Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.11967
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Forward Selection Attribute Reduction Technique for Optimizing Naïve Bayes Performance in Sperm Fertility Prediction

Ahmadi Irmansyah Lubis[1)*], Rudy Chandra[2)]
[1)]Politeknik Negeri Batam, Batam, Indonesia, [2)]Institut Teknologi Del, Sumatera Utara, Indonesia
[1)]ahmadi@polibatam.ac.id, [2)]rudy.chandra@del.ac.id

Submitted : Dec 4, 2022 | **Accepted** : Dec 23, 2022 | **Published** : Jan 1, 2023

Abstract: The problem of infertility between husband and wife is an important issue that destroys family harmony, and many people still consider infertility or infertility a female problem. However, about 7% of men of childbearing age suffer from infertility. The biggest factor causing male infertility is sperm quality problems. Sperm analysis can be the best predictor of male fertility potential. Machine learning and data mining techniques can be used to automate disease diagnosis. This study aims to obtain a regular form classification model from sperm sample data of 100 volunteers. This classification model can be used to predict male fertility levels into 2 classes, namely normal and alter (decreased fertility). This study uses a fertility dataset obtained from the UCI Machine Learning Repository. Before the data mining process, data preprocessing is required. The classification process is carried out using Naive Bayes and attribute reduction techniques using forward selection to see the increase in the accuracy of Naive Bayes performance. The Naive Bayes test without attribute reduction has an accuracy of 85%, while attribute reduction with forward selection has an accuracy of 88% in predicting sperm fertility. Therefore, by using forward selection with Naive Bayes to reduce attributes in this study, this study was able to increase accuracy by 3% and can be used to help predict sperm fertility.

Keywords: Fertility; Classification; Naïve Bayes; Attribute Reduction; Forward Selection

INTRODUCTION

Heredity is something that is very desirable for every married couple, so the level of male fertility is an important factor. In general, the male reproductive period is strongly influenced by the male's age (Khaira et al., 2020). When a man enters his 40s, the ability of his sperm to fertilize an egg decreases dramatically, and by the age of 55 the quality of his sperm will be very bad. The reason why it is said that 55 years old is the worst condition for a man is because the older the man, the weaker the sperm motility, and the harder it is for the sperm to reach the egg, so that in recent decades, male fertility has become a problem in the health department. In Indonesia, the infertility rate for married couples who have difficulty having children is around 10%. (Yepriyanto et al., 2014).

According to WHO research results, 50% of the causes of infertility are men, and the rest are influenced by sperm. Factors that affect the level of fertility itself, such as hormones, congenital diseases, and whether you have had surgery (Budianita et al., 2018). Other factors that affect fertility are unhealthy lifestyles such as smoking, alcohol consumption, obesity and a sedentary lifestyle. Smoking can damage a person's reproductive system and reduce their chances of having healthy children, according to a study conducted by HealthDay News (Arifin, 2020). Smoking also reduces the quality and quantity of human sperm (concentration, motility and sperm morphology). (Ma'mur, 2019).

One way to analyze male fertility is to use data mining techniques (Lubis et al., 2020). Data mining is data filtering by obtaining information from data using sizable data sets at various stages of the process. Data mining can be used for classification or prediction. Within the classification itself are the target categorical variables. A method or model developed by a researcher to solve a classification case. Predicting class labels and classifying data classification in data mining is to classify attributes and new data based on training data and class label values (Utomo et al., 2020).

*Ahmadi Irmansyah Lubis

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.11967
e-ISSN : 2541-2019
p-ISSN : 2541-044X

One of the algorithms that can be used for classification is Naive Bayes. Overall, Naive Bayes performs well compared to other classifiers because of its simplicity, low time complexity, small memory requirement and high prediction accuracy (Susanto, 2013). The weakness of Naive Bayes is that the prediction of running probability is not optimal, and features that are relevant to classification are not selected, so the accuracy is low. This can be overcome by selecting useful features to improve accuracy. To maximize the classification results to be obtained, the accuracy value is increased by attribute reduction (Noviati et al., 2015) using forward selection. Forward selection is a technique for reducing the dimensions of a dataset by removing irrelevant or redundant attributes. Forward selection is starting the model with zero variables and then entering the variables one by one until certain conditions are met (Laksono, 2008).

