Comparison Of Data Mining Methods For Recipient Prediction Poor Student Assistance (BSM) In MAN 2 North Lampung

Ovi Naeni[1], Resy Anggun Sari[2], Sriyanto[3]

Institute Informatics and Business Darmajaya, Indonesia

sriyanto@darmajaya.ac.id[3]

ABSTRACT

MAN 2 North Lampung is a State Madrasah Aliyah or equivalent to Senior High School which has implemented poor student assistance (BSM), system by considering the economic condition of the students or the criteria that have been set. Selection of BSM acceptance is a semi-structured problem type meaning that this process is not a routine agenda of a school but the agenda held at a certain time that is when students are in class X. Determined the BSM recipient candidate must collect the data file of candidate selection of BSM recipients from students' data coming from poor family to very poor family. So it takes a relatively long time, as well as high accuracy in making decisions. In predicting students who receive BSM, the authors apply the data mining process using the Naive Bayes method, Decision Tree, K-NN. The attributes used consist of siblings, Parent Occupation, parental income, smart indonesian card (KIP) recipients or not, the status of the family as an orphan or not. To perform the process of data mining in need of tools aids that is RapidMiner 5. The Implementation of data mining using a comparison of 3 methods can be seen based on the sample number of 393 students. the results of the precision value of the Naive Bayes method are better used for this study compared to other methods. While based on recall and accuracy values, Decision Tree is better used than other methods. But when viewed from the overall results of BSM receiver predictions, the most influential variable is parent income and receiver the KIP card.

Keywords: Data Mining , Decision Tree, Naive Bayes, and K-NN

1. Introduction

Underprivileged Student Assistance Program also known as BSM (Bantuan Siswa Miskin) introduced in 2008. *BSM* is a collection of tax-financed cash transfers to public school students from underprivileged households. *BSM* programs exist at all public schools across all levels of education and provide currently enrolled students from underprivileged households, selected by school administrators, with an annual cash transfer in one lumpsum installment [2]. One of the schools that implemented BSM program is Madrasah Aliyah Negeri 2 (MAN 2) North Lampung located in Jalan Taruna No 199 Padang Ratu, North Sungkai, District North Lampung.

Through the *BSM* Program it is expected that school aged children from poor households/families can continue to go to school, not dropping out of school, and in the future it is hoped that they can break the poverty chain that their parents are currently experiencing. The *BSM* program also supports the government's commitment to increase education participation rates in poor and remote districts/municipalities as well as to marginalized groups. It is ensured that with BSM every child can continue to go to school without thinking of any cost and can go to school for free.[2]

In the process of Prediction of BSM recipients in MAN 2 North Lampung, is a semi-structured problem type meaning this process is not a routine agenda of a school but the agenda held at a certain time that is when students are in grade X (ten). But there are still problems that often arise, that is less precise the distribution of underprivileged students assistance[1].

To solve the problem can use data mining approach, especially classification techniques. There are some commonly used algorithms such as k-NN, NB, SVM, NN and DT. In this research will compare of DT, NB and k-NN algorithm to get the most precisely algorithm in making prediction for BSM program beneficiaries [3].

Based on these probms the authors decided to conduct a study with Title Comparison 3 Methods In Data Mining for Prediction of *BSM* in State Senior High School 2 North Lampung. The data used as the sample is 393 students in Madrasah Aliyah Negeri 2 North Lampung.

Formulations Problems

The formulation of problem in this research is how to analyze prediction data of underprivileged student recipients with data mining using Naive Bayes method, Decision Tree (C4.5) *K-NN* in determining which criteria influence for BSM recipients.

Research Objectives

The purpose of this research is to compare decision Tree (C4.5), Naive Bayes, K-NN for predicting students who get underprivileged student assistance program.

Literature Review

1. Underprivileged Student Assistance

According to Juknis Year 2014 *BSM* is one of the four compensations that the Government will give to the community.
This program is a national program aimed at removing barriers for poor students participating in school. By helping underprivileged students gain access to appropriate education services, prevent dropping out, attract poor students to go back to school, help students meet the needs of learning activities, support the Nine Years Basic Education (even to upper secondary) school [4].

