Improved Decission Tree Performance using Information Gain for Classification of Covid-19 Survillance Datasets

Ivandari [1], Much. Rifqi Maulana [2], M. Adib Al Karomi [3*]

[1-3] *Computer Science, STMIK Widya Pratama Pekalongan, Indonesia*

Abstract— One of the most feared infectious diseases today is COVID-19. The transmission of this disease is quite fast. Patients also sometimes do not have the same symptoms. Overcoming the spread of the pandemic has been widely carried out throughout the world. Apart from the medical method, there are also many other methods, including computerization. Data mining is a discipline that can project data into new knowledge. One of the main functions of data mining is classification. Decision tree is one of the best models to solve classification problems. The number of data attributes can affect the performance of an algorithm. Feature selection is a process to remove data attributes that are not needed in the classification. The feature selection process can make the computational process better. Preprocessing with feature selection is also proven to improve the performance of the classification algorithm. One of the best feature selection algorithms is information gain. This study uses information gain to select the attribute features of the Covid-19 surveillance dataset. This study proves that there is an increase in the accuracy of the decision tree algorithm by adding information gain feature selection. Previously, the decision tree only had an accuracy rate of 65% for the classification of the Covid-19 surveillance dataset. After pre-processing using information gain, the accuracy rate increased to 75%.

Keywords— Decision tree, information gain, covid-19 surveillance, accuracy.

1. Introduction

Covid 19 is a disease that has taken the world by storm in recent years. This disease is new with a very fast spread rate. Recent studies have shown that there are several mutations of this virus that are more dangerous to humans. Most countries implement strict health protocols to keep their citizens from spreading this disease. Prevention is also carried out by many countries by vaccinating their citizens. Apart from the medical side, research on this disease is also carried out in various fields. In the computer field, algorithmic development is carried out to classify data and in the end it can become a new policy.

Data mining is a computer science that allows data extraction to gain new knowledge [1]. In data mining, the data used greatly affects the calculation performance. One of the calculation methods in data mining that is proven to be reliable is the decision tree. Decission tree is one of the best classification methods [2].. Besides having good accuracy, this method is also easier to understand because its use is more in line with human language [3].

In this study, the COVID-19 surveillance dataset is sourced from the UCI repository. Uci repository is a dataset provider portal that has been tested and is widely used by researchers to test algorithms. This dataset comes from the Ministry of Health of the Republic of Indonesia. This full version of dataset can be downloaded at https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance.

The classification of the Covid-19 survival dataset using the decision tree algorithm has been carried out and has an accuracy rate of 65% [4]. In this study all attributes are used for classification. The number of attributes can affect the classification results [3].

Feature selection is the process of selecting attributes that will be used in the classification process [5]. This process is done by removing unnecessary attributes from the dataset. The feature selection process can improve the computational process and can even improve accuracy. Information gain is a popular yahoo and is widely used in the feature selection process [6].

The more attributes that have no effect can also reduce the accuracy of an algorithm [7]. This study uses information gain for the selection of the attributes of covid-19 survillance. The decision tree produces an accuracy rate of 75%. This means that the use of information gain can increase the accuracy of the decision tree by 10% for the classification of the Covid-19 survival dataset.

2. Literature Review

2.1. Related Research

Similar research has been conducted using the KNN algorithm [8]. in this study the accuracy rate obtained was only 55%. In addition, similar research using the decision tree algorithm has also been carried out [4]. The decision tree classification in previous studies used all existing data attributes. In this study the decision tree obtained an accuracy rate of 65%.

2.2. Data Mining

Data is a commodity that is widely produced at this time. Data that is too much but has no meaning will only become garbage in our storage [9]. Data mining comes to solve this problem. The process in data mining allows data to be extracted to become new knowledge or information that was not previously known [10]. By recognizing existing data patterns, data mining methods can analyze and find new patterns [11]. In the existing dataset it is possible to create new rules, patterns or models that are different from the previous database [12].

The processes in data mining include collecting data, using historical data to find data linkages [13]. The relationship between these data can later become new information or knowledge. This new knowledge can later be used for decision making in an agency, company and even government.

2.3. Classification

One of the functions of data mining is classification. Classification is classified as supervised learning which requires training data and data testing. Past data is used to create new models or knowledge that can be applied to new data. Classification has proven to be used in the medical field [14], in education [15], is used in building engineering [16], and is widely used in other fields.

