Journal of Education and Science (ISSN 1812-125X)

www.edusj.mosuljournals.com

Comparison of K-Nearest Neighbor Classification Methods and Support Vector Machine in Predicting Students' Study Period

Enggar Novianto*(1), Suhirman Suhirman(2)

[1*,2]Master of Information Technology Study Program, University of Technology Yogyakarta, Indonesia

Article information

Article history:
Received November 23, 2023
Accepted December 22, 2023
Available online March 01, 2024

Keywords:
Classification
Student Study Period
SVM
KNN

Correspondence:
Enggar Novianto
6220211003.enggar@student.uty.ac.

Abstract

State Universities and Private Universities compete fiercely to produce quality students in line with the development of the world of education in Indonesia. Universities strive to improve quality and provide the best education to students and the number of students who graduate on time or not. In this research, a comparative test of the performance of the accuracy values of the K-Nearest Neighbor algorithm and the Support Vector Machine was carried out as a classification method for predicting the study period of students in the Bachelor of Law study program, Faculty of Law. Law, Sebelas Maret University, Surakarta, Indonesia using the RapidMiner application. In this study, a comparison of two classification methods was used, namely K-Nearest Neighbor and Support Vector Machine with 433 student data used. The data is divided into 70% training data and 30% test data. The test results for the highest K-NN prediction accuracy value were at K=5, namely 98.45%. While for the Support Vector Machine method, the accuracy value using the SVM model was 96.90%. Therefore, the results of this research are included in the good category in producing high accuracy, so that the contribution of the K-NN modeling research results using the value K=5 is getting the best accuracy compared to the SVM method using the SVM in predicting student study periods. class of 2021, Bachelor of Law study program, Faculty of Law, Sebelas Maret University.

DOI: 10.33899/edusj.2023.144865.1408,

1. Introduction

State Universities and Private Universities compete fiercely to produce high quality students in line with developments in the world of education in Indonesia. The university strives to improve the quality and provide the best education to students and the number of students who graduate on time or not [1]. A university is important because it provides timely information to many students, enabling them to make informed decisions and contribute to the growth of the university [2]. The study program fulfills the three goals of higher education: teaching, research, community service, and management of science and technology. The study program must have the ability to improve and guarantee the quality of the study program. In study programs, students are university assets, student graduation must be considered, and the impact of study time on university accreditation and study programs [3]. The current curriculum emphasizes creative learning, which aims to encourage students to apply relevant knowledge in different situations. The transition and development of knowledge is crucial for universities in developing countries [4]. The S1 Program at the Faculty of Humanities, Sebelas Maret University, Surakarta is a study program that aims to prepare students for the current academic year. Completion of studies on time is a crucial aspect for study programs to improve and increase the quality of education.

University success depends on students, who must meet certain criteria to complete their studies on time. Factors such as lack of interest and time spent studying can affect the duration students spend studying. To improve the quality of learning and increase research quality, a system that assesses students' potential is needed [5]. The performance of students in high-quality educational institutions is crucial, as it is a characteristic of high-quality universities. Many definitions of high-quality

students exist, and the quality of teaching is crucial. However, the quality of teaching cannot be predicted, and management must take steps to ensure the quality of teaching, as this will affect university accreditation [6]. Study program data can be processed quickly and accurately, which helps predict student study time, predict study periods based on data about students, predict graduation, help evaluate programs, enable universities and study programs to produce precise and high-quality graduates and data about graduating students can provide useful information for study programs if utilized optimally [7].

This paper is divided into several parts, namely the second part which contains the theoretical background, explains the classification model used, describes the research flow and literature review. Section three explains the results and discussion related to the implementation of the classification method and finally the conclusion of the results of this research.

In this research, a comparative test of the performance of the accuracy values of the K-Nearest Neighbor algorithm and the Support Vector Machine was carried out as a classification method for predicting the study period of students in the Bachelor of Law study program, Faculty of Law Sebelas Maret University, Surakarta, Indonesia. To apply this method, use the rapid miner tool to find the accuracy value so that it can produce a predicted class from the student's study time [8].

2. Literature Review

In their study, Asril and M. Isa [4] used the K-Nearest Neighbor (K-NN) algorithm to predict student learning periods based on their performance at the end of the study. The dataset used is an internal dataset from SQL Server. The study found that the K-NN algorithm had a significant effect on student performance, with 93.2% of students achieving a good learning status, 91.5% in total years, and 75.62% in total semesters. This suggests that the K-NN algorithm can be explained based on student performance.

S. Alfere and Y. Maghari [9] investigated student performance based on K-NN classifications modified like K-NN Cosine, K-NN Cubic, and K-NN Weighted. Data from Gaza high school students was used, with 30% for classification and 70% for training. The results showed that K-NN Weighted was the most accurate classification algorithm, with a 94.3% accuracy rate and a faster time compared to other algorithms. K-NN Cosine had an 81.3% accuracy rate in the other two algorithms.

Mailana et al [10] predicted student graduation timeliness using predictor variables like gender, marital status, employment status, and Grade Point Average. Data from 2016 to 2019 was used, with 97 data from departments and graduates. Results showed that SVM and C4.5 data mining techniques can accurately predict graduation timeliness for students at Al-Hidayah University, Bogor, with SVM achieving 85% accuracy and C4.5 achieving 80% accuracy.

In their study, Budiyantara et al [1] highlighted the STMIK Widuri campus addressed the issue of decreasing student graduation rates on time using data mining techniques. The research data comes from the STMIK Widuri Academic and Student Administration Bureau's archives, focusing on students majoring in Informatics Engineering and Information Systems during the academic years 2010/2011, 2011/2012, 2012/2013, and 2013/2014. The study used 242 data for instructions and 100 data for tests, comparing three methods: Decision Tree, Naive Bayes, and K-Nearest Neighbor. The results showed that Decision Tree had the highest accuracy at 98.04%, Naïve Bayes at 96.00%, and K-Nearest Neighbor at 90.00%.

