unplug Attribute analysis with classification algorithm on election participation

by Mochamad Alfan Rosid

Submission date: 09-Jan-2024 12:19PM (UTC+0700)
Submission ID: 2268247843
File name: Fitrani_2020_IOP_Conf._Ser.__Mater._Sci._Eng._821_012034.pdf (573.7K)
Word count: 4143
Character count: 20913

PAPER · OPEN ACCESS

Attribute analysis with classification algorithm on election participation

To cite this article: A S Fitrani *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **821** 012034

View the article online for updates and enhancements.

You may also like

- The analysis of the dynamics of the electorate system by using q-distribution-a case study
Dode Prenga, Klaudio Peqini and Rudina Osmani
- Backlash to fossil fuel phase-outs: the case of coal mining in US presidential elections
Florian Egli, Nicolas Schmid and Tobias S Schmidt
- Artificial Intelligence in Election Party of Broker Clientelism Joxzin (*Jogjakarta Islamic Never Die*)
Yeyen Subandi, Zuly Qodir, Hasse Jubba et al.

This content was downloaded from IP address 118.99.67.170 on 27/12/2022 at 02:37

IOP Conf. Series: Materials Science and Engineering **821** (2020) 012034 doi:10.1088/1757-899X/821/1/012034

Attribute analysis with classification algorithm on election participation

A S Fitrani*, M A Rosid, F Muharram and F L Kodriyah

Department of Informatics Engineering, Faculty of Engineering, Universitas Muhammadiyah Sidoarjo, Indonesia

*Email: asfjim@umsida.ac.id

Abstract. Stages of General Elections of the President, DPD, DPR, Provincial DPRD and Regency / City DPRD in Indonesia are determined by institutions namely the General Election Commission (KPU), where there is a measure of success in holding direct, general, and free. Another component of the implementation of elections is that there are contestants and voters. In the voter factor, this is also a measure for success in the overall process of implementation, namely success if high community participation in the administration of elections. However, vice versa, if community participation is low, one of them is the level of public confidence in the organizers (government) decreases. Data mining classification analysis and modification of attributes in prediction classes "Hadir" and "Tidak Hadir" on the final voter list (DPT). The number of datasets is 4249 instances, and the number of attributes is 11. The percentage results are 89.3417% for the Naive Bayes algorithm for prediction classes in the Presidential Election, DPD, DPR, Provincial DPRD and Regency/ City DPRD in 2019. Further analysis is done on eliminating some attributes to obtain information, whether it has a significant effect on the results of predictions. And in this analysis, for the ten attributes with the removal of "statusKK" (highest rank gain ratio) for the prediction class, the results are the worst. = After nine attributes, the removal of the "rt" and "tps" attribute (second and third rank) for the prediction class is the best result. There is the highest percentage difference for the prediction class on the classification algorithm for modified and or unmodified status attributes, the results of the percentages and classification algorithms are different.

1. Introduction

General Election, abbreviated as PEMILU, is a people's party in determining the representation of both legislative, senator (DPD) and executive (president, governor, and district/city). The implementation of the ELECTION of the President, DPD, DPR, Provincial DPRD, and Regency/ City DPRD is carried out in 2019. ELECTION is an important political event to determine leaders in a democratic country. The Permanent Voters List (DPT), which is then determined to be a DP4, is the right of citizens to be protected by law to choose their representation in the executive, senator and excesses. Residents who have been recorded to channel their aspirations to the General Election are conducted at the polling station (TPS). Requirements for citizens to have the right to vote are regulated by law. The voter segment group is

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

divided into three namely beginner voters, millennial voters and post menial voters. Where in the previous research the focus was on the scope of results, namely, predictions of legislative elections (C4.5 algorithm to predict the results of legislative selection of DKI Jakarta DPRD, Techno Nusa Mandiri Vol. IX No.1, March 2013). Here, we try to take the other side, which is related to voter participation in the democratic party process that takes place.