The classification process in the computer field is actually only to determine and then improve the accuracy of the algorithm. Furthermore, the classification model formed can be used in various fields according to the existing data. The level of accuracy of the algorithm is influenced by the dataset and the type of data used [17].

2.4. Decission Tree

Decision tree itself is a very powerful and well-known classification and prediction method [18]. In this decision tree, data in the form of facts is converted into a decision tree that contains rules and of course can be more easily understood with natural language. Decision tree models are widely used in cases of data with discrete-valued outputs[13]. Although it is possible it can also be used for data cases with numeric attributes.

Each node in the decision tree represents an attribute. While the branch of the node is the value of the attribute, and the leaf represents the class. The top node in the decision tree is called the root node. This root node has no input and may have no output and can even have more than one output. Internal root is a branching node that has only one input and has at least two outputs. Leaf node or terminal node is an end node that has only one input and no output.

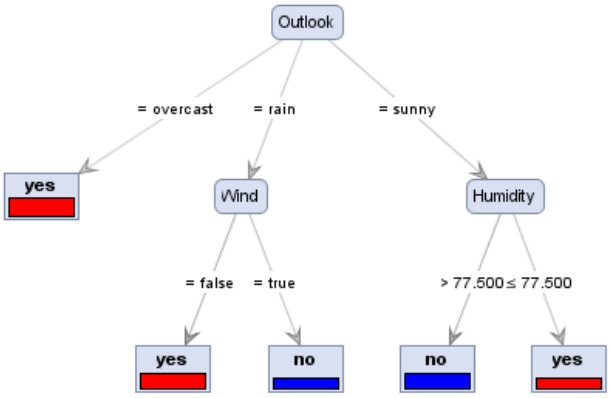


Figure 1. Golf data in decision tree

Figure 1 illustrates the decision to predict whether someone will play golf. The root node or root attribute is symbolized by the obtuse square at the top, namely outlook. The branch is symbolized by a line and the leaf node or terminal node is symbolized by a square with a tip that contains a label or destination, i.e. yes or no. While the internal node in Figure 1 is also symbolized by a tuple square located between the root node and the terminal node.

2.5. Information Gain

Information gain is a feature selection method that is widely used by researchers to determine the limits of the importance of an attribute [5] [19]. The information gain value is obtained from the entropy value before separation minus the entropy value after separation [20]. In calculating the previous information gain value, the total entropy value and the entropy value of the attribute concerned must first be known. The information gain value is the difference between the total entropy value and the attribute entropy to be calculated.

The measurement of the importance of an attribute was first pioneered by Claude Shannon in information theory [21] and is written according to the following equation:

$$info(D) = -\sum_{i=1}^{m} pi \; log_2(pi)$$

Information:
D : Set of cases
M : Number of partitions D
pi : Proportion of Di to D

While pi is the probability of a tuple in D that falls into class Ci and is estimated by |Ci,D| / |D|. The log function in this case uses log-based 2 because the information is encoded bit-based. The calculation of the entropy value after separation can be done using the following formula:

$$info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{ x } I\,(D_j)$$

Information:
D : Set of cases
A : Attribute
v : Number of partition attribute A
|Dj| : Number of cases on j .th partition

|D| : Number of cases in D
I (Dj) : Total entropy in partition

Meanwhile, to find the information gain attribute A, the following formula can be used:

$$Gain\ (A) = I\ (D) - I\ (A)$$

Information:
Gain (A) : Information gain attribute A
I (D): Total entropy
I (A): entropy A

In general, the stages of information gain are carried out in the following way:

1). Calculate the information gain value for all attributes in the original dataset.
2). Determine the desired threshold. With this limitation, it is possible for attributes that have weight equal to the limit or greater than the limit to be maintained and used in the classification stage. Then attributes with weights below the limit will be discarded and not used in the classification stage.
3). The dataset is updated using only the selected attributes.

3. Research Methods

The research method used in this research is experimental. The research uses public datasets and performs algorithmic calculations using the help of rapid miners. The stages in completing the research are as follows:

3.1. Method of collecting data

The data used in this study is the Covid-19 surveillance dataset. The main source of this data is the Ministry of Health of the Republic of Indonesia. This dataset has become public data and can be downloaded at: https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance. In this dataset there are 7 regular attributes and 1 label attribute. Table 1 is the dataset table used.

