Temporal Prediction on Students’ Graduation using Naïve Bayes and K-Nearest Neighbor Algorithm

Ahmad Marzuqi*, Kusuma Ayu Laksitowening, Ibnu Asror

Fakultas Informatika, Program Studi Informatika, Universitas Telkom, Bandung, Indonesia
Email: *ahmadmarzuqi@students.telkomuniversity.ac.id, ayu@telkomuniversity.ac.id, *asror@telkomuniversity.ac.id

Abstract—Akreditasi merupakan bentuk penilaian kelayakan dan mutu perguruan tinggi. Salah satu faktor penilaian akreditasi adalah persentase kelulusan tepat waktu. Jumlah kelulusan tepat waktu yang rendah dapat mempengaruhi penilaian akreditasi program studi. Prediksi kelulusan mahasiswa dapat menjadi solusi untuk masalah tersebut. Hasil prediksi dapat menunjukkan mahasiswa yang beresiko tidak lulus tepat waktu. Prediksi secara temporal memungkinkan mahasiswa maupun program studi melakukan penanganan yang diperlukan lebih dini. Prediksi kelulusan dapat menggunakan metode learning analytics, dengan menggunakan kombinasi algoritma naïve bayes dan k−nearest neighbor. Algoritma naïve bayes mencari mata kuliah yang paling mempengaruhi kelulusan. Algoritma k−nearest neighbor sebagai metode klasifikasi dengan batas atribut yang dipakai berjumlah 40% dari total atribut. Dataset yang digunakan adalah data mahasiswa S1 Teknik Informatika Universitas Telkom sebanyak empat angkatan dengan melibatkan data nilai tingkat 1, tingkat 2, tingkat 3, dan tingkat 4. Hasil yang didapatkan dari penelitian ini berupa 5 atribut yang paling berpengaruh terhadap kelulusan. Serta presentasi hasil algoritma kombinasi algoritma naïve bayes dan k−nearest neighbor. Dengan hasil persentase terbesar pada tingkat 1 75.40%, tingkat 2 82.08%, tingkat 3 81.91%, dan tingkat 4 90.42%.

Kata Kunci: Klasifikasi; Kelulusan; Mahasiswa; K-Nearest Neighbor; Naïve Bayes; Learning Analytics

Abstract—Accreditation is a form of assessment of the feasibility and quality of higher education. One of the accreditation assessment factors is the percentage of graduation on time. A low percentage of on-time graduations can affect the assessment of accreditation of study programs. Predicting student graduation can be a solution to this problem. The prediction results can show that students are at risk of not graduating on time. Temporal prediction allows students and study programs to do the necessary treatment early. Prediction of graduation can use the learning analytics method, using a combination of the naïve bayes and the k-nearest neighbor algorithm. The Naïve Bayes algorithm looks for the courses that most influence graduation. The k-nearest neighbor algorithm as a classification method with the attribute limit used is 40% of the total attributes so that the algorithm becomes more effective and efficient. The dataset used is four batches of Telkom University Informatics Engineering student data involving data index of course scores 1, level 2, level 3, and level 4 data. The results obtained from this study are 5 attributes that most influence student graduation. As well as the results of the presentation of the combination naïve bayes and k-nearest neighbor algorithm with the largest percentage yield at level 1 75.40%, level 2 82.08%, level 3 81.91%, and level 4 90.42%.

Keywords: Classification; Graduation; Student; K-Nearest Neighbor; Naïve Bayes; Learning Analytics

1. INTRODUCTION

Accreditation is a form of evaluation or assessment of the quality and feasibility of a higher education institution or study program conducted by an organization or the National Accreditation Board for Higher Education (BAN-PT) [1]. Higher education institutions that have better accreditation will be an option for prospective students. One of the National Accreditation Board for Higher Education accreditation assessments is the percentage of on-time graduation for each program [1]. Therefore, the passing rate is vital for accreditation assessments.

The percentage of graduation students on-time in college can be predicted with learning analytics. Learning analytics is a way to process data obtained from learning [2]. Learning analytics are starting to be used in higher education to improve the quality of education. It can provide benefits, namely increasing graduation, curriculum development, improving lecturer performance, time of admission after graduation, and increasing research in the field of education. One method that is often used in analytic learning is classification.

