Jurnal Mantik, 6 (2) (2022) 2369-2376

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

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

Prediction Of Election Participant With Malang City Demographic Data Using The K-Nn Algorithm

Nurtia Suryani[1], Arif Senja Fitrani[2]

[1*,2]Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo, Jawa Timur

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

ARTICLE INFO

Article history:
Received: Jul 15, 2022
Revised: Jul 30, 2022
Accepted: Aug 10, 2022

Keywords:
K-Nearest Neighbour,
Dataset Normalization,
Min Max Normalization,
Data Mining,
Prediction of Voter Participation

ABSTRACT

Election (General Election) is a step to choose and determine the figure of a leader. In general election activities, the higher the level of political participation indicates that the people understand the importance of democracy. On the other hand, if the level of participation is low, the people are less concerned about state problems. From the 2019 election activities in Malang City, the next step is connecting with demographic data sourced from the Central Statistics Agency (BPS). The demographic data includes aspects of Energy, Geographic, Education, Health, Population, Economy, Communication, Transportation, and Expedition, which are then integrated with election data. In the 2019 Presidential Election, the number of DPT (Permanent Voters List) was 623,185, while the number of citizens who exercised their right to vote was only 488,587. This study will look at the relationship of demographic data to public participation in implementing elections. Using the classification method for prediction with high and low label classes on the form of community participation at the polling station (TPS) level. In the preprocessing stage, the dataset model is determined by testing three types of normalization methods, then implemented in the K-Nearest Neighbor (K-NN) algorithm. From the test results, the highest level of accuracy obtained in predicting voter participation is 61.83%, and the F-1 score is 61.46%, with the Min Max normalization model occupying the best results.

1. Introduction

In simple terms, democracy is the "government of the people, by the people, and for the people." However, it is not easy to realize this meaning because democracy requires critical stages and processes that must be passed. In a country that adheres to the notion of democracy, elections are the primary key to creating democracy. Elections are a tangible form of people's means of expressing opinions to the state, based on the 1945 Constitution of the Republic of Indonesia and Pancasila. Elections In state activities, the higher the level of political participation indicates that the people understand the importance of democracy. On the contrary, if the level of participation is low, the people are less concerned about state problems [1].

Demography is an aspect of grouping individuals according to socio-economic characteristics or status of a social group. This status is measured by an index representing the seven main components of socioeconomic status: income, education, gender, age, neighborhood, place of birth, and political affiliation. From the point of view of election organizers, this demographic data is essential because the number of population calculations will be the basis for data collection for logistical and administrative planning [2].

Research on the prediction of election participation has also been carried out by [3] using the Naïve Bayes Algorithm method, obtaining an Accuracy value of 68.17%, with 80% training dataset split and 20% testing dataset split. While the 90% split training dataset and 10% split dataset testing get an Accuracy value of 75.56%. In conclusion, the Naïve Bayes algorithm has a good performance. In this study, the attributes used were five attributes with a categorical data set model. Furthermore, research by [4] uses the Naïve Bayes Algorithm with seven attributes with a categorical dataset model, which obtains an accuracy value of 78.95% with 70% split dataset training and 30% split dataset testing.

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

Previous research has applied the K-NN algorithm to classify such as student study periods[5]. This research uses 72 attributes with categorical datasets. From the six experimental scenarios, the highest accuracy value was obtained in the course attributes, with an accuracy rate of 75.95%. Another research that discusses the K-NN algorithm is about Indoor location tracking employees [6] by using six numerical scale attributes so that it is obtained that the system can detect employee positions with an accuracy of 86.18%. Research that discusses the comparison of dataset normalization using the K-NN Algorithm [7] with 11 numerically scaled attributes obtained accuracy using preprocessing with the normalization method is not better than the accuracy of previous research conducted by Arandika et al. 2014 with an accuracy rate of 68.75% and the Sacrament & Siahaan research. 2017 by 72.97%. In this study, we try to relate the impact of demographic conditions on the form of public participation in the 2019 general election. Demographic data through the Central Statistics Agency (BPS), which publishes demographic conditions at the village level every year, can be used as primary data in the study.

