Journal of Theoretical and Applied Information Technology
31st May 2020. Vol.98. No 10

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

NAIVE BAYES CLASSIFIER FOR PREDICTING THE FACTORS THAT INFLUENCE DEATH DUE TO COVID-19 in CHINA

[1]IVAN DIRYANA SUDIRMAN, [2]DIMAS YUDISTIRA NUGRAHA

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

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

E-mail: [1]ivan.sudirman@binus.edu, [2]dimas.nugraha@binus.edu

ABSTRACT

SARS CoV 2 or Coronavirus has spread rapidly throughout the world. Much remains unknown about this new virus. The rapid spread makes many countries must be vigilant in facing this new virus. This study uses the patient dataset in the early stage of the spread of SARS CoV 2 in China. Data is processed using data mining techniques with the Naive Bayes method. In addition the simulation process is carried out to find the optimal value of the findings.

Keywords: *Data Mining, Naive Bayes, Classification, Simulation, Covid-19*

1. INTRODUCTION

The spread of a new virus named SARS CoV 2 began in 2019 in China. Then months later it has spread beyond China. But when this research was written the spread of this new viruses in China began to be suppressed, unfortunately the spread of viruses outside of China began to increase. South Korea, Iran, and Italy are the three countries outside China with the most cases of Covid-19, the name of the disease caused by the SARS CoV 2 virus.

According to worldometers.info when this research was written there are 110.099 cases of Coronavirus, 3.831 deaths and 62.332 recovered. In China, there are 80.738 cases, South Korea 7.382 cases, Italy 7.375 cases and Iran 6.566 cases [1]. Even though there are a lot of cases around the world but according to WHO report the death rate is lower than SARS (Severe Acute Respiratory Syndrome) or MERS (Middle East Respiratory Syndrome) [2].

Although the death rate of this virus is relatively low when compared to SARS and MERS, the speed of spread of the SARS CoV 2 virus was apparently able to disturb the economies of affected countries. In Indonesia as for now, there are only 34 cases, however, the tourism sector is starting to feel the impact. Hotel occupancy rates drop by 40 percent. The retail industry has the potential to lose a turnover of USD 48 million.

More than 495 types of commodities or 13 percent of commodities for export to China will experience the impact. As a result of this virus, it is predicted that as many as 299 imported goods from China will decrease or even disappear from the Indonesian market [3].

SARS CoV 2 virus is a new virus so that at present there is no reliable drug to cure the diseases caused by the virus. Much is still unknown about this virus, how it infects other people, what factors are most influential to cause death, whether the risk of death can be predicted or not.

Lately, many researchers use data mining techniques to be able to find insights from data. The increased computer's ability to process data and the speed of the internet makes the use of data mining more widespread.[4]. Using data mining we can find patterns from data, There are four types of patterns that can be revealed by data mining techniques, namely association, prediction, cluster, sequential. In this research, we will conduct a classification data mining method which is the most frequently used data mining method. Classification is part of the machine learning family which also employs supervised learning.

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

There are many techniques for classification in data mining some of which are decision tree analysis, statistical analysis, neural networks, Bayesian classifiers in this study, researchers will use the Navies Bayesian technique and see how accurate the model is with the data. Simulations will be conducted after modeling is done to find the optimal results.

2. THEORY AND METHODOLOGY

Data mining is a nontrivial process in which the data in structured databases are defined with true, new, potentially helpful and ultimately understandable patterns . The study of numerical and categorical data in large and complicated data sets is part of data mining. The term is often used to refer to more advanced methods, such as text, web or spatial data[5].

Data mining deals with the analysis, identification, and establishment of existing data of associations and patterns. And thus data mining is characterized as a process for identifying patterns in data that could generate non-testing projections of unknown patterns [6].

