Jurnal TAM (Technology Acceptance Model) Volume 12, Nomor 2, Desember 2021 Hal. 149-157
p-ISSN : 2339-1103
e-ISSN : 2579-4221

COMPARISON OF TREE IMPLEMENTATION, REGRESSION LOGISTICS, AND RANDOM FOREST TO DETECT IRIS TYPES

Siti Mukodimah[1], Chairani Fauzi[2]
[1,2]Faculty of Computer Science, Informatics and Business Institute of Darmajaya, Lampung
[1]Departement of Information System, STMIK Pringsewu, Lampung
[1,2]Jl. Z.A. Pagar Alam No 93 Labuhan Ratu, Bandar Lampung, Lampung, Indonesia
[1]Jl. Wisma Rini No 09 Pringsewu, Lampung, Indonesia
E-mail: mukodimah97@gmail.com[1], chairani@darmajaya.ac.id[2]

Article history:
Received: August 20, 2021
Revised: October 4, 2021
Accepted: October 13, 2021

Abstract
Iris plant is a genus of Iridaceae plants that have 260-300 species of flowering plants that have a striking flower color and dominant color in each region. The name iris is taken from the Greek word for rainbow, which is also the name for the Greek goddess of the rainbow, Iris. The number of types of iris plants with almost the same physical characteristics, especially in the pistil and crown, causes the misdetection of iris plant types. Iris plants are deliberately used because data is already available digitally on the internet and software such as orange and is widely used as a material for classifying objects. This research was conducted to classify iris plant types using three algorithms, namely Tree algorithm, Regression Logistics, and Random Forest. Classification algorithms are a learning method for predicting the value of a group of attributes in describing and distinguishing a class of data or concepts that aim to predict a class of objects whose class labels are unknown. The results showed the largest AUC (Area Under Curve) value obtained by the Random Forest method. AUC accuracy is said to be perfect when the AUC value reaches 1,000 and the accuracy is poor if the AUC value is below 0.500. As for the precision value of the three models used Random Forest has the highest precision value. From the data tests that have been done training and testing can be seen that the level of accuracy of testing of the three models where the Random Forest model is superior as a method for classification of irises.

Keywords:
Iris Plant;
Tree Algorithm;
Logistic Regression;
Random Forest;
Classification.

1. INTRODUCTION

Iris plant or iris flower is a type of flower that is part of the Iridaceae family which consists of 300 species (www.orami.co.id). Iris flowers have a variety of striking colors. The many species of iris plants make researchers interested in researching iris plants. Previously, there have been many studies on iris plants, iris plants are deliberately used because the data is already available digitally on the internet and software such as orange and is widely used as material for classifying objects. Existing data based on research by Sir Ronald Aylmer Fisher in 1936, contains 50 data samples for 3 types of iris. The iris plants tested had almost the same physical appearance, only differing in the width of the petals, the length of the petals, the width of the crown, and the length of the crown as shown in Figure 1 below.

Figure 1. types of iris plants
Source: Google Pictute

The number of methods used to identify an object is based on the characteristics of the object. One way to classify the types of iris plants is by using the characteristics of each type of iris as the criteria for classifying the types of iris plants.

Based on the research by Diwahana Mutiara, the research was conducted to compare the clustering algorithm to determine the type of iris flower. The clustering algorithm techniques used include K-Means and K-Medoids. Davies Bouldin values and the number of clusters were examined using the rapidminer tool. The results show that the K-Means algorithm has the lowest davies Bouldin value of 0.167, while K-Medoids produces a davies Bouldin value of 0.291, but among the three algorithms, the K-Means algorithm is the most dominant and best algorithm in the comparison process of interest grouping. iris [1]. Further research was carried out by Fitria Febrianti in this study by comparing the K-Means and Fuzzy C-Means methods to cluster iris data. These two methods have differences. Not only in terms of algorithms, but in terms of calculating the value of the root mean square error (RMSE) it is also different. To calculate the RMSE value, two indicators are needed, namely training data and

checking data. From the discussion, the Fuzzy C-Means method has a smaller RMSE level than the K-Means method, namely 80 training data and 70 checking data with an RMSE value of 2.2122E-14. This shows that the Fuzzy C-means method has a higher level of accuracy than the K-Means method. [2], Research [3] conducted a classification of irises by applying logistic regression and random forest models. The study also used two feature extraction methods: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The results obtained from the performance comparison of both Machine Learning methods and both Dimensionality reduction methods show that LDA works much better than PCA, where using LDA Logistic Regression and Random Forest provides the same results. Research [4] Analyze by applying three methods. The KNN method outperforms other classification methods and shows that in the study there are no misclassifications based on the results obtained. Research [5] The results of this study suggest that iris species may be an important source of pharmacological active compounds such as flavonoids, isoflavones and xanthones. The study of the effects of environmental factors on the production and accumulation of secondary metabolites in Iris species is important.

