NAIVE BAYES CLASSIFIER FOR PREDICTING CORONA VIRUS

Mr. A. Hemanth Kumar[1], G. Pavani[2]

[1]Associate Professor, Dept of CSE, Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, AP, India.

[2]PG Scholar, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, AP, India.

Abstract: Coronavirus, often known as SARS CoV 2, has spread quickly over the world. There is still a lot to learn about this new pathogen. Numerous nations must be alert in battling this new virus due to its quick spread. The patient dataset from the initial SARS CoV 2 outbreak in China is used in this investigation. The Naive Bayes method of data mining is used to process the data. Additionally, the simulation method is used to determine the best value for the results.

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

1. INTRODUCTION

SARS CoV 2, a brand-new virus, started to circulate in China in 2019. After several months, it has expanded outside of China. However, at the time this study was produced, the transmission of these novel viruses in China started to be curbed; however, the spread of viruses outside of China started to pick up. The three nations outside of China with the greatest number of cases of Covid-19, the illness brought on by the SARS CoV 2 virus, are South Korea, Iran, and Italy.

When this study's findings were written, there were 110.099 Coronavirus cases, 3.831 fatalities, and 62.332 recoveries, according to worldometers.info. There are 80.737 cases in China, 7.382 in South Korea, 7.375 in Italy, and 6.566 in Iran [1]. Although there are numerous cases worldwide, the WHO data indicates that the death rate is lower than that of SARS or MERS (Middle East Respiratory Syndrome) [2].

Despite having a lower mortality rate than SARS and MERS, the SARS CoV 2 virus's rapid spread appears to have had an impact on the economies of the affected nations. Although there are currently only 34 cases in Indonesia, the tourism industry is beginning to feel the effects. occupancy rates in hotels fall by 40 percent. A 48 million dollar loss in revenue is possible for the retail sector.

The impact will be seen by more than 495 different commodity kinds, or 13% of all commodities exported to China. As many as 299 Chinese imports are expected to decline or even vanish from the Indonesian market as a result of this virus [3].

Since the SARS CoV 2 virus is a novel strain, there are currently no effective treatments for the illnesses it causes. There is still much to learn about this virus, including how it spreads to other

individuals, what causes death most frequently, and whether or not death risk may be forecast.

To extract insights from data, many academics now employ data mining techniques. Data mining is being used more frequently as a result of faster internet speeds and improved computer data processing capabilities. [4]. We can extract patterns from data using data mining techniques. These patterns fall into one of four categories: association, prediction, cluster, and sequential. Since classification data mining is the most used data mining technique, we shall apply it in this study. As a member of the machine learning family, classification makes use of supervised learning.

Data mining approaches for classifying data include a variety of methods, including decision tree analysis, statistical analysis, neural networks, and Bayesian classifiers. The Navies Bayesian approach will be used in this study to test the model's accuracy against the data. After modelling is completed, simulations will be run to determine the best outcomes.

2. THEORY AND METHODS

The data in structured databases are characterised with real, fresh, maybe beneficial, and eventually intelligible patterns through a process called data mining. Data mining is the process of analysing numerical and categorical data from huge, complex data collections. The phrase is frequently used to describe more sophisticated techniques, including text, online, or geographic data[5].

Analysis, identification, and development of relationships and patterns in existing data are all topics covered by data mining. The practise of finding patterns in data that could lead to non-testing projections of undiscovered patterns is thus described as data mining [6].

Regular statistical analysis and data mining serve different objectives. Traditional statistical approaches place a strong emphasis on testing stated hypotheses, whereas data mining methods examine a wide range of potential, largely untested ideas. [7]. The only way to gain more understanding and information from the always growing volume of digital data is to combine statistical and data mining approaches. such as Witten et al. [7]. The only way to gain more understanding and information from the always growing volume of digital data is to combine statistical and data mining approaches. As Witten and others. According to [6], analysing varied and complicated data in the future will require a fusion of disciplines and techniques such as pattern recognition, databases, artificial intelligence, and machine learning algorithms, in addition to a mix of data mining and statistical methods.

