Classifying the Parental Involvement on School From Home during Covid-19 using C4.5 Algorithm

Ilham A.E. Zaeni
Dept. of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
Corresponding author:
ilham.ari.ft@um.ac.id

Dessy Rif'a Anzani
Dept. of Psychology
UIN Maulana Malik Ibrahim
Malang, Indonesia
dr.anzani90@gmail.com

Dwi Sudarno Putra
Dept. of Electrical Engineering
Southern Taiwan University of Science and Technology
Taina, Taiwan
da72b205@stust.edu.tw

Mazarina Devi
Dept. of Industrial Technologyg
Universitas Negeri Malang
Malang, Indonesia
mazarina.devi.ft@um.ac.id

Laili Hidayati,
Dept. of Industrial Technologyg
Universitas Negeri Malang
Malang, Indonesia
laili.hidayati.ft@um.ac.id

Imam Sudjono
Dept. of Mechanical Engineering
Universitas Negeri Malang
Malang, Indonesia
imam.sudjono.ft@um.ac.id

***Abstract*— In implementing School from home (SFH) for kindergarten and elementary school children, the involvement of parents is needed. Parent involvement is behavior such as activities, support, decision making, advice, and example in the family to support the child's development. One of the factors that influence parental involvement is demographics. This study aims to classify the level of parental involvement based on demographic factors using the C4.5 algorithm. The research process starts with the preparation of the scale of parental involvement and then categorizes the parental involvement score into very high, high, medium, low, and very low. Demographic factors after conducting the data preprocessing are used as input in the C4.5 algorithm. The tree that has been developed shows that the root of the tree was the mother education level. The test results show that the accuracy of the C4.5 algorithm in classifying parental involvement is 86.67%. This result is good enough to be used in the classification process.**

Keywords—parental involvement, demographic, classification, C4.5 Algorithm

I. Introduction

Lately, the world has been rocked by the presence of the coronavirus or Covid-19. The virus also hit Indonesia so the government also appealed to the public to reduce activities outside the home. This effort aims to break the chain of the spread of the Covid-19 virus in Indonesia. School from home (SFH) is one of the programs carried out by the government in the Covid-19 pandemic to reduce the spread of the virus, so students are required to study at home [1]. However, the implementation of SFH for kindergarten and elementary school students has its own constraints. In implementing SFH for kindergarten and elementary school, parents are required to play an active role in the teaching and learning process. During the SFH, parental involvement is required.

Parental engagement is a variety of behavioral patterns such as direct activities with other people, support that provides benefits among parents, a sense of security given to families, decision making, the advice in the family environment, and how exemplary in the family[2]. Parental involvement (PI) in school is correlated with the success rate of children on academic success and social skills [3]. Parental involvement in parenting can strengthen the relationship between children and parents so that the child's growth and development process in social and educational aspects can run well [4].

Parental engagement consists of 3 aspects, namely: understanding as a parent, preparation for children's schooling, and parental participation in children's daily lives. As part of parental engagement, parents need to understand their role in children's development which includes cognitive, social, and emotional aspects as well as the child's physical growth [2]. Parents' attention to the readiness of children to learn in school must also be considered to facilitate the child in the process of adaptation and learning process to be more effective [5]. The role of parents in children's daily activities is important in the process of education and parenting.

Some researchers have examined the relationship between parental involvement and parental demographics [3], [6]–[8]. Factors related to demographics are parent and child age and gender, family structure and size, parent education, employment status, household income, government benefits, language spoken at home [6]. Research has been conducted to evaluate the relationship between demographic, parental, and cognitive, linguistic, and motor skills development in children with learning disabilities [7]. For this reason, a method is needed to classify the level of parental involvement based on parental demographic data.

One classification algorithm that is often used in research is the C4.5 algorithm. C4.5 is an algorithm developed from the ID3 algorithm [9]. C4.5 algorithm has been used for various purposes such as customer classification as a basis for granting credit [10], [11]. The C4.5 algorithm has also been used to classify the graduation predicate [12] and prediction of self-candidate new students in higher education [13]. In the health field, the C4.5 algorithm is used to diagnose diseases such as stroke, dengue fever, and diabetes patients [14]–[16].

