Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 9, No. 4, Desember 2018

ISSN 0216 – 0544
e-ISSN 2301– 6914

CAN K-NEAREST NEIGHBOR METHOD USED TO PREDICT SUCCESS IN INDONESIA STATE UNIVERSITY STUDENT SELECTION

[a]Harits Ar Rosyid, [a]Aris Maulana, [a]Utomo Pujianto*

[a,b,c]Electrical Engineering Department, State University of Malang, Indonesia
E-mail: [a] harits.ar.ft@um.ac.id, [b] maulanaaris221@gmail.com, [c]utomo.pujianto.ft@um.ac.id

Abstract

Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) is one of the selection pathways for student admissions to enter state universities (PTN) in Indonesia. This study aims to predict the chance of being accepted in the desired PTN and the lack of early monitoring of students for SNMPTN. The data source from the grades reports card of SMAN 1 Pakong, SMAN 8 Kediri, and SMAN 1 Pamekasan by using the average input of compulsory subjects, majors (Science / Social Sciences) and semester 1 to semester 5 which later the output to be accepted or not accepted An imbalanced dataset potentially affect the performance of the classification method used. Hence, we need to eliminate the imbalance class using SMOTE. Using 10-fold cross validation, this study compared K-Nearest Neighbor (KNN) without SMOTE and K-NN with SMOTE. The goal is to find the best prediction model between the two methods. The prediction model is applied to software for teachers to monitor student grades and ensuring students to pass the SNMPTN. The results show that KNN without SMOTE has higher accuracy than KNN with SMOTE. However, KNN with SMOTE outperform than KNN without SMOTE in precision and recall, KNN with SMOTE with K = 3 reached 80.08% Accuracy, 74.42% Precision and 91.68% Recall

Key words: SNMPTN, K-NN, SMOTE, CLASSIFICATION.

INTRODUCTION

Most of Indonesia high school student has the desire to continue their education at state universities (PTN). There are three pathways for high school students in Indonesia to be admitted to public state universities. They are SNMPTN, SBMPTN and special admission (Mandiri). One of the pathways for admission to student admissions to be accepted in public universities is SNMPTN and conducted simultaneously throughout Indonesia. Unlike the SBMPTN pathway or independent pathway, SNMPTN is an entrance for student admissions that is highly desired by high school students because this entrance does not require examinations or written tests in order to enter PTN. In 2011, of all students who enrolled only 20% are accepted through SNMPTN from all over Indonesia, for this reason, SNMPTN is a strict and prestigious entrance selection [1].

SNMPTN research has been conducted using the K-Nearest Neighbor (K-NN) method to predict SNMPTN acceptance for high school students in Indonesia. The dataset used is SMAN 8 Jakarta alumni from 2013 to 2017 in the form of alumni report cards and groups accepted and not accepted in SNMPTN. The report card score (each semester) is used for training data and making models that will be used to predict, the data used as input is the average value from semester 1 to 5. From the dataset used only those accepted at the University of Indonesia and the Institute of Technology Bandung as scoope research. The data labels are the information accepted at one of the faculties of the University of Indonesia, faculty of the Institute of Technology Bandung or not accepted. Furthermore, the dataset is divided into two groups, named Science and Social Sciences majors. In the Science Department, there are 200 alumni data while there are 100 alumni data for Social Studies majors which are used as training data. the results obtained in this study are acceptable, however, precision and recall it is not shown, and moreover can be optimized because the data used is not balanced [1]. So, there is a room for improvement such as using SMOTE to balance the data distribution.

In a study that compared the performance between naive bayes and K-NN. Using the dataset the nominal attribute of the study resulted in KNN having a better performance than Naive Bayes. Accuracy results of naive bayes 87.24% while K-NN has the best accuracy of 90.55%[2].

