Nusantara Hasana Journal
Volume 4 No. 8 (Januari 2025), Page: 32-39
E-ISSN : 2798-1428

COMPARISON OF THE PERFORMANCE OF REGRESSION-SPECIFIC AND MULTI-PURPOSE ALGORITHMS

Nasir Usman*[1], Darniati[2], Rosnani[3], Musdalifa Thamrin[4], Nurahmad[5], Nurdiansyah[6], Muhammad Faisal[7]

[1,2,3,4]STMIK Profesional Makassar, Indonesia
[5]Universitas Handayani Makassar, Indonesia
[6]Universitas Dipa Makassar, Indonesia
[7]Universitas Muhammadiyah Makassar, Indonesia

Email: nasirusman@stmikprofesional.ac.id

Abstract
Regression is a data science method for evaluating the relationship between independent and dependent variables. This study compares the performance of various regression algorithms using the Boston Housing Dataset, which consists of 506 samples divided into 80% for training and 20% for testing. Performance evaluation was conducted using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). All algorithms were implemented with default hyperparameter settings provided by the Scikit-learn library to ensure fair comparison. The results showed that versatile algorithms, particularly Gradient Boosting Machines (GBM) and Random Forest, achieved the best performance with R^2 values of 0.92 and 0.89, respectively, and lower errors. Conversely, regression-specific algorithms, such as Linear Regression and Ridge Regression, recorded R^2 values of approximately 0.67, while the k-Nearest Neighbors algorithm had the lowest performance with an R^2 of 0.65. Versatile algorithms proved to be more effective for datasets with complex non-linear patterns, while regression-specific algorithms were better suited for linear data patterns. These findings provide guidance for practitioners in selecting algorithms based on data characteristics and analysis objectives.
Keywords: *Regression-Specific, Multi-Purpose Algorithms, Comparison Technique, Boston Housing Dataset*

Abstrak
Regresi adalah metode analisis dalam data sains untuk mengevaluasi hubungan antara variabel independen dan dependen. Penelitian ini membandingkan kinerja berbagai algoritma regresi menggunakan Boston Housing Dataset, yang terdiri dari 506 sampel, dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian. Evaluasi kinerja dilakukan menggunakan metrik Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan Koefisien Determinasi (R^2). Semua algoritma diterapkan dengan pengaturan hiperparameter default yang tersedia pada library Scikit-learn untuk memastikan perbandingan yang adil. Hasil penelitian menunjukkan bahwa algoritma serbaguna, khususnya Gradient Boosting Machines (GBM) dan Random Forest, memiliki kinerja terbaik dengan nilai R^2 masing-masing 0,92 dan 0,89 serta error yang lebih rendah. Sebaliknya, algoritma regresi seperti Linear Regression dan Ridge Regression mencatat nilai R^2 sekitar 0,67. Algoritma k-Nearest Neighbors menunjukkan kinerja terburuk dengan nilai R^2 0,65. Algoritma serbaguna terbukti lebih efektif untuk dataset dengan pola non-linear yang kompleks, sedangkan algoritma regresi lebih cocok untuk data dengan pola linier. Temuan ini memberikan panduan bagi praktisi dalam memilih algoritma yang sesuai dengan karakteristik data dan tujuan analisis.

Nusantara Hasana Journal
Volume 4 No. 8 (Januari 2025), Page: 32-39
E-ISSN : 2798-1428

***Kata Kunci**: Regression-Specific, Multi-Purpose Algorithms, Comparison Technique, Boston Housing Dataset*

INTRODUCTION

In the era of big data and advancements in information technology, data analysis has become increasingly essential across various fields, including computer science, economics, and social sciences (Yang et al., 2023). One of the most commonly used analytical techniques is regression, which aims to understand the relationship between independent and dependent variables (Wahyuningsih et al., 2024). This study focuses on comparing the performance of specialized regression algorithms, including Linear Regression, Ridge Regression, Lasso Regression, and Support Vector Regression, with more general-purpose algorithms such as Decision Tree, Random Forest, Gradient Boosting Machines, and k-Nearest Neighbors in regression tasks. This research contributes to evaluating the effectiveness of each algorithm based on widely recognized evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2) (Botchkarev, 2019).