Data mining and usual statistical analyses have various purposes. Classical statistical methods emphasis primarily on verifying stated hypotheses, data mining approach look across many possible, mostly unknown hypotheses [7]. Combining statistical and data mining approach will be the only way to increase insight and knowledge from the always rising amount of digital data. As Witten et al. [7]. Combining statistical and data mining approach will be the only way to increase insight and knowledge from the always rising amount of digital data. As Witten et al. [6] referred to, examining various and complex data in the future will not merely need a combination of data mining and statistical approach but the blending of disciplines and techniques such as pattern recognition, databases, artificial intelligence, and machine learning algorithms.

In this research the dataset are evaluated using a Naïve Bayes Classifier. It is a classifier that generates a probability of a certain set observations belonging to a class[8], in this case the class is 'yes' or 'no' in death coulomn. Wang [9] stated that the classifier of Naive Bayes has been accepted as a straightforward probabilistic classifier based on clearly independent premises from the interpretation of the Bayesian theorem. In other words, a classification of Naive Bayes suggests that there is no connection between the existence of one specific function of a class and another. Naive Bayesian is a basic probabilistic classification built on the independence principle Bayesian theorem [10].

According to Zhang [11] In machine learning and data mining classification is a basic matter. In a classification, the purpose of a study algorithm is to create a classifier with class labels. The Naive Bayes approaches are a collection of supervised research algorithms which use the theorem of Bayes, predicated on the idea that each pair of features is independent of each class variable [12]. Bayes 'mathematical model sets out the following relationship, given the class variable y and the dependent function vector x_1 through x_2:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

With the naive expectation of conditional independence that:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

for all , this relationship is streamlined to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Because $P(x_1,\ldots,x_n)$ is constant, the following classification rule can be applied:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

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

So we can use the approximation of P(y) and P(xi|y) by Maximum A Posteriori (MAP), the latter of which is the relative y-frequency of the training set. The various naïve classificators in Bayes vary mainly in terms of the selection premises of P(xi|y).

Despite the seemingly under-simplified premises, in many actual-world circumstances, popular data classification and spam detection, naive Bayer classifiers have performed nicely. A relatively small amount of testing information is needed to calculate the appropriate parameters. Naive Bayes trainers and classifiers can be incredibly fast opposed to more advanced methods. Detaching the class conditional feature classes allows for a separate calculation of each distribution as a single dimensional distribution. This helps in effect mitigate issues resulting from the computational complexity curse. On the otherhand it is recognized that while Bayes is a decent clustering algorithm, it is a poor estimator so that estimate proba likelihood outcomes are not taken at face value [12].

A general method is usually followed in order to conduct data mining projects. Found on best practices, data mining researchers and practitioners suggested several procedures (workflows or simple step-by-step methods) to optimize the probability of success in conducting data mining projects. Cross-Industry Standard Process for Data Mining—CRISP-DM—was suggested in the mid-1990s by a European consortium of companies to guide as a nonproprietary standard practice for data mining [4], this is probably the most popular structured method.

The proposed procedure, which is an order of six stages that begins with a well understanding of the business and the need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that fulfills the particular business demand. Although these stages are sequential, a lot of backtracking usually happens.

As data mining is based on experience & testing, the whole process can be very iterative (i.e. one should expect to go back and forth through the steps a few times) and time-consuming, depending on the problem situation and the analyst's know-how. As later steps are based on the findings of earlier steps, special attention should be paid to earlier steps in order to ensure that the entire study does not proceed on an incorrect course.

Step 1: Business Understanding

Knowing what the analysis is, is the key element of any data mining project. Responding to this question starts with a detailed understanding of the business needs for new knowledge and a specific explanation of the business purpose of the report. Although the focus of the analysis in this study is not business but the spreading of Covid-19, the same approach is used. Therefore, the phenomenon relatied to the case of Covid-19 spread must be understood before processing the data.

The appropriate term for this paper would be case understanding. It all start in Wuhan, in a seafood market that sold not just seafood but also wild animals. Many speculation arrive regarding from where the virus jump from animal to human. Some research suggest that the virus come from bat. The virus then spread from human to human. China authority quickly lock down Wuhan.

