Jurnal Mantik, 6 (2) (2022) 2527-2534

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Demographic Attribute Selection Model For Prediction Of Election Participation Using Decision Tree

Linda Kushernawati[1], Arif Senja Fitrani[2], Metatia Intan Mauliana[3]

[12,3]Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo, Jawa Timur

E-mail: lindakushernawati@gmail.com, asfjim@umsida.ac.id, metatialiana@umsida.ac.id

ARTICLE INFO

Article history:
Received: Jul 25, 2022
Revised: Jul 30, 2022
Accepted: Aug 10, 2022

Keywords:
Classification,
Prediction,
Elections,
Demographics,
Decision Tree

ABSTRACT

Implementing a democratic general election is expected to produce people's representatives who can channel the people's aspirations. Demographic data is information that discusses a group of people with several related attributes and involves many factors. In this study, we will relate the relationship between the implementation of elections and the condition of demographic data with a benchmark for the form of public participation in the election. By utilizing 2019 election data and Bangkalan Regency demographic data from the Central Statistics Agency (BPS), it is expected to determine the relationship between the two conditions of the dataset on the form of public participation at the polling station (TPS) level. By starting with the Preprocessing step, it implement a classification method with the Decision Tree (DT) algorithm to predict community presence at the polling station level. There are three versions of the dataset that will be used in modeling, namely initial data that has not been selected for attributes (version 1), data that has been chosen using correlation-based attribute selection (version 2), and data that has been selected using chi-square attributes (version 3). The results show version 1 with a prediction of 81%, followed by version 2 with a prediction of 81%, and the last is version 3 with a prediction of 70%. The detachment model's formation with the selection attribute has a different impact, and the relationship between the election dataset and demographics has a significant effect, as indicated by the prediction results of version 2.

Copyright © 2022 Jurnal Mantik.
All rights reserved.

1. Introduction

Indonesia is a democratic country whose government is from the people, by the people, and for the people. With the implementation of a democratic general election as the embodiment of a means of democracy, it is hoped that it will produce people's representatives who can understand matters relating to the aspirations of the people, especially in terms of the process of formulating public policies with the existence of a rotating power system.

One crucial aspect of supporting the success of the election is demographic data. According to George W. Barclay (1970)[1], demography is a science that provides a fascinating picture of the population that is depicted statistically. In addition, demography studies overall behavior, not just individual behavior. Election implementation also needs to pay attention to the characteristics of the population and voters in an area to design the right strategy. Analyzing demographic data makes it possible to obtain a description of the behavior and determinants of voter decisions. According to the research results that predict election participation using the nave Bayes algorithm with 300 instances of 6 attributes, 97% accuracy is obtained [2]. Research that predicts public attendance in elections using the nave Bayes algorithm, the data used has 430 instances and consists of 9 attributes. The Accuracy value is 67% [3].

This study uses the Decision Tree (DT) Algorithm. Previous research that uses the DT algorithm is classifying student data with an accuracy of 97.63% [4]. The research on the prediction of election participation using the DT C4.5 algorithm with 592 instances of data obtained an accuracy of 92.91% [5].

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

In this study, we try to connect the form of community participation in election participation with the demographic conditions that occur in the Bangkalan Regency. This is based on several reasons, including the state of the BPS attribute and label class on the 2019 election recapitulation data. The BPS attributes involved consist of social, educational, health, trade, geographical, and potential village aspects. From various elements owned by Bangkalan Regency and connected to the recapitulation data, the aim is to determine the relationship between the two conditions of the dataset, which is determined by the prediction class on the high and low labels.

The low level of knowledge gained by the community regarding the implementation of elections and the increasing mobility of the people of an area. The low level of information held will impact people's ignorance of the importance of their participation in elections. So, it is essential to predict the level of community participation in the region so that there will be an even distribution of socialization in elections related to increasing the welfare of a part. For example, suppose an area has access to an adequate economic center and a high level of community participation. In that case, it can be concluded that this condition will affect a high level of election participation. This study uses the DT algorithm. There are several advantages of DT compared to algorithms that are easy to learn and interpret for data visualization. Others, namely (1) Can generate information quickly. (2) Can be processed with numeric or categorical scale attributes. (3) The selection of attributes is carried out automatically [6]. If attributes have no effect, it will not affect the final result even though there are correlated attributes.

