


# PCOS Disease Classification Using XGBoost Algorithm and Genetic Algorithm for Feature Selection

<sup>1,\*</sup>Enda Putri Atika, <sup>2</sup>Muh. Ilham Nadzirullah, <sup>3</sup>Alti Arindika 

<sup>1,2,3</sup> Department of Informatic Engineering, AMIKOM University, Special Region of Yogyakarta, Indonesia

\* Corresponding Author: endaputri@amikom.ac.id

**Abstract:** Polycystic Ovary Syndrome (PCOS) is an endocrine disorder that often occurs in women of reproductive age, with a global prevalence of 10-16%. The diagnosis of PCOS is still a challenge due to the uncertainty of the cause, which can worsen the patient's condition due to delayed detection. This study aims to develop a classification model to detect PCOS using a combination of SMOTE algorithm, genetic algorithm, and XGBoost. The dataset used is a public dataset from Kaggle entitled "Diet, Exercise, and PCOS Insights". A genetic algorithm was used to select the best 15 features, while SMOTE was applied to handle data imbalances. XGBoost is used for classification with a model accuracy of 82.86% and an F1-score of 88% for the PCOS negative class and 70% for the PCOS positive class. The results show that combining these algorithms can improve the accuracy of predictions and offer more efficient diagnosis solutions. This research is expected to contribute to developing early diagnosis methods for PCOS.

**Keywords:** Genetic Algorithm, SMOTE, XGBoost, PCOS, Classification



**Citation:** Atika, E. P., Nadzirullah, M. I., & Arindika, A. (2025). PCOS disease classification using XGBoost algorithm and genetic algorithm for feature selection. *Iota*, 5(1). <https://doi.org/10.31763/iota.v5i1.874>

Academic Editor: Adi, P.D.P

Received: January 05, 2025

Accepted: January 27, 2025

Published: February 06, 2025

**Publisher's Note:** ASCEE stays neutral about jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2025 by authors. Licensee ASCEE, Indonesia. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-Share Alike (CC BY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>)

## 1. Introduction

The development of digital technology today, especially artificial intelligence, has developed very rapidly in all fields, be it in the fields of education, transportation, agriculture, fisheries, tourism, and so on, without exception in the health sector. Artificial intelligence technology innovations in the health sector are very helpful for medical personnel in increasing the speed of diagnosing a disease (Borneo, 2023), as in the study (Purnama et al., 2024; Safitri et al., 2024) The use of machine learning algorithms in detecting diabetes and detecting potential health risks in pregnant women. The sooner a disease is detected, the faster treatment can be done.

One of the diseases that currently attracts attention because it often attacks the female reproductive organs at a productive age is PCOS (Polycystic Ovary Syndrome) (Saputra; Akbar Novan Dwi, 2019). PCOS is an endocrine disorder or syndrome that is very common in women of reproductive age with a spread rate of 10%-16% worldwide (Christiani et al., 2023). According to research (Aubuchon, 2020; Mareta & Amran, 2018), women with PCOS disorders have an 8.6 times greater risk of infertility, as well as a risk of serious complications such as type 2 diabetes, cardiovascular disorders and Obstructive Sleep Apnea (OSA).

Although PCOS is very common, the diagnosis is still quite difficult because the cause is not known for sure, based on the existing theory, high levels of the hormone insulin can be the main cause of PCOS disorders, but it is possible that other factors can also be the cause (Saputra; Akbar Novan Dwi, 2019). The causative inaccessibility of PCOS often leads to a delayed diagnosis, which can ultimately lead to a patient's condition worsening.

Therefore, based on the problems that have been explained earlier, this study wants to try to apply artificial intelligence technology in diagnosing complex diseases such as PCOS with the hope that this technology can later help medical personnel detect PCOS disorders quickly and accurately.

