Jurnal CoreIT, Vol.10, No.2, December 2024
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

The Implementation of Data Mining to Determine the Level of Students' Understanding in Utilizing E-Learning Using the K-Nearest Neighbor Method

Iwan Iskandar[1], Reski Mai Candra[1*]
Dept. of Informatics Engineering, Universitas Islam Negeri Sultan Syarif Kasim, Indonesia
iwan.iskandar@uin-suska.ac.id, reski.candra@uin-suska.ac.id

Abstract. The implementation of Information Technology is increasingly developing due to the growing demand. According to data obtained from the Indonesian Internet Service Providers Association (APJII) 2022 report, the number of internet users in Indonesia is 210.02 million, an increase of 27.9 million from the previous year. The application of E-Learning in various schools, campuses, and educational courses has been carried out. The utilization of e-learning media undoubtedly facilitates educators in transferring their knowledge to students. This research evaluates the level of understanding of each student who has used E-Learning during Covid-19 as a learning medium. In obtaining this level of understanding, the K-Nearest Neighbor (K-NN) method is applied. The data analyzed are based on assignment scores, quizzes, mid-term exams, and final exams from various related courses, namely Science and Mathematics Course Group, Programming Course Group, and Basic Informatics Course Group. A total of 1,627 data points were collected from the period between 2020 and 2021 when online learning was conducted using E-Learning. The data was processed using the KNN method with an 80:20 split between training and testing data. The analyzed K values were 3, 5, 7, 9, 11, 13, 15, 17, 19, and 21. The calculation results showed an accuracy of 75.69% at K=17 for the Basic Informatics Course Group, 77.61% at K=15 for the Science and Mathematics Course Group, and 96.20% at K=3 for the Programming Course Group.

Keywords: Course; E-Learning; Internet; K-Nearest Neighbor (KNN)

Received November 2024 / **Revised** November 2024 / **Accepted** December 2024

INTRODUCTION

The application of Information Technology is increasingly advancing driven by the growing demand. Information Technology infrastructure is established for the dissemination of communication technology that covers all regions worldwide, including Indonesia. Data obtained from the International Telecommunication Union (ITU) report states that the number of global Internet users in 2021 was 4.9 billion. Indonesia also experienced an increase in this number, as per the survey data from the Indonesian Internet Service Providers Association (APJII) in 2022 [1], with 210.02 million internet users, up by 27.9 million from the previous year. This number continued to rise from 196.71 million in the second quarter of 2019-2020 among Indonesia's population of 266.91 million people, signifying an increase of approximately 8.9% from the previous year, resulting in a 77.02% growth in Indonesia's internet penetration.

The recent surge in internet usage is closely linked to the impact of the Covid-19 pandemic, which necessitated individuals to conduct their activities from home. Consequently, all work and educational needs were fulfilled using the internet. Furthermore, the introduction of the Merdeka Belajar Kampus Merdeka (MBKM) curriculum by the Minister of Education, Culture, Research, and Technology, Mr. Nadiem Anwar Makarim, also contributed to the increased demand for online learning [2]. One of the 8 schemes of MBKM allows students to take courses outside their original campus, enabling them to engage in remote online learning from other campuses [3].

The implementation of E-Learning has been adopted by various schools, universities, and educational course providers. The utilization of e-learning media undoubtedly facilitates educators in transferring their knowledge to learners. Teaching processes, independent and group assignments, as well as evaluations through exams, can all be conducted using E-Learning platforms. However, despite the benefits of E-Learning, not all course materials can be fully comprehended by students. This remains a primary challenge

DOI: 10.24014/coreit.v10i2.33728

in the use of E-Learning. Not all lecturers or educators can effectively transfer knowledge to all students, evident from the examination scores obtained.

This research aims to assess the level of understanding of each student who has used E-Learning as a learning medium. To gauge this understanding, the research will employ the K-Nearest Neighbor (K-NN) method. This method was chosen due to its previously achieved high accuracy in several applied cases [4]. The K-NN model can be trained to prioritize correct predictions for underrepresented classes, thus improving overall accuracy [5] [6]. Another study on predicting students graduating on time stated that the KNN method has high accuracy [7] [8] [9] [10].

