Algoritm

by Putri Nasution

FILE	NAIBAYES.PDF (818.65K)		
TIME SUBMITTED	15-JAN-2021 01:06AM (UTC-0500)	WORD COUNT	3142
SUBMISSION ID	1487918922	CHARACTER COUNT	16661

International Journal of Advanced Science and Technology
Vol. 29, No. 8, (2020), pp. 2964-2974

NAIVE BAYES ALGORITHM TO PREDICT STUDENT SUCCESS IN A COURSE

*Evi Ramadhani [1], Shedriko [2], Ismail Husein [3], Ilka Zufria [4], Hamidah Nasution [5]

[1]*Department of Statistic, Universitas Syiah Kuala, Aceh, Indonesia*
[2]*Program Studi Informatika, Universitas Indraprasta PGRI, Indonesia*
[3]*Department of Mathematics, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia*
[4]*Department of Information System, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia*
[5]*Department of Mathematics, Universitas Negeri Medan, Medan, Indonesia*

email: evi.ramadhani@unsyiah.ac.id

Abstract.

Good coordination between lecturers is needed to ensure the relatively uniform quality of education from the delivery of material by the different lecturers. By knowing the success information from several classes, it is expected that the success of other classes can be predicted. This research is using a quantitative analysis method with Naive Bayes Algorithm methodology in passing decisions on IIT (Introduction to Information Technology) subject. The result gives the pattern of training data in mean and standard deviation towards three attributes, i.e. tasks, mid-semester and final exam score, which can classify or predict graduation for new data test. The pattern of this equation can also be used for other class classification, on the same or different subjects.

***Key Words** : naive, bayes, analysis, quantitative, graduation, prediction, classification.*

1. Introduction

Higher education that has improved its status from high school to university has a fairly unique vision, in addition to being an educational provider that is useful for the intellectual life of the nation so that it can later improve the lives of students, but also behaves as a college that helps alleviate poverty, namely by providing relatively inexpensive and affordable education costs to students. Cheap tuition fees do not necessarily make the university cheap with poor quality education and not promising. On the contrary, the quality of educators is constantly being improved, namely by increasing teaching discipline, conducting research and community service, and providing scholarship programs for continuing education of educators. The students are also constantly improved in quality and discipline, starting from attendance, assignment of assignments, activeness in class, and some other rules that must be obeyed by students.

The study was conducted at the Department of Informatics of Indraprasta PGRI University which consisted of about 45 classes per class. This condition makes a course that must be supported by around 11 lecturers per semester. Good coordination between lecturers is needed to ensure the relatively uniform quality of education from the delivery of material by the different lecturers. While the courses chosen for research purposes are Introduction to Information Technology (IIT). IIT is an introductory course in the world of information and communication technology. This course gives an outline of the essence of the Informatics

ISSN: 2005-4238 IJAST

Study Program whose substance is still understandable by lay people or by relatively new students. By knowing the success information from several classes, it is expected that the success of other classes can be predicted using the Naive Bayes Algorithm. The results of this study are expected to be a comparison with the pass level of the final semester students, such as in terms of knowing the level of achievement or crafts of one generation of students from the initial semester to the end of the course. Or it can be a good comparison for the same subject in other classes or different courses in the same semester or the next so that it can be a student performance controller for academic advisors.

2. Literature review

Atul Butte from Stanford University once said, "In a mound of data stored knowledge that can change the life of a patient or even change the world" (Dean, 2014: 1). This can be proven in the United States in the era of 2014 where around 235,000 women were diagnosed with breast cancer and 40,000 were estimated to die in the same year. Research in the field of pharmacy has found that Tamoxifen is a very effective drug for healing breast cancer. A few years ago when computer technology had developed so rapidly with large amounts of data processing, research on patient diagnoses from large data pools and the classification of DNA data sequencing was known that Tamoxifen was only effective in curing cancer in about 80% of patients, whereas 20 The remaining% will not get healing because of different DNA characters (Dean, 2014: 2). With the development of computer technology, big data and data mining and machine learning have become a reliable support in information technology as an inseparable part of human life (Smola & Vishwanathan, 2008: 3). Classification of email spam, the identification of certain entities (such as faces, fingerprints, sounds, etc.), classification of diseases to predictions related to many fields is carried out by three parts of the information technology (Smola & Vishwanathan, 2008: 4-20). Broadly speaking, functions in data mining can be divided into 2 categories (Han, Kamber & Pei, 2012: 15), namely:

1. Descriptive, in charge of characterizing or classifying certain data characteristics into a set of destination data

2. Predictive, in charge of inducing or stimulating a set of data to be able to do the prediction or forecast or forecast process

The predictive category in data mining has properties that overlap with the function of machine learning. While the data processing stage in data mining is divided into 6 parts (Han, Kamber & Pei, 2012: 88-120), namely:

1. Data Cleaning is the process of filling in the values that are missing from the data

2. Data Integration, combining data from various sources

3. Data Reduction, reducing data composition to be more effective and efficient

4. Data Transformation, changing data into forms that are suitable for mining needs

5. Data Discretization, changing numerical data by mapping values into intervals or certain label concepts

Machine learning is usually divided into 2 phases (Hertzmann & Fleet, 2012: 1):

1. Training: a model is learned from a collection of training data (data for learning)

ISSN: 2005-4238 IJAST

2. Application: the existing model is used to determine the results of a new test data set

Here are some types of machine learning (Hertzmann & Fleet, 2012: 2):

1. Supervised Learning, a type of learning where training data containing the correct answers are given first as a reference

2. Unsupervised Learning, a type of learning in which the data provided must be analyzed and classified in advance according to certain patterns before being declared correct

3. Reinforcement Learning, a type of learning in which an agent (in the form of a robot or controller) looks for certain learning patterns to provide optimal action based on previous experience.

Some basic algorithms which are Supervised Learning are Naive Bayes, Nearest Neighbors Estimators, A Simple Classifier, Perceptron, and K-Means (Smola & Vishwanathan, 2008: 20-36, and Simeone, 2018: 77-112). In this study, the algorithm used is Naive Bayes.

Naive Bayes algorithms is a probability framework for solving classification problems (Tzagarakis, 2010: 79). Classification is closely related to data that can be divided into 2 types (Smola & Vishwanathan, 2008: 14), namely discrete (uniform distribution) and continue (normal distribution). Naive Bayes algorithms for uniform distribution uses the formula (Smola & Vishwanathan, 2008: 18) as follows:

$$p(y|x) = \frac{p(x|y).p(y)}{p(x)} \quad (1)$$

$p(y) = prior\ probability\ of\ hypothess$

$p(x) = data\ evidence$

$p(x|y) = likelihood, data\ probability\ if\ event\ x\ is\ true$

$p(y|x) = posterior\ probability, probability\ event\ y\ on\ data\ in\ event\ x$

Whereas the normal distribution uses the Gauss Density formula (Smola & Vishwanathan, 2008: 18) as follows:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

$x = the\ value\ of\ x$

$\mu = mean$

$\sigma = deviation\ standard$

$p(x) = probability\ of\ x$

The mean value (Mean, 2018) uses the following formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (3)$$

$\bar{x} = mean$

$x_i = i - th\ sample\ value$

$n = number\ of\ sample$

and standard deviations (Variants and Standard Deviations, 2018) using the following formula:

$$s = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}} \quad (4)$$

$s = deviation\ standard$

$x_i = i - th\ value\ of\ x$

$\bar{x} = mean$

$n = number\ of\ sample$

Research methods

This research uses a quantitative analysis method (Kothari, 2004: 2-4) based on the data obtained to then process it with the Naive Bayes algorithms methodology. Research-based on analysis is carried out using facts or information that already exists, then analyzing it to make critical evaluations related to specific problems (Kothari, 2004: 3). While quantitative research is research based on numbers or quantum calculations of all phenomena related to numerics (Kothari, 2004: 3). So it can be concluded that this study was conducted by analyzing the numbers based on a particular formula on a particular object, in this case, the prediction of success of students in the Introduction to Information Technology course.

Results and Discussion

The study was conducted on students of the Department of Information at PGRI University who took the PTI course (Introduction to Information Technology), which is a beginner's course and introduction to the world of information technology and information systems. The total student population taken is 175 people from 4 classes that took place in the odd semester of the 2019-2020 school year. The students consist of men and women including new students who write on their Study Plan Card IIT courses for the first time and repeat students who take these courses for the second or more time. PGRI University emphasizes its lecturers to teach not only hard-skills but also soft-skills to students so that the expertise gained after graduation will be more complete. Therefore activities such as activities in class and quizzes become an assessment factor included in this study.