The involvement of the community for participation in the democratic process is determined by the presence of the community in the booth (TPS), where residents come to be able to vote. The absence of the public in the ELECTION process at polling stations was termed golput (white group) or indeed was not present at the time of the democratic party that was held. Many factors determine the absence of the public at the democratic party, among others, based on the age level that is identical to the millennial voters (ages 17 to 32), beginner voters (aged 17 to 20) who have no experience in a democratic party being held, apathy will representations that have been legally registered and other factors that generally allow the age level to not be able to participate in the democratic party this time.

The study took a sample of data in Wonokasian Village, Wonoayu District, Sidoarjo Regency, East Java Province at the 2019 ELECTION simultaneously in Indonesia. Classification method (Naïve Bayes algorithm, C4.5, and kNN), attribute analysis on attendance data and voter data is the focus of this research so that it can be known whether the determination of the effect of the prediction class is produced. Because there may be many attributes, but they have no influence on the prediction class (results) at all. Many attributes are irrelevant and will place unnecessary computational overhead on the data mining algorithm [1, 2]. At worst, it can cause the algorithm to be a poor outcome.

Classification methods in several algorithms are used in the implementation of research for quality attribute analysis to be tested on several different attributes [3, 4]. The hope is, can it provide a significant percentage change with the same number of instances.

2. Literature review

2.1. Classification

Classification method that is studying a set of data that can be produced by a new classification of data [5]. The classification process in data mining techniques is a set of data that can produce a classification model (target function). So, it requires a dataset on the set for the classification process. The dataset used is attributes and features using training data and testing data [6]. Classification techniques can be grouped into two categories. Namely, the classification technique globally calculates all training data and classification locally taking into account some training data [7].

2.2. C4.5 algorithm

At the learning stage of the data, the C4.5 algorithm constructs decision trees from training data [8, 9], in the form of cases or records (tuples) in the database. The three working principles of the C4.5 algorithm at the learning stage of the data are:

1. Making a decision tree.
2. The decision tree pruning and evaluation (optional).
3. Making rules from decision trees (optional).

The formula for the C4.5 algorithm is:

$$Gain(S,A) = Entropy(S)\sum_{i=1}^{n}\frac{Si}{S} * Entropy(Si) \quad (1)$$

where S is set of cases, A is attribute, n is number of partition attributes A, Si is number of cases on the i-partition, S is number of cases in S.

$$Entropy(S) \sum_{i=1}^{n} \frac{Si}{S} * pi * log2\, pi \quad (2)$$

where S is set of cases, A is feature, n is number of S partitions, pi is proportion of Si to S.

2.3. Naïve Bayes algorithm

Naïve Bayes algorithm is a classification method by calculating probabilities in determining the number of classes and values of a dataset [10]. The advantage of using the Naïve Bayes algorithm is that the small amount of training data can determine the required parameter estimates. Assume a simplification of the Naïve Bayes algorithm with attribute values that are mutually independent if given their output values. Naïve Bayes has a very strong level of accuracy and speed when applied to a database with large data [11]. Based on the Bayes theorem, which can classify the same method as the Decision Tree algorithm and Neural Network [12, 13]. In the Bayes theorem equation, the conditional probability is expressed as:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (3)$$

Where,

X : Data with unknown classes.
H : Data hypothesis with specific classes.
P (H) : Probability of H. hypothesis
P (X) : Probability X.
P (H | X) : Posterior probability H with condition X.
P (X | H) : Posterior probability X with the condition H.

2.4. k nearest neighbor classification

The K-Nearest Neighbor (NN) is the simplest method of machine learning [13, 14]. It is a type of instance-based learning in which object is classified based on the closest training example in the feature space. It implicitly computes the decision boundary; however, it is also possible to compute the decision explicitly. So the computational complexity of K NN is the function of the boundary complexity. The k-NN algorithm is sensitive to the local structure of the data set. The special case when k = 1 is called the nearest neighbor algorithm. The best choice of k depends upon the data set; larger values of k reduce the effect of noise on the classification but make boundaries between classes less distinct. The various heuristic techniques are used to select the optimal value of K. KNN has some strong, consistent results. As the infinity approaches data, the algorithm is guaranteed to yield an error rate less than the Bayes error rate.

