A Data-Driven Approach to Building a Student Graduation Map with the K-Nearest Neighbors Algorithm at Hayam Wuruk Perbanas University

Yudha Herlambang Cahya Pratama[1*], Laqma Dica Fitrani[2], Muhammad Septama Prasetya[3]

[1]Departement of Information System, Faculty of Engineering and Design, Hayam Wuruk Perbanas University, Surabaya, East Java, Indonesia
[2]Department of Digital Business, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, East Java, Indonesia
[3]Department of Information Systems, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, East Java, Indonesia
E-mail: [1*]yudha.herlambang@perbanas.ac.id, [2]laqma_dica.bd@upnjatim.ac.id, [3]muhammad_septama.si@upnjatim.ac.id

(Received: 17 Jun 2025, revised: 4 Aug 2025, accepted: 5 Aug 2025)

Abstract

This study addresses the underutilization of student academic data at Universitas Hayam Wuruk Perbanas Surabaya, despite a significant trend of over 50% of students graduating within 3.5 years—raising important questions about academic quality assurance. To respond to this, the study aims to develop a data-driven graduation map by applying the K-Nearest Neighbors (KNN) algorithm to classify students based on their likelihood of graduating on time or late. Using historical academic records from 316 students of the 2022 cohort, the methodology involved data preprocessing, feature selection, implementation in RapidMiner, and algorithm testing. The KNN model was evaluated using a 70:30 training-to-testing split. Results demonstrated strong predictive performance of ROC (Receiver Operating Characteristic), including 95.74% accuracy, 100% precision for predicting delayed graduation, and an AUC (Area Under Curve) of 0.969. While the model was highly effective in identifying on-time graduates, it exhibited slightly lower recall for late graduates. These findings offer practical implications for developing targeted academic support systems and contribute to the broader application of machine learning in higher education analytics in Indonesia. Limitations include the use of data from a single cohort and reliance on one algorithm; future research may explore multi-cohort data and compare multiple classification methods to enhance generalizability and robustness.

Keywords: Academic Data, Classification, Data Mining, Machine Learning, KNN.

I. INTRODUCTION

Higher education stands as a foundational pillar in nation-building, playing a pivotal role in cultivating a globally competitive and competent human resource [1]. The success of a higher education institution is measured not only by the quality of its teaching and research, but also by the efficiency and effectiveness of its learning-teaching processes, as reflected in the timely graduation rates of its students [2]. Delayed student graduation can lead to various negative consequences, both for individual students, such as career postponement and increased educational costs, and for the institution, including decreased reputation and inefficient resource allocation. Therefore, a profound understanding of student graduation dynamics and their influencing factors is crucial for developing targeted policies and interventions.

With the advancement of information technology and the abundance of available data, data-driven approaches have revolutionized various fields, including higher education [4]. The application of data mining and machine learning techniques offers significant potential for analyzing hidden patterns within student academic data. This analysis can then be leveraged to predict the risk of delayed graduation and to identify significant predictive factors [5]. By enabling early prediction of students who are likely to face challenges in completing their studies on time, institutions can design more personalized and effective academic and non-academic support strategies [6].

In the context of machine learning, the K-Nearest Neighbors (KNN) algorithm is a popular non-parametric classification method widely applied in various predictive studies within the field of education [7]. The fundamental principle of KNN is to classify new data points based on the majority class of their k-nearest neighbors within the feature space [8]. The advantages of KNN lie in its simplicity, its ability to handle high-complexity data without specific data distribution assumptions, and the relatively easy interpretability of its results [9]. Various prior studies have successfully applied the KNN algorithm to predict student academic performance [10], identifying students at risk of attrition [11], and even predicting student career choices [12].

Universitas Hayam Wuruk Perbanas Surabaya, a prominent higher education institution in East Java, possesses a valuable repository of student academic data that remains largely underutilized for strategic decision-making aimed at enhancing the quality of academic and student services. Interviews with the academic unit indicate that over 50% of students from the analyzed cohort successfully graduated within 3.5 years (seven semesters) this phenomenon of accelerated graduation warrants further investigation, particularly concerning its correlation with graduate quality. The researchers obtained data from 316 students of the 2022 cohort who are slated for graduation, representing a significant student population for further analysis regarding graduation patterns.

While accelerated graduation can be considered an indicator of study efficiency, it's crucial to ensure that this acceleration doesn't compromise the quality and competence of graduates [13]. This research has the potential to help ensure that students graduating within 3.5 years still meet the competency standards set by their study programs and remain relevant to workforce needs. By analyzing academic and non-academic data, this study can identify potential gaps in knowledge or skills that might arise from a shorter study path. This information will be invaluable for the university in taking corrective actions, such as curriculum adjustments, enhancing mentoring programs, or developing more comprehensive evaluation mechanisms.

Previous research has explored the application of classification algorithms, including KNN, in the context of predicting student graduation across various institutions and countries [14][15][16]. However, studies specifically focused on developing a "graduation map" for students using the KNN algorithm and leveraging real data from Universitas Hayam Wuruk Perbanas Surabaya over the last five years are still very limited. This graduation map is expected to provide a comprehensive visualization of student groups with specific characteristics who tend to graduate on time or late.

This study seeks to address that gap by implementing a KNN-based classification model to identify students likely to graduate on time or late. The research aims to develop a student graduation map using historical data from the 2022 cohort, providing both practical tools for academic support planning and broader methodological insights for similar institutions. Thus, the core problem this study investigates is: Can the KNN algorithm accurately classify student graduation timelines, and how can these predictions be used to enhance academic policy and intervention strategies?

II. RESEARCH METHOD

Preprocessing

A. Research Flow

This study employs a quantitative approach to analyze student graduation predictions using the K-Nearest Neighbor (KNN) algorithm. The quantitative method facilitates structured processing of numerical data to uncover patterns that contribute to student graduation outcomes [17]. The dataset includes academic variables such as Grade Point Average (GPA), attendance, and length of study, as well as additional demographic factors.

