Prediction System for the Spread of Corona Virus in Central Java with K-Nearest Neighbor (KNN) Method

1stFarid Fitriyadi, Muqorobin**\***2
Informatika Universitas Sahid Surakarta, Informatika ITB AAS Indonesia
*Universitas Sahid Surakarta,* ITB AAS IndonesiaSurakarta, Indonesia

email : faridfitriyadi@gmail.com, robbyaullah@gmail.com

*Abstract—*Corona Virus is currently spreading very rapidly in many parts of Indonesia, including Central Java Province. According to the current data of corona database in Central Java, today on 17th of August 2021, the number of confirmed cases is; Confirmed in Treatment (Active Cases): 16.344, Confirmed Recovered: 408.697, and Confirmed Dead: 29.148. Therefore, the total number of cases is 454.189, obtained from the sum of the number of being treated, recovered, and dead. Corona Virus is a collection of viruses that can infect the respiratory system, generally mild, such as common cold, although, some forms of diseases like; SARS, MERS, and COVID-19 are more deadly. In anticipating this case, the government has created some policies which include; limiting activities outside the house, having school activities done from home, working from home, and even having religious activities done from home too. The purpose of this study was to predict the possible rate of new cases in one of Central Java areas with confirmed cases of corona virus. Thus, it can be used as information material for the public to anticipate early. The research method applied in this research is problem analysis and literature study, data collection and implementation. The application of the K-Nearest Neighbor (KNN) method is expected to be able to predict the level of spread of COVID-19 in Central Java. The results of the research on testing the prediction system for the new cases level were tested in the Sragen area. Testing is carried out by taking samples for new cases, namely Kudu Regency/City, Confirmed: 17,599, Treated: 89, Recovered: 18,303, Died: 1,721, Suspected: 87 and Discarded Suspected: 1,711. After doing the prediction with K-NN algorithm, it showed the Condition: High.

Keywords : *K-Nearest Neighbor (KNN) Method*, *Corona Virus*

1. **INTRODUCTION**

The world is currently cautious of the spread of a virus known as the corona virus. Corona viruses (CoV) are part of a family of viruses which cause illnesses ranging from the flu to more severe diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). The disease caused by the corona virus known as COVID 19 is a new type that was discovered in 2019 and has never been identified to attack humans before (World Health Organization, 2019). The case of the corona virus appeared and attacked humans for the first time in the province of Wuhan, China. Initially, it was suspected that it was pneumonia, with flu-like symptoms in general. These symptoms include cough, fever, fatigue, shortness of breath and no appetite. However, unlike influenza, the corona virus can develop quickly to cause more severe infections and organ failure (Alvina Felicia Watratan, 2020).

This emergency condition especially occurs in patients with previous health problems. Due to the very fast transmission of the corona virus, the World Health Organization (WHO) established the corona virus as a pandemic on March 11, 2020. The status of a pandemic or global epidemic indicates that the spread of COVID-19 is occurring very quickly that almost no country in the world can ensure that they are safe from the corona virus (Widiyani, 2020).

COVID-19 has spread to 196 countries. There have been 414,179 positive confirmed cases and 18,440 dead from positive confirmed cases. Meanwhile, COVID-19 was first confirmed in Indonesia on March 2, 2020 as of March 25, there were cumulatively 790 confirmed positive cases (cumulatively), including 58 died, 31 recovered, and 701 were in treatment. Based on these conditions, Indonesia is in a state of alert to the threat of the corona virus and so far no vaccine for COVID-19 has been found, so a form of prevention from the spread of the virus can be done by breaking the chain of transmission. One method to break the chain of transmission is by doing social distancing. With the social restrictions, it is hoped that every community will not become infectious or infected because they do not make contact with anyone so that the rate of spread can decrease. In this study, we will apply the Naive Bayes algorithm to predict the level of spread of COVID-19 in Indonesia by implementing the Naive Bayes Algorithm, which is useful as a measure to anticipate the COVID-19 pandemic (Alvina Felicia Watratan, 2020).

