An Analysis of the Comparative Method of Classification in Determining Characteristics of Non-Active Students

[1]Fitra Luthfie Averroes, [2]JakaFitra, [3]Sriyanto
Departement of Informatics Engineering
Faculty of Computer Science
Informatic and Business Institute Darmajaya
Fitra.1721211010@mail.darmajaya.ac.id, jaka.1721211012@mail.darmajaya.ac.id, dan sriyanto@darmajaya.ac.id

ABSTRACT

Classification is a data mining technique that aims to know the data model being used. By using a data mining classification technique, we can classify existing information into their classes. Classification methods can also be applied in education, for example to group active and inactive students into higher education programs, and classify them based on their characteristics. This paper presents a comparison of several classification methods, which are: Naïve Bayes, k-NN and C 4.5. This paper uses the data from the inactive students of AMIK DCC campus C on 2013-2016 periods as the criteria to evaluate the group performance. The inactive students are divided into three groups: first-year, first-two-years, and three years inactive students. The results of this study indicate that the Bayes naive method provides higher accuracy than k-NN and C 4.5. The accuracy classification is Naïve Bayes 79.22%, while k-NN and C 4.5 are 77.21% and 74.94%, respectively.

Key word: Decision Tree, C4.5, Classification, Naïve Bayes, AMIK DCC Campus C

1. Introduction

A lot of public and private universities, public university and private college have a college information system that supports data management. Inactive student data can be one of the academic data, besides other human resources (such as lecturers, technicians, administrative personnel, etc.) and also equipment, finance and so on. The applications usually contain data for academic students, curriculum, lecturers, study plan cards, schedules, grades, and learning as well as alumni data.

One of the important informations is related to the factors of the inactivity of the students in college (PT). The factors can be seen from two sides, which are internal and external factors. While some external factors will be difficult or even impossible to control, and some can still be avoided by considering factors in the admission of prospective students, butthe internal factors can be controlled by appropriate policies.

Many studentsare not being active in their first year, and the percentage of active students are still high [15] and [16]. The management of non-active students on campus C is ineffective because it does not consider the Characteristics of the students. It is a challenge for the campus to find motives, determine student's characteristics and find the solutions to reduce the number of inactive students. The initial effort is to analyze the characteristics of non-active students, considering the high percentage of inactive students.

At the first year since student's registration, the percentage of inactive students is high, anddata mining techniques can be used to determine their characteristics that could affect their academic status. Based on the characteristics of inactive students, it is hoped that the campus could devise and find strategies and also solutions to handle and manage inactive students, so the high number of inactive students can be reduced.

In developing examples of data models, data Classification mining techniques will separate records into appropriate categories or classes.

This model is then used to classify the unknown data entries, and the number of algorithm classifications in the data mining, however not all can be done with classifying the categories of active or non-active students. This paper will compare several methods of classification algorithms: Baye naive, C4.5 and k-NN tree decisions.

The knowledgeandinformation gained from this study are expected to be:

1) Able to estimate the student's resignation rate in the coming years.
2) Able to reduce the student's resignation rate through appropriate policy-making based on the research results.
3) Used by the leaders as a reference to prepare:
   - New students' selection system.
   - System for learning and teaching.

2. Research Method

The data this research used was collected from the students ofcampus C who were registered in 2013 which have4 years academic records (8 semesters), up to 2016. The four-year period is used because the average duration of students completing the course is 3 years. The attributes used in creating the classification model are:

1) The study programs offered to students.
2) The classes taken, grouped into three categories: morning, night, and Saturday or Sunday.
3) The age of the student at the time of registration.
4) Gender: Male and Female
5) Marital status: married or unmarried.
6) Job status is grouped into workingor not working.
7) The amount of credits taken by students during their studies on campus.
8) The average number of courses enrolled each semester.
9) The GPA for three semesters: first-third-third encoded as, GPA1, GPA2 and GPA3.
10) The Targets, categorizing the status of non-active students and grouping them into three groups: C1 (non-active students in the first year, 2013-2014), C2 (inactive within the first two years, 2014-2015) and C3 (non-active in more than three years, 2015-2016).