3. Methods

Diagram for determining how to select characteristics by determining attendance class (present and absent) in the election process and determine the number of attributes for the best value that results in a comparison of the three conditions for the number of attributes (Figure 1):

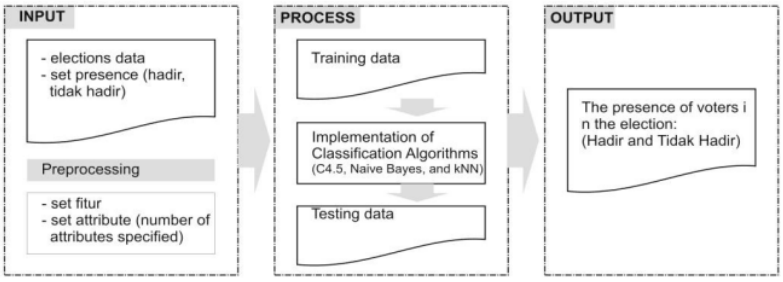

Figure 1. Classification diagram.

For the 'train and test' method, the available data is split into two parts called a training set and a test set (Figure 2). First, the training set is used to construct a classifier (decision tree, neural net, etc.). The classifier is then used to predict the classification for the instances in the test set. If the test set contains N instances of which C are correctly classified the predictive accuracy of the classifier for the test set is p = C/N. This can be used as an estimate of its performance on any unseen dataset. In cases where the dataset is only a single file, we need to divide it into a training set and a test set before using Method 1. This may be done in many ways, but a random division into two parts in proportions such as 1:1, 2:1, 70:30 or 60:40 would be customary.

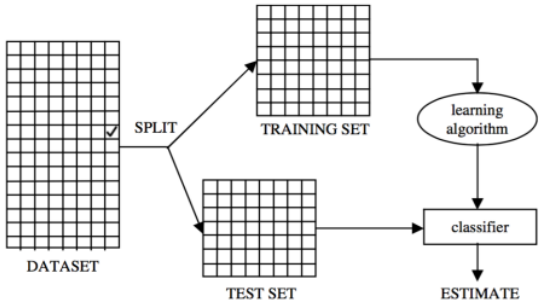

Figure 2. Train and test.

4. Data processing

Data warehouse originating from 2 data sources, namely C7 data (attendance list) and DP4 data (DPT data). From 2 references, a merger process is conducted (C7 data and DP4 data) based on the voter ID, to determine the presence of voters at the polling station. And changes in some attributes to produce new information, to produce better predictive results. Attributes from the results of the changes are the class, origin, category, and status of the KK. For the addition of a new attribute the location_TPS is raised based on the address (rt) of the voter and the name of the tips, so that it can know whether the location of the polling station is in the address "rt" or outside, this determines the distance from the voter to the polling station location.

The data warehouse has been set as many as 11 attributes, and the dataset is 4249 instances, then divided into two parts for further processing. Data train in the capacity of 70% (2973 instance) and test

data of 30 (1276 instances). Selecting data that is dismissed information to produce the best value in the testing process.

4.1. Data collection

The dataset is obtained from the final voter list (DPT) in Wonokasian village, Wonoayu sub-district, Sidoarjo, East Java. DPT is used in the General Elections of the President, DPD, DPR, Provincial DPRD and Regency/ City DPRD in 2019. The set datasets are 4249 instances, and 11 attributes are taken from form C7 (voter attendance list). From these datasets scattered in 15 polling stations (TPS), where polling stations are the polling implementation unit Table 1.

Table 1. Dataset description.

Dataset	No. of attribute	No. of instance
Form C7 at 15 polling stations, in the election of the President, DPD, DPR, Provincial DPRD and Regency / City DPRD in Wonokasian village, Wonoayu District, Sidoarjo Regency	11	4249

Form C7 data is sourced from the List of Electoral Potential Population (DP4), where the data summarizes the three conditions, namely probability data, geographic data, and attendance data. Personal and geographical data are found in DP4 data and attendance data in form C7 data. The description of processing 2 data sources (form C7 and DP4) is Table 2:

Table 2. 11 Attribute and future dataset.