The case that often arises in election participation is the rise and fall of the number of residents from year to year in an area, one of which is due to changing demographic conditions. In the 2014 Malang presidential election, the number of DPT reached 611,246 people. On the other hand, the number of people who exercised their right to vote only reached 441,425 people. Next, in the 2019 presidential election, the number of DPT came to 623,185 people. Meanwhile, the number of citizens who exercised their right to vote was only 488,587 [8]. To support the election's success, it is necessary to disseminate information about the importance of elections to the public. Find areas that need more socialization, which can be seen from the community's demographics. This is based on several conditions on the Class label on the election recapitulation data and BPS 2019 attributes. Related BPS attributes such as (1) Social and people's welfare, (2) Agriculture, Forestry, Livestock, and Fisheries, (3) Geography and Climate, (4) Government Position, (5) Tourism and Sports, (6) Economy and Trade, (7) Energy, (8) Communication, Transportation, and Infrastructure[9]. For example, the relationship between demographic data and participation is if the road level in an area is often traversed by 4-wheeled vehicles or more, and the level of community participation is also high. It can be concluded that the condition of the road atmosphere will affect high election participation.

This study uses the K-NN algorithm. The K-NN algorithm has advantages compared to other algorithms, namely (1) The training process is faster, (2) Simple and easy to learn, (3) Able to process data that has outliers and noise, and (4) it is Effective if the training data is large. The use of the K-NN algorithm is expected to help see the effect of demographic data on the level of public election participation in Malang City.

2. Method

2.1 Data Collect

Data Collect is the process of collecting, measuring, and analyzing various types of information using various methods such as interviews, observations, and literature studies. The purpose of data collection is to obtain information from experts in the field, and later the data will be processed to get helpful information[10].

2.2 Preprocessing

Data preprocessing is a set of methods applied to the database to remove noise, missing values, and inconsistent data. This preprocessing data is used because the real-time database data is often incomplete and inconsistent, causing data mining results to be inaccurate and inaccurate. Therefore, to improve the quality of the data to be analyzed, it is necessary to try the steps in preprocessing the dataset, starting from data cleaning, data integration, correlation analysis, and normalization.

a. Data Cleaning: It is a procedure to ensure the data's consistency, correctness, and usefulness. The trick is to detect corruption or error in the data. After that, delete or repair the data if needed[11].
b. Data Integration: An amalgamation of data from various databases into one new database. This data integration must be done carefully because errors in data integration produce distorted results [12].
c. Normalization: It is the process of making some variables have the same value range, not too big or too small, to make statistical analysis easier[13]. This normalization itself has three methods such as: Min – Max Normalization is a method that is relatively easy to implement because it includes a normalization method that is linear with the original data. However, this method can lead out of

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

bounds in some cases. Because of this deficiency, MinMax is not suitable for real-time analysis or evolving systems. MinMax is highly recommended for cases based on time frame analysis and forecasting.

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

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. Z-score is a method that is often used in various data mining or data science-based research. Z-score is a normalization method based on the data's mean (average value) and standard deviation (standard deviation). In addition, there are not many variables set in the calculation. Z-Score is very dynamic in performing normalization calculations[14]. The weakness of the Z-Score is that the process will repeat itself if new data is entered. In addition, the elements needed for the Z-Score analysis also require a reasonably long process, either the standard deviation or the average of each column.

$$x_{new} = \frac{x_{old} - \mu}{\sigma} \quad (2)$$

With a description of each symbol, namely, x_new is the result of normalization, x_old is the original x value. Is the population mean. Is the population standard deviation.