Hairani's research [11] aimed to improve the performance of the SVM method for classifying graduate students at Bumigora University using KNN and k-means-SMOTEe imputation methods. The study involved collecting student graduation data from 2009 to 2012, pre-processing it using KNNI and k-means-SMOTE, and testing its performance using f measurements. The results showed that the integration of these methods resulted in an accuracy of 83.9%, sensitivity of 81.3%, specificity of 86.6%, and f-measure of 83.5%, suggesting that the use of these methods can enhance the SVM method's performance.

In their research, Rachmatika and Bisri [12] used data from the 2017/2018 academic year at Pamulang University Teaching Information Technology Project and employed nine data mining methods. Two models, Random Forest (RF) and Generalized Bayes (GBT), were found to be more effective than each other. RF showed higher accuracy rates in evaluation using the AK-Reg.A and AKReg.B datasets, while GBT showed higher accuracy using the TI-Reg.A and TI-Reg.B datasets. This suggests that RF is a suitable solution for evaluating college students' academic performance.

Haryatmi's and Hervianti's research [2] used 2181 data taken from 2009 to 2016. Age, gender, GPA, IPS 1, IPS 2, IPS 3, IPS 4, major, credits, and study period are all used as student data attributes, student age is calculated from the date of birth until the date the student registers at the University. Data processing uses the WEKA application using the confusion matrix technique. The test results of the first group, which had 90% training data and 10% testing data, showed that the SVM algorithm provided a very good accuracy value, namely 94.4%.

Wiyono et al [13] Accreditation depends on student graduation timeliness, and timely graduation leads to better accreditation. Research on student performance using machine teaching algorithms like K-NN, SVM, and Decision Tree was conducted using 1530 data from students at Harapan Bersama Polytechnic. The R Studio application was used to build a model with training data, and the model was tested for accuracy. The comparison of the three algorithms showed that SVM had the best accuracy of 95%, K-NN 94.5%, and Decision Tree 93% in predicting student performance.

Zeniarja et al [6] A model is needed that can predict student graduation rates to be used as a basis for policy making. By comparing the highest level of accuracy of several classification algorithms, such as Naive Bayes, Random Forest, Decision Tree, K-Nearest Neighbor, and Support Vector Machine, the dataset used in the research amounted to 2293 records. This data comes from students who graduated from the Bachelor of Information Engineering program from 2012 to 2017. In this model, feature selection with the best features is used with twelve regular attribute features and one attribute as a label. It is shown that the model is classified with the selected Random Forest algorithm, which has the highest accuracy value of 77.35% compared to other algorithms.

A summary of the results of previous research can be seen in the Table 1.

Table 1. Summarize The Result

No	Reference	Classification Method	Accuracy
1	[4]	K-Nearest Neighbor	93.2%
2	[9]	K-Nearest Neighbor	K-NN Weighted : 94.3% K-NN Cosine : 81.3%
3	[10]	Support Vector Machine and C4.5	SVM : 85% C4.5 : 80%
4	[1]	Decision Tree, Naïve Bayes and K-Nearest Neighbor	Decision Tree : 98.04% Naïve Bayes :96.00% K-Nearest Neighbor: 90.00%
5	[11]	Support Vector Machine	83.9%
6	[12]	Decision Tree, Random Forest, Gradient Boosted Tree, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Neural Network, Auto Multilayer Perceptron, and Support Vector Machine	The study consists of six classification models, including DT, NB, kNN, LR, NN, MLP, and SVM, and two more powerful models, RF and GBT. Both models have similar performance but have different performance levels. RF has a higher performance in evaluation using AK-Reg.A and AKReg.B datasets, with 90% accuracy, while GBT has a higher
7	[2]	Support Vector Machine	94.4%
8	[13]	K-Nearest Neighbor, Support Vector Machine, Decision Tree	K-NN : 94.5% SVM : 95% DT : 93%
9	[6]	Naïve Bayes, Random Forest, Decision Tree, K-Nearest Neighbor, and Support Vector Machine	It is shown that the model is classified with the selected Random Forest algorithm, which has the highest accuracy value of 77.35% compared to other algorithms.

3. Methodology

This research proposes to compare the level of accuracy using the K-Nearest Neighbor and Support Vector Machine classification algorithms to predict student study periods. The methodology includes the use of student data sets to classify study period predictions. This research model produces a model that is evaluated using the accuracy performance measures shown in Figure 1.

