Classification Using C4.5 Algorithm in Election Participation Prediction

IOP Conf. Series: Materials Science and Engineering **874** (2020) 012016 doi:10.1088/1757-899X/874/1/012016

Classification Using C4.5 Algorithm in Election Participation Prediction

Arif Senja Fitrani[1], Mochamad Alfan Rosid[2], Cindy Taurusta[3], Indah Fauzia[4]

Universitas Muhammadiyah Sidoarjo, Sidoarjo, Indonesia

asfjim@umsida.ac.id[1], alfanrosid@umsida.ac.id[2], cindytaurusta@umsida.ac.id[3], Indahfauzia09@umsida.ac.idi[4]

Abstract. The General Election of the President, DPD, DPR, Provincial DPRD and Regency / City DPRD in 2019, in Indonesia, is carried out by an independent institution namely the General Election Commission (KPU), where there is a measure of the success of the holding by holding the principle of direct, general, free of secrecy. Another component of the election will be the implementation of contestants and voters. In the electoral factor, this is also a measure for success in the overall implementation process, which is a success if the participation of the community is high in the holding of elections. And conversely, if the community participation is low, one of them is the level of public trust in the organizer (government) decreases. Analysis of classification mining data with C4.5 algorithm places the prediction class "Present" and "Not Present" on the permanent voter list (DPT) taken from form C7 (voter attendance list form). C7 data was obtained from polling stations (TPS) located in the village of Wonokasian, Wonoayu District, Sidoarjo Regency, East Java. The number of datasets is 4249 instances, and the number of attributes is 10. With split datasets, for training data are 2965 instances, and testing data are 1284 instances. From testing, the testing data obtained the results of true or appropriate accuracy of 1151 instances (90%) and obtained incorrect or incompatible data as many as 134 instances (10%). This shows a good value for the results of 90% of the corresponding data in the prediction class in the Presidential Election, DPD, DPR, Provincial DPRD, and Regency / City DPRD in 2019.

Keywords: Data Mining, Classification, Algorithm Classification, Election

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 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. At worst, it can cause the algorithm to be a poor outcome [23].

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

2. Classification

Classification method that is studying a set of data that can be produced by a new classification of data. 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. 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 [6].

3. C4.5 algorithm

At the learning stage of the data, the C4.5 algorithm constructs decision trees from training data, 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: set of cases
A: attribute
n: number of partition attributes A
Si: number of cases on the i-partition
S: number of cases in S

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

Where :
S: set of cases
A: feature
n: number of S partitions
pi: proportion of Si to S

4. 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).:

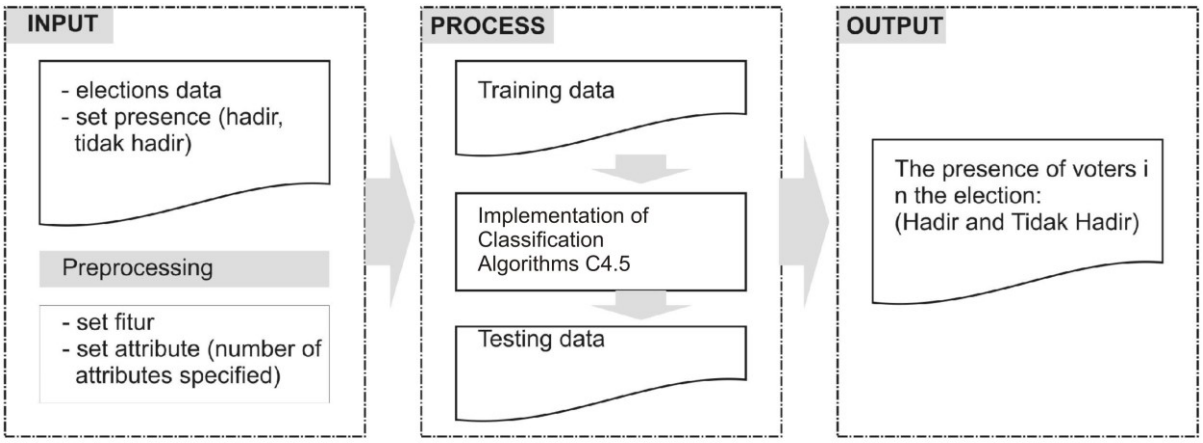

Figure 1. classification diagram.

