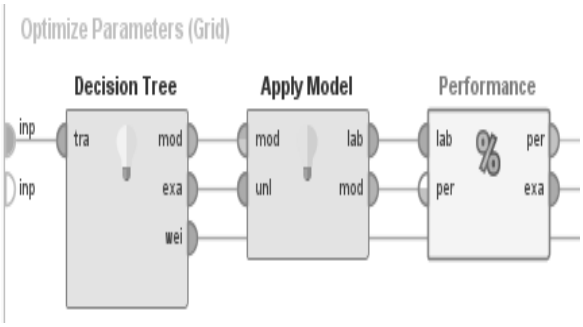

Figure 6. Optimize Parameter Subprocess.

In the Optimize Parameter subprocess in figure 6, the Decision Tree operator was used, and then the apply model operator, and the last one is the performance operator to find the performances of the model. After running the above process, the best criteria for Decision Tree using this dataset are as follows. Three criteria have the same accuracy, gain ratio, information gain, and Gini index. In information gain, entropies are determined for all attributes and the one with the least entropy is selected for break. This approach has a bias in choosing Attributes that have a significant number of values. While in Gain ratio, the information gain variant that varies the information gain for each attribute to allow the Attribute values to be large and standardized. And in Gini index, the calculation of inequalities between label character distributions. Splitting into a selected attribute results in a reduction of the resulting subsets' average gini index[16]. Thus, we agree with RapidMiner to use gain ratio criterion as shown in figure 7. Also, the gain ratio overcomes the issue by taking the number of branches that will arise before breaking the system into consideration. It corrects data gain by considering the inherent information of a break.

Table 3. Optimize Parameters (Grid)

iteration	Decisio...	accuracy
1	gain_ratio	0.950
2	informati...	0.950
3	gini_index	0.950
4	accuracy	0.933

Rapidminer proceeds the process using gain ratio criteria, and the result was as follows.

