Divorce Factor Classification Uses The C4.5 Algorithm Based On Particle Swarm Optimization

Endang Sri Palupi

Teknologi Informasi
Universitas Bina Sarana Informatika
endang.epl@bsi.ac.id

Abstract

Cases of household divorce increased in the West Java area during the Covid-19 pandemic. The pandemic has increased personal relationships and interactions between family members, and some families are using this opportunity to strengthen their relationships. However, increased family interaction can also result in increased conflict, leading to divorce. The author classifies divorce factors that have increased during the pandemic using the C4.5 Algorithm based on Particle Swarm Optimization (PSO). The main factors for divorce are economic factors that have hit during the pandemic coupled with unstable mental conditions resulting in poor communication and continuous fighting. So that the husband/wife leaves one of the parties, infidelity, and adultery, then domestic violence and ending in divorce. The dataset was taken from the West Java BPS website, and the author split the data, namely 80% training data and 20% testing data, to avoid overfitting. Research results on the classification of divorce factors during the pandemic using the C4.5 algorithm based on particle swarm optimization are an accuracy value of 87.50% and an AUC (Area Under Curve) value of 0.807, which is included in the good classification category.

Keywords: C4.5 Algorithm, classification, particle swarm optimization

Abstrak

Kasus perceraian rumah tangga meningkat di daerah Jawa Barat selama pandemic covid-19 melanda. Pandemic meningkatkan hubungan dan interaksi personal antar anggota keluarga, beberapa keluarga menggunakan kesempatan ini untuk mempererat hubungan. Namun meningkatnya interaksi keluarga juga dapat mengakibatkan meningkatnya konflik yang berujung pada perceraian. Penulis mengklasifikasi faktor – faktor perceraian yang meningkat selama pandemic menggunakan Algoritma C4.5 berbasis Particle Swarm Optimization (PSO), faktor utama perceraian adalah faktor ekonomi yang melanda selama pandemic ditambah oleh kondisi mental yang tidak stabil mengakibatkan komunikasi yang tidak baik hingga terjadi pertengkaran terus menerus sehingga suami/istri meninggalkan salah satu pihak, perselingkuhan dan zina, kemudian KDRT dan berakhir dengan perceraian. Dataset diambil dari website BPS Jabar, penulis melakukan splitt data yaitu data training 80% dan data testing 20% agar tidak terjadi overfitting. Hasil penelitian klasifikasi faktor perceraian pada masa pandemic menggunakan algoritma C4.5 berbasis particle swarm optimization adalah: nilai akurasi 87,50% dan nilai AUC (Area Under Curve) 0.807 yang termasuk dalam kategori good classification

Kata kunci: Algoritma C4.5, klasifikasi, particle swarm optimization

INTRODUCTION

The number of divorce cases during the COVID-19 pandemic in the West Java area has increased. These divorces occur due to economic factors, domestic violence, drunkenness, adultery, disputes, and so on. Based on data on the BPS website, divorces in the West Java area in 2021 increased significantly. Therefore, the author wants to classify divorce factors during the pandemic using the C4.5 algorithm, which is suitable for classification problems in machine learning and data mining. The C4.5 algorithm is the simplest classification and is easy to implement.(Iriadi & Nuraeni, 2016)

According to N. Al Ngizati and colleagues, in research, entitled Factors Causing an Increase in Divorce During Covid-19 which was written in 2023, stated the most common factors causing divorce during the pandemic were obstacles to getting a job and a reduction in the workforce so that they could

not make ends meet. Family needs. The second most common factors are not getting a living, continuous fighting, economic factors, and behavioural attitudes if one party is positive for the COVID-19 virus, an unpleasant family response, and a lack of humanization of the affected party—viruses, which cause quarrels and disputes. (Yasifah et al., 2023) The author of this study used data taken from the West Java Province BPS website. The main factors in the occurrence of divorce in the West Java region are economic, continuous disputes, one party leaving the partner, domestic violence, polygamy, drunkenness, gambling, sentenced to prison, and apostate. In general, there are similarities, namely that the most important factor in divorce cases is economic factors, which have become increasingly difficult during the Covid-19 pandemic.

The research entitled Application of Data Mining in the Divorce Prediction Process Using the Naive Bayes Algorithm in Central Aceh Regency produced a web-based divorce prediction system. This system uses data mining with the Naive Bayes algorithm method to predict divorce. Couples who want to divorce can make predictions independently on the system before filing a lawsuit with the Sharia court. Users will get prediction results using data mining. This research was written by Desi Laila Sari and colleagues in 2021. (Sari & Gemasih, 2021) Meanwhile, the authors wanted to classify divorce factors using the PSO-based C4.5 algorithm and the RapidMiner modelling framework in this research. Meanwhile, Desi Laila makes divorce predictions, the application of which can be done independently by the user.

