Classification of Covid-19 Survillance Datasets using the Decision Tree Algorithm

Ivandari [1], M. Adib Al Karomi [2]
[1-2] *Computer Science, STMIK Widya Pratama Pekalongan, Indonesia*

Abstract— Covid-19 is a new type of mutated virus that has been discovered and studied throughout the world. For the time being, no effective drug has been found to treat or prevent this disease. One way that governments around the world are doing is limiting physical contact with people with COVID-19. Data mining is a computer science to study data and perform extraction to get new knowledge. One technique in data mining is classification. C45 is one of the best classification algorithms. The result of the c45 algorithm can be a decision tree. Decision trees are used because the results can be well represented. and can be easily understood in human language. This study classified the Covid-19 surveillance dataset using the Decission tree. The Covid-19 surveillance dataset was obtained from a public data portal, namely the UCI machine learning repository. This study resulted in better accuracy than previous studies using the same dataset. The level of accuracy obtained by using the decision tree algorithm is 65%. Although in this study the accuracy value has increased by 10%, the level of accuracy is still relatively low. The low level of accuracy is due to the dataset used only has 7 attributes and 14 records.

Index Terms— Decision tree, covid-19 surveillance, accuracy.

1. Introduction

Coronavirus Diseases or often better known as covid-19 is a disease that is very easily transmitted with a wide spread and is a global issue today. In Indonesia, the first case of Covid-19 was discovered in early 2020. With a very fast spread in the first month, positive confirmed cases of Covid-19 reached 1790 people. Furthermore, at the end of 2020 this case expanded to areas with a total of half a million confirmed positive cases.

The government's policy by limiting crowds and providing vaccinations for public and elderly workers is one of the efforts to suppress the increase in Covid-19 cases. Data mining is a field of science to explore data and perform calculations to gain new knowledge from the data [1]. Classification is the most important part of data mining. Several classification algorithms are used to solve a problem. Decission tree is one of the best classification methods [2]. The representation of the use of decision trees is very easy to understand in human language [3].

Uci Machine Learning Repository is a collection of databases that are often used to test a method or algorithm. The data from the Uci Machine Learning Repository is public data that has been tested in all fields. For the health sector, there are several data, including data on people with diabetes, breast cancer, and cases of Covid-19, which are research trends in recent years. The Covid-19 dataset taken on the uci website is data on Covid-19 sufferers with various symptoms taken from the Ministry of Health of the Republic of Indonesia (https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance).

This study uses a decision tree algorithm to classify the COVID-19 dataset. Calculations are carried out using an auxiliary application, namely rapid miner. The calculation validation process is carried out using 10 folds cross validation. This validation model is the most widely used validation for classification calculations. The calculation of the accuracy of the algorithm is carried out using a confusion matrix. In this process it will be known the number of records for test data that are in accordance with the original results. Furthermore, it is divided by the whole calculation process so that the percentage value of the accuracy of an algorithm is obtained. The result of the representation of the decision tree can be easily understood in human language

2. Literature Review

2.1. Related Research

Previous research has carried out the classification of the Covid-19 surveillance dataset using the KNN algorithm [4]. In this study, the KNN algorithm obtained an accuracy rate of 55%. The dataset used in previous studies is the same as the dataset used in this study.

2.2. Data Mining

Data mining is an extraction process to obtain previously unknown information from data [1]. Data mining can analyze old cases to find patterns from data using pattern recognition techniques such as statistics and mathematics[5]. Data Mining or often also called

Knowledge Discovery in Database (KDD) is a field of science that discusses a lot about the pattern of a data. A series of processes to obtain knowledge or patterns from data sets is called data mining [1]. A big data can be useless and will only be garbage if we can't use it. Data mining answers this problem by analyzing the large data and then creating a certain rule, pattern, or model to recognize new data that is not in the stored data row [6].

Data mining is an activity that includes the collection and use of historical data to find regularities, patterns or relationships in data sets [7]. The output of data mining can be used to improve decision making in the future. Data ming has links with various other fields of science such as Machine Learning, Statistics, Visualization and databases.

