BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0403–0412
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0403-0412

PERFORMANCE COMPARISON OF DECISION TREE AND LOGISTIC REGRESSION METHODS FOR CLASSIFICATION OF SNP GENETIC DATA

Adi Setiawan[1*], Febi Setivani[2], Tundjung Mahatma[3]

[1]*Department of Data Science, Faculty of Science and Mathematics, Satya Wacana Christian University*
[2,3]*Department of Mathematics, Faculty of Science and Mathematics, Satya Wacana Christian University*
Diponegoro Street, Salatiga, Central Java, 50711, Indonesia

*Corresponding author's e-mail: * **adi.setiawan@uksw.edu***

Article History:

Received: 9th September 2023
Revised: 10th December 2023
Accepted: 9th January 2024

Keywords:

Accuracy;
Decision Tree;
Logistic Regression.

ABSTRACT

This research was conducted to compare the accuracy when decision tree and logistic regression methods are used on some data. Decision tree is one method of classification techniques in data mining. In the decision tree method, very large data samples will be represented as smaller rules, and logistic regression is a method that aims to determine the effect of an independent variable on other variables, namely dichotomous dependent variables. Both algorithms were written and analyzed using R software to see which method is better between the decision tree method and the logistic regression method applied to Single Nucleotide Polymorphism (SNP) genetic data, namely Asthma data. SNP Genetic Data was obtained from R software with the package name "SNPassoc" and the data name "asthma". Asthma data has 57 features, namely Country, Gender, Age, BMI, Smoke, Case control, and SNP genetic code. Comparative analysis was carried out based on the results of the accuracy values obtained in the two methods. Variations in the proportion of the test data used were 40%, 30%, 20%, and 10% and were simulated 1000 times on the grounds of obtaining a better accuracy value. The results obtained show that the decision tree method obtains an accuracy value of 0.5793, 0.5777, 0.5745, 0.5526, respectively, while the logistic regression method is 0.7696, 0.7729, 0.7763, 0.7788, respectively and they are achieved at the proportion of test data of 40%, 30%, 20%, 10%. Thus, it can be concluded that, in this case, the logistic regression method is better than the decision tree method in classifying Asthma data.

ok

How to cite this article:

A. Setiawan, F. Setivani and T. Mahatma., "PERFORMANCE COMPARISON OF DECISION TREE AND LOGISTIC REGRESSION METHODS FOR CLASSIFICATION OF SNP GENETIC DATA," *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0403-0412, March, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Classification is a process used to describe and differentiate a class of existing data. In classification modeling, the best model is needed to predict data [1]. There are two stages in the classification, namely data training and data testing [2]. In the first study, two methods will be used for classification; the decision tree method or decision tree is one of the classification methods applied. The decision tree is one of the methods used as a decision-making procedure for the problems included [3] so that decisions can be taken with careful consideration, which can lead to profitable results or decisions [4].

In addition to the decision tree, the logistic regression method is often applied to data classification. Logistic regression is one of the classical statistical methods used to solve classification problems. The logistic regression method that is often used for classification predicts the probability of each category of qualitative variables, which aims to determine the effect of a variable on other variables, namely, the independent variable with the dichotomous dependent variable [5]. Binary logistic regression is performed if the regression has a dichotomous dependent variable, namely a variable with two categories. Meanwhile, multinomial logistic regression is used when the independent variables used are categorical variables with more than two kinds of categories [6].

The second study was conducted by [7], a study that discussed the performance of a decision tree classification algorithm using WEKA software to evaluate biological data from patients. Comparisons were made to assess whether decision trees can serve as an effective tool for medical diagnosis in general by applying decision trees and regression as tools for the prognosis of medical conditions by taking optimal management from WEKA software. The decision tree is the most competent method for the data. Statistically, it is the most efficient method with an average result of 80% correct classification when executing data.

The third study was conducted by [8]. The research was conducted by applying the C4.5 algorithm, which is a method that is widely used because the classification rules follow human thought processes. It aims to introduce a decision tree-based method to the field of content marketing to efficiently improve marketing capabilities so that the decision tree method can provide reasonable results and accurate suggestions for content marketing.

