International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 06, Issue 02, June 2025
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Comparative Evaluation of Classification Algorithms for the Diagnosis of Polycystic Ovary Syndrome

Sri Wulandari
Politeknik Indonusa Surakarta
sriwulandari@poltekindonusa.ac.id

Abstract— *Polycystic Ovary Syndrome (PCOS) is a complex hormonal disorder that affects women's reproductive and metabolic health. Early detection is essential to prevent long-term complications. This study aims to analyze and compare the performance of four machine learning classification algorithms, namely Naive Bayes, K-Nearest Neighbor (KNN), Decision Tree, and Support Vector Machine (SVM), in assisting the diagnosis of PCOS based on clinical data. The dataset used contains 1000 patient data with five main features: age, body mass index (BMI), menstrual irregularities, testosterone levels, and antral follicle count. The data were divided using stratified sampling (80:20) and validated using the k-fold cross-validation technique (k=5). Model evaluation used accuracy, precision, recall, F1-score, and AUC metrics. The results showed that Decision Tree had the best performance (100% accuracy, AUC 0.997), followed by SVM (97% accuracy) and KNN (96%). Naive Bayes had the lowest accuracy (72%) and produced many false positives. Although Decision Tree is superior, there is a risk of overfitting, while SVM and KNN show more stable performance. This study confirms that classification algorithms, especially SVM and KNN, are effective for PCOS diagnosis based on clinical data. The practical implication of this finding is the development of accurate and efficient clinical decision support systems to improve women's healthcare.*

Keywords: *PCOS, Naive Bayes, KNN, Decision Tree, SVM.*

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a serious hormonal disorder that is often experienced by women of childbearing age in various parts of the world. This condition is characterized by hormonal imbalance, irregular ovulation, and the appearance of small cysts in the ovaries.[1] In addition to affecting fertility, PCOS is also closely related to metabolic disorders such as insulin resistance, obesity, dyslipidemia, and increased risk of type 2 diabetes and cardiovascular disease.[2]

The global prevalence of PCOS is estimated to range from 8% to 13%, depending on the diagnostic criteria used.[3] However, the actual number is likely higher because many cases go undetected due to varying symptoms and lack of awareness among both patients and healthcare providers. The diagnosis of PCOS generally refers to the Rotterdam criteria which include at least two of three indicators: chronic anovulation, clinical or biochemical hyperandrogenism, and polycystic ovaries detected by ultrasound.[1]

The long-term impact of PCOS is very significant, such as impaired fertility, psychological problems (depression and anxiety), and increased risk of endometrial cancer.[2] Therefore, early detection and accurate diagnostic methods are needed to prevent further complications. Technology, including machine learning approaches, plays an important role in the process of prediction and diagnosis based on medical data.

In today's digital era, machine learning methods have experienced rapid development and have been widely applied in the health sector, especially in predicting and diagnosing various diseases. Various classification algorithms have been proven to be able to recognize patterns from complex medical data.[4] These algorithms can identify patient characteristics and produce diagnostic predictions based on certain attributes. Some commonly used methods include Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression.[5]

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem with the assumption of independence between features. Although this assumption is not always met in real data, this algorithm still shows competitive performance in various medical classification studies due to its simplicity and efficiency in handling high-dimensional data.[6]-[10]

The K-Nearest Neighbors (KNN) method classifies new data based on its proximity to a set of nearest neighbor data. Although it does not require explicit training, the performance of KNN is highly dependent on the choice of k value and the distance measurement method. KNN is quite popular in diagnosis because it is able to capture non-linear relationships between features.[9]-[11]

The Decision Tree algorithm forms a prediction model in the form of a tree structure, where each internal node is a test of a feature, branches are the test results, and leaves represent class labels. This algorithm is intuitive and easy to understand, making it suitable for use in a medical context. However, without pruning techniques, this model is susceptible to overfitting.[7], [9], [12], [13]

Support Vector Machine (SVM) is a margin-based classification algorithm that works by finding the optimal hyperplane to maximally separate two classes. SVM is effective in handling high-dimensional and non-linear datasets with the help of kernel techniques. In the context of disease diagnosis, SVM is often used because of its stability and accuracy.[7], [13]

Although widely applied, evaluation of the performance of each model in the context of PCOS diagnosis is still needed. Therefore, this study aims to apply and compare the performance of four ML classification models—Naive Bayes, KNN, Decision Tree, and SVM—in diagnosing PCOS based on clinical data. It is expected that the results of this study can contribute to the development of a more accurate and efficient clinical decision support system.

