Naive Bayes Algorithm Analysis to Determine the Percentage Level of visitors the Most Dominant Zoo Visit by Age Category

Iin Parlina1, M. Yusuf Arno1, Nur Ahdina Febriati1, Raffiq Dewi1, Anjar Wanto4, Muhammad Ridwan Lubis1, Susiani1

1 AMIK Tunas Bangsa, Sudirman Street Pematangsiantar, Medan - Indonesia
2 Teknik Mesin, Universitas 45 Surabaya - Indonesia
3 Teknik Informatika, Universitas 45 Surabaya - Indonesia
4 STIKOM Tunas Bangsa, Sudirman Street Pematangsiantar, Medan - Indonesia

*iin@amiktunasbangsa.ac.id

Abstract. Classification is a method of data analysis that is used to create models that describe data classes that are considered important. For the classification of the classification process, the data used is THPS Visitor data which consists of 4 classes including Education, Gender, Age and Visit. The classification used as a comparison of results is the Naive Bayes Classifier. By classifying the number of visitors who are most dominant visiting by age category consisting of adults, adolescents and children. This study aims to classify the highest number of values for visitors by age category. This study was reviewed using the Naive Bayes algorithm. The results of this study indicate that the visitor data that is the most dominant visiting by age category is children who have 77% accuracy is the age of children. This accuracy value is the age that most often visits THPS.

1. Introduction

Rapid development of data mining is inseparable from the development of information technology that allows large amounts of data to accumulate. Along with data mining needs, there are several algorithms for processing large amounts of data[1]–[3], including the Naive Bayes Classifier is a classification method based on probability and Bayes's theorem with an independent assumption of attributes [4]. Grouping Visitors are made to determine the attributes based on age categories starting from 1-70 years whose accuracy value is 70% Children as the most dominant visitors visit. Predictions of visitors coming to THPS in the future can be seen based on the age category of previous visitors. The following are visitor data for 2018.

This study was conducted to calculate each class of visitor data variables aimed at calculating the probability value of each class of related data testing, where the results of the calculation of the probability value as a source of data to be compared between the age variable and the visit class. Then the results of the comparison value are utilized to calculate the accuracy value of the opportunities that "often" visit and "Ever" visit the Pematangsiantar City Animal Park. Using the datamining model[5]. The Naïve Bayes Classifier algorithm is an algorithm that can answer the problem where classification is done by calculating the probability value for each occurrence of the target attribute in each variable class. The purpose of this study is to produce the highest value that can provide the results of the
calculation of the Naive Bayes Classifier Algorithm that the highest value is the result of the classification of visitors that most often come to visit is called the dominant value [6]

The research that has been investigated is predicting the magnitude of the use of electric current using the Naïve Bayes method which is expected to be able to predict the amount of electricity used by each household to make it easier to regulate electricity usage. of the 60 household electricity usage data tested by the naïve Bayes method, the percentage of 78.3333% obtained for the accuracy of predictions, in which there were 47 household electricity usage data that were classified correctly.

2. Methodology

2.1. Collection Data

The data used are visitor data in Pematangsiantar City Animal Park 2018.

| Number | Education       | Gender | Age  | Visit  |
|--------|-----------------|--------|------|--------|
| 1      | Senior High School | Man    | Adult | ever   |
| 2      | Senior High School | Woman   | Adult | ever   |
| 3      | Primary school  | Man      | Children | ever   |
| 4      | Primary school  | Man      | Children | ever   |
| 5      | Primary school  | Woman   | Children | Often  |
| 6      | Children Class  | Man    | Children | Often  |
| 7      | Diploma         | Woman   | Adult | ever   |
| 8      | Senior High School | Woman   | Adult | Often  |

Source: from the Visitors of THPS

2.2. Data Mining concepts

There are several stages of data mining. The stages of Data Mining are as follows [5]:

1. Data Cleaning
   In this section, irrelevant data and inconsistent data will be eliminated.

2. Data integration (Data Integration)
   In this section the data will be integrated into the new database from various existing databases.

3. Data Selection
   In this section, data will be collected from the data base, which will then be analyzed and selected.

4. Data Transformation
   In this section the data mining process will be carried out by changing and merging the data into a suitable format.

5. Mining Process
In this section a large process will be carried out by applying methods to obtain new knowledge from unknown data. Several methods can be used based on Data Mining grouping.

6. Pattern evaluation
   To identify interesting patterns into the knowledge based found

7. Knowledge presentation
   In this section, visualization and presentation will be carried out to produce knowledge that will be obtained by the user with the method used.

c. Naive Bayes method
   Naive Bayesian Classifier is a classification with a statistical model to calculate the opportunities of a class that has each group of attributes that exist, and determine which class is the most optimal. In this method all attributes will contribute to decision making, with the same important weighting attributes and each attribute independent of one another [8]. The basis of the Naïve Bayes Classifier theorem used in programming is the Bayes formula as follows [9].