In 2023, Ihsan Ahmad Fauzi and colleagues wrote research entitled Implementation of Data Mining Clustering in Grouping Divorce Cases that Occurred in West Java Province Using the K-means Algorithm. This research groups data based on the same characteristics. The Davies Bouldin Index value is used to determine the number of clusters. The results of this research obtained the best cluster from grouping divorce cases, and there were 2 clusters, namely cluster 0, which contained five districts and nine cities. In contrast, in cluster 1, there were 13 districts, with a Davies Bouldin index value of 0.168 and an average centre of a mass distance value of 5.870. (Prihartono, Tohidi, Ahmad Fauzi, & Danar Dana, 2023) Meanwhile, researchers in this study classified divorce factors during the pandemic using the PSO-based C4.5 algorithm with results of an accuracy value of 87.50% and an AUC of 0.807, which is included in the good classification category.

Research written by M. Rayhan and colleagues entitled FP-Growth Algorithm to Understand Divorce Patterns During the Covid-19 Pandemic in Kab. Majalengka produced a combination of patterns for good association rules. From 97 processed data, 13 divorce pattern combination rules met the minimum support requirements of 30%, minimum confidence of 80%, and a lift ratio of more than 1. Economic factors and disputes are the leading causes of divorce in the Majalengka Regency. (Syarif, Purnamasari, & Mayasari, 2021) Meanwhile, the author classifies divorce factors in West Java during the pandemic using the PSO-based C4.5 algorithm. The main factor is economic, followed by continuous disputes and one party leaving the partner. The results of this research have an accuracy value of 87.50% and an AUC of 0.807, which is included in good classification.

The research entitled Application of C4.5 Based on PSO in Predicting Students Passing Higher Education Selection, written by Sulistiyanto in 2018, said that one of the classification techniques considered appropriate was C4.5, which was then optimized using the PSO. From the test results, the C4.5 algorithm produces an accuracy of 89.89%, while the accuracy of C4.5 with PSO is 94.56%. The experimental results prove that the PSO algorithm can improve the accuracy of C45 (Sulistiyanto, 2018). In this research, the author modelled classification using the PSO-based C4.5 algorithm. Still, there was no comparison of the accuracy results of the C4.5 algorithm using or not using PSO.

This research aims to classify the factors of divorce during the Covid-19 pandemic. The results of research into divorce factors during the Covid-19 pandemic are continuous disputes and quarrels, and these factors are caused by staying at home for too long, which results in various household frictions, including differences in understanding between husband and wife, couples who cannot balance relationships, time and activities.(Admaja, Hidayah, & Suparwi, 2021)

RESEARCH METHODS

The author used a descriptive quantitative method, namely research that describes variables as they are, supported by data in the form of numbers generated from the actual situation. According to (Sugiyono, 2018), the descriptive method is research carried out to describe independent variables, either only on one or more (stand-alone variables), without comparing and looking for that variable with other variables.
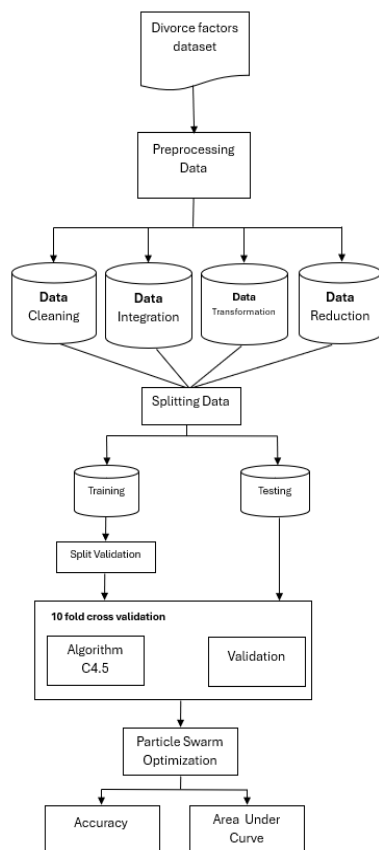
Figure 1. Research stages

