International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Enhance Data Scaling Through Standardization of Selected Variables in Data Science

[1]Mr. Mohammed Abdul Habeeb, [2]Ms.Shrooq Hamed Sulaiyem Alrubaii, [3] Dr. Shaik Abdul Azeem, [4]Mr. Syed Muhammed Shavalliuddin

[1]*College of Computing and Information Sciences, Dept of Information Technology, University of Technology and Applied Sciences* Al Musanna, Oman, Email: Mohammed.habeeb@utas.edu.om, habeeb4sa@gmail.com
[2]*College of Computing and Information Sciences., Dept of Information Technology, University of Technology and Applied Sciences* Al Musanna, Oman, Email: shurooq.alrubaii.act@utas.edu.om
[3] *College of Computing and Information Sciences. Dept of Information Technology, University of Technology and Applied Sciences* Al Musanna, Oman Email : azeem.shaikh@utas.edu.om
[4] *College of Computing and Information Sciences. Dept of Information Technology, University of Technology and Applied Sciences* Al Musanna, Oman, syed.shavalliuddin@utas.edu.om, syedmdshaval@gmail.com

ABSTRACT—

One of the most crucial preprocessing steps in data science is standardization. Standardization will bring data variables on a common scale which is necessary for most of the machine learning algorithms. Our research paper explains the importance of standardization and its effect on machine learning model performance and different techniques to achieve standardization [1] [5]. By studying various research articles and empirical analysis we focus on advantages and disadvantages of standardization in data science. Our research paper emphasizes the need of standardization in attaining the accuracy in projected data driven results. Theoretical concepts of standardization or z-score normalization [11]. By conducting standardization experiments on various datasets, we demonstrate how standardization solves the problems of scaling and enhances the machine learning models [3] [6]. This research paper directs data scientists with profound understanding and relevance of standardization with practical execution.

Keywords—standardization, datasets, variables, machine learning models, scaling

Introduction

Data processing is one of the fundamental steps in data science, which shows a significant impact on analyzing data and machine learning models [9]. Standardization is a preprocessing technique in which the data is transformed to maintain mean equal to zero and standard deviation equal to one. This research paper probes the concept of standardization its needs and its impact on improving various algorithms through standardization, variables are brought on common scale to block any impulses that may occur due to the difference in units and dimensions. Datasets containing variables with different units and scales are particularly dealt using standardization. Algorithms that depend on distant measures such as KNN-K Nearest Neighbor [2] and SVM-Support Vector Machine[1] [10] may give rise to one sided results if data sets are not standardized, our research paper targets to provide extensive understanding of standardization which includes theory and practical applications by canvasing both limits and boons of standardization, our approach is to offer a logical stand point on role of standardization in data preprocessing.

Problem Statement

Datasets in storage repositories mostly contain features or variables that are measured on different scales, which create difficult challenges during analysis of data which is to be used for various machine learning models [8]. Taking an example of data set which contain variables with distinctive scale like age calculated in years and profit calculated in rupees can generate hostile results from machine learning models. If we do not standardize variables machine learning models may tend towards the large scale variables, giving unnecessary importance to large scale variables which ultimately decrease the accuracy of our predictions. The principal problem our research paper tries to find and discuss is standardization of dataset variables and its effects[7]. Through this research paper we try to concentrate on how unstandardized data generate misleading outcomes, bring down the machine learning model validity. By focusing on those issues, we work towards underlying the significance of standardization in data processing channels. Our research paper also converse about the disadvantages of standardization, including lack of interpretation in some circumstances and recommends the contexts in which standardization of datasets need to be done effectively[6].