Previous studies using the SMOTE method to eliminate imbalance classes in the credit card fraud dataset. The dataset after the SMOTE process increase to 36,605 consisting of 23,347 positive classes and 13,258 negative classes. This study using the K-NN classification method to compare the performance of KNN with SMOTE and KNN without SMOTE. The K-NN method with unbalanced dataset produce poor performance of G-Mean and F-Measure. When the dataset is balanced using SMOTE, the classification performance improved. This study proved that the classification of SMOTE for the imbalance dataset highly recommended to the minimalized overfitting problem [3].

In this study, the prediction of the acceptance path of SNMPTN uses the average value of compulsory subjects, majors and the average semester 1 through semester 5. In Indonesia, there is still little research that helps high school students in predicting student opportunities to enter SNMPTN. Therefore, this study was made for the prediction of acceptance in the SNMPTN pathway. The dataset used is from SMAN 1 Pakong alumni report card data in 2013 and 2014, as well as an alumni data report of SMAN 1 Pamekasan in 2016 and SMAN 8 Kediri in 2016 and 2017. In the process of implementing the system, failure comes from the user, not the technical factors [4]. This study uses the K-NN algorithm without SMOTE and K-NN with SMOTE, the results of this research prediction will be labels accepted or not accepted.

MATERIAL AND METHODS

The field of research that develops and studies algorithms that can learn and make predictions from the data called machine learning [5]. In Machine Learning based on the method of applying Machine Learning has three divisions including Supervised learning, unsupervised learning, and Semi-supervised learning. Machine Learning methods that need help to run an algorithm are called Supervised learning. For unsupervised learning is a Machine Learning method that results from the actions of the computer itself. While semi-supervised learning is a Machine Learning method where not all data is labeled or has a label [6]. The results of machine learning can also be applied in educational games [7] and the fields of education [8]. Problems that can usually be

solved Machine learning include regression, clustering, and classification. the method of classifying data that has been determined by class is called classification [9]. For classification algorithms can use K-NN. The K-NN algorithm without SMOTE and KNN with SMOTE will be used in this study.

SMOTE

The simplest strategy that can be used in the case of unbalanced data is Random over-sampling, where the workings of this method balance the class by replicating the minority class to equal the majority class. Although this Random over-sampling method looks effective, this method can increase overfitting because data created duplicates from minority class data. To avoid overfitting, the SMOTE technique is conducted [10].

The SMOTE technique is used to solve a class imbalance problems. the workings of SMOTE in making new syntheses by using space features rather than duplicating data. The SMOTE technique produces a new synthesis by utilizing the distance between a sample of minorities and the nearest neighbor from a minority sample. The distance between the two samples is made as much new synthesis as needed so that the data becomes balanced [11].

Assume that the minority class dataset is a sample, the oversampling level is N and the nearest neighbor point is K. Calculation steps with SMOTE [12]:

1. Determine the K value of the nearest neighbor sample for each sample in the minority class sample dataset.
2. Select N samples randomly from each of the closest neighbors.
3. Calculate the new sample using formula (2.1) from the minority sample class and each sample N is a new synthesis and then a new synthesis is added to the sample data from the minority class.

$$SynthesisSample[newIndex] = Sample[i] + Random * (neighbor\ sample[i] - Sample[i]) \quad (1)$$

Description :

- SynthesisSample = new synthesis sample.
- Sample[i] = dataset samples of minority classes.
- i = the number of minority class samples.
- Random = random number value between [0,1].
- Neighbor sample[i] = sample the closest neighbor from the sample [i].

4. Repeat the process above until all minority class samples meet the requirements.

By applying this method, the selection of random points along the line segments between the two samples will be a new synthesis. Using this technique SMOTE can expand the decision area for minority classes [11]. Because in the case of SNMPTN datasets obtained are not balanced between accepted and not accepted, so this the data preprocessing stage is to balance the distribution of minority and majority label classes needed in this case. The next step for the dataset SMOTE results is calculated using K-NN.

