SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 141-148
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.609

Model Analysis of the 2019 Election Participation Level Against Demographics in Pamekasan Regency Using the Naive Bayes Method

Maulana Habib Firmansyah, Arif Senja Fitrani*, Azmuri Wahyu Azinar, Suhendro Busono

Department of Computer Science, Muhammadiyah University of Sidoarjo, Sidoarjo, Indonesia

*Corresponding Author: asfjim@umsida.ac.id

Abstract. This study aims to analyze the level of participation in the 2019 elections in Pamekasan Regency based on demographic data using the Naive Bayes classification method. The data used consisted of 189 instances and 208 predictor attributes obtained from the Central Statistics Agency (BPS) publication. The analysis process involved preprocessing, feature selection, and model evaluation stages. The test results showed a model accuracy of 66%, with the highest f1-score value in the high participation class. Further analysis also shows that most subdistricts and villages in Pamekasan have high participation rates. In addition, a very strong correlation was found between demographic attributes that have the potential to be important predictors of voter engagement. These findings provide an initial overview to understand the factors that influence community participation in elections.

Keywords: Elections, Prediction, Participation, Naïve Bayes

1 Introduction

General elections (PEMILU) are important political events that determine who will lead a democratic country. Elections are a cycle of political activities that fight for political interests to elect representatives and leaders in order to realize democracy. Elections are a cycle of political activities that accommodate the interests of the people, which are then packaged into various policies[1].

In elections, voter demographics are influential. Population is often associated with demographics. Statistical data on the population compiled based on classifications such as age, race, gender, religion, occupation, and education, as well as birth rates, death rates, population density, income levels, and so on, are called demographics. Demographic data is very important for government policy. The government often uses demographics to make policies and allocate resources[2].

Across Indonesia, public participation in the 2019 elections increased to 81.93% from the previously projected 77.5%. Many factors can influence this participation rate, one of which is regional demographics. Every year, the Central Statistics Agency (BPS) of Pamekasan Regency publishes data containing information on all aspects of each village in each sub-district. One of the aspects discussed in this data publication is demographics. Data on demographics can be linked to data on community participation in general elections[3].

Due to the rapid advancement in technology and data analysis today, data mining techniques can be used to find patterns and trends in election data. One of the classification methods used in data mining, the Naive Bayes Classifier, is a classification method used in data mining and is useful for predicting the value of target category variables[4].

2 Method

This study is qualitative in nature. Demographic data collected from the Central Statistics Agency (BPS) of Pamekasan Regency for this study aims to predict election participation. The data will be processed using the naive Bayes classification method to identify who is involved in the election[5]. The data is divided into two parts: training data and testing data. The training data is used to create a classification model using a classification algorithm, and the testing data is used to evaluate the classifier's ability to classify correctly[6].

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 141-148
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.609

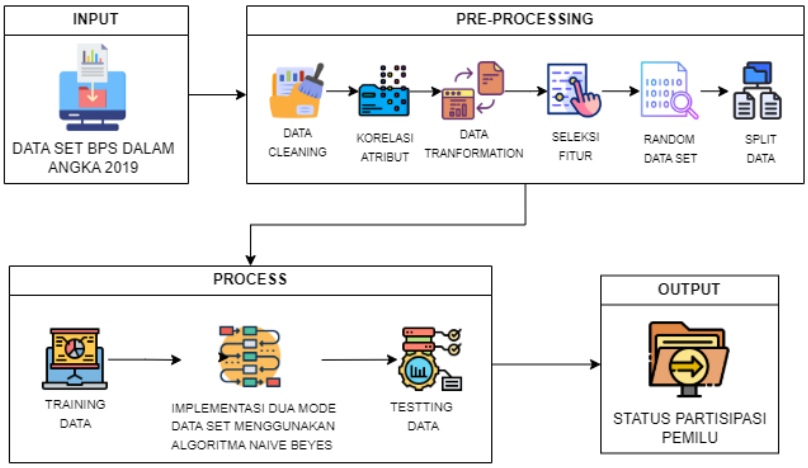

Figure 1. Research Flow

Figure 1 shows the stages of input, pre-processing, processing, and output. The following is an explanation of each stage.