In this study, a Nave Bayes Classifier is used to analyse the dataset. It is a classifier that determines the likelihood that a given collection of data belongs to a particular class[8], in this case, whether or not a person will die. According to Wang [9], the Naive Bayes classifier has gained acceptance as a simple probabilistic classifier based on Clearly separate premises from the Bayesian theorem interpretation In other words, a Naive Bayes classification implies that there is no relationship between the occurrence of one particular function of a class and another. A fundamental probabilistic categorization known as naive Bayesian is based on the Bayesian theorem's independence principle [10].

Zhang [11] claims [12] Classification is a fundamental component of machine data mining. The goal of a research algorithm in classification is to produce a classifier with class labels. The Naive Bayes techniques are a group of supervised learning algorithms that rely on the Bayes theorem and assume that each pair of features is independent of each class variable [12]. The mathematical Bayes model states the following relationship between the dependent function vectors x1 through x2 and the class variable y:

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

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

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

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

$$\Downarrow$$

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

Because $P(x_1,\ldots,x_n)$ is constant, the following classification rule can be applied:
Since the latter represents the relative y-frequency of the training set, we can apply the approximation of P(y) and P(xi|y) using Maximum A Posteriori (MAP). The selection premises of P(xi|y) are the main difference between the numerous naive classificators in Bayes.

Naive Bayer classifiers have done well in numerous real-world situations, including popular data classification and spam learning

detection, despite the ostensibly oversimplified underlying assumptions. To determine the proper settings, only a modest amount of testing data is required. In contrast to more sophisticated techniques, naive Bayes trainers and classifiers can be extremely quick. It is possible to calculate each distribution separately as a one-dimensional distribution by detaching the class conditional feature classes. This inadvertently lessens the effects of the computational complexity curse. On the other hand, while Bayes is a good clustering technique, it is acknowledged that it is a poor estimator, hence the results of estimate proba likelihood calculations are not taken at face value [12].

The typical approach used to carry out data mining projects is a broad one. To maximise the likelihood of success when executing data mining initiatives, data mining researchers and practitioners proposed a number of procedures (workflows or easy step-by-step approaches). The Cross-Industry Standard Process for Data Mining (CRISP-DM), which was proposed as a non proprietary standard practise for data mining by a European consortium of companies in the mid-1990s [4], is perhaps the most widely used structured approach.

The suggested process, which consists of six steps, starts with a thorough understanding of the company and the need for the data mining project (i.e., the application domain), and it concludes with the deployment of the solution that satisfies the specific business need. Although these steps are in order, there is frequently a lot of backtracking.

Considering that data mining is dependent on testing and experience, the entire

procedure might be time-consuming and iterative (i.e., one should anticipate repeating the stages several times), depending on the complexity of the problem and the analyst's skill level. The results of earlier steps serve as the foundation for following phases, thus more attention should be devoted to previous steps to make sure the entire study doesn't go in the wrong direction.

First: Business comprehension

Any data mining project must have a clear understanding of the analysis. Beginning with a thorough comprehension of the firm's demands for fresh information and a clear description of the report's business objective, one can respond to this question. The same methodology is employed even though the analysis in this study focuses on the distribution of Covid-19 rather than business. Therefore, before processing the data, the phenomenon related to the Covid-19 spread case must be understood.

Case understanding would be the proper phrase to use to describe this essay. Everything began in a seafood market in Wuhan that also sold wild animals. Regarding the source of the virus's transition from animals to humans, many theories have emerged. According to some research, bats are the source of the virus. The infection subsequently spread from person to person. Authorities in China rapidly secure Wuhan.