2.3. Classification

Classification is one of the main roles of data mining. Classification is included in supervised learning because in the classification process there is a learning process with past data. This process is used by algorithms to recognize patterns from data which can later be applied to new data whose groups are not yet known. Classification techniques are widely applied in the real world as well as in the medical world [8], education [9], building engineering [10], computer networks and are widely used in other fields.

The label in the classification or it can also be called the destination attribute is the attribute that will be searched for in the calculation of the data mining algorithm. For example in the medical world if there is a new patient with symptoms of a certain disease but the type of disease he is suffering from is not yet known. So classification can be a tool for making decisions. The existence of past data or what will be called training data will help a lot in the classification process. Because the amount of training data will affect the accuracy of the classification accuracy of a data mining algorithm. The number of attributes will also affect the performance of an algorithm[6], although too many attributes or commonly known as high-dimensional data will affect the time complexity of the algorithm. The more attributes that are used, the more expensive the computational process will be, or the longer the computation time will be. To overcome this, data attribute reduction can be done or also known as feature extraction and feature selection [7].

In doing a classification, it takes past data which will later be processed into a rule or new knowledge. The classification problem is basically as follows [11]:

1. Classification problems depart from the available training data.
2. Training data will be processed using a classification algorithm.
3. Classification problems end with the generation of knowledge which is represented in the form of diagrams, rules or knowledge.

Classification begins with the initial data that is used as algorithm learning data or also called training data. Of course, the training data in question is data with objective attributes or label attributes. What is meant by label is the final result of the data which will later be calculated using an algorithm. For example, there is student registration data with registration/unregistered labels. This data will be processed by the algorithm to find out patterns, rules or new knowledge from the data. Later this new pattern or knowledge can be used as a tool to predict if there are new records with unknown labels. The accuracy of the algorithm differs depending on the type of data it processes [12].

2.4. C45 Algorithm

C4.5 is the development of the ID3 algorithm [5] which was developed by Quinlan [13]. The C4.5 algorithm is widely used by researchers to perform classification tasks. The output of the C4.5 algorithm is a decision tree or often known as a decissin tree. In several studies the C4.5 algorithm has been the best choice compared to several other classification algorithms [9], [14]

Decision tree itself is a very powerful and well-known classification and prediction method [15]. In this decissin 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 the case of data with discrete-valued outputs [7]. 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.

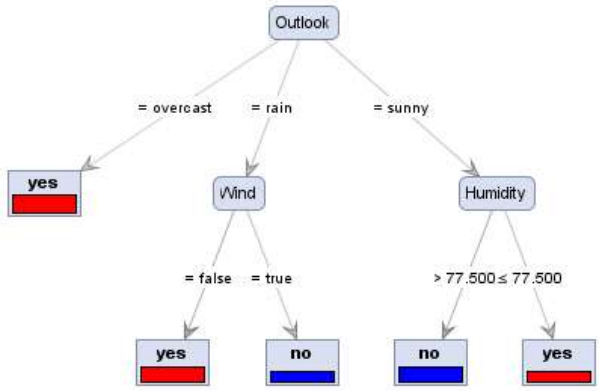

Figure 1. Illustration of Decission tree

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 point that contains a label or destination, namely yes or no. While the internal node in Figure 1 is also symbolized by a

tuple square that is between the root node and the terminal node.