Based on the description above, a study was conducted to classify parental involvement based on parents' demographics. The result of classification can be used for consideration in a campaign to improve the parental involvement. Using the classification result, the school can predict which family that should be more encouraged to participate in SFH activity by improving their parental involvement.

II. Methods

A. Parental Involvement Scale

To be able to measure the level of parental involvement, a parental engagement scale needs to be arranged. This scale is arranged based on the indicators that have been described previously, namely understanding as a parent, preparation for children's schooling, and parental participation in children's daily lives. The scale is applied in the form of a questionnaire. The number of questions compiled in this questionnaire was 36 questions. The blueprint of the Scale is shown in Table I.

The instrument must be used to measure what should be measured so that it can be stated that the instrument is valid [17]. The validity test on this instrument is carried out by applying the Content Validity Ratio (CVR) developed by Lawshe [18]. A validity test is carried out by 3 experts who are suitable in their field. If the CVR score is greater than 0.99, then the question items can be considered valid. The number of items tested for validity is 36 items. The CVR results obtained are there are 25 valid question items and there are 11 invalid question items.

The score results obtained through the parental involvement questionnaire are then categorized into 5 categories: very high, high, medium, low, and very low. This category will later be used as the target class in the classification process. The division of category is based on the criteria that are described in Table II. Where M is the mean and SD is the standard deviation.

B. Preprocessing Data

Preprocessing data or data preparation is a process or step taken to make raw data into quality data [15]. So that the data preprocessing stage in this study uses data transformation. Data transformation is used to assist in the data processing. This conversion aims to equalize the data types used (numeric). At this stage, the conversion is done by changing several attribute data into id or code. Transformation data for the attributes of father's education or mother's education are derived from parameters Elementary School/equivalent, Junior High School/equivalent, Senior High School /equivalent, Diploma, Bachelor, Master, and Doctoral to a value of 1 to 7. Transformation data for the attribute of the father's income or maternal income is made from Rp. 0 to Rp. 999,999, Rp. 1,000,000 to Rp. 1,999,999, Rp. 2,000,000, - to Rp. 2,999. 999, - Rp. 3,000,000, - to Rp. 3,999,999, - Rp. 4,000,000, - to Rp. 4,999,999, - and> Rp. 5,000,000, - becomes a value of 1 to 6. C4.5 Algorithm

C4.5 algorithm is a decision-making algorithm developed by Ross Quinlan[9]. The basic idea of this algorithm is to make a decision tree based on selecting attributes which have the highest priority or can be assumed to have the highest benefit value based on the entropy attribute value as the axis of classification attributes[11].

The stage of the C4.5 algorithm has two working concepts, namely the creation of a decision tree and the rule-making (rule model). The rules created from the decision tree will constitute a condition in the form of if-then rule [20]. In the C4.5 algorithm, there are four steps in the decision tree making process, including selecting attributes as root, creating branches for each value, dividing each case into a branch, repeating the process in each branch so that all branch cases are of the same class.

Calculation of the entropy value can be seen in (1) below:

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi \tag{1}$$

Where *S* is the set of cases. The *n* is the number of partitions *S*, and the *pi* is the proportion from *Si* to *S*.

During the construction of the decision tree, several branches in the training data may represent noise or outliers. Could be made to identify and cut branches by pruning trees. Trimmed trees should be smaller and more comprehensible.

Choose an attribute as the source, based on the current attributes' highest benefit value. Calculating the cost using the formula as define on equation (2)-(4) below: The C4.5 Algorithm using Gain ratio to determine which attribute will be chosen as the root or branch of the tree. The gain ratio can be calculated as:

$$Gain\ Ratio(S, A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \tag{2}$$

Where the S is space (data) sample used for training. The A is the attribute. The Split Info (S, A) is the split information on attribute A. The Gain (S, A) is the information gain at attribute A. The Gain can be calculated as:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \tag{3}$$

Where *S* is the set of cases and *A* is the attributes. The *n* is the number of attribute attributes *A*, the |*Si*| is the number of cases on the *i*-th partition, and the |*S*| is the number of cases in *S*. The Split Info can be calculated as:

$$Split\ Info\ (S, A) = \sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \tag{4}$$

C. Validation and Testing

The validation and testing methodology in this analysis employs Cross-Validation. Cross-validation is used for compiling and evaluating the amount of training data to be used and the amount of test data to be used.

