PROCEEDING OF INTERNATIONAL SEMINAR
GLOBAL ISSUES IN SOCIAL, SCIENCE, LAW, HEALTH AND BUSINESS ENVIRONMENT

Analysis of the Impact of Data Scaling Techniques on Classification Algorithms: KNN, SVM, and Logistic Regression Using Synthetic Data

Aghasi Yanda Wafa Azizah[1], Muhammad Radifan Asyauri[2], Yuliani Puji Astuti, S.Si., M.Si.[3]
[1-3] Data Science Study Program, Faculty of Mathematics and Natural Science, Universitas Negeri Surabaya
Corresponding Author Email: aghasiyanda@gmail.com

Abstract

Data preprocessing plays a crucial role in determining the performance of machine learning classification algorithms, particularly in handling feature scale variations (Ahsan et al., 2021; Pinheiro et al., 2025). . This study analyzes the impact of data scaling techniques on the performance of classification algorithms, namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression, using a controlled synthetic dataset. All experiments were implemented using Python in a Google Colaboratory environment to ensure reproducibility and ease of implementation. Three data scaling scenarios were evaluated: No Scaling, Min-Max Scaling, and Standard Scaling. Model performance was assessed using multiple evaluation metrics, including Accuracy, Precision, Recall, F1-Score, and Execution Time. The experimental results indicate that data scaling has a significant influence on the performance of distance- and margin-based classification algorithms (Amorim, Cavalcanti and Cruz, 2022). In particular, SVM combined with Standard Scaling achieved the best overall performance, yielding an accuracy of 0.940 and an F1-score of 0.933. KNN also demonstrated sensitivity to feature scaling, although its performance remained competitive in the absence of scaling due to the controlled characteristics of the synthetic dataset. In contrast, Logistic Regression exhibited relatively stable performance across all scaling techniques, indicating lower sensitivity to feature scale variations. These findings emphasize the importance of selecting appropriate data scaling techniques according to the characteristics of the classification algorithm (Sujon et al., 2024).. This study provides empirical insights that can serve as practical guidelines for preprocessing strategies in machine learning classification tasks.

Keywords: *Data Scaling, Classification Algorithms, KNN, SVM, Logistic Regression, Synthetic Data*

Abstrak

Pra-pemrosesan data memainkan peran penting dalam menentukan kinerja algoritma klasifikasi pembelajaran mesin, terutama dalam menangani variasi skala fitur (Ahsan et al., 2021; Pinheiro et al., 2025). Penelitian ini menganalisis dampak teknik penskalaan data terhadap kinerja algoritma klasifikasi, yaitu K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Regresi Logistik, dengan menggunakan kumpulan data sintetis yang terkontrol. Semua eksperimen diimplementasikan menggunakan Python dalam lingkungan Google Colaboratory untuk memastikan reproduktibilitas dan kemudahan implementasi. Tiga skenario penskalaan data dievaluasi: Tanpa Penskalaan, Penskalaan Min-Max, dan

Penskalaan Standar. Kinerja model dinilai menggunakan beberapa metrik evaluasi, termasuk Akurasi, Presisi, Recall, F1-Score, dan Waktu Eksekusi. Hasil eksperimen menunjukkan bahwa penskalaan data memiliki pengaruh signifikan terhadap kinerja algoritma klasifikasi berbasis jarak dan margin (Amorim, Cavalcanti and Cruz, 2022). Secara khusus, SVM yang dikombinasikan dengan Penskalaan Standar mencapai kinerja keseluruhan terbaik, dengan akurasi sebesar 0,940 dan F1-score sebesar 0,933. KNN juga menunjukkan sensitivitas terhadap penskalaan fitur, meskipun kinerjanya tetap kompetitif tanpa penskalaan berkat karakteristik terkontrol dari dataset sintetis. Sebaliknya, Regresi Logistik menunjukkan kinerja yang relatif stabil di seluruh teknik penskalaan, yang mengindikasikan sensitivitas yang lebih rendah terhadap variasi skala fitur. Temuan ini menekankan pentingnya memilih teknik penskalaan data yang sesuai dengan karakteristik algoritma klasifikasi (Sujon et al., 2024). Studi ini memberikan wawasan empiris yang dapat berfungsi sebagai pedoman praktis untuk strategi prapemrosesan dalam tugas klasifikasi pembelajaran mesin.