In this study, the authors will perform Naive Bayesian optimization using forward selection to classify male fertility using the dataset used, namely the Fertility dataset taken from the UCI Machine Learning Repository. The purpose of this study was to determine the accuracy of Naive Bayes and to determine the performance improvement of Naive Bayes after optimizing using forward selection in sperm fertility classification.

LITERATURE REVIEW

Research by (Lubis et al., 2021) researched on decision support systems in determining the level of sperm concentration with the Fertility Dataset using the MOORA method with the results obtained, namely the MOORA method was able to rank sperm fertility level concentration data. Research by (Lubis & Setiawan, 2022) researched on decision support systems in determining sperm concentration levels with the Fertility Dataset using ELECTRE and MOORA with the results obtained that ELECTRE is superior to MOORA in terms of execution time. Research by (Nurelasari, 2018) discusses the comparison of Naïve Bayes SVM and PSO in predicting fertility with the results obtained namely Naïve Bayes with an accuracy of 85% and SVM+PSO with an accuracy of 88%. Research by (Ubaedi & Djaksana, 2022) examined the C4.5 and Forward Selection methods in the classification of credit worthiness with the results obtained that C4.5 had an accuracy of 79.11% and the Feature Selection method succeeded in increasing the accuracy of C4.5 by 9.2% in predicting creditworthiness. Research by (Hermawanti & Safriandono, 2016) melakukan kombinasi K-Nearest Neighbor dan Forward Selection untuk the diagnosis of diabetes with the results obtained were that k-NN obtained an accuracy of 95.29% and K-NN with Forward Selection obtained an accuracy of 96.08%.

METHOD

This study uses experimental research methods and quantitative methods. This research was conducted according to the stages in Figure 1.

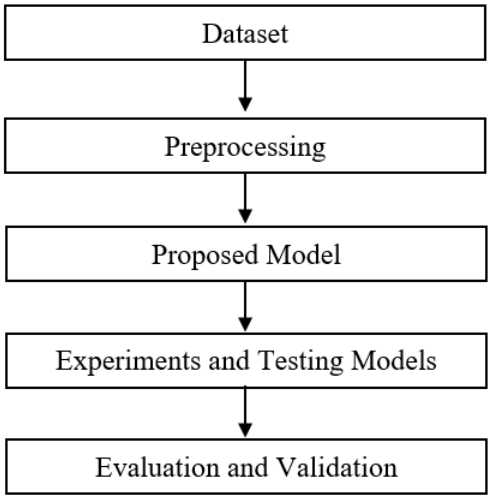

Fig 1. Framework of Research

Dataset

The dataset used in this study is the Fertility Dataset. The Fertility Dataset is the result of research conducted by David Gil, Jose Luis Girela, Joaquin De Juan, M. Jose Gomez-Torres, Magnus Johnsson in 2012. This dataset is published through the UCI Learning Machine Repository website. This Fertility Dataset consists of 100 records with 10 attributes. This dataset label consists of 2 classes, including N (normal) and O (altered). Where 88 records are in class N (normal) and 12 classes are in class O (altered).

*Ahmadi Irmansyah Lubis

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.11967
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1. Fertility Dataset Attribute Information

Attribute	Criteria	Value
The season in which the analysis was carried out	1) winter, 2) spring, 3) summer, 4) autumn	(-1, -0,33, 0.33, 1)
Age at time of analysis	18-36 years	Scale [0-1]
Childhood diseases (chickenpox, measles, mumps, polio)	Yes No	(0,1)
Serious accident or trauma	Yes No	(0,1)
Having surgery	Yes No	(0,1)
High fever in the last 1 year	Less than three months ago, more than three months ago, never	(-1, 0, 1)
Frequency of alcohol consumption	Several times a day, every day, a few times a week, once a week, almost never, or never	(0, 0.2, 0.4, 0.6, 0.8, 1)
Smoking habit	Never, once in a while, every day	(-1, 0, 1)
Number of hours spent sitting per day	Between 1 to 16 hours	Scale [0-1]
Diagnosis	Normal (N), Altered (O)	1,0