2. Method

2.1 Data Collection

Data Collection is the initial stage of collecting and measuring information about the targeted variables in a systematic way. Accurate data collection is essential in maintaining the integrity of research[7].

2.2 Preprocessing

This stage includes steps for data cleaning (such as dealing with the removal of noise and missing values), data integration (where multiple data sources are combined into one), Normalization (to avoid anomalies in the data or to reduce the memory size of the data set.), correlation (To identify the extent to which one variable is related to another variable), random dataset (random data collection).

a. Data Cleaning
   Data cleaning or data cleansing is a step that repairs damaged data and reduces unnecessary data details. Handling of missing data (Missing Values) and noise is included at this stage.
b. Data Integration
   The process of merging data from several data stores. This process must be done carefully to avoid redundancy and inconsistency in the data set [8].
c. Normalization
   The unit of measurement used can affect data analysis. All attributes must be expressed in the same unit of measure and use the same scale or range. Min-Max is one of the methods in Normalization. Min-max normalization is a normalization method by performing a linear transformation of the original data to produce a balance of comparison values between the data before and after the process [9]. The Normalization can be used in Equation 1:

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \quad (1)$$

   With a description of each symbol, namely, x_new is the result of normalization, x_old is the original x value, x_min is the minimum value of x, x_max is the maximum value of x.
d. Correlation
   Statistical methods were used to determine the linear relationship between the two variables.
e. Random Data
   Method for taking one or more data at random.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

2.3 Classification

Classification is identifying objects into a category, class, or group based on predetermined procedures, definitions, and characteristics. Classification aims to place objects assigned only to one of the categories called classes [10].

2.4 Prediction

Prediction (forecasting) is an attempt to predict or predict something that will happen in the future by utilizing relevant information from previous times (historically) using the scientific method. Predicting is to obtain information about what will happen in the future with the most significant probability of occurrence. Prediction methods can be done qualitatively through experts' opinions or quantitatively with mathematical calculations. One quantitative prediction method is to use time series analysis [11].

2.5 Decision Tree (DT) Algoritm

This method is in the form of a tree structure, such as a flow chart on a flowchart. Each internal node represents the test on the attribute, each branch represents the test result, and the node or leaf represents the class[12]. The concept of entropy is used to determine which attributes a tree will split. The higher the entropy of a sample, the more impure the sample is [13]. To calculate the Entropy of the sample S[14] is used Equation (2).

$$Entropy(S) = \sum_{i=1}^{n} - p_i * Log_2 p_i \quad (2)$$

Information:
S: case set
n: number of partitions S
pi: the proportion of Si to S

2.6 Evaluation Stage

The evaluation stage is carried out by comparing the accuracy and other testing aspects of the DT algorithm used in the modeling stage. The test results will later become a benchmark in determining whether the DT algorithm can accurately predict the effect of election participation in the Mojokerto district. The output generated from the DT algorithm will be in the form of a decision tree and a confusion matrix[15].

2.7 Research Framework

This study uses the DT algorithm, where the DT algorithm will predict all attributes to produce a decision tree. The dataset used is secondary data sourced from the website, namely the Bangkalan Regency Presidential Election Recapitulation data by the Bangkalan Regency KPU and Bangkalan Regency Demographic Data by the Bangkalan Regency BPS. The dataset consists of 3763 data contained in various characteristics. In comparison to the distribution of data in the dataset, 2522 has a high prediction class while 1242 has a low prediction class. The data will be used to predict the effect of election participation on demography in the Bangkalan Regency.

