AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 12, No. 1, 2023, 1309-1321

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

IMPLEMENTATION OF DECISION TREE AND SUPPORT VECTOR MACHINE ON RAISIN SEED CLASSIFICATION

Wardhani Utami Dewi[1], Khoirin Nisa[2*], Mustofa Usman[3]

[1,2*,3] Matematika, Universitas Lampung, Lampung, Indonesia
**Corresponding author. Jl. Prof. Dr. Ir. Sumantri Brojonegoro No.1, Gedong Meneng, Kec. Rajabasa, Kota Bandar Lampung, Lampung, Indonesia*
E-mail: *dewiutamiwardhani@gmail.com* [1)]
khoirin.nisa@fmipa.unila.ac.id [2)]
usman_alfha@yahoo.com [3)]

Received 27 December 2022; Received in revised form 19 February 2023; Accepted 10 March 2023

Abstract

In everyday life there are many complex and global problems, especially in terms of decision making. The purpose of this study is to apply the Decision Tree (DT) and Support Vector Machine (SVM) methods in classifying raisin seeds into the Besni and Kecimen classes. Then evaluate the model using the accuracy, sensitivity, specificity, and kappa levels. The method used is the Machine Learning (ML) method, namely the DT and SVM algorithms. The data used is secondary data with a total sample of 900 raisins. Algorithm processing is carried out using R-studio 4.2.1 software. The steps in the research, namely the raisin data were divided into training data (70%) and data testing (30%), and the evaluation of the two methods was carried out using data testing. The evaluation results are compared with the accuracy, sensitivity, specifications, and kappa of the DT and SVM algorithms. The results of the classification of the raisin seed data show that the SVM algorithm is superior to DT, so that the number of positive observations is more precise in predictions.

Keywords: Data Mining; Machine Learning; Supervised Learning

Abstrak

Dalam kehidupan sehari-hari terdapat banyak permasalahan yang kompleks dan global, terutama dalam hal pengambilan keputusan. Tujuan dari penelitian ini yaitu menerapkan metode Decision Tree (DT) dan Support Vector Machine (SVM) dalam mengklasifikasikan biji kismis ke dalam kelas Besni dan Kecimen. Kemudian melakukan evaluasi model menggunakan tingkat akurasi, sensitivitas, spesifisitas, dan kappa. Metode yang digunakan adalah metode Machine Learning (ML), yaitu algoritma DT dan SVM. Data yang digunakan adalah data sekunder dengan jumlah sampel 900 biji kismis. Proses algoritma dilakukan dengan menggunakan software R-studio 4.2.1. Langkah-langkah dalam penelitian yaitu data kismis dibagi menjadi data pelatihan (70%) dan data pengujian (30%), dan evaluasi kedua metode tersebut dilakukan dengan menggunakan data pengujian. Hasil evaluasi dibandingkan berdasarkan tingkat akurasi, sensitivitas, spesifisitas, dan kappa dari algoritma DT dan SVM. Hasil klasifikasi data biji kismis menunjukkan bahwa algoritma SVM lebih unggul dari DT, sehingga jumlah pengamatan positif lebih tepat dalam prediksi.

Kata kunci*: Data Mining; Machine Learning; Supervised Learning*

INTRODUCTION

In everyday life, humans always encounter problems in making systems and decisions. ML is interdisciplinary in a wide range of fields, building on the foundational concepts of computer science, statistics, mathematics, and many other disciplines (Soofi & Awan,

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

2017). ML has achieved great success in a variety of applications, particularly in supervised learning tasks such as classification and regression. In ML, typically, a predictive model that optimizes for a specific objective is learned from a set of training data, each associated with an event or object. ML is supervised learning and unsupervised learning are based on whether artificially assigned labels are present or absent (Su & Chiang, 2022).

Supervised learning involves predicting the response variable given the observed variable (Bzdok et al., 2018). Several types of supervised machine learning-based algorithms, namely, DT, logistic classification, random forest, SVM, and others (Mishra & Dasgupta, 2022). According to Ashoka et al. (2020); Charbuty & Abdulazeez (2021), DT is a common method for classification data in ML. The most significant feature of DT is its ability to transform a complex decision-making problem into a simple process, thus finding solutions that are understandable and easier to interpret (Gkikas et al., 2022). With this DT algorithm, it will be seen the things that affect the problem. Problems that are centralized and complex are broken down into specifics so that the cause of the problem is found and a solution can be found to achieve the goals of the system maker by reducing the risk of errors (Tariq et al., 2022).