Despite the Wuhan lockdown, the virus spread to other nations and is now officially considered a global pandemic[13]. This is a novel type of virus; first known as the Coronavirus, it is now officially known as SARS CoV 2 and the viral illness is known as Covid-19. According to what we know, the virus originated in Wuhan and soon spread throughout the country. The majority of the deaths, over 3.000 as of this writing, are in Wuhan, China.

Step 2: Understanding the Data

A well-defined task is required for a data mining analysis, and different business tasks call for different sets of data. The primary goal of the data mining method is to classify the pertinent data from the numerous databases that are available after business understanding. Several important factors need to be taken into account during the data identification and gathering procedure. To find the most pertinent data, the analyst must first define the data mining activity in a very clear and straightforward manner.

Step 3: Gathering Data

Data preparation for processing using the most effective data mining techniques is the aim. often referred to as data preparation. Take the data that was defined in the previous phase is the goal. Compared to other CRISP-DM phases, data pre-processing requires the most time and effort; many estimate that it takes up about 80% of the total time spent on the data mining project. Real-world data is typically unreliable (deficiencies in attributes of interest or data aggregates alone), messy (including abnormalities or outliers), and unclear, which is why this major project is being undertaken (containing inconsistencies in codes or names).

Step 4: Building the Model

Then, for a collection of data that has previously been configured, various modelling strategies are picked and applied to address the particular market criteria. Evaluation and comparison of the many built models are also part of the model construction step. Benefits include a range

of workable models as well as a clearly defined experiment and assessment methodology to discover the "correct" method for a particular task because the best technique or algorithm for a data-mining activity is not always known. Even with a single approach or algorithm, a number of factors must be tuned to get the greatest outcomes.

Testing and Evaluation in Step 5

Phase 5 involves testing and evaluating the generality and accuracy of the models that have been created. This stage examines the extent to which (i.e., should further models be developed and assessed?) and the degree to which this model (or models) satisfies the corporate goals. If time and financial restrictions allow, the created model(s) in a real-world scenario can also be tested. Other insights are frequently seen, which may not necessarily be related to the original business objectives but may also provide extra information and recommendations. Although the outcomes of the models are expected to correlate to the original business targets.

Step 6: Implementation

The process of data mining is not complete after modelling and simulation. Even though the model is meant to make the data simple to analyse, it is crucial to organise and present the information learned through this discovery so that end users can comprehend it and take advantage of it. The execution stage could be as simple as reporting Depending on the needs, a persistent data mining technique can be as complicated as a generation. The client, not the data analyst, often handles the implementation phases.

Because it is not a business case, we did not apply the deployment stage in this study. As a result, using the literature review mentioned above, the process is described in the next image.

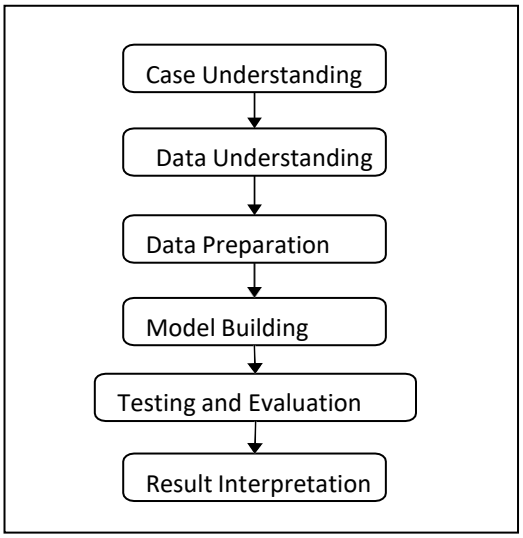


Figure 1. Methodology

This study makes use of a patient dataset obtained from Kaggle by SRK [14], who has gathered data on Covid-19 from a variety of sources. The researchers then use the RapidMiner tool for data mining operations in order to determine the pattern of Covid-19. Rapidminer is an ecosystem created by the same company that uses advanced data science, machine learning, profound learning, text mining, and predictive analytics. It applies to business and business applications, as well as research, training, fast prototyping, and application development, and it facilitates all levels of machine learning, including data processing, outcome analysis, model testing, and optimization [15].

