ISSN: 2460 – 7223

Prediction of Graduation of Students of the Lampung School of Technology Nusantara using the K-Nearest Neighbor and Naive Bayes Algorithm

1st Febri Sugandi
Faculty of Computer Science, Informatics and Business Institute of Darmajaya
fsugandi87@gmail.com

2nd Handoyo Widi Nugroho
Faculty of Computer Science, Informatics and Business Institute of Darmajaya
Lampung, Indonesia

3rd Idris Asmuni
Departement of STT Nusantara
Lampung, Indonesia
pakidris@gmail.com

***Abstract*— Predicting student graduation is the main factor for campuses to be able to assess the performance of each study program in learning achievement in each semester. Nusantara High School of Technology (STTN) Lampung has difficulty predicting graduation, so the machine learning approach with the K-Nearest Neighbor algorithm and the Naïve Bayes algorithm is very important in predicting graduation. In this paper, we discuss the K-Nearest Neighbor and Naïve Bayes methods which in research at STTN Lampung used the Rapidminer 9.1 application, with a total data of 372 student graduations which were then processed by previous data to obtain samples to be studied, then obtained a sample of 186 graduation students from the 2017 class. and 2019, for S1 Industrial Engineering and S1 Electrical. The results showed that the approach with the Naïve Bayes algorithm had a higher prediction accuracy of 89.09% compared to the K-Nearest Neighbor algorithm which obtained an accuracy rate of 74.77%. Further research can be carried out with other methods and samples from other year classes to produce more diverse predictions.**

Keywords — Prediction of student graduation, K-Nearest Neighbor Algorith,Naive Bayes Algorithm

I. INTRODUCTION

One of the benefits of the SIAKAD (Academic Information System) application in universities is to describe the achievement index of learning outcomes for each student to assess their activeness in taking lectures in each semester. Universities through special data processing institutions such as BAAK (Administrative Academic and Student Affairs Agency) usually provide Semester Results Reports which are presented to the Higher Education Service Institute (LLDikti) every semester, requiring accurate data complete about the results of student learning assessments. This need can be developed through research for campuses for future decision making, Each campus has a vision and mission that is in accordance with the field of study program it manages and will feel very satisfied, if the student activity reflected in the Graduation Index (IP) every semester is satisfactory because the better the IP, the student learning achievement every semester will be good. [1]

Therefore, to improve the quality of graduation and increase the accreditation of the Lampung School High of Technology Nusantara, it is necessary to predict students who will graduate on time. Prediction of student graduation is one of the most important and appropriate for forming patterns that may provide useful indications on student data, amounting to big. Therefore, a method is needed to solve this problem using the Data Mining method.

The previous research [2] with the title Students performance prediction using KNN and Naïve Bayesian, present study, by comparing the three evaluation parameters (Accuracy, Recall and Precision).

Research [3]. This study discusses the predictive model of student performance using K Nearest Neighbour and Naïve Bayes as classification techniques applied to the data set for secondary General certificates, which were collected from the ministry of education in the Gaza Strip. This study uses the K Nearest Neighbor and Nave Bayes algorithms, where the Naïve Bayes algorithm has the highest accuracy of 93.17% which means a strong relationship between features that affect student performance, and will help to predict student performance for next year.

Previous research conducted a measurement of student performance on the general certificate of secondary education in the Gaza Strip using the K Nearest Neighbor and Naïve Bayes algorithms. In this study, the authors tried to use the K Nearest Neighbor and Naïve Bayes methods to measure the graduation rate of students at the Nusantara College of Technology, Lampung. Lampung School of Technology Nusantara itself has never measured the graduation rate of students. Therefore, research will be conducted to measure student graduation at the Lampung School of Technology Nusantara using the K Nearest Neighbor and Naïve Bayes algorithms in classification using variables that can be used as criteria to determine student graduation. The K Nearest Neighbor and Nave Bayes algorithm methods are used because they are classification methods that can assist in measuring the level of accuracy to describe student graduation at the Lampung School of Technology Nusantara.

II. LITERATURE RIVIEW

A. Machine Learning

According to Andreas C. Muller & Sarah Guido (2017), Machine Learning is part of the field of science related to statistics, artificial intelligence and computer science that can predict the results of analysis and can learn data.[4]. According to Widodo Budiharto (2016), the Machine Learning process is the same as data mining. Both look for patterns, but Machine Learning uses data to improve understanding of the program itself.