Another algorithm of the supervised learning methods is SVM. This algorithm was first introduced by Vapnik, which showed its effectiveness (Reddy, 2021). Ma et al. (2018) is a binary classification model that creates a dividing hyperplane to minimize risk by maximizing the margin between data samples. Nedaie & Najafi (2018) SVM may not be the best choice for large datasets and the number of features used in classification can also impact its performance. The number of features can affect the classifier’s performance (Fachrurrozi et al., 2021).

Previous research by Imran et al. (2022) chose a better classifier using machine learning for COVID-19. The result of the research is that the model developed by DT is the most efficient classifier, with the highest accuracy percentage of 99.85%. Another related research was a comparative study of machine learning classifiers for modeling road traffic accidents by Bokaba et al. (2022) with the result that random forest combined with several imputations produced the best performance when compared to other combinations (Bokaba et al., 2022). Classification of raisin seeds using machine vision and artificial intelligence methods were previously done by Ilkay et al. (2020), namely using logistic regression, multilayer perceptron and SVM, while for evaluating the model they only used the level of accuracy.

Regarding those previous studies mentioned above, no one has applied the DT method for the raisin seed classification and they only used measurement statistics namely the level of accuracy in evaluating the model. Therefore, in this study we used DT method for raisin seed classification and compare it with the result using SVM. In addition, for model assessment we uses some evaluation measures instead of only their accuracy, i.e. the sensitivity, specificity and the kappa value. Therefore the most accurate method in classifying raisin seeds with some existing features can be obtained.

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

METHOD

The method used in this study is the DT C.50 algorithm and conventional SVM algorithms, and in this study the computation was done using *R-studio 4.2.1* software. The research was conducted at Computational Laboratory, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung, in 2022.

The raisin data were obtain from https://archive.ics.uci.edu/ml/datasets/Raisin+Dataset which was previously studied by Ilkay et al. (2020), they conducted data extraction through image acquisition and image processing to obtain morphological and shape characteristics. During the process of extracting features, several inferences about the features were made for each raisin identified in the image. Feature extraction was done by the feature morphology process. The study used a sample of 900 raisins, including 450 from both varieties, Besni and Kecimen. As many as 7 morphological characteristics are concluded for each raisin grain in Table 1.

Table 1. Features of the raisin seed dataset

No	Featured	Description
1	areas	Gives the number of pixels in raisin bounds
2	Perimeter	Measures the environment by calculating the distance between the raisin border and the pixels around it
3	Major Axis Length	Provides the length of the principal axis, i.e., the longest line that can be drawn on the raisin
4	Minor Axis Length	Gives the minor axis length, which is the shortest line that can be drawn on the raisin.
5	Eccentricity	gives a measure of elliptical eccentricity, which has the same moment as a raisin.
6	Convex Area	Gives the number of pixels of the smallest convex shell of the region formed by the raisin
7	extensions	Returns the ratio of the area formed by the raisins to the total pixels in the bounding box
8	class	Kecimen and Besni Raisins

The steps for raisin seed data classification using the DT and SVM methods are as follows:

1. Divide the data into two groups, i.e. 70% for training data and 30% for testing data.
2. Built classification model using DT and SVM based on training data
3. Create a confusion matrix for training data
4. Evaluate the model using testing data
5. Compare the level of accuracy of the two models

RESULT AND DISCUSSION

In this study, a model was created using DT and SVM to classify raisin seeds according to their features and compare the accuracy of the two models. Data is divided into two parts, namely data training and data testing. The training data will later be used to train the algorithm to find the appropriate model. At the same time, data testing will be used to test and find out the performance of the model obtained at the testing stage. The researcher determined 70% training data and 30% testing data from 900 raisin seeds, so there were 630 training data and 270 testing data. The classification process was assisted by R studio software. The following is the classification of raisin seed species based on the DT and SVM algorithms.

