Data Sharing Technique Modeling for Naive Bayes Classifier for Eligibility Classification of Recipient Students in the Smart Indonesia Program

Mustakim\textsuperscript{1,2,3,5}, Siti Syahidatul Helma\textsuperscript{2}, Ulya Ramadhani\textsuperscript{3}, Achmad Daengs GS\textsuperscript{4}, Rice Novita\textsuperscript{5}, Nuryanti\textsuperscript{6}, Sri Rahmawati Fitriati\textsuperscript{7}

\textsuperscript{1,2,3,5}Departement of Information System, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
\textsuperscript{1,2,3,5}Puzzle Research Data Technology, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
\textsuperscript{4}Faculty of Economic, Universitas 45 Surabaya, Surabaya, Indonesia
\textsuperscript{5}Departement of Syariah Economic, Faculty of Syariah and Law, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
\textsuperscript{6}Universitas PGRI Adi Buana Surabaya, Surabaya, Indonesia

*mustakim@uin-suska.ac.id

Abstract. The objective of Smart Indonesia Program (Program Indonesia Pintar: PIP) is to help school-aged people from poor / vulnerable / priority families to continue to receive education services to graduate from secondary education, both through formal and non-formal education channels. In its implementation, there are still many fraudulent in the process of nominating proposal PIP funds and there are still many prospective students who should not receive PIP because they do not meet the technical guidelines provided by the Ministry of Education and Culture to determine the eligibility of prospective recipients of PIP funds can be done by schools and stakeholders, one of them by using classification techniques. One algorithm that is widely used in classification is the Naive Bayes Classifier (NBC) algorithm. In this study three data sharing techniques were used, namely Hold Out 70% training data and 30% testing data, K-Means Clustering, and also 10 Fold Cross Validation. Determination of the best data sharing technique will be determined by looking at the value of Accuracy, Precision, and Recall and also the value of Area Under Curve (AUC) which is illustrated by the Receiver Operating Characteristic (ROC) curve so that the NBC algorithm is generated with 10 Fold Cross Validation has a very good classification level with the values of accuracy, precision, and recall respectively at 97.40%; 100%; and 76.14%.

1. Introduction
The Smart Indonesia Program (PIP) is a collaborative program between the Ministry of Education and Culture, the Ministry of Social Affairs, and the Ministry of Religion of the Republic of Indonesia that targets to help school-age people from poor/ vulnerable/ priority families to continue to receive education services until graduating from secondary education, both through formal and non-formal education channels [1]. PIP fund recipient students will get assistance once a year in the budget, in which the implementation also involves various agencies, including the Provincial Education Office, District/
City Service, Schools, Distribution Agencies and other agencies [2]. One way to assign students to receive PIP funds is based on similar data proposals involving schools or educational units that have been approved and validated by the Education Office. Schools have a role to identify students who have Smart Indonesia Cards and also identify, select, propose and determine students who do not have KIP are in accordance with the priority targets and then update the data on the prospective PIP funds recipients in the Dapodik application completely and correctly, then the proposal will be given to the Provincial Education Office for the levels of Senior High Schools and Vocational High Schools and to the District/ City Education Offices for the level of Primary and Secondary Schools [2]. In its implementation, there are still many fraudulent proposals for prospective recipients of PIP funds.

The Director General of the Ministry of Education and Culture stated that many prospective students should not receive PIP because they did not meet the technical guidelines provided by the Ministry of Education and Culture, this was based on data obtained, there were 17 thousand recipients of PIP funds who were misdirected and did not fulfill the requirements of receiving PIP funds. In 2016, the Education and Culture Office of Kepulauan Meranti District through the Head of the Department of Education, Suwandi said that there were students who received PIP funds from among the capable families. He said the PIP funds were not on target because of the lack of supervision from various parties. The school that submitted data on student candidates for PIP funds to be more selective in submitting PIP recipient data. If this continues, it is feared that it can lead to polemic in the society because many families who are economically obliged to be assisted and deserve PIP funds don't even get it. Data collection and management of funds is needed to determine who is truly eligible to get PIP assistance funds. He also said that the school has a very important role to play in the success of the Smart Indonesia Program, a school that can understand the daily situation of its students. Schools are expected to be able to record prospective recipients of PIP funds until data on recipients of PIP funds is truly valid. To determine the eligibility of prospective recipients of PIP funds, it can actually be done by schools and stakeholders, one of which is by using data mining techniques.