Some assessment criteria that are generally carried out on graduation of students in a course include the overall value of attendance, assignments, midterm, final semester exams and other supporting values such as activeness in class and quizzes, with details as follows:

1. The final value must reach a value greater than or equal to 56

2. Attendance must meet at least 80 percent

3. Minimum assignment value is 60

4. Attending a quiz that is held

5. Be active in class by asking or answering questions raised in class both raised by lecturers and students in the discussion session

However, some subjective assessments can occur with the assumption that a student may be careless at the time of the exam, even though in daily teaching and learning activities in the lecturer class, the student has considerable potential in the lecture. Here are some examples that might occur in the provision of assessment in class, namely:

1. Meet the minimum total score but absenteeism is less than 80 percent, so a graduation decision is seen at a high assignment score, such as greater or equal to 90

2. Meet the minimum total value but the attendance of less than 80 percent, so those graduation decisions are seen in high activity scores, such as more than 3 times asking questions or answering, plus always attending quizzes held in class

3. And so on

Assignment grades usually include the following values:

1. Attendance value Give a value with a portion of 50 percent of the value of the assignment if students attend all lectures which usually range from about 14 meetings

2. The value of activeness in class Usually the lecturer will give a portion of up to 30 percent of the value of the assignment if the student is active to ask or answer as many as more than or equal to 3 times

3. Quiz scores Grades with a portion of up to 20 percent of the grade of the assignment will be given if the student gets a very satisfactory value from the questions given

By getting the assignment scores in this study, it is assumed that the three assessments mentioned above are well represented so that they are expected to provide output values that tend to be accurate. The attributes used as evidence in the Naive Bayes Algorithm in this study consisted of 3 items, viz.

1. Value of assignment

2. Midterm scores

3. Final semester grades

Some sightings of training data where students are declared to have failed can be seen in Table 1 below.

Table 1. Some training data that states students did not pass

No	Name	Homework	MidTest	FinalTest	Result
5	M Nursyamsi	90	50	28	Failed
6	Sapta S Y	90	45	17	Failed
7	M Ryansah	0	30	0	Failed
9	Apriadi S	100	30	10	Failed
11	M Afitra P	0	0	0	Failed
13	Liftiana N S	100	50	18	Failed
14	Maher F S	80	50	18	Failed
16	Bima N S	90	30	42	Failed
21	Angga E	0	0	0	Failed
35	Lukaslie	0	68	0	Failed
36	Adelia F	90	62	28	Failed
37	Ibnu H	80	62	9	Failed
39	Deri D	90	50	21	Failed
45	Bintang S R	15	45	81	Failed
46	Sugeng R	15	68	68	Failed
47	Rita B	15	72	82	Failed
49	Laberta R	15	90	35	Failed
50	Rayindra D	15	85	56	Failed
53	Haris F A L	0	0	0	Failed
57	Nadia A S	80	80	60	Failed

Whereas some training data sightings where students have passed can be seen in Table 2 below.

Table 2. Some training data that states students have passed

No	Name	Homework	MidTest	FinalTest	Result
155	M Salya N	100	72	71	Passed
156	Dafit W	100	62	47	Passed
157	M Irsan	90	68	42	Passed
158	Fauzan A	100	80	62	Passed
159	Ilham T	90	50	63	Passed
160	Mailan	90	85	72	Passed
161	Fadri	90	50	61	Passed
162	Fikri F	100	72	63	Passed
163	M Abie A	90	72	62	Passed
164	Ivan P D A	100	72	75	Passed
165	Fitriani	100	68	49	Passed
166	Herry Y	100	72	63	Passed
167	Tony P	90	62	58	Passed
169	Raden R R Y	100	50	42	Passed
170	Alif F	80	55	50	Passed
171	M Helmi A	80	72	63	Passed
172	Adam M A	100	40	58	Passed
173	Fikri N	80	72	59	Passed
174	Anggara D S	90	68	59	Passed
175	Rachma M	100	62	48	Passed

From the tables that display the training data both passed and not passed, we can know that the evidence is continuous data (Meilani & Susanti, 2015) so that in the calculation using the Naive Bayes Algorithm we will choose the Gauss Density formula (2), while the mean and standard deviation each uses the formulas (3) and (4). Then obtained the results as in Table 3 below:

Table 3. Calculation of the mean and standard deviation of the sample

Status	Passed	Not Passed
Assignment (Mean)	92,94964029	41,38888889
Assignment (Str Dev)	6,858110338	42,38616875
Mid Test (Mean)	74,6618705	46,13888889
Mid Test (Str Dev)	15,0874523	28,91085093
Final Test (Mean)	66,69784173	25,97222222
Final Test (Str Dev)	16,34929182	27,85624948

Then to run the algorithm in accordance with the formulation that has been obtained, the testing data is taken as input as shown in table 4 below:

Table 4. Testing data provided

Examination Status	Assignment Remark	Mid Test Remark	Final Test Remark
?	85	80	45

By using the Gauss Density formula, we can see the results of the calculation as follows below.

