International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Robust Missing Data Handling using Intelligent Machine Learning Imputation Technique for Heterogeneous Dataset

[1]Sowmya Venkatesh, [2]Maragal Venkatamuni Vijay Kumar, [3]Ashoka Davanageri Virupakshappa

Submitted: 27/12/2023 **Revised**: 03/02/2024 **Accepted**: 11/02/2024

Abstract: In data analysis, the presence of missing values is a common challenge, especially in heterogeneous datasets that encompass a wide range of data types, including numerical, categorical, and unstructured data. Addressing missing data is crucial as it directly impacts the quality and reliability of subsequent analyses and modeling. This necessitates the development of robust imputation methods capable of handling diverse data types effectively. In light of the aforementioned requirement, this study presents a novel and pioneering methodology for forecasting and completing the imputed data variables throughout the dataset that contains multiple variables. The approach under consideration integrates Natural-Language-Processing (NLP) encoders, feature-extractors motivated by machine-learning, and sequential-regression imputation methods. To ascertain the practicality of the suggested approach, this study meticulously evaluates the way it performs using a well-established medical dataset of heart-disease sourced from the repository of UCI. The findings presented in this paper provide compelling evidence of the method's superiority over existing missing data imputation techniques, notably in terms of accuracy. This demonstration of practical viability and effectiveness addresses a significant concern in the field of data preprocessing and analysis, reaffirming the importance of robust imputation methods for enhancing the quality of data-driven decision-making processes.

Keywords: *Heterogenous Datasets, Missing Data, Natural Language Processing, Imputation.*

1. Introduction

Statistically speaking, imputation provides a way to fill in gaps in sets of data where actual values are unavailable. It's a crucial step in data preprocessing, as missing data can lead to biased analyses and reduced model performance in various fields. Imputation aims to improve the quality and completeness of a dataset, making it more suitable for analysis or Machine-Learning (ML) tasks [1].

Imputation is a widely employed statistical technique across diverse fields. In healthcare, it's crucial for medical research and clinical trials, filling gaps in patient data to ensure analyses and predictions are robust [2]. In finance, it aids in constructing complete time series data, essential for tracking stock prices and economic indicators [3]. Social sciences benefit from imputation when dealing with incomplete survey responses, enhancing the effectiveness of data analysis [4]. Retail uses imputation to estimate missing sales data, customer behavior, and inventory levels [5]. Similarly, in environmental science, imputation plays a vital role in estimating missing climate or pollution data, facilitating the creation of comprehensive datasets.

From the above, it is seen that imputation methods play a critical role in addressing missing data in various domains. These methods encompass several categories such as Mean/Median Imputation [6], which substitutes missing variables using the median or mean, assuming a normal distribution; Mode Imputation [6], used for categorical data by substituting missing values with the most frequent category; Regression Imputation [7], employing regression models to predict missing values, with linear regression [8] for numeric and logistic regression [9] for categorical data; K-Nearest Neighbors (KNN) Imputation [10], which estimates missing values based on their nearest neighbors in the dataset; and Multiple Imputation [11], involving the creation of multiple imputed datasets and separate analysis to account for the uncertainty in imputation, offering valuable tools for enhancing data completeness and analysis accuracy.

Handling missing values in heterogeneous datasets presents several challenges. Such datasets often combine diverse data types, including numerical, categorical, and text data, making it difficult for traditional imputation methods, which are often tailored to specific data types, to handle the diversity effectively [12]. Effective feature engineering becomes crucial to capture complex relationships between variables, but in heterogeneous datasets, identifying and engineering relevant features can be intricate due to less straightforward relationships. Moreover, high-dimensional datasets can suffer from overfitting and computational demands, and interactions between different data types may be overlooked by existing methods that treat data types

[1]*Research Scholar Dept. of Computer Science and Engineering Dr. Ambedkar Institute of Technology Bengaluru, Karnataka, India Affiliated to Visvesvaraya Technological University, Belagavi-590018 vsowmyaresearch@gmail.com*

[2]*Research Supervisor Dept. of Information Science and Engineering Dr. Ambedkar Institute of Technology Bengaluru, Karnataka, India Affiliated to Visvesvaraya Technological University, Belagavi-590018 dr.vijay.research@gmail.com*