Simple Feature Scaling is dividing each attribute value by the maximum attribute value.

$$x_{new} = \frac{x_{old}}{x_{max}} \quad (3)$$

With a description of each symbol, namely, x_new is the result of normalization, x_old is the original x value, x_max is the maximum x value.

d. Correlation Analysis: is a statistical evaluation method used to review the relationship between two continuous variables measured numerically [15].

e. A random dataset is a random dataset that aims to ensure each instance's representation[16].

2.3 Classification

Classification is a way of grouping objects based on the characteristics of the object of classification. In the process, classification can be done in many ways, either manually or with the help of technology. A classification done manually is a classification human do without using intelligent computer algorithms. While the classification uses technology, it has several algorithms, including Naïve Bayes, Support Vector Machine, Decision Tree, and K-NN[17].

2.4 Prediction

Prediction is estimating something that might happen in the future based on information that has been pocketed from the past and present in a systematic and coherent manner so that the error level can be minimized. In prediction, you do not have to have a correct and definite answer but must try to find a solution that is as accurate as possible regarding something that will happen in the future.

2.5 KNN Algoritm

The K-NN algorithm is a method of classifying objects based on the training data closest to the object. Nearest Neighbor is an approach to finding cases by calculating the proximity between new and old cases based on the appropriate weights on a set of existing features[18].

$$d_i = \sqrt{\sum_{i=1}^{p} (X_{2i} + X_{1i})^2} \quad (4)$$

2.6 Research Stage

This study uses an experimental model with a quantitative approach. The data collection comes from the agency's official website that is relevant to the topic raised. The data source used is secondary data obtained from the official website of the KPU and BPS agencies, and the data processing technique used is the data mining process. Research stage can be seen in Figure 1.

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

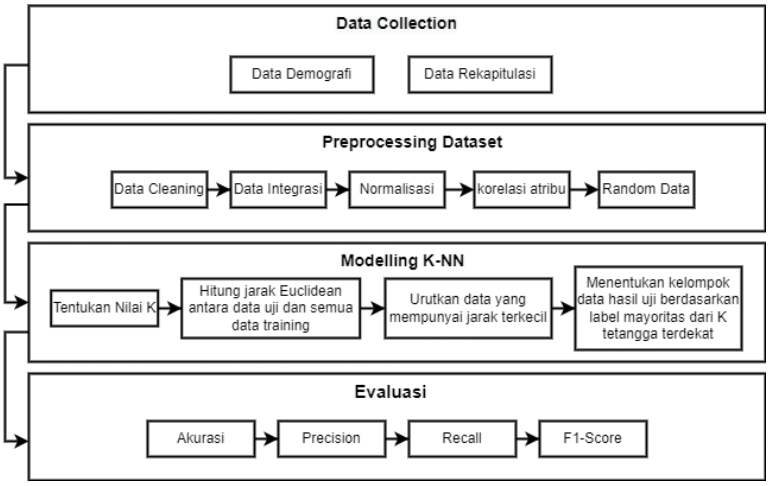

Figure 1. Flow of Stages of Predicting the Effect of Election Participation on Demographics

Based on Figure 1, it can be explained that the stages used in this research consist of data collect, data preprocessing, data modeling, and evaluation.

3. Result and Discussion

3.1 Data Collect

The data used for this research is the demographic data of the city of Malang obtained from the Malang City BPS website and the election recapitulation data obtained from the 2019 Election website. The BPS data has 58 records with 19 attributes consisting of (1) population, (2) natural disaster early warning system, (3) extraordinary tsunami early warning system, (4) safety equipment, and (5) signs and routes. Disaster evacuation, (6) construction, maintenance, or normalization of: rivers, canals, dams, ditches, drainages, reservoirs, beaches, etc., (7) minimarkets/supermarkets, (8) shops/grocery stalls, (9) stalls/shops food, (10) cooperatives, (11) the number of cell phone towers (BTS), (12) the number of cellular telephone communication service operators reaching out to villages/kelurahan, (13) the condition of the cell phone signal in most villages/kelurahan1, (14) types of transportation, (15) the existence of public transportation, (16) types of road surfaces, (17) can be passed by vehicles with four or more wheels motorized, (18) post offices/sub-posts/post houses, (19) companies/service agents private expedition.Furthermore, the Election Recapitulation data has 2365 records with six predictor attributes and one outcome attribute. Attributes consist of (1) sub-district, (2) village, (3) TPS, (4) DPT, (5) abstention, (6) percentage of abstention, (7) attendance, and (8) percentage of attendance.

