The Comparison of K-Nearest Neighbor (K-NN) Algorithm and Fuzzy Tsukamoto Logic in the Determination of SMA Students Majors in Banten

T Y Akhirina*, A Rusmadiana, D Yulistyanti and U Pauziah

Universitas Indraprastata PGRI, Jakarta, Indonesia

*azizahputriku@gmail.com

Abstract. Students’ majors at High School (SMA) in Banten is certainly different from the public ones which has already had its standardization and integrated computerized system in students’ majors. Most of private schools of SMA have not had a system and standardization in the students’ majors. Factors of majors ‘determination have not been based on students’ skill, interest, and aptitude. This present research aims at assessing two approaches in benefiting data of SMA majors. The first approach is by using K-Nearest Neighbor (K-NN) algorithm, and the second one is by using Fuzzy Tsukamoto method. From the data assessment in SMA students’ majors, both of approaches (i.e., K-Nearest Neighbor (K-NN) algorithm and Fuzzy Tsukamoto method) can be used for SMA students’ majors with a fairly-high percentage of accuracy, of which K-NN98.4% and Tsukamoto 99.2%. This points out that both methods can be used for SMA students’ majors in Banten. From those both methods, the comparison is conducted by using three factors, that is, final result, data processing, and the speed of process by looking at the condition in the field, then it is found that Fuzzy Tsukamoto method is better than K-NN method in SMA students’ majors in Banten.

1. Introduction

Nowadays, the SMA students’ majors is a routine activity which is conducted by school. The most different thing is the school’s status and its curriculum. In Banten, the SMA students’ majors for private schools has used a centralized system, that is, by using PPDB system (new students’ admission) by online at the level of SMA/SMK on the site http://siap-ppdb.com as the students’ majors is no longer as a take-home assignment for the school. However, students’ major determination at private schools is conducted annually. Based on the 14th interview result in Banten, most of major’s determinant is based on three factors. First is in line with students’ parents reference. Second, majors’ selection is based on friends’ interest and an up to date trend. The third factor is students’ own achievement. Majors’ determination that is based on those three factors will certainly make a regret for student whose majors are not in accordance with their aptitude, interest and skill against the majors. It has been conducted a research for SMA students’ majors in Banten by implementing K-Nearest Neighbor (K-NN) Algorithm and Fuzzy Tsukamoto method [1] where both methods can be implemented in SMA students’ majors in Banten. The present research aims at comparing the data sample testing result of students’ major with K-NN Algorithm and Fuzzy Tsukamoto Logika, to determine the chosen method in SMA students’ major determination in Banten. With the fuzzy logic, the human intelligent system can be implemented into machine language efficiently and easily. Fuzzy logic can be used to draw a conclusion from input data and output data [2].
2. Literature Review

2.1. Comparative Research
Comparison means to compare. According to Sudijono Anas, the comparative research is a research that tries to find out the similarities and differences of thing, person, working procedure, idea, criticism of person or group, idea and work procedure. Also, it can be used to compare the similarity of view and change of view of person, group or country towards cases, events or idea [3]. Meanwhile, Nana Syaedi Sukmadinata asserts that comparative research is directed to identify whether between two or more groups exists a difference in the aspect of variable that are overviewed [4]. This present research compares the result of SMA students’ majors concerning to K-NN algorithm and Fuzzy Tsukamoto method, then counts the percentage of its result of accuracy, whether it is suited or not from the result of manual calculation.

