Application of decision support system using the K-Nearest Neighbor and Weighted Product method for determining the recipients of low-income family scholarship (GAKIN) (case study: Poltekkes Kemenkes Semarang)

L A Nasher, N Bahtiar

Department of Computer Science, Faculty of Science and Mathematics, Diponegoro University Jalan Prof. H. Soedarto, SH. Tembalang Semarang 50275
E-mail: luthfiahmadnasher@student.undip.ac.id

Abstract. Poltekkes Kemenkes Semarang is a university of health that has a scholarship facility for underprivileged students or often referred as Beasiswa Keluarga Miskin (Gakin). At this time the determination of Gakin scholarship recipients is still in the form of manual selection. Decision-Making is a systematic approach to a problem with the gathering of facts, a mature determination of the alternatives to be faced, and the taking of actions that are calculated to be the most appropriate action. To simplify the process of determining a decision, present an application of decision support system with input and output required. This research makes an application of decision support system with K-Nearest Neighbor method to recommend the decision in the system. It aims to shorten the time and provide convenience to policymakers (decision makers) in Poltekkes Kemenkes Semarang in selecting recipients of Poor Family Scholarships (GAKIN). Based on existing data, KNN method can handle all new data from potential recipients with a value of neighbor is 5 to produce an accuracy of 99.61%, precision 99.85% and recall 99.42% with data of scholarship recipients last five years.

1. Introduction

Politeknik Kemenkes Semarang provides one special scholarship for low-income families. This scholarship is called the Gakin Scholarship (Low-Income Family). This scholarship category is sourced from the Semarang Ministry of Health Polytechnic Budget Implementation List (DIPA) every year. The students who receive this scholarship can be started when they enter the Semarang Ministry of Health Poltekkes as new students or when they become old students with certain conditions [1].

At present, Semarang Health Ministry Poltekkes conducts the selection of Gakin Scholarship recipients by selecting files without the help of an integrated application between one department and another. With the help of a decision support system application, it will increase the ease of selection and increase efficiency. In this case, it will shorten the stage of determining students who receive scholarships without losing the essence of selection as usual. Apart from that, applications like this also increase objectivity in the selection of Gakin scholarships.

As for making this decision support system application using two methods, first is K-Nearest Neighbor (KNN) and Weighted Product (WP). KNN is a method for classifying objects based on learning data that are the closest to the object. Learning data is projected to a large dimension of space, where each dimension represents the features of the data. This space is divided into sections based on the classification of learning data. The best k value for this algorithm depends on the data, in general, a
high k-value will reduce the effect of noise on the classification, but make the boundaries between each classification more blurred [2]. According to several journals entitled "Penerapan Metode K–Nearest Neighbor (KNN) dan Metode Weighted Product (WP) Dalam Penerimaan Calon Guru Dan Karyawan Tata Usaha Baru Berwawasan Teknologi (Studi kasus : Sekolah Menengah Kejuruan Muhammadiyah 2 Kediri)” by Nihru Nafi ’Dzikrulloh and friends, stated that the research using the KNN method on a simple data pattern was enough to overcome the problems that emerged so that the KNN method was taken to overcome the case of Gakins Scholarship acceptance in the Health Ministry Poltekes Semarang. Weighted Product method is used to adjust and complete the results of the classification selection from the previous method in minimizing uncertainty. This method can also determine the collection of classification results by the KNN method by ranking so that the best results can be taken. WP can be determined by assigning weight to each criterion to be ranked. So it can be expected to get a more accurate selection [2]. In this case, the WP method is used to rank the data collection that has been processed with KNN.

Based on the explanation of the problem and the basis for the method adopted, in this final project, an Application of Support System for the Decision on Gakins Scholarship with KNN and Weighted Product in the Ministry of Health Poltekes Semarang was built.

2. Literature review

2.1. K-Nearest Neighbor Method

The K-Nearest Neighbor method is one of the simplest classifiers. By memorizing all training data and classifying only if the attributes of the test data are by one example of training data. An obvious weakness of this approach is that many test records will not be classified because they do not match one of the training records. A more sophisticated approach is to find a group of objects k (the value of proximity to neighbors) in the training set that is closest to the test object, and base label assignments on the dominance of certain classes in this environment. So that by looking at the dominance of a particular class will become a new label on the new test data [3].