Passeed the exam

$p(assignment|passed) = 2{,}9719802\ x\ 10^{-2}$

$p(MidTest|passed) = 2{,}4843974\ x\ 10^{-2}$

$p(FinalTest|passed) = 1{,}011715\ x\ 10^{-2}$

The value of "posterior numerator"= product value = $7{,}47008\ x\ 10^{-6}$

Not Passed the exam

$p(assignment|not\ pass) = 5{,}545191\ x\ 10^{-3}$

$p(Mid\ Test|not\ pass) = 6{,}951582\ x\ 10^{-3}$

$p(Final\ Test|not\ pass) = 1{,}1344357\ x\ 10^{-2}$

The value of "posterior numerator" =product value = $4{,}87301\ x\ 10^{-7}$

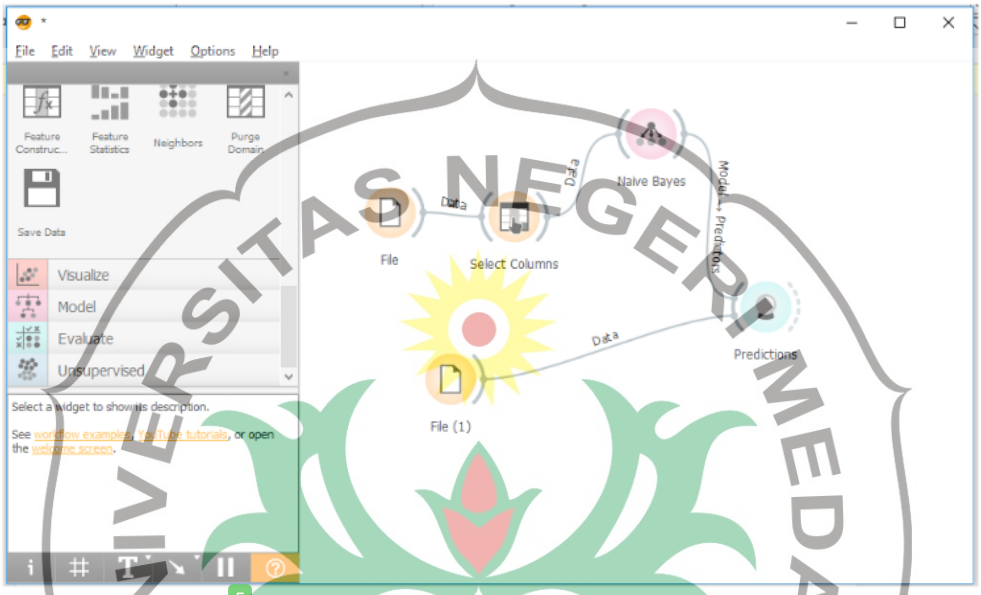
Status

Because the value of "passing" is greater than "not passing", the testing data mentioned above results or status is PASS

Testing

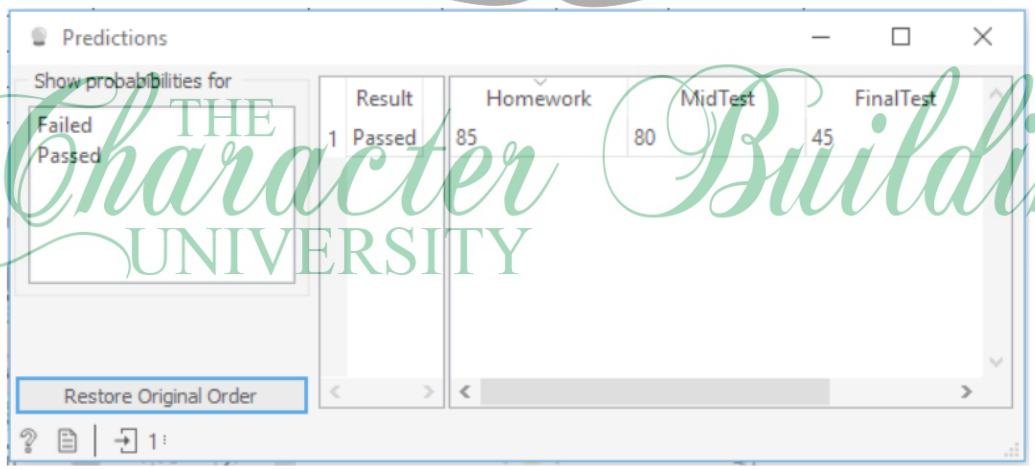
Testing is done using a software called Orange Data Mining. This software is a free data mining software that is quite powerful. Various reviews of this software show a high value compared to other free software. Figure 1 below shows the test results in the form of data models from the software below.

