Implementation of the K-Nearest Neighbor Method to determine the Classification of the Study Program Operational Budget in Higher Education

1st Gufron
Information System
(Diponegoro University)
Semarang, Indonesia
ghufronasadly@gmail.com

2nd Bayu Surarso
Mathematics
(Diponegoro University)
Semarang, Indonesia
bayusurarso@yahoo.com

3rd Rahmat Gernowo
Physics
(Diponegoro University)
Semarang, Indonesia
gernowo@yahoo.com

***Abstract*— Sultan Agung Islamic University annually designs operational costs for work programs in the study program and of course determines the amount of financial budget for the study program work program. K-Nearest Neighbor algorithm is needed to determine the required operational budget classification based on the number of active students, the number of financial admissions, the number of employees and the percentage of work program realization programs. The results of this study are to facilitate the leadership of higher education in the budget field to classify the amount of the budget required by study programs in the classification of up, or fixed. The purpose of this study is expected to facilitate the leadership of the financial budget department to classify the budget needed by the study program and as an awareness system in the work program of the study program with a classification value of 79.96% for the operational budget of the college study program.**

Keywords*— *Operational Finance, K-Nearest Neighbor, Study Program.

I. INTRODUCTION

Every college must have a budget for all study programs, the type of budget certainly has differences that are budgeted, namely the work program budget increases, stays or even falls[1]. To anticipate that the work program's financial budget is right on target according to the study program, a system is needed to make decisions so that the distribution of the budget can be on target, on time, and on the right amount. the development of information systems using the k-nearest neighbor method can be applied to classify data into one or several classes that have been defined, the use of classification values can produce finer decisions and provide probabilistic information [2]

The use of the right algorithm can improve the accuracy of the decisions taken. The k-nearest neighbor algorithm classification method is one of the better methods of classifying data and predictions, with a fast way of calculating accuracy and lightness in computing[3]. This algorithm is more effective in training large data and can produce more accurate data. Research concepts with the K-nearest neighbor algorithm have been carried out by many previous researchers including, namely, a single fault location scheme for parallel transmission channels using the k-nearest neighbor algorithm [4]. K-nearest neighbor is able to conduct training on the automatic feature learning dataset for monitoring nonlinear processes, an approach using auto-encoder denoising[5]. K-nearest neighbor is used to be able to reconstruct the position of control rods accurately, and modification strategies based on calibration factors used to improve the accuracy of monitoring the position of the rod when there is a mismatch between the actual physical factors and the physical factors modes [2]. The problem that often arises is the determination of budget costs in study programs that are not on target, so we need a system that can classify the amount of work program budgets in the study program based on training data taken from previous annual budget data (dataset). So that the university can overcome these problems early on. The use of data mining techniques with the k-nearest neighbor algorithm is expected to be able to provide useful information about the classification of budgeting techniques in the study program. The purpose of this study is to implement the k-nearest neighbor algorithm to support budget classification decisions in the study program. Build a decision support system application that can classify the amount of the study program work program budget using the k-nearest neighbor algorithm.

A. Data mining

Data Mining is a process that uses statistical techniques, mathematics, artificial intelligence, machine learning to extract and identify useful information and related knowledge from various large databases [6].

Data classification based on data expression is quite promising in the field of research in the field of data mining. two simple and efficient data classification techniques based on KNN are presented that are appropriate for high dimensional data. These new techniques use a new weighting strategy based on KNN. In experiments, using six benchmark datasets that are often used by researchers for high dimensional classification problems [7].

B. Decision-Making Process

Decision making in higher education is the result of policy agreements and the results of mutually agreed regulations. The outcome of the decision can be a statement agreed upon as an alternative or procedure to achieve certain objectives. The approach can be done through an approach above or below meaning that from the leadership level of subordinates or from the subordinate level to the leadership level. Decision making, basically proposing the various alternative actions chosen, the process through a mechanism, with the hope of producing the best decision.

C. Classification Methods

The classification method is the process of finding a model (function) that describes and distinguishes classes of

data or concepts that aim to be used to predict classes from objects whose class labels are unknown [8]. Classification is part of data mining, where data mining is a term used to describe new knowledge or information in a database. Data mining also uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various big data [9]..