The machine learning algorithm that will be used is the XGBoost Algorithm to carry out the task of classifying PCOS diseases because according to the journal (Kurniawan & Indahyanti, 2024; Rayadin et al., 2024) The XGBoost algorithm has a good performance in classifying or predicting biner data and this algorithm is also able to overcome the classification of unbalanced datasets (Septiana Rizky et al., 2022). The dataset used in this study is a public dataset available by Kaggle, there are approximately 30 features in this dataset, but not all of these features are relevant to be used in diagnosing PCOS disorders. Therefore, the Genertika Algorithm is used to select the most relevant features because in several previous studies such as (Nisyak et al., 2024; Sitanggang & Bahtiar, 2019; Wulandari et al., 2024) The use of machine learning algorithms combined with genetic algorithms in predicting the number of tourists, hepatitis prediction, and stroke prediction can produce a much better level of accuracy with relevant features.

In addition to the XGBoost Algorithm and the Genetics Algorithm, this study also uses the SMOTE Algorithm (Synthetic Minority Oversampling Technique) because the dataset used in this study is unbalanced data. Data imbalance occurs when the number of samples in one class is much larger than in another class in a dataset. This can cause machine learning models to tend to be biased toward the majority class because they ignore or do not recognize the minority class well. As a result, the model may show poor performance in detecting instances from minority classes (Nusantara, 2024).

One technique to overcome data imbalance is oversampling which increases the number of samples in minority classes. Some of the oversampling methods include the Synthetic Minority Over-sampling Technique (SMOTE), MDO (Majority-Driven Oversampling), and Random Oversampling (ROS). The use of oversampling can improve classification results by balancing the distribution of classes in the dataset (Cosmas Haryawan & Yosef Muria Kusuma Ardhana, 2023).

However, it is important to note that oversampling has drawbacks, such as increasing the risk of overfitting because the model can learn from over-added synthetic data. Therefore, proper selection of oversampling techniques and careful model evaluation are essential to ensure optimal model performance (Nusantara, 2024). In this study, the oversampling algorithm to be used is the SMOTE Algorithm because according to the journal (Amartha et al., 2025; Yanuari et al., 2024), SMOTE and XGBoost algorithms have good performance in oversampling data from unbalanced data.

The combination of SMOTE, XGBoost algorithms, and genetic algorithms in this study aims to create a more accurate and efficient classification model for PCOS diagnosis. By using genetic algorithms for feature selection, dataset dimensions can be significantly reduced without losing important information, thereby improving computational efficiency and model interpretability. Meanwhile, XGBoost is expected to maximize prediction accuracy by leveraging boosting capabilities to handle complex and unbalanced data.

This research has not only contributed to the development of better diagnostic methods for PCOS but also provided deeper insights into the key features associated with the disease. Thus, the results of this study are expected to support efforts for early detection, management, and prevention of PCOS complications more effectively in the future.

## 2. Theory

### 2.1 Dataset

The dataset used in this study is public in Kaggle, with the dataset title "Diet, Exercise and PCOS Insight". The data has 37 columns, 36 of which can be used as a feature to detect PCOS disorders in women, except for the "PCOS" column because the column will be labeled, the person has PCOS disorder or not. The total data in the dataset is 172 data, 80% will be used for training data and 20% for data testing. Each column in the dataset used can be seen in Table 1.

**Table 1.** Column descriptions in the dataset

It	Column Name
1	Age
2	Weight_kg
3	Height_ft
4	Marital_status
5	Family_History_PCOS
6	Menstrual_Irregularity
7	Hormonal_Imbalance
8	Hyperandrogenism
9	Hirsutism
10	Mental_Health
11	Conception_Difficulty
12	Insulin_Resistance
13	Diabetes
14	Childhood_Trauma
15	Cardiovascular_Disease
16	Diet_Bread_Cereals
17	Diet_Milk_Products
18	Diet_Fruits
19	Diet_Vegetables
20	Diet_Starchy_Vegetables
21	Diet_NonStarchy_Vegetables
22	Diet_Fats
23	Diet_Sweets
24	Diet_Fried_Food
25	Diet_Tea_Coffee
26	Diet_Multivitamin
27	Vegetarian
28	Exercise_Frequency
29	Exercise_Type
30	Exercise_Duration
31	Sleep_Hours
32	Stress_Level
33	Smoking
34	Exercise_Benefit
35	PCOS_Medication
36	BMI
37	PCOS

