Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Implementation of data mining predictions for student her-registration

Bosker Sinaga[1], Meman Marpaung[2], Fretty Wandani Ginting[3], Puspita Sari[4]
[1,2,3,4]Informatics Engineering, STMIK Pelita Nusantara, Medan City, North Sumatra, Indonesia

Article Info
Keywords:
data mining,
knn algorithm,
her-registration

ABSTRACT
Predictions of an event are carried out to see what will happen in the future, including predictions for students who will register. This is done so that the head of the study program or student affairs will immediately move to approach students who are predicted not to be likely to register. This research implements data mining predictions of student registration. There are often wrong predictions about who might not register. So that the student's approach to decision-making before the time of registration is not carried out by the study program or student affairs. Research Method, namely the survey research method, is a research method carried out using surveys or direct data collection from interested parties, namely the head of the study program. The algorithm used to analyze the data is the K-nearest neighbor algorithm. This research aims to apply data mining to predict student her-registration and to carry out analysis using data mining to predict student her-registration. The research results of applying data mining with the K-Nearest Neighbors algorithm can solve the problem of student enrollment prediction analysis. Of the 10 students who predicted their registration, 7 students would register and 3 students would not register.

Corresponding Author:
Bosker Sinaga
Informatics Engineering, STMIK Pelita Nusantara, Medan City, North Sumatra, Indonesia
boskersinaga@gmail.com

INTRODUCTION

Technological advances are developing so quickly, that their application can help the tasks of technology users. One application of technology that is widely used is education, especially higher education. Several large universities are assisted by the application of technology in decision-making [1] even in making predictions.

STMIK Pelita Nusantara is one of the best higher education candidates in North Sumatra which has five study programs [2] namely 1) Informatics Engineering (S1); 2) Software Engineering (S1); 3) Information Technology (S1); 4) Digital Business; 5) Network Computer Engineering Technology (D4); and 6) Information Management (D3). The vision of STMIK Pelita Nusantara is that STMIK Pelita Nusantara will become a superior higher education and computer education center that produces graduates with national competitiveness by 2024.

STMIK Pelita Nusantara at the beginning of every semester carries out her registration or what is often called re-registration. In this re-registration, there are often

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

wrong predictions about who might not register. So that the student's approach to decision-making before the time of registration is not carried out by the study program or student affairs. In this case, there was a decrease in the number of students re-registering from the previous semester. To solve this problem, STMIK Pelita Nusantara should apply technology, especially the application of data using methods/algorithms, so that the problem of predicting student re-admission can be solved before the time for re-registration.

Data mining is a scientific discipline that studies methods for extracting knowledge or finding patterns from data. Data mining is a data processing method to find hidden patterns in the data [3]. The K-Nearest Neighbor algorithm often called KNN is a classification algorithm that uses the close distance between data and other data. In the K-Nearest Neighbor algorithm for q-dimensional data, the distance from that data to other data can be calculated [4].

Several previous researchers used as a reference for this research is research [5] with the title Application of the K-Nearest Neighbor Algorithm to Predict the Quality of Consumable Water. The research results are a collection of data obtained from the Kaggle website, the modeling results are measured using the Confusion Matrix table to calculate accuracy. After being tested, this model has the highest accuracy level of 85.24% with a value of k (nearest neighbor) = 3.

Research [6] entitled Application for Student Graduation Prediction Based on K-Nearest Neighbor (K-NN). Based on tests carried out using K-fold cross-validation, it was found that the highest accuracy in the third model was 80% when the K-fold was 4th and 61% when the K value = 1. Meanwhile, testing using the Confusion Matrix obtained the highest accuracy of 98% at K=1 for the "On Time" classification, and 98% at K=2 for the "Not On Time" classification.

Research by [7] with research results obtained benefits in planning and scheduling the provision of clothing stock for the future accurately. By calculating data mining using classification techniques with the K-Nearest Neighbor algorithm which is the most dominant, it can be predicted that the number of sales in the next period will increase with an average monthly prediction of 14,900 and the most popular clothing brand is Cardinal.