2.2. K-NN Algorithm
K-Nearest Neighbor Algorithm is a method to classify the objects that are based on learning data of which the distance closest to the object. Nearest Neighbor is an approach to find out cases by calculating the closeness between new cases and old cases, that is based on the matching weight of a number of existing features [5]. K-Nearest Neighbor (KNN) is considered as an instance-based learning group. This algorithm is also as one of lazy learning techniques. KNN is carried out by looking for the object’s group in data training that is the closest (similar) to the object of new data or data testing [6]. KNN is one of the methods which is used in the classification. The principle of KNN is to find out the closest distance between data that will be evaluated by using its closest K-Nearest Neighbor in training data, the following is the formula to find out the distance using Euclidian formula (1).

\[ d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \]  

Where \( x_{\sim 1} \) = data training (1) that is a saved data in the system of SMA students’ major determination in the former year, \( x_{\sim 2} \) = data testing (1) that is a new student’s data which will be majored, \( i \) = data attribute (1) is an attribute of evaluation data in the major’s determination evaluation. \( d_i \) = range (1) the range of data training and data testing and \( p \) = data dimension (1)

2.3. Fuzzy Tsukamoto Logic
The stage of data processing phase using Fuzzy Tsukamoto comprises of some stages, first, Fuzzification, is the process of changing the input system that has firm value becomes linguistics variable linguistics uses the members’ function which is saved in the Fuzzy knowledge basis. In this case, there are two evaluation groups for IPA and IPS, each evaluation consists of three input variables and one output variable. Its input variables are skill, aptitude and interest, whereas its output variable is major. Second, basis formation of Fuzzy knowledge (Rule in the form IF…THEN), that is, a general form of fuzzy Tsukamoto method (X IS A) and (Y IS B) and (Z IS C), where A, B, and C are fuzzy’s compilation. Third, forming the inference machine, that is, a process by using the function of MIN implication to obtain \( \alpha \)-predicate of each rule \( (\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_n) \). Moreover, each of this value \( \alpha \)-predicate is used to calculate the output of inference result firmly (crisp) of each rule \( (z_1, z_2, z_3, \ldots z_n) \). Fourth, defuzzification phase, by using method of Average:

\[ Z = \frac{\Sigma(\alpha_nz_n)}{\Sigma\alpha_n} \]  

Where \( Z \) (2) is as the result of major’s average, and (2) is as a-prediction of each rule and \( z_n \) (2) is as the firm value (crisps) from each rule.
3. Research Methodology
The research methodology used in this present research can as follows:

3.1. Study of Literature
This phase reviews literature from books and journals that are related to multiple intelligence by using fuzzy logic method and data mining using KNN and conducts a consultation with experts or interviewees who are competent in SMA students’ major as well as other references that can be used to complete this research.

3.2. Data Collection
This is the stage of collecting data, information, and descriptions of variables which will be used in SMA students’ major determination and required theories in the application of Fuzzy Tsukamoto logic method and K-Nearest Neighbor (KNN) method to support the decision. After reviewing literature and collecting the data either from journal, book or students’ score collecting as well as students’ interview result, then it results variables in students’ major determination, that is, skill variable (the average score of IPA and IPS), the Aptitude Variable (test of IPA and test of IPS) and the Interest Variable (interview result with students and BP teacher).

3.3. Data Processing
The result of raw data collection[1], is taken about 125 datasamples which are considered as majors’ determinant variable.

3.3.1. K-NN Method
This research uses the data of students’ major at the XI classes, either IPA or IPS, which is obtained from schools in the region of Banten. All attributes value in categories as can be seen in table II where in the data training are 7 attributes, 6 are as the prediction attribute and 1 is as the result attribute.

The weighting that is carried out is based on the interview result with the school in determining the level of importance of each criterion of value weighting which is given is to normalize the data training and data testing. The weighting criterion to the calculation of KNN algorithm.

Data training is obtained from the result of IPA and IPS score grouping, that is based on the report result for skill and psychology test result for Aptitude and Interest that is based on the student’s questionnaire result. The data training normalization is conducted to minimize the range from the conventional calculation result. Data is converted concerning to the attribute weighting value.