The software used for processing classification data attributes is Rapidminer v5.3. The criteria used to evaluate performance using several methods of classification, the method used: naive Bayes, k-NN and c 4.5. Each classes of data collection should be represented in the appropriate proportion between training data and testing data. The Data is divided randomly in each class with the same ratio. To reduce the bias caused by a particular sample, the entire training and testing process is repeated several times with different samples. An error rate that is different from the average will be calculated to create an overall error rate.

2.1 Naïve Bayes

Naïve Bayes is a straightforward and powerful algorithm for the classification task which uses conditional probability and called as Bayes theorem. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge. For instance, over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables *F*1 through *Fn*, then the method use Bayes theorem [22]:

$$P(C|F_1,\ldots,F_n) = \frac{P(C)P(F_1,\ldots,F_n|C)}{P(F_1,\ldots,F_n)},$$

P (F1, ..., Fn) is the probability of a sample with F1, ..., Fn Characteristics present in class C, or widely known as Posterior. Mathematically, the Naïve Bayes classification is formulated as follows [22]:

$$C_{\mathrm{NB}} = \underset{c\in C}{\operatorname{argmax}}\, P(c)\prod_{i=1}^{n} P(f_i|c),$$

With c as the class variable included in a set of class C

2.2 *C4.5 Algorithm*

The C4.5 algorithm is the successor of the decision tree of the ID3 algorithm, introduced by Quinlan in 1979. Like the ID3 algorithm, the C4.5 algorithm uses the measurement of the Gain Information to determine the best attribute to divide the D data set at the time of decision tree formation. The value of the Gain information is calculated based on the value of the information needed to classify the records in the data set D before and after the data sets are separated by an attribute A. The information needed to classify a tuple in D is determined as follows [7]:

$$\mathrm{Info}(D) = -\sum_{i=1}^{m} p_i \log_2 p_i,$$

Where pi is the probability that the tuple on D has the class label Ci, and its value is estimated by | Ci, D | / |

D |. Info (D) is known as entropy D. It is assumed we want the partition record in D in attribute A having different value v, {a1, a2, ...,av}. The partition generated corresponds to the branches of node N. The information required to classify the tuples of D based on data separation by A is defined as Info A (D) by the following formula [7].

Where | Dj | / | D | is the partition weight to j.

The Profit Information formula is as follows

Gain (A) = Info (D) – InfoA (Dj).

Attribute A with the highest Gain Information value is selected as a separate attribute on node N. The decision tree forming process using C4.5 algorithm is performed in three main things step [20]:

- Generate decision tree using the ID3 algorithm.
- Change the decision tree into a collection of IF-THEN rules.
- Reduce any rules by removing the prerequisites if the accuracy of the rules increases without preconditions

2.3 *k-NN*

Thek-Nearest Neighbor algorithm is based onlearning by analogy, that is, by comparinga given test example with training examples that are similar to it. The training examples are described by an attribute. Each example represents a point in an n-dimensional space. In this way, all of the training examples are stored in an n-dimensional pattern space. When given an unknown example, a k-nearest neighbor algorithm searches the pattern space for the k training examples that are closer to the unknown example. These k training examples are the k "nearest neighbors" of the unknownexample. "Closeness" is defined in terms of a distance metric, such as the Euclidean distance.
Thek-nearest neighboralgorithmis amongstthesimplest of all machinelearning algorithms: an example is classified by a majority vote of its neighbors, with the example been assigned tothe class most commonlyamongst its k nearest neighbors (k is a positive integer, typically small).If k=1, then theexampleis simply assignedto theclass of its nearest neighbor.Thesamemethodcan be used for regression, by simply assigning the label value for the example to be the averageof the values of its k nearest neighbors. It can be useful to weight the contributions of theneighbors, so that the nearer neighbors contribute more to the average than the more distantones.
The neighbors are taken from a set of examples for which the correct classification (or, in thecase of regression, the value of the label) is known. This can be thought of as the training setfor the algorithm, though no explicit training step is required.