Figure 1: The research stages carried out are:
1. Data Preprocessing
   Text preprocessing is the initial processing stage of text to prepare the text into data that will be processed further. (Feldman & James, 2007) In this stage, the data is cleaned, and several datasets are created to make the analysis process easier and reduce unnecessary data.
   a. Data Cleaning: This is the process of identifying and correcting problems in the data, such as input errors, missing values , and duplication. The steps taken are deleting columns with many missing values , filling in the missing values with the average, median, or numerical data mode, correcting input errors, and identifying and deleting the same entries.
   b. Data Integration: combining data from various sources into a cohesive dataset. Combine different datasets while maintaining consistent data structure and format.
   c. Data Transformation: normalization, namely adjusting the scale of data values to a certain range, such as 0-1, to make comparisons easier.
   d. Data Reduction: data sampling selects a subset of the data representing the entire dataset to reduce the data size while maintaining important statistical characteristics.
2. Data Splitting
   Choosing a good split data composition can help avoid overfitting or underfitting so the classification model can produce better performance on data that has never been seen before.(Oktafiani, Hermawan, & Avianto, 2023) The author divides the data into 80% training and 20% testing data.
3. Building Model
   The k-fold cross-validation method obtains maximum accuracy results from the discriminant analysis function. K fold cross-validation estimates prediction errors in evaluating model performance. (Mardiana, Kusnandar, & Satyahadewi, 2022) Modelling uses the C4.5 Algorithm to classify divorce factors during the pandemic in the West Java area.
4. Particle Swarm Optimization
   Particle Swarm Optimization is a computational method that optimizes a problem by improving candidate solutions concerning certain quality measures iteratively. PSO can solve similar structured issues. (Rizki & Nurlaili, 2021) PSO is used in this research as an optimization to get better accuracy.
5. Model Evaluation
   The PSO-based C4.5 algorithm modelling results will display accuracy and AUC (Area Under Curve) values.

Time and Place of Research
The dataset was taken on the BPS website: https://jabar.bps.go.id/indicator/108/798/1/jumlah-perceraian-menurut-faktor.html. Divorce cases are based on factors that occurred in 2021 in West Java province. In total, there were 98088 divorce cases with 13 factors.

Types of research
The method used in this research is a descriptive quantitative method. The quantitative descriptive research method is a method that aims to create a picture or description of a situation objectively using numbers, starting from data collection, interpretation of the data, as well as the appearance and results. (Arikunto, 2006)

Research Target / Subject
There are three steps to descriptive analysis: formulating the problem, determining the type of information or data, determining data collection procedures, carrying out data processing, and making decisions based on the results of data analysis. Calculation of divorce factor analysis during the pandemic in West Java:

$$P = \frac{F}{N} \; x \; 100\% \dots\dots\dots\dots\dots\dots\dots\dots(1)$$

P = Percentage
F = number of divorces in 1 factor (for example, domestic violence factor, number of divorces 1500; adultery factor, number of divorces 940)
N = number of divorce cases with all factors.

Table 1. Divorce Factor Assessment Table

No	Factor	Interval
1	Economy	25%
2	Leaving one of the parties	20%
3	Domestic violence	12%
4	Constant disputes	11%
5	Sentenced to prison	8%
6	Adultery/Infidelity	7%
7	Gambling	4%
8	Forced marriage	3,4%
9	Polygamy	3,1%
10	Apostate	2,3%
11	Body defects	2,0%
12	Drunk	1,3%
13	Family	1,0%

In Table 1, the results of the assessment of divorce factors during the pandemic in the West Java area, the highest value is the economic factor, then the factor of leaving one of the parties, followed by the domestic violence factor, and so on up to 13 factors.

RESULTS AND DISCUSSION

The C4.5 algorithm forms decision trees, an influential and well-known classification and prediction method.
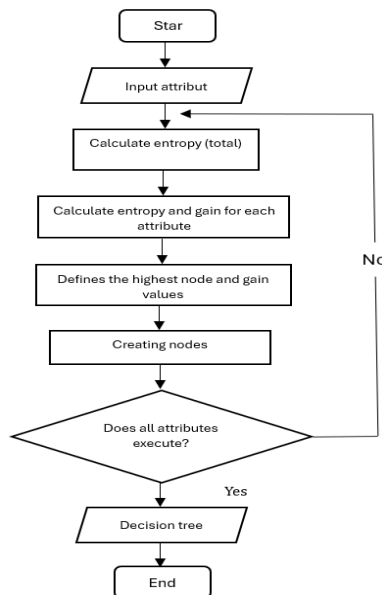

Figure 2. C4.5 algorithm flowchart

The C4.5 algorithm is a classification algorithm that creates prediction models based on categorized data. This algorithm is one of the most popular algorithms in data mining (machine learning) and is often used in industrial applications. (Kantinit, 2023) Figure 2 is a flowchart of how the C4.5 algorithm modelling is carried out.