Regardless the Wuhan lock down, the virus made their way to other countries and now offcially become global pandemic[13]. This is a new kind of virus, at first people named it Coronavirus, now it has official name SARS CoV 2 and this viral disease is called Covid-19. From what we understand, the virus start in Wuhan and spread quickly accross the nation. Death toll as for today is more than 3.000, most of the death numbers come from Wuhan, China.

Step 2: Data Understanding

A data mining analysis is unique to a well-defined task and various business tasks involve numerous sets of data. After understanding of business, the main business of the data mining process is to classify the relevant data from various databases available. In the data identification and collection process, several key points must be considered. The analyst should first of all be very clear and concise about the definition of the data mining activity in order to identify the most relevant data.

Step 3: Data Preparation

The goal is to prepare data for processing through data mining techniques, the most

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

commonly named data preprocessing. The objective is to take the data defined in the previous step. Data pre-processing takes more time and effort compared to other phases in CRISP-DM; many believe that this phase constitutes approximately 80% of the time spent on the data mining project overall. The explanation for this massive initiative is that real-world data is usually unreliable (deficiencies in attributes of interest or data aggregates only), messy (containing anomalies or outliers), and ambiguous (containing inconsistencies in codes or names).

Step 4: Model Building

Diverse modeling approaches are then chosen and used to address the specific market criteria for an already configured data set. The model construction phase also involves the assessment and comparative analysis of the various built models. Benefits include a variety of feasible models along with a well-defined experiment and evaluation approach to finding the "right" method for a given purpose because the optimal method or algorithm for a data-mining activity is not universally known. A number of parameters must be optimized in order to achieve the best results even with a single method or algorithm.

Step 5: Testing and Evaluation

In phase 5, the models built are tested and measured for their exactness and generality. This stage tests to what level (i.e., should more models be created and evaluated?) and the degree to which this model (or models) fulfills the business objectives. The built model(s) in a real-world scenario can also be evaluated if time and budget constraints permit. Although the outcomes of the models are supposed to correspond to the original business aims, other insights are often observed, which do not necessarily relate to the original business objectives but also may provide additional information and recommendations.

Step 6: Deployment

Modeling and simulation is not the end of the data mining process. Even if the model is intended to make the data easy to analyze, it is important to arrange and communicate information obtained from this discovery in such a manner that end-users understand and benefit. The implementation phase may be as easy as report generation or as complex as a persistent data mining method across the business, depending on the requirements. In many situations, the implementation phases are performed by the client, not the data analyst.

In this study we did not apply the deployment step because it is not a business case. Thus from the above literature study we can describe the methodology in the next picture.

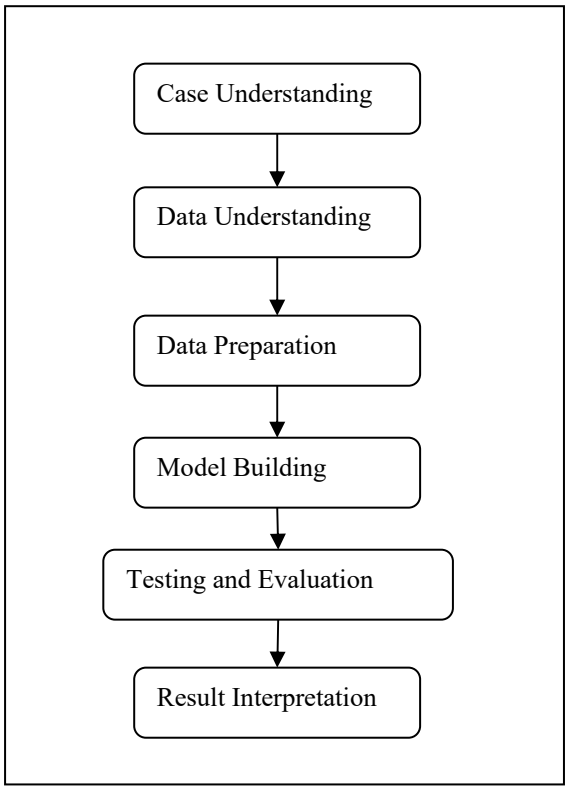

Figure 1. Methodology

To find out the pattern of Covid-19, this study uses a patient dataset taken from Kaggle by SRK [14] who has collected data on Covid-19 from various sources, Then for data mining processes researchers use RapidMiner application. Rapidminer is advanced data science, machine learning, profound learning, mining of text and predictive analytics ecosystem developed by the same name Company. This facilitates all levels of machine learning including data processing, outcome analysis, model testing, and optimization and applies to business and business applications, as well as to research, teaching, rapid prototyping, and application development [15].