2.1 Input

Data was collected at the initial stage of this study. This study used data published in 2019 from Pamekasan Regency and data summarizing the 2019 election results. This published data can be obtained directly from the official website of the Central Statistics Agency (BPS)[7]. The data consists of 85 predictor attributes and one target attribute, which are presented in Table 1.

Table 1. Data attributes from BPS

Attribute No.	Attribute Group	Attribute
X1 – X12	Geography and Climate	Area {X1}, Elevation above Sea Level {X2}, Coastline {X3}, Lowlands {X4}, Highlands/Mountains {X5}, River Name {X6}, River Length {X7}, Village Reservoir {X8}, Type and risk of disaster: Earthquake {X9}, Tsunami {X10}, Volcanic Eruption {X10}, Flood {X11}, Drought {X12}
X13 – X24	Government and Administration	Number of Hamlet Heads{X13}, Number of Modins{X14}, Number of BPD Members{15}, Distance to Subdistrict Office facilities{X16}, Distance to Police Station facilities{X17}, Distance to Hospital facilities{X18}, Distance to Community Health Center facilities{X19}, Number of Hamlets{X20}, Number of Neighborhood Units (RW){X21}, Number of Neighborhood Groups (RT){X22}, Early warning system for disasters{X23}, Signs and evacuation routes{X24}
X25 – X58	Population and Housing	Number of male residents{X25}, Number of female residents{X26}, Total number of residents{X27}, Sex ratio{X28}, Population density{X29}, Household density{X30}, Distribution of residents by age 00-04{X31}, Population distribution by age 04-09{X32}, Population distribution by age 10-14{X33}, Population distribution by age 15-19{X34}, Population distribution by age 20-24{X35}, Population distribution by age 25-29{X36}, Population distribution by age 30-34{X37}, Population distribution by age 35-39{X38}, Population distribution by age 40-44{X39}, Population distribution by age 45-49{X40}, Population distribution by age 50-54{X41}, Population distribution by age 55-59{X42}, Population distribution by age 60-64{X43}, Population distribution by age 65+{X44}, Type of private toilet{X45}, Type of shared toilet{X46}, Type of public toilet{X47}, No toilet{X48}, City gas{X49}, 3kg LPG{X50}, More than 3kg LPG{X51}, Kerosene{X52}, Firewood{X53},

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 141-148
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.609

		Bottled water{X54}, Refill{X55}, PDAM{X56}, Well{X57}, Spring water{X58}
X59- X79	Social	Facilities and ease of access to elementary schools{X59}, Facilities and ease of access to Islamic elementary schools{X60}, Facilities and ease of access to junior high schools{X61}, Facilities and ease of access to Islamic junior high schools{X62}, Facilities and ease of access to senior high schools{X63}, Facilities and ease of access to Islamic senior high schools{X64}, Facilities and ease of access to vocational schools (SMK), Facilities and ease of access to universities (PT), Number of elementary schools (SD), Number of Islamic elementary schools (MI), Number of junior high schools (SMP), Number of Islamic junior high schools (MTS), Number of senior high schools (SMA), Number of Islamic senior high schools (MA), Number of Vocational High Schools (SMK) {X73}, Number of Universities (PT) {X74}, Infants with poor nutrition {X75}, Soccer facilities {X76}, Volleyball facilities {X77}, Swimming facilities {X78}, Martial arts facilities {X79}
X80 – X89	Agriculture	Agricultural land area{X80}, Non-agricultural land area{X81}, Type of rice field{X82}, Type of non-rice field{X83}, Type of farmland{X84}, Type of community forest{X85}, Type of technical irrigation{X86}, Semi-technical irrigation type{X87}, Simple irrigation type{X88}, Rainfed irrigation{X89}
X90 – X96	Industry, Mining, and Energy	PLN electricity{X90}, non-PLN electricity{X91}, non-electricity{X92}, buildings and surrounding land{X93}, state forest{X94}, swamp{X95}, road{X96}
X97 – X102	Trade	Permanent market{X97}, Semi-permanent market{X98}, Minimarket{X99}, Grocery store{X100}, Restaurant{X101}, Hotel{X102}
X103 – X114	Transportation and Communication	Paved roads{X103}, Hardened roads{X104}, Dirt roads{X105}, Roads accessible to vehicles with 4 or more wheels{X106}, Land transportation facilities{X107}, Water transportation facilities{X108}, Air transportation facilities{X109}, BTS/cell towers{X110}, Signal conditions{X111}, Service providers{X112}, Post office{X113}, Courier services{X114}
X115 – X120	Finance	Government banks{X115}, private banks{X116}, BPR{X117}, KUD{X118}, Kopinkra{X119}, Kospin{X120}
Y	Elections	Level of Community Participation

