ALUMNI WAITING TIME ESTIMATION FOR GETTING JOB USING DATA MINING PREDICTION METHOD WITH NAIVE BAYES CLASSIFIER ALGORITHM

[1]**Nadiyah Maharty Ali**

Informatics Engineering of Muhammadiyah University of Yogyakarta

Email: mahartyali30@gmail.com

ABSTRACT

There is a lot of data from students and alumni from Muhammadiyah University of Yogyakarta, one of which is the data of the alumni in getting their job after completing their undergraduate study. There are several factors that make it difficult for alumni to get a job. This study aims to classify how long alumni can get a job by using the data mining process and using the Naive Bayes classification algorithm. The algorithm that is used, can predict sooner or later an alumnus gets a job and from the prediction results can be used as a base to a decision making, so that it can improve the quality of a university.

The support system in this study uses several parameters, such as their gender, faculty, GPA, graduation year, and the status of getting a job. The data used is 435, consisting of 7 faculties from 2011-2014. As the results of this study, the authors get the accuracy of the alumni time period to get a job is 71% and from the prediction results of the alumni Muhammadiyah of Yogyakata University from 2011-2014 get a job, there is more who get a job faster than those who are slow.

Keywords: predict the grace period to get a job, *data mining, Naive Bayes, RapidMiner.*

INTRODUCTION

A university definitely have alumni data such as the name of the student, study program that the student taken, gender and also the data of students who have graduated and who have already got a job. When data is not classified in certain groups or rules, it can cause other problems, such as slowing down the data search process. If we analyzed more deeply, classified data can provide a lot of information.

Alumni play an important role in improving the quality that has been achieved by a college, because the sooner the alumni get the job, it means that the system of teaching and learning at a university is good enough. From here the author requires data mining, By using data mining techniques can predict the length of alumni get a job after completing the undergraduate study program. Using Data Mining Prediction Method With Naive Bayes Classifier Algorithm.

BACKGROUND OF THE PROBLEM

The background of the problem in the study is that the data is not classified or grouped according to the information so that it cannot be used to predict the class of an object that in unknown class, meaning that it can cause other problems, such as slowing down the search for the data needed. So that the writer will calculate and predict the grace period for alumni of the Muhammadiyah University of Yogyakarta to get a job after completing the undergraduate study.

PURPOSE OF RESEARCH

The aim of this research is to predict the estimated grace period or how long the alumni of Muhammadiyah University of Yogyakarta get a job after completing their undergraduate studies.

By knowing the prediction of how long the alumni can get a job, prospective alumni or students who are still active in college can take advantage of being more diligent in studying to get a GPA above the average, and can graduate on time. In addition, by utilizing the department that has been undertaken, with the intention of getting a job that is in accordance with the department. And the benefits for the university are UMY's Alumni Data Managers no longer need to use manual methods in predicting how long alumni can get a job.

THEORETICAL FRAMEWORK

DATA MINING

Data mining is a knowledge discovery in database (KDD). Data Mining is one of the ways used to gain new knowledge by utilizing a very large amount of data. Several techniques have been developed and implemented to extract knowledge that might be useful for decision making. The techniques used to extract knowledge in data mining are pattern recognition, clustering, association, prediction and classification. [1]

There are several techniques that data mining has based on the tasks performed, such as [2]:

1. **Prediction**

Predictions are similar to estimates and classifications. It's just that the prediction of the results shows something that hasn't happened (maybe will happen in the future). It is necessary to estimate the Waiting Period for Alumni to Get a Job.

2. **Estimates**

Estimates are similar to classifications, except that the destination variable is more numerical than the category.

3. **Classification**

In variable classification, goals are categorical. For example, we will classify income in 3 classes, that is; high income, moderate income and low income.

4. **Clustering**

Clustering is more towards grouping records, observations, or cases in a class that has similarities.

5. **Association**

The task of the association is to identify the relationship between various events that occur at one time.