Based on five studies that have been carried out cluttering types of iris plants by applying clustering and comparative data mining methods, namely K-Means, Fuzzy C-Means, K-Medoids, KNN, LDA, PCA, while in the research that will be carried out a classification method is applied using three data mining methods, namely KNN, naïve Bayes, and Neural Network using data slices as a dataset that t available digitally on the internet and widely used as a material for the test object classification.

Iris plants that have many species become an interesting topic to be discussed in the research. In addition, the iris dataset has also been provided digitally, making it easier for researchers to obtain the iris dataset. Therefore, the authors are interested in removing iris plants by applying the KNN, Naïve Bayes, and Neural Network methods which will be used as methods to detect the types of iris plants. With this research, it is expected to find a new classification pattern that is different from previous research.

II. LITERATURE RIVIEW

2.1. Data Mining Concept

Efrain Turban (2005), Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases [6]. Daniel T. Larose (2004), Data Mining is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, using pattern recognition technology as well as statistical and mathematical techniques. [7]. Kusrini (2009), the terms data mining and knowledge discovery in databases are often used interchangeably to describe the process of extracting hidden information in a large database. Understanding the two terms have different concepts but are related to each other. One of the stages in the whole process of knowledge discovery in databases is data mining.

Based on some of the definitions of data mining above, it can be concluded that data mining is a process to find patterns using statistical techniques to explore hidden information in one large database. Knowledge discovery in databases, in general, can be in the picture right below.

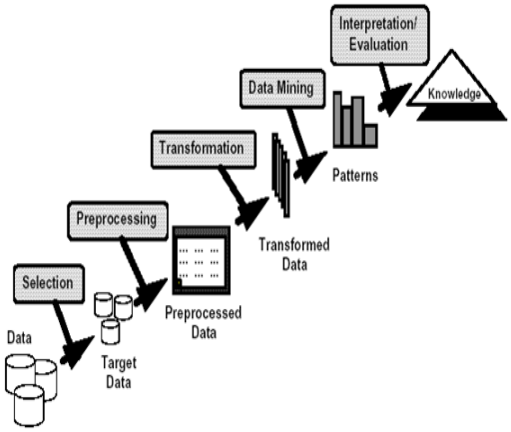


Figure 2. Stages of *Knowledge Discovery in Databases*
Source: Kusrini 2009

Data mining is divided into several groups based on the tasks that can be done, namely [7]:

1) Description
   Analytical research simply attempts to describe the patterns and trends contained in the data. Descriptions of patterns and tendencies often provide possible explanations for a pattern or tendency.
2) Estimate
   In variable estimation, the target estimate is more numerically directed than the category. The estimation model is built using a complete record that provides the value of the target variable as the predictive value. Furthermore, in the next review, the estimated value of the target variable is made based on the value of the prediction variable.
3) Prediction
   Predictions have similarities with estimation and classification. It's just that the predictions show something that hasn't happened yet.
4) Classification
   Variable classification is categorical. As we can classify income into three classes, namely high, medium, and low.
5) Clustering

Clustering is a grouping of records, based on observation and attention and forming a class of similar objects. A cluster is a collection of records that bear similarities to each other and have irregularities with records in other clusters. Clustering is different from classification in the absence of target variables in clustering.

6). Association
Identify the relationship between various events that occur at one time

2.2. Iris Plant

Iris plants are a genus of flowering plants that have diverse and unique flowers. The Iris genus has many species. Flowers Iris to be one genus of 260 to 300 species of flowering plants. The name Iris is taken from the Greek word meaning rainbow. Some authors state that the name iris refers to the many varieties of flower colors that the genus of plants has found among the many species. Three varieties of Iris were used in the Iris flower data described by Ronald Fisher in his 1936 paper. The use of various measurements in taxonomic problems is an example of linear discriminant analysis. Iris plants are ornamental plants that are cultivated because they have attractive flowers. In addition to being an ornamental plant iris plants are also plants that are commonly used as medicines. Parts of the plant that are used as medicinal ingredients include roots, rhizomes, and leaves. The leaves of the iris plant are a type of single leaf and the organs of the iris plant are lanset shaped and elongated with the tip of the iris plant leaves that tapered. Iris plants have stalks that reach 3 to 25 cm in length and iris plants are included in the type of perennial flower or one of the flowers that can live more than 2 (two) years. (www.melydaily.com).

III. RESEARCH METHODS

3.1. Data Collection Method

The data collection method is an important thing in research and is a strategy or way used by researchers in collecting data needed in their research. The data collection methods used in this study are:

a. Literature review
Library reviews are conducted by reading, quoting, and making notes sourced on library materials that support and relate to research in this regard regarding naïve Bayes mining data, Random Forest, and logistic regression.

b. Documentation
The data used in the study is taken from data already available digitally on the internet and software such as orange.