No	Attribute	Variable	Information
1	jmlKK	{dua,>dua,singgle}	Number of family members
2	asal	{SIDOARJO,LUAR,'LUAR PROV'}	Origin of voters
3	kategori	{Milenial,PascaMilenial,Pemula}	Type of society
4	kawin	{S,B,P}	Marriage Status (S = Already, B = Not yet and P = Ever)
5	jk	{P,L}	Gender (P = Female and L = Male)
6	alamat	{WONOKASIAN,LENGKONG,NGGODEK,NGEMPLAK,KRAMAT,NDOKOH,KRESAN,KLITIH}	Address Village and or hamlet in one village
7	rt	{rt01,rt02,rt03,rt04,rt05,rt06,rt07,rt08,rt09,rt10,rt11,rt12,rt13,rt14,rt16,rt15,rt17,rt19,rt18,rt20,rt21}	Address RT (Neighborhood) in one village
8	tps	{TPS-1,TPS-2,TPS-3,TPS-4,TPS-5,TPS-6,TPS-7,TPS-8,TPS-9,TPS-10,TPS-11,TPS-12,TPS-13,TPS-14,TPS-15,TPS-16}	Name of TPS (polling station) in one village
9	lokasi_TPS	{dalam,luar}	Distance between polling station locations and voter address
10	statusKK	{TSH,HS}	Attendance at polling stations in one family
11	hadir	{TH,H}	Individual attendance at TPS (class prediction)

4.2. Ratting attribute

To see the best weight of 11 attributes, it can be done by calculating the gain ratio, so that the best and worst attribute sequence can be known, namely (Table 3):

Table 3. Ranked attributes

Rangking	Gain Ratio	No	Attribute
1	0.390750	10	statusKK
2	0.054964	8	tps
3	0.036399	7	rt
4	0.018774	6	alamat
5	0.013883	9	lokasiTPS
6	0.004621	1	jmlKK
7	0.004580	4	kawin
8	0.001704	2	asal
9	0.001436	3	kategori
10	0.000126	5	jk

4.3. Attribute modify

Modification of attributes that are by the conditions in Table 4, is done to reduce the initial conditions which have several 11 attributes, then from the number 11 on some attributes not included in the testing process. This will be seen here, whether there is a change in the results of the steps for reducing the attribute that has been set. Starting with a crew condition of 11 attributes. Then ten attributes that try to delete one attribute (statusKK). In Table 3 "statusKK" ranks first regarding the quality attribute of the calculation of the gain ratio with a value of 0.390750 and for the condition of 9 attributes, delete two attributes, namely "rt and tps". Information from these two attributes, namely for "rt" is the location of the address of the voter domiciled and "tps" is the name of the place where the election is held. The quality of the two attributes ranks second and third after "statusKK".

Table 4. Three attribute conditions

No	11 attribute
1	jmlKK
2	asal
3	kategori
4	kawin
5	jk
6	alamat
7	rt
8	tps
9	lokasi_TPS
10	statusKK
11	hadir

No	10 attribute
1	jmlKK
2	asal
3	kategori
4	kawin
5	jk
6	alamat
7	rt
8	tps
9	lokasi_TPS
10	hadir

No	9 attribute
1	jmlKK
2	asal
3	kategori
4	kawin
5	jk
6	alamat
7	rt
8	statusKK
9	hadir

4.4. Condition dataset with deferent attribute

The dataset with condition class {TH, H} with a total of 4249 instances, training data 2973 instances and testing 1276 instances, displays datasets according to the discussion in Table 4, where changes directly

occur in each dataset presentation according to the attribute group used. Removal between "statusKK" for ten attribute categories and "rt and tps" for category nine attribute Tables 5, 6 and 7.

