Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since year 2017 to 2021 according to the decree No. 10/E/KPT/2019.

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

RESTI JOURNAL

(System Engineering and Information Technology)

Vol. 5 No. 1 (2021) 10 - 16 ISSN Electronic Media: 2580-0760

Classification of Malaria Complication Using CART (Classification and Regression Tree) and Naïve Bayes

Rachmadania Irmanita[1], Sri Suryani Prasetiyowati[2], Yuliant Sibaroni[3]
[1,2,3]School of Computing, Telkom University
[1]rachmadaniairmanita@student.telkomuniversity.ac.id, [2]srisuryani@telkomuniversity.ac.id, [3]yuliant@telkomuniversity.ac.id

Abstract

Malaria is a disease caused by the Plasmodium parasite that transmitted by female Anopheles mosquitoes. Malaria can become a dangerous disease if late have the medical treatment. The late medical treatment happened because of misdiagnosis and lack of medical staff, especially in the countryside. This problem can cause severe malaria that has complications. This study creates a system prediction to classify the severe malaria disease using Classification and Regression Tree (CART) method and the probability of malaria complication using Naïve Bayes method. The first step of this study is classifying the patients that have symptom are infected severe malaria or not based on the model that has been built. The next step, if the patient classified severe malaria then the data predicted if there any probability of complication by the malaria. There are 8 possibilities of complication malaria which are convulsion, hypoglycemia, hyperpyrexia, and the combinations of these four. The first step will evaluate by using F-score, precision and recall while the second step will evaluate by using accuracy. The highest result F-score, precision and recall are 0.551, 0.471 and 0.717. The highest accuracy 81.2% which predicted the complication is Hypoglycemia.

Keywords: Malaria, Classification and Regression Tree, Naïve Bayes.

1. Introduction

Indonesia is one of the tropical countries that have to deal with malaria. Malaria is a disease caused by the *Plasmodium* parasite and transmitter by female *Anopheles* mosquitoes. The disease is quite difficult to deal with due to the many types of disease have similar symptoms with malaria. It caused the ordinary people misdiagnose this disease and cause malaria suffers to get medical late treatment [1]. The delay of treatment causes severe malaria which has complications. Complications that can occur due to severe malaria are hyperpyrexia, convulsion and hypoglycemia [2].

Based on data and information of Indonesia's health profile in 2018 from the Indonesian Ministry of Health, there are reduction of malaria sufferers in several region in Indonesia. But in some endemic region such as Maluku and Papua, malaria cases that occur every year is relatively high. Based on the calculation of the Annual Parasite Incidence (API) of Malaria per 1000 population in 2017 and 2018 recorded the malaria sufferers in Maluku and Papua are still above 5‰ which means it is relative high malaria sufferers in these regions [3].

Nowadays, there are various studies that have been conducted to detect malaria using machine learning. The algorithm that were used to predicted the disease are decision tree, naïve bayes, neural network, K-NN (K-Nearest Neighbor), SVM (Support Vector Machine) and many more. This study builds a system to classify severe malaria and the probability of the complication.

In the previous study named malaria prediction using Bayesian and other machine learning technique by Hamisu Ismail Ahmad from African University of Science and Technology, were detecting the malaria using four classification algorithms [4]. The algorithms are naïve bayes, decision tree (J48), OneR and ZeroR. The accuracy with decision tree is 88.4% and 79.9% using Naïve Bayes. But, this study only predicts into positive malaria or negative malaria while there is possibility of malaria complication.

A study about predicting malaria utilized Artificial Neural Network (ANN) has approximately 85% results by using Back Propagation rules [5]. The dataset is use the symptom but it just classifies into affected and not affected. The other study using Bayesian Network has accuracy 81% to predict malaria disease [6]. This study uses the symptom of malaria and the environment of patient for the dataset. These two studies also not predict the complication. The other similar study is about expert

Accepted by editor: 20-12-2020 | Final Revision: 24-01-2021 | Online Publication: 13-02-2021

system to predict a fever-based disease by using K-Nearest Neighbor-Certainty Factor. The accuracy is 84.7%. But this study not only focusing in malaria because it is also predicted typhoid and dengue fever [7].