3.2. Naïve Bayes Method

Naïve Bayes is a simple probabilistic classification method that sums a set of frequency probabilities and combinations of values from existing datasets. Naïve Bayes is based on the simplifying assumption that attribute values are conditionally independent if given an output value. In other words, given the output value, the probability of observing together is the product of individual probabilities [8]. The advantage of using Naive Bayes is that this method only requires a small amount of training data to determine the parameter estimates needed in the classification process. Naive Bayes often performs much better in most complex real-world situations than expected [9]. The Naive Bayes Classifier is considered to work very well compared to other classifier models, namely the Naive Bayes Classifier has a better accuracy rate than other classifier models [10].

Naive Bayes Classifier included into learning supervised, Naive Bayes estimate the conditional class opportunities to assume that the attributes are conditionally independent given the label y, the conditional independence assumption can be expressed in the following form [11]. The calculation of Naive Bayes to the equation below.

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \qquad (3.1)$$

Information:
X = Data with unknown class (proof)
H = Hypothesis data X is a class specification
P(H|X) = Probability of hypothesis H true for condition X (posterior prob.)
P(H) = Hypothesis probability H (prior prob.)
PX = Probability prior to proof X

Classification by Naive Bayes works on probability theory that sees all the data as evidence in the probability. This gives a characteristic Naive Bayes as follows. (Indrawan, 2017).

1. Methods Naïve Bayes firm (robust) against the isolated data which typically represents data with different characteristics (outliner). Naive B ayes also can handle incorrect attribute values by ignoring the training data during the development process models and predictions.
2. Resistant to irrelevant attributes.
3. Attributes that are correlated bias de-declaration classification performance, Naïve Bayes, for that attribute independence assumption is not there.

The algorithm Naive Bayes has the advantages of being relatively easy to be implemented because it does not use numerical optimization, matrix calculations, and others, Efficient training, and use can use the data binary or polynomial because it is consumed independent then allow this method is implemented with a variety of dataset, relatively high accuracy. Algorithm Naive Bayes also have the disadvantage that the class probability estimates are inaccurate, and limits or thresholds must be determined manually and not by analysis.

3.3. Logistic Regression

Logistic Regression is a statistical method that is often used to analyze data describing the response variable and one or more predictive variables. The response variable from Logistic Regression is a dichotomy with only 1 (yes) and 0 (no) values, so the resulting response variable will follow the Bernoulli distribution with the following probability function [12], [13]

$$f(y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i)^{1-y_i} \quad (3.2)$$

With $y_i = 0,1$

Based on equation (3.2), the logistic regression equation model (3.3) is obtained.

$$\pi(x) = \frac{\exp(\beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)} \quad (3.3)$$

From equation (2) above we can transform commonly called the logit transformation $\pi(x)$ to obtain a function of g (x) which is linear in its parameters, thus simplifying predictive regression parameters defined by equation (3 .4)

$$g(x) = In\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad (3.4)$$

The Maximum Likelihood Estimator (MLE) method is a technique used to find a certain point so that it can maximize a function. This method is also used to predict the parameters contained in the Logistic Regression model. This method predicts by using the likelihood function. The likelihood function used is

$$L(\beta) = \ln\big(l(\beta)\big) = \sum_{j=0}^{p}\left[\sum_{i=1}^{n} y_i x_{ij}\right]\beta_j - \sum_{i=1}^{n} In\left[1 + exp\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)\right] \quad (3.5)$$

From Equation (3) we can differentiate against the β, after which it is equated to 0, but this way often obtains implicit results. To solve this problem, the Newton Rapson iteration method was performed to raise the likelihood function [14], [15].

The practice of testing parameters using the logistic regression method can be done simultaneously or in turn. In simultaneous testing, the test statistics used are G test statistics or commonly referred to as the Likelihood Ratio Test [15].

One of the methods used to interpret the predictive variable coefficients is by using the Odds ratio method. The odds ratio serves as an indicator of the comparison of the chances of appearing or not arising from an event. If the odds ratio value is less than 1, then there is a negative relationship between the predictive variable and the response variable every time the value of the predictive variable (X) changes and if the odds ratio value obtained is greater than 1, then between the predictive variable and the response variable, there is a positive relationship every time. times the change in the value of the predictive variable (X). The test statistic used in testing the suitability of the model is the Hosmer-Lemeshow Test statistic (Cˆ).

