Classification of the Number of Malaria Cases in Asahan Regency Using Random Forest Application

Naza Amarianda[1]*, Eva Darnila[2]*, Lidya Rosnita[3]*,
* Departement of Informatics, Universitas Malikussaleh, Aceh, Indonesia
naza.210170206@mhs.unimal.ac.id [1] , eva.darnila@unimal.ac.id [2] , lidyarosnita@unimal.ac.id [3]

Article Info

Article history:

Received 2025-06-25
Revised 2025-07-08
Accepted 2025-07-10

Keyword:

Classification,
Malaria,
Random Forest,
Accuracy.

ABSTRACT

This study aims to classify the number of malaria cases in Asahan Regency using the Random Forest method. This method was chosen because it is able to handle data with many and complex variables and reduce the risk of overfitting. Data were collected from the Asahan Regency Health Office. The research stages include data collection, preprocessing, model training, and model evaluation. The dataset used consists of 568 malaria case data from 25 sub-districts. The data is divided into 80% for training and 20% for testing. Of the total data, there are 109 data 19.2% in the low category, 334 data 58.8% in the medium category, and 125 data 22.0% in the high category. This classification aims to assist in mapping the level of malaria risk in the area. In this study, several variables were used for model training, including health centers, sub-districts, age, month, and gender. The results of the analysis showed that the most influential variables were health centers 47.53%, followed by sub-districts 43.77%, age 6.07%, months 2.18%, and gender 0.45%. The Random Forest model built was evaluated using accuracy, precision, recall, and F1-Score metrics. The evaluation results showed that the model was able to classify the number of malaria cases well, with an accuracy value of 0.97. With these results, Random Forest has proven effective as a classification method in malaria cases in Asahan Regency.

I. Introduction

Malaria is a disease spread by the bite of a female Anopheles mosquito infected with parasites from the genus Plasmodium [1]. There are four types of Plasmodium parasites that infect humans, namely: Plasmodium vivax, Plasmodium falciparum, Plasmodium malariae, and Plasmodium ovale. This disease is known as one of the main causes of a fairly high risk of death [2]. In 2018, an estimated 228 million cases were contaminated with a death toll of 405,000 people due to malaria globally, where children under the age of 5 years are the most vulnerable group, contributing 67% of deaths in the world. The most cases of malaria are found in the African region 93%, followed by the Southeast Asian region 3.4%, and the Eastern Mediterranean region 2.1% [3]. Almost all countries in Southeast Asia have reported cases of malaria, in 2018 the World Health Organization (WHO) estimated 8 million cases and 11,600 deaths caused by malaria in the Asian region.

In Indonesia itself, the highest number of malaria cases occurred in 2012 with a total of 417,819 cases and experienced a downward trend to 222,084 cases in 2018. The number of malaria cases continued to increase after 2018. The peak occurred in 2022, surpassing the number of cases in 2012 [4]. In 2016, the highest malaria cases in Sumatra Island were in North Sumatra Province. North Sumatra Province ranked fifth in the highest malaria cases in Indonesia after Papua, NTT, West Papua and Maluku Provinces. Malaria cases in North Sumatra Province were reported at 6,840 cases [5] . Environmental factors include humidity, rainfall, the presence of animals and plants, temperature, and deforestation. Tropical conditions in Southeast Asia amplify the influence of these factors in increasing the spread of malaria [6].

Asahan Regency is an endemic area for malaria in North Sumatra, the incidence of malaria based on routine reports has decreased, this can be seen from the Annual Parasite Incidence (API) figures, namely in 2015 the API value of Asahan Regency was 706,283 per 1,000 with a risk of 1.44%. In 2016, there were 712,684 per 1,000 with a risk of 0.96%. In 2017, there were 18,718 per 1,000 with a risk of 0.65%. And in 2018, there were 724,379 per 1,000 with a risk of 0.28% [7]. Behavioral and environmental factors play an important role in the spread of malaria. Behavioral factors include not using mosquito nets when sleeping, going out often at night, low employment, and low income. Environmental factors include humidity, rainfall, the presence of animals and plants, temperature, and deforestation. Tropical conditions in Southeast Asia amplify the influence of these factors in increasing the spread of malaria [8].

The Asahan District Health Office faces problems in controlling malaria, including unstable fluctuations in cases every year, especially in endemic areas. The available data is generally only a summary of the number of cases without in-depth analysis, making it difficult to determine intervention priorities. Limited resources, such as budget, medical personnel, and health facilities, also hamper targeted control efforts. Therefore, the application of data-based technology, such as the Random Forest algorithm for classifying areas based on risk levels, is needed. This solution will help the Health Office allocate resources more effectively, improve the quality of interventions, and optimize malaria control in Asahan.

