Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 2 (2024) 309 - 320 e-ISSN: 2580-0760

Hybrid Data Mining For Member Determination And Financing Prediction In Syariah Financing Saving And Loan Cooperatives

Ondra Eka Putra[1], Randy Permana[2]
[1]Department Information System, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang
[2]Department Informatics Engineering, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang
[1]ondraekaputra@upiyptk.ac.id, [2]randy_permana@upiyptk.ac.id

Abstract

Syariah Financing Saving And Loan Cooperatives (KSPPS) is an Islamic financial institution aimed at people who are on the lower middle scale to lift the economy of small communities through microfinancing programs. Problems that often occur in member recommendations to get KSPPS financing are often not on target. In addition, The amount of member financing is often problematic due to a lack of analysis, resulting in poor financing instalments. This research aims to present an analysis model for clustering and classification using hybrid data mining algorithms. This research method is using hybrid data mining Algorithms, namely K-Medoids, Naïve Bayes, and k-Nearest Neighbors (k-NN). This study uses the historical dataset of the last two years on KSPPS BMT Dadok Tunggul Hitam as a total of 70 data samples. The analysis parameters consist of income, business, residence Status, financing application, billing history, and balance amount. The best analysis Model will be obtained by comparing the results between Naïve Bayes with K-Medoids, and K-Nearest Neighbor (k-NN) with K-Medoids. The results of this research showed the best performance is using the hybrid Naïve Bayes data mining model with K-Medoids which has an accuracy of 90.91% for data split 70:30, while performance with K-fold cross-validation shows an accuracy of 93.49% using this algorithm. Overall, the results of this study can provide an effective analysis model to determine the status of the loan.

Keywords: *KSPPS; hybrid data mining;, k-medoids; naïve bayes; k-nearest neighbor (k-NN)*

How to Cite: O. E. P. Ondra Eka Putra and R. P. Randy Permana, "Hybrid Data Mining For Member Determination And Financing Prediction In Syariah Financing Saving And Loan Cooperatives", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 2, pp. 309 - 320, Apr. 2024.
DOI: https://doi.org/10.29207/resti.v8i2.5683

1. Introduction

Syariah Financing Saving And Loan Cooperatives (KSPPS) is a business entity that collects funds from the community and distributes them as financing to the community [1]. Financing provided by Sharia cooperatives by the provisions contained in law No. 10 of 1998 which explains the financing of Shariah principles. Financing provisions made by Sharia cooperatives must contain rules agreed upon by both parties to cause legal attachment by each party [2]. KSPPS carries out its activities based on Sharia principles and is a microfinance institution that has an important role in the development of small and medium enterprises [3]. Financing from KSPPS is intended for the development of lower-middle entrepreneurs based on Independence [4], so not all KSPPS members are recommended to get financing, it is necessary to do clustering of members. The practice that occurs in the field is that financing is given not only to members who are recommended to get it, and the amount of financing provided is not balanced with the ability of members to pay it off, causing bad credit problems, and programs that are implemented improperly.

The amount of financing approved by KSPPS members varies, very much found in the field the amount of financing approved does not match the ability of KSPPS members, so financing instalments often crash [5]. KSPPS financing value policy analyzes financing capability by asking about the remaining results of operations every day after deducting costs [6]. Based on these facts, it can be seen that this causes problems in the process of controlling kspps financing funds. Problems that occur due to the financing analysis process that is still done manually can cause errors in determining the amount of financing approved [7]. To overcome these problems, an analytical model is

Received: 20-01-2024 | Accepted: 26-04-2024 | Published Online: 28-04-2024

needed that can be used as an illustration in decision-making [8]. The analysis Model is presented in the classification process [9], which is used to determine the amount of kspps financing. The classification method is a process that can be used in determining a class of data, this method has the purpose of predicting the class of objects whose category is not yet known [10]. The process of this method is by forming a model that can distinguish data into different data classes based on certain functions and rules [11]. Research on the analysis of determinants of acceptance of credit data on Cooperative customers KSPPS BMT Lampung Tengah using Gradient Descent classification algorithm produces a good model in determining the acceptance of credit data KSPPS [12], this study has not used the real data used for testing.

Previous research on hybrid data mining has been done to get the best model for getting solutions from existing cases. The Hybrid Method is the ability to combine several individual classification models to increase the performance value of the model [13]. The research of hybrid data mining to identify Temporal effects on breast cancer survival using 1, 5, and 10 years. The results of this study show that 10-year breast cancer survival is much lower than 1 or 5 years [14]. Research conducted by Mohammad Taghi Sattari on the use of hybrid data mining methods for soil temperature estimation with meteorological parameters using Decision Tree (DT) and Gradient Boosted Trees (GBT) methods, research shows that the best estimation with DT models at soil depths of 10 and 20 cm [15]. The research of hybrid data mining to predict the hydraulic geometry of gravel bottom Rivers, results in the study indicate that hybrid models have higher predictive power than independent data mining models [16]. Research using hybrid machine learning algorithms based on data mining in the prediction of marketing decisions, this study resulted in a hybrid method used to improve manufacturing and marketing strategies [17]. Research on hybrid data mining for sales classification, the results showed that the hybrid data mining model has a significant influence in determining the classification results with an accuracy of 85.7143% [18]. Research on the optimization of population document services using hybrid data mining, this study shows the results of comparative analysis of K-NN has the highest accuracy of 97.14% for k=1 and k=2, and the lowest accuracy of 77.1% at k=11 and k=13, Naıve Bayes method has an accuracy of 94.2 [19].

Research on data clustering of worm children in Riau province using K-Means and K-Medoid algorithms. The results of the comparison of the K-Medoid algorithm and K-Means through the validity of the Silhouette Coefficient resulted in the value of the K-Means algorithm being 0.1443, and the value of the K-Medoid algorithm being 0.5009. This shows that the K-Medoids algorithm is better than the K-Means algorithm. The limitation of this study is the limited scope in Riau province, has not covered other regions [20]. The research on Trans Jakarta Corridor cluster determination research using the Majority Voting method using K-Medoids and K-Means Data Mining algorithms after and before the COVID-19 pandemic. The use of the k-Means algorithm before the pandemic produced a grouping of 3 clusters as the optimal number of clusters with a DBI value of 0.184, and during the pandemic produced a grouping of 2 clusters as the optimal number of clusters with a DBI value of 0.188. Meanwhile, the use of the K-Medoids algorithm produced 3 optimal clusters before the pandemic with a DBI value of 0.200 and produced 4 optimal clusters during the pandemic with a DBI value of 0.190 [21].