Regression is a statistical method used to predict the value of a dependent variable based on one or more independent variables. (Mądziel, 2024). In this context, specialized regression algorithms are designed to handle regression problems more efficiently and accurately compared to general-purpose algorithms. For example, Linear Regression is the simplest and most commonly used method but has limitations in handling non-linear data. (Senapati, 2023). On the other hand, algorithms such as Ridge and Lasso Regression offer solutions to issues like multicollinearity and feature selection, which are often challenges in regression analysis (Xin & Khalid, 2018). Meanwhile, Support Vector Regression (SVR) is capable of handling non-linear data by using appropriate kernels, providing greater flexibility in the model (Yan et al., 2020).

On the other hand, general-purpose algorithms like Decision Tree and Random Forest are known for their ability to handle large and complex datasets (Soekamto et al., 2023). Decision Tree splits data into subsets based on the values of specific features, while Random Forest combines multiple decision trees to improve accuracy and reduce overfitting (Elshazli et al., 2024). Gradient Boosting Machines (GBM) is also a popular algorithm that works by building models iteratively and optimizing the errors of previous models (Branco et al., 2017). k-Nearest Neighbors (k-NN) is a non-parametric algorithm that classifies data based on its proximity to other data points. (Yin et al., 2023) though it often requires longer computational time on large datasets. (Vieira et al., 2019).

The evaluation metrics used in this study, such as MAE, MSE, RMSE, and R^2, are crucial indicators for assessing the performance of regression models. MAE measures the average absolute error between predicted and actual values, providing a clear picture of how far the predictions deviate from reality (Botchkarev, 2019). MSE, on the other hand, imposes a greater penalty on larger errors, making it more sensitive to outliers (Zhang et al., 2019). RMSE is the square root of MSE and provides a measure of error in the same units as the dependent variable (Sfravara et al., 2024), Meanwhile, R^2 measures the proportion of variance in the dependent variable that can be explained by the independent variables (Njomaba et al., 2024).

Nusantara Hasana Journal
Volume 4 No. 8 (Januari 2025), Page: 32-39
E-ISSN : 2798-1428

By comparing specialized and general-purpose regression algorithms, this study contributes to providing deeper insights into the effectiveness of each approach in the context of regression tasks. The results of this research are expected to offer guidance for researchers and practitioners in selecting the most suitable algorithm for their specific applications, as well as contribute to the development of better data analysis methods in the future.

METHOD

This study adopts a quantitative approach with a comparative experimental design. The objective is to evaluate the performance of specialized regression algorithms and general-purpose algorithms in regression tasks. The performance of the algorithms is measured using evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2).

The dataset used is the Boston Housing Dataset, consisting of 506 housing data samples with 13 predictor features such as the number of rooms, crime rate, and distance to the city center (Botev et al., 2018). This dataset is a standard dataset frequently used in regression tasks. The data is divided into two groups: a) Training data (80%) to train the model; b) Testing data (20%) to evaluate the model's performance. The data splitting technique is performed randomly using a train-test split to ensure sample representativeness.

The research instruments include the implementation of the following algorithms: a) Specialized Regression Algorithms: Linear Regression, Ridge Regression, Lasso Regression, and Support Vector Regression (SVR); b) General-Purpose Algorithms: Decision Tree, Random Forest, Gradient Boosting Machines (GBM), and k-Nearest Neighbors (k-NN). The models are implemented using machine learning libraries such as Scikit-learn in Python, with all hyperparameters set to default values (no tuning) to ensure a fair comparison.

Research Procedure, namely: 1) Data Collection: Utilize the Boston Housing dataset available through the Scikit-learn library; 2) Data Preprocessing: Normalize the data, handle missing values (if any), and split the data into training and testing sets; 3) Model Training: Train each algorithm using the training data; 4) Model Evaluation: Use evaluation metrics to measure the performance of each algorithm on the testing data; 5) Result Analysis: Compare the performance outcomes between specialized regression algorithms and general-purpose algorithms.

The analysis is conducted by comparing the evaluation metric values. The best performance is determined based on the lowest MAE, MSE, and RMSE values, as well as the highest R^2 value. Statistical analysis is performed to evaluate the significance of performance differences between the algorithms.

The Boston Housing dataset was chosen because it is frequently used in regression studies as a benchmark, enabling direct performance comparisons with previous research. This dataset has the following characteristics: 1) Dataset Size: Large enough to support statistical analysis (506 data points); 2) Feature Diversity: 13 predictor features covering socio-economic characteristics and physical conditions; 3) Prediction Target: House prices (median value of owner-occupied homes), which is a continuous variable suitable for regression tasks. This dataset also facilitates the exploration of different machine learning algorithms without requiring complex preprocessing.

