INFLUENCE OF DEMOGRAPHICS FOR PREDICTION OF ELECTION PARTICIPATION USING LOGISTIC REGRESSION ALGORITHM

by Metatia Intan Mauliana

http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.3, Agustus 2022 ISSN : 2302-9706

INFLUENCE OF DEMOGRAPHICS FOR PREDICTION OF ELECTION PARTICIPATION USING LOGISTIC REGRESSION ALGORITHM

Rahmi Aulia Barlian[1*], Arif Senja Fitrani[2], Metatia Intan Mauliana[3]
[1*,2,3]Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo
barlianx88@gmail.com, asfjim@umsida.ac.id, metatialiana@umsida.ac.id

Abstract

Article Info
Received, 07/07/2022
Revised, 18/08/2022
Accepted, 20/08/2022

Indonesia is a democratic country for General Elections (Election) which are carried out directly, freely, confidentially, honestly, and fairly. Several stages of the election, among others, begin with compiling a permanent voter list (DPT), determination of polling stations (TPS), and recapitulating election results. Various factors, including the demographic factor, can affect citizen participation in the general election. Demographic data covers Energy, Geographic, Education, Health, Population, Economy, Communication, and Transportation factors. This study tries to combine election data with demographic data taken from the official website of the Central Statistics Agency (BPS) of Mojokerto Regency and data on the results of the 2019 Election calculations taken from the official website of the General Election Commission (KPU) of Mojokerto Regency. Preprocessing steps are data cleaning, data integration, and correlation attributes for a more optimal presentation of the dataset and the distribution of four split datasets (training data and testing data) to find the best results. Implementation of classification method with Logistic Regression (LR) algorithm to predict community participation at the TPS level. From the test results of four split datasets, the highest predictive value was 64.80% in composition 3 with a ratio of 80:20, where 127 data were labeled low, and 291 data were labeled high.

Keywords: Classification, Prediction, Elections, Demographics, Logistic Regression

1. Introduction

Elections are a means to implement people's sovereignty which is held directly, publicly, freely, confidentially, honestly, and relatively following Chapter 2 of Constitution of the Law on General Elections. The participation or voting rights of citizens in the administration of elections can affect the state's future in formulating new policies or reforming old ones.

The election stages begin with updating voter data and compiling a voter list following Article 4 of Law Chapter 2 of Constitution of the Law no. 10 of 2008. The preparation of the DPT becomes an important factor in the election's success. Many variants can affect the number of DPT changes so that there is a reduction or addition at any time. Among them are residents aged 17 years and over who are called novice voters, died, residents came, and residents left. In addition to changes in the number of DPT, geographical location also affects the success of the election [1].

Previous research [2], which predicts the status of public attendance in the gubernatorial election using the nave Bayes algorithm by applying seven variables to 76 data, obtained an accuracy value of 78.95%. In comparing the C4.5 algorithm and the neural network to predict the results of the DKI Jakarta legislative elections, the accuracy value is 97.84% for the C4.5 algorithm. In comparison, the neural network algorithm is 98.50% [3]. The Dhamasraya Regency community voter participation model in the 2014 election using the Bayesian logistic regression method obtained two independent variables that affect voter participation, including access to election information and candidate socialization [4]. Another study using a logistic regression algorithm was carried out by [5] for political classification in general elections with a model suitability test, which obtained a significance = 1, that is, accept H0, which means the model used is successful. The stratified binary logistic regression model was also used to predict women's economic participation in East Java Province. The urban strata binary logistic regression test results had three influential variables: marital status, family status, and education level. In contrast, the rural strata only had marital status and education level[6].

Mojokerto Regency is one of the regencies in East Java Province, where the use/utilization of residential areas is only 132,440 km2 out of 969,360 km2 of the total area of Mojokerto Regency. The southern part of Mojokerto Regency is a mountainous area with an average altitude of less than 500 meters above sea level. Only Pacet and Trawas Districts are the most considerable areas with an altitude of more than 700 meters above sea level. Life and the environment in mountainous areas encourage residents to prioritize things of economic value rather than thinking and being politically oriented. This is because conditions in mountainous regions are relatively complex in terms of transportation, communication, and sources of livelihood. It is proven from the results of the 2019 general election recapitulation that the Mojokerto Regency election participants only received 728,733 votes from the total DPT 839,517 votes. From the previous description, this study focuses on predicting the effect of demographic data on general election participation in Mojokerto Regency using the LR algorithm. The LR algorithm is a regression analysis to determine the relationship between categorical response variables, nominal and ordinal, and whether explanatory variables can be categorical or continuous[7].

