PENA TEKNIK
JURNAL ILMIAH ILMU - ILMU TEKNIK

Vol. 8, No.1, Maret (2023)
p-ISSN 2502-8952
e-ISSN 2623-2197
https://ojs.unanda.ac.id/index.php/jiit/index

Comparative Performance Results of Classification Algorithm in Data Mining to Identify Types of Glass Based on Refractive Index and Its Elements

[1] Rinto Suppa

[1]*Engineering, Andi Djemma University, Palopo, Indonesia*
E-mail: rintosuppa@gmail.com [1]

Keywords:

Performance evaluation; Classification algorithm; Data mining; Glass

Abstract

Data science is becoming familiar to the public and companies in the era of the Industrial Revolution 4.0. One part of data science is data mining. Data mining is the process of collecting information to see patterns from very large datasets and data discovery which is processed in such a way as to become knowledge based on the interpretation of the information obtained. This paper aims to compare the performance evaluation results of several classification algorithms in data mining (such as DT C-45, Neural Network, KNN, LDA, Naïve Bayes, SVM, and Rule Induction) for identifying types of glass based on its elements and the refractive index. The dataset used is a glass identification dataset from the *UCI Machine Learning Repository*. The results of the evaluation can be seen from the criteria like Accuracy and Kappa using *10-fold-cross validation*. As a result, the K-Nearest Neighbors (KNN) algorithm has the best Accuracy and Kappa values, namely 72.90% for Accuracy and 0.632 for Kappa values. To determine the significance of the accuracy value, the T-Test method is used.

1. INTRODUCTION

One of the most frequently used data mining functions is classification. The Classification method is *supervised learning* where the main objective is to find a model or characteristic of a class from a dataset and can predict or identify test data into a particular label or class (Larose, D. T., & Larose, C. D.2014). The classification algorithm that is often used in the classification process is a Decision Tree, Support Vector Machine (SVM),Neural Network (NN), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Linear Discriminant Analysis (LDA), Rule Induction (RI), and so on.

Each classification algorithm must have its advantages and disadvantages. Therefore, one algorithm cannot be certain that it is better than another. It may be that an algorithm is very suitable for a particular dataset but very bad for another dataset. For that, it is necessary to evaluate the performance of each algorithm in a dataset so that it can be concluded that the algorithm is best applied to the dataset.

In this article, the dataset used is the glass identification dataset taken from the *UCI Machine Learning Repository*. The importance of developing a classification method that can assist in identifying glass types effectively and efficiently. This dataset contains seven *types of glass* whose formation process is influenced by the *refractive index* and the composition of the elements, namely *Sodium (NA), Magnesium (MG), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba), and Iron (FE)*. For the *types of glass* variable, there are seven, namely *(1) vehicle windows non float processed, (2) vehicle windows float processed, (3) Building windows float processed, (4 building windows non float processed), (5) headlamps, (6)) tableware, (7) containers*. Therefore, seen from the existing attributes, it is very suitable to carry out the classification process for this dataset where the variable type of glass is used as the label/class while the refractive index variable and the composition of the elements are used as attributes.

There have been many classification algorithms developed, but it is still unclear which algorithm is the most effective and efficient in classifying glass types. Therefore, this study was conducted to compare the performance of several classification algorithms on glass identification datasets. By knowing which

classification algorithm is the most optimal, it is hoped that it can help improve the accuracy and efficiency of identifying glass types better. In addition, this study also aims to increase understanding of the use of classification algorithms in data processing and provide information about the performance of the classification model used in this study. Thus, it is hoped that the results of this research can be useful for the development of better classification methods in the future (Awopetu, O. A., & Olugbara, O. A. 2016).

The significances of this research are:

1. Improving the quality of glass products: With a more accurate and efficient classification algorithm, it is hoped that it can assist glass manufacturers in better-identifying types of glass to improve the quality of glass products.
2. Prevent accidents: Using the wrong type of glass can cause accidents, therefore a better classification algorithm is expected to prevent accidents caused by using the wrong type of glass.
3. Development of classification methods: This research can contribute to the development of classification methods in the field of data processing.
4. Development of technology: The results of this research can also be applied to other fields of technology such as the development of object recognition systems, facial recognition, and others that require classification methods.
5. Adding insight: The results of this study can also increase understanding and knowledge of the classification algorithm on the glass identification dataset and the performance evaluation technique of the classification algorithm.

