Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements

Peshawa J. Muhammad Ali

Department of Software Engineering, Faculty of Engineering, Koya University, Koya, Kurdistan Region – F.R. Iraq

***Abstract*—K-nearest neighbor (KNN) is a lazy supervised learning algorithm, which depends on computing the similarity between the target and the closest neighbor(s). On the other hand, min-max normalization has been reported as a useful method for eliminating the impact of inconsistent ranges among attributes on the efficiency of some machine learning models. The impact of min-max normalization on the performance of KNN models is still not clear, and it needs more investigation. Therefore, this research examines the impacts of the min-max normalization method on the regression performance of KNN models utilizing eight different similarity measures, which are City block, Euclidean, Chebychev, Cosine, Correlation, Hamming, Jaccard, and Mahalanobis. Five benchmark datasets have been used to test the accuracy of the KNN models with the original dataset and the normalized dataset. Mean squared error (MSE) has been utilized as a performance indicator to compare the results. It's been concluded that the impact of min-max normalization on the KNN models utilizing City block, Euclidean, Chebychev, Cosine, and Correlation depends on the nature of the dataset itself, therefore, testing models on both original and normalized datasets are recommended. The performance of KNN models utilizing Hamming, Jaccard, and Mahalanobis makes no difference by adopting min-max normalization because of their ratio nature, and dataset covariance involvement in the similarity calculations. Results showed that Mahalanobis outperformed the other seven similarity measures. This research is better than its peers in terms of reliability, and quality because it depended on testing different datasets from different application fields.**

***Index Terms*—K-nearest neighbor, Min-max, Normalization, Similarity, Mahalanobis.**

I. Introduction

The K-nearest neighbor (KNN) has been introduced as supervised learning for the First time by Fix and Hodges in 1951 (Fix and Hodges, 1951). Then, it has been developed by Thomas Cover in 1967 (Cover and Hart, 1976). The algorithm is considered one of the oldest machine learning (ML) algorithms used for classification and regression. The algorithm depends on the similarity or the distance measures between the unknown samples and the closest items in the training set. In regression, the output of the KNN is a value that came from a previously observed output of the closest neighbor called target or from averaging the value of a group of neighbors' target values.

The number of neighbors that may contribute to determining an accurate result for unknown result samples depends on the nature and the statistical properties of the dataset. There is no specific optimize the number of the neighbor that must be considered in the process of determining the results of KNN algorithm. Therefore, examining multiple tests with different numbers of neighbors are the only right process for setting this number. Another factor that also has an impact on the efficiency of the KNN is the type of the KNN itself. KNN comes in two types, equal weight KNN, and weighted KNN. With equal weight KNN, all participating neighbors will evenly contribute to computing the result on an equal base, whereas the contribution of the participated neighbors in the weighted based KNN is changing according to the weights assigned for each neighbor, and the value of each weight could be determined based on the neighbor's distance from the target.

As mentioned before, the core of the KNN's functionality depends on the similarity measurements. There are multiple similarity measuring methods (distance metrics) that used by KNN, such as City block, Euclidean, Chebychev, Cosine, Correlation, Hamming, Jaccard, and Mahalanobis (Cha, 2007). The accuracy of some methods is very sensitive to any change in the distance scale, such as City block, Euclidean, and Chebychev, whereas other methods, such as Mahalanobis and Jaccard, do not depend on the scale of the input features, and the accuracy of the results does not change with the change of the scale. Therefore, examining the impact of the data scale on the KNN performance is another objective of this work.

Data scaling is one of the preprocessing steps that come before the training phase of any ML algorithm. The aim of the scaling step is to accelerate the training phase and to improve the efficiency of any proposed model. The scaling preprocess

ARO-The Scientific Journal of Koya University
Vol. X, No.1 (2022), Article ID: ARO.10955, 7 pages
DOI: 10.14500/aro.10955
Received: 30 March 2022; Accepted: 07 June 2022
Regular research paper: Published: 25 June 2022
Corresponding author's e-mail: peshawa.jammal@koyauniversity.org

http://dx.doi.org/10.14500/aro.10955

is only working with numerical datasets. Although there are different data scaling (data normalization) techniques such as min-max normalization, z-score, soft-max, decimal scaling, max-abs scalar, and quantile transformer, this work focuses on the min-max normalization only.