The basic k-Nearest Neighbor algorithm is composed of two steps:

1. Find the k training examples that are closer to the unseen example.
2. Take the most commonly occurring classification for these k examples (or, in the case of

Regression, take the average of these k label values).

2.4 Evaluation

If confusion matrix expressed by the following matrix:

		Predicted class	
		Yes	No
Actual class	Yes	True positive = TP	False negative = FP
	No	False positive = FP	True negative = TP

Then the value true positive rate, false positive, success rate, and error rate stated as follow:

a) True positive rate = TP/(TP + FN)
b) False positive rate = FP/(FP + TN)
c) Success rate = (TP + TN)/(TP + TN + FP + FN)
d) Error rate = 1-success rate

Kappa statistics is a measure of performance improvements relative to random predictor.

$$\text{Kappa statistics} = \frac{D_{\text{observed}} - D_{\text{random}}}{D_{\text{perfect}} - D_{\text{random}}},$$

The formula mean absolute error, relative absolute error and root mean squared error of each stated below.

$$\text{Mean absolut error} = \frac{\sum_{i=1}^{n} |p_i - a_i|}{n},$$

$$\text{Root mean squared error} = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}},$$

$$\text{Relative absolut error} = \frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i}^{n} |\bar{a} - a_i|},$$

where*a* is the actual target value and *p* is the predictor target value.The AUC is widely recognized as the measure of a diagnostic test's discriminatory

power.The AUC value measures discriminative performance by estimating output probability selectedrandomly from the positive or negative population. The larger the AUC, the stronger the classificationused. AUC values ranged between 0.0 and 1.0. Meanwhile, three other values, namely,recall, Precision, and F-measure are the criteria used to evaluate the similarity among the attributesclassification. Recall is the success rate of recognizing classification results that need to berecognized. Precision is the degree of accuracy of classification results, whereas the F-measureis the value that represents the overall performance of the system and the combination of recalland precision values. Formulation for Recall, Precision, and F-measure, respectively:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

3. Results And Discussion

3.1. Data Processing

Description of this paper analyzes DCC campus non active students in two departments, non-active students D3 Management Informatics (MI) and Computer Accounting (KA). There are 351 students who are analyzed demographically and academically, i.e. active and inactive students at DCC campus C.

3.2. Implementation Result

3.2.1. Naive Bayes Methods

Data naive method Bayes, obtained as a result of matching data Campus C students using data 351 samples of students taken from 2013-2016 majoring in MI and KA, with the accreditation rate of 79.22%.

Table 1. *ConfusionMatriks Naïve Bayes*

	true NON AKTIF	true Aktif	class precision
pred. NON AKTIF	103	19	84.43%
pred. Aktif	54	175	76.42%
class recall	65.61%	90.21%	

In Naive Bayes method, one student of GPA <2.0 in semester 1 decreased the learning interest of the next semester shown in figures 1.

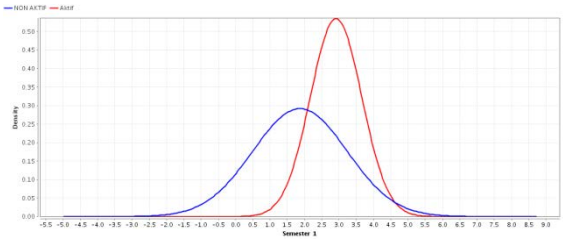


Figure 1. Indicates that on the first semester

Figure 2 indicates that on the second semester, the number of inactive students are still the same as the previous one.

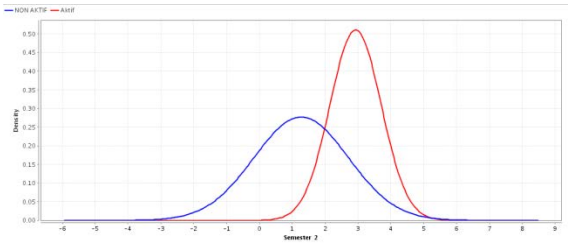


Figure 2. Indicates that on the second semester

Figure 3 indicates that on the third semester, there is a GPA decreasing on many students, which affected on the increasing number of inactive students in this semester.