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

3. RESULT AND DISCUSSION

The first step is to understand what's happening with the case and the data. All of these cases start from a City of Wuhan City, in Hubei Province of China, the Chinese WHO Country Office told the Chinese government on 31 December 2019 of cases of pneumonia of unknown etiology (unknown cause). A total of 44 patients with the unexplained etiological pneumonic disease have been reported by national authorities in China to WHO between 31 December 2019 and 3 January 2020. The causal agent was not identified during this reporting period. Detailed information was obtained from the China National Health Commission from the WHO on 11 and 12 January 2020 about the epidemic in the Wuhan area exposure sector.

On 7 January 2020, the Chinese authorities identified a new type of coronavirus isolated. The genetic code of the new coronavirus for use in countries in developing unique diagnostic kits was published by China on 12 January 2020. In its first imported case of new coronavirus, lab-confirmed (2019-nCoV) from Wuhan Province of Hubei, China, the Ministry of Public Health Thailand announced on 13 January 2020.

A laboratory-confirmed 2019 new coronavirus (2019-nCoV) introduced in Wuhan, Hubei Province, China was announced on 15 January 2020 by the Ministry of Health, Labor, and Welfare, Japan. [16].. 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.

The next steps is to study the data and prepared it for the modeling process. The data that was obtained by the researcher was then studied carefully. The format of the file is an excel file with 21 columns and having more than 1200 records. The label column is the age column, then the columns that are deemed not in accordance with the purpose of the study will not be used in the next step.

On the death column, there are many values but the value taken into this study is only number 1 which means the patient is deceased and 0 which is the patient is not death. Patient with number 1 is categorized as yes and patient with number 0 is categorized as no. There is only 42 records of patient who had died (yes). Thus in order to make the data balance between yes and no, the researcher add 58 more records with death value equal to 0, in total there are 100 data.

Some of the tables that contain data for use in the next step can be seen in Table.1 below. We delete several column that we think it is not related with the study such as time recorded, summary, etc. Our point of interest in this study is using Naive Bayes to predict death from age, sex, location, and other attributes.

Table. 1 Sample of The Dataset After Preparation

location	sex	age	Visit wuhan	From wuhan	death
Wuhan	male	61	no	yes	yes
Wuhan	male	69	no	yes	yes
Wuhan	male	89	no	yes	yes
Wuhan	male	89	no	yes	yes
Wuhan	male	66	no	yes	yes
Wuhan	male	75	no	yes	yes
Wuhan	female	48	no	yes	yes
Wuhan	male	82	no	yes	yes
Wuhan	male	66	no	yes	yes
Wuhan	male	81	no	yes	yes
Wuhan	female	82	no	yes	yes
Wuhan	male	65	no	yes	yes
Wuhan	female	80	no	yes	yes
Wuhan	male	53	no	yes	yes
Wuhan	male	86	no	yes	yes
Wuhan	female	70	no	yes	yes
Wuhan	male	84	no	yes	yes
Hubei	female	85	no	no	yes
Hubei	female	69	no	no	yes
Hubei	male	36	no	no	yes
Hubei	male	73	no	no	yes
Hubei	female	70	no	no	yes
Hubei	male	81	no	no	yes
Hubei	female	65	no	no	yes
Wuhan	male	70	no	yes	yes
Wuhan	female	76	no	yes	yes
Wuhan	male	72	no	yes	yes
Wuhan	male	79	no	yes	yes
Wuhan	male	55	no	yes	yes
Wuhan	male	87	no	yes	yes

