Decision Support Systems to Determining Programme for Students Using DBSCAN And Naive Bayes

Case Study: Engineering Faculty Of Universitas Nusantara PGRI Kediri

Erna Daniati
Engineering Faculty
Universitas Nusantara PGRI Kediri
Kediri, Indonesia
ernadaniati@unpkediri.ac.id

***Abstract*—Determination of programme interest has not been supported by adequate information. This causes students who have chosen programme to be less concentrated in college. This is because the chosen programme turns out to be not as expected to determine the programme. There is a system that is used to help make decisions. The system is a decision support system. The decision support system has several methods. The method is used to obtain modeling. There is a DBSCAN method that is used for grouping or clustering. This clustering is needed in order to decision maker is not input manually about preferences. Clustering contributes to add preferences automatically. So, decision maker can only focus on selecting alternatives decision. Furthermore, this alternative decision can be found using Naive Bayes Method. In addition, the Naive Bayes also has a fairly high accuracy. This is indicated by the results of testing in this research. The accuracy of Naïve Bayes has a high value compared to other classifiers.**

Keywords—Decision Support System, DBSCAN, Naïve Bayes

I. INTRODUCTION

Universitas Nusantara PGRI Kediri has several faculties. The location of the building from this faculty is in the city of Kediri, precisely in the Mojoroto area. This shows that the faculty's location is strategic and easy to reach from the city. One of the faculties that are in great demand is the engineering faculty. The faculty has adequate infrastructure in accordance with the programme. Every programme on this campus has been accredited by the Director General of Higher Education. This shows that this faculty has a programme that is ideal for the learning process of the students [1].

Every year, engineering faculties hold student admissions. Students who enroll in this campus are around 1000 students for each faculty [1]. This is a potential resource for the campus in carrying out the process of the academic community. Admission of these new students still uses the entrance test and the average score from the national final examination. The results of this entrance examination can also be used to classify students in several classes. Students who register on this campus also come from various regions. This has made the plurality of the UNP Kediri campus a major force. The Faculty of Engineering, Universitas Nusantara PGRI Kediri has 5 programmes. The programme is Mechanical, Electrical, Informatics (IT), Information Systems (IS), and Industrial Engineering. Each programme has a curriculum that has been standardized by the Director General of Higher Education in Indonesia. Every programme has its own graduation performance. This is intended to foster students who can be competent after graduating from college.

When new students begin for registratio, they can choose the programme freely. Students take the entrance examination process and verify the final examination score. Students choose the programme based on their interests. Determination of this interest has not been supported by adequate information. This led to students who had chosen the programme to be less concentrated in college. This is because the chosen programme was not as expected. Besides, some of the subjects were not as desired. Then, the socialization of the achievements of each programme is still incomplete in providing information. Students usually look for information on determining programmes through books or internet sites. This is because there is no system that is able to support students to determine programmes.

There is a system that is used to help make decisions. The system is a Decision Support System [2] [3]. This system is part of computer-based information systems (including knowledge-based systems or knowledge management) that are used to support decision making in an organization or company [4]. The Decision Support System also aims to solve various problems regarding managerial quality or company organization designed to develop work effectiveness and productivity [5].

The decision support system has several methods. The method is used to obtain modeling. This algorithm reduces processing time and increases grouping accuracy [6]. This algorithm classifies by finding the closest point or value from another point. The grouping of the DBSCAN method has a detailed level in determining the group members even though there are still members who are noise [7] . After the data are grouped, then the probabilistic value is sought to produce alternative decisions. This alternative decision can be searched by using the Naive Bayes method. The ranking of the probability of each data is used for decision support.

Decision making is one of the important events that need to be operational to the executive to deal with problems and choose the way in everyday life [8]. Therefore, many researchers have tried to understand and study decision-making processes in humans. If you have understood the whole decision making process in Humans, this has the ability to systematically control and manage everyone's decisions that lead to the path that is most suitable for each scenario.