## 2.2 PCOS

PCOS (Polycystic Ovary Syndrome) is an endocrine disorder that often occurs in women of reproductive age with a prevalence rate of about 10%-16% worldwide (Christiani et al., 2023). The cause of PCOS is not known for sure, but according to the primary theory, one of the causes of PCOS are insulin levels, irregular menstruation, hormonal imbalances, and many more (Saputra; Akbar Novan Dwi, 2019). If PCOS is not treated immediately, it can lead to more serious complications such as type 2 diabetes, cardiovascular disorders and *Obstructive Sleep Apnea (OSA)*, and an 8.6 times higher risk of infertility (Aubuchon, 2020; Mareta & Amran, 2018).

## 2.3 SMOTE

SMOTE (Synthetic Minority Oversampling Technique) is an algorithm used to handle unbalanced data by oversampling data on minority classes so that the amount of data is balanced with the majority class. This algorithm works by oversampling or creating sperm synthesis data for training in minority classes until the amount of data is the same as that of the majority class (Chawla et al., 2018).

The advantages of the SMOTE Algorithm over other oversampling algorithms have been proven in various journals. Such as research conducted by (Hasanah et al., 2024) The use of the SMOTE algorithm can significantly improve the performance of cardiovascular disease prediction models. According to research (WIJAYANTI et al., 2021) The use of the SMOTE algorithm does not cause any information to be lost, can prevent overfitting, and can improve the accuracy of predictions in minority classes.

## 2.4 XGBoost Algorithm

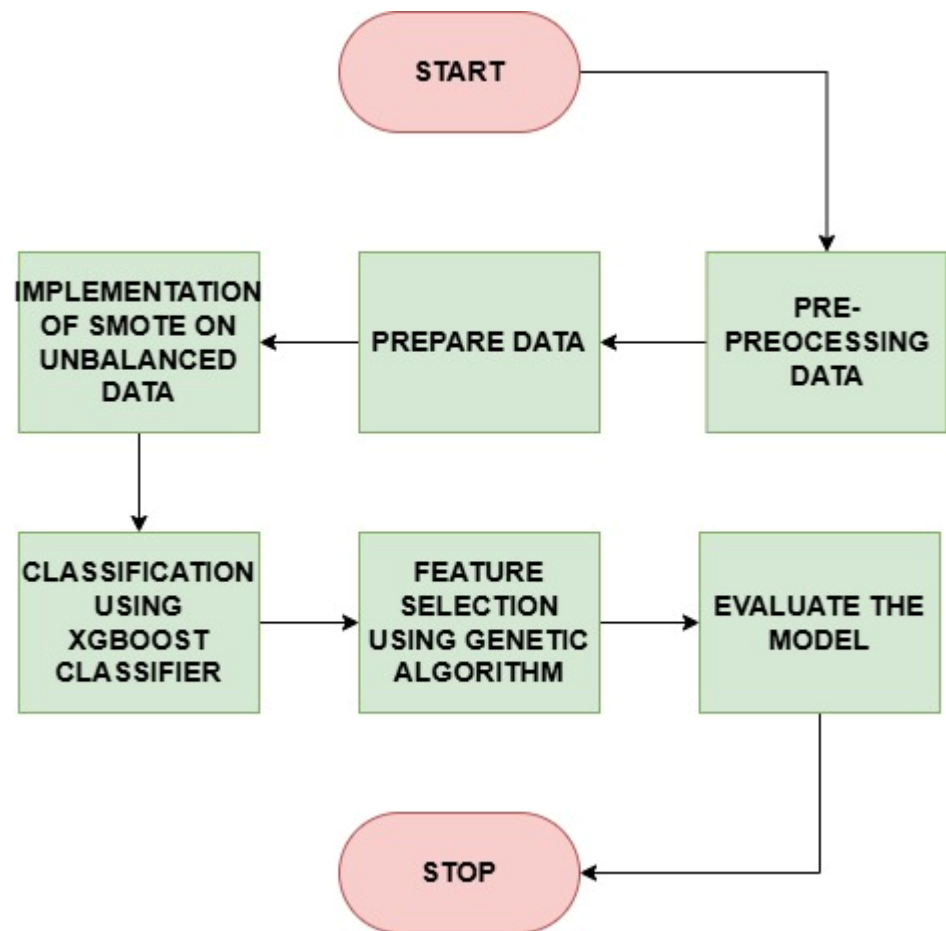
XGBoost (Extreme Gradient Boosting) is a machine learning algorithm that uses boosting techniques to improve prediction accuracy. The way it works involves iterative model building, where each new model attempts to correct the prediction errors of the previous model. This process starts with building a basic model, then the next model is focused on data that was difficult for the previous model to predict. By combining predictions from all models, XGBoost can generate powerful models with high accuracy. Additionally, XGBoost offers features such as regularization to prevent overfitting and the ability to handle lost data (Herni Yulianti et al., 2022). In addition, this algorithm also supports distributed learning, so it can be implemented on various computing platforms. Recent studies have shown that XGBoost can generate highly accurate models in a variety of applications, such as predicting infection risk in patients with limited clinical data (Zheng et al., 2023).