A previous study using a *glass identification dataset* conducted by Mus Mulyadi Baharuddin et all, only used the KNN algorithm to identify the type of glass and the results reached 64% accuracy, 63% precision, 71% recall, and 67% F-Measure where the K value the best is at K = 3 (Baharuddin, M. M., Azis, H., and Hasanuddin, T.2019). Research conducted by 1. E. Karlik and O. Baykan (2009) resulted that SVM can classify glass types properly. And for research conducted by A. Kumar and A. Kumar (2014), it was found that the KNN algorithm can identify the type of glass based on its constituent elements well, it's just that it is often wrong to identify the type of float glass by reading window glass. Research conducted by S. Sultana and S. Islam (2014) shows that the decision tree can be used to classify the type of glass from the Glass Identification dataset well.

For research on the comparison of classification algorithms that was carried out by Pandito Dewa Putra (Putra, P. D., & Rini, D. P.2020) in predicting heart disease, the best algorithm in the dataset used was the Naïve Bayes algorithm with an accuracy of 84.07%(Putra, P. D., & Rini, D. P. 2020). In addition, Kawsar Ahmed et al. (Ahmed, K., & Jesmin, T.2014)compared the performance of algorithms to predict diabetes, it was found that the best classification algorithm was the Bagging algorithm with an accuracy value of 85.2308%.

Based on the background above, this research aims is to compare the performance of several classification algorithms in identifying types of glass based on their refractive index and elemental composition.

2. METHODS

The application used in this study is Ms. Excel 2016 for collecting data and preprocessing data and RapidMiner Studio Version 7.6 (update 2022) for data processing such as modeling, validation, and significance test.

The flow chart in this study is

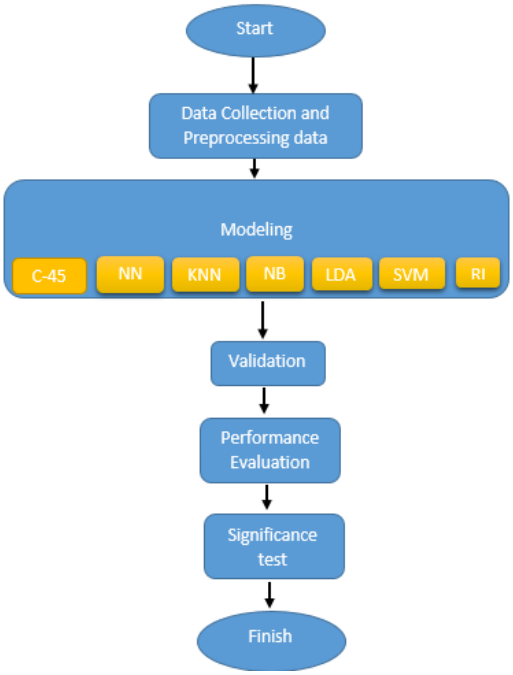


Figure 1. Flowchart

2.1 Data Collection and Preprocessing Data

The dataset used in this study was obtained from the *UCI Machine Learning Repository*, namely the glass identification dataset. The UCI Machine Learning Repository is an online repository that provides various sets of data that can be used for machine learning purposes. This repository is maintained by the University of California, Irvine (UCI) and provides a variety of data types, including numeric, textual, and image data. These data have been processed and organized so that they can be used for various purposes, such as developing prediction and classification models, data analysis, and developing artificial intelligence applications. Data sets in the UCI Machine Learning Repository are frequently used in academic and industrial research and machine learning competitions. (Kelleher, J. D., Tierney, B., & Tierney, B.2018)

All processes in this study started from data input to significance testing using the Rapidminer Studio 7.6 software. In this dataset, there are 10 variables, namely *type of glass*, *refractive index*, and composition of its elements, *Sodium (NA), Magnesium (MG), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba), and Iron (FE)*. For the *types of glass* variable, there are seven, namely *(1) vehicle windows non float processed, (2) vehicle windows float processed, (3) Building windows float processed, (4 building windows non float processed), (5) headlamps, (6)) tableware, (7) containers*.