KNN classification is a classification technique that is easy to understand and easy to implement. Despite its simplicity, it can perform well in many situations. In particular, the important results by Cover and Hart in their journals indicate that the errors of the rules of the nearest neighbor are limited above by two Bayes errors under certain assumptions that still make sense [3].

The KNN method algorithm is straightforward, works based on the shortest distance from the instance query to the training sample to determine the KNN. The training sample is projected to a multi-dimensional space, where each dimension represents the features of the data. This space is divided into sections based on the training sample classification. A point in this space is marked as c if class c is the most common classification found in the k of the nearest neighbor from that point.

The following is the algorithm stage of K-Nearest Neighbor) according to Xindong Wu in a paper entitled “Top 10 algorithm in data mining”: 

- a. Enter training data and test data
- b. Determine the parameter of “k” (the closest number of neighbors)
- c. Calculate the square of the Euclidean distance of the object against the given training data.
- d. Sort the results of the second stage in ascending order (sequentially from low to high)
- e. Collect categories from labels (nearest neighbor classification based on k value)
- f. By using the nearest neighbor category that is the most dominating, it is predicted as an object label.

Euclidean Distance is most commonly used to calculate the similarity of two vectors. The Euclidean Distance formula is the root of the 2 vector difference square (the root of square differences between 2 vectors) (Putra, 2010).

Where:

\[ D(a,b) = \sqrt{\sum_{k=1}^{d} (a_k - b_k)^2} \]  

1

\[ D(a,b) : \text{Distance of matrix } a \text{ to matrix } b \]

\[ k : \text{data index} \]

\[ d : \text{number of test data} \]
2.2. Weighted Product Method

Weighted Product Method (WP) is a method of decision making by multiplying to connect attribute values, where the value of each attribute must be raised first with the weight of the attribute in question. WP is one of the most popular multi-criteria decision analysis (MCDA). The MCDA provided is a limited set of alternative decisions that are explained regarding some decision criteria. Each alternative decision is compared with the other by multiplying the number of ratios, one for each decision criterion. Each ratio is raised to the power equivalent to the relative weight of the appropriate criteria [4].

WP requires a normalization process because this method multiplies the results of the assessment of each attribute. The multiplication results have not been meaningful if they have not been compared (shared) with the standard values. Weights for benefit attributes function as positive ranks in the multiplication process, while cost weights function as negative ranks. The steps in calculating the WP method are as follows:

1. Multiplying all attributes for all alternatives with a weight as a positive rank for the cost attribute.
2. The multiplication results are summed to produce values for each alternative.
3. Divide the value of V for each alternative with the value for each alternative.
4. Found the best alternative order that will be the decision.

Preferences for alternative indexes \( A_i \) are given as follows:

\[
w_{last_j} = \frac{w_{first_j}}{\sum w_{first_j}} \tag{2}\]

\[
S_i = \prod_{j=1}^{n} X_{ij}^{w_{last_j}} \tag{3}\]

Where:

- \( w_{first} \): initial weighting criteria
- \( w_{last} \): final criteria weight value (after normalization)
- \( j \): data index
- \( S \): alternative preferences are analogous as \( S \) vector
- \( n \): number of criteria
- \( X \): value of criteria
- \( i \): alternative index
- \( j \): criteria index

The relative preferences of each alternative are given as follows:

\[
V_i = \frac{\prod_{j=1}^{n} X_{ij}^{w_{last_j}}}{\prod_{j=1}^{n} X_{j,i}} \tag{4}\]

Where:

- \( V \): alternative preferences are analogous as \( V \) vector
- \( X \): the value of criteria.
- \( w \): weight of criteria
- \( i \): alternative index
- \( j \): criteria index
- \( n \): number of criteria

Simply interpreted as follows:

\[
V_i = \frac{S_i}{\sum S_i} \tag{5}\]

Where:

- \( V \): alternative preferences are analogous as \( V \) vector
- \( S \): value of \( S \) vector.
3. Analysis and design

3.1. Data Analysis

The data used in making the application is the last five years data from the recipient of the Gakin scholarship. In this case, the received data can be interpreted from 2013 to 2017. The names and majors of scholarship recipients are not permitted by the Semarang Health Ministry Poltekkes to be displayed because the data is not publicly available, so the author changes the name and department to "anonymous" according to the number of recipients available for the past five years.