Kata kunci: Penskalaan Data, Algoritma Klasifikasi, KNN, SVM, Regresi Logistik, Data Sintetis

Introduction

In the last decade, machine learning algorithms have become a fundamental component in intelligent decision-making systems across various industrial sectors. From healthcare and finance to manufacturing and smart agriculture, classification models are widely implemented to support predictive analytics and automated decision processes. One of the main tasks in machine learning is classification, in which a model is trained to predict category labels from new input data. Algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression are frequently selected due to their strong theoretical foundations, computational efficiency, and ability to model various data distributions.

However, optimal algorithmic performance does not depend solely on model selection, but is also strongly influenced by the quality and representation of the input data (Uddin and Lu, 2024). Proper feature representation plays a crucial role in both regression and classification tasks (Ahsan et al., 2021; Pinheiro et al., 2025). A common issue encountered in raw datasets is the variation in scale or numerical range among features. In many real-world datasets, attributes may have significantly different units or magnitudes, such as income measured in millions and age measured in small integers. Without proper adjustment, such disparities may unintentionally bias the learning process.

Scale disparity presents a particular challenge for distance-based and gradient-based learning algorithms (Singh and Singh, 2020). In the KNN algorithm, features with larger numerical ranges tend to dominate Euclidean distance calculations, potentially diminishing the contribution of other relevant attributes. Similarly, SVM, which seeks to maximize the separation margin between classes, and Logistic Regression, which relies on gradient descent optimization, are sensitive to inconsistent feature magnitudes. Unscaled data may distort the optimization landscape, slow convergence, and lead to suboptimal decision boundaries. To address these issues, data scaling techniques such as Min-Max

Normalization and Standardization (Z-score transformation) are commonly applied as standard preprocessing steps.

Recent large-scale empirical studies further emphasize that feature scaling should not be treated as a routine or optional procedure. Pinheiro et al. (2025) conducted a comprehensive benchmarking study involving multiple scaling techniques, machine learning algorithms, and datasets, revealing that scaling sensitivity varies significantly across models. Their findings indicate that distance-based and margin-based algorithms—including KNN and SVM—are particularly dependent on appropriate scaling, whereas tree-based ensemble models tend to be more robust to feature magnitude differences. This evidence highlights that scaling decisions must be aligned with the mathematical characteristics of the selected algorithm.

In addition, previous research underscores that preprocessing strategies directly affect model stability, generalization capability, and experimental reproducibility (Sujon et al., 2024). Improper handling of feature magnitudes may produce unstable classification boundaries and inconsistent performance across datasets. Therefore, systematic evaluation of scaling techniques is essential not only for improving predictive accuracy but also for ensuring reliability and robustness of classification systems.

Despite the growing body of research on feature scaling, most existing studies evaluate its impact using real-world datasets. Such datasets often contain confounding factors, including uncontrolled noise, extreme outliers, hidden correlations, or class imbalance, which may obscure the isolated effect of scaling techniques themselves (Dankar and Ibrahim, 2021; Rajotte et al., 2022). Consequently, it becomes difficult to determine whether performance differences arise from scaling methods or from inherent dataset characteristics.

To overcome this limitation, this study adopts a controlled experimental framework using synthetic data. Synthetic datasets provide full control over feature distribution, variance, redundancy, and class balance, allowing precise isolation of scaling effects. By minimizing external variability, this research aims to systematically evaluate and compare the performance of KNN, SVM, and Logistic Regression under three preprocessing scenarios: No Scaling, Min-Max Scaling, and Standard Scaling. The findings are expected to provide clearer empirical evidence regarding the significance of scaling techniques in improving classification accuracy, stability, and computational efficiency.

Research Method

Synthetic Data Generation

The dataset used in this study is synthetic data generated using the make_classification function from the Scikit-learn library. The use of synthetic data was chosen because it provides full flexibility in managing the characteristics of the dataset, such as the number of samples, the number of features, the level of redundancy, and the distribution of the class (Bae et al., 2025). With synthetic data, the influence of external variables that often appear in real-world data, such as uncontrolled noise or domain-specific biases, can be minimized (Dankar and Ibrahim, 2021). The generated dataset consists of 1,000 samples with 10 numerical features. Of the total features, five are redundant, and the rest are random (Uddin and Lu, 2024). The target variable consists of two classes, so the problem studied is included in binary

classification. This configuration was chosen to represent classification scenarios commonly used in basic machine learning experiments, while being complex enough to show performance differences between preprocessing techniques.