## 2.5 Genetic Algorithm

Genetic algorithms are optimization methods that mimic natural evolutionary processes, such as selection, crossover, and mutation, to find the optimal solution to a problem. The process begins by generating an initial population of randomly generated potential solutions. To evaluate the quality of each member of the population using objective functions. Individuals with the highest quality are selected for reproduction, where they undergo cross-over and mutation to produce the next generation. This cycle repeats until we get a solution or meet the termination criteria (Pane et al., 2019). The performance of the Genetic Algorithm has been proven in previous studies. For example, in research (Aisha & Lubis, 2024; Pangestu et al., 2023) The use of genetic algorithms can optimally handle the problem of creating subject schedules and optimizing network placement.

### 3. Method

The flow and research methods used in this research can be seen in Figure 1.



**Figure 1.** Research Flow

#### 3.1 Prepare Data

The first step that must be done is to prepare the data to be used. In this study, the data used is secondary data that can be downloaded through the Kaggle website with the title "Diet, Exercise and PCOS Insight".

#### 3.2 Pre-processing Data

Preprocessing the data used, the preprocessing carried out in this study is to delete data that is not needed such as in the PCOS column, one data is not detected by PCOS or not. Then convert categorical data into numerical data so that calculations can be made.

#### 3.3 Smote Implementation on Imbalanced Data

After preprocessing, the next step is to oversample the unbalanced data in the minority class using the SMOTE Algorithm.

#### 3.4 Feature Selection Using Genetic Algorithm

The next step that must be done is to select the relevant features or the best of the 36 existing features. Because not all features in this dataset are relevant to detecting PCOS disease disorders in women.

#### 3.5 Classification using XGBoost Classifier

After finding the best features of the results of the Genetic Algorithm processing, the next step is to classify the females who are PCOS positive and PCOS negative.

### 3.6 Evaluate the model with f1-score and accuracy.

The last stage is to evaluate the model made by looking at its accuracy value and F1-Score value, to see the performance of the model in classifying PCOS disease data.

## 4. Result and Discussion

As previously explained, there is an imbalance in the data in the dataset used in this study. The data of Class 0 or class "PCOS Negative" is more than Class 1 or class that is "PCOS Positive" with a ratio of 107:30. If not overcome, this dataset imbalance has the potential to cause bias in the model because the prediction results will tend to the majority class. Therefore, to overcome this, oversampling is carried out on minority classes using the SMOTE (Synthetic Minority Oversampling Technique) Algorithm. The SMOTE algorithm creates synthetic data by interpolating between samples of minority classes so that the amount of data in both classes is balanced. The results of this oversampling show that the amount of data in the "PCOS Negative" class and the "PCOS Positive" class becomes the same as in Figure 2 so that the distribution of the dataset becomes more even. This process ensures that the model can learn well from both classes without bias towards a particular class.

```
Sebelum SMOTE:
Counter({0: 107, 1: 30})

Setelah SMOTE:
Counter({0: 107, 1: 107})
```