Literature Review

To examine the effect of standardization, we make use of combined theoretical study and empirical analysis, we initiate by reexamining the basics of standardization such as z-score normalization, Standardization of data is one of the most important preprocessing step in machine learning, standardization scales features to a uniform range which leads to refine performance in machine learning model. [1] showed that using the correct standardization method increases SVM accuracy and decreases errors while the wrong choices gives poor results. [4] provided the standardized variable distance algorithm, with more than 90% accuracy achievement rate on datasets which is better than traditional algorithms. [2] proposed the standardized KNN algorithm for multimode fault detection, improving accuracy through scale information. [7] established standardization boosts effort estimation in software projects through random forest algorithm. [5] mapped ProteinNet, by standardizing protein data for biometrics. [6] Shows standardized imaging process in neuro-oncology for reconstruction. [3] Shown standardization is perfect for non –Gaussian data, improving model illustration. [11] equated normalization techniques, highlighting specific effects of the algorithm. [8] Practiced Hidden Markov Model to handle standardization backing data warehousing. [13] Used Machine Learning in Standardization of API data, improving security. [12] Boosted significance of standardization for data integration. Broadly, these studies reinforce the necessity of standardization to stabilize generalization and accuracy.

Methods

Equations

One way to standardize input column x_i through the formula of standardization is we create a new column x_i’ from xi, you will transform each value of column xi with the formula. [3] [10]

$x_i' = (x_i - \mu)/\sigma$ (1)

x_i’ is standardized column

x_i is the input column.

µ is mean of all data values in a column.

σ is standard deviation.

x_i’ is the new column after standardization.

$\mu = \sum_{i=1}^{n} x_i / n$ (2)

x_i is the input column.

µ is the mean.

n is number of values in the column in a dataset

To calculate the mean µ, we have to calculate the sum of all the values in selected dataset, count the number of values in the dataset then divide the sum of the values by number of values.

$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$ (3)

σ is standard deviation

x_i is the input column

µ is the mean of all the data values in a column

n is number of values in the column in a dataset

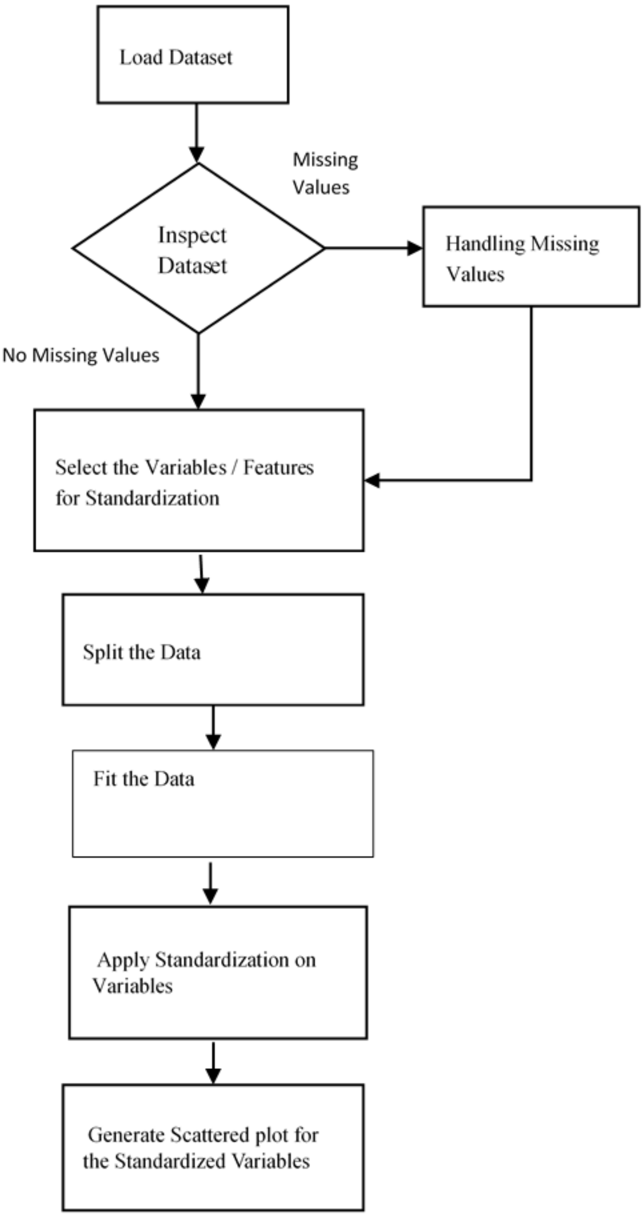
Now the good part we need to understand here is when you standardize with this formula (1), you will generate the new numerical values for each item of the column x_i, which will be x_i’.