3.2 Preprocessing

This study's auxiliary application used for processing and visualizing data uses Anaconda tools and the python programming language.

a. Data Cleaning

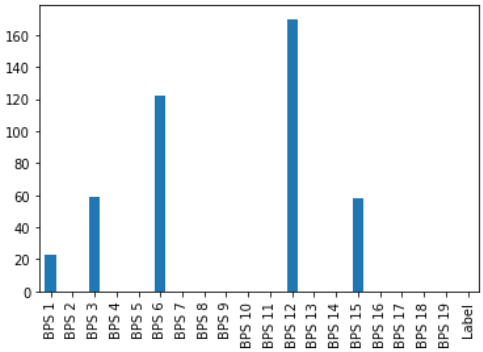

Figure 2. Checking the Missing Value on the Dataset

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

Figure 2 shows that several attributes have missing values below 30% in the BPS dataset, namely BPS 1, BPS 3, BPS 6, BPS 12, and BPS 15 attributes. Those attributes.

b. Data integration

In this study, two data from different agencies were used, namely the BPS and Election Dataset. It is necessary to combine data to obtain valuable information. The BPS dataset has a hierarchy of Village, District, and Regency, while the Election Dataset has a ranking, namely TPS, Village, District, and Regency. So from the two datasets that can be integrated, only the Village, District, and Regency attributes.

c. Attribute correlation

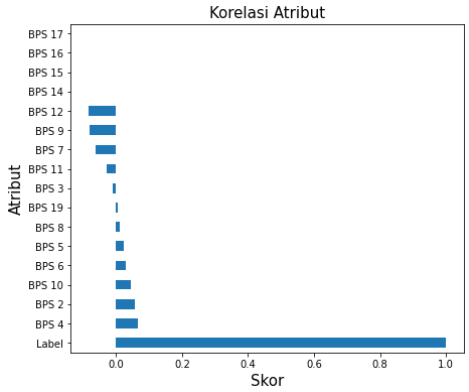

Figure 3. Check Attribute Correlation

Figure 3 shows that in the election participation dataset, it can be seen that the attributes that have a low influence are BPS 17, BPS 16, BPS 15, and BPS 14 attributes. And those that have a reasonably strong influence are BPS 2 and BPS 4.

d. Normalization

In this study, 3 data normalization models were carried out: model 1 Simple Feature Scaling, model 2 Min Max, and model 3 Z-score.

e. Random dataset

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

3.3 Modelling K-NN Method

At the preprocessing stage, it was explained that the purpose of this study was to predict the effect of election participation on the demographics of Malang City by applying the best normalization method so that a comparison or comparison of several data normalization models would be carried out. The algorithm used in this study is the KNN algorithm. To get the evaluation results, it is necessary to do modeling using Python.

3.4 Evaluation

This stage evaluates the model formed using the Confusion Matrix method to get accurate model information.

3.5 Application of K-NN Method

The implementation of K-NN uses the python programming language. The first step is to determine training and testing data distribution with a ratio of 70%:30%. After that, the application of KNN with the normalization model is carried out at the dataset normalization stage. Figure 3, Figure 4, and Figure 5 show the results of applying the KNN algorithm.