Data Splitting

After the Synthetic Data Generation process is completed, the dataset is divided into training data and test data using an 80:20 ratio. The data distribution is carried out randomly using certain random state values so that the results of the experiment are consistent and reproducible. The training data was used to train the classification model, while the test data was used to evaluate the model's generalization ability against data that had never been seen before (Rankin et al., 2020). The data splitting stage is important to ensure that the model's performance evaluation is not biased against the training data. With a clear separation between test data and training data, evaluation results can present more objective model performance.

Data Scaling Techniques

The data scaling stage is the core of this research. Three preprocessing scenarios were applied to compare their effect on the performance of the classification algorithm. The first scenario is *no scaling*, which is used as a baseline to determine the initial performance of the algorithm without preprocessing treatment. The second scenario uses Min-Max Scaling, which is a normalization technique that transforms the value of a feature into a range of [0, 1]. The third scenario uses Standard Scaling, which standardizes the data by subtracting the mean value and dividing it by the standard deviation.

Data scaling techniques are very important because differences in scale between features can lead to the dominance of certain features in the model learning process. Distance- and margin-based algorithms, such as KNN and SVM, are highly sensitive to differences in feature scale, so the selection of the right scaling technique can significantly improve model performance. In this study, the scaling process was carried out only on the training data. The scaling parameters obtained from the training data are then applied to the test data. This approach aims to prevent data leakage, which is a condition in which information from test data is accidentally used in the model training process, potentially causing the evaluation results to be invalid.

Classification Algorithm

This study uses three classic classification algorithms, namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. The selection of these algorithms is based on the difference in their working principles and the sensitivity of each algorithm to data scale. KNN is a distance-based algorithm that classifies data based on proximity to training data. Because it uses distance calculations, the performance of KNN is greatly influenced by the scale of the features. SVM is used as a representation of a margin-based algorithm that attempts to find the optimal hyperplane to separate classes. The distribution and scale of data greatly affect the formation of optimal margins. Logistic Regression is used as a linear model that is relatively more stable, but is still influenced by the scale of features in the gradient-based optimization process. All algorithms were run

using default parameters. This approach was chosen to keep the research focus strictly on the influence of data scaling techniques, rather than on the process of hyperparameter tuning.

Model Training and Evaluation

Each combination of data scaling techniques and classification algorithms is trained using the training data and tested using the test data. The training process is carried out systematically and repeatedly for each combination of methods to ensure consistency in the experimental results. The model's performance was evaluated using several evaluation metrics, namely Accuracy, Precision, Recall, and F1-score. The use of multiple metrics aims to provide a more comprehensive picture of performance, not only in terms of prediction accuracy, but also the balance between positive and negative class predictions (Ahsan et al., 2021). In addition to performance metrics, execution times are also recorded for each method combination. Execution time measurements are carried out to evaluate the computational efficiency of each approach. All experimental results were stored in a tabular format and analyzed comparatively to identify the pattern of influence that data scaling techniques have on model performance for each algorithm.

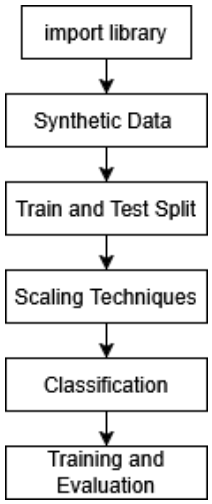

Figure 1. Workflow

Results and Discussion

Experimental Results

This study aims to analyze the effect of the application of data scaling techniques on the performance of classification algorithms, namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression, using synthetic datasets. The dataset is divided into training data and test data, then processed using three preprocessing scenarios, namely No Scaling, Min-Max Scaling, and Standard Scaling. Performance evaluation is carried out using five metrics, namely Accuracy, Precision, Recall, F1-Score, and Execution Time. A summary of the experimental results is presented in Table 1, which shows the

performance comparison of each combination of algorithms and scaling techniques. Based on Table 1, the combination of Standard Scaling and SVM produced the best overall performance, with an Accuracy value of 0.940 and an F1-Score of 0.933. These results show that the implementation of appropriate scaling is able to improve the model's ability to separate classes in high-dimensional feature spaces (Pinheiro et al., 2025; Amorim, Cavalcanti and Cruz, 2022).