**Figure 2.** SMOTE Algorithm Results

After the dataset is balanced, a feature selection stage is carried out using the Genetics algorithm to determine the best features that are most relevant in detecting PCOS. From the results of the feature selection carried out, the 15 best features were selected that covered various aspects, ranging from biological aspects such as age, family history with PCOS, menstrual irregularities, hormonal and lifestyle imbalances such as consumption of certain foods, and stress levels. The results obtained show that PCOS detection requires a multidimensional approach that not only focuses on physical health conditions but also individual lifestyles. From the results of the processing of the Genetic Algorithm, the 15 best features that are relevant in detecting PCOS can be seen in Table 2.

**Table 2.** Best Features

It	Feature Name	Explanation
1	Age	The age of the individual.
2	Family_History_PCOS	Family history with PCOS.
3	Menstrual_Irregularity	Menstrual irregularities.
4	Hormonal_Imbalance	Hormonal imbalance.
5	Hyperandrogenism:	High levels of androgens.
6	Conception_Difficulty	Difficulty getting pregnant.
7	Childhood_Trauma:	Childhood trauma.
8	Diet_Bread_Cereals:	Consume bread and cereals.
9	Diet_Fruits:	Consume fruits.
10	Diet_Vegetables:	Consume vegetables.
11	Diet_Sweets	Consume sweet foods.
12	Sleep_Hours	Sleep duration.
13	Stress_Level	Stress levels.
14	Smoking	Smoking habits.
15	PCOS_Medication	Use of PCOS drugs.

The performance of the classification model is evaluated using accuracy metrics and F1-Score. The model was built with an accuracy score of 82.86%, with an F1-Score of 88.86% for the "PCOS Negative" class and 70% for the "PCOS Positive" class. These results show that although the performance of the model in the majority class is better, the application of SMOTE has succeeded in improving the model's ability to recognize minority classes. These results prove that oversampling using SMOTE is an effective way to deal with data imbalances in PCOS classification.

From the results of this study, it can be seen that biological factors such as hormonal imbalance, menstrual irregularities, and hyperandrogenism have a great contribution to detecting PCOS which is in line with the medical literature (Salsabila et al., 2024). However, lifestyle factors such as the consumption of certain foods, stress levels, and smoking habits were also found to have a significant influence. These findings highlight the importance of a holistic approach to diagnosing PCOS, where biological and lifestyle factors need to be considered simultaneously.

Although this study has shown promising results, some limitations need to be noted. For example, the risk of overfitting due to oversampling with SMOTE needs to be further evaluated through testing on different datasets. In addition, model interpretation using techniques such as SHAP or LIME can provide additional insights into the influence of each feature on prediction, so that the results of this study can be more easily applied in medical practice. Thus, this research not only contributes to the development of AI-based PCOS detection models but also opens up opportunities for more in-depth follow-up research.

## 5. Conclusions

Based on the results obtained, it can be concluded that this study successfully overcomes the problem of data imbalance in the PCOS dataset using the SMOTE algorithm, which improves the distribution of training data to be balanced between the "PCOS Negative" and "PCOS Positive" classes. The application of genetic algorithms for feature selection resulted in the 15 best features that include biological and lifestyle factors, suggesting that PCOS detection requires a multidimensional approach. The classification model developed resulted in an accuracy of 82.86%, with an F1-Score of 88% for the majority class and 70% for the minority class, indicating an improvement in the model's ability to recognize minority classes after oversampling. These results suggest that factors such as hormonal imbalances, family history with PCOS, and lifestyle, including the consumption of certain foods as well as stress levels, have a significant contribution to the detection of PCOS. However, if you look at the F1-score value, the model still has limitations in detecting minority classes (PCOS), because the difference in the amount of data is too significant, so the majority of training data used in minority classes is synthetic data or artificial data. For more accurate results, future studies can add more data, and make the amount of data from both classes balanced by using original data, so that the results obtained are more relevant and accurate.