The study predicting malaria utilized convolutional neural network [8] from Gudlavalleru Engineering College in India has accuracy above 92%. But the dataset that used in this study other than the symptoms were used the image of malaria blood smear. Moreover this study only classifies the image into infected cell images or uninfected cell images. The other study predicting with image of blood smear is used support vector machine (SVM) and KNN has accuracy 99.23% [9].

The other similar study was predicted the malaria outbreak by using machine learning [10]. The algorithms were used are SVM and ANN with accuracy 92%. The other study about predict the malaria outbreak using Fuzzy Association Rule Mining [11] in South Korea. This study predicts that an outbreak which is classes as high 0.694 using F3 score validation. But the dataset for these two studies is the environment conditions instead of the symptoms of malaria disease.

Not only to predict the malaria but also there is a study about the prediction of artemisinin resistance in malaria using ensemble machine learning [12]. The accuracy is 0.6084 by using F1 score. The dataset is not the symptom of malaria but the malaria parasite genes. Besides predicting malaria, machine learning also used to predict the other disease such as DFH. In this research, predicting DFH disease spreading pattern were using inverse distance weight (IDW), ordinary and and universal kriging methods [13]. But this study just predicts the number of patients per year and predicts the DFH disease spreading pattern.

Based on the studies that have been mentioned above, in this study the dataset not only classify into positive malaria or negative malaria but also predict the malaria complication. Chapter 1 contains introduction about malaria, background of this study and the previous studies related. Chapter 2 discusses about the research method which include about the data and how it processed. Chapter 3 presented the result of the evaluation measure for each method. Chapter 4 discussed the conclusion.

2. Research Method

The algorithm that used are CART (Classification and Regression Tree) and Naïve Bayes. CART algorithm used to classify the dataset into “Yes” or “No”. “Yes” means positive severe malaria meanwhile “No” means negative severe malaria. If the data classified “Yes”, then the data continue to use to predict the possibilities of malaria complication.

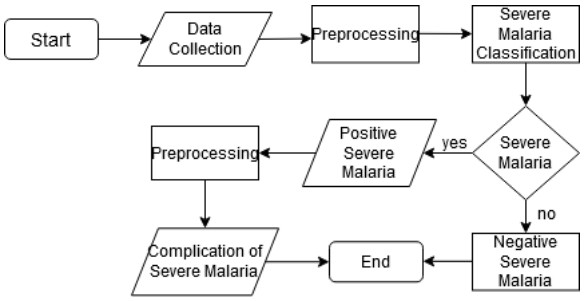

Figure 1.System Architecture

2.1. Data set

The dataset used consisting of 337 of malaria patients. The dataset used in this study are data of malaria patient in Nigeria [14]. The data is in csv form. This dataset consist of 17 attributes represented in Table 1. The class in this dataset was severe malaria. The age in this dataset is between 3 and 77 years old. The gender in this dataset encoded into form “1” for female and “0” for male. The rest attribute was encoded into form “1” for yes and “0” for no.

Table 1.Dataset of Malaria Patients

Attribute & Label	Notation	Values
age	X1	3-77 years old
gender	X2	Male: 0, Female: 1
fever	X3	Yes: 1, No: 0
cold	X4	Yes: 1, No: 0
rigor	X5	Yes: 1, No: 0
fatigue	X6	Yes: 1, No: 0
headache	X7	Yes: 1, No: 0
bitter-tongue	X8	Yes: 1, No: 0
vomiting	X9	Yes: 1, No: 0
diarrhea	X10	Yes: 1, No: 0
convulsion	X11	Yes: 1, No: 0
anemia	X12	Yes: 1, No: 0
jaundice	X13	Yes: 1, No: 0
cocacola-urine	X14	Yes: 1, No: 0
hypoglycemia	X15	Yes: 1, No: 0
prostration	X16	Yes: 1, No: 0
hyperpyrexia	X17	Yes: 1, No: 0
severe malaria	X18	Yes: 1, No: 0

2.2. Data Modeling

There are 2 targets needed in this study. The first target is the classification of the severe malaria. The second target is the predicting the complication of malaria. There are 8 possibilities of the complications represented in Table 2. These possibilities of complication were created as new class name complications.