This research combined the naïve Bayes and the k-nearest neighbor algorithm. Naïve bayes is used because it has the advantage of fast and simple calculations [3]. Naïve bayes also have a weakness the probability that the attribute cannot measure a prediction's accuracy [3]. The k-nearest neighbor algorithm is used because it can process large data but has the disadvantage of requiring high computation time [4]. To overcome the weaknesses and optimize the strengths of both algorithms, this study will combine the Naïve Bayes and the K-Nearest Neighbor algorithm. Combined method is done by selecting the most influential attributes with naïve bayes. Then the k-nearest neighbor algorithm includes only the selected attributes to build the classification model.

On this reseach the prediction is made temporally; it is done annually. Temporal prediction aims to be able to early predict either a student will graduate on time or not. So that it can take precautions against students who are at risk of not graduating on time or drop out. The dataset used for prediction is the grade index of students in Informatics Engineering Degree, Telkom University, class of 2008-2011. Using the attributes of courses and GPA, the classification will use the naïve bayes algorithm and the k-nearest neighbor.

Previous research [5] examined student graduation prediction using decision tree algorithms and artificial neural networks. On study, the decision tree algorithm's accuracy results are compared with the artificial neural network's results. The results showed that the artificial neural network's accuracy results were better than the decision tree algorithm.
network. It was found that the classification with artificial neural network has an accuracy of 79.74% and the decision tree has an accuracy of 74.51%. The conclusion in this study is that artificial neural network has a higher level of accuracy when compared to the decision tree method.

Another research [6] performed a combination of the naïve bayes algorithm and the k-nearest neighbor. The combined method on the study found the probability of the data attribute with naïve bayes and then classified the selected data using k-nearest neighbor. The study obtained better accuracy results and faster system running time when combining the nearest neighbor with naïve bayes compared to the algorithm without a combination.

Another approached has been discussed on other research [7]. It selected data attributes with bernoulli and multinomial naïve bayes. The study compared attribute selection with the recursive feature elimination (RFE) algorithm, 11-regularized support vector machine (SVM), SVM with RFE, LASSO. The selection of features with bernoulli naive bayes will be implemented in this study.

2. RESEARCH METHODOLOGY

2.1 System Overview

This research will create a system that combines the naïve bayes algorithm with the k-nearest neighbor to predict student graduation temporally. To run this system, the dataset needs to go through data preprocessing so that the data can be entered into the model that has been built. Table 1 lists the dataset used on our research.

| Class of Students | 2008       | 2009       | 2010       | 2011       |
|-------------------|------------|------------|------------|------------|
| Year              | 1st       | 2nd       | 3rd       | 4th       |
| Number of Students| 405 Data  | 369 Data  | 196 Data  | 284 Data  |
| 2008              | 16        | 16        | 16        | 16        |
| 2009              | 27        | 27        | 27        | 27        |
| 2010              | 34        | 34        | 34        | 34        |
| 2011              | 39        | 39        | 39        | 39        |

The table above is a description of the dataset used. The dataset consists of 4 different class, every class has a different number of students. Each class has 4 datasets from 1st year, 2nd year, 3rd year, 4th year that contain attributes (courses and GPA).

The flowchart above is the process of the whole system made in this research. Starting from the collection and data input of Telkom University S1 Informatics Engineering student batch 2008-2011. Then process the preprocessing data with Pentaho Data Integration and Microsoft Excel. After that, run the naïve bayes algorithm,
select the attributes of the courses that have been run naïve bayes with a limit of 40% of the total attributes. Finally, do the classification using the k-nearest neighbor algorithm and get the percentage of student graduation results.

2.2 Learning Analytics

Learning Analytics is the use of intelligent data, learned information, and analysis models to predict people’s learning and get ideas, and explore information and social relationships [8]. Learning Analytics focuses on education-related data such as student interactions with course material, grades, and lecturers with the aim of improving the quality of education [9]. This technique main goal is to improve learning outcomes and reduce attrition, especially among at-risk students. Learning analytics helps educators to ‘zoom in’ on individual students who need additional assistance.

2.2 Classification

Classification is a method for finding a model that explains and sort data classes by dividing the data into training data and test data. Classification is useful for predicting a class from a dataset that does not yet have a class [10]. There are many examples of classification algorithms: artificial neural networks, decision trees, naïve bayes classifiers, statistical analysis, rule-based methods, genetic algorithms, memory-based reasoning, rough sets, k-nearest neighbor, and support vector machines (SVM).