1. DT

DT is a technique that is widely used in data mining, namely a system that makes classifiers (Charbuty & Abdulazeez, 2021; Mienye et al., 2019)) According to Nikam (2015) and Gavankar & Sawarkar (2017), classification algorithms can process extremely huge amounts of data in data mining. Inside DT, there is entropy and information gain. Entropy is a metric for gauging the unpredictability or impurity of a dataset. The entropy value always lies between 0 and 1. The value is better when it is equal to 0 whereas it is worse when it is equal to 0, i.e., the closer the value is to 0, the better (Chen et al., 2019; Shang et al., 2013). If the targets are G with different feature values, the classification entropy of the set S with respect to c, is shown in equation (1):

$$Entropy(S) = \sum_{i=1}^{c} P_i \log 2^{P_i}, \qquad (1)$$

Where P_i is the ratio of the number of samples of the subset and the value of the th feature i. Information gain, also referred to as mutual information, is one of the segmentation metrics. It is clear from this how much information there is regarding random variable values (Liu et al., 2013). This is the opposite of entropy; the higher the value, the better. $Gain(S, A)$ is defined as follows on the definition of entropy (Taneja et al., 2014), as shown in equation (2).

$$Gain(S, A) = \sum_{V \epsilon V(A)} \frac{|S_V|}{S} Entropy\ (S_V), \qquad (2)$$

where the feature A range is $V(A)$, and Sv is a subset of the set S equal to the feature value of the feature v.

The DT algorithm used in this study is the C5.0 algorithm. Based on the application of the DT method to raisin seed data, with a sample of 630 training data and 8 features, the classification result are shown in Table 2.

Table 2. Summary of DT models

MajorAxisLength	>	422.2791:	Besni	(281/25)
MajorAxisLength	<=	422.2791:		
:...Perimeter	<=	1127.409:	Kecimen	(313/41)
Perimeter	>	1127.409:		
:...Extent	<=	0.7287691:	Besni	(23/7)
Extent	>	0.7287691:	Kecimen	(13/2)

If Major Axis Lengthmore than 422.2791 pixels, then group into the Besni class as many as 281 raisins. Meanwhile, if the Major Axis Length is less than or equal to 422.2791 pixels, then it is classified again based on perimeter features. If the perimeter is less than or equal to 1127.409 pixels, it will be classified into the Kecimen class, namely 313 raisins. On the other hand, the perimeter, which is larger than 112.409 must be seen based on the extent feature. Enter the Besni class if the extent is less than or equal to 0.7287691 pixels totaling 23 and vice versa if the extent is more than 0.7287691 then it is classified into the Kecimen class as many as 13 raisins. Therefore it can be concluded that in making a decision in the form of

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

classifying raisin seeds, it is necessary to pay attention to the magnitude of the Major Axis Length, Perimeter and Extent. More specifically, you are dominant in the Basic class if the Major Axis Length is greater than 422.2791 and the Extent is less than or equal to 0.7287691. On the other hand, entering the Kecimen class can be seen from the size of the perimeter and extent.

More details will be visualized using the DT presented in Figure 1.

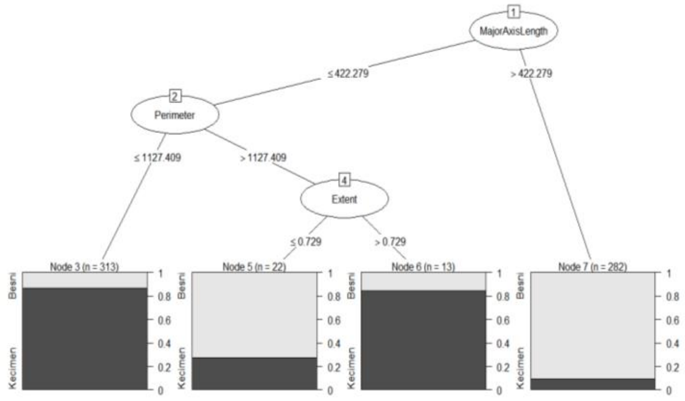


Figure 1. Raisin seed classification decision tree

Figure 1 shows that the model forming 4 trees along with their nodes. So it can be interpreted that there are only three features that influence the classification of raisin seeds into each class, namely Major Axis Length, Perimeter, and Extent. As for the Area, Minor Axis Length, and Convex Area features, they cannot be used as standards in determining the classification. Major Axis Length is the most important feature in this study, which is 100.00% influencing the classification of raisin seeds, followed by Perimeter and Extent.