K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) algorithm is a flexible method and a simple machine learning algorithm, although a is simple, it can classify test data into label classes by looking for data values that have characteristics similar to training data [13]. K-NN is also one of the best techniques for classifying data and can get high accuracy. The classification of this algorithm uses the distance between test data and training data[14]. The distance between test data and training will be calculated using Euclidean distance. Based on the similarity of characteristics between the test data and training data, the label will be determined. The following are the steps for classification using KNN [1]:

1. Specify value K.
2. Calculate the distance between datasets and training data using the formula.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (2)$$

Description :

- Xi = Training data value
- Yi = Test data value
- i = Data variable
- k = Data dimensions

3. Sort neighbor values based on the smallest value to the largest value.
4. Select as many neighbors as K from the sorted list.
5. Determine the value of the test data based on the most neighboring values.

Because the dataset used in this study has more numerical attributes, then K-NN is more suitable for processing numerical data.

Dataset

The data used in this study are collected by making a valued archive that will be registered when SNMPTN, the data is stored by the counseling teacher of SMAN 1 Pakong, SMAN 8 Kediri, and SMAN 1 Pamekasan. Data records values from semester 1 to semester 5, which are accompanied by information accepted or not accepted in SNMPTN registered by the student. This data is obtained from the grades of SMAN 1 Pakong alumni from the 2013 and 2014 classes, for SMAN 1 Pamekasan the data used in the class of 2016, while the last one was SMAN 1 Kediri class of 2016 and 2017. The dataset is 830 instances and 71 attributes, with the number of data classifications of SNMPTN accepted is 132 and those who are not accepted are 698. In the 72 attributes, there are 70 numeric attributes and 2 nominal attributes. The number of attributes to be used is 9 attributes with 7 numeric values and 2 nominal values. The un used data such as grades from semester 1 to 5 are incomplete. Data that has been collected will be processed by using the K-Nearest Neighbor algorithm. By simplifying the attributes that will be used by taking the average value of compulsory subjects, majors and each semester, the attributes used are only 9 as in Table 1. Table 1 is the attribute name used with the data type and a range of values.

Table 1. List Of Attributes In The Dataset

Attribute Name	Attribute Description	Data Type	Range of values
JM	Majoring in class	Nominal	IPA / IPS
AVG1	The average value of semester 1	Numeric	0 – 100
AVG2	The average value of semester 2	Numeric	0 – 100
AVG3	The average value of semester 3	Numeric	0 – 100
AVG4	The average value of semester 4	Numeric	0 – 100
AVG5	The average value of semester 5	Numeric	0 – 100
AVGCS	The average value of compulsory subjects	Numeric	0 – 100
AVGMA	Majors average value	Numeric	0 – 100
KET	Information accepted / not accepted at SNMPTN	Nominal	Accepted/ Not Accepted

This stage, data that initially has an imbalanced label attribute distribution but to be balanced using the SMOTE preprocessing method. The results of the SMOTE are shown in Table 2.

Table 2. Comparison Of Original Data and Data + SMOTE

	Amount accepted	Amount not accepted	Total Data
Original Data	**132**	**689**	**830**
Data+SMOTE	**698**	**689**	**1396**

Confusion Matrix

The evaluation phase of the classification results in this study uses Confusion Matrix. The confusion matrix is a table that contains the amount of the data tested in a study and records how often classified data is true or false [15]. Based on the results of the classification model, later it can show the results of the prediction and classification of this study. The Confusion Matrix model as follows :

Table 3. Confusion Matrix

		Prediction	
		+	-
Actual	+	TP	**FN**
	-	FP	**TN**

Description :

- TP (True Positive): the prediction in this case is TRUE and TRUE reality.
- TN (True negative): the prediction in this case is FALSE and FALSE reality.

- FP (False positive): the prediction in this case is TRUE and FALSE reality.
- FN (False negative): the prediction in this case is FALSE and TRUE reality.