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

Wuhan	female	66	no	yes	yes
Wuhan	male	58	no	yes	yes
Wuhan	male	66	no	yes	yes
Wuhan	male	78	no	yes	yes
Wuhan	male	67	no	yes	yes
Wuhan	male	65	no	yes	yes
Wuhan	male	58	no	yes	yes
Wuhan	female	67	no	yes	yes
Wuhan	female	82	no	yes	yes
Taiwan	male	65	no	no	yes
Kowloon	male	39	yes	no	yes
Hong Kong	male	70	no	no	yes
Guangdong	male	66	yes	no	no
Shanghai	female	56	no	yes	no
Zhejiang	male	46	no	yes	no
Tianjin	female	60	yes	no	no
Tianjin	male	58	no	no	no
Chongqing	female	44	no	yes	no
Sichuan	male	34	no	yes	no

After the data preparation stage, data processing using RapidMiner is carried out. RapidMiner is powerfull tools for data mining process without using any coding. RapidMiner use operator box and the user have to connect each of the operator box. Below is the example of the operator box used with the Naive Bayes algorithm.

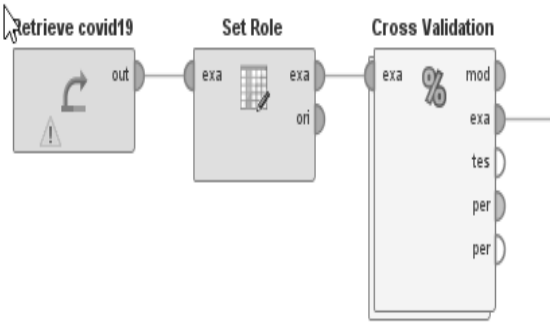

Figure 2. Process in RapidMiner

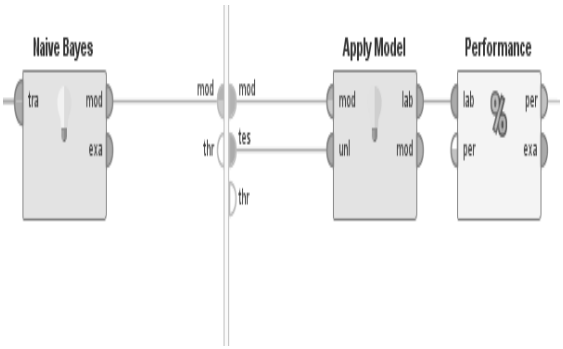

Figure 3. Cross Validation Process

We use auto model function in RapidMiner for simulation process. Descriptive results from the data indicate that the data processed is quite balanced which is 42 yes and 58 no.

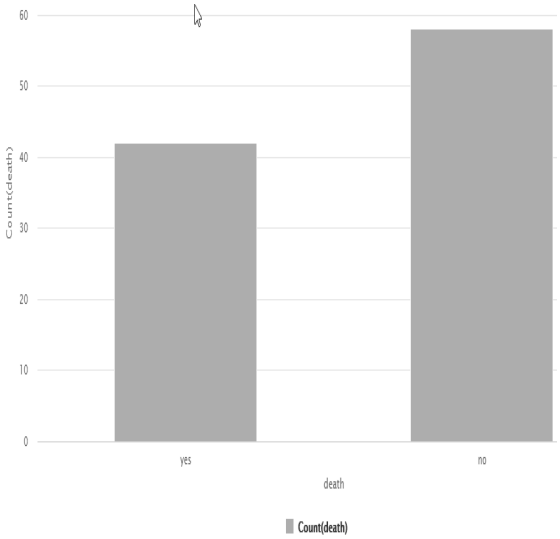

Figure 4. Death Column Distribution

Other descriptions of the data can be seen in the next image. We can see that attributes such as death, location, sex, visit_wuhan and from_wuhan is polynominal type and age is integer data type. Also there is no missing link in the data. This happens because the data cleaning process has been carried out using excel.