Data mining is a process of finding implied information that is useful and the process of identifying patterns that are meaningful in a large database using computational techniques from statistics, machine learning, and pattern recognition [13]. The essential problem in Data Mining is classification [11]. Classification is a method that can be used to estimate classes on an object whose class / label is unknown [20]. Gorunescu in his book states that the classification process is based on four fundamental components, namely Classes, Predictors, Training Data, and testing data [13]. One well-known algorithm is probabilistic classification based on the Bayes theorem or commonly called Naive Bayes Classifier (NBC) [22]. NBC is the most popular algorithm because it is simple, efficient, and has good performance in data sets [10] and also has a very high level of learning efficiency and can be used to estimate all probabilities in studying training data [7].

The study conducted by Hassan, S., et al in 2011 using the Naive Bayes method and Support Vector Machine (SVM) with 10-fold Cross Validation resulted in the NBC algorithm having a superior improvement over SVM with an improvement value of +28.78% while SVM is +6.36% [19]. Other study conducted by Xhemali, D., Hinde, C.J. and Stone, the 2009 RG for the classification of training web pages using the NBC, Decision Tree, and Neural Networks methods resulted in an accuracy of 95.20% and F-Measure values above 97% for the NBC algorithm which was superior to other algorithms, indicating that NBC algorithm is fast, consistent, easily maintained and accurate in classifying attribute data [23]. Based on the results of the above research, it can be concluded that the Naive Bayes Classifier algorithm is a classification algorithm that is superior when compared to other classification algorithms based on the value of accuracy produced.

The study is to find out how the Naives Bayes Classifier algorithm application to classify eligible PIP grant recipients, in this study also used three data sharing techniques, namely Hold Out 70% training data and 30% testing data, KMeans Clustering, and also 10 Fold Cross Validation. In 2015, Nasution, N, et al. Evaluating the academic performance of the students of the Faculty of Computer Science at Lancang Kuning University using the Naive Bayes Classifier algorithm to produce the highest accuracy in data sharing using Hold Out 70% training data and 30% data testing at 76.67% [21]. Goswami, et al.
(2014) conducted a study to classify documents on the DRDO website using 10 Fold Cross Validation data sharing techniques which produced the highest accuracy in document classification in the technology field with a value of 75.21% [9]. In 2017, Mustakim conducted a study on the comparison of training data and data testing techniques with the K-Means Clustering algorithm which resulted in an accuracy of 93.4% and K-Fold Cross Validation of 77.8% so that the data sharing technique uses the K-Means algorithm more good when compared to K-Fold Cross Validation for classification cases using the K-Nearest Neighbor classification algorithm [3]. Determination of the best data sharing technique will be determined by looking at the value of Accuracy, Precision, and Recall. This research is expected to be able to provide recommendations to schools and other stakeholders in finding solutions to determine the eligibility of prospective recipients of PIP funds, in addition it can find out in the case studied which data sharing techniques are the best to be implemented into a system.

2. Methodology
The methodology in this study to starting research, data collection, preprocessing, data sharing technique (hold out, K-Means and K-FCV), processing NBC algorithm and conclusion the best classification. The data used is data from students of Public High School SMAN 06 Pekanbaru in 2018 with data of 938 records, consisting of 9 main attributes that influence school recommendations towards students in determining the eligibility of students to receive funding from the Smart Indonesia Program. The parts of each attribute are Gender, Type of Residence, Transportation Equipment, Social Card Recipient, Parental Work, Parent Income.

