K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification

Zoelkarnain Rinanda Tembusai[1], Herman Mawengkang[2], *Muhammad Zarlis[3]
[1,2,3]Faculty of Computer Science and Information Technology, University of Sumatera Utara, Indonesia
[1]zoelkarnaintembusai@students.usu.ac.id, [2]hmawengkang@usu.ac.id, [3]m.zarlis@usu.ac.id*

Article Info

Article history:

Received Dec 18, 2020
Revised Dec 26, 2020
Accepted Jan 11, 2021

Keywords:

k-Nearest Neighbor
k-Fold Cross Validation
Analytic Hierarchy Process
Machine learning
Classification

ABSTRACT

This study analyzes the performance of the k-Nearest Neighbor method with the k-Fold Cross Validation algorithm as an evaluation model and the Analytic Hierarchy Process method as feature selection for the data classification process in order to obtain the best level of accuracy and machine learning model. The best test results are in fold-3, which is getting an accuracy rate of 95%. Evaluation of the k-Nearest Neighbor model with k-Fold Cross Validation can get a good machine learning model and the Analytic Hierarchy Process as a feature selection also gets optimal results and can reduce the performance of the k-Nearest Neighbor method because it only uses features that have been selected based on the level of importance for decision making.

Corresponding Author:

Muhammad Zarlis,
Department of Information Technology,
Faculty of Computer Science dan Information Technology
Sumatera Utara University,
Medan, Sumatera Utara.
Email: m.zarlis@usu.ac.id

1. INTRODUCTION

Data is a collection of various information or facts that contain information on a subject. The data classification leads to an artificial intelligence model that focuses on machine learning. Data classification is the grouping of objects into certain classes based on their values. Data classification is widely used in decision making.

Machine Learning is a very complex material [1]. There are many machine learning methods that can be used in the classification process, one of which is the k-Nearest Neighbor (k-NN). K-NN is a data classification method that does not require prior knowledge, where the new sample label is only determined by the nearest neighbor [2][3]. The k-NN method is also very simple and intuitive, easy to implement quickly, and one of the simplest and most popular of machine learning algorithms [4].

Li & Zhang in their research on music personalized recommendation, using the k-NN algorithm in collaborative filtering and taking advantage of the basic algorithm (k-NN) to be modified to make it more effective (k-NN improved) [5]. Meanwhile, Adege *et al.* in his research on indoor localization using the k-NN method and the backpropagation method obtained results that the k-NN method obtained better results than the backpropagation method [6].

In the classification process, the k-NN method requires features or criteria (features). Where each feature has its own value that defines certain classes, but too many features will slow down the performance of the k-NN method. The method proposed to reduce features is AHP (Analytic Hierarchy Process). Ren *et al.* in his research on the problem of selecting artificial intelligence strategies, the Analytic Hierarchy Process is proposed to complete the Multi-Criteria Group Decision-Making, so that many criteria in a problem can be weighted according to their importance in the decision making [7].

In data training, there is a method proposed to increase accuracy, namely k-Fold Cross Validation (k-FCV). Cross validation is statistical, it can be used in selecting a model to better predict predictive model test errors [8][9]. In their research, Caon *et al.* explained that k-Fold Cross Validation is the best technique that can be used in each case and completes the choice of method for further adaptation iterations [10].

From some of the explanations above, we will further analyze the performance of the k-Nearest Neighbor method with the k-Fold Cross Validation algorithm as a model evaluation by dividing training data and test data in order to obtain the best machine learning model and the Analytic Hierarchy Process method for feature selection from data classified based on the importance of each feature in decision making.

2. RESEARCH METHOD

2.1. Data Collection

Data collection was in the form of cervical cancer risk dataset (data) obtained from the UCI machine learning repository. This dataset focuses on the prediction of indicators or diagnosis of cervical cancer. These features include demographic information, habits, and historical medical records. In addition, literature studies and references to national and international journals are also needed to obtain additional knowledge related to theoretical foundations, analytical concepts, and methods in data classification.

