Online Version Available at: http://bsti.ubd.ac.id/e-jurnal

JOURNAL OF TECH-E

|2581-1916 (Online) | 2598-7585 (Printed) |

Article

Performance Analysis of Classification and Regression Tree (CART) Algorithm in Classifying Male Fertility Levels with Mobile-Based

Arif Rahman Hakim[1], Dewi Marini Umi Atmaja[2], Amat Basri[3], Andri Ariyanto[4]

[1, 2] Medika Suherman University, Digital Business, Jawa Barat, Indonesia
[3] Buddhi Dharma University, Information Systems, Banten, Indonesia
[4] Jenderal Achmad Yani University, Informatics, Jawa Barat, Indonesia

SUBMISSION TRACK

Recieved: May 16, 2023
Final Revision: July 17, 2023
Available Online: August 28, 2023

KEYWORD

Android, CART, Classification, Fertility, K-Fold Cross Validation

CORRESPONDENCE

E-mail: arif@medikasuherman.ac.id

A B S T R A C T

Fertility is the ability to produce offspring in a man or the ability of the reproductive organs to work optimally in fertilization. Fertility rates have declined drastically in the last fifty years. Machine Learning is a field devoted to understanding and building learning methods. This study will use machine learning algorithms to classify male fertility levels, namely the CART algorithm and the K-Fold Cross Validation validation method. The fertility dataset used in this study was obtained from the UCI Machine Learning website, with a total of 100 data and the variables used are Age, Childish diseases, Accident or serious trauma, Surgical intervention, High fevers in the last year, Frequency of alcohol consumption, Smoking habit, Number of hours spent sitting per day and Diagnosis. K-Fold Cross Validation can be used together with CART to measure the performance of the CART model on different data, so as to avoid overfitting or underfitting the CART model. Based on the calculation of the CART algorithm and the K-Fold Cross Validation validation method (K = 1 to K = 9), the average accuracy value for training data is 98.70% and the average accuracy value for testing data is 81.16%. The results of this study have proven that the CART algorithm can be used to classify the level of fertility in men well. In addition, the classification model formed can be implemented into a mobile application (android) so that it is easy to use and understand.

INTRODUCTION

Fertility is the ability to produce offspring in a man or the ability of the reproductive organs to work optimally in fertilization. The human reproductive system is very complex, so when pregnancy begins, ovulation and fertilization must occur properly and correctly [1]. Birth rates have been declining in many countries. As in European countries, the birth rate has fallen over the past 50 years [2].

The problem of infertility (infertility) is often directed at the woman, even though men are also very likely to be infertile. If a man experiences infertility, it will be difficult to get offspring. [1]. Several factors can affect male infertility, including environmental exposures and unhealthy or uncontrolled lifestyles, such as smoking, regular alcohol consumption, and time spent sitting each day. Other factors that affect male fertility include age, childhood diseases, accidents or trauma, and surgery. [3].

One way that can be used to overcome fertility problems is the use of machine learning techniques (Siradjuddin, 2020). Infertility can be prevented early by creating a system that can check male fertility as mentioned earlier. [4].

Classification and Regression Tree (CART) Algorithm is one of the algorithms in machine learning that can classify male fertility. This algorithm works like a tree consisting of leaves and branches that are used to make decisions [5]. In classification trees, the target variable is a discrete or categorical variable, such as a class label in classification. Whereas in regression trees, the target variable is a continuous-valued variable [6].

I. LITERATURES REVIEW

A similar study on the classification of male fertility quality has previously been conducted by [1] using an Artificial Neural Network (ANN) algorithm, which achieved the highest accuracy of 88.51%. The difference between the previous study and this research lies in the utilization of the CART algorithm and the K-Fold Cross Validation validation method.

The ultimate goal of this study is to develop an application that can be used to classify male fertility using the CART algorithm and the K-Fold Cross Validation validation method, with the hope of assisting medical professionals in making decisions for patients consulting about pregnancy programs.