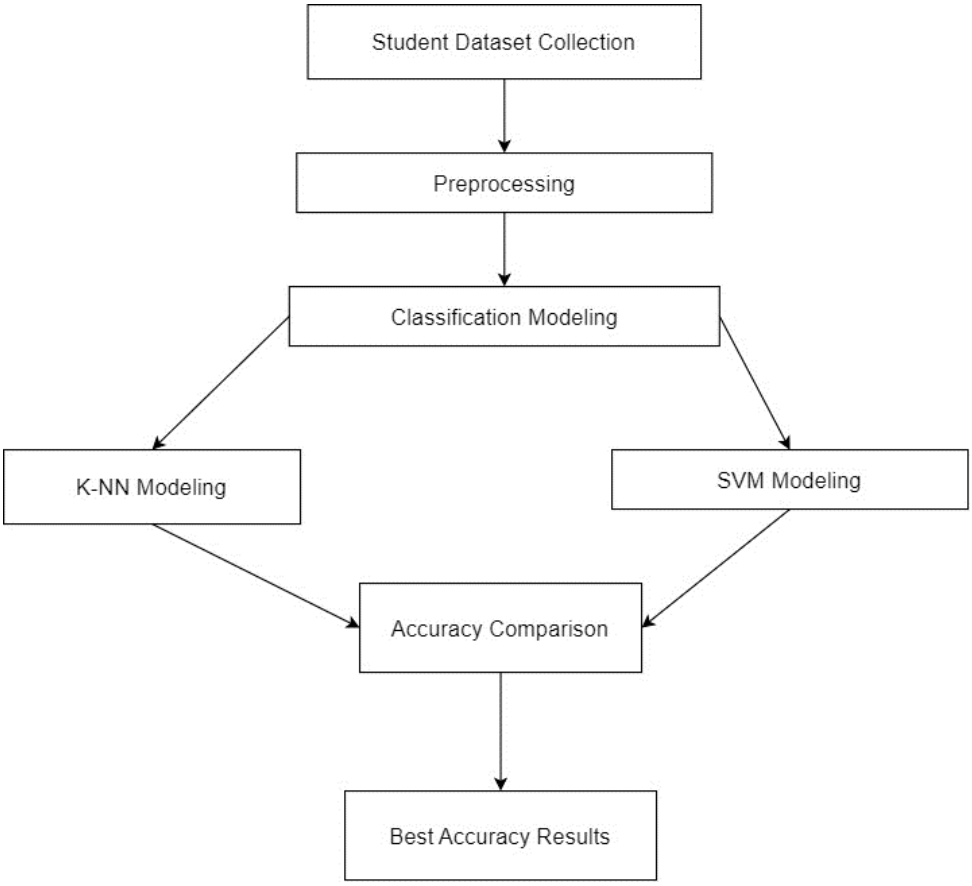

Figure 1. Proposed Methodology

3.1 Dataset Collection

The research method starts with data collection. Data collection was carried out in the Academic Section of the Bachelor of Law Study Program, Faculty of Law, Sebelas Maret University, class of 2021, totaling 433 data. The exam dataset has the attributes Student Name, Gender, 1st Semester IP, 2nd Semester IP, 3rd Semester IP, 4th Semester IP, 1st Semester Credits, 2nd Semester Credits, 3rd Semester Credits, 4th Semester Credits, Credits, GPA, and Status. The dataset attributes used in the name variable use the nominal data type, the gender variable uses the nominal data type, the Achievement Index for Semesters 1 - 4 with the real data type, Semester Units 1-4 with the Integer data type, and Status with the nominal data type with the correct class label. on time or late. GPA is the cumulative average points a student receives each semester. GPA is the average value of learning outcomes, with a value of 0 indicating the lowest score and a value of 4 indicating the highest score [14] SKS is a semester credit unit that has been taken by a student and status is a label for the class that contains accurate or late in predicting the student's study period. The following student data and data and attributes can be seen in Table 2. The 12 attributes that have not been preprocessed can be seen in Table 2.

Table 2. Student Dataset

Name	Gender	GP 1	GP 2	GP 3	GP 4	SKS 1	SKS 2	SKS 3	SKS 4	GPA	Status
Student1	Female	3.87	3.69	3.71	3.83	20	21	24	23	3.77	Appropriate
Student2	Female	3.83	3.73	3.89	3.86	20	21	24	23	3.83	Appropriate
Student3	Male	3.72	3.68	3.81	3.8	20	21	22	23	3.75	Late
Student4	Female	3.91	3.86	3.81	3.92	20	21	24	23	3.87	Appropriate
Student5	Female	3.94	3.74	3.8	3.95	20	21	24	23	3.86	Appropriate
……..	……	…….	…….	…….	…….	…….	……..	……	……..	…….	……
Student 429	Female	3.87	3.96	3.93	3.91	20	21	24	23	3.92	Appropriate
Student430	Female	3.76	3.87	3.89	3.95	20	21	24	23	3.87	Appropriate
Student431	Female	3.91	3.84	3.9	3.93	20	21	24	21	3.9	Late
Student432	Female	3.54	3.59	3.67	3.82	20	21	24	23	3.66	Appropriate
Student433	Female	4	4	4	4	20	21	24	23	4	Appropriate

Table 3. Attributes Before Preprocessing

No	Variable	Data Type
1	Name	Nominal
2	Gender	Nominal
3	GP 1	Real
4	GP 2	Real
5	GP 3	Real
6	GP 4	Real
7	SKS 1	Integer
8	SKS 2	Integer
9	SKS 3	Integer
10	SKS 4	Integer
11	GPA	Integer
12	Status	Nominal

3.2 Data Preprocessing

The data is prepared before analysis, the stage in which the data is prepared, which is used to test the algorithm model, which consists of the data cleaning stage, which is the process of characterizing the data structure [15]. Preprocessing is carried out to prevent inconsistent, erroneous, and imperfect data. Preprocessing is done to improve the quality of the data by removing noise [16]. Data preprocessing is used in realtime databases to improve data quality by cleaning data, a process of removing irrelevant, long, or irrelevant data. This process ensures the accuracy and relevance of the data being analyzed [17]. The data obtained was 433 with 70% for training data and 30% for test data. Data will be cleaned if there is empty data or duplicate data. This data processing stage is carried out to look for attributes that can be used as predictions. After data preprocessing is carried out, the data set used in this data mining process uses 11 attributes. The details of the attributes used can be seen in Table 4.

Table 4. Attributes After Preprocessing

No	Variable	Data Type
1	Gender	Nominal
2	GP 1	Real
3	GP 2	Real
4	GP 3	Real
5	GP 4	Real
6	SKS 1	Integer
7	SKS 2	Integer
8	SKS 3	Integer
9	SKS 4	Integer
10	GPA	Integer
11	Status	Nominal

3.3 Techniques and Methods Used in Classifying Student Study Periods

K-Nearest Neighbor (K-NN) is a classification algorithm that compares the distance between two points in training data and the data to be evaluated [3]. The algorithm classifies incoming data for regression problems by calculating the distance between previous instances in the database, selecting the k closest cases, and determining their meaning or position [17]. K-Nearest Neighbor (K-NN) is a non-parametric algorithm used for classifying large data sets based on the shortest distance between training data and evaluated information [18]. Supervised learning enhances KNN classification performance, as the model assigning observations class labels to training samples is closest to the classification prototype [19]. K-Nearest Neighbor (K-NN) is a machine learning algorithm that classifies very close objects by determining the shortest distance between the evaluated information and the training data [20].

