Integrated ECOD-KNN Algorithm for Missing Values Imputation in Datasets: *Outlier Removal*

Tsitsi Jester Mugejo[1]
Department of Cloud Computing
School of Information Science and Technology, Harare Institute of Technology
Harare, Zimbabwe

Weston Govere[2]
Department of Cloud Computing
School of Information Science and Technology,
Harare Institute of Technology
Harare, Zimbabwe

Abstract:- Missing data cause the incompleteness of data sets and can lead to poor performance of models which also can result in poor decisions, despite using the best handling methods. When there is a presence of outliers in the data, using KNN algorithm for missing values imputation produce less accurate results. Outliers are anomalies from the observations and removing outliers is one of the most important pre-processing step in all data analysis models. KNN algorithms are able to adapt to missing value imputation even though they are sensitive to outliers, which might end up affecting the quality of the imputation results. KNN is mainly used among other machine learning algorithms because it is simple to implement and have a relatively high accuracy. In the literature, various studies have explored the application of KNN in different domains, however failing to address the issue of how sensitive it is to outliers. In the proposed model, outliers are identified using a combination of the Empirical-Cumulative-distribution-based Outlier Detection (ECOD), Local Outlier Factor (LOF) and isolation forest (IForest). The outliers are substituted using the median of the non-outlier data and the imputation of missing values is done using the k-nearest neighbors algorithm. For the evaluation of the model, different metrics were used such as the Root Mean Square Error (RMSE), (MSE), R2 squared (R^2) and Mean Absolute Error (MAE). It clearly indicated that dealing with outliers first before imputing missing values produces better imputation results than just using the traditional KNN technique which is sensitive to outliers.

Keywords:- *Imputation; Outlier; Missing Values; Incomplete; Algorithm.*

I. INTRODUCTION

Missing data cause the incompleteness of data sets and can lead to poor performance of models which also can result in poor decisions, despite using the best handling methods. Analyses of datasets containing missing values can perpetuate deriving actions from a biased model. In this paper, we want to reveal the impact of how solving missing data using KNN algorithm may produce less accurate results, especially when there is a presence of outliers in the data. Additionally, there is a demonstration of how outliers can be identified using the Empirical-Cumulative-distribution-based Outlier Detection (ECOD), Local Outlier Factor (LOF) and isolation forest (IForest), how the outliers where substituted using the median of the non-outlier data and the imputation of missing values using KNN algorithm in a single model.

Outliers are anomalies from the observations and removing outliers is one of the important pre-processing step in all data analysis models (1). It is important to first identify the outliers, in this paper, using the Outlier Detection, to be able to remove or substitute them. Data is bound to have some noisy data or rather outliers which definitely affect the KNN missing value imputation process and the performance of the trained models (2). It is of most importance to filter all the noisy data from any training dataset. This step should be the first before imputing missing values to have more accurate results. Imputation result will not be good enough if performed before outlier handling.

Missing values are important when involving big data (3), which is very large amounts of data or large datasets which requires analysis and storage. Missing values generally pose a weakness to models (4) as they affect the quality of results especially with prediction systems. In the pre-processing stage of datasets with numeric values, it is noted that one of the main challenges is the processing of missing values. So it is important to deal with missing values in our datasets during pre-processing (5).

Challenges may also arise from choosing the wrong handling method of missing values (6) and this also affect the effectiveness of any model. Previous studies have provided information on imputation using the KNN algorithm and other various extensions of the algorithm, however failing to consider outlier detection and normalization before the missing value imputation process. The performance of the KNN imputation method can be greatly improved with solving outliers and normalization of the data (7). It has been proved that using normalization and imputation mean together is more accurate than the original mean and median methods (8).

This study is taking note of the outliers, by firstly detecting them using ECOD and substituting them using mean as it is more effective and then proceeding to impute the missing values in the datasets, for an improved accuracy of the imputation result. It has been noted that this combination has never been used in previous studies of imputation using KNN or other imputation methods, although it has been proved by this paper to improve the accuracy of the missing values imputation process.