**Acknowledgments:** We would like to thank AMIKOM University Yogyakarta for the support provided during this research. Thank you also to the Kaggle platform for the "Diet, Exercise, and PCOS Insights" dataset used in this study. The support of colleagues and family was very meaningful in completing this study.

**Author contributions:** The authors were responsible for building Conceptualization, Methodology, analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision of project administration, funding acquisition, and have read and agreed to the published version of the manuscript.

**Funding:** The study was conducted without any financial support from external sources.

**Availability of data and Materials:** All data are available from the authors.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Additional Information:** No Additional Information from the authors.

## References

1. Aisha, S., & Lubis, P. (2024). Application of Genetic Algorithms in Network Optimization Deployment. *Journal of Mathematics and Natural Sciences*, 2(1), 40–50. <https://doi.org/10.59581/konstanta-widyakarya.v2i1.2021>
2. Amarta, M. R., Wahyuni, R., & Irawan, Y. (2025). Optimization of C5. 0 Algorithm with Ensemble Boosting Technique for Improved Accuracy in the Classification of Public Reviews of BPJS Kesehatan Services. 5(1).
3. Aubuchon, M. (2020). Polycystic Ovary Syndrome and Obstructive Sleep Apnea. *Current Clinical Neurology*, 3(1), 177–202. [https://doi.org/10.1007/978-3-030-40842-8\\_13](https://doi.org/10.1007/978-3-030-40842-8_13)
4. Borneo, A. H. (2023). Development of Health Technology in the Digital Era. <https://stikeshb.ac.id/teknologi-kesehatan-di-era-digital/>
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Philip Kegelmeyer, W. S. (2018). *J Artif Intell Res*, 16, 16.
6. Christiani, Manubulu, C. P., & Arimas, E. (2023). Effects of curcumin on oxidative stress and metabolic profile of polycystic ovary syndrome patients. *Supplements*, 15, 1–11. <https://myjurnal.poltekkes-kdi.ac.id/index.php/hijp>
7. Cosmas Haryawan, & Yosef Muria Kusuma Ardhana. (2023). Comparative Analysis of Smote Oversampling Techniques on Imbalanced Data. *Journal of Informatics and Electronic Engineering*, 6(1), 73–78. <https://doi.org/10.36595/jire.v6i1.834>
8. Hasanah, U., Soleh, A. M., & Sadik, K. (2024). Effect of Random Under Sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models. 21(1), 88–102. <https://doi.org/10.20956/j.v21i1.35552>
9. Herni Yulianti, S. E., Oni Soesanto, & Yuana Sukmawaty. (2022). Application of the Extreme Gradient Boosting (XGBOOST) Method in Credit Card Customer Classification. *Journal of Mathematics: Theory and Applications*, 4(1), 21–26. <https://doi.org/10.31605/jomta.v4i1.1792>
10. Kurniawan, W., & Indahyanti, U. (2024). Predict the life expectancy of the population using the XGBoost method. *Indonesian Journal of Applied Technology*, 1(2), 18. <https://doi.org/10.47134/ijat.v1i2.3045>
11. Mareta, R., & Amran, R. (2018). The Relationship between Polycystic Ovary Syndrome (PCOS) and Infertility in the Private Practice of Palembang Obstetrics and Gynecology Doctors. *Sriwijaya Medical Magazine*, 50(2), 85–91. <https://ejournal.unsri.ac.id/index.php/mks/article/view/8552>
12. Nisyak, H., Fithriyah, N., & Fatimatu Zahra. (2024). Neural Network Optimization uses genetic algorithms to predict the number of tourists based on hotel occupancy. *JoMI: Journal of Millennial Informatics*, 2(1), 23–30. <https://journal.mudaberkarya.id/index.php/JoMI/article/view/114>