Preprocessing

Preprocessing or also known as data normalization serves to prepare data that is truly valid before being processed at the next stage (Erdiansyah et al., 2022). In this study, data normalization was carried out using the Min-Max method with the following formula (Lubis et al., 2022).

$$\frac{(Data-Min)*(NewMax-NewMin)}{(Max-Min)} + NewMin \quad (1)$$

Proposed Classification Model

Naive Bayes is a method that has no rules, Naive Bayes uses a branch of mathematics known as probability theory to find the greatest chance of a possible classification, by looking at the frequency of each classification in the training data. Naive Bayes is stated in the equation (Susanto, 2013):

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \quad (2)$$

Where Ci is the data hypothesis X, then X is data with an unknown class. For P(Ci|X) is the Probability of Ci based on X conditions, while P(X|Ci) is the Probability of X based on Ci conditions. Then P(Ci) is the probability of the hypothesis and Ci. P(X) is the probability of X data.

Forward Selection is an attribute reduction method that involves an empty set of attributes that need to be changed. Then, each attribute is evaluated individually, and the best attribute is selected with the highest possible amplification. Then, do the next iteration of the test continuously and stop until the attribute being tested does not have a significant impact on accuracy. Forward Selection is formulated as follows:

The first step is to do Determine the initial model, as in the following formula:

$$\hat{y} = b_0 \quad (3)$$

Enter the response variable with each predictor variable, for example $X1$, $X2$, …. Xn which is related to y. Suppose $X1$ so that it forms a model:

$$\hat{y} = b_0 + b_1X_1 \quad (4)$$

F test of the first variable selected. If *F*count < *F*table then the selected variable is discarded and the process is stopped. If *Fh*count > *F*table then the selected variable has a real influence on the related variable y; so it deserves to be taken into account in the model;

Enter the selected independent variables (the most significant) into the model. Suppose X2, thus forming a model:

$$\hat{y} = b_0 + b_1X_1 + b_2X_2 \quad (5)$$

F test, if Fcount <Ftable, then the process is stopped and the best model is the previous model. However, if Fhtung > Ftable, the independent variables are eligible to be included in the model and return to step C. The process will end if there are no remaining variables that can be entered into the model.

*Ahmadi Irmansyah Lubis

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.11967
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Stratified Sampling is a random sampling technique in the data population. The dataset is divided into several separate parts, then samples are taken randomly based on the strata that have been made. The Stratified Sampling stages are as follows (Ubaedi & Djaksana, 2022):

First, the population N is divided into sub-populations consisting of elements $N1$, $N2$, $N3$, ..., NL.
Then, between the sub-populations, there should be no overlap, so that $N1 + N2 + N3 + \cdots + NL = N$.
Finally, take a random sample from each sub-population with a proportional sample allocation.

Prior to sampling, determining the sample size is important. The sample taken must reflect the population. There are several ways to determine sample size. One of the most widely used and commonly used slovin theory is explained by the following formula:

$$n = \frac{N}{1+Ne^2} \quad (6)$$

Where n is the sample size, then N is the population size or the number of elements in the population and e is the precision value.

Experiments and Testing

In this stage, experiments are carried out on the data model that will be processed using the proposed method. Experiments in this study were carried out on the Fertility Dataset, after which the dataset will be divided into two parts, namely training data by 80% while data testing by 20%. The amount in the training data and data testing is determined by the split data process. Then an optimization algorithm will be applied using Forward Selection to calculate the weight value of each attribute before attribute reduction is carried out, so that it will increase the accuracy value in Naïve Bayes from the dataset.