2.2 Pre-Processing

At this stage, data *pre-processing* is carried out, as explained below.

A. Data Cleaning

Data cleaning is the process of finding errors such as duplication, inconsistency, and incomplete data. Then, decisions about the data are made, such as deleting inappropriate data or correcting it[8].

B. Attribute Correlation

Attribute 1	Attribute 2	Correlation
G136	G142	1.0
G32	G34	0.999999602423443
G233	G236	0.9999994618824921
G34	G52	0.9999968631680487
G32	G52	0.9999964357459555
G72	G73	0.9990849107531108
G53	G54	0.9987828735811418
G84	G85	0.9954482335602096
G78	G79	0.9948481387569174
G58	G59	0.9907389852669041

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 141-148
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.609

Correlation analysis between attributes is performed to evaluate the strength of the relationship between predictor attributes and target attributes, namely the level of election participation. Based on the table, there are several attribute pairs that have very high correlation values, such as G136 with G142 (1.0), G32 with G34 (0.999999), and G233 with G236 (0.999999). Correlation values close to 1 indicate that the attribute pairs have very high similarity in terms of the information they contain. This step aims to filter attributes that have a strong relationship with the target, while reducing attributes that are duplicative or less relevant. By eliminating attributes that do not contribute significantly, the classification process can run more efficiently and the results of the model built will be more accurate[9].

C. Data Transformation

```
G1      object
G2      object
G3      object
G4      object
G5      object
         ...
G265     int64
G266     int64
G267     int64
G268     int64
P       object
Length: 208, dtype: object
Hasil setelah Transformasi Data:
   G1   G2   G3   G4  G5   G7  G9  G10  G11  G12  ...  G260  G261  G262  G263  \
0   0  118  150  154  72  0.0  24   24   37   29  ...   1.0   1.0   0.0   0.0
1   0   32   63   43  67  0.0  24   24   37   29  ...   1.0   1.0   0.0   0.0
2   0   33   62   41  58  0.0  19   19   28   19  ...   1.0   1.0   0.0   0.0
3   0   34   65   46  23  0.0  25   25   41   30  ...   1.0   1.0   0.0   0.0
4   0  131  153  156  49  0.0  17   18   22   19  ...   1.0   1.0   0.0   0.0

   G264  G265  G266  G267  G268    P
0   0.0   0.0   0.0   0.0   0.0   71
1   0.0   0.0   0.0   0.0   0.0   47
2   0.0   0.0   0.0   0.2   0.0  172
3   0.0   0.0   0.0   0.0   0.0  142
4   0.0   0.0   0.0   0.0   0.0  177
```

In the data transformation stage, all attributes are converted into numerical format to suit the needs of the data mining process. As shown in the figure, the initial attributes that were previously object types (such as G1 to G5) have been converted into numerical values. The results of this transformation show that all columns now consist of numbers, including the target attribute "P", making it easier to further process the data using classification algorithms or statistical analysis[10].

