CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 9(2) (2024), Pages 222-236
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Comparative Analysis of Machine Learning Algorithms on Family Wellness Classification

Retno Budiarti*, Febri Hemarani, Mohammad Reza, Rindi Melati Mulyasari

[1] Department of Mathematics, Institut Pertanian Bogor, Indonesia

Email: retnobu@apps.ipb.ac.id

ABSTRACT

Family welfare is a state in which a family can experience happiness, have a decent quality of life, and be sufficient in meeting primary and secondary needs in family life. One factor that influences family welfare is the amount of per capita expenditure. This study aims to compare the performance of three machine learning algorithms, namely KNN (K-Nearest Neighbors), random forest, and naive Bayes, in classifying the status of families per province in Indonesia as prosperous or not prosperous. The data used in this study are 170 demographic and social statistics data from the years 2017-2021, obtained from the bps.go.id website. The first statistical analysis conducted is principal component analysis (PCA) with 9 predictor variables. PCA produces four principal components which are then used in the KNN, random forest, and naive bayes methods. The analysis results from the KNN yield an 63.46% accuracy, 62.07% precision, 69.23% recall, and 65.46% F1-score. The analysis results from the random forest yield an 69.23% accuracy, 70.83% precision, 65.38% recall, and 68.00% F1-score. The analysis results from the KNN yield an 57.69% accuracy, 54.35% precision, 96.15% recall, and 69.44% F1-score. Based on the values shown, the Random Forest method is the most suitable for classifying prosperous families.

Keywords: consumption expenditure; family welfare; KNN; naive bayes; random forest

INTRODUCTION

A family is a group of individuals within a small social system where each member depends on and influences one another. Every family has the same goal, which is to achieve well-being. Family well-being is a condition where a family can experience happiness, have a decent quality of life, and be sufficient in meeting the primary and secondary needs in family life [1].

Differences in family well-being are influenced by the quality of human resources. The level of education, the family's economic condition, and access to facilities and infrastructure are some factors that affect the quality of human resources [2]. Family well-being is also influenced by the family's economic condition, which is reflected in the head of the family's income, the family's social condition, depicted by the head of the family's education level and occupation, as well as the family's living conditions [3]. Sources of lighting, household consumption, the ability to use transportation facilities, and access to healthcare services are also factors that affect family well-being [4].

Submitted: February 02, 2024 Reviewed: July 25, 2024 Accepted: October 30, 2024
DOI: http://dx.doi.org/10.18860/ca.v9i2.28259

Based on these factors, family well-being can be determined by creating a predictive model. This model will be used to classify families as either well-off or not well-off. The methods used to classify the level of family well-being in this research are the K-Nearest Neighbors (KNN), K-Fold Cross Validation, and Bootstrap methods. The KNN method is chosen because the data used is secondary data, and the main purpose of this method is to classify new objects based on existing attributes and the training sample [5]. The KNN method predicts the category of new data that is added, in this case, families based on provinces, and determines whether they are closer to the well-off category or not. This KNN method is highly effective, productive, and easy to use for classifying data [6]. Additionally, this method is easy to implement, effective on large datasets, and fast in processing training data [7]. The Random Forest method is chosen because it can improve accuracy in situations where there is missing data and is robust against outliers. Additionally, Random Forest is efficient in terms of data storage. Another advantage is its ability in feature selection, which can enhance the performance of the classification model by selecting the best features. With this capability, Random Forest is suitable for handling big data with complex parameters efficiently [8]. Naive Bayes is a classification method that can estimate the probability of a class, allowing decision-making based on learned data. In the category of numerical-based approaches, the Naive Bayes method stands out due to its simplicity, speed, and high accuracy [9].

This study aims to compare the performance of three methods, namely KNN, Random Forest, and Naive Bayes Classifier, in classifying family well-being according to provinces in Indonesia based on underlying factors. This study also highlights the significant factors that are the main focus in determining the classification of family well-being. It is expected that the classification of family well-being will have positive implications for the formulation of effective policies. The state of the art in this field incorporates machine learning methods such as K-Nearest Neighbors (KNN), Random Forest, and Naive Bayes Classifier to develop robust predictive models. These methods are particularly effective for handling secondary data and complex datasets, enabling researchers to classify families as well-off or not based on criteria such as education level, economic status, access to healthcare, and living conditions.