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

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 95.00%
ConfusionMatrix:
True:   yes     no
yes:    57      3
no:     3       57
-----kappa: 0.900
ConfusionMatrix:
True:   yes     no
yes:    57      3
no:     3       57
-----weighted_mean_precision: 95.00%, weights: 1, 1
ConfusionMatrix:
True:   yes     no
yes:    57      3
no:     3       57
-----spearman_rho: 0.900
-----absolute_error: 0.083 +/- 0.186
-----relative_error: 8.33% +/- 18.63%
-----cross-entropy: 0.195
]
Decision Tree.criterion = gain_ratio
```

Figure 7. Parameters Set Result.

There were several parameters set used in this study to measure the model. From figure 7, the results of the Confusion Matrix, it appears that 57 predictions of yes are correct, and only three predicted yes, apparently it should be no. Moreover, vice versa, there are 57 predicted by no and correct, then there are three predicted by no but should be yes. The Accuracy for Decision Tree with gain ratio criteria was quite high, 95%.

The Kappa coefficient is a statistical inter-rater of reliability or agreement among raters, which is used to test qualitative records and to evaluate the agreement between two raters. The formula used to measure kappa is as follows:

$$K = PR(e),$$

Where Pr(*a*) is the agreement observed between raters, and Pr(*e*) is the hypothetical likelihood of raters suggesting an agreement by chance. The kappa coefficients are represented using the Landis and Koch guidelines [17], Where the intensity of the kappa coefficients is defined as follows: minor 0.01-0.20; decent 0.21-0.40; moderate 0.41-0.60; significant 0.61-0.80; almost perfect 0.81-1.00. The Kappa throughout this analysis is 0.9, which means nearly perfect.

Mean weight precision is also fine because the data are balanced. The relative error is also known as relative uncertainty or approximation error. In this study is 8,33%. Cross entropy is 0.195; the cross-entropy is the average number of bits needed to encode data coming from a source with distribution p when we use model q [18].

Table 4. Weight

attribute	weight
province	0.705
sex	0.217
age	0.078

The weight of the variable included in this study shows that the province has the highest weight while age is the lowest. This finding shows that Province or location is a better predictor of the death attribute, while age has a contribution to predicting death but not as much as Province. For the Tree in the Decision Tree, the cropping of the image is done so that it can be seen more clearly. The first division is based on age attributes with age attributes of the 20s, 30s, 40s, 60s, 70s, 80s, and 90s. The 20s, 30s, and 40s are highly likely to have no death classification.

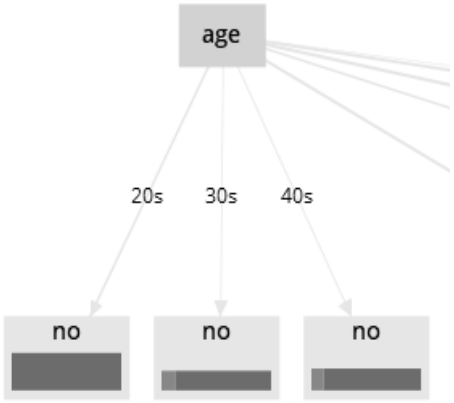

Figure 8. The tree for Age 20s, 30s, and 40s.

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

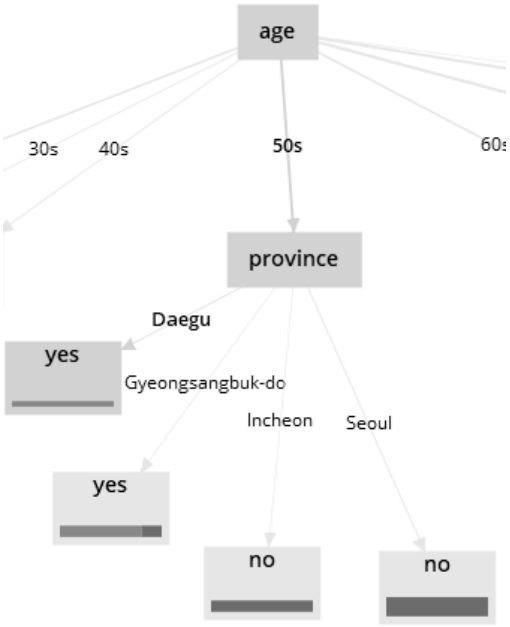

Figure 9. The Tree for Age 50s

While for the 50s, the probability to be categorized in yes or no is depend on the province. Daegu and Gyeongsangbuk-do have a higher probability for yes, while the age 50s in Incheon and Seoul have a higher probability for no.

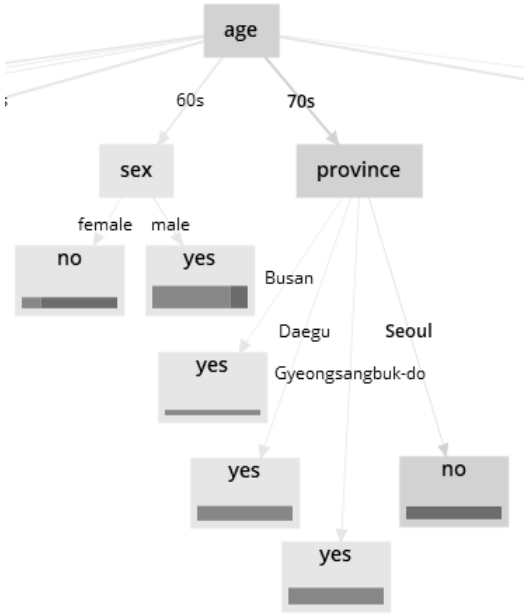

Figure 10. The Tree For Age 60s And 70s

For people in their 60s, gender is a determining factor to be grouped into the attributes of death, while females are more likely to survive than males. While people in the '70s, is like people in the '50s, both are more influenced by the location. Busan, Daegu, and Gyeongsangbuk – are more likely to be included in the yes group, while Seoul is more likely to be categorized as no. From figure 11, Gender is the first split for people in the 80s. Females are more likely to be categorized as yes and male to be categorized as yes, or no is more likely to be dependent in the province first. Daegu, Gyeongsangbuk-do have a higher probability of being categorized as yes, while Seoul has a high probability of being included in no group.

Therefore, in this study we can see that the factors that increase the chance of death are interrelated. Age is often considered the most important factor[2], but the chances of survival for someone above the age of 50 depend also on other factors such as gender and location. This study uses data from South Korea where at the time the data was obtained, South Korea succeeded in suppressing the death rate despite quite a lot of positive cases. So the amount of data used in this study is not large. In addition, the granularity of data is limited.

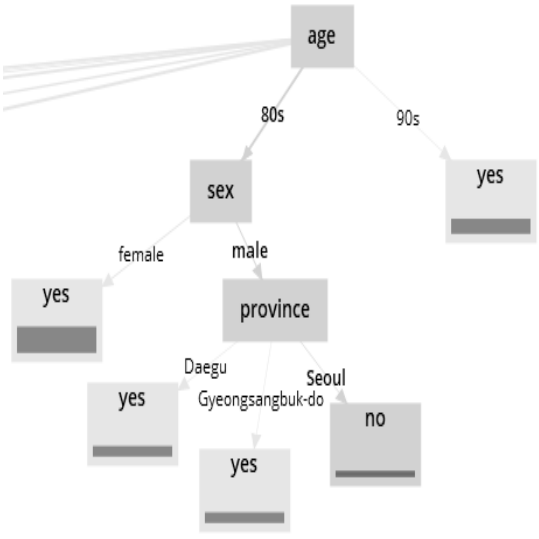

Figure 11. The Tree For Age 80s Dan 90s

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

4. CONCLUSION.

