International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-31, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

Hyperparameter Tuning of Random Forest Algorithm for Diabetes Classification

Fiqi Nur Rohman[1], Farikhin[2], Bayu Surarso[3]
[1,2,3]Faculty of Science and Mathematics, Diponegoro University

ABSTRACT: This study aims to optimize the hyperparameters of the Random Forest model in diabetes classification using the Pima Indian Diabetes dataset, given the importance of early diabetes diagnosis to mitigate serious health impacts. While Random Forest is a popular algorithm for classification due to its resistance to overfitting, the selection of the right hyperparameters significantly affects its performance. Therefore, this research utilizes Grid Search and Random Search techniques for hyperparameter tuning to improve model accuracy. The research methodology includes data collection, preprocessing, dataset splitting (80% for training and 20% for testing), feature scaling using Standard Scaler, and the application of the Random Forest algorithm with hyperparameter tuning and model evaluation based on accuracy, precision, recall, and F1-Score. The results show that Random Forest, when tuned with Grid Search and Random Search, significantly improved model performance, with Random Search yielding the best results, achieving an accuracy of 0.75, precision of 0.64, and recall of 0.69. This study demonstrates that hyperparameter tuning can significantly enhance the performance of the Random Forest model, contributing to the development of machine learning applications for medical diabetes diagnosis.

KEYWORDS: Classification, Diabetes, Hyperparameter Tuning, Machine Learning, Random Forest.

INTRODUCTION

Diabetes is a chronic disease of global concern, as it can lead to severe complications such as blindness, kidney failure, heart attacks, and even death. According to the International Diabetes Federation (IDF), over 90% of diabetes cases worldwide are classified as Type 2 diabetes, which occurs when the body fails to respond effectively to insulin or cannot produce enough insulin. This condition not only poses a significant threat to individual health but also imposes a substantial economic burden on healthcare systems globally (IDF Diabetes Atlas, 2021).

Early and accurate diagnosis of diabetes is crucial to mitigating its adverse effects. With advancements in technology, machine learning has emerged as a promising tool for automating disease diagnosis and classification. Among various machine learning algorithms, Random Forest has gained popularity for classification tasks due to its robustness, ability to handle high-dimensional data, and resistance to overfitting. However, the performance of a Random Forest model largely depends on the careful selection of hyperparameters, such as the number of trees (n_estimators), maximum tree depth (max_depth), and other configuration parameters.

While several studies have utilized Random Forest for diabetes prediction, the systematic optimization of hyperparameters is often overlooked or conducted manually, potentially limiting the model's performance. Therefore, this study aims to optimize the hyperparameters of the Random Forest model using Grid Search and Random Search techniques to enhance the accuracy of diabetes classification based on the Pima Indian Diabetes dataset.

This research seeks to provide insights into the influence of individual hyperparameters on model performance and identify the best combination of hyperparameters for improved predictive accuracy. Furthermore, the findings of this study can serve as a reference for advancing the application of machine learning in medical diagnostics.

RESEARCH METHOD

The stages of the research method consist of several steps, starting from data collection, preprocessing, dataset splitting, feature scaling, algorithm implementation, hyperparameter tuning, and evaluation. Preprocessing is a technique in data mining aimed at transforming raw data into a structured and comprehensible format [1]. Raw data is often incomplete or contains errors, so during preprocessing, incomplete data is removed from the dataset. After preprocessing, the next step is splitting the dataset into two parts: training data (train) and testing data (test). The split is typically done with a proportion of 80% for training data and 20% for testing

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-31, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

data. Following this, feature scaling is performed using the Standard Scaler method. Standard Scaler is a feature standardization technique that removes the mean and normalizes variance, ensuring no data points have values significantly larger than others [2]. After feature scaling, the subsequent steps involve applying algorithms, performing hyperparameter tuning, and evaluating the model. Figure 1 visually illustrates the research stages.

