Emerging Information Science and Technology
Vol. 1, No. 2, (2020), pp. 49-57
DOI: https://doi.org/10.18196/eist.127

# Applying the Naive Bayes Algorithm to Predict the Student Final Grade

Ronald Adrian[1], Muhammad Aldi Joko Satria Perdana[2], Asroni[2*] and Slamet Riyadi[2]

*[1]Universitas Gadjah Mada, Jl. Yacaranda, Sekip Unit IV, Yogyakarta, 55281, Indonesia*

*[2]Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia*
**Corresponding author: asroni@umy.ac.id*

## *Abstract*

*The teaching and learning process of the Faculty of Engineering of Universitas Muhammadiyah Yogyakarta has used e-learning intensively. One of the benchmarks in determining students' final grade is to take the values in e-learning. This study aims to predict students' final grades by utilizing the data mining process and the Naive Bayes algorithm. This study provides students and lecturers information to enhance the teaching and learning process to improve students' final grades and maintain satisfactory final grades until the lecture is complete. The research began with the literature study, data collection, data selection, data cleaning, data transformation and implementation with rapidminer and conclusion drawing. Based on the prediction of students' final grades, one course obtained many unsatisfactory grades with an accuracy rate of 93.75%. Thus, the higher the accuracy value, the closer the predicted final value to the actual value.*

***Keywords:*** *Naive Bayes, E-Learning, Classification, Final Grade*

## 1. Introduction

E-learning is an education system utilizing technology within the learning process. Universitas Muhammadiyah Yogyakarta (UMY), particularly the Engineering Faculty, has intensified the use of e-learning through online lectures. The web-based and intelligent tutoring systems collect large amounts of student academic data to analyze them as material for consideration in predicting students' final grades and improving the learning process effectiveness to meet the expected final grades.

However, through the e-learning process, several students obtained unsatisfactory final grades. Hence, an improvement is required in the learning process for students majoring in the Department of Informatics Engineering, Faculty of Engineering, Universitas Muhammadiyah Yogyakarta.

Academic analysis is a field using data mining and business intelligence for the benefit of the learning process. The process of analyzing e-learning data classified to take advantage of existing statistical information and find interesting patterns to be applied uses data mining techniques to study the factors affecting students' final grades. UMY has never conducted a study to predict students' final grades in using e-learning. Therefore, the writers used the data mining technique of the Naive Bayes Classifier Algorithm. We used this algorithm because it is proven to be effective in predicting academic related data [1], [2], [3], [4].

This study aims to predict students' final scores in existing subjects in the Department of Informatics Engineering, Faculty of Engineering at Universitas Muhammadiyah Yogyakarta, especially the 2015 class.

## 2. Method

This research started with a literature study, data mining, data collection, data selection, data processing, data testing, and conclusion drawing.

1. Literature Studies

   The first step in conducting this research was a literature study. It aimed to gather information by studying and reading literature about the Naive Bayes algorithm, especially to analyze the final grades of the subjects in the Department of Informatics Engineering, Faculty of Engineering, Universitas Muhammadiyah Yogyakarta.

2. Determining the Data Mining Method

   Common algorithms that are used in data mining are Naïve bayes and C4.5 Algorithm. Both algorithms has different way to make data prediction[6]. After conducting a literature study, the next stage was to determine the method in the classification technique. Naive Bayes Classifier is a classification method based on the Bayes theorem for predicting opportunities based on existing experiences using probability and statistical methods proposed by a British scientist named Thomas Bayes [5].

3. Data Collecting

   One of an important thing in this case study research type is collecting data as much as possible [7]. The data collection might affect the results of the application and conclusions if carried out improperly. The data were obtained from the Bureau of Information Systems at Universitas Muhammadiyah Yogyakarta in the form of e-learning data.

4. Data Selecting

   Not all data in the e-learning database of Universitas Muhammadiyah Yogyakarta were used for this study. The relevant data were filtered to be analyzed further. For example, the chosen data were the course table, quiz, grade quiz, and user. Furthermore, the inconsistent data and noise within the table were eliminated. In general, the data has imperfect and inconsistent contents such as invalid data, missing data and duplicates.

5. Data Processing

   The filtered data were then converted and combined into various formats suitable for data mining because the data mining method requires the appropriate data format.

6. Data Testing

   The data testing was by calculation with the Rapid Miner software accompanied by the Naive Bayes Algorithm.

7. Conclusion Drawing

   The conclusions were drawn based on the formulation of the problem and also the research objectives.