II. LITERATURE REVIEW

Many studies have been done to solve the issue of missing values in datasets. Incompleteness of data is handled depending on the type and requirements mainly. The two main methods of imputation are statistics and machine learning. These methods then generate values or approximations from the observable variables in order to replace the missing values (9). KNN is mainly used among other machine learning algorithms because it is simple to implement and have a relatively high accuracy. In the literature, various studies have explored the application of KNN in different domains.

This highlights the adaptability of KNN algorithms in addressing missing values and improving classification accuracy in diverse applications. In summary, the literature review showcases the significance of using the KNN algorithm for imputing missing values in datasets across various domains, including proteomics, recommendation systems, medical diagnostics, and other domains.

In the field of proteomics, (10) it is highlighted that there is a complexity of identifying the subcellular locations of proteins, especially when proteins can exist in multiple locations simultaneously (11). To address missing values in proteomic data, (12) the Cluster-based KNN (CKNN) imputation method was introduced, which incorporates local data clustering for improved quality and efficiency (13).

In the context of movie recommender systems, a comparative study was conducted (14) on pre-processing algorithms for Singular Value Decomposition (SVD) to help data managers choose the most suitable algorithm for their business needs (15). This study underscores the importance of selecting the right imputation method to enhance the accuracy and reliability of data analysis. Furthermore, in the medical field, (16) the study focused on missing value estimation methods for arrhythmia classification, emphasizing the significance of handling missing values in datasets to ensure accurate classification results (18).

Furthermore, a novel KNN variant (KNNV) algorithm was introduced (17) for accurate classification of COVID-19 based on incomplete heterogeneous data, showcasing improved results through experimental work (18). The KNNV algorithm addresses incompleteness by imputation and heterogeneity by converting categorical data into numerical values. Moreover, a hybrid missing data imputation method called KI was proposed (19), which combines k-nearest neighbors and iterative imputation algorithms to address missing values effectively (20). This approach leverages similarity learning techniques to impute missing data accurately.

Fig 1 shows the experimental design of the systems of most of the current studies (21). The studies show that they introduce missing values as the first step, if a dataset with no missing values is being used. An imputation algorithm is then picked for the imputation process. The imputed result is then evaluated using various different metrics.