The research was carried out through the following stages:

1. Problem Identification
   In this initial stage, the research identifies academic-related issues faced by the Academic Department at Universitas Hayam Wuruk Perbanas. The goal is to define the challenges in monitoring and supporting student graduation performance.
2. Data Collection
   Data were gathered from various sources including literature reviews, observations, and interviews. All collected data were documented systematically to serve as a foundation for the subsequent stages of analysis.
3. Data Analysis
   This stage comprises three key processes: data selection, data mining, and prediction results.
   a. Data selection involves data cleaning, transformation, and feature selection.
   b. Data mining includes dataset partitioning, implementation of the KNN algorithm, model validation, and testing.
   c. Prediction results cover the generation of graduation predictions, model evaluation, and performance analysis using the confusion matrix. The output of this process provides actionable insights into the primary factors influencing graduation outcomes.
4. Result Interpretation
   This stage focuses on interpreting the results generated by the KNN algorithm, assessing the model's performance, and drawing conclusions based on prediction accuracy and classification effectiveness.
5. Conclusion and Recommendations
   The final stage presents the conclusions derived from the prediction and analytical results. Based on the findings, the study offers practical recommendations to the Academic Department of Universitas Hayam Wuruk Perbanas to support evidence-based decision-making and improve student academic outcomes.

Data Collection

A. Type and Source of Data

This research adopts a quantitative approach to analyze student graduation predictions using the K-Nearest Neighbor

(KNN) algorithm. The method allows for structured numerical data processing to identify patterns that influence student graduation outcomes. The data utilized in this study comprises academic information such as GPA, attendance records, and study duration, as well as additional variables such as student demographic data. This study aims to provide a comprehensive overview of the effectiveness of KNN in predicting graduation outcomes and identifying key contributing factors.

The data used in this study are derived from both primary and secondary sources:

1. Primary data were obtained directly from the research subjects. This includes student academic records provided by the Head of Academic Department at Universitas Hayam Wuruk Perbanas. To complement the dataset, interviews were also conducted with the Head of Academic Department to gain contextual understanding and deeper insight into the graduation process and academic performance indicators.
2. Secondary data were collected from relevant documents and archival records, such as annual reports and academic performance reports issued by the Academic Department. These documents serve as supporting evidence to validate and enrich the primary data.

B. Research Period

The research was conducted from March 2025 to July 2025, covering the entire process from data collection to final analysis and interpretation.

The study was carried out at the Academic Department of Universitas Hayam Wuruk Perbanas. The selection of this location was based on the availability of relevant data and access to institutional stakeholders, ensuring that the research objectives could be effectively achieved. Conducting the study in a focused and well-defined environment helps maintain the clarity and relevance of the research scope.

C. Research Subject

The research subject refers to the informant who provides essential data or insights relevant to the study. In this study, the subject selection technique employed is purposive sampling, which involves the deliberate selection of individuals based on predefined criteria relevant to the research objectives [18]. This approach ensures that the selected subject possesses the necessary knowledge and authority to provide accurate and relevant data.

D. Data Collection Methods

This research employed several data collection techniques to ensure comprehensive and triangulated information gathering, as follows:

1. Literature Review
   The literature review involved a thorough examination of scholarly sources such as academic journals, books, articles, and previous research reports related to graduation prediction, machine learning, and academic performance. This method aimed to build a solid theoretical foundation, identify existing methodologies, and reveal research gaps that this study seeks to address. The review supports both the development of the research framework and the contextualization of findings within the broader academic discourse.
2. Observation
   The observation technique allows for the direct examination of relevant objects and phenomena in the research setting. According to Sugiyono, observation techniques can be categorized into participatory, overt, and unstructured observation. In this study, the overt observation method was employed, wherein the researcher transparently communicated the intention of conducting a study to the data sources. The researcher observed student academic performance data by reviewing Study Result Reports (KHS) available at the end of each semester across various study programs at Universitas Hayam Wuruk Perbanas.
3. Interview
   The interview method used in this study was semi-structured, which allows for a guided yet flexible conversation. This approach is well-suited for case study research and enables in-depth exploration of the interviewee's insights while maintaining focus on specific themes. The semi-structured interview was conducted with Ms. Rahayu Winarti, A.Md., to gather detailed information on academic processes, data management, and the institution’s efforts in supporting student graduation. The flexibility of this method also allows for the discovery of new insights during the discussion process.

Model Development

A. Data Analysis Technique

Once data were collected through literature review, observation, and interviews, the researcher proceeded with a systematic data analysis process to extract meaningful insights. In this study, the data analysis process uses the RapidMiner tool. RapidMiner was chosen because it provides an intuitive visual interface and supports efficient implementation of the KNN algorithm without requiring coding, making the data analysis and interpretation process more accessible. The analysis was structured into the following stages:

1. Data Selection
   This stage ensures that the data used for modeling are clean, consistent, and relevant:
   a. Data Cleaning: This involves identifying and handling missing values, duplicates, and anomalies. Irrelevant data points are removed, and missing values are addressed using methods such as mean or median imputation.
   b. Data Transformation: Numerical normalization (e.g., Min-Max Scaling) is applied to bring all attributes to the same scale, facilitating accurate distance computation in KNN. Categorical data are encoded into numeric form using techniques such as one-hot encoding.
   c. Feature Selection: Key attributes are selected to reduce dimensionality and ensure that only the most relevant features are used for model building.
2. Data Mining

DOI: 10.34148/teknika.v14i3.1282
TEKNIKA, Volume 14(3), November 2025, pp. 339-346
ISSN 2549-8037, EISSN 2549-8045

This phase involves the application of the K-Nearest Neighbor (KNN) algorithm to develop the graduation prediction model:

a. Dataset Splitting: The dataset is divided into training (70%) and testing (30%) subsets using stratified sampling to maintain class balance.

b. KNN Algorithm Implementation: The algorithm utilizes the training data to perform predictions. Parameters such as the number of nearest neighbors (K) and distance metrics (e.g., Euclidean Distance) are tested to determine the optimal configuration for the predictive model. The procedural stages of the K-Nearest Neighbor (KNN) algorithm are depicted in the Figure 1 below.