The fourth study was conducted to analyze COVID-19 data by [9]. To analyze the prediction of COVID-19 on several criteria such as accuracy, sensitivity, precision, and F1-score, decision tree and logistic regression techniques were used in the study, where the two methods are methods that are considered as the most widely used form of machine learning algorithms in classification, interpretability, and accuracy of the algorithm in producing predictive models with structures that are easy to understand and obtain relevant information.

In the fifth study conducted by [4] which compared the Naive Bayes, Decision Tree, and K-nearest Neighbors methods in predicting water quality, the results showed that K-nearest Neighbors had the highest accuracy compared to Naive Bayes and Decision Tree. In further research conducted by [10], which compared the Decision Tree method with the Ridge Logistic Regression in predicting datasets that have binary category features, the result was that the Ridge Logistic Regression method was better than the Decision Tree method, judging from the accuracy value, namely 84% and 81%.

Various studies on the use and comparison of the two methods above have been described, but not many have been used on SNP genetic data. Based on the background and supported by several previous studies, this research will be conducted using a classification method, namely the decision tree method and logistic regression in predicting SNP genetic data. In this study, it will be explained how the processing of SNP genetic data, from data preparation to knowing the accuracy of the decision tree model and logistic regression.

2. RESEARCH METHODS

This research was conducted by following the flow of data mining analysis. Data mining is the process of gathering information through large data. SNP genetic data used in this study was obtained from R software with the package name "SNPassoc" and the data name "asthma". Asthma data has 57 features, namely

Country, Gender, Age, BMI, Smoke, Case control, and SNP genetic code. Data mining is also interpreted as a process of searching for data patterns that were not previously known or suspected before [11].

There are many techniques and concepts that can be applied to the data mining process; therefore, in order to get the data as expected, several processes will be carried out with a number of steps. The processes that will be carried out are data import, data cleaning, data conversion, data sharing, calculation of data analysis results, and comparison of methods. The stages of the process to be carried out can be seen in Figure 1.

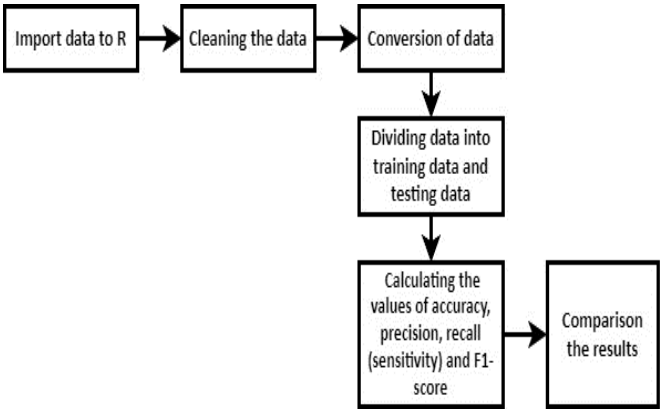

Figure 1. Research Methods

1) Import Data to R:
   This process is calling data that has been stored in R software for data analysis. The data used in this study are the features in the asthma data, namely Country, Gender, Age, BMI, Smoke, Casecontrol, and the SNP genetic code. The country feature contains Australia, Belgium, Estonia, France, Germany, Norway, Spain, Sweden, Switzerland, and the UK; the gender features are male and female; the age feature has an age range of 26.44 – 56.47 with a median of 43.93 and a mean 42.91, the BMI feature has a BMI (body mass index) with a range of 16.71 – 69.20 with a median of 24.93 and a mean of 25.53, the smoke and case control features are categories 0 and 1, and the genetic code features are 51 columns like rs4490198, rs4849332, rs1367179 and others where the code indicates the location of the genetic code.
2) Cleaning the Data:
   The data cleaning process is done by removing incomplete data or errors. Because the data obtained from the database has imperfect or invalid data such as missing data or NA (not available), so these data are better removed from the existing data set so as not to have a different effect on the dependent variable and not reduce accuracy of data analysis. After cleaning the data, there are 1076 rows of data remaining from the initial dataset, namely 1578 rows, and are ready to be taken to the next stage.
3) Conversion of Data:
   To facilitate data analysis, data conversion is carried out by changing the data from its original form according to needs, namely the data will be converted based on a scale of 0-1 and alphabetically with the aim that the data can have the same effect on the dependent variable, as the example in Table 1 below.