B. *K-Nearest Neighbour Method*

K-Nearest Neighbors Algorithm is the simplest Algorithm in Machine Learning. The way to use it is to build a machine learning model to store the training data set and make predictions for the new data point, then look for the nearest data point. (Andreas C. Muller & Sarah Guido, 2017). K-Nearest Neighbors Algorithm is a supervised learning algorithm. (Mustakim, Giantika Oktaviani, 2016). The principle of the K-Nearest Neighbors Algorithm is to use a non-parametric algorithm commonly used in classification and regression (Widodo Budiharto, 2016). The steps used in implementing the K-Nearest Neighbors Algorithm are as follows:

1. Collect data with various Algorithms.
2. Calculate the distance value (distance calculation).
3. In calculating the distance, you can use the Euclidean distance formula
4. Analyze data with various algorithms.
5. Processing training data
6. Calculating error rate.
7. Entering data and running the algorithm so that it can determine which class fits the data. (Widodo Budiharto, 2016).

Predicting students' academic performance using a modified kNN algorithm Moohanad Jawthari and Veronika Stoffov, 2021 Shows that the proposed algorithm has an accuracy of 14% better than the standard one, and is not sensitive to outliers.

Application of Data Mining Classification Method for Student Graduation Prediction Using K-Nearest Neighbor (K-NN) Algorithm by Mohammad Imron and Satia Angga Kusumah, 2018. This study aims to determine the level of accuracy that has been conveyed by the K-Nearest Neighbor (K-Nearest Neighbor) algorithm. NN) in predicting the graduation rate of students at Stmik Amikom Purwokerto. The results showed that the K-NN method produced a high accuracy of 89.04%.

Machine Learning Algorithms for Student Employability Prediction Using R G Vadivu and K.Sornalakshmi (2017), to predict job skills based on their regular performance. Using data the algorithm is applied to a data set of 250 students with 59 attributes. The results showed that the accuracy obtained after the analysis for KNN was 95.33% and for Nave Bayes was 97.67%.

C. *Naïve Bayes Method*

The Naive Bayes algorithm predicts future opportunities based on past experience, so it is known as Bayes' theorem. The main characteristic of this Nave Bayes Classifier is a very strong assumption (nave) of the independence of each condition/event. Predicting Students'Academic Performance Using Naïve BayesAbdullah Baz, Fatima Alshareef, Ebtihal Alshareef, Hosam Alhakami, Tahani Alsubait. The aim of their research is to predict students' academic performance at Umm Al-Qura University by using Naive Bayes method, one of the most known data mining classification algorithms. This classifier helps to predict the final GPA of students at early stages based on courses' grades in the first year. The classification algorithm called Naïve Bayes is employed on the dataset by using the WEKA tool. dataset is collected from Umm Al-Qura University database. This dataset consists of 138 records of students who graduated from College of Computer and information Systems in the year 2019, associated with 13 attributes including student ID, gender, eight courses' grades, GPA of both first and second semester in the first's year and the final GPA.Results achieved show that Naïve Bayes can be used for predicting students' academic performance at early stages in the first year with an accuracy of 72.46%

Students performance prediction using KNN and Naïve Bayesian Ihsan A. Abu Amra, Ashraf Yunis Maghari, 2017. This paper proposes a student performance prediction model using KNN and Nave Bayes as classification techniques applied to data sets for secondary General certificates, which were collected from the ministry of education. in the Gaza Strip. In our presented study, by comparing the three evaluation parameters (accuracy, Recall and Precision) for the two KNN and Naïve Bayes algorithms, the NaïveBayes algorithm has the highest accuracy of 93.17% which means a strong relationship between the features that affect student performance, and will help for predictions of student performance for the next year. Naïve Bayes is better than KNN, which means a strong relationship between the features that affect student performance, and it will help to predict student performance. Sometimes, KNN will be better than Naïve Bayes for other datasets and different IDEs. As future work, more classification algorithms can be applied to different educational data sets.

Text Classification for Student Data Set usingNaive Bayes Classifier and KNN ClassifierRajeswari R.P, Kavitha JulietDr.Aradhana. Theexperiment carriedout shows that Naives Bayes classifier is goodclassifier with accuracy of 66.67 than KNNclassifier with 38.89.Toemphasize on performance and accuracy of theseclassifiers using Rapid miner for Student Data Set.