According to (Muhamad et al. 2019) According to (Muhamad et al., 2019) in general, the C4.5 algorithm for building decision trees is as follows:
1. Select the attribute as root
2. Create a branch for each value
3. Divide cases within branches
4. Repeat the process for each branch until all cases in the branch have the same class.

The results of modelling using the C4.5 algorithm using a dataset of divorce factors in the West Java area during the pandemic are a decision tree shown in Figure 3 below.
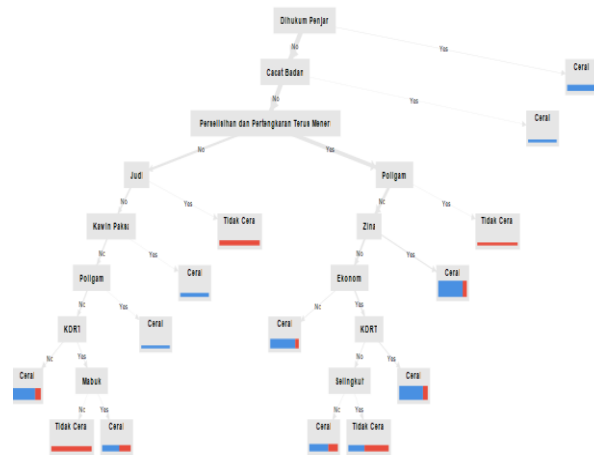

Figure 3. Divorce factors decision tree

Tree

```
Dihukum Penjara = No
|  Cacat Badan = No
|  |  Perselisihan dan Pertengkaran Terus Menerus = No
|  |  |  Judi = No
|  |  |  |  Kawin Paksa = No
|  |  |  |  |  Poligami = No
|  |  |  |  |  |  KDRT = No: Cerai {Cerai=8, Tidak Cerai=2}
|  |  |  |  |  |  KDRT = Yes
|  |  |  |  |  |  |  Mabuk = No: Tidak Cerai {Cerai=0, Tidak Cerai=4}
|  |  |  |  |  |  |  Mabuk = Yes: Cerai {Cerai=3, Tidak Cerai=2}
|  |  |  |  |  Poligami = Yes: Cerai {Cerai=2, Tidak Cerai=0}
|  |  |  |  Kawin Paksa = Yes: Cerai {Cerai=3, Tidak Cerai=0}
|  |  |  Judi = Yes: Tidak Cerai {Cerai=0, Tidak Cerai=4}
|  |  Perselisihan dan Pertengkaran Terus Menerus = Yes
|  |  |  Poligami = No
|  |  |  |  Zina = No
|  |  |  |  |  Ekonomi = No: Cerai {Cerai=7, Tidak Cerai=1}
|  |  |  |  |  Ekonomi = Yes
|  |  |  |  |  |  KDRT = No
|  |  |  |  |  |  |  Selingkuh = No: Cerai {Cerai=4, Tidak Cerai=2}
|  |  |  |  |  |  |  Selingkuh = Yes: Tidak Cerai {Cerai=2, Tidak Cerai=3}
|  |  |  |  |  |  KDRT = Yes: Cerai {Cerai=10, Tidak Cerai=2}
|  |  |  |  Zina = Yes : Cerai {Cerai=12, Tidak Cerai=2}
|  |  |  Poligami = Yes: Tidak Cerai {Cerai=0, Tidak Cerai=2}
|  Cacat Badan = Yes: Cerai {Cerai=2, Tidak Cerai=0}
Dihukum Penjara = Yes: Cerai {Cerai=5, Tidak Cerai=0}
```

Figure 4. Description of divorce factor decision tree

The following is a description of the decision tree in the figure 4 :