D. Feature Selection

Feature selection is the process of selecting the most relevant subset of features from a dataset for use in modeling[11]. This process involves:

1. Evaluating the importance of each feature
2. Removing redundant features
3. Selection of features that have a significant influence on classification results
4. Reducing the dimension of data to improve computational efficiency

E. Random dataset

Random datasets are used to randomize and maximize the representativeness of the weight of each data row on all attributes[12].

F. Data split

The processed data is then divided into 80% as *training* data and 20% as *testing* data.

2.3 Process

This stage involves the implementation of the Naive Bayes classification algorithm on data that has undergone pre-processing. The dataset is divided into two parts, namely 80% for training data (151 data) and 20% for testing data (38 data). The Naive Bayes method was chosen because of its efficiency in managing small to medium-sized data and its reliability in processing attributes that are assumed to be independent of each other.

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 141-148
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.609

At this stage, the classification model was built based on the training data to learn the relationship patterns between demographic attributes and community participation levels. After the model was formed, it was tested on previously unseen data to predict the participation category, namely high or low. This process included several important steps, such as converting all attributes into numerical format, separating the predictor and target variables, forming a classification model using the Naive Bayes probabilistic approach, and testing the model to generate predictions on new data.

2.4 Output

This stage produces the output from the classification process that has been carried out. The model that has been built is used to predict the participation category of 38 test data, which are grouped into two classes: high participation (1) and low participation (0). The classification results are then analyzed using a confusion matrix, which shows the distribution of model predictions against the actual labels. Based on these results, 16 data points were correctly identified as low participation, 9 data points were correctly classified as high participation, 2 data points were misclassified as low when they should have been high, and 11 data points were misclassified as high when they should have been low.

In addition, the model was also evaluated using a classification report that included metrics such as precision, recall, f1-score, and the number of data per class (support). These metrics provide a more in-depth picture of how well the model recognizes each category. To support the interpretation of the results, visualizations in the form of graphs were used to show classification patterns, thereby facilitating analysis of the model's performance in mapping community participation levels in the test data[14].

2.5 Analysis/Evaluation

An evaluation was conducted to assess the model's performance based on the classification results. The model produced an accuracy of 66%. For the high participation class, the f1-score reached 0.71 with a recall of 0.89, indicating fairly good performance. However, in the low participation class, the f1-score was only 0.58 due to low recall (0.45), even though the precision was high (0.82).

Overall, the model is more effective in recognizing high participation. In the future, it is recommended to balance the data and consider other classification methods to improve accuracy[15].

3 Results and Discussion

This study utilized 189 data instances with 208 features or attributes, where the target attribute was classified into two categories: class 1 indicated high participation, while class 0 represented low participation. The implementation was carried out using the Naïve Bayes algorithm with the Python programming language.

3.1 Accuracy and Precision

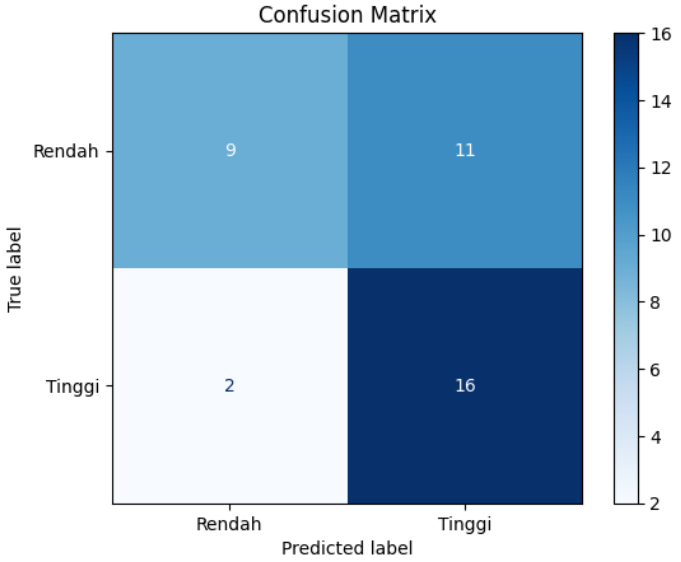

Figure 2. Confusion Matrix

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 141-148
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.609

Figure 2 shows the confusion matrix results of the Naive Bayes model. From the analyzed data, 2 data were misclassified as low participation, while 9 data were correctly classified as high participation. In addition, there were 16 data that were correctly classified as low participation, while 11 other data were misclassified as high participation.