While the label class used in the classification process is the label "Yes" which means Eligible receiving PIP funds and "No" which means Ineligible to receive PIP funds based on recommendations from the school. In the preprocessing stage, the Cleaning, Transformation and Normalization process of the data is carried out so that the final data used is clean, numerical, and has values ranging from 0 to 1.

2.1. Naïve Bayes Classifier
NBC is a classification algorithm based on the Bayes theorem with an independent assumption [7] which is used to predict data as accurately as possible [5]. In applying the Naive Bayes classifier method, the following equation is used [20]:

\[
(H|X) = \frac{P(X|H)P(H)}{P(X)} ..................................................(1)
\]

2.2. AUC and ROC
The ROC curve is a technique used to measure the performance of classifiers that cross entire range of class distributions and the error value generated by evaluating the ratings of positive and negative samples using values in the area under the ROC curve or Area Under Curve (AUC) [25]. AUC has values from the range 0-1 [4] Gorunescu (2011) states that the AUC value can be divided into several groups [13] as illustrated: Very good classification (0.90 – 1.00), Good classification (0.80 – 0.90), Fair classification (0.70 – 0.80), Poor classification (0.60 – 0.70) and Fail classification (0.50 – 0.60).

2.3. K-Means Clustering
K-Means is one of the most popular clustering algorithms [6]. K-Means clustering is an unsupervised algorithm that is used to group different objects into groups. The stages of K-Means clustering in general are (1) determining the desired value of K or group; (2) determine the centroid value; (3) determine the distance of each object to centroid; (4) grouping objects based on the closest distance [8].

2.4. Smart Indonesia Program
The Smart Indonesia Program (PIP) is one of the national programs [17] which is managed by the Office of Education and Culture which aims to increase access for children aged six to twenty-one years to be able to experience education services until graduating from secondary education units and preventing the public avoid the possibility of dropping out [2] [12]. Assistance provided to students in the form of cash [18] which can be used for education funding.
3. Result and Discussion

Preprocessing is a process that is carried out before the actual data analysis process begins [24] where at this stage a process aimed at cleaning / data cleaning, integration and data reduction, transmission, and data normalization stages [16]. After passing the cleaning or cleaning phase, the best dataset becomes 847 records.

Data Transformation can help improve efficiency of the algorithm applied [15]. Data transformation is done to convert data into forms that correspond to mining techniques [14]. In this study men were transformed into 1 and women were transformed into 2 for sex attributes, while in transportation equipment attributes namely public transportation/ bus/ pete-pete, walking, private car/ private vehicle, shuttle car/ shuttle bus, taxibike, bicycles, motorbikes, and others are transformed in a scale from 1 to 8. The same is done for other attributes. The normalization stage is used to change numerical data which has varied values so that it has values with the same range, namely in the range 0 to 1 using the min-max normalization equation [3] [15].

For the selection of data, in this study used 3 ways of selecting training data and testing data where each method of selecting data will be used in the classification process using the Naive Bayes Classifier algorithm. The first data selection is done by dividing the entire data set into 70% training data and 30% random data testing on the normalized dataset.

Subsequent data sharing was carried out using the Clustering K-Means technique. Data are grouped using K = 2 to K = 10 with a maximum iteration of 10 and 100. From the experimental results the data that will be used in the data sharing process is a cluster that has the least or smallest Davies-Bouldin Index validity test [24] . The following are the results of the experiment above.

![Figure 1. Value of Davies-Bouldin Index on the iteration of 10 and 100](image)

Based on Figure 1 above, the best grouping is grouping using K = 10 with a maximum of 100 iterations with a DBI value of 0.118. Then the results of each clustering are divided into 70% training data and 30% testing data. The following is the result of data sharing using the K-Means technique. The next data sharing technique is the K-Fold Cross Validation technique using K equal to 10 [9] the data is divided evenly with a ratio of 9: 1, where 9 for training data and 1 for testing data on each K. Selection of testing data and training hails data is different for each K.