Overall data from Gakin Scholarship recipients from 26 majors are detailed as follows:

a. In 2013 there were 276 registrants with as many as 150 students passed and not qualified as many as 126 students.
b. In 2014 as many as 272 registrants with pass details as many as 129 students and did not qualify as many as 143 students.
c. In 2015, there were 131 registrants with as many as 79 students passed and did not qualify as many as 52 students.
d. In 2016 as many as 282 registrants with details passed as many as 186 students and did not qualify as many as 96 students.
e. In 2017 there were 304 registrants with as many as 146 students passed and did not qualify as many as 158 students.

Based on the data above it was concluded that all scholarship recipient data from the last five years were 1265 students with details of 690 students having passed status and 575 students who were not qualified.

3.2. Application Description

Application of Decision Support System for Accepting Gakin Scholarship with KNN and WP methods in Poltekkes Kemenkes Semarang is one of the web-based software that functions as a helper in determining Gakin Scholarship recipients with the help of KNN and WP calculations. The calculation is used to sort the recommendations of prospective applicants from those who are most eligible to get scholarships until the least eligible to get a scholarship. The calculation is also based on new data entered by registrants through the application (web-based), by comparing the data of past scholarship recipients as training data.

The specifications of this application are based on functional needs and non-functional needs.

| NO | Explanation |
|----|-------------|
| 1  | Login and Logout |
| 2  | Management (editing and viewing) of Student Data conducted by Admin |
| 3  | Management (view, process registrant data, change status) Registrar data with Admin access rights |
| 4  | Management (add, view, delete and edit) Announcement Data with Admin permissions |
| 5  | Search data for students, registrants, and recipients with Admin permissions |
| 6  | Management (adding and editing) Data Officers with admin access rights |
| 7  | Print feature results with admin access rights |
| 8  | Management (registering, editing and viewing) Student data with student access rights |
| 9  | Manage (view) Announcement Data with student access rights |
| 10 | New account addition feature for students who do not have an account |

3.3. Modeling Analysis

Modeling analysis aims to find out what is needed in building an application. One of them used in this application is structured analysis [5].
Structured analysis is an analysis activity in the model building. The model was created to describe the flow of a load of information in the form of data and controls. Application analysis functions to find errors or weaknesses of an application so that solutions can be found. Application analysis is done after application planning and before application design. It is done to reduce application errors at a later stage.

The functional model is described through a data flow diagram that can show the dependence between processes, inputs, and outputs. The media used to describe the functional model of the Application of Support System for Decision on Acceptance of Poltekkes Gakin Scholarship Semarang is Data Context Diagram (DCD) and Data Flow Diagram (DFD).

![Data Context Diagram](image.png)

**Figure 1.** Data Context Diagram

From the picture above there are two external entities, namely students and admin. The DCD has one main process called **SPKBEAGAKIN** which in the main process there are other small processes in the decomposition section.
There are four decomposition process results from the main processes contained in DCD, namely the Login process, Student Management, Announcement Management, and Registrant Management.

3.4. Application Design
Application Design of Decision Support System for Gakin Scholarship Acceptance includes database design and algorithm design. The database contained in this application consists of five databases, namely student data, registrant data, announcement data, officer data, and WP data. These data are set using MySQL.

The algorithm for this application is explained in the following flowchart:
Figure 3. Flowchart of Application of Decision Support System for Determining Gakin Scholarship Recipients

Overall in determining the status of registrants as scholarship recipients is described in the flowchart above. The first step is to retrieve the data of all registrants in the year corresponding to the current year. Then repeated some new data to undergo the KNN process calculation.

After processing KNN, a recommendation appears according to the results of the calculation and is stored in the database. Then check whether all data already has recommendations, if not then return to the data retrieval process, if so then proceed with the WP process to get the sequence based on V Vector values v. After the WP calculation is done, the data sequence of the registrant with two parts appears, the first is the highest to the lowest v with the recommendation to pass and the second is the highest to the lowest v with the recommendation not to pass.

Then proceed with the process of determining the final status of the registrant, whether the status value will be taken the same as the recommendation given or different from the recommendation given due to other considerations from the decision maker.

The KNN algorithm and WP algorithm for this application will be explained in the following flowchart:
The first step is to take registration data this year and the last 5 years data. Then the status of each new registrant is determined by finding the neighbor’s value with $k = 5$ from the closest distance to Euclidean Distance. Then came the recommendations of each new registrant.