- faktor dihukum penjara = No, faktor dihukum penjara = Yes : cerai (cerai = 5, tidak cerai = 0)
- faktor cacat badan = No, faktor cacat badan = Yes : cerai (cerai = 2, tidak cerai = 0)
- faktor perselisihan dan pertengkaran terus menerus = No, faktor perselisihan dan pertengkaran terus menerus = Yes
- faktor judi = No, faktor judi = Yes : tidak cerai (cerai = 0, tidak cerai = 4)
- faktor kawin paksa = No, faktor kawin paksa = Yes : cerai (cerai = 3, tidak cerai = 0)
- faktor poligami = No, faktor poligami = Yes : cerai (cerai = 2, tidak cerai=0)
- faktor KDRT = No : cerai (cerai = 8, tidak cerai = 2)
- faktor KDRT = Yes : (cerai = 10, tidak cerai = 2)
- faktor mabuk = No : tidak cerai (cerai = 0, tidak cerai = 4)
- faktor mabuk = Yes : cerai (cerai = 3, tidak cerai = 2)
- faktor poligami = Yes : cerai (cerai = 2, tidak cerai = 0)
- faktor poligami = Yes : tidak cerai (cerai = 0, tidak cerai = 2)
- faktor kawin paksa = Yes, cerai (cerai = 3, tidak cerai = 0)
- faktor zina = No
- faktor zina = Yes : cerai (cerai = 12, tidak cerai = 2)
- faktor ekonomi = No : cerai (cerai = 7, tidak cerai = 1)
- faktor ekonomi = Yes
- faktor selingkuh = No : cerai (cerai = 4, tidak cerai = 2)
- faktor selingkuh = Yes : tidak cerai (cerai = 2, tidak cerai = 3).

The author uses the RapidMiner framework for modelling the PSO-based C4.5 algorithm. RapidMiner is one of the tools used in data mining. It has approximately 500 data mining operators, including input, output, data preprocessing, and visualization operators. RapidMiner is a stand-alone software for data analysis and a data mining engine that can be integrated into its products. (Maulid, 2023) The PSO-based C4.5 algorithm modelling results are as follows in Figure 5 with an accuracy result of 87.50%, with true positives of 908 records classified as selected divorce and false negatives of 139 records classified as Not Selected Divorce. Next, 59 true negative records were classified as selected Not Divorced, and 15 false negative records were classified as Not Divorced.

Table 2. Modelling accuracy results of the PSO-Based C4.5 Algorithm

Accuracy: 87,50%

	True Cerai	True Tidak Cerai	Class Precision
Pred. Cerai	908	139	85%
Pred Tidak Cerai	59	15	79,73%
Class recall	96.54%	32,14%	

The following Figure 5 is the AUC (Area Under Curve) modelling result of the PSO-based C4.5 algorithm with a result of 0.807 and is included in the good classification. AUC (area under the curve) calculation calculates and confirms which classifier is superior. AUC (area under the curve) is the area under the curve. laThe wide of AUC is always between the values 0 and 1. AUC is calculated based on the combined area of the trapezoidal points (sensitivity and specificity). (Qadrini, Seppewali, & Aina, 2021)
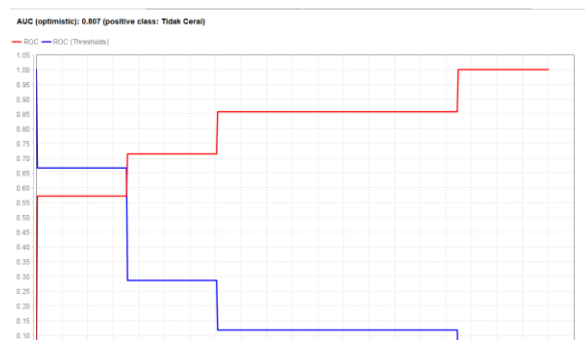

Figure 5. Area Under Curvenani

CONCLUSIONS AND SUGGESTIONS

Conclusion

This research aims to classify divorce factors during the 2021 pandemic in West Java Province using the PSO-based C4.5 algorithm so that the government can make a policy in the future if an outbreak occurs, especially from an economic and psychological perspective, because it can have an impact on household problems. The research results obtained an accuracy figure of 87.50% with an AUC value of 0.807 and was included in the good classification. Thus, the C4.5 algorithm is good for classifying mining data, and PSO is a good optimization method for increasing accuracy values.

Suggestion

For further research development, you can classify in other regions of Indonesia, use different algorithms, and use datasets with broader attributes to improve the results.

REFERENCES

.Admaja, M. D., Hidayah, N., & Suparwi. (2021). Faktor-Faktor Penyebab Perceraian Selama Masa Pandemi Covid-19 di Pengadilan Agama Sukoharjo. *Jurnal Penelitian Serambi Hukum*, *14*(2), 33–41. Retrieved from http://journal.uniba.ac.id/index.php/SH/article/view/705

Arikunto, S. (2006). *Prosedur Penelitian Suatu Pendekatan Praktek* (Revisi VI). Jakarta: Rineka Cipta.

