JRSSEM 2022, Vol. 02, No. 3, 311 – 318
E-ISSN: 2807 - 6311, P-ISSN: 2807 - 6494

JRSSEM
JOURNAL RESEARCH OF SOCIAL SCIENCE, ECONOMICS, AND MANAGEMENT

CLASSIFICATION OF RESEARCH PROPOSAL FUNDING USING NAÏVE BAYES AND DECISION TREE METHODS

Saifuddin[1]
E.I.H. Ujianto[2]
[1,2]Program Studi Magister Teknologi Informasi, Universitas Teknologi Yogyakarta, Indonesia
[*]e-mail: saifuddin@student.uty.ac.id, erik.iman@uty.ac.id
**Correspondence* : saifuddin@student.uty.ac.id

Submitted : 09 October 2022 *Revised* : 17 October 2022 *Accepted* : 25 October 2022

Abstract: In the selection process, determining a university funding research proposal at Tunas Pembangunan University (UTP) still has not fully used information technology to support related institutions, namely the UTP Institute for Research and Community Service (LPPM). So it has obstacles and requires a long time. So we need a system that is able to help these institutions to make it easier to determine recipients of research proposals that are worthy of funding. The application of data mining is a series of processes to explore added value in the form of knowledge that has not been known manually from a data set. This research has parameters, namely, NIDN, academic degree, track record, a proposed budget plan (RAB), and targeted outcomes. This is certainly less efficient because if a lecturer proposes a proposal, he must wait a long time to find out whether the results are accepted, accepted with improvements, or not. In addition, the assessment process has not used relevant methods so the results of the assessment of research proposal selection are not objective because the results of the assessment of the proposals obtained by the lecturer proposing the proposal are the final results in the form of a feasibility recommendation contained in a decision letter so that the application of classification with criteria in accordance with the selection needs is necessary. research proposal. By applying the data mining algorithm of the Naïve Bayes Method and the Decision Tree, it is hoped that it can simplify and accelerate the LPPM in determining recipients of research proposals that are eligible for funding at Tunas Pembangunan University.

Keywords: classification, research, data mining.

DOI: 10.36418/jrssem.v2i03.273

INTRODUCTION

In carrying out the realization of the Tri Dharma of higher education by educators (lecturers) are conducting research, community service, and teaching. Universities are obliged to carry out research and community service in addition to carrying out education as stipulated in Law Number 20 of 2003 concerning the National Education System Article 20. In line with this obligation, Law Number 12 of 2012 concerning Higher Education Article 45 stipulates that university research is directed at developing science and technology, as well as improving the welfare of the community and the competitiveness of the nation.

Therefore, every university, through the Institute for Research and Community Service (LPPM), is an institution that conducts research and community service for lecturers with full management, assessment and funding carried out professionally and proportionally. In conducting research and community service, it must be in accordance with the procedures set by LPPM.

From the existing procedures, one of which is the assessment of research proposals and community service that can be declared worthy or not feasible to be implemented and funded by universities. Every research and community service proposal submitted by a lecturer must go through a selection stage, both the administrative stage and the substance selection. The selection of incoming research proposals will be selected by a reviewer team determined by LPPM. This is also done by LPPM Tunas Pembangunan University (UTP) in selecting research proposals, but in assessing research proposals it still uses the manual method without a system that helps.

So far, the selection process carried out by LPPM UTP is in accordance with existing procedures, namely by carrying out the process of proposing a research proposal, then being assessed by the reviewer team based on the assessment sheet and predetermined criteria. However, this activity has not been supported by an information system so the research proposal selection process must be carried out by recapitulating the assessment data which requires a long assessment time.

This is certainly less efficient because if a lecturer proposes a proposal, they have to wait a long time to find out whether the results are accepted, accepted with corrections, or not. In addition, the assessment process has not used the relevant method so the results of the research proposal selection assessment are not objective because the results of the proposal assessment obtained by the proposing lecturer are the final result in the form of a feasibility recommendation contained in the decision letter, so it is necessary to apply a classification with criteria that are in accordance with the selection needs. research proposal.

So to overcome these problems it is necessary to have a classification of weighting criteria for research proposals submitted by lecturers in the UTP environment with data mining using the Naïve Bayes and Decision Tree methods.

facilitate the assessment team (reviewer) to complete the assessment of research proposals and community service by lecturers.

