The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency

A H Wibowo[1], T I Oesman[1]

[1] Department of Industrial Engineering, AKPRIND Institute of Science and Technology, Jl. Kalisahak 2, Daerah Istimewa Yogyakarta, Indonesia

E-mail: bagushind@akprind.ac.id

Abstract. Crime is an action which is considered as a violation of law and can harm others. Nowadays, crimes has increased with erratic patterns. Therefore, crime prevention is necessary since it will occur based on the historical data. Data mining is a technique that can be used to predict crimes that will occur. According to the previous researches, data mining techniques have several methods that can be used to predict crimes by utilizing the data of crimes that have occurred. Hence, it is necessary to conduct a comparative analysis of classification algorithms in order to obtain accurate prediction results based on the crime data in Sleman regency. The classification algorithms analyzed in this study were k-NN, Naive Bayes, and Decision Tree. Based on the three algorithms, the accuracy of k-NN with k = 5 was 57.88 percent, with k = 10 was 59.49 percent, with k = 15 is 59.38 percent, with k = 20 was 60.18 percent, and with k = 25 was 61.57 percent. Meanwhile, for the Naive Bayes algorithm, the accuracy reached 65.59 percent, and the Decision Tree algorithm reached 60.23 percent. In conclusion, the algorithm with the highest accuracy was owned by Naive Bayes.

1. Introduction

Nowadays, crimes have evolved rapidly, especially in Indonesia. Crime is an action which is considered as a violation of law and can harm others, and violates the prevailing norms. Crime is a socio-economic issue that affects people in the whole world and negatively impacts societies' welfare [1]. Hence, crime is a social disorder and can harm many people in various ways [2]. Crime is not systematic or totally random or cannot be predicted directly [3].

Researches on the use of data mining to predict crimes had also been conducted by [1, 3-7] which predicted crime rates with the aim of helping the police in preventing criminal actions that will occur. Those researches were conducted since the crime rates increased from year to year. Therefore, the prevention is necessary in order to minimize the level of crime that will occur. One technique that can be used to prevent crimes is data mining technique, by predicting the crime patterns that will occur based on the previous data. Data mining is a technique that will be accurate in predicting the crimes that will occur [3].

Conducting research by implementing data mining techniques to detect and predict the crime patterns can solve the criminal issues faster [8]. The initial test was conducted by implementing three methods, such as k-NN, Naive Bayes, and Decision Tree, in order to compare the accuracy.

2. Literature review

2.1. Data mining

Data mining is a data collection technique which is obtained from several sources, and then from the data, it will be converted into very useful information by using various methods. Data mining refers to an interdisciplinary scientific field that combines the techniques from machine learning, pattern recognition, statistics, databases, and visualizations in retrieving the information from large databases [9]. In general, data mining can be classified into 2 main categories, such as descriptive mining and predictive [10]. In conducting predictive, data processing techniques that already exist are needed to be collected and processed, and the most popular technique used is Data Mining [11]. Several methods in data mining techniques have been used to predict patterns, or in this case for predicting crime patterns. According to [11], data mining is the most popular technique used in the last ten years from 2000 to 2011. The information in data mining have different types [2].

2.2. k-Nearest Neighbors (k-NN)

k-Nearest Neighbor (k-NN) is a data classification technique based on the proximity of the data location to other data. The distance used is Euclidean Distance as stated in Equation (1)

$$dist\ (X_1,X_2) = \sqrt{\sum_{i=1}^{n} (X_{1i}-X_{2i})^2} \quad (1)$$

X1 : Sample Data
X2 : Testing
i : Variable Data
d : Distance
p : Dimension Data

k-NN makes an explicit prediction on the testing data based on a comparison of k-Nearest Neighbor. In order to calculate the proximity of data, the Euclidean Distance formula can be used.