## 3. Results

The research data were collected from the e-learning database of the Engineering Faculty, Universitas Muhammadiyah Yogyakarta, divided into four departments. The data received from Information System Bureau were in the form of MySQL format. The steps of the research are as follows:

1. Data Collection

   Data collection was performed by taking several tables from the e-learning database converted into Microsoft Excel format, covering the el_course, el_quiz, el_quiz_grades, and el_user tables. The data in the el_course table has two columns of id and fullname, while el¬_quiz encompasses three columns of id, course, and fullname. The el_quiz_grades table consists of the columns of id, quiz, userid, grade, course, quiz name, and the el_user table comprises 53 columns, not all used.

2. Data Selection

   The el_course table has 31 attributes with 737 records. The attributes to be retrieved are the id and full name. The el_quiz table consists of 41 attributes and has 2,044 records, then the attributes taken are the id, course, and name. The el_quiz_grades table has five attributes and 116,506 records. Attributes taken are id, quiz, userid, grade and the addition of 2 attributes in the el_quiz table. The el_user table comprising 53 attributes has 7,737 records. The attributes to be retrieved are id, username, firstname, lastname, and department.

3. Data Cleaning

   Data cleaning is a solution for main issue in the quality of data. The problem of quality data could occur in many system fileds [8], [9], [10]. Academic field is not an exception. After collecting and selecting data according to the required attributes, data cleaning was carried out to avoid duplication of data, data with missing values, and correct data with misprints. Then, all the data were entered into the SQL Server data warehouse.

4. Data Transformation

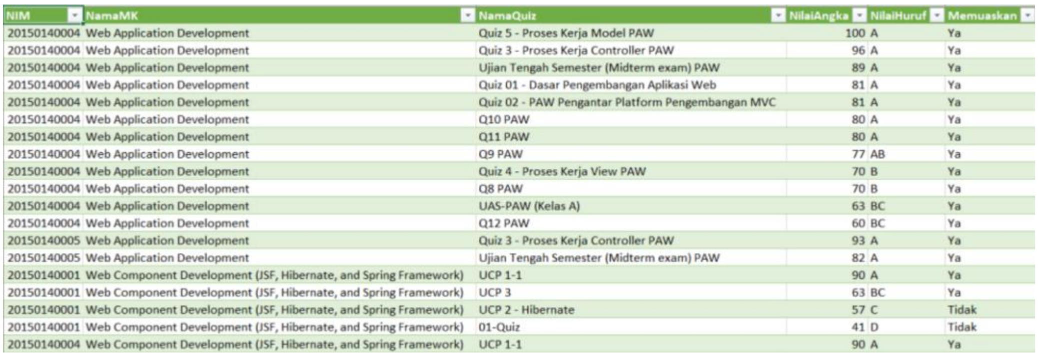
   This step focuses on transforming data into a suitable form for data mining. The data were collected from the 2015 class of the Informatics Engineering Department with 1043 records. The next step was adding letter value and satisfactory attributes based on numerical values to predict the data mining process, as presented in tables 1 and 2.

**Table 1. Scoring Categories**

| Numerical Scores | Satisfactory |
|---|---|
| 100 – 60 | Yes |
| 59 – 0 | No |

**Table 2. Letter Categories**

| Letter Scores | Numerical Scores |
|---|---|
| A | > 80 |
| AB | 75 < AB < 80 |
| B | 65 < B < 75 |
| BC | 60 < BC < 65 |
| C | 50 < C < 60 |
| D | 35 < D < 50 |
| E | E < 35 |