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

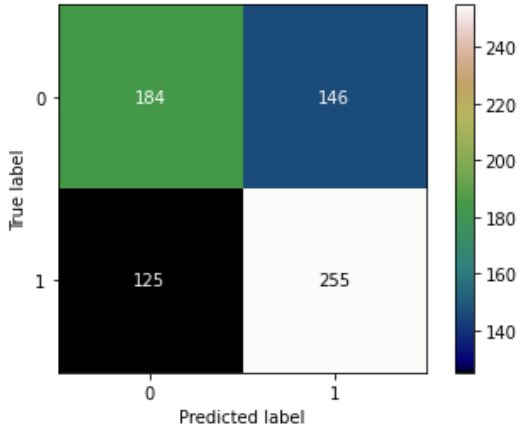

Figure 4. Implementation of K-NN Model 1

Figure 4 shows that the Low label was as low as 184, and the predicted high label was 255.

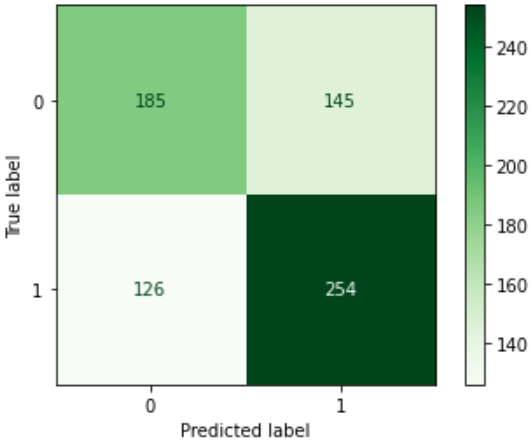

Figure 5. Implementation of K-NN Model 2

Figure 5 shows that 185 predicted Low labels as low and 254 predicted high labels.

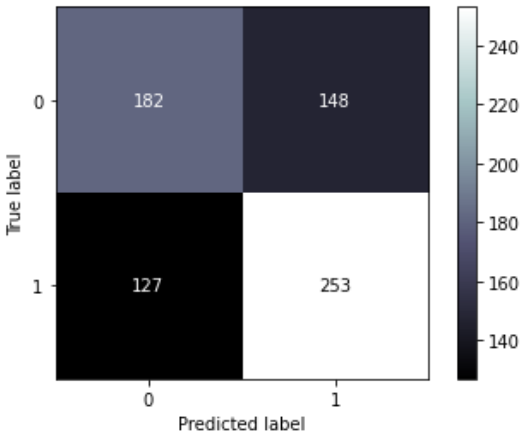

Figure 6. Implementation of K-NN Model 3

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

Figure 6 shows that the Low label predicted as low is 182, and the high label is predicted to be increased 253.

Table 1
Comparison of dataset model normalization accuracy

	Normalization Model		
K-NN	Model 1	Model 2	Model 3
Accuracy	61.83%	61.84%	61.27%
Precision	61.57%	**61.58%**	61%
Recall	61.43%	**61.45%**	60.87%
F-Score	61.44%	**61.46%**	60.87%

Table 1 shows that the normalization model that obtains the highest accuracy with a value of 61.83% is model 1 and model 2. The highest precision, with a value of 61.58%, is model 2. The highest recall is with a value of 61.45%, namely model 2. And the highest F1-Score is obtained by model 2 with a value of 61.46%.

4. Conclusion

Based on the results at the evaluation stage, it can be concluded that the Min-Max normalization model in model 2 is the best normalization model among the normalization models model 1 and model 3 in predicting the influence of election participation on the demographics of Malang City because it has the F1-Score with the highest value, so that indicates that model 2 has good precision and recall. From the data integration results, the value of the BPS dataset is less detailed because the dataset in the village scope position is more global. In this condition, further research is needed to integrate hierarchically parallel data to the concept of its attributes, which is the same as reviewing TPS.

References

[1] P. S. N. Wardhani, "Jurnal Pendidikan Ilmu-Ilmu Sosial Partisipasi Politik Pemilih Pemula dalam Pemilihan Umum," *JUPIIS J. Pendidik. Ilmu-Ilmu Sos.*, vol. 10, no. 1, pp. 57–62, 2018.

