Covid Symptom Severity Using Decision Tree

Naim Rochmawati
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
naimrochmawati@unesa.ac.id

Hanik Badriyah Hidayati
Department of Neurology
Universitas Airlangga
Surabaya, Indonesia
hanikhidayati@yahoo.com

Yuni Yamasari
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
yuniyamasari@unesa.ac.id

Wiyli Yustanti
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
wiyliyustanti@unesa.ac.id

Lusia Rakhmawati
Department of Electrical Engineering
Universitas Negeri Surabaya
Surabaya, Indonesia
Lusiarakhmawati@unesa.ac.id

Hapsari P. A. Tjahyaningtijas
Department of Electrical Engineering
Universitas Negeri Surabaya
Surabaya, Indonesia
hapsaripeni@unesa.ac.id

Yeni Anistyasari
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
yenian@unesa.ac.id

***Abstract*— Corona is a very contagious virus. In a pandemic like this, people often worry whether they are infected or not. When they cough, they often worry whether it is a sign of covid-19 or an ordinary cough. From the clinical symptoms can actually be known whether someone has Covid or not. In this study, a clinical symptom dataset will be used to classify the symptoms using a Decision Tree algorithm. The decision trees used in this research are J48 and Hoeffding Tree. Decision Tree is one of the most popular classification methods because it is easy to interpret by Humans. the prediction model uses a hierarchical structure. The concept is to convert data into decision trees or decision rules. the result of J48 were slightly better than the Hoeffding tree in terms of accuracy, precision, and recall. Meanwhile, from the tree view results, the Hoeffding Tree is simpler and the number of nodes is less than J48.**

Keywords*— decision tree, corona, covid, covid-19, corona, symptoms, prediction, decision rules

I. INTRODUCTION

Covid-19 or the coronavirus is a type of virus that attacks the human respiratory system where the virus is still related to the SARS and MERS viruses which have already infected the world's population. This virus was first discovered in the city of Wuhan, China [1] in December last year. Based on research, this virus is capable of causing death, although many have recovered from this disease. however, the number of people infected by the corona virus is still increasing every day.

The concern about this virus is increasing because it can result in death if infected. concerns arise because the coronavirus is very contagious. The symptoms of Covid-19 are similar to the symptoms of the common cold or flu.

According to the World Health Organization (WHO), COVID-19 or coronavirus causes respiratory disease and is spread through respiratory droplets and close contacts. Droplet transmission occurs when you have close contact (within one meter) with a person who has respiratory symptoms such as coughing or sneezing, which may spread these potentially infectious droplets, typically 5-10 microns in size, to your body [1].

This sometimes raises doubts and a lack of awareness for those with the flu. Having a common cold can be an early symptom of coronavirus infection. Symptoms that come are sometimes similar to other illnesses. The decision tree can answer this problem by producing rules so that it is clear whether a person has infected the coronavirus or not.

A decision tree is a classification method that uses a tree structure, where each node represents an attribute and the branch represents the value of the attribute, while the leaves are used to represent the class. The top node of the decision tree is called the root[2].

A. Decision Tree

The decision tree algorithm used in this study are:

- J48

  J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team [3]. This algorithm was developed from the C4.5 algorithm and implemented with java open source. The way J48 works is by swapping the nodes in the decision tree with leaves. J48 selects the best attribute and is included in the unprunned decision tree[4].

- Hoeffding Tree

  The Hoeffding tree is an incremental decision tree learner for large data streams, that assumes that the data distribution is not changing over time. It grows incrementally a decision tree based on the theoretical guarantees of the Hoeffding bound (or additive Chernoff bound). A node is expanded as soon as there is sufficient statistical evidence that an optimal splitting feature exists, a decision based on the distribution-independent Hoeffding bound. The model learned by the Hoeffding tree is asymptotically nearly identical to the one built by a non-incremental learner, if the number of training instances is large enough [5].