Figure 1 Research Flow

A. Data

The dataset utilized in this study is the Pima Indians Diabetes dataset, comprising data from 768 patients [3]. It contains several variables that represent various aspects of individual health information. The "Pregnancies" column records the number of pregnancies experienced by the patient, while the "Glucose" column documents the plasma glucose concentration measured two hours after an oral glucose tolerance test. Diastolic blood pressure, measured in mm Hg, is included in the "BloodPressure" column, and the "SkinThickness" column notes the triceps skinfold thickness in millimeters. The "Insulin" column records the two-hour serum insulin levels, and the "BMI" column provides the Body Mass Index value. The dataset also includes the "DiabetesPedigreeFunction" column, which represents a score estimating the likelihood of diabetes based on family history. Patient age in years is recorded in the "Age" column. Finally, the "Outcome" column indicates the presence or absence of diabetes, with a value of 1 denoting the patient has diabetes and 0 indicating otherwise.

B. Random Forest

Random Forest is a high-performing, scalable, and user-friendly classification model in machine learning. Given a dataset $S = \{(x_i, t_i)\}_{i=1}^{m}$ where each (x_i, t_i) consists of the feature vector x_i, represented by m features, and the target variable t_i, which is categorical, Random Forest functions as an ensemble model. This ensemble comprises multiple decision trees, with each tree trained on a bootstrapped subset of the data obtained through random sampling. For classifying a new data point *x*, the class label is determined using the equation:

$$y(x) = mode\{C_1(x), C_2(x), \dots, C_m(x)\}$$

Where *m* is the number of decision trees in the Random Forest, and mode selects the class with the highest frequency among the outputs of all decision trees.

In the case of binary classification, where the target variable $t_1 \in \{-1,1\}$, the decision is made based on the following rule:

$$(x) = sign\left[\sum_{i=1\to m} C_i(x)\right] = f(x) = \begin{cases} 1, & \sum_{i=1\to m} C_i(x) \geq 0 \\ -1, & otherwhise \end{cases}$$

The core concept of the ensemble model is to combine weak learners (models with limited predictive power) into a strong learner (a model with significantly improved performance). Random Forest leverages this principle to achieve robust and accurate classification by aggregating the predictions of individual decision trees [4].

C. Grid Search

Grid search is one of the most commonly used methods for hyperparameter tuning exploration. This method works by exhaustively searching through all combinations of hyperparameters predefined in the grid configuration. One advantage of grid search is its ability to run in parallel, where each experiment operates independently without being affected by execution order. As a result, the outcome of one experiment does not depend on the results of others. Additionally, grid search offers high flexibility in allocating computational resources [5].

D. Random Search

Random search is a fundamental improvement over grid search. This method involves randomly exploring hyperparameters using a specified distribution of possible parameter values. One of the key advantages of random search is its ability to run in parallel on computer architectures, enabling efficient and fast processing. Additionally, it provides resilience in handling system failures during experiments. With its flexibility and various strengths, random search proves to be an effective and reliable method for hyperparameter tuning [6].

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-31, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

E. Evaluation

The dataset is divided into 80% training data (614 samples) and 20% testing data (154 samples). To ensure the model does not overfit, a 10-fold cross-validation method is applied. The evaluation metrics used include accuracy, precision, recall, and F1-score. Additionally, a ROC curve analysis is conducted to compare the performance of the Random Forest model before and after hyperparameter tuning in distinguishing between positive and negative classes [7]. Metrics such as accuracy, precision, recall, and F1-score are employed to measure the model's performance [8].

RESULT

These results of experiment have been computed using Python programming on a specific computer setup and environment, which included the following specification

Table 1. System and Environment Specifications

Processor	AMD Ryzen 5
RAM	16 GB DDR4
Operating System	Windows 11
Python Environment	Google Colab (default python 3.10)

A. Exploration Data Analyst

The process of data exploration aims to gain a deep understanding of the data while identifying relationships between variables. In this context, descriptive statistics play a significant role by providing data summaries such as mean, median, and standard deviation, as well as presenting data visualizations like boxplots [9].

The descriptive statistics of this dataset provide an overview of information about diabetic patients. It includes various measured variables such as the number of pregnancies, blood glucose levels, blood pressure, and others. The types of descriptive statistics provided are count, mean, standard deviation (std), minimum value (min), first quartile (25%), median (50% or second quartile), third quartile (75%), and maximum value (max). Refer to Table 1 below.