2. METHOD

2.1 Prediction

Prediction is a way of estimating something very likely to happen in the future based on data that has and is happening. The meaning of prediction is similar to forecast or forecast. Predictions can be sourced from the scientific method or purely subjective. For example, weather predictions are obtained from the latest data and information based on satellite observations. Predictions such as sports events are usually obtained from subjective views or personal observations [8].

2.2 Classification

Classification is the placement of objects or data into one of several predefined categories. Classification involves learning the target function f, which associates each attribute set x with one of the predefined class labels y [9].

2.3 Logistic Regression Algorithm (LR)

It is an approach to producing a predictive model where the dependent variable is dichotomous or has two possible outcomes, namely Y=1 (success) and Y=0 (failure), with probabilities p and q = 1 – p, respectively[10]. The LR algorithm equation is formulated as Equation (1) and Equation (2) [11]:

$$Ln\left(\frac{P}{1}-P\right) = a + b_1x_1 + \cdots + \cdots + b_{19}x_{19} + e \quad (1)$$

$$logit(P) = a + b_1x_1 + \cdots + \cdots + b_{19}x_{19} + e \quad (2)$$

2.4 Research Stage

The model used in this study is a quantitative approach. The 2019 Mojokerto Regency Election Calculation Result Data and Mojokerto Regency Demographic Data are used. The data processing technique in this study uses a data mining process with prediction and classification methods. The research stage can be seen in Figure 1.

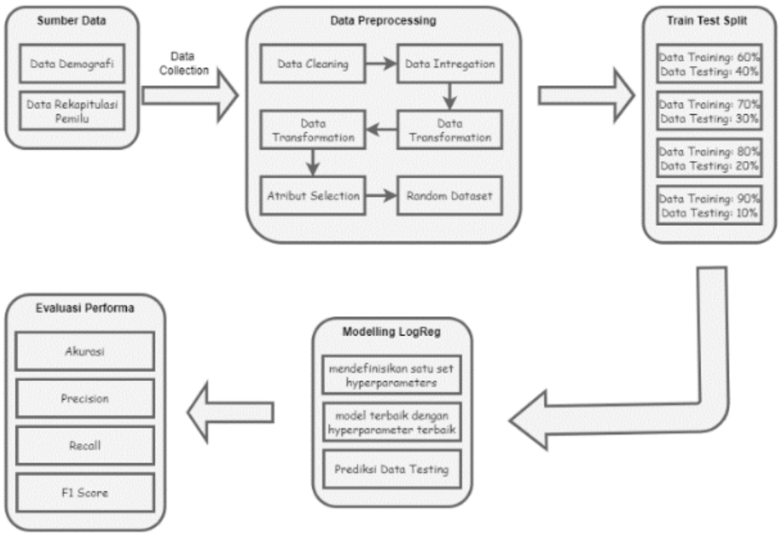

Figure 1. The flow of Stages of Predicting the Effect of Election Participation on Demographics

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

3. RESULTS AND DISCUSSION

3.1 Data Collect

The first step is data collection. The data used in this study are demographic data taken from the Mojokerto Regency BPS official website and 2019 election calculation results taken from the Mojokerto Regency KPU official website.The dataset has 3225 records, consisting of 20 attributes with 1 predictive label shown in Table 1.

Table 1. Dataset Attributes

Index	*Attribute Description*
B1	B1 population
B2	B2 natural disaster early warning system
B3	B3 special tsunami early warning system
B4	B4 safety equip
B5	B5 disaster evacuation signs and routes
B6	B6 construction, maintenance or normalization of: rivers, canals, embankments, ditches, drainages, reservoirs, beaches, etc.
B7	B7 mini market/supermarket
B8	B8 convenience store/grocery shop

B9	B9 food stall/shop
B10	B10 cooperative
B11	B11 number of cell phone towers (BTS)
B12	B12 number of cellular telephone communication service operators reaching out to villages or ward
B13	B13 Cellular phone signal conditions in most rural areas
B14	B14 type of transportation
B15	B15 the presence of public transportation
B16	B16 road surface type
B17	B17 can be passed by motorized vehicles with 4 or more wheels
B18	B18 post office/postal assistant/post house
B19	B19 private shipping company/agent
Label	Prediction class labels (low and high)

3.2. Preprocessing

The preprocessing stage is carried out before modeling the algorithm on the dataset using the python programming language. Pre-processing is necessary to reduce the risk that data may be incomplete or contain errors. The following are the stages of Pre-processing:

a) Data Cleaning

Data Cleaning is the process of filling in missing values, smoothing noise data, and overcoming data inconsistencies that aim to produce a good dataset before data mining modeling is carried out[12].