Table 1. Genetic Code Conversion

0	0.5	1
AA	AC	CC
AA	AG	GG
AA	AT	TT
CC	CG	GG
CC	CT	TT
GG	GT	TT

4) Dividing Data into Training Data and Testing Data:

   The data is divided into two, namely training and testing data, with the proportion of data testing respectively being 10%, 20%, 30% and 40%.

5) Calculating the values of accuracy, precision, recall (sensitivity) and F1-score based on the decision tree method and logistic regression method and the case. The definition of accuracy is shown in **Equation (1)** and precision is shown in **Equation (2)**. Recall is shown in **Equation (3)** and F1-Score is shown in **Equation (4) [12]**. Accuracy is the ratio of correct predictions to all data. Accuracy is formulated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

where *TP* = true positive, *TN* = true negative, *FP* = false positive and *FN* = false negative.
Precision is the ratio between a positive correct prediction and the overall positive predicted result. Precision is formulated as follows:

$$Precision = \frac{TP}{TP + FP}. \quad (2)$$

Sensitivity (recall) is the ratio between positive correct predictions and overall positive data. Sensitivity is formulated as follows:

$$Recall = \frac{TP}{TP + FN}. \quad (3)$$

F1 - *Score* is the average of the precision and recall values. F1-score is formulated as follows:

$$F_1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (4)$$

Procedure 1 until 5 are repeated 1000 times to find the median of the accuracy, precision, recall and F1-score.

6) Comparison:
The two methods will be compared based on calculating the values of accuracy, precision, sensitivity, and F1-score so that which method is better is obtained. In this case, the selection of the best method is done based on the accuracy value only.

2.1 Decision Tree

The Decision tree method is a method of classification techniques in data mining. In the decision tree method, very large data samples will be represented as smaller rules. Decision trees are also useful for exploring data, finding hidden relationships between a number of candidate independent variables and a dependent variable **[13]**. A decision tree is also called a classification method using a tree structure that has tree nodes representing the attributes that have been tested and each branch representing a distribution of test results and leaf nodes (leaf) representing certain class groups that describe possible outcomes. The top node level of a decision tree is the root node (root) depicting the node at the top of the tree representing the entire population or sample that has the greatest influence on a particular class. Each of these nodes represents a dividing node, where each of these nodes is one input and has at least two outputs. The leaf node is the last node, has only one input and has no output **[14]** as shown in **Figure 2**.

The process in the decision tree is to change the form of table data into a tree model. The tree model will produce simplified nodes, a series of nodes connected via branches. The tree structure in the decision tree moves down from the root node to the leaf nodes **[15]**. Development of a decision tree starts from the top down, namely from the root node which is in the highest position in the decision tree diagram. Branches on a decision tree diagram can enter a decision node or a leaf node based on an evaluation of all the attributes so that they have an outcome that might produce a branch.

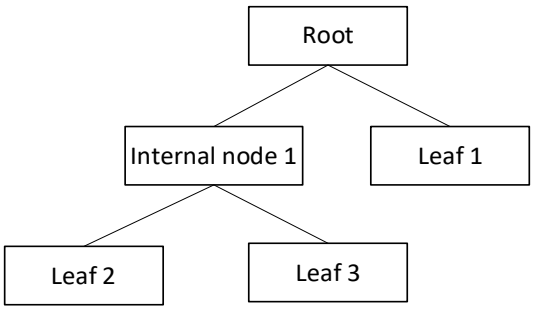

Figure 2. Decision Tree.

The basic concept of a decision tree is to convert data into a decision tree model, then convert the tree model into a rule and simplify the rule. The data in the decision tree is expressed in table form with attributes and records **[16].** Decision trees have the main benefit of being able to simplify the decision-making process from initially complex to simpler so that decision-makers can interpret solutions to the problems entered.