3.3.2. Fuzzy Tsukamoto Logic
Fuzzification. Based on the obtained data, the data separation between the evaluation of IPA and IPS, the used variables are skill, aptitude and interest. The group function of the skill and interest has the same limitation:

\[
\mu_{\text{Low Skill and } \mu_{\text{Aptitude}}} = \begin{cases} 
1 & \frac{72.5-x}{72.5-55} \\
\frac{55-x}{2} & 55 \leq x \leq 72.5 \\
0 & x \geq 72.5 
\end{cases}
\]

\[
\mu_{\text{Medium Skill and } \mu_{\text{Aptitude}}} = \begin{cases} 
1 & \frac{(x-55)}{(72.5-55)} \\
\frac{55-x}{90-x} & 55 \leq x \leq 72.5 \\
0 & x \geq 90, x \leq 55
\end{cases}
\]

\[
\mu_{\text{High Skill and } \mu_{\text{Aptitude}}} = \begin{cases} 
0 & \frac{x-72.5}{90-72.5} \\
\frac{72.5-x}{72.5} & 72.5 \leq x \leq 90 \\
1 & x \geq 90
\end{cases}
\]

The group function of Interest

\[
\mu_{\text{Low Interest}} = \begin{cases} 
1 & \frac{70-x}{70-50} \\
\frac{50-x}{50} & 50 \leq x \leq 70 \\
0 & x \geq 70
\end{cases}
\]
\[ \mu_{\text{High Interest}} = \begin{cases} 0 & \text{if } x \leq 50 \\ x - 50 & \text{if } 50 \leq x \leq 70 \\ 70 - 50 & \text{if } x \geq 70 \end{cases} \]

The rule formation of Fuzzy, from the three input variables and an output variable which has been defined, by analyzing data towards each compilation of Fuzzy’s boundary on each its variable, then 18 rules of Fuzzy are formed which will be used in this system, with the structure of rule IF Skill (K) AND Aptitude (B) AND Interest (M) THAN Result.

Inference, uses MIN implication function to obtain the score of \( \alpha \)-predicate of each rule \((\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_n)\). Then each of this value \( \alpha \)-predicate is used to calculate the output of inference result firmly (crisp) each rule \((z_1, z_2, z_3, \ldots z_n)\). Defuzzification, by using Average method (1).

3.4. Comparison
The result of data processing uses K-NN Algorithm and Fuzzy Tsukamoto Logic is carried out a comparison by comparing the percentage of the accuracy level result between manual calculation towards the method. Besides, it is conducted a comparison by seeing the condition at the field, that is, SMA in Banten.

3.5. Result
The final result points out the final result from comparison of two methods, which can be used by the schools in determining their students’ major determination.

4. Result and Discussion
4.1. Result
After having had the result of data processing which is obtained by the accuracy level of SMA determination through both methods\(^{(1)}\). From the total of sample data, 125 students are compared to manual calculation conducted by the school results the level of error as in table 1 as below:

| Table 1. The Comparison of Manual Prediction Result and Both Methods in SMA Student’s Determination |
|-----------------------------------------------|
| **Manual** | **K-NN** | **Tsukamoto** |
| Prediction of IPA | 70 | 68 | 69 |
| Prediction of IPS | 55 | 0 | 55 |

\[ X = \frac{\sum \text{XprediksiManual} - \text{XprediksiNotTrue}}{n} \times 100\% \] (3)

Results of K-NN Method:
\[ X = \frac{(70 - 2) + (55 - 0)}{125} \times 100\% = 98.4\% \]

Results of Tsukamoto Method:
\[
X = \frac{(70 - 1) + (55 - 0)}{125} \times 100\% \\
X = \frac{124}{125} \times 100\% = 99.2\%
\]

From the two methods, it releases the adequate high of accuracy percentage for K-NN with 98.4% and Tsukamoto 99.2%. This points out that both method can be used in the majors’ determination of SMA students in Banten.

4.2. Discussion
Determining the chosen method is not only based on the accuracy value, but also based on the condition at the field, by the two methods can be drawn a conclusion of the strength and weakness form those two methods concerning to final result, the data processing, and the speed of process that can be seen from table 2.

