Implementation bayes network based for classification study of completion

C Chotimah, Rachmat and T A Darsono
Informatics Engineering, Faculty of Engineering, Musamus University, Merauke, Indonesia
E-mail: cchotimah@unmus.ac.id

Abstract. Many factors influence the completion of the study. However, it is likely that some variables are dependent on each other. Method Bayes Network is used because it can illustrate the cause-effect relationship between completion of study with the factors that allegedly influenced as a graphical model representation so that more easily understood. With empirical data, the presence or absence of relationships among factors is calculated using mutual information. If the structure of the network has been established then subsequently calculated the probability of relationship for each event. The probability value obtained is shown to obtain a complete causality diagram. The results of the research on the classification process using student data in 2012 with Bayes Network method resulted in an accuracy of 71.09%.

1. Introduction
The amount of academic data continues to increase but the number of graduating students is not comparable to the number of students enrolled especially in completing studies on time. Student performance evaluation is needed for university management to take strategic steps in order to produce the right decisions. Completion of timely studies is one of the elements of accreditation assessment [1].

Data mining is a process of knowledge discovery from very large data [2]. Basically data mining aims to extract knowledge that is still hidden from very large data whose results can be understood by the user. For example, a marketing manager can classify by predicting whether a customer will buy a product or not, in the medical field to categorize tumor cells whether they are included in the category of malignant or benign, or by the institution of higher education to classify students based on specialization in their study program.

There are various kinds of classification techniques in data mining including Rule-Based Classifiers, Naive Bayes Classifiers, Bayes Network and Decision Tree [2]. The classification process usually uses training data sets where all objects are related to the known class label. Classification techniques will be used to analyze and retrieve the properties of training data to build a model. Then the model obtained will be used to classify new objects.

Many factors influence the completion of the study. Several studies have proven using various methods including Naive Bayes Classifier [3], and C4.5 Algorithm [4]. The characteristic of completing the study in the study is that not only academic factors influence the completion of studies but also are influenced by non-academic factors [3,4] assuming independence between factors. But there may be several interdependent variables.
This study uses the Bayes Network method, because the concept of this method is to be able to find the possibility of an event with interdependent variables. The meaning of the possibility of an event is the result of processing data in classifying the completion of the study. This approach is used because it can describe the causal relationship between completion of studies with factors that are thought to influence as a representation of graphic models so that they are easier to understand. Bayes Network is a graphical representation of causality relationships that are in a set of random variables [5]. The application of the Bayes Network method will make a model used in the study completion classification process.

2. Methods

2.1 Materials and Methods
Figure 1 is an activity activity carried out in the research that includes identification of factors that affect the completion and causality among factors from the literature study, perform dependency testing between factors with mutual information, calculate probability, implementation on case data and calculate the accuracy level.

![Diagram](image)

Figure 1. Research methods

2.2 Factors study of completion
Factors or variables used in terms of data availability are shown in Table 1 and change the form of data values to the appropriate format so that the data is ready for use.

| Variable (node) | Description of Variable | Value | Description |
|-----------------|-------------------------|-------|-------------|
| ps              | Completion of study      | tw    | On time ($\leq$ 4) |
|                 |                         | ttw   | Not timely ($> 4$) |
| jkel            | Gender                  | l     | Male        |
|                 |                         | p     | Women       |
| suku            | Ethnic                  | np    | All tribes in the State of Indonesian outside the provinces of Papua and West Papua |
|                 |                         | pa    | All tribes in the provinces of Papua and West Papua |
|                 |                         | ma    | Marind (native of Merauke district) |
2.3 Causality Analysis of Factors (variables)

The conceptual model is a description of the causality of a factor (variable) directly or indirectly, then uses empirical data on the interdependence test between factors using mutual information (MI). The causality test results are shown in Table 2.