Table 1. Covid-19 surveillance dataset

A01	A02	A03	A04	A05	A06	A07	Categories
+	+	+	+	+	-	-	PUS
+	+	-	+	+	-	-	PUS
+	+	+	+	-	+	-	PUS
+	+	-	+	-	+	-	PUS
+	-	-	-	-	-	+	PUS
+	+	+	-	-	-	+	PUS
+	+	-	-	-	-	+	PUS
+	+	+	+	-	-	-	PUS
+	-	-	+	+	-	-	PIM
-	+	-	+	+	-	-	PIM
+	-	-	+	-	+	-	PIM
-	+	-	+	-	+	-	PIM
-	+	-	-	-	-	+	PIM
-	-	-	-	-	-	+	PWS

3. 2. Algorithm Calculation Method

Algorithm calculation using rapid miner application. The first stage is feature selection on the COVID-19 surveillance dataset. This stage is to determine the importance of all existing attributes. Then determine the threshold, which is the attribute value limit that will be used in the next stage. Attributes with values below the threshold will not be used in the next classification process.

The next stage is validation. This stage uses 10 folds cross validation. The process is to divide all records in the dataset into 10 parts. 1 part is used as testing data and the other 9 parts are used as training data. This process is repeated until all records have one chance to become testing data.

The last stage is the evaluation of the results. This evaluation process uses a confusion matrix. This process is actually carried out simultaneously with the validation stage. All testing data is adjusted to actual conditions. The percentage of conformity will later be referred to as the accuracy level of the algorithm. Figure 2 is a diagram of the research process.

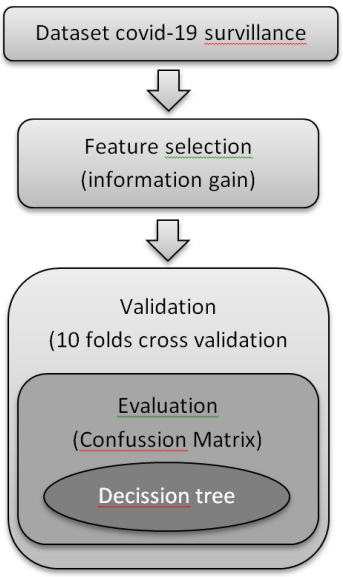


Figure 2. Research process

4. Results and Discussion

4. 1. Information gain feature selection

The COVID-19 surveillance dataset has 7 regular attributes. Feature selection using information gain can determine the level of importance of all data attributes. Table 2 is the level of importance of the Covid-19 surveillance attribute using information gain. The

importance value is between 0 and 1. The higher the value, the greater the importance.

Table 2. Information gain result

Atribute	Weight by information gain
A05	0.0
A06	0.0
A04	0.1818397563031609
A07	0.1818397563031609
A02	0.32999001598621563
A03	0.515387058257064
A01	1.0

4.2. Algorithm Calculation

The calculation of information gain feature selection and algorithm accuracy performance is done using a rapid miner application. In the rapid miner all the algorithms have been prepared and can be used instantly. The opening view of the rapid miner can be seen in Figure 3 below. While in Figure 4 is the welcome screen on the rapid miner.