3.4. Random Forest

The Random Forest (RF) method is a method that can improve accuracy results because generating child nodes for each node is done randomly. This method is used to build a decision tree consisting of root nodes, internal nodes, and leaf nodes by taking attributes and data randomly according to the applicable provisions. The root node is the topmost node, or commonly referred to as the root of the decision tree. An internal node is a branching node, where this node has at least two outputs and only one input. While the leaf node or terminal node is the last node that has only one input and no output. The decision tree begins by calculating the entropy value as a determinant of the level of attribute impurity and the value of information gain. To calculate the entropy value, the formula as in equation 1 is used, while the information gain value uses the equation. [15], [16]

$$Entropy\ (Y) = -\sum_i p(c|Y)\ log_2 p(c|Y)z \quad (3.6)$$

Where Y is the set of cases and p(c|Y) is the proportion of Y values to class c.

$$\text{Information Gain}(Y, a) = \text{Entropy}(Y) - \sum_{vE} \text{Values}(a)\frac{|Ya|}{|Ya|}\text{Entropy}\ (Y_V) \quad (3.7)$$

Where Values(a) are all possible values in the case set a. Yv is a subclass of Y with class v corresponding to class a. Yes are all values that correspond to a.

IV. RESULTS

In this study, the data used amounted to 150 data divided into two for training data and testing data. For training data used 70 percent of the data. And the rest of the data is used for testing. For testing, the system will take several data randomly from the dataset according to the number of input testing data entered in the form by the user. The training data is modeled using several parameters such as Tree, Random Forest, KNN, Naïve Bayes, and Neural Network. The result of the training is learning stored in the system which will be a reference for the system to determine input iris data.

p-ISSN : 2339-1103
e-ISSN : 2579-4221

Table 1. Test Data Table

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2
6	Iris-setosa	5.4	3.9	1.7	0.4
7	Iris-setosa	4.6	3.4	1.4	0.3
8	Iris-setosa	5.0	3.4	1.5	0.2
9	Iris-setosa	4.4	2.9	1.4	0.2
10	Iris-setosa	4.9	3.1	1.5	0.1
11	Iris-setosa	5.4	3.7	1.5	0.2
12	Iris-setosa	4.8	3.4	1.6	0.2
13	Iris-setosa	4.8	3.0	1.4	0.1
14	Iris-setosa	4.3	3.0	1.1	0.1
15	Iris-setosa	5.8	4.0	1.2	0.2
16	Iris-setosa	5.7	4.4	1.5	0.4
17	Iris-setosa	5.4	3.9	1.3	0.4
18	Iris-setosa	5.1	3.5	1.4	0.3
19	Iris-setosa	5.7	3.8	1.7	0.3
20	Iris-setosa	5.1	3.8	1.5	0.3
21	Iris-setosa	5.4	3.4	1.7	0.2
22	Iris-setosa	5.1	3.7	1.5	0.4
23	Iris-setosa	4.6	3.6	1.0	0.2
24	Iris-setosa	5.1	3.3	1.7	0.5
25	Iris-setosa	4.8	3.4	1.9	0.2
26	Iris-setosa	5.0	3.0	1.6	0.2
27	Iris-setosa	5.0	3.4	1.6	0.4
28	Iris-setosa	5.2	3.5	1.5	0.2
29	Iris-setosa	5.2	3.4	1.4	0.2
30	Iris-setosa	4.7	3.2	1.6	0.2
31	Iris-setosa	4.8	3.1	1.6	0.2

In the data training process, the dataset used is 70 percent of the total iris data set which is modeled as shown in Figure 1 below.