2.3. Naive Bayes

Bayes is one of the classification methods that can predict the probability of membership in a class. The value of a class in Naive Bayes method is independent, which independent on other attributes. This classification is conducted by using the following formula:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2)$$

X : data with unknown class
H : hypothesis of data X
$P(HIX)$: posteriori probability
$P(H)$: priori H
$P(XIH)$: probability X based on H hypothesis
$P(X)$: priori X

2.4. Decision tree

Decision Tree is a tree that exists in the analysis of problem solving and alternative solutions mapping which can be taken from the problem. Decision tree can also be called as one of the most popular classification algorithms since it is easy to interpret. Decision tree is suitable with the cases which

output are discrete values. The main benefit of using a decision tree is to examine and describe complex decision-making processes in order to make it simpler and easier to be interpreted. Decision tree is usually used to obtain the information which turns into decision. The decision tree begins with a node root (starting point) which is used by the user. Based on this node root, the user solves the leaf nodes according to the tree algorithm decision. The final result of composing the node root and leaf node is a decision tree with each branch that shows possible scenarios of the decision and the results [12]. Building a decision tree involves an initial stage in order to choose the attributes for roots, while it is based on the highest value of all existing attributes with Equation (3), as follows.

$$Gain(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \text{Entropy}(S_i) \quad (3)$$

S : Case sets
A : Atribut
n : The total number of partition A
$|S_i|$: The total number of cases in partition to i
$|S|$: The total number of cases S

The calculation to obtain entropy values can be seen in the Equation (4) as follows:

$$Entropy(S) = \sum_{i=1}^{n} -p_i \times \log_2 p_i \quad (4)$$

Pi: Proportion from Si to S

2.5. Confusion matrix

In data mining techniques, there are several ways to measure the performance of the model results, such as confusion matrix. Confusion matrix is a method used to calculate accuracy in the data mining concept. Confusion matrix contains information related to the actual classifications and predictions conducted by the classification system. The system performance is generally evaluated by using the data in the matrix [13].

Table 1. Confusion matrix table.

		Actual	
		Positive	**Negative**
Predicted	**Positive**	TP (True Positive)	FN (False Negative)
	Negative	*FP (False Positive)*	*TN (True Negative)*

According to [14] the Confusion matrix produces accuracy, precision, and recall values. Accuracy is defined as the proportion of the total number of correct predictions, which is determined by Equation (5).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

Precision is the accuracy measurement of certain class which has been predicted, while the precision is determined by the Equation (6).

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall is a measurement of the model ability to predict several issues from certain class which is obtained from the data collection, while recall is determined by the Equation (7).

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

The values of Precision and Recall can be given in the form of numbers by using percentage calculations (1-100%) or by using numbers between 0-1. The recommendation system will be considered good if the Precision and Recall values are high.

3. Methodology

3.1. Preprocessing stage

This study used crime data from the Sleman Regency's Police Department for 3 years, and this study chose 3 types of crime such as, theft, fraud, and embezzlement. The initial stage in this study was preprocessing. Preprocessing stage is the process of processing raw data in order to be processed, such as the data that is clean from outliers and redundancy. Data that has been processed at this preprocessing stage is called as preprocessing data. The preprocessing stage is also conducted in introducing the attributes that will be used in data mining process. This stage should be conducted before doing the data mining process. Preprocessing begins from selecting attributes, by removing attribute columns that have similar data records. On the other words, the deleted column is an attribute that has no relation with the data that are analyzed. Hence, an attribute that can be a predictor variable for other analysis processes should be separated. After that, the missing value should be filled with other values that often appear, and also the average value. Then, the inconsistent data which also contain errors should be selected. Therefore, in the initial stages of preprocessing, it produces consistent data. From the attributes that have been set, outlier detection is conducted. Outliers / anomalies are data which are considered to have different properties than the other data. This outlier should be avoided since the differences can cause the analysis to not reflect the actual results.

3.2. Testing the algorithm stage

Testing the algorithm requires training and testing data. Training data is preprocessing data that is free from outliers. Meanwhile, testing data is obtained from the training data that has its label removed. The classification in this study was implemented by applying three algorithms such as k-NN, Naive Bayes, and Decision Tree. Training and testing data were tested by applying these algorithms. It showed different classes of data in each algorithm. Therefore, in order to measure the accuracy of each algorithm, confusion matrix technique was implemeted. In k-NN algorithm, there was an optimal k value with the highest accuracy, rather than the k-NN algorithm classification with other k values. The optimal k value was obtained from the iteration of k-NN algorithm with training data, which was reduced by 5 for each iteration performed.