Based on this explanation, the objective of this research is to observe the entire process of student learning activities and evaluate efforts to improve the performance of the E-Learning application in the learning process. Data collection will involve gathering log data such as student grades for each course, the number of meetings, assignments, and the amount of teaching via streaming applications like Zoom, Google Meet, as well as the creation of educational video content on YouTube, and the E-Learning applications utilized. This data will be analyzed using the K-NN method to derive the level of student understanding. The results of this research are expected to serve as an evaluation material and contribute to the improvement and future actions in the E-Learning-based learning process. The outcomes of these actions will undoubtedly support data for the Program Accreditation process concerning new standards.

K- NEAREST NEIGHBORS

The K-NN (K-Nearest Neighbors) method is one of the classification methods that groups an object based on the learning data closest in distance to that object. The purpose of the K-NN method is to classify a new object based on attributes and training as samples. For example, there is a query point, which will then find a certain number, K, of objects or a training point closest to the query point. These neighbors will predict the value of the query based on classification [11] [12]. Before performing calculations with the K-NN method, the initial steps involve determining training data and test data. Then, the process involves calculating the shortest or nearest distance using Euclidean distance. Subsequently, the calculation step with the K-NN method will be carried out.

The K-NN method shares similarities with clustering techniques, which involve grouping new data based on the nearest neighbors or some data. The first step before determining the distance to neighboring data is to first establish the value of K neighbors. Then, to define a distance between two points, ($x1$, $y1$) in the training data and point ($x2$, $y2$) in the test data, Euclidean distance is used with the following formula [13] [14] [15] :

$$d(x,y)\sqrt{\sum i1(x_{ik}-y_{ik})^2} \quad (1)$$

Description:
$d(x,y)$ = Euclidean Distance
xik = the i-th value of variable k from x
yik = the i-th value of variable k from y

FOLD CROSS VALIDATION

K-Fold Cross Validation is a method used to determine the average success rate of a system. This method involves iterations through randomization of the dataset, testing the system with multiple randomly chosen input attributes. To assess the level of errors that occur, this technique can be employed. Each fold of the training data differs sufficiently from the original training data. In each iteration, one subset is reserved for testing data while the other subsets are used for training data.

In the process of the classification algorithm, new data is classified based on the proximity of the new data to the similar level of the closest existing data pattern. The value of K can be expressed as the determined number of nearest neighbors' data.

To estimate the K value effectively, the technique used is Cross Validation [16] [17]. Cross-validation is a validity test involving the use of comparative data to check the validity of an initial estimation. One of the

commonly used Cross Validation techniques to determine the optimal value for a classification algorithm is K-Fold Cross Validation. If a sufficient sample is provided, a larger K value will be more resistant to noise.

METHODS

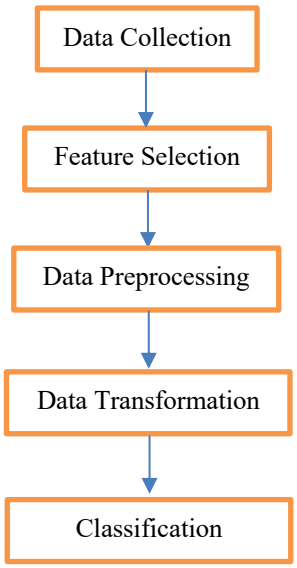

Fig. 1 Research Methodology

1) Data Collection

The data used in this study were obtained based on the scores of assignments, Quizzes, Mid-term Exams, and Final Exams The data were collected from several related courses, namely:

1. **Science and Mathematics Course Group**
   Includes the courses Calculus, Probability and Statistics, Discrete Mathematics, Linear Algebra, and Numerical Methods.
2. **Programming Course Group**
   Includes the courses Fundamental Programming, Algorithms and Programming (Alpro), Mobile Programming, Advanced Programming, Web Programming, and Web Application Development.
3. **Basic Informatics Course Group**
   Includes the courses Digital Systems, Databases, Theory of Languages and Automata, Interaction and Interface Design, Information Systems, Database Systems, Computer Networks, Operating Systems, Algorithm Strategies, and Programming Algorithms.

The total data comprises 1,627 records collected from the period 2020 to 2021, during which the learning process was conducted online using e-learning platforms.