Min-max data normalization method scales the data to specific ranges such as [0,1] or [−1,1] to eliminate the domination of some of the features over others in the ML techniques using similarity measurements like KNN. Assuming that features in a dataset may come with different ranges, the similarity measures assign more weight to features with larger ranges than those with small ranges. Therefore, min-max data normalization is used to equalize the weight of these features and make them have the same effect on the decision-making process.

There is a disagreement exists in the literature on the impact of min-max data normalization on the regression performance of KNN models with different similarity measures. There is also an ambiguity about how different types of KNN with different similarity measurements may respond to the min-max data normalization. The aim of this research is to study the combined effect of min-max data normalization with these eight similarity measurements on the regression performance of KNN.

The rest of this article is structured as follows: Section 2 explains the related works to this study. Section 3 is the methodology of this research work consisting of three stages: Adopting suitable datasets from the University of California Irvine (UCI) website, implementing min-max feature normalization, range [0,1], and applying and validating KNN on both the original and the normalized datasets. Section 4 summarizes all the observed results, and Section 5 discusses the observed results. Finally, Section 6 concludes this research work.

II. Related Works

Research works reported different results about the real impact of min-max data normalization on the performance of the KNN models. This disparity in the results is clearly seen in many articles and publications. Some studies recorded very little impact (Ambarwari, Adrian and Herdiyeni, 2020; Dadzie and Kwakye, 2021), whereas others showed a significant increase in the accuracy of the models (Ahsan, et al., 2021; Rajeswari and Thangavel, 2020). Although all mentioned articles utilized the Euclidean KNN, none of them justified the reasons behind this disproportion of the results.

In general, most of the research works that investigated the impact of data normalization were used benchmark datasets, such as the datasets of the repository dataset of the UCI (Ahsan, et al., 2021; Pires, et al., 2020; Bhardwaj, Mishra and Desikan, 2018; Dadzie and Kwakye, 2021; Jayalakshmi and Santhakumaran, 2011; Shorman, et al., 2018). This is because the research goal in those works was to determine the effects of the min-max scaling without considering the nature of the dataset application. In this work, five benchmark datasets from the UCI repository have been adopted.

The common comparison performance measure for regression models is the mean squared error (MSE) (Rajeswari and Thangavel, 2020; Singh, Verma and Thoke, 2015; Jayalakshmi and Santhakumaran, 2011; Shorman, et al., 2018; Bhardwaj, Mishra and Desikan, 2018). In addition to that, methods such as root mean square error (RMSE) (Prasetyo, et al., 2020) and the coefficient of determination (R2) (Aksu, Güzeller and Eser, 2019) can be used as well. However, this work utilizes the MSE method as a performance indicator. The mathematical expressions of the eight different similarity measurements used in this research are shown in Table I.

From the above literature review, it becomes clear that there is no strong vision available on the effects of min-max normalization on the performance of the different types of KNN algorithm. In other words, there is no clear answer to this question "What is/are the condition(s) that makes the performance of the KNN responds positively or negatively to a data scaling method (min-max normalized)?" To the best of our knowledge, there is no comprehensive study that can answer this question clearly. Therefore, the aim of this work is to answer the mentioned question clearly and precisely.