Evaluation and Validation

The evaluation process is carried out using the Confusion Matrix and the ROC (Receiver Operating Characteristic) curve. While the validation process will be carried out with Split Validation. The evaluation stage with Rapidminer automatically when the process is running. With the performance of the models to be compared, namely the performance of Naïve Bayes with Naïve Bayes with attribute reduction using Forward Selection. In measuring the results of the performance of the classification in this study carried out with the Confusion Matrix to obtain the results of Accuracy, Precision, Recall. Table 1 Confusion Matrix can be seen in Table 2 (Harafani & Maulana, 2019).

Tabel 2. Tabel Confusion Matrix

Actual Class	Assigned Class	
	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

The True Positive (TP) and True Negative (TN) are conditions in which the prediction results match the actual conditions that occur. Meanwhile, False Positive (FP) and False Negative (FN) are conditions in which the predicted results do not match the actual conditions. Then to calculate Accuracy, Precision, and Recall values can be calculated using the formula:

$$Acuraccy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$Precision = \frac{TP+TN}{TP+FP} \quad (9)$$

The ROC (Receiver Operating Characteristic) curve is a graph between Y-axis sensitivity and 1-X-axis specificity which seems to describe the correspondence between the Y-axis or sensitivity and X-axis or specificity. On the ROC curve, classification is a method of visualizing, organizing and selecting classifications based on algorithm performance. According to the accuracy of the AUC (Area Under ROC Curve) value in data mining classification it is divided into five groups namely (Ubaedi & Djaksana, 2022):

0.90 - 1.00 = very good classification
0.80 - 0.90 = good classification
0.70 - 0.80 = sufficient classification
0.60 - 0.70 = bad classification

*Ahmadi Irmansyah Lubis

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.11967
e-ISSN : 2541-2019
p-ISSN : 2541-044X

0.50 - 0.60 = incorrect classification

RESULT

In the application of Naïve Bayes with Forward Selection on the Fertility Dataset with the aim of knowing and getting results of better accuracy values in the classification of sperm fertility. The dataset that has been processed will then be calculated using RapidMiner Software to find out the results of the accuracy value of the method used in this study.

Naïve Bayes Test Results

This experiment was carried out on rapidminer, with the division of the dataset using split validation, namely the division of the amount of training and testing data was determined manually. In this study, the distribution of training and testing data was 80%:20%, namely 80% for training data or 80 data and 20% for testing data or 20 data. This training data is for model development and this testing data is for model testing. Then it is calculated using Naïve Bayes and produces a confusion matrix as shown in Table 3.

Table 3. Results of Confusion Matrix on Naïve Bayes

	True Normal	True Altered
Pred. Normal	85	12
Pred. Altered	3	0

From the calculation of the Confusion Matrix in Table 3. it can be measured the level of accuracy of the classification with Naïve Bayes without attribute reduction, namely as follows:

$$Acuraccy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{85 + 0}{85 + 12 + 3 + 0} = \frac{85}{100} = 0.85 \times 100\% = 85.00\%$$

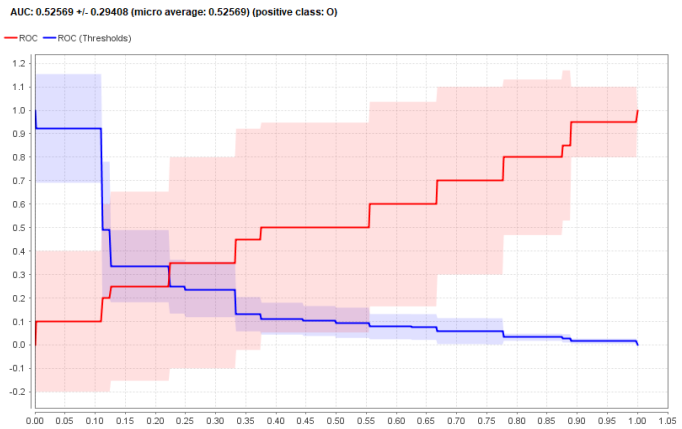

Fig 2. ROC Graph on Naïve Bayes Testing

Besides being able to produce a Confusion Matrix, this test also produces a ROC (Receiver Operating Characteric) curve as shown in Figure 2 with an AUC value of 0.52569.