METHODS

Data and Research Stages

The data used in this study is demographic and social statistics data from the years 2017-2021, obtained from the bps.go.id website. This data includes access to adequate sanitation, sources of lighting, ownership of BPJS (Social Security Administration for health), household consumption expenditure, educational attainment, the number of working heads of households, a source of drinking water, average housing area, access to healthcare services, and the number of motor vehicles. The total number of data in this study is 170, with 70% (119) used as training data and 30% (51) used as test data. The response variable (Y) used is the category of well-off and not well-off families according to provinces, grouped based on the household consumption expenditure per month in each province and 10 predictor variables as shown in Table 1.

Table 1. Predictor variables used

Code	Variables	Scale
X_1	Percentage of households by province with access to adequate sanitation	Ratio
X_2	Lighting Source	Ratio
X_3	Percentage of population with BPJS health insurance by province	Ratio
X_4	Level of Educational Attainment	Ratio
X_5	Percentage of Household Heads who are Employed	Ratio
X_6	Percentage of households by province, type of area, and source of adequate drinking water	Ratio
X_7	Percentage of households by province, type of area, and housing area	Ratio
X_8	Percentage of access to healthcare services	Ratio
X_9	Number of motor vehicles	Nominal

Data processing in this research utilizes RStudio and Microsoft Excel. The research steps are as follows:

1. Searching and collecting historical data in Table 1 from the bps.go.id website using Microsoft Excel.
2. Normalizing the data in Table 1 using RStudio.
3. Performing the Bartlett test and checking the suitability of variables with the KMO (Kaiser-Meyer-Olkin) test to assess multicollinearity in the data using RStudio.
4. Determining the correlation matrix, eigenvalues, and eigenvectors, and reducing data dimensions using Principal Component Analysis (PCA).
5. Modeling criteria for prosperous and non-prosperous families.
6. Classifying using KNN, Random Forest, and Naive Bayes methods.
7. Creating confusion matrices for each method.
8. Conducting an analysis by comparing the performance of each method by calculating accuracy, precision, recall, and F1-Score.

Principal Component Analysis

Principal component analysis is a technique used to explain a set of variables through several new variables that are mutually independent and linear combinations of the original variables. This analysis is used to reduce the dimensionality of data where the new variables are independent (uncorrelated) [10]. This analysis identifies which variables are most influential and thus capable of representing the data to be classified. The variables used are determined by the highest loading values in the principal component analysis.

Normalization of data

Normalization of data is the process of transforming data into a standardized or 'normal' form to facilitate data processing and analysis. Normalization is carried out by creating a matrix *x* of size $i \times j$where *i* is the amount of observation data and *j* is the number of variables. This activity transforms the original variables into standard normal form by calculating the mean and standard deviation of each variable using Equation (1):

$$Z_{ij} = \left[\frac{x_{ij} - \tilde{x}_j}{\sqrt{S_{jj}}}\right] \tag{1}$$

Description :

x_{ij} = Data value on the i^{th} row and j^{th} column

$\tilde{x}_j$ = Average of data in column j

$\sqrt{S_{jj}}$ = Standard deviation of data in column j

Determining the covariance matrix

The covariance matrix is a matrix used to calculate the extent of the relationship between two variables. The values of this covariance matrix will then be used to determine the eigenvalues and eigenvectors. The covariance can be seen in Equation (2):

$$S_{ab} = Cov\ (X_a, X_b) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ia} - \bar{x}_a)(x_{ib} - \bar{x}_b) \qquad a, b = 1,2, \ldots .p \tag{2}$$

From the equation above, the covariance matrix is obtained using Equation (3):

$$S = \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1p} \\ S_{21} & S_{22} & \ldots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \ldots & S_{pp} \end{bmatrix}_{p\times p} \tag{3}$$

Determining eigenvalues and eigenvectors

Eigenvalues are values that express the characteristics of a matrix, while eigenvectors are non-zero column matrices that, when multiplied by an $p \times p$ matrix, will result in multiples of the eigenvector itself. Eigenvalues can be calculated using the following Equation (4):

$$|S - \lambda I| = 0$$

$$\begin{vmatrix} S_{11} - \lambda & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} - \lambda & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} - \lambda \end{vmatrix} = 0 \tag{4}$$

Eigenvectors can be obtained using Equation (5):

$$(S - \lambda I)\ \boldsymbol{x} = \mathbf{0} \tag{5}$$

Determining principal components

Random vector of p, namely $X_1, X_2, \ldots .., X_p$, where the principal components of these variables are $PC_1, PC_2, \ldots .., PC_k$ obtained under the condition:

1. $k < p$
2. $\text{Var}(PC_1) > \text{Var}(PC_2) > \ldots > \text{Var}(PC_k)$
3. $\text{Cov}(PC_i, PC_j) = 0$, where $i \neq j$.