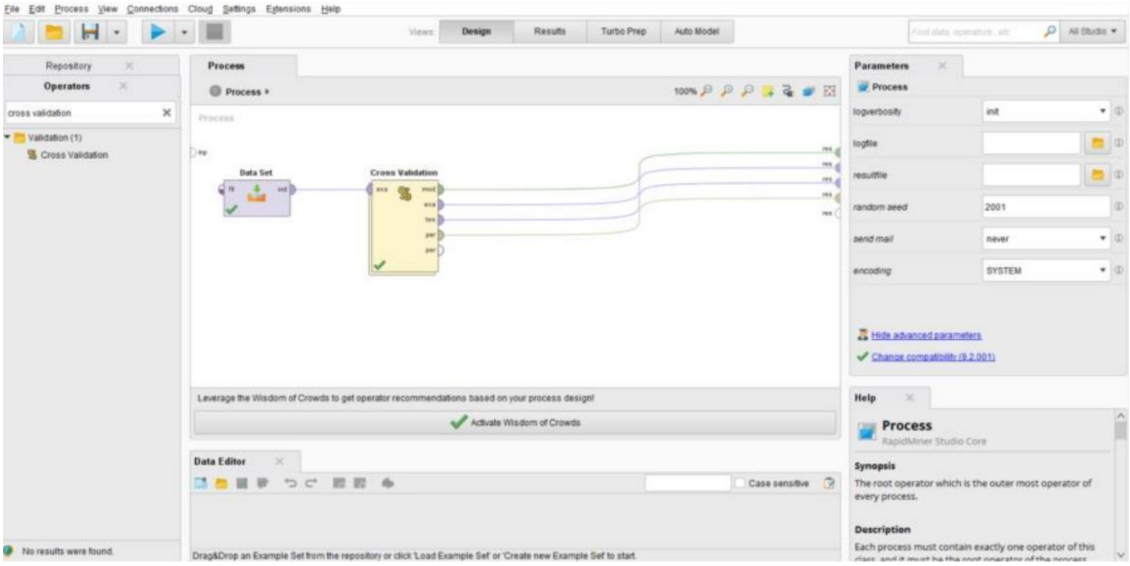


| NIM | NamaMK | NamaQuiz | NilaiAngka | NilaiHuruf | Memuaskan |
|---|---|---|---|---|---|
| 20150140004 | Web Application Development | Quiz 5 - Proses Kerja Model PAW | 100 | A | Ya |
| 20150140004 | Web Application Development | Quiz 3 - Proses Kerja Controller PAW | 96 | A | Ya |
| 20150140004 | Web Application Development | Ujian Tengah Semester (Midterm exam) PAW | 89 | A | Ya |
| 20150140004 | Web Application Development | Quiz 01 - Dasar Pengembangan Aplikasi Web | 81 | A | Ya |
| 20150140004 | Web Application Development | Quiz 02 - PAW Pengantar Platform Pengembangan MVC | 81 | A | Ya |
| 20150140004 | Web Application Development | Q10 PAW | 80 | A | Ya |
| 20150140004 | Web Application Development | Q11 PAW | 80 | A | Ya |
| 20150140004 | Web Application Development | Q9 PAW | 77 | AB | Ya |
| 20150140004 | Web Application Development | Quiz 4 - Proses Kerja View PAW | 70 | B | Ya |
| 20150140004 | Web Application Development | Q8 PAW | 70 | B | Ya |
| 20150140004 | Web Application Development | UAS-PAW (Kelas A) | 63 | BC | Ya |
| 20150140004 | Web Application Development | Q12 PAW | 60 | BC | Ya |
| 20150140005 | Web Application Development | Quiz 3 - Proses Kerja Controller PAW | 93 | A | Ya |
| 20150140005 | Web Application Development | Ujian Tengah Semester (Midterm exam) PAW | 82 | A | Ya |
| 20150140001 | Web Component Development (JSF, Hibernate, and Spring Framework) | UCP 1-1 | 90 | A | Ya |
| 20150140001 | Web Component Development (JSF, Hibernate, and Spring Framework) | UCP 3 | 63 | BC | Ya |
| 20150140001 | Web Component Development (JSF, Hibernate, and Spring Framework) | UCP 2 - Hibernate | 57 | C | Tidak |
| 20150140001 | Web Component Development (JSF, Hibernate, and Spring Framework) | 01-Quiz | 41 | D | Tidak |
| 20150140004 | Web Component Development (JSF, Hibernate, and Spring Framework) | UCP 1-1 | 90 | A | Ya |

**Figure 1. Data Transformation**

5. Implementation

   This research applied probability (prediction) with the application of the Naive Bayes algorithm. The data collected, selected and transformed, were then processed using the probability method. The records were then saved in .csv format, which could be immediately processed using the RapidMiner software with six Attribute Data of NIM, NamaMK, Quiz, NilaiAngka, Letter Value, and Satisfactory, as depicted in Figure 1. The operator cross-validation method with  5-folds and shuffled sampling was used in Figures 2 and 3. The attributes analyzed as labels were Letter Values, and as IDs were student numbers of 2015 class students majoring in the Department of Informatics Engineering. There were six subjects tested.



logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM

Hide advanced parameters
Change compatibility (9.2.001)

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

Process
RapidMiner Studio Core

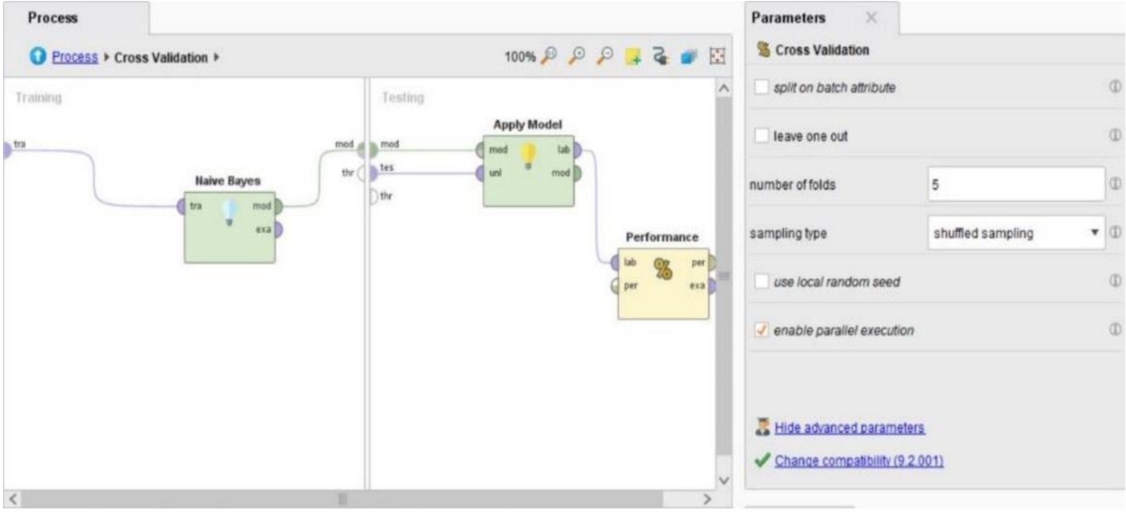
Synopsis
The root operator which is the outer most operator of every process.