2) Feature Selection

In this stage, a feature selection process is carried out to determine the variables used and those that are not used in calculations [18] [19]. The data obtained consists of several features: Name, ID Number (NIM), Assignments, Quizzes, Midterm Exams (UTS), and Final Exams (UAS). The features used in this research are grades from Assignments, Quizzes, Midterm Exams, and Final Exams.

3) Data Preprocessing

In this stage, data completeness is checked. If there are inconsistencies or outliers in the data, they are removed. After preprocessing, out of the total of 1.632 data points obtained, 1.627 data points remain. This reduction occurred due to some empty or missing data.

4) Data Transformation

This stage involves the transformation process to determine student understanding levels regarding passing a course. Table 1 below illustrates the levels of student understanding:

Table 1. Student Understanding Levels

RANGE	LETTER GRADE	NUMERICAL GRADE	DESCRIPTION		UNDESTANDING LEVEL
85 >	A	4	Pass	Very Good	Understand
80 - 85	A-	3,5	Pass	Good	
75 - 80	B+	3	Pass		
70 - 75	B	2,5	Pass		
65 - 70	B-	2	Pass		
60 - 65	C +	1,5	Pass	Fair	Fair
55 - 60	C	2	Pass		
50 - 55	D	1	Pass	Poor	Fail
<50	E	0	Fail	Fail	

Division of Training and Test Data, which is 80% and 20% respectively.

Table 2. Training Data 80 %

NO	HOME WORK	QUIZ	MID TEST	FINAL TEST	GRADE	UNDESTANDING LEVEL
1	100	90	45	60	B-	UNDERSTAND
2	100	65	45	45	C	FAIR
3	100	70	75	60	B	UNDERSTAND
4	100	85	60	50	B-	UNDERSTAND
5	100	70	45	60	C+	FAIR
….	100	75	60	50	C+	FAIR
….	100	40	55	45	C	FAIR
…..	100	50	70	40	C+	FAIR
1302	100	65	20	0	E	FAIL

Table 3. Testing Data 20 %

NO	HOME WORK	QUIZ	MID TEST	FINAL TEST	GRADE	UNDESTANDING LEVEL
1	100	75	80	90	A	UNDERSTAND
2	100	60	40	40	D	FAIR
…	100	60	40	40	D	FAIL
325	100	100	73	60	B+	UNDERSTAND

5) Data Mining

In this stage, the KNN (K-Nearest Neighbors) method is applied to determine students' understanding levels in the course.

a. Find the distance between neighbors using formula (1)

$$d(x,y)\sqrt{\sum i1}(x_{ik}-y_{ik})^2$$
$$= \sqrt{(100-100)^2+(90-75)^2+(45-80)^2+(60-90)^2}$$
$$= 48,4767986$$

Here is all the calculated data:

Table 4. Finding Euclidean Distance

NO	TESTING DATA 1	TESTING DATA 2	TESTING DATA 3	TESTING DATA 4
1	48,4767986	36,4005494	36,4005494	48,4767986
2	57,8791845	8,66025404	8,66025404	57,8791845
3	30,82207	41,5331193	41,5331193	30,82207
4	45,8257569	33,5410197	33,5410197	45,8257569
5	46,3680925	22,9128785	22,9128785	46,3680925
6	77,1362431	45	45	77,1362431
7	33,5410197	39,3700394	39,3700394	33,5410197
…	20,6155281	70,3562364	70,3562364	20,6155281
…	36,4005494	33,1662479	33,1662479	36,4005494

b. Enter the Value of K

Table 5. Enter the Value of K

Value of K					
Testing 1	**Testing 2**	**Testing 3**	**Testing 4**	**Testing 5**	**Testing 6**
3	5	7	9	11	13
			Understand		
			Understand		Understand
		Understand		Understand	Understand
				Understand	
		Understand	Understand	Understand	Understand
			Understand		Understand
Fail		Fail		Fail	Fail
		Understand		Understand	Understand
	Fair			Fair	
	Fail			Fail	
Fail		Fail			Fail
Fair		Fair		Fair	Fair
	Understand		Understand		
			Understand		Understand
	Fair		Fair		Fair
				Understand	
			Fair		Fair
		Fail		Fail	Fail
	Fair			Fair	Fair
			Understand		

RESULT AND DISCUSSION

In this research, an 80% - 20% split was used for training and testing data. Subsequently, experimentation was conducted with the values of K=3, K=5, K=7, K=9, K=11, K=13, K=15, K=17, K=19, K=21. The courses tested were divided into three major groups: Basic Informatics Courses, Science and Mathematics Courses, and Programming Courses. Table 6 explains the amount of training data and test data used in each category of courses.

