JOURNAL OF DATA SCIENCE | Vol.2024:04
eISSN:2805-5160

A Comparative Study of Z-Score and Min-Max Normalization for Rainfall Classification in Pekanbaru

Rahmad Ramadhan Laska[1], Anne Mudya Yolanda[2]

[1,2] Statistics Study Program, Department of Mathematics, Riau University, Pekanbaru, Indonesia

Email: annemudyayolanda@lecturer.unri.ac.id[2], rahmad.ramadhan3648@student.unri.ac.id[1]

Abstract

Data preprocessing plays a crucial role in enhancing the performance of machine learning algorithms for classification tasks. Among the essential preprocessing stages is data normalization, which aims to standardize data into a comparable range of values. This study focuses on normalizing rainfall data in Pekanbaru from 2019 to 2023. The objective is to compare various data normalization techniques, including Min-Max Normalization and Z-Score Normalization. The comparison of these particular strategies is justified because they are widely applied and have different approaches. Min-max normalization is an easy-to-implement technique that makes the data sensitive to outliers by scaling it to a specific range, often from 0 to 1. However, Z-Score Normalization, sometimes referred to as Standardization, standardizes the data by dividing by the standard deviation and subtracting the mean, maintaining the shape of the distribution and making it resistant to outliers. The findings demonstrate that applying normalization techniques effectively enhances classification performance compared to using unnormalized data. Specifically, the optimal classification performance is achieved through Z-Score Normalization, yielding accuracy, sensitivity, and specificity rates of 74.59%, 82.48%, and 63.92%, respectively.

Keywords

Z-Score Normalization, Min-Max Normalization, Classification, Support Vector Machine, Rainfall

Introduction

Classification, a fundamental process in machine learning, is designed to categorize data into pertinent groups or categories. Beyond mere data grouping, the classification function serves to predict patterns and relationships inherent in the dataset. Among the widely adopted classification algorithms, Support Vector Machine (SVM) holds prominence. SVM, a supervised machine learning algorithm, aims to identify the optimal hyperplane within N-dimensional space to effectively segregate data points belonging to different classes, ensuring maximal margin between the closest points of each class (Aggarwal, 2015).

In the context of classification applications, data preprocessing stands as a critical stride toward attaining optimal classification performance before the application of machine learning algorithms. This pivotal step encompasses a spectrum of tasks, comprising data discretization,

Submission: 18 April 2024; **Acceptance:** 2 May 2024

outlier and noise removal, integration of data from diverse sources, handling of incomplete data, and transformation of data into a dynamic range amenable to comparison or normalization. (Dougherty, 2012). Data normalization entails the transformation of features into a standardized range of values, thereby mitigating the potential bias arising from larger numerical values that might overshadow smaller numerical features. This process aims to foster equitable contribution among features, particularly in delineating pattern classes, by averting undue influence from numerical magnitudes (García et al., 2015).

Normalization concentrates on rescaling the values of data features to ensure each feature carries a balanced influence during classification. This practice fosters model convergence and ensures optimal algorithm performance (Henderi, 2021). There are several commonly used normalization techniques, namely Min-Max Normalization and Z-score Normalization (Singh & Singh, 2020). Normalization methods are acknowledged to exert a substantial influence on classification accuracy within multivariate datasets. In instances where data stems from two distributions with notably distinct means and variances, normalization emerges as pivotal in ensuring that individual variables do not introduce bias into predictions (Asesh, 2022).

Min-Max Normalization is a process whereby unnormalized data is linearly adjusted to a predetermined range, specifically from the minimum value to the maximum value (Han et al., 2012). This technique stands out as one of the most effective methods for enhancing classifier performance (Shantal et al., 2023). Min-max normalization, commonly referred to as feature scaling, applies a linear transformation to the original dataset, ensuring that all scaled data falls within the range of (0, 1).

A z-score represents a standardized rendition of a raw score (x), providing insights into the relative position of that score within its distribution (Cote et al., 2021). Z-scores integrate information regarding the location of the distribution (the mean/center) and its dispersion (the standard deviation/spread) to interpret a raw score (x). Specifically, they denote the deviation of the score from the mean in standard deviation units and its direction. The efficacy of this technique purportedly lies in its capacity to enhance model accuracy. Specifically, it involves transforming the dataset, initially comprising features with varying ranges, into a standardized range (Anggoro & Supriyanti, 2021).

The SVM algorithm is utilized for classifying rainfall data in Pekanbaru from 2019 to 2023. The objective of this study is to evaluate various data normalization methodologies, including Min-Max Normalization and Z-Score Normalization. Subsequently, performance evaluation is conducted employing metrics such as the confusion matrix, accuracy, sensitivity, and specificity.