For the type of data, only the *type of glass* variable is polynomial, while the other is real (numeric). Therefore, if we look at the data contained in this dataset, it is very suitable for a classification process, where the *type of glass* is the *label/class*, while the refractive index and elemental composition are *attributes*.



Name	Type	Missing	Filter (10 / 10 attributes):		
Label **Tipe**	Polynominal	0	Least 6 (9)	Most 2 (76)	Values 2 (76), 1 (70), ...[4 more]
RI	Real	0	Min 1.511	Max 1.534	Average 1.518
Na	Real	0	Min 10.730	Max 17.380	Average 13.408
Mg	Real	0	Min 0	Max 4.490	Average 2.685
Al	Real	0	Min 0.290	Max 3.500	Average 1.445
Si	Real	0	Min 69.810	Max 75.410	Average 72.651
K	Real	0	Min 0	Max 6.210	Average 0.497
Ca	Real	0	Min 5.430	Max 16.190	Average 8.957
Ba	Real	0	Min 0	Max 3.150	Average 0.175
Fe	Real	0	Min 0	Max 0.510	Average 0.057

Figure 2 Statistics from the *glass identification* dataset

By looking at Figure 2, it is very clear that there is no missing data and all data looks normal so that it can be continued to the next process, namely modeling.

2.2 Modeling

There are seven classification algorithms chosen for the modeling process in this study, namely: Decision Tree, Support Vector Machine (SVM),Neural Network (NN), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Linear Discriminant Analysis (LDA), Rule Induction (RI),. The seven algorithms were chosen because they are the most known algorithms so far for the classification process.

Each classification algorithm must have its advantages and disadvantages. Therefore, one algorithm cannot be certain that it is better than another. It may be that an algorithm is very suitable for a particular dataset but very bad for another dataset. For that, it is necessary to evaluate the performance of each algorithm in a dataset so that it can be concluded that the algorithm is best applied to the dataset.

2.2.1 Decision Tree C-45

Decision Tree C4.5 is the development version of Decision Tree ID3 initiated by Quinlan (Quinlan, J.R. 1996). The strong point is the Decision Tree Algorithm. C-4.5 is capable of dealing with features with numeric types, performing decision tree *pruning*, and obtaining rule sets. However, Algorithm C-4.5 also still relies on the Gain criteria, namely the Gain Ratio to determine the features that break up the nodes in the induced tree (Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., and DATA, M.2005)

The process of the Decision Tree C4.5 induction is as follows (Gorunescu, F.2011):

1. The initial process, namely determining Entropy (E) for the root node on the class/label composition

$$E(s) = -\sum_{i=1}^{m} p(w_1|s)\, log_2 p(w_1|s) \quad (1)$$

   Where $p(w_1|s)$ is the proportion of the ith class in all processed data at node s. $p(w_1|s)$ is obtained from the sum of all rows of data labeled class I divided by the number of rows of all data, while m is the number of different values in the data.
2. Furthermore, for numeric attributes, the best 'v' value must be determined to be used as a solver value by calculating its *Gain value*. The value with the highest Gain will be used as the Solver value for each attribute of the numeric type

$$G(s,j) = E(s) - \sum_{i=1}^{n} p(v_1|s)\, x\, E(s_i) \quad (2)$$

Where $p(v_1|s)$ is the proportion of values of v that appear in the class in the node. $E(s_i)$ is the entropy of composition of the value v from the *j-th* class in the *i-th* data node and n is the number of different values in that node.

