Grenze International Journal of Engineering and Technology, Jan Issue

Skewness based Outlier Elimination Algorithm for Pima Diabetes Dataset

Pushkar Joglekar[1], Tejaswini Katale[2], Surabhi Deshpande[3], Saurabh Kelgandre[4], Aishwarya Katale[5] and Aakash Chotrani[6]
[1-4]Vishwakarma Institute of Technology, Pune, India
Email: tejaswini.katale20@vit.edu
[5]Symbiosis International University, Pune, India
[6]LDRP Institute of Technology and Research, India
Email: {surabhi.deshpande20@vit.edu, saurabh.kelgandre20@vit.edu}

***Abstract*— An Outlier is a data point which appears to be distinct from the majority of the other datapoints within a dataset. Various factors, such as measurement errors, data entry problems, natural variability, and extreme events, can lead to outliers. They may lead to erroneous evaluation measures, biased data analysis, and incorrect model performance resulting in in-suitable outcomes. Hence, identification and elimination of outliers is a crucial pre-processing step. This research paper suggests an algorithm for correctly removing outliers from the Pima Diabetes Dataset. Ittakes into consideration the correlation values of the features with the outcome column, further arranging them in descending order. The normal distribution graphs for each feature in the dataset are plotted for data analysis. For the Diabetes dataset, the features are either Positive (Right) skewed or Symmetrical. For the Symmetrical curve features, it eliminates data points which are lying beyond three standard deviations i.e (μ - 3σ) and (μ + 3σ), whereas for Positively skewed features, it eliminates those data points which are greater than (μ + 2.5σ). Starting from the feature having the highest correlation value, based on the skewness it eliminates the outliers by iterating the features one by one based on the order of correlation values. The proposed algorithmis successful in elimination of 108 rows from Pima Diabetes Dataset. To check the efficacy ofthe algorithm, it is compared with four techniques - Local Outlier Factor, Mahalanobis Distance, Multivariate Normal Distribution (N Dimensional) and DBSCAN. Also, for better evaluation the number of outliers eliminated by each technique lies in the same range. Furthermore, after elimination of outliers, three Machine Learning are implemented - Logistic Regression, Random Forest and Decision Tree. Among all the techniques, logistic regression on the proposed algorithmgives the highest accuracy of 84%, highest precision of 0.88 and highest recall of 0.65.**

***Index Terms*— Outliers, Correlation Matrix, Normal Distribution Curve, Mean, Standard deviation.**

I. INTRODUCTION

Outlier is a sample value that differs notably from the mean of the measurement series, can be caused by

Grenze ID: 01.GIJET.10.1.158

many factors [14]. Outliers are the data points which deviate significantly from the majority ofthe average data points and the data's central tendencies like mean, median and mode. Median is a robust central tendency which represents the middle value of the data. It is not affected by the skewed distributions or extreme Outliers and hence for the proposed algorithm, the NA values within the dataset are replaced by the median of that column. Outliers in a data collection can impair or distort inferential processes like estimation, making traditional tests of hypotheses worthless [4]. One Outlier in a very large sample can totally skew the skewness of a distribution, making it a very unstablestatistic [3]. The outliers originally existed as the by-product of many clustering algorithms, and this directly inspired many scholars to improve some clustering algorithms to obtain the early major outlier detection algorithms [12]. There are various advantages to removing outliers from a dataset, including reduced noise, improved robustness, enhanced model performance, reduced risk of biasedness and improved data analysis and normalization. The proposed algorithm is implemented on the Pima Diabetes Dataset which is a binary Outcome dataset; 0: No Diabetes, 1: Diabetes. It is a chronic illness that develops either when the pancreas is not able to generate sufficient insulin or when the body does not utilize the insulin produced effectively [2]. The dataset used is taken from the National Institute of Diabetes and Digestive and Kidney Disease. It consists of 768 observations and 9 features of Integer, Continuous and Categorical data types. Correlation values between the characteristics are required in order to assess the significance and influence of the columns on the total dataset's result. Correlation matrix is a statistical measure which measures the strength and relationship between two continuous features. It lies between -1 to 1. A positive correlation means that if one variable increases, the other one also has a tendency to increase. The magnitude of the correlation value determines how strongly connected the features are; whereas in a negative correlation there is an inverse relationship between the variables; if a variable increases, the other tends to decrease. Gaussian Curve also known as Normal Distribution Curve is a well-known bell-shaped symmetric curve which is used to determine the skewness of the features. It takes into consideration the mean (μ) to determine the center of the data and the standard deviation (σ) to evaluate the spread and width of the distribution as shown in Fig.1. Earlier studies suggest that it is not always possible that the data is symmetrical by the mean which introduces the concept of Skewness. It is a measure of the asymmetry of the distribution of a variable [5]. Based on the distribution of the data values from the mean, the curve can be of three types: Positive (Right) Skewed, Symmetric and Negative (Left) Skewed. As displayed in Fig.2, for the positive skewed curve, most ofthe data points will lie on the left side of the central tendencies whereas for the negative skewed curve,most of the data values will lie on the right side. The mean in a normally distributed data represents the central tendency of the values of the data [6]. For the symmetric curve, all the central tendencies will be equal with the same number of data values on both sides of the mean.