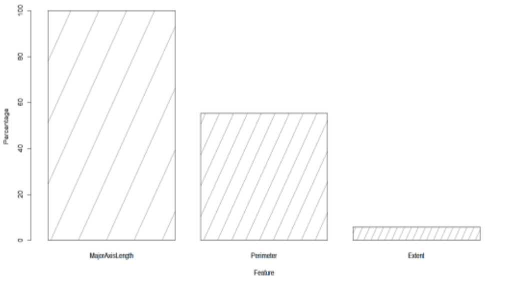


Figure 2. The histogram of the most important features in the classification of raisin seeds

a. DT classification result of training data

After applying the C5.0 model to the training data, the predicted results for the classification of raisin seeds are obtained which are presented in the form of a cross table in Table 3.

Table 3 . Cross table of DT classification of training data

Actual Class	**Predicted Class**		**Row**
	Besni	**Kecimen**	**Totals**
Besni	272	43	315
	0.432	0.068	
Kecimen	32	283	315
	0.051	0.449	
Columns Total	**304**	**326**	**630**

The total training data used was 630 raisin seeds, of which 315 entered the Besni class, and 315 entered the Kecimen class. If we look at the crosstable above, 272 seeds are correctly included in Besni, and the remaining 43 are misclassified; namely, they are included in the Kecimen class. Then, there were 283 seeds that entered the Kecimen class correctly, and the

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

remaining 32 were misclassified into the Besni class. So from a total of 630 training data, 555 were classified correctly into each class, and 75 were misclassified.

b. DT classification result of testing data

To evaluate the model that was developed using the training data, the testing data is used to test the model's performance. The results of using a DT on the testing data are displayed in cross table as shown in Table 4.

Table 4. Cross table DT testing data

Actual Class	**Predicted Class**		**Row**
	Besni	**Kecimen**	**Totals**
Besni	117	18	135
	0.433	0.067	
Kecimen	18	117	135
	0.067	0.433	
Columns Total	**135**	**135**	**270**

Based on the 270 testing data, it showed that there were 36 errors in the classification, namely 18 who should have entered Besni, but were classified in the Kecimen class. Then there were 18 who were supposed to enter the Kecimen class but entered Besni. So the model test predicts correctly that 234 raisin seeds are according to their class classification and 36 are not according to their class species. Broadly speaking, the accuracy level of the model obtained is 0.8667 and the error rate is 0.1333. Apart from the level of accuracy of the model, the following is a test statistic that can describe how good the model is using the DT C5.0 algorithm.

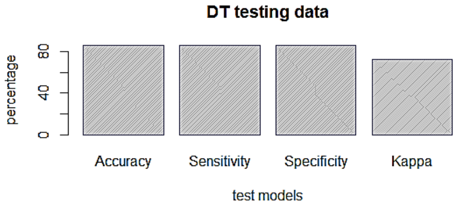


Figure 3. DT testing data

Based on the result of testing the DT model in Table 4 using data testing, it shows that the model is revealing because it can correlate the results with the features in the data used, which is 0.8667 or 86.67%. If we turn to measure the proportion of the number of positive observations that are correctly predicted, it is 86.67% (see the sensitivity level in Figure 3). Conversely, there is specificity, which measures the proportion of the number of negative observations that are correctly predicted, namely 0.8667 or 86.67%. In addition, there is a kappa of 0.7333 or 73.33%, meaning that the rows and columns of the training data are quite appropriate.

2. SVM

The SVM method is in the same class as the Artificial Neural Network **(**ANN) in terms of function and problem conditions that can be solved (Cervantes et al., 2020; Nalepa & Kawulok, 2019). This method is effective at building classifier that can predict labels for one or more feature vectors by creating a decision boundary between two classes (Huang et al., 2018). This closest point is called a support vector. Given a training data set labeled, $(x_1, y_1), \dots, (x_n, y_n), x_i \, \epsilon \, R^d$ and $y_i \, \epsilon \, (-1, +1)$. Where x_i is the feature vector representation and y_i is the class label (negative or positive) from training to i. The optimal hyperplane can then be defined as (Lantz, 2013):

$$\boldsymbol{wx^T} + b = 0,$$

where the weight vector x is the input feature vector and b is the bias. The w and b satisfies the following inequalities for all elements of the training set:

$$\boldsymbol{wx_i^T} + b \geq +1 \text{ if } y_i = 1$$
$$\boldsymbol{wx_i^T} + b \leq -1 \text{ if } y_i = -1,$$