Furthermore, the CART algorithm is a data analysis method used for segmentation and modeling [5]. In this method, a decision tree is constructed based on rules that categorize data into different categories. The selection of relevant features or attributes is crucial in forming the decision tree, enabling accurate classification.

The K-Fold Cross Validation validation method is used to measure the performance of the model developed in this study. The use of this method helps address issues of overfitting or underfitting that may occur in the classification model. In the K-Fold Cross Validation method, the data is divided into k overlapping subsets. The model is trained and tested multiple times using different subsets, and the test results from each subset are averaged. This provides a more objective assessment of the model's performance.

In the context of the expected application that can be developed from this research, the model generated from the CART algorithm with the K-Fold Cross Validation method is expected to provide predictions about the quality of male fertility based on specific attributes used as input. This application is anticipated to be used by medical professionals to support decision-making regarding pregnancy programs by providing more objective and accurate information.

II. FRAMEWORK

The dataset used in this research is fertility data sourced from the UCI Machine Learning Repository website with a total of 100 data. The framework used in this research includes 4 stages, including Preparation Stage, Data Collection, Machine Learning Model, Implementation Model in Android Application. More detailed information about the research method can be seen in Figure 1.

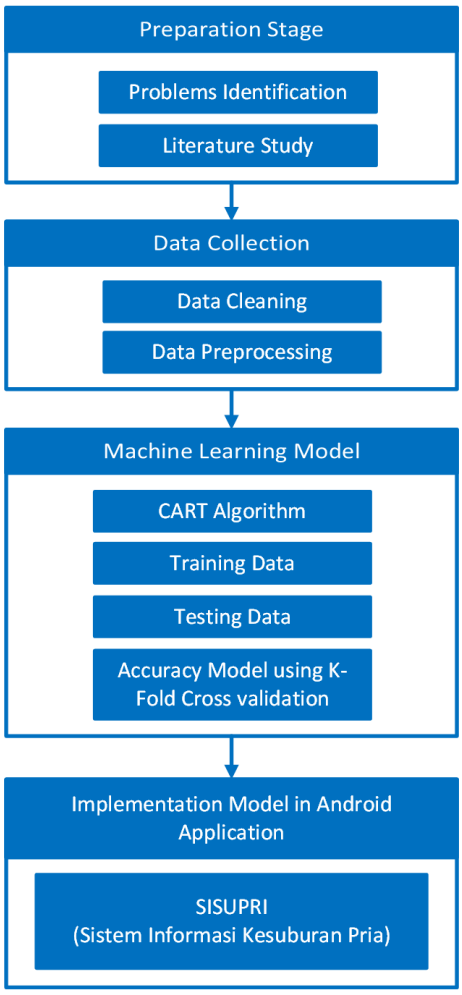

Figure 1. Research Flow Chart

III. METHODS

Based on Figure 1, the first stage carried out in this research is preparation, at this stage the steps taken are identifying problems and studying literature such as books, journals related to the theoretical basis to support this research. [7][8].

The next step taken in this study is to collect datasets, the dataset used in this study is fertility data obtained from UCI Machine Learning. The data used amounted to 100 data. The classes used in this study are Normal and Altered. The data used in this study can be seen in Table 1 and Table 2 below.

Table 1. Fertility Dataset

No.	Season	Age	Childish diseases	Accident or serious trauma	Surgical intervention
(0)	(1)	(2)	(3)	(4)	(5)
1	spring	30	no	yes	yes
2	spring	35	yes	no	yes
3	spring	27	yes	no	no
4	spring	32	no	yes	yes
5	spring	30	yes	yes	no
…	…	…	…	…	…
100	winter	30	no	yes	yes

Table 2. Fertility Dataset (Continued)

No.	High fevers in the last year	Freq of alcohol consumption	Smoking habit	Number of hours spent sitting per day	Diagnosis
(0)	(6)	(7)	(8)	(9)	(10)
1	more than 3 months ago	once a week	occasional	16	Normal
2	more than 3 months ago	once a week	daily	6	Altered
3	more than 3 months ago	hardly ever or never	never	9	Normal
4	more than 3 months ago	hardly ever or never	never	7	Normal
5	more than 3 months ago	once a week	never	9	Altered
…	…	…	…	…	…
100	more than 3 months ago	several times a week	never	3	Normal

After obtaining the Fertility dataset, the next stage is data cleaning and data pre-processing. The dataset has a variety of attribute ranges, so normalization will be carried out on several attributes in the dataset. The input variables that will be used can be described in Table 3 below.