Research by [8] with research results that the KNN algorithm can predict stroke based on gender, age, hypertension, history of heart disease, marital status, type of work, type of residence, average glucose levels, BMI, and smoking status with the accuracy obtained of 95% with a value of k=9.

Research by [9] with research results on the best k value for predicting shoe sales is as follows: For the 100th buyer, the best k value to use to predict the type of shoe that will be purchased is the value k = 9 with a success rate of 49%. For the 100th buyer, the best k value to use to predict the purchase of shoe types is k = 7 with a success rate of 74%. For the 100th buyer, the value k = 1 is the best k value to be used to predict a student's study period with a success rate of 89%.

Another research [10] entitled Application of Data Mining to Predict Rain Estimates Using the K-Nearest Neighbor Algorithm. The results obtained were that the data testing decision was NO. In other words, data mining and the K-Nearest Neighbor algorithm can

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

help the problem-solving process. Based on previous research, it is very suitable for predictions to use the K-Nearest Neighbor algorithm.

METHODS

Research methods are important for a researcher to achieve a goal and be able to find answers to the problems posed. The research stages start from identifying the problem to publishing scientific articles, as in the following fishbone diagram:

Figure 1. Research Stages

Identifying Problems

The first step in this research is identifying the problem to find out what problems exist in student registration so that the researcher better understands the problem to be studied. From the observations made by researchers so far, there are often incorrect predictions about who might not be registered so the head of the study program or student affairs does not first approach students who are predicted not to register.

Proposal Preparation

After the researchers knew about the problem, the research team then prepared a research proposal and RAB to be submitted to LPPM STMIK Pelita Nusantara to be funded in implementing the tri dharma of higher education, one of which was odd semester TA research. 2022/2023. In preparing the research proposal, all teams were involved to provide input. Both from the lecturer team and the student team.

Data collection

The steps taken in data collection are research data sources which are divided into 2, namely primary data sources and secondary data sources (Sugiyono, 2015). The primary data in this research is drug sales data and drug inventory data, attribute data that will be used in the selection. The secondary data used is by searching for journals that support the research to be conducted and are by the research topic.

Analyzing Data

In analyzing the data that has been obtained, the researcher then analyzes the data using KNN algorithm calculations to get predictions for student registration.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Update of Analysis Results

Data updating is carried out so that the results of data analysis using the KNN algorithm can be synchronized with the understanding of STMIK Pelita Nusantara as the manager. Accredited National Journal Publication. In the final stage, the research output is the publication of the Accredited National Journal, Sinta.

Preparation of reports

At this stage, the researcher prepares a final research report to be submitted to LPMM STMIK Pelita Nusantara as a final accountability report by the schedule determined by LPMM STMIK Pelita Nusantara.

RESULTS AND DISCUSSION

Data analysis

In this research, in the process of calculating the data used as a sample of Her-Registration data for the odd semester 2023/2024 STMIK Pelita Nusantara. The criteria for this research can be seen in table 1.

Table 1. Attribute

No	Attribute	Weight
1	IPK	25
2	Tuition fee	25
3	Lecture Attendance	20
4	Less familiar with the application	20
5	Status	10

Table 2. Raw data

No	Name	NIM	Sem	study program	Force	Last GPA
1	Maria Magdalena	190121088	7	TIF	2019	3
2	Santa Elisa Br. Tarigan	200121001	7	TIF	2020	4
3	Muhammad Aizad	200121002	7	TIF	2020	1.2
...	...	...	...	...	...	...
514	Ariani N. Sianturi	220121101	3	TIF	2022	3.1
515	Verawati S. Hutabarat	220111001	3	MIF	2022	3
518	Elpika Bulele	220161005	3	BDG	2022	3.71
...	...	...	...	...	...	...
555	Angeline M. Br Ginting	220161052	3	BDG	2022	2.96
556	Fathia A. Siregar	220141002	3	TGI	2022	4
...	...	...	...	...	...	...
637	Siwa Nesen	220141056	3	TGI	2022	3

Data Selection

The research data obtained is data in raw form which must undergo a selection stage to be used in research. Not all research data can be used, in this study the researcher took the variables needed for the research, namely GPA, Tuition fee, Lecture Attendance, Lack of Understanding of the Application, and Status. The selected data is to be used as research

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

data so that it is by the research targets. In Table 3 below are the results of the data selection.

