Available online at www.coins-indocompt.org

COINS JOURNAL
Computing and Information System Journal

COmputing and INformation Systems Journal

IndoCompt PUBLISHER

E-ISSN: 3109-3248

Vol. 1, No. 3, December 2025

CLASSIFICATION OF THE SOCIOECONOMIC STATUS OF PROSPECTIVE GROOMS USING THE MODIFIED K-NEAREST NEIGHBOR (MKNN) ALGORITHM

Rosi Ernita Sari[a*], Lutfi Ali Muharom[b], Ginanjar Abdurrahman[c]

[a, b, c]*Muhammadiyah University of jember, Jember 68121, indonesi*

[*] rosiernita19@gmail.com

Abstract

Marriage is an important moment in life that is influenced not only by emotional aspects, but also by socioeconomic factors. The socioeconomic status of the prospective groom can affect the harmony of the household that will be built. This study aims to classify the socioeconomic status of prospective grooms using the Modified K-Nearest Neighbor (MKNN) algorithm and evaluate its performance through accuracy, precision, and recall measurements. The dataset used consists of 200 data points on prospective grooms obtained from the BKKBN (National Family Planning Agency) of Bondowoso District, with attributes including occupation, source of income, and income value. The classification process involves data pre-processing, Euclidean distance calculation, validation of training data, weighted voting, and K-fold Cross Validation. The test results showed that MKNN was able to provide good classification performance, with the highest accuracy of 88%, precision of 91.60%, and recall of 88% in a specific K-Fold scenario. This study shows that the MKNN algorithm is effective in classifying the socioeconomic status of prospective grooms and can be used as a reference for further research.

Keywords: Classification, socioeconomic status, prospective brides and grooms, modified k-nearest neighbor, MKNN

1. Introduction

Marriage is a social institution influenced not only by emotional factors but also by the socioeconomic conditions of each individual. One important factor often overlooked in the process leading up to marriage is the economic readiness of the prospective couple, particularly the prospective groom. Data from the Central Statistics Agency (BPS, 2024) shows that approximately 25% of divorce cases in Indonesia are caused by economic problems, highlighting the importance of financial stability in maintaining family harmony. Low socioeconomic conditions can directly affect family well-being, limiting access to education, reducing job opportunities, and lowering income levels. Therefore, mapping the socioeconomic status of prospective spouses can serve as a preventive measure to reduce potential domestic conflicts. Previous studies. have attempted to classify socioeconomic status using the Naïve Bayes algorithm with satisfactory results. However, to achieve more optimal classification performance, this study proposes the use of the Modified K-Nearest Neighbor (MKNN) algorithm. MKNN extends the standard K-Nearest Neighbor (KNN) method by incorporating training data validity and weighted voting, making it more adaptive to unbalanced data distributions. Similar approaches have been explored in various domains [1] demonstrated the effectiveness of MKNN in handling imbalanced public health datasets, [2] Springer Journal of Ambient Intelligence and Humanized Computing) showed the applicability of KNN-based methods in socioeconomic data analysis. In addition,[3] introduced MKNN as a robust alternative to KNN, improving accuracy through a combination of distance measurement and data validity assessment. Using job attributes, income sources, and income values from 200 prospective grooms in Bondowoso, this study evaluates the accuracy, precision, and recall of the MKNN algorithm. The results are then compared with those of the Naïve Bayes algorithm to determine which approach is more effective for socioeconomic classification. Novelty of the Study This research is the first to apply MKNN for classifying the socioeconomic status of prospective grooms using official BKKBN data from Bondowoso. Unlike previous studies, this work integrates both training data validity and weighted voting in a single framework to address class imbalance, while also conducting a direct performance

comparison with Naïve Bayes on the same dataset. This combination offers a new perspective on applying MKNN in the social domain and provides insights that can inform policymaking in family and social welfare planning.