Table I
Similarity Measurements Used with KNN Model

Similarity Metrics	Mathematical Expression	Notes
Cityblock	$sim(x,y)=\sum_{i=1}^{n}\lvert x_i-y_i\rvert$	
Euclidean	$sim(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$	
Chebychev	$sim\ (x,\ y)=max_i\ \{\lvert x_i-y_i\rvert\}$	
Cosine	$sim(x,y)=1-\frac{x.y^t}{\sqrt{(x.x^t).(y.y^t)}}$	x^t and y^t are transpose vectors of the vectors x and y. The dot (.) represents dot product.
Correlation	$sim(x,y)=1-\frac{(x-\bar{x}).(y-\bar{x})^t}{\sqrt{(x-\bar{x}).(x-\bar{x})^t}\sqrt{(y-\bar{y}).(y-\bar{y})^t}}$	$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$
Hamming	$sim(x,y)=\frac{\#(x_i\neq y_i)}{n}$	# is the counting number
Jaccard	$sim(x,y)=\frac{\#[(x_i\neq y_i)\cap((x_i\neq 0)\cup(y_i\neq 0))]}{\#[(x_i\neq 0)\cup(y_i\neq 0)]}$	# is the counting number
Mahalanobis	$sim(x,y)=\sqrt{(x-y)C^{-1}(x-y)^t}$	C^{-1} is the inverse covariance matrix

x and y are two different records (vectors) that have the same number of attributes n, and $sim\ (x,\ y)$ is the similarity measure between them. x_i and y_i are feature values belonging to the record x and y

III. Methodology

The main aim of this work is to test the efficiency of the KNN algorithm against the min-max data normalization method. To achieve that, KNN has been operated using eight different similarity measurement methods (City block, Euclidean, Chebychev, Cosine, Correlation, Hamming, Jaccard, and Mahalanobis). Through each operation, five benchmark datasets have been fed separately to a regression-based KNN algorithm. For every benchmarked dataset, the accuracy of the KNN will be examined against the scaled and non-scaled datasets. The methodology can be summarized as follows:

- Step 1: Adopting suitable datasets from the UCI website,
- Step 2: Implementing min-max feature normalization, range [0,1],
- Step 3: Applying and validating KNN on both the original and the normalized datasets.
- Step 4: Comparing results and making conclusions.

Fig. 1 explains the methodology of the research work.

A. Step 1: The UCI Datasets

Five different benchmark datasets were downloaded from the ML repository website of the UCI (Dua and Graff, 2019). The reason behind selecting these datasets is the existing variation in the ranges of records among all attributes. Some of the datasets have big differences in their ranges like in the airfoil self-noise dataset (Table II) or very similar ranges like in power plant dataset (Table III). Such a variation is expected to have a role in explaining the impact of min-max normalization on the regression performance of the KNN algorithm with different similarity measurements. The datasets belong to the real applications of physics, life sciences, engineering, and business (Table IV). The dataset's statistical properties are shown in Tables II, III, V-VII.

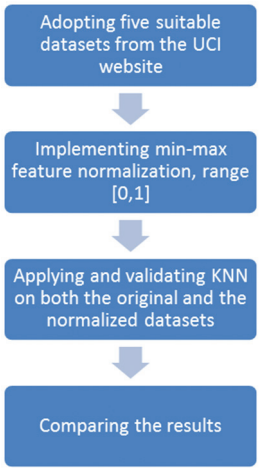

Fig. 1. The methodology of the research work.

Table II
Statistical Properties of the Airfoil Self-noise Dataset (Brooks, Pope and Marcolini, 1989)

Feature	Max. value	Min. value	Range	Mean	Standard deviation
F1	20,000	200	19,800	2886.380572	3152.573137
F2	22.2	0	22.2	6.782302063	5.918128125
F3	0.3048	0.0254	0.2794	0.136548237	0.093540728
F4	71.3	31.7	39.6	50.86074518	15.5727844
F5	0.0584113	0.000400682	0.058010618	0.01113988	0.013150234
Target	140.987	103.38	37.607	124.8359428	6.898656622

B. Step 2: Implementing Min-max Feature Normalization, Range [0,1]

As shown in Equation (1), a normalized data sample *x'* could be obtained from the original data sample *x*. For an attribute, it is mostly dependent on instances with the maximum and minimum values in the same attribute. In this normalization method, the original data sample component values will be transformed to [0,1] range.

$$x' = [(\frac{x - oldMin}{oldMax - oldMin}) * (newMax - newMin)] + newMin \quad (1)$$

Where *x'* is the normalized data sample, *x* is the original data sample, *oldMin* is the minimum data among any attribute of the original dataset, *oldMax* is the maximum data among any attribute of the original dataset, *newMin* is the minimum of the normalized dataset, and *newMax* is the maximum of the normalized dataset.

