MATICS : Jurnal Ilmu Komputer dan Teknologi Informasi ISSN : 1978-161X(p); 2477-2550(e)
Volume 17, No. 2(2025), pp 63-66 DOI : 10.18860/mat.v17i2. 30256
Received : March 28th 2025; **Accepted** : August 28th 2025 ; **Avalaible Online** : September 29th 2025

Comparison of Linear Regression, Decision Tree Regression, and Random Forest Regression Algorithms in Predicting Baldness Risk

Sebastianus Adi Santoso Mola, Alfonsus Maria De Liguori Goru, Christian Jaquelino Lamapaha, Yoseph Kurubingan Bekayo

Abstract— **Baldness is a prevalent condition affecting both men and women, influenced by various factors including age, hormonal fluctuations, and genetic predispositions. Accurate prediction of baldness risk is crucial for early diagnosis and effective prevention strategies. This study evaluates the performance of three regression algorithms: Linear Regression (LR), Decision Tree Regression (DTR), and Random Forest Regression (RFR) in forecasting baldness risk. Additionally, the study incorporates advanced methods such as Gradient Boosting Regressor (GBR), Multi-Layer Perceptron Regressor (MLP), and Extreme Gradient Boosting (XGBoost) for comparative analysis. Utilizing a dataset comprising 5,925 samples with variables such as age, gender, stress levels, and lifestyle factors, the models were assessed based on Mean Squared Error (MSE) and R-squared (R²) metrics. The results demonstrate that XGBoost significantly outperformed all other models, achieving the lowest MSE of 0.0840 and the highest R² of 0.9190, followed by Gradient Boosting and Random Forest Regression. These findings underscore the efficacy of advanced ensemble learning techniques and neural networks in managing complex datasets for precise predictions.**

Index Terms—**Baldness, Linear Regression, Decision Tree Regression, Random Forest Regression**

I. Introduction

H serves a vital function in protecting the scalp from sun exposure and is an important aesthetic feature for both men and women. However, issues such as hair loss (effluvium) and baldness (alopecia) are prevalent concerns [1]. Baldness, particularly when triggered by age, hormonal changes, and genetic predispositions, can lead to permanent hair loss and is more frequently observed in men. Epidemiological data indicate that approximately 50% of men aged 50 years and 25% of men under the age of 21 experience some form of baldness, with the majority of cases attributed to Androgenetic Alopecia (AGA) [2].

Predicting the risk of baldness is crucial for early diagnosis and preventive treatment. Data-driven technologies, such as data mining, present opportunities to uncover patterns that can enhance diagnostic accuracy and inform treatment strategies. This study aims to compare the performance of three regression algorithms—Linear Regression (LR), Decision Tree Regression (DTR), and Random Forest Regression (RFR)—in predicting baldness risk, with the goal of identifying the most accurate and effective method. In addition to these three algorithms, advanced regression models such as Gradient Boosting Regression (GBR), Multi-Layer Perceptron Regression (MLP), and Extreme Gradient Boosting (XGBoost) will also be employed for comparative analysis.

Numerous studies have explored the efficacy of LR, DTR, and RFR methods in various predictive contexts. Research has demonstrated that the RFR algorithm often outperforms other methods in terms of accuracy and error reduction. For instance, studies [3], [4], [5] have shown that RFR excels in predicting house prices, estimating information system project cost budgets, and forecasting housing prices in Boston, respectively.

Conversely, other studies have indicated that DTR can be superior in certain scenarios. For example, research [6] and [7] highlight instances where the DTR algorithm provides more accurate predictions compared to its counterparts. In the study referenced in [6], the DTR algorithm demonstrated significant advantages in specific applications, suggesting that its performance may vary depending on the dataset and context of the

Manuscript received March 28, 2025. (Write the date on which you submitted your paper for review.) This work was supported in part by Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana.

Sebastianus Adi Santoso Mola, Ilmu Komputer, Fakuktas Sains dan Teknik, Universitas Nusa Cendana email: adimola@staf.undana.ac.id

Yoseph Kurubingan Bekayo, Ilmu Komputer, Fakuktas Sains dan Teknik, Universitas Nusa Cendana email: bekayokurnia@gmail.com

Alfonsus Maria De Liguori Goru, Ilmu Komputer, Fakuktas Sains dan Teknik, Universitas Nusa Cendana email: marfingoru@gmail.com