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

The goal of training the SVM model is to find $\boldsymbol{w}$ such that the hyperplanes separate the two classes and maximize the margins $\frac{1}{||w||^2}$. An alternative use for SVM is kernel methods, which allow us to create higher dimensional nonlinear models. Kernel functions can help perform certain computations more quickly that would otherwise require high-dimensional computations. It is defined as b.

$$K(x,y) = < f(x), f(y) >.$$

SVM method is used to identify the best classifier function that can distinguish between two sets of data from two distinct classes. The use of this machine learning technique is because of its convincing performance in predicting classes from new data. This study uses the SVM method to classify raisin seeds into two classes, namely Besni and Kecimen. So that the dependent variable is class. The model was formed using 630 training data. It is known that the SVM model on training data is C-Classification type, with a radial kernel.

a. SVM classification result of training data

The results obtained from the application of the SVM classification method using training data are presented in Table 5 in the form of a cross table.

Table 5. Cross table of SVM classification of training data

Actual Class	Predicted Class		Row
	Besni	**Kecimen**	**Totals**
Besni	262	53	315
	0.416	0.084	
Kecimen	31	284	315
	0.049	0.451	
Columns Total	**293**	**337**	**630**

The results obtained from the SVM classification method using training data, namely 262 seeds or 41.6% were classified correctly into the Besni class and the remaining 53 seeds or 8.4% were misclassified into the Kecimen class. Then with a probability of 0.451 or 284 raisins seeds right into the Kecimen class, the remaining 31 raisins were misclassified.

b. SVM classification result of testing data

The SVM model was tested again using 270 testing data which showed that 111 raisin seeds were rightly included in the Besni class and the remaining 24 were misclassified into the Kecimen class. Furthermore, 128 raisin seeds were correctly indicated as belonging to the Kecimen class, and the remaining 8 raisin seeds were incorrectly classified as belonging to the Besni group. When compared with the results of SVM for training data, the use of SVM in testing data is more accurate, although not very much significant. The results of applying SVM to classify raisin seeds in the testing data presented in crosstable form.

Table 6. Cross table of SVM classification of testing data

Actual Class	Predicted Class		Row
	Besni	**Kecimen**	**Totals**
Besni	111	24	135
	0.411	0.089	
Kecimen	7	128	135
	0.026	0.474	
Columns Total	**118**	**152**	**270**

More specifically, the level of the four accuracy measurements of the model are presented as bar diagram in Figure 4.

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

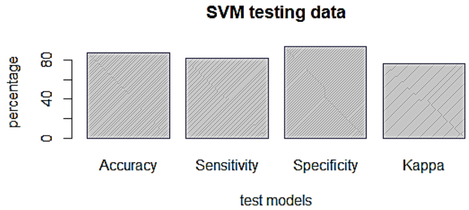
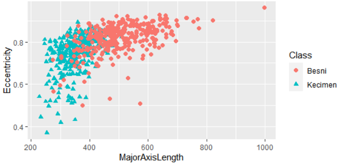
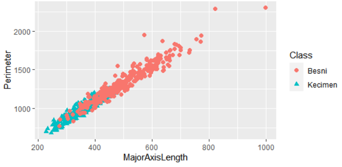
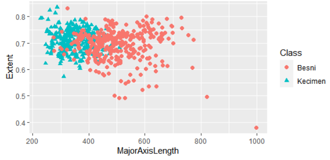
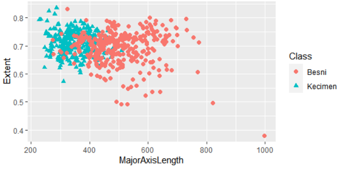
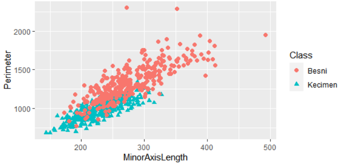
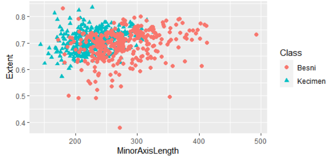
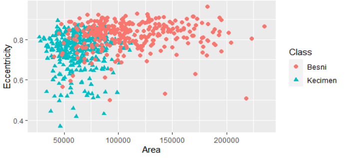
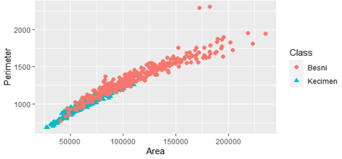
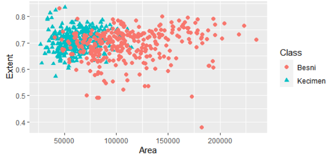