C. Step 3: Applying and Validating KNN on the Datasets

In this work, eight similarity measurements (City block, Euclidean, Chebychev, Cosine, Correlation, Hamming, Jaccard, and Mahalanobis) were adopted with the KNN algorithm as a regression method. For each similarity measure, 100 tests were conducted on the both of the original and the normalized datasets, with a total number of 1600 tests. The target of applying all tests is to determine the impact of min-max data normalization on the regression performance of KNN by comparing the MSE results observed from the original dataset and the normalized versions. The reason for adopting the eight different similarity measures of KNN is their diversity in dealing with the datasets and their different methodology in determining the nearest neighbors. In all the tests, MSE is computed as a performance measure and 10-fold as a validation technique.

Two types of KNN models were adopted, equal weight KNN and weighted KNN. For each of the type, until 10 neighbors were considered. The tests were adopted on the five datasets for both original and the normalized data. The total number of tests was 1600 tests, as follows: Eight similarity measures, two types of models (weighted and equal weight), until 10 neighbors, two cases raw data and normalized data, for each of the five datasets. The total is (8 × 2 × 10 × 2 × 5 = 1600 tests).

Table III
Statistical Properties of the Combined Cycle Power Plant Dataset (Tüfekci, 2014)

Feature	Max. value	Min. value	Range	Mean	Standard deviation
F1	37.11	1.81	35.3	19.65123119	7.45247323
F2	81.56	25.36	56.2	54.30580372	12.707893
F3	1033.3	992.89	40.41	1013.259078	5.938783706
F4	100.16	25.56	74.6	73.30897784	14.60026876
Target	495.76	420.26	75.5	454.3650094	17.066995

Table IV
Datasets Properties

Dataset	Number of features	Type of features	Number of instances	Supervised learning	Application
Airfoil self-noise dataset (Brooks, Pope and Marcolini, 1989)	5	Real	1503	Regression	Physics
Physicochemical properties of protein tertiary structure dataset (Rana, 2013)	9	Real	9146	Regression	Life sciences
Combined cycle power plant dataset (Tüfekci, 2014)	4	Real	9568	Regression	Energy
Concrete compressive strength dataset (Yeh, 1998)	8	Real	1030	Regression	Civil engineering
Real estate valuation dataset (Yeh and Hsu 2018)	6	Real	414	Regression	Business

Table V
Statistical Properties of the Physicochemical Properties of Protein Tertiary Structure Dataset (Rana, 2013)

Feature	Max. value	Min. value	Range	Mean	Standard deviation
F1	32,240.2	2783.15	29,457.05	9873.68162	4011.808135
F2	11,787.1	403.5	11,383.6	3016.435929	1450.041879
F3	0.56848	0.09362	0.47486	0.302155567	0.062784658
F4	343.239	10.6891	332.5499	103.4039974	54.9395949
F5	4,467,324.7	374,315.5155	4,093,009.223	1,369,092.965	558385.2823
F6	470.897	33.6462	437.2508	145.5447009	69.30473494
F7	83,153.57	1108.9	82,044.67	3987.14593	1880.513854
F8	337	0	337	70.04286027	56.50548747
F9	47.4559	15.5049	31.951	34.48790348	5.930509868
Target	20.981	0	20.981	7.833154384	6.120956974

Table VI
Statistical Properties of the Concrete Compressive Strength Dataset (Yeh, 1998)

Feature	Max. value	Min. value	Range	Mean	Standard deviation
F1	540	102	438	281.1656311	104.5071416
F2	359.4	0	359.4	73.89548544	86.27910364
F3	200.1	0	200.1	54.18713592	63.99646938
F4	247	121.75	125.25	181.5663592	21.35556707
F5	32.2	0	32.2	6.20311165	5.973491651
F6	1145	801	344	972.9185922	77.75381809
F7	992.6	594	398.6	773.5788835	80.1754274
F8	365	1	364	45.66213592	63.16991158
Target	82.599225	2.331807832	80.26741697	35.81783583	16.70567917

Table VII
Statistical Properties of the Real Estate Vluation Dataset (Yeh and Hsu, 2018)