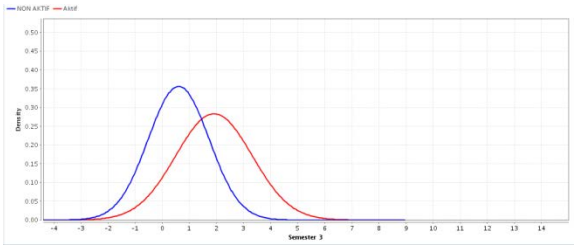


Figure 3. Indicates that on the third

3.2.3 Method k-NN

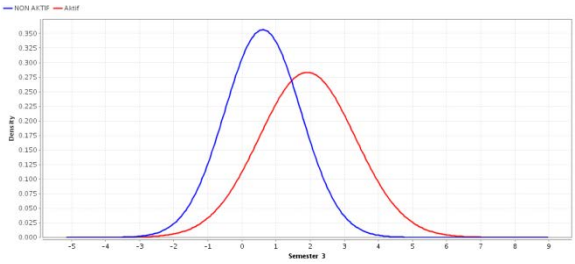
The K-NN method, obtained as a result of data matching between the test results of the proposed method using the 315 weather data tested with the accuracy rate of 77.21%. Accuracy is stated that the value of the ratio of the number of student data classified in the class correctly (true positive) and the amount of data belonging to the class that is otherwise (true negative).

The results of the test data matching are then inserted into the confusion matrix table. Based on Table 2 and the value of accuracy can be calculated using the following calculation:

Table 2. *ConfusionMatriks k-NN*

	True Non Aktif	True Aktif	Class Precision
Pred. NON	112	35	76.19%
Pred. Aktif	45	159	77.94%
Class recall	71.34%	81.96%	

Figure 4. Result of data execution above that is



3.2.4 Decision Tree Method

Method of Decision Tree got the result of matching test data between theresults of the method proposed by using student data which tested 351 that is accuracy 73,51%. Accuracy is stated that the value of the ratio of the amount of non-active student data are classified correctly and the amount of classified data is true (true negative) with all classified non-active student data.

The results of the test data matching are then inserted into the confusion matrix table. Based on Table 3, the accuracy value can be calculated using the following calculation:

Table 3. *Confusion Matriks Decision Tree*

	true NON AKTIF	true Aktif	class precision
pred. NON AKTIF	95	31	75.40%
pred. Aktif	62	163	72.44%
class recall	60.51%	84.02%	

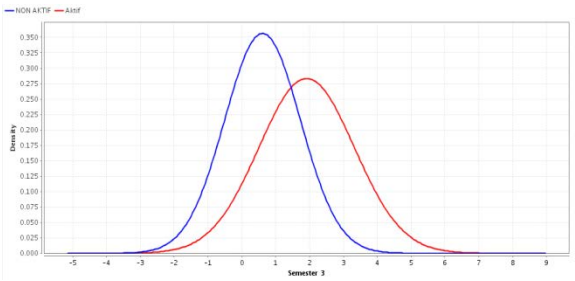


Figure 5. Results of decision tree

From the results obtained from the decision tree method, but the results of the can combine the technique of C4.5 has a number of conclusions, therefore, this technique is used as to classify students who are not active at Campus C based on their characteristics.

4. Conclusion

The description in this paper analyzes the data of the inactive D3 students in two study areas, Information Management (MI) and Computer Accounting (KA), where on campus C there are 351 registered students. The demographic and academic distribution profiles ofinactive students on campus C, show that inactive students generally range between the ages of 19-22 years. The fact that this factor can affect the status of active students at campus C, that the age of students affects the level of readiness of students in studying on campus.The more mature students, the higher the level of readiness of students to learn independently. ± 87% of students are actively working. This confirms the characteristics of college studentsC, who have jobs, and chose to go to college for part-time college. In general, they prefer to learn while working. Through the classes of night and Saturday or Sunday, even thoughthe learning process would be more complex, because of their age, job and family.