2. Data Mining

Data Mining is often called Knowledge discovery in database *KDD*is an activity that includes the collection of usage data, historically, to find regularities, patterns or relationships in large data sets.[1] Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions [5]

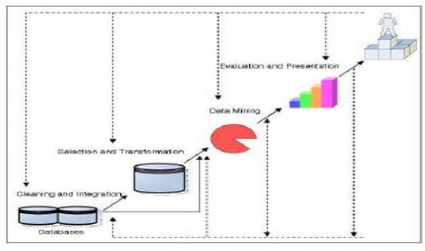

Figure 1. Data Mining is a step in in the KDD process (Han and Kamber, 2001)

3. Classification

Classification is the separation or ordering of objects into classes[7]. There are two phases in classification algorithm:first the algorithm tries to find a model for the class attribute as a function of other variables of the datasets. Next, it appliespreviously designed model on the new and unseen datasets for determining the related class of each record.[8]

4. Naive Bayes

The naive bayes classification is a statistical classification that can be used to estimate the probability of class membership. The Naive Bayes classification is based on the Bayes theorem that has a classification capability similar to decesion trees and neural networks. Nnaive Bayes classification is proven to have high speed and accumulation when applied to the database with good data.[6]

5. Decision Tree (C4.5)

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.[9]
The popular Decision Tree algorithms are ID3, C4.5,CART.The ID3 algorithm is considered as a very simple decision tree algorithm. It uses information gain as splitting criteria. C4.5 is an evolution of ID3. It uses gain ratio assplitting criteria [10].

6. K- Nearest Neighbor *(k-NN)*

The *k-NN* algorithm is a method that uses a supervised algorithm in which the results of a new instance query are classified by the majority of the categories in *k-NN*. The purpose of this algorithm is to classify new objects based on attributes and training samples.[12].

6. Evaluation Technique

a) **Confusion matrics**
is a method that uses a matrix table as in Table 1,if the data set consists of only two classes, one class is considered positive and the other negative [7].

Table 1. Model *Confusion Matrics*

		True Class	
		Positive	Negative
Predicted Class	Positive	True positives count (TP)	False negatives count (FP)
	Negative	False positves count (FN)	True negatives count (TN)

True positives are the number of positive records that are classified as positive, false positives are the number of negative records classified as positive, false negatives is the number of

positive records classified as negative, true negatives is the number of negative records that are classified as negative, then enter the test data.

After the test data is inserted into the confusion matrix Once the data has entered into the confusion matrix then it can be calculated values of sensitivity (recall), specificity, precision and accuracy. To calculate used equation below [7].

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

b) Classifier Evaluation Metrics: Accuracy, Error Rate, Sensitivity and Specificity

Table 2. Classifier Evaluation *Matrics*

A\P	C	-C	
C	TP	FN	P
-C	FP	TN	N
	P'	N'	All

1. Classifier Accuracy, or recognition rate: percentage of test set tuples that are correctly classified
   Accuracy = (TP + TN)/All
2. Error rate: 1 – accuracy, or
   Error rate = (FP + FN)/All
3. Class Imbalance Problem:
   - One class may be rare, e.g. fraud, or HIV-positive
   - Significant majority of the negative class and minority of the positive class
4. Sensitivity: True Positive recognition rate, Sensitivity = TP/P
5. Specificity: True Negative recognition rate Specificity = TN/N [8]

2.7.3 Classifier Evaluation Metrics: Precision and Recall, and F-measures

1. Precision: exactness – what % of tuples that the classifier labeled as positive are actually positive
$$Precision = \frac{TP}{TP + FP}$$
2. Recall: completeness – what % of positive tuples did the classifier label as positive?
$$Recall = \frac{TP}{TP + FN}$$
3. Perfect score is 1.0
4. Inverse relationship between precision & recall
5. F measure (F1 or F-score): harmonic mean of precision and recall
6. Fß: weighted measure of precision and recall assigns ß times as much weight to recall as to precision[8]

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

c) *ROC* Curve

ROC (Receiver Operating Characteristics) curve is a test based on performance. ROC express confusion matrix. The value of the *ROC* curve only consists of 0 to 1. The *ROC* curve value closer to 1 then the better as shown in[7]

Table3.Classfying the accuracy of diagnostic tes

0.90 – 1.00	Excellent Classification
0.80 – 0.90	Good Classification
0.70 – 0.80	Fair Classification
0.60 – 0.70	Poor Classification
0.50 – 0.60	Failure

2. Research Method

This research uses several methods that aim to solve the problems that exist in this research. The method proposed in this study is as follows: Classification is the task of mining data that maps data into class groups. Classification techniques classify data items to predefined class labels, construct the included data classification model, build models used to predict future data trends. Commonly used algorithms include K-Nearest Neighbors, Naïve Bayes Classification, Decision Tree.[11]