The principal components are obtained through a linear combination of the eigenvector matrix and the data matrix as shown in Equation (6):

$$\begin{aligned} PC_1 &= \sum_{j=1}^{p} a_{j1}X_j = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ PC_2 &= \sum_{j=1}^{p} a_{j2}X_j = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ &\vdots \\ PC_k &= \sum_{j=1}^{p} a_{jp}X_j = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p \end{aligned} \tag{6}$$

From equation (6), it can be expressed in matrix form as shown in equation (7):

$$\begin{bmatrix} PC_1 \\ PC_2 \\ \vdots \\ PC_k \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}_{p\times p} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}_{p\times 1} \tag{7}$$

Modeling Family Welfare

The Poverty Line is a minimum value of expenditure on food and non-food needs that must be met to avoid being categorized as poor. Based on this poverty line, Indonesian citizens with incomes below Rp 535,547 per capita are classified as unable [11].

However, because the data used in this study is household consumption expenditure data per province and there are outliers within it, family welfare can be modeled using the median, which is Rp 1,124,474 as the threshold between prosperous and non-prosperous families. A family is considered prosperous if its expenditure exceeds Rp 1,124,474. Meanwhile, a family is considered non-prosperous if its expenditure is less than Rp 1,124,474.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a method used to classify new objects based on old attributes and training samples. This method classifies objects based on data that is closest to the object. KNN belongs to supervised learning, where the result of a new query instance is classified based on the majority, and the class of the classification result is based on the class that appears most frequently [12]. The proximity in this method typically uses Euclidean distance with Equation (8):

$$D_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{8}$$

Description:

D_{xy} : Distance x and y

x_i : Training data

y_i : Testing data

n : The number of individual attributes

Random Forest

Random forest is a supervised learning that is an extension of the decision tree method where each tree in this method depends on the random vector values that are sampled freely and evenly. This method is commonly used for classification, regression, and so on [13]. Random forest is a method that can produce lower errors, is simple, and easily parallelizable, resulting in good classification outcomes. In addition, random forest can also be applied to big data with complex parameters as it can improve accuracy in cases of missing data, resist outliers, and is efficient for data storage [14]. This method is a modification of bagging by selecting *m* variables or adding random sub-sampling. Here is the random forest algorithm [15]:

1. Create a bootstrap sample Z of size *N* from the dataset.
2. Randomly select m variables from the p variables, where $m \leq p$. The value of m is chosen by approximating the square root of the total p variable.
3. Grow a random forest tree without pruning.
4. Repeat steps 1-3 *n* times to form a classification of *n* trees.
5. Once the trees are formed, the next step is to calculate the error rate to find the optimal mtry.
6. The class prediction is made based on the majority vote from the *n* trees.

The value of p represents the number of predictor variables used as splitters in forming the classification tree, where a higher value leads to a greater correlation.

Naive Bayes

Naive Bayes is a classification method discovered by Thomas Bayes in the 18th century. This method uses supervised techniques to predict the probability of a class for a characteristic that will be used in a particular dataset. This method has a high level of accuracy with simple calculations [9]. The data to be classified using naive Bayes follows Equation (9):

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \tag{9}$$

Description:

X : The characteristic resulting from the testing of the dataset

H : Hypothesis data X to be input into class

$P(H\|X)$	: Probability of hypothesis H given condition X
$P(X\|H)$	: Probability of X given condition H
$P(X)$	: Probability of X
$P(H)$	: Probability of hipótesis H

Confusion matrix

Accuracy calculation in data mining can be done using the confusion matrix method. This method consists of test boundaries predicted correctly or incorrectly by the data performed by the classification model. The confusion matrix is created by classifying into two classes, namely:

Table 2. Confusion matrix.