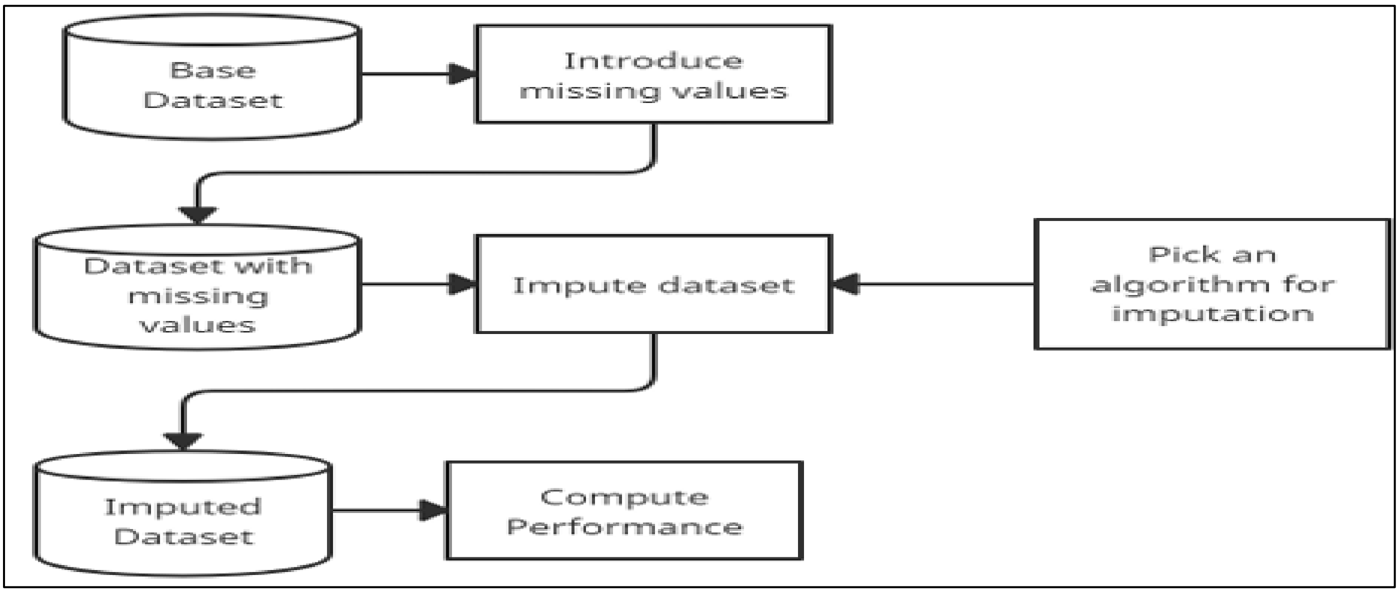

Fig 1 Block Diagram of the Proposed System Experimental Design.

Overall, the literature highlights the significance of the KNN algorithm in imputing missing values in datasets. Researchers have developed novel approaches, such as the MKDF-WKNN classifier and KNNV algorithm, to enhance the accuracy of classification models when dealing with incomplete data (22). Additionally, hybrid methods like KI have been proposed to improve missing data imputation by incorporating k-nearest neighbors and iterative algorithms. The studies discussed emphasize the importance of selecting appropriate imputation methods, which all involve KNN, to enhance data quality, analysis accuracy, and classification performance. However, all the KNN algorithms used by different researchers fail to address the issue that KNN is sensitive to outliers and can affect the result of the missing values imputation process.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1459

III. METHODOLOGY

The first step was data exploration. Pandas was used for the data frames which makes it easy to work with structured data. To support the large datasets and calculate the average arithmetic set of values in the datasets, Numpy was also used and for plotting graphs for visual comprehension Matplot was also used. The outliers are identified using a combination of the Empirical-Cumulative-distribution-based Outlier Detection (ECOD), Local Outlier Factor (LOF) and isolation forest (IForest). The outliers are substituted using the median of the non-outlier data and the imputation of missing values is done using the k-nearest neighbors algorithm. It identifies the 'k' nearest data points to the missing value based on a distance metric. For numerical data, the mean of these neighbors is used to replace the missing value. For categorical data, the most frequent category (mode) among the neighbors is used. This approach leverages the similarity between data points to provide a more accurate imputation compared to simple mean or mode imputation. Fig 2 illustrates the experimental design of the proposed system.