Data Preparation/Preprocessing

In this study, there are several steps that must be done on the data that has been obtained. Data or notes and attributes are not all used, but must go through several stages of initial data processing (data preparation). So to get quality data, there are some techniques done as follows:

a) Data Cleaning

Good data and quality are the key to producing quality data, nois data that is still outlier or error, data attribute loss data incomplete value, and inconsistent data inconsistent data in charging attributes.
Stages in data cleaning:

1. Eliminate or identify outliers
2. Completing incomplete or missing values, naive bayes algorithms have their own features that can handle incomplete or missing data,
3. Inconsistent data is fixed.
4. Unravel redundancy caused by data interrogation.

b) Data intergration and transformasion

The next step is integration techniques used to analyze correlation data, redundant attributes and duplicate data, and transformation is used to improve the accuracy and efficiency of the algorithm. Excess naive bayes algorithm is able to process data that is nominal, continuous, and ordinal. Therefore the value of each attribute contained in the dataset does not have to be transformed.

c) Data reduction

Data reduction is the process of reducing the dataset by reducing the number of attributes or records that are not needed for less but still be informative. Obtain representations in reduced data volumes but still obtain similar or similar analytical results and data descriptions that are part of data reduction, an important part of numerical data.

d) Naive Bayes Algorithm

Classification can be defined in detail as an activity of learning or training on the function f maps the vector x into one of several available class y labels. The activity will give the result of a model which is then stored as model [14]. Models created in later learning can be used to determine the label of the class. During the learning process in making the model, we need a learning algorithm such as: *k-NN*, Naive Bayes, Decission Tree.

Naive Bayes Classifier *(NBC)* also referred to as Bayesian Classification is a method of classifying statistics useful for the process of determining the probability of a membership of a class. Bayes's theorem underlies the Naive Bayes Classifier that has similar classification capabilities to the Decision Tree and Neural Network. *NBC* also efficiently, effectively, and reliably handles data noise such as irrelevant attributes. NBC can also overcome large datasets with both variable and continuous attributes [13].

3. Discussion

The following is the result of initial data processing (data preparation) that is Data Cleaning and Data Reduction. then generate the 9 attribute dataset consisting of 3 numeric attributes, 5 category attributes and 1 target label of the 393 student data.

The next step is to collect all BSM receiver prediction datasets into the data test. Based on the overall student dataset where there were 393 student data, with 100% composition for the data test, composite comparison was based on the study.[9] So that the test data owned as many as 393 records.

Decision Tree Implementation

a) Tree

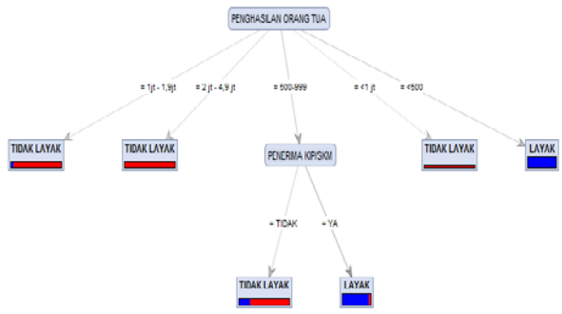

Figure 2. Decision tree

Based on the decision of the tree, it can be concluded that there are 2 classes that are "Eligible" and "Not Eligible" with the following explanation:

Tree

```
PENGHASILAN ORANG TUA = 1jt - 1,9jt: TIDAK LAYAK {LAYAK=3, TIDAK LAYAK=44}
PENGHASILAN ORANG TUA = 2 jt - 4,9 jt: TIDAK LAYAK {LAYAK=0, TIDAK LAYAK=48}
PENGHASILAN ORANG TUA = 500-999
|   PENERIMA KIP/SKM = TIDAK: TIDAK LAYAK {LAYAK=13, TIDAK LAYAK=43}
|   PENERIMA KIP/SKM = YA: LAYAK {LAYAK=120, TIDAK LAYAK=8}
PENGHASILAN ORANG TUA = <1 jt: TIDAK LAYAK {LAYAK=0, TIDAK LAYAK=9}
PENGHASILAN ORANG TUA = <500: LAYAK {LAYAK=105, TIDAK LAYAK=0}
```

Figure 3.Explanation of the decision tree

b) Decision Tree Accuracy

accuracy: 94.15% +/- 2.54% (mikro: 94.15%)

	true LAYAK	true TIDAK LAYAK	class precision
pred. LAYAK	228	10	95.80%
pred. TIDAK LAYAK	13	142	91.61%
class recall	94.61%	93.42%	