Table 3. Classification Report

	precision	recall	f1 - score	support
0	0.82	0.45	0.58	20
1	0.59	0.89	0.71	18
accuracy			0.66	38
macro average	0.71	0.67	0.65	38
weighted average	0.71	0.66	0.64	38

Table 3 shows that the model performs quite well in classifying class 1 (high participation), with a precision value of 0.59, recall of 0.89, and f1-score of 0.71. This indicates that the model is able to recognize most of the data from class 1 despite its moderate precision. Conversely, for class 0 (low participation), the model's performance is less than optimal. Although the precision value is quite high (0.82), the recall value is only 0.45, which means that the model often fails to detect data that actually belongs to this class. The f1-score value of 0.58 indicates moderate performance for this class. Overall, the model achieved an accuracy of 66%, with an average macro F1-score of 0.65, which indicates balanced but not yet optimal performance in both classes. The fairly balanced amount of data between class 0 (20 data points) and class 1 (18 data points) shows that class imbalance is not a major factor in the variation in the performance of this model.

3.2 Participation Rate Percentage

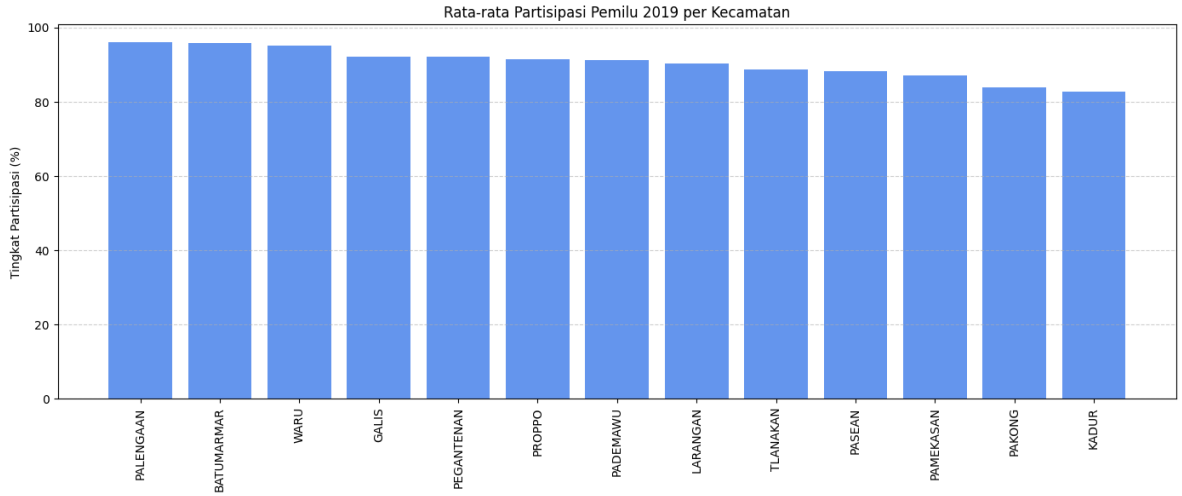

Figure 3. Percentage of Participation by Subdistrict

Figure 3 shows the average election participation rate by subdistrict in Pamekasan Regency. From the data presented, all subdistricts have a high participation rate, with an average participation percentage ranging from 82.87% to 96.03%. These results indicate that all subdistricts in Pamekasan Regency have high participation rates in elections. These findings provide a general picture that, overall, community participation in Pamekasan Regency is very good.