[3]*Co-Supervisor Dept. of Information Science and Engineering JSS Academy of Technical Education Bengaluru, Karnataka Affiliated to Visvesvaraya Technological University, Belagavi-590018 dr.dvashoka@gmail.com*

independently. Moreover, scalability issues may also arise when dealing with large datasets, and assumptions made by imputation methods may not hold true, potentially introducing bias into imputed values [13]. The selection of the appropriate imputation model is challenging, requiring careful consideration for each missing value. Handling text data for imputation, often necessitating specialized techniques like Natural Language Processing (NLP), can be complex, and evaluating imputation performance in heterogeneous datasets poses difficulties with traditional metrics often falling short. Lastly, generalizing imputation models across different heterogeneous datasets proves challenging due to variations in data structure and content [14].

When it comes to handling data that is missing across heterogeneous-datasets, which comprises a combination of unstructured and structured data, an integrated strategy which incorporates ML feature-extraction, sequential-regression imputation and NLP encoders appears to be quite helpful [15]. NLP encoders excel at processing unstructured text data, uncovering valuable insights related to missing values within documents or user-generated content. ML-driven feature extraction captures intricate relationships between variables, including those with missing values, generating pertinent features for imputation models. Sequential regression imputation models are then applied to the combined dataset, incorporating structured and unstructured data as well as engineered features, enhancing their ability to predict missing values accurately. This holistic approach leverages the strengths of NLP for text data, feature engineering for structured data, and sequential regression imputation, ultimately enhancing data analysis, ML, and decision-making across diverse domains. Hence, the contribution of this work are as follows

- The present study presents a novel methodology for the prediction and imputation of missing-data variables within multifaceted datasets, which consist of a wide range of data types including unstructured, categorical and numerical data. The proposed methodology effectively integrates NLP encoders, feature-extractors motivated by ML, and sequential-regression imputation methods. This combined strategy presents an extensive and flexible method for effectively managing missing information across different kinds of data.
- The suggested approach is thoroughly validated using a benchmarked medical dataset of heart-disease retrieved through the repository from UCI. This validation ensures that the method is empirically complete and can be confidently applied to real-world scenarios, particularly in healthcare settings.
- The results presented in the paper highlight a crucial contribution—the proposed method outperforms existing missing data imputation techniques in terms of accuracy while also significantly reducing the computational time required for imputation. This demonstrates the practical viability and effectiveness of the approach, addressing a pressing concern in data preprocessing and analysis.

2. Literature Survey

In [16], they employed five distinct methods, namely the mode, median, mean, linear regression and Decision-Tree (DT) based regression methods, for estimating the absence of clinical features values from real complicated Hepatocellular-Carcinoma (HCC) information. The evaluation of the findings was conducted using two distinct ML techniques, namely DT and Naïve-Bayes (NB) classifier. These ML techniques were employed to forecast outcomes related to survival. The performance standards selected for evaluation in this study were sensitivity, accuracy, f-score, specificity and precision. The most optimal performance in predicting survival rates for HCC was achieved through the utilization of DT-based classification and regression techniques. In [17], a comprehensive analysis was carried out to investigate the influence of imputation techniques upon the concept of equality within the domain of graph-data, specifically focusing on node characteristics. Various neural-network and embedding methodologies were employed to assess the aforementioned effect. The present study undertakes a comprehensive investigation on six distinct datasets, aiming to shed light on the intricate matter of equality in graph-node categorization. The research specifically focuses on the challenges posed by missing information and the diverse array of imputation methods employed to address this problem. Through a series of thorough tests, the study meticulously examined and highlighted the different problems that arise in the context of equality within this domain. The researchers have discovered that the selection of the imputation technique has a significant impact in terms of accuracy and equality. The findings of the research they conducted offer significant contributions towards the field of equality in ML applied to graph-data, shedding light on effective strategies for addressing absence of information in graph-data with efficiency.

