International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 2931-2937 www.ijaem.net ISSN: 2395-5252

Naïve Bayes Model for Student Data Analysis

Rofilde Hasudungan[1]
[1]Lecturer, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia

Submitted: 10-07-2021 Revised: 20-07-2021 Accepted: 23-07-2021

ABSTRACT: Student data analysis is an important task to discover phenomena in educational sector. This task involves association, classification, and clustering. Classification is popular technique to discover performance of students and other. Naive Bayes classifier is one of classification algorithm. This algorithm has advantages in the simplicity, speed of execution and accuracy. This paper aim to use Naïve Bayes Classifier to predict students' performance in quiz of subject fundamental programming. One of characteristic of Naïve Bayes is that all variables are equals, however in the real-world data may have many attributes and not all attributes are needed. Furthermore, some of attributes reduce result of classification and it is costly. To overcome this problem, we used rough set-based attribute selection, where the advantage of this technique is that can find a set attributes that have the same power using all attributes. The experiment shown that the combination of rough set-based feature selection and naïve bayes classifier has better result than naïve bayes.

KEYWORDS: Naïve Bayes Classifier, Rough Set Theory, Feature Selection, QuickReduct, Educational Data Mining.

I. INTRODUCTION

Educational data mining (EDM) is an emerging trend that concerned to develop techniques for exploring and analysing data that come from the educational context. In recent years, EDM has proven to be more successful at many of these educational statistics problems, due to enormous computing power and data mining algorithms [1]. There are lot of techniques used to discover hidden pattern in educational data such as Decision Tree [2], [3][4][5][6]; Support Vector Machine and Artificial Neural Network[7]; Naïve Bayes and Bayesian Network [1][2][8] .

Naïve Bayes classifier is a classification method that applied for many fields such as economy, astronomy, image recognition, etc. This method based on bayes theorem where the probability of certain event is based on posterior event. This method considered fast and robust especially when dealing with big data. Naïve Bayes consider all attribute as equal, and that why it called Naïve. However, using all attributes for classification may affect the result such as accuracy, due to some attribute does not contribute to the classification. Furthermore, some attributes may consider as superfluous that make classification process costly.

There are lot of techniques for select the best attributes for example by using decision tree and its variance, Particle Component Analysis (PCA) and Rough Set [9]. This research aims to use rough set as feature selection to select the best attributes from data. Rough set is mathematical tools proposed by Pawlak dealing with vague and inconsistent data [9] as well as incomplete data [10]. The main concept of rough set theory is indiscernible relation and set approximation where this concept enables rough set to approximate the concept of data. The advantages of this method compared to others method such as fuzzy is there are no parameters required since the information about data is gathered from that data itself. In literature, there are lot of technique based on this method to select the best feature from the data, either for classification, or clustering.

II. RELATED WORK

Data mining or Knowledge Discovery in Database (KDD) is a tool used to discover information hidden in data. There are several data mining method such as classification, clustering, outlier detection, and association rules. Fig 1 shown the process to extract data into knowledge.

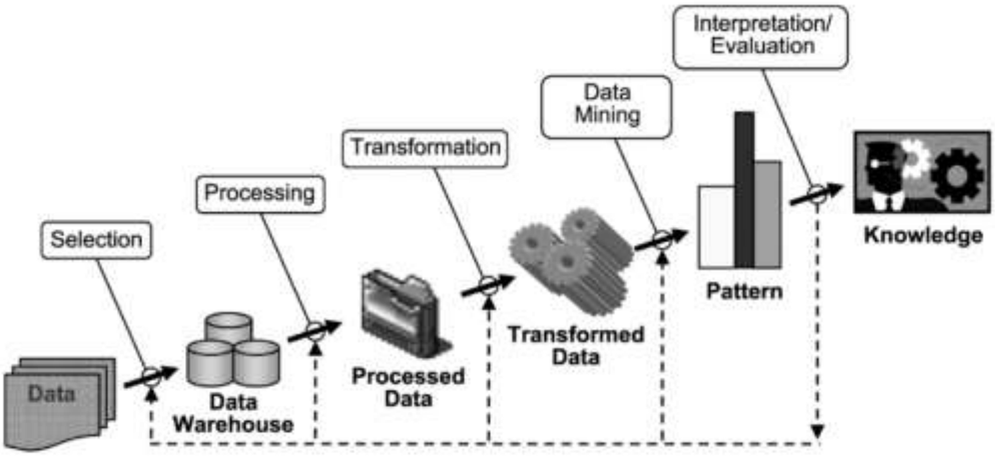


Fig. 1. Data Acquisition

The application of data mining is enormous. It uses in many fields such as astronomy, economy, social, ecommerce, sheath, and education. In education sector, data mining mainly uses for clustering and classification. In clustering, data mining used to group students based on characteristic such as student learn behaviour, personal information, social economy factor, etc. Meanwhile, for classification, data mining mainly used to predict such as GPA, Pass or fail in certain subject, etc. It also used to discover the factor that affects student performance. Data mining in education sector mainly call data mining in education (EDM). This term not only because the source of data, but the objective of EDM is different so that data mining techniques cannot directly applied.