2.2. Research

At this stage, the dataset is analyzed to obtain knowledge about the algorithms and methods analyzed, namely the k-Fold Cross Validation algorithm, the k-Nearest Neighbor method, and the Analytic Hierarchy Process method in classifying cervical cancer risk data and to determine the accuracy of the method used. used. The flow or process of this research, namely:

a. Data cleaning
   A machine learning model cannot directly process data found from multiple sources. There is a term Garbage In - Garbage Out which means the results of machine learning will be bad if the input is also bad. General things that can be done at the data cleaning stage include format consistency, data scale, data duplication, missing value, and skewness.

b. Data preparation
   Generally, some machine learning models cannot process categorical data, so it is necessary to convert categorical data into numeric data. This is called data preparation.

c. Data storage
   Processed data is entered into certain data stores so that it can be processed again at a later time, with the concept of the Relational Database Management System (RDBMS).

d. Data evaluation
   The evaluation data used in this study is the k-Fold Cross Validation. In cross validation, the dataset is divided by *k* folds. Where at each iteration each fold is used once as test data and the remaining fold is used as training data, the process is repeated until all data is evaluated.

e. Data classification
   After obtaining the distribution of training data and test data from the evaluation of the model with k-Fold Cross Validation, the data is classified with k-NN to get the accuracy of the model being built. This is repeated until the k-Fold is 10 or until it gets the highest level of accuracy.

f. Feature selection
Feature selection will remove features or attributes that don't really affect machine learning performance. The feature selection proposed in this research is the Analytic Hierarchy Process method.

g. Analysis
Then an analysis is carried out whether the model built and the results of the feature selection can improve the performance and accuracy of machine learning for data classification or not.

2.3. Analysis Method

The dataset used has 858 data and 36 attributes. With 30,888 records, the missing value was 3,622 records. After the data cleaning process is complete, the number of attributes that initially amounted to 36 became 34 attributes including labels (target). Two attributes have been omitted due to too much missing data.

a. Evaluate the k-NN model with k-Fold Cross Validation
Overview of the k-FCV process as an evaluation model in this study can be seen in Figure 1.

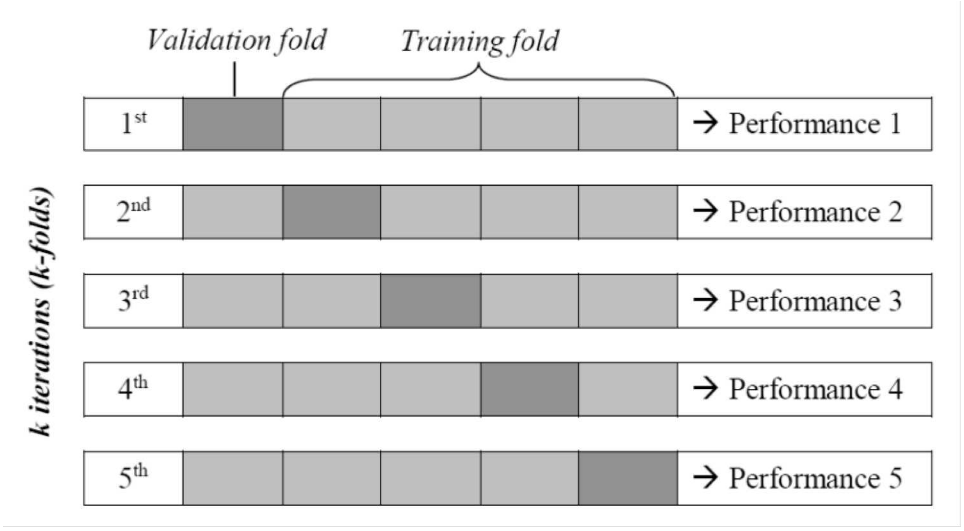


Figure 1. Model evaluation with k-Fold Cross Validation

In Figure 1 it can be seen the evaluation process of the k-NN method with the k-FCV. The number of folds that will be used in this study is 5 folds. Each fold will be tested using k-value k-NN, which varies from 3, 5, 7, and 9. Thus, the dataset will be divided into training data and testing data resulting from model evaluation with k-FCV.

After determining the training data and test data by evaluating the k-FCV algorithm, the next step is to start the data classification process using the k-NN method. The following are the research stages in the k-NN method:

1) Determine the value of *k*.
2) Calculate the Euclidean distance of the test data with the training data in the dataset.
3) Display the Euclidean distance ascending.
4) Take the smallest distance of *k*.
5) The results of data classification using the k-NN method.

b. Selection of features (attributes) with AHP
The steps taken in selecting the optimal attributes (features) to be used as a continuation of testing in this study are as follows:

1) Create a pairwise comparison matrix.