From the Confusion Matrix results, accuracy (3), precision (4), and recall (5) can be calculated with the following formula :

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)\ x\ 100\% \quad (3)$$

$$Precision = TP/(TP + FP)\ x\ 100\% \quad (4)$$

$$Recall = TP/(TP + FN)\ x\ 100\% \quad (5)$$

Usually to calculate the effectiveness and evaluate the performance of classification methods can use accuracy [16]. However to calculate the proportion of true positive predictive cases TP can use precision, while recall is used to calculate the proportion of TP cases that are correctly predicted[17]. The last is the error rate used to calculate the ratio of the amount of data classified incorrectly from the sum of all data [16].

RESULT AND DISCUSSION

Research Result

After going through various stages, the final step that must be done is the process of evaluating the results of classification. therefore In this evaluation process, the results of the KNN without SMOTE will be compared to the KNN with SMOTE. After that, The classification results that have a higher performance will indicate a better algorithm for the SNMPTN acceptance classification based on Score report. A comparison of accuracy can be seen in Figure 1.

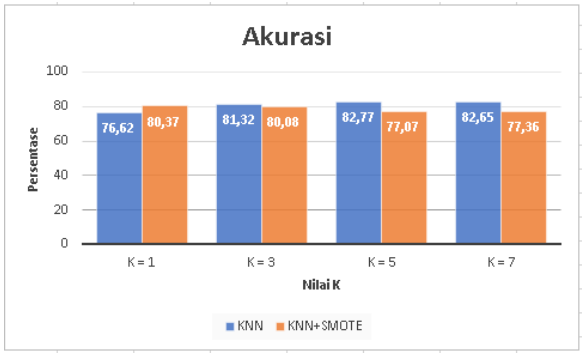

Figure 1. Comparison of Accuracy K-NN without SMOTE with K-NN using SMOTE.

For results Figure 1 shows that KNN without SMOTE has better accuracy than KNN with SMOTE. KNN without SMOTE has the best accuracy of 82.77% when K = 5, while KNN with SMOTE has the best accuracy when K = 1 with a value of 80.37.

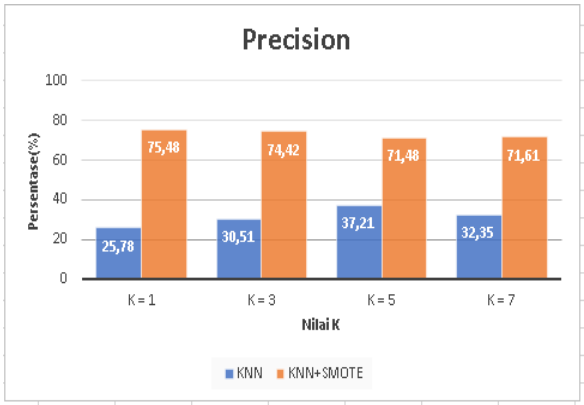

Figure 2. Comparison of Precision K-NN without SMOTE with K-NN using SMOTE.

Conversely results Figure 2 shows that KNN with SMOTE has better precision than KNN without SMOTE. KNN with SMOTE has the best precision of 75.48% when K = 1, while KNN without SMOTE has the best precision when K = 5 with a value of 37.21%. But here KNN without SMOTE has poor performance because the value of precision does not exceed 50%.

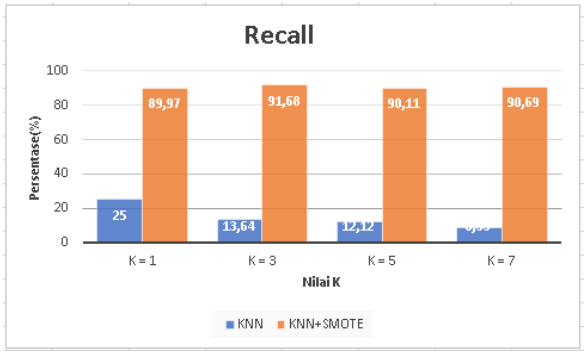

Figure 3. Comparison of Recall K-NN without SMOTE with K-NN using SMOTE.

