Holdout Validation for Comparison Classification Naïve Bayes and KNN of Recipient Kartu Indonesia Pintar

*F Tempola, R Rosihan, R Adawiyah

Department of Informatics, Faculty of Engineering, Universitas Khairun, Indonesia

*Corresponding author: firman.tempola@unkhair.ac.id

Abstract. Kartu Indonesia Pintar (KIP) is one of the leading programs in the field of Education in the current government era. At present in the provision of smart Indonesian cards often occur not on target. The selection process for smart Indonesian cards is not transparent. For this reason, a transparent selection process is needed so that it does not cause jealousy among students. One way to make the selection process more objective is by applying existing classification methods to data mining. The study applies two methods to see the comparison, namely naïve bayes and k-nearest neighbor. In addition, the algorithm validation stage uses holdout validation. The purpose of algorithm validation is done so that each data has the same opportunity in the training and testing process, in this study also applied 150 dataset, the results showed that the system with no validation algorithm naïve Bayes accuracy is better that is an average accuracy of 85.66% and an average accuracy of k-nn 84.89%. However, if algorithm validation is applied, the k-nn accuracy is better at 88.7% compared to naïve bayes which is only 81.3%.

Keywords: Holdout Validation, Naïve Bayes, KNN, Kartu Indonesia Pintar

1. Introduction

The Smart Indonesia Program (PIP) through the Kartu Indonesia Pintar (KIP) is the provision of assistance in cash of education to children aged school (ages 6-21 years) who comes from a family of poor, vulnerable poor: the owner of the Family Welfare Card (KKS), participants of the Family Hope (PKH) program, orphaned strays, persons with disabilities, victims of natural disasters/disaster [1]. The accuracy to determine the recipient of the card Indonesia smart is accurate and not false targets requires a time that is long, so it takes the algorithm to classify accepted or rejected. Algorithms are included in the classification of the data mining such as k-nearest neighbour and naive Bayes are famous with the level of accuracy that is good. For it needs to be done comparing the two methods in order to know the algorithm that is more high level of accuracy in KIP recipient classification with some criteria that work fathers, work mothers, earnings fathers, income mothers, the status of stay and the status of children to how the family.

Some research assistance from the government for Education by utilizing information technology in this case algorithms in the field of computer science has been done by several researchers before. [2] with the title Methods Naïve Bayes for Determination of Recipients Scholarship Bidikmisi University of Mulawarman with five criteria that work the old, earning the elderly, the number of dependents, power electricity, and the value of the exam nationwide. Research that do have a degree of accuracy of 85.56%. [3] applied case-based reasoning to determine the eligibility of scholarship recipients. [4] applies case-based reasoning by utilizing case adaptations to determine the eligibility of students to receive scholarships. [5] with the title of System Recommendations Providing Assistance Fund PIP Using the K-Nearest Neighbour Method (Study Case: UPTD SMP Negeri 1 Berbek) with 4 criteria are
dependents of the parents, work the old, earning the elderly, and achievement. Testing is done on 1 new student testing data on 10 training data that results in the classification of students getting PIP funding assistance. For a comparative study of the two methods carried out by [6] The results showed that the accuracy of the system by applying Naive Bayes was better than k-NN.

Data mining as a process to get useful information from large database warehouses. Data Mining can also be interpreted as extracting new information taken from large chunks of data that help in making decisions [7].

The Nearest Neighbour algorithm (sometimes called the K-Nearest Neighbour) is an algorithm that classifies based on the proximity of the location (distance) of a data with other data. Bayes is a simple probabilistic based prediction technique based on the application of the Bayes theorem (or Bayes' rule) with strong (naive) independence assumptions. In other words, in Naïve Bayes, the model used is an "independent feature model" [8].

Based on this in this study will be made a comparison of the classification between naïve Bayes and k-NN for classification recipient Kartu Indonesia Pintar. On the other hand, using stage validation using holdout validation. process validation ever by [9].

2. Method

Data mining is the process mining in there can possess techniques of statistics, mathematics, intelligence system, and machine learning to extract and identify the information that is useful and knowledge that is relevant from a variety of large databases. The term data mining has a nature as a scientific discipline whose main purpose is to find, explore, or mine knowledge from the data or information that we have. Data mining often also referred to as Knowledge Discovery in Database (KDD). KDD is the activity which includes the collection, data usage, the historical to find regularities, patterns or relationships in the large data set size [10].

2.1 Classification

Classification is a job evaluating data objects to put them in a certain class from a number of available classes. In the classification there are two processes that do that is to build a models to be stored as memory and using models such to conduct the introduction or classification or prediction on some of the data that is known in the class where the object of the data are entered based on a model that has been stored in the memory [11].

