International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 7, No. 3, (2023), pp. 211-216

Prediction of Student Graduation using the K-Nearest Neighbors Method

Embun Fajar Wati[1], Elvi Sunita Perangin-Angin[2], Anggi Puspita Sari[3]
[1,2,3]Universitas Bina Sarana Informatika, Indonesia
Email: [1]embun.efw@bsi.ac.id, [2]elvi.evt@bsi.ac.id, [3]anggi.apr@bsi.ac.id

Abstract

Predictions on the accuracy of student graduation are designed to support study programs in guiding students so that they can graduate on time. The number of student graduations will influence the university's accreditation score. Graduation predictions can provide very useful information in decision-making; therefore, research was conducted on student graduation data. This data will be processed using the K-Nearest Neighbor method. The dataset used consisted of 150 students majoring in informatics engineering. The variables included gender, age, marital status, grade, and job status. The research methodology used in this study consists of 6 stages: Data Collection, Data Selection, Preprocessing, Transformation, Testing, and Evaluation. In the preprocessing or cleaning stage, the data can be fully utilized because all fields have been filled in correctly. Meanwhile, in the transformation stage, the data is categorized as follows: age (young: 19-24, old: 25-50) and grade (large: 3-4, small: 1-2.9). The K-Nearest Neighbor (KNN) method can predict student graduation rates. The KNN method, processed with the RapidMiner 9.9 tool, obtained an average accuracy of 100%. Based on the results of 100% accuracy and an AUC value of 1, it can be concluded that the KNN method is highly accurate in classifying graduation data for the 150 students.

Keywords: *K-Nearest Neighbor, Graduation, On time, Accreditation, Students*

1. Introduction

In the education system, students are a valuable asset for educational institutions [1]. One way to evaluate the effectiveness of an educational institution in cultivating high-caliber individuals is by examining the academic performance of its students [2]. Students are commonly recognized as a cohort with more extensive intellectual attributes than individuals of their same age who are not enrolled in educational institutions or those in younger age brackets [3]. In order to cultivate capable, insightful, competitive, and creative human resources, educational institutions must offer high-quality education to their students [4].

Every year, the Admission of New Students at the tertiary level is a crucial activity that marks the beginning of the new academic year. However, a significant issue arises as not all students are able to graduate on time as per the prepared curriculum, ultimately impacting the accreditation of higher education institutions [5]. Every year, the growing number of students leads to a significant accumulation of student data. To ensure study programs effectively guide students towards timely graduation, accurate predictions regarding their graduation status are crucial. By monitoring students' predicted graduation outcomes throughout their lectures, the study program, with the assistance of academic supervisors, can prioritize special attention towards those who are projected to graduate late. This proactive approach enables these students to enhance their academic performance each semester, ultimately enabling them to graduate on time [6]. The amount of time students spend studying is indicative of their level of academic performance [7]. In order to address this issue, it is necessary to develop a model that can accurately predict student graduation. One of the data mining classification methods, specifically K-Nearest Neighbor, can be utilized for this purpose [8]. Predicting student graduation can yield valuable insights from extensive student data [9].

ISSN : 2580-7250

Information Technology (IT) refers to the use of technology for processing data. This includes tasks such as obtaining, compiling, storing, and manipulating data in order to generate high-quality information. The information produced through IT should be relevant, accurate, and timely, serving personal, business, governmental, and strategic purposes. It plays a crucial role in decision-making processes. [10]. The need for data mining arises from the abundance of big data, which can be effectively utilized to extract valuable information and insights, ultimately generating new knowledge [11]. Data mining is a form of data exploration and analysis that involves the automatic extraction of patterns and features from extensive amounts of data [12]. Data mining involves the utilization of specific algorithms to extract patterns from data. It employs mathematical algorithms to segment data and assess the likelihood of various predetermined outcomes [13]. The K-Nearest Neighbor algorithm, utilizing the Euclidean Distance method for distance measurement, proves to be highly beneficial in addressing time efficiency concerns and effectively predicting graduation outcomes [14]. The k-nearest neighbor algorithm is a powerful data classification technique that involves searching for similar cases by calculating the proximity between new and old cases using matching weights. [6].

2. Research Methodology

The research methodology employed in this study consists of six stages, namely:

a. Data Collection
b. Data Selection
c. Preprocessing
d. Transformation
e. Testing
f. Evaluation

2.1. Data Collection

The data collection process involved gathering information from students enrolled in the Informatics Engineering program. The data collection methods were tailored to meet the specific requirements of this research.

2.2. Data Selection

The data selection stage is crucial for obtaining relevant information to accurately predict student graduation. In order to predict student graduation, it is necessary to gather data specifically from students who have already graduated. During the data selection stage, a total of 150 data points were obtained, which will be utilized as training data. The student data is divided into two categories: personal data (including gender, age, marital status, and job status) and academic data (grades). According to the graduation data, 203 students graduated on time, while 97 students graduated late [15].