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

Name	Type	Missing
Label death	Polynominal	0
location	Polynominal	0
sex	Polynominal	0
age	Integer	0
visit_wuhan	Polynominal	0
from_wuhan	Polynominal	0

Figure 5. Data Desciprtion

We are sure enough that the dataset above can be used for the next step. Data processing is performed using the classification method with the group of concern is "yes" in the death column. Thus we can predict several factors that lead to death based on the dataset above. We also use simulation to improve the prediction.

Model development/testing and model evaluation/implementation are the commonest two-step classification form prediction methodology. A selection of input data, including current labels, is used in the model development phase. Upon teaching a model the model is tested for accuracy testing against the holdout sample and eventually used for practical use wherever it needs to be used.[4].

In contrast to the predictive precision of two or more techniques, one can use a technique called k-fold cross-validation in order to minimize the bias associated with a random sampling of the training and holdout test samples. The complete data set is randomly divided into k-exclusive sub-sets of approximately equal size in k-fold cross-validation, also called rotational estimation.

The models are trained and tested k times. K days. It is conditioned every time on just one fold and checked every time on the remaining single fold. The average accuracy estimation of a model is determined by merely combining the k individual precision measurements as shown in the next equation.

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i$$

Where CVA stands for cross-validation accuracy, k is the number of folds used, and A is the accuracy measure (e.g., hit rate, sensitivity, specificity) of each fold. In this research, we use 10 fold cross-validation and the result is shown in a confusion matrix.

Density of location attribute described in the next figure. Wuhan is the highest density this is because the data depict the early stage of the spreading phase.

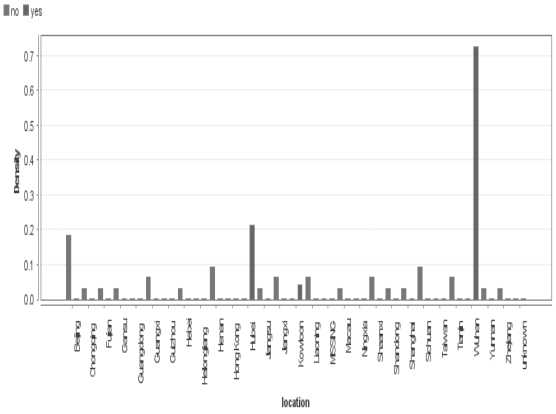

Figure 6. Location Density

Also if we look at the age distribution between age and density is shown in figure 3 below.

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

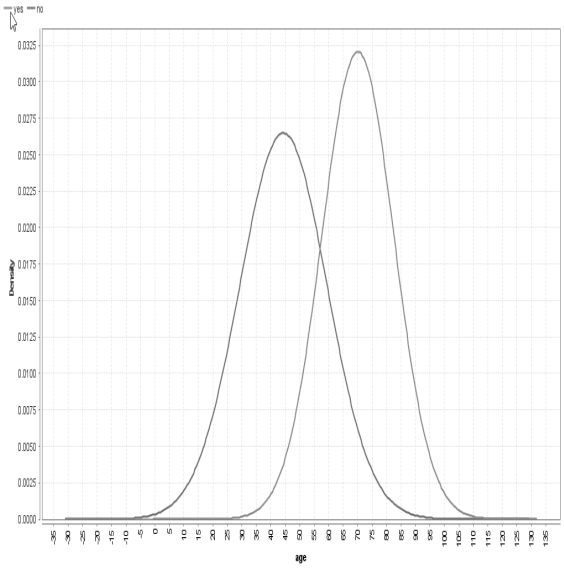

Figure 7. Age Density Distribution

The confusion matix is shown in the next figure.

Accuracy: 93%

	true yes	true no	class precision
pred yes	41	6	87,23%
pred no	1	52	98,11%
class recall	97,52%	89,56%	

Figure 8. Confusion Matrix

The result of the cross-validation using RapidMiner is shown in the confusion matrix. As we can see, the Naive Bayes classifier able to predict 41 yes with 6 that actually no but predicted as yes, while able predict 52 no and 1 yes but actually no. Thus we got the model accuracy for 93% which is good enough.