4. Results and discussion

The study was conducted on a data set with the total number 1,735 crimes data for 3 years, which consisted of 15 attributes, such as Day, Season, Time, TKP, District, Gender of the Victim, Occupation of the Victim, Age, Residence, Sex of the Offender, Occupation of the Offender, Age of the Perpetrators, the Residence, number of the perpetrators, and types of crimes which were classified into 3 classes, such as theft, fraud, and embezzlement. Many previous studies have explained the ratio used in determining training and testing sets. As much as 60% of the total data was used for training data, and 40% was used as testing data. In the initial stage of preprocessing, 15 attributes were obtained in the process of analyzing data, which made classification by using k-NN, Naive Bayes, and Decision Tree algorithms. The results of the accuracy test for each algorithm were summarized in the Table 2.

Table 2. The result of algorithm accuracy test.

Algorithm Type	Accuracy	Percentage
k-NN (5)	503	57.88 %
k-NN (10)	517	59.49 %
k-NN (15)	516	59.38 %
k-NN (20)	523	60.18 %
k-NN (25)	535	61.57 %
Naive Bayes	570	65.59 %
Decision Tree	524	60.30 %

From the Table 2 it can be concluded that Naive Bayes was an algorithm that reached the highest accuracy. Meanwhile, the optimal k value for k-NN algorithm was 25.

5. Conclusions

Performance measurement of a data mining algorithm can be conducted based on several aspects such as accuracy, computational speed, robustness, scalability, and interpretability. This research only measured the performance of the data mining algorithm based on the accuracy aspects. The results of algorithm accuracy test by implementing confusion matrix could be seen from the accuracy values. Therefore, the percentage of the Naive Bayes algorithm of 65.59% achieves a higher level of accuracy than the k-NN (5) algorithm of 57.88%, k-NN (10) of 59.49%, k-NN (15) of 59.38%, k -NN (20) 60.18%, k-NN (25) 61.57% and Decision Tree 60.30%.

6. References

[1] Thongsatapornwatana U 2016 *2016 Second Asian Conference on Defence Technology* (*ACDT*) 123–8
[2] Kaur S 2017 *International Journal of Advanced Research in Computer Science* **8** 1336–42
[3] Wu J, Meziane F, Saraee M, Aspin R and Hope T 2016 *2016 International Workshop on Computational Intelligence for Multimedia Understanding* (*IWCIM*) 1–4
[4] Yu C-H, Ward M W, Morabito M and Ding W 2011 *2011 IEEE 11th International Conference on Data Mining Workshops* 779–86
[5] Sathyadevan S *et al*. 2014 *2014 First International Conference on Networks & Soft Computing* (*ICNSC2014*) 406–12
[6] Nasridinov A, Ihm S-Y and Park Y-H 2013 *A Decision Tree-Based Classification Model for Crime Prediction* (Dordrecht: Springer,) pp 531–8
[7] Tayal D K, Jain A, Arora S, Agarwal S, Gupta T and Tyagi N 2015 *AI & society* **30** 117–27
[8] Nath S V 2006 *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops* 41–4
[9] Larose D T 2005 *Discovering knowledge in data : an introduction to data mining* (New Jersey: Jhon Willey & Sons)
[10] Tan P-N, Steinbach M and Kumar V 2006 *Introduction to data mining* (Boston: Pearson Addison Wesley)
[11] Liao S-H, Chu P-H and Hsiao P-Y 2012 *J. Expert Systems With Applications* **39** 11303–11
[12] Agrawal S and Agrawal J 2015 *Procedia Computer Science* **60** 708–13
[13] Palaniappan S, Mustapha A, Mohd Foozy C F and Atan R 2017 *International Journal of Informatics Visualization* **1** 214
[14] Chaurasia V, Pal S and Tiwari B 2018 *Journal of Algorithms & Computational Technology* **12** 119–26

Acknowledgments
The authors would like to thank Kepolisian Polres Sleman Daerah Istimewa Yogyakarta which had helped the researchers in collecting the useful data for this study.