Table 2. The Comparison of Manual Prediction Result and Both Methods in SMA Student’s Determination

| Description         | Tsukamoto                                      | K-NN                                          |
|---------------------|------------------------------------------------|------------------------------------------------|
| Final Result        | Final result can directly give a certainty based on the result of evaluation of student’s skill, aptitude, and interest based on the percentage of result. | Final result is really influenced by amount of its data training, if its data training is big, then its accuracy level is also higher, besides that, the closeness range determination with data testing has not had a certainty parameter to arrange, if its data testing is good then the result will also be good, unless the decision will not be properly drawn |
| Data Processing     | Data processing that is based on several stages of process, more input variables will be more possibilities of the calculated inference result. This data processing is proper for the schools which already have or not have valid data of a certain benchmark in the determination. | New data can only be processed if there is data training considered to be a benchmark of data closeness. This data processing is really suited to the school which conducts majors routinely and has a valid data. |
| The Speed of Process| The speed of process depends on the amounts of input variable, more input variable then the complexity level of data processing is higher which influences the speed of process | The speed of process and the sufficient computation budget as it is required a range calculation from each query instance on the whole training sample. |

In accordance to the description in table V and the most proper method in SMA students’ major’s determination is by using Fuzzy Tsukamoto method, the reason is that the schools’ condition of SMA
in Banten, particularly private schools of SMA in the rural area which have not yet had good data, of which data training or evaluation benchmark standard of student’s majors. Besides, the interest of SMA students is still low that causes the least amounts of students. Another factor is the variable of flexibility level which can be adjusted by each school’s boundaries as it can be applied at schools with different evaluation boundary standard.

5. Conclusion and Suggestion
In line with the research result, some conclusions drawn as below:

a. Tsukamoto and K-NN method can be implemented as the decision support in SMA students’ majors determination that is based on students’ skill, interest and aptitude. Tsukamoto method employs majors’ determination by calculating a percentage of majors’ recommendation concerning to the Fuzzy logic. K-NN method uses method that is based on the proximity distance between data testing to data training with the parameter of value $k = 5$. Where the accuracy level is 98,4% for K-NN and 99,2% for Fuzzy TsukamotoLogic.

b. Tsukamoto method is the chosen method in SMA students’ majors determination in the region of Banten, this comparative result is based on the schools’ condition of SMA in the region of Banten that are not equal, either from evaluation standard, students’ skill and their amounts of students.

c. It is required a more-depth study in determining the chosen method, as it can be applied not only in the region of Banten but also in other regions for it is only properly applied and is limited in the region of Banten.

Acknowledgement
Best gratitude to Kemenristek Dikti and LPPM of Indraprasta PGRI University both for the material and spiritual supports as this research could be well completed.

References
[1] Ana Rusmardiana, at all 2017 Analisis Penentuan Jurusan Sekolah Menengah Atas Berdasarkan Metode Fuzzy Tsukamoto dan Algoritma K-Nearest Neighor (K-NN) The 6th University Research Colloquium 2017 Universitas Muhammadiyah Magelang Proceeding 122
[2] Andik S, Siti N A. 2018 Development of Decision Support System for Ordering Goods using Fuzzy Tsukamoto International Journal of Electrical and Computer Engineering (IJECE)81182
[3] Anas S 2009 Pengantar Statistik Pendidikan (Jakarata: Rajawali Pers) p 194
[4] Sukmadinata Nana S 2005 Metode Penelitian Pendidikan (Bandung : Remaja)
[5] Rosdakarya.Turban, Efraim, Jay E. Aronson, Ting Peng Liang 2005 Decision Support Systems and Intellegent System (7th ed.) (New Jersey : Prentice – Hall)
[6] Santosa and Budi 2007 Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. (Yogyakarta: Graha Ilmu)
[7] Kusumadewi and S Purnomo, H 2004 Aplikasi Logika Fuzzy untuk Pendukung Keputusan, (Yogyakarta: Graha Ilmu)