2.2 Pre-processing Data

Pre-processing is done for the dataset is normalized value of the attribute which has spread that the big of which is to attribute that is worth numerically by using the normalization formula of data, as shown in Equation 1.

\[ \hat{x} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]  

(1)

2.3 K-nearest Neighbour

K-Nearest Neighbour (KNN) belongs to the classification algorithm group. This algorithm is also one of the supervised learning techniques. K-NN is done by searching for groups of k objects in the training of data that is most close (Similar) to the object on the Data of new or the Data testing. Algorithm K-Nearest Neighbour is a method to perform the classification of the object based on the Data of learning that distance is most close to the object it. Nearest Neighbour is the approach to look for cases to calculate the closeness between the case of the new and the case of the old that is based on matching the weight of a number of features that exist.

To define the distance between two points, namely the point in the training data (xi) and the point in the testing data (xj), the Manhattan Distance formula is used, as shown in Equation 2.

\[ d(x, y) = \sum_{k=1}^{n} |x_i - y_i| \]  

(2)
2.4 Naïve Bayes

Naïve Bayes is a probabilistic method of simple classification based on the Bayes Theorem where classification is done through training sets of data efficiently. Naïve Bayes assumes that the value of an input attribute in a given class is independent of other attribute values. Bayes’ own theorem was put forward by the British scientist Thomas Bayes, which predicts future opportunities based on past experiences so that it is known as the Bayes Theorem [12].

Naïve Bayes Stages [13]:

1. Calculate the mean and standard deviation of each numerical data, formula for finding the mean can be seen in Equation (3).

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]  

The formula for finding the standard deviation can be seen in Equation (4)

\[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \]  

2. Calculating the probability of each discrete data can be seen in equation (5)

\[ P(E) = \frac{x}{N} \]  

3. Calculating Gauss Density can be seen in equation (6)

\[ f(x_i) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]  

4. Calculating Likelihood can be seen in equation (7)

\[ L(x) = f(x_1) \times f(x_2) \times ... \times f(x_n) \ldots \ldots \]  

5. Calculating the likelihood probability can be seen in equation (8)

\[ P(L_i) = \frac{L_i}{\sum_{i=1}^{n} L_i} \]  

2.5 Laplacian Smoothing

In a large dataset, the selection of training data by random will lead to the possibility of zero values in the probability model. This, zero value will cause Naïve Bayes Classifier to not be able to classify and input data. Therefore, we need a smoothing method that can avoid the zero value in the probability model. Laplacian Smoothing is a smoothing method commonly used in Naïve Bayes Classifier. Laplacian Smoothing is commonly known as add one smoothing, because in its calculation, each variable in each parameter is added 1 [14]. Laplace Smoothing equation is written in Equation (9)

\[ P(X_k \mid C) = \frac{P(X_k \mid C) + 1}{P(C) + N} \]  

2.6 Performance measurement system

Accuracy calculation is needed as a benchmark for evaluation in the system. Measurement of accuracy can be used in various ways one of which uses recognition rate. The recognition rate equation used can be seen in the following formula can be seen in Equation 10. [15]

\[ Accuracy = \frac{\sum_{Correct}}{\sum_{Sample}} \times 100\% \]
2.7 Holdout Validation
The validation process is very important to do, the goal is that each data has the opportunity as training data and testing data. There are several model validations, one of which is holdout validation. Holdout validation is done by dividing the data between the two parts by 50% or 70% and 30%, respectively. But in principle, holdout validation provides an opportunity for each data to become training data and testing data. So that the holdout validation test is applied to measure the accuracy will occur twice. For example the data is divided into two groups of data A and B. First test data A as training data and data B as testing data. Then test both data B as training data and data A as testing data. The accuracy of each test is then summed and divided by 2. As shown in Equation 11.

\[
\hat{x} = \frac{Testing_A + Testing_B}{2}
\]  

(11)

3. Result and Discussions
3.1 Testing method classify without validation holdout
In this test the data validation process is not carried out but tested from the lowest to the highest amount of data. For the k-NN algorithm, k initialization is performed 3 times namely 3, 5, and 7. The test results are shown in Figure 1 below.

Figure 1. Result of testing without validation
Based on Figure 1. a, the test is carried out on training data 1 to 149 and test data 149 to 1 using the value of K = 3 get the lowest accuracy of 48% and the highest accuracy of 100% with an average total accuracy of 84.68%. Based on Figure 5. b, testing is carried out on training data 1 to 149 and test data 149 to 1 using the value of K = 5 to get the lowest accuracy of 0% and the highest accuracy of 100% with an average total accuracy of 84.08%. Based on Figure 5. c, testing is carried out on training data 1 to 149 and test data 149 to 1 using the value of K = 7 get the lowest accuracy of 0% and the highest accuracy of 100% with an average total accuracy of 84.89%. Based on Figure 5. d, the test is carried out on the training data to 1 to 149 and the test data to 149 to 1 get the lowest accuracy of 51% and the highest accuracy of 100% with an average total accuracy of 85.66%.