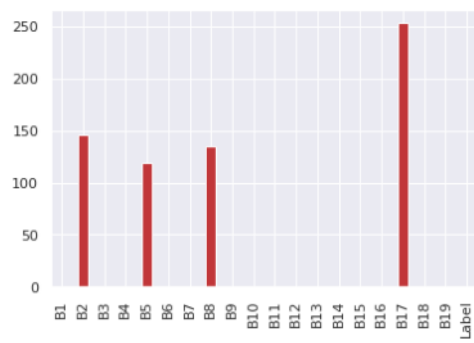

Figure 2. Missing Value in Dataset

It can be seen in Figure 2 that the dataset has several missing values below 25% for attributes B2, B5, B8, and B17. Based on this situation, it is necessary to handle it, namely to improve the value by filling in the missing values using the median for each attribute that has a missing value.

b) Data Integration

The next stage is to combine data that will be used for data analysis. The dataset comes from a combination of Mojokerto Regency demographic data and Mojokerto Regency election recapitulation data. This study uses a concept hierarchy scheme, namely BPS data at the village level while election data at the TPS level. This causes BPS data to make election data a prediction class, namely attribute labels.

c) Data Transformation

The third stage is mapping the entire set of attributes to a new replacement value, namely dataset normalization. This study uses Z-Score normalization.

d) Attribute Correlation

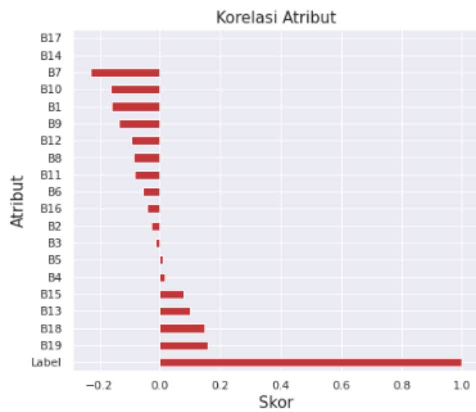

Figure 3. Check Attribute Correlation

It can be seen in Figure 3 that the correlated attributes are pretty low, namely attributes B17, and B14. Therefore, these two attributes must be removed 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 Logistic Regression Algorithm (LR)

Data mining is carried out at the modeling stage using Jupyter software with the python programming language. The algorithm used is Logistic Regression with a comparison of Training data, and Testing data can be seen in Table 2.

Table 2. Comparison of Train and Test Split

Index	*Data Sharing*
Composition 1	Training Data: 60%
	Testing Data: 40 %
Composition 2	Training Data: 70%
	Testing Data: 30 %
Composition 3	Training Data: 8%
	Testing Data: 20 %
Composition 4	Training Data: 90%
	Testing Data: 10 %

3.4. Data Test Evaluation

After the testing process has been carried out on the overall composition of the data distribution, the results are compared to obtain the composition of the data distribution with the best value.

The test is carried out based on the composition of the data division by using the confusion matrix as the test method.

Table 3. Composition Test Results 1

http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.3, Agustus 2022 ISSN : 2302-9706

Confusion Matrix	*Prediction Label*	
	Low	High
Low	264	330
High	145	551

Based on Table 3, it can be seen that the implementation of the algorithm with composition one successfully predicts 264 data with low labels and 551 data with high labels correctly.

Table 4. Composition Test Results 2

Confusion Matrix	*Prediction Label*	
	Low	High
Low	196	246
High	104	422

Based on Table 4, it can be seen that the implementation of the algorithm with composition 2 successfully predicts 196 low-label data and 422 high-labeled data correctly.

Table 5. Composition Test Results 3

Confusion Matrix	*Prediction Label*	
	Low	High
Low	127	168
High	59	291

Based on Table 5, it can be seen that the implementation of the algorithm with composition 3 successfully predicts 127 low-labeled data and 291 high-labeled data correctly.

Table 6. Composition Test Results 4

Confusion Matrix	*Prediction Label*	
	Low	High
Low	58	80
High	38	147

Based on Table 6, it can be seen that the implementation of the algorithm with composition 4 successfully predicted 58 data with low labels and 147 data with high labels correctly.