Table 1. Measures of Each Variable

	Scaling Method	Model	Accuracy	Precision	Recall	F1-Score	Execution Time (s)
1	Standard Scaling	SVM	0.94	0.922	0.943	0.933	0.031
2	Min-Max Scaling	SVM	0.935	0.921	0.932	0.927	0.043
3	No Scaling	SVM	0.92	0.909	0.909	0.909	0.018
4	No Scaling	KNN	0.915	0.949	0.852	0.898	0.013
5	Min-Max Scaling	KNN	0.9	0.947	0.818	0.878	0.011
6	Standard Scaling	KNN	0.895	0.947	0.807	0.871	0.008
7	Min-Max Scaling	Logistic Regression	0.835	0.802	0.83	0.816	0.006
8	No Scaling	Logistic Regression	0.835	0.809	0.818	0.814	0.02
9	Standard Scaling	Logistic Regression	0.835	0.809	0.818	0.814	0.004

Analysis of the Influence of Data Scaling on Support Vector Machine (SVM)

The SVM algorithm shows a high sensitivity to changes in the scale of the data (Singh and Singh, 2020). In the no-scaling scenario, SVM obtained an F1-Score of 0.909. However, after Min-Max Scaling and Standard Scaling were applied, the F1-Score values increased to 0.927 and 0.933, respectively. This performance improvement occurs because SVM relies on hyperplane margin calculation in separating classes. Standard Scaling, which normalizes data based on mean and standard deviation, results in a more balanced distribution of features, helping SVMs determine optimal separation margins (Sujon et al., 2024). This indicates that the implementation of Standard Scaling is highly recommended when using SVM, especially on datasets with different feature scale variations.

Analysis of the Influence of Data Scaling on K-Nearest Neighbors (KNN)

In the KNN algorithm, data scaling also plays an important role because KNN relies on the calculation of distances between data points (Pagan, Zarlis and Candra, 2023). Without scaling, KNN produced an F1-Score of 0.898, which indicates a fairly good performance. However, after applying Min-Max Scaling and Standard Scaling, the F1-Score values decreased slightly to 0.878 and 0.871. This decrease shows that in the synthetic dataset used, the distribution of features is relatively balanced so that scaling does not provide significant gains. In other words, when the dataset is controlled and does not have extreme scale differences, KNN is still able to work well without additional preprocessing. These findings confirm that the effectiveness of data scaling is highly dependent on the characteristics of the data used (Ahsan et al., 2021; Uddin and Lu, 2024).

The Effect of Data Scaling on Logistic Regression

Logistic Regression shows the most stable performance compared to the other two algorithms. The Accuracy and F1-Score values were relatively consistent across all three preprocessing scenarios, with the F1-Score being in the range of 0.814–0.816. This stability is due to the nature of Logistic Regression, which is based on linear models and gradient optimization, so it is not very sensitive to minor differences in feature scale (Pinheiro et al., 2025). These results indicate that the application of data scaling to Logistic Regression plays more of a role in the aspect of computational efficiency than in the improvement of model accuracy.

Execution Time Analysis

In terms of execution time, all algorithms showed efficient performance because the experiments were run in a Google Colaboratory environment with a moderate dataset size. Logistic Regression has the fastest execution time, especially with Standard Scaling at 0.004 seconds, while SVM has a relatively higher execution time due to the complexity of the optimization process. This analysis of execution time shows that the selection of algorithms depends not only on accuracy but also on the need for computational efficiency, especially for implementation on systems with limited resources.

Conclusion and Recommendation

This study aims to analyze the impact of the application of data scaling techniques on the performance of classification algorithms, namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression, using a controlled generated synthetic dataset. The main focus of this study is to evaluate the extent to which the stages of data preprocessing, especially data scaling, affect the performance of the classification model in terms of accuracy, prediction balance, and computational efficiency (Pinheiro et al., 2025).

Based on the results of the experiments that have been conducted, it can be concluded that the data scaling technique has a significant influence on classification algorithms based on distance and margin, especially KNN and SVM. The SVM algorithm shows the most consistent performance improvement when combined with Standard Scaling, which results in the highest Accuracy and F1-Score values compared to other scenarios. This shows that normalization of data distribution based on averages and standard deviations is able to help SVM in forming a more optimal class separation margin (Sujon et al., 2024).