Table 2. Description of Variable Statistics in The Diabetes Dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
Count	768	768	768	768	768	768	768	768	768
Mean	3.85	120.89	69.11	20.53	79.80	31.99	0.47	33.24	0.35
Std	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11,76	0.47
Min	0	0	0	0	0	0	0.07	21	0
25%	1	99	62	0	0	27.3	0.24	24	0
50%	3	117	72	23	30.5	32	0.37	29	0
75%	6	140.25	80	32	127.25	36.6	0.62	41	1
Max	17	199	122	99	846	67.1	2.42	81	1

Table I provides an overview of various measured variables, such as the number of pregnancies, blood glucose levels, blood pressure, skinfold thickness at the tricep area, insulin levels in the serum, body mass index (BMI), family history of diabetes, patient age, and the target variable indicating whether the patient is diagnosed with diabetes or not. Each statistic, such as mean, standard deviation, quartiles, minimum value, and maximum value, offers valuable insights into the distribution and characteristics of the data, which can be used for further analysis and understanding of the dataset.

Additionally, data visualization plays a crucial role in data analysis, as it helps uncover patterns, trends, and anomalies that may not be immediately apparent from raw numbers. In the context of diabetes prediction using machine learning, data visualization can be employed to explore the distribution of variables, relationships between them, and differences in characteristics between individuals with and without diabetes. The relationships between several variables in diabetes are illustrated in Figures (2) and (3).

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-31, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

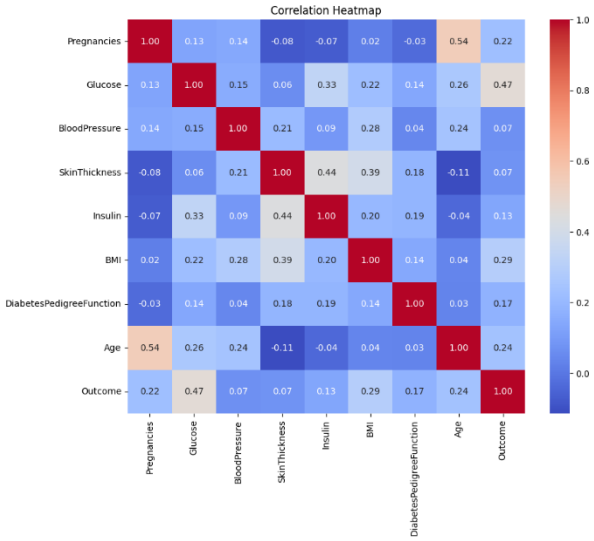

Figure 2 Correlation Values and Heatmap Between Variables

Figure 2 illustrates the relationships between variables. Correlation is used to assess the strength and direction of the linear relationship between two variables, with values ranging from -1 to 1. The analysis reveals that blood glucose levels (Glu) have a significant positive correlation with diabetes risk (Outcome), with a correlation value of 0.47. This indicates that higher glucose levels are associated with a greater likelihood of diabetes. Additionally, age (Age) shows a moderate positive correlation with diabetes risk (0.24), suggesting that the risk increases with age. BMI has a weaker positive correlation (0.29), indicating that a higher BMI may be linked to an increased risk of diabetes, though not as strongly as glucose or age. Other factors, such as family history of diabetes (DPF), show a lower positive correlation (0.17), suggesting an influence on diabetes risk, though not as strong as variables like glucose and age. Next, refer to Figure 3 below.