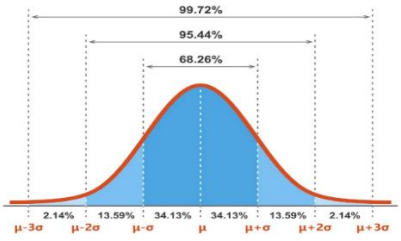

Figure 1: Gaussian Distribution curve [21]

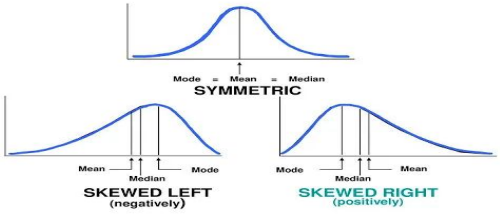

Figure 2: Skewed Distribution Curves[22]

The Normal Distribution curves for the features are shown in Fig.3, Fig.4, Fig.5, Fig.6, Fig.7, Fig.8, Fig.9, Fig.10. For the Diabetes dataset, the skewness of the features is of two types: Positive and Symmetrical.

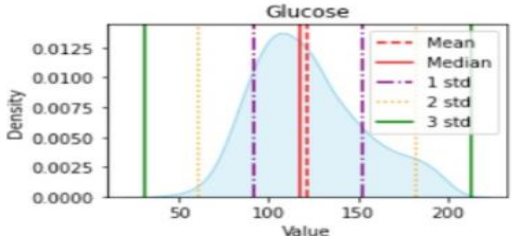

Figure 3: Gaussian Curve of Glucose

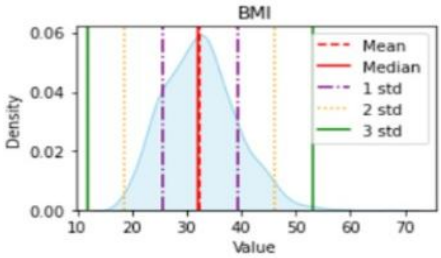

Figure 4: Gaussian Curve of BMI

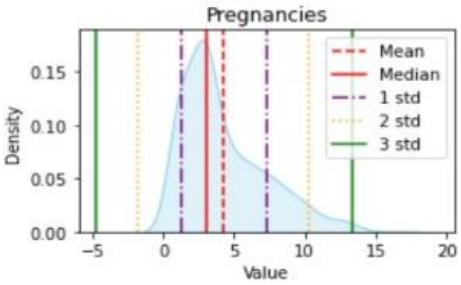

Figure 5: Gaussian Curve of Pregnancies

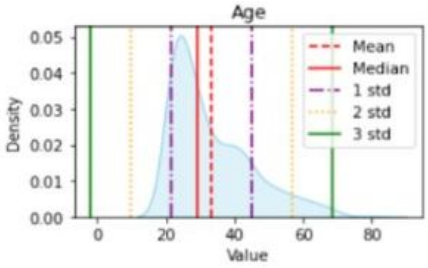

Figure 6: Gaussian Curve of Age

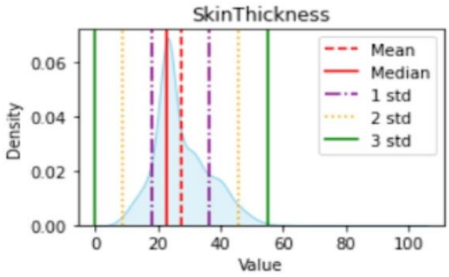

Figure 7: Gaussian Curve of SkinThickness

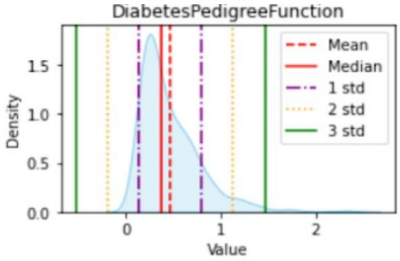

Figure 8: Gaussian Curve of DiabetesPedigree

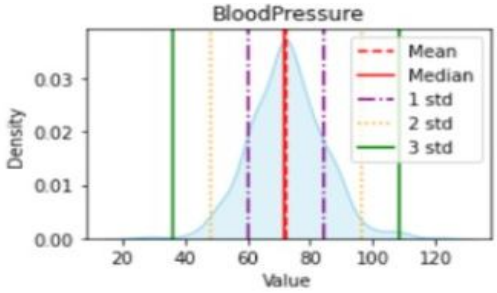

Figure 9: Gaussian Curve of BloodPressure

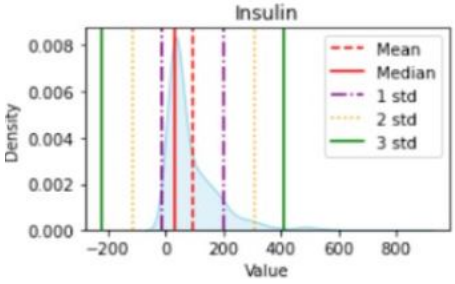

Figure 10: Gaussian Curve of Insulin

TABLE I. CORRELATION VALUES AND TYPE OF SKEWNESS

Columns	**Correlation Value with Outcome column**	**Skewness**
Glucose	0.49	Right
BMI	0.31	Normal
Pregnancies	0.24	Right
Age	0.23	Right
SkinThickness	0.18	Right
DiabetesPedigreeFunction	0.17	Right
BloodPressure	0.16	Normal
Insulin	0.14	Right

II. Literature Review

Paper [1] introduces a distance based multivariate Outlier elimination technique. It uses the ideologyof Principal Component Analysis. They also claim that unlike other algorithms, it does not requireprior definition of any operation and also results in fast processing time for calculation of distances.The authors of Paper [4] propose a method for multiple Outliers identification in Linear Regression which includes inward step wise procedure and the outward step wise procedure. They have concluded that the proposed methodology is suitable for high, large and complex datasets. In Paper [7], the