Nusantara Hasana Journal
Volume 4 No. 8 (Januari 2025), Page: 32-39
E-ISSN : 2798-1428

The algorithms used are divided into two main categories: 1) Specialized Regression Algorithms: Specifically designed for regression tasks, such as Linear Regression, Ridge Regression, Lasso Regression, and Support Vector Regression (SVR). These algorithms provide theoretical and efficient solutions for regression, leveraging assumptions of linearity and regularization (in Ridge and Lasso); 2) General-Purpose Algorithms: Capable of handling both regression and classification tasks, such as Decision Tree, Random Forest, Gradient Boosting Machines (GBM), and k-Nearest Neighbors (k-NN). These algorithms offer flexibility in managing non-linear relationships and complex datasets. The selection of algorithms aims to evaluate the effectiveness of specialized algorithms compared to general-purpose algorithms in regression tasks.

Evaluation metrics are chosen to ensure comprehensive performance analysis of algorithms: 1) MAE (Mean Absolute Error): Measures the average of absolute errors, directly indicating the degree of deviation between predictions and actual values; 2) MSE (Mean Squared Error): Gives higher penalties for large errors, suitable for identifying models that consistently make precise predictions; 3) RMSE (Root Mean Squared Error): The square root of MSE, providing error values on the original data scale; 4) R^2 (Coefficient of Determination): Assesses how well the variance in the data is explained by the model. These metrics are selected to provide a holistic perspective on the model's accuracy and fitness.

RESULTS AND DISCUSSION

The results section should provide details of all of the experiments that are required to support the conclusions of the paper. The section may be divided into subsections, each with a concise subheading.

Testing Results

Performance of Regression-Specific Algorithms

Four regression-specific algorithms were evaluated in this study: Linear Regression, Ridge Regression, Lasso Regression, and Support Vector Regression (SVR). The performance of each algorithm was measured using the selected evaluation metrics, and the results are as follows:

Algorithm	MAE	MSE	RMSE	R^2
Linear Regression	3.19	24.29	4.93	0.67
Ridge Regression	3.13	24.48	4.95	0.67
Lasso Regression	3.25	24.41	4.94	0.67
Support Vector Regression (SVR)	2.73	25.67	5.07	0.65

Performance of Versatile Algorithms

Four versatile algorithms were evaluated: Decision Tree, Random Forest, Gradient Boosting Machines (GBM), and k-Nearest Neighbors (k-NN). The evaluation results indicate the following insights:

Algorithm	MAE	MSE	RMSE	R^2
Decision Tree	2.39	10.42	3.23	0.86
Random Forest	2.04	7.90	2.81	0.89
Gradient Boosting Machines (GBM)	1.91	6.21	2.49	0.92
k-Nearest Neighbors (k-NN)	3.66	25.86	5.09	0.65

Performance Comparison

Gradient Boosting Machines (GBM) demonstrated the best performance

Nusantara Hasana Journal
Volume 4 No. 8 (Januari 2025), Page: 32-39
E-ISSN : 2798-1428

with the highest R^2 value (0.92) and the lowest errors across all metrics. Random Forest ranked second, achieving an R^2 of 0.89. In contrast, regression-specific algorithms like Linear Regression and Ridge Regression performed less effectively ($R^2 \approx 0.67$), while k-Nearest Neighbors (k-NN) exhibited the weakest performance, with an R^2 of 0.65.

Discussion

Advantages of Versatile Algorithms

Versatile algorithms such as Gradient Boosting Machines (GBM) and Random Forest excel in regression tasks, particularly with complex datasets. Their primary strength lies in their ensemble approach, which combines predictions from numerous base models to reduce variance and mitigate overfitting. For instance, GBM employs a boosting method that iteratively corrects errors from the previous model, leading to exceptional predictive performance.

Although Decision Tree models offer high interpretability, they are less competitive than Random Forest and GBM due to their susceptibility to overfitting when used as standalone models. Nonetheless, they remain valuable for applications requiring visual explanations.

Challenges in Regression-Specific Algorithms

Regression-specific algorithms such as Linear Regression, Ridge Regression, and Lasso Regression are limited by their inherent assumption of linearity. This reduces their effectiveness when dealing with datasets featuring nonlinear relationships or complex interactions among features. For example, while Ridge Regression addresses multicollinearity issues, its performance still falls short compared to versatile algorithms in this study.

Support Vector Regression (SVR) demonstrates advantages in handling high-dimensional datasets but requires careful tuning of parameters like the kernel and margin of error to optimize performance.

