KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 4, No. 1, Januari (2023), pp. 73-79

Model Comparison of Random Forest and Logistic Regression Algorithms in PCOS Disease Detection

Khoirun Nisa[1], Purwono[2], Bala Putra Dewa[3], Sony Kartika Wibisono[4]*
[1,2,3,4]Universitas Harapan Bangsa, Purwokerto
Email: [1]khoirunnisa@uhb.ac.id, [2]purwono@uhb.ac.id, [3]bdewa@uhb.ac.id, [4]wibisono.sony@gmail.com

Abstract

PCOS or Polycystic Ovary Syndrome is a hormonal imbalance affecting egg cells' growth, making them remain small and not develop into large and mature egg cells to be fertilized by sperm cells. It is an endocrinopathy disease occurred in 10-15% of productive-aged women worldwide. The study aims to find the most suitable algorithm to be used in the optimization of PCOS detection. Thus, a performance comparison between random forest and logistic regression algorithms needs to be conducted in order to find the best performance in terms of accuracy. The research used a dataset containing 40 features. According to comparison results, the random forest algorithm was superior to logistic regression, with an accuracy of 91 %.

Keywords: *PCOS, Random Forest, Logistic Regression*

1. Introduction

PCOS or Polycystic Ovary Syndrome is a hormonal imbalance affecting egg cells' growth, making them remain small and not develop into large and mature egg cells to be fertilized by sperm cells. It is an endocrinopathy disease occurred in 10-15% of productive-aged women worldwide [1]. The World Health Organization (WHO) predicted that 117 million women had suffered PCOS or approximately 3.5% of the population of women worldwide [2]. Judging from the high number, PCOS has become a big problem affecting not only the health condition of PCOS patients but also the social views and judgments of their personalities [3]. PCOS is caused by abnormal metabolism of androgen and estrogen and abnormal production control of androgen. PCOS can also be associated with peripheral insulin resistance and hyperinsulinemia, causing suppression of the production of sex hormone-binding globulin (SHBG) to increase androgens [4]. Moreover, women with ovary dysfunction are associated with hypertension, increased risk of cardiovascular diseases, obesity, gynecological cancer, type-2 diabetes mellitus, and a higher risk of miscarriage in the first trimester of pregnancy [5]. Therefore, preliminary detection of PCOS disease is vital to prevent health risk complications.

Along with rapid technological advancement, technology applications can be helpful for early disease detection. Technology also improves many aspects of health sectors, such as treatment management, medical assistance, health information systems, teleconsultations, and the prevention of many diseases. Some researchers have proposed many technologies for diagnosing PCOS. For example, a certainty factor with five parameters was applied to diagnose PCOS, resulting in sufficient accuracy [6]. Meanwhile, three novel classification methods in ultrasound images were proposed to detect PCOS follicles [7]. The three methods were based on: Neural Network-Learning Vector Quantization (LVQ) method, KNN - euclidean distance, and Support Vector Machine (SVM). The best accuracy in the study was obtained from the SVM algorithm, which was 82.55%. In [8], the backpropagation algorithm was modified to detect PCOS and was performed in ultrasound images.

In this study, a classification system was built to predict and analyze whether someone is a PCOS patient or not. Random forest and logistic regression algorithms were both applied in PCOS detection, and their performance results would be compared. Random

ISSN: 2720-992X

forest is an algorithm in machine learning used to classify large amounts of data [9]. A recent study in [10] found that the random forest algorithm resulted in an accuracy of 98.2% while being used for medicine classification. A classification based on the random forest algorithm was also applied in [11] to detect diseases in rice cultivars. Dataset used in this study had three classes. Meanwhile, logistic regression describes data that explains the relationship between the dependent binary variable and the nominal, ordinal, interval, and high-level independent variables. Another study in [12] found that the logistic regression method had an accuracy of 87.50%. Moreover, a prediction based on logistic regression to detect diabetes performed in [13] had 82% accuracy.