authors implemented 2 feature selection algorithms on HTRU - S Pulsar Stars Dataset - Grid Searchand Recursive feature elimination (RFE). In Paper [8], the authors have a done comparison of three Outlier Elimination Techniques - Cluster Based, Density Based and Distance based. Further they have discussed Statistical based, distance based and high dimensional approaches. In Paper [9], the authors have proposed a Global High Dimensional Outlier Detection Algorithm which outperformsother existing Outlier Elimination techniques. The authors of Paper [10] aims to improve the accuracy of fraud detection dataset for which they have implemented hybrid oversampling techniques and random under-sampling. The conclusion of the authors is that the proposed algorithm outperformed the hybrid classifier except for KNN classifier. Paper [11] is a comparison study of accuracies of 6 classification techniques - Linear Discriminant Analysis (LDA), K-Nearest Neighbor (K-NN), Linear SVM, Radial Basis Function Support Vector Machine (RBF SVM), Sigmoid SVM, Logistic Regression(LR), Naive Bayes (NB), Classification and Regression Trees (CART) which was evaluated on Pima Diabetes Dataset. Highest accuracy of 87.01% was given by SVM. In Paper [13], the proposed methodis an integration of Random Forest, Interquartile Range (IQR), Deep Learning algorithms and dataset class detection for Medical Internet of Things (MIOT) wearable sensors. The final model accuracy results in 99.71 % for Area Under Curve (AUC). Paper [15] involves a comprehensive survey of outlier techniques from 2009 - 2019 categorizing them as distance- based, clustering-based and learning basedmethods further discussing their pros and cons. They have also mentioned ensemble techniques like isolation forest, bagging, boosting and extreme gradient boosting. They further mention anti hub techniques, heuristic approaches and natural neighbor concepts for efficient identification of outliers.