Research in reducing credit risk is carried out to avoid the destruction of a finance company in solving the problem of credit risk analysis [22], research on credit analysis conducted by customer business feasible (feasible), mark criteria (marketable business results), profitable, with the k-NN method with an average value of 68% precision and recall value of 51% [23]. Another study predicted prospective financing customers using the comparison of k-NN, Decision Tree, Naive Bayes, and logistic regression methods in the classification of Non-Performing Financing. The data set used was 80% for training and 20% for testing. The results showed Naive Bayes classification is the best classification result with 80% data distribution for training and 20% for testing with a sensitivity of 58.25%, accuracy value of 84.69%, and specificity of 90.16% [24]. Research on the classification of Sharia Cooperative customer financing approval using the Naïve Bayes algorithm, Decision Tree, and SVM, results showed the highest accuracy using the Support Vector Machine (SVM) at 89.86%, while the Naïve Bayes algorithm at 77.29%, and Decision Tree 89.02% [25]. Research on the approval of Islamic Cooperative customer financing using a PSO-based SVM classification algorithm, the results of this study showed an accuracy of 90.91% SVM with PSO, and 89.86% SVM without PSO [26].

Based on the previous explanation, the study produced a pattern to analyze the amount of financing for members of Syariah Financing Saving And Loan Cooperatives (KSPPS) using hybrid data mining methods. The novelty of this study presents a maximum analysis model using the parameters of income, business, residence Status, financing application, billing history, and balance amount, before classifying the amount of financing, a clustering of recommended members is carried out to obtain financing. With this novelty, the classification model optimally provides certainty in determining the amount of financing for KSPPS members. In general, this study is also able to make a major contribution to KSPPS in making decisions to determine the amount of financing for KSPPS members in the next period.

2. Research Methods

Research on determining the suitability of cooperative members in proposing loans and determining the loan amount that can be obtained through a hybrid data mining approach. The proposed Hybrid approach combines the Clustering method to group members and the classification method to provide a classification of the amount of loan that can be obtained by a member. Classification is an activity in grouping an object into a class [27]. At the classification stage, analysis would carried out using two classification algorithm methods, called Naïve Bayes and Nearest Neighbour (K-NN). Performance measurements of the classification results of the two algorithms would carried out to obtain the best classification model that suits the analyzed case. Performance measurements are carried out in the majority of studies to compare an algorithm or model [28]. The research workflow can be described in Figure 1.

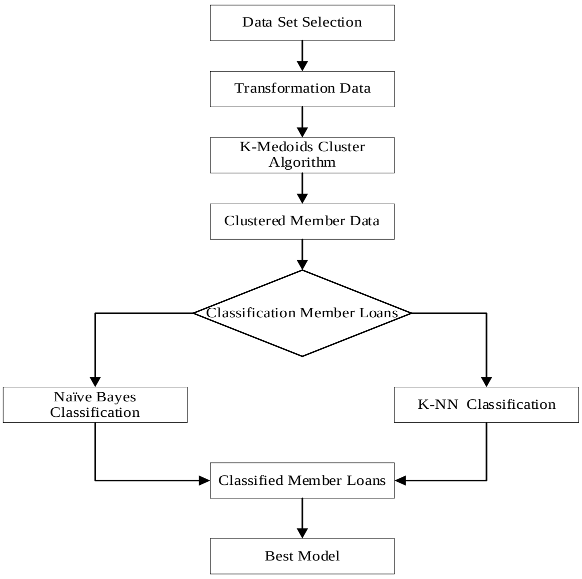


Figure 1. Research Workflow

Based on Figure 1, the following stages were contained in the hybrid data mining model based on Clustering and Classification methods.

The first stage of this model is selecting the dataset used in the first process, called the clustering method of the proposed hybrid data mining model. The Clustering method is used to reduce the number of candidate members who fall into the recommendation category to obtain loan approval. Based on the analysis process carried out by the needs of the cooperative, several variables contained in the dataset were determined to be used in the clustering process, including income, businesses owned and the status of members' residences.

Selected variables from the dataset will undergo data normalization first. Normalization carried out to form numerical data can improve the training process with consistent data distribution [29]. The data normalization concept used is Min-max, where this normalization will produce data within a certain range of values. The Min-Max formula is shown in Formula 1.

$$\text{Xnormalization} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \tag{1}$$

The K-Medoids algorithm is a quite popular algorithm for grouping data besides the K-Means algorithm. The fundamental difference between these two algorithms lies in determining convergent clusters, where K-Means cluster centres are updated by calculating the total average of all points in the cluster, while K-Medoids selects Medoid (one of the data points) as the centre of the cluster by calculating the minimum total distance to other points contained in the cluster. The advantage of K-Medoids over K-Means is based on reducing differences between data objects by selecting more representative objects [32]. Differences in data objects that are very different or called outliers can significantly affect the results of clustering. The K-Medoids algorithm applies Euclidean distance to measure the distance between two points in a multidimensional space by calculating the square root of the sum of the squares of the differences between each element in the two vectors.

The K-Medoids clustering stage is carried out by initiating the selection of initial medoid data points. Based on the selected medoids, distance calculations are carried out using distance calculation metrics such as Minkowski or Manhattan. New medoid groups are determined based on the smallest total distance from each medoid group. The next step is to update the medoids and carry out the distance calculation and medoid grouping steps again. The final result of the medoids is medoids that have converged (no changes in the position of the Medoids) based on the last grouping and the previous grouping. The Euclidean distance for the K-Medoids algorithm is shown in Formula 2.

$$\text{the j} = \sum_{i=1}^{k} \sum_{j=1}^{m} d(Oj, Cj) \tag{2}$$

There are classes of k clusters, where each cluster has its own centre point which is symbolized as $C1$, $C2$, ..., Ck. The objects contained in each cluster are expressed as $O1$, $O2$, ..., Om. The function $d(Oj, Ci)$ describes the distance from an object Oj to the centre point Ci of the class cluster.

The cluster results from K-Medoids will group cooperative members into two groups. The first group with the highest score will be the group recommended for loan applications, while the second group will be the group of members who are not approved for borrowing. The recommendation cluster is saved and then returned to the dataset selection stage as in the previous stage.

The classification in this model is to determine the amount of loan that can be obtained by a cooperative member. Source The dataset used is data sourced from the results of the data group recommended in the K-Medoids cluster process. Based on the results of K-Medoids, all variables related to the recommendation data group will be included as a data set for classification activities of loan amounts that can be obtained by cooperative members.