The next step is to take registrant data from new registrants this year. Then the initial weighting is determined for each criterion. The initial weight is as follows:
Table 2. Initial Weight Value

| No | Criteria            | Initial Weight Value |
|----|---------------------|----------------------|
| 1. | Wall type           | 5                    |
| 2. | Floor type          | 4                    |
| 3. | Clean water         | 4                    |
| 4. | Parent Education    | 4                    |
| 5. | Building area       | 3                    |
| 6. | Investation         | 2                    |
| 7. | Kitchen Fuel        | 2                    |
| 8. | Toilet              | 1                    |
| 9. | Lighting            | 1                    |
| 10.| Chicken Per Week    | 1                    |
| 11.| Clothes per year    | 1                    |
| 12.| Puskesmas fees      | 1                    |
| 13.| Eat Per Day         | 1                    |

After determining the initial weight, then the weight normalization process by giving the initial weight value of each criterion with the sum of all initial weight values.

Table 3. Final Weight Value

| No | Criteria            | Final Weight Value |
|----|---------------------|--------------------|
| 1. | Wall type           | 0.16667            |
| 2. | Floor type          | 0.13333            |
| 3. | Clean water         | 0.13333            |
| 4. | Parent Education    | 0.13333            |
| 5. | Building area       | 0.10000            |
| 6. | Investation         | 0.06667            |
| 7. | Kitchen Fuel        | 0.06667            |
| 8. | Toilet              | 0.03333            |
| 9. | Lighting            | 0.03333            |
| 10.| Chicken Per Week    | 0.03333            |
| 11.| Clothes per year    | 0.03333            |
| 12.| Puskesmas fees      | 0.03333            |
| 13.| Eat Per Day         | 0.03333            |

After the normalization process ends, proceed with the calculation to get the S Vector value from each new data.

Then check whether all new data already has S Vector value, if not, then calculate the S Vector value again, if all the S Vector values are added and stored into the new "sumS" variable which will be used to calculate V Vector Value.

Then proceed with the calculation of V Vector Value. The V Vector Value is the normalization of the S Vector Value, with the calculation of the S Vector Value of each new data divided by the total number of S Vector values.

After calculating the V V Vector value is complete, then it is checked first whether all data already has V Vector Value. If not then it will return to the calculation of V Vector Value and if it is then S Vector Value and V V Vector value are entered into the database according to the registrant's id appropriate.

When all data has been entered into the database, the results of the sequence of prospective registrants are issued by V Vector Value. The ordering of applicant candidates is made into two parts, the first is
the highest to lowest V Vectors with recommendations to pass and the second is the highest V Vectors to the lowest with recommendations not qualify. This result is used to overcome the quota if it is not sufficient by looking at the eligibility of registrants from the value of V Vector.

4. Result and Discussion
5. Implementation
Following is the process of calculating the determination of scholarship recipients:

Table 4. Registered Data Tables

| Criteria /ID | 1266 | 1267 | 1268 | 1269 | 1270 |
|-------------|------|------|------|------|------|
| A           | 5    | 6    | 4    | 6    | 6    |
| B           | 2    | 4    | 1    | 4    | 2    |
| C           | 1    | 2    | 1    | 1    | 1    |
| D           | 1    | 3    | 1    | 2    | 1    |
| E           | 1    | 2    | 2    | 2    | 1    |
| F           | 2    | 2    | 4    | 4    | 2    |
| G           | 1    | 3    | 1    | 4    | 2    |
| H           | 1    | 3    | 3    | 1    | 2    |
| I           | 1    | 4    | 1    | 3    | 1    |
| J           | 2    | 2    | 5    | 2    | 3    |
| K           | 1    | 3    | 1    | 4    | 2    |
| L           | 1    | 1    | 2    | 1    | 2    |
| M           | 1    | 4    | 1    | 5    | 2    |
| N           | 1    | 2    | 4    | 4    | 2    |
| Status      | BD   | BD   | BD   | BD   | BD   |

Where for each code above means the following variables:

Table 5. Determinant Variables

| NO | Code | Variable                                                  |
|----|------|-----------------------------------------------------------|
| 1  | A    | Source of Income for the Head of the Family               |
| 2  | B    | Floor type                                                |
| 3  | C    | Toilet Facilities                                         |
| 4  | D    | Investment Ownership                                     |
| 5  | E    | Source of lighting                                       |
| 6  | F    | Kitchen Fuel                                              |
| 7  | G    | The frequency of meat/chicken per week                    |
| 8  | H    | Building area                                             |
| 9  | I    | Wall type                                                 |
| 10 | J    | Frequency Buy clothes in a year                           |
| 11 | K    | Source of Clean Water                                    |
| 12 | L    | Ability to Pay Health Center Health Costs                 |
| 13 | M    | The Highest Education of the Family Head                  |
| 14 | N    | The frequency of Eating in a Day                          |
| 15 | BD   | Unprocessed                                               |

In this case, variable A is not used in the calculation because the variable has rules if it is worth (TNI = 2, POLRI = 3, PNS = 4) then it is confirmed: "NOT QUALIFIED".