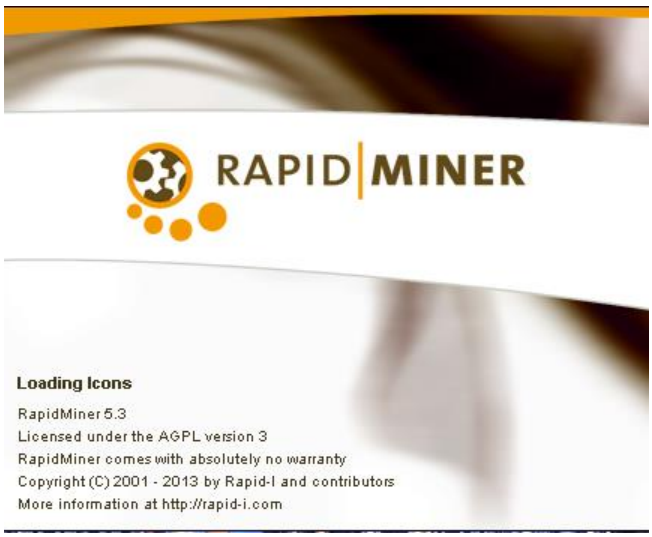


Figure 3. Display of rapid miner

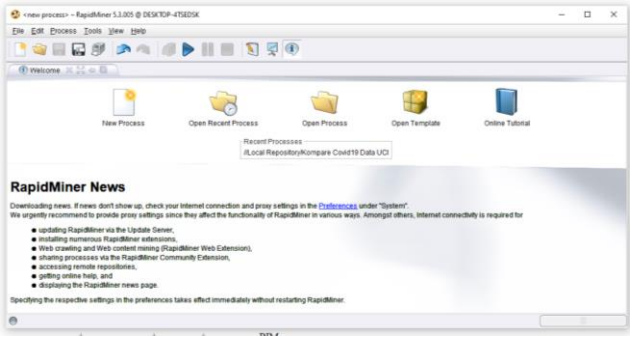


Figure 4. welcome screen rapid miner

The calculation process is carried out by placing the prepared dataset into the worksheet provided by the rapid miner. Figure 5 is a display of the rapid miner worksheet.

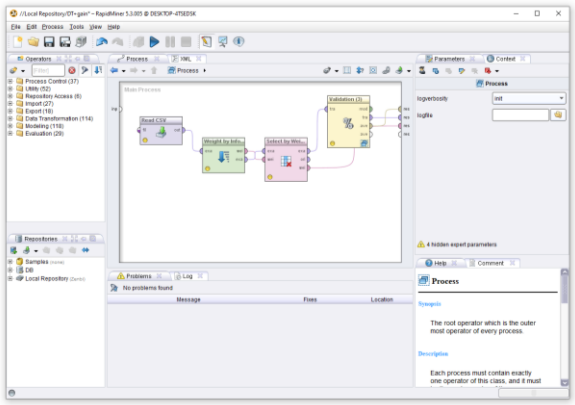


Figure 5. Rapid Miner Worksheet

In Figure 5 there are 4 stages carried out in completing the research. The first stage is the prepared dataset. Dataset preparation including attribute selection. The selection can be done automatically by the rapid miner system. To overcome errors in this study, manual selection of attributes was carried out. The selection of attributes includes the type of attribute and which attribute function will be used as a label.

The second stage is weighting using information gain. This process runs automatically. In this process, the importance of all regular attributes is calculated. The output of the weighting using this information gain can be seen in table 2.

The third stage is the determination of the threshold. In this study used the threshold value is 0.5. This means that all attributes with a gain value of less than 0.5 will not be used in the next classification process. What is used in the next classification process is an attribute with a gain value equal to or greater than 0.5.

The fourth stage is the validation stage. In this stage, 10 fold cross validation is used. In this validation process there is an evaluation. The process of evaluating and calculating the accuracy of the algorithm uses a decision tree. Figure 6 is a process in validation.

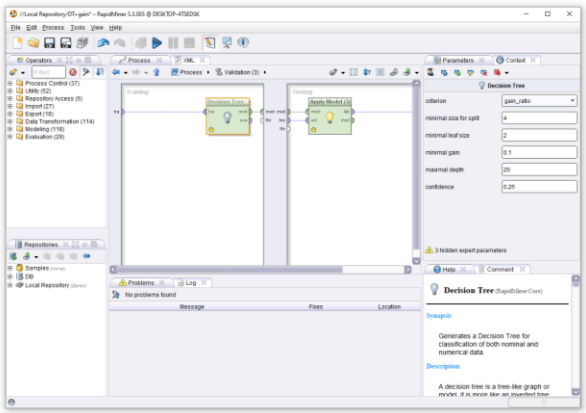
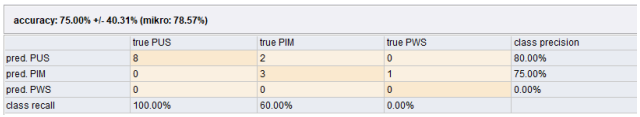


Figure 6. Algorithm calculation process

Calculations are carried out to determine the level of accuracy. Figure 7 is the confusion matrix formed. This confusion matrix is the proportion of label conformity that corresponds to actual conditions. From Figure 7, it can be seen that the accuracy of the decision tree is 75%.



accuracy: 75.00% +/- 40.31% (mikro: 78.57%)

	true PUS	true PIM	true PWS	class precision
pred. PUS	8	2	0	80.00%
pred. PIM	0	3	1	75.00%
pred. PWS	0	0	0	0.00%
class recall	100.00%	60.00%	0.00%	

Figure 7. Calculation results

4.3. Discussion

The decision tree accuracy rate in this study is 75%. Previously with the same data and algorithm the decision tree only obtained an accuracy rate of 65% [4]. an increase of 10% occurs because the attributes used are appropriate. Some attributes are not included in the classification process. The attributes used are only selected attributes in the information gain feature selection process. The right number of attributes can improve the performance of an algorithm. And vice versa, the number of attributes that have no effect on the classification can make the algorithm's performance decrease.