Feature	Max. value	Min. value	Range	Mean	Standard deviation
F1	2013.5833	2012.666667	0.9166666	2013.148953	0.281995327
F2	43.8	0	43.8	17.71256039	11.39248453
F3	6488.021	23.38284	6464.63816	1083.885689	1262.109595
F4	10	0	10	4.094202899	2.945561806
F5	25.01459	24.93207	0.08252	24.96903007	0.012410197
F6	121.56627	121.47353	0.09274	121.5333611	0.015347183
Target	117.5	7.6	109.9	37.98019324	13.6064877

Because our focus in this research work is on variance in the results happening by adopting different similarity measures, therefore, the effect of number of the neighbors (k) did not get too much attention. Instead, the minimum MSE among all of the adopted experiments in each of the similarity measures is observed to be used for the comparison purposes. Determining the effect of number of neighbors (k) on the performance of the KNN is not in the scope of this research. It is noticed that each one of the datasets is responded differently to the increase in the number of neighbors until 10 neighbors, this is because the variance in the nature of the datasets and their statistical properties.

IV. Results

As mentioned previously, the performance indicator that utilized by this work for checking the efficiency of the proposed KNN is MSE. Throughout the experimental tests, eight models of KNN have been tested each model uses a specific type of similarity measurements. Each model of KNN has been trained and tested with five benchmarked datasets. For each dataset, the efficiency of the proposed KNN was computed in two situations; when the KNN is trained with the original dataset, and second, when the KNN is trained with normalized dataset. All testes were passed through MSE checking. As a result, the overall tests that have been conducted by this work are 1600 tests.

The minimum MSE results of KNN models using different similarity measurements on the five datasets are shown in Table VIII, Figs. 2-6. Each number of the MSE results shown in the Table VIII is the minimum of 20 tests, in other words, 10 tests including until 10 neighbors, and this done for the two cases weighted and equal weight ($10 \times 2 = 20$). Minimum MSE means the best result among all the 20 tests.

V. Discussion

The MSE results that obtained from five types of the KNN (City block, Euclidean, Chebychev, Cosine, and Correlation) that trained and tested with two datasets (airfoil self-noise and physicochemical) showed a significant improvement in the KNN's efficiency, look at Figs. 2 and 3. However, no significant improvement obtained when the same type of KNN is trained with power plant dataset, Fig. 4. On the other side, the efficiency of the mentioned type of the KNN has been degraded significantly when the KNN trained and tested with concrete strength and house valuation datasets, Figs. 5 and 6. It is clearly observed that performances of three types of the KNN (Hamming, Jaccard, and Mahalanobis) have not been changed. The disparity behavior of the KNN types against scaling and normalizing the training dataset is going back to the mathematical process or concept that each similarity measurement method has following whereas they do data processing. The five similarity measures (City block, Euclidean, Chebychev, Cosine, and Correlation) are sensitive

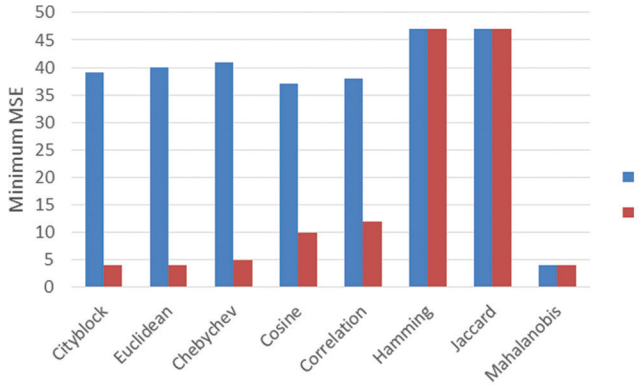

Fig. 2. Minimum MSE values of the KNN models with airfoil self-noise dataset using different similarity measurements (original vs. normalized).