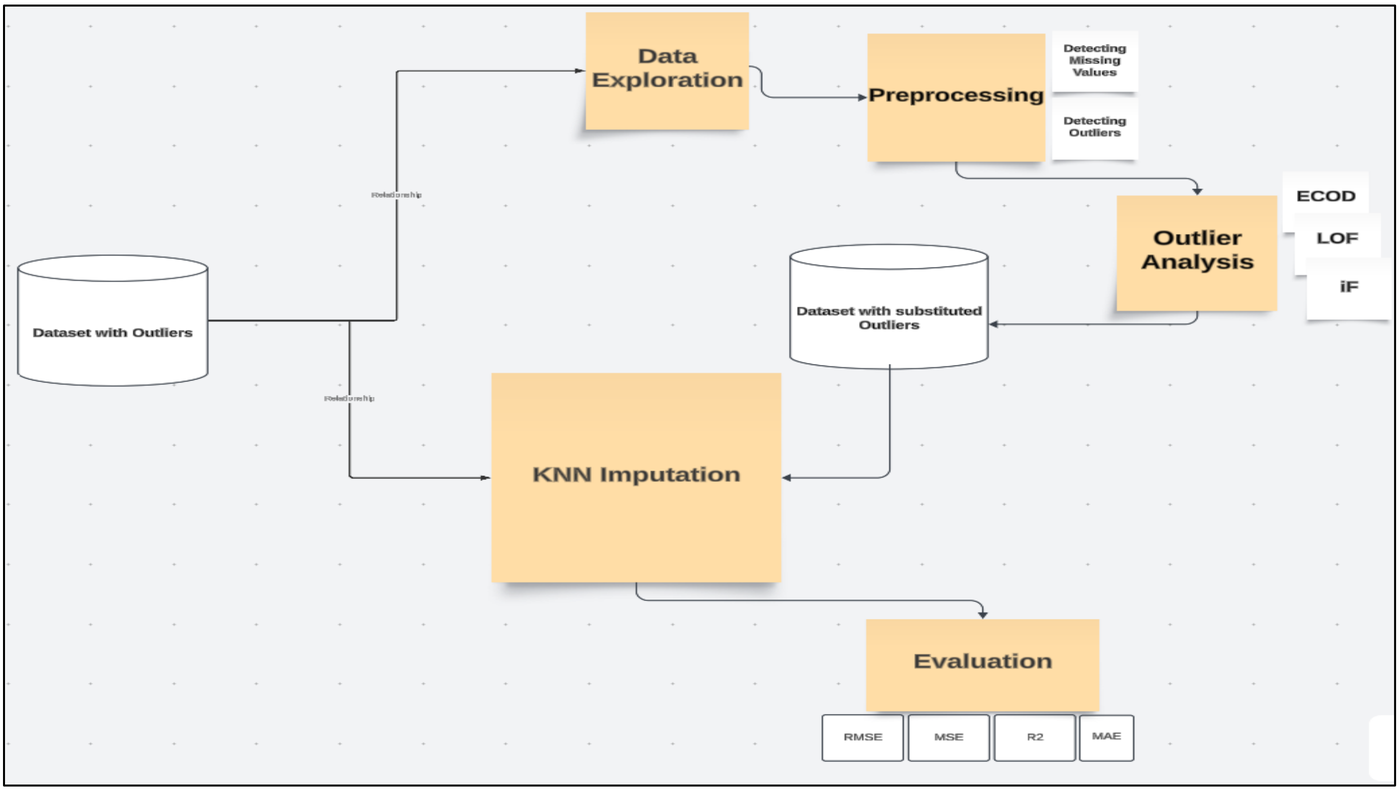

Fig 2 Block Diagram of the Proposed System Experimental Design.

➢ *Dataset Preparation*

This experiment was implemented using five datasets from the Kaggle Website tabulated in Table 1. The datasets were loaded for feature extraction and standardization of the features. Preprocessing to check if outliers and missing values were present was done. This would then lead to next step of Outlier Analysis where the emphasis of the experiment is. This would involve detecting outliers and substituting them with the results from the analysis.

➢ *Evaluation Criteria*

For the evaluation of the model, different metrics were used such as the Root Mean Square Error (RMSE), (MSE), R2 squared (R^2) and Mean Absolute Error (MAE).

- RMSE metric is used in machine learning to compute the difference between the observed value and the imputed value.
- MSE is used to measure the average of the squared differences between the predicted values and actual target values. The lower the MSE, the closer the models results are to the true values.

Table 1 Details of the Datasets used

Dataset No.	**Details of the datasets used**		
	Dataset Name	***Rows***	***Attributes***
1	Dissolved O2 River Water	3500	37
2	Crop Recommendation	1470	20
3	Online Course Engagement	4650	12
4	Health Care Diabetes	1460	6
5	Amazon cell phone and accessories	10448570	12

- R^2 or the Coefficient of Determination is another metric used to evaluate the model's goodness. It is known as the Goodness of fit. R2 squared score move towards one, which means that regression line moves towards perfection. It was used in this study because of its ability to measure variability.
- MAE, which is the mean absolute error matches the error value units to the predicted target value units. MAE has changes that are intuitive and it penalize large errors more. The square of the error value increases or inflates the mean error value.