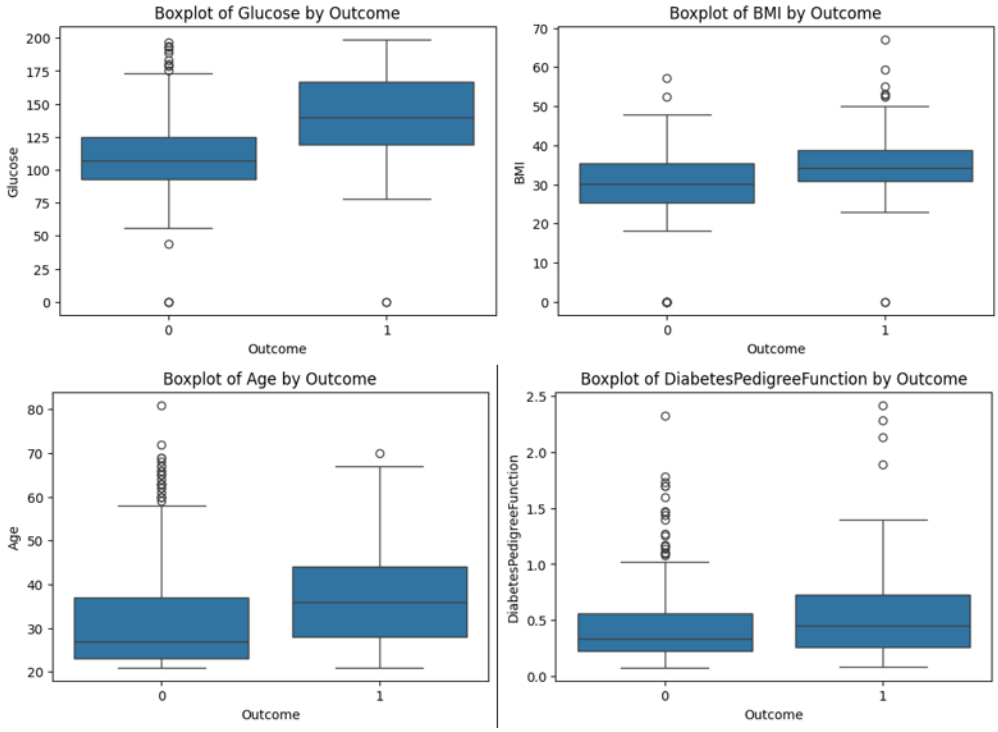

Figure 3 Boxplot of the Relationship Between Variables (Glucose, BMI, Age, and DPF)

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-31, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

Figure 3 displays the relationship between several variables (Glucose, BMI, Age, and DPF) and the outcome variable (diabetes risk) in the diabetes dataset. For example, the comparison between Glucose and Outcome shows a significant difference in glucose levels between the non-diabetic (0) and diabetic (1) groups. The median glucose level in the diabetic group is higher, indicating a tendency for higher glucose levels in this group, with some outliers recording extremely high or low glucose levels. BMI also shows a difference, although not as pronounced as glucose, with the median in the diabetic group slightly higher and a wider range of values, along with notable outliers. Age, as shown in the boxplot, has a higher median in the diabetic group, highlighting age as a significant risk factor. DPF, reflecting a family history of diabetes, also shows a clear difference between groups, with a higher median in the diabetic group and a more varied range of values. Overall, this boxplot illustrates that Glucose, BMI, Age, and DPF have significant relationships with diabetes risk, with the diabetic group tending to have higher values for these variables compared to the non-diabetic group.

B. Model Prediction

The dataset is divided into a training set of 80% and a testing set of 20%. The Random Forest classification model is then trained using the training set, and its performance is evaluated using the testing set. Below are the results of the confusion matrix.

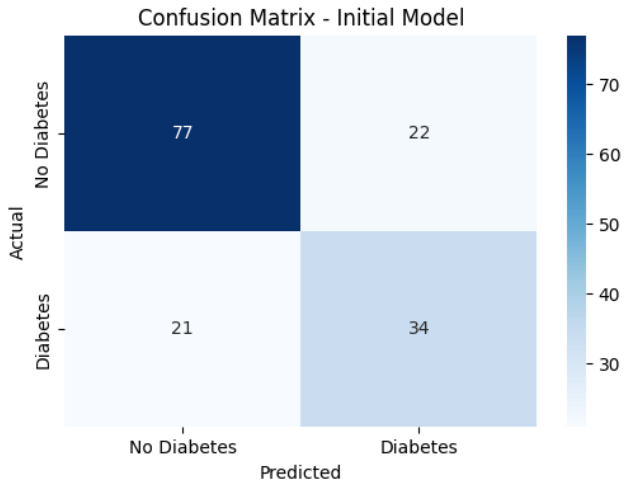

Figure 4 Random Forest Confusion Matriks

Figure 4 displays the confusion matrix for the Random Forest model. This confusion matrix provides an overview of how the model classifies data into two classes: "No Diabetes" and "Diabetes". In the confusion matrix for Random Forest, it is shown that the model correctly classifies 77 data points as "No Diabetes" and 34 data points as "Diabetes". However, the model also produces 22 false positive errors (No Diabetes classified as Diabetes) and 21 false negative errors (Diabetes classified as No Diabetes). The total classification errors for this model are 43.