Table 1. The pairwise comparison matrix

Attribute	Age (1)	Number of sexual partners (2)	First sexual intercourse (3)	…	Citology (33)
(1)	1	3	5	…	3
(2)	0.3333	1	3	…	7
(3)	0.2	0.3333	1	…	3
…	…	…	…	…	…
(33)	0.3333	0.1429	0.3333	…	1

2) The sum of each attribute (feature).

Table 2. The sum of each attribute

Attribute	Age (1)	Number of sexual partners (2)	First sexual intercourse (3)	…	Citology (33)
Sum	11.1904	12.9476	18.1714	…	113

3) Criteria (attribute) value matrix.

Table 3. Attribute value matrix

Attribute	Age (1)	Number of sexual partners (2)	First sexual intercourse (3)	…	Citology (33)
(1)	0.0894	0.2317	0.2752	…	0.0265
(2)	0.0298	0.0772	0.1651	…	0.0619
(3)	0.0179	0.0257	0.055	…	0.0265
…	…	…	…	…	…
(33)	0.0298	0.011	0.0183	…	0.0088

4) Find the average of the row.

Table 4. Priority weight value

Attribute	Calculation	Weight
(1)	(0.0894 + 0.2317 + 0.2752 + … + 0.0265) / 33	**0.074**
(2)	(0.0298 + 0.0772 + 0.1651 + … + 0.0619) / 33	**0.069**
(3)	(0.0179 + 0.0257 + 0.055 + … + 0.0265) / 33	**0.059**
…	…	…
(33)	(0.0298 + 0.011 + 0.0183 + … + 0.0088) / 33	**0.008**

Table 5. Weight sum vector

Attribute	Calculation	Weight
(1)	(1*0.074) + (3*0.069) + ... + (3*0.008)	3.433
(2)	(0.3333*0.074) + (1*0.069) + ... + (7*0.008)	3.424
(3)	(0.2*0.074) + (0.3333*0.069) + ... + (3*0.008)	2.99
…	…	…
(33)	(0.3333*0.074) + (0.1429*0.069) + ... + (1*0.008)	0.347

Table 6. Consistency vector

Attribute	Calculation	Weight
(1)	(3.433 / 0.074)	46.5
(2)	(3.424 / 0.069)	49.96
(3)	(2.99 / 0.059)	50.45
…	…	…
(33)	(0.347 / 0.008)	45.63
	Rata-rata (*ƛ max*)	47.118

After the consistency vector value has been determined, it is necessary to calculate the values of two other things, namely lamda (X) and consistency index (CI) before the final consistency ratio can be calculated. The lamda value is the average value of the consistency vector.

CI ((*ƛ max* - *N*) / *N-1*) = ((47.118 - 33) / 33 - 1) = 0.4412

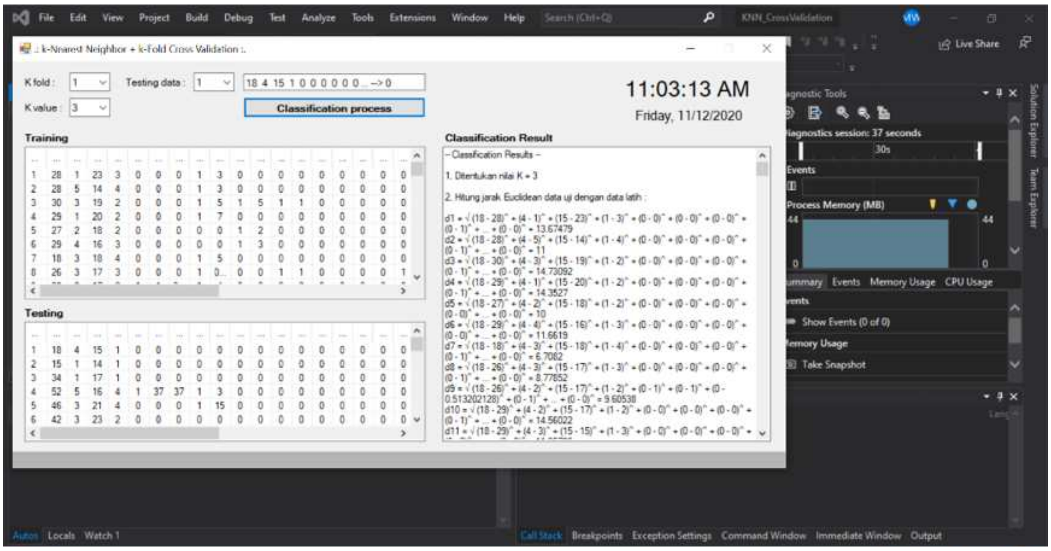
CR (*CI* / *IR*) = 0.4412 / 5.79 = 0.0762

Because CR <0.1, the consistency ratio of the calculation is acceptable. Therefore, the priority weight values in Table 4 can be used as parameter values for feature (attribute) selection.