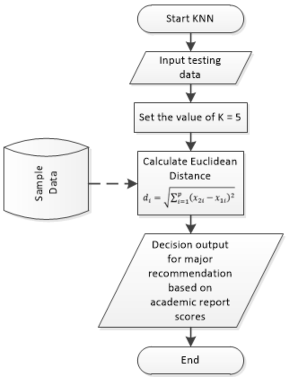


Figure 1. KNN Algorithm Flowchart

Based on Figure 1, the value of K is set to 5. A small K value (e.g., K=1) can make the model too sensitive to training data (overfitting), since predictions rely on just one neighbor that may be an outlier. On the other hand, a large K value can make the model too general (underfitting), as it considers too many neighbors from different classes. Setting K=5 provides a good balance, offering stable and accurate results. It is also commonly used as a starting point in many KNN studies and applications.

c. Model Validation: To evaluate the robustness of the model on training data and to determine the optimal value of the K parameter in the KNN algorithm, split-validation was conducted. This technique partitions the training data into several subsets, using some for training and others for validation in multiple iterations. The performance results from each iteration are averaged to provide a more reliable estimate of the model's generalization capability.

d. Model Testing: After the model is trained and validated, model testing is performed using a reserved test set that has not been used in either the training or validation processes. This step provides an unbiased evaluation of how well the model performs in predicting student graduation outcomes based on unseen data. The output of this stage indicates the model's real-world applicability and predictive accuracy.

3. Prediction Results

a. Graduation Prediction
The final KNN model provides a classification output indicating whether a student is predicted to graduate or not graduate. This classification is based on key academic and demographic features previously defined during the data preprocessing stage.

b. Model Evaluation
To assess the performance of the KNN model, several evaluation metrics were used:
- Accuracy: The proportion of correct predictions to the total number of predictions made.
- Precision: The proportion of true positive graduation predictions among all students predicted to graduate.
- Recall: The ability of the model to correctly identify all students who actually graduated.
- ROC Curve: A graph used to evaluate the model's ability to distinguish between positive and negative classes based on predicted probabilities.

c. Confusion Matrix Analysis
A confusion matrix was utilized to present a detailed breakdown of the model's predictions, distinguishing between:
- True Positives (TP): Students correctly predicted to graduate.
- True Negatives (TN): Students correctly predicted not to graduate.
- False Positives (FP): Students predicted to graduate but did not.
- False Negatives (FN): Students predicted not to graduate but did.

This matrix provides insight into the types of errors made by the model and helps in refining the predictive process.

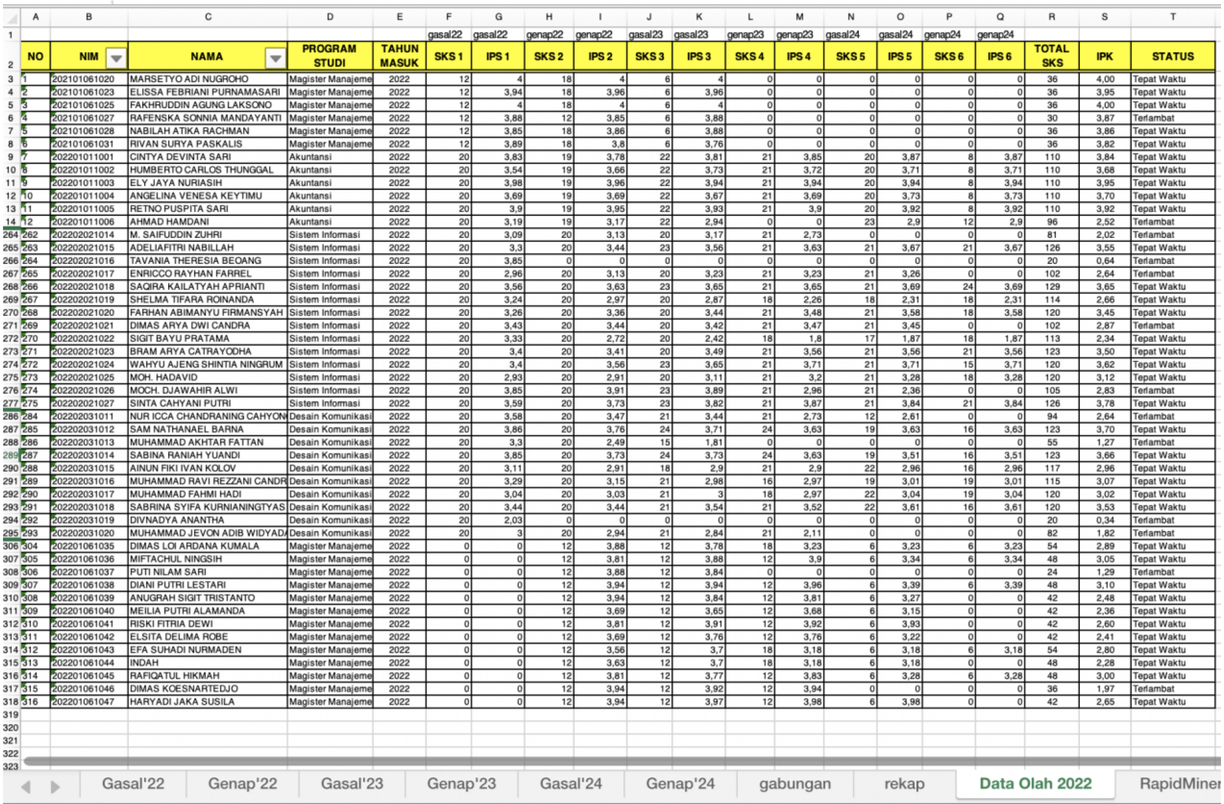
III. RESULT AND DISCUSSION

This chapter presents the results and discussion of the application of the KNN algorithm to predict student graduation based on academic data such as GPA, IPS, and number of credits. The analysis includes model evaluation using accuracy, precision, recall, and confusion matrix metrics. The results of data processing will be displayed in the form of tables, graphs and narrative explanations to provide an overview of the performance of the KNN algorithm.

A. Preprocessing Data

In the process of implementing the K-Nearest Neighbors (KNN) algorithm, the data preprocessing step is a very important stage to ensure that the data used meets the requirements and is relevant for analysis. Preprocessing is carried out to overcome various problems in the dataset, such as incomplete data, data imbalance, or different value scales for each attribute. This stage aims to improve data quality so that the prediction results produced by the KNN algorithm are more accurate and reliable. The results of the data preprocessing process can be seen in Figure 2 below.