Naïve Bayes classification is a probabilistic classification method based on Bayes' theorem, which assumes that each feature used for classification is independent of the other. The basis of Naive Bayes is Bayes' Theorem, which provides a way to calculate the posterior probability (the probability of a class after looking at the data) based on prior probability (the probability of the class before looking at the data) and likelihood (the probability of the data given the class). The class probabilities for the Naive Bayes method are shown in Formula 3.

$$\mathrm{P(Class|data)} = \frac{P(data|Class)xP(Class)}{P(data)} \quad (3)$$

Based on the Bayes theorem equation, it gives Naïve Bayes the advantage of being able to classify the estimated parameters with a small number [30].

Meanwhile, K-Nearest Neighbors (K-NN) classification is a non-parametric classification method which is based on the principle that similar objects tend to be close to each other. K-NN uses the distance metric Euclidean Distance to determine a new data point class based on the majority of labels from the number of nearest neighbors determined by the K parameter. The formula of the K-NN method is shown in Formula 4.

$$d(xi.xj) = \sqrt{}\sum_{r=1}^{n}(\left(ar(xi) - \left(ar(xj)\right)\right)^2 \quad (4)$$

Once the nearest neighbors are identified, K-NN takes the majority class of those neighbors, and the new data point is classified according to the majority labels of its K nearest neighbors [31].

Each classification method, both Naïve Bayes and K-NN, will provide classification results for the amount of loans approved to cooperative members. At this stage, an evaluation of the two classification models will be carried out.

The results of two classification methods involve the process of comparing the performance and effectiveness of the two methods based on various evaluation metrics. Accuracy testing includes a series of steps to evaluate the extent to which the results produced by a model or system meet expectations or requirements. Accuracy, Precision and Recall are used to evaluate the model [32].

Accuracy is a test indicator related to accuracy or the extent to which a classification method model is correct in making predictions. The accuracy formula is shown in Formula 5.

$$\mathrm{Accuracy} = (\mathrm{TP} + \mathrm{TN}) / (\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}) \quad (5)$$

Precision Measurement measures the proportion of instances predicted as positive that are actually positive. The precision formula is shown in Formula 6.

$$\mathrm{Precision} = (\mathrm{TP}) / (\mathrm{TP} + \mathrm{FP}) \quad (6)$$

Recall is a measurement of the extent to which the model can identify positive value instances. The recall formula is shown in Formula 7.

$$\mathrm{Recall} = (\mathrm{TP}) / (\mathrm{TP} + \mathrm{FN}) \quad (7)$$

Determination of the best model is decided based on performance measurements with the highest value of each model.

3. Results and Discussions

In doing data mining, follow the steps of Knowledge Discovery From Database (KDD) rules, starting from the data selection stage to the evaluation stage. The selection stage, using historical data of the last two years on KSPPS BMT Dadok Tunggul Hitam was a total of 70 sample data. The attributes of selection data consist of Income, Business, Population Status, Financing Application, Billing History, Total Balance, and Approved Financing. The dataset used is shown in Table 1.

Table 1. Dataset KSPPS BMT Dadok Tunggul Hitam

Member	Income	Business	Residence Status	Financing Application	Billing History	Balance Amount	Financing Approved
Member 1	High	Medium	Rent House	High	Problem	Medium	Medium
Member 2	High	Not Aviable	Rent House	Low	No Problem	Medium	Low
Member 3	Medium	Small	Own House	Low	No Problem	Medium	Medium
Member 4	Low	Not Aviable	Own House	Medium	No Problem	Low	Medium
Member 5	High	Not Aviable	Own House	High	Problem	High	Low
Member 6	High	Medium	Own House	High	No Problem	High	High
...	...	...	...	...	...	...	...
Member 68	Low	Medium	Other	Medium	No Problem	Low	Medium
Member 69	High	Not Aviable	Rent House	Low	No Problem	Medium	Medium
Member 70	Low	Not Aviable	Other	High	Problem	Low	Low

3.1 K-Medoids Process

In this study, we applied the K-medoids algorithm to perform a cluster analysis of cooperative members who are recommended to obtain cooperative financing. This algorithm was chosen because of its ability to group cooperative members into homogeneous clusters based on income, business and residence Status. The data mining process with the K-Medoids algorithm is used to cluster members who are recommended to get KSPPS financing because the determination of recommendations for members who get KSPPS financing is seen from several parameters to maintain

KSPPS financial stability. The use of the K-medoids algorithm in cooperative member clustering analysis has advantages over other algorithms such as k-means. This is due to the ability of k-medoids to determine which medoid is stronger than each cluster so that the cluster results become more stable and interpretable. Data sets used in this study consisted of 70 samples in the KSPPS BMT Dadok Tunggul Hitam.

In the process of K-medoids, the following steps are performed:

Data normalization is a primary element of data mining to keep the datasets consistent. In the normalization process, it is necessary to transform the data or transform the original data into a format that allows efficient data processing. The main goal of data normalization is to eliminate data redundancy (repetition) and standardize information for better data workflows. Data normalization is carried out to scale the data of an attribute so that it is in a smaller range, such as -1 to 1 or 0 to 1. Normalization of data using Min-Max normalization in Formula 1. The results of data normalization for K-medoids are shown in Table 2.

Table 2. Data Normalization

Member	Income	Business	Residence Status
Member 1	0.67	1	0
Member 2	0.67	0	0
Member 3	0.33	0.5	1
Member 4	0.00	0	1
Member 5	0.67	0	1
Member 6	0.67	1	1
...	...	...	...
Member 68	0.00	1	0.5
Member 69	0.67	0	0
Member 70	0.00	0	0.5

Determine the k value according to the number of clusters, namely 2, loan recommendation and non-recommendation loan clusters.

Define medoids or centroid of cluster. Medoid randomly selected as many as 2 points, the centroid for each cluster shown in Table 3.

Table 3. Centoid of Cluster

Cluster	Income	Business	Residence Status
Cluster 1	0.00	0.00	0.50
Cluster 2	0.00	1.00	0.50

Calculate the distance of each data object to each medoid formed. Below are some calculations of the distance between the data object and the first medoid to the object's centroid.

$$d_1 = \sqrt{((0-0)^2 + (0-0)^2 + (0,5-0,5)^2)}$$

Calculation of data object distance with the second medoid:

$$d_2 = \sqrt{((0-0.67)^2 + (0-1)^2 + (0.5-0)^2)} = 1.30$$

Select the closest distance of each data object to the medoid, and insert the data object into the cluster. The results of the first iteration calculation are shown in Table 4.