III. Methodology

1. After loading the dataset, the NA values are replaced by the median of the column. Since the median is a robust central tendency, only the relative location of the sorted data will have an impact on it rather than the values of Outliers.
2. Correlation values of all the features with the Outcome column are calculated.
3. The features are organized in descending order according to the correlation values because itis presumably that the columns with higher correlation are believed to be more essential and contributing to the prediction of outcome.
4. Furthermore, Gaussian curves for all the features are visualized.
5. For this dataset, the features are either Normal skewed and right skewed.
6. For the Normal skewed features, Outliers are those data points which lie beyond 3 standard deviations. Hence the data points which are lying beyond $(\mu - 3\sigma)$ and $(\mu + 3\sigma)$ are eliminated.
7. For the Right skewed features, the majority of the data points lie on the left side of the graph,so in this scenario, the midpoint of $(\mu + 2\sigma)$ and $(\mu + 3\sigma)$ which is equal to $(\mu + 2.5\sigma)$ is calculated. Hence the data points which are greater than $(\mu + 2.5\sigma)$ are eliminated.
8. The elimination of Outliers is implemented by the order of Correlation Values from Table.1

(a) Starting from Glucose, being a Right Skewed feature, implementing Pt. 6 which resulted in identification of 2 Outliers [768 - 2] = 766.

(b) Further for BMI, Pt.5, resulted in elimination of 5 Outliers [766 - 5] = 761.

(c) For Pregnancies, Age, SkinThickness and DiabetesPedigreeFunction; as all are Right skewed implementing Pt.6.

Pregnancies: [761 - 23] = 738

Age: [738 - 21] = 717

SkinThickness: [717 - 9] = 708

DiabetesPedigreeFunction: [708 - 19] = 689

(d) For BloodPressure, being a symmetrical feature, Pt.5 is implemented, [689 - 6] = 683
(e) For Insulin, Pt.6 is implemented; [683 - 23] = 660
(f) As a result, total [768 - 660] = 108 rows are considered as Outliers.
(g) Machine learning algorithms are used to compare effectiveness - Logistic Regression, Random Forest and Decision Tree. The flowchart of the proposed algorithm is shown in Fig.11