Description
Each process must contain exactly one operator of this class, and it must be the root operator of the process.

Drag&Drop an Example Set from the repository or click 'Load Example Set' or 'Create new Example Set' to start.

No results were found.

**Figure 2. Cross-Validation Connection**



☐ split on batch attribute
☐ leave one out
number of folds: 5
sampling type: shuffled sampling
☐ use local random seed
☑ enable parallel execution

Hide advanced parameters
Change compatibility (9.2.001)

**Figure 3. Cross-Validation Process**

Table 3 displays the research results on six subjects in the Informatics Engineering Department with 48 data records based on the number of students in class 2015. The test results using cross-validation with 5-folds on the six subjects in Table 3 indicate that the subjects with the highest accuracy level can measure the closeness between the actual value and the predicted value. Meanwhile, the precision and recall values with high percentage results can be predicted correctly. It implies that the level of success and accuracy in predicting data following the actual and expected values is relatively high.

**Table 3. Subject Values**

| Business Intelligence System | | | Web Application Development | | | Object-Oriented Analysis Design | | |
|---|---|---|---|---|---|---|---|---|
| Accuracy | 70.83% | | Accuracy | 81.25% | | Accuracy | 93.75% | |
| Letter Value Prediction | Precision | Recall | Letter Value Prediction | Precision | Recall | Letter Value Prediction | Precision | Recall |
| AB | 66.67% | 40.00% | A | 90.00% | 81.82% | B | 91.67% | 91.67% |
| B | 66.67% | 87.50% | AB | 40.00% | 66.67% | BC | 90.00% | 90.00% |
| BC | 85.71% | 66.67% | B | 90.00% | 94.74% | C | 100.00% | 91.67% |
| C | 80.00% | 92.31% | C | 76.92% | 100.00% | D | 93.33% | 100.00% |
| **Web Content Development** | | | **Software Testing and Quality Assurance** | | | **Web Component Development** | | |
| Accuracy | 77.08% | | Accuracy | 68.75% | | Accuracy | 89.58% | |
| Letter Value Prediction | Precision | Recall | Letter Value Prediction | Precision | Recall | Letter Value Prediction | Precision | Recall |
| A | 100.00% | 80.00% | AB | 33.33% | 25.00% | A | 100.00% | 91.67% |
| AB | 50.00% | 50.00% | B | 58.82% | 83.33% | AB | 87.50% | 87.50% |
| B | 78.95% | 78.95% | BC | 75.00% | 75.00% | B | 84.21% | 100.00% |
| BC | 76.92% | 90.91% | C | 86.67% | 86.67% | BC | 100.00% | 85.71% |
| C | 75.00% | 85.71% | D | 60.00% | 60.00% | C | 75.00% | 75.00% |