2.3. Preprocessing

This stage involves a data cleaning process to ensure the usability of previously obtained data by eliminating duplication, errors, and ensuring appropriate validation rules. Additionally, missing values and noise are addressed to prevent any issues during processing, ensuring accurate results. The preprocessed data consists of a total of 150 complete entries.

2.4. Transformation

This stage involves transforming the table contents to match the selected data, ensuring that the selected data aligns with the process being carried out. All the data

is complete and there are no empty values. The transformed data includes age categories (young: 19-24, old: 25-50) and grade levels (large: 3-4, small: 1-2.9).

2.5. Testing

The data was tested using the k-Nearest Neighbor method. The new dataset was divided into two parts, namely training data and testing data, using 10-fold cross-validation. The training data was then classified using K-Nearest Neighbor to determine the accuracy in classifying students' graduation rates. The testing phase utilized validation techniques to obtain accuracy values. The dataset was tested with the proposed method using the RapidMiner 9.9 application.

2.6. K-Nearest Neighbor (K-NN)

The KNN algorithm is a classification algorithm that falls under the category of supervised learning algorithms. It operates by classifying an object based on the distances to its nearest neighbors. It is worth noting that multiple similarities can exist within the data, allowing KNN to classify both a set of k similar data points and data points with numerous similarities. The distance metric employed by KNN is known as the Euclidean Distance. In addition to its simplicity in determining the shortest distance between data points, this algorithm possesses the advantage of being able to generalize from a relatively small training data set [5].

2.7. Evaluation

The test results from the K-Nearest Neighbor method reveal the average level of accuracy. The accuracy level is influenced by the clustering of data, which ultimately helps achieve the goal of predicting graduation rates for students based on various categories.

3. Results and Discussion

Data collection was conducted using a dataset comprising 150 informatics engineering students. The variables employed in data processing through the K-Nearest Neighbor (K-NN) approach include gender, age, marital status, grade, and job status.

The data selection stage is crucial for obtaining relevant information to accurately predict student graduation. To effectively predict student graduation, it is essential to gather data specifically on students who have successfully graduated [1]. During the data selection stage, a total of 150 data points were obtained. These 150 data points will be utilized as training data.

The preprocessing stage, also referred to as clearing, is a crucial data cleaning process. It is performed to ensure that previously acquired data is usable, devoid of duplicates, errors, and adheres to validation rules [1]. Out of the 150 data points used, all were complete with no empty fields, ensuring that all the data was utilized.

The subsequent phase involves transformation, which refers to altering the table's format to match the selected data, ensuring its alignment with the ongoing process. [1]. The obtained data is currently in its raw form and needs to be processed before further analysis. A total of 150 data points were transformed into two categories: age (young: 17 years - 21 years, old: 22 years - 40 years) and grade (high: 3 - 4, low: 1 - 2.9). The transformed data is presented in table 1 [15].

Table 1. Student Dataset

Gender	Age	Marital Status	Grade	Job Status
Male	Young	Married	Low	Employee
Female	Young	Single	Low	Employee

Male	Young	Married	Low	Unemployee
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
Male	Old	Married	Low	Employee
Male	Young	Single	Low	Employee
Female	Young	Married	Low	Unemployee

The aim of this research was to determine the accuracy value of student performance through graduation rates using the K-Nearest Neighbor (KNN) method. RapidMiner software was utilized for data processing and implementation of the KNN method. The research involved testing the KNN method multiple times, specifically 10 times, with 10-fold Cross Validation [4]. The following text describes data processing using the KNN method with RapidMiner 9.9 tools.