Data that has been divided into training data and testing data is then processed using the Naive Bayes Classifier algorithm and an evaluation is carried out with accuracy, precision and recall. The evaluation value for each data sharing technique is figure 2.

From the figure 2, the classification results using the Naive Bayes Classifier algorithm result in the evaluation of Accuracy, Precision, and Recall values in the Hold Out 70% data sharing training data and 30% testing data which are 95.67% respectively; 100%; and 59.26%. While the data sharing technique using K-Means Clustering obtained an Accuracy value of 96.06%, Precision of 100%, and Recall value of 66.67%. In the K-Fold Cross Validation data sharing technique using K equal to 10, the value of accuracy, Precision, and Recall is 97.40%; 100%; and 76.14%. While the most optimal K value is K = 6 and K = 7 with a value of 97.41%. One way to look at other evaluation values is to evaluate the the...
Area Under Curve (AUC) value which can be seen through the ROC curve. and the Y axis represents true positive [13].

![Evaluation of NBC](image)

**Figure 2.** Value of Naïve Bayes Classifier Evaluation

The figures above illustrates the AUC value of each data sharing technique using Hold-Out 70% training data and 30% testing data of 0.881, this value is generated from the AUC Optimistic average of 0.882 and Pessimistic AUC is 0.879. which means that the classification falls into the good category, the data is divided using the K-Means Clustering technique of 0.921 resulting from the AUC Optimistic average of 0.922 and Pessimistic AUC 0.919 or in other words the classification is very good, while using the 10 Fold Cross Validation model is 0.937 with AUC Optimistic at 0.938 and Pessimistic AUC 0.936 with a description of the AUC value that is very good classification.

4. Conclusion

Based on the results of the study conducted using the Naive Bayes Classifier Algorithm, it can be accomplished that this algorithm can be used to classify the eligibility of prospective recipients of the Smart Indonesia Program using the 10-Fold Cross Validation data sharing technique because it has a higher value of accuracy, precision and recall if compared to other data sharing techniques that are Hold out and K-Means Clustering techniques. As well as having an AUC value described through the ROC curve of 0.937 or in other words the Naive Bayes Classifier classification algorithm using a 10 Fold Cross Validation data sharing technique can be categorized as Very Good / Excellent and can be applied into an application.

References

[1] Kementerian Pendidikan dan Kebudayaan, “Program Indonesia Pintar,” Available: http://indonesiapintar.kemdikbud.go.id/, 2016.

[2] “Peraturan Direktur Jenderal Pendidikan Dasar dan Menengah tentang Petunjuk Pelaksanaan Program Indonesia Pintar pada Jenjang Pendidikan Dasar dan Menengah,” 2018.

[3] Mustakim, “Effectiveness of K-Means Clustering to Distribute Training Data and Testing Data On K-Nearst Neighbor Classification,” *J. of Theoretical and Applied Information Technology*, vol. 95, no. 21, pp. 5693-5700, 2017.

[4] Singh P, “Quantifying Concept Drifting in Network Traffic using ROC Curves from Naive Bayes Classifiers,” *Nirma University Int. Conf. on Engineering (NUiCONE)*, pp. 1-5, 2013.

[5] Soria D, Garibaldi J, Biganzoli E, and Ellis I, “A Comparison of Three Different Methods for Classification of Breast Cancer Data,” *Seventh Int. Conf. on Machine Learning and Applications*, pp. 619-624, 2008.

[6] Zhong S, Lin S, Xu G, and Shi K, “The Expansibility Research of K-means Algorithm under the GPU,” *IEEE*, pp. 734-737, 2016.

[7] An Y, Sun S, and Wang S, “Naive Bayes Classifiers for Music Emotion Classification Based on Lyrics,” *IEEE Computer society*, vol. 978, no.1, pp. 635-638, 2017.
[8] Kapil S, Chawla M, and Ansari D, “On K-means Data Clustering Algorithm with Genetic Algorithm,” Fourth Int. Conf. on Parallel, Distributed and Grid Computing (PDGC), pp. 202-206, 2016.