Table 4. First Iteration Cluster Results

Cluster	Total
Cluster 1	37
Cluster 2	33

After all the data is filed into each cluster, both data from the calculation results of the first iteration and subsequent iterations. Calculate the closest distance value from the data and determine the deviation value.

The sum of the first iteration's shortest distances,

= 0.83+0.83+0.78+...+0 = 48.78

The sum of the nearest distances of the second iteration, =1+0+1,13+...+0,83 = 57,5

The second deviation resulting from the sum of the closest distances per iteration is 8.72. Based on these calculations, the value S>0 is acquired, and then the process is stopped.

The calculation and testing of the K-medoids algorithm also uses a data mining tool, Rapidminer software shown in Figure 2.

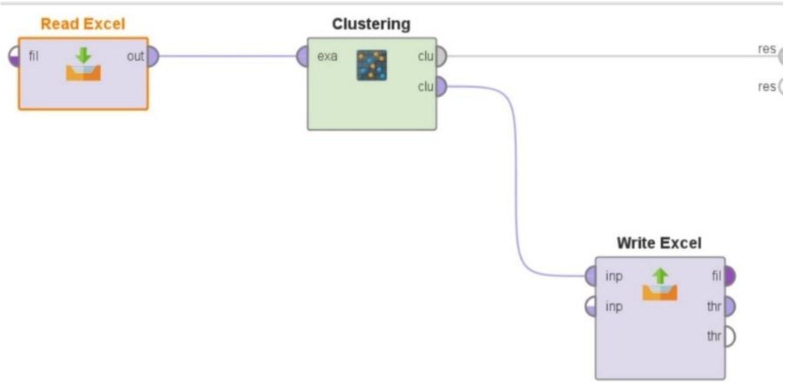
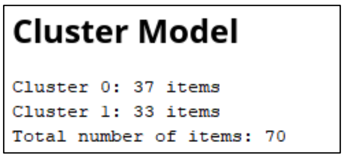


Figure 2. K-Medoids Process with RapidMiner

Figure 2 is an architectural form of the K-Medoids algorithm using Rapidminer Software in the process of clustering cooperative members who are recommended to obtain financing. The results of the cluster are shown in Figure 3.



Cluster Model

```
Cluster 0: 37 items
Cluster 1: 33 items
Total number of items: 70
```

Figure 3. The result of K-Medoids Cluster with RapidMiner

The results of the cluster using the K-Medoids algorithm are 2 clusters because it is determined the number of K=2 is the number of clusters produced. Cluster 0 (Recommendations) has 37 items, and Cluster 1 (Not Recommended) has 33 items from a total dataset of 70 items. Each Cluster 0 and Cluster 1 has a medoid that represents the centre point of the cluster. This Medoid is a data point that represents the overall characteristics of the cluster in a representative manner. Based on the results of the cluster can be seen that the distribution of data between the two clusters is relatively balanced. This information can help cooperative management to design services that better suit the needs of each member cluster, thereby increasing overall member engagement and satisfaction.

3.2 Naïve Bayes Process

In this study, we used the Naïve Bayes algorithm to perform a classification of financing approved by the cooperative. This algorithm was chosen because of its efficiency in handling categorical data and its ability to classify data based on the conditional probability of observed features. The Naïve Bayes process executes after the grouping stage using the K-Medoids algorithm. The attributes used for the naïve Bayes process are Income, Business Status and Residence, Financing Application, Billing History, Balance Amount and Approved Financing. The dataset used for the Naïve Bayes process is the cluster 0 category data group from the K-medoids process or the KSPPS financing application recommendation cluster, which is 37 items. The dataset of the Naïve Bayes process is shown in Table 5.

Table 5. Dataset After K-Medoids Process

Member	Income	Business	Residence Status	Financing Application	Billing History	Balance Amount	Financing Approved
Member 1	High	Medium	Rent House	High	Problem	Medium	Medium
Member 6	High	Medium	Own House	High	No Problem	High	High
Member 8	Very High	Medium	Own House	High	No Problem	Medium	High
Member 9	Medium	Medium	Own House	Low	No Problem	Medium	High
Member 10	Medium	Medium	Own House	Medium	Problem	Medium	Low
Member 12	Medium	Medium	Rent House	High	No Problem	High	Medium
...	...	...	...	...	...	...	...
Member 65	High	Small	Other	Medium	No Problem	Medium	Medium
Member 66	Medium	Medium	Other	High	Problem	High	Low
Member 67	High	Medium	Other	Medium	No Problem	Medium	Medium

Naïve Bayes has a logical approach to the chance of occurrence of data in a dataset, so it uses a calculation of the number of occurrences of each attribute in the dataset used. The calculation of the number of occurrences of each attribute in the dataset used is shown in Table 6. Calculate the probability of the appearance of the value of the class label, called by class of large, medium and low categories. The calculation of the occurrence of each class and the probability value of each class contained in the dataset are shown in Table 7.

Calculate the probability value of each attribute to the occurrence of a class of existing parameters. The value of the probability of occurrence of the attribute to the occurrence of the class is shown in Table 8.

The results of the classification using the Naïve Bayes algorithm show that approved financing can be classified into three categories of approved financing: low, medium, and High. This result is based on the conditional probability of the observed features.

Multiply all the results in step *3.2 b* with the testing data. The test data used is new data without labels. The Naïve Bayes algorithm can classify testing data through class and attribute probability values. The Test Data is shown in Table 9.

Table 6. Number of Data Occurrences of Each Attribute

Parameters	Attributes	Total
Income	Very High	3
	High	20
	Medium	14
	Low	0
Business	Medium	27
	Small	9
	Not Aviable	1
Residence Status	Own House	4
	Rent House	23
	Other	10
Financing Application	High	11
	Medium	13
	Low	13
Billing History	No Problem	28
	Problem	9

Table 7. Probability Value of Class

Class	Total	Probability Value
High Class	8	0.22
Medium Class	23	0.61
Low Class	6	0.16
Total Class	37	1.00

Tabl 8. The Probability Value of the Attribute Appearing Against the Class Appearing

Parameters	Attributes	Total	Probability Against Class		
			High	Medium	Low
Income	Very High	3	0	0.13	6
	High	20	0.5	0.52	0.33
	Medium	14	0.5	0.26	0.67
	Low	0	0	0	0
Business	Medium	27	0.75	0.74	0.67
	Small	9	0.25	0.22	0.33
	Not Aviable	1	0	0.04	0
Residence Status	Own House	4	0.25	0.09	0
	Rent House	23	0.75	0.57	0.67
	Other	10	0	0.35	0.33
Financing Application	High	11	0.63	0.13	0.33
	Medium	13	0.5	0.39	0.17
	Low	13	0	0.43	0.5
Billing History	No Problem	28	0.75	0.96	0
	Problem	9	0.25	0.04	1
Balance Amount	High	11	0.75	0.13	0.33
	Medium	23	0.25	0.78	0.5
	Low	3	0	0.09	0

Table 9. Testing Data of Naïve Bayes

Member	Income	Business	Residence Status	Financing Application	Billing History	Balance Amount	Financing Approved
Member Test	Low	Medium	Other	Medium	No Problem	Low	?