	Predicted Positives	**Predicted Negative**
Actual Positives Instances	Number of True Positives instances (TP)	Number of False Negatives instances (FN)
Actual Negatives Instances	Number of False Positives instances (FP)	Number of True Negatives instances (TN)

Explanation:

1. True Positives (TP) is the number of positive data classified as positive.
2. False Positives (FP) is the number of negative data classified as positive.
3. False Negatives (FN) is the number of positive data classified as negative.
4. True Negatives (TN) is the number of negative data classified as negative.

The values obtained through this method will be used to calculate the accuracy of the data to be tested [16]. Accuracy is a measure used to see the percentage of correctly classified or predicted data by the algorithm. This accuracy value can be calculated by comparing the number of correctly predicted data to the total number of data available using Equation (10):

$$\begin{aligned} accuracy &= \frac{\sum correct\ data}{(TP + TN + FP + FN)} \times 100\% \\ &= \frac{TP + TN}{(TP + TN + FP + FN)} \times 100\% \end{aligned} \tag{10}$$

Precision is a metric used to evaluate how many true positive predictions the model makes when predicting positive cases [17]. The precision value can be calculated using Equation (11):

$$Precision = \frac{TP}{(TP + FP)} \times 100\% \tag{11}$$

Recall is used to assess how often the model successfully predicts positive cases out of all actual positive examples [17]. The recall value can be calculated using Equation (12):

$$Recall \;=\; \frac{TP}{(TP\; +\; FN)} \times 100\% \tag{12}$$

F1-Score is used to evaluate the performance of a classification model on a dataset that has an imbalance between positive and negative examples. Precision and recall may provide less accurate information in such cases, but the F1-Score can provide a more balanced measurement between the two [17]. The F1-Score value can be calculated using Equation (13):

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{TP}{Precision\; +\; \; Recall}\right) \times 100\% \tag{13}$$

The initial understanding of the classification method's performance based on the values in the confusion matrix has a practical reference known as the rule of thumb for the confusion matrix, which includes:

1. If the number of TP and TN is high, while the number of FP and FN is low, it represents that the model has good performance.
2. If the number of FP generated is very high, it represents that the model has more false positive predictions or classifies too many positive labels.
3. If the number of TP generated is low, it represents that the model has more false negative predictions or failures that should be positive but are predicted as negative.
4. If the number of TP generated is low, it represents that the model correctly identifying positive labels has problems and requires improvement.
5. If the number of TN generated is low, it represents that the model correctly identifying negative labels has problems and requires improvement.

The rule of thumb for the confusion matrix provides only a rough understanding of the classification method's performance, so more detailed and accurate evaluation is needed to consider the context and purpose of using the method [17].

RESULTS AND DISCUSSION

Dimension Reduction with Principal Component Analysis

This study uses a total of 170 data points, of which 70% (119) are used as training data and 30% (51) as test data. Additionally, there are 10 predictors utilized in this research, the number of predictor variables is reduced using Principal Component Analysis (PCA) to make the data classification process more efficient. Before conducting Principal Component Analysis (PCA), a check is performed to see if there is multicollinearity using Bartlett's test of sphericity and the suitability of variables using the Kaiser-Meyer-Olkin (KMO) test. Bartlett's test of sphericity produces a value of $p - value\; <\; 2.2\; \times 10^{-16}$, the value is less than 0.05, indicating that there is multicollinearity in the variables and PCA analysis is needed. The KMO test produces a value of $p - value\; =\; 0.6106017$, the value is greater than 0.05, indicating that the data is sufficient for factor analysis and the model is appropriate.

Multicollinearity identified through Bartlett's Test of Sphericity can be addressed using the Principal Component Analysis (PCA) method. In this process, the number of factors to be retained is based on the eigenvalues obtained, where factors with eigenvalues greater than 1 will be prioritized. This approach allows for the simplification of the data structure by eliminating excessive correlations among variables, resulting in uncorrelated principal components that can maximize data representation. The following are the eigenvalues from the principal component analysis.