Following are the recommendations with the KNN calculation using the value of 5:
Table 6. Recommendation Results with KNN

| NO | ID   | Status      |
|----|------|-------------|
| 1  | 1266 | “TIDAK LOLOS” |
| 2  | 1267 | “LOLOS”     |
| 3  | 1268 | “TIDAK LOLOS” |
| 4  | 1269 | “LOLOS”     |
| 5  | 1270 | “TIDAK LOLOS” |

The following are the calculation results from the data above:

Table 7. The calculation results

| NO | ID   | S Value | V Value | Status      |
|----|------|---------|---------|-------------|
| 1  | 1266 | 1.176   | 0.119   | “TIDAK LOLOS” |
| 2  | 1267 | 3.021   | 0.305   | “LOLOS”     |
| 3  | 1268 | 1.417   | 0.143   | “TIDAK LOLOS” |
| 4  | 1269 | 2.655   | 0.268   | “LOLOS”     |
| 5  | 1270 | 1.647   | 0.167   | “TIDAK LOLOS” |

Based on the results of the above calculations, the level of eligibility to register for a scholarship is sorted:

Table 8. Results of Ordering Feasibility in Getting Scholarships

| NO | ID | Vector V    | Recommendation |
|----|----|-------------|----------------|
| 1  | 1267 | 0.304682   | “LOLOS”       |
| 2  | 1269 | 0.267796   | “LOLOS”       |
| 3  | 1270 | 0.166073   | “TIDAK LOLOS” |
| 4  | 1268 | 0.142887   | “TIDAK LOLOS” |
| 5  | 1266 | 0.118563   | “TIDAK LOLOS” |

6. Testing

Application of Support System for Decision on Acceptance of Gakin Scholarships Poltekkes Semarang has conducted some tests to test the appropriateness of the application. The test results have been received so that it can be concluded that this information system has fulfilled all the functional requirements agreed upon earlier.

By comparing all scholarship recipients data in the last five years with the determination of system recommendations using Confusion Matrix, the following results are obtained in Table 9. Based on the results of the comparison above obtained precision of 99.85%, recall of 99.42% and accuracy of 99.61%.
Table 9. Results of All Data Confusion Matrix

|                | All Data     |
|----------------|--------------|
| Accuracy       | 99.61%       |
| Correct        |              |
| Classification |              |
|                | +            |
|                | -            |
| Classified As  |              |
|                | +            | 687 | 99.85% |
|                | -            | 4   | 99.31% |
| Class Precision|              |
|                | 99.42%       |
|                | 99.83%       |

7. Conclusion

The conclusions that can be taken in the execution of the final project Application of Decision Support System for the Determination of GAKIN Scholarship Acceptance is the application of Support System for Decision on Gakin Scholarship Acceptance as intended by the Semarang Health Ministry Poltekkes. Based on the comparison of the data of the last five years with the results of system calculations obtained precision of 99.85%, recall of 99.42% and accuracy of 99.61%. If there is a new acceptance variable, it is recommended to add a variable addition feature. Several decision-making variables have little influence on the calculation so that it is necessary to evaluate the decision-making variables for scholarship acceptance the following year.

Reference

[1] Semarang P K 2012 Sejarah Politeknik Kesehatan Kemenkes Semarang (Semarang: s.n).
[2] Dzikrulloh N, Indriati and Setiawan B 2017 Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 1 378.
[3] Wu X, Kumar V, Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan G, Angus, Liu B, Yu P, Zhou Z, Steinbach M, Hand D, Steinberg D 2008 Knowl. Inf. Syst 14 1
[4] Basri, 2017. Metode Weighted Product (Wp) Dalam Sistem Pendukung Keputusan Penerimaan Beasiswa Prestasi. (Article)
[5] Kendall K and Kendall J 2003 Analisis dan Perancangan Sistem (Jakarta: PT. Prenhallindo).