[9] Goswami S, Bhardwaj P, and Kapoor S, Naive Bayes Classification of DRDO Tender Documents, IEEE, pp. 593-597, 2014.

[10] Jabbar M.A and Samren S, “Heart Disease Prediction System Based on Hidden Naive Bayes Classifier”.

[11] Zhang, H, “The Optimality of Naive Bayes,” Proceedings of 17th Int. Florida Artificial Intelligence Research Society Conf., Menlo Park, pp. 562-567, May. 2004.

[12] Zulvia S, “Evektivitas Pemanfaatan Dana Program Indonesia Pintar Siswa di Desa Ngrayun Untuk Pemenuhan Wajib Belajar 9 tahun,” Kajian Moral dan Kewarganegaraan, vol. 5 no. 3, pp. 716-734, 2017.

[13] F. Gorunescu, “Data Mining Concepts, Models and Techniques,” Berlin,Heidelberg: Springer Berlin Heidelberg, vol. 12, 2011.

[14] S.S. Baskar, Dr. L. Arockiam, and S.Charles, “A Systematic Approach on Data Preprocessing In Data Mining COMPUSOFTWARE,” An int. j. of advanced computer technology, vol. 2, no. 11, pp. 335-339, 2013.

[15] Shalabi L, and Shaaban Z, “Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix,” Int. Conf. on Dependability of Computer Systems, 2006.

[16] Suad A, Alasadi and Bhaya W, “Review of Data Preprocessing Techniques in Data Mining,“ J. of Engineering and Applied Sciences, vol. 12, no.16, pp. 4102-4107, 2017.

[17] Sartika, Safitri P, and Edison, “Implementasi Program Kartu Indonesia Pintar di Kota Tanjung Pinang,”.

[18] Sraswati N, “Implementasi Kebijakan Program Indonesia Pintar (PIP) Pada jenjang Sekolah Dasar di Kecamatan Sungai Pining Kota Samarinda,” J. Administrasi Negara, vol. 5, no. 4, pp. 6737-6750, 2017.

[19] Hassan S, Rafi M, and Shaikh M, “Comparing SVM and Naive Bayes Classifiers for Text Categorization with Wikitext as knowledge Enrichment,” IEEE, pp. 31-34, 2011.

[20] Marlina L, Muslim, and Siahaan A, “Data Mining Classification Comparison (Naive Bayes and C4.5 Algorithms),” Int. J. of Engineering Trends and Technology (IJETT), vol. 38, no.7, pp. 380-383, 2016.

[21] Nasution N, Djahara K, and Zamsuri A, “Evaluasi Kinerja Akadeimi MahasiswaMenggunakan Algoritma Naive bayes (Studi Kasus: Fasilkom Unilak),” Jurnal Teknologi Informasi & Komunikasi Digital Zone, vol. 6, no no. 2, pp. 1-11, 2006.

[22] Dey L, Chakraborty S, Biswas A, Bose B, and Tiwari S, “Sentiment Analysis of Review Datasets using Naive Bayes’ and K-NN Classifier,” Int. J. of Information Engineering and Electronic Business, vol. 8, no. 4, pp. 54-62, 2016.

[23] Xhemali, D., Hinde, C.J. And Stone, R.G, “Naive Bayes vs. Decision Trees vs. Neural Networks in the classification of training web pages,” Int. J. of Computer Science Issues, vol. 4, no. 1, pp. 16-23, 2009.

[24] A. Famili , Shen W, Weber R, and Simoudis E, “Data Preprocessing and Intelligent Data Analysis,” Intelligent Data Analysis, vol. 1, no. 1, pp. 3-23, 1997.

[25] Fujino A, and Ueda N, “A Semi-supervised AUC Optimization Method with Generative Models,” IEEE computer society, pp. 883-888, 2016.