Classification is included in supervised learning because the class label in each tuple has been provided. The algorithm that implements the classification method is called a classifier. The algorithm used in this study uses the naïve bayes algorithm and k-nearest neighbor.

2.3 K-Nearest Neighbor

K-nearest neighbor is a classification method for new data based on the closest distance to existing data. At initiation, it will determine K, and this K is the number of points closest to the data points being tested. Finding the best K value is a process that can improve the classification model. Generally, a higher K value reduces the effect of noise but makes the boundaries between each classification even further [11]. The sensitivity of this algorithm is high because it is very susceptible to noise in training data. There are many formulas for calculating distance, namely types of formulas, namely euclidean, mahalohins distance, cosine, City block, chebychev, correlation (correction), hamming, jaccard, minkowski, seuclidean, and spearman [12]. The k-nearest neighbor algorithm usually uses the euclidean distance formula in calculations for test data and training data. The following is the euclidean distance formula:

$$\text{dist}(x, y) = \sqrt{\sum_{i=1}^{n} (x^i - y^i)^2}$$

(1)

2.4 Naïve Bayes

Naïve bayes is a simple classification algorithm that calculates probability based on the number of occurrences and the combination of values from an existing dataset. This algorithm strongly assumes the independence of attributes or not depending on the value of the class variable. Naïve bayes is a classification algorithm with probability science and statistics derived from the Bayes theorem. The naïve bayes algorithm has a very minimum error rate [13] and is known for its simple, fast, and highly accurate calculations [14]. The equation of the Bayes theorem is [15]:

$$P(H | X) = \frac{P(X|H)P(H)}{P(X)}$$

(2)

3. RESULTS AND DISCUSSION

3.1 Results of Testing the Most Influential Attributes with Naïve Bayes

The results of this test were obtained from the students’ final data entered into the naïve Bayes algorithm to select attributes. The selection of this attribute uses a bernoulli naïve bayes algorithm, so the data is first converted into binary form (0 & 1) by using the median dataset used. These results were obtained by using the 2008 student dataset level 1 to 4; this dataset was chosen because it has more student data than other datasets. These test results are in the form of subject attributes that have the most influence on student graduation. Here are the five attributes that most influence graduation.

| No | Attribute                           |
|----|------------------------------------|
| 1  | Probability and Statistics          |
| 2  | Object Oriented Engineering         |
| 3  | Software Engineering (RPL)          |

Ahmad Marzuqi, Copyright ©2021, MIB, Page 684
Submitted: 11/03/2021; Accepted: 22/04/2021; Published: 25/04/2021
Based on the results of testing the table above to find the attributes that most influence student graduation. There are five course attributes that most influence student graduation. The most influence graduation attributes are Probability and Statistics, Object Oriented Engineering, Software Engineering (RPL), Computer Architecture Organization, Algorithm Analysis Design. Probability and Statistics courses significantly affect graduation because these courses become one of the most difficult courses.

3.2 Attribute Selection Result Testing on Accuracy

This test compares the accuracy obtained between the algorithm that passes the attribute selection with the algorithm without passing the attribute selection. This test displays the k-nearest neighbor algorithm classification percentage results that pass attribute selection with an attribute limit of 40% and the k-nearest neighbor algorithm without passing attribute selection. The following is a table of comparison test results obtained.

Table 2. Comparison of the Classification Results of Attribute Selection with KNN