Table 3. Input Data

No.	Variable	Unit Value	Description
1.	Season	-1 = winter -0.33 = spring 0.33 = summer 1 = fall	Season
2.	Age	18-36 years old	Age18-36
3.	Childish diseases	0 = yes 1 = no	Infectious diseases, for example: Measles, chickenpox, mumps, etc.
4.	Accident or serious trauma	0 = yes 1 = no	Severe trauma accident
5.	Surgical intervention	0 = yes 1 = no	History of surgical operation
6.	High fevers in the last year	-1 = last 3 months 0 = more than the last 3 months 1 = no	Febrile fever in the past year
7.	Frequency of alcohol consumption	1 = several times per day 2 = every day 3 = several times per week 4 = once a week 5 = rarely or never	Large amount of alcohol consumption
8.	Smoking habit	-1 = never 0 = sometimes 1 = every day	The habit of smoking
9.	Number of hours spent sitting per day	1-16 hours	Length of sitting duration per day 0-24 hours

After defining the input data, the CART algorithm must also define the output class. The desired output category in this study is the male fertility diagnostic result category shown in Table 4 below.

Table 4. Fertility Status Class

Unit Value	Description
1	Normal
0	Altered

The next step that must be done is the data normalization process, in data normalization, it is determined that the data range varies between 0 and 1, so that it will be easier to complete the classification steps. The output class is symbolized by 1 for the Normal class, while 0 for the Altered class. Data that has been normalized will then be used in this research.
After the data pre-processing stage is complete, the next stage is model building using the CART algorithm. At this stage the data will be processed in accordance with the CART algorithm flow and divided into training data and testing data. The next step is to calculate the accuracy of the model, the method used for the model accuracy calculation process of the CART algorithm is the K-fold Cross Validation method. The final step in this model building stage is to obtain the accuracy value of the CART algorithm in a data classification model that will later be used in making applications. After the preparation, data collection, and model building stages are complete, the last step is to implement the data classification model into an android application with the application name SISUPRI (Sistem Informasi Kesuburan Pria / Male Fertility Information System), then test the application functionally.

Data Analysis Technique

The K-fold Cross Validation method performed in this study is 8 times with K = 2, K = 3, K = 4, K = 5, K = 6, K = 7, K = 8 and K = 9. As an example of the K-fold Cross Validation stage with a value of K = 9 can be illustrated as Table 5 below.

Table 5. Illustration of K-Fold Cross Validation Stages

Eksperiment to-	Training Data	Testing Data
1	K2, K3, K4, K5, K6, K7, K8, K9	K1
2	K1, K3, K4, K5, K6, K7, K8, K9	K2
3	K1, K2, K4, K5, K6, K7, K8, K9	K3
4	K1, K2, K3, K5, K6, K7, K8, K9	K4
5	K1, K2, K3, K4, K6, K7, K8, K9	K5
6	K1, K2, K3, K4, K5, K7, K8, K9	K6
7	K1, K2, K3, K4, K5, K6, K8, K9	K7
8	K1, K2, K3, K4, K5, K6, K7, K9	K8
9	K1, K2, K3, K4, K5, K6, K7, K8	K9

From Table 5 above, if K=9 the data is divided into 9 groups with 8 parts for training data and 1 part for test data.