Table 2. Results causality test using mutual information

|   | ipk       | jkel      | suku      | St      | as        | js        | ps        |
|---|-----------|-----------|-----------|---------|-----------|-----------|-----------|
| ipk| 0.00025*  | 0.00122*  | 0.00297*  | 0.00003*| 0.00193*  | 0.01820** |
| jkel| 0.00025* | 0.00005*  | 0.00000*  | 0.00025*| 0.00054*  | 0.01313** |
| suku| 0.00122* | 0.00005*  | 0.000076* | 0.00030*| 0.00038*  | 0.01526** |
| St | 0.00297*  | 0.00000*  | 0.00076*  | 0.00027*| 0.00016*  | 0.01219** |
| as | 0.00003*  | 0.00025*  | 0.00030*  | 0.00027*| 0.02056** | 0.00004*  |
| js | 0.00193*  | 0.00054*  | 0.00038*  | 0.00016*| 0.02056** | 0.01212** |

*Value below threshold

**Value as per threshold

Based on MI values in table 2, the factors of GPA, gender, ethnicity, marital status, and school majors have direct interaction with the completion of studies which can be seen from the results of the mutual value of information having a value greater than the threshold value. However, high school factors do not have interaction with the completion of studies directly, it can be seen from the value of mutual information that is less than the threshold value. Factors from the school have direct interaction with the mayor school. Based on the results of the causality test, Figure 2 shows a picture of the conceptual model for completion of the study at Universitas Musamus Merauke.

![Figure 2. Conceptual model study of completion](image-url)
2.4 Calculation of probability values
Based on the conceptual model completion of studies that are formed which are network structures. The next step is to calculate the probability of each factor (node) using empirical data. Probability calculation for each variable is the number of frequencies in the attribute values of each variable (node) and also the attribute value of the combination of variables that influence [6]. The probability values are exemplified in the ipk variables shown in Table 3 and the ps variable values are shown in table 4.

| Table 3. Probability of grade point average |
|--------------------------------------------|
| Grade-Point Average | ipk |
| Satisfactory (mm) | 0.17 |
| Very Satisfactory (sm) | 0.76 |
| Cumlade (cum) | 0.06 |

| Table 4. Conditional probability study of completion |
|-----------------------------------------------------|
| Combination node | P(ps=tw) | P(ps=ttw) |
|------------------|----------|-----------|
| mm l np bl ipa   | 0.36     | 0.64      |
| mm l np bl ips   | 0.33     | 0.67      |
| mm l np bl lain  | 0.33     | 0.67      |
| sm p np bl ipa   | 0.51     | 0.49      |
| sm p np bl ips   | 0.55     | 0.45      |
| sm p np bl lain  | 0.47     | 0.53      |
| cum p np bl ipa  | 1.00     | 0.00      |
| cum p np bl ips  | 1.00     | 0.00      |
| cum p np bl lain | 1.00     | 0.00      |

The diagram with probability value of each variable (node) can be seen in Figure 3. This diagram is used as the basis to classify the prediction function if the evidence of one or several variables in accordance with the observation facts.
3. Results

3.1. Probability test study of completion
The probability test for completion of the study is the process of calculating the probability of completion of the study using the model in Figure 3. The calculation results show that the probability of completion of the study on time "tw" of 49% can be seen in Table 5. The calculation results indicate that the probability of completion of the study is not timely "ttw" is larger.

Table 5. Probability study of completion

| Penyelesaian Studi | Probability |
|--------------------|-------------|
| On Time            | 0.49        |
| Not Timely         | 0.51        |

3.2. Testing by setting the value of the evidence
Scenario testing is by profiting the value of evidence for each variable (node) so that it will be able to calculate the probability of completion of the study based on the parameters inputted. The results of the test in Table 6 by doing 25 scenarios, the probability of completion of the study is not timely "ttw" is greater that is 52% (13 records) while 48% (12 records) for data on the class of completion of the study on time "tw".

Table 6. Test results with scenario (setting of evidence)

| Node | ps |  |  |
|------|----|---|---|
| Jkel | ethnic | st | as | js | ipk | tw | ttw |
| L    | np  | bl | sma | ipa | mm | 0.0089728493 | 0.0159517320 |
| L    | pa  | me | sma | ipa | sm | 0.0006261573 | 0.0009392360 |
| P    | ma  | bl | sma | ipa | mm | 0.0000000000 | 0.0006207271 |
| L    | ma  | bl | smk | ipa | cum | 0.0000387763 | 0.0000000000 |
| P    | np  | me | sma | ips | sm | 0.0081616941 | 0.0043947584 |
3.3. Classification of Test Data

In Figure 4 is the result of the classification process using student data of 2012 taken at random amounts to 128 test data with the class label is already known so that the prediction will be produced correctly or wrongly. The results of data testing provide accuracy for the classification process of completion of studies using the bayes network that is with the accuracy of 71.09%. Summary of test results in confusion matrix is shown in Table 7.