978-1-5386-8448-1/19/$31.00 ©2019 IEEE

II. Decision Making Phase

Decision making is the process of choosing one of the many guidelines available to achieve the desired goal. This process has four phases, namely Intelligence, Design, Choosing, and Implementation [8][9][10]. These phases are shown in Figure 1. The decision-making process is the process of deciding which items to choose between alternatives. The Simon Model is proposed to present the phase of the decision making process. The first is the intelligence phase, which is the identification phase of the organization's goals and objectives related to a problem. A decision maker needs to define the problem and its characteristics. The second is the design phase, it is the phase of development, testing and validation of the decision making model. The third is the choice phase, which is the phase of finding, evaluating, and recommending solutions to problems that are appropriate. The last phase is implementation, which consists of placing the chosen solution to take action to solve the problem.

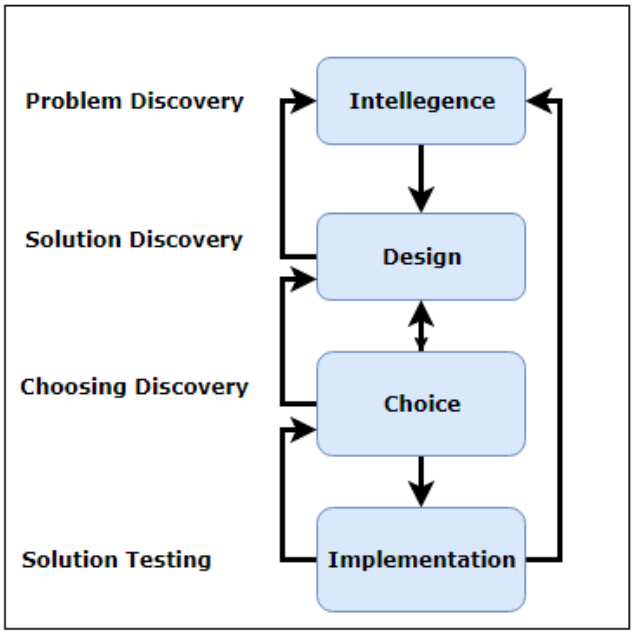

Fig 1. Decision Making Phases

III. Clustering Using DBSCAN

Cluster analysis or clustering is the process of partitioning a set of data objects (or observations) into subsets [11]. Each subset is a cluster, so objects in clusters are similar to each other, but different from objects in other clusters. The cluster set resulting from cluster analysis can be referred to as clustering. In this context, different grouping methods can produce different groupings in the same data set. Partitions are not carried out by humans, but by clustering algorithms. Therefore, grouping is useful because it can lead to the discovery of previously unknown groups in the data. Cluster Analysis has been widely used in many applications such as business intelligence, image pattern recognition, web search, biology, and security. In business intelligence, grouping can be used to organize large numbers of customers into groups, where customers in a group have the same strong characteristics. This facilitates the development of business strategies to improve customer relationship management.

This algorithm needs to choose a more prominent distance and size consumption characteristics, in terms of data clusters to be grouped, different areas are diversified [12]. To meet the increasing demand the distance between two points reflects the density between points, and indicates whether points and points can be grouped into the same class. The DBSCAN algorithm requires users to enter 2 parameters: One parameter is radius (Epsilon), which represents the range of circle environments centered around a fixed point P; the other parameter is the number of minimum points in the environment with a point P as the center (Minimum Points). Illustration of clustering results using DBSCAN is shown in Figure 2.

Fig 2. Ilustration Clustering Result Using DBSCAN

IV. Decision Alternatives Using Naïve Bayes

Bayes's theorem is named after Thomas Bayes, an inappropriate English priest who did initial work in probability theory and decisions during the 18th century. Let X be a data tuple. In Bayesian terms, X is considered "proof" [11]. As usual, this is explained by measurements made on set n attributes. Let H be a number of hypotheses such as that the tuple X data belongs to a specified class C. For classification problems, we want to determine P (H | X), the probability that the H hypothesis has by providing "proof" or observed tuple data X. In other words , we look for the probability that tuple X belongs to class C, since we know the description of attribute X.