Table 3. Data Processed

No	Name	Last GPA	Tuition fee	Lecture Attendance	Lack of understanding of the application	Student Status
1	Ninda Lestari	4	Paid off	Full	Understand	No Scholarship
2	Santa Elisa Br. Tarigan	4	Paid off	Full	Understand	KIP
3	Muhammad Aizad	1.2	Paid off	Not Full	Understand	No Scholarship
4	Fachrizal Husaini	0.6	Paid off	Not Full	Understand	No Scholarship
5	Tri Adri Anov	3.6	Belum Paid off	Full	Understand	No Scholarship
6	Nurul Amalia	3.6	Belum Paid off	Full	Understand	No Scholarship
7	Meisya Tiara Fernanda	3	Paid off	Full	Understand	No Scholarship
8	Sari Mutiara Dewi	3.4	Belum Paid off	Full	Understand	No Scholarship
9	Aura Nissa Galuh Suanda	3.6	Belum Paid off	Full	Understand	No Scholarship
10	Juanda	3.4	Paid off	Full	Understand	No Scholarship

Data Preprocessing

After data selection, the next step is data preprocessing. Data preprocessing is changing data by changing the data variables used into variable codes. The variables that will be used are GPA, Tuition fee (UK), Lecture Attendance (KP), Lack of understanding of the application (KPA), and Student Status (SM).

Table 4. Data Preprocessing

No	Code	IPK	UK	KP	KPA	SM
1	A1	5	5	5	5	5
2	A2	5	5	5	5	3
3	A3	1	5	3	5	5
4	A4	1	5	3	5	5
5	A5	4	3	5	5	5
6	A6	4	3	5	5	5
7	A7	3	5	5	5	5
8	A8	3	3	5	5	5
9	A9	4	3	5	5	5
10	A10	3	5	5	5	5

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

The data obtained in this study were student registration data for the odd semester of TA. 2023/2024 at STMIK Pelita Nusantara. The attributes used are Last GPA, Tuition fee, Lecture Attendance, Lack of understanding of the application and status.

Data Transformation

Next, carry out data transformation techniques to place the prediction results obtained from the research site and then carry out the calculation process using the KNN method. The following is a data transformation table which can be seen in Table 5 below:

Table 5. Data Preprocessing

TEST DATA							
No	Kode	IPK	UK	KP	KPA	SM	Information
1	A1	5	5	5	5	5	HER-REG
2	A2	5	5	5	5	3	HER-REG
3	A3	1	5	3	5	5	HER-REG
4	A4	1	5	3	5	5	HER-REG
5	A5	4	3	5	5	5	HER-REG
6	A6	4	3	5	5	5	HER-REG
7	A7	3	5	5	5	5	HER-REG
8	A8	3	3	5	5	5	HER-REG
9	A9	4	3	5	5	5	HER-REG
10	A10	3	5	5	5	5	HER-REG
11	A11	5	3	3	5	5	??

Based on the test data on the 5 students above, it will be predicted whether these students will register in the even semester of TA. 2023/2024 with this value.