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 1, February 2025, Pages 141-148
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i1.609

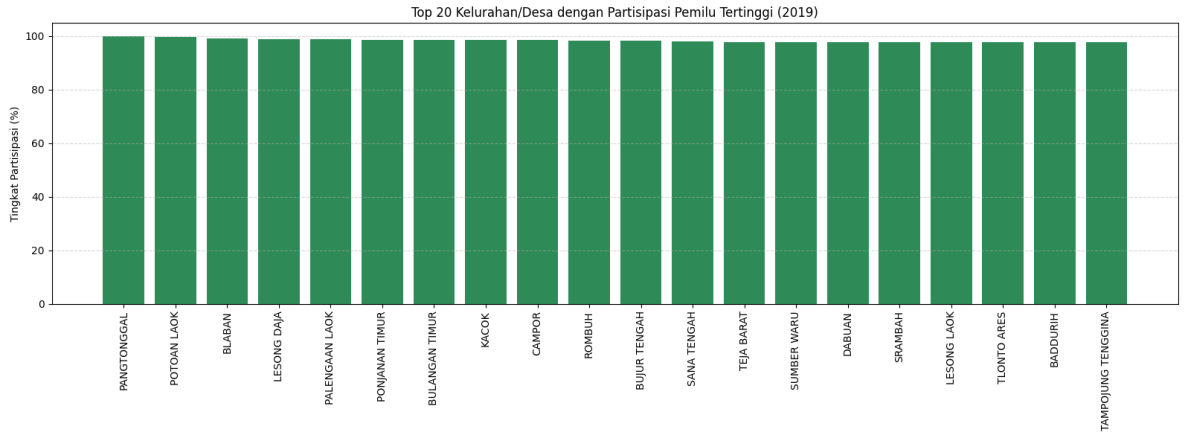

Figure 4. Percentage of Participation in the 20 Highest Villages

Figure 4 shows the 20 subdistricts/villages with the highest election participation rates in Pamekasan Regency. The data shows that all villages/subdistricts on this list had participation rates above 97%, reflecting the community's high enthusiasm for participating in the elections. These results indicate that in some areas, community participation was very high, demonstrating the success in increasing political awareness and democratic participation at the local level.

4 Conclusion

The results of this study conclude that the Naïve Bayes algorithm can be used to classify election participation rates based on demographic factors in Pamekasan Regency. This method is capable of producing an accuracy of 66%, with fairly good performance in recognizing the high participation category. Model evaluation shows that certain demographic attributes, such as the number of male and female residents and Muslims, have a very strong correlation and can be used as significant predictor variables in influencing community participation rates. In general, these results indicate that election participation in the Pamekasan region is at a good level, and demographic factors play an important role in determining patterns of community involvement in elections.

Further research is recommended to develop and compare the performance of other classification algorithms, such as Decision Tree, Random Forest, or SVM, to improve the accuracy of election participation predictions. In addition, further exploration of other demographic variables that have not been studied in depth is needed. The use of a broader and more representative dataset from various regions can also provide more comprehensive insights into voter participation patterns. The results of this study can be used as a reference by the government and election organizers in formulating strategies to increase participation, especially in areas with low participation rates.

5 Acknowledgments

The researchers would like to thank all parties involved for assisting them in the smooth running of this research.

References

[1] A. S. Fitriani, "JTAM (Journal of Theory and Application of Mathematics) Application of Data Mining Using the Naïve Bayes Classification Method to Predict Participation in Gubernatorial Elections," vol. 3, no. 2, pp. 98–104, 2019, doi: 10.31764/jtam.v3i2.995.

[2] F. Setiawan, A. S. Fitrani, and A. Eviyanti, "National Seminar & Call Paper Faculty of Science and Technology (SENASAINS 5 th," 2022. [Online]. Available: https://pemilu2019.kpu.go.id

[3] Y. Raharja, A. Senja Fitrani, R. Dijaya, and F. Science and Technology, "CLASSIFICATION OF ELECTION PARTICIPATION LEVELS BASED ON INDUSTRY SECTORS USING THE NAÏVE BAYES ALGORITHM," *Jurnal TEKINKOM*, vol. 7, no. 1, 2024, doi: 10.37600/tekinkom.v7i1.1204.