Support Vector Machine (SVM) is a new learning method based on statistical theory, used to solve classification problems of different types. It was first introduced by Boser, Guyon, and Vapnik in 1992 and is a data-driven technique that uses statistical methods for prediction [21]. SVM uses non-linear transformation and relevant kernel function to transform vector input into tinggi dimensions, enabling accurate product operation in dimensions [22]. SVM is a classification of linear and non-linear data, the input data represents predictor variables, and the output is an interdependent target variable, which finds an efficient function to distinguish members of two classes [23] Binary SVM models are trained to classify sprain or nonsprain motions from data sensors. It maps data points onto a high-dimensional space and finds the optimal hyperplane, which divides the data into two classes [24]. SVM is an algorithm that uses nonlinearity to convert data into larger dimensions, ensuring data from two axes can be represented by hyperplane, enhancing its performance in various functions [25]. Backing A well-known and straightforward machine learning and classification algorithm is called Vector Machine. The SVM approach is applies to both linear and non-linear data sets. The decision boundary, or optimal separation hyperplane, is merely a line that is drawn to divide the two classes based on the various classification criteria [26].

Data mining is a data analysis method used to discover hidden patterns in large data sets. It is a scientific discipline that combines machine learning techniques, pattern recognition, statistics, databases, and visualization to address the problem of retrieving information from large databases [17]. The process of extracting and identifying valuable patterns from massive amounts of data through the use of artificial intelligence, machine learning, and statistics is known as data mining. Numerous tasks, including description, prediction estimate, classification, grouping, and association, can be assigned to it. Important information is extracted from massive databases, or "big data," using a sophisticated process that combines machine learning, statistics, mathematics, and artificial intelligence. Extraction and identification of data for particular information pertaining to a huge data or large database is the aim of data mining [27]. Using and processing data to produce new and valuable information is the goal of data mining [28].

Classification is a method used to categorize each item in a data set into a set of classes or groups, aiding linear programming, decision trees, artificial neural networks, and statistics [29]. Classification is a supervised learning method because it requires training data to create rules that classify testing data into certain classes or groups. Classification techniques also known as classifier techniques are systematic approaches to creating classification models from a set of input data [30]. Classification studies have been carried out in various scientific disciplines with different data. The purpose of classification is to facilitate the process of making decisions by searching for models, characters, and concepts from the dataset for each class [31].

3.4 Matrix of Confusion

The research evaluated the model's accuracy by comparing K-NN and SVM classification methods. The Confusion Matrix was used to calculate the accuracy of data mining concepts, displaying the number of correctly and incorrectly classified test data [17]. The matrix of confusion, which is often used in binary classification problems, provides additional information on the classifier's performance in addition to displaying the quality of ranking for the items in the validation set. By displaying the difference between the true and predicted values [18]. One visual aid that is frequently used in supervised learning is the confusion matrix. Every row in the matrix denotes an event in the real class, and every column in the matrix is an illustration of a prediction class [32]. The primary purpose of evaluation metrics is to evaluate the performance of a classifier. This performance is verified by means of mathematical formulas that compare the predictions made by a model with the actual values in the database. Evaluation metrics are used to assess a classifier's performance, comparing predictions with actual data. Precision measures the percentage of correctly classified samples, calculated using the ratio of correct classifications to total classifications, or true positives to positives [33]. The Confusion Matrix can be seen in Table 5 [5].

Table 5. Confusion Matrix

Prediction	Actual	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Samples with positive indications that are appropriately expected to be positive are known as true-positives (TP). False-positive (FP) samples are those that are intended to be positive but are actually negative. Samples with a negative rating even if they were correctly predicted to be negative are known as true negative (TN). Samples that were expected to be negative but turned out to be positive are known as false-negatives (FN) [18].

Accuracy

One indicator of machine learning operations accuracy is the overall percentage of correctly classified samples. The following is how to calculate the accuracy formula [5] :

$$\text{Accuracy} : \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision

A machine learning job performance measure called precision compares the quantity of samples properly classified to the total quantity of samples [18]. One divides the total number of correct categories by the total number of classifications made [34]. The following is how to calculate the precision formula [5] :

$$\text{Precision} : \frac{TP}{TP+FP} \quad (2)$$

Recall

The percentage of positive marker samples that were correctly anticipated is known as recall. Recall, also known as true positive sensitivity or rate, quantifies how well a classifier predicts real positive samples [18]. The following is how to calculate the recall formula [5] :

$$\text{Recall} : \frac{TP}{TP+FN} \quad (3)$$

4. Results And Discussion

This research uses the K-NN and SVM classification methods as a prediction of student study periods and to measure the system's accuracy value by comparing the dataset with factual data on actual results. The first test used the K-NN algorithm on an internal student dataset which was used as test data totaling 433 data with 11 attributes. The second test of this research was to determine which method got the highest accuracy value. The second test uses a classification of student study periods using the SVM method. This classification is used to predict a student's study period. The results of the data preprocessing will be divided into two data, namely 70% training data, or 304 data, and 30% test data, or 129 data, then after that calculations will be carried out on the data using the K-NN and SVM classification algorithms by calculating the proximity to the data existing training. The classification results of the K-NN and SVM methods can be seen in Table 6 and Table 7.