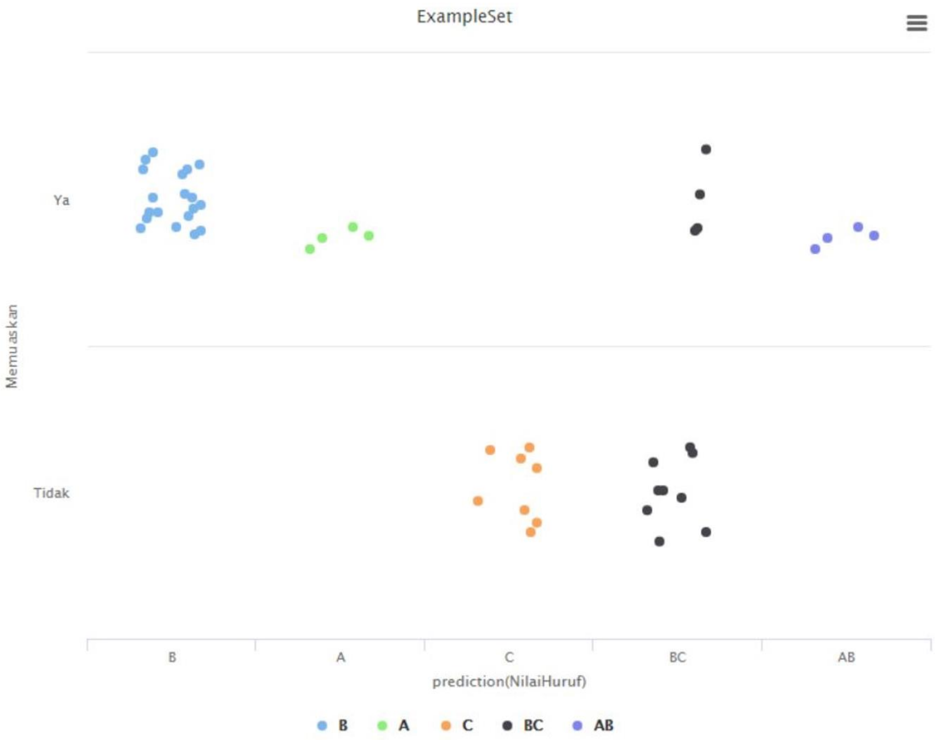


**Figure 4. Business Intelligence System Scatter Graphic**

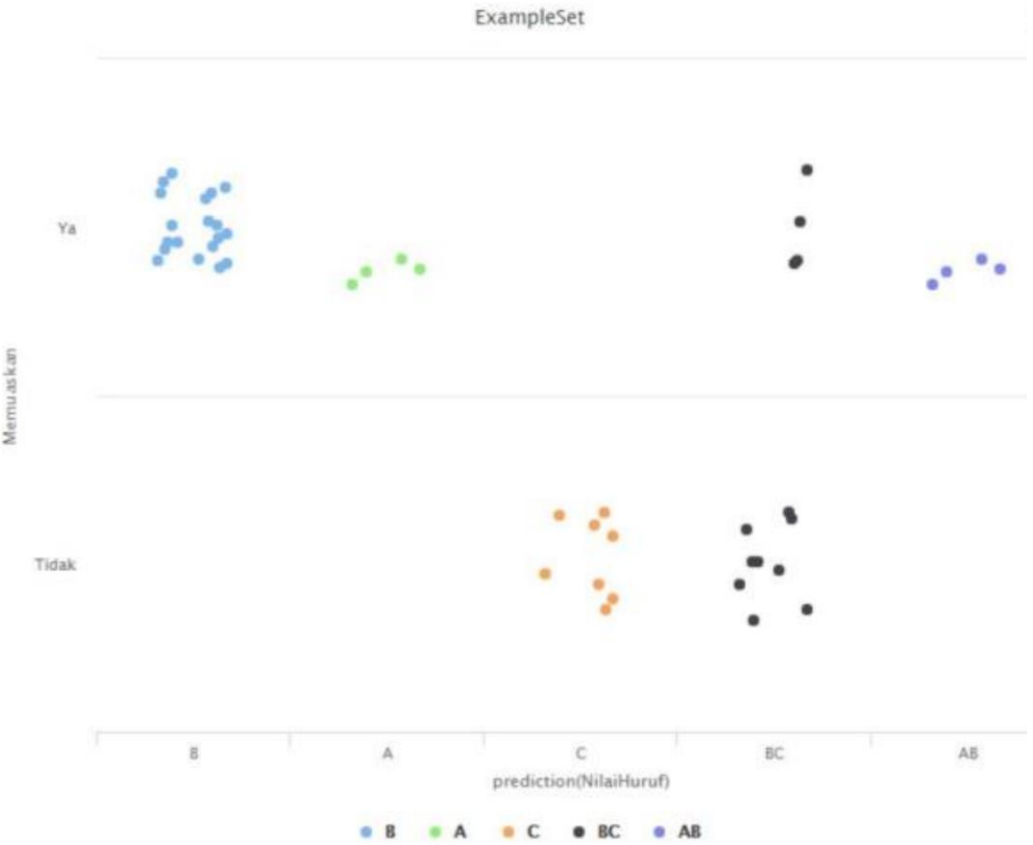


**Figure 5. Web Content Development Scatter Graphic**

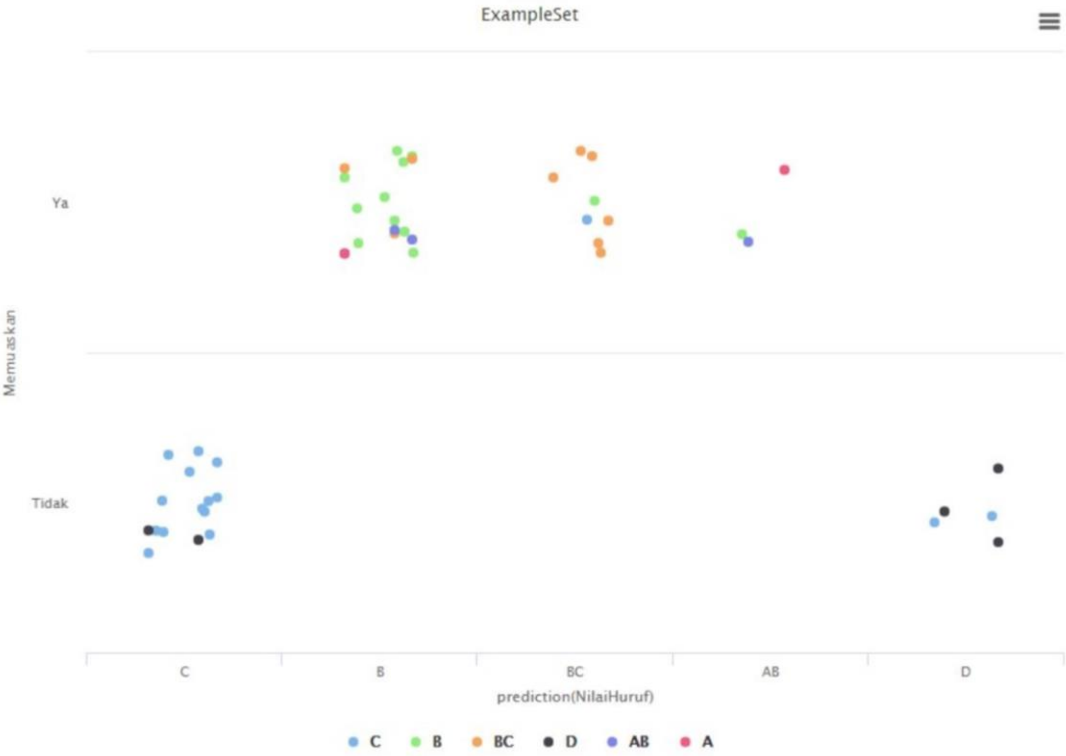


**Figure 6. Software Testing and Quality Assurance Scatter Graphic**

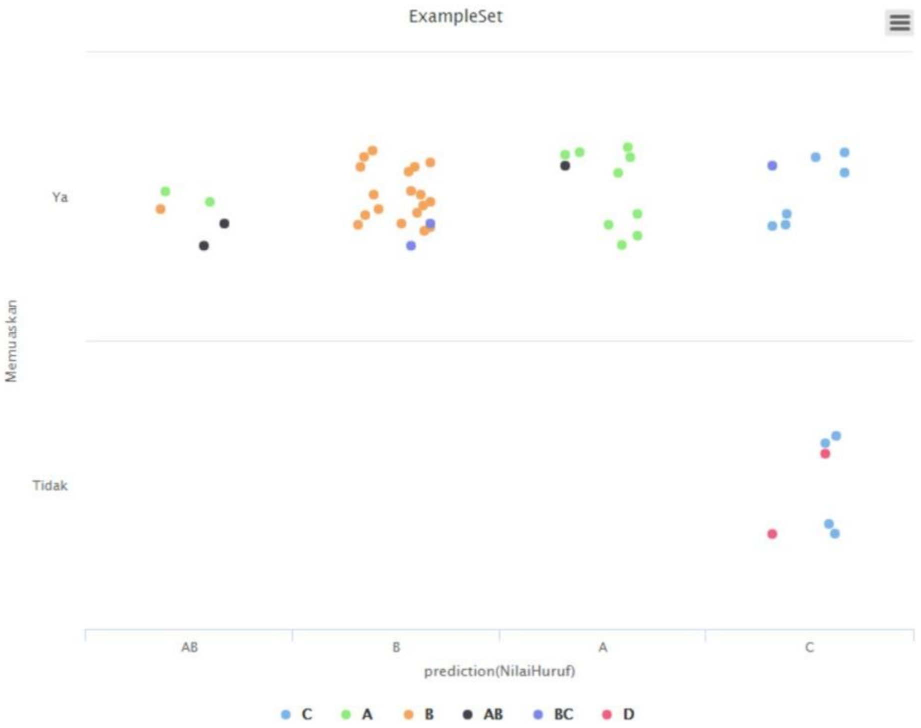


**Figure 7. Web Application Development Scatter Graphic**

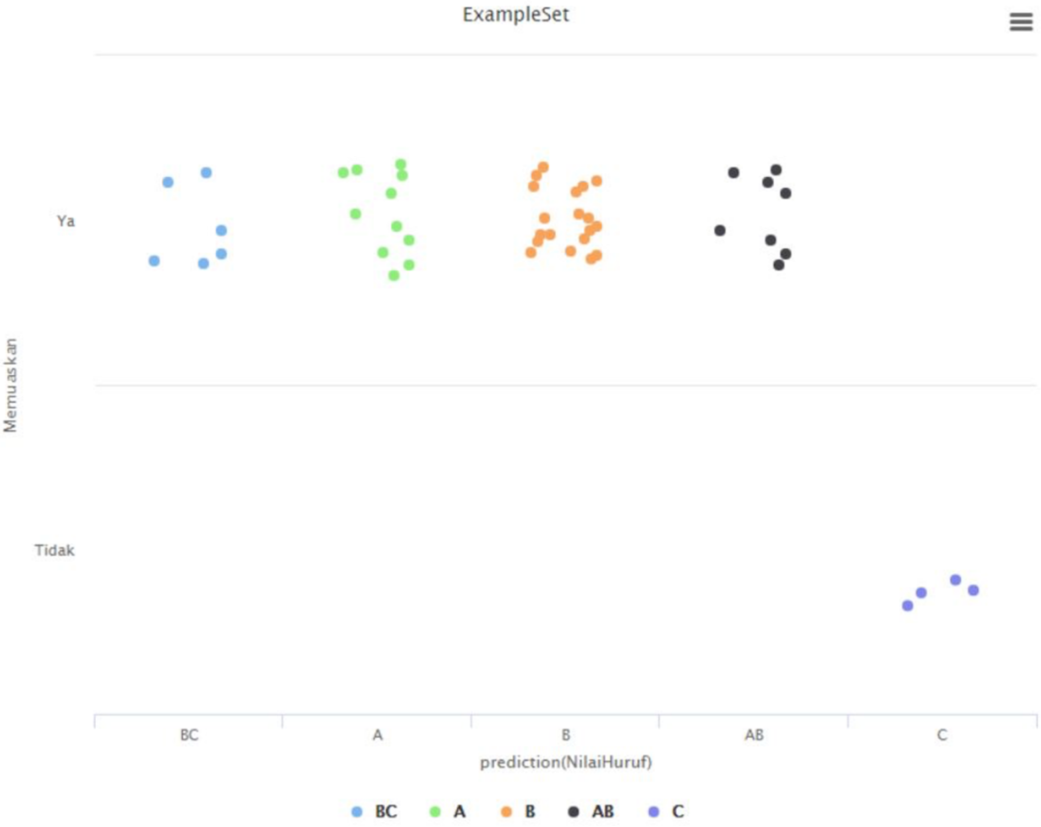


**Figure 8. Web Component Development Scatter Graphic**

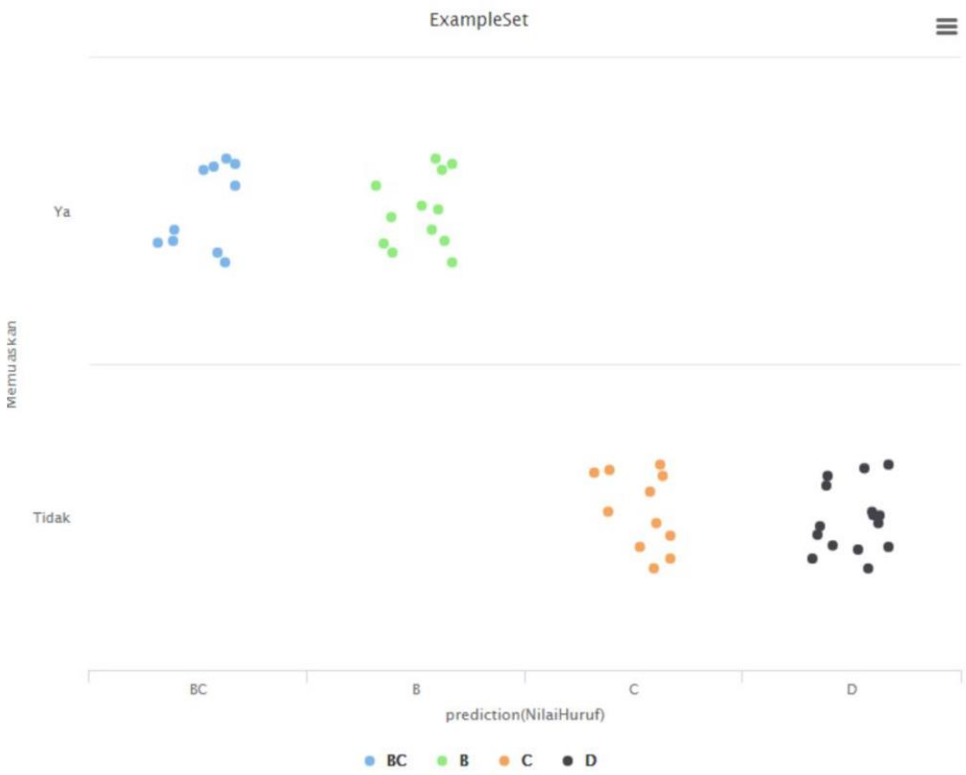


Figure 9. Object-Oriented Analysis Design Scatter Graphic

Figure 4 depicts that more students obtain satisfactory final scores, 31 against 17. Figure 5 shows that 31 students possess satisfactory final scores, while 17 others obtain unsatisfactory final scores. Figure 6 presents a thin comparison between students obtaining satisfactory and unsatisfactory final scores, 28 students versus 20 students. Figure 7 demonstrates 42 students with satisfactory final scores and six other students with unsatisfactory final scores. Figure 8 displays that more students get satisfactory final scores, as many as 44 students, while the other six students obtain unsatisfactory final scores. In Figure 9, 22 students get unsatisfactory, while 