Feldman, R., & James, S. (2007). *The Text Mining Handbook* (Cambridge, Ed.). New York.

Iriadi, N., & Nuraeni, N. (2016). Kajian Penerapan Metode Klasifikasi Data Kelayakan Kredit Pada Bank. *Jurnal Teknik Komputer AMIK BSI*, *II*(1), 132–137. Retrieved from https://ejournal.bsi.ac.id/ejurnal/index.php/jtk/article/view/371/

Kantinit. (2023). Algoritma C4.5: Pengertian, Cara kerja dan Contoh Implementasi. Retrieved from KantinIT website: https://kantinit.com/kecerdasan-buatan/algoritma-c4-5-pengertian-cara-kerja-dan-contoh-implementasi/

Mardiana, L., Kusnandar, D., & Satyahadewi, N. (2022). Analisis Diskriminan Dengan K Fold Cross Validation Untuk Klasifikasi Kualitas Air Di Kota Pontianak. *Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster)*, *11*(1), 97–102. Retrieved from https://jurnal.untan.ac.id/index.php/jbmstr/article/view/51608

Maulid, R. (2023). RapidMiner : Tools Data Science Andalan Data Experts. Retrieved from DQLab website: https://dqlab.id/rapidminer--tools-data-science-andalan-data-experts

Muhamad, M., Windarto, A. P., & Suhada, S. (2019). Penerapan Algoritma C4.5 Pada Klasifikasi Potensi Siswa Drop Out. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1), 1–8. https://doi.org/10.30865/komik.v3i1.1688

Oktafiani, R., Hermawan, A., & Avianto, D. (2023). Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning. *Jurnal Sains Dan Informatika*, *9*(April), 19–28. https://doi.org/10.34128/jsi.v9i1.622

Prihartono, W., Tohidi, E., Ahmad Fauzi, I., & Danar Dana, R. (2023). Implementasi Data Mining Clustering Dalam Mengelompokan Kasus Perceraian Yang Terjadi Di Provinsi Jawa Barat Menggunakan Algoritma K-Means. *Kopertip : Jurnal Ilmiah Manajemen Informatika Dan Komputer*, *7*(3), 64–70. https://doi.org/10.32485/kopertip.v7i3.328

Qadrini, L., Seppewali, A., & Aina, A. (2021). Decision Tree dan Adaboost pada Klasifikasi Penerima Program Bantuan Sosial. *Jurnal Inovasi Penelitian*, *2*(7), 1959–1966. https://doi.org/10.47492/JIP.V2I7.1046

Rizki, A. M., & Nurlaili, A. L. (2021). Algoritme Particle Swarm Optimization (PSO) untuk Optimasi Perencanaan Produksi Agregat Multi-Site pada Industri Tekstil Rumahan. *Journal of Computer, Electronic, and Telecommunication*, *1*(2), 1–9. https://doi.org/10.52435/complete.v1i2.73

Sari, D. L., & Gemasih, H. (2021). *Penerapan Data Mining Dalam Proses Prediksi Perceraian Menggunakan Algoritma Naive Bayes di Kabupaten Aceh Tengah*. *4*(1), 23–35. https://doi.org/10.14716/ijtech.v0i0.0000

Sugiyono. (2018). *Metode Penelitian Kuantitatif* (Bandung, Ed.). Bandung: Alfabeta.

Sulistiyanto, S. (2018). Penerapan C4.5 Berbasis Particle Swarm Optimization (PSO) dalam Memprediksi Siswa Lolos Seleksi Perguruan Tinggi. *Prosiding Seminar Nasional Darmajaya*, *1*(1), 162–170. Retrieved from https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/1242

Syarif, M. R. R., Purnamasari, I., & Mayasari, R. (2021). Algoritma Fp-Growth untuk Mengetahui Pola Perceraian Selama Masa Pandemi COVID-19 di Kab. Majalengka. *Journal of Information Technology and Computer Science (INTECOMS)*, *4*(2). Retrieved from https://journal.ipm2kpe.or.id/index.php/INTECOM/article/view/2782

Yasifah, N. A. N. N., Azahra, B. A., Ridha, Z. S., Nurhasanah, H., Amaliah, N., Setianing, I. D., & Mukarromah, S. (2023). Faktor-Faktor Penyebab Meningkatnya Perceraian Pada Masa Covid-19. *Jurnal Yustitia*, *24*(1). https://doi.org/10.53712/yustitia.v24i1.1969