Naïve Bayes + Forward Selection Test Results

At this stage, the experimental results will be explained using the Forward Selection method as an attribute weight optimization which is meant to add the weight value of each attribute in a dataset for sperm fertility classification, which will be calculated using RapidMiner and produce weight values as Table 4.

Table 4. Results of Calculation of the Attribute Weights of the Fertility Dataset

Attribute	Weight
Season	1
Age	1
Childish Diseases	0
Accident or Serious Trauma	0
Surgical Intervention	0
High Fever in the last year	0

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.11967
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Frequency of Alcohol Consumption	0
Smoking Habit	0
Number of Hours Spent Sitting Per Day	0

The results of calculating the weight of this attribute are symbolized by the numbers 1 and 0. The attribute weight with a value of 1 is likely to significantly affect the accuracy results. There are 2 attributes with a weight of 1, namely Season and Age. Meanwhile, attributes with a weight of 0 indicate that these attributes do not affect the accuracy value and are reduced to these attributes.

Table 5. Results of Confusion Matrix on Naïve Bayes + Forward Selection

	True Normal	True Altered
Pred. Normal	44	6
Pred. Altered	0	0

From the calculation of the Confusion Matrix in Table 5. the accuracy of the classification with Naïve Bayes + Forward Selection can be measured as follows:

$$Acuraccy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{44 + 0}{44 + 0 + 6 + 0} = \frac{44}{50} = 0.88 \times 100\% = 88.00\%$$

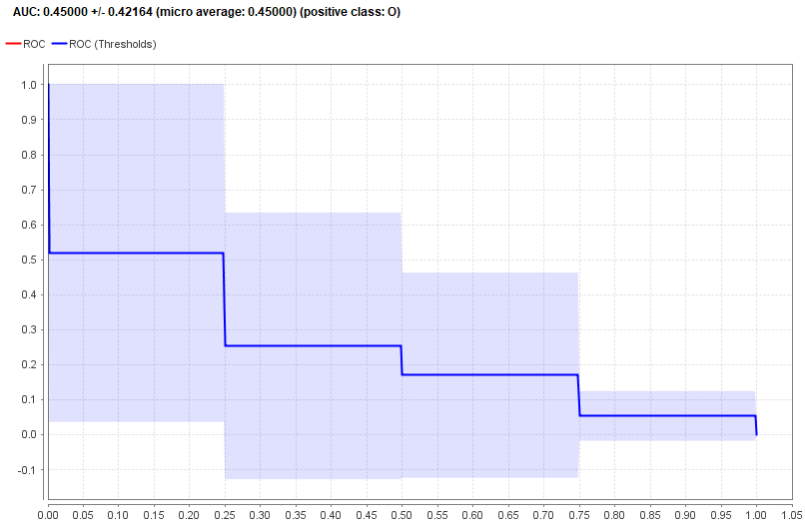

Fig 3. Grafik ROC Pada Pengujian Naïve Bayes + Forward Selection

The ROC (Receiver Operating Characteristic) Curve as shown in Figure 3 with an AUC value of 0.450

DISCUSSION

From the results of the research that has been done, an evaluation will be carried out at this stage so that it is known that the increase in the accuracy value of the Naïve Bayes classification before attribute reduction and Naïve Bayes is carried out after optimizing attribute weights using Forward Selection by calculating testing data in the form of accuracy values and ROC graphs. The following is a comparison of the accuracy values of Naïve Bayes before optimizing attribute weights and Naïve Bayes after reducing attributes using Forward Selection.