2. Research Methods

2.1. Dataset

This research used a dataset taken from the Kaggle repository. The dataset contains 44 features, including age, body weight, body height, blood type, pulse rate, Hb, LH, FSH, HIP, AMH, hip size, waist size, PRL, Vitamin D3 content, hair-fall condition, fast food consumption, weight gain, acne, and hair follicles in particular spots. As many as 541 records were collected from the dataset, which was grouped into two categories: 364 were normal records, and 177 were PCOS patients' records, as can be seen in Table 1. Some features or attributes of the data samples contained missing values; thus, the preprocessing stage is crucial to solving the missing values issue.

Table 1. Data Samples

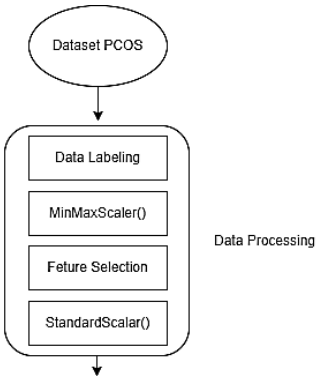
	PCOS (Y/N)	Age (yrs)	Weight (Kg)	Height (Cm)	BMI	…..	Reg.Exercise (Y/N)	BP _Systolic (mmHg)	BP _Diastolic (mmHg)	Follicle No. (L)
0	0.0	28.0	44.6	152.0	19.300.000	…..	0.0	110.0	80.0	3.0
1	0.0	36.0	65.0	161.5	24.921.163	…..	0.0	120.0	70.0	3.0
2	1.0	33.0	68.8	165.0	25.270.891	…..	0.0	120.0	80.0	13.0
…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..
541	0.0	25.0	52.0	161.0	20.060.954	…..	0.0	120.0	80.0	3.0

2.2. Methods

This research employed model comparison by applying the random forest and the logistic regression algorithms. The programming language in the machine learning implemented in this study was Phyton. Concisely, the research conducted three main contributions as follows:

a) Applying machine learning algorithms to important features of the PCOS dataset;
b) Model comparison, especially the accuracy of each applied algorithm.

The proposed research consists of three modules: image preprocessing methods, data split using K-Fold Cross Validation, and data classification using random forest and logistic regression algorithms, which were evaluated using the confusion matrix for model comparison. The block diagram of the proposed systems is presented in Figure 1 below.



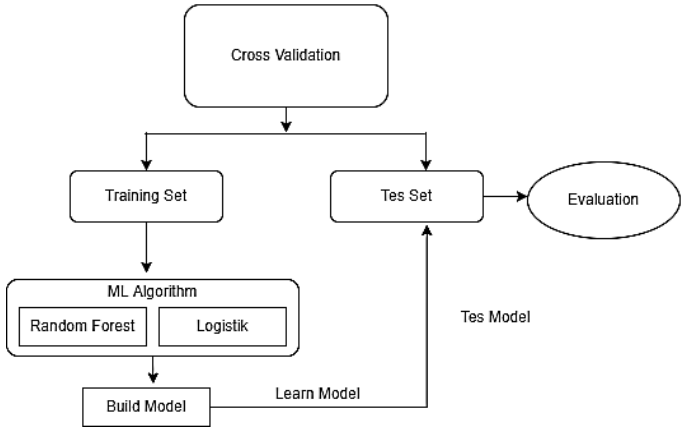


Figure 1. Research block diagram

According to Figure 1, this research has several methodological steps. The research began with inputting features and attributes of the PCOS dataset. Then, in data processing, data labeling and feature selection were performed on the data. The processed data was then continued with the cross-validation process. The data was then ready, and training would be performed using random forest and logistic regression. The training results were modeled and then evaluated using the confusion matrix.

2.2.1. Preprocessing

Preprocessing is an important step in data mining. The data that will be used for processing are not always in the best condition. Sometimes there were issues in the dataset that would affect the result of the mining process, i.e., missing values, data overload, outliner, or incompatible data format. Therefore preprocessing stage is necessary. It can compensate for these troublesome issues, and good performance of data classification can be obtained.