2.2 Logistic Regression

Regression analysis is one of the data analyzes that aims to determine the effect of a variable on another variable, namely, the independent variable with the dichotomous dependent variable. The logistic regression model was created to describe the probability of the dependent variable between 0 stating "disagree" and 1 stating "agree", logistic regression has accurate classification accuracy **[5]**. Independent variables symbolized by X are variables whose values do not depend on other variables and are used to explain the values of other variables. Meanwhile, a variable whose value depends on other variables is the dependent variable, usually symbolized by Y.

The results of observations of the dependent random variable (y) have two categories, namely 0 and 1, so that they follow the Bernoulli distribution with the probability density function **[21]**:

$$P(Y = y) = \pi^{y}(1-\pi)^{1-y}; y = 0,1. \tag{5}$$

In this case, if y = 0 then $P(y=0) = 1-\pi$ and if y =1 then $P(Y=y) = \pi$.
The logistic regression function can be written as follows **[22]**:

$$f(z) = \frac{1}{1+e^{-z}}. \tag{6}$$

where

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_k x_k \, .$$

The z value is between -∞ and +∞ so that the f(z) value lies between 0 and 1, indicating that the logistic model describes the probability or risk of an object. In general, the logistic regression model is written in the form:

$$\pi(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_k x_k)}} \, . \tag{7}$$

Estimation of the parameters of the logistic regression model can be described using the logit transformation of $\pi(x)$ i.e.

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right).$$

Because

$$\frac{\pi(x)}{1-\pi(x)} = e^{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_k x_k)}}. \tag{8}$$

then

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_k x_k \, . \tag{9}$$

The likelihood function that will be used as an estimate of the maximum β value is as follows:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \, . \tag{10}$$

where
n : the number of observations,

x_i : the value of the independent variable for the i-th observation,

y_i : the value of the dependent variable for the i-th observation. For more information about the theory of logistic regression is explained **[17] [18] [19]** and **[20].**

2.3 Classification

One of the data mining methods commonly used is the classification method. Classification is a data analysis process that separates one data class from another to determine which object belongs to a certain category [23]. The data classification process can be carried out by means of a training process and a testing process. These processes produce training data and testing data.

- The training process is carried out by entering sample data into tables prepared for the calculation process. The table includes attributes, the amount of data that has been classified based on predetermined targets.
- The testing process is carried out by entering test data and predictive data. The attributes used in the testing process must match the attributes in the training process [24].

3. RESULTS AND DISCUSSION

3.1 Method of Decision Tree

The study was conducted using the decision tree method for classification of SNP genetic data to determine the results of predicting asthma or not based on Country, Gender, Age, BMI, and Smoke as independent variables and Case Control as the dependent variable.

In this study, the data used was SNP genetic data, namely asthma data, which had 1578 lines and 57 features, and data cleaning was carried out so that the data became 1076 lines and 57 features. Furthermore, the data is divided into two, namely training data and testing data, with variations in the proportion of test data, respectively 40%, 30%, 20%, and 10%. This selection is based on training data which is used more to build models compared to testing data so that in this study 40%, 30%, 20% and 10% were used respectively. By carrying out a simulation of 1000 times, it will produce accuracy, precision, recall, and F1-Score values as in Table 2. From Table 2, it can be seen that using a proportion of test data of 40% gets better results compared to the proportions of other test data.

Table 2. Results of Decision Tree Method

Proportion	Accuracy	Precision	Recall	F1-Score
40%	0.5793	0.8676	0.5523	0.6750
30%	0.5777	0.8655	0.5514	0.6742
20%	0.5745	0.8621	0.5512	0.6734
10%	0.5526	0.8596	0.5433	0.6666

From the data analysis that has been carried out by simulating 1000 times and the proportion of test data is 40%, 30%, 20% and 10%, the histogram is shown in Figure 3, namely the histogram with the proportion of test data 40%, Figure 4, namely the histogram with the proportion of test data 30 %, Figure 5 is a histogram with a proportion of test data of 20%, and Figure 6 is a histogram with a proportion of test data of 10%. By looking at the acquisition of accuracy, precision, sensitivity (recall), and F1-Score values on the histograms, it is known that the histograms tend to be skewed to the left so that the data tends not to be normally distributed, so in Table 2 the median values are taken. In addition, precision histograms tend to be smoother, while accuracy, recall and F1-score histograms tend to be less smooth.