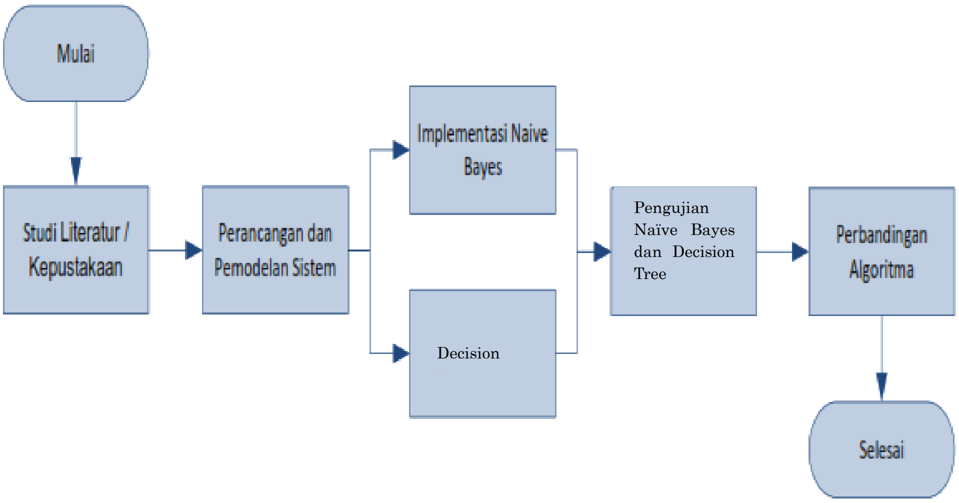
In several previous studies, many classification methods have been implemented in real life. Some algorithms that are very popular today are Naive Bayes and C.45. Naive Bayes is a classification algorithm with a simple formula and is easy to apply as presented by (Jadhav et al., 2016) and (Maryamah et al., 2016), while C.45 algorithm in several studies using decision tree classification, such as research (Saxena & Sharma, 2016) provides a high level of accuracy. Another study that examines the comparison of the performance of several data mining classification methods has previously been carried out (Santra & Jayasudha, 2012), in this study, used the Naive Bayes algorithm for the classification technique, whereas previous studies used the C4.5 algorithm. Another study (Dimitoglou et al., 2012) tested the ability of data mining and machine learning methods to accurately predict the survival of patients diagnosed with lung cancer. This study compares the effectiveness of the naive Bayes algorithm and decision tree C4.5 which are implemented to predict a person's survival due to certain diseases. The results obtained indicate that the naive Bayes algorithm is superior to the C4.5 decision tree for this case. (Ashari et al., 2013) proposed a new method for finding alternative designs by using the classification method. The methods used in this study include nave Bayes, decision tree, and k-nearest neighbor. The experimental results show that the decision tree excels in the calculation speed process followed by naive Bayes and k-nearest neighbors. Data Mining also known as Knowledge Discovery in Database (KDD) is defined as the extraction of potential, implicit and unknown information from a set of data. The Knowledge Discovery in the Database process involves the results of the data mining process (the process of extracting the tendency of a data pattern), then converting the results accurately into information that is easy to understand (Siburian, 2014). The terms data mining and Knowledge Discovery in Databases (KDD) are often used interchangeably to describe the process of extracting hidden information in a large database. Actually, these terms have different concepts but are related to each other and one of the stages in the whole KDD process is data mining (le Cam et al., 2016). Data mining refers to the process of searching for previously unknown information from a large data set (Ginting et al., 2014). Another definition of Data Mining is a series of processes that employ one or more computer learning techniques to analyze and extract knowledge automatically or a series of processes to explore added value from a data set in the form of knowledge that has not been known manually (Sijabat, 2015). Data mining is a term used to describe the discovery of knowledge in databases (Kusrini & Taufiq, 2009).

The results of the author's observations, research on "Classification of Determination of Internal Funding Research Proposals Using the Naïve Bayes Method and Decision Tree (Case Study: Tunas Pembangunan University)" with the aim of holding this classification is expected to find out which method has greater accuracy. Thus, the best method between the two is used in determining the research proposal, which is between eligible to be funded or not eligible to be funded and there are several similar topics that have been carried out, including the following:

1. Comparison of C4 Data Mining Algorithm Classification Methods. 5 and Naive Bayes for Hepatitis Disease Prediction (Septiani, 2017).
2. Comparison of Naive Bayes classifier and C4. 5 algorithms in predicting student study period (Gerhana et al., 2019).
3. Comparison of Data Mining Model C4 Classification Algorithm. 5 and Naive Bayes for Diabetes Disease Prediction (Fatmawati, 2016).
4. Comparison of C4 Algorithm Performance. 5 and Naive Bayes for the Accuracy of Student Concentration Selection (Supriyanti et al., 2016).
5. Classification of posts Twitter traffic jam the city of Jakarta using algorithm C4. 5 (Hajrahnur et al., 2018).

MATERIALS AND METHODS)

This study, uses a methodological step consisting of several stages as follow:


Mulai
Studi Literatur / Kepustakaan
Perancangan dan Pemodelan Sistem
Implementasi Naive Bayes
Decision
Pengujian Naïve Bayes dan Decision Tree
Perbandingan Algoritma
Selesai

Picture 1. Research Methodology

Literature Study

At this stage, a search for literature studies on research material is carried out which includes data mining, classification, Naive Bayes algorithm, decision tree, and testing methods using precision, recall, and accuracy. The search is based on previous research on theories related to the research conducted as well as theories that are currently being developed.