The dataset was divided into two, which are training and testing data. The holdout method used train_test_split() function in the scikit-learn library was applied to divide the dataset. Data separation to obtain training and testing data was done using the validation method with cross_val_score() function taken from the same library, scikit-learn. Then, the data records were categorized into two classes, namely 0 and 1, where 0 was used for non-PCOS women (normal) and 1 for PCOS women. The standard transformation formula is shown in equation (1).

$$p' = \frac{P - Pmin}{Pmax - Pmin} \tag{1}$$

Here, P denotes the original value of the feature, $Pmin$ is the minimum value, $Pmax$ is the maximum value, and p' is the normalization value. MinMaxScale was used to change the feature by scaling each feature. The feature scale in data processing becomes an important step in combining independent variables or feature ranges.

2.2.2. Random Forest Algorithm

The random forest approach proposed by Breiman is a machine learning algorithm with many decision trees. Random forest is a combination of Bagging methods and Random Subspaces. This method has proven its success in regression and classification problems in recent years and is one of the best machine learning algorithms used in various fields [14]. Dibandingkan dengan algoritma lain seperti jaringan syaraf tiruan dan

support vector machine, random forest memiliki parameter yang lebih sedikit ditentukan saat berjalan. Kumpulan pengklasifikasi yang diatur oleh pohon individu dapat direpresentasikan seperti dalam persamaan berikut [15].

$$\{RF(y, a_n), n = 1,2, \dots i, \dots\} \quad (2)$$

Di mana, RF adalah pengklasifikasi, $\{a_n\}$ singkatan dari identik independen mendistribusikan vektor acak, dan setiap decission tree memiliki suara untuk yang paling terkenal kelas pada variabel input y. Sifat dan dimensi tergantung pada proses yang digunakan pada pembangunan decission tree.

2.2.3. Logistic Regression Algorithm

Logistic regression is a part of statistics that is also commonly called the generalized linear model. The logistic regression model was used when the response variable referred to two values. For instance, logistic regression is used when a subject is an inanimate object or non-living thing, whether it has unique characteristics, etc. The response variable is assumed as y; then, a subject/event is assumed as y=1 if it has a particular characteristic, and y=0 when it does not have the characteristic [16]. The formula equation for logistic regression is expressed in equation (2).

$$F(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

2.2.4. Confusion Matrix

The confusion matrix is a matrix that shows the number of correct and incorrect predictions made by the classification model compared to the actual results (target values) in the data. The matrix has $n \times n$ size, where n denotes the number of the target values (class). The confusion matrix used in the research and its data description is provided in Table 2 below.

Table 2. Confusion Matrix

		Actual	
		Not Churn	Churn
Prediction	Not Churn	a	b
	Churn	c	d

Description:
TP (True Positive) = d/(c+d) x 100%
FP (False Positive) = b/(a+b) x 100%
TN (True Negative) = a/(a+b) x 100%
FN (False Negative) = c/(c+d) x 100%
Precision = d/(b+d) x 100%
Accuracy = (a+d)/(a+b+c+d) x 100%

3. Result And Discussion

The research used a dataset collected from Kaggle. The dataset was then processed in data preprocessing and separated in the cross-validation. The logistic regression algorithm was applied to the dataset that had been processed and separated. Then, the first experiment's results would be evaluated by its accuracy and Area Under Curve (AUC) score as performance results. Adapun beberapa perubahan transformasi. The classification results using logistic regression are shown in Table 3.

.

Table 3. Classification Results of Logistic Regression

	class	precision	recall	f1-score	support
Non-PCOS women	0.0	0.87	0.89	0.88	109
PCOS women	1.0	0.77	0.74	0.75	54
Accuracy		84%			163

Meanwhile, the resulting confusion matrix of applying logistic regression is shown in Figure 2.