Figure 4. SVM classification result of testing data

Based on the results in Table 6, the classification of testing data using the model is also revealing. The level of accuracy of the prediction results with testing data is 0.8852 or 88.52%, meaning that the model is very good at correlating results with features in the data. Furthermore, the sensitivity value was obtained at 0.8222 or 82.22%, meaning that the proportion of positive observations is right on the prediction. While the specificity value is 0.9481 or 94.81%, indicating the proportion of the number of negative observations that is correct in the prediction. Then for the balance accuracy value of 0.8852 or 88.52%. Used to measure the accuracy of the proportion of the number of positive class observations that are right predictions. In addition, it can also be seen based on the kappa value, which is 0.7704 or 77.04%, which means that there is a good match between the rows and columns.

Table 7. SVM Plots

Feature	Eccentricity	Perimeter	Extensions
Major Axis Length			
Minor Axis Length			
areas			

If raisin seeds are plotted using SVM, even though they are good in use, it is quite difficult to classify them using 2 dimensions because a hyperplane is not formed. Some data still look overlapping. But we can still classify it from the color of the two, the red color indicates that the raisin seeds are in the Besni class and the Tosca color is in the Kecimen class. Only a few features are displayed because the classification results obtained are not too different.

Pay attention to the Eccentricity feature, when it is correlated with the Major Axis Length, Minor Axis Length, and Area, it will form the same classification. The greater the magnitude of these three features, the raisin seeds are classified into the Besni class, and vice versa if the lower the value of the three features is, then it will enter the Kecimen class. This also applies to the Perimeter feature. It is different with the Perimeter feature, it is

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

clear that the plot formed is linear, if the Major Axis Length, Minor Axis Length, and Area are large, then it will be classified as Large, and low will be classified as Kecimen.

Based on the results from DT and SVM, we obtained that both methods are very suitable in classifying raisin seeds into two classes, Besni and Kecimen. The advantages of DT here can explain important features in classifying raisin seeds. The most important feature is Major Axis Length, followed by Perimeter and Extent. At the same time, other features such as Area, Minor Axis Length, and Convex Area cannot be used as a standard in determining the classification. In the model formed from training data and tested using data testing, raisin seeds that are correctly classified with the DT data testing model are 234, and the misclassification errors are 36 seeds. To measure the model, several measurement statistics are used, namely the level of accuracy, sensitivity, specificity, and kappa value. Based on these measurement statistics, the results meet the standards. This means that the use of the DT method in the classification of raisin seeds with 7 features is very good.

The second classification method used is SVM. The working algorithm uses nonlinear mapping to convert the original training data to higher dimensions of the performance of SVM. The algorithm with this linear classification method uses the kernel to be able to handle nonlinear data. The concept of SVM as a classification method is to separate data using a hyperplane. SVM techniques are generally very useful for data with no known distribution. If you already have data labels, SVM can be used to generate one or more separator hyperplanes so that the data is separated into several segments, and each segment contains only one type of data. In the research on the classification of raisin seeds using the SVM data training model, type C-Classification, with radial kernels. Radial kernel or radial basis function (RBF) is a popular kernel function used in various kernel learning algorithms. The kernel parameter has a value of one or can be called Cost, the gamma parameter value is 01.142, and the number of supper vectors is 249. To see the grouping of raisin seeds with the SVM plot is quite good, however, it cannot be described in a 2-dimensional plane because the hyperplane is not formed, so the division of the segments is not very clear. It would be better to see the grouping based on the crosstable of the training data.

After the SVM model was formed based on training data, it was tested using data testing with 239 correct classification results and 31 misclassification errors. So the accuracy level was quite high compared to DT. In accordance with the research objectives, the level of accuracy of the DT and SVM methods will be compared. As seen in the following is a comparison of the measurement statistics presented in Table 6.