Now if we calculate the mean (µ) of these new column x_i’ transformed values, your calculated result for the mean value µ will be equal to zero(0) ie µ=0 and the calculated standard deviation of new column xi’ will be one ie σ =1. This is what happens in standardization, here you will scale or standardize any value by substituting it in the given formula (1), and after scaling the values we get new series of scaled values who definitely acquire the following characteristics ie mean(µ)=0 and standard deviation σ =1. are particularly going to demonstrate this with programming code.

Methodology

To examine the effect of standardization, we make use of combined theoretical study and empirical analysis, we initiate by reexamining the basics of standardization such as z-score normalization, standardization will be clarified in detail, including its benefits and drawbacks [12], our research paper will provide in and out understanding of standardization, which helps the data scientists to gain the knowledge required to choose correct standardization method for the particular case study. We are going to carry a set of practically demonstrating experiments using the datasets of real-world repositories such as Kaggle, to know how standardization works on datasets, we will generate scattered plots depicting before and after scaling of datasets, we plot the PDF-probability function for before and after scaling of datasets. If the data scientists work on algorithms such as linear regression, logistic regression, K-Means Clustering, K-Nearest Neighbor and Support Vector Machine, in these algorithms you calculate distance [1] [2]. The results generated through standardization will help data scientists to analyze the improvement in machine learning models.

Flow Chart


Load Dataset
Missing Values
Inspect Dataset
Handling Missing Values
No Missing Values
Select the Variables / Features for Standardization
Split the Data
Fit the Data
Apply Standardization on Variables
Generate Scattered plot for the Standardized Variables

3.1 Methodology

Result and Analysis

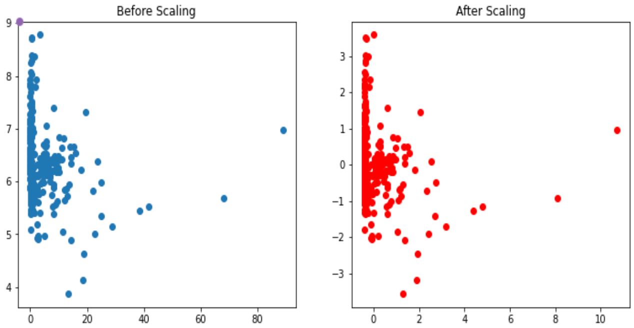
Real Estate and Urban Economics

We have selected a dataset named Boston Housing from real estate and urban economics domain whose source is Kaggle dataset repository. Dataset contains the following features or columns whose abbreviations are crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, b, lstat, medv. We selected three columns crim ie per capita crime rate by town, zn proportion of residential land zoned for lots for over 25000 square feet. and rm ie average number of rooms per dwelling.

We performed the following activities first we converted the crim and rm columns to numeric in order to handle the nonnumeric values, second we performed train test split on crim and rm columns dropping the column zn from the initially selected columns, third we fit and transformed the training set and only transformed the testing set with the help of standard scalar method, fourth we generated the scattered plot for trained dataset before and after scaling. [3] [10]

Standardization of Dataset variables

(Real Estate and Urban Economics)



The above first figure shows the scattered plot created for the original trained columns crim and rm of dataset before scaling and second figure shows the scattered plot created for the scaled trained columns crim and rm of dataset after scaling. We observe that the value of the mean (2) will be zero and value of the standard deviation (3) will be 1