Table 4. Pearson Correlation Results on Each Attribute

	JM	AVG1	AVG2	AVG3	AVG4	AVG5	AVGCS	AVGMA
JM	1,00	-0,19	-0,43	-0,41	-0,46	-0,43	-0,48	-0,36
AVG1	-0,19	1,00	0,80	0,57	0,53	0,53	0,65	0,55
AVG2	-0,43	0,80	1,00	0,75	0,70	0,65	0,79	0,72
AVG3	-0,41	0,57	0,75	1,00	0,95	0,86	0,94	0,90
AVG4	-0,46	0,53	0,70	0,95	1,00	0,89	0,94	0,88
AVG5	-0,43	0,53	0,65	0,86	0,89	1,00	0,88	0,79
AVGCS	-0,48	0,65	0,79	0,94	0,94	0,88	1,00	0,86
AVGMA	-0,36	0,55	0,72	0,90	0,88	0,79	0,86	1,00

The last results Figure 3 shows that KNN with SMOTE has a better recall than KNN without SMOTE. KNN with SMOTE has the best recall of 91.68% when the value of K = 3, while KNN without SMOTE has the best recall when K = 1 with a value of 25.00%. Moreover, KNN without SMOTE has a poor performance because the recall value does not exceed 30%.

Finally, Table 4 shows the result of the Pearson correlation between two attributes (see Table 1 for abbreviation details). The closer the value to one or minus one, the stronger correlation between the pair attributes. In contrast, the closer the value to zero the smaller the correlation between these attributes.

The results in Table 4 indicates that school majors (JM) do not have a strong correlation with the average score of compulsory subjects, majors or each semester. While the average value of each semester with the following semester grades has a strong correlation, showed by the values constantly above or equal 0.7.

For the average value of compulsory subjects must have a strong correlation with the average value of semester 3 and semester 4 with results more than 0.9. While for the average value of major subjects have a strong correlation with the average value of semester 3, semester 4 and semester 5. Conclusions for the characteristics of the dataset, there must be an increase between the average value of the semester with the value of the next semester, because the results of Pearson correlation calculations have a strong correlation and vice versa JM in the Pearson correlation calculation, does not have a strong correlation at each semester average value.

Discussion

The result is almost the same because only the values of TP and TN are used, thus it has high accuracy results between balanced and unbalanced data. So that FP and FN are not used in calculating accuracy, therefore to prove the comparison of the performance of KNN without SMOTE and KNN with SMOTE the calculation of precision and recall needed to verify the best method. The precision calculation (2.3) uses the TP formula divided by TP plus FP which used to calculate the proportion of positive prediction cases (FP) which a truly worth (TP), hence that when the data not balanced and more data False (Not accepted), the prediction model will be more TN and will give small precision results.

The SMOTE function here used to add data that's worth True (Accepted) which results in a large number of TP values so that precision results will be better. And finally in the Recall calculation (2.4) using the formula TP divided by TP plus FN which used to calculate the proportion of TP cases predicted correctly, hence that when the data not balanced and more data False (Not accepted), the prediction of the model will be more TN and will provide a

lower recall result. The SMOTE function here the still same used to add data that worth True (Accepted) which results in a large number of TP values so that precision results will be higher.

From the results above it can be concluded that KNN without SMOTE has better accuracy than KNN with SMOTE, but the performance of KNN without SMOTE is very bad because the value of precision and recall is very small compared to KNN with SMOTE. This is because the KNN without SMOTE has unbalanced data distribution in the label class, so that more often classifying label classes are not accepted compared to the label class accepted. It is different from results KNN with SMOTE, where the results have a higher value of precision and recall because the SMOTE method makes balanced data distribution by increasing minority classes (Accepted) so that the KNN method can classify data correctly and balanced.