B. Related Work

The Decision tree is usually applied for classification and prediction. Many other researchers have researched on decision trees. Among them is a study comparing several methods about decision trees [6][7]. Research using this decision tree method can be applied in many fields. J48 is used in research related to Power Quality Disturbance [8]. J48 research with several public datasets was also carried out. Heard disease classification for absence or presence was performed using the J48 [9]. Several studies J48 were

implemented also to detect network intrusion[10], hepatitis data modeling[11], and deagnosis of dental xray[12]. Research on Hoeffding has also been conducted by several researchers. The Hoeffding Tree algorithm has been implemented using a streaming dataset [13]. Hoeffding is also used to classify diabetes mellitus[14].

Research on covid related to artificial intelligence is also widely conducted. one of them is about predicting mortality risk in patients with Covid-19[15]. Symptom severity classification research was conducted using gradient tree boosting [16]. Corana virus diagnosis to classify PDP, ODP and OTG has been done [17]

II. Research Method

This research has a sequence of steps in solving the problem. Starting from the preparation of the dataset that must be used then preprocessing the dataset. To implement the classifier, setting the test data must be done first by doing cross-validation so that the results are more accurate. After setting the test data, the classifier can be implemented. The final step is to compare the J48 algorithm with the Hoeffding Tree because these two algorithms will be used to solve the problem. The steps can be explained as in Fig. 1.

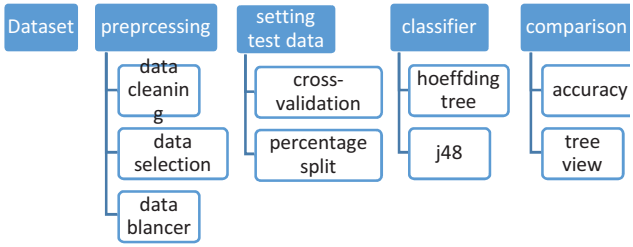

Fig. 1 Research Method Diagram

The following is a more detailed explanation of Fig.1. Each of the steps will be explained in sub-chapters.

A. Dataset

This research was conducted using the Weka application. The dataset used in this study is the public dataset from the Kaggle website [18].

B. Preprocessing

Preprocessing consists of several steps, namely cleaning the data then selecting the attributes to be used. Attributes that are deemed insignificant will be removed. The original attribute consisted of 16 attributes. After being selected and deleted, the attributes used are 13 attributes. Input attributes are fever, tiredness, dry-cough, difficulty-in-breathing, afternoon-throat, none-symptoms, pains, nasal-congestion, runny-nose, diarrhea, none-experiencing, and age. Meanwhile, the target attribute is severity. The severity attribute has 4 values, namely none, mild, moderate, and severe. The amount of data after preprocessing was 31,740.

The values of the attributes in the dataset used in this study are not balanced. Some have a value of almost 90%, while others have no more than 10%. So the values must be equalized first and then another process can be done. The Databalancer filter is applied to balance the values in the dataset.

C. Setting Test Data

In this section, there are two steps:

1. Cross-validation

   Cross-validation is a model validation technique for assessing how statistical analysis results will generalize to independent data sets. This technique is mainly used to make model predictions and estimate how accurate a predictive model will be when implemented. In a prediction problem, a model is usually given a set of data (dataset) that is known to be used in running the training (training dataset), as well as a set of unknown data (or data that is first seen) against the model being tested (test dataset). The purpose of cross-validation is to define a dataset to "test" a model in the training phase (i.e., data validation), to limit problems such as overfitting, providing insight into how the model will generalize independent of the dataset (i.e., unknown datasets, for example from the real problem), etc. [19]. Cross-validation In this study, cross-validation will be carried out 3 times for each algorithm. they are 2 fold, 10 fold, and 20 fold. Fold selection is done randomly to measure the results, whether the results are significant or not.