13. Nusantara, U. B. (2024). Oversampling VS Undersampling in Dealing with Data Imbalance. Bina Nusantara University. <https://sis.binus.ac.id/2024/11/01/oversampling-vs-undersampling-dalam-mengatasi-ketidakseimbangan-data/>
14. Pane, S. F., Maulana Awangga, R., Rahmadani, E. V., & Permana, S. (2019). Implementation of Genetic Algorithms for Population Service Optimization. *Journal of Incentive Tech*, 13(2), 36–43. <https://doi.org/10.36787/jti.v13i2.130>
15. Pangestu, L. A., Suryawan, S. H., & Latipah, A. J. (2023). Application of Genetic Algorithms in Scheduling Subjects. *Journal of Informatics*, 10(2), 194–205. <https://doi.org/10.31294/inf.v10i2.16701>
16. Purnama, J. J., Hikmawati, N. K., & Rahayu, S. (2024). Analysis of classification algorithms to identify potential health risks of pregnant women. *Journal of Applied Computer Science and Technology ( Jacost )*, 5(1), 50–55.
17. Rayadin, M. A., Musaruddin, M., Saputra, R. A., & Isnawaty, I. (2024). Implementation of Ensemble Learning XGBoost and Random Forest Methods for Battery Replacement Time Prediction. *BIOS: Journal of Information Technology and Computer Engineering*, 5(2), 111–119.
18. Safitri, E., Rofianto, D., Purwati, N., Kurniawan, H., & Karnila, S. (2024). Diabetes Mellitus Disease Prediction using Machine Learning Algorithms. 12(4), 760–766. <https://doi.org/10.26418/justin.v12i4.84620>
19. Salsabila, W. Q., Adyani, K., & Realita, F. (2024). Literature Review: Risk Factors for Polycystic Ovary Syndrome in Adolescents. *Journal of Health (JoH)*, 11(02), 164–174 <https://doi.org/10.30590/joh.v11n2.832>
20. Saputra; Akbar Novan Dwi. (2019). Polycystic Ovary Syndrome (PCOS) in Adolescents. Ministry of Health Sardjito Hospital. <https://sardjito.co.id/2019/09/30/polycystic-ovary-syndrome-pcos-pada-remaja/>
21. Septiana Rizky, P., Haiban Hirzi, R., & Hidayaturohman, U. (2022). Comparison of LightGBM and XGBoost Methods in Handling Data with Unbalanced Classes. *J Statistics: Scientific Journal of Theory and Applications of Statistics*, 15(2), 228–236. <https://doi.org/10.36456/jstat.vol15.no2.a5548>
22. Sitanggang, L. O., & Bahtiar, N. (2019). Application of data mining to detect hepatitis disease using the Polynomial Support Vector Machine (SVM) method (Case Study: Indian Liver Patient Data). *Journal of Informatics Society*, 10(1), 20–27. <https://doi.org/10.14710/jmasif.10.1.31490>
23. WIJAYANTI, N. P. Y. T., N. KENCANA, E., & SUMARJAYA, I. W. (2021). Smote: Potential and Drawbacks on Surveys. *E- Journal of Mathematics*, 10(4), 235. <https://doi.org/10.24843/mtk.2021.v10.i04.p348>
24. Wulandari, S., Mukti, Y. I., & Susanti, T. (2024). Optimization of the Artificial Neural Network Algorithm with Genetic Algorithm in Stroke Prediction. *Synchron*, 8(2), 1056–1063. <https://doi.org/10.33395/sinkron.v8i2.13609>



- 
25. Yanuari, E. D. D., Yudhianto, R. B., Ulfia, R. R., & Sartono, B. (2024). COMPARATIVE STUDY OF SVM-SMOTE AND SMOTE-TOMEK METHODS IN OVERCOMING IMBALANCE CLASS USING THE XGBOOST MODEL ON THE CLASSIFICATION OF KUR RECIPIENT HOUSEHOLDS. 5(3), 2266–2283.
  26. Zheng, J., Li, J., Zhang, Z., Yu, Y., Tan, J., Liu, Y., Gong, J., Wang, T., Wu, X., & Guo, Z. (2023). Clinical Data based XGBoost Algorithm for infection risk prediction of patients with decompensated cirrhosis: a 10-year (2012–2021) Multicenter Retrospective Case-control study. *BMC Gastroenterology*, 23(1), 1–10. <https://doi.org/10.1186/s12876-023-02949-3>