Table 3. Eigenvalues and proportion obtained

	Eigen vector								
	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9
Eigen values	3.08	1.62	1.19	1.00	0.66	0.56	0.47	0.27	0.16
Proportion	34.2%	18%	13.2%	11.2%	7.3%	6.2%	5.3%	3%	1.8%
Cumulative proportion	34.2%	52.2%	65.4%	76.6%	83.9%	90.1%	95.4%	98.4%	100%

From Table 3, it can be seen that the analysis shows there are four components with eigenvalues greater than 1, namely 3.08, 1.62, 1.19, and 1.00. These eigenvalues indicate the contribution of each component to the total variance explained in the data. The proportions generated by these four components are 34.2% for the first component, 18% for the second component, 13.2% for the third component, and 11.2% for the fourth component. Thus, the cumulative proportion produced by these four components reaches 76.6%, which is clearly above the 70% threshold commonly used in research. This indicates that these four main components are capable of capturing most of the information present in the initial data. Therefore, the number of components to be used in the subsequent analysis will be these four main components. It is expected that these four components are sufficiently representative to describe all initial predictor variables related to Family Welfare based on Province in Indonesia. With this approach, it is hoped that the analysis can be conducted more efficiently, without losing important information that could affect the prediction results. The loading values of the four main components can be seen in Table 4.

Table 4. Formulation of four main components

Initial Predictor Variables	Loading of Principal Component 1	Loading of Principal Component 2	Loading of Principal Component 3	Loading of Principal Component 4
X_1	0.50252*	-0.09659	-0.01521	0.17808
X_2	-0.02297	-0.65718*	0.22849	0.03384
X_3	0.02757	-0.17443	-0.75972*	0.27192
X_4	0.46927	0.01018	0.07365	0.27085
X_5	-0.43099	0.00357	0.14325	0.20743
X_6		0.06838	-0.21250	-0.00803

Initial Predictor Variables	Loading of Principal Component 1	Loading of Principal Component 2	Loading of Principal Component 3	Loading of Principal Component 4
X_7	0.44431	0.57647	-0.25476	0.19911
X_8	0.16401	0.31760	0.47705	0.53697
X_9	0.26109	0.30064	0.08246	-0.66990*

*Largest absolute value

In Principal Component Analysis (PCA), the term "loading" refers to the weight or contribution of each initial predictor variable in the formation of principal components. It reflects the extent to which the initial predictor variables influence the variation in the data. Each initial predictor variable is assigned a loading for each principal component generated by PCA. Loadings depict the importance of these variables in forming the principal components. The range of loading values is from -1 to 1, with the sign indicating the direction of the relationship between the initial predictor variables and the principal components. High loadings indicate significant contributions, while loadings close to zero indicate low or insignificant contributions. Variables with high and positive loadings on the first principal component have a significant influence in explaining the variation in the first principal component, while high and negative loadings indicate opposite effects. By analyzing loadings, the most important variables in explaining the variation in the data can be identified. Variables with high loadings are considered to have a significant influence on shaping the patterns and structure of the data.

Coefficients for each initial predictor variable are calculated based on the loading values on each principal component. High coefficients indicate that the corresponding loading values are also high, and vice versa. The greater the coefficient, the greater the influence of that variable in this classification process. The computation of coefficients PC_1, PC_2, PC_3, and PC_4 for each data point, this process is carried out sequentially following equation (6), which then results in the formation of Equations (14), (15), (16), (17):

$$PC_k = \sum_{j=1}^{p} a_{jp}X_j = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p \tag{6}$$

$$\begin{aligned} PC_1 = {} & 0.5025X_1 - 0.0230X_2 + 0.0276X_3 + 0.4693X_4 - 0.4310X_5 \\ & + 0.4443X_6 - 0.2185X_7 + 0.1640X_8 + 0.2611X_9 \end{aligned} \tag{14}$$

$$\begin{aligned} PC_2 = {} & -0.0966X_1 - 0.6572X_2 - 0.1744X_3 + 0.0102X_4 + 0.0036X_5 + 0.0684X_6 \\ & + 0.5765X_7 + 0.3176X_8 + 0.3006X_9 \end{aligned} \tag{15}$$

$$\begin{aligned} PC_3 = {} & -0.0152X_1 + 0.2285X_2 - 0.7597X_3 + 0.0736X_4 + 0.1433X_5 - 0.2125X_6 \\ & - 0.2548X_7 + 0.4770X_8 + 0.0825X_9 \end{aligned} \tag{16}$$

$$\begin{aligned} PC_4 = {} & 0.1781X_1 + 0.0338X_2 + 0.2719X_3 - 0.2709X_4 - 0.2074X_5 - 0.0080X_6 \\ & + 0.1991X_7 + 0.5370X_8 - 0.6699X_9 \end{aligned} \tag{17}$$