In [18], an adequate imputation method for handheld medical information was identified. The researchers utilized a Portable-Health-Clinic (PHC) dataset, which had been meticulously gathered throughout a span of twelve years through multiple places in Bangladesh. Upon analysis, it was observed that around twenty percent of the information turned out to be absent or lacking. A comparison study was conducted to assess the efficacy of eight missing value managing techniques. These techniques were applied to 5 state-of-the-art ML techniques, utilizing the PHC Dataset. The evaluation of imputation efficiency for every instance was conducted by considering two key metrics: F-

measure and accuracy. The Multiple-Imputation through Chained-Equations (MICE) imputation method demonstrated superior performance in terms of f-measure and accuracy across every scenario. The present research provides evidence supporting the superiority of the MICE method in effectively handling missing values across various ML processes and computations. In [19], they put forward a novel approach termed Multi-Scaled Deep-Networks. This strategy involved the utilization of Variational-Auto-Encoders (VAEs) to address missing information through alignment, in addition to employing non-linear-regression techniques with activated-neurons and kernel matrix-structures employing the Scaled-Exponential Linear-Unit (SELU) technique. This approach specifically targeted the second-stage of VAEs, that encompasses decoding. The suggested approach, which was known as Multi-kernel Scaled-Deep Time-Series imputation (MSDTSI), demonstrated higher accuracy in comparison with traditional and previous ML techniques. Its primary objective was to impute the missing information within the Physionet-Challenge dataset, thereby enabling accurate prediction of mortality-rates for patients. The approach suggested in this study exhibited outstanding results compared to other approaches, as evidenced by its Area-Under Receiver-Operating-Characteristic (AUROC) result of 74.8 and Mean-Absolute-Error (MAE) value of 0.44.

In [20], the primary emphasis was found upon the bankruptcy-related dataset of Polish organizations. The collection of data consisted of statistical characteristics that were meticulously selected, with the inclusion of synthetic characteristics. The initial step involved conducting preliminary-processing and preliminary-analysis, which encompassed the imputation of values that are absent. This was accomplished through the utilization of well-established imputation methods, including K-Nearest-Neighbor (KNN) Imputation, Expectation Maximization Imputation and MICE. In order to mitigate the challenge of data-imbalance, the researchers employed the Synthetic-Minority-Oversampling technique (SMOTE) to oversample the minority-class labels. The informational modeling process involved the utilization of k-fold cross-validation upon the previously mentioned imputation techniques, in addition to resampled and imputed-datasets. After conducting a comprehensive analysis, a total of 36 distinct evaluations were obtained for 9 distinct approaches across four imputed-datasets. Subsequently, the efficiency of these techniques was meticulously assessed and evaluated using validation datasets, allowing for the subsequent ranking of each model based on their respective performance metrics. In [21], employed various ML algorithms, including Random-Forest (RF), Logistic-Regression (LR), XGBoost (XGB), AdaBoost (ADA), Support-Vector-Machine (SVM), KNN, NB, and DT. These algorithms were utilized for the purpose of forecasting heart-disease, utilizing a freely available diabetic dataset. In the conducted study, the researchers observed that the implementation of ADA algorithm, coupled with the utilization of median-imputation technique and recurrent feature removal, yielded the most favorable outcomes in terms of performance evaluation. The achieved results were quantified through various metrics, including f-measure, precision, recall and accuracy. Notably, the achieved accuracy-score was determined to be 84%, indicating the proportion of correctly classified instances. Furthermore, the precision-score, which measures the accuracy of positive predictions, was calculated to be 82%. The recall-score, which gauges the ability to identify positive instances, was found to be 88%. Lastly, the F1-score, which combines both recall and precision, was determined to be 85%. These findings highlight the effectiveness of the aforementioned approach in achieving excellent performance within the given context. Furthermore, the findings of their study indicate that each of the four imputation approaches, namely median and mean imputation, MICE and KNN imputation, demonstrate effective performance when employed on their dataset for the purpose of heart-disease prediction. In relation to the utilization of Recursive-Feature-Elimination (RFE), it was observed that NB demonstrates a discernible enhancement in efficiency, irrespective of the imputation method employed. Similarly, ADA exhibited enhanced efficiency when utilizing KNN and median imputation in conjunction with RFE. Conversely, the remaining algorithm configurations do not exhibit a significant rise in performance indicators (including f-measure, accuracy, recall and precision) when RFE was applied.