3. RESULTS AND DISCUSSION

3.1. Test Result

In this test using 150 test data with varying *k* values, namely 3, 5, 7, and 9, as well as varying fold values, namely 1, 2, 3, 4, and 5. The attributes or features used in the test also vary, namely 33 features (default dataset) and 4 and 7 features (selected by AHP). The total test data are all 9000 times tested. The 4 features used in testing the selection results using AHP are age, number of sexual partners, first sexual intercourse, and num of pregnancies. While the 7 features of AHP selection results are age, number of sexual partners, first sexual intercourse, number of pregnancies, smokes, smokes (years), and smokes (packs / year). Testing is assisted by self-programmed applications. Figure 2 shows a display of the application used.



:: k-Nearest Neighbor + k-Fold Cross Validation ::

K fold : 1 Testing data : 1 18 4 15 1 0 0 0 0 0 0 ... --> 0

K value : 3

Classification process

11:03:13 AM

Friday, 11/12/2020

Training

```
1  28  1  23  3  0  0  0  1  3   0  0  0  0  0  0  0  0  0
2  28  5  14  4  0  0  0  1  3   0  0  0  0  0  0  0  0  0
3  30  3  19  2  0  0  0  1  5   1  5  1  1  0  0  0  0  0
4  29  1  20  2  0  0  0  1  7   0  0  0  0  0  0  0  0  0
5  27  2  18  2  0  0  0  0  0   1  2  0  0  0  0  0  0  0
6  29  4  16  3  0  0  0  0  0   1  3  0  0  0  0  0  0  0
7  18  3  18  4  0  0  0  1  5   0  0  0  0  0  0  0  0  0
8  26  3  17  3  0  0  0  1  0... 0  0  1  1  0  0  0  0  1
```

Testing

```
1  18  4  15  1  0  0   0   0  0   0  0  0  0  0  0  0  0  0
2  15  1  14  1  0  0   0   0  0   0  0  0  0  0  0  0  0  0
3  34  1  17  1  0  0   0   0  0   0  0  0  0  0  0  0  0  0
4  52  5  16  4  1  37  37  1  3   0  0  0  0  0  0  0  0  0
5  46  3  21  4  0  0   0   1  15  0  0  0  0  0  0  0  0  0
6  42  3  23  2  0  0   0   0  0   0  0  0  0  0  0  0  0  0
```

Classification Result

```
-- Classification Results --

1. Ditentukan nilai K = 3

2. Hitung jarak Euclidean data uji dengan data latih :

d1 = √(18 - 28)^ + (4 - 1)^ + (15 - 23)^ + (1 - 3)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 1)^ + ... + (0 - 0)^ = 13.67479
d2 = √(18 - 28)^ + (4 - 5)^ + (15 - 14)^ + (1 - 4)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 1)^ + ... + (0 - 0)^ = 11
d3 = √(18 - 30)^ + (4 - 3)^ + (15 - 19)^ + (1 - 2)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 1)^ + ... + (0 - 0)^ = 14.73092
d4 = √(18 - 29)^ + (4 - 1)^ + (15 - 20)^ + (1 - 2)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 1)^ + ... + (0 - 0)^ = 14.3527
d5 = √(18 - 27)^ + (4 - 2)^ + (15 - 18)^ + (1 - 2)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 0)^ + ... + (0 - 0)^ = 10
d6 = √(18 - 29)^ + (4 - 4)^ + (15 - 16)^ + (1 - 3)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 0)^ + ... + (0 - 0)^ = 11.6619
d7 = √(18 - 18)^ + (4 - 3)^ + (15 - 18)^ + (1 - 4)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 1)^ + ... + (0 - 0)^ = 6.7082
d8 = √(18 - 26)^ + (4 - 3)^ + (15 - 17)^ + (1 - 3)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 1)^ + ... + (0 - 0)^ = 8.77852
d9 = √(18 - 26)^ + (4 - 2)^ + (15 - 17)^ + (1 - 2)^ + (0 - 1)^ + (0 - 1)^ + (0 -
0.513202128)^ + (0 - 1)^ + ... + (0 - 0)^ = 9.60538
d10 = √(18 - 29)^ + (4 - 2)^ + (15 - 17)^ + (1 - 2)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
(0 - 1)^ + ... + (0 - 0)^ = 14.56022
d11 = √(18 - 29)^ + (4 - 3)^ + (15 - 15)^ + (1 - 3)^ + (0 - 0)^ + (0 - 0)^ + (0 - 0)^ +
```