Figure 4. Decision Tree Accuracy

a. Based on the above data can be concluded that the result of accuracy of Decision Tree algorithm that is equal to 94,15%

b. Based on the above data, Precision for BSM Eligible Receiver prediction is 95.80%. This is because the number of students selected as many as 228 students with predicted number of students will be elected 238 students. Here's the calculation:

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

$$Precision = \frac{228}{228 + 10}$$

$$Precision = \frac{228}{238}$$

$$Precision = 95{,}80\ \%$$

c. Based on the above data, Recall for selected Caleg is equal to 94.61%. This is because the number of selected siswayang is 228 students with the number of students who actually selected 241 students. Here's the calculation

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

$$Recall = \frac{228}{228 + 13}$$

$$Recall = \frac{228}{241}$$

Recall = 94,61 %

c) Decision Tree *AUC*

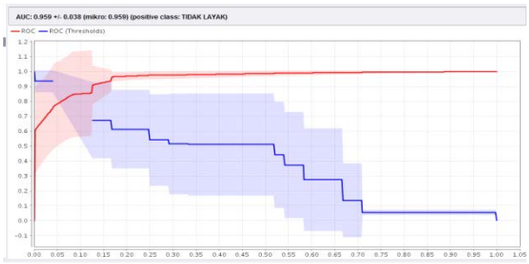

Figure 5.Decision Tree A*UC*

Based on the above data, Decision Tree AUC test shows the number 0.959 so that this test belong to Excellent Classification

AUC

AUC: 0.959 +/- 0.038 (mikro: 0.959) (positive class: TIDAK LAYAK)

Figure 6. *AUC* results

d) Implementation of Naive Bayes using RapidMiner 5

1. Naive Bayes Accuracy

accuracy: 92.12% +/- 4.32% (mikro: 92.11%)

	true LAYAK	true TIDAK LAYAK	class precision
pred. LAYAK	232	22	91.34%
pred. TIDAK LAYAK	9	130	93.53%
class recall	96.27%	85.53%	

Figure 7. Naive Bayes Accuracy

a. Accuracy

Based on the above data can beconcluded that the accuracy of Naive
Bayes algorithm is 92.12%.

b. Precision

Based on the above data, Precision for BSM Receiver prediction is 91.34%.
This is because the number of students selected as many as 232 students with a predicted number of students selected is 254 students. Here's the calculation

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

$$Precision = \frac{232}{232 + 22}$$

$$Precision = \frac{232}{254}$$

Precision = 91,34 %

2. Class Recall

Based on the above data, Recall for Candidate is selected for 96.27%. This is because the number of students selected is 232 students with the actual number of students as many as 241 students. Here's the calculation:

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

$$Recall = \frac{232}{232 + 9}$$

$$Recall = \frac{232}{241}$$

Recall = 96,27 %

3. Naive Bayes AUC

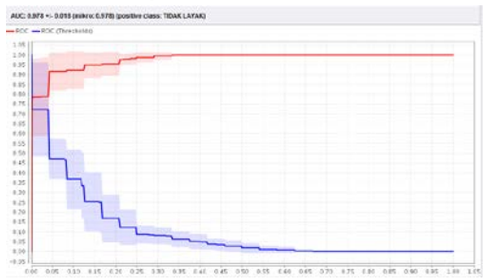

Figure 8. Naive Bayes *AUC*

Based on the data above, testing Naive Bayese AUC shows the number of 0.978 so that this test belongs to Excellent Classification

AUC

AUC: 0.978 +/- 0.018 (mikro: 0.978) (positive class: TIDAK LAYAK)

Figure 9. *AUC*Results

e) Implementation Results K-Nearest Neighbor *(K-NN)* using RapidMiner 5

1. K-Nearest Neighbor *(K-NN)*

accuracy: 91.60% +/- 4.43% (mikro: 91.60%)

	true LAYAK	true TIDAK LAYAK	class precision
pred. LAYAK	227	19	92.28%
pred. TIDAK LAYAK	14	133	90.48%
class recall	94.19%	87.50%	

Figure 10.K-Nearest Neighbor (*K-NN*) Accuracy

- **Accuracy**
  Based on the above data can be concluded that the result of accuracy algorithm K-Nearest Neighbor (K-NN) that is equal to 91,60%.