3. RESPONSE AND CONVERSATION

Understanding what is going on with the case and the data is the first step. According to a report from the Chinese WHO Country Office to the Chinese government on December 31, 2019, all of these cases of pneumonia with an

unknown cause originate in the city of Wuhan in the Chinese province of Hubei (unknown cause). Between December 31, 2019, and January 3, 2020, national authorities in China reported to WHO a total of 44 cases of pneumonic illness with no known aetiology. During the reporting period, the causative agent was not found. On January 11 and 12, 2020, the WHO provided the China National Health Commission with comprehensive information about the outbreak in the Wuhan area exposure sector.

The Chinese government discovered an isolated new type of coronavirus on January 7, 2020. China released the genetic code for the brand-new coronavirus on January 12 so that other nations might use it to create their own diagnostic tools. The Ministry of Public Health Thailand said on January 13, 2020, that it has imported its first case of a novel coronavirus, lab-confirmed (2019-nCoV) from Wuhan Province of Hubei, China.

The Ministry of Health, Labor, and Welfare of Japan identified a laboratory-confirmed 2019 new coronavirus (2019-nCoV) that was introduced in Wuhan, Hubei Province, China, on January 15, 2020. [16].. Since then, the sickness known as Covid-19 caused by the virus known as SARS CoV 2, which has since spread to several nations outside of China.

The data will then need to be studied and prepared for the modelling procedure. The researcher then carefully examined the data that she had collected. There are more than 1200 records in the file, which is formatted as an excel file with 21 columns. The age column in the label column is followed by any other columns that are regarded to not be relevant to the study's objectives.

There are other values in the death column, but only numbers 1 and 0 are used in this study, indicating that the patient is dead and not. Patients with a 1 are classified as yes patients, while those with a 0 are classified as no patients. Only 42 records of patients who passed away are available (yes). Consequently, in order to balance the data between yes and no, the There are now 100 data when the researcher adds 58 more records with death values equal to 0.

Table.1 below shows a few of the tables that provide information to be used in the subsequent stage. We remove a number of columns from the database that we believe have no bearing on the study, such time recorded, summary, etc. Our focus in this study is on applying Naive Bayes to predict death based on factors like age, sex, location, and other factors.

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
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

Following the stage of data preparation, RapidMiner is used to process the data. RapidMiner is an effective tool for data mining without the need for coding. RapidMiner uses operator boxes, and each operator box needs to be connected by the user. The operator box used with the Naive Bayes method is demonstrated here.

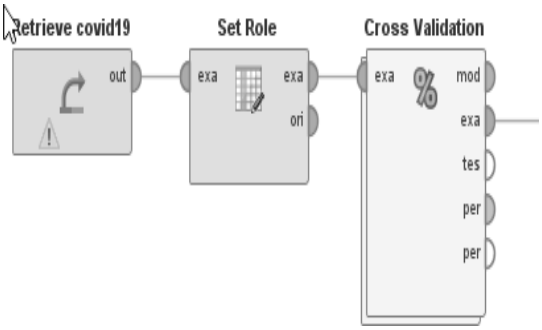


Figure 2. Process in RapidMiner

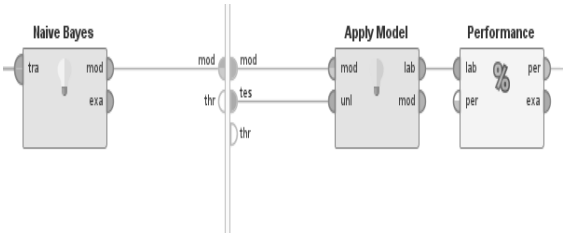


Figure 3. Cross Validation Process

For the simulation process, we use RapidMiner's auto model feature. According to descriptive results from the data, which are 42 yes and 58 no, the processed data is pretty evenly balanced