II. RESEARCH METHODS

1. **Research Design**
   This study was conducted with a quantitative approach using computational experiments. The purpose of this approach is to comparatively evaluate the performance of several classification algorithms in diagnosing Polycystic Ovary Syndrome (PCOS) based on medical data.

2. **Data and Data Sources**
   The data used is a public dataset on PCOS downloaded from the Kaggle.com site. This dataset includes information from 1000 female patients of reproductive age who experience hormonal disorders. There are five main attributes in this dataset that are often associated with a PCOS diagnosis.

Table 1. PCOS Dataset Variables and Information

Feature	Description
Age	The age range of patients is between 18 to 45 years
BMI	Body Mass Index with a range of 18–35 kg/m²
Menstrual Irregularities	Irregular menstrual cycle (0 = No, 1 = Yes)
Testosterone Levels	Testosterone levels in the patient's blood (20–100 ng/dL)
Antral Follicle Count	Number of follicles detected by ultrasound (5–30)
Target (PCOS Diagnosis)	Patient diagnosed with PCOS (0 = No, 1 = Yes)

3. **Data Pre-processing**
   Data pre-processing stages include:
   - Data cleaning: remove empty data and outlier values
   - Data transformation: normalization or standardization of numeric features
   - Categorical variable encoding: one-hot encoding is performed if necessary.
   - Data division: data is divided into training data and test data with a proportion of 80:20 using stratified sampling method.
4. **Classification Algorithm**The four classification algorithms evaluated in this study are:
   1. Naive Bayes
   2. K-Nearest Neighbor (KNN)
   3. Decision Tree
   4. Support Vector Machine (SVM)

   Each model is trained using training data, then tested using test data.
5. **Model Evaluation**The performance of each algorithm is evaluated using the following metrics:
   1) Accuracy

2) Precision
3) Recall (Sensitivity)
4) F1-Score
5) Area Under the Curve (AUC)

The aim of this evaluation was to determine the best model in classifying PCOS diagnosis.

6. **Model Validation**
To ensure the generalization ability of the model, the k-fold cross-validation technique was used with k = 5. The average of the results from each fold was used to obtain a more stable final evaluation value.

7. **Tools and Programming Languages**
The entire analysis process was performed using Python on the Google Colab platform. Some of the libraries used include Pandas, NumPy, Scikit-learn, and Matplotlib/Seaborn for data analysis, modeling, and visualization of results..

III. RESULT AND ANALYSIS

The dataset used in this study is open data from Kaggle related to PCOS diagnosis. After pre-processing, 1000 clean data were obtained consisting of 801 patients who did not have PCOS and 199 patients diagnosed with PCOS.

Table 2. Polycystic Ovary Syndrome (POCS) Diagnosis Dataset

No	Age	BMI	Men-struation Irre-gularity	Testos-terone Level	Antral Follicle Count	PCOS Diag-nosis
1	24	34.7	1	25.2	20	0
2	37	34.7	0	25.2	25	0
3	32	34.7	0	25.2	28	0
4	28	34.7	0	25.2	26	0
5	25	34.7	1	25.2	8	0
…..	…..	…..	…..	…..	…..	…..
…..	…..	…..	…..	…..	…..	…..
996	34	34.7	1	25.2	23	0
997	45	34.7	1	25.2	7	0
998	37	34.7	0	25.2	28	0
999	41	34.7	0	25.2	9	0
1000	22	34.7	1	25.2	7	0

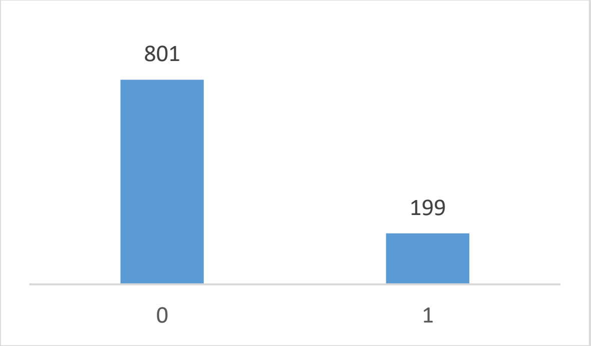
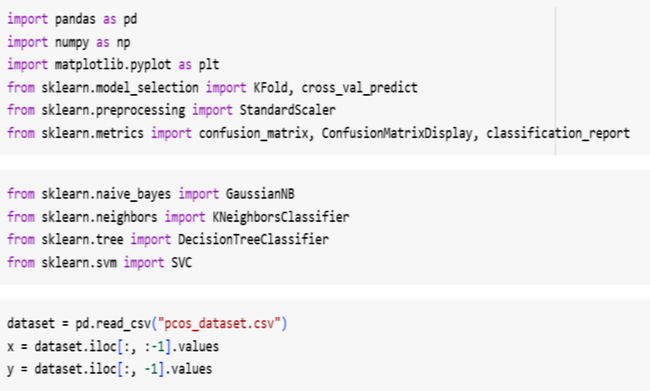