There are several previous studies with the same method, namely random forest classification. in previous research, conducted by Hidayat et al with a case study of heart disease obtained high accuracy results of 94%. The process includes data preprocessing, normalization, data split, classification, and evaluation [9]. Then the research conducted by Hadi et al., using a case study of diabetes produced an accuracy of 74.78% [10].

Based on the research description above, it can be concluded that random forests are able to provide a high level of accuracy in classifying data. The purpose of this study is to implement the Random Forest method in classifying the number of malaria cases in Asahan Regency into risk level categories, namely, low, medium and high. Assisting the Health Office and related agencies in making more appropriate decisions in malaria control, such as determining intervention priorities in high-risk areas, based on the classification results.

II. METODOLOGY

Research methods are the steps used by researchers to collect, analyze, and interpret data to answer questions or solve research problems [11]. Figure 1 presents a research process scheme that shows each part and stage of the overall procedure. This illustration provides a comprehensive picture of the designed plan, and facilitates understanding of the reciprocal relationship between components that support each other in achieving the main objectives of the research.

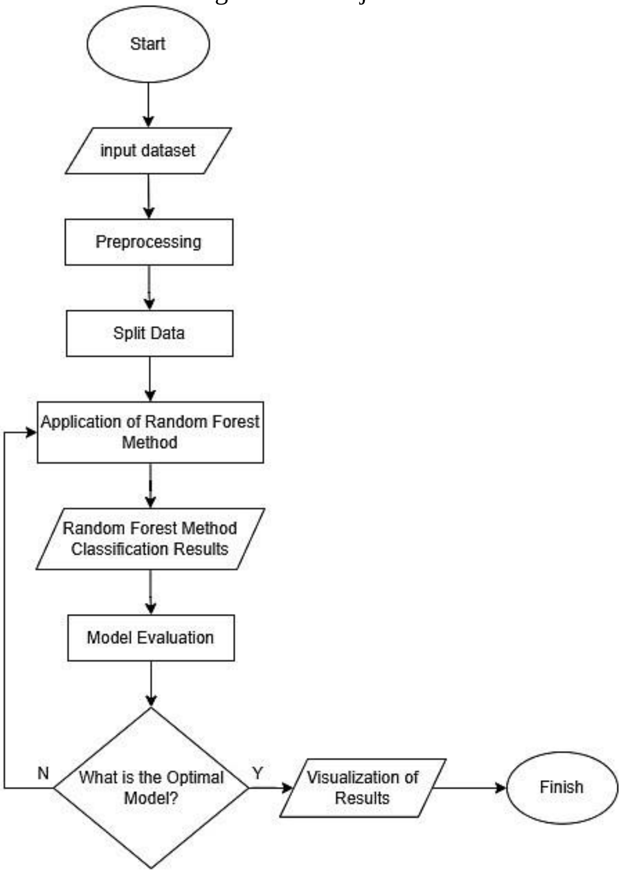

Start
input dataset
Preprocessing
Split Data
Application of Random Forest Method
Random Forest Method Classification Results
Model Evaluation
What is the Optimal Model?
N
Y
Visualization of Results
Finish

Figure 1. Flowchart of the System Work Process Using the Random Forest Algorithm

A. Data collection

In this initial process, the author looks for a dataset that is in accordance with the topic of this research. In this study, the dataset used is primary data obtained directly from the Asahan Regency Health Office. The data includes information on the number of malaria cases in 2023 that are relevant to the research objectives. The dataset consists of 5 variables, namely month, sub-district, gender, age and health center, by grouping them into 3 categories, namely low, medium and high based on the number of cases in 25 sub-districts in Asahan Regency with a total dataset of 568 data.

B. Preprocessing

Data preprocessing is the initial stage in the data analysis process that includes cleaning, transforming, and preparing raw data so that it can be used optimally in further analysis processes, such as data mining, machine learning, and statistical analysis. The main purpose of this preprocessing is to improve data quality so that the analysis results are more accurate and efficient, reduce the presence of noise and irrelevant data that can affect the final results, and adjust the data to suit the requirements of the algorithm that will be applied in the modeling process [12].

The stages in the data preprocessing process that will be discussed include [13]:

1. *Data Cleaning*

   This stage includes the process of identifying and handling incomplete data (such as blank values), outliers, and invalid data. Missing values can be handled by deleting them or replacing them with values that are considered appropriate. Meanwhile, outliers can be handled by deleting them or using statistical approaches such as trimming and winsorizing. In addition, at this stage, duplicate data is also deleted.