III. RESEARCH METHODOLOGY

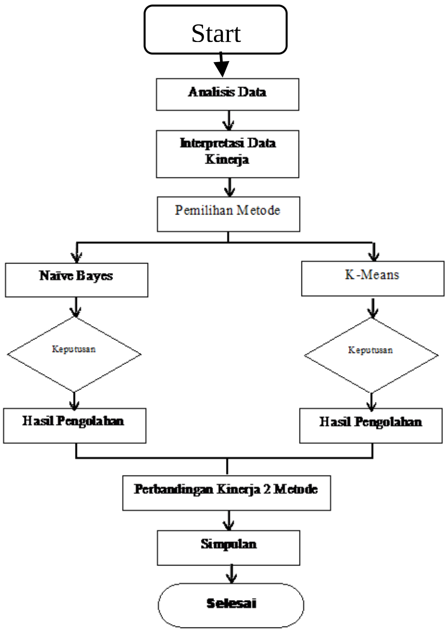


Figure 2. Research Methodology

Preprocessing is one of the important stages for data in the mining process. The data used in the mining process is not always in an ideal condition for processing. Sometimes in the data there are various problems that can interfere with the results of the mining process itself, such as missing values, redundant data, outliers, or data formats that do not match the system.

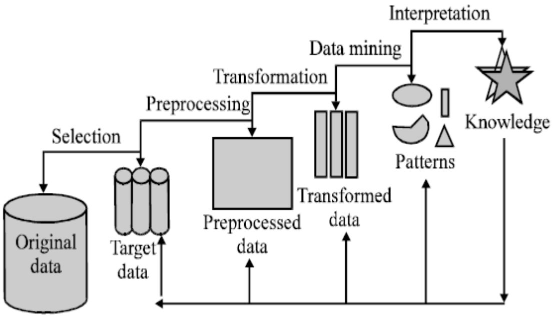


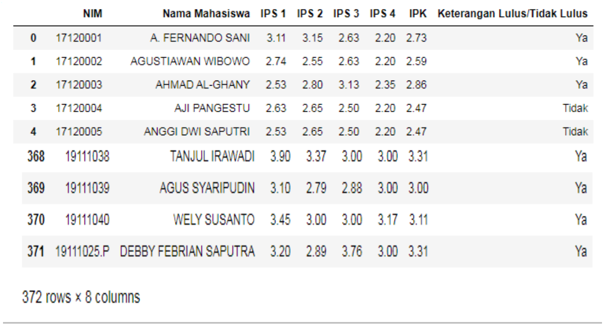
Figure 3. The position of data preprocessing in data mining

The amount of data to be processed is 186 data. In The amount of data to be processed is 186 data. In the data retrieval process on Rapidminer for the Naïve Bayes and K-Nearest Neighbor algorithms directly in csv formatfrom the transformed data. Furthermore, cross validation is carried out for the data that has been taken. K-Fold Cross validation is a statistical method that can be used to evaluate the performance of a classification model where the data is separated into two parts, namely training process data and test data. k-Fold Cross Validation is used because it can reduce computation time while maintaining the accuracy of the estimate. According to (Jiang, Ping., 2017) K-Fold Cross Validation is a type of cross validation test that serves to assess the process performance of an algorithm method by dividing data samples randomly and grouping the data as much as the K k-fold value.

IV. DISCUSSION

A. *Analysis of Data Requirements*

Before carrying out the process of calculating the algorithm, data collection is first carried out. The data used in this study is data for undergraduate students of Industrial Engineering and Electrical Engineering at the Nusantara Lampung High School of Technology class 2017 to 2019. The student data used are 372 student data consisting of 228 Industrial Engineering Bachelors and 144 Electrical Engineering Bachelors. The research taken has an input attribute of semester achievement index (IPS) 1 to 4 and a Grade Point Average (GPA) and the output attribute is Graduation. The research data obtained can be seen in table 1 as follows:



	NIM	Nama Mahasiswa	IPS 1	IPS 2	IPS 3	IPS 4	IPK	Keterangan Lulus/Tidak Lulus
0	17120001	A. FERNANDO SANI	3.11	3.15	2.63	2.20	2.73	Ya
1	17120002	AGUSTIAWAN WIBOWO	2.74	2.55	2.63	2.20	2.59	Ya
2	17120003	AHMAD AL-GHANY	2.53	2.80	3.13	2.35	2.86	Ya
3	17120004	AJI PANGESTU	2.63	2.65	2.50	2.20	2.47	Tidak
4	17120005	ANGGI DWI SAPUTRI	2.53	2.65	2.50	2.20	2.47	Tidak
368	19111038	TANJUL IRAWADI	3.90	3.37	3.00	3.00	3.31	Ya
369	19111039	AGUS SYARIPUDIN	3.10	2.79	2.88	3.00	3.00	Ya
370	19111040	WELY SUSANTO	3.45	3.00	3.00	3.17	3.11	Ya
371	19111025.P	DEBBY FEBRIAN SAPUTRA	3.20	2.89	3.76	3.00	3.31	Ya