In literature, there are lot of researchers used data mining for analyse education data such as Mokhairi et al. [8] use Naïve Bayes to identify the hidden information between subjects that affected the performance of students in Sijil Pelajaran Malaysia (SPM). This study used to classify students' performance in early stages of second semester. There 488 students involved in this study, where the data collected from 2011 to 2014. This model has accuracy 73.4%. Khasanah and Harwati[2] used Bayesian Network and Decision Tree to identify students' performance (drop or not). This study used 12 attribute and involving 178 students, where after data cleaning there are only 104 students available where 13 students classified as drop and 91 students classified as not. In this study used five attribute selections algorithm to filter attributes. Furthermore, the attributes that appears in most attribute selection algorithm are selected (8 attributes). In this experiment shown that Bayesian Network has accuracy 98.08%. This result better than decision tree with accuracy 94.23%. Baby et al. [7] use three methods to investigate the influence of lecturer pedagogue to student performance. The aim of this study is to give feedback to the instructor so he/she can improve pedagogue skill. Based on experiment found that the decision tree (C4.5) has better accuracy that Support Vector Machine (SMO) and Artificial Neural Network (MPL). The accuracy of C4.5, SMO, and MLP are 94.37%, 90.85%, and 92.96%, respectively.

Al-Barrak and Al-Razgan[3] predict student final GPA based on their grade in previous courses by using decision tree (J48) at King Saud University. By using decision tree found the courses that affect student GPA. The courses are Java1, Database principles, software engineering 1, information security, computer ethic and project 2. Ahmed and Elaraby [4] predict final score for subject by using factors such as assignments, homework, mid semester, seminar, participation, and attendance. By using decision tree, we extract rules from data and found that mid semester is the main factors that affect final score. Lakshmi et al. [5]compare three decision tree method (1) C4.5, (2) ID3, and (3) CART to predict student performance. In this research, indicators such as (1) parent's education, (2) living location, (3) economy status, and (4) friends and family support, (5) resources accessibility, and (5) attendance are used to classify student into first, second, third or fail. This study found that the accuracy for CART, C4.5 and ID3 are 55.83%, 54.17% and 50%, respectively. Adhatrao[6]predict performance of enrolled students by using prior knowledge in classes X and XII in high school by using C4.5 and ID3. By using this model one can predict promising students and improve those who would probably get lower grades. In this study found that either C4.5 and ID3 have same accuracy 74.145%, however C4.5 has advantages for faster execution time that ID3 where for 182 students C4.5 executes in 39.1 millisecond compares 47.6 millisecond for ID3. Giap et al. [11] investigate several influence factors such as family, school, and community toward student performance by using decision tree. There are 33 factors, and 425 students involves in this research. Based on experiment found that 12 factors from 33 factors affects student performance. The factors are sex, mother education, father education, family size, motivation, study time, failure in math test, family support, math course, join nursery course, alcohol consumption in weekend, and math score in second years

III. RESEARCH METHOD

To solve the problem, this study develops research methodology as depicted in **Error! Reference source not found.**. There are several stages in this methodology: (1) Data collection, (2) Feature selection, (3) Data Analysis and (4) Evaluation.

A. Data Collection

This study uses a dataset that obtained from a questioner for subject Fundamental Programming in department of informatic, faculty science and technology, Universitas Muhammadiyah Kalimantan Timur. The questioner contains a set of questions that ask the student about performance of the lecturer that teach the subject. Furthermore, the data taken from questioner are combined with the quiz result of each student.

B. Feature Selection

Feature selection used to reduce features by eliminating unnecessary or superfluous features. In this stage, we employed rough set. Rough set is mathematical tools proposed by Pawlak [9] for deal with vague and uncertainty. This tool based on indiscernibility relation and set approximation.