Table 6. Example of K-NN Classification Results

Gender	GP 1	GP 2	GP 3	GP 4	SKS 1	SKS 2	SKS 3	SKS 4	GPA	Status	Confidence Appropriate	Confidence Late	Prediction
Female	3.87	3.69	3.71	3.83	20	21	24	23	3.77	Appropriate	1.000	0	Appropriate
Female	3.94	3.74	3.8	3.95	20	21	24	23	3.86	Appropriate	1	0	Appropriate
Male	3.94	3.93	3.64	3.92	20	21	20	23	3.86	Late	0	1	Late
Female	3.94	3.89	3.87	3.86	20	21	24	23	3.89	Late	1	0	Appropriate
Male	3.64	3.22	3.8	3.85	20	21	24	24	3.7	Late	1	0	Appropriate

Table 7. Example of SVM Classification Results

Gender	GP 1	GP 2	GP 3	GP 4	SKS 1	SKS 2	SKS 3	SKS 4	GPA	Status	Confidence Appropriate	Confidence Late	Prediction
Female	3.87	3.69	3.71	3.83	20	21	24	23	3.77	Appropriate	1	0	Appropriate
Female	3.94	3.74	3.8	3.95	20	21	24	23	3.86	Appropriate	1	0	Appropriate
Male	3.94	3.93	3.64	3.92	20	21	20	23	3.86	Late	0	1	Late
Female	3.94	3.89	3.87	3.86	20	21	24	23	3.89	Late	1	0	Appropriate
Male	3.64	3.22	3.8	3.85	20	21	24	24	3.7	Late	0	1	Late

Testing the accuracy of the predictions produced using the K-NN and SVM classification methods using the RapidMiner application. The RapidMiner application is an application used to process mining data in machine learning. The software called RapidMiner was developed by the same company and is used for business and commercial applications, as well as for research, training, education, prototyping, and application development. It provides an integrated environment for machine learning, data mining, text mining, predictive analytics, and business analytics. It also supports all steps of the data mining process, including data preparation, the RapidMiner Basic Edition visualization can be downloaded under the AGPL license and is built on the open core model [35]. Here's how to process data using the K-NN and SVM methods in RapidMiner:

1. The next step is to model with K-Nearest Neighbor. Next, select the test data and training data that have been imported, and select the Retrieve Student Data, Split Data, k-NN, Apply Model, and Performance operators. The appearance of the K-NN model can be seen in Figure 2.

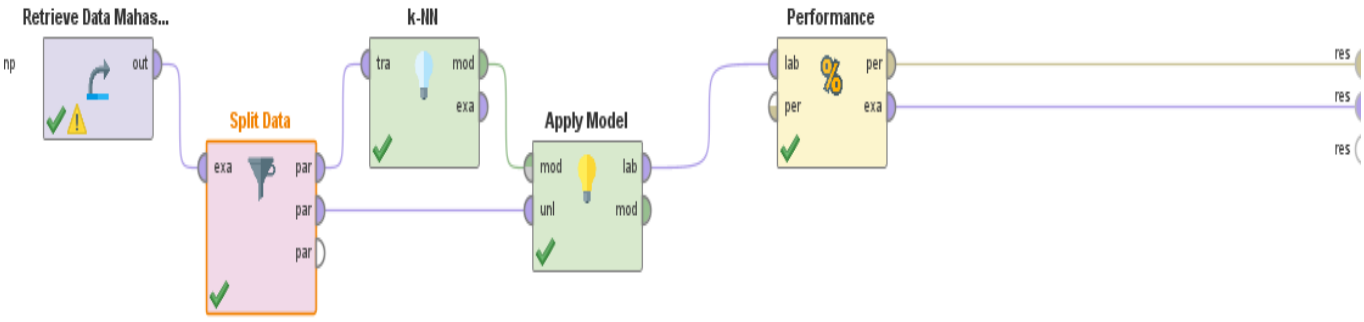

Figure 2. K-NN Model

Classification results by dividing 70% of the data as training data and 30% as testing data using the K-NN method using the attributes student gender, IPS 1, IPS 2, IPS 3, IPS 4, SKS 1, SKS 2, SKS 3, SKS 4 and GPA and status get predicted results for the student's study period with accuracy of 98.45. The classification results of the K-NN method using the value K = 5 can be seen in Table 8.

Table 8. Confusion Matrix K-NN

Prediction	Actual	
	Positive	Negative
Positive	91	2
Negative	0	36

Accuracy: 98.45%, Classification Error: 1.55%

2. The next process is to create modeling using the Support Vector Machine method. Next, select the test data and training data that have been imported, and select the operator Retrieve Student Data, Nominal to Numerical, Split Data, SVM, Apply Model, and Performance. The appearance of the SVM model can be seen in Figure 3.

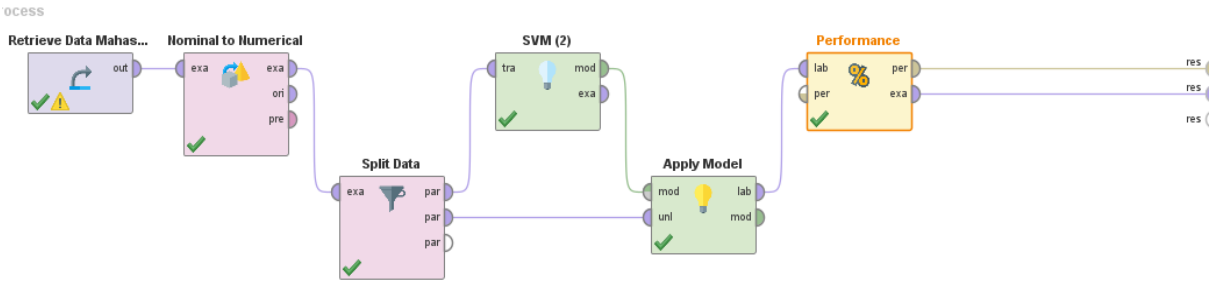

Figure 3. SVM Model

Classification results by dividing 70% of the data as training data and 30% as testing data using the SVM method using the attributes of student gender, IPS 1, IPS 2, IPS 3, IPS 4, SKS 1, SKS 2, SKS 3, SKS 4 and GPA and status get predicted results for the student's study period with an accuracy of 97.67%. The classification results of the SVM method can be seen in Table 9.