Figure 1. Data Model Using Orange Data Mining



While the prediction results can be seen in the following figure 2 below.

Figure 2. Prediction Results with Orange Data Mining



From testing with the same test data obtained the same calculation results with manual calculation, namely "PASS". It's just that, from a number of times the test results using the same value for Mid Test and Final Test values and different assignment values, the results obtained are somewhat different, namely from the manual calculation of graduation threshold values starting from 75 and above while calculating with the threshold value the software starts from a value of 85 upwards. This is possible because the calculation using software will be more detailed than the manual. For example in giving the value of the mathematical constant "e" which is calculated manually using a relatively limited decimal number.

Conclusion

The results of this study indicate the high value obtained in using the Naive Bayes Algorithm in predicting student graduation. This can be seen in the evaluation of tests and scores in Orange Data Mining Software where the value of accuracy, precision, and recall have given is very high. Likewise, the results of the pattern formulation of this study can be used as a basis for predicting the success of students in other subjects. This research can also be used as a reference for further research using various other supervised machine learning algorithms to compare the results obtained with the results of this study. Some terms in this research can also be a development for the focus of discussion in further research.

Reference

[1] Algoritma Naive Bayes, 2020, https://informatikalogi.com/algoritma-naive-bayes/, February 16, 2020.
[2] Dean, Jared, 2014, *Big Data, Data Mining and Machine Learning*, John Wiley & Sons Inc.
[3] Hertzmann, Aaron & Fleet, David, 2012, *Machine Learning and Data Mining Lecture Notes*, Computer Science Department, University of Toronto.
[4] Han, Jiawei & Kamber, Micheline & Pei, Jian, 2012, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Inc.
[5] Kothari, C. R., 2004, *Research Methodology, Methods & Techniques*, New Age International (P) Ltd.
[6] Mean, 2018, https://www.rumusstatistik.com/2013/07/rata-rata-mean-atau rataan.html, February 16, 2020.
[7] Meilani, Budanis D & Susanti, Nofi, 2015, *Aplikasi Data Mining Untuk Menghasilkan Pola Kelulusan Siswa Dengan Metode Naive Bayes*, Jurnal Ilmiah Nero Vol. 1 No. 3.
[8] Simeone, Osvaldo, 2018, A Brief Introduction to Machine Learning for Engineers, King's Collage London.
[9] Smola, Alex & Vishwanathan, S. V. N., 2008, *Introduction to Machine Learning*, Cambridge University Press.
[10] Tzagarakis, Manola, 2010, *Managing Big Data, Classification: Alternative Methods*, College Slide, Department of Economics, University of Patras.
[11] Varian dan Standar Deviasi, 2018, https://www.rumusstatistik.com/2013/07/varian-dan-standar-deviasi-simpangan.html, 16 Februari 2020.
[12] Syah Rahmad, M K M Nasution, Ismail Husein, Marischa Elveny, "Optimization Tree Based Inference to Customer Behaviors in Dynamic Control System", International Journal of Advanced Science and Technology, pp. 1102 – 1109,2020.
[13] Husein Ismail, Rahmad Syah, "Model of Increasing Experiences Mathematics Learning with Group Method Project", International Journal of Advanced Science and Technology, pp. 1133-1138, 2020.

[14] Syah Rahmad, Mahyuddin K.M Nasution, Ismail Husein, "Dynamic Control Financial Supervision (OJK) for Growth Customer Behavior using KYC System", International Journal of Advanced Science and Technology, pp. 1110 – 1119, 2020.

ISSN: 2005-4238 IJAST

Algoritm

ORIGINALITY REPORT

%28	%27	%8	%13
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	jurnal.uinsu.ac.id Internet Source	%13
2	Submitted to Universitas Pendidikan Ganesha Student Paper	%7
3	sersc.org Internet Source	%6
4	Dean, . "Introduction", Big Data Data Mining and Machine Learning, 2014. Publication	%1
5	es.scribd.com Internet Source	%1
6	ro.scribd.com Internet Source	<%1
7	opus-htw-aalen.bsz-bw.de Internet Source	<%1
8	healthservices.boisestate.edu Internet Source	<%1

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	OFF		

THE Character Building UNIVERSITY