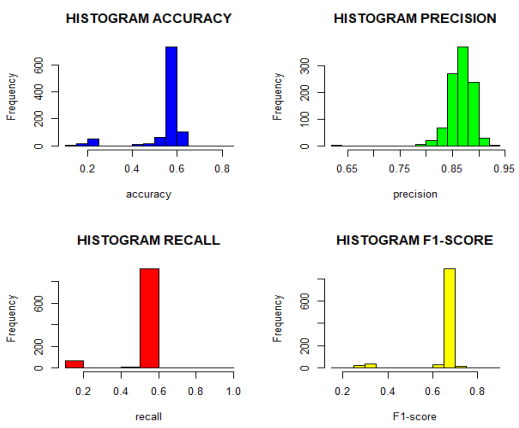

Figure 3. Proportion of Testing Data : 40%.

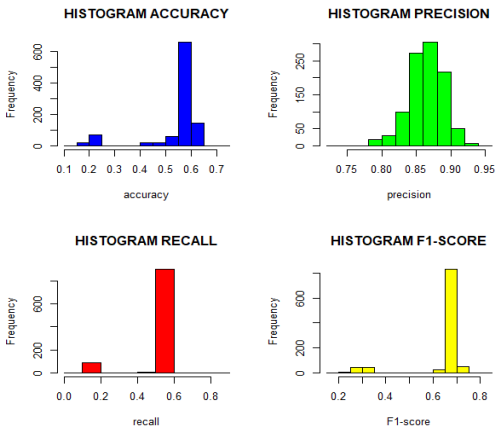

Figure 4. Proportion of Testing Data : 30%.

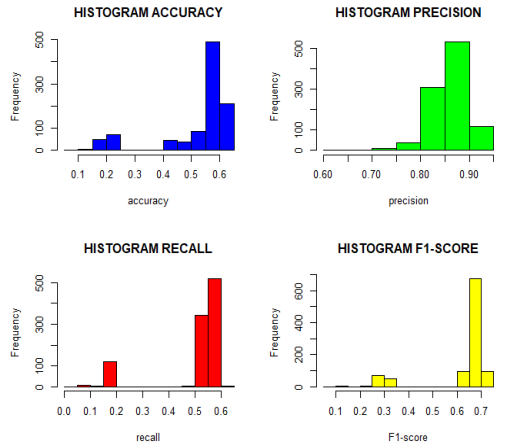

Figure 5. Proportion of Testing Data 20%.

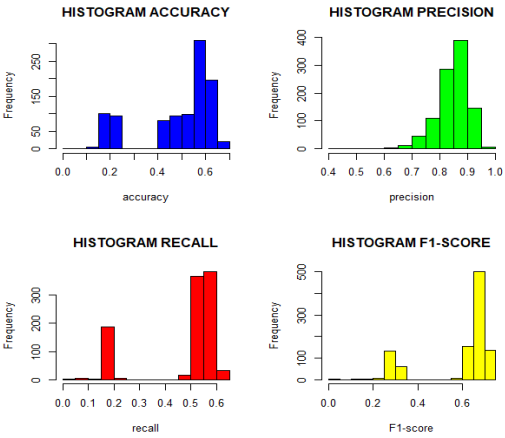

Figure 6. Proportion of Testing Data 10%.

3.2 Method of Logistic Regression

Data analysis using the logistic regression method was carried out in the same way and data as the decision tree method. The results obtained are as in **Table 3**, it can be seen if the highest value is obtained in the proportion of test data of 10%.

Table 3. Results of Logistic Regression Method

Proportion	Accuracy	Precision	Recall	F1-Score
40%	0.7696	0.8035	0.9379	0.8653
30%	0.7729	0.8026	0.9480	0.8679
20%	0.7763	0.8028	0.9540	0,8706
10%	0.7788	0.8020	0.9565	0.8731

From the results of the analysis of the logistic regression method, histograms were obtained with 1000 simulations and different proportions of test data. The histogram looks asymmetrical or tends not to have a normal distribution, so that the values taken in **Table 3** are the median of the analysis results obtained using the logistic regression method. The histogram obtained looks as follows, seen in **Figure 7**, namely the histogram with the proportion of test data 40%; **Figure 8**, namely the histogram with the proportion of test data 30%; **Figure 9**, namely the histogram with the proportion of test data 20%; and **Figure 10**, namely the histogram with the proportion of data test 10%. In contrast to the histograms obtained from the results of using the decision tree method (**Figure 7** until **Figure 6**), in these histograms (**Figure 7** until **Figure 10**), both accuracy, precision, recall, and F1-score histograms tend to be smooth.