3. Then calculate the *Entropy* of each attribute against the class/label, then calculate the *Gain* for each attribute. The attribute that has the highest Gain value will be used as the root node.
4. Next, the split position is a count for the attribute that has become the root node by counting the *gain ratio*. To count the *gain ratio*, first, count the *split info*.

$$Splitinfo(s,j) = -\sum_{i=1}^{k} p(v_1|s)\, log_2 p(v_1|s) \quad (3)$$

$$GainRasio(s,j) = \frac{G(s,j)}{Splitinfo(s,j)} \quad (4)$$

Where *k* represents the number of fractions, in the same way in Eq. 4, we get the *gain ratio* for the other options. The highest *gain ratio* will be the next branch node under the *root node.*

5. For the next node, repeat steps 1-4 by calculating the *Entropy*, *Gain*, and *Gain ratio* using the remaining data from the node process above.

Continue recursively on each of the next branch nodes by repeating steps 1-5 until all data in each node provides only one *label class*. The nodes that cannot be broken down are called the *leaf nodes* containing the decisions (*label classes*).

Excess of Decision Tree C-45:

1. Easy to understand and interpret by humans because the resulting model is in the form of a decision tree that can be easily explained.
2. Able to handle numeric and categorical data.
3. C-45 can handle incomplete data by deleting data or using the average value of the missing attribute.
4. C-45 can be used to build more complex models by combining multiple decision trees.

Lack of Decision Tree C-45:

1. Overfitting: The C-45 can produce very large and complex decision trees, which have the potential to overfit the training data.
2. Sensitive to noise: The C-45 is sensitive to noise in data, which may affect model performance.
3. Limitations in dealing with continuous problems: C-45 cannot deal with continuous problems effectively without preprocessing first.
4. Difficulty in handling minority classes: C-45 tends to produce decision trees that ignore minority classes, which can reduce the model's performance in classifying unequal data.
5. Limitations in handling linear relationships between features: C-45 cannot handle linear relationships between features well and requires more complex algorithms to solve them.

2.2.2 Neural Network (NN)

The Neural Network (NN) algorithm is a type of machine learning algorithm that is used to classify data. NN works by adopting the workings of the human brain which processes information in a parallel and distributed way. (E. Prasetyo, 2014).

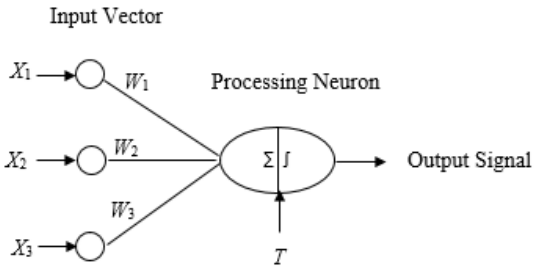


Figure 3. NN design in general

From Figure 3, The NN algorithm for classification consists of several layers consisting of neurons (x). The neurons are connected by a weight that can be adjusted during the model training process. The Activation Function (T) is a mathematical function that is used for each neuron to produce output based on the input received and the weights set. There are three types of layers used in NN, namely:

1. Input Layer: The first layer on the NN functions to receive input from the input data.
2. Hidden Layer: The middle layer in the NN which functions to carry out the transformation process on the input received from the previous layer.
3. Output Layer: The last layer in the NN which produces output in the form of class predictions from the data provided.

The advantages of the Neural Network (NN) algorithm for classification (Brownlee, J.2021):

1. Ability to process complex and unstructured data: NN can handle unstructured data such as images, text or sound well, making it very suitable for classification tasks on complex data.
2. Ability to classify large data: Compared to traditional classification methods, NN can process larger data more quickly and accurately.
3. Ability to learn independently: NNs can learn on their own from the data provided and improve the performance of their models automatically.
4. Ability to solve multicollinearity problems: NN can solve multicollinearity problems in data which means the independent variables have a correlate with each other.

Disadvantages of the Neural Network (NN) algorithm for classification (Li, M., Deng, X., Du, X., Hu, J., & Yang, J. 2020):

1. Algorithm complexity: The NN algorithm is complex algorithm and requires considerable time and resources to learn, implement and execute.
2. Reliance on training data: NNs require large and representative training data in order to learn and improve their models properly. If the data is unrepresentative and incomplete, the performance of the Neural Network model may decrease.
3. Overfitting: NNs can overfit if there are too many trainers with training data, affecting model performance on data that has never been seen before.
4. Difficulty in the interpretation of results: The results of the NN algorithm can be difficult to interpret and explain due to the complexity of the algorithm. It becomes a challenge for the user to explain the results produced by the NN model to others.