CONCLUSION

For the results of the correlation between attributes in the dataset, it has a strong correlation between the average value of the semester with the next semester, majors do not have a strong correlation between grades each semester. Therefore, there must be an increase between the average value of each semester with the value of the next semester because the results show a strong correlation.

The comparison between KNN without SMOTE and KNN with SMOTE for the case predicts the SNMPTN acceptance pathway in this study, More suitable using the KNN with SMOTE and with value K = 3 the results obtained are good because the accuracy reaches 80.08%, Precision 74.42% and Recall 91.68%. For further research, it can be developed using other algorithms and is expected to be able to obtain higher accuracy results, so that the implementation of predicting SNMPTN acceptance pathways can be used by students as a consideration when the select state universities (PTN).

REFERENCES

[1] A. T. Wibowo and D. Fitrianah, "a K-Nearest Algorithm Based Application To Predict Snmptn Acceptance for High School," *Int. Res. J. Comput. Sci.*, vol. 5, no. 1, pp. 9–20, 2018.

[2] M. J. Islam, Q. M. J. Wu, M. Ahmadi, and M. A. Sid-Ahmed, "Investigating the Performance of Naïve- Bayes Classifiers and K- Nearest Neighbor Classifiers," *J. Converg. Inf. Technol.*, vol. 5, no. 5, pp. 133–137, 2010.

[3] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE Dan K-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[4] P. A. Santoso, A. P. Wibawa, and U. Pujianto, "Internship recommendation system using simple additive weighting," *Bull. Soc. Informatics Theory Appl.*, vol. 2, no. 1, pp. 15–21, 2018.

[5] M. Vahdat, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, "Can Machine Learning explain Human Learning ?," *Neurocomputing*, 2015.

[6] H. Y. Chen, C. H. Chuang, Y. J. Yang, and T. P. Wu, "Exploring the risk factors of preterm birth using data mining," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5384–5387, 2011.

[7] H. Ar Rosyid, M. Palmerlee, and K. Chen, "Deploying learning materials to game content for serious education game development : A case study," *Entertain. Comput.*, vol. 26, no. March 2017, pp. 1–9, 2018.

[8] M. D. Jaelani, A. P. Wibawa, and U. Pujianto, "Technology acceptance model of student ability and tendency classification system," *Bull. Soc. Informatics Theory Appl.*, vol. 2, no. 2, pp. 47–57, 2018.

[9] A. S. B. Asmoro, W. S. G. Irianto, and U. Pujianto, "Perbandingan Kinerja Hasil Seleksi Fitur pada Prediksi Kinerja Akademik Siswa," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, pp. 84–89, 2018.

[10] R. A. Mollineda, V. Garcia, J. S. Sanchez, and R. Martin-felez, "Surrounding

neighborhood-based SMOTE for learning from imbalanced data sets," *Prog Artif Intell*, vol. 1, pp. 347–362, 2012.

[11] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res. 16*, vol. 16, pp. 321–357, 2002.

[12] H. Li, D. Pi, and C. Wang, "The Prediction of Protein-Protein Interaction Sites Based on RBF Classifier Improved by SMOTE," *Math. Probl. Eng.*, vol. 2014, pp. 1–7, 2014.

[13] H. S. Khamis, K. W. Cheruiyot, and S. Kimani, "Application of k- Nearest Neighbour Classification in Medical Data Mining," *Int. J. Inf. Commun. Technol. Res.*, vol. 4, no. 4, pp. 121–128, 2014.

[14] A. Giri, M. V. V. Bhagavath, B. Pruthvi, and N. Dubey, "A Placement Prediction System using k-nearest neighbors classifier," *Proc. - 2016 2nd Int. Conf. Cogn. Comput. Inf. Process.*, pp. 3–6, 2016.

[15] P.-N. Tan, M. Steinbach, and Vipin Kumar, *Introduction to data mining*. 2006.

[16] M. Junker, R. Hoch, and A. Dengel, "On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy," *Proc. Fifth Int. Conf. Doc. Anal. Recognit.*, 1999.

[17] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.