Based on Table 9, after multiplying the probability value of the class with all the attribute values of the testing data, the probability of a High Class = 0, Medium Class = 0.005201409, and Low Class = 0, so the classification of the testing data member is in the Medium class category. Calculation and testing of the Naïve Bayes algorithm also use a data mining tool, Rapidminer software shown in Figure 4.

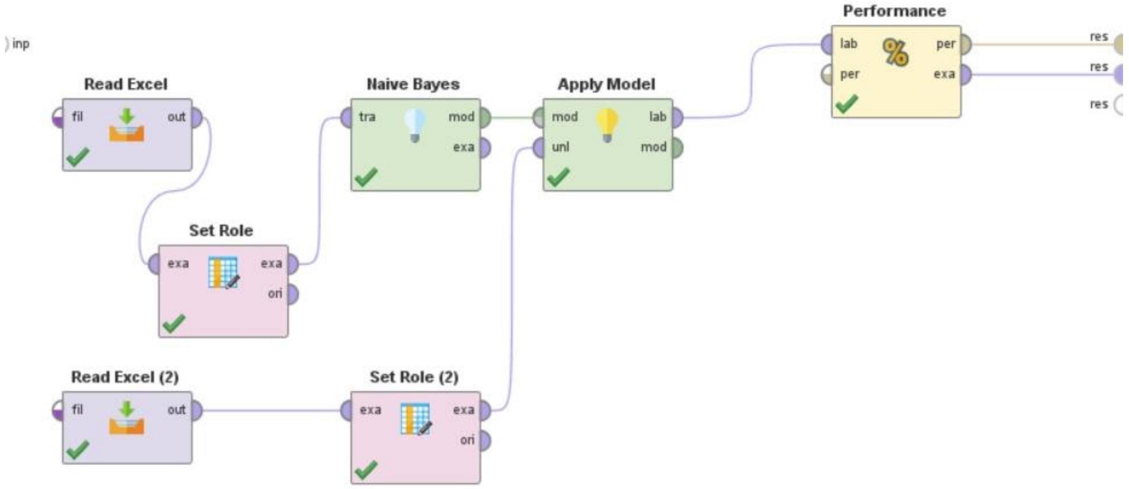


Figure 4. Naïve Bayes Process With RapidMiner

Figure 4 is an architectural form of the Naïve Bayes algorithm using Rapidminer Software in the classification process for the prediction of approved financing for cooperative members. The results of classification based on testing data are shown in Table 10.

Table 10. The Result of Naïve Bayes Clasification with RapidMiner

Member	Income	Business	Residence Status	Financing Application	Billing History	Balance Amount	Financing Approved
Member Test	Low	Medium	Other	Medium	No Problem	Low	Medium

3.3 k-Nearest Neighbour Process

In this study, we also used the k-NN algorithm to classify financing approved by the cooperative. This algorithm was chosen because of its ability to handle high-dimensional data and can classify data based on the closest distance to the nearest neighbor in the feature space. k-Nearest neighbor (k-NN) algorithm is used to classify financing approval submitted by cooperative members. The attributes used for the k-NN process are income, business and residence Status, financing application, billing history, balance amount and approved financing. The dataset used for the k-NN process is the cluster dataset from the K-medoids process, which is in the cluster 0 category or the recommended cluster for submitting kspps financing of 37 items, using the dataset in Table 5.

In the k-NN stage, several attributes are added that help in classifying classes, the amount of financing approved by the cooperative includes: financing application, billing history, and the amount of balance owned by cooperative members. Lingustik value for the large attribute of the financing application, the balance amount, and approved financing are obtained based on the value interval that has been set. Financing submission is the amount of financing submitted by members of the cooperative to the cooperative. The amount of balance is a recapitulation or the amount of balance owned by members since becoming a member of the cooperative. Financing approved is the nominal amount of financing approved by the cooperative from the analysis carried out by the cooperative based on the track record of the activities of cooperative members. The linguistic value of the financing application attribute, the balance amount, and the approved financing are shown in Table 11.

Table 11. Linguistic Value of Financing Application, Balance Amount and Approved Financing

No	Linguistic Value	Interval Value
1	High	> 10.000.0000
2	Medium	> 4.000.000 Up to <=10.000.000
3	Low	<= 4.000.000

The k-NN algorithm works by using values from numerical attributes, then the dataset is transformed using weighting. The data transformation results for the k-NN process are shown in Table 12.

The k-NN algorithm works by finding the closest distance to several K from a neighbor as a reference point in determining the class of the new data. Before determining the number K, the distance between the data needs to be calculated first using the Euclidean distance formula.

Table 12. K-NN Algorithm Data Transformation

Member	Income	Business	Residence Status	Financing Application	Billing History	Balance Amount	Financing Approved
Member 1	3.00	3.00	1.00	3.00	1.00	2.00	Medium
Member 6	3.00	3.00	3.00	3.00	2.00	3.00	High
Member 8	4.00	3.00	3.00	3.00	2.00	2.00	Medium
Member 9	2.00	3.00	3.00	1.00	2.00	2.00	Medium
Member 10	2.00	3.00	3.00	2.00	1.00	2.00	High
Member 12	2.00	3.00	1.00	3.00	2.00	3.00	High
...	...	...	...	...	...	...	...
Member 65	3.00	2.00	2.00	2.00	2.00	2.00	Medium
Member 66	2.00	3.00	2.00	3.00	1.00	3.00	Low
Member 67	3.00	3.00	2.00	2.00	2.00	2.00	Medium

Table 13 shows the new member data as test data that has the value of each attribute used. Based on the value of the attribute contained in Table 13, performed the calculation process of the search distance, which displayed the results in Table 14.