Table 2.Possibilities of Complication in the Second Target

Possibilities	Complication
C1	Hyperpyrexia
C2	Hypoglycemia
C3	Convulsion
C4	Hypoglycemia, Hyperpyrexia
C5	Convulsion, Hyperpyrexia
C6	Convulsion, Hypoglycemia
C7	Convulsion, Hypoglycemia, Hyperpyrexia
0	No complication

DOI: https://doi.org/10.29207/resti.v5i1.2770

The attributes for the possibilities complications were not used for attribute to classify severe malaria. There are 2 datasets. The first dataset for classification of severe malaria were divided into 70% data train and 30% data test. The second dataset is from the result of first dataset. It is use for prediction the possibilities of malaria complication and divided into 85% data train and 15% data test. The second dataset were divided into different number because the lack number of dataset. The algorithm need use many data as possible for the data train to make the accuracy better.

Moreover, there are 2 types of dataset in this study named symptom dataset and original dataset. The symptom dataset consist of the symptom attribute where the attribute age and gender is not used. The symptom that used as the attribute is fever, cold, headache, diarrhea, cocacola-urine, and prostration. These attributes chosen based on the correlation. The symptoms which have highest correlation with the severe malaria and symptoms that have lowest correlation with the complications are chosen. The lowest correlation were choose because it is the most suitable for predicting using naïve bayes which need attributes that non-dependent [15]. The symptom dataset used to compare with the original dataset.

The correlation of each attribute presented as Figure 3. The measure of correlation was represented in Figure 2. The darkest color is the highest correlation, while the lightest color is the lowest correlation. The original dataset consist of the entire attributes. This dataset used for comparison between the symptom dataset and the original dataset.

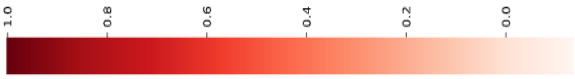

Figure 2. The Measure of Correlation

The dataset was clear from missing values and encoded the categorical attribute into numerical form. But the attribute age has number between 3 and 77 meanwhile the other data only 0 or 1. This attribute need to be transformed into the other form in order to make the data suitable for the algorithms. There are several ways to transform the data using normalization, including min-max normalization.

The min-max normalization scales the numerical values into range between minimum and maximum values of an attribute with the following formula [16].

$$newValue = \frac{originalValue - oldMin}{oldMax - oldMin} \quad (1)$$

Where the $oldMin$ and $oldMax$ were the original minimum and maximum values of an attribute. The $newValue$ is the transformation of $originalValue$.

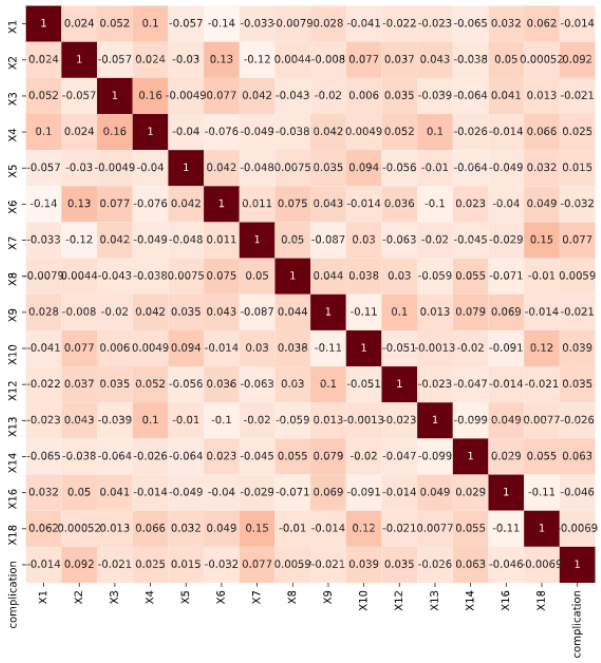

Figure 3. Attribute's Correlation

2.3. Classification and Regression Tree (CART)

Classification and Regression Tree (CART) algorithm was developed by Leo Breimn, Charles Stone, Jerome Friedman and Richard Oslen in 1984. CART is one of decision tree algorithm that combines classification tree and regression tree [17].