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 [23].

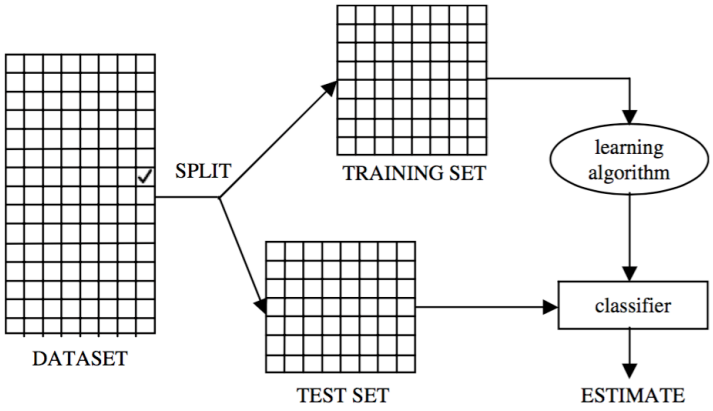

Figure 2. split train and test.

5. 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.

5.1 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 10 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	10	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 1:

Table 2. 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)

5.2 Dataset

At the preprocessing stage, a dataset is determined, which states the data is in normal condition. To prepare for the next process, split two data conditions, namely, training data and testing data with prediction classes present and not present, Table 5.

Table 3. dataset, 10 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
…	…

5.3 Ratting Attribute

To see the best weight of 10 attributes, it can be done by calculating the gain ratio, so that the best and worst attribute sequence can be known, namely:

Table 4. 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

5.4 Training Data

The dynamic condition characteristics of the dataset for the prediction class are present and not present, specifying the instance conditions for each attribute and variable. In figure 3, it also shows visually the number of each variable, where each instance in each variable shows a dynamic condition with 70% training data (2973 instances) out of a total of 4249 data sets. The condition in the prediction class for features is present in 614 instances (20.65%), and 2359 instances (79.35%) are absent.

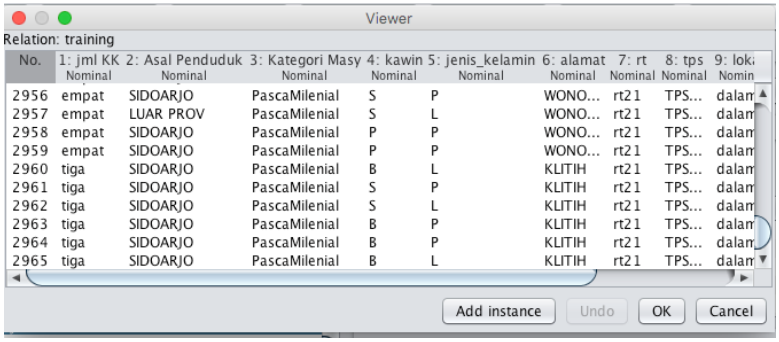

Figure 3. Training data.

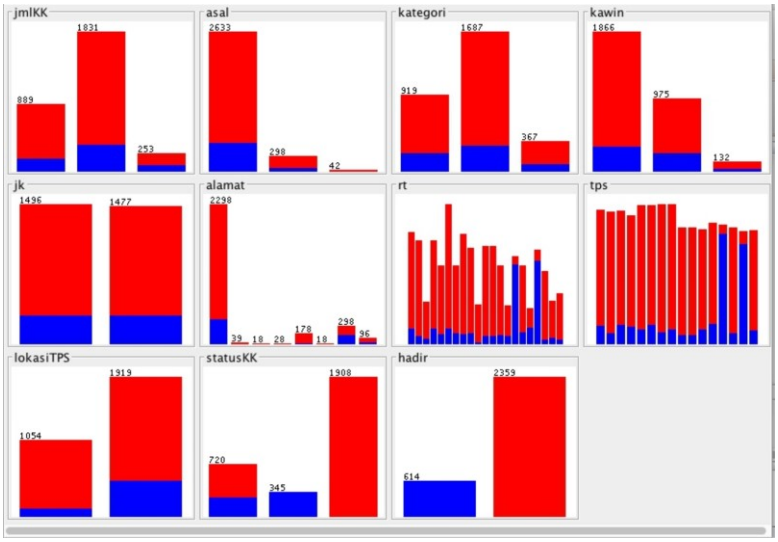

Figure 4. Visual all attribute.