IV. RESULTS AND DISCUSSION

The proposed model consists of outlier removal and imputation whilst the other KNN imputation techniques does not take outlier removal in consideration. The results in table 2-6 below show both the proposed model and basic KNN evaluation results for comparison.

Of all the metrics used, the proposed model seem to have better results compared to KNN, as shown from the above tables of various datasets. RMSE evaluation had the worst results for both models and in all the datasets, but better for the proposed model based on the simulation results.

Table 2 Dissolved O2 River Water Results

Dissolved O2 River Water Results		
Metric	***ECOD-KNN(Proposed System)***	***KNN***
RMSE	1.775	1.958
MSE	3.246	3.834
R^2	0.542	0.548
MAE	1.320	1.425

Table 3 Online Course Engagement Data Dataset

Online Course Engagement Data Dataset		
Metric	***ECOD-KNN(Proposed System)***	***KNN***
RMSE	1.523	2.518
MSE	0.223	1.331
R^2	0.742	0.948
MAE	1.202	1.282

Table 4 Crop Recommendation Dataset

Crop Recommendation Dataset		
Metric	***ECOD-KNN(Proposed System)***	***KNN***
RMSE	0.923	1.210
MSE	0.389	1.765
R^2	0.427	0.812
MAE	1.897	3.423

Table 5 Health Care Diabetes Dataset

Health Care Diabetes Dataset		
Metric	***ECOD-KNN(Proposed System)***	***KNN***
RMSE	1.302	1.838
MSE	0.482	0.935
R^2	0.923	1.275
MAE	1.193	1.585

Table 6 Amazon of cell phone and accessories Product Ratings Dataset

Amazon of cell phone and accessories Product Ratings Dataset		
Metric	***ECOD-KNN(Proposed System)***	***KNN***
MSE	0.0153	0.0529
R^2	0.9645	0.9742
MAE	0.046	0.245

V. CONCLUSION

Important information goes missing when a dataset has missing values. Missing values have to be imputed to avoid such scenarios. Imputing missing values insures that the dataset is complete and this will help the various models to produce accurate results where decision making is based on. KNN is widely used to impute missing values among other techniques. However one of the disadvantages is that, it sensitive to outliers, which was the focus of this study. The study focused on detecting outliers using a combination of Local Outlier Factor (LOF), isolation forest (IForest) and ECOD. After averaging the detectors, for the outliers result, they are replaced in dataset with median of the non-outlier data. The k-nearest neighbor's algorithm is then used to impute the missing values. After testing the model with more than 5 different datasets, the evaluation criteria using RMSE, MSE, R^2 and MAE clearly indicated that dealing with outliers first before imputing missing values produces better imputation results than just using the traditional KNN technique which is sensitive to outliers. Despite the good performance of the proposed ECOD-KNN model, there are other missing value imputation techniques that can perform better. Also, KNN operates by memorizing the entire dataset which can be a disadvantage.

REFERENCES

[1]. H. Nugroho, N.P Utama, and K. Surendro, "Normalization and outlier removal in class center-based firefly algorithm for missing value imputation," Open Access, J Big Data, (2021)8:129.

[2]. D. Chehal, P. Gupta, P. Gulati, and T. Gupta, "Comparative Study of Missing Value Imputation Techniques on E Commerce Product Ratings," Informatica 47 (2023) 373–382.

[3]. A.F. Sallaby, Azlan, "Analysis of Missing Value Imputation Application with K-Nearest Neighbor (K-NN) Algorithm in Dataset," (International Journal of Informatics and Computer Science) Vol 5 No 2, July 2021, Page 141-144.