System design and modeling

At this stage, two systems are designed, namely the system that applies the Naive Bayes algorithm and the system that applies the Decision Tree. The system built applies 2 data mining classification algorithms in the selection of research proposals at Tunas Pembangunan University. The data in this case is data on proposals received from the Institute for Research and Community Service, Universitas Tunas Pembangunan. Data obtained from sources as much as 140 data. The data used as consideration for determining the recipient of the proposal are:

a. dependent variable (bound)

Variable Y: Acceptance of the proposal (Accepted / Not Accepted)

b. Independent variable (not bound)

Variable X: age, last education, RAB submission, number of proposals, number of publication outputs

Implementasi Naive Bayes

The step taken at this stage is to translate the design that has been formed into a system that applies the Naive Bayes algorithm. (Jadhav et al., 2016) stated that the Naïve Bayes Classifier is an independent model that discusses simple classification based on the Bayes theorem. Naïve Bayes is an algorithm that can classify a certain variable using probability and statistical methods. Broadly speaking, the Naïve Bayes algorithm can be explained as follows:

$$P(R|S) = \frac{P(R)P(S|R)}{P(S)} \tag{1}$$

Implementasi *Decision Tree*

The concept of a decision tree or decision tree is to convert data into decision rules. The main benefit of using a decision tree is its ability to break down complex decision-making processes into simple ones so that decision-making will make it easier to solve a problem (Rismayanti, 2018). Decision Tree is one of the most popular classification methods because it is easily interpreted by humans (Wahyuningsih & Utari, 2018). A Decision Tree is used for pattern recognition and is included in statistical pattern recognition (Rosandy, 2016). The Decision Tree uses 2 calculations, the first is the Gain calculation in Equation 2 and the Entropy calculation in Equation 3 (Nugraha et al., 2016).

Testing Precision, Recall and Accuracy Naive Bayes and *Decision Tree*

Testing an algorithm requires standards and test equipment (Kurniawan & Kurniawan, 2018). Comparing 2 algorithms must have the same standard so that the best algorithm can be known from the comparison. At this stage, testing is carried out by calculating the value of precision, recall, and accuracy from Naive Bayes and Decision Tree. The initial step in this stage is to divide the data in each case into 2, namely training data or training data and testing data or test data. Training data is used as reference data in the calculation of each algorithm, while testing data is used to assess the predictions and determinations made by each algorithm are correct or not. In dividing the data into training data and testing data, several comparisons were made.

Algorithm Comparison

At this stage, a comparison of the values of precision, recall, and accuracy is carried out for each algorithm in each case. After that, the results of each algorithm are recapitulated so that conclusions can be drawn regarding the best algorithm for each case.

RESULTS AND DISCUSSION

After carrying out several stages and research procedures from pre-processing or preparation to a process that includes several stages including data cleaning, data collecting, determining criteria, determining probabilities, and testing, the following are the results of the research:

Naïve Bayes Methode

The research stages that will be carried out are prepositional data, namely processing raw data from data on research funding receipts from 2016 to 2022 with a total dataset of 140.

In the Naïve Bayes method, constant string or categorical data is divided into two types, namely continuous numeric data, so that The resulting difference will be seen when determining the probability value of each criterion, either criterion with string data values or criteria with numeric data. The stages carried out are as follows:

1. Data Collection, namely data that is used as training and testing data, in this case, is data on receipt of research funds. The criteria for determining training data and testing data are 70% training data and 30% testing data.
2. Data Cleaning, namely at this stage there is a criterion that is eliminated because these criteria have no effect on the results of the classification accuracy of the Naive Bayes method. From the total dataset of 140, after eliminating the dataset, the number of datasets is 27. While the number of attributes used is 15.
3. Determining Criteria, at the stage of determining these criteria, data criteria are used based on the data that has been collected.
4. Determine the Probability of Each Criterion, at this stage determine criteria used as a reference in classifying recipients of research funds
5. Testing, at this stage is the stage of applying the Naive Bayes method with some data that is ready to be tested.

   The dataset used is divided into two parts with a ratio of 70% used for training data with 98 data records and 30% testing data with 42 data records from the total number of records from the 140 datasets. And the accuracy results obtained from the experimental design above using the method Naive Bayes is an accuracy value of 92.70%.

Decision Tree Method

In testing using the Decision Tree method, there are several criteria used, including gain_ratio, information_gain, gini_index, and accuracy. This criterion is one of several operators that will produce an estimate of how accurately a model will appear.