CART has several advantages. The first advantage, this algorithm can control outlier because the outliers will be separated during the splitting algorithm. The next advantage, CART will determine the variables that are important and eliminate those that not important. The other advantage mention that CART computing is fast [18]. This algorithm has decision tree basic's algorithm.

CART algorithm start with create a node *N* that represent the input of dataset *D*. If all of the tuple in *D* has the same class then *N* becomes the leaf and labeled with the class of tuple. To create the branch, this algorithm uses a function to determine the splitting criterion. After the best splitting criterion found, then node *N* labeled with the splitting criterion. This process iterative until there is no branch could be created [19].

This algorithm uses *gini index* to find the splitting criterion. The *gini index* measures the impurity of the set of training tuple with the following formula [19]:

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (2)$$

Where P_i is the probability of tuple in D belongs to which class then the sum is computed over m class. After get the probability of each partition from each

DOI: https://doi.org/10.29207/resti.v5i1.2770

attribute then calculate the *gini Index* of each attribute with the following formula [19].

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (3)$$

Where $Gini_A(D)$ is the *gini index* of an attribute while the D_1 and D_2 is the *gini index* of each partition of each attribute.

2.4. CART Model Selection

The model in CART based on the first dataset input. The first dataset input as the dependent variable are consist of 6 attributes which are fever, cold, headache, diarrhea, cocacola-urine, and prostration. The class or independent variable is severe malaria which has label "1" for "Yes" and 0 for "No". The maximum depth of the tree is 4 with *gini index* as the splitting criterion. The maximum depths were 4 to make the tree less complex.

2.5. Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm that utilizes Bayes' Theorem. Bayes' theorem named after Thomas Bayes [20]. This theorem calculated the probability of hypothesis, denoted by *H*, by the evidence which denoted by *X* with the following fomula [19].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (4)$$

Where $P(H|X)$ is the posterior probability, which means of H conditioned on X. $P(X|H)$ is the posterior probability, which means of X conditioned on H. $P(H)$ is the prior probability of of H. $P(X)$ is the prior probability of X.

This algorithm has several advantages. The first advantages are having endurance to face the missing value. It is because Naïve Bayes use the entire attributes to do prediction which means if one of the attribute values is missing the information from others attribute still used. The next advantages, Naïve Bayes is efficient computationally. The other advantage, this algorithm has low variance because there is no search function [21]. Naive bayes works as follow [19].

First let a dataset D that consist of the tuples that represented by *n*-dimensional attribute vector, $X = (x_1, x_2, \dots, x_n)$. Then suppose there are m label. With given a tuple *X*, Naïve Bayes will that *X* belong to which label that have the highest posterior probability conditioned on X utilize the Bayes' theorem on equation (4). The predicted label is the class which have the maximum of X conditioned on class multiplies the probability of class.

2.6. Naïve Bayes Model Selection

The model in Naïve Bayes based on second dataset input. The second dataset input consist of 9 attribute which are age, gender (sex), fever, cold, headache, diarrhea, cocacola-urine, prostration and prediction of severe malaria. The class is complication with label as Table 2.

2.7. Model Testing

The model that were built need to be evaluated. The evaluation need to measure how accurate the algorithm predicts the label. There are several measures such as accuracy, recall, F-score, sensitivity, specificity and precision [19].

This study utilizes accuracy, recall and F-score for classification using CART and precision for predicting using Naïve Bayes as the evaluation measures. The evaluation measures are different in each algorithm because in CART the dataset is imbalance while in Naïve Bayes the target class is multiclass. The accuracy can calculate by following formula [22].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

Accuracy can be summarized into a table called Confusion Matrix. Confusion matrix is a table size *m* by *m* where *m* is the number of given class. The table is represented as Table 3 following [22].

Table 3.Confusion Matrix

	Actual Class (+)	Actual Class (-)
Predicted Class (+)	TP	FP
Predicted Class (-)	FN	TN

Precision is the percentage of tuples labeled as positive is actually positive. Recall is the percentage of positive label is labeled as positive. F-score is the harmonic mean of recall and precision. The precision, recall and F-score (F) can calculate by following formula [23].