Table 13. Test Data

No	Parameters	Value
1	Income	1.00
2	Business	3.00
3	Residence Status	2.00
4	Financing Application	2.00
5	Billing History	2.00
6	Balance Amount	1.00

Table 14. The Results of Search Distance Data

Member	Distance	Financing Approved
Member 1	2,83	Medium
Member 6	3,16	High
Member 8	3,46	Medium
Member 9	2,00	Medium
Member 10	2,00	High
Member 12	2,65	High
...	...	...
Member 65	2,45	Medium
Member 66	2,65	Low
Member 67	2,24	Medium

$$d(1) = \sqrt{(1-3)^2 + (3-3)^2 + (2-1)^2 + (2-3)^2 + (2-1)^2 + (1-2)^2} = 2{,}83$$

$$d(2)=\sqrt{(1-3)^2 + (3-3)^2 + (2-3)^2 + (2-3)^2 + (2-2)^2 + (1-3)^2} = 3{,}16$$

$$d(3)=\sqrt{(1-4)^2 + (3-3)^2 + (2-3)^2 + (2-3)^2 + (2-2)^2 + (1-2)^2} = 3{,}46$$

$$d(4)=\sqrt{(1-2)^2 + (3-3)^2 + (2-3)^2 + (2-1)^2 + (2-2)^2 + (1-2)^2} = 2{,}00$$

$$d(5)=\sqrt{(1-2)^2 + (3-3)^2 + (2-3)^2 + (2-2)^2 + (2-1)^2 + (1-2)^2} = 2{,}00$$

The number of K values used in this process is even, namely k=4, k=6, k=8, and k=10 because the number of classifications determined there are 3 classes, namely Low, Medium, and Large. The most dominant class group will then be the class of the data being predicted. After the process of calculating the distance of the nearest neighbor with the value of k=4, k=6, k=8, and K=10 from all data, the data obtained with the closest distance is shown in Table 15.

Table 15. Total of Class Category

Value of K	Class Category	Total Data
	Low	2
K=4	Medium	7
	High	2
	Low	4
K=6	Medium	17
	High	4
	Low	6
K=8	Medium	18
	High	5
	Low	6
K=10	Medium	21
	High	7

Based on Table 15 with a value of k=4 obtained the number of large class category data is as much as 2 data, the medium class is as much as 7 data and the low clis ass as much as 2 data. value k=6 obtained the number of large class category data as much as 4 data, medium class as much as 17 data and low class as much as 4 data. Value of k=8 obtained the amount of large class category data as much as 5 data, medium class as much as 18 data and low class as much as 6 data. Parameter value k=10 obtained the number of large class category data as much as 7 data, medium class as much as 21 data and low class as much as 6 data. Based on the amount of each data from the feature values that are processed using parameter values k=4, k=6, k=8, and k=10, the prediction results obtained for the next data is the medium class category.

Calculation and testing of k-NN algorithms also use data mining tools, called Rapidminer software shown in Figure 5.

Figure 5 is an architectural form of the k-NN algorithm using Rapidminer Software in the classification process for the prediction of approved financing for cooperative members. The results of classification based on testing data are shown in Table 16.

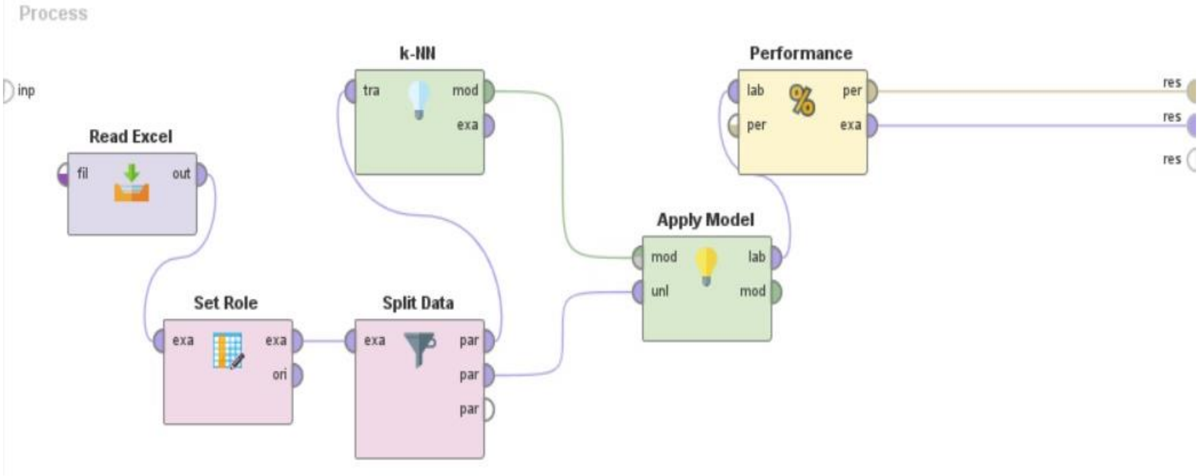


Figure 5. k-NN Process With RapidMiner

Table 16. The result of k-NN Classification with RapidMiner

Member	Income	Business	Residence Status	Financing Application	Billing History	Balance Amount	Financing Approved
Member Test	Low	Medium	Other	Medium	No Problem	Low	Medium

3.4 Evaluation Model

Performance measurement of classification performed by the k-NN algorithm, is necessary to do with the measurement of accuracy to give an idea of how accurate the model is in predicting the entire class, precision to measure how appropriate the model is in predicting a particular class and Recall to measure how well the model in finding a positive case using equation (5), (6), and (7). In the test, the calculation of the performance of the Naïve Bayes and K-NN models using data split of 80:20 and 70:30, and K-fold cross-validation. The results of model performance testing are shown in Table 17 and Table 18.

Table 17. The Results of the Performance Model Using Data Split

Data Split	Metode	K-Parameter	Accuracy	Precision			Recall		
				High	Medium	Low	High	Medium	Low
80:20	K-NN + K-Medoids	4	75.00%	50.00%	83.33%	0.00%	50.00%	100.00%	0.00%
		6	75.00%	50.00%	83.33%	0.00%	50.00%	100.00%	0.00%
		8	75.00%	50.00%	83.33%	0.00%	50.00%	100.00%	0.00%
		10	62.50%	0.00%	71.43%	0.00%	0.00%	100.00%	0.00%
	Naïve Bayes + K-Medoids	-	85.71%	00%	83.33%	100.00%	00%	100.00%	100.00%
70:30	K-NN + K-Medoids	4	72.73%	50.00%	77.78%	0.00%	50.00%	100.00%	0.00%
		6	81.82%	66.67%	87.50%	0.00%	100.00%	100.00%	0.00%
		8	63.64%	0.00%	70.00%	0.00%	0.00%	100.00%	0.00%
		10	63.64%	0.00%	70.00%	0.00%	0.00%	100.00%	0.00%
	Naïve Bayes + K-Medoids	-	90.91%	100.00%	87.50%	100.00%	50.00%	100.00%	100.00%