Table of trained and scaled columns

	crim	*rm*
Count	353.0	349.0
Mean	0.0	0.0
std	1.0	1.0
min	-0.4	-3.6
25%	-0.4	-0.6
50%	-0.4	-0.1
75%	0.0	0.5
max	10.7	3.6

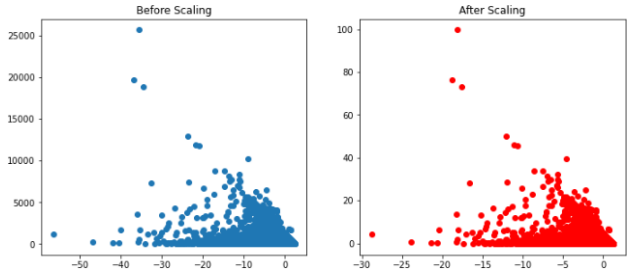
Financial and Cybersecurity

We have selected a dataset named credit card fraud detection from financial and cybersecurity domain whose source is Kaggle dataset repository. Dataset contains the following features or columns time, v1 to v28, class and amount, v1 to v28 are the features resulting from principal component analysis (PCA), class with column value 0 and 1, 0 indicates genuine transaction and 1 indicates fraudulent transaction and Amount is the transaction amount.

We performed the following activities first we converted the v1 and amount columns to numeric in order to handle the nonnumeric values, second we perform the train test split on v1 and amount columns dropping the time column from the initially selected columns, third we fit and transformed the training set and only transformed the testing set with the help of standard scalar method, fourth we generated the scattered plot for trained dataset before and after scaling. [3] [10]

Standardization of Dataset variables

(Financial and Cybersecurity)



The above first figure shows the scattered plot created for the original trained columns v1 and amount of dataset before scaling and second figure shows the scattered plot created for the scaled trained columns v1 and amount of dataset after scaling. We observe that the value of the mean (2) will be zero and value of the standard deviation (3) will be 1

Table of trained and scaled columns		
	V1	**Amount**
Count	199365.0	199365.0
Mean	0.0	0.0
std	1.0	1.0
min	-28.8	-0.3
25%	-0.5	-0.3
50%	0.0	-0.3
75%	0.7	-0.0
max	1.3	99.7

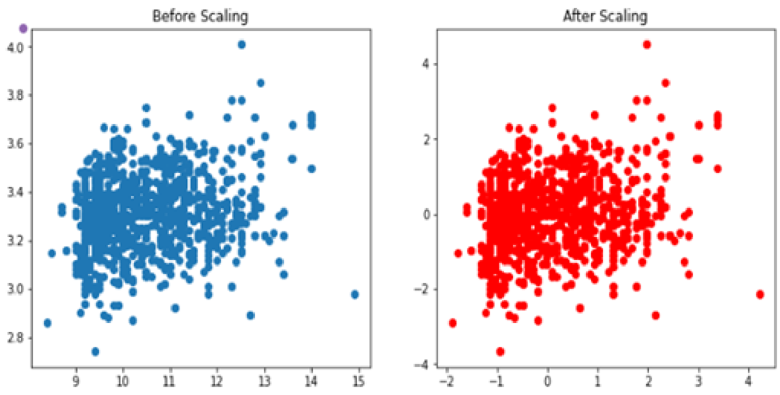
Food and Beverage Business.

We have selected a dataset named wine quality red from food and beverage business domain whose source is UCI university of califonia, irvin machine learning dataset repository. Dataset contains the following features or columns fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.

We performed the following activities first we selected the columns alcohol and pH, second we performed the train test split on alcohol and pH columns, third we fit and transformed the training set and only transformed the testing set with the help of standard scalar method, fourth we generated the scattered plot for trained dataset before and after scaling. [3] [10]

Standardization of Dataset variables

(Food and Beverage Business)



The above figure1 shows the scattered plot created for the original trained columns alcohol and pH of the dataset wine quality red before scaling and second figure shows the scattered plot created for the scaled trained columns alcohol and pH after scaling. We observe that the value of the mean (2) will be zero and the value of the standard deviation (3) will be 1.