PerformanceVector

```
PerformanceVector:
accuracy: 93.00% +/- 6.75% (micro average: 93.00%)
ConfusionMatrix:
True:   yes     no
yes:    41      6
no:     1       52
root_mean_squared_error: 0.172 +/- 0.133 (micro average: 0.213 +/- 0.000)
correlation: 0.875 +/- 0.117 (micro average: 0.863)
```

Figure 9. Performance Vector

The model performance vector is good enough with 0.172 root mean squared error and 87,5% correlation. Thus we can sure enough that Naive Bayes can explain the data quite well.

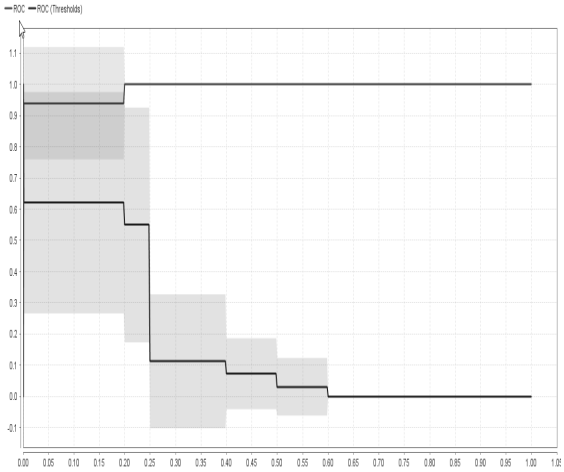

Figure 10. ROC

The ROC in the figure 10 above shows us that the model is on the above the ROC threshold. The area under curve atau AUC is 0.93 which is good.

Attribute	Weight
location	0.370
from_wuhan	0.044
visit_wuhan	0.020
age	0.011
sex	0.009

Figure 11. Attribute Weight

Based on figure 11, the most important factor is location, from_wuhan, visit_wuhan, age and lastly is sex. The above explains that the location, especially the center of the spread of the disease and also the visit to the center of the spread is the most important determinant before age and sex. Thus is logical because the dataset we use is on the early stage of the virus spread. However, this clearly shows that based on our findingsm stay away from the center of the epidemic is the safest way, also containing the infected area is the best decision to stop the virus spreading further.

The next step is we tried to find the optimum solution using simulation function on

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

RapidMiner. First, we see what happened if a person's age is around 55, from Wuhan, and male (figure 7a).

Figure 7a. Attribute on RapidMiner Simulation.

With the attributes mentioned before, the probability of that person falls into the yes category is 83% and no is 17%, shown in figure 12.

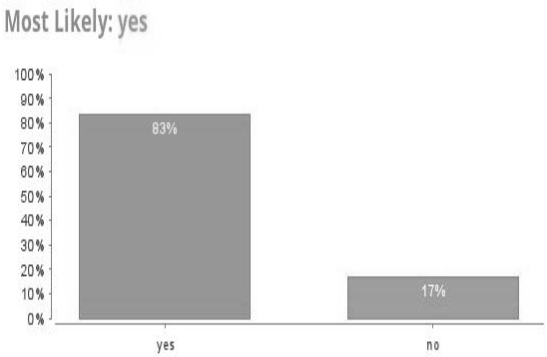

Figure 12. Simulation Output

Then we try to optimize in the group no with the following results, 100% belongs to the no group. This optimization is one of the superior feature from RapidMiner.

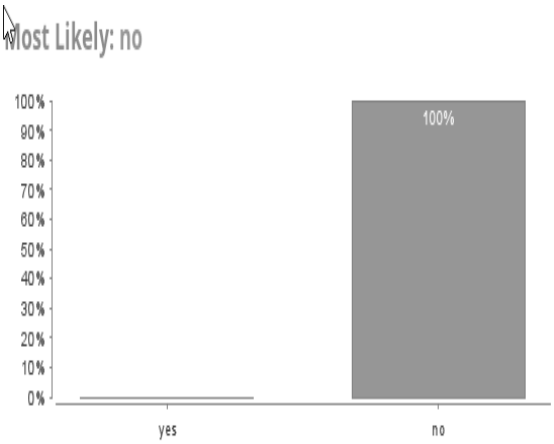

Figure 13. Simulation Output for 'No'

In order to have a 100 percent probabilty of no, the simulation shown that the important factors for optimum no is shown in figure 14.

