IMPLEMENTATION OF NAÏVE BAYES CLASSIFICATION METHOD TO PREDICT GRADUATION TIME OF IBI DARMAJAYA SCHOLAR

Ketut Artaye[1]
Informatics Engineering IBI Darmajaya Lampung
Z.A. Pagar Alam Street No.93 Bandar Lampung
email : ketutartaye@gmail.com

ABSTRACT

Quality of a university can be seen in the average of how long a collage student take to graduate, how long for graduates to have a job also can be seen in study time in university. Every university study time will variate with different student in the university.Every major responsible for study development of scholars.They also have a task to predict study time length of every college student to decide and anticipate student who skip class that will cause a bad result in major performance.

It is important to be done, especially in Informatics and Business Institute Darmajaya to maintain quality and performance of each major. Because of that reason, writer did research with title "**Implementation of Naïve Bayes Classification Method to Predict Graduation Time of IBI Darmajaya Scholar"**to understand about average study time of each college student.

Key Words: *Naïve Bayes Classification*, Predict, Scholar Graduation.

1. INTRODUCTION

Informatics and Business Institute Darmajaya is one of leading private college institution in LAMPUNG province, founded in 1995. The credibility given by society, local government and central government (DIKTI), IBI Darmajaya has growth and developed to be a big college and have a good reputation as education institution.

To receive and to maintain that predicate is not an easy task, because of that needed a strategy to maintain quality that has been achieve throughout the process. Quality of a college can be seen from graduation time and also getting a job also can be seen from study time of the scholar.Every college will have variation of graduation time from different scholar. The scholar is one of important aspect in evaluation of college's major accomplishment. Monitoring the student's entry, progress of the student, student's achievement, the graduation ratio of he number of student passing, and competency of graduates should be getting a serious attention to receive trust from *stakeholder* in appreciate and alumnus requirement. Informatics Engineering is one of major program at IBI Darmajaya and one of favorite major choices in 2008 until 2011 with average of 250 scholars each year.However in 2012-2014 this major program going through a degradation seen from interest quantity of scholar whom took the major program. Furthermore average level of graduation is decreasing. Major Program responsible to monitoring study progress of the scholar. They also have task to predict study time for each student to decide and anticipating student from skip class which reason of bad major perform.

2. LITERATURE STUDY

1. Data Mining

Based on Kusrini (2009) *Datamining*isa term used to find hidden knowledge in database.*Datamining*i s a semi-automatics process using statistics, math, artificial intelligence, and *machinelearning*to extract and identify

potential knowledge information and become a benefit in bigger database.

Data mining is not an entirely new field. One of the difficulties of defining data mining is the fact that data mining inherited many aspects and techniques from the different fields of science that are already well established in advance. Data mining has long roots from the field of science such as artificial intelligence (artificial intelligent), machine learning, statistical, database and also information retrieval.

2. Naïve Bayes Classification

The bayes theory is the fundamentals statistic approach in the introduction of a pattern recognition. This approach based on the quantification of trade-off between classification various decisions made by the use of probability and the charges caused in those decisions.

Bayesianclassification is statistic classification that can be used to predict probability of membership of a class.Bayesianclassification based on Bayes theory that has classification ability same with decisiontreeand neuralnetwork.Bayesianclassification proved had accuracy and high velocity when applied in database with massive data(Kusrini, 2009). Bayes T h e o r y have general form :

X = Data with unknown class

H= Hypothesisd a t a fromXwith specific class

P(H|X)=Probability of H Hypothesis based on x fact (posteriori prob.)

P(H) = Probability of H Hypothesis (prior prob.)

P(X|H)=Probability ofXbased on current condition

P(X) = Probability of X

3. RESEARCH METHOD

In this research will be implemented *Naïve Bayes Classifier* Method to predict time of graduation which this research can determine exactness of study time in informatics engineering major.