One method of cross-validation will be used in this analysis, with as many as experiments on classification k. The k value that will be used in this research is 10-fold, which is the best choice for a very precise validation process. The output of the C4.5 algorithm was calculated using a confusion matrix to find out precision, recall, and accuracy.

TABLE I. BLUEPRINT OF PARENTAL INVOLVEMENT SCALE

Variable	Indicator	Number of Question
Parental Involvement	1. The concept of understanding as a parent	14
	2. Preparation of children's schooling. In the field of education	9
	3. Parental participation in children's daily lives	13

TABLE II. CRITERIA FOR CATEGORIZING THE PARENTAL INVOLVEMENT SCORE [19]

Category	Criteria
Very High	Score>M+1.5*SD
High	M+1.5*SD ≥Score>M+0.5*SD
Medium	M+0.5*SD ≥Score>M-0.5*SD
Low	M-0.5*SD ≥Score>M-1.5*SD
Very Low	M-1.5*SD ≥Score

III. Result and Discussion

A. *Parental Involvement*

The parental involvement questionnaire consist of 25 questions with the score of each question is ranged from 1 to 4. Based on the parental involvement score of all the subjects, it can be known that the total score ranged between 32 to 90. The average score of all the subjects is 68.28 and the standard deviation is 15.97. Based on these scores, the parental involvement of each subject can be categorized into very high, high, medium, low, and very low. Table III shows the distribution of each category of parental involvement score.

The parental involvement category that is shown in Table III is not balanced. Most of the subject has a high and medium score of parental involvement. There is no subject that has a very high parental involvement score. There are 8% and 12 of the subject that has a low and very low score, respectively.

B. *Demographic Data*

The demographic data that has been collected is including the child age and gender, parents age, parent education level, parent occupation, parent earning, number of siblings, and number of family members at home. The demographic data is then used as the input attribute of the classification algorithm. The example of demographic data that has been collected in this study can be shown on Fig. 1 to Fig 4.

Fig 1 to Fig 4 shows the scatter plot of each category of parental involvement based on the demographic data. Fig 1 shows the position of the high, medium, low, and very low category based on the father's and mother's education level. This image shows that the category of parental engagement is mixed in all areas of the graph, so it's not easy to be categorized. The other image on Fig. 2 to Fig. 4 also shows the same result. Each category is overlapped on the entire graph. Therefore, the demographic data cannot be easily used for classification by using one or two attributes.

C. *Classification using C4.5 Algorithm*

The demographic data is then used to classify the category of parental involvement level. The classification process is using the C4.5 algorithm. The tree that has been developed shows that the number of leaves is 24. The size of the tree is 47. The root of the tree is the mother's education level. The tree that has been developed is shown in Fig 5.

D. *Algorithm Performance*

The C.45. algorithm performance was evaluated by using the confusion matrix. The parental involvement was categorized into 5 categories, but there are no subject that can be categorized as very high parental involvement. Therefore, only 4 categories were included into confusion matrix. The confusion matrix that has been obtained during the experiment is shown in Table IV.

TABLE III. DISTRIBUTION OF PARENTAL INVOLVEMENT CATEGORY

	Number of Subject	Percentage
Very High	0	0%
High	63	42%
Medium	57	38%
Low	12	8%
Very Low	18	12%

Based on the result that is shown on Table IV, it can be concluded that most of the data can be correctly classified. There are 130 instances that can be correctly classified. Therefore, the accuracy of the algorithm is 86.67%. There are 20 instances that can be incorrectly classified, or the error rate is 13.33%. The detailed accuracy is shown on Table V.

