Implementation of Naive Bayes for selection of green space areas

N Pandiangan and M L C Buono
Information System, Faculty of Engineering Musamus University, Merauke, Indonesia
E-mail: Lela24pandiangan@gmail.com

Abstract. City development, population growth, and climate influence triggering a flood disaster. Green open space is one of the alternative flood prevention in urban areas. However, in determining the area of green open space is one of the alternative flood prevention in urban areas. However, in determining the area construction of the green open spaces in urban areas because it has undergone changes in the city, many parameters need to be considered for assessment, a simple way to solve it is by utilizing current technology and using data mining techniques. Data mining is an option that can be used to determine the area of green open space development based on existing data patterns. The Naïve Bayes Algorithm is the algorithm used in this research, the algorithm produces an average accuracy value is 53.94094% and gets the highest accuracy value at a data ratio of 80:20 with an accuracy value is 61%

1. Introduction
Floodings is one of the natural disasters caused by the development of the city, the increasing population and climate influences. Various efforts have been made, one of them is the construction of green open spaces in cities is one of the good solutions for handling floods in terms of ecosystems [1]. Another benefit of green open space is can provide a sense of comfort, support public health and have aesthetic value in the development of the city. But in its construction, many things affect the selection of areas for building green open spaces, including population density, land prices [2], distance from the center of the settlement, proximity to public facilities, proximity to the roadside [3]. So it needs to be held analysis regarding these parameters, to choose a good area for the construction of the green open spaces. However, at this time there is a simple way that can be done by using technology with data mining techniques to make it easier to determine something based on data patterns that have ever existed before [4]. Data mining can be used for cases of prediction, classification, filtering, and grouping data to find hidden patterns from the dataset [5]. Data mining is one of the most widely used techniques in data mining and has been developed to overcome the problem of processing data in various fields including the environment [6], health [7], education [8]. Many algorithms are included in data mining, but in this case, the author chooses Naïve Bayes as a method for predicting the feasibility of green open space areas by analyzing a simple probabilistic set of the dataset [9]. Its nature is able to process complex of data into the right data, clear, predictive model and very efficient for data classification and supported with good accuracy [10] be the best reason for choosing this algorithm for our case.
2. Methodology

2.1. Naïve Bayes Classifier

Naïve Bayes is an effective and efficient classification algorithm. Naïve bayes works on the basis of Bayes Theory and independent attributes [11]. The class attribute values in Naïve Bayes are independent of other attribute values [12]. Naïve Bayes which is also called the probabilistic method, it is assumed that

In Naïve Bayes as also usually called probabilistic, it is assumed that there are variables which are defined as variables \( x_1 \) until \( x_d \) and this \( x \) variable with each other is conditionally independent. So, variables are not related to each other and do not affect the response value [13].

Naïve Bayes was introduced by Thomas Bayes, someone who worked on the fields of probability and decision. The Naïve Bayes Algorithm equation as follows:

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

where, \( X \) is a dataset which is considered “proof”.

Naïve Bayes measures the dataset as much \( n \) attributes. For example, \( H \) representing the hypothesis of a data \( X \) belongs to the class specified by \( C \). In classification problems, \( P(H|X) \) interpreted the probability of a hypothesis \( H \) with “proof” or data \( X \) which has been observed. In other words, what you will look for is the probability of the data \( X \) belongs to class \( C \) [14].

2.2. Dataset

The data used in this research is artificial sample data the selection of green environment space area which will be divided into two data, namely training data and test data for the calculation of Naïve Bayes Classifier. The data consists of 500 data with 5 parameter attributes, as follows:

- The extensive area of green open space
- Total population
- Distance from the center of the settlement
- The land price
- Proximity to public facilities
- Proximity to river banks
- Proximity to the roadside

2.3. Naïve Bayes Classifier in this research

In this research, Naïve Bayes will classify the training data from the dataset, so that get the data pattern and will be tested using the data from the dataset to determine the accuracy of the Naïve Bayes Algorithm in determining the appropriate area to be a green open space. The process carried out by the Naïve Bayes Algorithm can be seen in figure 1. The steps for classifying with Naïve Bayes Algorithm in this research.

![Figure 1. The steps for classifying with Naïve Bayes Algorithm in this research.](image)
In this research, the dataset will be divided into training data and test based on ratio 60:40, 65:35, 70:30, 75:25 and 80:20. After that, the data will be analyzed using Weka Workbench.

2.4. Evaluation
In working with the Naïve Bayes Classifier, it is necessary to test the model to determine the performance of the algorithm in the data mining process. At this stage, Naïve Bayes will be tested for accuracy in classifying the selection of green open spaces using a confusion matrix. The confusion matrix can be seen in table 1. Confusion Matrix Table.