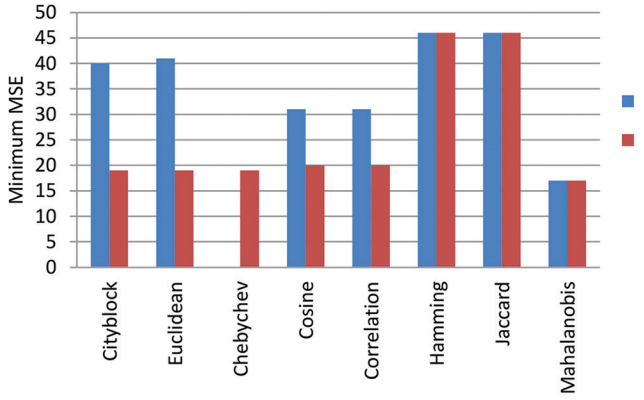

Fig. 3. Minimum MSE values of the KNN models with physicochemical dataset using different similarity measurements (original vs. normalized).

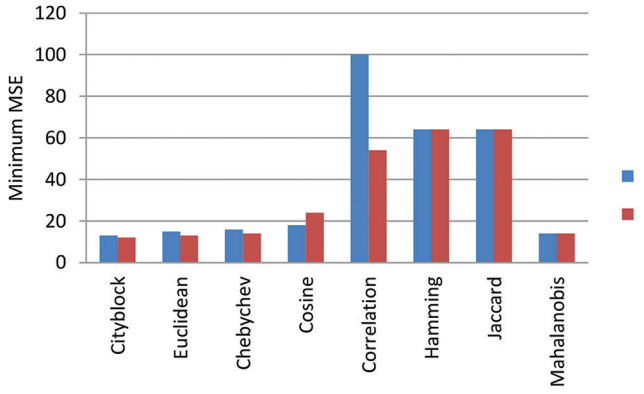

Fig. 4. Minimum MSE values of the KNN models with power plant dataset using different similarity measurements (original vs. normalized).

to the differences exist in the feature ranges, whereas the other three measurements (Hamming, Jaccard, and Mahalanobis) are not sensitive for difference in feature ranges. Hamming and Jaccard are ratio-based similarity measures that cannot be affected by min-max normalization; therefore, their results remained unchanged in all the five datasets, Figs. 2-6. Mahalanobis similarity measurement involves the covariance of the training dataset in the calculations of similarity, which eliminates the effect of min-max normalization.

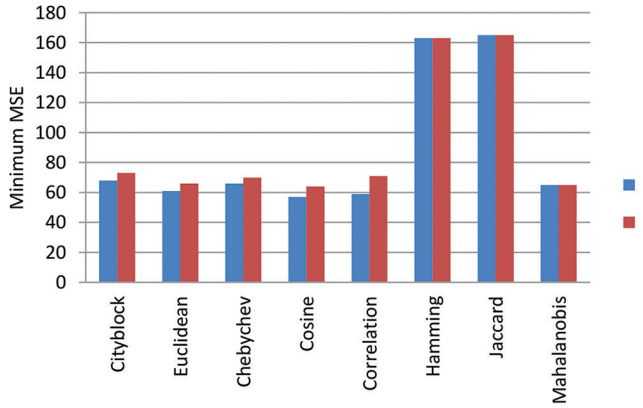

Fig. 5. Minimum MSE values of the KNN models with concrete strength dataset using different similarity measurements (original vs. normalized).

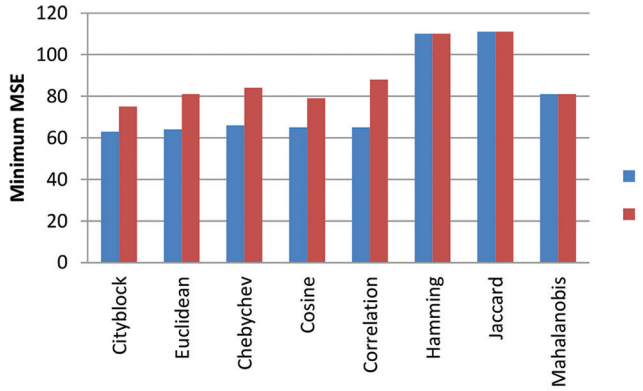

Fig. 6. Minimum MSE values of the KNN models with real estate valuation dataset using different similarity measurements (original vs. normalized).