In the KNN algorithm, the influence of data scaling was also seen, although the increase was not always significant. The results of the experiment show that KNN is still able to achieve competitive performance in no-scaling scenarios, mainly due to the relatively balanced and controlled characteristics of the synthetic dataset. These findings indicate that the effectiveness of data scaling in KNN is highly dependent on the distribution and range of feature values in the dataset used. In contrast to KNN and SVM, Logistic Regression showed relatively stable performance in all preprocessing scenarios. Changes in data scaling techniques do not provide significant performance improvements against key evaluation metrics. This is due to the nature of Logistic Regression which is based on a linear model, so it is not too sensitive to scale differences between features. However, the application of data scaling to Logistic Regression still provides advantages in terms of optimization stability and computational time efficiency (Ahsan et al., 2021; Pinheiro et al., 2025).

Overall, this study confirms that the data preprocessing stage, especially data scaling, is an important component in building an optimal classification system. The selection of the right scaling technique must be adjusted to the characteristics of the algorithm and the data used (Amorim, Cavalcanti and Cruz, 2022). Standard Scaling is recommended as the most consistent approach, especially for SVM algorithms. As a direction of future work, this research can be developed by using real-world datasets, evaluating more preprocessing techniques, and comparing more complex classification algorithms. In addition, testing on larger data scales and different computing environments can provide more comprehensive insights into the performance and efficiency of classification algorithms.

References

Ahsan, M., Mahmud, M., Saha, P., Gupta, K. dan Siddique, Z. (2021) 'Effect of data scaling methods on machine learning algorithms and model performance', *Technologies*, 9(3).

Allorerung, P., Erna, A., Bagussahrir, M. dan Alam, S. (2024) 'Analisis performa normalisasi data untuk klasifikasi K-Nearest Neighbor pada dataset penyakit', *JISKA*.

Amorim, L., Cavalcanti, G. dan Cruz, R. (2022) 'The choice of scaling technique matters for classification performance', *Applied Soft Computing*, 133, p. 109924.

Bae, W., Alkobaisi, S., Horak, M., Bankar, S., Bhuvaji, S., Kim, S. dan Park, C. (2025) 'Synthetic data generation and evaluation techniques for classifiers in data starved medical applications', *IEEE Access*, 13.

Dankar, F. dan Ibrahim, M. (2021) 'Fake it till you make it: Guidelines for effective synthetic data generation', *Applied Sciences*, 11.

Klein, J. et al. (2024) 'Synthetic data at scale: A development model to efficiently leverage machine learning in agriculture', *Frontiers in Artificial Intelligence*. doi: 10.3389/fpls.2024.1360113.

Maheshwari, D., Sierra-Sosa, D. dan Garcia-Zapirain, B. (2022) 'Variational quantum classifier for binary classification: Real vs synthetic dataset', *IEEE Access*, 10, pp. 3705-3715.

Pagan, M., Zarlis, M. dan Candra, A. (2023) 'Investigating the impact of data scaling on the k-nearest neighbor algorithm', *Computer Science and Information Technologies*.

Pinheiro, J., De Oliveira, S., Silva, T., Saraiva, P., De Souza, E., Ambrosio, L. dan Becker, M. (2025) 'The impact of feature scaling in machine learning: Effects on regression and classification tasks', *IEEE Access*.

Rajotte, J., Bergen, R., Buckeridge, D., Emam, K., Ng, R. dan Strome, E. (2022) 'Synthetic data as an enabler for machine learning applications in medicine', *iScience*, 25.

Rankin, D., Black, M., Bond, R., Wallace, J., Mulvenna, M. dan Epelde, G. (2020) 'Reliability of supervised machine learning using synthetic data in health care', *JMIR Medical Informatics*, 8.

Singh, D. dan Singh, B. (2020) 'Investigating the impact of data normalization on classification performance', *Applied Soft Computing*, 97, p. 105524.

Sujon, K., Hassan, R., Towshi, Z., Othman, M., Samad, M. dan Choi, K. (2024) 'When to use standardization and normalization: Empirical evidence from machine learning models and XAI', *IEEE Access*, 12.

Uddin, S. dan Lu, H. (2024) 'Dataset meta-level and statistical features affect machine learning performance', *Scientific Reports*, 14.