![Confusion Matrix Table](image)

The Confusion matrix consists of 4 values that influence testing, they are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These values will determine the precision value of a data or the level of accuracy produced by the system in the classification process. Whereas, there are also a data sensitivity values. That is the value that indicates the inaccuracy of the system in determining the data in the output class should be and the sensitivity value also shows how accurate the proximity of the predictive value and actual value is. Equations in calculating these values can be seen in the following equation:

\[
\text{Precision} = \frac{TP}{TP+FP} \quad (2)
\]

\[
\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3)
\]

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)
\]

3. Result and Discussion
The classification results using the Naïve Bayes Algorithm are performed using Weka Workbench with 500 datasets which will be tested for validation based on ratios data 60:40, 65:35, 70:30, 75:25 and 80:20 can be seen in table 2. Result Experiment.

![Table 2. Result of the experiment](image)
The following is a graph of the accuracy Naïve Bayes Algorithm Classification in selection location of the green open space, which can be seen in Figure 2. Accuracy graph for Naïve Bayes Algorithm in This research.

Figure 2. Accuracy graph for Naïve Bayes Algorithm in This research

Based on the graph, the highest accuracy can be seen in the 80:20 data ratio with an accuracy value of 61% and the smallest accuracy obtained in testing the data ratio of 70:30. Based on the graph, the highest accuracy can be seen in the 80:20 data ratio with an accuracy value of 61% and the smallest accuracy obtained in testing the ratio of 70:30. With an average for the overall ratio reaching 53.94094%. Many training data don’t guarantee to produce high accuracy, this can be seen from the graph, the more comparison of training data and test data, does not indicate an increase in the value of accuracy. This can be seen in the graph, the value of accuracy at a ratio of 65:35 produces an accuracy value of 52.5714% but when the ratio increases to 70:30, the accuracy drops to 49.3333%.

4. Conclusion
The Naïve Bayes Algorithm can predict the feasibility of selecting open green spaces with an average accuracy is 53.94094%. Many data does not guarantee a high value of accuracy. It is expected that in the development of further research, the data used is real and more diverse parameters and characteristics, so that can be seen how is good implementation from Naïve Bayes Algorithm in a variety of the other research.

5. References
[1] Schuch G, Serrao-neumann S, Morgan E and Low D 2017 Land Use Policy Water in the city: Green open spaces, land use planning and flood management – An Australian case study *Land use policy* **63** 539–50
[2] Humaida N, Budi L and Badriyah S 2016 Priority assessment method of green open space (case study: Banjarbaru City) *33* 354–64
[3] Alifia N 2016 Identifikasi Letak Dan Jenis Ruang Permukiman Perkotaan *3* 25–38
[4] Amra I A A 2017 Naïve Bayesian 909–13
[5] Rady E A and Anwar A S 2019 Prediction of kidney disease stages using data mining algorithms *Informatics Med. Unlocked* 100178
[6] Gibert K, Izquierdo J, Sánchez-marrè M, Hamilton S H and Holmes G 2018 SC Environ. Model. Softw.
[7] Amin M S, Chiam Y K and Varathan K D 2018 Abstract Cardiovascular disease is one of the biggest cause for morbidity and mortality among the Telemat. Informatics

[8] Aldowah H, Al-samarraie H and Fauzy W M 2019 Educational Data Mining and Learning Analytics for 21st century higher education: A Review and Synthesis Telemat. Informatics

[9] Harahap F, Yugo A, Harahap N, Ekadiansyah E and Utama U P 2018 Implementation of Naïve Bayes Classification Method for Predicting Purchase 2018 6th Int. Conf. Cyber IT Serv. Manag. 1–5

[10] Cui S, Zhao L, Wang Y, Dong Q, Ma J, Wang Y, Zhao W and Ma X 2018 Republic of China SC Injury

[11] Andrejiova M and Grincova A 2018 Author’s Accepted Manuscript Wear

[12] Orphanou K, Dagliati A, Sacchi L, Stassopoulou A, Keravnou E and Bellazzi R 2018 Incorporating repeating temporal association rules in Naïve Bayes classifiers for coronary heart disease diagnosis J. Biomed. Inform. 81 74–82

[13] Borrotti M, Minervini G, Lucrezia D De and Poli I 2016 Naïve Bayes ant colony optimization for designing high dimensional experiments Appl. Soft Comput. J. 1–10

[14] Fatahillah N R 2017 Implementation Of Naive Bayes Classifier Algorithm On Social Media (Twitter ) To The Teaching Of Indonesian Hate Speech 128–31