2. Literature Review

2.1. Social economy

Socioeconomic status reflects a person's position in society based on their occupation, education, and income. Differences in socioeconomic status influence individuals' perspectives and attitudes toward certain issues. Social factors such as occupation and education are closely related to income, while favorable economic conditions can improve social status. Thus, socioeconomic status reflects an individual's ability to meet their daily needs.[4].

2.2. Classification

Classification is a technique for mapping unlabeled data into classes based on previously classified data [5]. The process involves three main stages: (1) designing a model using training data, (2) implementing the model to classify test data, and (3) evaluating the model to assess the accuracy and effectiveness of the classification results. This technique is important in data processing for pattern-based decision making.

2.3. K-Neares Neighbor

K-Nearest Neighbor (KNN) is a machine learning algorithm used to classify data based on the closest distance from a number of neighbors in the training data [6]. This algorithm identifies the K Nearest Neighbors of the test data using a distance metric, such as Euclidean Distance, to determine its class [7]. KNN can also be used for estimation and prediction in various data mining applications.

2.4. Modified K-Neares Neighbor

Modified K-Nearest Neighbor (MKNN) is an extension of the KNN algorithm that adds two main stages: validity calculation and weighting through the Weight Voting method [5]. MKNN determines the class of test data based on its proximity to validated training data, using distance methods such as Euclidean Distance or Cosine Similarity. The process begins by determining the value of K, calculating the distance between data points, validity, and Weighted Voting to determine the final classification [8]. This technique improves classification accuracy, especially when the data is imbalanced.

2.5. Training Data Validity Value

In the MKNN algorithm, each training data is validated against its nearest neighbors to obtain a validity value that is used in the classification process. This value becomes the basis for weighting when determining the test data class [5].

$$Validity\ (x) = \frac{1}{H} \sum_{i=1}^{H} s\ (lbl)(x), lbl\ (Ni(x)) \quad (1)$$

With:

Validity = Validity between training data

H = Number of closest neighbors

I = Best value is 1

lbl (x) = Class x label

lbl (Ni (x)) = Label the class closest to x

The S function is used to calculate the similarity between the test data and its nearest neighbor [9].

$$S\ (a, b) = \begin{cases} 1\ a = b \\ 0\ a \neq b \end{cases} \quad (2)$$

With:

S = Similarity

a = Class a in training data

b = Class other than a in training data

2.6. Euclidean Distance Calculation

Distance calculation in MKNN uses the Euclidean Distance formula between training data and test data. This method is used to determine the degree of similarity between data [10].

$$(xi, yi) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (3)$$

With:

d (x, y) = Euclidean distance between training data point x and test data point y
xi = Training data sample
yi = Test data
n = Attribute dimensions

2.7. Weighted Voting

In the MKNN method, the weight of each neighbor is calculated using formula 1/(de+0.5), where the Euclidean distance is. These weights are then multiplied by the validity value of each training data point. This process ensures that data points with closer distances and higher validity have a greater influence. This approach is effective for improving classification accuracy, especially in imbalanced data [8].

$$W(i) = \text{Validity}(i) \times \left(\frac{1}{d_e+0.5}\right) \quad (4)$$

With:

W = Weight of test data with training data i
i = Amount of training data
validity = Validity of training data
d_e = Training data distance

2.8. K-Fold Cross Validation

K-Fold Cross Validation is used to evaluate model performance by dividing the data into k equal parts [11]. Each iteration uses k–1 part as training data and one part as test data, performed alternately until all parts are used. For example, if the data is divided into D1, D2, and D3, each subset will take turns becoming the test data [12]. This approach makes the evaluation more balanced and accurate because all the data is utilized fairly in both training and testing.

2.9. Confusion Matrix

A *Confusion Matrix* is used to measure the accuracy of a classification model by presenting the number of correctly and incorrectly classified data in tabular form. This method is commonly used in data mining to evaluate model performance. Its components include True Positive, True Negative, False Positive, and False Negative, which help calculate metrics such as accuracy, precision, and recall [13].