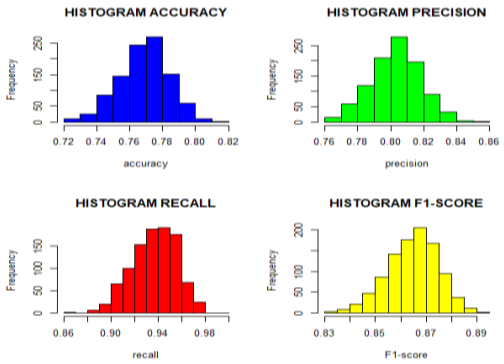

Figure 8. Proportion Testing Data: 40%.

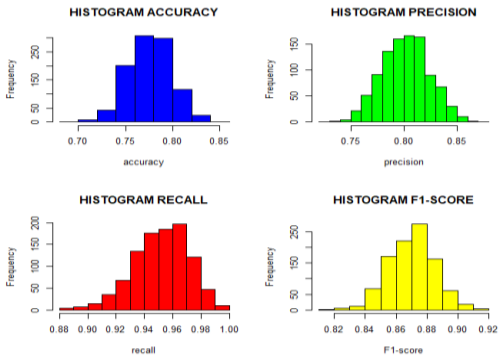

Figure 9. Proportion Testing Data: 30%.

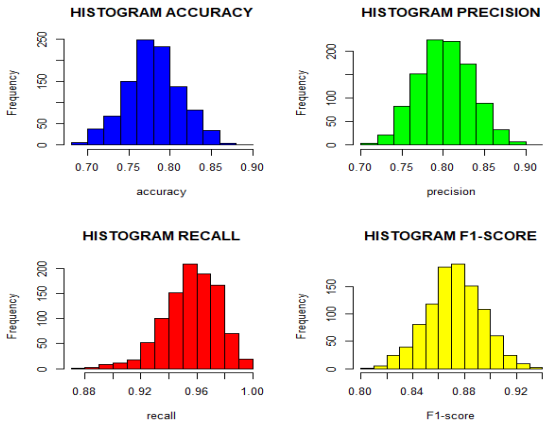

Figure 10. Proportion of Testing Data: 20%.

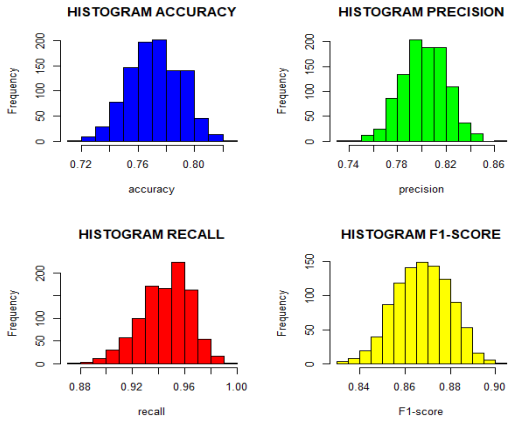

Figure 11. Proportion of Testing Data: 10%.

Based on the results that have been obtained from the analysis using the decision tree and logistic regression methods for classification, these results will be compared to determine which method is better in classification for Genetic Single Nucleotide Polymorphism data on asthma data. It can be seen from **Table 2** and **Table 3** that it is known that the decision tree method has the highest results at the proportion of test data of 40% and logistic regression has the highest results at the proportion of test data of 10%. If a comparison of the two methods is carried out, especially in the proportion of test data 20% for the accuracy value using the Kolmogorov-Smirnov test, the p-value is 2.2×10^{-16} because the p-value < 0.05, the two accuracy values have significant difference.