Methodology

The dataset used in this research comprises rainfall data for Pekanbaru from 2019 to 2023, sourced from the Meteorological, Climatological, and Geophysical Agency. It encompasses six variables, including five independent variables and one dependent variable denoted as Y, representing the rainfall category. The independent variables encompass average temperature, average humidity, duration of sunshine, wind direction at maximum speed, and average wind speed. Further details regarding the dataset can be found in Table 1.

Table 1. Rainfall Data of Pekanbaru in 2019-2023

Date	X_1	X_2	X_3	X_4	X_5	Rainfall
01-01-2019	28.00	77.00	4.60	360.00	4.00	0.00
02-01-2019	28.20	75.00	7.10	340.00	3.00	0.00
03-01-2019	28.10	74.00	8.80	320.00	3.00	0.00
04-01-2019	27.20	81.00	NA	320.00	1.00	8.70
05-01-2019	28.50	77.00	2.70	230.00	2.00	NA
06-01-2019	27.20	85.00	8.00	360.00	1.00	11.30
07-01-2019	27.50	87.00	2.10	60.00	2.00	29.60
08-01-2019	27.00	87.00	5.40	270.00	2.00	2.20
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31-12-2023	25.60	96.00	1.80	340.00	2.00	30.30

The analysis process conducted in this research encompasses several stages. Initially, data collection is performed, followed by handling missing data through linear interpolation. Linear interpolation is an approach that assumes the relationship between two data points can be approximated linearly or in a straight line between them (Huang, 2021). This method is employed to estimate values between two known data points.

In this study, categorical data is labeled and encoded, with the designation of 0 assigned to indicate the absence of rain (not rainy), and 1 allocated to signify the presence of rain (rainy). Subsequently, descriptive analysis is conducted. The data are then normalized, and subsequently partitioned into training and testing datasets.

In this analysis, SVM models are trained using both original and normalized data. SVM is a supervised learning system that employs linear functions in a high-dimensional feature space to classify data. Developed to enhance classification accuracy, SVM offers benefits such as explicit model dependence on a subset of data points and support vectors aiding in model interpretation (Ovirianti et al., 2022). Although SVM initially operates on a linear principle, it has evolved to handle non-linear problems through the introduction of the kernel concept. The core of the SVM method lies in finding the best hyperplane as a class separator, maximizing the margin between data classes (Quan & Pu, 2022).

Performance evaluation of the methods is carried out, followed by a comparison of results based on the evaluation metrics using a confusion matrix (Zeng, 2020). Confusion matrix is a table comparing model predictions to the actual values of the target attribute, used to obtain model performance metrics such as accuracy, sensitivity, and specificity (Larner, 2021). After acquiring these performance metrics, an analysis is conducted to compare the outcomes of both normalization methods. Finally, conclusions are drawn based on the findings.

Results and Discussion

Following data collection, it was identified that there were missing values, necessitating the implementation of linear interpolation to address this issue. Linear interpolation involves

estimating values between two known data points. Once the dataset has been effectively cleared of missing values, the analysis can proceed accordingly. The subsequent stage involves labeling and coding categorical data. This process is imperative as classification analysis necessitates labeled or categorized data. For the rainfall variable, adhering to the Meteorological, Climatological, and Geophysical Agency guidelines, the label 'not rainy' is assigned to values ranging from 0 to 0.5, whereas the label 'rainy' is designated to values exceeding 0.5. However, due to software limitations in directly processing textual labels, it has become essential to convert the labels into numerical representations. In this study, 'not rainy' is coded as 0, while 'rainy' is coded as 1.

Before proceeding to the analysis stage, it is crucial to conduct an overview of the data utilized. Descriptive analysis serves to furnish fundamental insights into the data and elucidate its distribution. The ensuing outcomes depict the findings of the descriptive analysis conducted on the dataset.

Table 2. Descriptive Analysis

	X_1	X_2	X_3	X_4	X_5
Number of data	1808.00	1808.00	1808.00	1808.00	1808.00
Average	27.12	82.46	4.53	194.60	1.54
Standard Deviation	1.09	6.05	2.57	103.50	0.68
Minimum Value	23.40	65.00	0.00	0.00	0.00
Quartile 1	26.40	78.00	2.60	140.00	1.00
Quartile 2	27.20	82.00	4.60	180.00	2.00
Quartile 3	27.90	87.00	6.60	300.00	2.00
Maximum Value	30.50	100.00	10.40	360.00	4.00

Table 2 presents the descriptive statistics for the variables in the dataset. The average temperature variable (X1) exhibits an average value of 27.12°C, accompanied by a standard deviation of 1.09°C. Likewise, the average humidity variable (X2) portrays an average of 82.46%, with a standard deviation of 6.05%. The sunshine duration variable (X3) demonstrates an average of 4.53 hours, with a standard deviation of 2.57 hours. Additionally, the wind direction at maximum speed variable (X4) showcases an average of 194.6°C, with a standard deviation of 103.5°C. Lastly, the average wind speed variable (X5) manifests an average of 1.54 m/s, accompanied by a standard deviation of 0.68 m/s. Further details regarding the distribution of the rainfall category variables are illustrated in Figure 2.