2.2.3 K-Nearest Neighbor (KNN)

The K-Nearest Neighbors (KNN) algorithm is one of the machine learning algorithms used for classification or regression. In classification, the goal is to predict the class of data based on the classes in the training data. (Gorunescu, F.2011).

The KNN algorithm works by calculating the distance between test data and existing training data. The training data that has the closest distance to the test data will be used to predict the class of the test data.

KNN assumes that data that are closer together tend to have similar characteristics. Therefore, KNN is often referred to as the "nearest neighbor" method. (Kelleher, J. D., Tierney, B., & Tierney, B. 2018)

The KNN algorithm process for classification can be explained as follows:

1. Determines the value of K (number of nearest neighbors) be used.
2. Calculate the distance between the test data and all training data using distance metrics such as Euclidean distance, Manhattan distance, or Minkowski distance.
3. Take the K training data that has the closest distance to the test data.
4. Voting to determine the class of test data. If K = 1, then the class of the test data is taken from the training data class that has the shortest distance. If K > 1, then the class of the test data is determined based on the majority of the classes from the closest K training data.
5. Returns the class predicted from the test data.

The advantages of the K-NN for classification are:

1. KNN can be used for multiclass classification.
2. KNN can be used on unstructured data, such as image data or text.
3. KNN can be used on small datasets or with a small number of features.
4. KNN is easy to understand and implement.

However, there are several drawbacks to the KNN algorithm, including:

1. KNN requires a long computational time to calculate the distance between data during training, especially on large datasets.
2. KNN requires optimal K parameter settings to produce accurate results. Incorrect setting of the K parameter may result in underfitting or overfitting.
3. KNN is susceptible to data that contains noise or outliers, which can affect classification significantly.
4. KNN cannot generalize to data that has never been seen before. This means that KNN cannot be used to make accurate predictions on data that has never been seen before. (Yadav, S. K., Kumar, P., & Singh, V. K.2021)

2.2.4 Naïve Bayes (NB)

The Naive Bayes algorithm is a probabilistic classification method based on Bayes' theorem. This algorithm is often used in text processing and data classification, especially in email spam classification and sentiment analysis.

Basically, the Naive Bayes algorithm works by calculating the probability that input data belongs to a certain class based on the frequency of occurrence of certain features in the data. (Zhang, H., & Liu, Y. 2021).

Mathematically, to carry out a classification using the Naive Bayes algorithm, it is first necessary to calculate the posterior probability, namely the class probability (y) for the given features (x). Then, the class with the highest probability will be selected as the output class.

The Naive Bayes algorithm is considered "naive" because it assumes that the features in the input data are independent of one another. Although this assumption is often not met in real data, the Naive Bayes algorithm can still provide good classification results in many cases.

The Naive Bayes algorithm can be used in various applications, such as text classification, image classification, and numeric data classification. The advantages of the Naive Bayes algorithm are that it is simple, fast, and can be used for large datasets. However, the downside is the assumed independence between attributes, which can result in lower accuracy in some cases. (Raschka, S., & Mirjalili, V.2021)

2.2.5 Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysis (LDA) algorithm is a data analysis technique used to project multi-dimensional data into a lower-dimensional space for classification purposes. LDA tries to find a line or linear plane that can distinguish between different classes of data. (Zhang, Z., Chen, J., Wang, J., & Li, C. 2022)

The general steps in LDA are as follows:

1. Compute the mean and covariance matrices of each class.
2. Calculate the scatter matrix between classes and the scatter matrix within classes.
3. Calculate the Fisher matrix which is the inverse of the within-class scatter matrix multiplied by the inter-class scatter matrix.
4. Calculate the eigenvalues and eigenvectors of the Fisher matrix.
5. Sort the eigenvectors based on their associated eigenvalues and take the top few eigenvectors to form a transformation matrix.
6. Use a transformation matrix to project data into a lower-dimensional space.
7. Classify the projected data using simple classification algorithms such as k-nearest neighbor (KNN) or logistic regression.