When viewed from the overall variation in the proportion of test data, the results obtained using the logistic regression method have better values, seen in the values of accuracy, sensitivity (recall), and F1-score, which have significant differences in results with the results obtained using the decision tree method. In contrast, the precision value obtained from the decision tree method has a better precision value than the logistic regression method and has a significant difference.

In the histogram obtained from the decision tree and logistic regression methods, it can be seen that the histogram of the logistic regression is more stable than the histogram of the decision tree. It can be seen that there are several values in the decision tree histogram that are empty or the values generated in this method have a large enough difference in values.

From the results obtained by comparing the results of the accuracy value in the proportion of test data 20%, the result is that the logistic regression method is better when compared to the decision tree seen from the accuracy values of each method, namely 77.63% and 57.45%.

In this study, scoring was used as stated in **Table 1.** The results obtained could be different if different values were used. For this reason, it is necessary to carry out further research using dominant, recessive, or other genetic factors that do not conflict with the principles of genetics.

Other related and recent research can be explained as follows. In the research conducted by Marji and Handoyo, which compared the decision tree method with the Ridge Logistic Regression in predicting datasets

that have binary category features, the result was that the Ridge Logistic Regression method was better than the decision tree method, as seen from the accuracy values of 84% and 81 % **[10]**. In this case, we get the same result; judging from the accuracy value, we get that the logistic regression method is better than the decision tree. Research on the comparison of logistic regression and decision trees was also carried out by Arista in predicting Covid-19 data, getting the result that decision trees were better than logistic regression with accuracy values of 98% and 97% **[9]**. It is different from this research because the accuracy value is better in the decision tree method**.**

CONCLUSIONS

Based on the research that has been done, it can be concluded that the logistic regression method is better than the decision tree method in predicting SNP (single nucleotide polymorphism) genetic data. The results obtained show that the decision tree method obtains an accuracy value of 0.5793, 0.5777, 0.5745, 0.5526, respectively, while the logistic regression method is 0.7696, 0.7729, 0.7763, 0.7788, respectively and they are achieved at the proportion of test data of 40%, 30%, 20%, 10%. Thus, it can be concluded that the logistic regression method is better than the decision tree method. The overall accuracy values between the decision tree and logistic regression methods have a significant difference using the Kolmogorov-Smirnov test with a smaller *p*-value. Suggestions for further research are to replace the methods used, such as SVM, random forest, or other methods to obtain a higher level of accuracy. In addition, you can also use other datasets.

REFERENCES

[1] H. Sulistiani and A. A. Aldino, "Decision Tree C4.5 Algorithm for Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia)," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, pp. 40–50, 2020, doi: 10.21107/edutic.v7i1.8849.

[2] S. A. Zega, "Penggunaan Pohon Keputusan untuk Klasifikasi Tingkat Kualitas Mahasiswa Berdasarkan Jalur Masuk Kuliah," *Semin. Nas. Apl. Teknol. Inf. Yogyakarta*, pp. 7–13, 2014.

[3] V. Anestiviya, A. Ferico, and O. Pasaribu, "Analisis Pola Menggunakan Metode C4.5 Untuk Peminatan Jurusan Siswa Berdasarkan Kurikulum (Studi Kasus : Sman 1 Natar)," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 1, pp. 80–85, 2021, [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[4] A. Tangkelayuk, "The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 1109–1119, 2022, doi: 10.35957/jatisi.v9i2.2048.

[5] P. Juwita, Sugiman, and P. Hendikawati, "Ketepatan Klasifikasi Metode Rergresi Logistik dan Metode Chaid dengan Pembobotan Sampel," vol. 45, no. 1, pp. 1–8, 2021, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/JM/article/view/32699/12083

[6] E. D. Anggara, A. Widjaja, and B. R. Suteja, "Prediksi Kinerja Pegawai sebagai Rekomendasi Kenaikan Golongan dengan Metode Decision Tree dan Regresi Logistik," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 1, pp. 218–234, 2022, doi: 10.28932/jutisi.v8i1.4479.