[4] A. W. Anggraeni, A. S. Fitrani, and A. Eviyanti, "Application of the Support Vector Machine Algorithm to Predict Election Participation Levels on Education Quality," *Edumatic: Journal of Informatics Education*, vol. 8, no. 1, pp. 21–27, Jun. 2024, doi: 10.29408/edumatic.v8i1.24838.

[5] D. Mizta Chulloh and A. Senja Fitrani, "Accuracy Test of K-Means in Predicting Election Participation in the Demographics of Pasuruan District Region."

[6] F. Harahap, N. E. Saragih, E. T. Siregar, and H. Sariangsah, “Fitriana Harahap Application of Data Mining with the Naive Bayes Classifier Algorithm in Predicting Paint Purchases KEYWORDS Data Mining Purchase of paint Naive Bayes. CORRESPONDENCE.”

[7 D. E. Safitri and A. S. Fitrani, “IMPLEMENTATION OF CLASSIFICATION METHODS WITH SUPPORT VECTOR MACHINE KERNEL GAUSSIAN RBF ALGORITHMS FOR PREDICTING ELECTION PARTICIPATION IN SURABAYA CITY DEMOGRAPHY,” *Indonesian Journal of Business Intelligence (IJUBI)*, vol. 5, no. 1, p. 36, Jun. 2022, doi: 10.21927/ijubi.v5i1.2259.

[8] W. Indah, S. Sinaga, R. Buaton, H. Sembiring, and S. Kaputama, “Classification of Population Data in General Elections in Binjai City Using the K-Means Algorithm (Case Study: KPU Binjai City),” 2023.

[9] R. I. Borman and M. Wati, “Application of Data Mining in Classifying Data on Members of the Sejahtera Bandarlampung Credit Union Using the Naïve Bayes Algorithm.”

[10] W. I. Rahayu, A. Anindita, and M. N. Fauzan, "DETERMINATION OF VOTER DATA VALIDATION AND CLASSIFICATION OF THE RESULTS OF THE BONE REGENCY DPRD ELECTIONS TO PREDICT THE WINNING PARTY USING THE NAIVE BAYES METHOD, D4 Informatics Engineering Study Program 123, Indonesia Post Polytechnic 123," 2022.

[11] M. Younus, Suswanta, and Muchamad Zaenuri, “Public-Private Collaboration to Overcome the Digital Divide in Digital Transformation of Government,” *Digital Zone: Journal of Information and Communication Technology*, vol. 15, no. 1, pp. 28–41, May 2024, doi: 10.31849/digitalzone.v15i1.17027.

[12] I. M. A. A. D. Putra, I. M. G. Sunarya, and I. G. A. Gunadi, “Comparison of Naive Bayes Algorithm Based on Feature Selection Gain Ratio with Conventional Naive Bayes in Predicting Hypertension Complications,” *JTIM : Journal of Information Technology and Multimedia*, vol. 6, no. 1, pp. 37–49, Apr. 2024, doi: 10.35746/jtim.v6i1.488.

[13] Y. Nanda Khoiril Umat *et al.*, "Some rights reserved BY-NC-SA 4.0 International License ANALYSIS OF FACTORS AFFECTING THE SELECTION OF THE GOVERNOR OF THE SPECIAL REGION OF JAKARTA USING THE NAIVE BAYES ALGORITHM AND LOGISTIC REGRESSION 1)," vol. 9, no. 2, pp. 211–224, 2024, doi: 10.36341/rabit.vxix.xxx.

[14] H. Rusli, “Predicting General Election Results Based on Social Media Data Using Naive Bayes Data Mining Techniques,” 2025. [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR

[15] S. Yanah, W. Purbaratri, S. Paylina, A. Novita, I. Safitri, and N. K. Tachjar, “SWADHARMA (JEIS) ELECTION APPLICATION SENTIMENT ANALYSIS USING THE NAIVE BAYES ALGORITHM”.