Table 5. Classifier model (full training set)

No	Rule	No	Rule
1	hadir kel = THS: TH (329.0)	17	| | | | jenis_kelamin = P: TH (2.0)
2	hadir kel = HS: H (1906.0)	18	| | | tps = TPS-6: H (3.0/1.0)
3	hadir kel = TSH	19	| | | tps = TPS-7: H (3.0/1.0)
4	| kawin = S: H (425.0/118.0)	20	| | | tps = TPS-8: TH (0.0)
5	| kawin = B	21	| | | tps = TPS-9: H (3.0/1.0)
6	| | jml KK = tiga: TH (80.0/28.0)	22	| | | tps = TPS-10: TH (0.0)
7	| | jml KK = dua: TH (26.0/11.0)	23	| | | tps = TPS-11: TH (6.0/2.0)
8	| | jml KK = empat	24	| | | tps = TPS-12: H (9.0/2.0)
9	| | | Kategori Masy = PascaMilenial: TH	25	| | | tps = TPS-13: TH (6.0)
10	(10.0/5.0)	26	| | | tps = TPS-14
11	| | | Kategori Masy = Milenial: TH (66.0/30.0)	27	| | | | jenis_kelamin = L: H
12	| | | Kategori Masy = Pemula: H (15.0/5.0)	28	(4.0/1.0)
13	| | jml KK = lima	29	| | | | jenis_kelamin = P: TH (2.0)
14	| | | tps = TPS-1: TH (6.0/2.0)	30	| | | tps = TPS-15: TH (0.0)
15	| | | tps = TPS-2: TH (2.0/1.0)	31	| | | tps = TPS-16: TH (0.0)
16	| | | tps = TPS-3: H (1.0)	32	| | jml KK = satu: TH (2.0)
	| | | tps = TPS-4: H (3.0/1.0)		| | jml KK = enam: H (23.0/7.0)
	| | | tps = TPS-5		| | jml KK = tujuh: TH (0.0)
	| | | | jenis_kelamin = L: H (6.0/2.0)		| kawin = P: H (27.0/12.0)

From the process for calculating training data, the rule base is obtained with Leaves 30 and Size of the tree 37, and Table 6 sets the rule base, which will lead to all the testing processes carried out for test data in addition to the results obtained for the rule base. Figure 4, visual classification tree visual, encourages to make it easier to visually understand the rule base on the results of the Algorithm C4.5.

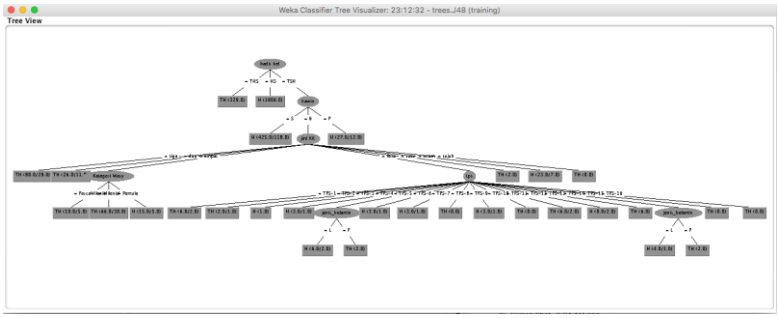

Figure 5. Tree view.

6. Testing Data Results

The results of the classification algorithm C4.5 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. The best ranking is based on two groups of attributes (9 attributes and ten attributes), namely on nine attributes.

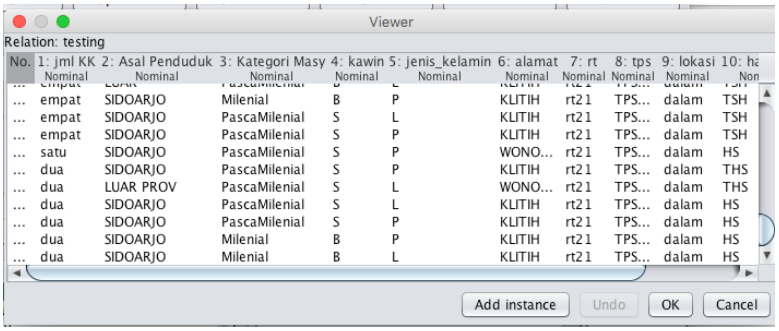

Figure 6. Testing data.