Hyperparameter Tuning will be performed to enhance the performance of the Random Forest model. This process aims to identify the optimal combination of parameters, such as the number of trees (n_estimators), tree depth (max_depth), and other parameters, to improve the accuracy of the model's predictions. By optimizing the hyperparameters, the goal is to reduce overfitting or underfitting, thereby achieving better performance on unseen data. Techniques like grid search or random search will be employed to find the best parameter combination by evaluating the model's performance across different hyperparameter sets.

C. Hyperparameter Tuning Performance

Hyperparameter Tuning will be performed to enhance the performance of the Random Forest model. This process aims to identify the optimal combination of parameters, such as the number of trees (n_estimators), tree depth (max_depth), and other parameters, to improve the accuracy of the model's predictions. By optimizing the hyperparameters, the goal is to reduce overfitting or underfitting, thereby achieving better performance on unseen data. Techniques like grid search or random search will be employed to find the best parameter combination by evaluating the model's performance across different hyperparameter sets. Below are the results of the confusion matrix.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-31, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

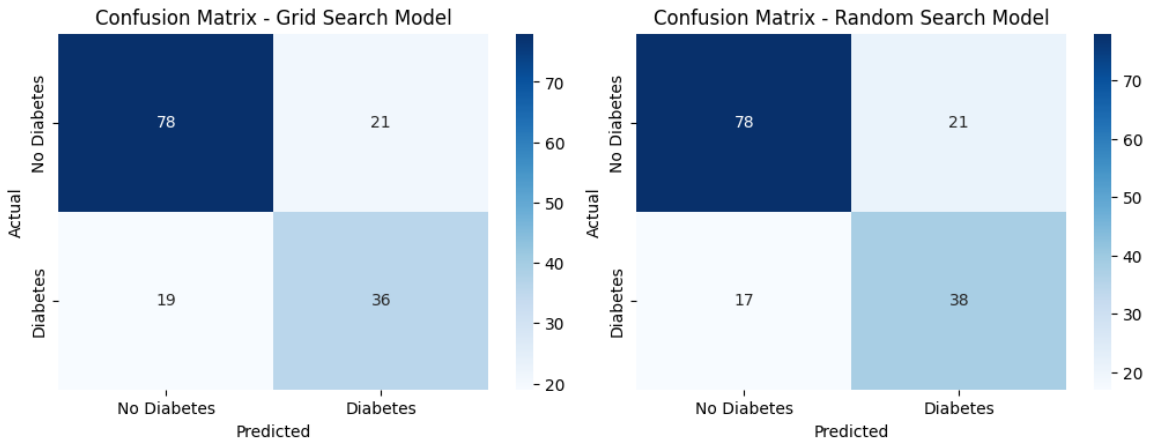

Figure 5 Grid Search and Random Search Confusion Matriks

Figure 5 displays the confusion matrix of Grid Search and Random Search as the results of hyperparameter tuning. Grid Search correctly classified 78 data points as "No Diabetes" and 36 data points as "Diabetes." However, Grid Search also resulted in 21 false positive errors and 19 false negative errors. The total errors for Grid Search are 40. Random Search correctly classified 78 data points as "No Diabetes" and 38 data points as "Diabetes." However, Random Search also resulted in 21 false positive errors and 17 false negative errors. The total errors for Random Search are 38.

In classification analysis, the Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) are essential tools for evaluating model performance. The figure below shows the ROC Curve for the Random Forest model and the results of Hyperparameter Tuning using Grid Search and Random Search. The AUC for each model is calculated based on its ability to predict whether an individual has diabetes or not. The ROC Curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) for each model. TPR (or sensitivity) measures the proportion of true positives correctly identified, while FPR measures the proportion of negatives incorrectly identified as positives (false positives). The higher the TPR and the lower the FPR, the better the model's performance. Refer to the figure below for further analysis.