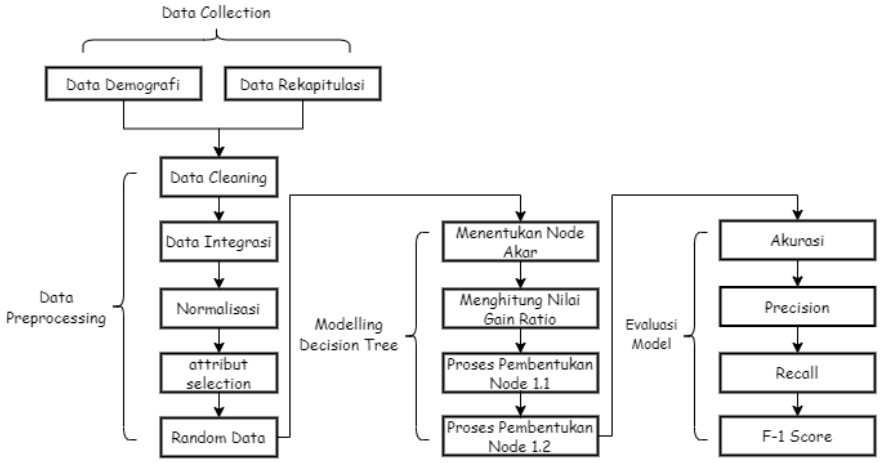

Figure 1. Research Stage

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Figure 1 shows four flow stages applied in this research: Data Collect, Data Preprocessing, Data Modeling, and Evaluation.

3. Result and Discussion

3.1 Data Collect

In the data collection stage, several secondary data sources are used, namely data that is already available and can be obtained from the website of the Bangkalan Regency BPS agency and election recapitulation data. In the dataset, some attributes influence public participation in elections. The dataset consists of 3763 rows and 19 attributes, and there is one attribute as a prediction class. The attributes used are shown in Table 1.

Table 1
Dataset Attributes

Index	Attributes Description
A1	Population
A2	natural disaster early warning system
A3	extraordinary tsunami early warning system
A4	safety equipment
A5	signs and disaster evacuation routes
A6	manufacture, maintain or normalize: rivers, canals, embankments, ditches, drainages, reservoirs, beaches, etc.
A7	mini market/supermarket
A8	convenience store/grocery shop
A9	food stalls/shops
A10	cooperative
A11	number of Base Transceiver Station (BTS)
A12	number of cellular telephone communication service operators reaching out to villages or ward
A13	Cellular phone signal conditions in Most rural areas
A14	type of transportation
A15	presence of public transportation
A16	road surface type
A17	can be passed by vehicles with four or more wheels
A18	post office/postal assistant/post house
A19	private shipping company/agent
Label	Prediction Class Labels (low and High)

3.2 Preprocessing

The dataset used for implementing the algorithm must be of good quality data. To ensure good quality data, the preprocessing stage is first carried out. This stage will begin with Data Cleaning, Integration, Normalization, Attribute Correlation, and Random Datasets.

a. Cleaning Dataset
   This stage is carried out to clean the data from missing values so that the data to be processed will be relevant as needed so that the results obtained can be optimal.

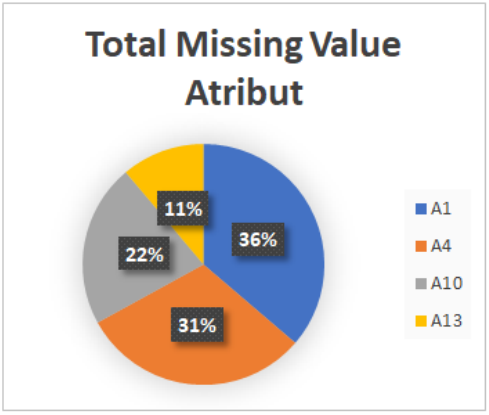

Figure 2. Missing Value Attribute

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

In Figure 2 the dataset has missing values for attributes A1 36%, A4 31%, A10 22%, and A13 11%. To overcome this, handling is needed by replacing the missing value with the attribute's median value.

b. Dataset Integration

Dataset integration combines Bangkalan Regency Demographic Data and Bangkalan Regency Election Recapitulation Data into one dataset. This study uses a concept hierarchy scheme, namely BPS data at the village level while the election data is at the TPS level. This resulted in BPS data making the election data a prediction class, namely the Label attribute. The distribution of data distribution in the election participation dataset can be seen in Table 2.