Table 2.1 *Confusion Matrix*

Confusion Matrix		Prediction Class	
		1	2
Class Current	1	TP	FP
	2	FN	TN

Description:

- True Positive (TP) is the number of positive data correctly classified by the system.
- True Negative (TN) is the number of negative data correctly classified by the system.
- False Positive (FP) is the number of positive data incorrectly classified.
- False Negative (FN) is the number of negative data incorrectly classified by the system.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (5)$$

$$Presisi\ Weighted = \sum_{i=1}^{n}\left(\frac{S\,i}{S} \times \frac{TP_i}{TP_i+FP_i}\right) \times 100\% \quad (6)$$

$$Recall\ Weighted = \sum_{i=1}^{n} \left(\frac{S^i}{S} \times \frac{TP_i}{TP_i + FP_i}\right) \times 100\% \quad (7)$$

3. Method

3.1. Research Stages

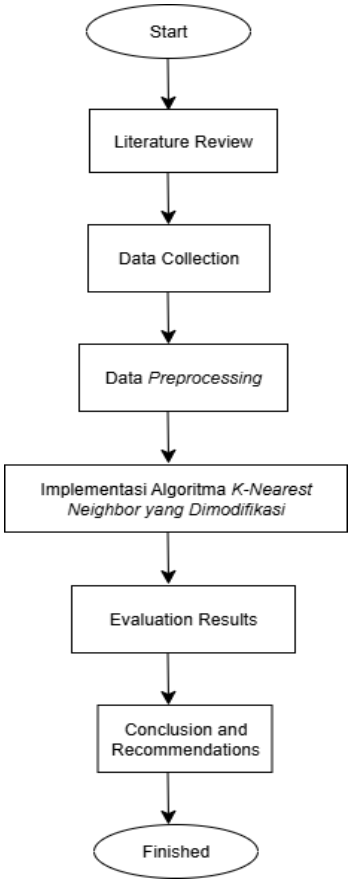

Figure 3.1. Research Stages

3.2. Literature Review

A literature review was conducted to collect and analyze relevant references in order to understand previous research developments. The aim was to identify research gaps and find a theoretical basis to support the application of the Modified K-Nearest Neighbor algorithm in this study.

3.3. Data Collection

This study uses secondary data from Firdaus & Muharrom (2024) in Bondowoso Regency, consisting of 238 data points, with attributes of occupation, source of income, and income value. The analysis was conducted using the Modified K-Nearest Neighbor (MKNN) algorithm.

3.4. Implementation of a *Modified K-Nearest Neighbor Algorithm*

The following is a flowchart illustrating the Modified K-nearest Neighbor Algorithm flow that will be used for classification.

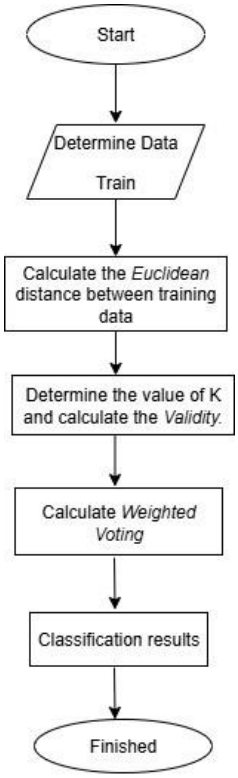

Figure 3.2. Implementation of a *Modified K-Nearest Neighbor Algorithm*

3.4.1 Determine Data Train

From a total of 238 data points, the researcher took 14 data points, namely from numbers 210 to 224, as examples for the calculations to be used.

3.4.2 Calculating the Euclidean distance between training data

Calculating the Euclidean distance between training data means determining the degree of similarity or difference between data points based on their attribute values.

3.4.3 Calculating the validity of training data

The validity of the training data is calculated using the S function, which has a value of 1 if the data class label and its neighbors are the same, and 0 if they are different, with K as the number of closest neighbors in the distance calculation.