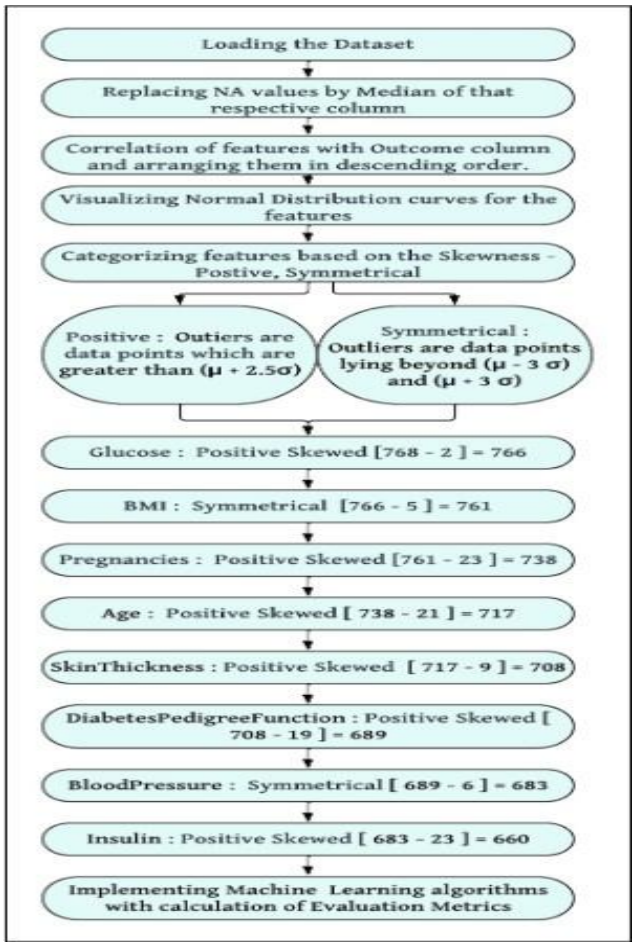

Figure 11: Flowchart of Proposed Algorithm

IV. RESULTS

To check the effectiveness of the proposed algorithm, it is compared with 4 Outlier Elimination algorithms, taking accuracy, precision and recall values, resulting from Logistic Regression and Random Forest into account. As shown in Table. 2, to check the efficacy of the algorithm, for all the four algorithms, a similar number of Outliers are eliminated by giving a certain threshold value. - Proposed Algorithm (108), Local Outlier Factor (91), Mahalanobis Distance (113), Multivariate Normal Distribution (N Dimensional) (107) and DBSCAN (102). After comparison, the proposed algorithm gives better accuracy, precision and recall values for all the 3 Machine Learning Algorithms - Logistic Regression, Random Forest and Decision Tree. For Logistic Regression, highest accuracy is 84%, highest precision is 0.88 and highest recall value is 0.65 resulted by the proposed algorithm. For Random Forest highest accuracy is 81%, highest precision is 0.80 and highest recall value is 0.60 resulted by proposed algorithm. For Decision Tree, highest accuracy is 75%, highest precision is 0.67 and highest recall value is 0.54 resulted by proposed algorithm. Thus, the proposed algorithm successfully identifies and eliminates 108

rows from Pima Diabetes Dataset with the highest accuracy of 84%, highest precision of 0.88 and highest recall of 0.65 resulted by Logistic Regression.

TABLE II. COMPARISON OF EVALUATION METRICS WITH PROPOSED ALGORITHM

Algorithms		**Proposed Algorithm**	**Local Outlier Factor**	**Mahalanobis Distance**	**Multivariate Normal Distribution (N Dimensional)**	**DBSCAN**
No of Outliers		**108**	**91**	**113**	**107**	**102**
Logistic Regression	Accuracy	**0.84**	0.77	0.78	0.78	0.79
	Precision	**0.88**	0.79	0.77	0.82	0.64
	Recall	**0.65**	0.52	0.48	0.5	0.51
Random Forest	Accuracy	**0.81**	0.72	0.77	0.76	0.73
	Precision	**0.80**	0.71	0.71	0.75	0.48
	Recall	**0.60**	0.45	0.53	0.47	0.45
Decision Tree	Accuracy	**0.75**	0.65	0.71	0.72	0.67
	Precision	**0.67**	0.54	0.58	0.61	0.39
	Recall	**0.54**	0.47	0.48	0.52	0.42

IV. CONCLUSION

In conclusion, analysing and retrieving huge Datasets can be difficult. Identification and elimination of correct Outliers is a crucial pre-processing step. For handling, they also need sophisticated software programmes and powerful hardware. Hence, the suggested approach can be applied to various datasets by taking skewness and correlation into account.

V. ACKNOWLEDGEMENT

The tremendous guidance and support of my mentor Sir. Pushkar Joglekar made it feasible to complete this research study. Their advice, help, and unceasing encouragement made it possible for the study project to be completed successfully. I sincerely appreciate and am grateful for their participation and significant contribution.

REFERENCES

[1] Ugah Tobias Ejiofor, Arum Kingsley Chinedu, Charity Uchenna Onwuamaeze, Everestus Okafor Ossai, Henrrietta Ebele Oranye, Nnaemeka Martin Eze, Mba Emmanuel Ikechukwu, Ifeoma Christy Mba, Comfort Njideka Ekene-Okafor, Asogwa Oluchukwu Chukwuemeka, Nkechi Grace Okoacha, "A New Procedure for Multiple Outliers Detection in Linear Regression", Article in Mathematics and Statistics · August 2023, DOI: 10.13189/ms.2023.110416

[2] Atiek Iriany, Henida Ratna Ayu Putri, Harry Maringan Tua, "K-means Non-Hierarchical Cluster and DBSCAN Outlier Detection in the Grouping of Stock Issuers", Journal of Management Information and Decision Sciences

[3] J. Carmona, I. Lopez, J. Mateo, L. Jimenez, and E. Aldana, "A Distance-Based Method for Outlier Detection on High Dimensional Datasets", IEEE Latin America Transactions, Vol. 18, No. 3, March 2020

[4] Hanaa Torkey, Elhossiny Ibrahim, EZZ El Din Hemdan, Ayman El Sayed, Marwa A. Shouman, "Diabetes classification application with efficient missing and outliers' data handling algorithms", Complex & Intelligent Systems (2022) 8:237–253 https://doi.org/10.1007/s40747-021-00349-2