System analysis presented in this paper is a whole description of the obstacles in application in Naïve Bayes Classification algorithm in deciding study time of IBI Darmajaya scholar. As for attribute used in predicting time of graduation cover:

a. Gender

   Variable gender only have two possibilities, which is male and female.In research by Purwanto (2007) in Lampung with 230 scholar samplewhom worn in different position and*grade*on plantation show that their salary depend on gender.

b. Hometown

   Hometown variable divide into Bandar Lampung City and Outer from Bandar Lampung City. Whom the hometown is in Bandar Lampungthen classified the data inn“DALAM KOTA” while the others is “LUAR KOTA”.

c. Type of Shcool

   Type of Scholl variable contain all the possibility of school typebefore university entrance. Value

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

   that determined on the software depend on the classification result, which is for senior high school classified to general and the others to vocational.

d. School Location

   School location variable classified become from inner city of Bandar Lampung or outside Bandar Lampung City.

   If the school location is in inner city of Bandar Lampung then classified the data in “DALAM KOTA”other than that classified in ”LUAR KOTA”.

e. Economic

Economicis a variable contain about family welfare. Choices in this software are divided into three parts, which is high, middle, and lower.

Table 3.1 Family Welfare

No	Penghasilan	Keterangan
1	Penghasilan <= Rp. 1.500.000/ Bulan	Rendah
2	Penghasilan >Rp.1500.000 dan < Rp.3.000.000/Bulan	Sedang
3	Penghasilan > Rp. 3.000.000/Bulan	Tinggi

f. Grade-Point Average

GPA variable is Grade-Point Average of every semester already taken by student. The size of the GPA of every student take effect to amount of SKS in the next semester. Thereforethe amount of SKS will take big effect to student study time. GPA variable classified into 3 parts.

Table 3.2 Student GPA

No	IPK	Keterangan
1	IPK > =3	3
2	IPK >= 2 dan IPK < 3	2
3	IPK < 2	1

g. Decision

Decision variable is a data that functioning to decide the result.In data classification already fixed, so there is no mistake in calculation of the software. Decision data only have three value, **fast**, **on time**, and **late**.

4. RESULT AND DISCUSSION

In this stages, begins with getting data samples from student whom already graduate to be used as data training. Used data already clean up and transformed into category.In this test data sample collected from 2011-2012 generation that already graduate. From 191 student data, 50 record taken as data training.Based on data processing from the sample data, it classified into fast category is 21 students, on time category 6 student, and late category 23 students.

In testing process, the data divide into two parts, training and test data.In *Naïve Bayes Classification* algorithm, traning data used to create table of probabilities, and test data used to test the probability table. The datacan be seen as below on table 4.1.

Table 4.1 Student Data Training

JENIS KELAMIN	KOTA LAHIR	TIPE SEKOLAH	KOTA SEKOLAH	IPK	EKONOMI	WAKTU KELULUSAN
L	DALAM KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
P	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	TEPATWAKTU
L	LUAR KOTA	KEJURUAN	LUAR KOTA	3	TINGGI	CEPAT
L	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT
L	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	2	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
L	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	2	SEDANG	TEPATWAKTU
L	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT

L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	TELAT
L	DALAM KOTA	UMUM	DALAM KOTA	2	SEDANG	TELAT
P	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	TELAT
L	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	2	SEDANG	TELAT
L	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT
P	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
P	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	TINGGI	TEPATWAKTU
L	LUAR KOTA	KEJURUAN	LUAR KOTA	2	SEDANG	TELAT
L	DALAM KOTA	UMUM	LUAR KOTA	3	SEDANG	TEPATWAKTU
L	DALAM KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
L	DALAM KOTA	UMUM	DALAM KOTA	2	SEDANG	CEPAT
P	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TEPATWAKTU
L	DALAM KOTA	KEJURUAN	DALAM KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	TELAT
L	LUAR KOTA	KEJURUAN	DALAM KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	TELAT
P	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	2	SEDANG	CEPAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TELAT
L	DALAM KOTA	KEJURUAN	LUAR KOTA	3	SEDANG	TEPATWAKTU
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT
L	LUAR KOTA	UMUM	DALAM KOTA	2	TINGGI	CEPAT
P	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
P	LUAR KOTA	UMUM	DALAM KOTA	3	SEDANG	CEPAT
P	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	KEJURUAN	DALAM KOTA	2	SEDANG	CEPAT
L	LUAR KOTA	KEJURUAN	DALAM KOTA	2	SEDANG	CEPAT
L	DALAM KOTA	UMUM	DALAM KOTA	2	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TELAT
L	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	TELAT
P	LUAR KOTA	UMUM	LUAR KOTA	3	SEDANG	CEPAT