NAIVE BAYES ALGORITHM

Naive Bayes is a classification of statistics that can be used to predict the probability of membership of a class. Bayesian classification is based on Bayes's theorem which has a classification capability similar to a decision tree and neural netrok. Bayesian classfication has proven high accuracy and speed when applied to databases with large data. [3] *Naive Bayes* has the following equation::

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

X = Data with unknown classes
H = The data hypothesis X is a specific class
P(H|X) = Proportion of hypothesis H based on condition x (posteriori prob.)
P(H) = Probalance of hypothesis H (prior prb.)
P(X|H) = Probalance X based on the condition P
(X) = Probalance of X

METHODOLOGY OF RESEARCH

In this study, the object of the study was the alumni of Yogyakarta Muhammadiyah University from 2011-2014, with 7 faculties.

The steps or stages in this research are as follows:

1. Literature study, namely the initial process in which the authors collect relevant information and are needed in this study.
2. Data Collection, in this process the author gets 689 preliminary data, with many attuts to the obtained Excel file.
3. Data Selection, the process by which the author selects from the many attributes that exist, then the author uses 4 attributes as an ordinary class, and 1 main class as a label that is gender, faculty, GPA, year of graduation and also the date of starting work. Because the information contained in it already represents the information needed to be used as an indicator of research.

jenis_kelamin	id_prodi	tahun_lulus	ipk	tanggal_mulai_kerja
L	32	2003	NULL	NULL
L	11	2009	3.14	NULL
L	52	2003	3.04	NULL
L	104	NULL	NULL	NULL
L	104	NULL	NULL	NULL
L	104	NULL	NULL	NULL
L	104	NULL	NULL	NULL
L	104	NULL	NULL	NULL
L	104	NULL	NULL	NULL
L	104	NULL	NULL	NULL
L	61	2011	3.06	2013-11-01 00:00:00.000
L	11	2011	2.34	NULL

Image 1 Data Selection

4. Data cleaning, the process by which the author deletes and removes data that is still null (0). After the author has cleaned the data, the data used for this study is 435 data. And it will be processed into two processes, namely training data and testing data.
5. Data Transformation & Data Initialization, which is the stage of converting data into an appropriate form for this research process. That was changed, that is, at first id_prodi was changed to faculty, the start_work_date was changed to start_work_status, and the GPA and Start_Work_Statuus's value was initialized.

jenis_kelam	Fakultas	ipk	tahun_lulus	status_mulai_kerja
L	HUKUM	3.00	2011	2013-11-01 00:00:00.000
L	TEKNIK	3.48	2011	2012-04-14 00:00:00.000
L	KEDOKTERAN	3.04	2011	2015-12-01 00:00:00.000
P	KEDOKTERAN	3.46	2011	2015-10-01 18:01:00.000
L	KEDOKTERAN	3.04	2011	2015-09-01 00:00:00.000
P	KEDOKTERAN	2.99	2011	2015-12-15 00:00:00.000
L	KEDOKTERAN	3.04	2011	2016-01-01 00:00:00.000
P	KEDOKTERAN	3.01	2011	2015-02-02 00:00:00.000
L	KEDOKTERAN	2.94	2011	2016-03-28 00:00:00.000
P	KEDOKTERAN	2.77	2011	2016-04-01 00:00:00.000
P	EKONOMI	3.48	2011	2013-01-01 00:00:00.000
P	EKONOMI	3.24	2011	2012-06-21 00:00:00.000
L	TEKNIK	3.51	2011	2012-01-04 00:00:00.000
L	TEKNIK	3.12	2011	2015-02-02 00:00:00.000
L	PENDIDIKAN AGAMA	3.36	2011	2012-07-31 00:00:00.000
L	ISIPOL	3.34	2012	2014-10-06 00:00:00.000
L	ISIPOL	2.65	2012	2016-01-04 00:00:00.000
L	ISIPOL	3.14	2012	2015-08-10 21:21:00.000