[4]. P. Mishra, K.D. Mani, P. Johri, and D. Arya, " FCMI: Feature Correlation based Missing Data Imputation"

[5]. I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[6]. F. E. Harrell, Jr., "Regression Modeling Strategies," Nashville, TN, USA July 2015, ISSN 2197-568X

[7]. C. K. Enders, "Applied Missing Data Analysis," Second Edition, 2022 pp1-43,

[8]. M. Tannous, M. Miraglia, F. Inglese, L. Giorgini, F. Ricciardi, R. Pelliccia, M. Milazzo, and C. Stefanini, "Haptic-based Touch Detection for Collaborative Robots in Welding Applications", ROBOTICS COMPUT. INTEGR. MANUF., 2020. (IF: 3)

[9]. L.Y. Wang, D. Wang; Y.H. Chen, "Prediction Of Protein Subcellular Multisite Localization Using A New Feature Extraction Method", GENETICS AND MOLECULAR RESEARCH : GMR, 2016

[10]. F. Pirotti, R. Ravanelli, F. Fissore, and A. Masiero, "Implementation and Assessment of Two Density-based Outlier Detection Methods Over Large Spatial Point Clouds", OPEN GEOSPATIAL DATA, SOFTWARE AND STANDARDS, 2018. (IF: 3).

[11]. P. Keerin, W. Kurutach, and T. Boongoen, "Cluster-based KNN Missing Value Imputation for DNA Microarray Data", 2012 IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN, AND ..., 2012. (IF: 3)

[12]. K.M. Fouad, M.M. Ismail, A.T. Azar, and M.M Arafa, "Advanced Methods for Missing Values Imputation Based on Similarity Learning," PEERJ. COMPUTER SCIENCE, 2021. (IF: 3).

[13]. S. Patra, and B. Ganguly; "Improvising Singular Value Decomposition By KNN for Use in Movie Recommender Systems", JOURNAL OF OPERATIONS AND STRATEGIC PLANNING, 2019.

[14]. N. Rabiei, A.R. Soltanian, M. Farhadian, and F. Bahreini; "The Performance Evaluation of The Random Forest Algorithm for A Gene Selection in Identifying Genes Associated with Resectable Pancreatic Cancer in Microarray Dataset: A Retrospective Study", CELL JOURNAL, 2023.

[15]. F. Yang, J. Du, J. Lang, W. Lu, L. Liu, C. Jin, and Q. Kang; "Missing Value Estimation Methods Research for Arrhythmia Classification Using The Modified Kernel Difference-Weighted KNN Algorithms", BIOMED RESEARCH INTERNATIONAL, 2020. (IF: 3)

[16]. Z. Zhang, "Introduction To Machine Learning: K-nearest Neighbors", ANNALS OF TRANSLATIONAL MEDICINE, 2016. (IF: 7)

[17]. A. Hamed, A. Sobhy, and H. Nassar; "Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data Using A KNN Variant Algorithm", ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING, 2021. (IF: 3)

[18]. N. Rabiei, A.R. Soltanian, M. Farhadian, and F. Bahreini; "The Performance Evaluation of The Random Forest Algorithm for A Gene Selection in Identifying Genes Associated with Resectable Pancreatic Cancer in Microarray Dataset: A Retrospective Study"CELL JOURNAL, 2023,

[19].] M. Zaki, Shao-jie Chen, Jicheng Zhang, Fan Feng, Liu Qi, M.A. Mahdy, and Linlin Jin, "Optimized Weighted Ensemble Approach for Enhancing Gold Mineralization Prediction", APPLIED SCIENCES, 2023.

[20]. S. Sheikhi; M.T. Kheirabadi, and A. Bazzazi; "A Novel Scheme for Improving Accuracy of KNN Classification Algorithm Based on The New Weighting Technique and Stepwise Feature Selection", 2020.

[21]. M. Zhang, and W. Xu; "Study on An Improved Lie Group Machine Learning-based Classification Algorithm", 2020 IEEE 3RD INTERNATIONAL CONFERENCE OF SAFE PRODUCTION ..., 2020.

[22]. E.Y. Boateng, J. Otoo, and D.A. Abaye; "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review", 2020. (IF: 4)