The classification process is based on four components:

1. Class

The dependent variable is in the form of a categorical that represents the 'label' contained in the object. For example risk of heart disease, credit risk.

2. Predictor

The independent variable is represented by the characteristics (attributes) of the data.

3. Training the dataset

A data set containing the values of the two components above is used to determine the suitable class based on predictors.

4. Testing the dataset

Contains new data to be classified by the model that has been made and classification accuracy

D. K-Nearest Neighbor Algorithm

In the k-neighbor paper to improve the accuracy of cluster-based location estimates from wireless nods the k-Nearest Neighbor (k-NN) algorithm is one of the easiest and simplest machine learning algorithms. The KNN algorithm uses all data elements for training. The data selection criteria are based on the closest data set from the focus point. The k value varies based on the data set of elements available for training. Each data element behaves as a center point until the frequency of events is obtained for all data sets. The weights are then multiplied by the frequency to calculate the weighted average. The weighted average leads to the estimated position of the data set [10]

K-nearest neighbor is a method for classifying objects based on learning data that is the closest distance to the object. Learning data is projected into multi-dimensional 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 reduces the effect of noise on the classification but makes the boundaries between each classification more blurred. There are many ways to measure the proximity between new data and old data (training data), including Euclidean distance and Manhattan distance (city block distance), the most commonly used is Euclidean distance [11].

that is:

$$\mathbf{d(x,y)} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{m}} (\mathbf{x_i} - \mathbf{y_i})^2}$$

To weight variables using the pair comparison method, the weight of the variable is determined by normalizing the eigenvector, which is associated with the maximum eigenvalue in a ratio matrix before comparing the scale of influence or significance values between variables.

E. Financial Budget Study Program

The study program is an integrated learning plan as a guideline for organizing academic or professional education which is organized on the basis of a curriculum and is intended so that students can master knowledge, skills, and attitudes in accordance with curriculum objectives, the Entrepreneurial Hybrid curriculum can be applied to academics who can combine entrepreneurial orientation with academic values and norms of science, budgetary participation has a negative effect on budgetary slack Organizational Commitment reinforces the negative influence of budgetary participation on budgetary slack. Love of money does not moderate the effect of budgetary participation on budgetary slack within the scope of work units [1].

F. Evaluation

In the evaluation step, the performance of the classifier is calculated using a confusion matrix [12] Accuracy is a popular matrix. This refers to the ability of the model to correctly predict class labels, and that is proportions. defined as follows:

$$accuracy = \frac{\sum_{t=1}^{n} N_{tt}}{\sum_{n=1}^{n} \sum_{t=1}^{n} N_{tj}}$$

Table 3. Confusion Matrix Model

Classification True	Classification as	
	+	-
+	*True positive*	*False negative*
-	False Positiv	True Negative

Table description:

1. True positives are the number of positive records which is classified as positive,
2. False positives are the number of negative records which is classified as positive,
3. False negatives are the number of positive records which is classified as negative,
4. True negatives are negative record numbers which is classified as negative.

II. METHOD

The steps in this procedure can be explained as follows

A. Problem Identification and Formulation

At this stage, the identification of parameters or attributes that can be used to measure the determination of the budget classification of the study program work program. This stage can also determine the analysis model that will be used in determining the classification of prospective scholarship recipients both quantitatively and qualitatively for decision

support systems. The problem formulation is made after identifying the problem.

B. Research purposes

This stage is used to develop information systems that are able to support decision making in classifying study program work program budgets using the k-nearest neighbor algorithm.

C. Literature Study

At this stage, the process of collecting literature such as international and national journals in accordance with the research topic, also in the form of articles relevant to the research topic.