It is noticeable that the impact of the min-max normalization for the five similarity measures (City block, Euclidean, Chebychev, Cosine, and Correlation) is not always positive, it depends on the nature of the dataset and the differences in the range of the features. An improvement in the results of airfoil self-noise and psychochemical datasets is observed, no significant impact on the results of power plant dataset is observed, and a degradation impact on the results of concrete strength and house valuation datasets is observed. This proves that these measurements are sensitive to the nature of the dataset, more precisely to the differences in the feature ranges. In addition, when the dataset has a homogeneous feature range, it will not be affected by any of the similarity measurements like in the power plant dataset, Tables III, VIII, and Fig. 4. This is also another proof that distance-based metrics of KNN are very sensitive to the differences in the feature ranges.

KNN with Mahalanobis measure provided best results among all the five datasets that have been tested in this research work. Same results observed for the original and the normalized datasets (Table VIII), also, Figs. 2-6.

Comparing the results observed in this research work with the results collected from literatures showed that this

 http://dx.doi.org/10.14500/aro.10955

Table VIII
Minimum MSE Values of the KNN Models Using Different Similarity Measurements (original vs. Normalized)

No.	KNN models using different similarity measurements	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Orig.	Norm	Orig.	Norm	Orig.	Norm	Orig.	Norm	Orig.	Norm
1	City block	39	4	40	19	13	12	68	73	63	75
2	Euclidean	40	4	41	19	15	13	61	66	64	81
3	Chebychev	41	5	42	19	16	14	66	70	66	84
4	Cosine	37	10	31	20	18	24	57	64	65	79
5	Correlation	38	12	31	20	100	54	59	71	65	88
6	Hamming	47	47	46	46	64	64	163	163	110	110
7	Jaccard	47	47	46	46	64	64	165	165	111	111
8	Mahalanobis	4	4	17	17	14	14	65	65	81	81

Table IX
Comparing the Reliability of this Research with other Previous Literatures

No.	Research work	Performance of KNN model	Similarity measurement (distance metric)	Number of datasets used	Type of application of datasets
1	Dadzie and Kwakye, 2021	No significant improvement observed	Not mentioned – probably Euclidean	1	1
2	Ambarwari, Adrian and Herdiyeni, 2020	No significant improvement observed	Not mentioned – probably Euclidean	1	1
3	Ahsan, *et al*., 2021	Improved	Not mentioned – probably Euclidean	1	1
4	Rajeswari and Thangavel, 2020	Improved	Not mentioned – probably Euclidean	5	1
5	This research work	Depends on the type of the similarity measure	Eight types of similarity measures are used	5	5

research work implemented a better research methodology and analysis, also, the results are more precise and accurate (Table IX). The previous literatures tested the KNN only with Euclidean distance method without considering other similarity measurements, whereas this research analyzed the KNN model results of each of the similarity measurements individually, without generalizing the results. Results showed that some of the KNN models could not respond to the min-max normalization. In some cases, a performance degradation recorded, which has not been mentioned in any of the previous studies. Furthermore, most of the previous literatures depended on testing only one dataset or in best cases depended on using different datasets of one application like in Rajeswari and Thangavel, 2020. This research work depended on five datasets belongs to five different real applications and different in the number of attributes. Therefore, the conclusions of this research are considered more reliable.

The KNN models utilizing Hamming, Jaccard, and Mahalanobis are not responsive to min-max data normalization but may respond positively with other normalization techniques which lay outside the scope of this research work such as z-score, soft-max, decimal-scaling, max-abs-scalar, robust scalar, and quantile transformer.

VI. Conclusion

It has been concluded from the experiments that min-max normalization may cause performance degradation to the KNN models utilizing similarity measure (City block, Euclidean, Chebychev, Cosine, and Correlation). Therefore, testing datasets with and without min-max data normalization are recommended before considering their results. Attaching min-max with KNN models utilizing (Hamming, Jaccard, and Mahalanobis) is not recommended, because it has no effect on the performance of these models.