It can be observed that in PC_1, variable X_1, has the largest influence with a positive coefficient of 0.502518, whereas in PC_2, variable X_2, has the greatest influence with a negative coefficient of 0.65718. In PC_3, variable X_3 is the most influential with a negative coefficient of 0.75972. Additionally, in PC_4, variable X_9 is the most influential with a negative coefficient of 0.6699. Therefore, the variables that have the most significant impact on the classification process with the four principal components are the percentage of households by province with access to adequate sanitation, lighting source, the percentage of population with BPJS health insurance by province, and the number of motor vehicles.

Classification with K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) method utilizes the first four principal components from the independent variables to classify prosperous families, namely X_1, X_2, X_3, and X_9. In the application of this method, the value of K was set to 14, chosen based on an initial analysis showing that this value produced the lowest error rate among other K values in the range of 1 to 15, specifically 0.3653846. The use of $K = 14$ demonstrates effectiveness in determining the nearest neighbors, thus providing higher accuracy in classification. After the classification process is completed, the KNN method generates a confusion matrix that provides an overview of the model's performance in classifying the data. This matrix reflects the number of correct and incorrect predictions for each category and enables further evaluation of the effectiveness of the KNN method in the context of classifying family welfare. The resulting confusion matrix is as follows:

Table 5. Confusion matrix resulting from K-Nearest Neighbor (KNN)

Actual	Prediction results		Number of rows
	The family is not prosperous	**Prosperous family**	
The family is not prosperous	18	8	26
Prosperous family	11	15	26
Number of colums	29	23	52

Table 5 shows that there are 18 data points classified as TP (True Positive), meaning they were predicted to belong to the category of non-prosperous families and indeed belong to that category based on the actual classification. There are 8 data points classified as FN (False Negative), where the actual category is non-prosperous families but they were predicted as prosperous families. There are 11 data points classified as FP (False Positives), where the actual category is prosperous families but they were predicted as non-prosperous families. Additionally, there are 15 data points classified as TN (True Negatives), predicted to belong to the category of prosperous families and indeed belong to that category based on the actual classification.

Classification with Random Forest

The random forest method uses the first four principal components of the independent variables to classify prosperous families, namely X_1, X_2, X_3 dan X_9. This method yields an error of 0.3077. After the classification process is completed, the random forest method generates a confusion matrix that provides an overview of the model's performance in classifying the data. This matrix reflects the number of correct and incorrect predictions for each category and enables further evaluation of the effectiveness of the random forest method in the context of classifying family welfare. The resulting confusion matrix is as follows:

Table 6. Confusion matrix resulting from Random Forest

Actual	Prediction results		Number of rows
	The family is not prosperous	**Prosperous family**	
The family is not prosperous	17	9	26
Prosperous family	7	19	26
Number of colums	24	28	52

Table 6 shows that there are 17 True Positive (TP) data points, meaning the predicted category is prosperous families, which aligns with the actual category of prosperous families. There are 9 False Negative (FN) data points where the actual category is non-prosperous families but predicted as prosperous families. There are 7 False Positive (FP) data points where the actual category is prosperous families but predicted as non-prosperous families. Lastly, there are 19 True Negative (TN) data points predicted as non-prosperous families, which align with the actual category of non-prosperous families.