2. Split Percentage

   The dataset used will be used as training data and testing data. For this reason, it is necessary to share data so that the portion for testing and training can be more certain. In this study, the data was divided into the percentage of data by 66% training data and 34% testing data. This number is usually given the default weka

TABLE I. Comparison between J48 and holding tree

J48							
Fold	Accuracy	Precision	Recall	F-measure	MCC	Roc Area	PRC Area
2	83.60%	0.859	0.836	0.831	0.791	0.939	0.831
10	83.50%	0.858	0.835	0.830	0.790	0.938	0.834
20	83.53%	0.858	0.835	0.830	0.790	0.937	0.832
Hoeffding Tree							
Fold	Accuracy	Precision	Recall	F-measure	MCC	Roc Area	PRC Area
2	82.65%	0.846	0.827	0.822	0.777	0.937	0.825
10	83.07%	0.849	0.831	0.825	0.782	0.937	0.833
20	82.84%	0.847	0.828	0.823	0.779	0.937	0.833

D. Classifier

There are two types of classifiers used in this study, namely J48 and Hoeffding Tree.

E. Comparison

After the classification process is complete, a comparison of the results of the accuracy and the tree view of each classifier will be carried out.

III. RESULT AND DISCUSSION

This research will discuss the accuracy comparison between j48 and Hoeffding tree. Besides, the results from the tree view will be displayed.