- **Precision**
  Based on the above data, Precision for BSM Worthy Receiver's prediction is 92.28%. This is because the number of students selected as many as 227 students with the number of students who are predicted to be selected is 246 students. Here's the calculation:

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

$$Precision = \frac{227}{227 + 19}$$

$$Precision = \frac{227}{246}$$

Precision = 92,28 %

- **Class Recall**
  Based on the above data, Recall for Caleg elected is equal to 94.19% This is because the number of students selected are 227 Students with the number of students who actually selected 241 students. Here's the calculation

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

$$Recall = \frac{227}{227 + 14}$$

$$Recall = \frac{227}{241}$$

Recall = 94,19 %

2. K-Nearest Neighbor *(K-NN)AUC*

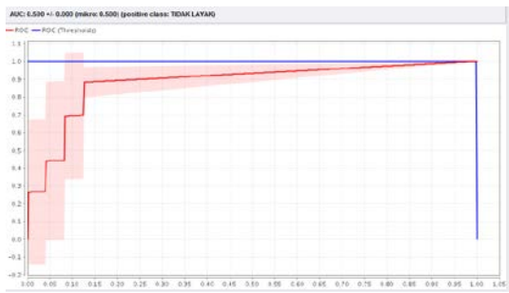

Figure 11. K-Nearest Neighbor (*K-NN) AUC*

Based on the above data, testing 2 k-Nearest Neighbor (*K-NN) AUC* shows the number of 0,500 so that this test belong to Failure.

Figure 12. *AUC*Results

4. Conclusion

BSM Receiver Prediction uses 3 algorithms, yielding Accuracy value as follows: Decision Tree (C4.5) get accuracy value 94,15%, k-NN get accuracy value 91,60% and naive bayes 92,12%.
From the results of research and testing, the method proposed in this research by using the method of Decision Tree (C4.5) can be used to determine the beneficiaries of underprivileged students. and the most influential criteria in determining BSM recipients are the income of parents and the KIP card member.

Bibliographies

[1] Ahmad Ashari. 2013 "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" Universitas Ghajah Mada. Yogyakarta.

[2] AzzahroNadyaEklyma.2016."Penggunaan Dana Bantuan Siswa Miskin (BSM) oleh siswa SMA dan SMK di Kabupaten temanggung",Skrpsi. Fakultas Ilmu pendidikan, Universitas Yogyakarta.

[3] D.T.Wahyuni, T. Sutojo dan A. Luthfiarta, "Prediksi Hasil Pemilu Legislatif Dki Jakarta Menggunakan Naïve Bayes Dengan Algoritma Genetika Sebagai Fitur Seleksi," 2014.

[4] Fried H. manJerome "Data Mining and Statistics: What's the Connection?"Department of Statistics and Stanford Linear Accelerator Center. Stanford University,Stanford, CA 94305.

[5] Jajuli Mohamad, Defiyanti Sofi. 2015, "Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining".Teknik Informatika Fakultas Ilmu Komputer,Universitas Singaperbangsa Karawang. Karawang

[6] Jananto,Arief."Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa" . Volume 18, No.1, Januari 2013 : 09-16 ISSN : 0854-9524.Program Studi Sistem Informasi, Universitas Stikubank

[7] J .Han and M. Kamber, "Data Mining: Concepts and Techniques." Morgan-Kaufmann Publishers, San Francisco, 2001.

[8] Kamber Micheline, Han Jiawei, Pei Jian 2011 "Data Mining: Concepts and Techniques Model Evaluation". University

of Illinois at Urbana-Champaign & Simon Fraser University.

[9] Maimon, O and Rokach, L. “Data Mining and Knowledge Discovery.” Springer Science and Business Media, 2005.

[10] Mahendra, D.C dan Kurniawan, A.W. “Klasifikasi Data Debitur Untuk Menentukan Kelayakan Kredit Dengan Menggunakan Metode Naive Bayes,” 2015.

[11] Nugroho,Sulistyo,Yusuf.2015,”Perbandinga n 3 Metode Dalam Data Mining Untuk Prediksi penerima Beasiswa Berdasarkan Prestasi Di SMA Negeri 6 Surakarta”, Fakultas Komunikasi dan Informatika, Universitas Muhammadiyah Surakarta, Surakarta.

[12] P-N. Tan, M. Steinbach, V. Kumar, “Introduction to Data Mining.” Addison Wesley Publishing, 2006.

[13] Prasetyo, “DATA MINING - Mengolah Data Menjadi Informasi Menggunakan Matlab”, Yogyakarta: ANDI, 2014

[14] Wanghiston.2012,” Bantuan Siswa Miskin Cash Transfers For underprivileged Students social Assistance Program And Public Expenditure Review 5”, Indonesia Stock Exchange Building Tower II/1 2th Floor JI. Jend. Sudirman Kay. 52-53 Jakarta 12910.