Image 2 Data Transformation

Table 1. 1 GPA Initialization

GPA	INITIALIZATION
GPA <3	LESS THAN 3
GPA 3>3.50	3 TO 3.50
GPA >3.50	MORE THAN 3.50

Table 1. 2 Initialization of Start Working Status

Start Working Status	INITIALIZATION
From the year of graduation– 2 years	FAST
More Than 2 Years	SLOW

Here is the following data image that has been transformed and has been initialized:

	jenis_kelamin	fakultas	ipk	tahun_lulus	tanggal_mulai_kerja
3	P	PENDIDIKAN AGAMA	DARI 3 SAMPAI 3.50	2012	CEPAT
4	L	TEKNIK	LEBIH DARI 3.50	2011	CEPAT
5	P	ISIPOL	LEBIH DARI 3.50	2012	CEPAT
6	P	ISIPOL	DARI 3 SAMPAI 3.50	2012	CEPAT
7	P	EKONOMI	DARI 3 SAMPAI 3.50	2012	CEPAT
8	L	PENDIDIKAN AGAMA	DARI 3 SAMPAI 3.50	2012	CEPAT
9	L	TEKNIK	DARI 3 SAMPAI 3.50	2011	CEPAT
10	P	EKONOMI	DARI 3 SAMPAI 3.50	2012	CEPAT
11	P	EKONOMI	LEBIH DARI 3.50	2012	CEPAT
12	L	ISIPOL	LEBIH DARI 3.50	2012	CEPAT
13	P	PENDIDIKAN AGAMA	LEBIH DARI 3.50	2012	CEPAT
14	L	TEKNIK	DARI 3 SAMPAI 3.50	2012	CEPAT
15	P	ISIPOL	DARI 3 SAMPAI 3.50	2012	CEPAT
16	L	TEKNIK	LEBIH DARI 3.50	2012	CEPAT
17	P	EKONOMI	DARI 3 SAMPAI 3.50	2011	CEPAT
18	L	TEKNIK	KURANG DARI 3	2012	CEPAT
19	L	ISIPOL	DARI 3 SAMPAI 3.50	2012	CEPAT
20	L	PENDIDIKAN AGAMA	LEBIH DARI 3.50	2012	CEPAT

Image 3 Data Initialization

6. Implementation is a process of determining information from the data used. The technique used is the prediction method with the Naive Bayes algorithm and with the RapidMiner software. The process is by importing training data files and testing data on the software used and then connecting with RapidMiner tools operators.

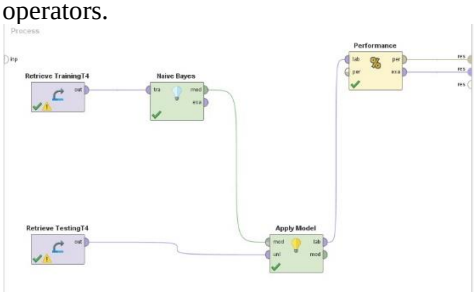

Image 4 Linking Performance Operators

After running, the prediction results will appear in the view result. Looks like in image 5 below.

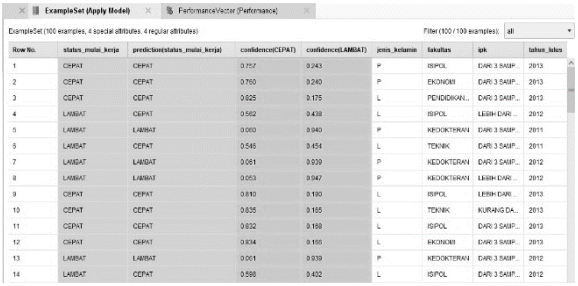

Image 5 The Calculation Results(ExampleSet)

In the picture above, it provides information about alumni who get a job faster or slower. In Image 6 is a display of accuracy, class prediction and prediction accuracy, and class recall.