1. Indiscernible Relation. Suppose we have R a family of binary relation call equivalence relation, indiscernible relation can be defined as equation (1).

$$\mathrm{IND}(B) = \{x, y \in U \mid \forall a \in B \rightarrow a(x) = a(y)\} \quad (1)$$

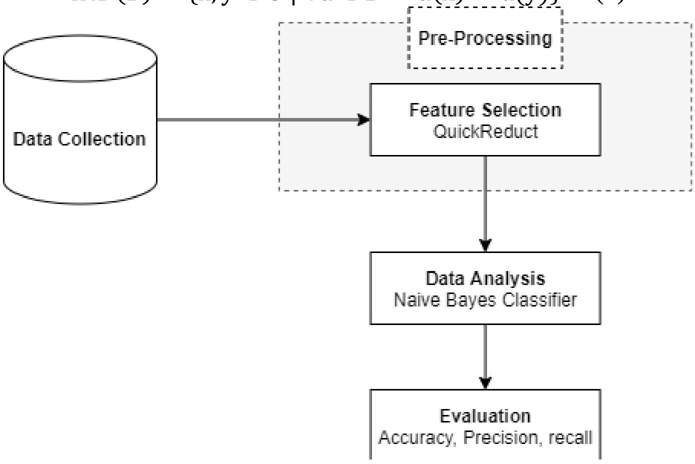


Fig. 2. Research Methodology

where U is finite set of objects called universe, $x, y \in U$, and B is set of attributes.

2. Set Approximation. Set approximation are used to “approximate” a certain concept. There are two approximations: (1) lower approximation and (2) upper approximation that defined in equation (2) and equation (3), respectively.

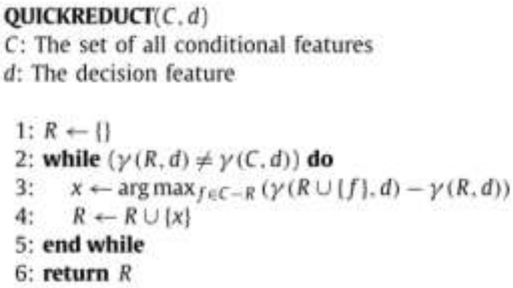
$$\underline{B(X)} = \{x \in U \mid [x]_B \subseteq X\} \quad (2)$$

$$\overline{B(X)} = \{x \in U \mid [x]_B \cap X \neq \emptyset\} \quad (3)$$

3. Attribute Dependency. Through set approximation attribute dependency in equation (4).

$$\gamma_D^C = \frac{|POS_C|}{|U|} = \sum_{X^* \subseteq X} \frac{\underline{\underline{B(X^*)}}}{|U|} \quad (4)$$

4. Reduct. Suppose where have conditional attributes R, and there are subsets $R \subseteq C$. Reduct used to find R where attribute dependency $\gamma_D^R = \gamma_D^C$. There are lot of techniques of reduct and this study usedQuickReduct algorithm as shown in Fig. 2.



```
QUICKREDUCT($C, d$)
$C$: The set of all conditional features
$d$: The decision feature

R ← {}
while ($\gamma(R, d) \neq \gamma(C, d)$) do
    $x \leftarrow \arg\max_{f \in C-R} (\gamma(R \cup \{f\}, d) - \gamma(R, d))$
    $R \leftarrow R \cup \{x\}$
end while
return R
```

Fig. 2. QuickReduct Algorithm

C. Data Analysis

Naïve bayes classifier is a classification method that can be used to predict the probability the membership of the class. This method based on Bayes theorem that provided a way to calculate the probability of a prior event by using another subsequent event has occurred. The main formula of the Bayes theorem is given as bellow:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (5)$$

Where X is data with unknown class, H is the hypothesis of X data is a specific class, $P(H|X)$ the probability of hypothesis H is based on X condition, $P(H)$ is H hypothesis probability, $P(X|H)$ is probability X under these conditions, $P(X)$ is the probability of X.

Naïve Bayes classifier is one of the most simple but sophisticated technique based on Bayes theorem. This technique assumes that all features all independence to each other that why it called Naïve Bayes. Naïve Bayes classifier has several stages as follows ([12]):

1) Let D be training set of tuples and their associated class labels.
2) Suppose that there are m classes, $C_1, C_2, \dots, C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, condition on X. Naïve bayes classifier predict that object X belongs to class C_i if only if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, n \neq i$. $P(C_i|X)$ is calculated by using following equation:
$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (6)$$
3) As $P(X)$ is constant for all classes, only $\mathrm{P(X|C_i)P(C_i)}$ need to be maximized.
4) Calculate probability of $\mathrm{P(X|C_i)}$ by using following equation:
$$\mathrm{P(X|C_i)} = \prod_{\mathrm{k=1}}^{\mathrm{n}} \mathrm{P(x_k|C_i)} \qquad (7)$$
5) To predict the class label of X, $\mathrm{P(X|C_i)P(C_i)}$ is evaluated for each class $\mathrm{C_i}$. The classifier predicts that the class label of tuple X is the class $\mathrm{C_i}$ if and only if $\mathrm{P(X|C_i)P(C_i)} > \mathrm{P\big(X\big|C_j\big)P(C_j)}$ for $\mathrm{i \leq j \leq m, j \neq i}$.