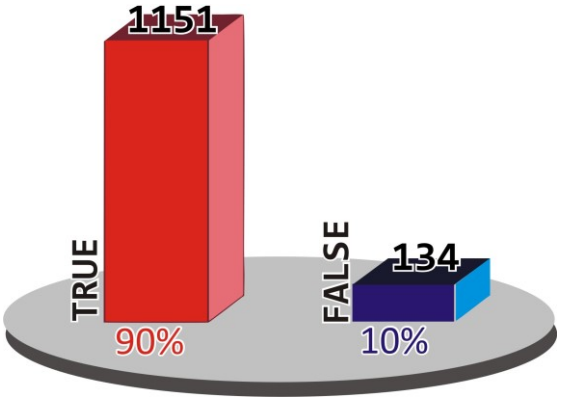

Figure 7. Graph of prediction class results

Table 6. Result Testing Data

Attribute	Status	C45	
11 attribute	True	1151	90 %
	False	134	10 %

5. Conclusion

Analysis of the three C4.5 algorithms in the test contained ten attributes and with the Present and Not Present classes of predictions. The total dataset uses 4249 instances, which are divided into 2965 instances of training data and 1284 instances of testing data. The dataset used is C7 form in the presidential election, DPD, DPR, Provincial DPRD, and Regency / City DPRD held by KPU in Wonokasian village, Wonoayu sub-district, Sidoarjo Regency, East Java. Testing the classification method on the C4.5 algorithm obtained prediction results that are as many as 1151 instances or 90% of the conditions declared correct and 134 instances or 10% of the terms said inappropriate or incorrect. Then from the results of the true percentage value, which is significant, the testing process in this dataset can prove the level of voter attendance in the election implementation and the presentation model of attributes that support or have considerable weight as well. It is hoped that the testing process can be carried out in the implementation of the upcoming elections.

Acknowledgments

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

References

[1] M. (School of C. of P. Bramer, *Data for Data Mining. In: Principles of Data Mining*. London: Springer, 2016.
[2] J. Zhang, S. O. Williams, and H. Wang, "Intelligent computing system based on pattern recognition and data mining algorithms," *Sustain. Comput. Informatics Syst.*, vol. 20, pp. 192–202, 2018.
[3] J. Xie and F. Burstein, "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. 526–537, 2011.
[4] S. S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms," *Int. J. Mod. Trends Eng. Res.*, vol. 4, no. 7, pp. 58–63, 2017.
[5] G. S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.
[6] J. Novaković, P. Strbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugosl. J. Oper. Res.*, vol. 21, no. 1, pp. 119–135, 2011.
[7] M. A. Rosid, G. Gunawan, and E. Pramana, "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, 2015.
[8] I. Abdallah *et al.*, "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. 3053–3062, 2018.
[9] Y. Yang and W. Chen, "Taiga: Performance optimization of the C4.5 decision tree construction algorithm," *Tsinghua Sci. Technol.*, vol. 21, no. 4, pp. 415–425, 2016.
[10] G. D. Rashmi, A. Lekha, and N. Bawane, "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. 108–113, 2016.
[11] S. L. B. Ginting, J. Adler, Y. R. Ginting, and A. H. Kurniadi, "The Development of Bank Application for Debtors Selection by Using Naïve Bayes Classifier Technique," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, no. 1, 2018.
[12] K. Balasaravanan and M. Prakash, "Detection of dengue disease using artificial neural network based classification techniquetion," *Int. J. Eng. Technol.*, vol. 7, no. 1, pp. 13–15, 2018.
[13] S. D. Jadhav and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *Int. J. Sci. Res.*, vol. 5, no. 1, pp. 1842–1845, 2016.
[14] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," *Proc. - IEEE Symp. Comput. Commun.*, no. Iscc, pp. 204–207, 2017.
[15] A. P. Windarto *et al.*, "Analysis of the K-Means Algorithm on Clean Water Customers Based

on the Province," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012001.

[16] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, "C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject," *J. Phys. Conf. Ser.*, vol. 1255, no. 012005, pp. 1–7, 2019, doi: 10.1088/1742-6596/1255/1/012005.

[17] Sudirman, A. P. Windarto, and A. Wanto, "Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018, doi: 10.1088/1757-899X/420/1/012089.