Christian Jaquelino Lamapaha, Ilmu Komputer, Fakuktas Sains dan Teknik, Universitas Nusa Cendana email: christianlamapaha2711@gmail.com

analysis. In research [6], the DTR algorithm demonstrated exceptional performance in predicting rice prices, achieving a remarkable accuracy of 100% across various test data ratios: 90:10, 80:20, 70:30, and 60:40. Additionally, study [7], which aimed to predict the composite stock price index, revealed that the DTR algorithm produced the lowest Mean Squared Error (MSE) value of 1268.242. Furthermore, other studies have indicated that the LR method also exhibits strong performance, as evidenced in [8] and [9]. Research. Research [8], which focused on predicting TikTok music trends, found that the LR method achieved the lowest MSE, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), with an RMSE of 0.177 and an MAE of 0.118, thereby establishing it as the most effective model for the dataset. Similarly, research [9], which examined the prediction of online motorcycle taxi transactions, demonstrated that the LR method outperformed the RFR model in terms of prediction error metrics, yielding a lower RMSE of 1.6, an MSE of 2.6, and a reduced Mean Absolute Percentage Error (MAPE).

Overall, these findings indicate that no single algorithm is universally superior across all scenarios. Random Forest Regression generally yields the best results for complex datasets characterized by non-linear patterns, while Linear Regression demonstrates superior performance in contexts involving simpler and more interpretable data. Recognizing the strengths of each regression method in different applications, this study aims to compare the performance of LR, DTR, and RFR in predicting baldness risk.

II. Research Methods

This research follows a systematic approach comprising several stages to predict the risk of baldness utilizing Linear Regression, Decision Tree Regression, and Random Forest Regression methods. The detailed stages of the research process are illustrated in Figure 1. Each stage is designed to ensure a comprehensive analysis and accurate prediction of baldness risk, facilitating a robust comparison among the selected regression methods.

A. Data sources and types

This research utilizes data sourced from Kaggle [10], a platform that provides publicly available datasets for research and machine learning applications. The datasets were selected based on their relevance to the research objectives and the completeness of the variables necessary for analysis.

The type of data employed in this study is secondary data, which refers to information that has already been collected and made available by other parties, thus facilitating immediate use for further analysis. This dataset encompasses variables such as age, gender, occupation, and other relevant factors that contribute to the prediction of baldness risk.

B. Research steps

The stages of this research involve a systematic series of steps aimed at comparing the performance of LR, DTR, and RFR algorithms in predicting the risk of baldness.

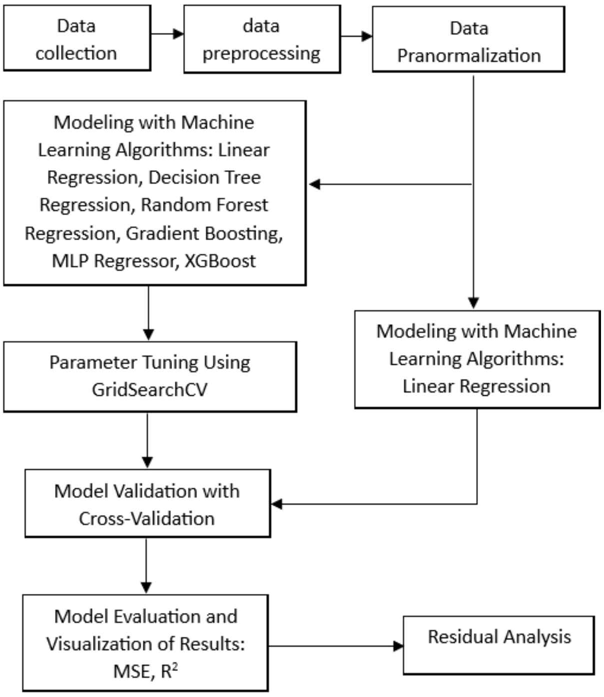
Data collection
data preprocessing
Data Pranormalization
Modeling with Machine Learning Algorithms: Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting, MLP Regressor, XGBoost
Modeling with Machine Learning Algorithms: Linear Regression
Parameter Tuning Using GridSearchCV
Model Validation with Cross-Validation
Model Evaluation and Visualization of Results: MSE, R²
Residual Analysis

Figure 1.Steps of the Research Method

This research employs an experimental approach, utilizing modelling and performance evaluation of regression models through machine learning techniques. The specific steps involved in the research are:

a. Data Collection

The data utilized in this research is represented by input variables (features) that are employed to predict target or output values (target variables). The dataset is subsequently divided into two subsets: a training set comprising 80% of the overall data and a testing set consisting of the remaining 20%. This division is executed using the train-test split method, with the split performed randomly to ensure a balanced distribution between the two subsets.

b. Data preprocessing

The data preprocessing steps undertaken in this research are as follows:

1. **Removal of Irrelevant Columns**: Columns containing provincial data were excluded from the dataset, as they were deemed unnecessary for the analysis and prediction models.
2. **Handling Missing Values**: Data entries with blank or NaN values were removed to ensure that all inputs could be processed by the model without interference.
3. **Conversion of Data Types**: Columns with object data types (string or categorical) were converted to numeric data types through appropriate encoding or mapping techniques, facilitating their integration into the predictive models.

c. Data Normalization

Normalization is employed to standardize the dataset, ensuring that it maintains a mean of zero and a standard deviation of one. This process is essential in machine learning modeling, as it helps to prevent the dominance of one variable over others that may operate on different scales. By normalizing the data, the model can more effectively learn from all input features, leading to improved performance and accuracy in predictions.

d. Modeling with Machine Learning Algorithms

The model used in this study consists of several regression algorithms:

- LR: This algorithm is designed to identify a linear relationship between independent variables (features) and dependent variables (targets). The relationship is mathematically represented by a straight line equation, as shown in [11]:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

  where: y is the target variable (the output to be predicted), X_i is the independent variable, β_0, β_1 …, β_n are regression coefficients (showing the contribution of each feature), ϵ is the error or residual that represents the variation not explained by the model.
- DTR: This algorithm partitions data into nodes based on input variables to predict target values. Its working principle involves repeatedly dividing the data using "if-then" rules, resulting in a tree structure [12]. This tree structure consists of:
  - **Root Node**: The initial node that encompasses all the data.
  - **Decision Nodes**: Internal nodes that separate the data based on specific conditions.
  - **Leaf Nodes**: The terminal nodes that provide the predicted target value for the group of data associated with that node.
- RFR: This model comprises a large number of decision trees, where each tree is trained on a randomly drawn subset of data (with replacement) from the overall dataset [13]. This ensemble approach enhances predictive accuracy and robustness by aggregating the results of multiple trees, thereby reducing the risk of overfitting.

e. Parameter Tuning Using Grid Search

The Grid Search method is employed to optimize the parameters of the DTR and RFR models. Grid Search operates by exploring various combinations of these parameters and selecting the combination that yields the lowest MSE value on the training data. This systematic approach ensures that the models are fine-tuned for optimal predictive accuracy.

f. Model Validation with Cross-Validation

The cross-validation method, utilizing either 10 or 20 folds, is employed to assess model performance across different subsets of the data. The primary objective of this technique is to mitigate overfitting and achieve more stable estimates of model performance. During each fold, the MSE and R² metrics are calculated to capture the variation in model performance, providing a comprehensive evaluation of the algorithms' effectiveness across the dataset.

g. Model Evaluation and Visualization of Results

Once the best model is obtained from Grid Search, it is then evaluated on testing data. Some of the metrics calculated include:

- MSE is a metric used to measure the average squared difference between predicted and actual values in a dataset. MSE calculates the magnitude of the prediction error on average and magnifies larger errors due to the use of squares, The MSE formula is as follows [17]:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  where n is the amount of data in the dataset, y_i is the actual value for the i-th data, and $\hat{y}_i$ is the predicted value for the i-th data.
- R², or the coefficient of determination, is a metric that indicates what proportion of the variation in the dependent variable can be explained by the regression model. R² measures how well the model predicts or explains the observed data, The R squared formula is [18]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

  Where y_i is the actual value for the i-th data, $\hat{y}_i$ is the predicted value for the i-th data, $\bar{y}$ is the average of the actual values, and n is the amount of data.

  The evaluation results are displayed as per-fold graphs for each model, showing the MSE and R² for each fold.

h. Residual Analysis

The residuals of the test data predictions are plotted for each model to analyze the prediction errors. Residual plots serve as a diagnostic tool to identify any systematic errors within the models that may require correction. By examining these plots, researchers can detect patterns or trends in the residuals, which may indicate issues such as non-linearity or heteroscedasticity that could affect the model's predictive accuracy.

i. Conclusion

After evaluating the performance of each model using various metrics, researchers can select the best-performing model for predicting the target variable, focusing on achieving the lowest MSE and the highest R² values.

III. Results and Discussion

A. Dataset

The baldness prediction data used is obtained from the Kaggle website. The total amount of data is 7917 data.