	Alcohol	**pH**
Count	1119.0	1119.0
Mean	-0.0	0.0
Std	1.0	1.0
Min	-1.9	-3.7
25%	-0.9	-0.6
50%	-0.3	0.0
75%	0.6	0.6
Max	4.2	4.5

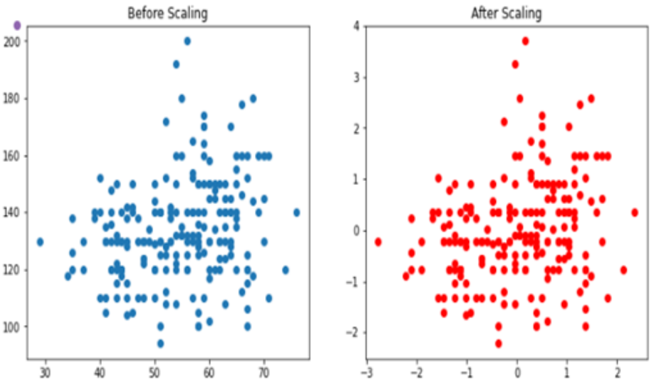
Health and Medicine.

We have selected a dataset named heart disease from health and medicine domain whose source is UCI university of California, Irvin machine learning dataset repository. Heart disease dataset contains the following features or columns age, sex, cp chest pain, trestbps resting blood pressure, chol serum cholesterol, FBS fasting blood sugar, restecg resting electrocardiographic results, thalach maximum heart rate achieved, exang exercise induced angina, oldpeak depression induced by exercise relative to rest, slope slope of the peak exercise, ca number of major vesssels, thal thalassemia.

We performed the following activities first we dropped the missing values from all the columns, second we selected the two appropriate features or columns ie age and trestbps for standardization, third we performed the train test split action on the age and trestbps columns, fourth we fit and transformed the training set and only transformed the testing set with the help of standard scalar method, fifth we generated the scattered plot for trained dataset before and after scaling. [3] [10]

Standardization of Dataset variables

(Health and Medicine) Heart Disease



The above figure shows the scattered plot created for the original trained columns age and trestbps of the dataset heart disease before scaling and the second figure shows the scattered plot created for the scaled trained columns age and trestbps after scaling. We observed that the value of the mean (2) will be zero and the value of the standard deviation (3) will be 1.

	Age	**Trestbps**
Count	207.0	207.0
Mean	-0.0	-0.0
Std	1.0	1.0
Min	-2.8	-2.2
25%	-0.8	-0.8
50%	0.1	-0.2
75%	0.8	0.5
Max	2.3	3.7

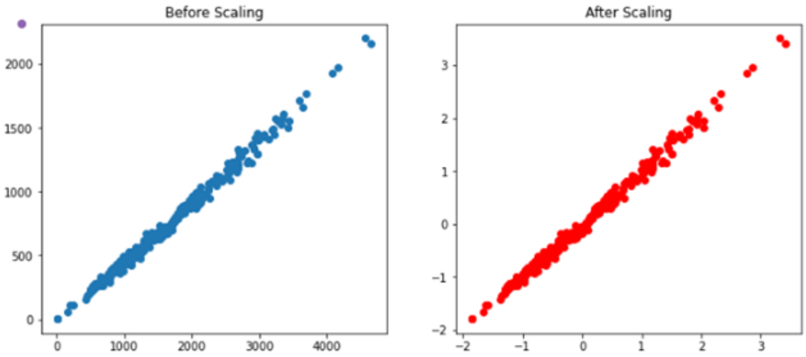
Health and Medicine

We have selected a dataset named diabetes_all_2016 from health and medicine domain whose source is Data.gov repository which is united states government's open data site. Diabetes_all_2016 dataset contains the following features or columns ct is a unique identifier, bpad blood pressure average diastolic, bpan blood pressure average normal, bpan2 blood pressure average normal second measurement, bwad body weight average diastolic, bwan body weight average normal, bwan2 body weight average normal second measurement, bmad body mass average diastolic, bman body mass average normal, bman2 body mass average normal second measurement.