2. *Data Transformation*

   Some attributes in the dataset may need to be scaled or distributed to suit the needs of the analysis or modeling. Techniques often used in this stage include normalization and standardization to adjust the scale of the data. In addition, logarithmic transformation can also be applied to change the distribution of attributes that have an exponential pattern.

3. *Feature Selection*

   At this stage, the most relevant attributes are selected and provide significant information for the analysis objectives. Proper feature selection can increase the effectiveness and accuracy of the model. Attributes that are less relevant or have high correlation with other attributes can be removed from the dataset to avoid redundancy.

4. *Encoding Categorical Variables*

   If there are categorical variables in the dataset, then an encoding process is needed to convert them into numeric form so that they can be processed by the machine learning algorithm. Common encoding techniques include one-hot encoding and label encoding.

C. Algorithm Random Forest

The random forest algorithm approach was proposed by an expert named Breiman. Random forest is a machine learning algorithm that has many decision trees as a combined base classifier. This algorithm is a combination of the Random Subspaces and Bagging methods [14]. Random Forest is a development of the Classification and Regression Tree (CART) method, which uses the bag or bootstrap aggregation method and random feature selection [15].

The process of implementing the Random Forest method involves several steps, namely the process begins with bootstrap sampling, which is taking random samples from training data with replacement. Each subset formed has the same size as the initial data, but one data can be selected more than once [16]. This subset is used to build decision trees individually with the formula (1).

$$\mathrm{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$
$$D_i : Bootsrap\ (D), i = 1, 2, \ldots, t \qquad (1)$$

With
n : Amount of data
t : Number of trees in the forest

When forming a tree, random feature selection is performed at each node. This aims to reduce the correlation between trees using the formula (2).

$$m = \sqrt{p} \qquad (2)$$

With
p: Total number of features
m: The number of features selected to determine the split at each node.

Each decision tree is built based on a subset of the data and a selected subset of features. At each node, the best split is determined by measuring impurity using the Gini index or entropy, which respectively measure how pure the data is. The Gini Index value can be calculated using formula (3) as follows.

$$\text{Gini Index} = 1 - \sum_{i=1}^{k} {p_i}^2 \qquad (3)$$

The Gini index is used to determine the best features in each separation process by minimizing the level of impurity, resulting in a more efficient decision tree and being able to separate data more accurately. The smaller the Gini value, the better because it indicates that the data in the group is more homogeneous. Meanwhile, entropy is used to evaluate the quality of separation by calculating the weighted average of the Gini value in each subset formed. The Gini Gain value can be calculated using the formula shown in equation (4).

$$\text{Entropy} = -\sum_{i=1}^{k} p_i\, Log_2(p_i) \qquad (4)$$

With
p_i: is the proportion of samples included in class - i
K: is the number of classes in the target variable

After all the trees are formed, voting is done to determine the final classification result. Each tree provides a prediction and the final result is determined based on the majority vote of all the trees in the forest using a formula (5).

$$\text{Class} = ArgMax_k \sum_{i=1}^{T} \delta(y_i - k) \qquad (5)$$

With
T: Number of trees in the forest
y_i: Class prediction from tree to tree-i
k: Number of classes being calculated
$\delta(y_i - k)$: is 1 if the tree-i prediction is class k, and 0 otherwish . The class with the most votes will be chosen as the final prediction.

The use of the Random Forest method in the classification process has been proven to provide good performance based on the results of various previous studies. The positive performance of this method is achieved through a series of structured and systematic steps, as explained in the reference literature. The stages in the Random Forest calculation process can be seen in Figure 2.

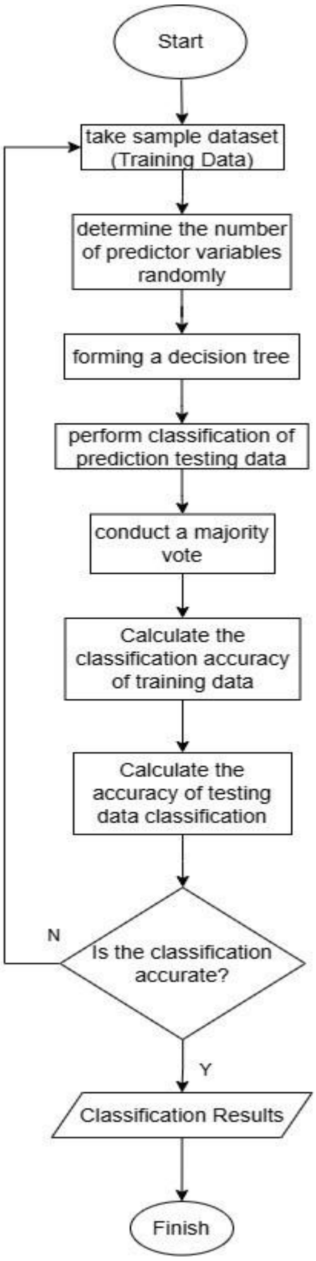

Start
take sample dataset (Training Data)
determine the number of predictor variables randomly
forming a decision tree
perform classification of prediction testing data
conduct a majority vote
Calculate the classification accuracy of training data
Calculate the accuracy of testing data classification
N
Is the classification accurate?
Y
Classification Results
Finish