Table 1.Sample of Baldness Prediction Data from Kaggle

	Age	Gender	Job_role	Province	salary	Is_married	Is_Hereditary	Weight	Height	Shampo	Is_smoker	Education	Stress	Bald_prob
0	27	Female	Government Employee	Bengkulu	7957452.757417303	1	0	54.315053	170.428542	Pantone	1	Bachelor Degree	5	**0.605974**
1	53	Female	Government Employee	Bandung	7633002.755486635	1	0	72.873404	165.530097	Pantone	0	Bachelor Degree	7	**0.532860**
2	37	Female	Employee	Bandung	6637624.864428122	1	0	46.321533	154.599388	Moonsilk	0	Bachelor Degree	4	**0.418442**
3	36	Female	Jobless	Palu	3624871.391361162	1	0	51.539781	167.340481	Deadboy	1	Elementary School	9	**0.804050**
4	38	Male	NaN	Palangkaraya	6031807.52048343	1	0	60.726909	165.514773	Merpati	1	Magister Degree	1	**0.368371**

Table 1 is a detail related to the prediction data of the possibility of baldness with features including age, gender, job role, province, salary, is married, is hereditary, weight, height, shampoo, is smoker, education, and stress.

Table 2 is a detailed dataset used in testing the data mining model. At this stage, the missing value and province columns were removed, so that the total data became 5925, with 12 features.

Table 2.Sample of data after removing missing values and province column

	Age	Gender	Job_role	salary	Is_married	Is_Hereditary	Weight	Height	Shampo	Is_smoker	Education	Stress	Bald_prob
0	27	0	1	7957452.757417303	1	0	54.315053	170.428542	3	1	0	5	**0.605974**
1	53	0	1	7633002.755486635	1	0	72.873404	165.530097	3	0	0	7	**0.532860**
2	37	0	0	6637624.864428122	1	0	46.321533	154.599388	2	0	0	4	**0.418442**
3	36	0	2	3624871.391361162	1	0	51.539781	167.340481	0	1	2	9	**0.804050**
4	55	0	1	9213032.163715528	1	1	54.287045	179.235145	3	0	0	1	**0.732562**

Table 3 is the result of the baldness prediction dataset that has been normalized, the normalized data is divided into training data and testing data. The training data used is 80%, namely 4740 data and 20% testing data, namely 1185 data.

Figure 2 shows the distribution of baldness by sex. Of the total data, there were 4069 data on male sex and 3176 of them experienced baldness. Meanwhile, the data on the female sex amounted to 1856 and 1045 of them experienced baldness. Overall, this graph shows that baldness is more common in men than women.

Figure 3 shows the distribution of baldness by age. Based on the figure, baldness is more common at the age of 30 to 59 years. Meanwhile, children, adolescents, and the elderly over 60 years old have fewer cases of baldness.

Table 3.Sample of data after normalization

Age	Gender	Job_role	salary	Is_marr	Is_Heredi	Weight	Height	Shampo	Is_smok	Education	Stress	Bald_prob

					ied	tary				er			
0	-1.270961	-1.480648	0.731369	-0.194745	0.140689	-0.507373	-0.163588	0.296347	0.70326	1.006604	-0840844	-0.156204	**0.044364**
1	1.361530	-1.480648	0.731369	-0.267228	0.140689	-0.507373	1.790828	-0.153664	0.70326	-0.993439	-0840844	0.543877	**-0.396413**
2	-0.258464	-1.480648	-0.9740088	-0.489598	0.140689	-0.507373	-1.005401	-1.157846	0.0000	-0.993439	-0840844	-0.506244	**-1.086194**
3	-0.359714	-1.480648	2.436746	-1.162657	0.140689	-0.507373	-0455857	0.012653	-1.40562	1.006604	0.069859	1.243959	**1.238490**
4	1.564029	-1.480648	0.731369	0.085756	0.140689	1.970935	-0.166537	1.105392	0.70326	-0.993439	-0840844	-1.556366	**0.807515**