Figure 1. Number of PCOS Diagnosis

Classification analysis was performed using four algorithms: Naive Bayes, KNN, Decision Tree, and SVM, with training and testing data split at 80% and 20%. Implementation was done with Python on the Google Colab platform.



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import KFold, cross_val_predict
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, classification_report
```

```
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
```

```
dataset = pd.read_csv("pcos_dataset.csv")
x = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
```

Figure 2. Script Python on the Google Colab

3.1 Model Performance Evaluation

The classification results of each model produce a confusion matrix which is presented in Figures 3-6 and a summary of model performance assessment metrics in Table 3.

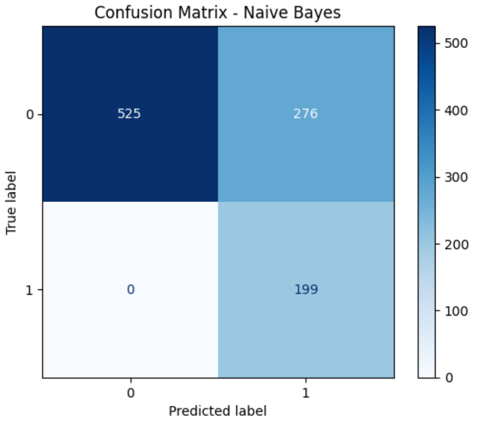


Figure 3. Confusion Matrix Naïve Bayes

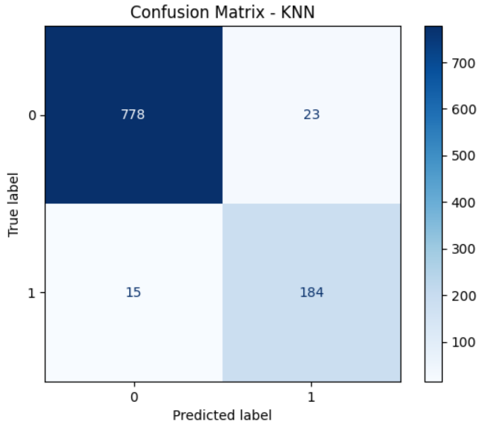


Figure 4. Confusion Matrix KNN

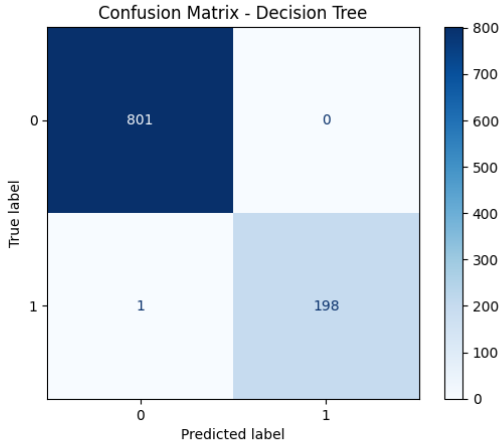


Figure 5. Confusion Matrix Decision Tree

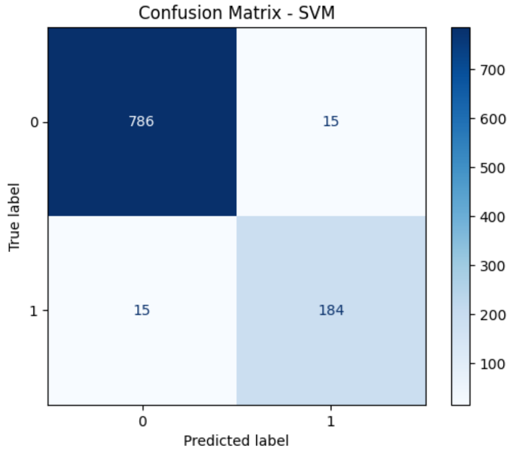


Figure 6. Confusion Matrix SVM

Table 3. Model Evaluation

Metric	Naive Bayes	KNN	Decision Tree	SVM
Pecision	0.71	0.93	1	0.95
Recall	0.83	0.95	1	0.95
F1-score	0.69	0.94	1	0.95
Accuracy	0.72	0.96	1	0.97
AUC	0.987	0.986	0.997	-

The evaluation results show that:

1) Naive Bayes achieved an accuracy of 72% with a precision of 0.71, a recall of 0.83, and an F1-score of 0.69. Although the recall value was quite high, many false positive classification errors occurred, indicating overprediction of PCOS cases. The Naïve Bayes model showed a perfect recall rate (1.00) for detecting the positive class (PCOS), meaning all PCOS cases were successfully identified. However, its low precision value (0.42) indicated many errors in classifying non-PCOS individuals as PCOS patients. This situation can reduce the level of confidence in the prediction results. This algorithm prioritizes sensitivity over accuracy, making it suitable for use when detecting all cases of the disease is a priority, even though it sacrifices accuracy.
2) KNN showed very good results with 96% accuracy, 0.93 precision, 0.95 recall, and 0.94 F1-score. The K-Nearest Neighbor (KNN) algorithm shows stable and proportional performance. The combination of high recall and precision values indicates a low error rate, both in identifying and classifying PCOS and non-PCOS patients. The large F1 score also indicates that this model is quite reliable, making it a good choice for real-world diagnostic applications.
3) Decision Tree obtained perfect results on the test data (accuracy, precision, recall, and F1-score = 1) and an AUC of 0.997. The Decision Tree model produced near-perfect accuracy with only one misclassification. Almost all PCOS and non-PCOS patients were correctly predicted. However, this overly ideal result could be an indication of overfitting, especially if the training and test data are not adequately separated or cross-validation is not performed optimally
4) SVM yielded 97% accuracy, with precision, recall, and F1-score of 0.95 each. The number of misclassifications was relatively low and the model showed balance in predicting the two classes. Support Vector Machine (SVM) shows superior performance with a balance between high precision and sensitivity values. A large F1 value indicates that this model is able to recognize PCOS cases effectively with