DOI: 10.34148/teknika.v14i3.1282
ISSN 2549-8037, EISSN 2549-8045



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1						gasal22	gasal22	genap22	genap22	gasal23	gasal23	genap23	genap23	gasal24	gasal24	genap24	genap24			
2	NO	NIM	NAMA	PROGRAM STUDI	TAHUN MASUK	SKS 1	IPS 1	SKS 2	IPS 2	SKS 3	IPS 3	SKS 4	IPS 4	SKS 5	IPS 5	SKS 6	IPS 6	TOTAL SKS	IPK	STATUS
3	1	202101061020	MARSETYO ADI NUGROHO	Magister Manajeme	2022	12	4	18	4	6	4	0	0	0	0	0	0	36	4,00	Tepat Waktu
4	2	202101061023	ELISSA FEBRIANI PURNAMASARI	Magister Manajeme	2022	12	3,94	18	3,96	6	3,96	0	0	0	0	0	0	36	3,95	Tepat Waktu
5	3	202101061025	FAKHRUDDIN AGUNG LAKSONO	Magister Manajeme	2022	12	4	18	4	6	4	0	0	0	0	0	0	36	4,00	Tepat Waktu
6	4	202101061027	RAFENSKA SONNIA MANDAYANTI	Magister Manajeme	2022	12	3,88	12	3,85	6	3,88	0	0	0	0	0	0	30	3,87	Terlambat
7	5	202101061028	NABILAH ATIKA RACHMAN	Magister Manajeme	2022	12	3,85	18	3,86	6	3,88	0	0	0	0	0	0	36	3,86	Tepat Waktu
8	6	202101061031	RIVAN SURYA PASKALIS	Magister Manajeme	2022	12	3,89	18	3,8	6	3,76	0	0	0	0	0	0	36	3,82	Tepat Waktu
9	7	202201011001	CINTYA DEVINTA SARI	Akuntansi	2022	20	3,83	19	3,78	22	3,81	21	3,85	20	3,87	8	3,87	110	3,84	Tepat Waktu
10	8	202201011002	HUMBERTO CARLOS THUNGGAL	Akuntansi	2022	20	3,54	19	3,66	22	3,73	21	3,72	20	3,71	8	3,71	110	3,68	Tepat Waktu
11	9	202201011003	ELY JAYA NURIASIH	Akuntansi	2022	20	3,98	19	3,96	22	3,94	21	3,94	20	3,94	8	3,94	110	3,95	Tepat Waktu
12	10	202201011004	ANGELINA VENESA KEYTIMU	Akuntansi	2022	20	3,69	19	3,69	22	3,67	21	3,69	20	3,73	8	3,73	110	3,70	Tepat Waktu
13	11	202201011005	RETNO PUSPITA SARI	Akuntansi	2022	20	3,9	19	3,95	22	3,93	21	3,9	20	3,92	8	3,92	110	3,92	Tepat Waktu
14	12	202201011006	AHMAD HAMDANI	Akuntansi	2022	20	3,19	19	3,17	22	2,94	0	0	23	2,9	12	2,9	96	2,52	Terlambat
264	262	202202021014	M. SAIFUDDIN ZUHRI	Sistem Informasi	2022	20	3,09	20	3,13	20	3,17	21	2,73	0	0	0	0	81	2,02	Terlambat
265	263	202202021015	ADELIAFITRI NABILLAH	Sistem Informasi	2022	20	3,3	20	3,44	23	3,56	21	3,63	21	3,67	21	3,67	126	3,55	Tepat Waktu
266	264	202202021016	TAVANIA THERESIA BEOANG	Sistem Informasi	2022	20	3,85	0	0	0	0	0	0	0	0	0	0	20	0,64	Terlambat
267	265	202202021017	ENRICCO RAYHAN FARREL	Sistem Informasi	2022	20	2,96	20	3,13	20	3,23	21	3,23	21	3,26	0	0	102	2,64	Terlambat
268	266	202202021018	SAQIRA KAILATYAH APRIANTI	Sistem Informasi	2022	20	3,56	20	3,63	23	3,65	21	3,65	21	3,69	24	3,69	129	3,65	Tepat Waktu
269	267	202202021019	SHELMA TIFARA ROINANDA	Sistem Informasi	2022	20	3,24	20	2,97	20	2,87	18	2,26	18	2,31	18	2,31	114	2,66	Tepat Waktu
270	268	202202021020	FARHAN ABIMANYU FIRMANSYAH	Sistem Informasi	2022	20	3,26	20	3,36	20	3,44	21	3,48	21	3,58	18	3,58	120	3,45	Tepat Waktu
271	269	202202021021	DIMAS ARYA DWI CANDRA	Sistem Informasi	2022	20	3,43	20	3,44	20	3,42	21	3,47	21	3,45	0	0	102	2,87	Terlambat
272	270	202202021022	SIGIT BAYU PRATAMA	Sistem Informasi	2022	20	3,33	20	2,72	20	2,42	18	1,8	17	1,87	18	1,87	113	2,34	Tepat Waktu
273	271	202202021023	BRAM ARYA CATRAYODHA	Sistem Informasi	2022	20	3,4	20	3,41	20	3,49	21	3,56	21	3,56	21	3,56	123	3,50	Tepat Waktu
274	272	202202021024	WAHYU AJENG SHINTIA NINGRUM	Sistem Informasi	2022	20	3,4	20	3,56	23	3,65	21	3,71	21	3,71	15	3,71	120	3,62	Tepat Waktu
275	273	202202021025	MOH. HADAVID	Sistem Informasi	2022	20	2,93	20	2,91	20	3,11	21	3,2	21	3,28	18	3,28	120	3,12	Tepat Waktu
276	274	202202021026	MOCH. DJAWAHIR ALWI	Sistem Informasi	2022	20	3,85	20	3,91	23	3,89	21	2,96	21	2,36	0	0	105	2,83	Terlambat
277	275	202202021027	SINTA CAHYANI PUTRI	Sistem Informasi	2022	20	3,59	20	3,73	23	3,82	21	3,87	21	3,84	21	3,84	126	3,78	Tepat Waktu
286	284	202202031011	NUR ICCA CHANDRANING CAHYON	Desain Komunikasi	2022	20	3,58	20	3,47	21	3,44	21	2,73	12	2,61	0	0	94	2,64	Terlambat
287	285	202202031012	SAM NATHANAEL BARNA	Desain Komunikasi	2022	20	3,86	20	3,76	24	3,71	24	3,63	19	3,63	16	3,63	123	3,70	Tepat Waktu
288	286	202202031013	MUHAMMAD AKHTAR FATTAN	Desain Komunikasi	2022	20	3,3	20	2,49	15	1,81	0	0	0	0	0	0	55	1,27	Terlambat
289	287	202202031014	SABINA RANIAH YUANDI	Desain Komunikasi	2022	20	3,85	20	3,73	24	3,73	24	3,63	19	3,51	16	3,51	123	3,66	Tepat Waktu
290	288	202202031015	AINUN FIKI IVAN KOLOV	Desain Komunikasi	2022	20	3,11	20	2,91	18	2,9	21	2,9	22	2,96	16	2,96	117	2,96	Tepat Waktu
291	289	202202031016	MUHAMMAD RAVI REZZANI CANDR	Desain Komunikasi	2022	20	3,29	20	3,15	21	2,98	16	2,97	19	3,01	19	3,01	115	3,07	Tepat Waktu
292	290	202202031017	MUHAMMAD FAHMI HADI	Desain Komunikasi	2022	20	3,04	20	3,03	21	3	18	2,97	22	3,04	19	3,04	120	3,02	Tepat Waktu
293	291	202202031018	SABRINA SYIFA KURNIANINGTYAS	Desain Komunikasi	2022	20	3,44	20	3,44	21	3,54	21	3,52	22	3,61	16	3,61	120	3,53	Tepat Waktu
294	292	202202031019	DIVNADYA ANANTHA	Desain Komunikasi	2022	20	2,03	0	0	0	0	0	0	0	0	0	0	20	0,34	Terlambat
295	293	202202031020	MUHAMMAD JEVON ADIB WIDYAD	Desain Komunikasi	2022	20	3	20	2,94	21	2,84	21	2,11	0	0	0	0	82	1,82	Terlambat
306	304	202201061035	DIMAS LOI ARDANA KUMALA	Magister Manajeme	2022	0	0	12	3,88	12	3,78	18	3,23	6	3,23	6	3,23	54	2,89	Tepat Waktu
307	305	202201061036	MIFTACHUL NINGSIH	Magister Manajeme	2022	0	0	12	3,81	12	3,88	12	3,9	6	3,34	6	3,34	48	3,05	Tepat Waktu
308	306	202201061037	PUTI NILAM SARI	Magister Manajeme	2022	0	0	12	3,88	12	3,84	0	0	0	0	0	0	24	1,29	Terlambat
309	307	202201061038	DIANI PUTRI LESTARI	Magister Manajeme	2022	0	0	12	3,94	12	3,94	12	3,96	6	3,39	6	3,39	48	3,10	Tepat Waktu
310	308	202201061039	ANUGRAH SIGIT TRISTANTO	Magister Manajeme	2022	0	0	12	3,94	12	3,84	12	3,81	6	3,27	0	0	42	2,48	Tepat Waktu
311	309	202201061040	MEILIA PUTRI ALAMANDA	Magister Manajeme	2022	0	0	12	3,69	12	3,65	12	3,68	6	3,15	0	0	42	2,36	Tepat Waktu
312	310	202201061041	RISKI FITRIA DEWI	Magister Manajeme	2022	0	0	12	3,81	12	3,91	12	3,92	6	3,93	0	0	42	2,60	Tepat Waktu
313	311	202201061042	ELSITA DELIMA ROBE	Magister Manajeme	2022	0	0	12	3,69	12	3,76	12	3,76	6	3,22	0	0	42	2,41	Tepat Waktu
314	312	202201061043	EFA SUHADI NURMADEN	Magister Manajeme	2022	0	0	12	3,56	12	3,7	18	3,18	6	3,18	6	3,18	54	2,80	Tepat Waktu
315	313	202201061044	INDAH	Magister Manajeme	2022	0	0	12	3,63	12	3,7	18	3,18	6	3,18	0	0	48	2,28	Tepat Waktu
316	314	202201061045	RAFIQATUL HIKMAH	Magister Manajeme	2022	0	0	12	3,81	12	3,77	12	3,83	6	3,28	6	3,28	48	3,00	Tepat Waktu
317	315	202201061046	DIMAS KOESNARTEDJO	Magister Manajeme	2022	0	0	12	3,94	12	3,92	12	3,94	0	0	0	0	36	1,97	Terlambat
318	316	202201061047	HARYADI JAKA SUSILA	Magister Manajeme	2022	0	0	12	3,94	12	3,97	12	3,98	6	3,98	0	0	42	2,65	Tepat Waktu