IV. RESULT

Data Processing Using CART

At the data processing stage using the CART algorithm, researchers began coding the algorithm using the python programming language as shown in Figure 2 below.

```
[ ]: df = pd.read_csv('fertility.csv')

[ ]: df.head()
```

Figure 2. Code for reading datasets

Figure 2 is the code to read the csv file that will be used as a dataset. In the code above, the program will open the fertility.csv file and will save it into a df variable that will be used in the next process.

Model Classifier (Normalization)

```
[ ]: def minmaxscaler(x, var_min, var_max):
         x_std = (x - var_min) / (var_max - var_min)
         return x_std

[ ]: df_normalize = df.copy()
     df_normalize = df_normalize.drop(["Diagnosis"], axis = 1)
     df_normalize['Age'] = df_normalize.apply(lambda x: minmaxscaler(x['Age'], 18, 36), axis=1)
     df_normalize['Frequency of alcohol consumption'] = df_normalize.apply(lambda x: minmaxscaler(x['Frequency of alcc
     df_normalize['Number of hours spent sitting per day'] = df_normalize.apply(lambda x: minmaxscaler(x['Number of hc

[ ]: # Variabel independen
     X = df_normalize
     # Variabel dependen
     y = df["Diagnosis"]

[ ]: # pisahkan data menjadi data lati dan data uji dari keseluruhan data
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
     print("total data train = %d" %len(X_train)) # menampilkan banyak data latih
     print("total data test = %d" %len(X_test)) # menampilkan banyak data uji
```

Figure 3. Code for Normalization Min-Max Data

Figure 3 is a code snippet used for data normalization. The coding function uses min-max normalization.

Modelling

```
[ ]: tree_clf = DecisionTreeClassifier()
     path = tree_clf.cost_complexity_pruning_path(X_train, y_train)
```

Figure 4. Model Building Code

Figure 4 is the code used for the formation of classification models using Decision Tree, the algorithm used by default is CART.

Accuracy Results Using K-Fold Cross Validation

After the data is processed using the CART algorithm, an accuracy calculation is then carried out using K-fold Cross Validation

[9][10]. The accuracy calculation result is considered good or accurate if the accuracy result is close to or equal to 1, otherwise the algorithm is considered bad or inaccurate if the accuracy result is close to zero (0). In this study, 8 (eight) K-fold Cross Validation processes were carried out with K=2, K=3, K=4, K=5, K=6, K=7, K=8 and K=9.

To get the accuracy value, the formula is used:

$$Accuracy = \frac{Number\ of\ Corresponding\ Data}{Total\ Data} \quad (1)$$

Meanwhile, to get the average accuracy value, the formula is used:

$$Average\ Accuracy = \frac{Total\ Number\ of\ Accuracy}{Number\ of\ Iterations} \quad (2)$$

V. DISCUSSION

Based on the calculation results using formulas (1) and (2), the average accuracy of the CART algorithm using the K-fold Cross Validation method is obtained as much as K=2, K=3, K=4, K=5, K=6, K=7, K=8 and K=9 for training data. The following Table 6 is a comparison of the accuracy value of the CART algorithm for training data

Table 6. Comparison of Accuracy Value of CART Algorithm for Training Data

Number of K	Average Accuracy of Training Data (%)
K=2	100.00
K=3	99.29
K=4	99.52
K=5	98.93
K=6	98.57
K=7	98.10
K=8	98.37
K=9	96.79
Overall Average	98.70

Table 6 is a comparison of the average accuracy results of the CART algorithm using the K-fold Cross Validation method for training data. At the value of K=2, the average accuracy value is 100.00%, for K=3, the average accuracy value is 99.29%, for K=4, the average accuracy value is 99.52%, for K=5, the average accuracy value is 98.93%, for K=6, the average accuracy value is 98.57%, for K=7, the average accuracy value is 98.10%, while when the value of K=8, the average accuracy value is 98.37%, finally, when the value of K=9, the average accuracy value is 96.79%. Based on these results, the performance of the CART algorithm has an average accuracy value of 98.70%. While the highest average accuracy value is when the value of K=2 is 100.00%.