Poor Performance of k-NN

Despite its simplicity in implementation, k-Nearest Neighbors (k-NN) exhibited the poorest performance among the tested algorithms. The primary reason for this is its sensitivity to data dimensionality, which necessitates dimensionality reduction or normalization techniques to enhance accuracy. Without adequate data preprocessing, k-NN yields high error rates and low coefficients of determination.

Research Implications

The findings of this study offer valuable insights for machine learning practitioners in selecting regression algorithms. For tasks involving complex data with nonlinear relationships, versatile algorithms such as GBM and Random Forest are recommended. However, when interpretability is a priority, Linear Regression or Decision Tree models may be more suitable.

Research Limitations

This study was conducted using the default parameter settings provided by the Scikit-learn library for all algorithms, without any hyperparameter tuning. While this approach ensures consistency and fairness in comparing the algorithms, it may limit the potential performance of some models, especially those that are highly sensitive to parameter adjustments. Future research could explore hyperparameter optimization techniques, such as grid search or random search, to better capture the full capabilities of each algorithm.

Additionally, the evaluation was limited to a single dataset, the Boston

Nusantara Hasana Journal
Volume 4 No. 8 (Januari 2025), Page: 32-39
E-ISSN : 2798-1428

Housing Dataset. Although this dataset is widely used as a benchmark in regression studies, relying on a single dataset restricts the generalizability of the findings. Future studies should incorporate a variety of datasets with different characteristics, such as complex non-linear relationships, higher dimensionality, or imbalanced distributions, to provide a more comprehensive and robust evaluation of the algorithms.

The study's findings highlight that versatile algorithms, particularly Gradient Boosting Machines (GBM) and Random Forest, outperform regression-specific algorithms such as Linear Regression, Ridge Regression, and Lasso Regression in regression tasks. This is evidenced by the highest coefficients of determination (R^2) achieved by GBM (0.92) and Random Forest (0.89), compared to regression-specific algorithms with average R^2 values below 0.7.

This research confirms previous findings, such as those by (Natekin & Knoll, 2013), which state that ensemble models like GBM are more flexible in handling datasets with nonlinear and complex patterns than simple linear models. Additionally, the strong performance of Decision Trees (R^2: 0.86) aligns with literature emphasizing their model interpretability and adaptability to various data types.

However, versatile algorithms are not always the optimal choice in every scenario. For instance, Support Vector Regression (SVR), specifically designed for regression, demonstrates strengths in handling high-dimensional and nonlinear data. Despite its relatively low R^2 (0.65), this aligns with the literature that highlights SVR's potential in complex data analysis but notes the necessity of intensive parameter tuning to achieve optimal performance.

The study also observed that k-Nearest Neighbors (k-NN) had lower performance compared to GBM and Random Forest, with an R^2 of only 0.65. This may be attributed to k-NN's reliance on the number of neighbors and the chosen distance metric, as discussed in the k-NN tutorial by (Cunningham & Delany, 2022).

Comparisons of the results with the literature suggest that the superiority of GBM and Random Forest lies in their ensemble nature, which reduces overfitting and enhances generalization by aggregating outputs from multiple base models. Additionally, penalized regression methods like Ridge and Lasso Regression, despite their lower performance, remain relevant in contexts requiring multicollinearity reduction—a common challenge in regression datasets.

CONCLUSION

This study compared the performance of regression-specific algorithms and versatile algorithms in regression tasks using evaluation metrics such as MAE, MSE, RMSE, and R^2. The findings indicate that versatile algorithms, particularly Gradient Boosting Machines (GBM) and Random Forest, consistently outperformed regression-specific algorithms such as Linear Regression, Ridge Regression, Lasso Regression, and Support Vector Regression (SVR). Gradient Boosting Machines achieved the highest performance with an R^2 value of 0.92, followed by Random Forest with an R^2 value of 0.89. In contrast, regression-specific algorithms demonstrated lower performance, with the highest R^2 (0.67) recorded by Linear Regression. The superior performance of versatile algorithms can be attributed to their ensemble nature, which effectively handles complex data

patterns and improves model generalization. Additionally, the study highlights that algorithms like SVR offer specific advantages in handling high-dimensional and nonlinear data, even though their results were not as strong as those of ensemble algorithms. These findings underscore the importance of selecting algorithms that align with the dataset's characteristics and the analysis goals. In conclusion, this research reinforces the notion that algorithm selection should be tailored to the dataset's properties and the objectives of the analysis. GBM and Random Forest, with their consistently high performance, provide valuable insights for practitioners in choosing regression algorithms for complex and diverse applications.