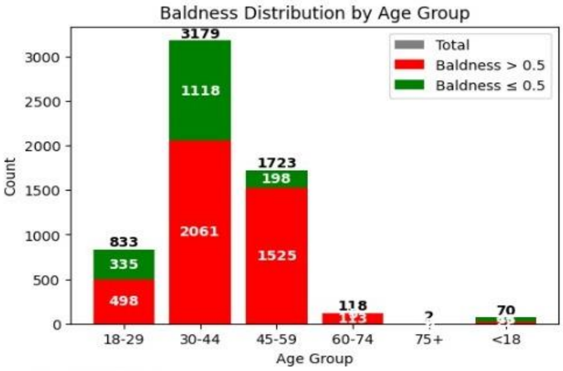

Figure 2. Distribution of baldness by gender

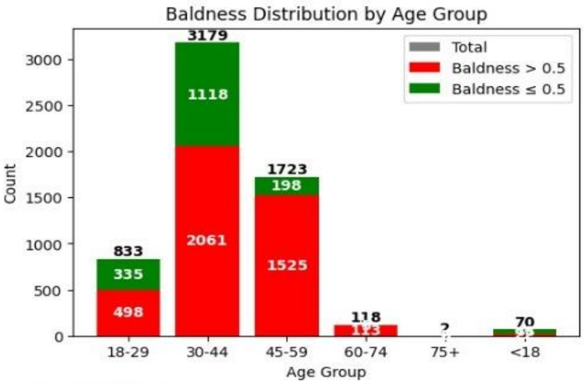

Figure 3. Distribution of baldness based on age

B. Prediction Model Testing

In the model testing process, modeling is conducted using LR, DTR, and RFR. The objective of this research includes the processing of datasets through the testing of various methods to develop prediction models that achieve the highest accuracy. Additionally, the evaluation of error rates for each testing method is a critical aspect of the analysis. The methodologies employed in this research encompass Grid Search for hyperparameter optimization, Cross Validation for assessing model performance, and Residual Analysis for diagnosing prediction errors.

- Grid Search

In this reseach, Grid search is used to find the best (optimal) parameter combination that maximizes model performance. During this search process, K-Fold Cross-Validation is used by Grid Search to evaluate the model performance for each hyperparameter combination. Thus, Grid Search automatically uses K-Fold Cross-Validation on each tested hyperparameter combination to generate an average evaluation score.

Grid Search is employed on tree-based models such as DTR and RFR. The parameters optimized for these models include:

- **max_depth**: Controls the maximum depth of the tree, helping to prevent overfitting.
- **min_samples_leaf**: Specifies the minimum number of samples required to be at a leaf node, which can also help in reducing overfitting.
- **min_samples_split**: Determines the minimum number of samples required to split an internal node, influencing the model's complexity.

For the RFR model, an additional hyperparameter, **n_estimators**, is included to control the number of trees in the forest, thereby impacting both the complexity and performance of the model.

Table 4. Hyperparameters in DTR and RFR

Hyperparameter	**DTR**	**RFR**
max_depth	5,10,15,20	5, 10, 15, 20
min_samples_leaf	1,2,4	1,2
min_samples_split	2, 5, 10	2, 5
n_estimator	-	50, 100, 150

The optimal hyperparameters obtained from the Grid Search for the models are as follows: For the DTR, the optimal parameters are a maximum depth of 10, a minimum samples leaf of 4, and a minimum samples split of 10. In contrast, for the RFR, the optimal hyperparameters include a maximum depth of 10, a minimum samples leaf of 2, a minimum samples split of 5, and an n_estimators value of 150.

- Cross Validation

Cross-validation is a statistical technique employed to evaluate model performance by partitioning the dataset into multiple folds. In this research, the dataset is divided into 20 subsets (folds). During each iteration, one of the subsets is designated as the test data (validation set), while the remaining subsets serve as the training set. This process is repeated 20 times, ensuring that each fold acts as the test data once. An evaluation score is calculated for each iteration, and the average score is often used to provide an indication of the model's performance.

Table 5. Average MSE and R^2 for LR, DTR and RFR

Method	Average		Time
	MSE	R^2	(s)
LR	0.2074	0.7920	0.27
DTR	0.2085	0.7898	0.98
RFR	0.1703	0.8286	95.07

From Table 5, it is evident that RFR method yields the best prediction results, exhibiting the lowest MSE value and the highest R² value when compared to both LR and DTR methods. This superior performance is consistent across both the training and testing processes, highlighting the effectiveness of the RFR model in accurately predicting the target variable. However, when considering the time efficiency, LR demonstrates a clear advantage due to its simplicity and faster computation. In contrast, the RFR method requires significantly more time to execute, making it less suitable for scenarios where computational efficiency is a priority. This trade-off between accuracy and processing time is an important consideration in model selection.

As a comparison of model performance, we also explore more complex models based on boosting and neural networks. The methods included are GBR, MLP Regressor, and XGBoost, as presented in Table 6. The results indicate that the boosting methods generally provide superior prediction accuracy compared to LR and DTR. However, it is important to note that these boosting methods tend to be slower in execution than both LR and DTR, highlighting a trade-off between predictive performance and computational efficiency.

Table 6. Average MSE and R^2 GBR, MLP Regressor, and XGBoost

Method	Average		Time
	MSE	R^2	(s)
GBR	0.1526	0.8465	40.04
MLP Regressor	0.1763	0.8228	59.81
XGBoost	0.1520	0.8471	3.36

Figures 4 and 5 illustrate the MSE and R² graphs for all methods evaluated. From both graphs, the superiority of XGBoost is clearly evident across all folds, as indicated by its consistent low MSE and high R² values. The RFR method follows closely in second place, demonstrating strong performance in nearly all folds. In contrast, Linear Regression exhibits the worst performance, characterized by higher errors and lower R² values. This visual representation reinforces the findings from the numerical evaluations, highlighting the effectiveness of these models in predicting the target variable.