Table 6. Amount of Training Data and Test Data

COURSE	TRAINING DATA	TESTING DATA
Basic Computer Science Course	720	181
Science and Mathematics Courses	265	67
Programming Course	314	79

Accuracy testing using the confusion matrix method. The following is the formula for confusion matrix used for each test:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}} X\ 100\%$$

Below are the accuracy results for each course group:

Table 7. Accuracy Results

VALUE OF K	ACCURACY		
	Basic Computer Science Course	Science and Mathematics Course	Programming Course
3	70,71%	56,71 %	96,20 %
5	73,48%	61,19 %	94,93 %
7	75,13%	62,68 %	89,87 %
9	75,13%	67,61 %	93,67 %
11	75,13%	65,67 %	93,67 %
13	75,13%	73,13 %	93,67 %
15	74,58%	77,61 %	93,67 %
17	75,69%	77,61 %	92,40 %
19	75,13%	77,61 %	91,13 %
21	75,13%	77,61 %	91,13 %

Here, Figure 2 illustrates the testing using the Confusion Matrix to determine the highest accuracy value 75,69% with K=17 for **Basic Computer Science Course**

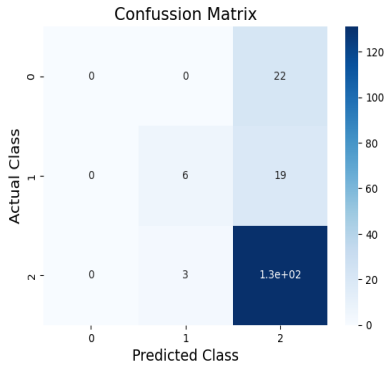

Fig 2. Confusion Matrix test used for Basic Computer Course

Here, Figure 3 illustrates the testing using the Confusion Matrix to determine the highest accuracy value 77,61% with K = 15 for **Science and Mathematics Course**

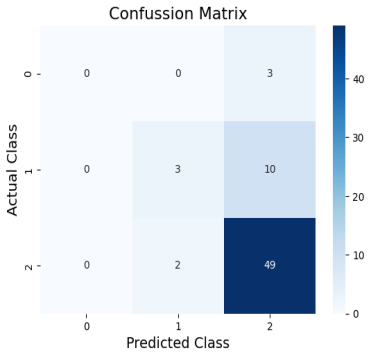

Fig 3. Confusion Matrix test used for Science and Mathematics Course

Here, Figure 4 illustrates the testing using the Confusion Matrix to determine the highest accuracy value 96,20% width K = 3 for **Programming Course**

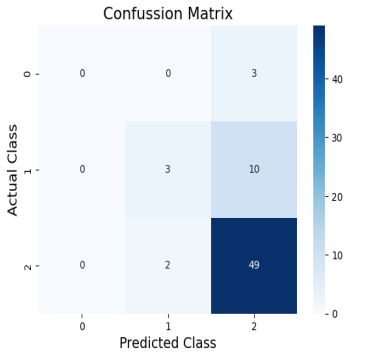

Fig 4. Confusion Matrix test used for Programming Course

CONCLUSION

From 30 experiments conducted on 1,627 data, it was concluded that the highest accuracy value was found in the Programming Course Group at K=3, with an accuracy of 96.20% and a training-to-testing data ratio of 80%: 20%. From these experiments, it can be observed that the KNN method has high accuracy in determining the level of student understanding in the courses. It can also be concluded that, based on the obtained data, e-learning has a high failure rate in the Programming course group. This is certainly a point of concern for the program, as these courses require special attention to improve the learning process.

REFERENCES

[1] A. P. j. I. I. (APJII), "Laporan Survei Internet APJII 2019-2020 Q2," APJII, Jakarta, 2021.