Table 18. The Results of the Performance Model Using K-fold Cross

Method	K-Fold	Accuracy	Precision			Recall		
			High	Medium	Low	High	Medium	Low
Naïve Bayes + K-Medoids	1	70.27%	44.44%	76.92%	100.00%	86.96%	57.14%	28.57%
	2	70.27%	44.44%	76.92%	100.00%	86.96%	57.14%	28.57%
	3	93.49%	100.00%	88.62%	89.71%	60.14%	98.65%	88.71%
	4	81.08%	80.00%	80.77%	83.33%	57.14%	91.30%	71.43%
	5	81.08%	80.00%	81.48%	80.00%	57.14%	95.65%	57.14%
	6	83.78%	84.00%	80.00%	85.71%	57.14%	91.30%	85.71%
	7	83.78%	80.00%	84.00%	85.71%	57.14%	91.30%	85.71%
	8	75.68%	57.14%	79.17%	83.33%	57.14%	82.61%	71.43%
	9	81.08%	80.77%	80.77%	83.33%	57.14%	91.30%	71.43%
	10	83.78%	80.00%	84.00%	85.71%	57.14%	91.30%	85.71%
K-NN + K-Medoids	1	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	2	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	3	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	4	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	5	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	6	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	7	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	8	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	9	62.16%	33.33%	70.00%	0.00%	28.57%	91.30%	0.00%
	10	64.86%	40.00%	70.97%	0.00%	28.57%	95.65%	0.00%

Based on Table 17, the performance measurement value of the best K-Medoids with the k-NN model obtained an accuracy of 81.82% for data split 70:30 with parameter value K=6. The K-Medoids with Naïve Bayes model obtained the best accuracy of 90.91% for data split 70:30. Based on Table 18, the performance measurement value of the algorithm using K-fold cross-validation, the best model using the K-Medoids algorithm with Naïve Bayes obtained an accuracy of 93.49% at K-fold = 3. From the accuracy of the results obtained, this study is noteworthy when compared with previous studies. A comparison of studies is shown in Table 19.

Table 19. Comparison of Research Findings with Previous Studies

Research	Algorithm	Accuracy
Ondra (Author)	Naïve Bayes + K-Medoids	93.49%
S. Harlina	K-NN	68%
R.F. Putra	Naïve Bayes	84.69%
Nurajijah	SVM+PSO	90.91%

Based on Table 19, the model of this study can be used by the management of the cooperative to design a financing program that suits the needs and capabilities of members, as well as manage financing risk more effectively with the best accuracy value of 93.49%.

4. Conclusions

Hybrid data mining research by combining clustering and classification methods in the case of Sharia cooperatives (KSPPS) produces a model that is capable of classifying the number of loans approved to cooperative members. The clustering carried out by the K-Medoids algorithm divides members into two groups, called recommendations and non-recommendations members. The classification model will then carry out classification based on the recommended member data. In the 80:20 split data, the Naïve Bayes classification model obtained an accuracy of 85.71%, while for the K-NN classification model with a parameter set of K = 6 an accuracy of 75.00%

was obtained. The 70:30 split data of the Naïve Bayes classification model obtained an accuracy of 90.91%, while for the K-NN classification model with a parameter set of K = 6 an accuracy of 81.82% was obtained. The performance measurement value of the algorithm uses K-fold cross-validation, the best model using the K-Medoids algorithm with Naïve Bayes, obtaining an accuracy of 93.49% at K-fold = 3. So it is concluded that the combination of K-Medoids and Naïve Bayes in hybrid data mining can classify well the loan amounts approved to members of Syariah Financing Saving And Loan Cooperatives (KSPPS).

Acknowledgements

Thanks to the chairman of the foundation of Higher Education Computer padang Dr. Zerni Melmusi, MM., AK., CA., who gave full support to the study. Thanks to the chairman of KSPPS BMT Tunggul Hitam who provided understanding and data related to the research conducted.

References

[1] A. Riyani, G. Pratama, and S. Surahman, "Analisis Sistem Pengelolaan Keuangan Pembiayaan Syariah Dengan Akad Murabahah," *Ecobankers J. Econ. Bank.*, vol. 3, no. 1, p. 1, 2022, doi: 10.47453/ecobankers.v3i1.672.

[2] R. Muhamad, "Kegiatan Usaha Bank Perkreditan Rakyat Ditinjau Dari Undang-Undang Nomor 10 Tahun 1998 Tentang Perbankan," *Lex Priv.*, vol. 8, no. 1, pp. 66–77, 2020.

[3] I. Nurlaeli, "Analisis Akad Qardhul Hasan (Studi Kasus di KSPPS BMT Mentari Bumi Purbalingga)," *Islam. J. Pemikir. Islam*, vol. 23, no. 2, pp. 239–253, 2022.

[4] A. Studies, "Implementasi Sistem Bagi Hasil dan Perlakuan Akuntansi pada Pembiayaan Mudharabah Di KSPPS BMT An-Nuur Jombang," *JFAS J. Financ. Account. Stud.*, vol. 3, pp. 55–71, 2021.

[5] A. Nugroho, D. I. Astanti, and D. Septiandani, "PENYELESAIAN PEMBIAYAAN MACET DENGAN JAMINAN HAK TANGGUNGAN DI KOPERASI SIMPAN PINJAM DAN PEMBIAYAAN SYARIAH (KSPPS) HUDATAMA CABANG SEMARANG BARAT," *Semarang Law Rev.*, vol. 1, no. 1, pp. 46–58, 2020.

[6] K. F. Hana and E. A. Chodlir, "ELABORASI ANALISIS PEMBIAYAAN DALAM MEMINIMALISIR NON PERFORMING FINANCE (NPF) PADA LEMBAGA KEUANGAN SYARIAH," *MALIA J. Islam. Bank. Financ.*, vol. 5, no. 2, pp. 121–132, 2021.

[7] S. Arlis *et al.*, "POLA PENENTUAN STATUS PEMINJAMAN DENGAN ALGORITMA PERCEPTRON," *SEBATIK*, pp. 619–623, 2018.

[8] A. D. Mining and I. Pendahuluan, "Penerapan Metode Principal Component Analysis (PCA) Pada Klasifikasi Status Kredit Nasabah Bank Sumsel Babel Cabang KM 12 Palembang Menggunakan Metode Decision Tree," *Generic*, 2022.