3. Methodology

The architecture of proposed work is given in Figure 1. The architecture begins with the dataset as its foundation. Initially, the dataset undergoes a series of crucial data processing steps, including data visualization for a comprehensive understanding of its characteristics and train-test splitting to facilitate model evaluation. Subsequently, the preprocessed data is directed to the data encoding phase, where various techniques such as normalization, one-hot encoding, and string sequencing are applied to transform the data into a suitable format for further analysis. Once encoded, the data proceeds to the feature extraction stage, where sophisticated methods like neural-networks, embedding, and Recurrent-Neural-Networks (RNNs) are employed to distill essential information and patterns from the dataset. Following feature extraction, the data enters the imputation phase, leveraging sequential-regression techniques for predicting and filling missing values, ensuring data completeness. Finally, after imputation, the architecture culminates in the realm of predictive analytics, where ML classifiers utilize the enriched dataset to make informed predictions and insights, thereby empowering data-driven decision-making processes.

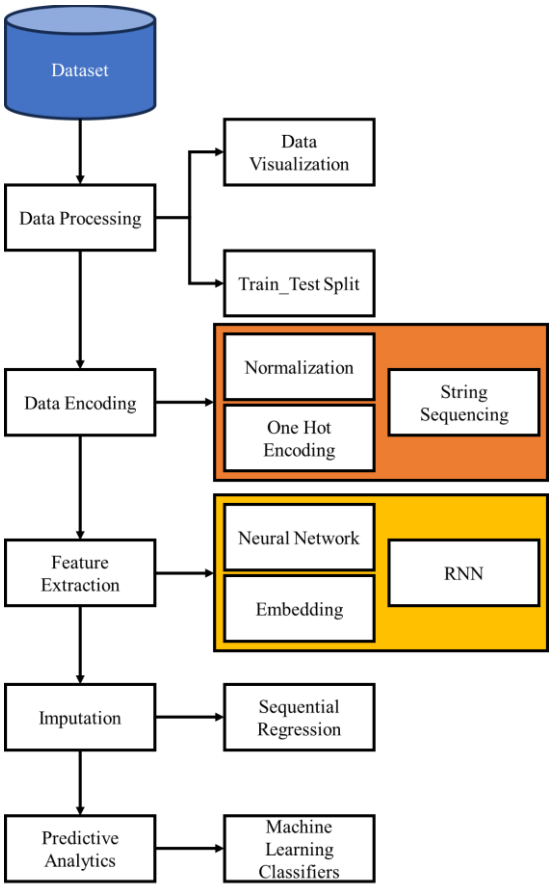

Fig 1. Architecture of Proposed Work

Further, in Figure 2, the complete process of data encoding is shown. In the data encoding process, we employ a systematic approach to transform the diverse data types within our heterogeneous dataset into a format conducive to further analysis. The process of normalizing numerical information involves the application of a technique that rescales the variables to a standardized range. This range was typically set from 0 to 1, or alternatively, using a mean of 0 along with a standard-deviation of 1. This step eliminates potential biases stemming from the differing scales of numerical features. For categorical data, the one-hot encoding technique is employed, wherein each categorical variable is transformed into a binary vector where each category is represented by a binary indicator variable. This allows the ML models to interpret categorical data effectively. Lastly, for text data, string sequencing is utilized. Here, textual information is tokenized and converted into numerical sequences, enabling the application of numerical-based algorithms to text data. These encoding methods collectively enable the dataset to be in a format that can be processed and analyzed by ML models, ensuring that valuable information from diverse data types is retained while achieving data uniformity.

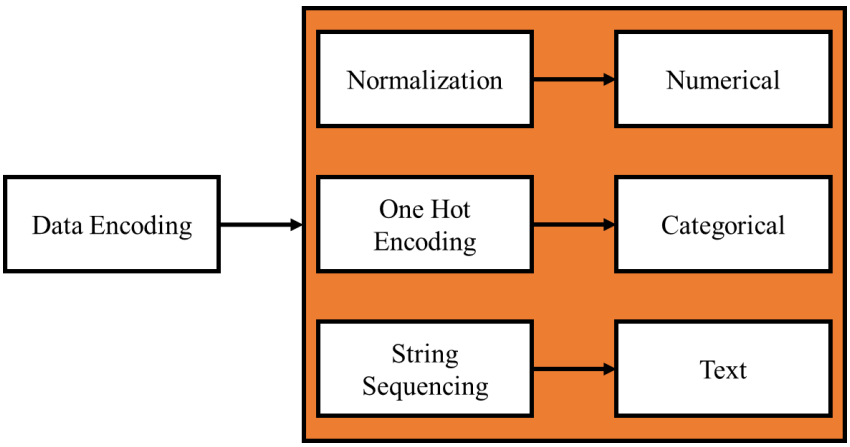

Fig 2. Data Encoding Process.