D. Evaluation

To measure proposed method, this study employed tenfold cross-validation which means that the experiment will be conducted in ten iterations. By using this schema, we will find True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Furthermore, these values will be used to calculate performance of the model such as accuracy, precision, and recall. The accuracy, precision and recall define by equation (8), equation (9), and equation (10) respectively.

$$\mathrm{accuracy} = \frac{\mathrm{TP + TN}}{\mathrm{TP + TN + FP + FN}} \qquad (8)$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP + FP}} \qquad (9)$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP + FN}} \qquad (10)$$

IV. RESULT AND DISCUSSION

A. Rough Set as Feature Selection

By using rough set, the selecting the condition attributes are conducted. Based experiment found that attributes dependency γ_D by using all attributes is =1. Furthermore, QuickReduct is run to find the subset of attributes that have similar value with γ_D. Based on QuickReduct, we found there are 6 attributes can be used to represent all attributes as described in Table 1.

Table 1: Attributes after QuickReduct

Role	Attributes
Object id	NIM
Condition attributes	P1
	P2
	P3
	P6
	P7
	P14
Decision attributes	Result

B. Classification using Naïve Bayes Classifier

In this process, we are using RapidMiner to run Naïve Bayes. To validate our approaches, cross-validation is employed, with k=10. Based on these experiments, two confusion matrices are built as shown in

Table 2and Table 3 for Rough Set + Naïve bayes and Naïve bayes, respectively.

Table 2: Confusion Matrix for Naive Bayes Classifier with QuickReduct Feature Selection

Variable	True Fail	True Pass
Pred. Fail	33	5
Pred. Pass	4	5

Table 3: Confusion Matrix for Naïve Bayes Classifier

Variable	True Fail	True Pass
Pred. Fail	27	5
Pred. Pass	10	5

As shown in Table 3 and Table 4, there is improvement for students who are predicted fail in quiz from 27 to 33 students. However, for students who are predicted pass still same. Detail comparison for the value of confusion matrix shown in Figure 2.

C. Evaluation

Based on value in confusion matrix for each model, we calculated accuracy, class precision, and recall by using equation (8), equation (9), and equation (10) respectively. The comparison between Rough Set+Naïve Bayes and Naïve Bayes shown in Fig. 3. Fig. 3 shown that by selecting features and give them to the Naïve Bayes Classifier it can increase the accuracy significantly from 68.09% to 80.85%, and class precision fail from 84.38 to 86.64%. The improvement is due to the number of students that predicted fail increase from 27 to 33 closer to real data. However, the students who are predicted pass does not change.