The steps to make a decision tree from the C4.5 algorithm are as follows [13]:

1. Preparing training data, training data is data taken from historical data that has happened before or is called past data and has been grouped into certain classes.
2. Determine the root of the tree. The root of the tree is determined by calculating the highest GainRatio of each attribute. Before calculating the GainRatio, first calculate the Total Entropy before looking for each Entropy class, while the formula for finding Entropy is as below:

$$\boldsymbol{Entropy\ (S) = \sum_{i=1}^{n} -pi * log_2 pi}$$

Information:
S = Set of cases
n = number of partitions S
pi = proportion of Si to S

Where log2pi can be calculated by:

$$\boldsymbol{log(X) = \frac{ln\ (X)}{ln\ (2)}}$$

3. Calculates the GainRatio value as the root of the tree, but previously calculated the Gain and SplitEntropy (SplitInfo), the formula to calculate the Gain is as below:

$$\boldsymbol{Gain(S,A) = Entropy\ (S) - \sum_{i+1}^{n} \frac{|\ S_i\ |}{|\ S\ |} * Entropy\ (S_i)}$$

The formula to calculate SplitEntropy, as below:

$$\boldsymbol{SplitEntropy_A(S) = -\sum_{i+1}^{n} \frac{|S_i|}{|\ S\ |} * log_2 \left(\frac{|S_i|}{|S\ |}\right)}$$

The formula to calculate the GainRatio is below:

$$\boldsymbol{GainRatio\ (A) = \frac{Gain\ (A)}{SplitEntropy\ (A)}}$$

Information:
S = Case Set
A = Attribute
n = number of attribute partitions A
|Si| = number of cases on partition i
|S| = number of cases in S

4. Repeat steps 2 and 3 until all tuples are partitioned
5. The decision tree partitioning process will stop when:
   a. All tuples in node N get the same class
   b. No attributes in partitioned tuples anymore
   c. There are no tuples in the empty branch

2.5. Cross Validation

Cross validation is an act of proving a method or the performance of an algorithm. In the process of testing data mining, the most widely used is cross validation. Cross validation is proof by dividing the data partly as training data and partly as testing data with a certain composition. The most widely used division in data mining classification research is dividing the data randomly into 10 parts. One part is used as testing data and 9 parts are used as training data. This kind of validation is also known as 10fold cross validation [1].

In this study, testing was carried out with 10 folds cross validation, namely by dividing the data randomly into 10 parts, then 9 parts were used as training data and 1 part became testing data. This treatment is repeated up to 10 times repeatedly by replacing the testing data and training data until evenly distributed. Then the results of the accuracy of the 10 times the experiment was taken on average.

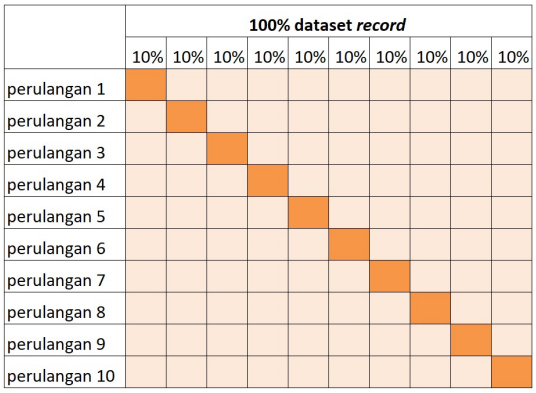

Figure 2. Illustration of 10 folds cross validation

Figure 2 shows an illustration of 10 folds cross validation. Where the data illustrated in purple is divided into 10 parts randomly. Then one part of the data (10%) is taken as testing data (yellow color) and the rest becomes training data. The experiment was repeated up to 10 times until the entire data section got a turn as testing data. The results of all experiments are taken for accuracy and then from the overall accuracy level (a1 to a10) the average is taken to be used as a benchmark for accuracy of 10-fold cross validation.

2.6. Confussion Matrix

The Confusion Matrix is the result of the evaluation of a data mining classification which is embodied in a table [16]. The convussion matrix contains information about the labels of the classification results with the actual labels.

The general form of the confusion matrix can be seen according to Figure 2. The entire data record will be seen with the appropriate classification data and inappropriate classification data. This is what will be used as a measuring tool for the performance of an algorithm. The final result of the calculation is the level of accuracy taken from the classification division that corresponds to the total amount of data.