Table 5. Comparison Results of Testing the Proposed Method

Methods	Accuracy (%)	ROC
Naïve Bayes	85.00	0.52569
Naïve Bayes + Forward Selection	88.00	0.45000

*Ahmadi Irmansyah Lubis

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.11967
e-ISSN : 2541-2019
p-ISSN : 2541-044X

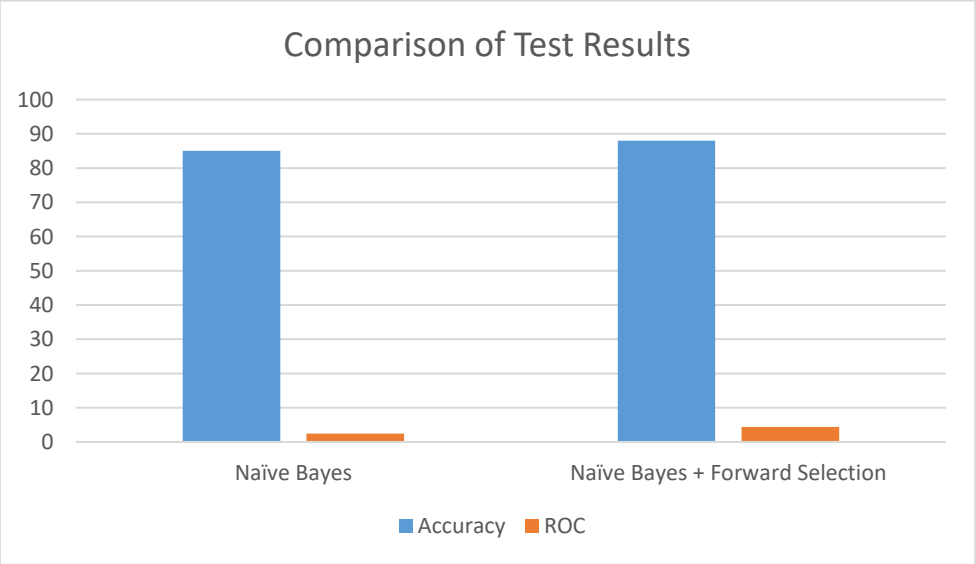

Fig 4. Comparison Results of Testing the Proposed Method

In Figure 4 it can be seen that the classification results of the Naïve Bayes method have increased accuracy by 3% after attribute reduction using Forward Selection. So it is proven that Forward Selection optimization can improve the performance of Naïve Bayes in sperm fertility classification. This research is a continuation of previous studies using the same data, namely the Fertility Dataset.

CONCLUSION

Based on the test results of the proposed model, namely the classification of Fertility Dataset with Naïve Bayes and Forward Selection, results in increased accuracy and performance. The results of accuracy with Naïve Bayes on the Fertility Dataset produce a confusion matrix accuracy value of 85.00% and an AUC of 0.769 after optimization using Forward Selection for attribute reduction in the Fertility Dataset and the results of the accuracy of Naïve Bayes increase to 88.00% and AUC of 0.793. so it can be concluded that Forward Selection can improve the accuracy of Naïve Bayes in Fertility Dataset.

REFERENCES

Arifin, T. (2020). Optimasi Decision Tree menggunakan Particle Swarm Optimization untuk klasifikasi sel Pap Smear. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *7*(3), 572–579. https://doi.org/10.35957/jatisi.v7i3.361

Budianita, E., Hustianto, F. R., Syafria, F., & Nasir, M. (2018). Implementasi Algoritma Jaringan Syaraf Tiruan (JST) Hopfield untuk Klasifikasi Kualitas Kesuburan Pria. *Seminar Nasional Teknologi Informasi, Komunikasi Dan Industri (SNTIKI-10)*, *November*, 137–142.

Erdiansyah, U., Lubis, A. I., & Syahputra, G. (2022). *Klasifikasi Penyakit Diabetic Retinopathy Menggunakan Multilayer Perceptron*. 1–6.

Harafani, H., & Maulana, A. (2019). Penerapan Algoritma Genetika pada Support Vector Machine Sebagai Pengoptimasi Parameter untuk Memprediksi Kesuburan. *Jurnal Teknik Informatika STMIK Antar Bangsa*, *V*(1), 51–59.

Hermawanti, L., & Safriandono, A. N. (2016). PENGGABUNGAN ALGORITMA FORWARD SELECTION DAN K-NEAREST NEIGHBOR UNTUK MENDIAGNOSIS PENYAKIT DIABETES DI KOTA SEMARANG/Combining of Forward Selection Algorithm and K-Nearest Neighbor To Diagnose Diabetes Disease in Semarang City. *Momentum*, *12*(2), 28–31.