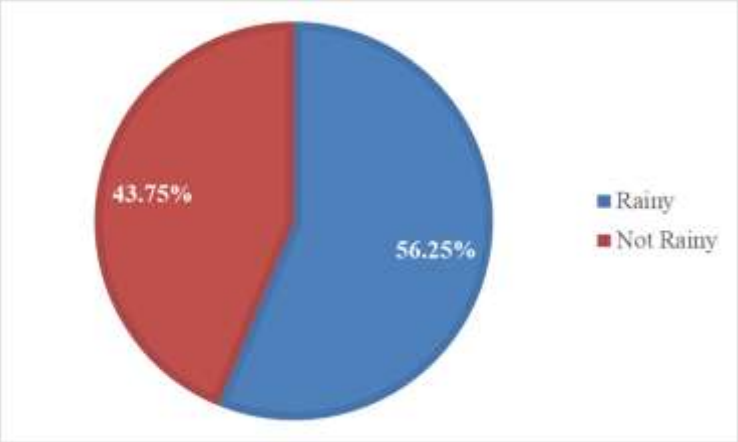

Figure 2. Percentage Distribution of Rainfall Categories in Pekanbaru (2019-2023)

In Figure 2, it can be seen that the rainy category surpasses the not rainy category. The rainy category comprises 1017 data points, accounting for 56.25% of the total dataset, whereas the not rainy category encompasses 791 data points, representing 43.75% of the total dataset.

The next step is to normalize the data using Min-Max Normalization and Z-Score Normalization. After normalizing the data, it is partitioned into two segments: training and testing datasets. This partitioning is performed randomly while preserving the proportion of classes present in the overall dataset, utilizing the stratify function. Additionally, the random_state function is employed to establish a consistent random order during the data partitioning process, ensuring unbiased and reproducible data division with consistent outcomes. The training dataset comprises 1446 instances, constituting 80% of the total data, while the testing dataset comprises 362 instances, representing 20% of the total data.

In this study, the SVM model was applied to three types of data: without normalization, with Min-Max Normalization, and with Z-Score Normalization. This comparison aimed to assess the classification performance between normalized and unnormalized data. Given that the data used are continuous and non-linear, the most appropriate SVM method is utilizing the Gaussian Kernel or Radial Basis Function (RBF). The SVM algorithm operates by identifying a hyperplane capable of effectively segregating data based on the rainy and not rainy classes.

The evaluation of method performance seeks to quantify the effectiveness of the classification model in predicting labels for new data. This evaluation utilizes testing data derived from various normalization techniques applied earlier in the classification analysis, encompassing both normalized and unnormalized data. The outcomes of the confusion matrix for data without normalization are presented in Table 3.

Table 3. Confusion Matrix for data without normalization

Actual	Prediction	
	Not Rainy	Rainy
Not Rainy	67	91
Rainy	32	172

Table 3 illustrates that in the data without normalization, the model accurately predicted the "not rainy" category 67 times, while incorrectly predicting "not rainy" when it was actually "rainy" 32 times. Additionally, the model accurately predicted the "rainy" category 172 times, but erroneously predicted "rainy" when it was actually "not rainy" 91 times. The results of the confusion matrix using Z-score normalization are presented in Table 4.

Table 4. Confusion Matrix for Z-Score Normalization data

Actual	Prediction	
	Not Rainy	Rainy
Not Rainy	101	57
Rainy	35	169

Table 4 shows that in the normalized data using z-score normalization, the model correctly predicted not rainy in 101 cases, and correctly predicted rainy in 169 cases. However, there were 35 cases where the model predicted not rainy, but it actually rained. In addition, there were 57 cases where the model predicted rainy, but there was actually not rainy. The confusion matrix results using min-max normalization are shown in Table 5.