SARS CoV 2 has been a pandemic for weeks to come. Several countries are trying to stop the spread of this new virus. South Korea is one of the countries that has succeeded in suppressing the spread of this virus. The South Korean government has taken a different path, namely, through extensive testing, contact tracing, and disclosure of data.

Based on the dataset, age, gender, and provincial attributes were selected to predict mortality. The results give us a glimpse of the use of decision trees to study the spread of this virus.

The best criteria for the decision tree are the gain ratio, the gain in information, and the Gini index. The gain ratio is used in this study. The province or location is the most critical factor in determining the death of this virus. This finding reinforces the policy of reducing the spread of viruses through isolation or lock-down.

Gender and age also have an essential role to play in determining the chances of being included in the "yes" or "no" category of the death attribute. For those in their 60s and 80s, it seems that sex is more decisive than location. Those in their 60s who were female were more likely to survive than women in their 80s. While men in their 60s have a lower chance of survival, men in their 80s still seem to depend on their location. Age 20s, 30s and 40s have an excellent opportunity to be part of the group no. The ages of the 50s and 70s both depend on the province. For these age groups, the provinces with many COVID-19 cases, are more likely to be included in the yes category.

The findings of this study reinforce the notion that to prevent the spread of this virus, such as home isolation or lockdown. They are thus lowering the risk of getting infected from a hazardous location, especially for those in the 50s and 90s. It should be noted, however, that this study has several notes to consider. Only 120 records from South Korea were used in this study. The balanced attribute is only the attribute of death. This study is also to show a glimpse of the use of data mining in the COVID-19 dataset in South Korea. More data and more attributes are needed to get a better understanding.

REFERENCES

[1] "Coronavirus Update (Live): 275,125 Cases and 11,376 Deaths from COVID-19 Virus Outbreak - Worldometer." https://www.worldometers.info/coronavirus/ (accessed Mar. 21, 2020).

[2] I. Chakraborty and P. Maity, "COVID-19 outbreak: Migration, effects on society, global environment and prevention," *Sci. Total Environ.*, vol. 728, p. 138882, Aug. 2020, doi: 10.1016/j.scitotenv.2020.138882.

[3] D. Lee and J. Lee, "Testing on the move: South Korea's rapid response to the COVID-19 pandemic," *Transp. Res. Interdiscip. Perspect.*, vol. 5, p. 100111, May 2020, doi: 10.1016/j.trip.2020.100111.

[4] R. Sharda, D. Delen, and E. Turban, *Business intelligence, analytics, and data science: a managerial perspective*, Fourth edition. New York, NY: Pearson, 2018.

[5] V. Kotu and B. Deshpande, *Predictive analytics and data mining: concepts and practice with RapidMiner*. Amsterdam: Elsevier/Morgan Kaufmann, Morgan Kaufmann is an imprint of Elsevier, 2015.

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

[7] 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.

[8] I. D. Utama and T. Inayati, "Brand Post Analysis and Categorization in Automobile's Instagram Accounts," in *2019 International Conference on Information Management and Technology (ICIMTech)*, Aug. 2019, vol. 1, pp. 12–17, doi: 10.1109/ICIMTech.2019.8843753.

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

[10] J. Jotheeswaran and D. Y. S. Kumaraswamy, "OPINION MINING USING DECISION TREE BASED FEATURE SELECTION THROUGH MANHATTAN HIERARCHICAL CLUSTER MEASURE," . *Vol.*, vol. 58, p. 9, 2005.

[11] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, p. 2825−2830, Oct. 2011.

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

[12] "The big lesson from South Korea's coronavirus response - YouTube." https://www.youtube.com/watch?v=BE-cA4UK07c (accessed Apr. 11, 2020).
[13] "Data Science for COVID-19 (DS4C)." https://kaggle.com/kimjihoo/coronavirusdataset (accessed Mar. 21, 2020).
[14] M. Hofmann and R. Klinkenberg, Eds., *RapidMiner: Data Mining Use Cases and Business Analytics Applications*, 0 ed. Chapman and Hall/CRC, 2016.
[15] "Data Mining Tools," *RapidMiner*, May 28, 2019. https://rapidminer.com/glossary/data-mining-tools/ (accessed Apr. 13, 2020).
[16] "Decision Tree - RapidMiner Documentation." https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_decision_tree.html (accessed May 13, 2020).
[17] R. J. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, pp. 159–174, 1977.
[18] K. P. Murphy, *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series): Murphy, Kevin P.: 9780262018029: Amazon.com: Books*, 1st ed. .