We performed the following activities, first we selected all the relevant features or columns for standardization, second we split the data into training and testing datasets, third we fit and transformed the training dataset and only transformed the testing dataset with the help of standard scalar method, fourth we generated the scattered plot for trained dataset before and after scaling. [3] [10]

Standardization of Dataset variables

(Health and Medicine) Diabetes_all_2016



The above figure shows the scattered plot created for all the original trained columns of the dataset diabetes_all_2016 before scaling and after scaling for all the scaled trained columns. We observed that the value of mean (2) will be zero and the value of the standard deviation (3) will be 1.

	CT	BPAD	BPAN	BPAN2	BWAD	BWAN	BWAN2	BMAD	BMAN	BMAN2
Count	273.0	273.0	273.0	273.0	273.0	273.0	273.0	273.0	273.0	273.0
Mean	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
Std	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Min	-2.1	-1.9	-1.8	-1.9	-1.9	-1.8	-1.9	-1.8	-1.7	-1.7
25%	-0.7	-0.8	-0.8	-0.7	-0.8	-0.7	-0.7	-0.8	-0.8	-0.7
50%	0.4	-0.2	-0.1	-0.1	-0.2	-0.2	-0.1	-0.2	-0.1	-0.2
75%	0.6	0.6	0.6	0.7	0.7	0.5	0.6	0.6	0.6	0.6
Max	3.5	3.4	3.8	3.7	3.4	4.1	4.0	3.5	3.9	3.9

Conclusion

Our Research paper explores the standardization of five different datasets boston housing and credit card detection datasets from Kaggle data set repository, wine quality and heart disease datasets from UCI dataset repository, diabetes_all_2016 dataset from the data.gov dataset repository. Datasets used for standardization are related to different domains such as business, health, medicine and finance, these datasets provide deep insight into their respective fields.

We standardized datasets in order to have consistency in scale among various features or columns to facilitate accuracy and efficiency in data analysis. In boston housing dataset through standardization we can improve the performance of regression models which are used in predicting house prices, by standardizing credit card detection dataset we can improve the process of detecting the fraudulent transactions, in wine quality datasets standardization helped in improving the comparison of physiochemical properties to anticipate wine quality, heart disease dataset after standardization improves the performance of classification algorithms which will aid in exact diagnosis of heart disease, in diabetes_all_2016 dataset through standardization we can generate uniform analysis on diabetes related measures which are most important for diabetes health research. The results generated after standardizing various datasets show that standardization is a crucial preprocessing step which redefines implementation and performance of machine learning models.

References

Luor and Dah-Chin, “A comparative assessment of data standardization on support vector machine for classification problems,” Department of Applied Mathematics, I-Shou University, Taiwan., Intelligent Data Analysis, vol. 19, no. 3, pp. 529-546, 2015, https://www.researchgate.net/publication/281424643_A_comparative_assessment_of_data_standardization_on_support_vector_machine_for_classification_problems, DOI: 10.3233/IDA-150730.

Song Bing, Tan Shuai, Shi Hongbo and Zhao Bo, “Fault detection and diagnosis via standardized k nearest neighbor for multimode process,”, University of Science and Technology, Shanghai 200237, China, Version of Record 20 January 2020, https://www.sciencedirect.com/science/article/abs/pii/S1876107019303736, https://doi.org/10.1016/j.jtice.2019.09.017.

Khaled Mahmud Sujon, Rohayanti Binti Hassan, Zeba Tusnia Towshi, Manal A. Othman, MD Abdus Samad and Kwonhue Choi, “When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI,” IEEE Access (Volume: 12), pp. 135300 - 135314, 17 September 2024, 1963, https://ieeexplore.ieee.org/abstract/document/10681438, ISSN: 2169-3536.