Data mining is a process of extracting or mining large data and information, which have not been known previously, but can be understood and useful from large databases and are used to make very important business decisions. Data mining describes a collection of techniques with the aim of finding unknown patterns in the data that has been collected. Data mining allows users to find knowledge in database data that is impossible for the user to know the existence of. Data mining is a process of automatically searching for useful information in large data storage places. As a series of processes, data mining can be divided into several stages of the process. These stages are interactive; the user is directly involved or through a knowledge base.

The stages of Data Mining are:

1. Data Cleaning

Data cleaning is the process of removing noise and inconsistent or irrelevant data.

1. Data Integration

Data integration is the merging of data from various databases into one new database.

1. Data Selection

The data in the database is often not used entirely, therefore only data that is suitable for analysis will be retrieved from the database.

1. Data Transformation

Data is converted or merged into a format which is suitable for processing in Data Mining.

1. Mining Process

It is a major process when methods are applied to find valuable and hidden knowledge from data. Several methods can be used based on data grouping. Pattern Evaluation is to identify interesting patterns in the knowledge based that are found.

1. Knowledge Presentation

It is a visualization and presentation of knowledge about the methods used to obtain the knowledge obtained by users.

K-Nearest Neighbor is one of the methods in making decisions using supervised learning where the results of the new input data are classified based on the closest in the value data [3].

The K-Nearest Neighbor (KNN) algorithm is a method for classifying objects based on the learning data that is closest to the object. KNN is a supervised learning algorithm where the results of the new query instance are classified based on the majority of the categories in the KNN algorithm. Where the class appears the most will be the class resulting from the classification [4].

Closeness is defined in terms of metric ranges, such as Euclidean range. Euclidean range [5] can be found using equation 1 below.



D : range of closeness

x : data training

y : data testing

n : the number of individual attributes between 1 to n

f : similarity function between case X and case Y

i : individual attributes between 1 and n

The steps to calculate the K-Nearest Neighbor method [6] include;

1. Determine the parameter K (number of closest neighbors).

2. Calculate the square of the Euclidean range (query instance) of each object to the given sample data using equation 1.

3. Then sort the objects into groups that have the smallest Euclidean range.

4. Collect category Y (Nearest Neighbor Classification)

5. By using the category of Nearest Neighbor which is the majority, the value of the calculated query instance can be predicted.

PHP according to Anhar (2020: 3) is a server-side web programming language that is open source, PHP is also a script that is integrated with HTML and is located on the server (server side HTML embedded script). PHP is also a script used to create very dynamic web pages, dynamic means that the display page that will be displayed is created when the page is requested by the client. PHP was first created by Rasmus Lerdorf, a reliable C programmer from Greenland Denmark in 1995, PHP was named FI (Form Interpreted) which was used to manage from the web. In its development, the codes used were released to the public so that many programmers around the world began to develop them. In 1997 PHP was released with version 2.0, this version was already integrated with the C programming language and was equipped with modules so that the quality of PHP's work was significantly improved. In the same year, a programming company called Zend re-released this version of PHP with a better, cleaner and faster version. Along with the development of the era in 1994, PHP version 4.0 began to be released and this version was most widely used in the early 21st century because this version of PHP was able to build complex webs with high speed stability. In 2004, the programming company Zend released PHP again with its latest version 5.0 which the core of the PHPH interpreter underwent major changes. This version also incorporated an object-oriented programming model into PHP to answer the development of programming languages towards an object-oriented paradigm (Anhar, 2020).

Previous research by Rizki Tri Prasetio and Sari Susanti, from BSI University, 2019 with the entitled "Life Expectancy Prediction of Lung Cancer Patients Post Thoracic Surgery Using Boosted k-Nearest Neighbor" concluded that lung cancer ranks sixth out of the ten leading causes of death in Indonesia. Factors causing lung cancer are dominated by cigarette smoke. Thoracic surgery is one of the main solutions for lung cancer. However, there are many risks and complications after thoracic surgery that can lead to death. In this study, the life expectancy of lung cancer patients will be predicted after living one year of life after thoracic surgery using computer aided diagnosis (CAD). This prediction is done by analyzing the patient's condition before and after surgery. The data used in the study was secondary data which contained 470 data with a distribution of 400 patient data that were alive (survival) and 70 data of patients who died (die). Adaptive Boost was used as a level optimization algorithm on the k-nearest neighbor algorithm. The results of the study showed that the proposed method produces an accuracy of prediction of life expectancy of 85.11% using 10-fold cross validation with the k parameter in the k-nearest neighbor algorithm with a value of 5.