The process of producing alternative decisions is done by the Naive Bayes method. This method corresponds to equation 1. In Equation 1, there is a set of D training data related to the class or label [5]. Next, the columns in the training data table are represented as $X = (x_1, x_2, ..., x_n)$. Then, the existing classes are represented as $C_1, C_2, ..., C_m$. Classifier will predict the X attribute included in the class where the posterior probability is high mentioned in X. Classifier Naive Bayes predicts the data line X that belongs to the C_i class if and only if $P(C_i | X) > P(C_j | X)$ for $1 < j < m, j \neq i$. The data training process flow that produces a probability table is shown in Figure 2.

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \quad \text{.....(1)}$$

V. Methodology

In this study, data collection was initially carried out. The collected data is then grouped using DBSCAN. The grouped data is useful for determining the modeling of each programme. Then, alternative decisions are generated by finding the probability of each existing programme data. The probability is sought using the Naive Bayes method. The general description of this decision-making process is shown in Figure 3. Then, the determination of attributes uses several questionnaires with the following questions:

1. What is your choice of main interest in the programme at the UNP Kediri Faculty of Engineering?
2. What is your second choice of interest in the programme at the UNP Kediri Faculty of Engineering?
3. Do you know the profile of graduates from the chosen programme?
4. What type of school do you come from in vocation school or high school?
5. At previous school, what did you take majors or concentration?
6. Mention one of the subjects that you like the most before?
7. What is the average score of the National Final Examination (UAN)?

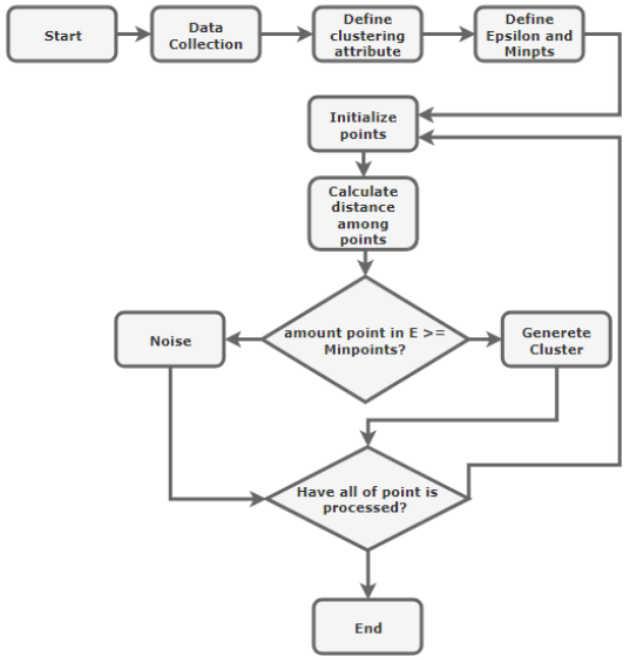

Fig 3. Clustering Using DBSCAN

Then, in Table I, the questionnaire results are followed by 81 prospective new students. The table contains 7 columns. Each row of data will be grouped according to the steps of DBSCAN in Figure 1. There are term that use for a meaning. IT is abbreviation of Information Technology and IS is Information System

TABLE I. THE ANSWER OF QUESTIONNAIRE

No	Main-interest	Second-interest	Profile	Origin-school	Major	Subject	UAN Score
1	industrial	IT	no	vocation school	busness-management	economy	6.2
2	industrial	IT	no	vocational school	busness-management	economy	6.3
3	electrical	IT	yes	high school	natural science	physic	6.3
4	electrical	IT	yes	high school	natural science	physic	6.1
5	IT	electrical	no	high school	other	other	8.3
6	IT	electrical	no	high school	other	other	8.5
7	IS	industrial	yes	vocational school	busness-management	accounting	9
8	IS	industrial	yes	vocational school	busness-management	accounting	9
9	mechanical	mechanical	no	vocational school	IT	math	7.5
10	mechanical	mechanical	no	vocational school	IT	math	7.7