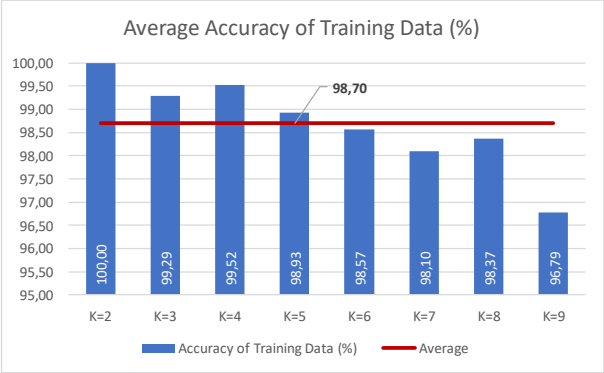

Figure 5. Graph of Average Accuracy of Training Data

Figure 5 is a graph of the average accuracy of the CART algorithm using the K-fold Cross Validation method for training data in units of percent, the blue bar graph is the average accuracy of training data for each K, while the red line graph is the overall average of K-fold Cross Validation (K=1 to K=9) with an average accuracy value for training data of 98.70%.
While the average accuracy of the CART algorithm using the K-fold Cross Validation method of K=2, K=3, K=4, K=5, K=6, K=7, K=8 and K=9 for testing data can be seen in Table 7 below.

Table 7. Comparison of Accuracy Value of CART Algorithm for Testing Data

Number of K	Average Accuracy of Testing Data (%)
K=2	80.00
K=3	79.95
K=4	72.63
K=5	84.29
K=6	82.83
K=7	82.86
K=8	84.03
K=9	82.74
Overall Average	81.16

Table 7 is a comparison of the average accuracy results of the CART algorithm using the K-fold Cross Validation method for testing data. At the value of K=2, the average accuracy value is 80.00%, for K=3, the average accuracy value is 79.95%, for K=4, the average accuracy value is 72.63%, for K=5, the average accuracy value is 84.29%, for K=6, the average accuracy value is 82.83%, for K=7, the average accuracy value is 82.86%, while when the value of K=8, the average accuracy value is 84.03%, finally, when the value of K=9, the average accuracy value is 82.74%. Based on these results, the performance of the CART algorithm has an average accuracy value of 81.16%. While the highest average accuracy value is when the K=5 value is 84.29%.

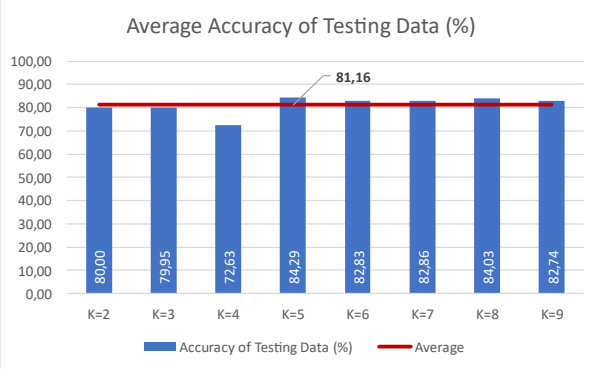

Figure 6. Graph of Average Accuracy of Testing Data

Figure 6 is a graph of the average accuracy of the CART algorithm using the K-fold Cross Validation method for testing data in units of percent, the blue bar graph is the average accuracy of testing data for each K, while the red line graph is the overall average of K-fold Cross Validation (K=1 to K=9) with an average accuracy value for testing data of 81.16%.

Based on the above discussion, it is evident that CART has proven to be a powerful tool for nonparametric data analysis and prediction. This method can overcome complexity in data by dividing variables into mutually exclusive subgroups [11]. In addition, CART is a flexible and easy-to-interpret method, making it suitable for modeling in various fields such as health, medicine, finance, and social sciences [12].

Despite its ability to handle complex data, CART has a disadvantage in handling data that has many variables, as it can produce trees that are very complex and difficult to interpret [13]. In addition, CART tends to overfitting the training data, especially when the tree grows very deep [14].