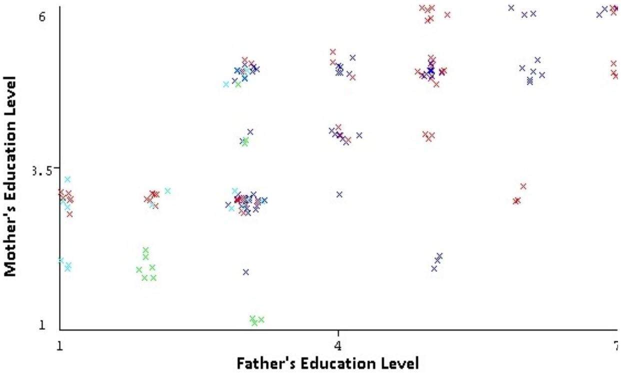

Fig. 1. The scatter plot of Father's Education vs Mother's Education

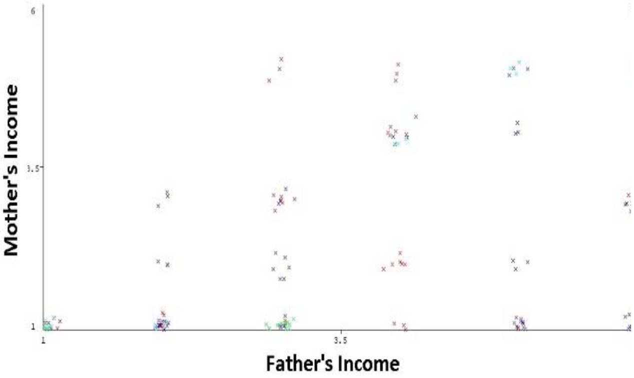

Fig. 2. The scatter plot of Father's Income vs Mother's Income

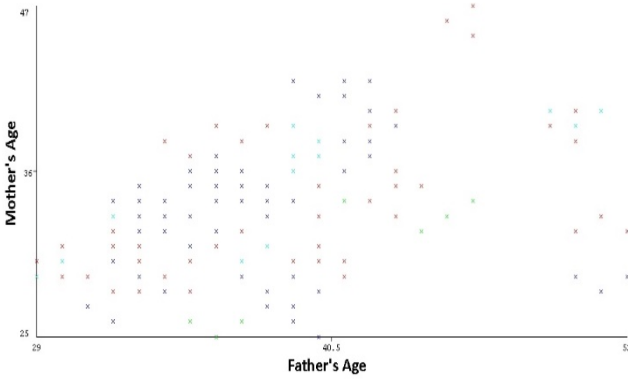

Fig. 3. The scatter plot of Father's Age vs Mother's Age

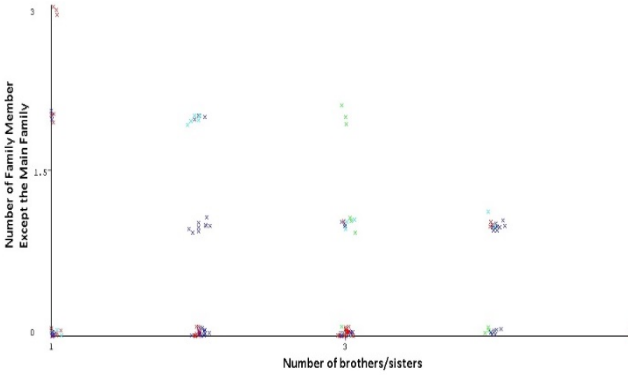

Fig. 4. The scatter plot of Number of Siblings vs Number of Family member except main Family

```
Mother's Education <= 2
| Father's Education <= 1: Very Low (3.0)
| Father's Education> 1
| | Gender of Children <= 1: Low (10.0 / 1.0)
| | Child Gender> 1: High (3.0)
Mother's Education> 2
| Father's Occupation <= 1
| | Mother's Education <= 3: Medium (6.0)
| | Mother's Education> 3: Low (3.0)
| Father's Occupation> 1
| | Average Father Earnings <= 1
| | | Father's age (in years) <= 35: High (2.0)
| | | Father's age (in years)> 35: Very Low (6.0)
| | Average Father Earnings per month> 1
| | | Father's Education <= 2
| | | | Family members except the main family <= 0: Medium (9.0)
| | | | Family members except the main family > 0: Very Low (3.0)
| | | Father's Education> 2
| | | | Average Father Earnings <= 2: High (15.0)
| | | | Average Father Earnings per month> 2
| | | | | Father's Education <= 3
| | | | | | Gender of Children <= 1
| | | | | | | Mother's age (in years) <= 28: High (3.0)
| | | | | | | Mother's age (in years)> 28
| | | | | | | | Mother's Occupation <= 1: Very Low (3.0)
| | | | | | | | Mother's Work> 1
| | | | | | | | | Average Father Earnings <= 4: Medium (9.0)
| | | | | | | | | Average Father Earnings > 4: Very Low (3.0)
| | | | | | Child Gender> 1: High (5.0)
| | | | | Father's Education> 3
| | | | | | Gender of Children <= 1
| | | | | | | Mother's age (in years) <= 32
| | | | | | | | Average Father Earnings <= 4: Medium (3.0)
| | | | | | | | Average Father Earnings > 4
| | | | | | | | | Mother's Occupation <= 1: Medium (3.0)
| | | | | | | | | Mother's Work> 1
| | | | | | | | | | Age of Child (in years) <= 9: High (8.0)
| | | | | | | | | | Child Age (in years)> 9: Medium (4.0 / 1.0)
| | | | | | | Mother's age (in years)> 32: High (15.0)
| | | | | | Gender of Children> 1
| | | | | | | Number of siblings <= 3
| | | | | | | | Number of siblings <= 1
| | | | | | | | | Average Father Earnings <= 3: High (4.0)
| | | | | | | | | Average Father Earnings > 3: Medium (3.0)
| | | | | | | | Number of siblings> 1: Medium (21.0)
| | | | | | | Number of siblings> 3: High (6.0)
```