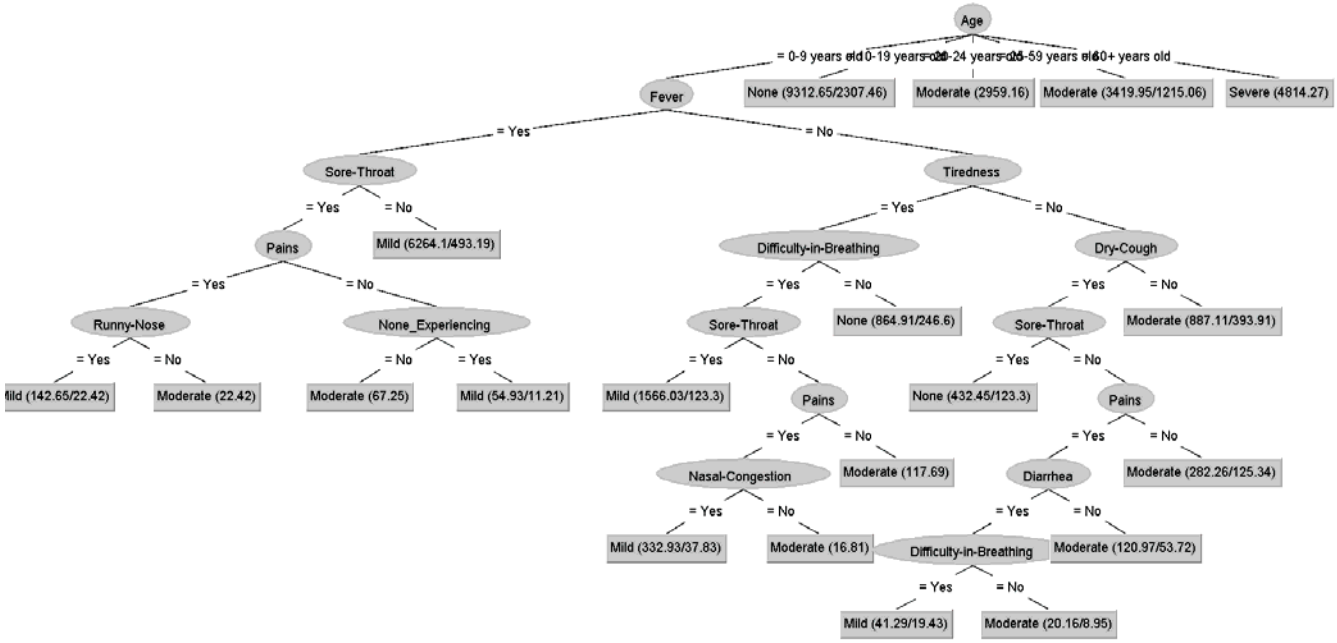

Fig. 2. Tree View of J48 using 10 fold cross-validation

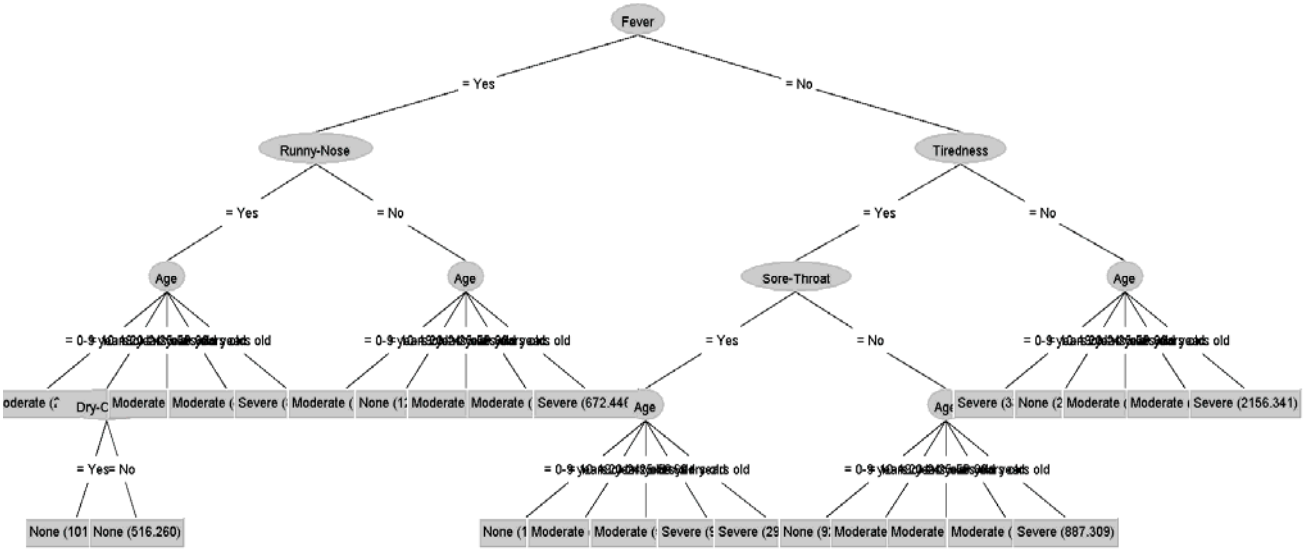

Fig. 3. Tree View of Hoeffding Tree using 10 fold cross-validation

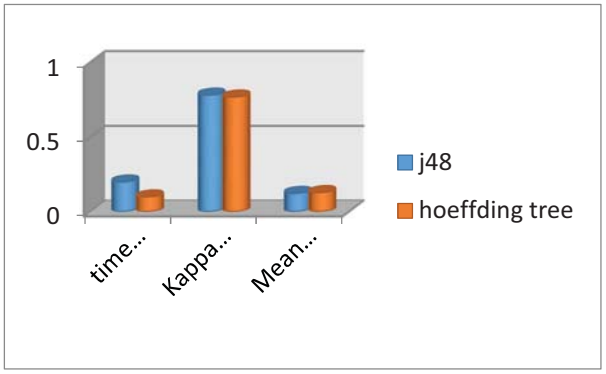

Fig. 4. Comparison of J48 and Hoeffding Tree with cross-validation 2 fold

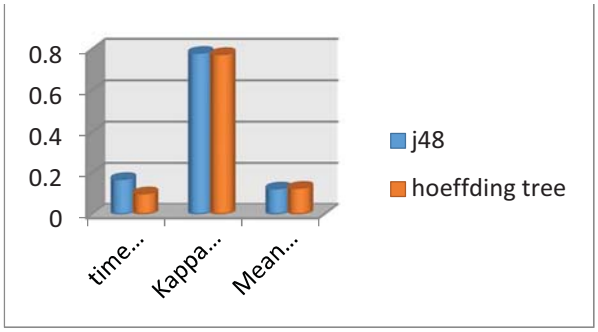

Fig.5. Comparison of J48 and Hoeffding Tree with cross-validation 10 fold

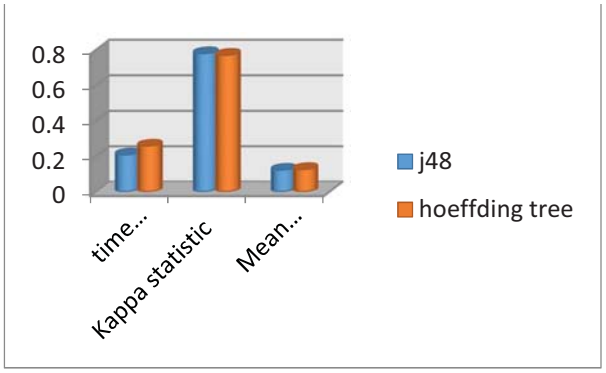

Fig. 6. Comparison of J48 and Hoeffding Tree with cross-validation 20 fold

From Table 1, it can be seen that the difference in the fold in the case in this study is not too significant in the difference in results both when cross-validation is 2 folds, 10 folds, and 20 folds. The average accuracy, precision, and recall results of j48 are slightly higher than the Hoeffding tree. But the difference is not very significant.

Fig. 2 and Fig. 3 show the difference in results where the main node which is the root node of J48 is age while the root node of the Hoeffding Tree is Fever. The resulting tree view of Hoeffding Tree has fewer nodes than that of J48. This means that the tree view Hoeffding is simpler and easier to understand than the tree view generated by J48. Other comparisons that must be considered are the time required to carry out the execution process, the kappa statistic, and the mean absolute error.

Fig. 4, Fig. 5, and Fig. 6 show the comparison of time taken, kappa statistic, and mean absolute error between J48 and the Hoeffding Tree. The blue color is for the J48 algorithm and the orange one is for the Hoeffding Tree algorithm.