Table 5. Confusion Matrix for Min-Max Normalization data

Actual	Prediction	
	Not Rainy	Rainy
Not Rainy	101	57
Rainy	37	167

According to Table 5, the confusion matrix results for Min-Max normalization reveal 101 correct predictions for the "not rainy" condition, alongside 37 incorrect predictions (predicting "not rainy" when it is actually "rainy"). Additionally, the model accurately predicted the "rainy" condition 167 times, but there were 57 instances where the model incorrectly predicted "rainy" when it was actually "not rainy". Table 6 presents the accuracy, specificity, and sensitivity metrics along with their corresponding results.

Table 6. SVM classification performance evaluation results

Data	Evaluation		
	Accuracy	Sensitivity	Specificity
Without Normalization	66.02%	84.31%	42.41%
Z-score Normalization	74.59%	82.84%	63.92%
Min-Max Normalization	74.03%	81.86%	63.92%

From Table 8, it is evident that the performance of the model utilizing all data normalization techniques surpasses that of using data without normalization. The accuracy of the model without normalization stands at 66.02%, with a sensitivity of 84.31% and specificity of 42.41%. Conversely, employing Z-score normalization yields an accuracy of 74.59%, sensitivity of 82.48%, and specificity of 63.92%. Similarly, Min-max normalization demonstrates favorable outcomes with an accuracy of 74.03%, sensitivity of 81.86%, and specificity of 63.92%. Based on

the analysis, it can be concluded that the z-score normalization technique yields the best performance.

Conclusion

Based on the conducted analysis, it can be concluded that the implementation of normalization in the classification of rainfall in Pekanbaru from 2019 to 2023 has effectively enhanced classification performance compared to the method without normalization. Among the normalization techniques utilized, z-score normalization exhibited the most favorable classification performance. The resulting model achieved accuracy, sensitivity, and specificity of 74.59%, 82.48%, and 63.92%, respectively. For future research about data normalization, it is advisable to select data with significantly different scales across variables, considering multi-class classification. Moreover, increasing the amount of utilized data, exploring other normalization techniques that may yield better performance, and considering the use of alternative classification algorithms sensitive to scale differences are recommended.

References

Aggarwal, C. C. (2015). Data classification: Algorithms and applications. In *Data Classification: Algorithms and Applications*. CRC Press. https://doi.org/10.1201/b17320

Anggoro, D. A., & Supriyanti, W. (2021). *Improving Accuracy by applying Z-Score Normalization in Linear Regression and Polynomial Regression Model for Real Estate Data of Trends Improving Accuracy by applying Z-Score Normalization in Linear Regression and*. *August*. https://doi.org/10.30534/ijeter/2019/247112019

Asesh, A. (2022). Normalization and Bias in Time Series Data. *Proceedings of MIDI'2021 – 9th Machine Intelligence and Digital Interaction Conference*, *440*, 90. https://link.springer.com/10.1007/978-3-031-11432-8

Cote, L. R., Gordon, R. G., Randell, C. E., Schmitt, J., & Marvin, H. (2021). *Introduction to Statistics in the Psychological Sciences* (O. E. R. Collection. (ed.)). University of Missouri - St. Louis. https://irl.umsl.edu/oer/25

Dougherty, G. (2012). Pattern recognition and classification: An introduction. In *Encyclopedia of Earth Sciences Series*. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-36699-9_69

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. In *Intelligent Systems Reference Library* (Vol. 72). Springer Cham.

Han, J., Micheline Kamber, & Pie, J. (2012). *Data mining: Concepts and techniques (3rd ed).*

Henderi, H. (2021). Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *IJIIS: International Journal of Informatics and Information Systems*, *4*(1), 13–20. https://doi.org/10.47738/ijiis.v4i1.73

Huang, G. (2021). Missing data filling method based on linear interpolation and lightgbm. *Journal of Physics: Conference Series*, *1754*(1). https://doi.org/10.1088/1742-6596/1754/1/012187

Larner, A. J. (2021). The 2x2 matrix: Contingency, confusion, and the metrics of binary classification. In *Neuropsychological Neurology*. Springer. https://doi.org/10.1017/cbo9780511545009.001

Ovirianti, N. H., Zarlis, M., & Mawengkang, H. (2022). Support Vector Machine Using A Classification Algorithm. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *7*(3), 2103–

2107. https://doi.org/10.33395/sinkron.v7i3.11597

Quan, Z., & Pu, L. (2022). An improved accurate classification method for online education resources based on support vector machine (SVM): Algorithm and experiment. *Education and Information Technologies*, *0123456789*. https://doi.org/10.1007/s10639-022-11514-6

Shantal, M., Othman, Z., & Bakar, A. A. (2023). A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization. *Symmetry*, *15*(12). https://doi.org/10.3390/sym15122185

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics - Theory and Methods*, *49*(9), 2080–2093. https://doi.org/10.1080/03610926.2019.1568485