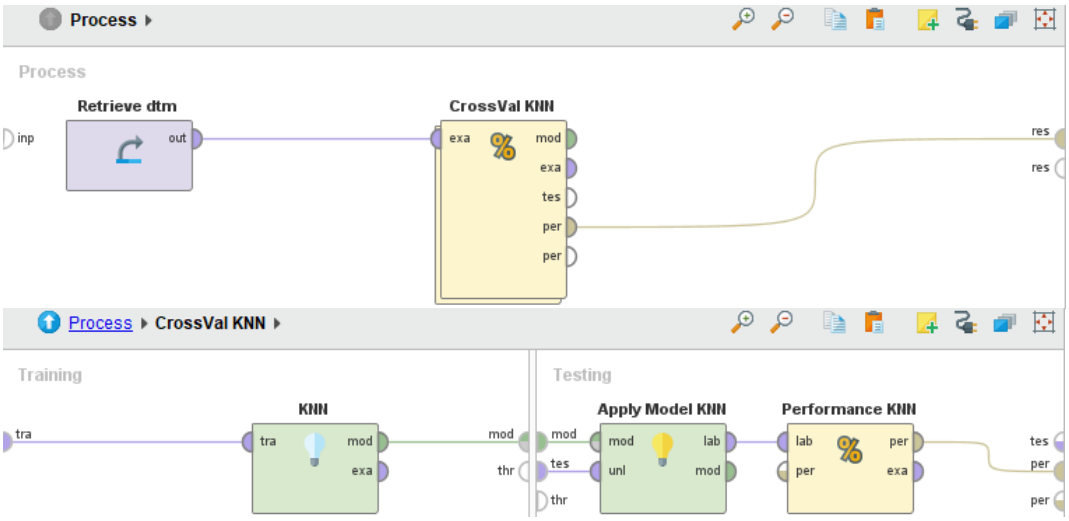

Figure 1. Data processing in RapidMiner using the K-Nearest Neighbors (KNN) method

The analysis of the results of the K-Nearest Neighbor (KNN) Model with Cross Validation and Confusion Matrix is presented in Figure 2, showcasing the accuracy results of the K-Nearest Neighbor (KNN) algorithm. Additionally, Figure 3 displays the results of the ROC curve. The K-Nearest Neighbor (KNN) model exhibits a remarkable accuracy level of 100% and the ROC curve demonstrates an AUC value of 1.

accuracy: 100.00% +/- 0.00% (micro average: 100.00%)

	true Tepat	true Terlambat	class precision
pred. Tepat	101	0	100.00%
pred. Terlambat	0	49	100.00%
class recall	100.00%	100.00%	

Figure 2. Confusion Matrix

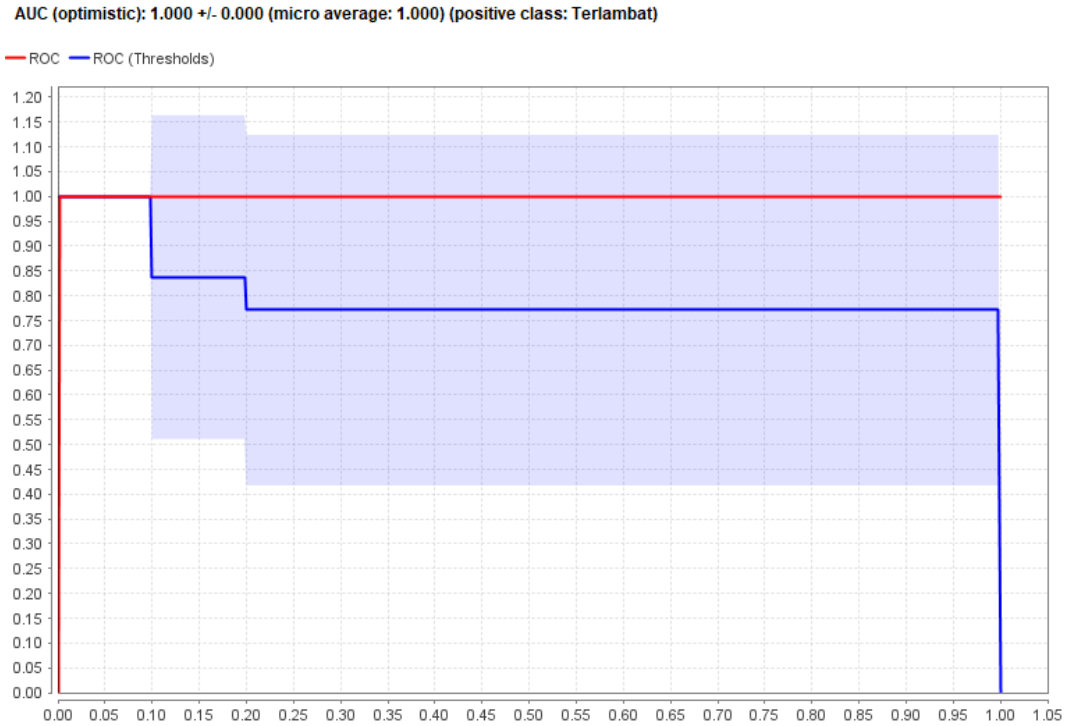

Figure 3. ROC curve

4. Conclusion

The K-Nearest Neighbor (KNN) method has shown promising results in predicting student graduation rates. By utilizing the RapidMiner 9.9 tool, the KNN method achieved an impressive average accuracy of 100%. This high accuracy is further supported by the AUC value of 1, indicating the method's exceptional ability to classify graduation data for a sample of 150 students.

Moreover, it is important to note that the accuracy of testing the student graduation model using the K-Nearest Neighbor (K-NN) algorithm is influenced by the extent of data clustering. Therefore, it is recommended for future research to conduct a comparative analysis between the classification method and the clustering method. This will help determine the accuracy value associated with utilizing both algorithmic approaches.