a low error rate. This model is very suitable for implementation in the medical field, especially if further optimized, such as by activating the probability option to calculate the AUC value.

Although Decision Tree performed dominantly on all evaluation metrics, additional validation is still needed to ensure the model does not experience overfitting. On the other hand, KNN and SVM showed stable performance and are feasible to apply. Meanwhile, the use of Naive Bayes is recommended only when detecting all PCOS cases is a top priority, although it carries a fairly high risk of positive misclassification

3.2 K-Fold Validation

The validation value of each model with k-fold cross validation is shown in table 4.

Table 4. Model Validation

Fold	Naive Bayes	KNN	Decision Tree	SVM
K=1	0.735	0.965	0.999	0.955
K=2	0.69	0.97	1	0.975
K=3	0.735	0.965	1	0.96
K=4	0.735	0.975	1	0.99
K=5	0.725	0.935	1	0.97
Average	0.724	0.962	0.999	0.97

To test the stability of model performance, five-fold cross validation (k=5) was used. The validation results show that:

1) The performance of Naive Bayes is relatively low with a fairly large variation in accuracy between folds. This instability indicates that the model is less consistent in dealing with variations in training data and may not be able to accommodate the complexity of the relationship between features in PCOS diagnosis. This may be due to the assumption of independence between features which is not always met in medical data.
2) KNN shows consistent performance with high average accuracy and low standard deviation. This model is able to maintain stability across multiple folds, thus providing reliable classification results, especially when the data distribution represents the population proportionally.
3) Decision Tree recorded almost perfect results with very high accuracy values and minimal deviation. This performance indicates the dominance of the model over the training data. However, this also has the potential to indicate overfitting, especially if the model is not equipped with a pruning mechanism or additional validation.
4) SVM performs competitively with high accuracy and low inter-fold fluctuation. This stability indicates that SVM is a robust and effective model for classification. Its advantage in separating classes with maximum margin makes it ideal for applications in clinical diagnosis, especially when features have sharp boundaries between one class and another.

Overall, the results show that although Decision Tree provides the highest accuracy, SVM and KNN models are more stable and tend to have better generalization ability on new data. Meanwhile, Naive Bayes is less suitable for use in this case because of its relatively low performance.

VI. CONCLUSION

Based on the evaluation results, the Support Vector Machine (SVM) model was considered the most balanced in terms of accuracy, precision, recall, and stability of results. SVM showed consistent performance in testing and had a low misclassification rate, making it a reliable choice for PCOS diagnosis. Although Decision Tree showed excellent results on training and testing data, this needs to be further validated using external data to avoid overfitting. On the other hand, KNN also showed very good performance and can be a promising alternative in supporting the diagnosis process.