Table 2
Check Data Distribution In Prediction Class

Low (0)	High(1)	Total Data
1242	2522	3763

c. Attribute Scalling

The third stage is attributed to scaling or better known as dataset normalization. This study uses Min Max Normalization.

d. Attribute Selection

In the attribute selection stage, the attribute analysis process is carried out from the dataset. This is useful in reducing the number of attributes involved in determining the prediction class because redundant and less relevant attributes can affect the computational and evaluation process. In this study, attribute selection analysis uses correlation-based attribute selection and chi-square. The results of correlation-based feature selection on the dataset can be seen in Table 3.

Table 3
Correlation-Based Feature Selection

Attribute	Rank	Attribute	Rank
Label	1.000.000	A6	-0.061633
A13	0.147798	A10	-0.074542
A18	0.072359	A11	-0.107012
A19	0.032270	A1	-0.118499
A4	0.026770	A9	-0.119511
A16	0.016110	A12	-0.130740
A15	0.012966	A8	-0.140068
A2	-0.000516	A3	-0.140894
A5	-0.022787	A7	-0.217859
A17	-0.036122	A14	0
Label	1.000.000	A6	-0.061633

Based on Table 3, A14 is an attribute that does not correlate with the prediction class. Therefore, these attributes need to be dropped to optimize the computational process.

e. Random dataset

The last step is to do a random dataset to ensure each instance's representation.

3.3 Modelling Decision Tree (DT) Algorithm

Data mining is carried out at the modeling stage using Jupyter software with the python programming language. The algorithm used is DT with a comparison of training data and testing data, which is 80:20. The DT modeling process does not use the pruning process. Three versions of the data will be used in the DT algorithm modeling: initial data that has not been selected for attributes, data that has been chosen using correlation-based attribute selection, and data that has been selected using chi-square attributes.

3.4 Data Test Evaluation

In the testing process, the data used are obtained from the initial data from the attribute selection process with Correlation-based Feature Selection and chi-square. The results obtained will be in the form of a confusion matrix.

3.5 Application of Decision Tree (DT) Algorithm

The DT modeling process does not use the pruning process. Tests were carried out using the Confusion Matrix evaluation. These results will form a matrix of true positive and true negative. Testing is carried out based on a predefined version. The results of confusion are shown in Table 4.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Table 4
Confusion Matrix Version 1

Prediction	Correct(Low)	Correct(High)
Pred (Low)	178	70
Pred (High)	74	431

Based on Table 4, version 1 produces predictions of true (low) 178, False (High) 70, False (Low) 70, True (High) 431. So from these results, the values of accuracy, precision, recall, and F1-Score can be determined in Table 5.

Table 5
Evaluation Result Version 1

Evaluation	Version 1
Accuracy	81%
Precision	78%
Recall	79%
F1-Score	78%

Table 5 shows that the accuracy is 81%, precision is 78%, recall is 79%, and F1-Score is 78%.

Table 6
Confusion Matrix Version 2

Prediction	Correct(Low)	Correct(High)
Pred (Low)	180	70
Pred (High)	74	431

Based on Table 6, version 2 produces predictions true (low) 180, False (High) 74, False (Low) 68, True (High) 431. So from these results, it can be determined the values of accuracy, precision, recall, and F1-Score in Table 7.

Table 7
Evaluation Result Version 2

Evaluation	Version 2
Accuracy	81%
Precision	79%
Recall	79%
F1-Score	79%

Table 7 shows that the accuracy is 81%, precision is 79%, recall is 79%, and F1-Score is 79%.

Table 8
Confusion Matrix Version 3

Prediction	Correct(Low)	Correct(High)
Pred (Low)	60	70
Pred (High)	74	466

Based on Table 8, version 3 produces predictions true (low) 60, False (High) 39, False (Low) 188, True (High) 466. So from these results, it can be determined the values of accuracy, precision, recall, and F1-Score in Table 9.

Table 9
Evaluation Result Version 3

Evaluation	Version 3
Accuracy	70%
Precision	66%
Recall	58%
F1-Score	57%

Table 9 shows that the accuracy is 70%, precision is 66%, recall is 58%, and F1-Score is 57%.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Based on the tests that have been carried out, it shows that the dataset that was selected using the attribute version 2 has a difference of 2 data that can be predicted lower than the dataset that is not selected. This resulted in the results of the evaluation of testing using version 2 to be high.