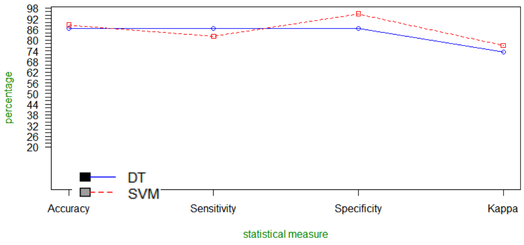


Figure 6. Comparison of accuracy rates

In Figure 6 the blue line indicates DT, and the dotted red line describes SVM. If we look at the level of accuracy of the two methods, the SVM

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

model is more accurate than the DT, which is 88.52%. This means that the use of the SVM algorithm produces a model that is very good at correlating results with features in the data. On the other hand, with sensitivity, the proportion of the number of positive observations is right on the prediction, and the DT algorithm is superior, which is equal to 86.67%. As with accuracy, the SVM algorithm with statistical specificity measures is much better than DT, meaning that 94.81% of the proportion of negative observations (misclassification error) is correct in the prediction. Then the use of Kappa where Kappa accuracy is highly recommended because in calculating its accuracy, it uses all the elements in the confusion matrix. The kappa value in SVM is higher than in DT, so in this study, the best algorithm used to classify raisin seeds into the Besni and Kecimen classes is the SVM algorithm.

DT and SVM methods are two methods that are feasible to use in decision making. But in this case, the DT method sometimes describes illogical results. Unlike the case with SVM, by forming a new hyperplane, it will make it easier to classify objects into two groups. For model evaluation, it is better to use more than one measurement statistic for being able to assess the result from different perspectives.

This is in line with previous research by Khojastehnazhand & Ramezani (2020) that SVM provides more accurate results in terms of classification of raisin seeds. Likewise with the research of Guo et al. (2022), the results of the experiments that have been carried out show that even though MCNN and AlexNet achieve good prediction results, SVM has a better classification effect on raisin skins.

The implication of the results of this study is that the ML method will make it easier to classify raisin seeds into the Besni and Kecimen classes with the existing features. Helping food engineering researcher to be able to use the SVM method as a classification method that has been proven to have a high level of accuracy.

CONCLUSION AND SUGESSTION

The SVM algorithm has a superior accuracy level compared to DT in terms of building a model so as to obtain the right number of positive observations and predictions.

This is indicated by several statistical measurements, namely the accuracy rate of SVM is 88.52% while DT is 86.67%. This means that the SVM algorithm produces a model that is very good at correlating results with features in the data. SVM is much better than DT in terms of specificity, which is 94.81%, the proportion of the number of misclassification errors is right in the prediction. Then for the use of the Kappa accuracy level it is highly recommended because in calculating its accuracy it uses all the elements in the confusion matrix. SVM kappa value of 77.04% is superior to DT of 73.33%.

As for suggestions for further research, that is to be able to compare all classification methods such as Naive Bayes, ANN, DT, and SVM on raisin seed data. Then determine the most accurate method for classifying raisin seeds into the Besni or Kecimen class, with the same model evaluation statistical measurements as this study.

REFERENCE

Ashok, P., Jackermeier, M., Jagtap, P., KÅetínský, J., Weininger, M., & Zamani, M. (2020). DtControl: Decision tree learning algorithms

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

for controller representation. *HSCC 2020 - Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control ,Part of CPS-IoT Week*. https://doi.org/10.1145/3365365.3382220

Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents. In *Applied Sciences* (Vol. 12, Issue 2). https://doi.org/10.3390/app12020828

Bzdok, D., Krzywinski, M., & Altman, N. (2018). Machine learning: supervised methods. *Nature Methods*, *15*(1), 5–6. https://doi.org/10.1038/nmeth.4551

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/https://doi.org/10.1016/j.neucom.2019.10.118

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01 SE-), 20–28. https://doi.org/10.38094/jastt20165

Chen, X., Yang, Z., & Lou, W. (2019). Fault Diagnosis of Rolling Bearing Based on the Permutation Entropy of VMD and Decision Tree. *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, 1911–1915. https://doi.org/10.1109/EITCE47263.2019.9095187

Fachrurrozi, S., Muljono, Shidik, G. F., Fanani, A. Z., Purwanto, & Zami, F. A. (2021). Increasing Accuracy of Support Vector Machine (SVM) By Applying N-Gram and Chi-Square Feature Selection for Text Classification. *2021 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 42–47. https://doi.org/10.1109/iSemantic52711.2021.9573210

Gavankar, S. S., & Sawarkar, S. D. (2017). Eager decision tree. *2017 2nd International Conference for Convergence in Technology (I2CT)*, 837–840. https://doi.org/10.1109/I2CT.2017.8226246

Gkikas, D. C., Theodoridis, P. K., & Beligiannis, G. N. (2022). Enhanced Marketing Decision Making for Consumer Behaviour Classification Using Binary Decision Trees and a Genetic Algorithm Wrapper. *Informatics*, *9*(2), 45.