Therefore, to reduce the risk of overfitting, this research uses K-Fold Cross Validation in order to provide a better estimate of the model's performance in various data situations, as it involves testing on several different subsets of test data. K-Fold Cross Validation is an effective method to measure model performance without wasting valuable data. It helps reduce the risk of overfitting and provides a more stable estimate of a model's predictive ability [15].

Model Implementation on Android

After the research is conducted and the research results are obtained in the form of accuracy values, the CART algorithm can be used to classify male fertility. The following is a display of the Android application using a machine learning model that has previously been formed with the CART algorithm and the K-fold Cross Validation method with a value of K=5 to classify or predict male fertility.

Figure 7. Splash Screen Page

Figure 8. Main Menu Page

Figure 7 is a splash screen page that appears when the user first opens the application, the

splash screen display will appear for a few seconds with the words SISUPRI (Sistem Informasi Kesuburan Pria / Male Fertility Information System) before entering the application's home page. After the splash screen disappears, the home page will appear as in Figure 8. The home page contains four menus including the Prediction menu, About Fertility menu, About Application menu and Exit menu.

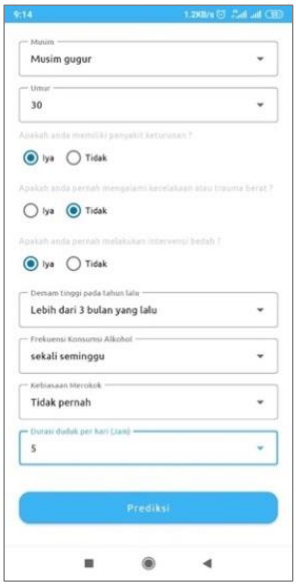

Figure 9. Prediction Form Page

Figure 10. Prediction Results Page

Figure 9 is a display form for predicting male fertility (fertility) which consists of several attributes that must be filled in, including season, age, hereditary diseases, accidents or severe trauma, surgical interventions, fever in the last year, frequency of alcohol consumption, smoking habits and duration of sitting per day. Based on the data covered by the user, the system will display the prediction of male fertility (fertility) based on the calculation of the CART algorithm as shown in Figure 10.

Figure 11. About Fertility Page

Figure 11 is an explanation page about the definition of fertility in men.

Figure 12. Factors affecting fertility page

Figure 13. How to increase fertility page

Figure 12 is a page explaining the factors that affect male fertility, while Figure 13 contains an explanation of ways to improve male fertility.

Figure 14. About Application Page

Figure 14 is a page about the application that contains an explanation of the creation of the application, the source of the dataset, the algorithm used and the accuracy results obtained.

Figure 15. App exit pop up page

Figure 16. Splash Screen Page When Exiting the App

Figure 15 is a pop up page that displays a question if the user wants to close the SISUPRI application. While Figure 16 is a splash screen page when the user chooses to close / exit the application.

VI. CONCLUSION

Based on the discussion described above, it can be concluded that the performance of the CART algorithm using the K-fold Cross Validation test method with the value of K=2, K=3, K=4, K=5, K=6, K=7, K=8 and K=9 obtained the best accuracy value at the value of K=5 with an average accuracy value for training data of 98.93% and an average accuracy value for testing data of 84.29%. The overall average of K-fold Cross Validation (K=2 to K=9) obtained an average accuracy value for training data of 98.70% and for testing data obtained an average accuracy value of 81.16%. CART algorithm has been able to classify the level of fertility in men and can be implemented through android applications.

REFERENCES

[1] E. Budianita, F. R. Hustianto, F. Syafria, and M. Nasir, "Implementasi Algoritma Jaringan Syaraf Tiruan (JST) Hopfield untuk Klasifikasi Kualitas Kesuburan Pria," *Semin. Nas. Teknol. Informasi, Komun. dan Ind.*, no. November, pp. 137–142, 2018.

[2] S. W. A. Adi, "Komparasi Metode Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Dan Random Forest (RF) Untuk Prediksi Penyakit Gagal Jantung," *MATHunesa J. Ilm. Mat.*, vol. 9, no. 2, pp. 437–446, 2021.