[9] T. A. Putra, P. Ayu, W. Purnama, R. Afira, and Y. Elva, "Optimization Analysis Model Determining PNMP Mandiri Loan Status Based Based on Pearson Correlation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 158, pp. 5–12, 2022.

[10] S. Nasional, T. Elektro, S. Informasi, and T. Informatika, "Klasifikasi Status Pinjaman Calon Nasabah Koperasi Simpan Pinjam Menggunakan Metode Bayesian Network (Studi Kasus: Koperasi Simpan Pinjam BTM Nasyiah 1 Bojonegoro)," *SNESTIK -Seminar Nas. Tek. Elektro, Sist. Informasi, dan Tek. Inform.*, pp. 409–414, 2022.

[11] C. Fadlan, S. Ningsih2, and A. P. Windarto, "PENERAPAN METODE NAÏVE BAYES DALAM KLASIFIKASI KELAYAKAN KELUARGA PENERIMA BERAS RASTRA," *JUTIM*, vol. 3, no. 1, pp. 1–8, 2018.

[12] A. Swasono, "Analisis Faktor - Faktor Penentu Penerimaan Data Kredit Pada Nasabah Koperasi KSPPS BMT Lampung Tengah Menggunakan Gradient Descent," *J. Ilmu Data*, vol. 2, no. 10, pp. 1–9, 2022.

[13] A. Gumilar, S. S. Prasetiyowati, and Y. Sibaroni, "Performance Analysis of Hybrid Machine Learning Methods on Imbalanced Data (Rainfall Classification)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 158, pp. 481–490, 2022.

[14] S. Simsek and A. Dag, "A Hybrid Data Mining Approach for Identifying the Temporal Effects of Variables Associated with Breast Cancer Survival," *Dep. Inf. Manag. Bus. Anal. Fac. Scholarsh. Creat. Work.*, 2020.

[15] M. T. Sattari, A. Avram, and H. Apaydin, "Soil Temperature Estimation with Meteorological Parameters by Using Tree-Based Hybrid Data Mining Models," *Mathematics*, 2020.

[16] K. Khosravi, Z. Sheikh, and J. R. Cooper, "Predicting stable gravel-bed river hydraulic geometry : A test of novel , advanced , hybrid data mining algorithms," *Environ. Model. Softw.*, vol. 144, no. August, p. 105165, 2021, doi: 10.1016/j.envsoft.2021.105165.

[17] T. S. Kumar, "Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm," *J. Artif. Intell. Capsul. Networks*, vol. 02, no. 03, pp. 185–193, 2020.

[18] K. Penjualan, D. E. Putri, E. Praja, and W. Mandala, "Hybrid Data Mining berdasarkan Klasterisasi Produk untuk," *J. KomtekInfo*, vol. 9, pp. 68–73, 2022, doi: 10.35134/komtekinfo.v9i2.279.

[19] I. Riadi, A. Yudhana, and M. R. Djou, "Optimization of Population Document Services in Villages using Naive Bayes and k-NN Method," *Int. J. Comput. Digit. Syst.*, vol. 1, no. 1, 2024.

[20] D. Marlina, N. F. Putri, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak," *J. CoreIT*, vol. 4, no. 2, pp. 64–71, 2018.

[21] R. Sistem, A. Wibowo, M. Makruf, I. Virdyna, and F. C. Venna, "Penentuan Klaster Koridor TransJakarta dengan Metode Majority Voting," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, pp. 565–575, 2021.

[22] S. Harlina, "DATA MINING ON CREDIT FEASIBILITY DETERMINATION USING K-NN ALGORITHM BASED ON FORWARD SELECTION," *Raharja Open J. Syst.*, vol. 11, no. 2, pp. 236–244, 2018.

[23] S. Harlina and M. O. Kadang, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi kelayakan Calon Nasabah Kredit Berbasis Web," *Pros. Semin. Nas. Tek. Elektro, Inform. Sist. Inf.*, 2022.

[24] R. F. Putra and I. D. Ratih, "Comparison of K-Nearest Neighbor , Naive Bayes Classifier , Decision Tree , and Logistic Regression in Classification of Non- Performing Financing," *Int. J. Adv. Sci. Comput. Appl.*, vol. 2, pp. 69–76, 2023, doi: 10.47679/ijasca.v2i2.35.

[25] N. Nurajijah and D. Riana, "Algoritma Naïve Bayes, Decision Tree, dan SVM untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah," *J. Teknol. dan Sist. Komput.*, vol. 7, no. 2, pp. 77–82, 2019, doi: 10.14710/jtsiskom.7.2.2019.77-82.

[26] Nurajijah, F. Amsury, I. Saputra, Frieyadie, D. N. Sulistyowati, and B. Rifai, "Approval of Sharia Cooperative Customer Financing Using PSO-Based SVM Classification Algorithm," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012047.

[27] D. Priyanto, A. Robbiul, and D. Jollyta, "Naïve Bayes and K-Nearest Neighbor Approaches in Data Mining Classification of Drugs Addictive Diseases," *Ilk. J. Ilm.*, vol. 15, no. 2, pp. 262–270, 2023.

[28] A. Martono and G. Maulani, "The Effect of The Prediction of The K-Nearest Neighbor Algorithm on Surviving COVID-19 patients in Indonesia," *Ilk. JurnalIlmiah*, vol. 15, no. 2, pp. 240–249, 2023.

[29] C. Fu, M. Quintana, Z. Nagy, and C. Miller, "Filling time-series gaps using image techniques: Multidimensional context autoencoder approach for building energy data imputation," *Appl. Therm. Eng.*, 2023.

[30] E. Setiawati, U. D. Fernanda, and S. Agesti, "Implementation

of K-Means , K-Medoid and DBSCAN Algorithms In Obesity Data Clustering," *IJATIS Indones. J. Appl. Technol. Innov. Sci.*, vol. 1, no. February, pp. 23–29, 2024.

[31] R. Mutia and J. A. Ariani, "Performance Comparison K-Nearest Neighbor , Naive Bayes , and Decision Tree Algorithms for Netflix Rating Classification," *IJATIS Indones. J. Appl. Technol. Innov. Sci.*, vol. 1, no. February, pp. 16–22, 2024.

[32] G. Varone, W. Boulila, M. Driss, S. Kumari, and M. Khurram, "Finger pinching and imagination classification : A fusion of CNN architectures for IoMT-enabled BCI applications," *Inf. Fusion J.*, vol. 101, no. May 2023, 2024.