Fig. 4 shows the comparison of J48 and Hoeffding Tree when cross-validation is 2 fold. Fig. 5 shows the comparison of J48 and the Hoeffding Tree at the cross-validation of 10 folds. Fig. 6 shows the comparison of J48 and Hoeffding Tree at cross-validation of 20 folds. The three pictures above show that the time taken by Hoeffding Tree in processing data faster than the J48. But when the 20-fold cross-validation, it takes much longer than j48. The average Kappa Statistic J48 is bigger than Hoeffding tree. Likewise with Mean Absolute Error where the Hoeffding tree is bigger than J48.

IV. Conclusion and Future Work

Using j48 and the Hoeffding Tree can produce clear rules whether someone is exposed to mild, moderate, severe, or not COVID. The comparison between the algorithm j48 and the Hoeffding Tree is not very significant. It's just that in this study, the results were slightly better than the Hoeffding tree in terms of accuracy, precision, and recall. Meanwhile, from the tree view results, the Hoeffding Tree is simpler and the number of nodes is less than J48.

In the future, the research work can also be carried out using the same dataset with different preprocessing. Research can also be done with more variations in cross-validation and split percentages of training and testing data.

Acknowledgment

The author would like to gratitude for the contribution provided by the Department of Informatics, the State University of Surabaya to this research.