[2] L. P. Martha, "Hubungan Karakteristik Demografis Masyarakat Dengan Tingkat Partisipasi Politik (Studi Kasus Pilpres 2014 Di Kecamatan Cibinong Bogor)," *Repos. Univ. Pakuan*, 2014.

[3] D. Hakim, A., & Suherman, "Prediksi Kehadiran Masyarakat Dalam Pemilihan Umum Dengan Menggunakan Metode Naïve Bayes Classification," *J. Ilm. Komputasi*, vol. 3, no. 1, 2019.

[4] M. Simanjuntak, N. Nurfalinda, and M. R. Rathomi, "PENERAPAN METODE NAIVE BAYES UNTUK MEMPREDIKSI STATUS KEHADIRAN MASYARAKAT DALAM PEMILIHAN GUBERNUR.," *Student Online J. Umr. - Tek.*, vol. 3, no. 1, pp. 152–163, 2022.

[5] I. Nikmatun, I. A., & Waspada, "IMPLEMENTASI DATA MINING UNTUK KLASIFIKASI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR," *J. SIMETRIS*, vol. 10, no. 2, 2019.

[6] S. Ramadona, S., Diono, M., Susantok, M., & Ahdan, "Indoor location tracking pegawai berbasis Android menggunakan algoritma k-nearest neighbor," *J. Ilm. Telekomun.*, vol. 1, no. 1, pp. 51–58, 2021.

[7] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE MENGGUNAKAN ALGORITMA K-NN," *J. Comput. Eng. Syst. Sci.*, vol. 4, no. 1, pp. 78–82, 2019.

[8] Komisi Pemilihan Umum, "SERTIFIKAT REKAPITULASI HASIL PENGHITUNGAN PEROLEHAN SUARA PASANGAN CALON PRESIDEN DAN WAKIL PRESIDEN DARI SETIAP KECAMATAN DALAM WILAYAH KABUPATEN/KOTA PEMILIHAN UMUM TAHUN 2019," 2019.

[9] Badan Pusat Statistik Kota Malang, *Kecamatan Blimbing Dalam Angka 2020*. Kota Malang, 2020.

[10] A. Yuniarti, A. Yasin, and Y. A. Nugroho, "Efektifitas Algoritma Data Mining dalam Menentukan Pendonor Darah Potensial," *Syntax J. Inform.*, vol. 11, no. 1, pp. 12–22, 2022.

[11] Algoritma Data Science Education Center, "DATA CLEANING," *Algoritma Data Science Education Center*, 2022.

[12] D. Kurniasari and A. Widya, "Integrasi Data: 3 Cara Penggabungan Data Untuk Hasil Analisis Yang

Optimal," *DQLab*, 2021.
[13] U. Afifah, "3 Metode Normalisasi Data (Feature Scaling) di Python," *IlmudataPy*, 2022.
[14] M. L. Radhitya and G. I. Sudipa, "PENDEKATAN Z-SCORE DAN FUZZY DALAM PENGUJIAN AKURASI PERAMALAN CURAH HUJAN," *SINTECH (Science Inf. Technol. J.*, vol. 3, no. 2, pp. 149–156, 2020.
[15] R. Hayati, "Metode Penelitian Ilmiah Pengertian Analisis Korelasi, Jenis, dan Contohnya," 2022.
[16] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika, 2107.
[17] Fatmawati, "PERBANDINGAN ALGORITMA KLASIFIKASI DATA MINING MODEL C4.5 DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT DIABETES," *J. Techno Nusa Mandiri*, vol. 13, no. 1, 2016.
[18] Y. Harun, R., Chandra Pelangi, K., & Lasena, "PENERAPAN DATA MINING UNTUK MENENTUKAN POTENSI HUJAN HARIAN DENGAN MENGGUNAKAN ALGORITMA K NEAREST NEIGHBOR (KNN)," *J. Manaj. Inform. Sist. Informasi)*, vol. 3, no. 1, 2020.

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