Figure 2. Display data classification process

In Figure 2, you can see the display of the data classification process as a method of testing process. The test results will be presented in several tables. Table 7 shows the comparison of the results of the evaluation of the k-NN model with the k-FCV on the fold-3 and the results of feature selection with AHP (4 and 7 features).

Table 7. Fold-3 test result

Test to	*k* value	k-NN + k-FCV (default)	k-NN + k-FCV + AHP (4 features)	k-NN + k-FCV + AHP (7 features)
1	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
2	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
3	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
4	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
5	3	*True*	*True*	*True*

	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
6	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
7	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
8	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
9	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
10	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
11	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
12	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
13	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
14	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
15	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
…	…	…	…	…
150	3	*True*	*True*	*True*
	5	*True*	*True*	*True*
	7	*True*	*True*	*True*
	9	*True*	*True*	*True*
	Correct amount:	**571 records**	**569 records**	**569 records**
	Wrong amount:	**29 records**	**31 records**	**31 records**
	Percentage:	**95.2%**	**94.8%**	**94.8%**

In Table 7 it can be seen that the highest level of accuracy is also in the test with the default k-NN and k-FCV, namely with an accuracy rate of 95.2%, but the difference in the level of accuracy with 4 and 7 features is also not too far away, namely only 0.4% with 4 or 7 features.

3.2. Discussion

From the results of research and testing carried out 9000 times with varying *k* values and varying folds, it produces good accuracy. This shows that the k-NN model evaluation with the k-FCV algorithm is quite effective. As well as feature selection using the AHP method is also quite effective because it produces a level of accuracy that is not much different from the standard k-NN + k-FCV, but the advantage of having fewer features will optimize the performance of the k-NN method because the computation process can run lighter. Comparison of the level of accuracy of the test results has been visualized in graphical form in Figure 3.

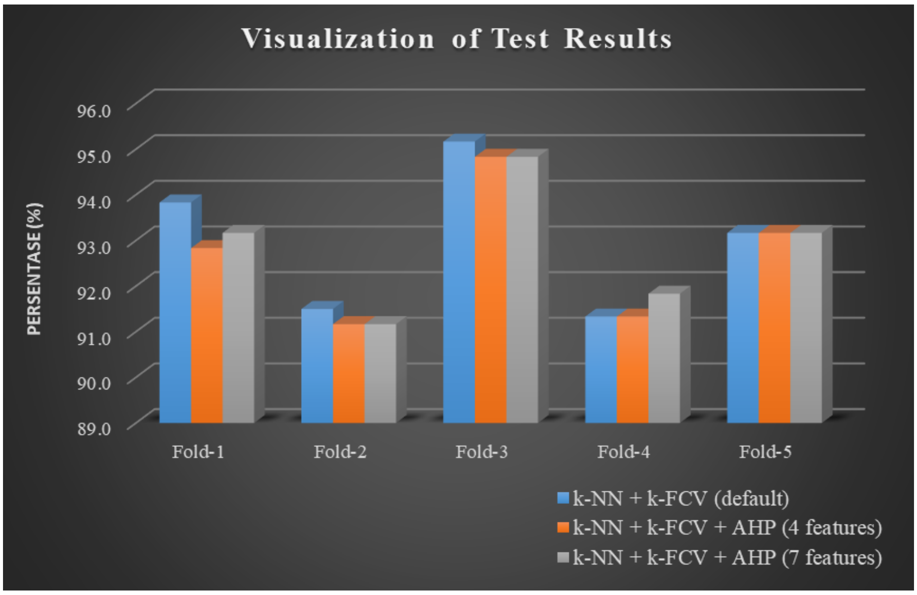


Figure 3. Comparison Graph of k-NN + k-FCV Accuracy Level with k-NN + k-FCV + AHP

In Figure 3 you can see the level of accuracy in each test with varying folds and *k* values. Assigning the correct *k* value is quite influential in the classification process, although it is not always successful in all test data. The highest level of accuracy is in the fold-3 which identifies that the fold-3 test is the best machine learning model in this study.