Fig. 5. The Tree that Has Been Developed

Based on Table V, it can be known that the weighted average of True positive rate, Precision, Recall, and F-measure are 0.867, 0.869, 0.867, and 0.867, respectively. The best result is obtained on the Low category of parental involvement with the True positive rate, Precision, Recall, and F-measure are 1, 0.923, 1, and 0.96, respectively.

Compared classification result on other case using the same algorithm, such as university student satisfaction analysis [21], this result is lower than the other result. The university student satisfaction analysis using C4.5 algorithm yields an accuracy of 95%. Compared to other study which evaluate students satisfaction based on student sentiment in Twitter Social Media using Naïve Bayes algorithm that yield an accuracy of 84% [22], this study gives a better result. This study yield accuracy of 86.67% which is can be categorized as good classification[21].

TABLE IV. DETAILED ACCURACY BY CLASS

Class	TP Rate	Precision	Recall	F-Measure
High	0.825	0.881	0.825	0.852
Medium	0.877	0.877	0.877	0.877
Low	1	0.923	1	0.96
Very Low	0.889	0.762	0.889	0.821
Weighted Avg.	**0.867**	**0.869**	**0.867**	**0.867**

IV. CONCLUSION

The study has been conducted to classify the parental involvement level using C4.5 algorithm. The study has been conducted by several step, including compiling the parental involvement questionnaire, collecting the demographic data, preprocessing the data, and classifying the parental involvement level based on demographic data using C4.5 algorithm. The tree that has been developed has 24 leaves with tree size is 47. The root of the tree is mother education level. Based on the discussion about the performance of C4.5 Algorithm on classifying the parental involvement level, it can be concluded that the accuracy of the algorithm is 86.66%. This result is good enough. Therefore, the C4.5 algorithm can be used to classify the parental involvement level. The classification output may be used for consideration in a parental participation enhancement program. Using the classification result, the school can predict which family should be encouraged to improve their parental involvement. By using the classification result, the parental involvement enhancement program could be effectively conducted.

REFERENCES

[1] K. P. dan K. R. Indonesia, "Mendikbud Terbitkan SE tentang Pelaksanaan Pendidikan dalam Masa Darurat Covid-19." [Online]. Available: https://www.kemdikbud.go.id/main/blog/2020/03/mendikbud-terbitkan-se-tentang-pelaksanaan-pendidikan-dalam-masa-darurat-covid19. [Accessed: 26-Jun-2020].

[2] S. M. Sheridan, L. L. Knoche, K. A. Kupzyk, C. P. Edwards, and C. A. Marvin, "A randomized trial examining the effects of parent engagement on early language and literacy: The Getting Ready intervention," *J. Sch. Psychol.*, vol. 49, no. 3, pp. 361–383, Jun. 2011, doi: 10.1016/j.jsp.2011.03.001.

[3] G. O. Kohl *et al.*, "Parent involvement in school conceptualizing multiple dimensions and their relations with family and demographic risk factors," *J. Sch. Psychol.*, vol. 38, no. 6, pp. 501–523, Nov. 2000, doi: 10.1016/S0022-4405(00)00050-9.