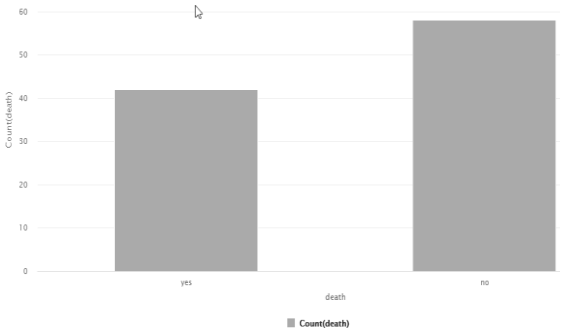


Figure 4. Death Column Distribution

The next slide shows additional descriptions of the data. We can see that while age is an integer data type, attributes like death, location, sex, visit wuhan, and

from wuhan are polynominal types. Additionally, the data does not contain any gaps. This occurs as a result of utilising Excel to carry out the data cleansing process.

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 confident enough to proceed with using the aforementioned dataset. The categorization method is used to process the data, and the death column's category of concern is marked as "yes." So, using the data from the dataset above, we can forecast a number of factors that cause death. Additionally, simulation is used to enhance the forecast.

The two steps most frequently used in classification for prediction methodologies are model development/testing and model evaluation/implementation. During the model development stage, a variety of input data are employed, including the most recent labels. After being taught, a model is checked for accuracy using the holdout sample before being applied practically wherever it is needed. [4].

In order to reduce the bias associated with a random sampling of the training and holdout test samples, one can utilise a technique called k-fold cross-validation, as opposed to the predictive precision of two or more procedures. In a process known as rotational estimation, the entire data set is randomly split into k-exclusive subsets of roughly equal size.

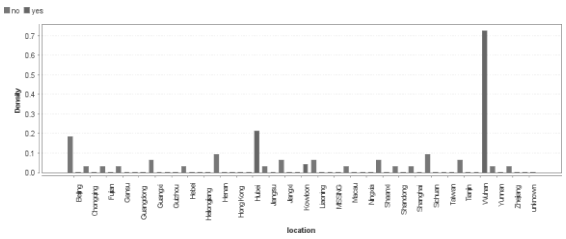
K periods are spent training and testing the models. K days. Every time, only one fold

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

is used to condition it, and the remaining fold is used to check it. Simply adding the k distinct precision measurements yields the average accuracy estimate for a model, as indicated in the following equation.

Where k is the number of folds utilised, A is the accuracy metric (such as hit rate, sensitivity, or specificity) of each fold, and CVA stands for cross-validation accuracy. The results of this study's 10 fold cross-validation are displayed in a confusion matrix.

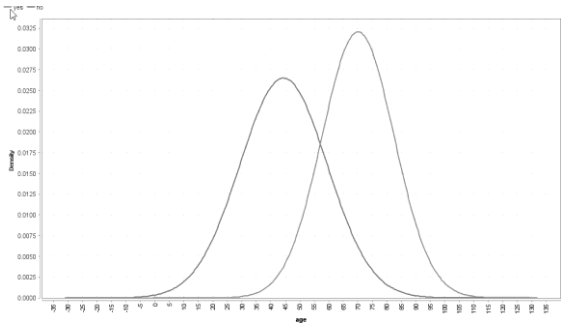
The following figure describes the density of location attribute. Because the data show the early stages of the spreading



phase, Wuhan has the highest density.

Figure 6. Location Density

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



The confusion matix is shown in the next figure.

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

Figure 8. Confusion Matrix

The confusion matrix displays the findings of the RapidMiner cross-validation. As we can see, the Naive Bayes classifier was able to correctly predict 52 no and 1 yes while actually correctly predicting 41 yes and 6 no. As a result, we achieved a 93% model accuracy, which is acceptable.