Naive Bayes is less recommended in this context due to its low accuracy rate and tendency to produce high false positive predictions. In the early diagnosis process, a model with high recall value is very important to avoid undetected cases. Therefore, models such as KNN, SVM, and Decision Tree are more appropriate because they are able to recognize PCOS cases effectively.

The results of this study are expected to provide a real contribution to the development of a more accurate and efficient artificial intelligence-based clinical decision support system, especially for improving women's reproductive health services.

REFERENCES

[1] H. J. Teede *et al.*, "HHS Public Access Author manuscript Fertil Steril. Author manuscript; available in PMC 2020 January 02. Published in final edited form as: Fertil Steril. 2018 August ; 110(3): 364–379. doi:10.1016/j.fertnstert.2018.05.004. Recommendations from the interna," *Heal. Hum. Serv.*, vol. 110, no. 3, pp. 364–379, 2018, doi: 10.1016/j.fertnstert.2018.05.004.Recommendations.

[2] S. Pililis *et al.*, "The Cardiometabolic Risk in Women with Polycystic Ovarian Syndrome (PCOS): From Pathophysiology to Diagnosis and Treatment," *Med.*, vol. 60, no. 10, 2024, doi: 10.3390/medicina60101656.

[3] Y. Che, J. Yu, Y. S. Li, Y. C. Zhu, and T. Tao, "Polycystic Ovary Syndrome: Challenges and Possible Solutions," *J. Clin. Med.*, vol. 12, no. 4, 2023, doi: 10.3390/jcm12041500.

[4] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, 2019, doi: 10.1056/nejmra1814259.

[5] K. Dissanayake and M. G. M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Appl. Comput. Intell. Soft Comput.*, vol. 2021, 2021, doi: 10.1155/2021/5581806.

[6] M. Alwateer, A. M. Almars, K. N. Areed, M. A. Elhosseini, A. Y. Haikal, and M. Badawy, "Ambient healthcare approach with hybrid whale optimization algorithm and Naïve Bayes classifier," *Sensors*, vol. 21, no. 13, pp. 1–21, 2021, doi: 10.3390/s21134579.

[7] T. Muthia and Y. E. Putra, "Perbandingan Akurasi Model Pembelajaran Mesin SVM , KNN , Decision Tree , dan Naive Bayes pada Klasifikasi Gangguan Kesehatan Mental."

[8] D. Derisma and F. Febrian, "Perbandingan Teknik Klasifikasi Neural Network, Support Vector Machine, dan Naive Bayes dalam Mendeteksi Kanker Payudara," *Bina Insa. Ict J.*, vol. 7, no. 1, p. 53, 2020, doi: 10.51211/biict.v7i1.1343.

[9] M. Abdul Jabbar, E. Hasmin, C. Susanto, W. Musu, and I. Artikel, "Komparasi Algoritma Decision Tree, Naive Bayes, dan K-Nearest Neighbors dalam Klasifikasi Kanker Payudara Comparison of Decision Tree Algorithms, Naive Bayes, and K-Nearest Neighbors in Breast Cancer Classification," *Oktober*, vol. 14, no. 3, pp. 258–270, 2022, [Online]. Available: https://www.doi.org/10.22303/csrid.14.3.2022.258-270.

[10] K. Akmal, A. Faqih, and F. Dikananda, "Perbandingan Metode Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 470–477, 2023, doi: 10.36040/jati.v7i1.6367.

[11] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-10358-x.

[12] N. A. Maulidiyyah, T. Trimono, A. T. Damaliana, and D. A. Prasetya, "Comparison of Decision Tree and Random Forest Methods in the Classification of Diabetes Mellitus," *JIKO (Jurnal Inform. dan Komputer)*, vol. 7, no. 2, pp. 79–87, 2024, doi: 10.33387/jiko.v7i2.8316.

[13] M. M. Nishat *et al.*, "A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 7, no. 29, pp. 1–12, 2021, doi: 10.4108/eai.13-8-2021.170671.