26 students have satisfactory final scores. Based on the scatter graph, the object-oriented analysis design course has unsatisfactory final grades, while other subjects produce satisfactory final grades. In a nutshell, based on Figures 4-9, an evaluation of the subjects with unsatisfactory final scores can be carried out to increase the number of students obtaining satisfactory final scores.

## 4. Conclusions

The following conclusions were drawn based on testing and analysis:

1. The Naive Bayes Algorithm can be used to predict students' final grades at Universitas Muhammadiyah Yogyakarta to improve their final grades in using e-learning.
2. The Naive Bayes algorithm in predicting students' final grades with the cross-validation and shuffled sampling method with 5-folds have the following accuracy levels:
   a. Object-Oriented Analysis Design with an accuracy of 93.75%
   b. Web Component Development with an accuracy of 89.58%
   c. Web Application Development with an accuracy of 81.25%
   d. Web Content Development with 77.08% accuracy
   e. Business Intelligence System with 70.83% accuracy
   f. Software Testing and Quality Assurance with an accuracy of 68.75%

   The higher the accuracy value, the more accurate it will be in predicting the final grades.
3. This study provides information that a subject obtained special attention because it had the highest accuracy, in which students receiving less satisfactory final scores were more than

those obtaining satisfactory final scores. Meanwhile, the other five subjects had more satisfactory final scores compared to unsatisfactory

## References

[1] Murtopo, Aang Alim. 2016. ‘Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naive Bayes [Timely Graduation Prediction of STMIK YMI Tegal Students Using the Naive Bayes Algorithm]’. CSRID (Computer Science Research and Its Development Journal) 7 (3): 145. https://doi.org/10.22303/csrid.7.3.2015.145- 154. 2013

[2] Mustafa, M. Syukri, Muh Rizky Ramadhan, and Angelina P. Thenata. ‘Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier [Implementation of Data Mining for Evaluating Student Academic Performance Using the Naive Bayes Classifier Algorithm]’. Creative Information Technology Journal 4 (2): 151. https://doi.org/10.24076/citec.2017v4i2.106. 2018

[3] Putri, Astrid Novita. ‘Penerapan Naive Bayesian Untuk Perankingan Kegiatan Di Fakultas Tik Universitas Semarang [Implementation of Naive Bayesian for Ranking Activities at the Faculty of Tik Universitas Semarang]’. Simetris : Jurnal TeknikMesin, Elektro dan Ilmu Komputer 8 (2): 603. https://doi.org/10.24176/simet.v8i2.1545. 2017