5. TESTING

This testing intend to understand the *Naïve Bayes Classification*algorithm in data classification into class which has been specified.On this test, training data given to create probability table.The next step will be given the test data to test the table of probability.

Based on training data on table 4.1, it can be classified if the student data given input in gender, school type, School

location, economic welfare, and GPA using *Naive Bayes Classification* algorithm.This is the example of student data of IBI Darmajaya majoring Information Engineering that the class are unknown.

Gender = Male
Hometown = Inner city
School type = Vocational
School Location = Inner city
GPA = 3
Economic = Medium

Based on the test data, it can be decide by this few steps :

1. Calculating Number of class / label

   P (fast) = 21/50 "The amount of data***Fast***in data training divided by all amount of data".

   P (On Time) = 6/50 "The amount of data***On Time***in data training divided by all amount of data".

   P (Late) = 23/50 "The amount of data***late***in data training divided by all amount of data".

2. Calculating the amount of the same case with the same class :

   P (Gender = Male| Y=Fast) = 15/21

   P (Gender = Male|Y=On Time) = 4/6

   P (Gender = Male|Y=Late) = 20/23

   P(Hometown= Inner City | Y= Fast)=5/21
   P(Hometown= Inner City | Y= On Time)=3/6
   P(Hometown= Inner City | Y= Late)=5/23
   P(School Type= Vocational | Y=Fast)=5/21
   P(School Type= Vocational | Y=On Time)=2/6
   P(School Type= Vocational | Y=late)=7/23
   P(School Location= Inner City | Y=Fast)=11/21
   P(School Location= Inner City | Y= On Time)=1/6
   P(School Location= inner City | Y= Late)=8/23
   P(IPK= 3 | Y=Fast)=16/21
   P(IPK= 3 | Y= On Time)=5/6
   P(IPK= 3 | Y=Late)=14/23
   P(Economic= Middle| Y=Fast)= 19/21
   P(Economic= Middle| Y=On Time)=5 /6
   P(Economic= Middle| Y=late)=23/23

3. Multiplied all the result of fast variable, on time variable, and late variable.

   P(Male\Fast) x P(Inner City\Fast) x P(vocational\Fast) x P(2.8\Fast) x P(Inner City\ Fast) x P(Middle\ Fast).

   $=\frac{15}{21} x \frac{5}{21} x \frac{5}{21} x \frac{11}{21} x \frac{16}{21} x \frac{19}{21}$

   = 0.7143 x 0.2381 x 0.2381 x 0.5238 x 0.7619 x 0.9048

   =0.0146

   P(Male \On time) x P(Inncer City\on time) x P(vocational\on time) x P(2.8\ on time) x P(inner city\ on time) x P(middle\ on time).

   $=\frac{4}{6} x \frac{3}{6} x \frac{2}{6} x \frac{1}{6} x \frac{5}{6} x \frac{5}{6}$

   =0.0667 x 0.5000 x 0.3333 x 0.1667 x 0.8333 x 0.8333

   = 0.0129

   P(Male\late) x P(inner city\late) x P(vocational\late) x P(2.8\ late) x P(Inner city\late) x P(middle\late).

   $=\frac{20}{23} x \frac{5}{23} x \frac{7}{23} x \frac{8}{23} x \frac{14}{23} x \frac{23}{23}$

   =0.8696 x 0.2174 x 0.3043 x 0.3478 x 0.6087 x 1

   = 0.0122

4. Compare the result from fast, on time, and late.

   From the result above, we can see the highest probability value belongs to class (P|Fast), so we can conclude that the student graduate fast.