The dataset used is divided into two parts with a ratio of 70% used for training data with 98 data records and 30% testing data with 42 data records from the total number of records from 140 datasets. Testing training and testing data using the Decision Tree method obtain accurate results. 43.54%.

CONCLUSIONS

To classify the research proposal funding using the Naïve Bayes method and the Decision Tree method, it is better to use the Naïve Bayes method because the accuracy obtained is 92.70%.

REFERENCES

Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *4*(11).

Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. *ArXiv Preprint ArXiv:1206.1121*.

Fatmawati, F. (2016). Perbandingan Algoritma Klasifikasi Data Mining Model C4. 5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes. *Techno Nusa Mandiri: Journal of Computing and Information Technology*, *13*(1), 50–59.

Gerhana, Y. A., Fallah, I., Zulfikar, W. B., Maylawati, D. S., & Ramdhani, M. A. (2019). Comparison of naive Bayes classifier and C4. 5 algorithms in predicting student study period. *Journal of Physics: Conference Series*, *1280*(2), 022022.

Ginting, S. L. B., Zarman, W., & Hamidah, I. (2014). Analisis Dan Penerapan Algoritma C4. 5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik. *PROSIDING SNAST*, 263–272.

Hajrahnur, S., Nasrun, M., Setianingsih, C., & Murti, M. A. (2018). Classification of posts Twitter traffic jam the city of Jakarta using algorithm C4. 5. *2018 International Conference on Signals and Systems (ICSigSys)*, 294–300.

Jadhav, A., Pandita, A., Pawar, A., & Singh, V. (2016). Classification of unstructured data using naïve bayes classifier and predictive analysis for RTI application. *An International Journal of Engineering & Technology*, *3*(6), 1–6.

Kurniawan, D. A., & Kurniawan, Y. I. (2018). Aplikasi Prediksi Kelayakan Calon Anggota Kredit Menggunakan Algoritma Naïve Bayes. *Jurnal Teknologi Dan Manajemen Informatika*, *4*(1).

Kusrini, E. T. L., & Taufiq, E. (2009). Algoritma data mining. *Yogyakarta: Andi Offset*.

le Cam, M., Daoud, A., & Zmeureanu, R. (2016). Forecasting electric demand of supply fan using data mining techniques. *Energy*, *101*, 541–557.

Maryamah, M., Asikin, M. F., Kurniawaty, D., Sari, S. K., & Cholissodin, I. (2016). Implementasi Metode Naïve Bayes Classifier Untuk Seleksi Asisten Praktikum Pada Simulasi Hadoop Multinode Cluster. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK) FILKOM UB*, *3*(4), 273–278.

Nugraha, P., Aribawa, I. W., Priyana, I. P. O., & Indrawan, G. (2016). Penerapan Metode Decision Tree (Data Mining) Untuk Memprediksi Tingkat Kelulusan Siswa Smpn1 Kintamani. *Seminar Nasional Vokasi*.

Rismayanti, R. (2018). Decision Tree Penentuan Masa Studi Mahasiswa Prodi Teknik Informatika (Studi Kasus: Fakultas Teknik dan Komputer Universitas Harapan Medan). *Query: Journal of Information Systems*, *2*(1).

Rosandy, T. (2016). Perbandingan Metode Naive Bayes Classifier Dengan Metode Decision Tree (C4. 5) Untuk Menganalisa Kelancaran Pembiayaan (Study Kasus: KSPPS/BMT Al-Fadhila. *Jurnal Teknologi Informasi Magister*, *2*(01), 52–62.

Santra, A. K., & Jayasudha, S. (2012). Classification of web log data to identify interested users using Naïve Bayesian classification. *International Journal of Computer Science Issues (IJCSI)*, *9*(1), 381.

Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. *Procedia Computer Science*, *85*, 962–969.

Septiani, W. D. (2017). Komparasi Metode Klasifikasi Data Mining Algoritma C4. 5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis. *Jurnal Pilar Nusa Mandiri*, *13*(1), 76–84.

Siburian, B. R. (2014). Aplikasi Data Mining Untuk Menampilkan Tingkat Kelulusan Mahasiswa Dengan Algoritma Apriori. *Teknik Informatika STMIK Budi Darma Medan*.

Sijabat, A. (2015). Penerapan Data Mining Untuk Pengolahan Data Siswa Dengan Menggunakan Metode Decision Tree (Studi Kasus: Yayasan Perguruan. *Vol. V*, 7–12.

Supriyanti, W., Kusrini, K., & Amborowati, A. (2016). Perbandingan Kinerja Algoritma C4. 5 dan Naïve Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa. *Jurnal Informa: Jurnal Penelitian Dan Pengabdian Masyarakat*, *1*(3), 61–67.

Wahyuningsih, S., & Utari, D. R. (2018). Perbandingan Metode K-Nearest Neighbor, Naive Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit. *Konferensi Nasional Sistem Informasi (KNSI) 2018*.