[4] Rahman, Fathur, and Muhammad Iqbal Firdaus. ‘Penerapan Data Mining Metode Naive Bayes Untuk Prediksi Hasil Belajar Siswa Sekolah Menengah Pertama (SMP) [Application of Data Mining the Naive Bayes Method to Predict Learning Outcomes of Junior High School Students (SMP)]’. Al Ulum Jurnal Sains dan Teknologi 1 (2). 2016

[5] A. D. Rachmatsyah and B. Wijaya, “Data Mining Predicts The Graduation of Students of STMIK Atma Luhur Information System Using Neive Bayes Algorithm,” Jurnal Mantik, vol. 4, no. 3, pp. 2100–2105, 2020.

[6] L. Marlina, M. Muslim, A. U. Siahaan, and P. Utama, “Data Mining Classification Comparison (Naïve Bayes and C4. 5 Algorithms),” Int. J. Eng. Trends Technol, vol. 38, no. 7, pp. 380–383, 2016.

[7] Z. Zainal, “Case study as a research method,” Jurnal kemanusiaan, vol. 5, no. 1, 2007

[8] K. Natarajan, J. Li, and A. Koronios, “Data mining techniques for data cleaning,” in Engineering Asset Lifecycle Management, Springer, 2010, pp. 796–804.

[9] E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches,” IEEE Data Eng. Bull., vol. 23, no. 4, pp. 3–13, 2000.

[10] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” Applied artificial intelligence, vol. 17, no. 5–6, pp. 375–381, 2003.