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

With a 0.172 root mean squared error and an 87.5% correlation, the model performance vector is adequate. Thus, we may be certain that Naive Bayes does a good job of explaining the data.

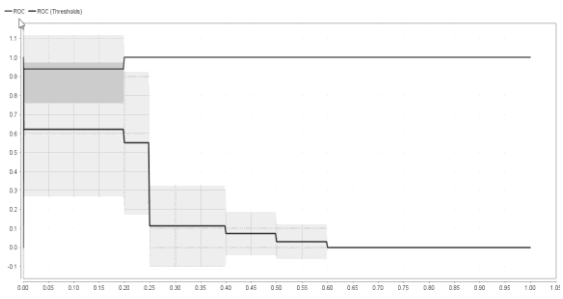


Figure 10. ROC

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.

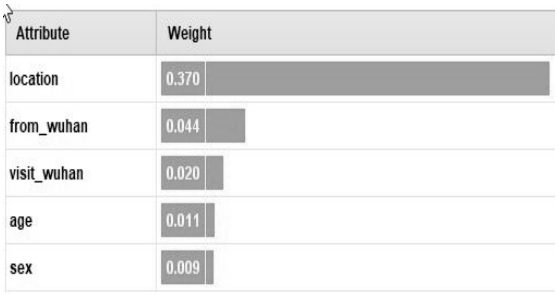
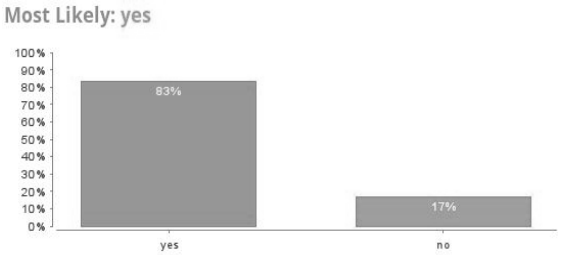


Figure 11. Attribute Weight

According to figure 11, location, from Wuhan, visit Wuhan, age, and sex are the most crucial variables. The information above demonstrates that geography, particularly the disease's centre of transmission and visits to that region, are the most significant determinants before age and sex. Therefore, it makes sense given that the dataset we utilise focuses on the early stages of viral propagation. However, this demonstrates unequivocally that, in light of our results, avoiding the epidemic's epicentre is the safest course of action. Additionally, isolating the contaminated region is the greatest course of action to prevent the virus from spreading further.

The next step was an attempt to use the simulation feature on RapidMiner. First, we examine what transpired in the case of a man, around 55-year-old, Wuhan

resident Figure 7a. Attribute on



RapidMiner Simulation

Figure 12. Simulation Output

Then, using the following outcomes, we attempt to optimise in the group no, with 100% belonging to the no group. This optimization is one of RapidMiner's best features.

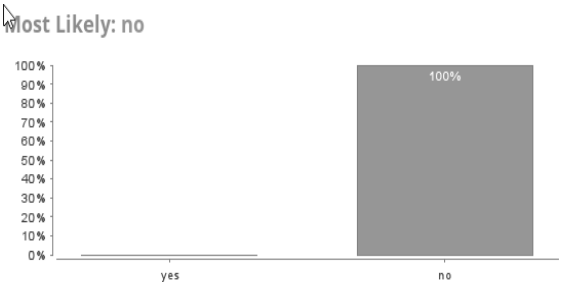


Figure 13. Simulation Output for 'No'

The simulation revealed that the critical elements for the best no are depicted in figure 14 in order to have a 100 percent probability of no.

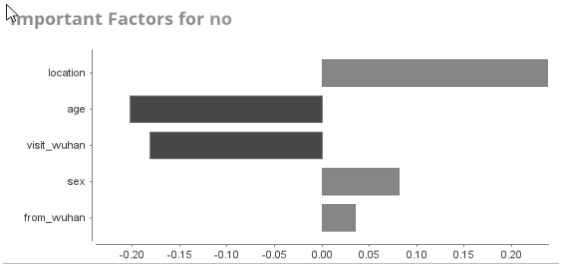


Figure 14. Optimum Attribute

Location, sex, and being from Wuhan are the criteria that favour no group the most. Age and visiting Wuhan are the only variables that conflict for any group.