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F = \frac{2 \times precision \times recall}{precision + recall} \quad (8)$$

TP is true positive which mean the positive tuples labeled as positive correctly. TN is true negative which mean the negative tuples labeled as negative correctly. FP is false positive which mean the negative tuple mislabeled as positive. FN is false negative which mean the positive tuple mislabeled as negative [19].

3. Result and Discussion

3.1. Classification of Severe Malaria using CART

The CART algorithm in this study use 4 as the maximum depth. The maximum depth has chosen 4 to make the tree smaller and less complex. So, it makes the tree easier to understand. The dataset that used is the original

DOI: https://doi.org/10.29207/resti.v5i1.2770

which include attribute age and gender (sex). The result of this classification as confusion matrix as Table 4.

Table 4.Confusion Matrix of Severe Malaria Classification

Product	Actual Class (+)	Actual Class (-)
Predicted Class (+)	18 (17.65%)	18 (17.65%)
Predicted Class (-)	14 (17.73%)	52 (50.98%)

17.65% data were labeled as positive correctly while 50.98% data labeled as negative correctly. The rest data 17.65% data mislabeled as positive and 17.73% data mislabeled as negative. Based on the confusion matrix, the results of evaluation measures as Table 5 following.

Table 5.Results of Evaluation Measure for Severe Malaria Classification

Evaluation Measure	Result
Precision	0.4
Recall	0.717
F-Score	0.529

The tree of this algorithm is generated by using the library from *sklearn* library for CART algorithm in python as Figure 4 and Figure 5. The attribute that had been chosen as root is headache. Headache attribute has the best *gini index* to splitting. If the next attribute is lower than 0.015, the attribute become a new child node in the left. If the next attribute higher than 0.015, it become a new child node in the right that represented in Figure 5.

The attribute for the left nodes that have been chosen is diarrhea. Then it will generate the algorithm to choose which attribute that has the best *gini index* to splitting. It is repeated until get the classification. Samples mean the total data that use to prediction. Value means the number of data that predict to classify into positive severe malaria or negative severe malaria. It also applies for the child in right nodes.

3.2. Prediction the Complication of Severe Malaria using Naïve Bayes

The data from CART algorithm is reused for prediction the complication of malaria using Naïve Bayes. The dataset were adding the classification result that predicted positive severe malaria. It makes the number for Naïve Bayes dataset is lower than dataset for CART algorithm. The number of dataset in this study is only 18 based on the result of classification after using CART. The data test only predicted 4 possibilities of complication from Table 2 as Table 6.