Overall, LDA is a powerful algorithm for data classification with well-defined features and a sizeable amount of data. However, LDA can become less effective when there is a lot of noise or when there are too few data samples in the training set. (Liu, B., Zhu, Y., & Zhang, L.2021)

2.2.6 Support Vector Machine (SVM)

The SVM (Support Vector Machine) algorithm is a machine learning method used for classification and regression. SVM works by building a model that can separate two different data classes with a hyperplane (hyper field) that has the largest margin (distance) between these classes.. (Javed, K., Gani, A., & Khan, S.U.2021)

Following are the general steps of the SVM algorithm for classification:

1. Entering data and labels into the SVM model.
2. SVM will find the best hyperplane to separate the two classes. This hyperplane will be selected based on the largest margin between the two classes, thereby minimizing misclassification.
3. Then, SVM will take into account several support vectors, namely data that is closest to the hyperplane. This support-vector will affect the location of the hyperplane and the resulting margins.
4. Once the hyperplane and margins are found, SVM will classify the new data by entering it into the model and seeing where it is relative to the hyperplane. The new data will be classified into the appropriate class according to its position relative to the hyperplane. (Figueiredo, M.A., & Gama, J.2021)

The advantages of SVM in the classification of data mining are as follows:

1. Able to solve classification problems on complex data with high dimensions.
2. Very effective in overfitting the training data.
3. Able to handle cases of unbalanced data (imbalanced data) well.
4. SVM has strength in separating contiguous or contiguous classes.
5. Can use different types of kernels to project data to different feature spaces.

Disadvantages of SVM in the classification of data mining, namely:

1. Not effective in dealing with non-linear data without using the right kernel.
2. It takes a long time to train models on large data.
3. Very sensitive to parameter settings, such as kernel parameters and C parameters.
4. SVM tends to require more memory and computation compared to other classification algorithms. (Deng, Y., Qian, Y., Chen, Y., & Xie, C.2021)

2.2.7 Rule Induction(RI)

The Rule Induction Algorithm is one of the methods used in machine learning to classify data. This algorithm focuses on creating rules that can be used to distinguish between different classes or categories in the dataset.

The main process of the Rule Induction algorithm is to look for significant patterns in the dataset and create rules based on these patterns. In this stage, the algorithm will group data based on relevant features and find recurring patterns. . (Tan, M., & Li, J. 2021)

After the patterns are found, the algorithm will generate rules that can be used to classify data that have not been seen before. These rules can be in the form of a conclusion-condition set (IF-THEN) or a set of rule sets. The rules generated by the Rule Induction algorithm will continue to be refined and evaluated using techniques such as pruning and cross-validation to improve classification accuracy. This algorithm can be used in various types of data, such as numeric, categorical, and incomplete data. (Guo, H., Huang, X., & Sun, X.2021)

The advantages and disadvantages of the RI algorithm are:

Advantages:

1. Easy to interpret and understand. The resulting rules can be used to explain the relationship between the attributes in the data and the classes contained in the data.
2. Efficient in data processing with a small number of attributes and a relatively small amount of data.
3. Able to capture complex patterns in data and create rules based on these patterns.
4. It can be used in various types of classification problems in Data Mining (such as binary and multiclass classification).

And Disadvantages:

1. Less effective in dealing with data that has a very large number of attributes or too much data.
2. The resulting rules tend to be rigid and inflexible in dealing with data variations that may occur in the future.
3. The ability to find patterns and make rules depends on the quality of the training data used.
4. Dependence on the algorithm used to find patterns in the training data. The algorithm used can have weaknesses and affect the quality of the generated rules.

2.3 Validation

In evaluating the performance of a data mining method, a systematic way must be needed (Gorunescu, F.2011). For this reason, data validation is needed to determine the best type of learning scheme to be used, based on training data to train learning schemes to maximize data usage (Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., and DATA, M.2005). In this study, the technique used to validate is 10-cross-fold validation. This means that the test is carried out 10 times because the existing data is divided into 10 parts. Of the 10 sections, each section will make 9% training data and 1% for test data. The measurement result is the average value of the 10 times the test.