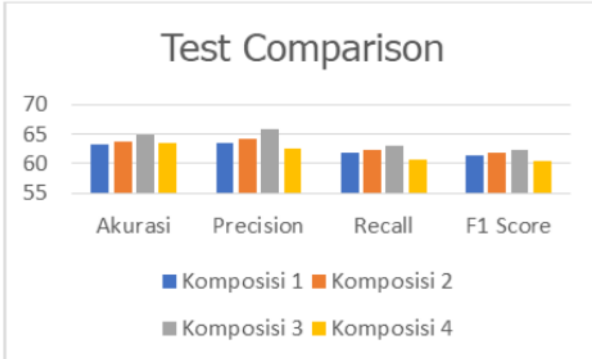

Figure 4. Data Sharing Test Results

Based on Figure 4, it can be seen that the results of the composition 1 test resulted in an accuracy of 63.18%, Precision 63.54%, Recall 61.80%, F1 Score 61.27%. composition 2 produces an accuracy of 63.84%, Precision 64.25%, Recall 62.29%, F1 Score 61.75%. composition 3 produces an accuracy of 64.80%, Precision 65.83%, Recall 63.10%, F1 Score 62.38%. composition 4 produces an accuracy of 63.47%, Precision 63.47%, Recall 60.74%, F1 Score 60.74%.

4. Conclusion

Based on the results of the modeling evaluation that has been carried out on the dataset on predictions of election participation in Mojokerto Regency using four different compositions of Training data distribution and Testing data using the LR algorithm, the highest level of accuracy is obtained at 64.80% in composition 3 with a ratio of 80:20. Based on testing data from the composition with the highest evaluation value, composition 3 has a total testing data of 645 records, and it is found that attributes B7 and B9, namely minimarkets/supermarkets and food stalls/shops, have a high enough influence in predicting election participation in Mojokerto Regency. These results are based on attributes B7 and B9, which have 350 records with a data class labeled high.

Reference

[1] N. Widhiastini, N. Subawa, N. Sedana, and N. P. Permatasari, "Analisis Faktor-Faktor Yang Mempengaruhi Partisipasi Masyarakat Dalam PILKADA Bali," *Publik J. Ilmu Adm.*, vol. 8, p. 1, Oct. 2019, doi: 10.31314/pjia.8.1.1-11.2019.

[2] M. Simanjuntak, N. Nurfalinda, and M. R. Rathomi, "PENERAPAN METODE NAIVE BAYES UNTUK MEMPREDIKSI STATUS KEHADIRAN MASYARAKAT DALAM PEMILIHAN GUBERNUR," *Stud. Online J. SOJ UMRAH - Tek.*, vol. 3, no. 1, Art. no. 1, Mar. 2022.

[3] M. Badrul, "PERBANDINGAN ALGORITMA C4.5 DAN NEURAL NETWORK UNTUK MEMPREDIKSI HASIL PEMILU LEGISLATIF DKI JAKARTA," *J. Pilar Nusa Mandiri*, vol. 10, no. 2, Art. no. 2, Sep. 2014, doi: 10.33480/pilar.v10i2.470.

[4] S. Wulandari, F. Yanuar, and H. Yozza, "MODEL PARTISIPASI PEMILIH MASYARAKAT KABUPATEN DHAMASRAYA PADA PEMILU 2014 DENGAN MENGGUNAKAN METODE REGRESI LOGISTIK BAYESIAN," *J. Mat. UNAND*, vol. 6, no. 1, Art. no. 1, 2017, doi: 10.25077/jmu.6.1.128-133.2017.

[5] V. M. Santi, "Pengembangan Model Regresi Logistik Multinomial untuk Klasifikasi Politik pada Pemilihan Umum," *J. Stat. Dan Apl.*, vol. 2, no. 1, pp. 37–43, Sep. 2018, doi: 10.21009/JSA.02105.

[6] M. K. Kotimah and S. P. Wulandari, "Model Regresi Logistik Biner Stratifikasi Pada Partisipasi Ekonomi Perempuan Di Provinsi Jawa Timur | Kotimah | Jurnal Sains dan Seni ITS", Accessed: Aug. 08, 2022. [Online]. Available: http://ejurnal.its.ac.id/index.php/sains_seni/article/view/6096

[7] D. Yuniarti, "Comparison of Classification Methods Between Logistic Regression and Artificial Neural Network (Case Study: Selection of Language and Social Studies Depertement at SMAN 2 Samarinda academic year 2011/2012)," vol. 4, p. 8, 2013.

[8] H. Utari, M. Mesran, and N. Silalahi, "PERANCANGAN APLIKASI PERAMALAN PERMINTAAN KEBUTUHAN TENAGA KERJA PADA PERUSAHAAN OUTSOURCHING MENGGUNAKAN ALGORITMA SIMPLE MOVING AVERAGE," *J. TIMES*, vol. 5, no. 2, Art. no. 2, Dec. 2016.

[9] Wahyudin, "Metode Iterative Dichotomizer 3 (ID3) Untuk Penyeleksian Penerimaan Mahasiswa Baru," vol. 2, no. 2, p. 11, 2009.