In Table I there are 10 rows of data which are a sample of 81 new prospective students who fill in the quison data. In the UAN colon there are scores in numerical form. The score must be changed in nominal form. Changing the score follows the following rules:

score >= 8.0 → high

score >= 6.5 and score < 8.0 → medium

score < 6.5 → low

In Table II the results of the change in score are shown before the results.

TABLE II. DISCREATIZATION RESULT OF UAN SCORE

No	Main-Interest	Second-Interest	profile	Origin-school	Major	Subject	UAN Score
1	industrial	IT	no	vocational school	busness-management	economy	low
2	industrial	IT	no	vocational school	busness-management	economy	low
3	electrical	IT	yes	high school	natural science	physic	low
4	electrical	IT	yes	high school	natural science	physic	low
5	IT	electrical	no	high school	other	other	high
6	IT	electrical	no	high school	other	other	high
7	IS	industrial	yes	vocational school	busness-management	accounting	high
8	IS	industrial	yes	vocational school	busness-management	accounting	high
9	mechanical	mechanical	no	vocational school	IT	math	average
10	mechanical	mechanical	no	vocational school	IT	math	average

Then, from the overall data the quisoner data is clustered with DBSCAN with parameters as follows:

Epsilon: 1

Minpts: 2

From 81 quisoner data, there are 71 data noise and 10 data clustered. There are 5 data cluster that each cluster has 2 member of data. The clustering results are shown in Table III. The grouping results are in the form of Cluster 0, Cluster 1, Cluster 2, Cluster 3, and Cluster 4. Then, the name of the cluster is replaced with the name of programme according to the main interest and as a class label. This is shown in Table IV. Data Table IV also called model. Data is carried out model to obtain probabilities. This is done using the Naive Bayes method.

TABLE III. CLUSTERING RESULT USING DBSCAN

No	Main-Interest	Second-Interest	profile	Origin-school	Major	Subject	UAN Score	Cluster
1	industrial	IT	no	vocational school	busness-management	economy	low	Cluster 0
2	industrial	IT	no	vocational school	busness-management	economy	low	Cluster 0
3	electrical	IT	yes	high school	natural science	physic	low	Cluster 1

4	electrical	IT	yes	high school	natural science	physic	low	Cluster 1
5	IT	electrical	no	high school	other	other	high	Cluster 2
6	IT	electrical	no	high school	other	other	high	Cluster 2
7	IS	industrial	yes	vocational school	busness-management	accounting	high	Cluster 3
8	IS	industrial	yes	vocational school	busness-management	accounting	high	Cluster 3
9	mechanical	mechanical	no	vocational school	IT	math	average	Cluster 4
10	mechanical	mechanical	no	Vocational school	IT	math	average	Cluster 4

TABLE IV. RESULT OF SEPARATING NOISE AND DATA TRANING

No	Main-Interest	Second-Interest	profil	Origin-school	Major	Subject	UAN Score	Cluster
1	industrial	IT	no	vocational school	busness-management	economy	low	industrial
2	industrial	IT	no	vocational school	busness-management	economy	low	industrial
3	electrical	IT	yes	high school	natural science	physic	low	industrial
4	electrical	IT	yes	high school	natural science	physic	low	electrical
5	IT	electrical	no	high school	other	other	high	electrical
6	IT	electrical	no	high school	other	other	high	IT
7	IS	industrial	yes	vocational school	busness-management	accounting	high	IT
8	IS	industrial	yes	vocational school	busness-management	accounting	high	IS
9	mechanical	mechanical	no	vocational school	IT	math	average	IS
10	mechanical	mechanical	no	vocational school	IT	math	average	mechanical