REFERENCES

Botev, Z., Chen, Y.-L., LrEcuyer, P., MacNamara, S., & Kroese, D. P. (2018). Exact Posterior Simulation From The Linear Lasso Regression. *2018 Winter Simulation Conference (WSC)*, 1706–1717. https://doi.org/10.1109/WSC.2018.8632237

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). A Survey of Predictive Modeling on Imbalanced Domains. *ACM Computing Surveys*, *49*(2), 1–50. https://doi.org/10.1145/2907070

Cunningham, P., & Delany, S. J. (2022). k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys*, *54*(6), 1–25. https://doi.org/10.1145/3459665

Elshazli, M. T., Hussein, D., Bhat, G., Abdel-Rahim, A., & Ibrahim, A. (2024). Advancing infrastructure resilience: machine learning-based prediction of bridges' rating factors under autonomous truck platoons. *Journal of Infrastructure Preservation and Resilience*, *5*(1), 5. https://doi.org/10.1186/s43065-024-00096-x

Mądziel, M. (2024). *Energy Modeling for Electric Vehicles Based on Real Driving Cycles: An Artificial Intelligence Approach for Microscale Analysis*. https://doi.org/10.20944/preprints202402.0120.v1

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

Njomaba, E., Ofori, J. N., Guuroh, R. T., Aikins, B. E., Nagbija, R. K., & Surový, P. (2024). Assessing Forest Species Diversity in Ghana's Tropical Forest Using PlanetScope Data. *Remote Sensing*, *16*(3), 463. https://doi.org/10.3390/rs16030463

Senapati, A. (2023). *Correlation Coefficient-based Breakpoint detection @Piecewise Linear Regression*. https://doi.org/10.21203/rs.3.rs-2917422/v1

Sfravara, F., Barberi, E., Bongiovanni, G., Chillemi, M., & Brusca, S. (2024). Development of a Predictive Model for Evaluation of the Influence of Various Parameters on the Performance of an Oscillating Water Column Device. *Sensors*, *24*(11), 3582. https://doi.org/10.3390/s24113582

Soekamto, Y. S., Chandra, M., Wiradinata, T., Tanamal, R., & Saputri, T. R. D. (2023). *Property Category Prediction Model using Random Forest Classifier to Improve Property Industry in Surabaya* (pp. 256–265). https://doi.org/10.2991/978-94-6463-144-9_24

Vieira, J., Duarte, R. P., & Neto, H. C. (2019). kNN-STUFF: kNN STreaming Unit for Fpgas. *IEEE Access*, *7*, 170864–170877. https://doi.org/10.1109/ACCESS.2019.2955864

Wahyuningsih, T., Iriani, A., Dwi Purnomo, H., & Sembiring, I. (2024). Predicting students' success level in an examination using advanced linear regression and extreme gradient boosting. *Computer Science and Information Technologies*, *5*(1), 29–37. https://doi.org/10.11591/csit.v5i1.p29-37

Xin, S. J., & Khalid, K. (2018). Modelling House Price Using Ridge Regression and Lasso Regression. *International Journal of Engineering & Technology*, *7*(4.30), 498. https://doi.org/10.14419/ijet.v7i4.30.22378

Yan, L., Wu, C., & Liu, J. (2020). Visual Analysis of Odor Interaction Based on Support Vector Regression Method. *Sensors*, *20*(6), 1707. https://doi.org/10.3390/s20061707

Yang, Y., Gong, H., & Zang, J. (2023). The U.S. Opinion on China's Climate Issue During the Biden Administration from the Perspective of Big Data Software WordSmith 8.0. In *Proceedings of the 2022 3rd International Conference on Big Data and Informatization Education (ICBDIE 2022)* (pp. 23–30). Atlantis Press International BV. https://doi.org/10.2991/978-94-6463-034-3_4

Yin, Q., Ye, X., Huang, B., Qin, L., Ye, X., & Wang, J. (2023). Stroke Risk Prediction: Comparing Different Sampling Algorithms. *International Journal of Advanced Computer Science and Applications*, *14*(6). https://doi.org/10.14569/IJACSA.2023.01406115

Zhang, M., Hu, R., & Jiang, L. (2019). Three-dimensional sound reproduction in vehicle based on data mining technique. *Concurrency and Computation: Practice and Experience*, *31*(4). https://doi.org/10.1002/cpe.4936