Previous research by Rahandanu Rachmat from Jendral Ahmad Yani University, 2020 with the title "Water Quality Prediction System in Tirta Raharja Regional Drinking Water Company Using K – Nearest Neighbors (K – NN)" concluded that PDAM (Regional Drinking Water Company) Tirta Raharja is one the only Regional Owned Enterprise having task of providing clean water services for the people of Cimahi City. Clean water is a basic need that is always consumed by the community and affects community activities. The development of the city of Cimahi is currently quite rapid, with the plan to build a smart city, causing the domestic demand for clean water to increase. K – Nearest Neighbor (KNN) is a classification algorithm that considers several supporting parameters to carry out the classification process which results in ease of calculation time and power. KNN can be considered as one of the most well-known non-parametric models. In the research and implementation process of data mining on determining the feasibility of water quality at PDAM Tirta Raharja using K – Nearest Neighbors, it can be concluded that K – Nearest Neighbors which is implemented in the process of testing the feasibility of drinking water at PDAM Tirta Raharja, produces an accuracy of 93% for testing with the Eligible label. Drinkable, and 98% for accuracy testing with the label Not Worth Drinking with a K value of 14 where the K value is the most ideal amount that has been processed through K – Fold Cross Validation from a total of 1,818 data.

1. **RESEARCH METHODS**

The research method is a scientific process or method to obtain data that will be used for research purposes. The research was carried out in stages starting from planning, determining the research focus, research time, data collection, analysis, and presentation of research results.

The research methods applied in this study are as follows:

1. Problem Analysis and Literature Study

This stage is the first step to determine the formulation of the research problems, which in this case is observing problems related to the level of spread of COVID-19 that occurs especially in Indonesia. The existing problems are then analyzed to find out how to solve them and determine the scope of the problems to be studied. The theoretical basis from various literatures regarding the application of the Naive Bayes method, concepts and theories of data mining and prediction of the rate of spread of COVID-19 in Indonesia, from journals are studied in order to obtain a knowledge base to conduct further research.

2. Collecting Data

The systematic procedure used to collect data is quantitative research methods. The implementation of quantitative research methods focuses on the use of numbers, tables, graphics and diagrams to display the results of the data obtained and those will be used for data analysis of the K-Nearest Neighbor method. As for data collection, researchers took data directly on the official web address: https://corona.jatengprov.go.id/data. After the data is collected, data analysis is carried out to adjust the data process to be processed in the K-Nearest Neighbor method.

3. Implementation

In accordance with data processing, the implementation stage is about how the data processing is applied in a system. As for testing this system, the researcher built a prototype using the PHP programming language to implement the Prediction of the Spread of Corona Virus Cases in Central Java. Further testing was done to determine whether the research conducted was in accordance with the expected objective which is to predict the level of spread of COVID-19 in Central Java Province.

1. **RESULT AND Discussion**

In this discussion, the researcher wrote down the stages of the data processing process to produce a prediction system for the spread of the corona virus in Central Java.

1. Data on the spread of the corona virus in Indonesia increased in July. Meanwhile, in August it has started to experience a decline



Figure 1. Condition of the Spread of the Corona Virus in Indonesia

1. At this initial stage, researcher obtained the data on the spread of corona cases in Central Java by from official website: https://corona.jatengprov.go.id/data
2. After that, the data was taken and data processing was carried out as needed in the K-Nearest Neighbor Algorithm.
3. The data that researcher obtained from the official website is still in the form of original data as shown in Figure 2.



Figure 2. Data Source of the Confirmation of the Spread of the Corona Virus in Central JavaProvince

1. After that, the researcher took and processed the data for the calculation process using the K-Nearest Neighbor method.
2. In the data, there was no classification label, so the researcher added a column to determine the classification of labeling at the case level.
3. As for giving classification labels, the researchers made four case categories, namely Very High, High, Medium and Low.
4. The way to determine these categories is to carry out a data normalization process based on the highest, average and lowest number of cases, then the rank of range between cases is obtained based on case data dated 17 August 2021. Then researchers can retrieve data as shown in table 1.