Figure 2. Result of Prepocessing Data

The results of this preprocessing have gone through several stages, including data cleaning, data transformation, and normalization. Data cleaning is done to handle empty values or invalid data. Data transformation is needed to ensure that all attributes in the dataset have a uniform format and are suitable for analysis [19]. Normalization is done to align the attribute scale so that no feature dominates the distance calculation results in the KNN algorithm, which uses metrics such as Euclidean distance.

B. Selection Data

The selection of appropriate data ensures that the information used in the analysis is relevant to the research objectives, so that the model can produce accurate and meaningful predictions. The data used must include attributes that have a correlation with the target variable, such as student graduation in the context of the research. The data selection process can be seen in Table 1 below.

Table 1 Data Selection

No	Attribute	Information
1	NIM	✓
2	Name	✓
3	Study Program	X
4	Year of Entry	X
5	Credits 1	X
6	GPA 1	X
7	Credits 2	X
8	GPA 2	X
9	Credits 3	X
10	GPA 3	X
11	Credits 4	X
12	GPA 4	X
13	Credits 5	X
14	GPA 5	X
15	Credits 6	X
16	GPA 6	X
17	Total Credits	✓
18	Cumulative GPA	✓
19	Status	✓

The data selection stage includes identifying and determining the most relevant features from the dataset, such as GPA, Total Credits, and completion status. Data selection results in 5 selected attributes from a total of 15 attributes available in the data. These attributes are selected based on the assumption that they have a significant influence on the likelihood of student graduation status. The selection of the right attributes can also reduce the dimensionality of the data, thereby reducing the complexity of the KNN algorithm calculation without sacrificing the quality of the prediction.