The possible degradation impact of using min-max data normalization on the KNN models utilizing similarity measurements (City block, Euclidean, Chebychev, Cosine, and Correlation) return to eliminating the natural domination of one of the attributes by the min-max normalization, and this leads to performance degradation like in the two datasets; concrete strength and house valuation. Therefore, it is better to test the KNN model with both the original dataset and the normalized dataset before deciding if the min-max data normalization is useful or not.

It makes no sense to use min-max normalization with KNN models utilizing (Hamming, Jaccard, and Mahalanobis), because they use the common variance of the dataset in the similarity calculations.

Furthermore, it is not necessary to use min-max data normalization with homogeneous datasets whatever the similarity measurement is, like with power plant dataset.

References

Ahsan, M.M., Mahmud, M.A.P, Saha, P.K., Gupta, K.D. and Siddique, Z., 2021. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9, p.52.

Aksu, G., Güzeller, C.O. and Eser, M.T., 2019. The effect of the normalization method used in different sample sizes on the success of artificial neural network model. *International Journal of Assessment Tools in Education*, 6(2), pp.170-192.

Ambarwari, A., Adrian, Q.J. and Herdiyeni, Y., 2020. Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 4(1), pp.117-122.

Bhardwaj, C.A., Mishra, M. and Desikan, K., 2018. *Dynamic Feature Scaling for K-Nearest Neighbor Algorithm*. Available from: https://www.arxiv.org/ftp/arxiv/papers/1811/1811.05062.pdf [Last accessed on 2022 Feb 01].

Brooks, T.F., Pope, D.S. and Marcolini, M.A., 1989. *Airfoil Self-noise and Prediction (NASA Reference Publication). Technical Report 1218*. National Aeronautics and Space Administration, United States.

Cha, S.H., 2007. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 4(1), pp.300-307.

Cover, T. and Hart, P., 1976. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), pp.21-27.

Dadzie, E. and Kwakye, K., 2021. *Developing a Machine Learning Algorithm-Based Classification Models for the Detection of High-Energy Gamma Particles*. Available from: https://www.hal.archives-ouvertes.fr/hal-03425661 [Last accessed on 2022 Feb 01].

Dua, D. and Graff, C., 2019. *UCI Machine Learning Repository*. University of California, School of Information and Computer Science, Irvine, CA, p.27. Available from: http://www.archive.ics.uci.edu/ml [Last accessed on 2022 Feb 01].

Fix, E. and Hodges, J.L., 1951. *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, United States.

Jayalakshmi, T. and Santhakumaran, A., 2011. Statistical normalization and back propagationfor classification. *International Journal of Computer Theory and Engineering*, 3, pp.89-93.

Pires, I.M., Hussain, F., Garcia, N.M., Lamesk, P. and Zdravevski, E., 2020. Homogeneous data normalization and deep learning: A case study in human activity classification. *Future Internet*, 12, pp.1-14.

Prasetyo, J., Setiawan, N.A. and Adji, T.B., 2020. Improving normalization method of higher-order neural network in the forecasting of oil production. *E3S Web of Conferences*, 200, p.02016.

Rajeswari, D. and Thangavel, K., 2020. The performance of data normalization techniques on heart disease datasets. *International Journal of Advanced Research in Engineering and Technology*, 11, pp.2350-2357.

Rana, P.S., 2013. *Physicochemical Properties of Protein Tertiary Structure Data Set. UCI Machine Learning Repository*. Available from: https://www.archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure

Shorman, A.R.A., Faris, H., Castillo, P.A., Merelo, J.J. and Al-Madi, N., 2018. The influence of input data standardization methods on the prediction accuracy of genetic programming generated classifiers. In: *Proceedings of the 10th International Joint Conference on Computational Intelligence*. SciTePress, Portugal, pp.79-85.

Singh, B.K., Verma, K. and Thoke, A.S., 2015. Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification. *International Journal of Computer Applications*, 116(19), pp.975-8887.

Tüfekci, P., 2014. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power and Energy Systems*, 60, pp.126-140.

Yeh, I.C. and Hsu, T.K., 2018. Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65, pp.260-271.

Yeh, I.C., 1998. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12), pp.1797-1808.