[2] B. R. d. T. R. I. Kementerian Pendidikan, "Kemdikbud Republik Indonesia," 2022. [Online]. Available: https://pusatinformasi.kampusmerdeka.kemdikbud.go.id/hc/id/articles/4417185050777-Apa-itu-Kampus-Merdeka. [Accessed 16 12 2023].

[3] K. R. d. T. Kementerian Pendikan, "Merdeka Belajar Kampus Merdeka," 2024. [Online]. Available: https://dikti.kemdikbud.go.id/wp-content/uploads/2024/06/Buku-Panduan-Merdeka-Belajar-Kampus-Merdeka-MBKM-2024.pdf.

[4] A. M. T. A. M. Hazem, "Efficient Computational Cost Reduction in KNN through Maximum Entropy Clustering," in *icci*, 2024.

[5] R. U. A. A. R. U. R. Debarshi, "A Comprehensive Study of the Performances of Imbalanced Data Learning Methods with Different Optimization Techniques," in *Communications in computer and information science*, 2024, pp. 209-228.

[6] R. A. T. T. M. S. F. S. Tiara, "Model algoritma knn untuk prediksi kelulusan mahasiswa stikom cki," *Jurnal Ilmiah Informatika & Komputer,* vol. 29, no. 2, p. 11803, 2024.

[7] A. J. S. I. N. S. I. G. A. G. Gd., "Improving k-nearest neighbor performance using permutation feature importance to predict student success in study," *ndonesian Journal of Electrical Engineering and Computer Science,* vol. 35, no. 3, pp. 1835-1844, 2024.

[8] A. H. R. K. K. Shandy, "Implementation of Data Mining for Predicting Student Graduation Using the K-Nearest Neighbor Algorithm at Jambi Muhammadiyah University," *Indonesian Journal of Artificial Intelligence and Data Mining,* vol. 7, no. 1, p. 26150, 2024.

[9] A. M. K. Iqlimah, "Comparison of Classification Algorithms for Predicting Graduation of Informatics Engineering Students with Orange Data Mining," *Indonesian Journal of Computer Science,* vol. 13, no. 2, p. 3796, 2024.

[10] T. N. Văn, " Using Machine Learning models to predict the on-time graduation status of students," *Tạp chí Khoa học và Đào tạo Ngân hàng,* p. 2506, 2023.

[11] N. N. S. B. D. Dzikrulloh, "Penerapan Metode K – Nearest Neighbor (K-NN) dan Metode Weighted Product (WP) Dalam Penerimaan Calon Guru Dan Karyawan Tata Usaha Baru Berwawasan Teknologi (Studi Kasus : Sekolah Menengah Kejuruan Muhammadi," 2017.

[12] D. A. S. O. S. W. Saputri, "Implementasi Data Mining Menggunakan Metode K-Nearest Neighbor Untuk Menentukan Stok Obat Obatan Pada Apotek: Studi Kasus Apotek Salaam," *Dinamika Informatika,* 2016.

[13] R. L. M. L. A. K. Arif, "Optimization of distance formula in K-Nearest Neighbor method," *Bulletin of Electrical Engineering and Informatics,* vol. 9, no. 1, p. 1464, 2020.

[14] Z. E. B. D. P. Enrico, "A quantum k-nearest neighbors algorithm based on the Euclidean distance estimation," *Quantum Machine Intelligence,* 2024.

[15] W. I. N. P. T. S. L. S. M. F. D. W. Wahyono, "Perbandingan penghitungan jarak pada k-nearest neighbour dalam klasifikasi data tekstual," *Jurnal Teknologi dan Sistem Komputer,* pp. 54-58, 2020.

[16] I. A. W. M. N. U. S. Urwah, "Examining the Impact of Different K Values on the Performance of Multiple Algorithms in K-Fold Cross-Validation," 2023.

[17] F. O. Opeoluwa, "Determining the optimal number of folds to use in a K-fold cross-validation: A neural network classification experiment," *Research in mathematics,* 2023.

[18] T. P. R., "Nonhypothesis-Driven Research: Data Mining and Knowledge Discovery," *Computers in health care,* 2023.

[19] M. G. A. H. Mai, "Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects," *The international journal of construction management,* 2020.