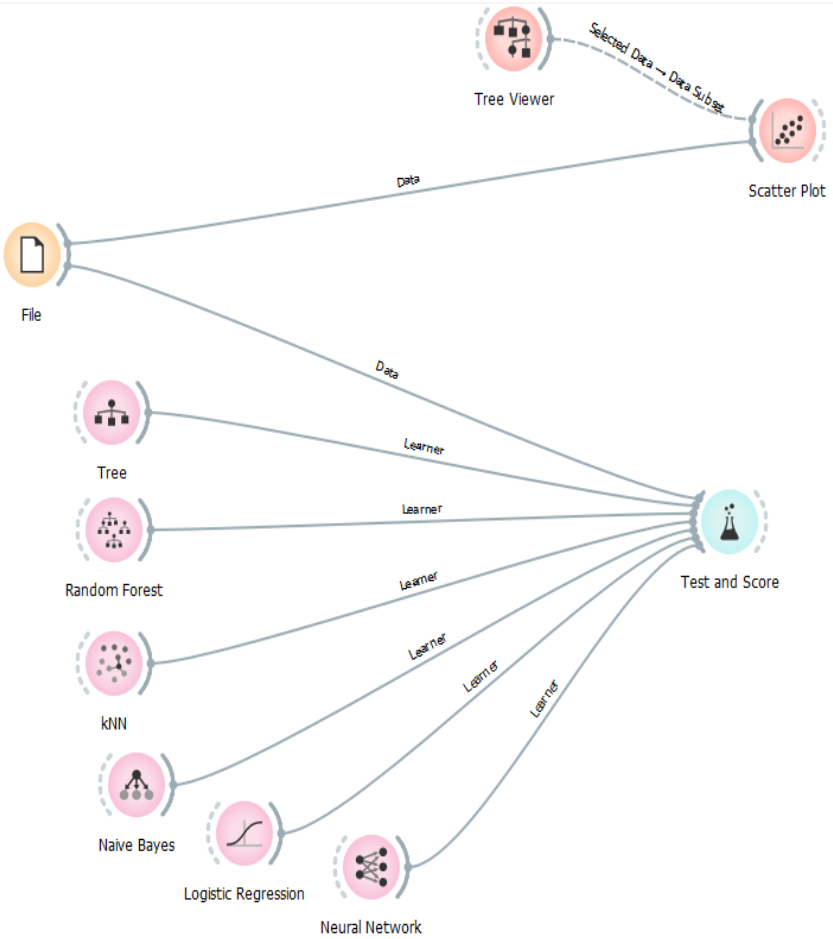


Figure 2. Training set process.

From the training set process carried out using 7 parameters as the test model applied in the ORANGE application, the results of the training set test can be seen in Figure 3 below.

Model	AUC	CA	F1	Precision	Recall
kNN	0.988	0.953	0.953	0.953	0.953
Tree	0.965	0.953	0.953	0.953	0.953
Random Forest	0.989	0.947	0.947	0.947	0.947
Neural Network	0.993	0.947	0.947	0.948	0.947
Naive Bayes	0.981	0.893	0.893	0.894	0.893
Logistic Regression	0.998	0.967	0.967	0.967	0.967

Model Comparison by AUC

	kNN	Tree	Random Fo...	Neural Net...	Naive Bayes	Logistic Reg...
kNN		0.889	0.382	0.320	0.863	0.173
Tree	0.111		0.039	0.028	0.170	0.031
Random Forest	0.618	0.961		0.435	0.787	0.319
Neural Network	0.680	0.972	0.565		0.857	0.342
Naive Bayes	0.137	0.830	0.213	0.143		0.051
Logistic Regression	0.827	0.969	0.681	0.658	0.949	

Figure 3. Model Comparison

From the figure, it can be seen that the precision results of the 7 (seven) algorithms are applied as test parameters. Good precision is where the precision number obtained is close to 1 (one). Of the seven parameters applied, the highest precision value is shown by the logistic regression algorithm with a precision number of 0.967.

After the data training process is carried out using 7 (seven) models, then data testing is carried out using three models with the highest precision during training, the data used at the highest time is 30%. The data testing process is carried out by applying the dataset to the ORANGE application which can be seen in Figure 4. below.

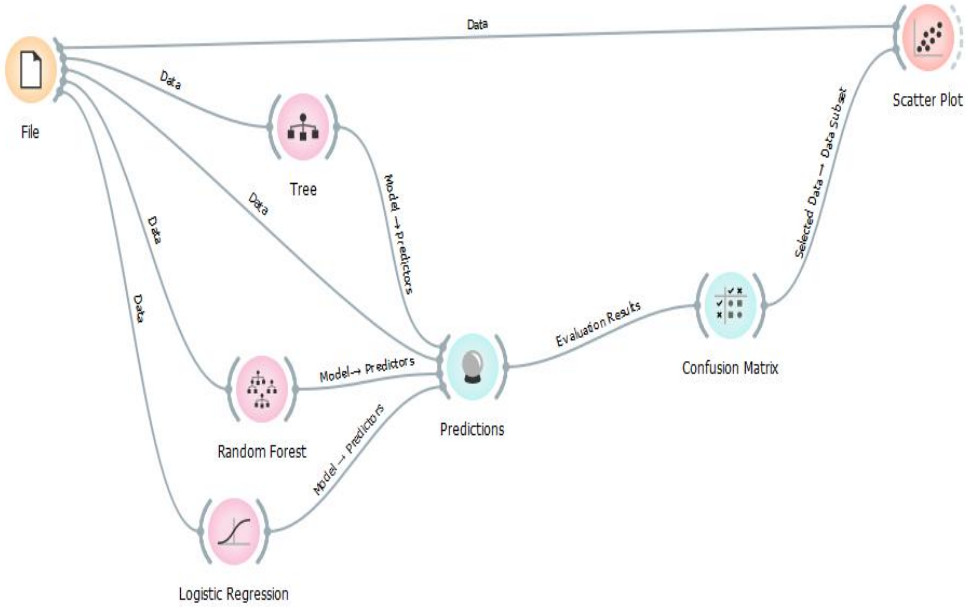
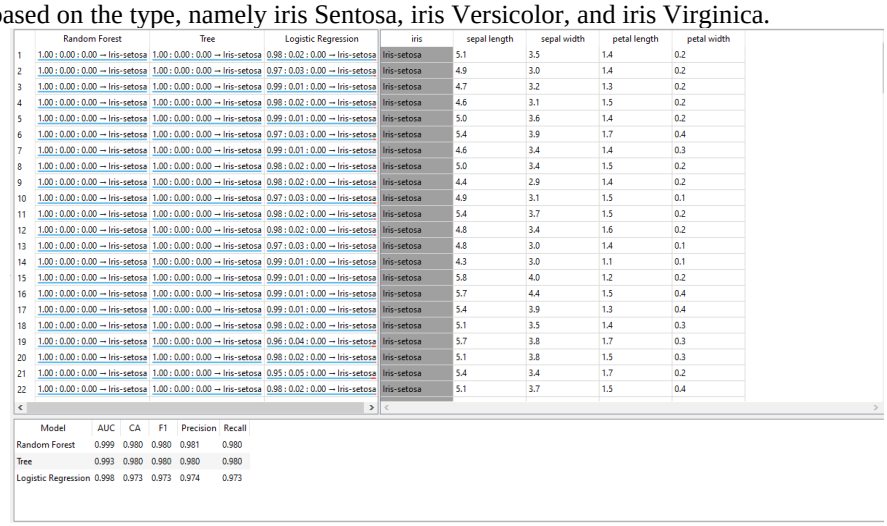