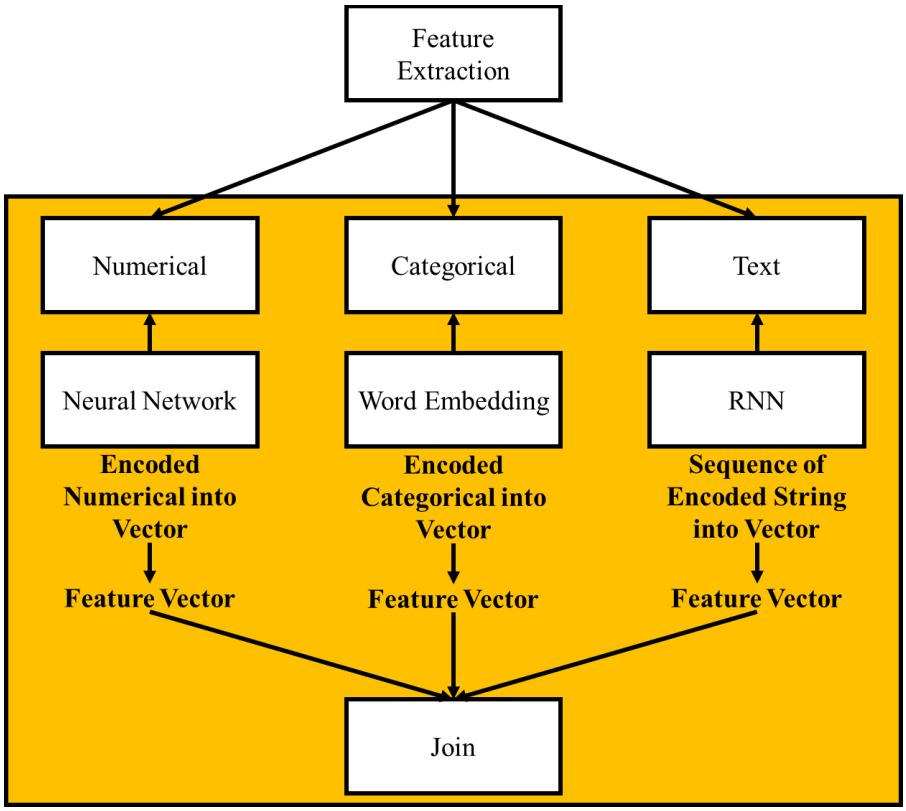

Fig 3. Feature Extraction Process.

Further, in Figure 3, the feature extraction process has been given. In the feature extraction phase, our methodology harnesses the power of various techniques to distill essential information from the diverse data types present in the heterogenous dataset. Numerical data are subjected to a neural network-based feature extraction process, where intricate patterns and relationships within the numerical variables are captured, resulting in a numerical feature vector. Categorical data, in contrast, undergo word embedding, which transforms categorical values into vector representations, preserving their semantic meaning. Textual data are processed using Recurrent Neural Networks (RNNs), allowing us to uncover context and variations within unstructured text, generating a sequence of encoded vectors. These distinct feature vectors, each representing a particular data type, are then harmoniously joined to create a comprehensive feature representation of the entire dataset. This amalgamation of feature vectors enables us to capture both the individual characteristics of each data type and their collective contributions to our subsequent analysis, facilitating richer insights and more accurate imputations in our heterogeneous dataset.

In the imputation phase of our data analysis pipeline, we employ a Sequential-Regression-Imputation approach to address missing-values within our dataset. This approach is particularly effective in capturing complex dependencies and relationships between variables, allowing us to predict and replace missing values sequentially. By iteratively updating missing values based on the information obtained from other variables, the Sequential Regression Imputation technique enhances the completeness and accuracy of our dataset. It proves especially valuable in scenarios where the missing data is not missing completely at random (MCAR) but exhibits patterns and relationships that can be leveraged for more informed imputations. This technique contributes significantly to the overall data quality, ensuring that downstream analyses and modeling are based on a more comprehensive and reliable dataset.