Figure 2. The Flowchart of Classification Malaria Cases Using Random Forest

D. Evaluation and Validation

The next stage is to analyze and evaluate the classification results obtained from the performance of the Random Forest algorithm. This evaluation aims to understand how the algorithm can improve accuracy in processing datasets regarding academic stress levels. One of the tools used in the evaluation process is the Confusion Matrix, which is a table that describes the comparison between the model's prediction results and actual data. It consists of four cells: True positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [17].

Confusion Matrix dibentuk melalui pelatihan dan model prediksi, lalu hasil klasifikasi setiap sampel ditabulasi. Akurasi tinggi ditunjukkan oleh nilai dominan pada diagonal utama, sedangkan kesalahan ditandai oleh nilai besar pada sel False Positive (FP) atau False Negative (FN). The overall effectiveness level of the model performance is indicated by the accuracy value. Accuracy represents the proportion of correct predictions compared to all data analyzed in the study [18]. Formula (6) describes how to calculate the accuracy value.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision measures the proportion of correct positive predictions to all positive predictions, and indicates the accuracy of the model in classifying positive data [18]. Formula (7) shows how to calculate the precision value.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall is a measure of the average of true positives used to determine how well the model is able to recognize or detect the normal class [18]. The calculation of the recall value is shown in formula (8).

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

F1-Score is used to evaluate the performance of a classification model by considering the balance between precision and recall. Formula (9) shows how to calculate F1 Score.

$$F1 - score = 2 \times \frac{Precision \ \times recall}{Precision + recall} \tag{9}$$

III. Result and Discussion

A. Dataset Retrieval

The data used in this study came from the Asahan Regency Health Office in 2023. The data obtained were 568 data which had five research variables which were used as the basis for the analysis, namely month, sub-district, gender, age and health center. Can be seen in Table 1.

TABLE I
DATASET

No	Bulan	Kecamatan	Jenis Kelamin	Umur	Puskesmas
1	Januari	Tanjung Balai	Laki-laki	42	Sei Apung
2	Januari	Tanjung Balai	Laki-laki	9	Sei Apung
3	Januari	Air Batu	Perempuan	26	Hessa Air Genting
4	Januari	Kisaran Barat	Laki-laki	2	Sidodadi
5	Januari	Kisaran Barat	Laki-laki	14	Sidodadi
6	Januari	Kisaran Barat	Perempuan	53	Sidodadi
7	Januari	Kisaran Barat	Perempuan	24	Sidodadi
…	…	…	…	…	…
564	Desember	Tanjung Balai	Perempuan	20	Sei Apung
565	Desember	Tanjung Balai	Perempuan	20	Sei Apung
566	Desember	Tanjung Balai	Laki-laki	20	Sei Apung
567	Desember	Simpang Empat	Perempuan	34	Simpang Empat
568	Desember	Simpang Empat	Perempuan	34	Simpang Empat

B. *Preprocessing Data*

1. *Class Label Creation (Labeling)*

Labeling is done by calculating the number of malaria cases based on sub-district. Each row in the dataset represents one case of malaria, so the number of cases per sub-district is calculated based on the frequency of occurrence of the sub-district name. Then, the number of cases is categorized into three labels, namely:

a. Low: if the number of malaria cases ≤ 20 cases
b. Medium: if the number of malaria cases > 20 to 50 cases
c. High: if the number of malaria cases > 50 cases

The provisions for the category limits for the number of cases refer to the guidelines and considerations of the P2 (Prevention and Control) Disease Division of the Asahan District Health Office, which technically understands the distribution pattern and level of alertness to malaria cases in the Asahan District area. The following are the labeling results of the number of cases per sub-district, which can be seen in table 2 below.