Classification with Naive Bayes Classifier

The Naive Bayes Classifier method uses the first four principal components of the independent variables to classify prosperous families, namely X_1, X_2, X_3 dan X_9. This method yields an error of 0. After the classification process is completed, the naive bayes classifier method generates a confusion matrix that provides an overview of the model's performance in classifying the data. This matrix reflects the number of correct and incorrect predictions for each category and enables further evaluation of the effectiveness of the naive bayes classifier method in the context of classifying family welfare. The resulting confusion matrix is as follows:

Table 7. Confusion matrix resulting from Naive Bayes Classifier

Actual	Prediction results		Number of rows
	The family is not prosperous	**Prosperous family**	
The family is not prosperous	25	1	26
Prosperous family	21	5	26
Number of colums	46	6	52

Table 7 shows that there are 25 data points classified as TP (True Positive), meaning the data predicted to belong to the category of non-prosperous families aligns with the actual category of non-prosperous families as well. There are 1 data points classified as FN (False Negative), where the actual category is non-prosperous families but predicted as prosperous families. There are 21 data points classified as FP (False Positives), where the actual category is prosperous families but predicted as non-prosperous families. Then, there are 5 data points classified as TN (True Negatives), predicted to belong to the category of prosperous families, aligning with the actual category of prosperous families as well.

Method Performance Comparison

The most effective method for classifying prosperous families can be identified by examining high values of accuracy, precision, recall, and F1 Score, which serve as key indicators of model performance. These values are obtained by calculating the elements of the confusion matrix resulting from the application of various classification methods, namely K-Nearest Neighbor (KNN), Random Forest, and Naive Bayes. This calculation provides a deep understanding of how well each method can accurately identify categories and the level of error present in predictions. To facilitate performance comparison, the accuracy, precision, recall, and F1 Score values from these three methods are summarized and presented in Table 8. This table offers a comprehensive overview of the strengths and weaknesses of each method in classification, helping to determine the most optimal method within the context of this study.

Table 8. Method Performance Comparison

Measure	K-Nearest Neighbor (KNN)	Random Forest	Naive Bayes Classifier
Accuracy	63.46%	69.23%	57.69%
Precision	62.07%	70.83%	54.35%
Recall	69.23%	65.38%	96.15%
F1- Score	65.46%	68.00%	69.44%

The accuracy, precision, recall, and F1 score values presented in Table 8 can be calculated using Equations (10), (11), (12), and (13) as follows:

$$\begin{aligned} Accuracy\ (KNN) &= \frac{18+15}{18+15+11+8} \times 100\% \\ &= 63.46\% \\ Precision\ (KNN) &= \frac{18}{18+11} \times 100\% \\ &= 62.07\% \\ Recall\ (KNN) &= \frac{18}{18+8} \times 100\% \\ &= 69.23\% \\ \frac{1}{F1} &= \frac{1}{2}\frac{18}{precision + recall} \times 100\% \\ &= 65.46\% \end{aligned}$$

According to Table 8, the highest accuracy value is obtained using the Random Forest methods, both at 69.23%. The highest precision value is obtained using Random Forest at 70.83%. The highest recall and F1-Score values are obtained using the Naive Bayes Classifier method at 96.15% and 69.44%, respectively. Based on the values shown, the Random Forest method is the most suitable for classifying prosperous families.

CONCLUSIONS

This study successfully determined the best method for classifying prosperous families based on household consumption expenditure per province and predictors obtained from the Badan Pusat Statistik (BPS) also known as Central Bureau of Statistics. In this research, 9 predictors were then reduced to four principal components determined from eigenvalues. For classifying test data, the analysis results from the KNN yield an 63.46% accuracy, 62.07% precision, 69.23% recall, and 65.46% F1-score. The analysis results from the random forest yield an 69.23% accuracy, 70.83% precision, 65.38% recall, and 68.00% F1-score. The analysis results from the KNN yield an 57.69% accuracy, 54.35% precision, 96.15% recall, and 69.44% F1-score. Random Forest method emerged as the best method among K-Nearest Neighbor (KNN) and Naive Bayes Classifier because it yielded higher values in accuracy and precision. For further research, we recommend using Artificial Neural Networks (ANN) method as it can handle data complexity, making it suitable for intricate classification tasks.

REFERENCES

[1] O. Sukmana, “Konsep dan Desain Negara Kesejahteraan (Welfare State),” *Jurnal Sosial Politik*, vol. 2, no. 1, hlm. 103, Sep 2017, doi: 10.22219/sospol.v2i1.4759.

[2] Y. Zebua, P. K. Wildani, A. Lasefa, dan R. Rahmad, “Faktor Penyebab Rendahnya Tingkat Kesejahteraan Nelayan Pesisir Pantai Sri Mersing Desa Kuala Lama Kabupaten Serdang Bedagai Sumatera Utara,” *Jurnal Geografi*, vol. 9, no. 1, hlm. 88–98, Feb 2017, doi: 10.24114/jg.v9i1.6923.