4. CONCLUSION

Based on the results of testing and analysis of the k-Nearest Neighbor method with the k-Fold Cross Validation algorithm as an evaluation model, it is proven that it can share training data and test data quite well. As well as the Analytic Hierarchy Process method for feature selection from classified data, it also gets an accuracy level that is almost the same as the classification without feature selection, namely by testing from fold-1 to fold-5, it always produces an accuracy rate above 90%. The best test is on fold-3, which is getting an accuracy rate of 95%. So it can be concluded that the evaluation of the k-Nearest Neighbor model with k-Fold Cross Validation can obtain a good machine learning model and the Analytic Hierarchy Process as a feature selection also gets optimal results and can reduce the performance of the k-Nearest Neighbor method because it only uses attributes (features) which have been selected based on the level of importance for decision making.

REFERENCES

[1] W.M. Lee, “Python® Machine Learning”, *Publish by John Wiley & Sons, Inc. ISBN: 978-1-119-54563-7*, pp. 1-296, 2019.

[2] D. Pan, Z. Zhao, L. Zhang, & C. Tang, “Recursive Clustering K-Nearest Neighbors Algorithm and the Application in the Classification of Power Quality Disturbance”. *IEEE Conference on Energy Internet and Energy System Integration (EI2)*, pp. 1-5, 2017. https://doi.org/10.1109/EI2.2017.8245652

[3] H. Jaafar, N. Mukahar, & D.A. Ramli, “Methodology of Nearest Neighbor: Design and Comparison of Biometric Image Database”, *IEEE Student Conference on Research and Development (SCOReD)*, pp. 1-6, 2016. https://doi.org/10.1109/SCORED.2016.7810073

[4] S. Du & J. Li, “Parallel Processing of Improved KNN Text Classification Algorithm Based on Hadoop”, *IEEE 7th International Conference on Information, Communication and Networks*, pp. 167-170, 2019. https://doi.org/10.1109/ICICN.2019.8834973

[5] G. Li & J. Zhang, “Music Personalized Recommendation System Based On Improved KNN Algorithm”, *IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 777-781, 2018. https://doi.org/10.1109/IAEAC.2018.8577483

[6] A.B. Adege & H. Lin, “Indoor Localization using K-Nearest Neighbor and Artificial Neural Network Backpropagation Algorithms”, *IEEE 27th Wireless and Optical Communications Conference (WOCC)*, pp. 1-2, 2018.

[7] Z. Ren, Z. Xu, & H. Wang, “The Strategy Selection Problem on Artificial Intelligence with An Integrated VIKOR and AHP Method under Probabilistic Dual Hesitant Fuzzy Information”, *IEEE International Journal*, pp. 1-23, 2017.

[8] P. Tamilarasi & U. Rani, “Diagnosis of Crime Rate Against Women using k-Fold Cross Validation through Machine Learning Algorithms”, *IEEE 4th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1034-1038, 2020. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000193

[9] M.R. Wayahdi, D. Syahputra, & S.H.N. Ginting, “Evaluation of k-Nearest Neighbor Model with k-Fold Cross Validation on Image Classification”, *Journal INFOKUM, Vol. 9, No. 1, ISSN: 2302-9706*, pp. 1-6, 2020.

[10] D.R.S. Caon, A. Amehraye, J. Razik, G. Chollet, R.V. Andreao, & C. Mokbel, “Experiments on Acoustic Model Supervised Adaptation and Evaluation by k-Fold Cross Validation Technique”, *IEEE International Journal*, pp. 1-4, 2010. https://doi.org/10.1109/ISVC.2010.5656264

[11] H. Mudia, "Back Propagation Neural Network for Controlling Coupled Water Tank", *Bulletin of Comp. Sci. & Electr. Eng.*, vol. 1, no. 1, pp. 12–18, Jun. 2020. https://doi.org/10.25008/bcsee.v1i1.4