In the realm of predictive analytics, we leverage the power of two ML classifiers, namely SVM and NB, to extract meaningful insights and make informed predictions from our enriched dataset. SVM is a robust and versatile algorithm known for its ability to handle complex data by finding optimal hyperplanes that best separate classes in high-dimensional spaces. On the other hand, NB, based on probabilistic principles, excels in modeling the probability distribution of data and is particularly effective for text and categorical data. By employing these two diverse classifiers, we ensure a comprehensive approach to predictive analytics, accommodating a wide range of data types and structures. This strategy enables us to harness the strengths of both SVM and NB, improving the accuracy and robustness of our predictive models and ultimately empowering data-driven decision-making processes across various domains. The results of the prediction using the SVM and NB classifiers has been discussed in the next section.

4. Results and Disscussions

For evaluating the results and executing the proposed work, this work considered python programming language. The

system on which the work was executed consisted of 16 GB RAM, Windows Operating System and 1TB hard disk. This work was executed using the Annacoda which has built in Python Terminal. For evaluating this work, the Heart Disease UCI Dataset was used [22]. The confusion matrix achieved by the SVM classifier using original dataset and SVM classifier using imputed dataset is given in Figure 4 and Figure 5 respectively. Further, the confusion matrix achieved by the NB classifier using original dataset and NB classifier using imputed dataset is given in Figure 6 and Figure 7 respectively.

Comparing the performance of SVM and NB classifiers on both the original and imputed datasets reveals interesting insights. In the original dataset, SVM achieved a true positive (TP) count of 17, indicating its ability to correctly identify positive instances, while it had 9 false positives (FP), indicating instances incorrectly classified as positive. Additionally, it displayed 3 false negatives (FN) and 32 true negatives (TN). On the other hand, NB achieved a TP count of 18 and 8 FP, showcasing its aptitude for true positive identification. However, it had 6 FN and 29 TN. When considering the imputed dataset, SVM showed improvement in TP, now at 18, and reduced FP to 8, suggesting enhanced precision. However, there were 5 FN, potentially indicating some loss in recall. NB also showed improved precision, with TP at 15 and FP at 3, while FN increased to 11 and TN remained at 32.

It is observed that imputation generally led to better precision for both classifiers, with fewer false positives. However, SVM exhibited more consistent performance across the two datasets, maintaining a similar balance between precision and recall. In contrast, NB showed higher precision but lower recall in the imputed dataset, suggesting that imputation may have affected its ability to capture all positive instances. These findings emphasize the importance of considering imputation strategies carefully, as they can influence classifier performance and the trade-off between precision and recall, depending on the specific context and dataset characteristics

.

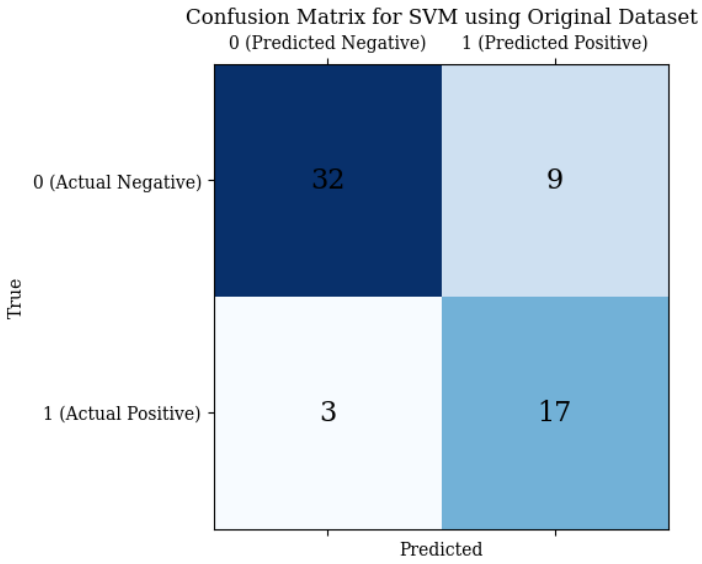

Fig 4. Confusion-Matrix for SVM using original dataset.