KNN algorithm

1. First, we determine the parameter K. First, we must determine the value of K first. There is no exact formula for determining the K value. However, one tip that can be considered is that if the number of classes is even then the K value should be odd, conversely if the number of classes is odd then the K value should be even. For example, we make the number of nearest neighbors K = 3.
2. Second, we calculate the distance between the new data and all the training data. We use Euclidean distance. We calculate based on the formula:

$$dis = \sqrt{\sum_{i=0}^{n} (x_{1i} \quad x_{2i})^2 + (y_{1i} \quad y_{2i})^2 + \cdots}$$

Sample calculation using Table 4. above with the concept of per column minus the test data in red:

$$A1 = \sqrt{(5-5)^2 + (3-5)^2 + (3-5)^2 + (5-5)^2 + (5-5)^2} = 2.83$$
$$A2 = \sqrt{(5-5)^2 + (3-5)^2 + (3-5)^2 + (5-5)^2 + (3-5)^2} = 3.48$$
$$A3 = \sqrt{(5-1)^2 + (3-5)^2 + (3-3)^2 + (5-5)^2 + (5-5)^2} = 4.47$$
$$A4 = \sqrt{(5-1)^2 + (3-5)^2 + (3-3)^2 + (5-5)^2 + (5-5)^2} = 4.47$$
$$A5 = \sqrt{(5-4)^2 + (3-3)^2 + (3-5)^2 + (5-5)^2 + (5-5)^2} = 2.24$$

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

$$A6 = \sqrt{(5-4)^2 + (3-3)^2 + (3-5)^2 + (5-5)^2 + (5-5)^2} = 2.24$$
$$A7 = \sqrt{(5-3)^2 + (3-5)^2 + (3-5)^2 + (5-5)^2 + (5-5)^2} = 3.46$$
$$A8 = \sqrt{(5-3)^2 + (3-5)^2 + (3-5)^2 + (5-5)^2 + (3-5)^2} = 2.83$$
$$A9 = \sqrt{(5-4)^2 + (3-3)^2 + (3-5)^2 + (5-5)^2 + (5-5)^2} = 2.24$$
$$A10 = \sqrt{(5-3)^2 + (3-5)^2 + (3-5)^2 + (5-5)^2 + (5-5)^2} = 3.46$$

3. Third, sort the distances formed (in ascending order) and determine the nearest neighbors to K.

Table. 6. KNN Process Data

No	Code	IPK	UK	KP	KPA	SM	Euclidean distance	Distance Order (Small to Large)	Does it include nearest neighbor (k=3)?
1	A1	5	5	5	5	5	2.83	4	Tidak
2	A2	5	5	5	5	3	3.46	8	Tidak
3	A3	1	5	3	5	5	4.47	9	Tidak
4	A4	1	5	3	5	5	4.47	10	Tidak
5	A5	4	3	5	5	5	2.24	1	Ya
6	A6	4	3	5	5	5	2.24	2	Ya
7	A7	3	5	5	5	5	3.46	6	Tidak
8	A8	3	3	5	5	5	2.83	5	Tidak
9	A9	4	3	5	5	5	2.24	3	Ya
10	A10	3	5	5	5	5	3.46	7	Tidak

4. Determine the category of nearest neighbors. With a K value <=3, it is found that rows 5,6,9 in the table are included in the Yes category (K<=3) and the rest are not.

Table 7. Data Results 3 Closest Distance KNN method

No	Code	IPK	UK	KP	KPA	SM	Euclidean distance	Distance Order (Small to Large)	Does it include nearest neighbor (k=3)?
1	A1	5	5	5	5	5	2.83	4	No
2	A2	5	5	5	5	3	3.46	8	No
3	A3	1	5	3	5	5	4.47	9	No
4	A4	1	5	3	5	5	4.47	10	No
5	A5	4	3	5	5	5	2.24	1	Yes
6	A6	4	3	5	5	5	2.24	2	Yes
7	A7	3	5	5	5	5	3.46	6	No
8	A8	3	3	5	5	5	2.83	5	No
9	A9	4	3	5	5	5	2.24	3	Yes
10	A10	3	5	5	5	5	3.46	7	No

Look for a number of K data with the closest distance, then determine the class of the new data. The value of K is 3, so we need to take 3 data with the closest distance to the new data. The closest distances are the 5th data, 6th data, and 9th data.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

5. Data results of students who will register in the even semester of TA. 2023/2024 with Last GPA, Tuition fee, Lecture Attendance, Lack of understanding of the application, student status is as in the table below.