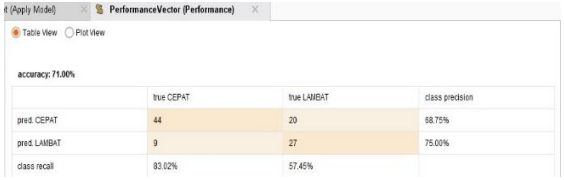

Image 6 The Calculation Result (PerformanceVector)

From the above results, it can be seen that the accuracy of the data used is 71%, and on FAST class prediction is 68.75% and SLOW of 75.00%. As for the class recall itself for the FAST class of 83.02% and for the SLOW class of 57.45%.

In general, Precision, Recall, and Accuracy can be formulated as:

- For the FAST class:

$$\text{precition} = \frac{44}{44+20} = \frac{44}{64} = 0{,}6875 = 68{,}75\%$$

$$\text{recall} = \frac{44}{44+9} = \frac{44}{53} = 0{,}83018868 = 83{,}02\%$$

- For the SLOW Class

$$\text{precition} = \frac{27}{27+9} = \frac{27}{36} = 0{,}75 = 75\%$$

$$\text{recall} = \frac{27}{27+20} = \frac{27}{47} = 0{,}57446809 = 57{,}45\%$$

- For Accuracy

$$accuracy = \frac{44+27}{44+27+20+9} = \frac{71}{100} = 0{,}71 = 71\%$$

TESTING NAIVE BAYES ALGORITHM

The author uses 435 training data and 100 testing data.

Table 2 1 Data Training

No	Jenis_kelamin	fakultas	ipk	Tahun_lulus	Status_mulai_kerja
1	P	PENDIDIKAN AGAMA	DARI 3 SAMPAI 3.50	2012	CEPAT
2	L	TEKNIK	LEBIH DARI 3.50	2011	CEPAT
3	P	ISIPOL	LEBIH DARI 3.50	2012	CEPAT
4	P	ISIPOL	DARI 3 SAMPAI 3.50	2012	CEPAT
5	P	EKONOMI	DARI 3 SAMPAI 3.50	2012	CEPAT
6	L	PENDIDIKAN AGAMA	DARI 3 SAMPAI 3.50	2012	CEPAT
7	L	TEKNIK	DARI 3 SAMPAI 3.50	2011	CEPAT
8	P	EKONOMI	DARI 3 SAMPAI 3.50	2012	CEPAT
9	P	EKONOMI	LEBIH DARI 3.50	2012	CEPAT
10	L	ISIPOL	LEBIH DARI 3.50	2012	CEPAT
.....	P	KEDOKTERAN	DARI 3 SAMPAI 3.50	2012	LAMBAT
.....	L	TEKNIK	KURANG DARI 3	2014	LAMBAT
.....	P	ISIPOL	LEBIH DARI 3.50	2014	LAMBAT
.....	L	HUKUM	DARI 3 SAMPAI 3.50	2013	LAMBAT
435	L	EKONOMI	LEBIH DARI 3.50	2013	LAMBAT

Table 2 2 Data Testing

No	Jenis kelamin	fakultas	ipk	Tahun_lulius	Status_mulai_kerja
1	P	ISIPOL	DARI 3 SAMPAI 3.50	2013	CEPAT
2	P	EKONOMI	DARI 3 SAMPAI 3.50	2013	CEPAT
3	L	PENDIDIKAN AGAMA	DARI 3 SAMPAI 3.50	2013	CEPAT
4	L	ISIPOL	LEBIH DARI 3.50	2012	LAMBAT
5	P	KEDOKTERAN	DARI 3 SAMPAI 3.50	2011	LAMBAT
6	L	TEKNIK	DARI 3 SAMPAI 3.50	2011	LAMBAT
7	P	KEDOKTERAN	DARI 3 SAMPAI 3.50	2012	LAMBAT
8	P	KEDOKTERAN	LEBIH DARI 3.50	2012	LAMBAT
...	L	KEDOKTERAN	DARI 3 SAMPAI 3.50	2011	LAMBAT
100	P	HUKUM	DARI 3 SAMPAI 3.50	2013	LAMBAT