Our research supports Wang's et al. [17], who claim that the majority of the first fatalities were senior adults, who may have rapid illness development. The median number of days between the onset of the first symptom and death was 14.0 (range 6–41); it tended to be shorter in those who were 70 years or older (11.5 [range 6–19] days) than in those who were younger (20 [range 10–41] days; P =.033).

4.CONCLUSION.

Numerous nations have been impacted by the SARS Co V 2 virus' rapid spread. Despite the fact that 3,4% of deaths attributed to this virus are fatal, the whole population is in fear. Numerous events have been cancelled due to the disruption of global economies; in fact, the Italian league's series A has to be postponed.
This article demonstrates that Naive Bayes can accurately explain the dataset with a 93% confidence level. Location, from and visit Wuhan, age, and finally sex should be used to classify the dead group because they are the most crucial factors. These results support earlier research that connected sex and age to virus-related deaths. Additionally, we are more and more confident that location and distance are significant factors. require attention
It may be inferred from the simulations that age, in addition to location, is the determining factor of death from this virus. A person is less likely to die from this virus the younger they are. Distance from Wuhan and locations that have never been there might lessen the likelihood of being put into the death gangs. Data show that women have a higher survival rate than males.

5.REFERENCES

[1] According to Worldometer, the COVID-19 Wuhan China Virus Outbreak has resulted in 110,099 Cases and 3,831 Deaths. https://www.worldometers.info/coronavirus/ (retrieved March 9, 2020).

[2] Noah B. L. J. Higgins "WHO reports a higher-than-expected worldwide coronavirus fatality rate of 3.4%," 2020 Mar. 03, CNBC. According to this source, the global coronavirus fatality rate is 3 point 4 percent greater than previously believed. (retrieved March 9, 2020).

[3] According to Katadata.co.id, "Efek Domino Virus Corona ke Industri Penunjang Pariwisata," Mar. 02, 2020.efek- domino-virus-corona-ke-industri-penunjang-pariwisata (katadata.co.id/berita/2020/03/02) (retrieved March 9, 2020).

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

Data mining and connected open data: New approaches for data analysis in environmental research, Ecol. Model., vol. 295, pp. 5-17, Jan. 2015, doi: 10.1016/j.ecolmodel.2014.09.018.

[6] Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, by I. H. Witten, E. Frank, and M. A. Hall. Morgan Kaufmann, 2011; Burlington, MA

[7] Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition, by I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. Morgan Kaufmann, 2016. Amsterdam

[8] "Nave Bayes Classifier And Fuzzy Logic System For Computer - Aided Detection And Classification Of Mammamographic Abnormalities," by M. S. Sequera, S. A. Guirnaldo, and I. D. P. Jr. Vol., P. 12, 2005.

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

The article "Detection of Pathological Brain in MRI Scanning Based on Wavelet-Entropy and Naive Bayes Classifier, Bioinformatics and Biomedical Engineering, Cham, 2015, pp. 201-209, doi:10.1007/978-3- 319-16483-0 20.

The Optimality of Naive Bayes, by H. Zhang, p. 6.

[12] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning, vol. Learning Research, Vol. 12, Oct. 2011, p. 2825–2830.

[13] "WHO Director-opening General's remarks at the COVID-19 - 11 March 2020 media briefing." Opening statements by the director general of WHO at the COVID 19 media briefing may be found at this link: 11 March 2020 (accessed Mar. 12, 2020).

[14] “Novel Corona Virus 2019 Dataset.” Corona virus new 2019 data collection is available at https://kaggle.com/sudalairajkumar (accessed Mar. 09, 2020).

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

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

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