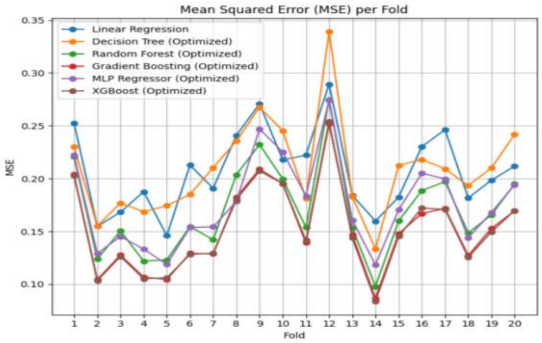

Figure 4. MSE result with Cross Validation

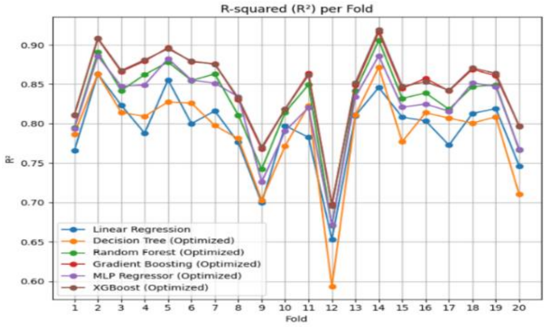

Figure 5. R^2 result with Cross Validation

- Residual Analysis

The Residual Analysis method is employed to assess the extent to which the developed model can effectively capture patterns in the data. Residuals, defined as the differences between actual values and predicted values, play a crucial role in this analysis. By examining the residuals, researchers can identify patterns that the model may not have captured or detect potential issues with the model's assumptions. Residuals are calculated by subtracting the predicted values from the actual values. In this analysis, the dataset is divided into 80% training data and 20% testing data, allowing for a comprehensive evaluation of the model's performance.

Table 5.Results of training and testing residual analysis

Method	Training		Testing	
	MSE	R^2	MSE	R^2
LR	0.2062	0.7964	0.1924	0.7972
DTR	0.1096	0.8918	0.1971	0.7923
RFR	0.0863	0.9147	0.1579	0.8336
GBR	0.1298	0.8719	0.1369	0.8557

MLP Regressor	0.1565	0.8455	0.1521	0.8398
XGBoost	0.1258	0.8757	0.1343	0.8584

Table 9 presents the results of the training and testing phases of the overall residual analysis for all models utilized in this research, including LR, DTR, RFR, GBR, MLP Regressor, and XGBoost. Additionally, graphical representations of the residual analysis are illustrated in Figures 6 to 11, providing a visual insight into the performance and error patterns of each model.

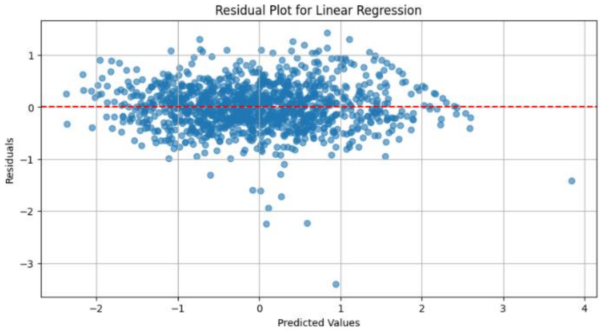

Figure 6. LR Results with Residual Analysis

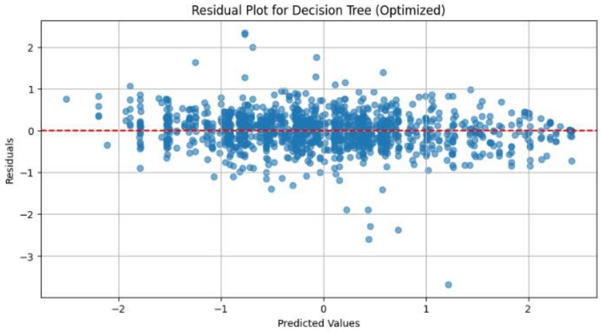

Figure 7. DTR Results with Residual Analysis

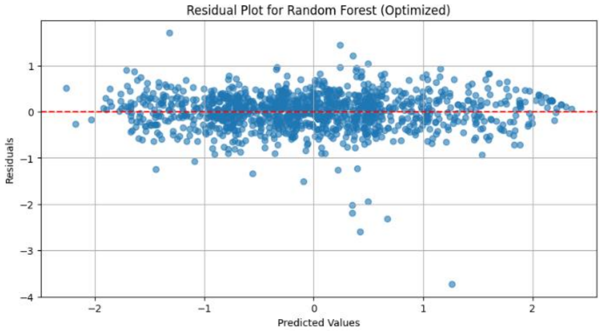

Figure 8. RGR Results with Residual Analysis

From the LR, DTR, and RGR models, as illustrated in Figures 6 to 8, it is evident that the LR model exhibits a more dispersed visualization of its actual values. Conversely, the RGR model demonstrates a closer alignment with the actual values. This observation is consistent with the finding that RGR has the smallest MSE among the three models.