6. RESULT

According to the implementation result using 50 training data which from every class with total percentage 42%

Fast class, 12% on time class and 46 late class. From three class, late class got the highest point. Next step is doing the testing for 20 test data then obtained 20% fast class, 35% on time class and 45% late class. From every class, the subclass was taken as comparison, such as gender subclass and school city subclass.

According to the testing data, the result is woman tend to graduate faster or on time rather than male that more late dominant. The same result as school city, where university student from another region/city tend to be late rather than university student that come from the city. The student university that come from the city tend to be fast or on time to graduate.

7 CONCLUSION

According to the problem background and the discussion on the previous section, the conclusions are :

1. Training data percentage that used is 42% Fast class, 12% on time class and 46% late class. The determination of training data can be the rule for data testing.
2. According to testing result using 20 testing data with gender subclass obtained : male that graduate fast 1 person, on time 4 person and late 8 person. On the other side, female that graduate fast 3 person, on time 3 person and late 1 person. From that result, female has bigger opportunity to graduate fast or on time.
3. *Naïve Bayes* algorithm is supported by probabilistic science and statistic science, especially in using guide data to support the decision of classification. In *Naïve Bayes* algorithm, every attribute will give contribution in decision making with attribute integrity that equally important and every attribute independent one and another.
4. The period of study or in this case accuracy of study period every university student can be predicted according to the high school they attend before, genders and academic data and also personality in university.

SUGGESTION

1. Total data that used as training data or testing data can be added until the result obtain a better algorithm function.
2. For the future developing, there's possibility to do more trials with another algorithm and the result can be compared and analyzed.
3. Predictor variable that used can be added more and the data value variation can be more and data consistency can be noted.

REFERENCES

[1] Han, J., Kamber, M. (2000). Data mining: *Concepts and Techniques New York: Morgan- Kaufman*.

[2] Budi S.(2007). Data mining: *Teknik Pemanfaatan data untuk keperluan bisnis*. Teori & Aplikasi. Garha Ilmu: Surabaya.

[3] Sri Kusumadewi (2009).CommIT, Vol. 3 No. 1 Mei 2009, hlm. 6 – 11. *"Klasifikasi status gizi menggunakan Naive bayesian classification"*.

[4] Amir Hamzah (2012).SNASTPeriode III *"Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis"*, Yogjakarta.

[5] Fithri, A.L(2013): Jurnal SIMETRIS, Vol 4 No 1 Nopember 2013 *"Sistem Pendeteksian Penyimpangan Tingkah Laku Anak Usia 0 sampai 3 Tahun Dengan Metode Bayesian"*.

[6] Bustami,(2013)TECHSI: Jurnal Penelitian Teknik Informatika *"Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi"*.

[7] Salim,Y. (2012): Media SainS, Volume 4 Nomor 2, *"Penerapan Algoritma Naive Bayes Untuk Penentuan Status Turn-Over Pegawai"*

[8] Rosmala D, Wulandari.(2014). Konferensi Nasional Sistem Informasi 2014, No Makalah : 050 *"Implementasi Crisp-Dm Dan Naïve Bayes Classifier Pada Datamining Churn Prediction"*.

[9] Kusrini , L. Taufiq Emha . 2009. *Algoritma Data Maining*, Edisi Pertama. Andi Yogyakarta.

[10] S.Yeffrianjah (2012). Media SainS, Volume 4 Nomor 2, Oktober 2012 "*Penerapan algoritma naive bayes Untuk penentuan status turn-over pegawai"*

[11] Sutrisno, Widianto, Afriyudi.(2013).Jurnal Ilmiah Teknik Informatika Ilmu Komputer Vol.x No.x, 4 November 2013 : 1-11 *"Penerapan data mining pada penjualan menggunakan metode clustering study kasus pt. Indomarco Palembang"*.