Table 5. Dataset, 11 attribute

No.	Data election
1	dua,SIDOARJO,Milenial,S,P,WONOKASIAN,rt01,TPS-1,dalam,TSH,TH
2	dua,SIDOARJO,Milenial,S,L,WONOKASIAN,rt01,TPS-1,dalam,TSH,H
3	>dua,LUAR,Milenial,B,L,WONOKASIAN,rt01,TPS-1,dalam,TSH,TH
4	singgle,SIDOARJO,Milenial,S,L,WONOKASIAN,rt01,TPS-1,dalam,HS,H
…	…

Table 6. Dataset, 10 attribute

No.	Data election
1	dua,SIDOARJO,Milenial,S,P,WONOKASIAN,rt01,TPS-1,dalam,TH
2	dua,SIDOARJO,Milenial,S,L,WONOKASIAN,rt01,TPS-1,dalam,H
3	>dua,LUAR,Milenial,B,L,WONOKASIAN,rt01,TPS-1,dalam,TH
4	>dua,LUAR,PascaMilenial,S,P,WONOKASIAN,rt01,TPS-1,dalam,TH
…	…

Table 7. Dataset, 9 attribute

No.	Data election
1	dua,SIDOARJO,Milenial,S,P,WONOKASIAN,dalam,TSH,TH
2	dua,SIDOARJO,Milenial,S,L,WONOKASIAN,dalam,TSH,H
3	>dua,LUAR,Milenial,B,L,WONOKASIAN,dalam,THS,TH
4	>dua,LUAR,PascaMilenial,S,P,WONOKASIAN,dalam,THS,TH
…	…

5. Testing data results

The results of the classification algorithm (C4.5, Naïve Bayes and kNN) to determine the prediction class "Hadir and Tidak hadir". The results of Table 8, explain the comparison of class predictions in both the number of attributes and algorithm groups. The highest value is in 11 attributes, namely, in the Naïve Bayes algorithm with a percentage value of 89.3417% or 1140 instances. In this analysis, try to see the results in the reduction of attributes. Subtraction is done for attributes that have the first ranking "statusKK" (10 attributes), second "TPS" and third rank "rt" (9 attributes). For the highest ranking of Gini ratio Table 3. For prediction class 9 attributes, where the results of the prediction class show the highest value of 3 algorithms namely C4.5 with a percentage of 99.6364% or 1131 instances, then the highest percentage prediction class for ten attributes is 10.7555 % or 1107 instances with algorithms that have the same nine attributes, namely C4.5 (Figure 3). The best ranking is based on two groups of attributes (9 attributes and ten attributes), namely on nine attributes.

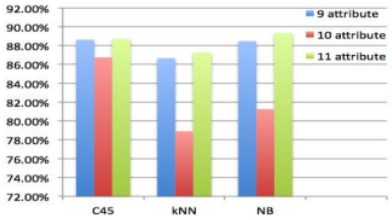

Figure 3. Graph of prediction class results.

Table 8. Result testing data

Attribute		C45		kNN		Naïve Bayes	
9 attribute	True	1131	**88.6364 %**	1106	86.6771 %	1129	88.4796 %
	False	145	11.3636 %	170	13.3229 %	147	11.5204 %
10 attribute	True	1107	**86.7555 %**	1007	78.9185 %	1037	81.2696 %
	False	169	13.2445 %	269	21.0815 %	239	18.7304 %
11 attribute	True	1132	88.7147 %	1114	87.3041 %	1140	**89.3417 %**
	False	144	11.2853 %	162	12.6959 %	136	10.6583 %

6. Conclusion

Analysis of three algorithms (C4.5, Naive Bayes and k-NN) and changes to attributes, which uses a dataset of 4240 instances which are divided into training data 2973 instances and testing 1276 instances. The presentation of the results obtained for the Naïve Bayes algorithm (for 11 attributes) ranks first for the prediction class from the attendance list of the President, DPD, DPR, Provincial DPRD and Regency / City DPRD in 2019, with a percentage of 89.3417%. But here try to look at the analysis for the reduction of the number of attributes, namely the group of 9 attributes (attributes that are reduced by "rt" and "TPS") and ten attributes (attributes that are reduced by "statusKK"). The results show that for nine attributes, it is better with the percentage value for the predictive class, which is 88.6364%. This shows that for the reduction of attributes with the highest gain ratio (statusKK) in the ten attributes, it has a significant effect, so the predictive value of the ten attributes is bad. The results of the attributes that are not deleted (11 attributes) and deleted attributes (10 and 9 attributes) indicate that the results of predictions are different, ie for the best not deleted is Naïve Bayes and the best-deleted algorithm is C4.5. In the end it can be seen that the weight level of an attribute for the value of the gain ratio affects the results of the predictions both on the percentage value of the prediction class and especially on the best value for the algorithm used.