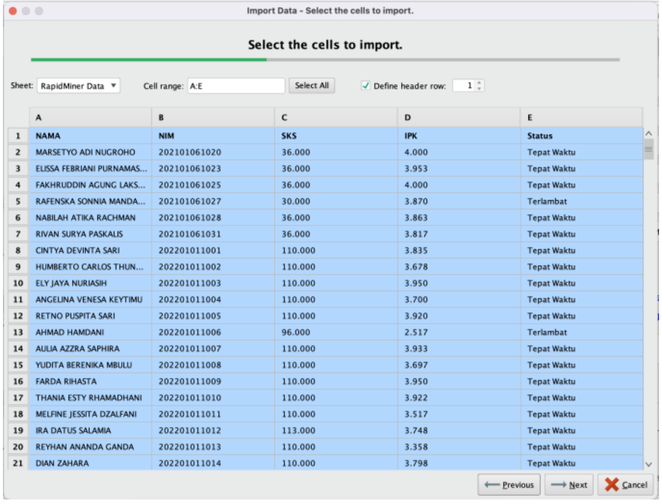
C. Implementation RapidMinner

RapidMinner as an analysis tool to build a prediction model using the K-Nearest Neighbors (KNN) algorithm. This stage aims to describe in detail the technical process carried out starting from data processing to producing an optimal model.

DOI: 10.34148/teknika.v14i3.1282
TEKNIKA, Volume 14(3), November 2025, pp. 339-346
ISSN 2549-8037, EISSN 2549-8045

a) Import Data

Importing data is the initial process in implementation, which is entering the dataset into RapidMiner. At this stage, data attribute settings are made such as determining the target label and ensuring the data format is in accordance with the analysis needs. The imported data is data that has passed the selection stage, namely 316 data with 5 attributes. The results of the data import can be seen in Figure 3 below.



	A	B	C	D	E
1	NAMA	NIM	SKS	IPK	Status
2	MARSETYO ADI NUGROHO	202101061020	36.000	4.000	Tepat Waktu
3	ELISSA FEBRIANI PURNAMAS...	202101061023	36.000	3.953	Tepat Waktu
4	FAKHRUDDIN AGUNG LAKS...	202101061025	36.000	4.000	Tepat Waktu
5	RAFENSKA SONNIA MANDA...	202101061027	30.000	3.870	Terlambat
6	NABILAH ATIKA RACHMAN	202101061028	36.000	3.863	Tepat Waktu
7	RIVAN SURYA PASKALIS	202101061031	36.000	3.817	Tepat Waktu
8	CINTYA DEVINTA SARI	202201011001	110.000	3.835	Tepat Waktu
9	HUMBERTO CARLOS THUN...	202201011002	110.000	3.678	Tepat Waktu
10	ELY JAYA NURIASIH	202201011003	110.000	3.950	Tepat Waktu
11	ANGELINA VENESA KEYTIMU	202201011004	110.000	3.700	Tepat Waktu
12	RETNO PUSPITA SARI	202201011005	110.000	3.920	Tepat Waktu
13	AHMAD HAMDANI	202201011006	96.000	2.517	Terlambat
14	AULIA AZZRA SAPHIRA	202201011007	110.000	3.933	Tepat Waktu
15	YUDITA BERENIKA MBULU	202201011008	110.000	3.697	Tepat Waktu
16	FARDA RIHASTA	202201011009	110.000	3.950	Tepat Waktu
17	THANIA ESTY RHAMADHANI	202201011010	110.000	3.922	Tepat Waktu
18	MELFINE JESSITA DZALFANI	202201011011	110.000	3.517	Tepat Waktu
19	IRA DATUS SALAMIA	202201011012	113.000	3.748	Tepat Waktu
20	REYHAN ANANDA GANDA	202201011013	110.000	3.358	Tepat Waktu
21	DIAN ZAHARA	202201011014	110.000	3.798	Tepat Waktu

Figure 3. Import Data Processing

b) Model Process Validation

The Validation Process Model discusses the methods used to evaluate model performance. Validation techniques, such as dividing data into training and testing or applying cross-validation, are implemented to measure the performance of the model in predicting data. This step aims to ensure that the model built is not only accurate on the training data, but also reliable in handling data outside the training sample [20]. The configuration process in rapidminner can be seen in Figure 4 below.

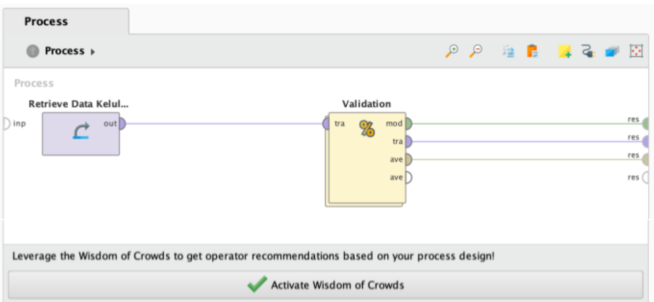


Figure 4. Konfigurasi Process Validation

c) Algoritma KNN

This section describes the application of the K-Nearest Neighbors algorithm in building a prediction model. At this stage, important parameters are set, such as the K value and the distance metric used, to optimize the prediction results. This process includes the process of training the model with training data and testing the model with testing data to produce a comprehensive performance evaluation. The process configuration in the KNN algorithm can be seen in Figure 5 below.

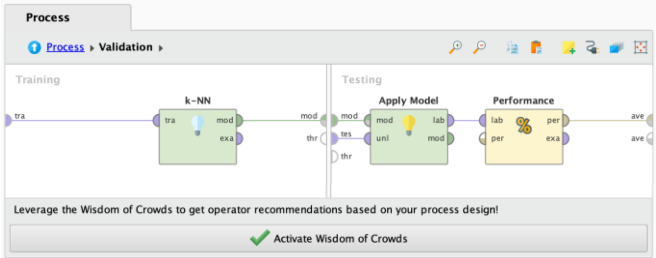


Figure 5. KNN Algorithm Process Configuration

Figure 5 explains that data splitting is done by dividing the data into 2 parts, namely 70% for training data and 30% for testing data. In the training data split, the KNN algorithm operator process is carried out, while in the testing data split, the apply model operator and performance operator are carried out. The KNN operator is used to train predictions on training data. The Apply Model operator is used to test the model that has been trained by the KNN operator. The performance operator is used to evaluate the prediction results. The next process is to run the process model to see the prediction results of the KNN algorithm with a value of K=5.