TABLE II
LABELING RESULT

No	Bulan	Kecamatan	Jenis Kelamin	Umur	Puskesmas	Kategori
1	Januari	Tanjung Balai	Laki-laki	42	Sei Apung	Tinggi
2	Januari	Tanjung Balai	Laki-laki	9	Sei Apung	Tinggi
3	Januari	Air Batu	Perempuan	26	Hessa Air Genting	Sedang
4	Januari	Kisaran Barat	Laki-laki	2	Sidodadi	Sedang
5	Januari	Kisaran Barat	Laki-laki	14	Sidodadi	Sedang
6	Januari	Kisaran Barat	Perempuan	53	Sidodadi	Sedang
7	Januari	Kisaran Barat	Perempuan	24	Sidodadi	Sedang
…	…	…	…	…	…	…
563	Desember	Tanjung Balai	Laki-laki	20	Sei Apung	Tinggi
564	Desember	Tanjung Balai	Perempuan	20	Sei Apung	Tinggi
565	Desember	Tanjung Balai	Perempuan	20	Sei Apung	Tinggi
566	Desember	Tanjung Balai	Laki-laki	20	Sei Apung	Tinggi
567	Desember	Simpang Empat	Perempuan	34	Simpang Empat	Rendah
568	Desember	Simpang Empat	Perempuan	34	Simpang Empat	Rendah

2. *Data Transformation*

The data transformation is carried out as follows: categorical values are converted into numeric using the label encoding method, so that they can be processed by the random forest classification algorithm. In this study, the data consists of five variables and is classified into three categories, namely category 0 (low), category 1 (medium), and category 2 (high). The following are the results of data encoding can be seen in table 3 below.

TABLE III
DATA TRANSFORMATION

No	Bulan	Kecamatan	Jenis Kelamin	Umur	Puskesmas	Kategori
1	0	2	0	42	3	2
2	0	2	0	9	3	2
3	0	5	1	26	6	1
4	0	11	0	2	14	1
5	0	11	0	14	14	1
6	0	11	1	53	14	1
7	0	11	1	24	14	1
8	0	12	1	12	15	2
…	…	…	…	…	…	…
564	11	2	1	20	3	2
565	11	2	1	20	3	2
566	11	2	0	20	3	2
567	11	4	1	34	5	0
568	11	4	1	34	5	0

After successfully changing the data type, the author can calculate the number of target data from the category variable. The goal is to find out how much data is included in the low, medium and high categories from all available data. From the results of manual calculations, the following data were obtained: Low category as many as 109 data, medium category as many as 334 data and high category as many as 125 data. With a total of 568 data. So that the following visualization results are obtained.

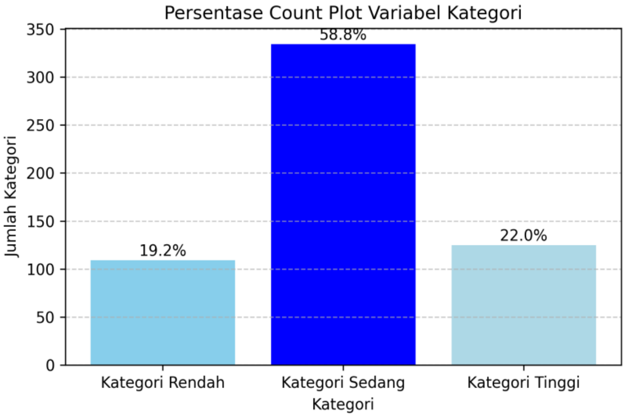
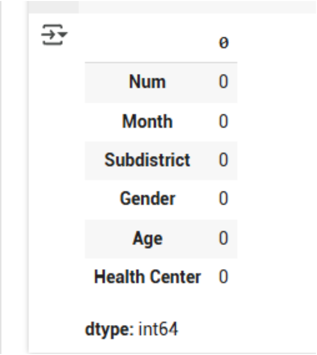


Figure 3. Visualization of categorical variables

Figure 3 shows the distribution of the number of malaria cases classified into three categories, namely Low Category, Medium Category, and High Category. Based on the graph, it can be seen that most of the data is in the Medium Category with a percentage of 58.8%, followed by the High Category at 22.0%, and the Low Category at 19.2%. Although there is a slight difference in the proportion between categories, overall the data distribution is still quite balanced. There is no extreme dominance in one class that can cause the data to be significantly unbalanced (imbalanced dataset). Therefore, there is no need for data balancing techniques such as oversampling or undersampling at the preprocessing stage. Thus, the data can be used directly in the model training process without the risk of bias towards one of the classes.

3. *Handling Missing Value*

This stage aims to check for the presence of missing values in the dataset. If empty values are found, then the data cleaning process will be carried out. However, based on figure 4, it is known that the dataset does not contain missing values.



```
                0
Num             0
Month           0
Subdistrict     0
Gender          0
Age             0
Health Center   0

dtype: int64
```

Figure 4. Missing Value

Based on Figure 4, it can be concluded that there are no missing values in the dataset. Each attribute shows a number of empty values of zero, indicating that all data has been filled in completely. Thus, the dataset is ready to be used for further analysis stages without the need for additional preprocessing related to missing values. In this study, the data normalization process was not carried out because the algorithm used is resistant to the presence of outliers and differences in scale between features. Normalization is generally applied to standardize the scale of variables in order to improve algorithm performance. However, in this context, algorithms such as Random Forest used are included in the type of decision tree, which divides data based on feature values without considering the absolute scale. Therefore, normalization does not have a significant effect on model performance.