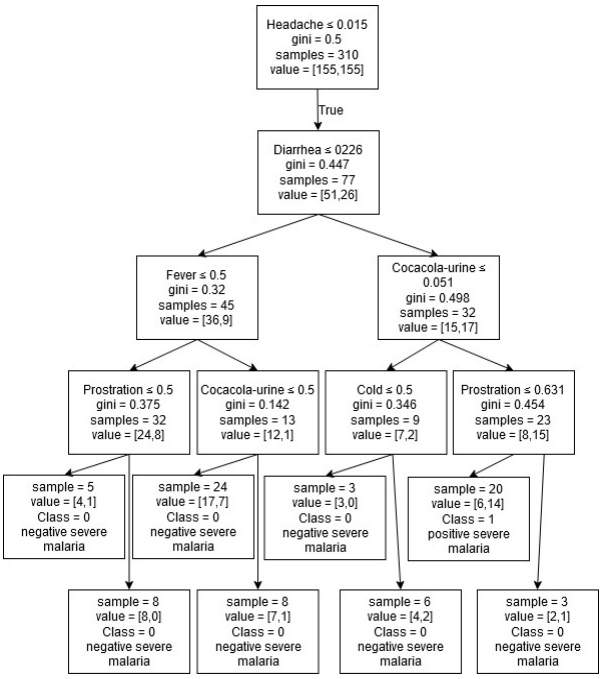

Figure 4. Left Nodes of CART Tree for Classification the Severe Malaria

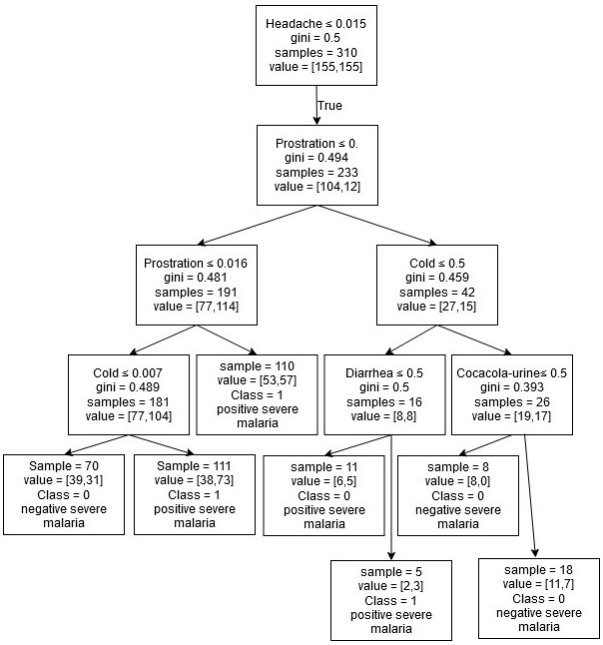

Figure 5. Right Nodes of CART Tree for Classification the Severe Malaria

Table 6.Possibilities of Complication Predicted

Possibilities	Complication	Prediction
C1	Hyperpyrexia	0%
C2	Hypoglycemia	81.2%
C6	Convulsion, Hypoglycemia	0%
0	No complication	0%

DOI: https://doi.org/10.29207/resti.v5i1.2770

The possibilities of complication were labeled correctly only the C2. The rest were mislabeled into the others possibilities. The accuracy of this algorithm is 81.25%.

3.3. Classification and Prediction the Symptom dataset

This section is the result of experiment for classification and prediction using the symptom dataset. The attribute in this dataset only the symptoms of malaria which mean attribute age and gender (sex) is not included. This experiment has been tried 3 times which has 5 iterations for each test. The result as Table 7, Figure 6, Figure 7, and Figure 8 following.

Table 7.The Best Accuracy of Each Attempts

Attempts	Accuracy	
	Symptom Dataset	Original Dataset
1	81.2%	50%
2	75%	50%
3	81.2%	56.2%

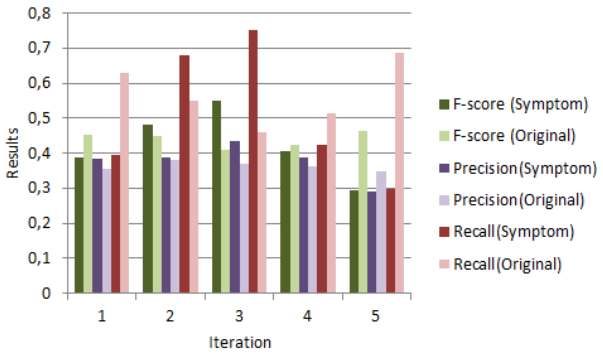

Figure 6. First Attempt of Experiment

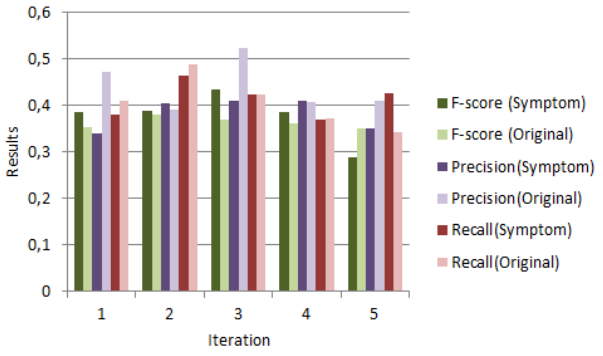

Figure 7. Second Attempt of Experiment

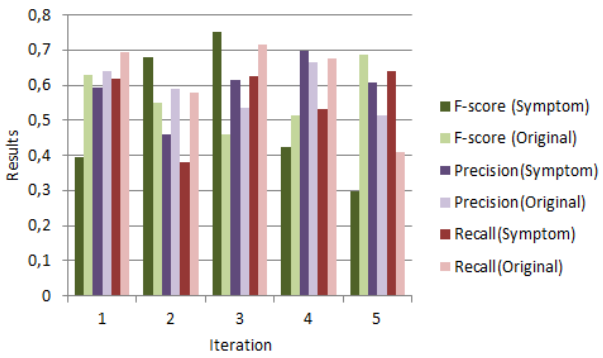

Figure 8. Third Attempt of Experiment

The comparison result between symptom dataset and original dataset, the F-Score is tends to be higher with the original dataset than the F-Score with the symptom dataset. But the percentage of accuracy is not dependent by the F-score of classification.