[3] R. Astika dan L. Harudu, “Faktor-Faktor Yang Mempengaruhi Tingkat Kesejahteraan Keluarga,” *Jurnal Penelitian Pendidikan Geografi*, vol. 8, no. 4, hlm. 2502–2776, 2023, doi: 10.36709/jppg.v8i4.94.

[4] D. P. Sari, W. Astuti, dan N. Dzulfikry, “Indikator dan Tingkat Keluarga Sejahtera menurut Dinas P3AP2KB Kabupaten Sambas,” *Ekodestinasi*, vol. 1, no. 1, hlm. 47–54, Mar 2023, doi: 10.59996/ekodestinasi.v1i1.38.

[5] Y. A. Setianto, K. Kusrini, dan H. Henderi, “Penerapan Algoritma K-Nearest Neighbour Dalam Menentukan Pembinaan Koperasi Kabupaten Kotawaringin Timur,” *Creative Information Technology Journal*, vol. 5, no. 3, hlm. 232–241, Sep 2018, doi: 10.24076/citec.2018v5i3.179.

[6] K. Taunk, S. De, S. Verma, dan A. Swetapadma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification,” dalam *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, Mei 2019, hlm. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[7] N. Bhatia dan V. Vandana, “Survey of Nearest Neighbor Techniques,” 2010. doi: 10.48550/arXiv.1007.0085.

[8] R. Supriyadi, W. Gata, N. Maulidah, dan A. Fauzi, “Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah,” *E-Bisnis : Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 13, no. 2, hlm. 67–75, Nov 2020, doi: 10.51903/e-bisnis.v13i2.247.

[9] A. Syarifah dan A. Muslim, “Pemanfaatan Naïve Bayes Untuk Merespon Emosi Dari Kalimat Berbahasa Indonesia,” *UJM*, vol. 4, no. 2, hlm. 147–156, 2015, [Daring]. Tersedia pada: http://journal.unnes.ac.id/sju/index.php/ujm

[10] J. R. Beattie dan F. W. L. Esmonde-White, “Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra,” *Appl Spectrosc*, vol. 75, no. 4, hlm. 361–375, Apr 2021, doi: 10.1177/0003702820987847.

[11] BPS, “Profil Kemiskinan di Indonesia Maret 2023,” Jakarta, Jul 2023.

[12] A. J. T, D. Yanosma, dan K. Anggriani, “Implementasi Metode K-Nearest Neighbor (Knn) Dan Simple Additive Weighting (Saw) Dalam Pengambilan Keputusan Seleksi Penerimaan Anggota Paskibraka,” *Pseudocode*, vol. 3, no. 2, hlm. 98–112, Jan 2016, doi: 10.33369/pseudocode.3.2.98-112.

[13] S. Basu, K. Kumbier, J. B. Brown, dan B. Yu, “Iterative random forests to discover predictive and stable high-order interactions,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, hlm. 1943–1948, Feb 2018, doi: 10.1073/pnas.1711236115.

[14] R. Supriyadi, W. Gata, N. Maulidah, dan A. Fauzi, “Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah,” *E-Bisnis : Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 13, no. 2, hlm. 67–75, Nov 2020, doi: 10.51903/e-bisnis.v13i2.247.

[15] S. Mahmuda, “Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube,” *Jurnal Jendela Matematika*, vol. 2, no. 01, hlm. 21–31, Jan 2024, doi: 10.57008/jjm.v2i01.633.

[16] P. Romadloni, B. Adhi Kusuma, dan W. Maulana Baihaqi, “Komparasi Metode Pembelajaran Mesin Untuk Implementasi Pengambilan Keputusan Dalam Menentukan Promosi Jabatan Karyawan,” *Jati (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 2, hlm. 622–628, Sep 2022, doi: 10.36040/jati.v6i2.5238.

[17] N. K. K. Ardana *dkk.*, “Perbandingan Metode KNN, Naive Bayes, dan Regresi Logistik Binomial dalam Pengklasifikasian Status Ekonomi Negara,” *Jambura Journal of Mathematics*, vol. 5, no. 2, hlm. 404–428, Agu 2023, doi: 10.34312/jjom.v5i2.21103.