D. Algorithm Testing

Testing the K-Nearest Neighbors (KNN) algorithm is one of the important steps in implementing a prediction algorithm. Testing aims to evaluate the performance of the model that has been built based on the available data. This stage is carried out using several evaluation techniques including split validation, confusion matrix, and ROC (Receiver Operating Characteristic) curves.

a) Split Validation

Split validation is a method of dividing a dataset into training data and testing data. This technique is to ensure that the model is tested on data that has never been used in training, so that the evaluation results are more objective. In this study, the Split Validation operator was applied with a proportion of 70:30. The settings on rapidminner can be seen in Figure 6 below.

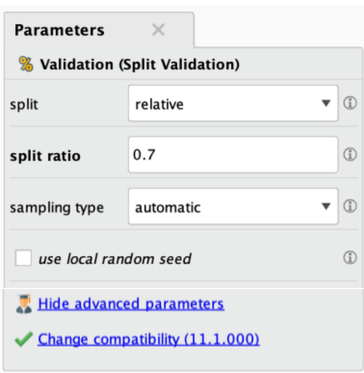


Figure 6. Parameters Split Validation

DOI: 10.34148/teknika.v14i3.1282

In Figure 6 it is explained that the split data type is set with relative data. The split ratio is set to 0.7, meaning using the rule of 70% training data and 30% testing data. The sampling type used is automatic.

b) Confution Matrix

Confusion Matrix discusses how to analyze the results of model predictions. Confusion Matrix provides detailed information about the distribution of prediction results, including correct predictions (True Positive and True Negative) and incorrect predictions (False Positive and False Negative) [21]. The results of the confusion matrix from this study can be seen in Figure 7 below.

accuracy: 95.74%

	true Tepat Waktu	true Terlambat	class precision
pred. Tepat Waktu	77	4	95.06%
pred. Terlambat	0	13	100.00%
class recall	100.00%	76.47%	

Figure 7. Confusion Matrix for KNN

Figure 7 shows the prediction results of the KNN algorithm. It can be seen in the figure that 77 data were predicted to be on time, and the results were correct on time. There were 0 data predicted to be late, but the actual results were on time. There were 4 data predicted to be on time, but it turned out to be late and 13 data predicted to be late, the actual results were late. Through this analysis, performance metrics of 95.74% accuracy, 100% precision, and 76.47% recall were obtained which helped us understand the strengths and weaknesses of the KNN model in classifying data. The accuracy value of 95.74%, the model showed very good performance. This shows that the KNN algorithm can utilize data patterns very effectively to make predictions. The model has high precision for both classes, especially for "Late" which reaches 100%. However, the recall for the "Late" class is lower than the "On Time" class. This shows that the model is more effective in detecting students who are on time than those who are late.

c) ROC Curve

The ROC curve evaluates the model's ability to distinguish between target classes (positive and negative) at various prediction thresholds. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR), while the Area Under Curve (AUC) value is used as a quantitative measure of model performance. The results of the ROC curve in this study can be seen in Figure 8 below.

Figure 8 shows the red curve value of the model performance at each threshold. The closer to the upper left corner, the better the performance. The ROC curve shows that the KNN model has excellent classification performance in distinguishing between on-time and delayed graduates, with an AUC value of 0.969. An AUC close to 1 indicates that the model is highly accurate and sensitive in identifying the "Delayed" class. This suggests that the KNN algorithm used is highly effective and reliable for predicting student graduation outcomes and has strong potential to be utilized as a supporting tool in academic early warning systems.

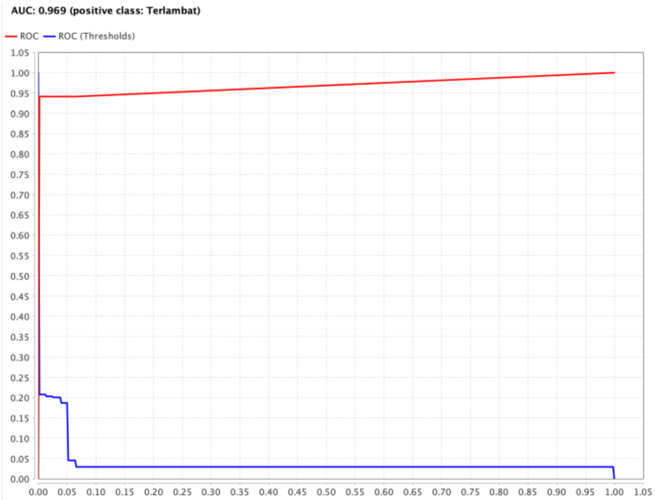


Figure 8. KNN Algorithm ROC Curve Graph

IV. CONCLUSION

The results of the study show that the K-Nearest Neighbor (KNN) algorithm is able to provide excellent performance in predicting student graduation. Based on the evaluation results, the model produces an accuracy of 95.74%, which shows that almost all predictions made by the KNN algorithm are in accordance with the actual data. Based on the KNN algorithm's high accuracy in predicting student graduation, university management is advised to implement an early warning system. This system enables early identification of students at risk of delayed graduation, allowing timely and targeted academic interventions such as additional guidance or intensive monitoring. Academic advisors can use the classification results to provide more focused mentoring to students categorized as at risk.

Furthermore, the KNN analysis can serve as a basis for evaluating and adjusting academic policies, including curriculum structure, semester credit loads, and graduation regulations. If common patterns are found among students who graduate late—such as low GPA or high course loads in the final semesters—management can design a more adaptive academic strategy. Non-academic factors should also be considered; the university can provide support programs such as counseling services, soft skills training, or motivation workshops to help students overcome personal or psychological barriers affecting their academic performance.

To maximize the usability of prediction results, it is recommended to integrate KNN data into an academic monitoring dashboard accessible to management, study programs, and academic advisors. This dashboard enables real-time, data-driven tracking of students' academic progress, supporting swift, measurable, and targeted decision-making and interventions. These steps not only encourage timely graduation but also enhance the overall quality of educational services.