Table 8. Final Calculation Results

No	Code	IPK	UK	KP	KPA	SM	Euclidean distance	Distance Order (Small to Large)	Does it include nearest neighbor (k=3)?	Prediction
1	A1	5	5	5	5	5	2.83	4	No	Her-Registrasi
2	A2	5	5	5	5	3	3.46	8	No	Her-Registrasi
3	A3	1	5	3	5	5	4.47	9	No	Her-Registrasi
4	A4	1	5	3	5	5	4.47	10	No	Her-Registrasi
5	A5	4	3	5	5	5	2.24	1	Yes	Not Her-Reg
6	A6	4	3	5	5	5	2.24	2	Yes	Not Her-Reg
7	A7	3	5	5	5	5	3.46	6	No	Her-Registrasi
8	A8	3	3	5	5	5	2.83	5	No	Her-Registrasi
9	A9	4	3	5	5	5	2.24	3	Yes	Not Her-Reg
10	A10	3	5	5	5	5	3.46	7	No	Her-Registrasi

CONCLUSION

From the research process that has been carried out, our research concludes that the application of data mining with the K-Nearest Neighbors algorithm can solve the problem of student enrollment prediction analysis. Of the 10 students who predicted their registration, 7 students would register and 3 students would not register.

REFERENCE

[1] D. J. P. ismail Husein, “Decision Support System for Determining,” vol. 1, no. 1, hal. 11–21, 2017.

[2] J. Infokum, “Election Of The Head Of The Study Program By Applying The SAW Method (Case Study STMIK Pelita Nusantara),” vol. 9, no. 1, hal. 91–97, 2020.

[3] Y. Yahya dan W. Puspita Hidayanti, “Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada ‘Lombok Vape On,’” *Infotek J. Inform. dan Teknol.*, vol. 3, no. 2, hal. 104–114, 2020, doi: 10.29408/jit.v3i2.2279.

[4] D. A. M. Reza, A. M. Siregar, dan Rahmat, “Penerapan Algoritma K-Nearest Neighbord Untuk Prediksi Kematian Akibat Penyakit Gagal Jantung,” *Sci. Student J. Information, Technol. Sci.* , vol. III, no. 1, hal. 105–112, 2022.

[5] H. Said, N. H. Matondang, dan H. N. Irmanda, “Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi,” *Techno.Com*, vol. 21, no. 2, hal. 256–267, 2022, doi: 10.33633/tc.v21i2.5901.

[6] L. A. R. Hakim, A. A. Rizal, dan D. Ratnasari, “Aplikasi Prediksi Kelulusan Mahasiswa Berbasis K-Nearest Neighbor (K-NN),” *JTIM J. Teknol. Inf. dan Multimed.*, vol. 1, no. 1,

hal. 30–36, 2019, doi: 10.35746/jtim.v1i1.11.

[7] A. Pratama, F. Ali Ma, dan A. Rizki Rinaldi, "Klasifikasi Penerima Beasiswa Dengan Menggunakan Algoritma K Nearest Neighbor," 2021.

[8] M. N. Maskuri, Harliana, K. Sukerti, dan R. M. H. Bhakti, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 4, no. 1, hal. 130–140, 2022.

[9] B. Hardiyanto dan F. Rozi, "Prediksi Penjualan Sepatu Menggunakan Metode K-Nearest Neighbor," *JOEICT(Jurnal Educ. Inf. Commun. Technol.*, vol. 04, no. 02, hal. 13–18, 2020.

[10] N. Nursobah, S. Lailiyah, B. Harpad, dan M. Fahmi, "Penerapan Data Mining Untuk Prediksi Perkiraan Hujan dengan Menggunakan Algoritma K-Nearest Neighbor," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, 2022, doi: 10.47065/bits.v4i3.2564.