Based on the accuracy from Table 7 it can be compared the accuracy using CART and continued Naïve Bayes is better with accuracy 81.2% than using OneR with 79.8% or using ZeroR that have accuracy 60.9% [4]. The number of dataset that used by the OneR and ZeroR study is 699 instances with 5 attributes such as FileNo, fever, headache, nausea, and vomiting. This dataset has one class that labeled into Yes and No. OneR is a classification that generates the one rule of one predictor in the data and selects the total smallest error as the rule. Meanwhile ZeroR is a classification which relies on the target and ignores all predictors. The study will compare the result of classification using OneR and ZeroR.

But if the accuracy in Table 7 compared with ANN it show that ANN have better accuracy with approximately 85% [5]. The number of dataset that used in ANN study is 376 patients with symptoms such as fever, shivering, vomiting, fatigue, and headache with classify into affected or not affected malaria. This dataset was divided into 256 data for data train and 120 data for data test. The Neural Network consists of units (neuron) in arranged layer which convert an input into output. The study utilizes the symptom as input to do the classification using ANN.

3.4. Discussion

Based on the result, the percentage of CART model that positive malaria was classified as positive malaria (precision) is 40%. The 71.7% percentage of tuples that classify as positive malaria was the actual is positive malaria (recall). By using the Naïve Bayes model, the possibilities of malaria complication that has been predicted only 4 possibilities from the entire possibilities in Table 2 and the result only predicted 1 possibility correctly, Hypoglycemia [2]. It is because the dataset that utilize in this study is small to divide into two targets. From 120 tuples from the first dataset of data test only the tuples that predicted positive severe malaria that were utilized.

The evaluation measure, accuracy, has significant different result between symptom dataset and original dataset. From table 7 can be seen the accuracy is tends to be higher with the symptom dataset than the accuracy with the original dataset. The most significant result is 81.2% for symptom dataset and 50% for original dataset. It is because the symptom dataset only use the symptom of severe malaria while original dataset use the entire attribute include age and sex.

Based on Figure 3 the attribute sex has low correlation, 0.0005 from range 0-1, to severe malaria which cause the lack classification of malaria. This lack classification

DOI: https://doi.org/10.29207/resti.v5i1.2770

causes the lack prediction of complication of severe malaria. Attribute sex has high correlation, 0.092 from range 0-1, to predict complication. This high correlation makes the accuracy of prediction become low. It is because the algorithm for prediction is naïve bayes that requires the non-dependent attribute [15].

4. Conclusion

Based on the result, it can be concluded the attribute that have been chosen affected the accuracy of the complication prediction. The symptom dataset have higher accuracy, than the original dataset. The F-score is tends to be higher using the original dataset. It is because based on Figure 3 each attribute has correlation around 0.1 or -0.1. Even though the correlation is low it affected the F-score and accuracy. The dependencies of each attribute affected the evaluation measures.

The symptom dataset have bigger accuracy because it has more correlation for each attribute. The original dataset have less accuracy because the attribute sex has lack correlation to classify the severe malaria. This is affected the accuracy prediction of probabilities that can occurred.

The amount of tuples of dataset also affected the number of possibilities that can be predicted. As mentioned in the result section, the second dataset only use the tuples that have been predicted positive severe malaria. The amount is different depends by the result of data test from first data set. However, the amount is not enough to make the algorithm choose all the possibilities to be predicted. So, only some of the possibilities can be predicted.