DOI: 10.34148/teknika.v14i3.1282

Recommendations for future work, the development of prediction models can be done by exploring other algorithms, such as Random Forest (RF) or Support Vector Machine (SVM) to compare performance with the KNN algorithm. In addition, the use of feature engineering techniques to add new variables, such as student organization participation activities and social background, can provide richer insights into graduation predictions. Integration with real-time data using machine learning can be implemented to improve prediction efficiency and support earlier preventive actions for students who are predicted to be late.

REFERENCES

[1] OECD, "Higher Education," 2024. https://www.oecd.org/en/topics/higher-education.html.

[2] R. Pratama, R. Herdiana, R. Hamonangan, and S. Anwar, "Analisis Prediksi Kelulusan Mahasiswa Menggunakan Metode Artificial Neural Network," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 687–693, 2024, doi: 10.36040/jati.v8i1.8762.

[3] A. Marwah, "A Study Of Factors Influencing Students In Selecting Higher Education Institutions -A Literature Review," *Dogo Rangsang Res. J.*, vol. 122, no. 12, pp. 162–171, 2022.

[4] S. S. Akbar *et al.*, *Pengantar Bisnis Strategi Inovatif di Era Digital*, Pertama. Purbalingga: Eurika Media Aksara, 2024.

[5] Supriyono *et al.*, *Sistem Informasi Manajemen*. Purbalingga: Eureka Media Aksara, 2024.

[6] J. M. Polgan *et al.*, "Optimasi Prediksi Kelulusan Mahasiswa Menggunakan Random Forest untuk Meningkatkan Tingkat Retensi," *Jurnal Minfo Polgan,* vol. 13, pp. 2364–2374, 2025. doi :10.33395/jmp.v13i2.14472

[7] R. Sakti and A. Daulay, "Analisis Kritis dan Pengembangan Algoritma K-Nearest Neighbor (KNN): Sebuah Tinjauan Literatur," *Jurnal Pendidikan Sains dan Komputer*, vol. 4, no. 2, pp. 131–141, 2024. doi: 10.47709/jpsk.v4i02.

[8] Z. Hadiansyah, Z. Rozikin, and M. Fatchan, "Implementasi Algoritma K-Nearest Neighbor Dalam Klasifikasi Penyakit Kanker Paru Paru," *Journal of Computer System and Informatics*, vol. 6, no. 1, pp. 96–106, 2024, doi: 10.47065/josyc.v6i1.6195.

[9] M. Utami and V. Ayumi, "Prediksi Penjualan Produk Terlaris Menggunakan Algoritma K-Nearest Neighboard (KNN)," *Journal Computer Science and Information Syetem*, vol. 1, no. 2, pp. 43–47, 2024. doi:10.61567.

[10] M. Jawthari and V. Stoffová, "Predicting students' academic performance using a modified kNN algorithm," *Pollack Period.*, vol. 16, no. 3, pp. 20–26, 2021, doi: 10.1556/606.2021.00374.

[11] A. Suzana, B. Lomi, A. A. Pekuwali, and R. T. Abineno, "Pengelompokan Mahasiswa Berpotensi Drop Out Pada Program Studi Teknik Infromatika Menggunakan Metode K-Means Clustering," pp. 340–351, 2024. https://ojs.unkriswina.ac.id/index.php/semnas-FST.

[12] F. R. Yudana, M. Suyanto, and A. Nasiri, "Model Klasifikasi Untuk Menentukan Kesiapan Kerja Mahasiswa Dan Kelulusan Tepat Waktu Dengan Metode Machine Learning," *IJITECH Indones. J. Inf. Technol.*, vol. 1, no. 1, pp. 1–12, 2023, doi: 10.37680/ijitech.v1i1.xx.

[13] R. Mubarak, M. Hanafi, and D. Sasongko, "Komparasi Performa Naive Bayes Gaussian dan K-NN Untuk Prediksi Kelulusan Mahasiswa dengan CRISP-DM," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 6, pp. 2982–2991, 2024, doi: 10.30865/klik.v4i6.1924.

[14] S. A. Mohamed Yusoff, J. Othman, A. Mohd Mydin, W. A. Wan Mohamad, E. J. Johan, and A. F. Mohamed Yusoff, "Students' Academic Performance: Prediction using Machine Learning Approaches," *Int. J. Acad. Res. Progress. Educ. Dev.*, vol. 13, no. 3, pp. 1778–1792, 2024, doi: 10.6007/ijarped/v13-i3/22164.

[15] L. K. Smirani, H. A. Yamani, L. J. Menzli, and J. A. Boulahia, "Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths," *Sci. Program.*, vol. 2022, 2022, doi: 10.1155/2022/3805235.

[16] O. Ojajuni *et al.*, *Predicting Student Academic Performance Using Machine Learning*, vol. 12957 LNCS, no. September. Springer International Publishing, 2021. 10.1109/ISRITI48646.2019.9034585.

[17] R. Rismala, I. Ali, and A. Rizki Rinaldi, "Penerapan Metode K-Nearest Neighbor Untuk Prediksi Penjualan Sepeda Motor Terlaris," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 585–590, 2023, doi: 10.36040/jati.v7i1.6419.

[18] 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, pp. 169–173, 2023, https://tunasbangsa.ac.id/ejurnal/index.php/jurasik. doi: 10.30645/jurasik.v8i1.552.

[19] U. Putra and I. Yptk, "Penerapan algoritma klasifikasi untuk prediksi tingkat kelulusan mahasiswa menggunakan rappidminer," vol. 6, no. 1, pp. 376–388, 2025, doi: 10.46576/djtechno.

[20] S. Sumarlin and D. Anggraini, "Implementasi K-Nearest Neighbord Pada Rapidminer Untuk Prediksi Kelulusan Mahasiswa," *High Educ. Organ. Arch. Qual. J. Teknol. Inf.*, vol. 10, no. 1, pp. 35–41, 2018, doi: 10.52972/hoaq.vol10no1.p35-41.

[21] N. K. Di *et al.*, "Gudang Jurnal Multidisiplin Ilmu Prediksi Kelulusan Siswa Menggunakan Algoritma K-Nearest," vol. 2, no. November, pp. 110–115, 2024. doi: 10.59435/gjmi.v2i11.1050.