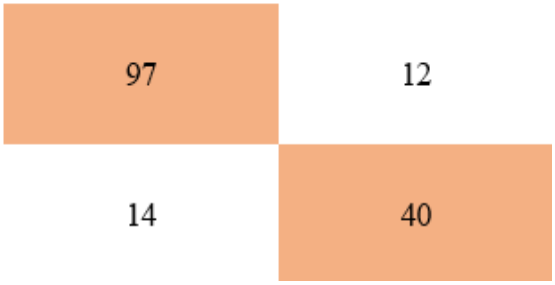


Figure 2. Confusion Matrix of Logistic Regression

According to the data results in Table 3, the accuracy can be calculated by adding diagonally-aligned values and then dividing the result by the total number of testing data used in the experiment, which can be expressed as (97+40)/163 = 84%. In comparison, the performance results of applying random forest for data classification are shown in Table 4.

Table 4. Classification Results of Random Forest

	class	precision	recall	f1-score	support
Non-PCOS women	0.0	0.90	0.97	0.93	107
PCOS women	1.0	0.93	0.76	0.84	51
Accuracy		91%			158

The resulting confusion matrix of applying the random forest algorithm for data classification can be seen in Figure 3 below.

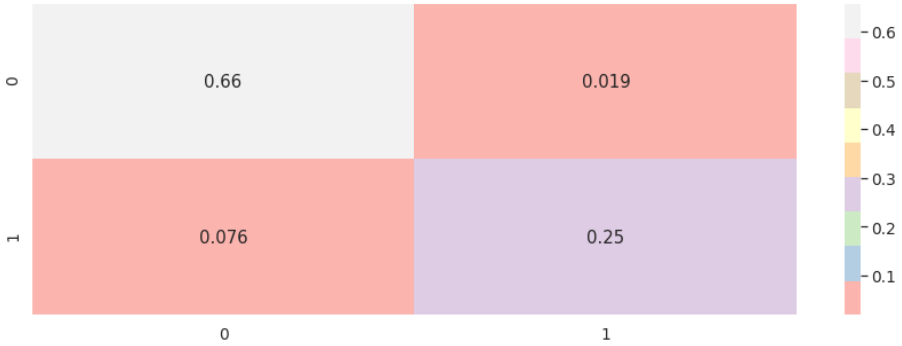


Figure 3. Confusion Matrix of Random Forest

4. Conclusions

Based on experiment results, it can be concluded that random forest algorithm could classify data for PCOS detection with promising accuracies using 30% testing data ratio. In both implementation of random forest and logistic regression algorithms, with 41 data attributes and parameter specifications as param_grid=parameters, hidden_layer_sizes = (15.5), the resulting accuracy was considerably good with >80% accuracy. The accuracy

of the random forest algorithm in PCOS detection using dataset classification was 91%, more prominent than the logistic regression with 84% accuracy. The expected further work is an application of parametric optimization using other machine learning methods and adding more testing data so that the detection system can perform with the best optimal results.

References

[1] E. A. Greenwood, L. A. Pasch, K. Shinkai, M. I. Cedars, and H. G. Huddleston, "Clinical course of depression symptoms and predictors of enduring depression risk in women with polycystic ovary syndrome: Results of a longitudinal study," *Fertil. Steril.*, vol. 111, no. 1, pp. 147–156, 2019, doi: 10.1016/j.fertnstert.2018.10.004.

[2] T. Vos, A. Flaxman, and M. Naghavi, "HHS Public Access Global Burden of Disease Study 2010," *Lancet*, vol. 380, no. 9859, pp. 2163–2196, 2012, doi: 10.1016/S0140-6736(12)61729-2.Years.

[3] M. O. Goodarzi, D. A. Dumesic, G. Chazenbalk, and R. Azziz, "Polycystic ovary syndrome: etiology, pathogenesis and diagnosis," *Nat. Rev. Endocrinol.*, pp. 219–231, 2011, doi: https://doi.org/10.1038/nrendo.2010.217.