Figure 4. Data Testing Process

All tests were carried out using ORANGE by utilizing three classification models with the highest precision when training data was carried out using seven models as parameters. The inputs used include four features and one output. The four characteristics used as input are the length of the petals, the width of the petals, the length of

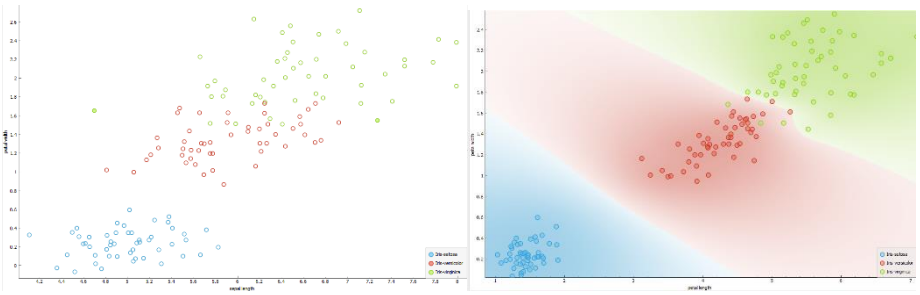
the crown, and the width of the flower crown. After the input process, the resulting output is in the form of iris classification based on the type, namely iris Sentosa, iris Versicolor, and iris Virginica.



	Random Forest	Tree	Logistic Regression	iris	sepal length	sepal width	petal length	petal width
1	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.5	1.4	0.2
2	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.97 : 0.03 : 0.00 → Iris-setosa	Iris-setosa	4.9	3.0	1.4	0.2
3	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.99 : 0.01 : 0.00 → Iris-setosa	Iris-setosa	4.7	3.2	1.3	0.2
4	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	4.6	3.1	1.5	0.2
5	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.99 : 0.01 : 0.00 → Iris-setosa	Iris-setosa	5.0	3.6	1.4	0.2
6	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.97 : 0.03 : 0.00 → Iris-setosa	Iris-setosa	5.4	3.9	1.7	0.4
7	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.99 : 0.01 : 0.00 → Iris-setosa	Iris-setosa	4.6	3.4	1.4	0.3
8	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	5.0	3.4	1.5	0.2
9	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	4.4	2.9	1.4	0.2
10	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.97 : 0.03 : 0.00 → Iris-setosa	Iris-setosa	4.9	3.1	1.5	0.1
11	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	5.4	3.7	1.5	0.2
12	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	4.8	3.4	1.6	0.2
13	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.97 : 0.03 : 0.00 → Iris-setosa	Iris-setosa	4.8	3.0	1.4	0.1
14	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.99 : 0.01 : 0.00 → Iris-setosa	Iris-setosa	4.3	3.0	1.1	0.1
15	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.99 : 0.01 : 0.00 → Iris-setosa	Iris-setosa	5.8	4.0	1.2	0.2
16	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.99 : 0.01 : 0.00 → Iris-setosa	Iris-setosa	5.7	4.4	1.5	0.4
17	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.99 : 0.01 : 0.00 → Iris-setosa	Iris-setosa	5.4	3.9	1.3	0.4
18	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.5	1.4	0.3
19	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.96 : 0.04 : 0.00 → Iris-setosa	Iris-setosa	5.7	3.8	1.7	0.3
20	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.8	1.5	0.3
21	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.95 : 0.05 : 0.00 → Iris-setosa	Iris-setosa	5.4	3.4	1.7	0.2
22	1.00 : 0.00 : 0.00 → Iris-setosa	1.00 : 0.00 : 0.00 → Iris-setosa	0.98 : 0.02 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.7	1.5	0.4