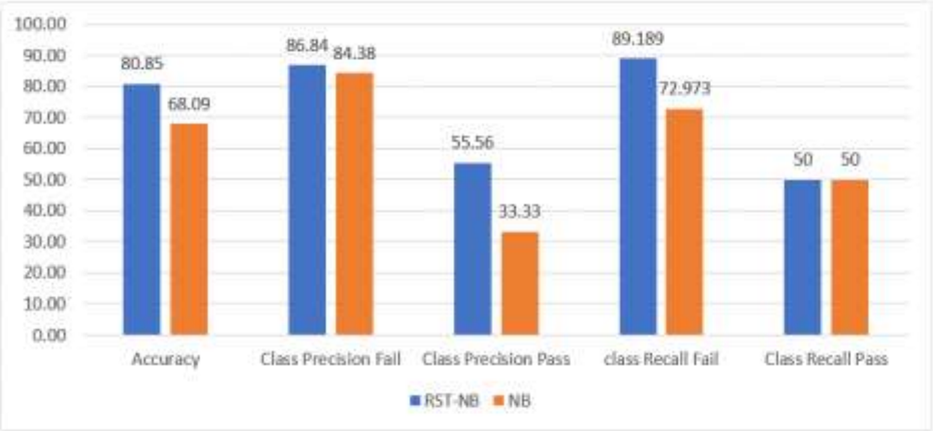


Fig. 3. Comparison between RST-Naive Bayes and Naive Bayes

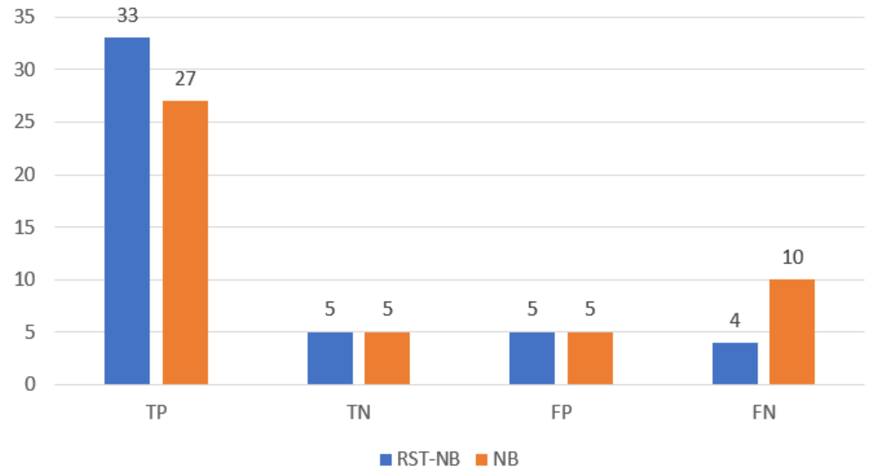


Fig. 4. Confusion Matrix Values Comparison

V. CONCLUSION

In this study, rough set and naïve bayes classifier are proposed to analyse student data. There 47 students involve in this study, where the data are collected by using questioner and quiz result for subject fundamental programming. There are 28 attributes used by a student in questioner to measures the lecturer that teach the subject and there is one attribute contains score for that student in that subject. By using rough set, attributes are reduced from 28 attributes to 6 attributes. Furthermore, naïve bayes classifier is used to analyse the dataset, and through experiment shown that the proposed model has significant performance (accuracy, precision, and recall).

REFERENCES

[1] H. Shaziya, R. Zaheer, and G. Kavitha, "Prediction of Students Performance in Semester Exams using a Naïve bayes Classifier," Int. J. Innov. Res. Sci. Eng. Technol., vol. 4, no. 10, pp. 9823–9829, 2015, doi: 10.15680/IJIRSET.2015.0410072.

[2] A. U. Khasanah and Harwati, "A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques," IOP Conf. Ser. Mater. Sci. Eng., vol. 215, no. 1, p. 12036, Jun. 2017, doi: 10.1088/1757-899X/215/1/012036.

[3] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," Int. J. Inf. Educ. Technol., vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/ijiet.2016.v6.745.

[4] A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method," World J. Comput. Appl. Technol., vol. 2, no. 2, pp. 43–47, Feb. 2014, doi: 10.13189/wjcat.2014.020203.

[5] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, Jun. 2013, doi: 10.5815/ijmecs.2013.05.03.

[6] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, and V. Honrao, "Predicting Students' Performance Using ID3 And C4.5 Classification Algorithms," Int. J. Data Min. Knowl. Manag. Process, vol. 3, no. 5, pp. 39–52, Oct. 2013, doi: 10.5121/ijdkp.2013.3504.

[7] A. M. Baby, "Pedagogue Performance Assessment (PPA) using Data mining Techniques," IOP Conf. Ser. Mater. Sci. Eng., vol. 396, no. 1, p. 012024, Aug. 2018, doi: 10.1088/1757-899X/396/1/012024.

[8] M. Mokhairi, H. Nawang, and S. N. Wan, "Analysis on Students Performance Using Naïve," J. Theor. Appl. Inf. Technol., vol. 31, no. 16, pp. 3993–4000, 2017, [Online]. Available: www.jatit.org.

[9] Z. Pawlak, "Rough sets," Int. J. Comput. Inf. Sci., vol. 11, no. 5, pp. 341–356, 1982, doi: 10.1007/BF01001956.

[10] M. Kryszkiewicz, "Rough set approach to incomplete information systems," Inf. Sci. (Ny)., vol. 112, no. 1–4, pp. 39–49, 1998, doi: 10.1016/S0020-0255(98)10019-1.

[11] Y. C. Giap, N. Leonardi, B. Waseso, and R. Rahim, "Data Mining of Family, School, and Society Environments Influences to

Student Performance," IOP Conf. Ser. Mater. Sci. Eng., vol. 420, p. 012090, Oct. 2018, doi: 10.1088/1757-899X/420/1/012090.

[12] S. Agarwal, Data mining: Data mining concepts and techniques. MK, 2014.