Table 1. Data Analysis of Corona Cases in Central Java

|  |  |  |
| --- | --- | --- |
| **Description** | **Case** | **Total number** |
| Average | 12.734 | 12.734 |
| MIN | 3.802 | 3.802 |
| MAX | 33.710 | 33.710 |
| Range - High | 33.710 - 12.734 = 20.976 | 20.976 |
| Range- Average | 12.734 - 3.802 = 8.932 | 8.932 |

1. Based on the table above, a category of case conditions can be made, as shown in table 2

Table 2. Number of Level Category of

Corona Virus Cases in Central Java

|  |  |
| --- | --- |
| **Number of Cases** | **Category** |
| 20.976 | - | 33.710 | Very high |
| 12.734 | - | 20.976 | High  |
| 8.932 | - | 12.734 | Average  |
| 3.357 | - | 8.932 | Low |

1. After that, the data of corona cases in Central Java will change with the addition of column of the condition for the level of virus spread, so that the data can be interpreted as training data, this can be seen in table 3.

Table 3. Training Data of Corona Virus

with Training Category

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Regency/City | Confirmed | Suspects | Discarded Suspects | CONDITION RESULT |
| 1 | Semarang Ci | 33.710 | 526 | 11.348 | Very high |
| 2 | Klaten | 26.716 | 992 | 3.788 | Very high |
| 3 | Banyumas | 26.475 | 1.186 | 6.224 | Very high |
| 4 | Jepara | 17.562 | 395 | 3.155 | High |
| 5 | Magelang | 17.336 | 108 | 2.125 | High |
| 6 | Kebumen | 16.911 | 206 | 2.232 | High |
| 7 | Kudus | 16.685 | 49 | 2.711 | High |
| 8 | Cilacap | 16.161 | 713 | 1922 | High |
| 9 | Kendal | 15.815 | 215 | 3676 | High |
| 10 | Purworejo | 15665 | 403 | 2001 | High |
| 11 | Surakarta | 15323 | 244 | 2079 | High |
| 12 | Sragen | 14776 | 180 | 1430 | High |
| 13 | Karanganyar | 14.465 | 614 | 2236 | High |
| 14 | Tegal | 14.390 | 671 | 4613 | High |
| 15 | Semarang | 14288 | 170 | 1850 | High |
| 16 | Wonosobo | 12816 | 114 | 2875 | High |
| 17 | Demak | 11682 | 129 | 3996 | Average |
| 18 | Pemalang | 11434 | 144 | 3929 | Average |
| 19 | Blora | 11181 | 326 | 2113 | Average |
| 20 | Temannggung | 10939 | 130 | 1452 | Average |
| 21 | Boyolali | 10579 | 251 | 1368 | Average |
| 22 | Sukoharjo | 10401 | 621 | 2772 | Average |
| 23 | Wonogiri | 10089 | 94 | 1980 | Average |
| 24 | Purbalingga | 9456 | 300 | 1468 | Average |
| 25 | Banjarnegara | 9034 | 483 | 3626 | Average |
| 26 | Brebes | 8245 | 313 | 1466 | Rendah |
| 27 | Batang | 7560 | 31 | 1018 | Rendah |
| No | City | Confirmed | Suspects | Discarded Suspects | CONDITION RESULT |
| 28 | Pati | 7359 | 31 | 2236 | Low |
| 29 | Pekalongan  | 7210 | 127 | 1520 | Low |
| 30 | Rembang | 6600 | 11 | 889 | Low |
| 31 | Grobogan | 6138 | 74 | 2188 | Low |
| 32 | Magelang | 5591 | 22 | 682 | Low |
| 33 | Pekalongan | 5086 | 12 | 340 | Low |
| 34 | Salatiga | 4219 | 65 | 609 | Low |
| 35 | Tegal | 3802 | 124 | 512 | Low |

1. After that, the researcher conducted a calculation on the K-Nearest Neighbor method. The formula is as follows:



1. As for system testing, researcher will use data testing, which is providing new training data by making case examples as shown in table 4.
2. Based on the testing data, the researchers performed calculations by calculating using the K-Nearest Neighbor (KNN) method to predict classification based on past data.