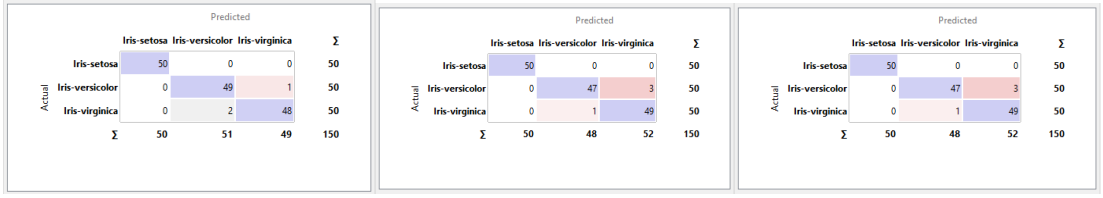
Model	AUC	CA	F1	Precision	Recall
Random Forest	0.999	0.980	0.980	0.981	0.980
Tree	0.993	0.980	0.980	0.980	0.980
Logistic Regression	0.998	0.973	0.973	0.974	0.973

Figure 5. iris prediction table



(a) Sepal (b) Petal

The picture above is an image of the results of testing the iris flower dataset which was tested using three models, namely tree, random forest, and logistic regression. Based on the picture above, it can be seen that the test was carried out to classify the sepals and petals. In the picture (a) which shows the classification results of sepal petals, it can be seen that the blue group which shows the iris Sentosa petals separates completely, this happens because the characteristics of the iris Sentosa petals are very different from the iris virginica and the color version of the iris. As for the red and green groups, there are several points that are mixed, this is because the physical characteristics of the virginica iris petals and the color version of the iris have little in common so that during the testing process there are prediction errors.



Predicted

Actual	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Iris-setosa	50	0	0	50
Iris-versicolor	0	49	1	50
Iris-virginica	0	2	48	50
Σ	50	51	49	150

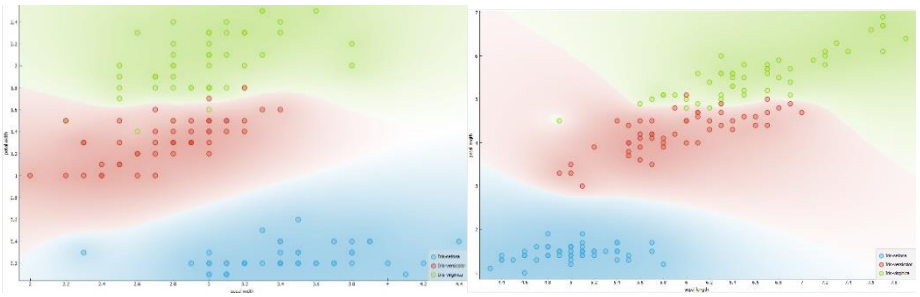
Predicted

Actual	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Iris-setosa	50	0	0	50
Iris-versicolor	0	47	3	50
Iris-virginica	0	1	49	50
Σ	50	48	52	150

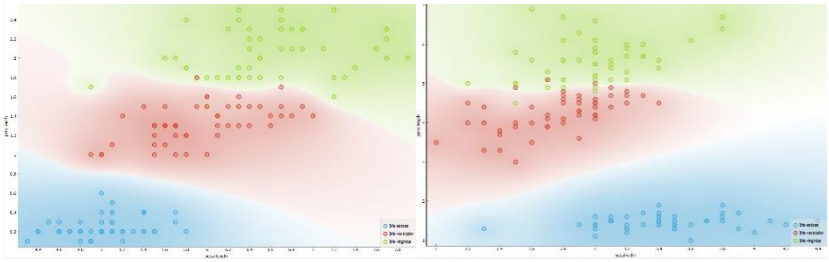
Predicted

Actual	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Iris-setosa	50	0	0	50
Iris-versicolor	0	47	3	50
Iris-virginica	0	1	49	50
Σ	50	48	52	150

(a) Tree (b) RF (c) LR



(a) S-P-W (b) S-P-L



(c) SL-PW (d) SW-PL

Figure 6. the results of testing data processing using ORANGE

Having done testing data using three models classification it can be seen in Table 4. 2. the level of accuracy of the tree, random forest and logistic regression methods.

Table 2. Table of analysis results

	Tree	**Random Forest**	**Logistic Regression**
Precision	0,980	0,981	0,974
AUC	0,993	0,999	0,998
Prediction Error	Iris Sentosa: 0 Iris Versicolor: 1 Iris Virginica: 2	Iris Sentosa: 0 Iris Versicolor: 3 Iris Virginica: 0	Iris Sentosa: 0 Iris Versicolor:3 Iris Virginica: 1

Table 2 shows a table with the value of AUC (Area Under the Curve) with the Tree of 0. 9 93, Random Forest of 0, 999, and to model Logistic Regression of 0.998. AUC accuracy is said to be perfect if the AUC value reaches 1,000 and poor accuracy if the AUC value is below 0.500. while the precision value of the three models used by Random Forest has the highest precision value of 0.981. of test data that has been carried out training and testing can be seen that the accuracy of the testing of Tri model where the model Random Forest is superior as a method for classification iris.