Guo, J., Chen, C., Chen, C., Zuo, E., Dong, B., Lv, X., & Yang, W. (2022). Near-infrared spectroscopy combined with pattern recognition algorithms to quickly classify raisins. *Scientific Reports*, *12*(1), 7928.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics*, *15*(1), 41–51.

Ilkay, C., Murat, K., & Sakir, T. (2020). Classification of Raisin Grains Using Machine Vision and

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

Artificial Intelligence Methods. *Gazi Muhendislik Bilimleri Dergisi*, *6*(3), 200–209. https://dergipark.org.tr/tr/download/article-file/1227592

Imran, M., Sattar, M. U., Khan, H. W., Ghaffar, A., & Mushtaq, H. (2022). Selecting a Better Classifier Using Machine Learning for COVID-19. *International Journal of Computing and Digital Systems*, *11*(1), 955–962. https://doi.org/10.12785/ijcds/110178

Khojastehnazhand, M., & Ramezani, H. (2020). Machine vision system for classification of bulk raisins using texture features. *Journal of Food Engineering*, *271*, 109864.

Lantz, B. (2013). Machine Learning with R: Learn how to use R to apply powerful machine learning methods and gain an insight into real world applications. In *Livery Place*. Packt Publishing Ltd., Packt Publishing Ltd.

Liu, Y., Hu, L., Yan, F., & Zhang, B. (2013). Information Gain with Weight Based Decision Tree for the Employment Forecasting of Undergraduates. *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, 2210–2213. https://doi.org/10.1109/GreenCom-iThings-CPSCom.2013.417

Ma, G., Chao, Z., Zhang, Y., Zhu, Y., & Hu, H. (2018). The application of support vector machine in geotechnical engineering. *IOP Conference Series: Earth and Environmental Science*, *189*(2). https://doi.org/10.1088/1755-1315/189/2/022055

Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: a review. *Procedia Manufacturing*, *35*, 698–703.

Mishra, A., & Dasgupta, A. (2022). Supervised and Unsupervised Machine Learning Algorithms for Forecasting the Fracture Location in Dissimilar Friction-Stir-Welded Joints. In *Forecasting* (Vol. 4, Issue 4, pp. 787–797). https://doi.org/10.3390/forecast4040043

Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: a review. *Artificial Intelligence Review*, *52*(2), 857–900. https://doi.org/10.1007/s10462-017-9611-1

Nedaie, A., & Najafi, A. A. (2018). Support vector machine with Dirichlet feature mapping. *Neural Networks*, *98*, 87–101. https://doi.org/https://doi.org/10.1016/j.neunet.2017.11.006

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science and Technology*, *8*(1), 13–19.

Reddy, M. R. (2021). Implementation of SVM machine learning Algorithm to predict lung And Breast Cancer. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(12), 3050–3060.

Shang, C., Li, M., Feng, S., Jiang, Q., & Fan, J. (2013). Feature selection via maximizing global information gain for text classification. *Knowledge-Based Systems*, *54*, 298–309.

DOI: https://doi.org/10.24127/ajpm.v12i1.6873

https://doi.org/https://doi.org/10.1016/j.knosys.2013.09.019

Soofi, A. A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, *13*, 459–465.

Su, Q.-H., & Chiang, K.-N. (2022). Predicting Wafer-Level Package Reliability Life Using Mixed Supervised and Unsupervised Machine Learning Algorithms. In *Materials* (Vol. 15, Issue 11). https://doi.org/10.3390/ma15113897

Taneja, S., Gupta, C., Goyal, K., & Gureja, D. (2014). An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering. *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, 325–329. https://doi.org/10.1109/ACCT.2014.22

Tariq, A., Yan, J., Gagnon, A. S., Riaz Khan, M., & Mumtaz, F. (2022). Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. *Geo-Spatial Information Science*, 1–19.