3.4.4 Calcuate Weight Voting

Weight voting is a method of determining the final class by assigning weights to each neighbor, where the weight is calculated based on Euclidean distance and the validity value of the training data. Neighbors that are closer and more valid have a greater influence on the classification results.

4. Results And Discussion

4.1. Classification Results

In this classification process, the *Modified K-Nearest Neighbor* (MKNN) algorithm is used. Next is the process of calculating the *Modified K-Nearest Neighbor* classification results. This classification is used to determine the accuracy, precision, and recall values. With *K-Fold Cross Validation Fold* 2, *Fold* 4, *Fold* 5, and *Fold* 8. Using the values K=6, K=7, K=8, K=9, and K=10.

Table 4. 1 Overall Results of Accuracy, Precision, and Recall Values in K Testing

K-*Fold*	Nilai K														
	6			7			8			9			10		
	Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*
2-*fold scenario* 1	83.00%	77.95%	83.00%	82.00%	74.98%	82.00%	82.00%	74.98%	82.00%	83.00%	77.95%	83.00%	82.00%	74.98%	82.00%
2-*fold scenario* 2	75.00%	83.45%	75.00%	76.00%	84.89%	76.00%	74.00%	84.46%	74.00%	75.00%	85.93%	75.00%	76.00%	84.89%	76.00%
4-*fold scenario* 1	82.00%	70.29%	82.00%	82.00%	70.29%	82.00%	82.00%	70.29%	82.00%	82.00%	70.29%	82.00%	82.00%	70.29%	82.00%
4-*fold scenario* 2	84.00%	86.61%	84.00%	82.00%	77.33%	82.00%	82.00%	77.33%	82.00%	82.00%	77.33%	82.00%	82.00%	77.33%	82.00%
4-*fold scenario* 3	82.00%	83.06%	82.00%	80.00%	82.28%	80.00%	80.00%	82.28%	80.00%	80.00%	82.28%	80.00%	86.00%	83.65%	86.00%
4-*fold scenario* 4	70.00%	82.86%	70.00%	72.00%	83.24%	72.00%	82.00%	85.56%	82.00%	82.00%	85.56%	82.00%	82.00%	85.56%	82.00%
5-*fold scenario* 1	80.00%	67.69%	80.00%	80.00%	67.69%	80.00%	80.00%	67.69%	80.00%	80.00%	67.69%	80.00%	80.00%	67.69%	80.00%
5-*fold scenario* 2	80.00%	64.00%	80.00%	80.00%	64.00%	80.00%	80.00%	64.00%	80.00%	80.00%	64.00%	80.00%	80.00%	64.00%	80.00%
5-*fold scenario* 3	82.50%	87.40%	82.50%	82.50%	87.40%	82.50%	85.00%	88.04%	85.00%	82.50%	87.40%	82.50%	85.00%	88.04%	85.00%
5-*fold scenario* 4	85.00%	83.57%	85.00%	82.50%	78.22%	82.50%	82.50%	68.06%	82.50%	82.50%	68.06%	82.50%	82.50%	68.06%	82.50%
5-*fold scenario* 5	77.50%	83.25%	77.50%	80.00%	83.73%	80.00%	80.00%	83.73%	80.00%	80.00%	83.73%	80.00%	80.00%	83.73%	80.00%
8-*fold scenario* 1	80.00%	64.00%	80.00%	84.00%	86.67%	84.00%	80.00%	64.00%	80.00%	80.00%	64.00%	80.00%	80.00%	64.00%	80.00%
8-*fold scenario* 2	84.00%	77.00%	84.00%	84.00%	77.00%	84.00%	84.00%	77.00%	84.00%	84.00%	77.00%	84.00%	84.00%	77.00%	84.00%
8-*fold scenario* 3	76.00%	57.76%	76.00%	76.00%	57.76%	76.00%	76.00%	57.76%	76.00%	76.00%	57.76%	76.00%	76.00%	57.76%	76.00%
8-*fold scenario* 4	84.00%	84.00%	84.00%	80.00%	76.52%	80.00%	80.00%	76.52%	80.00%	80.00%	76.52%	80.00%	80.00%	76.52%	80.00%
8-*fold scenario* 5	68.00%	86.17%	68.00%	72.00%	91.60%	72.00%	72.00%	85.82%	72.00%	68.00%	85.17%	68.00%	72.00%	85.82%	72.00%
8-*fold scenario* 6	88.00%	89.50%	88.00%	88.00%	89.50%	88.00%	88.00%	89.50%	88.00%	88.00%	89.50%	88.00%	88.00%	89.50%	88.00%
8-*fold scenario* 7	80.00%	82.00%	80.00%	80.00%	82.00%	80.00%	84.00%	84.00%	84.00%	84.00%	84.00%	84.00%	84.00%	84.00%	84.00%
8-*fold scenario* 8	80.00%	89.00%	80.00%	84.00%	89.62%	84.00%	84.00%	89.62%	84.00%	84.00%	89.62%	84.00%	84.00%	89.62%	84.00%