In Figure 4, the training and data testing process is displayed. Data training starts with the availability of training data. The purpose of this data training is to produce a model in the form of a probability score list of each data. In this data training, the UAN column must be discrete. If it is still in nominal form or the number continues, no calculation can be made

This model is used to classify test data that does not yet have a class or label. However, for the purposes of decision support, the determination of this classification is stated. So, it only suffers from probable ranking. In Table V the results of data training are shown using the Naive Bayes method. Then, the amount of data used for training data is 10 data. This is because the previous data is noise.

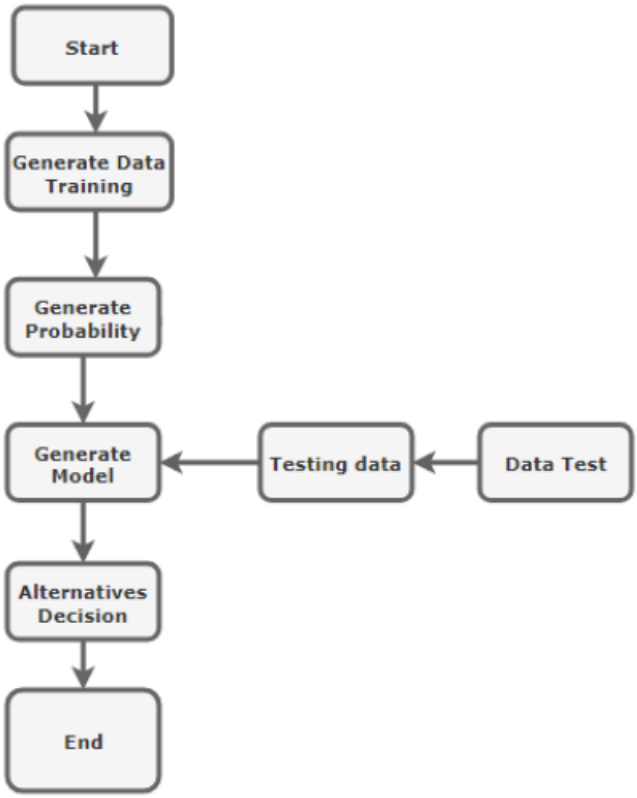

Fig 4. Training and Testing Data

TABLE V. MODEL AS A RESULT OF TRAINING DATA USING NAÏVE BAYES

Attribute		Class/Label				
		IT	IS	electrical	industrial	Mechanical
	Proba-bilitas Prior	0.2	0.2	0.2	0.2	0.2
Main Interest	electrical	1	1	3	1	1
	industrial	1	1	1	3	1
	mechanical	1	1	1	1	3
	IS	1	1	1	1	1
	IT	3	3	1	1	1
	total	7	7	7	7	7
Second Interest	IT	1	1	3	3	1
	IS	1	1	1	1	1
	mechanical	1	1	1	1	3
	industrial	1	3	1	1	1
	electrical	3	1	1	1	1
	Total	7	7	7	7	7
profile	no	3	1	1	3	3
	yes	1	3	3	1	1
	Total	4	4	4	4	4
Origin-school	vocational school	1	3	1	3	3
	high school	3	1	3	1	1
	Total	4	4	4	4	4
Major	busness-management	1	3	1	3	1
	electronic	1	1	1	1	1
	Natural science	1	1	3	1	1
	Social	1	1	1	1	1
	Other	3	1	1	1	1
	Mechanical	1	1	1	1	1
	Automotive	1	1	1	1	1
	Electrical-Power	1	1	1	1	1
	IT	1	1	1	1	1
	Total	11	11	11	11	11