References

[1] "WHO says coronavirus is spread by respiratory droplets, not through air." https://www.livemint.com/news/world/who-says-coronavirus-is-spread-by-respiratory-droplets-not-through-air-11585908641833.html (accessed Aug. 13, 2020).
[2] L. Rokach and O. Maimon, "DECISION TREES."
[3] UOC, "J48 decision tree - Mining at UOC." .
[4] P. Kapoor, R. Rani, and R. JMIT, "Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning," *Int. J. Eng. Res. Gen. Sci*, vol. 3, no. 3, pp. 1613–1621, 2015.
[5] "Hoeffding Decision Trees – streamDM." .
[6] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, pp. 130–135, 2015, doi: 10.11919/j.issn.1002-0829.215044.
[7] V. Kapoor, M. Madan, and M. Dave, "The Analytical Comparison of ID3 and C4.5 using WEKA," *Int. J. Comput. Appl.*, vol. 167, no. 11, pp. 1–4, 2017, doi: 10.5120/ijca2017914286.
[8] F. A. S. Borges, R. A. S. Fernandes, A. M. Lucas, and I. N. Silva, "Comparison Between Random Forest Algorithm and J48 Decision Trees Applied to the Classification of Power Quality Disturbances," *Proc. ...*, pp. 146–148, 2016, [Online]. Available: http://search.proquest.com/openview/c79b355e6e441a51c86016d928e67985/1?pq-origsite=gscholar&cbl=1976357.
[9] L. Devasena and I. S. B. Hyderabad, "Proficiency Comparison of Random Forest and J48 Classifiers for Heart Disease Prediction," *Int. J. Comput. Acad. Res.*, vol. 5, no. 1, pp. 46–55, 2016.

[10] S. Sahu and B. M. Mehtre, "Network intrusion detection system using J48 Decision Tree," *2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015*, pp. 2023–2026, 2015, doi: 10.1109/ICACCI.2015.7275914.
[11] M. Ramasamy, S. Selvaraj, and M. Mayilvaganan, "An empirical analysis of decision tree algorithms: Modeling hepatitis data," *ICETECH 2015 - 2015 IEEE Int. Conf. Eng. Technol.*, no. March, pp. 18–21, 2015, doi: 10.1109/ICETECH.2015.7275013.
[12] J. Nuansanong, S. Kiattisin, and A. Leelasantitham, "Diagnosis and interpretation of dental X-ray in case of deciduous tooth extraction decision in children using active contour model and J48 tree," *2014 Int. Electr. Eng. Congr. iEECON 2014*, no. type 2, pp. 48–51, 2014, doi: 10.1109/iEECON.2014.6925902.
[13] A. Muallem, S. Shetty, J. W. Pan, J. Zhao, and B. Biswal, "Hoeffding Tree Algorithms for Anomaly Detection in Streaming Datasets: A Survey," *J. Inf. Secur.*, vol. 08, no. 04, pp. 339–361, 2017, doi: 10.4236/jis.2017.84022.
[14] S. S. Subha, "STREAMING CLASSIFICATION HOEFFDING TREE OF DIABETES," vol. 118, no. 18, pp. 1857–1865, 2018.
[15] M. Pourhomayoun and M. Shakibi, "Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making," vol. 19, no. February, 2020, doi: 10.1101/2020.03.30.20047308.
[16] Y. Liu *et al.*, "Symptom severity classification with gradient tree boosting," *J. Biomed. Inform.*, vol. 75, pp. S105–S111, 2017, doi: 10.1016/j.jbi.2017.05.015.
[17] W. Wiguna and D. Riana, "Diagnosis of Coronavirus Disease 2019 (Covid-19) Surveillance Using C4.5 Algorithm," *J. Pilar Nusa Mandiri*, vol. 16, no. 1, pp. 71–80, 2020, doi: 10.33480/pilar.v16i1.1293.
[18] "COVID-19 Symptoms Checker | Kaggle." https://www.kaggle.com/iamhungundji/covid19-symptoms-checker (accessed Aug. 13, 2020).
[19] "Validasi- Silang (statistik) - Wikipedia bahasa Indonesia, ensiklopedia bebas." https://id.wikipedia.org/wiki/Validasi-_Silang_(statistik) (accessed Aug. 14, 2020).