Classification		*Predicted class*	
		Class: YES	*Class*: NO
Observed class	*Class*YES	***a*** ***True Positive (TP)***	***b*** ***False Negative (FN)***
	*Class*NO	***c*** ***False Positive (FP)***	***d*** ***True Negative (TN)***

Figure 3. Confusion matrix [16]

Figure 3 is a confusion matrix with two labels (positive and negative). The level of accuracy of an algorithm can be calculated by the following equation:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{a+d}}{\boldsymbol{a+b+c+d}} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}$$

Information:
a : positive classification result with positive actual class
b : the result of the classification is negative with the actual class being positive
c : positive classification result with negative actual class
d : the result of the classification is negative with the actual class being negative

3. Research Methods

This study uses an experimental method to obtain the best results for the classification of the Covid-19 survival dataset. The assistive application used is rapid miner. The stages of the research carried out are as follows:

3.1. Data Collection

The data collection process is carried out using public data. In the data published by the uci machine learning repository, one dataset is obtained with 7 regular attributes and 1 destination attribute or label attribute. The dataset with the name Covid-19 surveillance can be viewed and downloaded from the url: https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance.

3. 2. Algorithm Calculation

This stage is the main stage of research. In this stage, it is divided into several parts, namely the validation stage and the evaluation stage.

3. 2.1. 10 fold cross validation

This stage is carried out by using an auxiliary application, namely rapid miner. The validation process is carried out by dividing the dataset into 10 parts, then 1 part is used as test data with 9 other parts as comparison. This process is repeated up to 10 times where the other part is turned into test data in the next iteration. Furthermore, the iteration process ends when all records have been part of the test data once.

3. 3.2. Confusion matrix

confusion matrix is used to determine the label of the testing data in accordance with the training data. Next, the level of accuracy will be obtained from the number of test data record labels that match the labels on the training data. This process uses a rapid miner application.

4. Results and Discussion

4. 1. Data Attribute Analysis

The initial step is to obtain the Covid-19 surveillance database which was previously obtained from the url: https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance. In the dataset there are 7 regular attributes and 1 label attribute or destination attribute, namely categories. All regular attribute names are intentionally coded to maintain dataset objectivity. There are 14 records in this dataset. Table 1 is the dataset used in this study.

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

From table 1 it can be seen that the existing data uses nominal types. This nominal type is all entries from data records that cannot be compared using a scale with other record entries. To handle data with nominal types, various classification methods can be used, including decision tree [17].

4.2. Algorithm Calculation

The calculation process is carried out using an auxiliary application, namely rapid miner. This application is widely used for the process of measuring the accuracy of a data mining method. Figure 4 and Figure 5 are the main display on the rapid miner.

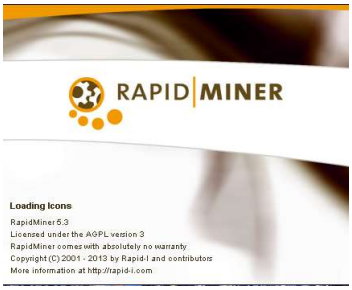

Figure 4. Initial view of rapid miner

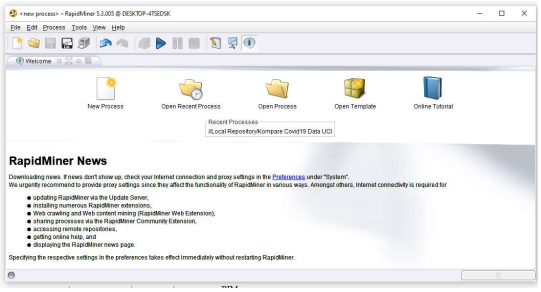

Figure 5. Rapid miner main page

In the validation process, it is carried out by utilizing 10 folds cross validation and using a confusion matrix to evaluate and calculate algorithms using the decision tree method. Figure 6 is a series of processes in a rapid miner application.