C. *Split Data*

After the preprocessing stage is complete, the next step is to divide the data into training data and testing data. Training data is used to build a model by optimally studying the patterns and characteristics of the data. Meanwhile, test data is used to evaluate the performance of the model in generalizing to data that has never been seen before.

In this study, the dataset was divided into 80:20 ratios, where 80% was used as training data and 20% as testing data. From a total of 568 data, 454 data were randomly selected to train the model, while the remaining 114 data were used to test the performance of the model that had been built. This stage is a crucial step before entering the further analysis process.

D. *Classification with Random Forest*

This study uses the Random Forest method to classify the number of malaria cases. The accuracy of the system in classifying is highly dependent on the model training process. The Random Forest algorithm has a number of parameters that are usually adjusted or tested to improve performance and optimize the results of the model being built.

Some of these parameters include [19]:

1. *n_estimators*

   The number of decision trees used in a Random Forest ensemble affects the performance of the model. The more trees used, the greater the chance of the model being more robust against overfitting. However, this also results in increased computational time and resources required.

2. *max_depth*

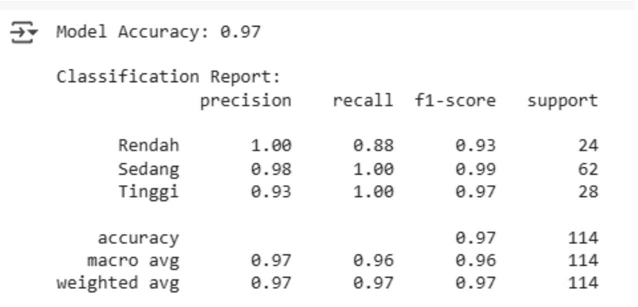
   The max_depth parameter on each decision tree serves to set the level of model complexity. If this parameter is not specified, the tree will continue to grow until each leaf contains a completely homogeneous class or until it reaches the minimum number of samples allowed on each leaf

3. *random_state*

   It is an integer number used to set the initialization of random numbers in the model. By setting this value, the model training results will remain consistent every time it is run, making it easier to compare between different models.

E. *Evaluation Method*

At this stage, the development of the machine learning model is carried out by randomly dividing the dataset into training data and testing data. The purpose of this step is to perform cross-validation in order to obtain an optimal level of prediction accuracy. This study uses the Random Forest algorithm as a classification method. In the initial process, the data is divided with a proportion of 80% for training and 20% for testing, which is used to assess model performance. The classification results using the Random Forest algorithm show very good performance with accuracy, precision, recall, and f1-score of 97% each. This stage includes data division, feature extraction, model training, performance evaluation, and confusion matrix creation. The following are the results of the model evaluation with the application of the Random Forest algorithm.



```
Model Accuracy: 0.97

Classification Report:
              precision    recall  f1-score   support

      Rendah       1.00      0.88      0.93        24
      Sedang       0.98      1.00      0.99        62
      Tinggi       0.93      1.00      0.97        28

    accuracy                           0.97       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.97      0.97       114
```

Figure 5. Random Forest Classification Results

From the prediction results of the Random Forest algorithm, it can be analyzed that out of 114 testing data, there are 3 data that are predicted incorrectly, and there are 111 data that are predicted correctly. To find out more clearly about how accurate the prediction results are made by the Random Forest method, a model evaluation is carried out from the prediction results that have been produced by the Random Forest method by calculating the level of accuracy of the prediction data that has been produced using Accuracy, precision, recall, and F1-Score. The following is the visualization result of the confusion matrix.

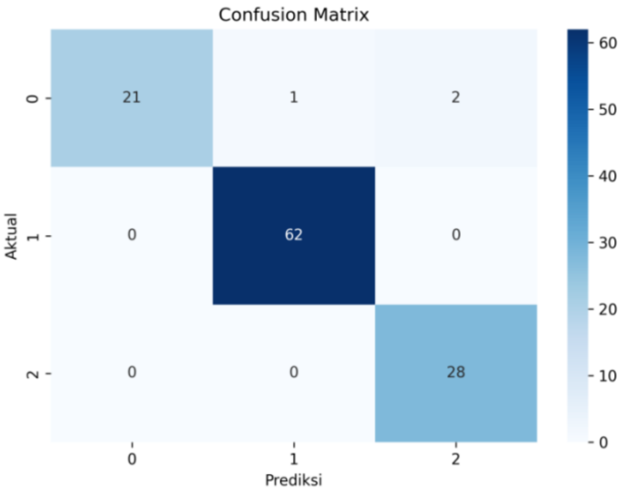


Figure 6. Confusion matrix visualization

Figure 6 shows the confusion matrix of the classification results using the Random Forest algorithm. This confusion matrix is used to evaluate the performance of the model in predicting each class, namely Low (0), Medium (1), and High (2). Based on the evaluation results, the model successfully classified most of the data very well.