[4] M. A. Sanchez-Garrido and M. Tena-Sempere, "Metabolic dysfunction in polycystic ovary syndrome: Pathogenic role of androgen excess and potential therapeutic strategies," *Mol. Metab.*, vol. 35, no. February, p. 100937, 2020, doi: 10.1016/j.molmet.2020.01.001.

[5] A. S. Laganà, S. G. Vitale, M. Noventa, and A. Vitagliano, "Current management of polycystic ovary syndrome: From bench to bedside," *Int. J. Endocrinol.*, vol. 2018, 2018, doi: 10.1155/2018/7234543.

[6] Fitri Handayani, A. Fauzi, and A. Sihombing, "Penerapan Metode Certainty Factor dalam Mendiagnosa Penyakit Kanker Nasofaring," *J. Pharm. Heal. Res.*, vol. 6, no. 1, pp. 255–262, 2020, [Online]. Available: http://ejurnal.seminar-id.com/index.php/jharma/article/view/345%0Ahttp://ejurnal.seminar-id.com/index.php/jharma/article/download/345/215.

[7] B. Purnama, U. N. Wisesti, Adiwijaya, F. Nhita, A. Gayatri, and T. Mutiah, "A classification of polycystic Ovary Syndrome based on follicle detection of ultrasound images," 2015, doi: 10.1109/ICoICT.2015.7231458.

[8] U. N. Wisesty, J. Nasri, and Adiwijaya, "Modified Backpropagation Algorithm for Polycystic Ovary Syndrome Detection Based on Ultrasound Images," in *Recent Advances on Soft Computing and Data Mining*, 2017, pp. 141–151.

[9] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 187–192, 2021, doi: 10.29207/resti.v5i1.2813.

[10] P. Purwono, A. Wirasto, and K. Nisa, "Comparison of Machine Learning Algorithms for Classification of Drug Groups," *Sisfotenika*, vol. 11, no. 2, p. 196, 2021, doi: 10.30700/jst.v11i2.1134.

[11] A. Purnamawati, W. Nugroho, D. Putri, and W. F. Hidayat, "InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan Attribution-NonCommercial 4.0 International. Some rights reserved Deteksi Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree, Random Forest, Naïve Bayes, SVM dan KNN," vol. 5, no. 1, 2020, [Online]. Available: https://doi.org/10.30743/infotekjar.v5i1.2934.

[12] J. J. Pangaribuan, H. Tanjaya, and Kenichi3, "Mendeteksi Penyakit Jantung Menggunakan Machine Learning Dengan Algoritma Logistic Regression," *Inf. Syst. Dev.*, vol. 6, no. 2, 2021, [Online]. Available: https://books.google.ca/books?id=EoYBngEACAAJ&dq=mitchell+machine+learning+1997&hl=en&sa=X&ved=0ahUKEwiomdqfj8TkAhWGslkKHRCbAtoQ6AEIKjAA.

[13] Erlin, Yulvia Nora Marlim, Junadhi, Laili Suryati, and Nova Agustina, "Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 11, no. 2, pp. 88–96, 2022, doi: 10.22146/jnteti.v11i2.3586.

[14] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *Sistemasi*, vol. 10, no. 1, p. 163, 2021, doi: 10.32520/stmsi.v10i1.1129.

[15] I. H. Hassan, M. Abdullahi, M. M. Aliyu, S. A. Yusuf, and A. Abdulrahim, "An improved binary manta ray foraging optimization algorithm based feature selection and random forest classifier for network intrusion detection," *Intell. Syst. with Appl.*, vol. 16, no. August, p. 200114, 2022, doi: 10.1016/j.iswa.2022.200114.

[16] T. M. Jawa, "Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia," *Alexandria Eng. J.*, vol. 61, no. 10, pp. 7995–8005, 2022, doi: 10.1016/j.aej.2022.01.047.