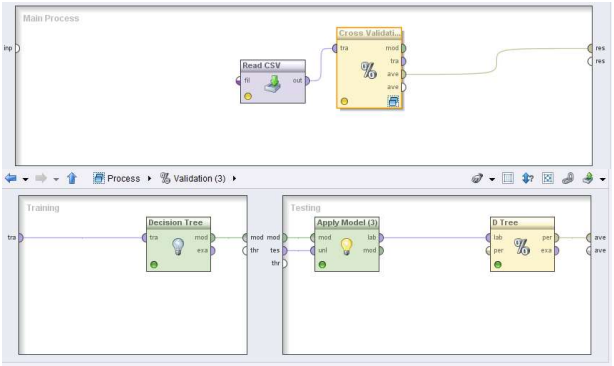

Figure 6. Calculation process

From the series of processes carried out as shown in Figure 6, the results of the confusion matrix are obtained as shown in Figure 7.

accuracy: 65.00% +/- 45.00% (mikro: 71.43%)

	true PUS	true PIM	true PWS	class precision
pred. PUS	7	2	0	77.78%
pred. PIM	1	3	1	60.00%
pred. PWS	0	0	0	0.00%
class recall	87.50%	60.00%	0.00%	

Figure 7. Calculation results from rapid miners

The results of the accuracy of the algorithm calculations using rapid miner show the decision tree accuracy rate is 65% with details as shown in Figure 7.

4.3. Discussion

The level of accuracy obtained from the previous calculation process is relatively low. This is because the Covid-19 surveillance dataset only has 14 records. The decision tree algorithm is very sensitive to data. The number of records and attributes can affect the learning process in the decision tree algorithm, the results of which can also affect the gain value in the decision tree algorithm [18]. Besides being influenced by the number of existing records, the low level of accuracy can also be influenced by the low variance of the label attribute or the destination attribute. In the Covid-19 surveillance dataset, there are 3 variants of labels with details: 8 records (PUS), 5 records (PIM), and 1 record (PWS). This number clearly affects the classification results because one label variant is very dominant compared to the other variants.

5. Conclusion

The results of this study are as follows:

1. The classification of the Covid-19 survival dataset using the decision tree algorithm obtained an accuracy rate of 65% which is classified as low accuracy.
2. The low level of accuracy is caused by the lack of existing records, and there are variants of the dominant label attribute.

References

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

[2] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2007.

[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, "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.

[5] D. T. Larose, *Discovering Knowledge in Data*. John Wiley & Sons, 2005.

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

2012.
[7] B. Santosa, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Edisi Pert. Yogyakarta: Graha Ilmu, 2007.
[8] A. Christobel and D. . Sivaprakasam, "An Empirical Comparison of Data Mining Classification Methods," vol. 3, no. 2, pp. 24–28, 2011.
[9] 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.
[10] 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.
[11] S. Susanto and D. Suryadi, *Pengantar Data Mining: Menggali Pengetahuan dari Bongkahan Data*. Yogyakarta: Andi Offset, 2010.
[12] 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.
[13] J. Han and M. Kamber, *Data Mining: Concepts and Techniques Second Edition*. Elsevier, 2006.
[14] D. Widiastuti, J. S. Informasi, and U. Gunadarma, "ANALISA PERBANDINGAN ALGORITMA SVM , NAIVE BAYES , DAN DECISION TREE DALAM MENGKLASIFIKASIKAN SERANGAN (ATTACKS)," pp. 1–8, 2007.
[15] Kusrini and L. E. Taufiq, *Algoritma Data Mining*. Yogyakarta: Andi Offset, 2009.
[16] F. Gorunescu, *Data Mining: Concept, Models and Techniques*, Vol 12. Berlin: Heidelberg: Springer Berlin Heidelberg, 2011.
[17] M. A. Alkaromi, "Komparasi Algoritma Klasifikasi untuk dataset iris dengan rapid miner," *IC Tech*, vol. XI, no. 2, 2014.
[18] I. Indrayanti, S. Devi, and M. A. Al Karomi, "Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus," *IC-TECH*, vol. XIII, no. 2, pp. 1–6, 2017.