Abdullah Elen and Emre Avuclu, “Standardized Variable Distances: A distance-based machine learning method,” https://www.sciencedirect.com/science/article/abs/pii/S1568494620307936, https://doi.org/10.1016/j.asoc.2020.106855.

Mohammed AlQuraishi, “ProteinNet: a standardized data set for machine learning of protein structure,” 11 June 2019, Volume 20, article number 311, (2019), https://link.springer.com/article/10.1186/s12859-019-2932-0, https://doi.org/10.1186/s12859-019-2932-0.

Xiao Tian Li and Raymond Y Huang, “Standardization of imaging methods for machine learning in neuro-oncology,” Neuro-Oncology Advances, Volume 2, Issue Supplement_4, December 2020, Pages iv49–iv55, 23 January 2021, https://academic.oup.com/noa/article/2/Supplement_4/iv49/6117779, https://doi.org/10.1093/noajnl/vdaa054.

Pinkashia Sharma and Jaiteg Singh, “Machine Learning Based Effort Estimation Using Standardization,” 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 28-29 September 2018, IEEE Xplore: 28 March 2019, https://ieeexplore.ieee.org/abstract/document/8674908?casa_token=ZADwqakn3BMAAAAA:GR-61rwi6h-LPSBrYXsPnw68PUiJMnlbSk4uraUots28c2t6c-25MhTm55wFuPwIeciHh0KgxbLt, DOI: 10.1109/GUCON.2018.8674908.

Abdul Kaleem , Khawaja Moyeezullah Ghori, Zahra Khanzada and M. Noman Malik, “Address Standardization using Supervised Machine Learning, ” 2011 International Conference on Computer Communication and Management Proc .of CSIT vol.5 (2011) © (2011) IACSIT Press, Singapore, https://www.researchgate.net/profile/Muhammad-Malik-17/publication/283706723.

Irina Kalinina, Aleksandr Gozhyj, Peter Bidyuk, Victor Gozhyi, Maksym Korobchynskyi and Vasiliy Nadraga, “A Systematic Approach to Data Normilization and Standardization in Machine Learning Problems,” Springer Nature Switzerland AG 2025 S. Babichev and V. Lytvynenko, ISDMCI 2024, LNDECT 244, pp. 206-219, 2025, https://doi.org/10.1007/978-3-031-88483-2_11.

Peshawa Jamal Muhammad Ali and Rezhna Hassan Faraj, “Data Normalization and Standardization: A Technical,” The Machine Learning Lab. at Koya University, Erbil, Iraq. Machine Learning Technical Reports, 2014, 1(1), pp 1-6, https://docs.google.com/document/d/1x0A1nUz1WWtMCZb5oVzF0SVMY7a_58KQulqQVT8LaVA/edit#.

Kelsy Cabello-Solorzano, Isabela Ortigosa de Araujo, Marco Pena, Luis Correia and Antonio J. Tallon-Ballesteros, “The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis.” Online: 31 August 2023, pp 344–353, : Lecture Notes in Networks and Systems LNNS,volume 750, https://link.springer.com/chapter/10.1007/978-3-031-42536-3_33.

T. Aditya Sai Srinivas, Y. Sravanthi, Y. Vinod Kumar and I.V. Dwaraka Srihith, “Data Standardization: Key to Effective Data Integration,” https://www.researchgate.net/publication/375120903_Data_Standardization_Key_to_Effective_Data_Integration, November 2023, Volume 6 Issue 1, DOI: https://doi.org/10.5281/zenodo.10060920.

Bharath Bhushan Sreeravindra and Anoop Gupta, “Machine Learning Driven API Data Standardization,” International Journal of Global Innovations and Solutions (IJGIS), 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5026776, DOI:10.21428/e90189c8.4d24b759.