| Dataset     | K  | K-Nearest Neighbor + Naïve Bayes (40% Attribute) | K-Nearest Neighbor Without Attribute Selection |
|-------------|----|-------------------------------------------------|-----------------------------------------------|
|             |    | Accuracy                                        | Accuracy                                     |
| Class of    |    |                                                 |                                               |
| 2008        | I  | Level 1                                         | 23                                            |
|             |    |                                                 | 6                                             |
|             |    |                                                 | 70.89%                                        |
|             |    |                                                 | 16                                            |
|             |    |                                                 | 67.91%                                        |
|             | II | Level 2                                         | 19                                            |
|             |    |                                                 | 10                                            |
|             |    |                                                 | 81.34%                                        |
|             |    |                                                 | 27                                            |
|             |    |                                                 | 82.08%                                        |
|             | III| Level 3                                         | 17                                            |
|             |    |                                                 | 13                                            |
|             |    |                                                 | 81.34%                                        |
|             |    |                                                 | 34                                            |
|             |    |                                                 | 79.85%                                        |
|             | IV | Level 4                                         | 21                                            |
|             |    |                                                 | 15                                            |
|             |    |                                                 | 88.05%                                        |
|             |    |                                                 | 39                                            |
|             |    |                                                 | 92.53%                                        |
| Class of    | V  | Level 1                                         | 20                                            |
| 2009        |    |                                                 | 6                                             |
|             |    |                                                 | 75.40%                                        |
|             |    |                                                 | 16                                            |
|             |    |                                                 | 72.95%                                        |
|             | VI | Level 2                                         | 2                                             |
|             |    |                                                 | 10                                            |
|             |    |                                                 | 80.32%                                        |
|             |    |                                                 | 27                                            |
|             |    |                                                 | 75.40%                                        |
|             | VII| Level 3                                         | 22                                            |
|             |    |                                                 | 13                                            |
|             |    |                                                 | 80.32%                                        |
|             |    |                                                 | 34                                            |
|             |    |                                                 | 81.14%                                        |
|             | VIII| Level 4                                        | 5                                             |
|             |     |                                                 | 15                                            |
|             |     |                                                 | 82.78%                                        |
|             |     |                                                 | 39                                            |
|             |     |                                                 | 82.78%                                        |
| Class of    | IX | Level 1                                         | 19                                            |
| 2010        |     |                                                 | 6                                             |
|             |     |                                                 | 73.84%                                        |
|             |     |                                                 | 16                                            |
|             |     |                                                 | 58.46%                                        |
|             | X  | Level 2                                         | 10                                            |
|             |     |                                                 | 10                                            |
|             |     |                                                 | 76.92%                                        |
|             |     |                                                 | 27                                            |
|             |     |                                                 | 70.76%                                        |
|             | XI | Level 3                                         | 12                                            |
|             |     |                                                 | 13                                            |
|             |     |                                                 | 76.92%                                        |
|             |     |                                                 | 34                                            |
|             |     |                                                 | 72.30%                                        |
|             | XII| Level 4                                         | 18                                            |
|             |     |                                                 | 15                                            |
|             |     |                                                 | 81.53%                                        |
|             |     |                                                 | 39                                            |
|             |     |                                                 | 81.53%                                        |
| Class of    | XIII| Level 1                                        | 18                                           |
| 2011        |     |                                                 | 6                                             |
|             |     |                                                 | 69.14%                                        |
|             |     |                                                 | 16                                            |
|             |     |                                                 | 61.70%                                        |
|             | XIV| Level 2                                         | 5                                             |
|             |     |                                                 | 10                                            |
|             |     |                                                 | 70.21%                                        |
|             |     |                                                 | 27                                            |
|             |     |                                                 | 73.40%                                        |
|             | XV | Level 3                                         | 11                                            |
|             |     |                                                 | 13                                            |
|             |     |                                                 | 81.91%                                        |
|             |     |                                                 | 34                                            |
|             |     |                                                 | 80.85%                                        |
|             | XVI| Level 4                                         | 24                                            |
|             |     |                                                 | 15                                            |
|             |     |                                                 | 90.42%                                        |
|             |     |                                                 | 39                                            |
|             |     |                                                 | 90.42%                                        |

Figure 2. Comparison Graph Attribute Selection Accuracy
In the graph of the test results above, the results obtained between the algorithm that passes the attribute selection with the algorithm without passing the attribute selection are not much different. Classification by selecting attributes with naïve Bayes only by using 40% of its attributes results in approximately the same accuracy as the algorithm without passing attribute selection. This model can be used as a more effective and efficient model because the KNN algorithm does not need to calculate the entire dataset in a shorter time.