The Medium class is the class with the highest accuracy, where all 62 actual data were successfully predicted correctly without any misclassification true positive = 62. This shows that the model has a recall rate of 100% and very high precision for this class. For the High class, as many as 28 actual data were also successfully predicted correctly, so that the recall also reached 100%. However, there were two data from the Low class that were incorrectly classified as High, so that the precision of the High class decreased slightly to 93.3%.

Meanwhile, in the Low class, out of a total of 24 actual data, only 21 data were predicted correctly, while 3 other data were incorrectly classified: one data to the Medium class and two data to the High class. This causes the recall value for the Low class to drop to 87.5%, although the precision remains high. Overall, the Random Forest model shows a total accuracy of 97.4%, with very good performance in recognizing patterns from each class. Although there are few errors in the Low class, these results show that the model has strong and reliable classification capabilities in the context of the data used.

In addition, to understand the contribution of each feature to the classification process, a feature importance analysis was performed. This analysis provides information on which features are most influential in decision making by the Random Forest model, thus providing additional insight into the data patterns and characteristics that are most important in predicting the number of malaria cases.

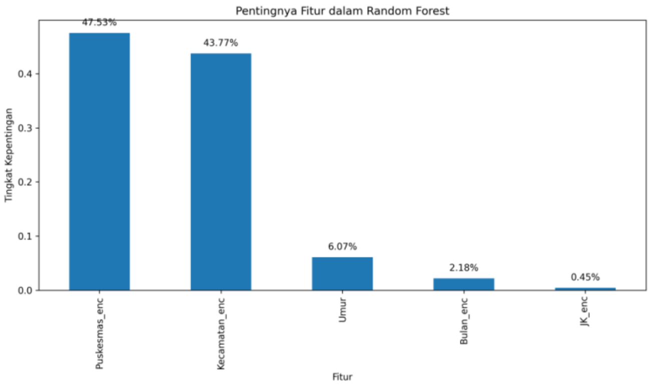


Figure 7. Fiture Importances

Figure 7 is the level of feature importance. Of all the variables, the health center is the most influential variable in this study, namely the health center with a result of 47.53%, 43.77% for the District, 6.07% for age, 2.18% for months and 0.45% for gender.

IV. CONCLUSION

Based on the analysis conducted using the dataset of the number of malaria cases that have 3 categories, it can be concluded that the Random Forest method can classify the number of malaria cases in Asahan Regency very well. By using a dataset of 568 data, which is divided into 80% for training data with a total of 454 data, and 20% for test data with a total of 114 data. From a total of 568 data, consisting of 25 sub-districts in Asahan Regency, 3 categories were produced, namely the low category with 12 sub-districts, the medium category with 11 sub-districts and the high category with 2 sub-districts. Of all the variables, the health center is the most influential variable in this study, namely the health center with a result of 47.53%, 43.77% for the Sub-district, 6.07% for age, 2.18% for months and 0.45% for gender. Then, there are 109 data 19.2% for the low category, 334 data 58.8% for the Medium category and 125 data 22.0% for the high category. The Random Forest model managed to achieve an accuracy level of 97%, precision of 97%, recall of 97%, and F1-Score of 97%. There are several research techniques that help find the right model conditions so that they produce good accuracy when the classification process is carried out, namely data preprocessing to correct errors in the dataset, then classification and model parameter settings are carried out.

REFERENCES

[1] D. A. Rokhayati, R. C. Putri, N. A. Said, and D. S. S. Rejeki, “Analisis Faktor Risiko Malaria di Asia Tenggara,” *Balaba J. Litbang Pengendali. Penyakit Bersumber Binatang Banjarnegara*, vol. 18, no. 1, pp. 79–86, 2022, doi: 10.22435/blb.v18i1.5002.

[2] P. V. K. Tchuenkam *et al.*, “Distribution of non-falciparum malaria among symptomatic malaria patients in Dschang, West Region of Cameroon,” *medRxiv*, p. 2025.03.24.25324523, Mar. 2025, doi: 10.1101/2025.03.24.25324523.