4.1.1 Overall Results of Accuracy, Precision, and Recall Values in K Testing

Accuracy is the ratio of correct predictions to all data. The best results were achieved in scenario 6 for all K values (6–10) with an accuracy of 88%. Precision, which is the ratio of correct positive predictions to all positive predictions, is highest at K=7 in scenario 5 at 91.60%. Recall, which is the ability to detect all positive data, is highest in scenario 6 for all K values (6–10) at 88%.

Table 4.2 Test Results for K=6 to K=10

Value K	Akurasi	Presisi	Recall
6	88.00%	89.50%	88.00%
7	88.00%	91.60%	88.00%
8	88.00%	89.62%	88.00%
9	88.00%	89.62%	88.00%
10	88.00%	89.62%	88.00%

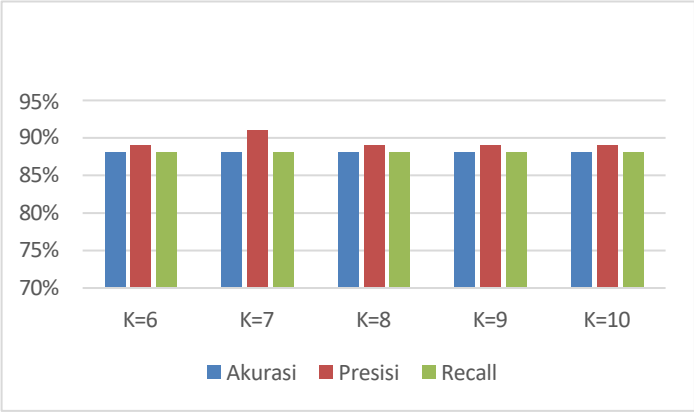

Figure 4.1 Classification Results

Based on the table of Accuracy, Precision, and Recall results for the socioeconomic status dataset of prospective grooms using the Modified K-Nearest Neighbor algorithm with K-Fold Cross Validation. By using various K values, namely K=6, K=7, K=8, K=9, and K=10, the search was able to produce an Accuracy percentage of 88%, Precision of 91.60%, and recall of 88%. In the experiment with K=7, the system used the classification results from the Confusion Matrix to achieve the highest values.

5. Conclusion

This study aims to classify the socioeconomic status of prospective grooms using the Modified K-Nearest Neighbor (MKNN) algorithm and compare its performance with the Naive Bayes Classifier algorithm. The test results show that the MKNN algorithm is capable of providing fairly good classification results, with the highest accuracy of 88% at K=6 in scenario 6. The highest precision was achieved at 91.60% at K=7 in scenario 5, while the highest recall reached 88% in the same scenario with the highest accuracy, namely K=6 in scenario 6.