4. CONCLUSION

The conclusion from the implementation and analysis of the overall combination model of the naïve bayes and the k-nearest neighbor algorithm shows that the attributes of the courses have the most influence on student graduation (Table 1). There are five course attributes that most influence student graduations are Probability and Statistics, Oriented Engineering, Software Engineering (RPL), Computer Architecture Organization, Algorithm Analysis Design. Each attribute of this course is expected to help Telkom University’s Undergraduate Informatics study program provide lessons to students to get maximum and better results in courses to increase the percentage of student graduation. The combination algorithm of naïve bayes and k-nearest neighbor is more effective and efficient than the algorithm without attribute selection. The classification percentage has increased every time it rises. From the temporal prediction model of student graduation from class 2008 to 2011, the most considerable percentage results were obtained at level 1 75.40%, level 2 82.08%, level 3 81.91%, and level 4 90.42%. With these results, it is hoped that the study program can predict student graduation early to know the possibility of graduating on time or not to take preventive action if it is possible that students who are at risk of not graduating on time. Suggestions for further research from all the implementation and analysis can consider other attributes such as elective courses, non-academic attributes, and curriculum. Create a model with a newer dataset, use a classification method other than k-nearest neighbor, and use another selection attribute method.

REFERENCES

[1] Badan Akreditasi Nasional Perguruan Tinggi, “Akreditasi Perguruan Tinggi Kreteria dan Prosedur 3.0,” pp. 1–18, 2019.
[2] M. Waluyo and L. Pramudhitastari, “Learning Analytics untuk meningkatkan kualitas pembelajaran,” Semin. Nas. Ketua Pendidik. Berkemajuan dan Mengembirakan (The Second Progress. Fun Educ. Semin., 2017.
[3] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Choli sodion, “Optimali Naive Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris,” J. Teknol. Inf. dan Ilmu Komput., 2017, doi: 10.25126/jilik.20174325.
[4] T. A. Setiawan, R. S. Wahono, and A. Syukur, “Integrasi Metode Sample Bootstrapping dan Weighted Principal Component Analysis untuk Meningkatkan Performa K Nearest Neighbor pada Dataset Besar,” J. Intell. Syst., 2015.
[5] E. Rohmawan, “Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Desicion Tree Dan Artificial Neural Network,” J. Ilm. Matrik, 2018.
[6] M. K. Sari, Ernawati, and Pranowo, “Kombinasi Metode K-Nearest Neighbor,” Semin. Nas. Teknol. Inf. dan Multimed., 2015.
[7] A. Askaari, A. D’Aspremont, and L. El Ghaoui, “Naive feature selection: Sparsity in naive bayes,” arXiv, 2019.
[8] A. D. Sarayçılınkaya, H. Karal, F. Altinay, and Z. Altinay, “Reflections on Adaptive Learning Analytics,” no. March, pp. 61–84, 2021, doi: 10.4018/978-1-7998-7103-3.ch003.
[9] E. Suhartono, “Systematic Literature Review ( SLR ) : Metode , Manfaat , Dan Tantangan Learning Analytics Dengan Metode Data Mining di Dunia Pendidikan Tinggi,” J. Ilm. INFOKAM, vol. 13, no. 1, pp. 73–86, 2017.
[10] J. Han, M. Kamber, and J. Pei, Data Mining Concept and Techniques, third edition (3rd ed.), 2012.
[11] I. A. Angreni, S. A. Adisasmita, M. I. Ramli, and S. Hamid, “PENGARUH NILAI K PADA METODE K-NEAREST NEIGHBOR (KNN) TERHADAP TINGKAT AKURASI IDENTIFIKASI KERUSAKAN JALAN,” Rekayasa Sipil, 2019, doi: 10.22441/js.rs.2018.v07.i2.01.
[12] R. Arian, A. Hariri, A. Mehridehnavi, A. Fasshihi, and F. Ghasemi, “Protein kinase inhibitors’ classification using K-Nearest neighbor algorithm,” Comput. Biol. Chem., 2020, doi: 10.1016/j.compbiolchem.2020.107269.
[13] R. Wijayatun and Y. Sulistyo, “Prediksi Rating Film Menggunakan Metode Naive Bayes,” J. Tek. Elektro Unnes, vol. 8, no. 2, pp. 60–63, 2016, doi: 10.15294/jte.v8i2.7764.
[14] A. Syarif and M. A. Muslim, “Pemanfaatan Naive Bayes Untuk Memesan Emosi Dari Kalimat Berbahasa Indonesia,” Unnes J. Math., vol. 4, no. 2, 2015, doi: 10.15294/ujm.v4i2.9706.
[15] H. Annur, “Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes,” Ilk. J. Ilm., 2018, doi: 10.33096/ilkom.v10i2.303.160-165.