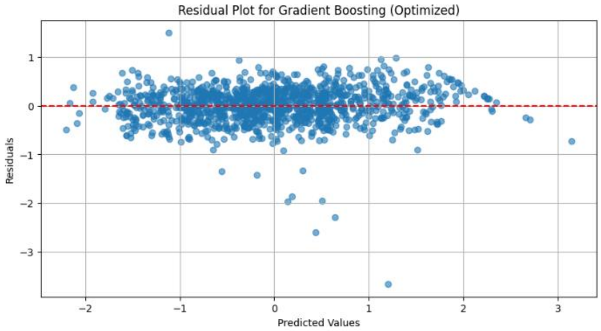

Figure 9. GBR with Residual Analysis

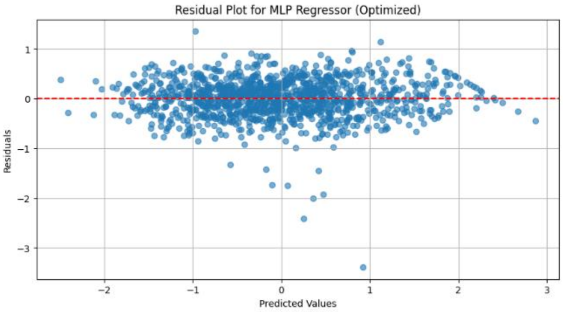

Figure 10. MLP Regressor Results with Residual Analysis

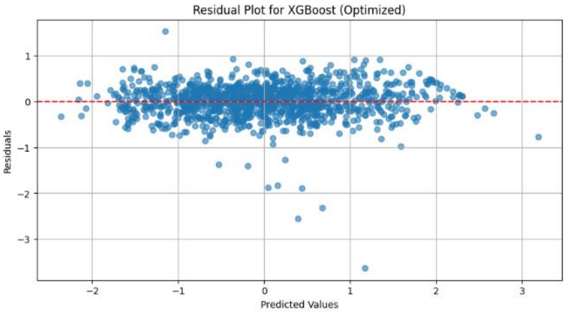

Figure 11. XGBoost Results with Residual Analysis

When compared with boosting and neural network models, as illustrated in Figures 9 to 11, the boosting models such as Gradient Boosting Regressor (GBR) and XGBoost exhibit more compact residual distributions. The prediction data that deviates significantly from the actual values is considerably reduced in these models compared to others. Notably, the density of points around the actual values is particularly pronounced in the XGBoost method, indicating its exceptional prediction performance. This visualization underscores the effectiveness of boosting techniques in minimizing prediction errors.

From the results obtained, XGBoost emerges as the model with the best performance, exhibiting the lowest MSE and the highest R^2 values during data testing. The GBR also serves as a strong alternative, demonstrating performance that is closely aligned with that of XGBoost. Additionally, the RFR model provides commendable results and is easier to interpret compared to GBR or XGBoost. In contrast, LR and DTR models are less optimal, as they exhibit lower accuracy and a tendency to overfit the data.

IV. Conclusion

This research evaluated three regression models: LR, DTR, and RFR. The evaluation results, obtained through K-Fold Cross Validation, demonstrated that RFR outperformed both LR and DTR, as indicated by the smallest MSE value and the largest R^2 value. However, it is important to note that RFR incurs significant time costs due to the necessity of repeatedly building multiple trees, which can impact its practicality in time-sensitive applications.

By comparing the three methods—LR, DTR, and RFR—with boosting methods such as GBR and XGBoost, as well as neural network methods like the MLP Regressor, it was found that the RFR method remains superior to the MLP Regressor. However, it is outperformed by both XGBoost and GBR, indicating that while RFR is effective, the boosting methods provide enhanced predictive performance. The GBR and RFR demonstrated strong performance across the evaluated datasets, while LR performed adequately for simpler datasets. The inclusion of the MLP Regressor underscored the potential of neural networks in regression tasks; however, it exhibited slightly lower accuracy compared to the ensemble methods. These findings suggest that advanced machine learning techniques, such as XGBoost, are particularly effective in handling complex datasets for accurate predictions, providing valuable insights.