In the case in table 2.1 and table 2.2 it will be predicted to determine which alumni in table 2.2 get a job quickly or slowly using probability. To calculate the probability value or prediction of alumni who get a job sooner or later by using the method as follows::

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

First, calculate the amount from Fast and Slow from the training data table. From the training data table, the following results are obtained:

- FAST = 237
- SLOW = 198

After knowing the number of alumni who is faster and slower to get a job, the next step is to calculate each attribute from the data testing as follows:

- *Gender* = P, *Faculty* = ISIPOL, *GPA* = 3 TO 3.50, *Graduation Year* = 2013
- Step 1 : Calculate the number of classes/labels
  P(start working status | amount of data)
  P(FAST / 435) = 237/435 = 0,54462759
  P(SLOW / 435) = 198/435 = 0,45517241

- Step 2 : Calculate the number of *record atribut* with the same class/label
  ➢ P (Gender | amount of data (fast & slow)
  P (P / FAST = 95/237 = 0,4008438819

  (the above calculation is the number of "female" sex data with "fast" information divided by the number of fast data)

  P (P / SLOW = 126/198 = 0,5151515152

  (the above calculation is the number of "female" sex data with "slow" information divided by the amount of slow data)

  ➢ P (Faculty = ISIPOL | FAST) = 101/237 = 0,4261603376
  (the above calculation is the amount of data on the "ISIPOL" faculty with "fast" information divided by the number of fast data)
  P (Faculty = ISIPOL | SLOW) = 72/198 = 0,2424242424
  (the above calculation is the amount of data on the "ISIPOL" faculty with "slow" information divided by the number of slow data)
  ➢ P (GPA = 3 TO 3.50 | FAST) = 131/237 = 0,552742616
  (the above calculation is the amount of GPA data "3 to 3.50" with the "fast" information divided by the number of fast data)
  P (IPK = 3 TO 3.50 | SLOW) = 108/198 = 0,5454545455
  (the above calculation is the amount of GPA data "3 to 3.50" with the "slow" information divided by the number of slow data)
  P (Graduation Year = 2013 | FAST) = 128/237 = 0,5400843882
  (the above calculation is the amount of data with the year of graduation "2013" with the information "fast" divided by the number of fast data)

➢ P (Graduation Year = 2013 | SLOW) = 57/198 = 0,2878787879
(the calculation above is the amount of data with the year of graduation "2013" with the statement "slow" divided by the amount of slow data)

- Step 3 : Multiplied all the results of the FAST & SLOW variable
  ➢ P (Gender = P | FAST) x P(Faculty = ISIPOL | FAST) x P(GPA = 3 TO 3.50 | FAST) x P(Graduation Year = 2013 | FAST)
  = 0,4008438819 x 0,4261603376 x 0,552742616 x 0,5400843882
  = 0,0509956181
  ➢ P (GENDER = P | SLOW) x P(Faculty = ISIPOL | SLOW) x P(GPA = 3 TO 3.50 | SLOW) x P(Graduation Year = 2013 | SLOW)
  = 0,5151515152 x 0,2424242424 x 0,5454545455 x 0,2878787879
  = 0,0095079153

- Step 4 : Compare the results of FAST & SLOW classes
  FAST = 0,0509956181
  SLOW = 0,0095079153

Table 2 3 Manual Calculation Results Table

Jenis_ Kelamin	Fakulta s	ipk	Tahun_ lulus	Status_mulai_kerja	Hasil prediksi
P	ISIPOL	Dari 3 sampai 3.50	2013	CEPAT	*CEPAT*

Because the results (P | FAST) are greater than (P | SLOW) the results of the prediction for alumni of the ISIPOL faculty of female class of 2013 with GPA from 3 to 3.50 are "FAST".