Attribute		Class/Label				
		IT	IS	electric al	industrial	Mechani cal
	Proba-bilitas Prior	0.2	0.2	0.2	0.2	0.2
Subject	Accounting	1	3	1	1	1
	Economy	1	1	1	3	1
	Math	1	1	1	1	3
	Physic	1	1	3	1	1
	Language	1	1	1	1	1
	Other	3	1	1	1	1
	Total	8	8	8	8	8
UAN Score	low	1	1	3	3	1
	average	1	1	1	1	3
	high	3	3	1	1	1
	Total	5	5	5	5	5

In Table IX is the result of the calculation of probability for electrical Informatic (IT), Information System, Electrical, Industrial, and Machine Engineering programmes. This shows that students who enter the previous criteria are highly recommended to determine the programme is Electrical Engineering because of the highest probability. However, the number closes the possibility of choosing another programme. The main decision is in the student's hands not from the system because it is only to support the decision.

VI. Result and Discussion

The next step is to do the prepared test data. This test data is data entered by students. The student input the data shown in table VI supayes to produce alternative decisions. This alternative decision is expected to enable students to make decisions in determining the programme. This test data will be processed with a model obtained from calculations using the Naïve Bayes Method. The calculation process to produce this alternative decision is shown in Table VII.

The data entered on the decision support system produces alternative decisions as shown in Table VII. The alternative decision contains several programmes that are sorted according to the highest probability score. In Table VIII, the highest probability score is an electrical programme. So, students who input the test data are recommended to choose an electrical programme. However, the overall decision was with the student. Decision support systems only provide assistance to decide.

Testing the accuracy of the Naïve Bayes method is shown in table IX. The Naïve Bayes method is compared to the Linear SVM kernel and C4.5. The comparison table shows that the best accuracy is in the Bayes Method and the SVM Linear kernel method.

TABLE VI. Sample Of Testing Data

Main Interest	Second Interest	Profile	Origin-school	Major	Subject	UAN	Class
IT	IS	yes	high school	natural science	math	average	?

TABLE VII. The Rank Of Alternatives Decision Based On Probabilities

No.	Alternateives Decision	score
1	electrical	1.5654 x 10-5
2	IT	5.218 x 10-6
3	IS	5.218 x 10-6
4	mechanical	5.218 x 10-6

TABLE VIII. Comparisson Of Accuracy Evaluation

Method	MAE	RMSE	RAE (%)	RRSE (%)	Error (%)	Acuracy (%)
Naïve Bayes	0.024	0.0318	7.0011	7.4108	0	100
SVM Kernel Linear	0	0	0	0	0	100
C4.5	0.1067	0.2667	31.1111	62.2222	40	60

Description of Table VIII:

MAE: Mean Absolue Error

RMSE: Root Mean Squared Error

RAE: Root Absolute Error

RRSE: Root Relative Squared Error

I. Conclusion

In this study there are several things that can be summarized as follows:

1. Determination of criteria in this support system using the DBSCAN method so that criteria can be formed automatically and no need for involvement of decision makers.
2. The DBSCAN method can produce groups or clusters with the same number of clusters. However, the number of group no can be determined at the beginning of the process.
3. The process of producing a decision in this study uses the Naïve Bayes method.
4. Alternative Decisions with the highest Naïve Bayes probability score are used as recommendations.
5. The accuracy test results of the Naïve Bayes method have in common with the SVM method of the Linear kernel.

Acknowledgment

Our thanks go to Universitas Nusantara PGRI Kediri for all the support that has been given so that this research can be carried out successfully.

References

[1] E. Daniati and A. Nugroho, "K-Means Clustering With Decision Support System using SAW," IEEE Int. Conf. Control Syst. Comput. Eng., vol. 6, no. November, pp. 25–27, 2016.

[2] E. Turban, J. E. Aronson, and T.-P. Liang, "Decision Support Systems and Intelligent Systems," Decis. Support Syst. Intell. Syst., vol. 7, p. 867, 2007.

[3] V. L. Sauter, Decision Support Systems for Business Intelligence. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2011.