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

Where
\[
P(H \mid X) = \text{Probability posterior } H \text{ in } X
\]
\[
P(X \mid H) = \text{Probability Posterior } X \text{ in } H
\]
\[
P(H) = \text{Probability Prior of } H
\]
\[
P(X) = \text{Probability Prior of } X
\]

2.2. Normalization Data

![Flowchart Research](image)

**Figure 1. Flowchart Research**
3. Results And Discussion

Testing visitor training data by age category is done by testing accuracy, precision and recall. Testing of visitor data based on this age category is done to determine the level of accuracy, precision and recall of the output produced by the calculation of the classification.

The data training is used to determine the probability of data for the most dominant THPS ending data. The visitor data are as follows:

| Number | Education          | Gender | Age   | Visit |
|--------|--------------------|--------|-------|-------|
| 1      | Senior High School | Man    | Adult | ever  |
| 2      | Senior High School | Woman  | Adult | ever  |
| 3      | Primary School     | Man    | Children | ever |
| 4      | Primary School     | Man    | Children | ever |
| 5      | Primary School     | Woman  | Children | Often |
| 6      | Children Class     | Man    | Children | Often |
| 7      | Diploma            | Woman  | Adult | ever  |
| 8      | Senior High School | Woman  | Adult | Often |
| 9      | Diploma            | Woman  | Adult | Often |

In the calculation or determine the Probability of each Data Variable Class Visitors must use the Bayes rule, as for the rules of the bay as follows:

\[
P(A) = \sum_{i=1}^{2} (P(B_i)P(A|B_i))
\]

\[
= P(B_1)P(A|B_1) + P(B_2)P(A|B_2)
\]

Table 3. Probability Calculation Results on Education Class variables

| P (EDUCATION = | …          | Often | ever |
|-------------|-------------|-------|------|
| Children Class | 8%          | 6%    |
| Primary School     | 33%         | 11%   |
| Junior high school | 17%         | 6%    |
| Senior High School | 33%         | 22%   |
| Diploma            | 8%          | 33%   |
| Bachelor degree    | 0%          | 22%   |
| Total              | 100%        | 100%  |

Table 4. Results of Gender Class Probability Calculation

| P (GENDER = | …          | Often | ever |
|-------------|-------------|-------|------|
| Man         | 42%         | 44%   |
| Women       | 58%         | 59%   |
| Total       | 100%        | 103%  |

Table 5. Results of Calculation of Probability of Age Classes

| P (Age = | | Often | ever |
|---------|----------|-------|------|
| Adult   | 67%      | 56%   |
Calculating Prediction values (Precision) Precision process to find the proportion of visitors with positive diagnosis results. Formula for calculating precision values as follows:
Precision = TP / (TP + FP) (3)

Table 6. Results of Comparative Values

| Prediction | Class | Often | ever |
|------------|-------|-------|------|
| Often      |       | 11    | 6    |
| ever       |       | 1     | 12   |

Thus the results of the accuracy of the calculation of the comparison of each class variable is worth 77%. Where is the 77% accuracy value as the best value.

4. Conclusion
Based on research conducted by the author, it can be concluded that Naïve Bayes Classification is a fairly easy, interesting and simple method that can be used for prediction. By comparing the values between classes of visits with the product class, the Naïve Bayes method is used to obtain an accuracy value of 77% as the Age of Children who often visit the animal park.

Reference

[1] T. R. Patil, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” International Journal Of Computer Science And Applications, ISSN: 0974-1011, vol. 6, no. 2, pp. 256–261, 2013.
[2] T. Calders and S. Verwer, “Three naive Bayes approaches for discrimination-free classification,” Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277–292, 2010.
[3] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, “Application of Classification Method C4 . 5 on Selection of Exemplary Teachers,” in IOP Conference Series, 2018, pp. 1–6.
[4] X. Zhou et al., “Detection of Pathological Brain in MRI Scanning Based on Wavelet-Entropy and Naive Bayes Classifier,” International Work Conference on Bioinformatics and Biomedical Engineering, vol. 9043, pp. 201–209, 2015.
[5] A. McCallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification,” AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41–48, 1998.
[6] T. Li, J. Li, Z. Liu, P. Li, and C. Jia, “Differentially private Naive Bayes learning over multiple data sources,” Information Sciences, vol. 444, pp. 89–104, 2018.
[7] S. Sudirman, A. P. Windarto, and A. Wanto, “Data Mining Tools | RapidMiner: K-Means Method on Clustering of Rice Crops by Province as Efforts to Stabilize Food Crops In Indonesia,” IOP Conference Series: Materials Science and Engineering, vol. 420, no. 12089, pp. 1–8, 2018.
[8] G. Kaur and A. Chhabra, “Improved J48 Classification Algorithm for the Prediction of Diabetes,” International Journal of Computer Applications, vol. 98, no. 22, pp. 13–17, 2014.
[9] M. Faisal, K. & Chowdhury, M. Rahman, A. Hossain, and K. Dahal, “Enhanced Classification Accuracy on Naive Bayes Data Mining Models,” International Journal of Computer Applications, vol. 28, no. 3, pp. 975–8887, 2011.