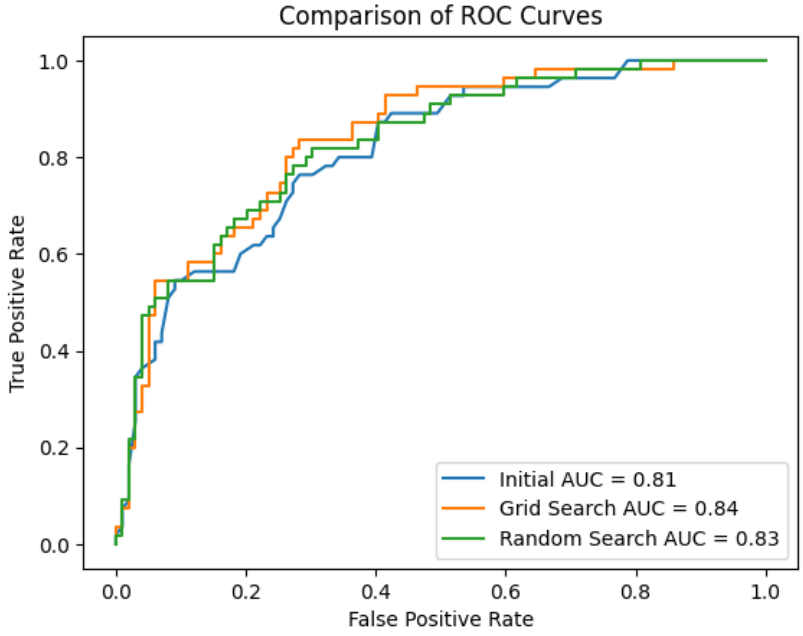

Figure 6 ROC Curve for Random Forest and Hyperparameter Tuning Using Grid Search and Random Search

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-31, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

Figure 6 shows a comparison of ROC curves for the Random Forest model, including the initial model, the results from tuning with Grid Search, and the results from tuning with Random Search. The ROC curve for the initial model shows an AUC of 0.81, indicating decent performance but with room for improvement. After hyperparameter tuning with Grid Search, the AUC increased to 0.84, showing a significant improvement in the model's ability to predict diabetes. Meanwhile, Random Search yielded an AUC of 0.83, slightly lower than Grid Search but still showing an improvement over the initial model. Overall, these results suggest that both Grid Search and Random Search effectively enhanced the model's performance, with Grid Search contributing more significantly to the improvement in the model's accuracy.

D. Model Performance

After performing Hyperparameter Tuning using Grid Search and Random Search, the dataset is divided into a training set of 80% (614 data) and a testing set of 20% (154 data). The Random Forest model is trained using the training data, and its performance is evaluated using the testing data. The performance evaluation results of the Random Forest model after tuning with Grid Search and Random Search are shown in Table II.

Table 3. Hyperparameter Tuning Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.72	0.60	0.61	0.61
Grid Search	0.74	0.63	0.65	0.64
Random Search	0.75	0.64	0.69	0.66

Table II shows good performance in predicting diabetes, with an accuracy above 70%. Although all three models show good accuracy, there are some classification errors that need further analysis. The Random Forest model, after hyperparameter tuning with Grid Search and Random Search, has a fairly high precision of 0.60, 0.63, and 0.64, indicating that both tuning techniques successfully improved the model's accuracy in classifying patients with diabetes. However, there is a relatively high false positive rate in Random Forest, meaning some patients who should be classified as "no diabetes" are instead categorized as "diabetes." This error can occur due to high glucose or BMI values in patients without diabetes. To reduce this error, threshold tuning on the model's prediction probability can be applied, making the model more cautious in classifying someone as diabetic based solely on insufficiently significant indicators. An alternative would be to add other features that could provide more context to the data, such as patients' diet patterns and physical activity.

In addition, Grid Search and Random Search show an improvement in recall compared to the initial model, with Grid Search having a recall of 0.65 and Random Search 0.69. This indicates that the hyperparameter tuning techniques helped the model become more sensitive in detecting patients who truly have diabetes. To further improve recall and reduce false negatives, oversampling techniques or the use of SMOTE (Synthetic Minority Over-sampling Technique) can be applied to address the class imbalance in the positive class, making the model more sensitive to diabetes cases.

Overall, the use of Grid Search and Random Search for hyperparameter tuning on the Random Forest model successfully improved the model's performance, with Random Search providing the best results in terms of precision and recall.

CONCLUSION

This research focuses on Hyperparameter Tuning of the Random Forest algorithm for diabetes disease classification using the Pima Indians Diabetes dataset. The study shows that Hyperparameter Tuning on the Random Forest algorithm can improve the model's performance in diabetes disease classification. The analysis results indicate that the use of Grid Search and Random Search for hyperparameter tuning successfully improved the performance of the Random Forest model compared to the initial model. Although Random Forest has an accuracy of 0.72, Grid Search and Random Search provide higher accuracies of 0.74 and 0.75, respectively. In addition, both Grid Search and Random Search also show improvements in precision and recall, with Random Search yielding the best results with a precision of 0.64 and recall of 0.69.