[4] G. Hornby and R. Lafaele, "Barriers to parental involvement in education: an explanatory model," *Educ. Rev.*, vol. 63, no. 1, pp. 37–52, Feb. 2011, doi: 10.1080/00131911.2010.488049.

[5] J. P. Spillane, "Educational Leadership," *Educ. Eval. Policy Anal.*, vol. 26, no. 2, pp. 169–172, Jun. 2004, doi: 10.3102/01623737026002169.

[6] N. J. Hackworth *et al.*, "What Influences Parental Engagement in Early Intervention? Parent, Program and Community Predictors of Enrolment, Retention and Involvement," *Prev. Sci.*, vol. 19, no. 7, pp. 880–893, Oct. 2018, doi: 10.1007/s11121-018-0897-2.

[7] R. M. Vilaseca *et al.*, "Demographic and parental factors associated with developmental outcomes in children with intellectual disabilities," *Front. Psychol.*, vol. 10, no. APR, p. 872, Apr. 2019, doi: 10.3389/fpsyg.2019.00872.

[8] N. Choi, M. Chang, S. Kim, and T. G. Reio, "A Structural Model of Parent Involvement with Demographic and Academic Variables," *Psychol. Sch.*, vol. 52, no. 2, pp. 154–167, Feb. 2015, doi: 10.1002/pits.21813.

[9] J. R. Quinlan, *C4. 5: Programs for machine learning*. Morgan Kaufmann Publishers, 1987.

[10] L. N. Rani, “Klasifikasi Nasabah Menggunakan Algoritma C4 . 5 Sebagai Dasar Pemberian Kredit,” 2016.

[11] S. A. Lusinia, S. Kom, M. Kom, and I. Komputer, “Algoritma C4.5 Dalam Menganalisa Kelayakan Kredit(Studi Kasus Di Koperasi Pegawai Republik Indonesia (KPRI) Lengayang Pesisir Selatan, Painan, Sumatera Barat),” *J. KomTekInfo Fak. Ilmu Komput.*, vol. 1, no. 2, Feb. 2014.

[12] Y. S. Nugroho, “Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta,” in *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST)*, 2014.

[13] E. Darmawan, “C4.5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education,” *J. Online Inform.*, vol. 3, no. 1, p. 22, Jun. 2018, doi: 10.15575/join.v3i1.171.

[14] L. Amini *et al.*, “Prediction and control of stroke by data mining,” *Int. J. Prev. Med.*, vol. 4, no. Suppl 2, pp. S245–S249, 2013.

[15] A. Andriani, “Klasifikasi Berbasis Algoritma C4 . 5 untuk Deteksi Kenaikan Case Fatality Rate Demam Berdarah,” pp. 70–75, 2017.

[16] U. Pujianto, A. L. Setiawan, H. A. Rosyid, and A. M. M. Salah, “Comparison of Naïve Bayes Algorithm and Decision Tree C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement,” *Knowl. Eng. Data Sci.*, vol. 2, no. 2, 2019, doi: http://dx.doi.org/10.17977/um018v2i22019p58-71.

[17] Sugiyono, *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta, 2014.

[18] C. H. Lawshe, “A Quantitative Approach to Content Validity,” *Pers. Psychol.*, vol. 28, no. 4, pp. 563–575, Dec. 1975, doi: 10.1111/j.1744-6570.1975.tb01393.x.

[19] S. Azwar, *Metode Penelitian*. Yogyakarta: Pustaka Pelajar, 2012.

[20] H. Widayu, S. Darma Nasution, and N. Silalahi, “Data Mining Untuk Memprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5,” *J. MEDIA Inform. BUDIDARMA*, vol. 1, no. 2, Jun. 2017.

[21] F. Aldi and A. Ade Rahma, “University Student Satisfaction Analysis on Academic Services by Using Decision Tree C4.5 Algorithm (Case Study : Universitas Putra Indonesia ‘YPTK’ Padang),” *J. Phys. Conf. Ser.*, vol. 1339, p. 012051, Dec. 2019, doi: 10.1088/1742-6596/1339/1/012051.

[22] F. C. Permana, Y. Rosmansyah, and A. S. Abdullah, “Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media,” in *Journal of Physics: Conference Series*, 2017, vol. 893, no. 1, doi: 10.1088/1742-6596/893/1/012051.