Table 7. Result Classification

| Student Registration Number | Completion of studies (actual) | Probability value on time | Probability value Right on time | Prediction result | Appropriate predictions |
|-----------------------------|--------------------------------|---------------------------|-------------------------------|-------------------|------------------------|
| 201220201002                | ttw                            | 0.0077001047              | 0.0156335459                  | ttw               | True TTW               |
| 201220201010                | ttw                            | 0.0000000000              | 0.0023944378                  | ttw               | True TTW               |
| 201220201016                | ttw                            | 0.0083326334              | 0.0130331089                  | ttw               | True TTW               |
| 201220201017                | ttw                            | 0.0018248911              | 0.0000000000                  | ttw               | True TTW               |
| 201220201019                | ttw                            | 0.0000000000              | 0.0025615526                  | ttw               | True TTW               |
| 201220201020                | ttw                            | 0.0615637158              | 0.0591494524                  | ttw               | True TTW               |
| 201220201021                | ttw                            | 0.0017205053              | 0.0030586761                  | ttw               | True TTW               |
| 201220201022                | ttw                            | 0.0000000000              | 0.0023944378                  | ttw               | True TTW               |
| 201220201027                | ttw                            | 0.0000000000              | 0.0001098662                  | ttw               | True TTW               |
| 201220201029                | ttw                            | 0.0045898632              | 0.0584164805                  | ttw               | False TTW              |
| 201220201030                | ttw                            | 0.0005303102              | 0.0000000000                  | ttw               | False TTW              |
| 201220201001                | ttw                            | 0.0013979106              | 0.0028381821                  | ttw               | True TTW               |
| 201220201014                | ttw                            | 0.0000000000              | 0.0000671405                  | ttw               | True TTW               |
| 201220201015                | ttw                            | 0.0000483411              | 0.0005559232                  | ttw               | True TTW               |

The classification process will produce predictions correctly or incorrectly. The classification results tabulated with confusion matrix are shown in Table 8.

Table 8. Results of the bayes network classification with confusion matrix
| Prediction       | On time | Not on time |
|------------------|---------|-------------|
| On Time          | 48      | 18          |
| Not on time      | 19      | 43          |

Accuracy = \( \frac{48 + 43}{48 + 18 + 19 + 43} \times 100\% = 71.09\% \)

The results of the study of completion classification process using Bayes Network have an accuracy value of 71.09%.

4. Conclusions
The structure of the bayes network generated is strongly influenced by the condition of the dataset used, then the resulting structure of a dataset may not necessarily apply to any other dataset. The quantity of the dataset and the distribution of variable values in the dataset used for the construction of the bayes network structure significantly influence the structure and accuracy of the resulting predictions. Testing system model using 128 data samples with bayes network method resulted in higher value accuracy value that is equal to 71.09%.

References
[1] Anonim.2008. Buku VI Matriks Penilaian Instrumen Akreditasi Program Studi Sarjana. Badan Akreditasi Nasional Perguruan Tinggi. Jakarta.
[2] Han, J., Kamber., Pei, J., 2012. Data Mining: Concepts and Techniques 3rd edition. Elsivier Inc. USA.
[3] Ridwan, M., Suyono, H. dan Sarosa, M., 2013. Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier, Jurnal EECCIS Vol.7, No. 1.
[4] Jaman, J. H., 2013. Prediksi Kelulusan Mahasiswa Dengan Metode Algoritma C4.5. SYNTAX. Vol. 2 No. 2.
[5] Jananto, A., 2010, Prediksi Kinerja Mahasiswa Baru Menggunakan Teknik Data Mining, Tesis, Prodi S2/S3 Ilmu Komputer, Universitas Gadjah Mada, Yogyakarta.
[6] Kjærluff, U. B. dan Madsen, A. L., 2013. Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis. Second Edition. Springer New York.