This research has limitations, including a relatively small dataset size, which may affect the model's generalization. Additionally, the data used only includes basic medical variables without considering other lifestyle factors that may be influential,

such as diet and physical activity. These limitations may lead to a lack of sensitivity in the model to variations in broader individual health conditions.

For future research, it is recommended to use a larger and more diverse dataset to improve model generalization. Further hyperparameter tuning techniques, as well as the exploration of other optimization methods such as Grey Wolf Optimization, are also recommended to enhance model performance. Additionally, the use of techniques for handling imbalanced data, such as SMOTE, may help reduce classification errors in groups with fewer sample sizes.

REFERENCES

1. Lingga Aji A, Pratiwi Amalia Nur A, Respatiwulan R. Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifiere. *Indones J Appl Stat*. 2019;2(1):34-41.
2. Purwanto A, Masduki A, Fahlevi M, et al. Impact ofWork From Home(WFH) on Indonesian Teachers Performance During the Covid-19 Pandemic : An Exploratory Study. *Int J Adv Sience Technol*. 2020;29(5):6235-6244.
3. Jack S, BS, JE E, MD, MPH, WC Dickson. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In: *Annual Symposium on Computer Applications in Medical Care*. ; 1998:261-265.
4. Heryadi Y, Wahyono T. *Machine Learning Konsep Dan Implementasi*. Cetakan I. (Turi, ed.). Gava Media; 2020.
5. Yu T, Zhu H. Hyper-Parameter Optimization: A Review of Algorithms and Applications. *Arxiv*. Published online 2020:1-54.
6. Andonie R. Hyperparameter optimization in learning systems. *J Membr Comput*. 2019;1(4):279-291.
7. Pavel L, Patrick D, Christin S, Rieck K. Learning Instruction Detection: Supervised or Unsupervised? In: *Lectures Note in Computer Science*. ; 2005:50-57.
8. Pratap C Sen, Mahimarnab H, Mitadru G. Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: *Advances in Intelligent System and Computing*. ; 2020:99-111.
9. Shreffler J, Heucker MR. *Exploratory Data Analysis: Frequencies, Descriptive Statistics, Histograms, and Boxplots*. StatPearls Publishing; 2023.
10. Ali M, Haider MN, Lashari SA, Sharif W, Khan A, Ramli DA. Stacking Classifier with Random Forest functioning as a Meta Classifier for Diabetes Diseases Classification. *Procedia Comput Sci*. 2022;217(1877-0509):3453-3462.
11. Jain N, Jana PK. LRF: A logically randomized forest algorithm for classification and regression problems[Formula presented]. *Expert Syst Appl*. 2023;213(September 2021). doi:10.1016/j.eswa.2022.119225
12. Liaw A, Wiener M. Classification and Regression by Random Forest. *R News*. 2022;2(3):18-22.
13. Guo Z, Guo R, Lin S. Multi-factor fuzzy prediction model of concrete surface chloride concentration with trained samples expanded by random forest algorithm. *Mar Struct*. 2022;86(September):103311. doi:10.1016/j.marstruc.2022.103311
14. Biau G, Scornet E. Rejoinder on: A random forest guided tour. *Test*. 2016;25(2):264-268. doi:10.1007/s11749-016-0488-0
15. Breiman L. Random Forest. *Mach Learn*. 2001;45(1):5-32.
16. Billah M, Islam AKMS, Mamoon W Bin, Rahman MR. Random forest classifications for landuse mapping to assess rapid flood damage using Sentinel-1 and Sentinel-2 data. *Remote Sens Appl Soc Environ*. 2023;30(February):100947. doi:10.1016/j.rsase.2023.100947
17. Roberts JF, Mwangi R, Mukabi F, et al. Pyeo: A Python package for near-real-time forest cover change detection from Earth observation using machine learning. *Comput Geosci*. 2022;167(February):105192. doi:10.1016/j.cageo.2022.105192
18. Han X, Zhu X, Pedrycz W, Li Z. A three-way classification with fuzzy decision trees. *Appl Soft Comput*. 2023;132:109788. doi:10.1016/j.asoc.2022.109788