5. Conclusion

From the results of previous studies it can be concluded that:

1. The use of information gain feature selection in the classification of the Covid-19 surveillance dataset can increase the accuracy of the decision tree.
2. An increase in decision tree accuracy of 10% is obtained when the attribute threshold value used is 0.5.

References

[1] Ian H Witten. Eibe Frank. Mark A Hall, *Data Mining 3rd*. 2011.

[2] X. Wu, *The Top Ten Algorithms in Data Mining*. New York: Taylor & Francis Group, LLC, 2009.

[3] M. A. Alkaromi, "Information Gain untuk Pemilihan Fitur pada Klasifikasi Heregistrasi Calon Mahasiswa dengan Menggunakan K-NN," 2014.

[4] Ivandari and M. A. Al Karomi, "Classification of Covid-19 Survillance Datasets using the Decision Tree Algorithm," *Jaict*, vol. 6, no. 1, pp. 44–49, 2021, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/2896.

[5] H. Deng and G. Runger, "Feature Selection via Regularized Trees," Jan. 2012, Accessed: Oct. 16, 2014. [Online]. Available: http://arxiv.org/abs/1201.1587v3.

[6] M. A. Al Karomi, M. R. Maulana, S. J. Prasetiyono, Ivandari, and Arochman, "Strengthening campus finance by analyzing attribute attributes for student registration classifications." p. 1, 2019, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/1431.

[7] Ivandari, T. T. Chasanah, S. W. Binabar, and M. A. Al Karomi, "Data Attribute Selection with Information Gain to Improve Credit Approval Classification Performance using K-Nearest Neighbor Algorithm," *IJIBEC*, vol. I, pp. 15–24, 2017.

[8] Ivandari and M. A. Al Karomi, "Algoritma K-NN untuk klasifikasi dataset Covid-19 survillance," *IC Tech*, vol. 16, no. 1, pp. 12–15, 2021, [Online]. Available: https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/137.

[9] M. A. Alkaromi, "Komparasi Algoritma Klasifikasi untuk dataset iris dengan rapid miner," *IC Tech*, vol. XI, no. 2, 2014.

[10] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier, 2011.

[11] D. T. Larose, *Discovering Knowledge in Data: an Introduction to Data Mining*. John Wiley & Sons, 2005.

[12] E. Prasetyo, *Data Mining Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: Andi Offset, 2012.

[13] B. Santosa, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Edisi Pert. Yogyakarta: Graha Ilmu, 2007.

[14] A. Christobel and D. . Sivaprakasam, "An Empirical Comparison of Data Mining Classification Methods," vol. 3, no. 2, pp. 24–28, 2011.

[15] A. H. M. Ragab, A. Y. Noaman, A. S. Al-Ghamdi, and A. I. Madbouly, "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining," *Proc. 2014 Work. Interact. Des. Educ. Environ. - IDEE '14*, pp. 106–113, 2014, doi: 10.1145/2643604.2643631.

[16] A. Ashari, I. Paryudi, and A. M. Tjoa, "Performance Comparison between Naïve Bayes , Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," vol. 4, no. 11, pp. 33–39, 2013.

[17] D. R. Amancio *et al.*, "A systematic comparison of supervised classifiers," Oct. 2013, Accessed: Oct. 20, 2014. [Online]. Available: http://arxiv.org/abs/1311.0202v1.

[18] Kusrini and L. E. Taufiq, *Algoritma Data Mining*. Yogyakarta: Andi Offset, 2009.

[19] J. Novakovic, "The Impact of Feature Selection on the Accuracy of 1DwYH Bayes Classifier," vol. 2, pp. 1113–1116, 2010.

[20] B. Azhagusundari and A. S. Thanamani, "Feature Selection based on Information Gain," no. 2, pp. 18–21, 2013.

[21] R. G. Gallager and L. Fellow, "Claude E . Shannon : A Retrospective on His Life , Work , and Impact," vol. 47, no. 7, pp. 2681–2695, 2001.