Bibliographies

[1] Anonim, Indonesia Open University Portofolio, Indonesia Open University Publishing Center, Jakarta, 2010.

[2] Anonim, Indonesia Open University Catalogue 2013, 2nd ed., Ministry of Education and Culture, South Tangerang, Indonesia Open University, 2013.

[3] J.P. Bean, Student attrition, intentions, and confidence: Interaction effect in a path model, Res. High. Educ. 17 (1982), pp. 291–320.

[4] R.A. Berk, Statistical Learning from a Regression Perspective, Springer Science + Business Media, New York, NY, 2008.

[5] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996), pp. 123–140.

[6] Computer Center-IOU, Data of Open University Students of Registration Period of 2004.1–2012.2, Indonesia Open University, Jakarta, 2013.

[7] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., The Morgan Kaufmann series in data management systems, Morgan Kaufmann, San Francisco, 2006.

[8] S. Islam, Readiness for self learning of Open University and High School Students in Open and Distance Learning Higher Education System in Indonesia, J. Open Dist. Edu. 11(1) (2010), pp. 1–14.

[9] D. Keegan, Theoretical Principles of Distance Education, Routledge, London, 1993.

[10] R. Kohavi and C. Kunz, Option Decision Trees with Majority Votes, Proceeding of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann, 1998, pp. 148–156.

[11] K. Machova, F. Barcak, and P. Bednar, A bagging method using decision trees in the role of base classifiers, Acta Polytech. Hung. 3(2) (2006), pp. 121–132.

[12] R. Manclin and D. Opitz, An Empirical Evaluation of Bagging and Boosting, Proceeding of the Fourteenth International Conference on Machine Learning, AAAI Press/MIT Press, Cambridge, MA, 1997, pp. 546–551.

[13] M.G. Moore, Theory of Transactional Distance, Routledge, New York, NY, 1993.

[14] S. Orr, The organizational determinants of success for delivering fee-paying graduate courses, Int. J. Educ. Manag. 14 (2000), pp. 54–61.

[15] D.P. Rahayu, An analysis of characteristics of OU's non-active students with cluster Encamble approach, unpublished thesis, postgraduate Program, Bogor Agricultural University, Bogor, 2009.

[16] D.J. Ratnaningsih, A. Saefuddin, and H. dan Wijayanto, An analysis of drop-out students' survival in distance higher education, J. Open Dist. Educ. 9(2) (2008), pp. 101–110.

[17] L. Rokach, Ensemble Methods for Classifier in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Springer Science + Business Media, New York, NY, 2008.

[18] R. Schuemer, Some Psychological Aspects of Distance Education, Hagen, Institute for Research into Distance Education, (ED 357 266), Germany, 1993.

[19] N. Soleiman, Continuity of registration and its relation to the examination results, Research Report, Indonesia Open University, Jakarta, 1991.

[20] M. Stephen, Machine Learning, an Algorithmic Perspective, Chapman & Hall/CRC machine learning & pattern recognition series, Boca Raton, FL, 2009.

[21] U.T. Sufandi, Development of student learning progress system prediction based artificial neural networks: Case in Open University, Unpublished thesis, postgraduate Program, Bogor Agricultural University, Bogor, 2007.

[22] S. Taheri and M. Mammadov, Learning the Naïve Bayes classifier with optimization models, Int. J. Appl. Math. Comp. Sci. 23(4) (2013), pp. 787–795.

[23] M.F. Zaman and H. Hirose, Classification performance of bagging and boosting type ensemble methods with small training sets, New Generat. Comput. 29 (2011), pp. 277–292.

[24] DewiJuliahRatnaningsiha andImasSukaesihSitanggang, Comparative analysis of classification methods in determining non-active student characteristics in Indonesia Open University, 2015, pp. 0266-4763