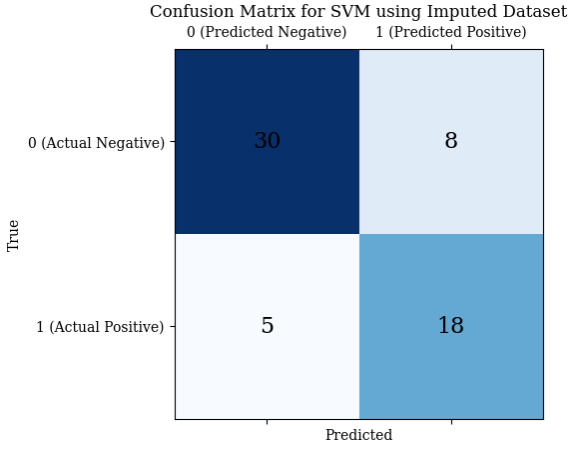

Fig 5. Confusion-Matrix for SVM using imputed dataset.

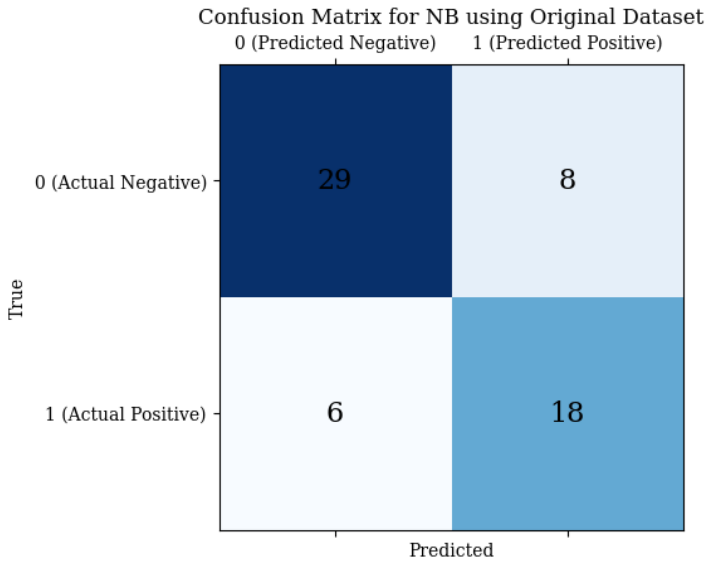

Fig 6. Confusion-Matrix for NB using original dataset.

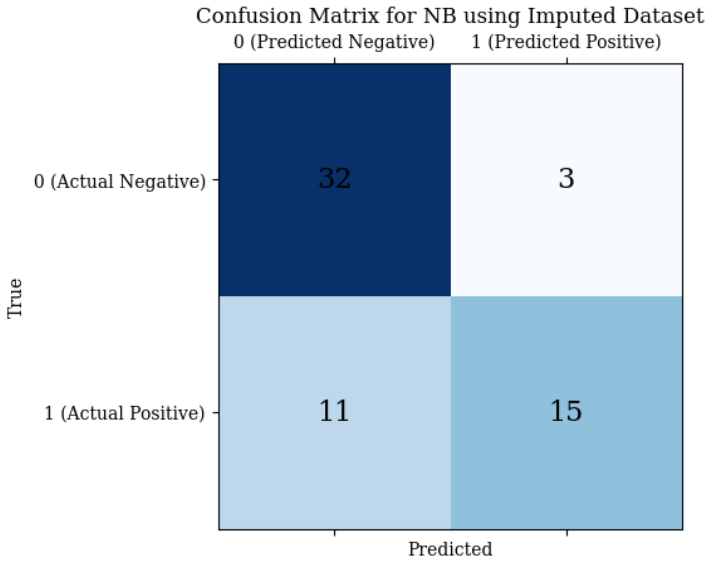

Fig 7. Confusion-Matrix for NB using imputed dataset.

In Figure 8 and Figure 9, the accuracy performance using the SVM and NB classifier is given. In the "SVM using Original dataset" scenario, during training, the SVM classifier achieved an accuracy of 90.57%, and during testing, it achieved an accuracy of 80.32% when applied to the original dataset. In the "SVM using Imputed dataset" scenario, the SVM classifier achieved an accuracy of 86.77% both during training and testing when applied to the imputed dataset. In the "NB using Original dataset" scenario, during training, the NB classifier achieved an accuracy of 89.98%, and during testing, it achieved an accuracy of 77.04% when applied to the original dataset. In the "NB using Imputed dataset" scenario, during training, the NB classifier achieved an accuracy of 83.58%, and during testing, it achieved an accuracy of 75.41% when applied to the imputed dataset.