V. CONCLUSION

The conclusions obtained from this study are the Tree Method, Random Forest, and Logistic Regression can be used to detect iris type based on petal length and width, and crown length and width. The use of three classification models to detect iris type provides fairly good results in the AUC (Area Under Curve) value with a Tree of 0.993, Random Forest of 0.999, and for logistic regression model of 0.998. AUC accuracy is said to be perfect when the AUC value reaches 1,000 and the accuracy is poor if the AUC value is below 0.500. While for the precision value of the three models used Random Forest has the highest precision value which is 0.981. From the data tests that have been done training and testing can be seen that the level of accuracy of testing from the three models where the Random Forest model is superior as a method to detect the type of iris flowers.

Based on the results of this study, it can be a contribution to the relevant institutions, but there are several things that the authors can suggest for further research, namely further exploration of iris data using other models that have many features that need to be tried, considering the problem is more complex than just identifying three iris type.

REFERENCES

[1] D. M. C. Hermanto, "Analisis Algoritma Clustering," *J. Media Apl.*, vol. 9, no. 2, pp. 72–84, 2017.

[2] F. Febrianti, M. Hafiyusholeh, and A. H. Asyhar, "Perbandingan Pengklusteran Data Iris Menggunakan Metode K-Means Dan Fuzzy C-Means," *J. Mat. "MANTIK,"* vol. 2, no. 1, p. 7, 2016, doi: 10.15642/mantik.2016.2.1.7-13.

[3] D. Rana, S. P. Jena, S. K. Pradhan, C. Engineering, D. Rana, and S. P. Jena, "Performance Comparison of PCA and LDA with Linear Regression and Random Forest for IRIS Flower Classification," vol. 17, no. 9, pp. 2353–2360, 2020.

[4] B. Taha Chicho, A. Mohsin Abdulazeez, D. Qader Zeebaree, and D. Assad Zebari, "Machine Learning Classifiers Based Classification For IRIS Recognition," *Qubahan Acad. J.*, vol. 1, no. 2, pp. 106–118, 2021, doi: 10.48161/qaj.v1n2a48.

[5] T. Yabuya, T. Imayama, T. Shimomura, R. Urushihara, and M. Yamaguchi, "New types of major anthocyanins detected in Japanese garden iris and its wild forms," *Euphytica*, vol. 118, no. 3, pp. 253–256, 2001, doi: 10.1023/A:1017562518106.

[6] B. E. Turban, J. E. Aronson, and T. Liang, *Decision Support System and Intelegent System*, 7th Ed. Ji. Yogyakarta: Penerbit Andi Yogyakarta, 2005.

[7] D. T. Larose, *Discovering Knowledge in Data An Introduction to Data Mining*. Canada: Published simultaneously in Canada,

2005.

[8] H. N. Ahmad, V. Suhartono, and I. N. Dewi, "Penentuan Tingkat Kelulusan Tepat Waktu Mahasiswa Stmik Subang Menggunakan Algoritma C4.5," *J. Teknol. Inf.*, vol. 13, no. 1, pp. 46–56, 2017.

[9] H. Willa Dhany and F. Izhari, "Analisis Algorithms Support Vector Machine Dengan Naive Bayes Kernel Pada Klasifikasi Data," vol. 6, pp. 595–598, 2019.

[10] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009, [Online]. Available: http://cogprints.org/6708/.

[11] H. Hermanto, A. Mustopa, and A. Y. Kuntoro, "Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 2, pp. 211–220, 2020, doi: 10.33480/jitk.v5i2.1181.

[12] A. Bimantara and T. A. Dina, "Klasifikasi Web Berbahaya Menggunakan Metode Logistic Regression," *Annu. Res. Semin.*, vol. 4, no. 1, pp. 173–177, 2019.

[13] R. D. Tobias, "Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to The Annals of Statistics. ® www.jstor.org," *Ann. Stat.*, vol. 14, no. 2, pp. 590–606, 1986.

[14] D. G. Kleinbaum, *Modeling Strategy Guidelines*. 1994.

[15] S. Y. dan N. emiliyawati Nugroho, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," *J. Tek. Elektro*, vol. 9, no. 1, pp. 24–29, 2017, doi: 10.15294/jte.v9i1.10452.

[16] K. Schouten, F. Frasincar, R. Dekker, and M. Riezebos, "Heracles: A framework for developing and evaluating text mining algorithms," *Expert Syst. Appl.*, vol. 127, pp. 68–84, 2019, doi: 10.1016/j.eswa.2019.03.005.