Table 9. Confusion Matrix SVM

Prediction	Actual	
	Positive	Negative
Positive	88	1
Negative	3	37

Accuracy: 96.90%, Classification Error : 3.10%

From the results of comparative testing of the K-Nearest Neighbor and Support Vector Machine methods, the results obtained can be seen in Table 10.

Table 10. Accuracy Testing Results

Method	Accuracy
K-NN	98.45%
SVM	96.90%

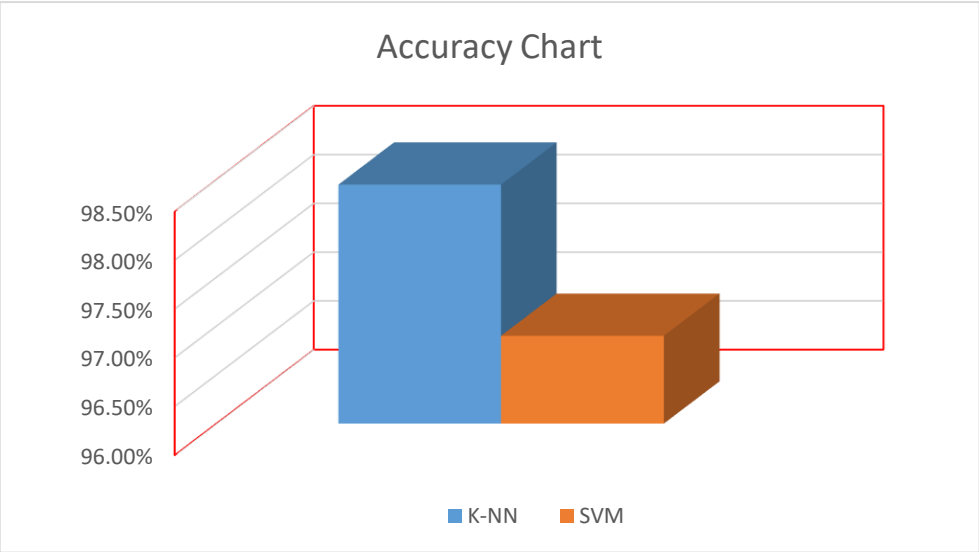

Figure 4. Accuracy Comparison Chart

As shown in Table 10 and Figure 4, the comparison of the accuracy testing results of the K-NN and SVM methods shows that for K = 5 it is good to apply to K-NN by producing the best accuracy results of 98.45%, while for the SVM which produces a value accuracy of 96.90%.

5. Comparison with Related Works

This comparison was made by comparing the results of previous research. Comparison of accuracy values by looking at chapter two regarding literature review. The resultss of the research model obtained good accuracy values from the results of previous research. Table 11 shows a comparison of the results of this study with previous relevant research.

Table 11. Comparison of Research Performance

Author	Accuracy
Asril & M. Isa, 2020	K-NN : 93.2%
S. Alfere & Y. Maghari, 2018	K-NN Weighted : 94.3% K-NN Cosine : 81.3%
Mailana. et al., 2021	SVM : 85% C4.5 : 80%
Budiyantara. et al., 2020	Decision Tree : 98.04% Naïve Bayes : 96.00% K-Nearest Neigbor : 90.00%
Hairani, 2021	SVM : 83.9%
Rachmatika & Bisri, 2020	Random Forest : 90%
Haryatmi & Hervianti, 2021	SVM : 94.4%
Wiyono, et al., 2020	K-NN : 94.5% SVM : 95% DT : 93%
Zeniarja, et., 2022	77.35%
Proposed System	**K-NN : 98.45%** **SVM : 96.90%**

6. Conclusions

In this research, a comparison of two classification methods, namely K-Nearest Neighbor and Support Vector Machine, was used, with 433 student data used. The data is divided into 70% training data and 30% test data. The test results for the highest prediction accuracy value for K-NN were found at K=5, namely 98.45%. For the Support Vector Machine method, the accuracy was 96.90%. Therefore, the results of this research are included in the good category in producing high accuracy, so that the contribution of the K-NN modeling research results using the value K=5 is getting the best accuracy compared to the SVM method using the SVM in predicting student study periods. class of 2021, Bachelor of Laws study program, Faculty of Law, Sebelas Maret University. The results of these predictions can be used to provide special guidance or treatment for students who receive predictions that are late in completing their studies.

7. Acknowledgments

Writing the results of this research in this paper always received support from my thesis supervisor, namely Ass. Prof. Suhirman, M.Kom., Ph.D at the Master of Information Technology Study Program, Yogyakarta Technology University. He provided many suggestions and input regarding the writing of this paper and expressed his thanks to the management of the Journal of Education and Science for providing the opportunity to submit this paper so that it can be published can provide benefits and knowledge to all readers.

8. References

[1] A. Budiyantara, I. E. Prenngki, P. A. Pratama dan N. Wiliani, "Komparasi Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Memprediksi Mahasiswa Lulus Tepat Waktu," *Jurnal Ilmu Pengetahuan Dan Teknologi Komputer,* vol. 5, no. 2, pp. 265-270, 2020. https://doi.org/10.33480/jitk.v5i2.1214.

[2] E. Haryatmi dan S. P. Hervianti, "Penerapan Algoritma Support Vector Machine Untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu," *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi),* vol. 5, no. 2, pp. 386-392, 2021 https://doi.org/10.29207/resti.v5i2.3007.

[3] I. A. Nikmatun dan I. Waspada, "Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *Jurnal SIMETRIS,* vol. 10, no. 2, pp. 421-432, 2019. https://doi.org/10.24176/simet.v10i2.2882.

[4] T. Asril dan S. M. Isa, "Prediction of Students Study Period using K-Nearest Neighbor Algorithm," *International Journal of Emerging Trends in Engineering Research,* vol. 8, no. 6, pp. 2585-2593, 2020. https://doi.org/10.30534/ijeter/2020/60862020.