[3] WHO, “World malaria report 2019,” in *Genewa: World Health Organization*, france, 2019, p. Hal. 232.

[4] A. R. M. Nazhid and S. Wulandari, “Mengulas Eliminasi Malaria,” *Bul. APBN*, vol. Vol. VIII, no. No. 23, p. Hal. 2-13, 2023.

[5] KemenKesRI, *Profil Kesehatan Indonesia 2018*, vol. No. 1227. Jakarta: Kementrian Kesehatan Republik Indonesia, 2017. [Online]. Available: website: http://www.kemkes.go.id

[6] H. Agustina Br Ginting *et al.*, “Analisis Faktor Risiko dan Upaya Pencegahan Malaria di Kecamatan Medan Labuhan Analysis of Risk Factors and Malaria Prevention Efforts in Medan Labuhan District,” *J. Kolaboratif Sains*, vol. 8, no. 3, pp. 1428–1436, 2025, doi: 10.56338/jks.v8i3.6918.

[7] Dinkes Kab.Asahan, “Profil Pemerintah Kabupaten Asahan.” Accessed: Nov. 19, 2024. [Online]. Available: https://portal.asahankab.go.id/2018/

[8] W. M. Essendi *et al.*, “Epidemiological Risk Factors for Clinical Malaria Infection In The Highlands Of Western Kenya,” *Malar. J.*, vol. Vol. 18, no. No.1, pp. 1–7, 2019, doi: 10.1186/s12936-019-2845-4.

[9] H. Hidayat, A. Sunyoto, and H. Al Fatta, “Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier,” *J. SisKom-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. Vol. 7, no. No. 1, p. Hal. 31-40, 2023, doi: 10.47970/siskom-kb.v7i1.464.

[10] D. A. Hadi and D. A. N. Sirodj, “Metode Random Forest untuk Klasifikasi Penyakit Diabetes,” *Bandung Conf. Ser. Stat.*, vol. 3, no. 2, pp. 428–435, 2022, doi: 10.29313/bcss.v3i2.8354.

[11] R. Dewi, “EPIDEMIOLOGI PENYAKIT MALARIA DI WILAYAH KERJA PUSKESMAS LABUHAN RUKU KABUPATEN BATU-BARA TAHUN 2020 SKRIPSI Diajukan Sebagai Salah Satu Syarat,” Universitas Islam Negeri Sumatra Utara, Medan, 2021.

[12] F. Alghifari and D. Juardi, “Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes,” *J. Ilm. Inform.*, vol. 9, no. 02, pp. 75–81, 2021, doi: 10.33884/jif.v9i02.3755.

[13] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, “Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru,” *J. Informatics Comput. Sci.*, vol. 05, pp. 97–100, 2023, doi: 10.26740/jinacs.v4n03.

[14] G. Louppe, “Understanding random forests: From theory to practice,” *arXiv Prepr. arXiv1407.7502*, vol. 3, no. 2, p. 225, 2014, doi: 10.48550/arXiv.1407.7502.

[15] Syaidatussalihah and Abdurrahim, “Klasifikasi Status Kemiskinan Menggunakan Algoritma Random Forest,” *J. Mat.*, vol. 5, no. 1, pp. 38–44, 2022, doi: 10.29303/emj.v5i1.133.

[16] M. L. Suliztia, “Penerapan Analisis Random Forest pada Prototype Sistem Prediksi Harga Kamera Bekas Menggunakan Flask,” *Univ. Islam Indones.*, vol. Vol. 3, no. NO. 12, p. Hal. 122, 2020, doi: /dspace.uii.ac.id/123456789/23969.

[17] N. Novianti, S. P. A. Alkadri, and I. Fakhruzi, “Klasifikasi Penyakit Hipertensi Menggunakan Metode Random Forest,” *Progresif J. Ilm. Komput.*, vol. 20, no. 1, pp. 380–392, Feb. 2024, doi: 10.35889/PROGRESIF.V20I1.1663.

[18] S. D. Amalia, M. A. Barata, and P. E. Yuwita, “Optimization of Random Forest Algorithm with Backward Elimination Method in Classification of Academic Stress Levels,” *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 633–641, Jun. 2025, doi: 10.30871/jaic.v9i3.9280.

[19] D. M. U. Atmaja, A. R. Hakim, A. Basri, and A. Ariyanto, “Klasifikasi Metode Persalinan pada Ibu Hamil Menggunakan Algoritma Random Forest Berbasis Mobile,” *JRST (Jurnal Ris. Sains dan Teknol.*, vol. Vol. 7, no. No. 2, p. Hal. 161, 2023, doi: 10.30595/jrst.v7i2.16705.