3.2 Testing method classify using validation holdout

As the concept of holdout validation is data divided into two parts. In this study the data is shared equally at 50% each. With a dataset of 150 data, then data A 75 data and data B 75 data. Furthermore, it is applied in each classification model. For the application of k-NN the value of k used is k = 3. Obtained results for testing 1 with data A as training data and data B as testing data obtained 86.7% accuracy. Furthermore, data A as testing data B as training data obtained an accuracy of 90.7%. With an average accuracy of the k-NN method of 88.7%, while for the application of Naive Bayes in test 1 an accuracy of 85.3% was obtained. And the second test obtained 77.3% accuracy, with an average accuracy for the application of the Naive Bayes method is 81.3. Thus, for the case of smart Indonesian card recipients the accuracy by applying k-nn is better than naive Bayes.

4. Conclusion

Based on the results of tests conducted for the application of the k-NN method of initiation k = 3, the accuracy is better than k = 5 or k = 7 if the data is shared equally for the testing process in the case of receipt of a Kartu Indonesia Pintar. While the average accuracy for each initialization process k = 7 has the highest accuracy that is equal to 84.89% compared to k = 5 and k = 3. However, when compared with the average accuracy of applying the Naive Bayes method, the accuracy of Naive Bayes is higher compared to k-NN that is obtained an average accuracy of 85.66%. In contrast to testing using holdout validation, k-NN accuracy is better than naive Bayes, namely k-NN 88.7% and naive Bayes with an average accuracy of 81.3%.

References

[1] Kemendikbud, 2016.“Program Indonesia Pintar,” http://indonesiapintar.kemdikbud.go.id. http://indonesiapintar.kemdikbud.go.id (accessed Feb. 02, 2019).
[2] D. M. Dahri, D., Agus, F., & Khairina,2016. “Metode Naive Bayes Untuk Penentuan Penerima Beasiswa Bidikmisi Universitas Mulawarman,” J. Inform. Mulawarman, vol. 11, no. 2.
[3] F. Tempola and S. Do Abdullah, 2018.”Case-Based Reasoning ( CBR ) untuk Penentuan Kelayakan Mahasiswa Penerima Beasiswa,” J. PROtek, vol. 05, no. 2, pp. 1–5, doi: 10.13140/RG.2.2.15460.01928.
[4] F. Tempola, A. Musdholifah, and S. Hartati, 2018, “Case Based Reasoning for Determining the Feasibility of Scholarship Grantees Using Case Adaptation,” Proc. - 2018 5th Int. Conf. Inf. Technol. Comput. Eng. ICTACEE 2018, pp. 370–374, doi: 10.1109/ICTACEE.2018.8576898.
[5] F. Aprilia, 2016,”Sistem Rekomendasi Pemberian Bantuan Dana PIP Menggunakan Metode K-Nearest Neighbor ( Study Kasus : UPTD SMP NEGERI 1 BERBEK ),” Simki-MTEKsain, vol. 1, no. 8.
[6] F. Tempola, M. Muhammad, and A. Khairan, 2018, “Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation Comparison of Classification Between Knn and Naive Bayes At the Determination of the Volcanic Status With K-Fold Cross,” J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 5, pp. 577–584, doi: 10.25126/jitiik20185983.
[7] M. Bramer, Principles of Data Mining, 1st ed. London: Springer-Verlag, 2007.
[8] P. HARRINGTONT, Machine Learning In Action, 1st ed. USA: manning publication, 2012.
[9] C. Tantithamthavorn, S. Member, and S. McIntosh, 2017, “An Empirical Comparison of Model Validation Techniques for Defect Prediction Models,” IEEE Trans. Softw. Eng., vol. 43, no. 1, pp. 1–18.

[10] F. Gorunescu, Data Mining: Concepts, Models and Techniques. London: Springer, 2011.

[11] Budi Santosa dan Ardian Umam, Data Mining and Big Data Analytics. Theories and Implementation Using Python and Apache Spark. 1st ed. Yogyakarta: Spreader Media Library, 2018.

[12] F. Tempola, M. Muhammad, and A. Khairan, 2018, “Naive Bayes Classifier for Prediction of Volcanic Status in Indonesia,” Proc. - 2018 5th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2018, pp. 365–369, doi: 10.1109/ICITACEE.2018.8576966.

[13] A. Anas, F Tempola dan A Khairan, 2019, “Hybrid fuzzy dan Naive Bayes Dalam Penentuan Status UKT,” Protek, vol. 06, no. 1, pp. 6–11.

[14] S. W. Cahyanti, A. F., Saptono, R., & Sihwi, 2015, “Penentuan Model Terbaik pada Metode Naive Bayes Classifier dalam Menentukan Status Gizi Balita dengan Mempertimbangkan Independensi Parameter,” J. ITSMART, vol. 4, no. 1, pp. 28–35.

[15] F. Tempola, 2019, “IMPLEMENTED PSO-NBC AND PSO-SVM TO HELP DETERMINE STATUS OF VOLCANOES IMPLEMENTASI PSO-NBC DAN PSO-SVM UNTUK,” J. Penelit. pos dan Inform., vol. 9, no. 2, pp. 97–103, doi: 10.17933/jppi.2019.090202.