[5] Raghad Sehly, Mohammad Mezher, "Comparative Analysis of Classification Models for Pima Dataset", 2020 International Conference on Computing and Information Technology, University of Tabuk, Kingdom of Saudi Arabia

[6] Haitao Lin, Xiangru Li, Ziying Luo, “Pulsars Detection by Machine Learning with Very Few Features”, Pulsars detection by machine learning with very few features, Article in Monthly Notices of the Royal Astronomical Society · April 2020 DOI:10.1093/mnras/staa218
[7] Cuie Zheng, Yun Lio, Dajun Sun, Yun Zhao, “The Outlier Elimination Methodology of UltraShort BaseLine Positioning Data”, 2019 International Conference on Control, Automation and Information Sciences (ICCAIS)
[8] R. S. Sonawane, S. S. Banait, “A Review on Outlier Detection and Removal Techniques”, IJRAR
[9] Hongzhi Wang, Mohamed Jaward Bah, Mohamed Hammad, “Progress in Outlier Detection Techniques: A Survey”, IEEE Access, DOI: 10.1109/ACCESS.2019.2932769
[10] Sihem Jebari, Abir Smiti, Aymen Louati, “AF-DBSCAN: An unsupervised Automatic Fuzzy Clustering method based on DBSCAN approach”, IEEE International Work Conference on Bioinspired Intelligence, July 3-5, 2019, Budapest, Hungary
[11] Shubin Su, Limin Xiao, (Member, IEEE), Li Ruan, Fei Gu, Shupan Li, Zhaokai Wang, Rongbin Xu, “An Efficient Density-Based Local Outlier Detection Approach for Scattered Data”, IEEE Access, DOI: 10.1109/ACCESS.2018.2886197
[12] Md. Maniruzzaman, Md. Jahanur Rahman, Md. Al-MehediHasan, Harman S. Suri, Md. Menhazul Abedin Ayman El-Baz, Jasjit S. SurI, “Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers”, Journal of Medical Systems (2018) https://doi.org/10.1007/s10916-018-0940-7
[13] Harshada C. Mandhare, S. R. Idate, “A Comparative Study of Cluster Based Outlier Detection, Distance Based Outlier Detection and Density Based Outlier Detection Techniques”, International Conference on Intelligent Computing and Control Systems
[14] T. Maruthi Padmaja, Narendra Dhulipalla, Raju S. Bapi, P. Radha Krishna, “Unbalanced Data Classification Using extreme outlier Elimination and Sampling Techniques for Fraud Detection”, 15th International Conference on Advanced Computing and Communications, DOI: 10.1109/ADCOM.2007.74
[15] Parmeet kaur, “Outlier Detection using K Means and Fuzzy Min Max Neural Network in Network Data”, 2016 8th International Conference on Computational Intelligence and Communication Networks, DOI: 10.1109/CICN.2016.142
[16] Nidhi Nigam, Tripti Saxena, Vineet Richhariya, “Global High Dimension Outlier Algorithm for Efficient Clustering & Outlier Detection”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN)
[17] Prasanta Gogoi, D.k. Bhattacharyya, B. Borah, Jugal K. Kalita, “A Survey of Outlier Detection Methods in Network Anomaly Identification”, Published by Oxford University Press on behalf of The British Computer Society, doi:10.1093/comjnl/bxr026
[18] Denis Cousineau, Sylvain Chartier, “Outliers detection and treatment: a review”, Article in International Journal of Psychological Research · June 2010 DOI: 10.21500/20112084.844
[19] R. Vijayabhanu, V.Radha, “Recognition and Elimination of Missing Values and Outliers From An Anaerobic Wastewater Treatment System Using K-Means Cluster”, 2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)
[20] Davaadorjin Monhor, Shuzo Takemoto, “Understanding the concept of outlier and its relevance to the assessment of data quality: Probabilistic background theory”, Article in Earth Planets and Space · November 2005 DOI: 10.1186/BF03351881
[21] https://www.simplypsychology.org/normal-distribution.html
[22] https://lamduy-nguyen.medium.com/skewness-definition-problem-and-reducing-methods-a5ee25a8b6ca