2.4 Performance Evaluation

Evaluation of the performance of the classification model is based on testing true and false objects (Wu, X., & Kumar, V.2009). In this study, the results of performance evaluation can be seen from the accuracy and Kappa value. Accuracy is the percentage of accuracy in predicting or identifying test data, while Kappa is the consistency of the assessment by considering all aspects in the matrix.

2.5 Significance Test

The significance test aims to test how far the performance evaluation results differ from each classification algorithm. If it is not significant, it means that the performance evaluation of these algorithms is practically the same. In this study, the significance test method applied was the T-Test. (Sawyer, T. L., & Kotrlik, J. W.2021).

3. RESULT AND DISCUSSION

This section discusses analysis results that can be displayed in a diagram, figure, table, or other illustration which is easy to understand and communicate.

The purpose of this research is to compare performance evaluations in the form of accuracy and Kappa of each classification algorithm model used to identify types of glass based on the refractive index and composition of the elements contained in the glass identification dataset. The results can be read in Table 1 below

Table 1. Accuracy and Kappa Values

	Accuracy	**Kappa**
C45	67,29 %	0,544
KNN	72,90 %	0,632
NN	69,16 %	0,576
LDA	59,35 %	0,425
NB	45,79 %	0,291
SVM	46,26 %	0,207
RI	69,16 %	0,576

When viewed from table 1, it is very clear, the algorithm that has the best Accuracy and Kappa values is the KNN algorithm with an Accuracy value of 72.90% and Kappa of 0.632. Meanwhile, the algorithm with the lowest performance evaluation is the NB algorithm with an accuracy of 45.79% and Kappa of 0.291. However, to ascertain whether an algorithm model can be said to be better than other algorithm models in a dataset, a significance test is needed. If the results are significant, then it can be said that the algorithm is better or worse than other algorithms, if the results are not significant, then it can be said that there is no difference between one algorithm and another algorithm in its use (Shrestha, R., et al. 2019). The results of the significance test with the T-Test method in this study can be read in the following table

Table 2. The results of the significance test using the T-Test

	C-45	**KNN**	**NN**	**LDA**	**NB**	**SVM**	**RI**
C-45		0,068	0,662	0,023	0	0	0,696
KNN			0,393	0	0	0	0,417
NN				0,037	0	0	0,988
LDA					0,002	0	0,052
NB						0,862	0
SVM							0
RI							

According to statistics, if the T-Test value is more than 0.05, it can be concluded that there is no significance between the two algorithms compared. In table 2, for yellow cells, the value is less than 0.05 and for white ones, the value is greater than 0.05. This describes to us that between the C-45, KNN, NN, and RI algorithms, the performance evaluation can be said to be no different because the results are not significant. As for the Naive Bayes and SVM algorithms, it can also be seen that the performance between these two algorithms is the worst used in this dataset.

4. CONCLUSIONS

Based on the discussion above, it can be concluded:

1. The results of the performance evaluation to identify the type of glass based on the refractive index and the composition of its elements on the *glass identification dataset*, the largest of which is the KNN algorithm with an accuracy of 72.90% and a Kappa of 0.632, and the smallest is owned by NB algorithm with an accuracy of 45.79 % and Kappa 0.291
2. Based on the significance test, the C-45, KNN, NN, and RI algorithms are best used, while the NB and SVM algorithms are the worst to be applied as classification models on the *glass identification dataset*.
3. It is necessary to develop or improve existing algorithms to improve performance in classifying glass types.
4. With a more accurate and efficient classification algorithm, it is hoped that it can assist glass manufacturers in better-identifying types of glass to improve the quality of glass products.

5. REFERENCES

Ahmed, K., & Jesmin, T. (2014). Comparative analysis of data mining classification algorithms in type-2 diabetes prediction data using weka approach. *International Journal of Science and Engineering*, *7*(2), 155-160.