[5] A. Putri, C. S. Hardiana, E. Novfuja, F. T. P. Siregar, R. Y. Fatma dan R. Wahyuni, "Comparison of K-NN, Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction," *MALCOM : Indonesian Journal of Machine Learning and Computer Science,* vol. 3, no. 1, pp. 20-26, 2023. DOI : https://doi.org/10.57152/malcom.v3i1.610.

[6] J. Zeniarja, A. Salam dan F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," *Jurnal Rekayasa Elektrika,* vol. 18, no. 2, pp. 102-108, 2022. DOI : https://doi.org/10.17529/jre.v18i2.24047.

[7] N. Hidayati dan A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," *JEATech : Journal of Engineering and Applied Technology,* vol. 2, no. 2, pp. 86-91, 2021. DOI : 10.21831/jeatech.v2i2.42777.

[8] M. R. Qisthiano, "Klasifikasi Terhadap Prediksi Kelulusan Mahasiswa Dengan Menggunakan Metode Support Vector Machine (SVM)," dalam *Seminar Nasional Teknologi Dan Multidisiplin Ilmu SEMNASTEKMU 2022*, Semarang, 2022.

[9] S. S. Alfere dan A. Y. Maghari, "Prediction of Student's Performance Using Modified KNN Classifiers," dalam *International Conference on Engineering & Future Technology (ICEFT 2018),* Gaza, Palestine, 2018.

[10] A. Mailana, A. A. Putra, S. Hidayat dan A. Wibowo, "Comparison of C4.5 Algorithm and Support Vector Machine in Predicting the Student Graduation Timeliness," *JOIN (Jurnal Online Informatika),* vol. 6, no. 1, pp. 11-16, 2021. DOI:10.15575/join.v6i1.608.

[11] Hairani, "Peningkatan Kinerja Metode SVM Menggunakan Metode KNN Imputasi dan K-Means-Smote untuk Klasifikasi Kelulusan Mahasiswa Universitas Bumigora," *Jurnal Teknologi Informasi dan Ilmu Komputer,* vol. 8, no. 4, pp. 713-718, 2021. DOI: https://doi.org/10.25126/jtiik.2021843428.

[12] R. Rachmatika dan A. Bisri, "Perbandingan Model Klasifikasi untuk Evaluasi Kinerja Akademik Mahasiswa," *JEPIN Jurnal Edukasi & Penelitian Informatika,* vol. 6, no. 3, pp. 417-422, 2020. DOI: http://dx.doi.org/10.26418/jp.v6i3.43097.

[13] S. Wiyono, D. S. Wibowo, M. F. Hidayatullah dan D. , "Comparative Study of KNN, SVM and Decision Tree Algorithm for Student's Performance Prediction," *IJCSAM (International Journal of Computing Science and Applied Mathematics),* vol. 6, no. 2, pp. 50-53, 2020. DOI: http://dx.doi.org/10.12962/j24775401.v6i2.4360.

[14] S. Wiyono dan T. Abidin, "Comparative Study Of Machine Learning KNN, SVM, And Decision Tree Algorithm To Predict Student's Performance," *INTERNATIONAL JOURNAL OF RESEARCH - GRANTHAALAYAH,* vol. 7, no. 1, pp. 190-196, 2019. DOI: https://doi.org/10.29121/granthaalayah.v7.i1.2019.1048.

[15] R. P. S. Putri dan I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Informatika Jurnal Ilmu Komputer dan Informatika,* vol. 4, no. 1, pp. 1-7, 2018. DOI: https://doi.org/10.23917/khif.v4i1.5975.

[16] I. N. Switrayana, D. Ashadi, H. dan A. Aminuddin, "Sentiment Analysis and Topic Modeling of Kitabisa Applications using Support Vector Machine (SVM) and Smote-Tomek Links Methods," *International Journal of Engineering and Computer Science Applications (IJECSA),* vol. 2, no. 2, pp. 81-91, 2023. DOI: 10.30812/IJECSA.v2i2.3406.

[17] O. W. Yuda, D. Tuti, L. S. Yee dan S. , "Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest," *SATIN Sains dan Teknologi Informasi,* vol. 8, no. 2, pp. 122-131, 2022. DOI: https://doi.org/10.33372/stn.v8i2.885.

[18] A. M. Ali dan N. N. Saleem, "Classification of Software Systems attributes based on quality factors using linguistic knowledge and machine learning: A review.," *Journal of Education and Science,* vol. 31, no. 3, pp. 66-90, 2022. DOI: 10.33899/edusj.2022.134024.1245.

[19] K. S. T. P. Vital dan K. K. Kumar, "Student Classification Based on Cognitive Abilities and Predicting Learning Performances using Machine Learning Models," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 8, no. 6, pp. 3554-3569, 2020. DOI:10.35940/ijrte.F8848.038620.

[20] H. A. dan S. Alim, "Implementasi Orange Data Mining Untuk Klasifikasi kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes," *NERO Networking Engineering Research Operation,* vol. 6, no. 2, pp. 133-144, 2021. DOI: http://dx.doi.org/10.21107/nero.v6i2.237.

[21] N. W. E. R. Dewi, I. G. A. Gunadi dan G. Indrawan, "Detection of Class Regularity with Support Vector Machine methods," *Lontar Komputer Jurnal Ilmiah Teknologi Informasi,* vol. 11, no. 1, pp. 20-31, 2020. DOI: https://doi.org/10.24843/LKJITI.2020.v11.i01.p03.

[22] J. J. Purnama, H. M. Nawawi, S. Rosyida, R. dan R. , "Klasifikasi Mahasiswa HER Berbasis Algoritma SVM dan Decision Tree," *Jurnal Teknologi Informasi dan Ilmu Komputer,* vol. 7, no. 6, pp. 1253-1260, 2020. DOI: 10.25126/jtiik.202073080.