V. Bibliography

[1] I. Sina *et al.*, “HAIR LOSS AND ALOPECIA,” vol. 20, no. 2, 2021.

[2] N. Sa, P. S. Biologi, and U. N. Padang, “Literature Review : Peran Hormon Testosteron terhadap Androgenetic Alopecia,” vol. 8, pp. 30472–30482, 2024.

[3] E. Fitri, “Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah,” *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 58–64, 2023, doi: 10.52158/jacost.v4i1.491.

[4] Farhanuddin, Sarah Ennola Karina Sihombing, and Yahfizham, “Komparasi Multiple Linear Regression dan Random Forest Regression Dalam Memprediksi Anggaran Biaya Manajemen Proyek Sistem Informasi,” *J. Comput. Digit. Bus.*, vol. 3, no. 2, pp. 86–97, 2024, doi: 10.56427/jcbd.v3i2.408.

[5] F. Hidayah, S. J. Angesti, and Y. P. Widyastuti, “Prediksi Harga Rumah di Boston Menggunakan Metode Linear Regression, SVR, Decision Tree dan Random Forest Regression,” pp. 1–9.

[6] M. A. Sembiring, F. W. Sembiring, and S. Informasi, “Analisa Kinerja Model Regresi Dalam Machine Learning,” vol. 8, no. 1, pp. 144–152, 2024.

[7] D. Eko Waluyo *et al.*, “Implementasi Algoritma Regresi pada Machine Learning untuk Prediksi Indeks Harga Saham Gabungan,” *Univ. Dian Nuswantoro, Semarang Jln. Imam Bonjol*, vol. 9, no. 1, pp. 12–17, 2024.

[8] N. S. Soraya and H. Hendry, “Komparasi linear regression, random forest regression, dan multilayer perceptron regression untuk prediksi tren musik TikTok,” *Aiti*, vol. 20, no. 2, pp. 191–205, 2023, doi: 10.24246/aiti.v20i2.191-205.

[9] D. Pramesti and Wiga Maulana Baihaqi, “Perbandingan Prediksi Jumlah Transaksi Ojek Online Menggunakan Regresi Linier Dan

Random Forest," *Gener. J.*, vol. 7, no. 3, pp. 21–30, 2023, doi: 10.29407/gj.v7i3.20676.

[10] I. H. Maula, "Kemungkinan Kebotakan | Kaggle." Accessed: Dec. 05, 2024. [Online]. Available: https://www.kaggle.com/datasets/itsnahm/baldness-probability

[11] I. Nurdin, Sugiman, and Sunarmi, "Penerapan Kombinasi Metode Ridge Regression (RR) dan Metode Generalized Least Square (GLS) untuk Mengatasi Masalah Multikolinearitas dan Autokorelasi," *J. Mipa*, vol. 41, no. 1, pp. 58–68, 2018.

[12] G. Chairunisa *et al.*, "Life Expectancy Prediction Using Decision Tree, Random Forest, Gradient Boosting, and XGBoost Regressions," *J. Sintak*, vol. 2, no. 2, pp. 71–82, 2024, doi: 10.62375/jsintak.v2i2.249.

[13] Diana Tri Susetianingtias, Eka Patriya, and Rodiah, "Model Random Forest Regression Untuk Peramalan Penyebaran Covid-19 Di Indonesia," *Decod. J. Pendidik. Teknol. Inf.*, vol. 2, no. 2, pp. 84–95, 2022, doi: 10.51454/decode.v2i2.48.

[14] R. Dahlia and C. I. Agustyaningrum, "Perbandingan Gradient Boosting dan Light Gradient Boosting Dalam Melakukan Klasifikasi Rumah Sewa," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 6, pp. 1016–1020, 2022, doi: 10.32672/jnkti.v5i6.5460.

[15] U. Riyanto, "Penerapan Algoritma Multilayer Perceptron (Mlp) Dalam Menentukan Kelayakan Kenaikan Jabatan: Studi Kasus Pt. Abc - Jakarta," *JIKA (Jurnal Inform.*, vol. 2, no. 1, pp. 58–65, 2018, [Online]. Available: http://jurnal.umt.ac.id/index.php/jika/article/view/5481

[16] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 21–26, 2022, doi: 10.31605/jomta.v4i1.1792.

[17] H. Nuha, "Mean Squared Error (MSE) dan Penggunaannya," *Papers.Ssrn.Com*, vol. 52, pp. 1–1, 2023, [Online]. Available: https://ssrn.com/abstract=4420880

[18] H. Hernandez, "Vol. 8, 2023-10," vol. 8, pp. 1–43, 2023, doi: 10.13140/RG.2.2.26570.13769.