References

[1] E. S. Susanto, Kusrini, And H. Al Fatta, "Prediksi Kelulusan Mahasiswa Magister Teknik Informatika Universitas Amikom Yogyakarta Menggunakan Metode K-Nearest Neighbor," *J. Teknol. Inf.*, Vol. Xiii, No. 2, Pp. 67–72, 2018.

[2] S. R. Rani, S. R. Andani, And D. Suhendro, "Penerapan Algoritma K-Nearest Neighbor Untuk Prediksi Kelulusan Siswa Pada Smk Anak Bangsa," In *Prosiding Seminar Nasional Riset Information Science (Senaris)*, 2019, Pp. 670–676.

[3] L. A. R. Hakim, A. A. Rizal, And D. Ratnasari, "Aplikasi Prediksi Kelulusan Mahasiswa Berbasis K-Nearest Neighbor (K-Nn)," *J. Teknol. Inf. Dan Multimed.*, Vol. 1, No. 1, Pp. 30–36, 2019.

[4] E. Purwaningsih And E. Nurelasari, "Penerapan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kelulusan Pada Siswa," *Syntax J. Inform.*, Vol. 10, No. 1, Pp. 46–56, 2021.

[5] S. P. Nabila, N. Ulinnuha, And A. Yusuf, "Model Prediksi Kelulusan Tepat Waktu Dengan Metode Fuzzy C-Means Dan K-Nearest Neighbors Menggunakan Data Registrasi Mahasiswa," *J. Ilm. Nero*, Vol. 6, No. 1, Pp. 38–46, 2021.

[6] A. Y. Saputra And Y. Primadasa, "Penerapan Teknik Klasifikasi Untuk Prediksi

Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour," *Techno.Com*, Vol. 17, No. 4, Pp. 395–403, 2018.

[7] E. Novianto, A. Hermawan, And D. Avianto, "Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, Decision Tree Untuk Prediksi Status Kelulusan Mahasiswa S1," *Rabit J. Teknol. Dan Sist. Inf. Univrab*, Vol. 8, No. 2, Pp. 146–154, 2023.

[8] D. Safitri, S. S. Hilabi, And F. Nurapriani, "Analisis Penggunaan Algoritma Klasifikasi Dalam Prediksi Kelulusan Menggunakan Orange Data Mining," *Rabit J. Teknol. Dan Sist. Inf. Univrab*, Vol. 8, No. 1, Pp. 75–81, 2023.

[9] F. Sugandi, "Prediksi Kelulusan Mahasiswa Stmik Dharmawacana Menggunakan Algoritma K-Nearest Neighbors," *Sienna*, Vol. 4, No. 1, Pp. 20–26, 2023.

[10] J. Astri, J. Karman, And N. K. Daulay, "Prediksi Kelulusan Mahasiswa Menggunakan Metode K-Nearest Neigbor (Knn) Pada Fakultas Ilmu Teknik, Univeritas Bina Insan," *J. Ris. Sist. Inf. Dan Tek. Inform.*, Vol. 8, No. 1, Pp. 169–173, 2023.

[11] D. P. Sari, S. S. Hilabi, And A. Hananto, "Penerapan Data Mining Metode K-Nearest Neighbor Untuk Memprediksi Kelulusan Siswa Sekolah Menengah Pertama," *Smartics J.*, Vol. 9, No. 1, Pp. 14–19, 2023.

[12] R. Situmorang, W. I. Rahayu, And R. N. S. Fathonah, "Model Algoritma K-Nearest Neighbor (K-Nn) Dan Naïve Bayes Untuk Prediksi Kelulusan Mahasiswa," *Jati (Jurnal Mhs. Tek. Inform.*, Vol. 7, No. 1, Pp. 250–254, 2023.

[13] I. Ramdhani, "Perbandingan Metode Data Mining Model Klasifikasi Naive Bayes, Decission Tree Dan K-Nearest Neighbour Dalam Memprediksi Ketepatan Kelulusan Mahasiswa Prodi Teknik Informatika Di Universitas Pamulang," *J. Ilmu Komput. Jik*, Vol. Vi, No. 1, Pp. 88–94, 2023.

[14] S. Mulyati, S. M. Husein, And Ramdhan, "Rancang Bangun Aplikasi Data Mining Prediksi Kelulusan Ujian Nasional Menggunakan Algoritma (Knn) K-Nearest Neighbor Dengan Metode Euclidean Distance Pada Smpn 2 Pagedangan," *J. Tek. Inform.*, Vol. 4, No. 1, Pp. 65–73, 2020.

[15] E. F. Wati, A. P. Sari, E. T. Alawiah, M. H. Siregar, And B. Rudianto, "Particle Swarm Optimization Comparison On Decision Tree And Naive Bayes For Pandemic Graduation Classification," In *2nd International Conference On Advanced Information Scientific Development (Icaisd)*, 2021, Pp. 1–11.