[3] D. C. P. Buani, "Optimasi Algoritma Naïve Bayes dengan Menggunakan Algoritma Genetika untuk Prediksi Kesuburan (Fertility)," *Rev. Bras. Ergon.*, vol. 3, no. 2, pp. 80–91, 2016.

[4] H. Harafani and A. Maulana, "Penerapan Algoritma Genetika pada Support Vector Machine Sebagai Pengoptimasi Parameter untuk Memprediksi Kesuburan," *J. Tek. Inform. STMIK Antar Bangsa*, vol. V, no. 1, pp. 51–59, 2019.

[5] S. Mata, P. Di, and E. S. Pare, "Pemanfaatan Classification and Regression Trees (Cart) Untuk Memprediksi Kelulusan Siswa Pada," pp. 16–23, 2011.

[6] A. Purnamawati, M. N. Winnarto, and M. Mailasari, "Analisis Cart (Classification and Regression Trees) Untuk Prediksi Pengguna Sepeda Berdasarkan Cuaca," *J. Teknoinfo*, vol. 16, no. 1, p. 14, 2022, doi: 10.33365/jti.v16i1.1478.

[7] A. R. Hakim, D. Marini, U. Atmaja, D. Haryadi, and N. Suwaryo, "Twitter Sentiment Analysis Terhadap Pengguna E-Commerce Menggunakan Text Mining," *SNTEM Semin. Nas. Teknol. Energi dan Miner.*, vol. 1, no. November, pp. 1227–1237, 2021.

[8] A. Perdana, M. Tanzil Furqon, and Indiriati, "Penerapan Algoritma Support Vector Machine (SVM) Pada Pengklasifikasian Penyakit Kejiwaan Skizofrenia (Studi Kasus: RSJ. Radjiman Wediodiningrat, Lawang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 3162–3167, 2018, [Online]. Available: http://j-ptiik.ub.ac.id.

[9] H. S. Wafa *et al.*, "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)," *Informatics Digit. Expert*, vol. 1, pp. 40–45, 2022.

[10] E. Carlsen, A. Giwercman, N. Keiding, and N. E. Skakkebaek, "Evidence for decreasing quality of semen during past 50 years," *Obstet. Gynecol. Surv.*, vol. 48, no. 3, pp. 200–202, 1993, doi: 10.1097/00006254-199303000-00023.

[11] A. L. S. Chemex *et al.*, *Classification and Regression Trees by Leo Breiman*, no. January. 1999.

[12] M. Kuhn and K. Johnson, *Applied predictive modeling*. 2013.

[13] T. Hastie, R. Tibshirani, G. James, and D. Witten, *An Introduction to Statistical Learning, Springer Texts*, vol. 102. 2006.

[14] T. et. all. Hastie, "Springer Series in Statistics The Elements of Statistical Learning," *Math. Intell.*, vol. 27, no. 2, pp. 83–85, 2009, [Online]. Available: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf.

[15] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2019, O'reilly)*. O'Reilly Media, 2017.

BIOGRAPHY

Arif Rahman Hakim, Graduated from the Informatics Engineering study program at Pelita Bangsa University in 2017 (S1), and graduated from the Informatics Engineering Masters at the President University in 2020. Currently working as a lecturer at Medika Suherman University.

Dewi Marini Umi Atmaja, Graduated from Informatics Study Program, Jenderal Achmad Yani University in 2018 (S1), and graduated from the Informatics Engineering Masters at the President University in 2020. Currently working as a lecturer at Medika Suherman University.

Amat Basri, Graduated in the Information Technology Study Program (S1) in April, 2002, continued his Masters in Information Systems in Januari, 2015 and graduated in September, 2016. Currently, I am a lecturer in information systems program at Buddhi Dharma University.

Andri Ariyanto, Graduated from the Informatics Study Program at Jenderal Achmad Yani University in 2018 (S1).