From the calculation above, it is known that alumni with the category of Women, Faculty of ISIPOL, with GPA from 3 to 3.50, year of graduation in 2013, and the status of starting work quickly predicted to get a fast job because the results of the Fast class calculation is greater than the Slow class with a probability of 0,0509956181. Looks like in picture 7

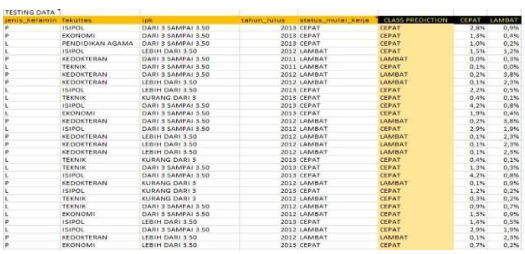

image 7 Prediction Results in Excel

DISCUSSION

The following are the results of the tests that have been carried out on RapidMiner and the Naive Bayes Algorithm:

1. In this study the author uses 435 data as training data and 100 data as testing data as manual testing in Excel and on testing using RapidMiner software.

2. Manual testing on Excel using the Naive Bayes model is that in 2011 more alumni were slower to get jobs, with a comparison of 3% fast - 5% slow, in 2012 more alumni were slower to get jobs, with a comparison of 37% fast - 66% slow, in 2013 more alumni were faster to get jobs, with a comparison of 54% fast - 29% slow, and in 2014 more faster to get a job, with a comparison of 6% fast - 1% slow.

P/tahun_lulus	CEPAT	LAMBAT
2011	3%	5%
2012	37%	66%
2013	54%	29%
2014	6%	1%
	100%	100%

Image 1 Calculation of the Naive Bayes Manual Model

(Graduation Year)

3. The Class Prediction section uses 100 data in the testing data. There were data that matched between fast class and fast predicted class, which were 44 data and slow data that matched the slow class and slow predicted class was 27 data. Whereas for improper prediction is 29 data.

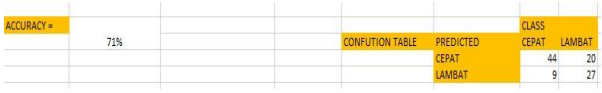

image 2 Class Prediction's Calculation Results (confuntion table)

4. The accuracy level in manual calculation in Excel and testing using the RapidMiner software is the same, which is 71%. For the FAST class, the RapidMiner prediction is 68.75%, and the recall in the FAST class is 83.02%. Whereas prediction for SLOW class is 75% and recall in SLOW class is 57.45%. (as in image 4.24)

5. The factors that influence the FAST and SLOW end result of the RapidMiner software for alumni in getting a job because of the manual calculation of the Naive Bayes model on Excel are higher percent in the fast class, which consists of attributes of gender, faculty, GPA, year of graduation and work start status.

CONCLUSION & SUGGESTION

CONCLUSION

After testing and analyzing the author gets the following conclusions:

1. The Naive Bayes algorithm can be used to predict the grace period or the length of alumni of the Muhammadiyah University of Yogyakarta in getting a job.
2. The Naive Bayes algorithm in predicting the grace period or length of alumni getting a job has a accuracy level of performanceVector that is 71%, class precision that is FAST 68%, SLOW 75%, and for class recall is; FAST 83.02% while SLOW 57.45 %.
3. Factors that influence the FAST and SLOW end result of RapidMiner software, alumni can get jobs because from the manual calculation of the Naive Bayes model on Excel has a higher percent in the fast class, which consists of gender, faculty, GPA, year of graduation, and start working status.

4. The information obtained in this study is that Muhammadiyah University of Yogyakarta alumni are predicted to get a job sooner after completing their S1 study.

SUGGESTION

The suggestion given by the author is

1. It is expected that the UMY database is updated every year, so it will making it easier for subsequent research.

2. If there are researchers who want to continue this research, they should use a different algorithm as a comparison of accuracy.

REFERENCES

[1] S. Defiyanti, "Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining," *Konf. Nas. Inform.*, no. March 2017, pp. 39–44, 2015.

[2] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Eeccis*, vol. 7, no. 1, pp. 59–64, 2013.

[3] A. Jananto, "Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa," *Teknol. Inf. Din.*, vol. 18, no. 1, pp. 9–16, 2013.