[7] F. Hajjej, M. A. Alohali, M. Badr, and M. A. Rahman, "A Comparison of Decision Tree Algorithms in the Assessment of Biomedical Data," vol. 2022, pp. 1–9, 2022, doi: 10.1155/2023/9810245.

[8] Y. Liu and S. Yang, "Application of Decision Tree-Based Classification Algorithm on Content Marketing," *J. Math.*, vol. 2022, 2022, doi: 10.1155/2022/6469054.

[9] A. Arista, "Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19," *Sinkron*, vol. 7, no. 1, pp. 59–65, 2022, doi: 10.33395/sinkron.v7i1.11243.

[10] Marji and S. Handoyo, "Performance of Ridge Logistic Regression and Decision Tree in the Binary Classification," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 15, pp. 4698–4709, 2022.

[11] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017, doi: 10.22202/ei.2016.v2i2.1465.

[12] A. Setiawan, "Perbandingan Penggunaan Jarak Manhattan, Jarak Euclid, dan Jarak Minkowski dalam Klasifikasi Menggunakan Metode KNN pada Data Iris," *J. Sains dan Edukasi Sains*, vol. 5, no. 1, pp. 28–37, 2022, doi: 10.24246/juses.v5i1p28-37.

[13] F. Dwi Meliani Achmad, Budanis, Slamat, "Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree," *J. IPTEK*, vol. 16, no. 1, pp. 18–23, 2012, [Online]. Available: http://jurnal.itats.ac.id/wp-content/uploads/2013/06/3.-BUDANIS-FINAL-hal-17-23.pdf

[14] Robianto ; Sampe Hotlan Sitorus ; Uray Ristian, "Penerapan Metode Decision Tree Untuk Mengklasifikasikan Mutu Buah Jeruk BerdasarkanFitur Warna Dan Ukuran," *J. Komput. dan Apl.*, vol. 9, no. 01, pp. 76–86, 2021.

[15] S. Bahri and A. Lubis, "Metode Klasifikasi Decision Tree Untuk Memprediksi Juara English Premier League," *J. Sintaksis*, vol. 2, no. 1, pp. 63–70, 2020.

[16] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, "Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5," *Prism. Pros. Semin. Nas. Mat.*, vol. 3, pp. 64–71, 2020.
[17] F. A. Novianti and S. W. Purnami, "Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasarkan Hasil Mamografi," *J. SAINS dan Seni ITS*, vol. 1, no. 1, pp. D147–D152, 2012.
[18] A. Z. Z. Hariro, I. Sabina, and M. Jannah, "Analisis Regresi Pada Pembelajaran Statistik Ilmu Sosial," *J. Bakti Sos.*, vol. 1, no. 1, pp. 7–13, 2022, [Online]. Available: https://jurnal.asrypersadaquality.com/index.php/baktisosial/article/view/130/171
[19] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.
[20] J. M. Hilbe, *Practical guide to logistic regression*. 2016. doi: 10.18637/jss.v071.b03.
[21] E. Sofha, H. Yasin, and R. Rahmawati, "Klasifikasi Data Berat Bayi Lahir Menggunakan Probabilistic Neural Network dan Regresi Logistik (Studi Kasus di Rumah Sakit Islam Sultan Agung Semarang Tahun 2014)," *J. Gaussian*, vol. 4, pp. 815–824, 2015, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian
[22] S. R. Diaprina and S. Suhartono, "Analisis Klasifikasi Kredit Menggunakan Regresi Logistik Biner Dan Radial Basis Function Network di Bank 'X' Cabang Kediri," *J. Sains dan Seni ITS*, vol. 3, no. 2, pp. D218–D223, 2014, [Online]. Available: https://ejurnal.its.ac.id/index.php/sains_seni/article/view/8139%0Ahttps://ejurnal.its.ac.id
[23] R. Ariadni and I. Arieshanti, "Implementasi Metode Pohon Keputusan untuk Klasifikasi Data dengan Nilai Fitur yang Tidak Pasti," *ResearchGate*, no. June, pp. 3–5, 2015.
[24] M. Y. Firmansyah, "Penerapan Algoritma Itterative Dechotomiser 3 (ID3) Untuk Klasifikasi Penyakit Tifoid," vol. 3, pp. 1–6, 2019.