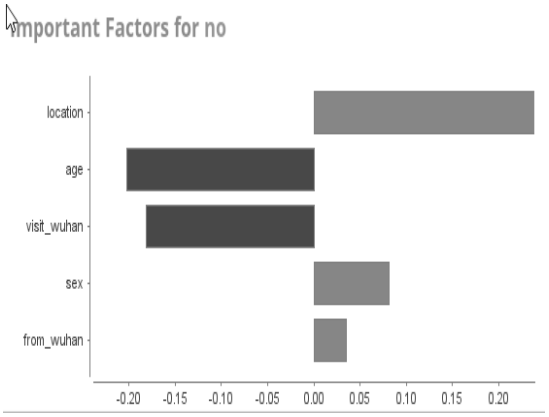

Figure 14. Optimum Attribute

The factors that support optimum of no group is location, sex, and from Wuhan. The factors that contradict for no group is age and visit Wuhan.

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

Figure 11. Attributes from Optimum Simulation

Our study is consistent with Wang's et al [17] which state that the first occurred deaths were majorly elderly people who might have faster disease progression. The median number of days from the occurence of the first symptom to death was 14.0 (range 6-41) days, and it tended to be shorter among people aged 70 years or more (11.5 [range 6-19] days) than those aged less than 70 years (20 [range 10-41] days; P = .033).

4. CONCLUSION.

The rapid spread of the SARS CoV 2 virus has affected many countries. Although according to WHO the death rate due to this virus is 3,4% but the world community is experiencing panic. The economies of countries have been disturbed, various events have been cancelled, as a matter of fact the Italian league, series A has to be postponed.

This article shows that Naive Bayes can explain the dataset well enough with an accuracy of 93%. the most important attribute to be classified in the dead group is location, from and visit Wuhan, age and lastly sex. These findings reinforce other findings that linking deaths from the virus with age and sex. We are also increasingly convinced that distance and location are also important and need attention.

From the simulations that have been carried out, it can be concluded that in addition to location, age is the determining factor of death from this virus. The younger a person is, the less likely to die from this virus. Locations that are further away and have never been to Wuhan can reduce the possibility of being categorized into the death groups. Based on data, women have a greater chance of survival than men.

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

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

[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] M. S. Sequera, S. A. Guirnaldo, And I. D. P. Jr, "Naïve Bayes Classifier And Fuzzy Logic System For Computer – Aided Detection And Classification Of Mammamographic Abnormalities," . *Vol.*, P. 12, 2005.

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

[9] Y. Dai and H. Sun, "The naive Bayes text classification algorithm based on rough set in the cloud platform," *J. Chem. Pharm. Res.*, vol. 6, pp. 1636–1643, Jan. 2014.

[10] X. Zhou *et al.*, "Detection of Pathological Brain in MRI Scanning Based on Wavelet-Entropy and Naive Bayes Classifier," in *Bioinformatics and Biomedical Engineering*, Cham, 2015, pp. 201–209, doi: 10.1007/978-3-319-16483-0_20.

[11] H. Zhang, "The Optimality of Naive Bayes," p. 6.

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

[13] "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020." https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed Mar. 12, 2020).

[14] "Novel Corona Virus 2019 Dataset." https://kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset (accessed Mar. 09, 2020).

[15] M. Hofmann and R. Klinkenberg, Eds., *RapidMiner: Data Mining Use Cases and Business Analytics Applications*, 1 edition. Boca Raton: Chapman and Hall/CRC, 2013.

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

[17] W. Wang, J. Tang, and F. Wei, "Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China," *J. Med. Virol.*, vol. 92, no. 4, pp. 441–447, 2020, doi: 10.1002/jmv.25689.