[23] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C45, Naive Bayes, KNN Dan SVM," *Jurnal Tekno Insentif,* vol. 13, no. 1, pp. 16-25, 2019. DOI: https://doi.org/10.36787/jti.v13i1.78.

[24] N. Francis, H. Suhaimi dan P. E. Abas, "Classification of Sprain and Non-sprain Motion using Deep Learning Neural Networks for Ankle Sprain Prevention," *The International Journal of Computing Journal,* vol. 22, no. 2, pp. 159-1169, 2023. DOI 10.47839/ijc.22.2.3085.

[25] H. A. Mustopa dan A. Y. Kuntoro, "Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa," *Jurnal Ilmu Pengetahuan dan Teknologi Komputer,* vol. 5, no. 2, pp. 211-220, 2020. DOI: https://doi.org/10.33480/jitk.v5i2.1181.

[26] M. J. H. Mughal, "Data Mining: Web Data Mining Techniques, Tools and Algorithms : An Overview," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 9, no. 6, pp. 208-215, 2018. .

[27] A. F. Sallaby dan A. , "Analysis of Missing Value Imputation Application with K-Nearest Neighbor (K-NN) Algorithm in Dataset," *The IJICS (International Journal of Informatics and Computer Science),* vol. 5, no. 2, pp. 141-144, 2021. DOI 10.30865/ijics.v5i2.3185.

[28] A. S. Fitrani, F. dan W. Novarika, "Implementation of Data Mining Using Naïve Bayes Classification Method To Predict Participation of Governor And Vocational Governor Selection in Jemirahan Village, Jabon District," *The IJICS (International Journal of Informatics and Computer Science),* vol. 3, no. 2, pp. 66-79, 2019. DOI 10.30865/ijics.v3i2.1391.

[29] M. Raharjo, R. J. L. Putra dan T. A. A. Sandi, "Implementasi Metode Decision Tree Klasifikasi Data Mining Untuk Prediksi Peminatan Jurusan Robotika oleh Mahasiswa," *Jurnal Teknik Komputer AMIK BSI,* vol. V, no. 2, pp. 161-166,

2019. DOI: https://doi.org/10.31294/jtk.v5i2.4852.

[30] E. E. Barito, J. T. Beng dan D. Arisandi, "Penerapan Algoritma C4.5 Untuk Klasifikasi Mahasiswa Penerima Bantuan Sosial COVID-19," *Jurnal IlmuKomputer dan Sistem Informasi,* vol. 10, no. 1, pp. 1-9, 2022. DOI: https://doi.org/10.24912/jiksi.v10i1.17819.

[31] A. S. dan C. R. Sari, "Data Mining Klasifikasi Kelulusan Mahasiswa Menggunakan Metode Naive Bayes," *Journal Peqguruang : Conference Series,* vol. 4, no. 1, pp. 423-428, 2022.

[32] A. dan O. Pahlevi, "Data Mining Model for Designing Diagnostic Applications Inflammatory Liver Disease," *Sinkron : Jurnal dan Penelitian Teknik Informatika,* vol. 5, no. 1, pp. 51-57, 2020. DOI: 10.33395/sinkron.v5i1.10589.

[33] E. D. Canedo dan B. C. Mendes, "Software Requirements Classification Using Machine Learning Algorithms," *MDPI,* vol. 22, no. 9, pp. 1-20, 2020. https://doi.org/10.3390/e22091057.

[34] D. Chicco dan G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification," *BMC Genomics,* vol. 21, no. 6, pp. 1-13, 2020. https://doi.org/10.1186/s12864-019-6413-7.

[35] D. Arunadevi, S. dan M. Raja, "A Study of classification algorithms using Rapidminer," *International Journal of Pure and Applied Mathematics,* vol. 119, no. 12, pp. 15977-15988, 2018.

مقارنة طرق تصنيف الجيران الأقرب Kوآلة المتجهات الداعمة في التنبؤ بفترة دراسة الطلاب

إنجار نوفيانتو[1*]، سوهيرمان [2]

[2،*1] برنامج ماجستير دراسة تكنولوجيا المعلومات، جامعة التكنولوجيا يوجياكرتا، إندونيسيا

المستخلص

تنافس الجامعات الحكومية والجامعات الخاصة بشدة لإنتاج طلاب ذوي جودة عالية بما يتماشى مع تطور عالم التعليم في إندونيسيا .تسعى الجامعات إلى تحسين الجودة وتوفير أفضل تعليم للطلاب وعدد الطلاب الذين يتخرجون في الوقت المحدد أم لا .في هذا البحث تم إجراء اختبار مقارن لأداء قيم دقة خوارزمية K-Nearest Neighbor وآلة ناقل الدعم كطريقة تصنيف للتنبؤ بفترة دراسة الطلاب في برنامج دراسة بكالوريوس الحقوق، كلية الحقوق .قانون .القانون، جامعة سيبيلاس ماريت، سوراكارتا، إندونيسيا باستخدام تطبيق.RapidMiner في هذه الدراسة، تم استخدام المقارنة بين طريقتين للتصنيف، وهما-K أقرب جار وآلة ناقل الدعم مع استخدام بيانات 433 طالبًا .تنقسم البيانات إلى 70% بيانات تدريب و 30%بيانات اختبار .كانت نتائج الاختبار لأعلى قيمة دقة تنبؤ K-NN عند5 = K ، وهي .98.45 بينما بالنسبة لطريقةSupport Vector Machine ، كانت قيمة الدقة باستخدام نموذج .96.90% SVM لذلك تم إدراج نتائج هذا البحث ضمن الفئة الجيدة في إنتاج دقة عالية، بحيث أن مساهمة نتائج بحث النمذجة K-NN باستخدام القيمة K=5 تحصل على أفضل دقة مقارنة بطريقة SVM باستخدام SVM في التنبؤ بفترات دراسة الطالب .دفعة2021 ، برنامج دراسة البكالوريوس في القانون، كلية الحقوق، جامعة سيبيلاس ماريت