4. Conclusion

Based on testing from 3 different data versions, it shows that the dataset with attribute selection using version 2 (Correlation-based Feature Selection) produces a higher value than version 1 (dataset that does not perform attribute selection), and the dataset with attribute selection using version 3 (chi-square). Of the 19 attributes in the dataset, which were selected using Correlation-based Feature Selection, 18 were used to determine the prediction of the effect of election participation. It can be concluded that using Correlation-based Feature Selection can select attributes well. Based on the results of testing data, which amounted to 753, attributes A7 (minimarket), A10 (Cooperative), A16 (Type of Road Surface), and A17 (Roads can be passed by four wheels or more) have a high influence on the process of predicting the impact of political participation in Bangkalan Regency. Based on this, it can be concluded that if the availability of economic places and roads that can be passed by four or more wheeled vehicles in Bangkalan Regency are excellent and high in number, it influences community participation in elections. This is inversely proportional to attributes A1 (Number of Population) and A8 (Grocery Stores). According to the data testing results, these two attributes have a low influence on predicting the influence of political participation in the Bangkalan Regency. Based on these data, it can be concluded that the greater the number of grocery stores and a large number of residents, the lower public participation in elections.

References

[1] G. W. Barclay, A. J. Coale, M. A. Stoto, and T. J. Trussell, "A reassessment of the demography of traditional rural China," *Popul. Index*, pp. 606–635, 1976.
[2] A. Chowiyah, "Penerapan Data Mining Mengguankan Metode Klasifikasi Naive Bayes untuk Memprediksi Partisipasi pemilihan Gubernur dan Wakil Gubernur di Desa Jemirahan Kecamatan Jabon," *Univ. Muhammadiyah Sidoarjo*, 2019.
[3] A. Hakim, "Prediksi Kehadiran Masyarakat Dalam Pemilihan Umun Dengan Menggunakan Metode Naïve Bayes," *Pros. SeNTIK*, vol. 3, no. 1, 2019.
[4] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, vol. 14, no. 2, pp. 217–224, 2018.
[5] D. Nurrahman, "Algoritma klasifikasi c4. 5 berbasis particle swarm optimization untuk prediksi hasil pemilihan legislatif dprd karawang," *II*, pp. 28–37, 2017.
[6] N. P. Volkova, N. O. Rizun, and M. V Nehrey, "Data science: opportunities to transform education," 2019.
[7] A. K. Fauziyyah, "Analisis sentimen pandemi Covid19 pada streaming Twitter dengan text mining Python," *J. Ilm. SINUS*, vol. 18, no. 2, pp. 31–42, 2020.
[8] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*, vol. 72. Springer, 2015.
[9] T. T. Hanifa, S. Al-Faraby, F. Informatika, and U. Telkom, "Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging," *vol*, vol. 4, pp. 3210–3225, 2017.
[10] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, "Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris," *J. Teknol. Inf. dan Ilmu Komput. (JTIIK), p-ISSN*, pp. 2355–7699, 2017.
[11] N. Nurmahaludin, "ANALISIS PERBANDINGAN METODE JARINGAN SYARAF TIRUAN DAN REGRESI LINEAR BERGANDA PADA PRAKIRAAN CUACA," *J. INTEKNA Inf. Tek. dan Niaga*, vol. 14, no. 2, 2014.
[12] S. D. Jadhav and H. P. Channe, "Comparative study of K-NN, naive Bayes and decision tree classification techniques," *Int. J. Sci. Res.*, vol. 5, no. 1, pp. 1842–1845, 2016.
[13] D. Oktafia and D. L. Pardede, "Perbandingan Kinerja Algoritma Decision Tree dan Naïve Bayes dalam Prediksi Kebangkrutan," *J. Ilm. Ilmu Komput. Progr. Stud. Sist. Inf.*, 2010.
[14] F.-J. Yang, "An extended idea about decision trees," in *2019 International Conference on*

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Computational Science and Computational Intelligence (CSCI), 2019, pp. 349–354.
[15] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf. Sci. (Ny).*, vol. 340, pp. 250–261, 2016.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).