When compared to the results of a previous study by Tito Alif Firdaus (2024) using the Naive Bayes Classifier algorithm, the accuracy was 82%, precision was 86%, and recall was 100%. This shows that the MKNN algorithm has advantages in accuracy and precision, while Naive Bayes is superior in recall. Thus, the MKNN algorithm is considered more balanced and adaptive to variations in the socioeconomic data of prospective male brides, making it a good alternative for classifying data with similar characteristics.

6. References

[1] M. R. Hunafa and A. Hermawan, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor pada Imbalace Class Dataset Penyakit Diabetes," *Media Online*, vol. 4, no. 3, pp. 1551–1561, 2023, doi: 10.30865/klik. v4i3.1486.

[2] M. Suyal and P. Goyal, "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning," *Int. J. Eng. Trends Technol.*, vol. 70, no. 7, pp. 43–48, 2022, doi: 10.14445/22315381/IJETT-V70I7P205.

[3] H. Parvin, H. Alizadeh, and B. Minaei-bidgoli, "MKNN: Modified K-Nearest Neighbor," no. May 2014.

[4] C. J. P. et al. Pangi, Joris, Jouke J. Lasut, "Kehidupan Sosial Ekonomi Petani di Desa Maliku Satu Kecamatan Amurang Timur Kabupaten Minahasa Selatan," vol. 13, no. 1, 2020.

[5] A. I. Pradana and V. Atina, "Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa," pp. 239–248, 2024, doi: 10.33364/algoritma/v.21-1.1618.

[6] E. Enjelina and V. P. Rantung, "Penerapan Algoritma K-Nearest Neighbor Untuk Clustering Kebutuhan Obat Berdasarkan Mutasi Laporan Bulanan Pada Dinas Kesehatan Kabupaten Minahasa," *Innov. J. Soc. Sci. ...*, vol. 3, pp. 6834–6841, 2023, [Online]. Available: http://j-innovative.org/index.php/Innovative/article/view/7048

[7] R. Rahmadini, Enjel Erika LorencisLubis, Aji Priansyah, Yolanda R.W.N, and Tuti Meutia, "Penerapan Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor," *J. Mhs. Akunt. Samudra*, vol. 4, no. 4, pp. 223–235, 2023, doi: 10.33059/jmas. v4i4.7074.

[8] I. Cholissodin, F. M. Evanita, J. J. Tedjasulaksana, and K. Wicaksono, "Klasifikasi Tingkat Laju Data Covid-19 Untuk Mitigasi Penyebaran Menggunakan Metode Modified K-

Nearest Neighbor the Classification of Covid-19 Data Rate for Distribution Mitigation Using Modified K-Nearest Neighbor (MKNN)," vol. 8, no. 3, pp. 595–600, 2021, doi: 10.25126/jtiik.202184400.

[9] N. Made, A. Pranasanthi, and I. M. Widiartha, "Implementasi Algoritma Modified K-Nearest Neighbor untuk Klasifikasi Indeks Kualitas Udara Perkotaan di Berbagai Negara," vol. 3, pp. 395–404, 2025.

[10] S. I. Sari, A. A. Suryanto, and A. Haryoko, "Pangan Non Tunai Dengan Menggunakan Metode Knn (K-Nearest Neighbor)," vol. 5, no. 1, pp. 10–21, 2024.

[11] A. L. Susetyo *et al.*, "Gender Di Indonesia Menggunakan Metode," 2024.

[12] B. Maal, G. Boosting, B. M. T. A. Permata, B. M. T. A. Permata, and K. Kunci, "Gradient Boosting Optimization with Pruning Technique for Prediction of," pp. 719–727, 2024.

[13] D. A. Rozzaq and L. A. Muharom, "Klasifikasi Kesehatan Calon Pengantin Menggunakan Algoritma Modified K-Nearest Neighbor (MKNN)," vol. 6, no. 2, pp. 103–109, 2024.