D. Field Study

The field study was conducted at Sultan Agung Islamic University in Semarang by observing and interviewing the speakers (the finance department). Field studies were conducted to obtain data relating to the research theme in the form of student data in the work program study program. From the observation, results found some study program budget work data in the 2018 budget year. Study programs with an increase in 13 study programs while 11 study programs remain,

E. Determination of Variables

At this stage, we will determine what variables/parameters are used in data processing on the system that will be built with the k-nearest neighbor algorithm. These variables will be obtained from the study program's budget history data which are in the form of Number of Active Students, Total Cost of Acceptance (UKT), Number of Lecturers, Percentage of Work Program Realization (%). Determination of variables is done by selecting the attributes that influence the process of determining the cost of the study program budget, giving parameter values / variables adjusted to the raw data (training data). Attributes/variables and variable values can be seen in

Table 2 Rating Variables

No	Variabel	Value of Variabel	
1	Number of Active Students Total Cost Receipt (UKT)	1	<500
		2	>500
2	Number of Lecturers	1	<200
		2	>200
3	Number of Active Students Total Cost Receipt (UKT)	1	<=6
		2	>6
4	Number of Lecturers	1	<95
		2	>95

F. Determination of Variable Weight

To measure the distance between attributes, the weighting of the attributes will be carried out. The attribute distance weights are given values between 0 to 1, the variable weighting is done by the method of comparing pairs of variables with the scale specified in Tables 3 and 4.

Table 3. The weighting of the Increased Budget Variables

1	Number of Active Students	0,32
2	Total Cost Receipt (UKT)	0,21
3	Number of Lecturers	0,25
4	Percentage of Work Program Realization (%)	0,21

Table 4. The weighting of Budget Variables Down

1	Number of Active Students	0,32
2	Total Cost Receipt (UKT)	0,21
3	Number of Lecturers	0,27
4	Percentage of Work Program Realization (%)	0,2

III. RESULT

This study uses 23 training data consisting of 13 data from the increased budget study program, and 11 fixed budget study program data. To find out the results of the analysis of the k-nearest neighbor algorithm, a manual calculation is performed using sample data as follows:

Training data:

1. Number of Active Students = 224, Total Cost of Acceptance (UKT) = Rp220,500,000 Number of Lecturers = 7, Percentage of Work Program Realization (%) = 100 (Budget Increase)

2. Number of Active Students = 252, Total Cost Revenue (UKT) = Rp 639,325,000 Number of Lecturers = 9, Percentage of Realization of Work Programs (%) = 97 (Fixed Budget)

3. Number of Active Students = 1893, Total Cost Acceptance (UKT) = Rp 1,039,500,000 Number of Lecturers = 21, Percentage of Realization of Work Programs (%) = 97 (Fixed Budget)

Testing data as follows:

Number of Active Students = 350, Total Cost of Acceptance (UKT) = Rp 200,500,000 Number of Lecturers = 6, Percentage of Realization of Work Programs (%) = 95

To calculate the closeness of a case between training data and testing data above, it is used

equation (3). To find out whether the testing data is included in the Upward budget classification or a fixed budget, the following steps can be taken [4]

1. Calculating the closeness between new data and data number 1

. Number of Active Students = 350 and 224, Total Cost of Acceptance (UKT) = Rp 200,500,000 and Rp 220,500.00 Number of Lecturers = 6 and 7, Percentage of Work Program Realization (%) = 97 and 100

2. Calculating the closeness between new data and data number 2

. Number of Active Students = 350 and 252, Total Cost Acceptance (UKT) = Rp. 200,500,000 and Rp. 639,325,000 Number of Lecturers = 6 and 9, Percentage of Work Program Realization (%) = 97 and 97

3. Calculating the closeness between new data and data number 3

. Number of Active Students = 350 and 1893, Total Cost Acceptance (UKT) = Rp 200,500,000 and Rp 1,039,500,000 Number of Lecturers = 6 and 21, Percentage of Realization of Work Programs (%) = 97 and 97

4. Choose the case with the closest proximity. From steps 1, 2 and 3 it can be seen that the lowest value is case number 2. This means the closest case to the new case is case number 2 in the training data.

5. Using the classification of cases with the closest proximity. Based on the results

obtained in step 3, the classification of 3 cases will be used to classify new cases. Namely the possibility of study programs will get a fixed budget.