372 rows × 8 columns

Table 1 Student Data for S1 Industrial Engineering and S1 Electrical Engineering, Nusantara Technological College Lampung, 2016 to 2019

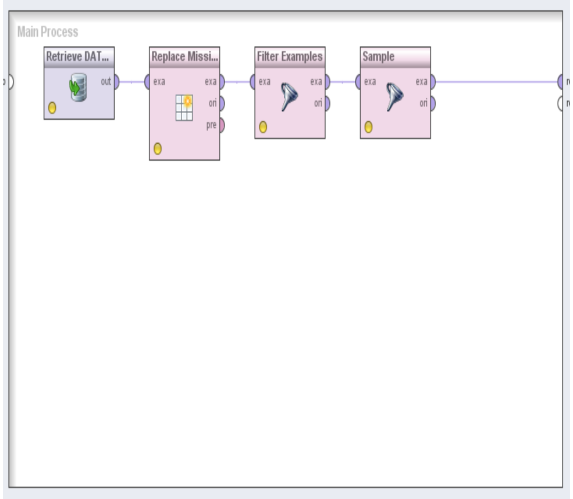


Figure 3. Data preprocessing stage using Rapid Miner 5.3

B. *Data Transformation*

After the data needs analysis process, the next step is to carry out the data transformation process. Based on the data from the 2017 to 2019 batch of the Lampung Nusantara High School of Technology, the data that will be transformed are the Grade Point Average (GPA) and a statement of pass or fail. The data is data that will be used as input attributes that will be used for analysis with the Naïve Bayes Algorithm and k-nearest neighbor. The transformation process is done by making a classification on the input attribute. The classification of input attributes can be seen in table 4.3 as follows:

Table 2. Classification of Student Graduation Predictions

Atribut	Klasifikasi
IPK	< 2,50 (Tidak) > 2,50 (Ya)

C. *Test Results*

Based on 186 student data from the 2017 to 2019 batches that have been tested, the results of the calculation of accuracy and error for each algorithm are obtained. The results of testing each algorithm are known that the performance of the Naïve Bayes Algorithm is better than the K-Nearest Neighbor Algorithm. However, classification accuracy cannot achieve perfect results in the absence of errors. This is influenced by the amount of test data and training data used from the preprocessing stage. For the nave Bayes algorithm, the accuracy reaches 96.77% which is quite good, this is because of the advantages of the Naive Bayes Algorithm itself, which is capable of classifying even though it has training data that is little for parameter estimation. Meanwhile, the K-Nearest Neighbor Algorithm produces a low 89.25% accuracy, this is because the algorithm is not effective if the amount of training data is small.

K-Nearest Neighbour Algoritma

The amount of data to be processed is 186 data. The data retrieval process in Rapidminer for the K-Nearest Neighbor Algorithm is directly in csv format. The data processing process uses the K-Nearest Neighbor Algorithm as shown below

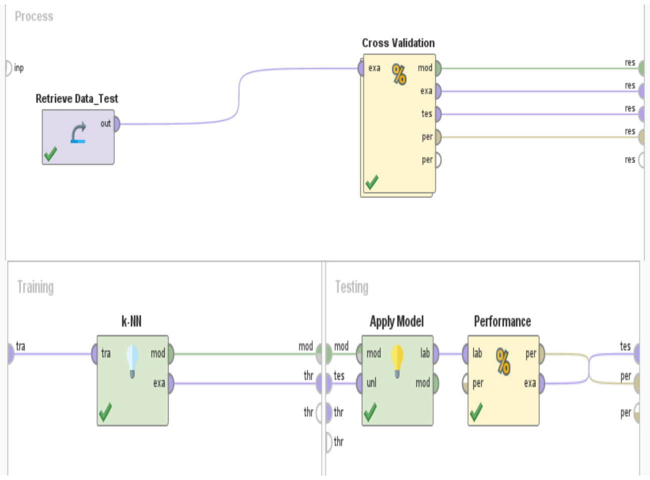


Figure 4. The K-Nearest Neighbor Algorithm data processing process

Nave Bayes Algorithm

In this data processing using the data retrieval process in Rapidminer 9.1 for the Naïve Bayes Algorithm directly in csv format. The data processing process uses the Naïve Bayes Algorithm as shown in Figure 4 below:

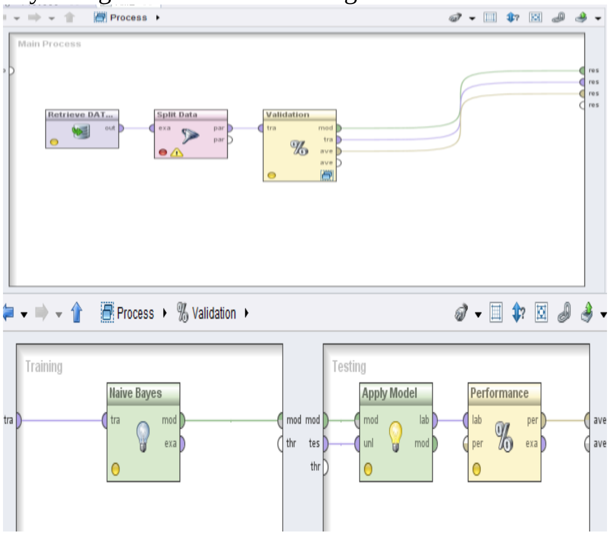


Figure 5 Nave Bayes Algorithm data processing process

V. CLOSING

A. Conclusion

At the testing, analysis and design stage of the student graduation rate prediction system using data from students of the Lampung Nusantara High School of Technology class 2017 to 2019 by comparing k-Nearest Neighbors and Naive Bayes, it can be concluded:

1. After applying the Naive Bayes and K-Nearest Neighbor methods to classify the graduation rates of 2017 to 2019 students, it is known that the performance of the Naive Bayes method is superior to the K-Nearest Neighbor method.
2. The results of the comparison of accuracy from the evaluation of the algorithm using metrics Accuracy and 10-fold Cross Validation obtained an accuracy of 89.25% for K-Nearest Neighbor while Naïve Bayes reached 96.77%., in other words the Nave Bayes Algorithm is Better than the Algorithm Naive Bayes.
3. The Naïve Bayes model with 10 fold Cross Validation achieves the highest accuracy compared to the K-Nearest Neighbor algorithm.

B. Suggestion

From the results of this study, several suggestions were obtained that could be considered for further research, including:

1. Considerations in further development, it is hoped that it can predict other things, such as the eligibility for scholarships, determining the cost of the Per-Semester Development Unit to the length of the study period that students may take.
2. The feature extraction process can remove unnecessary variables that can affect the accuracy of the training, testing process and the results obtained.
3. The student achievement prediction system can be used for the Nusantara Lampung High School of Technology to predict student graduation rates well.

REFERENCES

[1] Alpaydin, Ethem. 2010. *Introduction to Machine Learning. Second Edition. The MIT Press: Cambridge, Massachusetts.*

[2] Bonaccorso, Giuseppe. 2017. *Machine Learning Algorithm. Packt: Birmingham-Mumbai.*

[3] Budiharto, Widodo. 2016. *Machine Learning & Computational Intelligence*. Penerbit Andi: Yogyakarta.

[4] Cormen, Thomas H. dkk. 2009. *Introduction to Algorithm Third Edition*. The MIT Press: United States of America.

[5] Guido, Sarah. & Andreas C. Muller. 2017. *Introduction to Machine Learning with Python.O'Reilly Media*: United States of America.

[6] Hunt, John. 2019. *A Begineers Guide to Python 3 Programming. Springer*: United Kingdom.

[7]. Hurwitz, Judith. & Daniel Kirsch. 2018. *Machine Learning for Dummies*. IBM Limited Edition: United States of America.

[8] Igual, Laura & Santi Segui. 2017. *Introduction to Data Science*. Springer: Spanyol.

[9] Kataria, Aman. & M. D. Singh. 2013. *A Review of Data Classification Using K-Nearest*

[10] Neighbors Algorithm. International Journal of emerging Technology and Advanced Engineering, Vol. 3, Issue 6, June 2013, 1-7.

[11] Kubat, Miroslav. 2017. *An Introduction to Machiine Learning Second Edition*. Springer: University of Miami, United State of America.

[12] Muhardi & Eka Sabna, 2016. Penerapan Data Mining untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan dan Hasil Belajar. Journal CoreIT. Vol. 2, No. 2, Desember 2016.