[4] D. M. Khairina, F. Ramadhani, S. Maharani, and H. R. Hatta, "Department recommendations for prospective students Vocational High School of information technology with Naïve Bayes method," in ICITACEE 2015 - 2nd International Conference on Information Technology, Computer, and Electrical Engineering: Green Technology Strengthening in Information Technology, Electrical and Computer Engineering Implementation, Proceedings, 2016, pp. 92–96.

[5] F. Burdi, A. H. Setianingrum, and N. Hakiem,

[6] "Application of the naive bayes method to a decision support system to provide discounts (Case study: PT. Bina Usaha Teknik)," in Proceedings - 6th International Conference on Information and Communication Technology for the Muslim World, ICT4M 2016, 2017, pp. 281–285.

[7] M. Li, D. Meng, S. Gu, and S. Liu, "Research and Improvement of DBSCAN Cluster Algorithm," in Proceedings - 2015 7th International Conference on Information Technology in Medicine and Education, ITME 2015, 2016, pp. 537–540.

[8] S. Sumathi and S. N. Sivanandam, Introduction to Data Mining and its Applications, vol. 29. 2006.

[9] L. Ploywattanawong, "Decision Support Systems Model for Admission to Higher Education," in 2016 2nd IEEE International Conference on Computer and Communications Decision, 2016, pp. 1359–1363.

[10] J. Chanwijit, W. Lomwongpaiboon, O. Dowjam, and P. Tangworakitthaworn, "Decision Support System for Targeting Higher Education," in Proceedings of the 2016 5th ICT International Student Project Conference, ICT-ISPC 2016, 2016, pp. 154–157.

[11] E. Madyestmadja and M. T. Oktavia, "Decision Support System In Determining The Programme Concentration In Higher Education," in 2016 Fourth International Conference on Information and Communication Technologies (ICoICT) DECISION, 2016, vol. 4, no. c.

[12] H. Jiawei, M. Kamber, and J. Pei, Data Mining Concept and Techniques, 3rd ed. Waltham, USA: Morgan Kaufman, 2012.

[13] L. Zhang, S. Deng, and S. Li, "Analysis of Power Consumer Behavior Based on The Complementation of K-Means and DBSCAN," in IEEE Conference on Energy Internet and Energy System Integration (EI2), 2017, no. 1.

TABLE IX. CALCULATION OF PROBABILITY USING NAÏVE BAYES

Probability /Class	**IT**			**IS**			**electrical**			**industrial**			**mechanical**		
	Count	**Total**	**Probability**	**Count**	**Total**	**Probability**	**Count**	**Total**	**Probability**	**Count**	**Total**	**Probability**	**Count**	**Total**	**Probability**
Class	2	10	0.2	2	10	0.2	2	10	0.2	2	10	0.2	2	10	0.2
Main-Interest: TI	3	7	0.428571429	3	7	0.42857143	1	7	0.14285714	1	7	0.142857143	1	7	0.142857143
Second-Interest: IS	1	7	0.142857143	1	7	0.14285714	1	7	0.14285714	1	7	0.142857143	1	7	0.142857143
UAN: average	1	5	0.2	1	5	0.2	1	5	0.2	1	5	0.2	3	5	0.6
Profile: yes	1	4	0.25	3	4	0.75	3	4	0.75	1	4	0.25	1	4	0.25
Origin-school: high school	3	4	0.75	1	4	0.25	3	4	0.75	1	4	0.25	1	4	0.25
Major: natural science	1	11	0.090909091	1	11	0.09090909	3	11	0.27272727	1	11	0.090909091	1	11	0.090909091
Subject: math	1	8	0.125	1	8	0.125	1	8	0.125	1	8	0.125	3	8	0.375
Naïve Bayes Probability			5.218×10^{-6}			5.218×10^{-6}			1.5654×10^{-5}			5.79777×10^{-7}			5.218×10^{-6}