A. Verification of Ms. Calculation Results Excel and System

Calculation of the proximity of old cases in training data with new cases in testing data, it is known from 20 data records, 9 classified as an increased budget, 15 data classified as fixed, 20 data classes are not classified accordingly. The accuracy level of the application of the k-nearest neighbor algorithm is 85.56%, 90.57% precision and 92.90% recall.

B. Algorithm Testing

Calculation of the proximity of old cases in training data with new cases in testing data, it is known from 20 data records, 9 classified as an increased budget, 15 data classified as fixed, 20 data classes are not classified accordingly. The accuracy level of the application of the k-nearest neighbor algorithm is 85.56%, 90.57% precision, and 92.90% recall.

IV. CONCLUSION

This research was conducted by implementing the k-nearest neighbor algorithm in the study program's work budget data. To find quality data, preprocessing is carried out before applying it to the algorithm. The closeness between new cases and old cases is done to determine which class the new cases will be classified in. By building a decision support system to classify the work budget of a study program that is determined or analyzed using the k-nearest neighbor algorithm. There are four variables used, namely the Number of Active Students, Number of Acceptance Fees (UKT), Number of Lecturers, Percentage of Realization of Work Programs (%). Verification test results show that the decision support system made using the k-nearest neighbor analysis algorithm produces the same output as the manual calculations performed with Microsoft Excel, where the system output is the lowest value used as the closest case to classify new cases as the results Microsoft Excel Calculation To measure the performance of the k-nearest neighbor algorithm, the Confusion Matrix method is used in this study using a 10-fold cross-validation for the classification of study program budgets. Of the 12 data sets recorded in the work program study program budget reached 77.96% and included in the classification is very good.

REFERENCES

[1] I. G. P. Pundarika and A. A. N. . Dwirandra, "The Effect of Budget Participation on Budgetary Slack with Organizational Commitments and Love of Money as Moderation," *Int. J. Sci. Res.*, vol. 8, no. 2, pp. 491–496, 2019.

[2] X. Peng, Y. Cai, Q. Li, and K. Wang, "Control rod position reconstruction based on K-Nearest Neighbor Method," *Ann. Nucl. Energy*, vol. 102, pp. 231–235, 2017.

[3] T. Sathish, S. Rangarajan, A. Muthuram, and R. P. Kumar, "Analysis and modelling of dissimilar materials welding based on K-nearest neighbour predictor," *Mater. Today Proc.*, no. xxxx, 2019.

[4] A. Swetapadma and A. Yadav, "A novel single-ended fault location scheme for parallel transmission lines using k-nearest neighbor algorithm," *Comput. Electr. Eng.*, vol. 69, no. April 2017, pp. 41–53, 2018.

[5] Z. Zhang, T. Jiang, S. Li, and Y. Yang, "Automated feature learning for nonlinear process monitoring – An approach using stacked denoising autoencoder and k-nearest neighbor rule," *J. Process Control*, vol. 64, pp. 49–61, 2018.

[6] E. Turban, J. E. Aronson, and T. Liang, "Decision Support Systems and."

[7] S. M. Ayyad, A. I. Saleh, and L. M. Labib, "Gene expression cancer classification using modified K-Nearest Neighbors technique," *BioSystems*, vol. 176, no. December 2018, pp. 41–51, 2019.

[8] J. Han, M. Kamber, and Jian Pei, *Data Mining Concepts and Techniques*. 2012.

[9] C. di M. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*, First. Italy, 2009.

[10] A. Muhammad, M. S. Mazliham, P. Boursier, and M. Shahrulniza, "K-Nearest Neighbor Algorithm for Improving Accuracy in Clutter Based Location Estimation of Wireless Nodes," *Malaysian J. Comput. Sci.*, vol. 24, no. 3, pp. 146–159, 2011.

[11] M. of C. Bramer, *Principles of Data Mining*, Second. London, 2013.

[12] C. H. Cheng, C. P. Chan, and Y. J. Sheu, "A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction," *Eng. Appl. Artif. Intell.*, vol. 81, no. March, pp. 283–299, 2019.