Acknowledgments

We hereby thank you to Universitas Muhammadiyah Sidoarjo for supporting the publication of this research.

References

[1] Bramer M 2016 *Data for Data Mining:In Principles of Data Mining* London: Springer.
[2] Zhang J, Williams S O and Wang H 2018 Intelligent computing system based on pattern recognition and data mining algorithms *Sustain. Comput. Informatics Syst.*vol **20** pp 192–202.
[3] Xie J and Burstein F 2011 Using machine learning to support resource quality assessment: An adaptive attribute-based approach for health information portals *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* vol **6637** LNCS pp 526–37.
[4] Nikam S S 2017 A comparative study of classification techniques in data mining algorithms *Int. J. Mod. Trends Eng. Res.* vol **4** no 7 pp 58–63.
[5] Xia G S, Hu J, Hu F, Shi B G, Bai X, Yanfei Z, Lu X, and Zhang L 2017 AID: A benchmark data set for performance evaluation of aerial scene classification *IEEE Trans. Geosci. Remote Sens.* vol **55** no 7 pp 3965–81.
[6] Novaković J, Strbac P and Bulatović D 2011 Toward optimal feature selection using ranking methods and classification algorithms *Yugosl. J. Oper. Res.* vol **21** no 1 pp 119–35.
[7] Rosid M A, Gunawan G, and Pramana E 2015 Centroid based classifier with TF – IDF – ICF for classfication of student's complaint at appliation e-complaint in Muhammadiyah University of Sidoarjo vol **1** no 1.
[8] Abdallah I, Dertimanis V, Mylonas C, Tatsis K, Chatzi E, Dervilis N, Worden K and Maguire A E 2018 Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data *Saf. Reliab. - Safe Soc. a Chang. World - Proc. 28th Int. Eur. Saf. Reliab. Conf. ESREL 2018* no December pp 3053–62.
[9] Yang Y and Chen W 2016 Taiga: Performance optimization of the C4.5 decision tree construction algorithm *Tsinghua Sci. Technol.*vol **21** no. 4 pp 415–25.
[10] Rashmi G D, Lekha A, and Bawane N 2016 Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for Breast Cancer dataset *2015 Int. Conf. Emerg. Res. Electron. Comput. Sci. Technol. ICERECT 2015* pp 108–13.
[11] Ginting S L B, Adler J, Ginting Y R, and Kurniadi A H 2018 The development of bank application for debtors selection by using Naïve Bayes Classifier technique *IOP Conf. Ser. Mater. Sci. Eng.*, vol **407** no 1.
[12] Balasaravanan K and Prakash M 2018 Detection of dengue disease using artificial neural network based classification techniquetion *Int. J. Eng. Technol.*vol **7** no 1 pp 13–5.
[13] Jadhav S D and Channe H P 2016 Comparative study of K-NN, Naive Bayes and Decision Tree Classification techniques *Int. J. Sci. Res.* vol **5** no 1 pp 1842–45.
[14] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, and Gutierrez J 2017 A comprehensive investigation and comparison of Machine Learning techniques in the domain of heart disease *Proc. - IEEE Symp. Comput. Commun.* no. Iscc pp. 204–7.

unplug Attribute analysis with classification algorithm on election participation

ORIGINALITY REPORT

6%	6%	4%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	M. C. Mantovani, H. V. Ribeiro, M. V. Moro, S. Picoli, R. S. Mendes. "Scaling laws and universality in the choice of election candidates", EPL (Europhysics Letters), 2011 Publication	1%
2	eduvest.greenvest.co.id Internet Source	1%
3	export.arxiv.org Internet Source	1%
4	archive.org Internet Source	1%
5	core.ac.uk Internet Source	1%
6	archive-ouverte.unige.ch Internet Source	1%
7	Submitted to Universitas Mercu Buana Student Paper	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		