A. Kumar and A. Kumar. 2014.Glass Identification Using K-Nearest Neighbor (KNN) Algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 7.

Awopetu, O. A., & Olugbara, O. A. (2016). Fuzzy-logic based decision support system for the classification of glass samples. *Measurement*, 94, 218-226.

Brownlee, J. 2021. Neural Network Architectures for Beginners. Machine Learning Mastery.

Baharuddin, M. M., Azis, H., & Hasanuddin, T. (2019). Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca. *ILKOM Jurnal Ilmiah*, *11*(3), 269-274.

Cohen, W. (2015). Fast Effective Rule Induction Machine Learning. In *the Proceedings of the 12th International Conference*.

Deng, Y., Qian, Y., Chen, Y., & Xie, C. 2021. A survey of support vector machine: Concepts, applications, and challenges. *IEEE Access*, 9, 18379-18407.

E. Karlik and O. Baykan. 2009.Feature Selection and Classification of Glass Types Using Support Vector Machines." Expert Systems with Applications, vol. 36, no. 2,

E. Prasetyo. (2014).Data Mining – Mengolah Data Menjadi Informasi menggunakan Matlab. Yogyakarta: Andi

Figueiredo, M. A., & Gama, J. 2021. Data imbalance and concept drift in support vector machines: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(2), e1401.

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media.

Guo, H., Huang, X., & Sun, X. 2021. An Enhanced Rule Induction Algorithm for Data Classification. *Journal of Intelligent & Fuzzy Systems*, 41(2), 2467-2478.

Javed, K., Gani, A., & Khan, S. U. 2021. A comprehensive review of support vector machines for data classification. *Journal of Ambient Intelligence and Humanized Computing*, 12(1), 101-125.

Kelleher, J. D., Tierney, B., & Tierney, B. 2018. *Data science: an introduction*. CRC press.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons.

Li, M., Deng, X., Du, X., Hu, J., & Yang, J. 2020. Recent Advances in Convolutional Neural Networks. *IEEE Access*, 8, 191298-191324.

Liu, B., Zhu, Y., & Zhang, L. 2021. A novel hybrid linear discriminant analysis approach based on manifold learning for high-dimensional data classification. *Knowledge-Based Systems*, 222, 106956.

Putra, P. D., & Rini, D. P. (2020, February). Prediksi Penyakit Jantung dengan Algoritma Klasifikasi. In *Annual Research Seminar (ARS)* (Vol. 5, No. 1, pp. 95-99).

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71-72.

Raschka, S., & Mirjalili, V. 2021. *Python Machine Learning*, Third Edition. Packt Publishing.

Sawyer, T. L., & Kotrlik, J. W. 2021. Basic statistical concepts for single-group and two-group research designs. *Journal of Agricultural Education*, 62(2), 79-95.

Shrestha, R., et al. 2019. Comparative analysis of classification algorithms: Naïve Bayes, logistic regression, decision tree, random forest, gradient boosting, and XGBoost. Journal of Big Data, 6(1), 1-26.

S. Sultana and S. Islam. 2014. Glass Identification Using Decision Tree. *International Journal of Computer Applications*, vol. 97, no. 22,

Tan, M., & Li, J. 2021. Improved Rule Induction Algorithm for Multi-Class Classification. *IEEE Access*, 9, 5164-5174.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005, June). Practical machine learning tools and techniques. In *Data Mining* (Vol. 2, No. 4).

Wu, X., & Kumar, V. (Eds.). (2009). *The top ten algorithms in data mining*. CRC press.

Yadav, S. K., Kumar, P., & Singh, V. K. (021. A Survey of k-Nearest Neighbor Methods for Machine Learning. *Journal of Intelligent Systems*, 30(1), 1-20.

Zhang, H., & Liu, Y. 2021. A Novel Naive Bayes Method Based on Mutual Information for Multi-label Classification. *IEEE Access*, 9, 78816-78825.

Zhang, Z., Chen, J., Wang, J., & Li, C. 2022. Multi-source feature learning with linear discriminant analysis for human activity recognition. *Neurocomputing*, 485, 78-89.