[10] A. C. Delima, F. Yanuar, and H. Yozza, "PENERAPAN METODE REGRESI LOGISTIK ORDINAL BAYESIAN UNTUK MENENTUKAN TINGKAT PARTISIPASI POLITIK MASYARAKAT KOTA PADANG," *J. Mat. UNAND*, vol. 8, no. 3, p. 1, Dec. 2019, doi: 10.25077/jmu.8.3.1-8.2019.

[11] I. Nahib, "PREDIKSI SPASIAL DINAMIKA AREAL TERBANGUN KOTA SEMARANG DENGAN MENGGUNAKAN MODEL REGRESI LOGISTIK," *Maj. Ilm. GLOBE*, vol. 18, no. 2, p. 95, Dec. 2016, doi: 10.24895/MIG.2016.18-2.421.

[12] H. Bhalekar, S. Kumbhar, H. Mewada, P. Pokharkar, S. Patil, and M. R. Gound, "Pra pemrosesan data using ID3 classifier," vol. 1, no. 3, p. 6, 2015.

INFLUENCE OF DEMOGRAPHICS FOR PREDICTION OF ELECTION PARTICIPATION USING LOGISTIC REGRESSION ALGORITHM

ORIGINALITY REPORT

17%	8%	13%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	revista-agroproductividad.org Internet Source	3%
2	Ilham Rizky Widianto, Wowon Priatna, Hendarman Lubis. "Penerapan Algoritma FIFO Untuk System Pemesanan Kaos Sablon", Jurnal Kajian Ilmiah, 2023 Publication	2%
3	Tiara, , Mira Kania Sabariah, and Veronikha Effendy. "Sentiment analysis on Twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program", 2015 3rd International Conference on Information and Communication Technology (ICoICT), 2015. Publication	2%
4	V M Santi, A Kurnia, K Sadik. "Modelling of the number of malarias suffers in Indonesia using	1%

Bayesian generalized linear models", Journal of Physics: Conference Series, 2019
Publication

5	Arif Senja Fitrani, Nikko Enggaliano Pratama, Agus Budi Raharjo, Yudhi Purwananto, Diana Purwitasari. "A Comparative Study on Machine Learning based Prediction Models for Public Participation Rate in an Election Voting", 2022 3rd International Conference on Electrical Engineering and Informatics (ICon EEI), 2022 Publication	1 %
6	Budi Sunarko, Uswatun Hasanah, Ulfah Mediaty Arief, Feddy Setio Pribadi et al. "Prediction of Student Satisfaction with Academic Services Using Naive Bayes Classifier", 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2022 Publication	1 %
7	proceeding.researchsynergypress.com Internet Source	1 %
8	Stavros Karathanasis. "Using data-mining techniques for PM10 forecasting in the metropolitan area of Thessaloniki, Greece", 2007 International Joint Conference on Neural Networks, 08/2007	1 %

Publication

9	Vianti Widyasari, Arief Senja Fitrani. "Application of Data Mining with Classification Methods for Promotion of New Student Admissions at Muhammadiyah University of Sidoarjo Using Web-Based Naïve Bayes Algorithm", Procedia of Engineering and Life Science, 2021 Publication	1 %
10	hdl.handle.net Internet Source	1 %
11	ejurnal.ung.ac.id Internet Source	<1 %
12	ijels.com Internet Source	<1 %
13	infor.seaninstitute.org Internet Source	<1 %
14	Nur Ilmayasinta. "FORECASTING ARRIVAL OF FOREIGN TOURISTS USING SEASONAL ARIMA BOX-JENKINS", BAREKENG: Jurnal Ilmu Matematika dan Terapan, 2021 Publication	<1 %
15	bestjournal.untad.ac.id Internet Source	<1 %
16	jurnal.unprimdn.ac.id Internet Source	<1 %

17	ojs.ukb.ac.id Internet Source	<1 %
18	Padrul Jana, Dedi Rosadi, Epha Diana Supandi. "COMPARISON OF ROBUST ESTIMATION ON MULTIPLE REGRESSION MODEL", BAREKENG: Jurnal Ilmu Matematika dan Terapan, 2023 Publication	<1 %
19	Yufri Isnaini Rochmat Maulana, Tessy Badriyah, Iwan Syarif. "Influence of Logistic Regression Models For Prediction and Analysis of Diabetes Risk Factors", EMITTER International Journal of Engineering Technology, 2018 Publication	<1 %
20	en.wikipedia.org Internet Source	<1 %
21	soj.umrah.ac.id Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		