Irmansyah Lubis, A., Setiawan, F., & Lusiyanti, L. (2021). Penentuan Peringkat Konsentrasi Tingkat Kesuburan Sperma Menggunakan Metode MOORA. *Digital Transformation Technology*, *1*(2), 62–68. https://doi.org/10.47709/digitech.v1i2.1116

Khaira, U., Syarief, N., & Hayati, I. (2020). Prediksi Tingkat Fertilitas Pria Dengan Algoritma Pohon Keputusan Cart. *Program Studi Sistem Informasi, Fakultas Sains Dan Teknologi, Universitas Jambi*, *5*(1), 35–42.

Laksono, P. J. T. (2008). Penerapan Forward Selection Pada Support Vector Machine Untuk Klasifikasi Kanker Payudara. *Ilmukomputer.Com*, 1–27.

*Ahmadi Irmansyah Lubis

Lubis, A. I., Erdiansyah, U., & Siregar, R. (2022). Komparasi Akurasi pada Naive Bayes dan Random Forest dalam Klasifikasi Penyakit Liver. *Journal of Computing Engineering, System and Science (CESS)*, *7*(1), 81–89.

Lubis, A. I., & Setiawan, F. (2022). *Komparasi Kinerja ELECTRE dan MOORA dalam Menentukan Konsentrasi Tingkat Kesuburan Sperma*. Jurnal *Infotekmesin*, *13*(01), 99–105. https://doi.org/10.35970/infotekmesin.v13i1.1012

Lubis, A. I., Sihombing, P., & Nababan, E. B. (2020). Comparison SAW and MOORA Methods with Attribute Weighting Using Rank Order Centroid in Decision Making. *MECnIT 2020 - International Conference on Mechanical, Electronics, Computer, and Industrial Technology*, *February 2022*, 127–131. https://doi.org/10.1109/MECnIT48290.2020.9166640

Ma'mur, K. (2019). Analisis Penerapan Algoritma ID3 dalam Mendiagnosis Kesuburan Pria. *Jurnal Informatika Universitas Pamulang*, *4*(2), 35. https://doi.org/10.32493/informatika.v4i2.2274

Noviati, N., Fauziati, S., & Hidayah, I. (2015). Analisis Pengaruh Seleksi Fitur pada Klasifikasi Konsentrasi … (Noviati dkk.). *Snst*, 160–165.

Nurelasari, E. (2018). Komparasi Algoritma Naive Bayes Dengan Support Vector Machine Berbasis Particle Swarm Optimization untuk Prediksi Kesuburan. *Bina Insani ICT Journal*, *5*(1), 61–70.

Susanto, B. M. (2013). Komparasi Algoritma Naive Bayes Dan C4. 5 Dalam Mendeteksi Kesuburan. *Snit 2013*, 69–73. http://seminar.bsi.ac.id/snit/index.php/snit-2013/article/view/268%0Ahttps://seminar.bsi.ac.id/snit/index.php/snit-2013/article/view/268/264

Ubaedi, I., & Djaksana, Y. M. (2022). Optimasi Algoritma C4.5 Menggunakan Metode Forward Selection Dan Stratified Sampling Untuk Prediksi Kelayakan Kredit. *JSiI (Jurnal Sistem Informasi)*, *9*(1), 17–26. https://doi.org/10.30656/jsii.v9i1.3505

Utomo, D. P., Sirait, P., & Yunis, R. (2020). Reduksi Atribut Pada Dataset Penyakit Jantung dan Klasifikasi Menggunakan Algoritma C5.0. *Jurnal Media Informatika Budidarma*, *4*(4), 994–1006. https://doi.org/10.30865/mib.v4i4.2355

Yepriyanto, R., Kustanto, & Utami, Y. R. W. (2014). Sistem Diagnosa Kesuburan Sperma Dengan Metode K-Nearest Neighbor (K-Nn). *Jurnal Ilmiah SINUS*, 33–44.

*Ahmadi Irmansyah Lubis