Overall this research can be used as a reference to avoid misdiagnose of severe malaria. Moreover, this it give the prediction of the complication if the patient have severe malaria with accuracy 81.2%. So it can help the patient to get the treatment as soon as possible. For the future work, it can be tested by using more dataset than 300 datasets for further experiment and using the other models also the other methods as the comparison to conduct the classification and prediction to get the better accuracy.

References

[1] Biantong, T.R., Furqon, M.T. and Soebroto, A.A., 2018. Implementasi Metode Support Vector Machine Untuk Klasifikasi Jenis Penyakit Malaria. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, *2548*, p.964X.

[2] World Health Organization, 2015. *Guidelines for the treatment of malaria*. World Health Organization.

[3] Kurniawan, R. ed., 2019. *Profil kesehatan Indonesia tahun 2018*. Kementerian Kesehatan RI.

[4] Hamisu, I.A., 2019. *Malaria Prediction using Bayesian and other Machine Learning Techniques* (Doctoral dissertation).

[5] Parveen, R., Jalbani, A.H., Shaikh, M., Memon, K.H., Siraj, S., Nabi, M. and Lakho, S., 2017. Prediction of Malaria using Artificial Neural Network. *IJCSNS*, *17*(12), p.79.

[6] Parveen, R., Song, W., Qiu, B., Bhatti, M.N., Hassan, T. and Liu, Z., 2021. Probabilistic Model-Based Malaria Disease Recognition System. Complexity, 2021.

[7] Shofia, E.N., Putri, R.R.M. and Arwan, A., 2017. Sistem Pakar Diagnosis Penyakit Demam: DBD, Malaria dan Tifoid Menggunakan Metode K-Nearest Neighbor–Certainty Factor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, e-ISSN*.

[8] Sujith, C., Ramya, G., Sree, B.L.J., Teja, G. and Nagamani, T., MALARIA DETECTION USING CONVOLUTIONAL NEURAL NETWORK.

[9] Fuhad, K.M., Tuba, J.F., Sarker, M., Ali, R., Momen, S., Mohammed, N. and Rahman, T., 2020. Deep Learning Based Automatic Malaria Parasite Detection from Blood Smear and Its Smartphone Based Application. Diagnostics, 10(5), p.329.

[10] Sharma, V., Kumar, A., Lakshmi Panat, D. and Karajkhede, G., 2015. Malaria outbreak prediction model using machine learning. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *4*(12).

[11] Buczak, A.L., Baugher, B., Guven, E., Ramac-Thomas, L.C., Elbert, Y., Babin, S.M. and Lewis, S.H., 2015. Fuzzy association rule mining and classification for the prediction of malaria in South Korea. *BMC medical informatics and decision making*, *15*(1), p.47.

[12] Ford, C.T. and Janies, D., 2020. Ensemble machine learning modeling for the prediction of artemisinin resistance in malaria. *F1000Research*, *9*(62), p.62.

[13] Prasetiyowati, S.S. and Sibaroni, Y., 2018. Prediction of DHF disease spreading patterns using inverse distances weighted (IDW), ordinary and universal kriging. *JPhCS*, *971*(1), p.012010.

[14] Adeboye, N.O., Abimbola, O.V. and Folorunso, S.O., 2020. Malaria patients in Nigeria: Data exploration approach. *Data in brief*, *28*, p.104997.

[15] Chakrabarti, S., Cox, E., Frank, E., Güting, R.H., Han, J., Jiang, X., Kamber, M., Lightstone, S.S., Nadeau, T.P., Neapolitan, R.E. and Pyle, D., 2008. *Data mining: know it all*. Morgan Kaufmann.

[16] Roiger, R.J., 2017. *Data mining: a tutorial-based primer*. CRC press.

[17] Ma, X., 2018. *Using classification and regression trees: A practical primer*. IAP.

[18] Timofeev, R., 2004. Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin*, pp.1-40.

[19] Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

[20] EMC Education Services, 2015. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley.

[21] Sammut, C. and Webb, G.I. eds., 2011. *Encyclopedia of machine learning*. Springer Science & Business Media.

[22] Torres-Moreno, J.M. ed., 2014. *Automatic text summarization*. John Wiley & Sons.

[23] Brownlee, J., 2020. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery

DOI: https://doi.org/10.29207/resti.v5i1.2770