Table 5. K-NN The Result of K-NN Method Calculation.



1. After that, the researcher took the closest range from the calculation results and categorized it according to the condition category.
2. Based on the calculation of the K-NN Algorithm, the researchers took the value of K = 3. There, it can be seen that there are more number of "Highest" condition categories, so the prediction results state that the testing data produces the following condition: Highest.

Tabel 7. Hasil Klasifikasi Metode K-NN The Classification Results of K-NN Method



1. Based on the results of the classification above, the system testing is stated for the following condition: Highest.

Table 8. The Prediction Result of the Condition of Corona Virus Spread

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Regency | Confirmed | Dead | Suspects | Discarded Suspects | Result |
| Kudus | 17.599 | 1.721 | 87 | 1711 | The highest |

1. To make the implementation of the system easie, the researchers made a prototype system to run the system flow using the K-NN Algorithm method
2. The interface is still made simple using the PHP programming language, the following is a display of the Input Data Testing Interface:



Figure 3. The Implementation of Prediction System

1. After inputting the system, it will produce the following data.



1. The test is stated valid because the results of the appointment are the same as the tests carried out on manual calculations and the prototype program obtains the same results, the system is stated valid.

**VI. CONCLUSION**

Based on study on the Prediction of the Spread of the Corona Virus in Central Java Province with the K-NN Method, the followings are the conclusions that can be drawn:

* 1. The K-NN method utilizes training data to generate the probability of each criteria for a different class, so that the probability values of these criteria can be optimized to predict the level of Covid-19 spread in Central Java based on the classification process which is carried out by the Naive Bayes method itself.
	2. Based on the data on COVID-19 Cases Per Regency/City which was used as training data, the K-NN method successfully classified those which were tested. Hence, the K-NN method is successful in predicting the number of COVID-19 cases per Regency/City.
	3. The implementation of the system by making a prototype of the PHP programming language has been able to test the system by producing test results which are in accordance with the algorithm calculation.

**REFERENCES**

1. Mona, Nailul. 2020. Konsep Isolasi Dalam Jaringan Sosial Untuk Meminimalisasi Efek Contagious (Kasus Penyebaran Virus Corona Di Indonesia). Jurnal Sosial Humaniora Terapan (JSHT), 2(2), pp. 117. doi: https://doi.org/10.7454/jsht.v2i2.86.
2. Saleh, Alfa. 2015. Implementasi Metode Klasifikasi Naive Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. Creative Information Technology Journal (Citec Journal), 2(3), pp. 208-216. doi: https://doi.org/10.24076/citec. 2015v2i3.49.
3. Bustami. 2013. Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. TECHSI (Jurnal Penelitian Teknik Informatika), 3(2),pp. 127-146. doi: https://doi.org/10.29103/techsi.v5i2.154.
4. Ridwan, M., Suyono, H., Sarosa, M. 2013. Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. Jurnal EECCIS, 1 (7), pp. 59-64.
5. Mujiasih, S. 2011. Pemanfaatan Data Mining Untuk Prakiraan Cuaca. Jurnal Meteorologi dan Geofisika, 12(2), pp. 189-195. doi: http://dx.doi.org/10.31172/jmg.v12i2.100
6. Yunus, N. R., & Rezki, Annisa. 2020. Kebijakan Pemberlakuan Lockdown Sebagai Antisipasi Penyebaran Corona Virus Covid-19. SALAM; Jurnal Sosial & Budaya Syar-i, 7(3), pp. 227-238. doi: <https://doi.org/10.15408/sjsbs.v7i3.15083>.
7. Muqorobin, M., & Rais, N. A. R. (2020). Analysis of the Role of Information Systems Technology in Lecture Learning during the Corona Virus Pandemic. International Journal of Computer and Information System (IJCIS), 1(2), 47-51.
8. Hikmah, I. N., & Muqorobin, M. (2020). Employee Payroll Information System On Company Web-Based Consultant Engineering Services. International Journal of Computer and Information System (IJCIS), 1(2), 27-30.