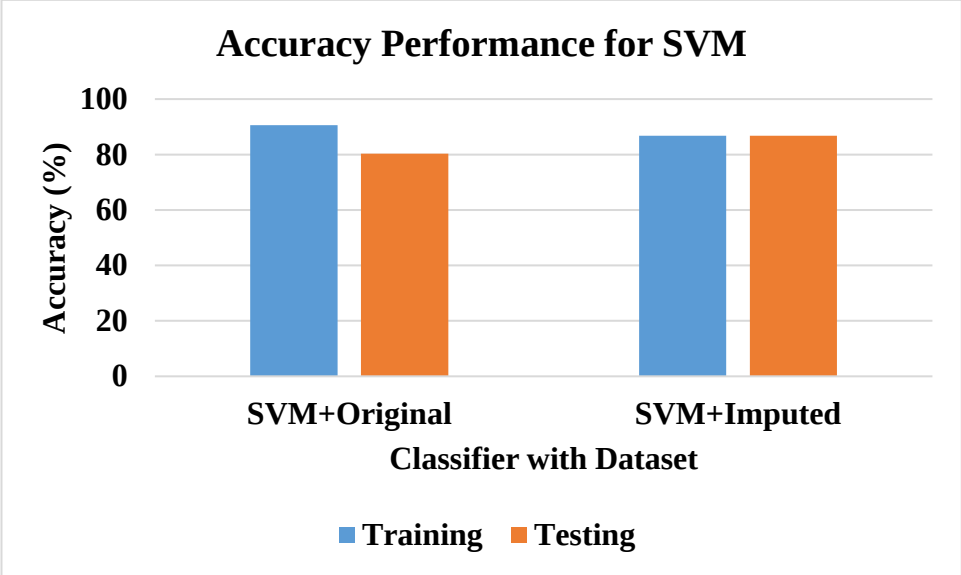

Fig 8. Accuracy Performance using SVM Classifier.

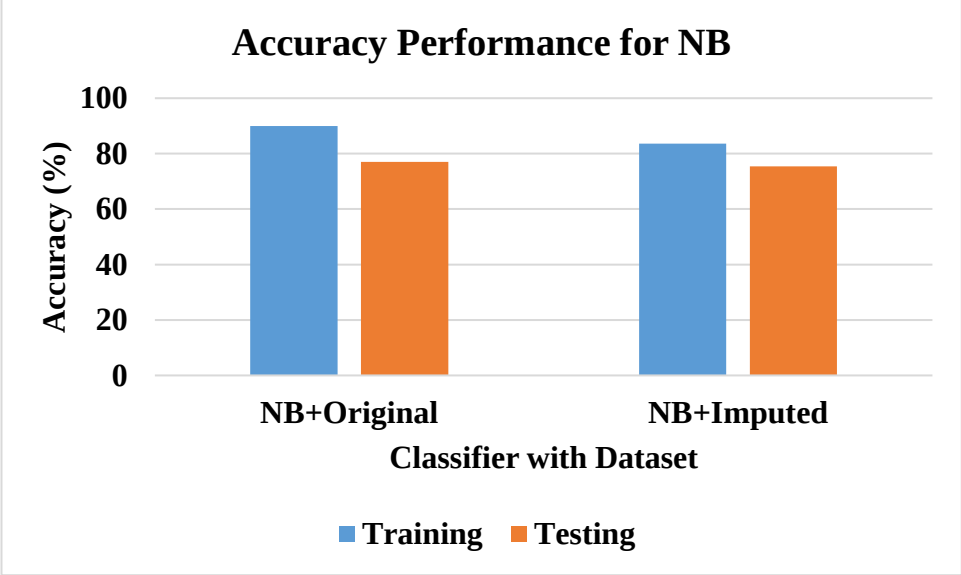

Fig 9. Accuracy Performance using NB Classifier

In the case of SVM, the classifier exhibited higher training accuracy with the original dataset (90.57%) compared to the imputed dataset (86.77%). However, when it came to testing, SVM's accuracy remained consistent (around 80%) for both datasets. This suggests that SVM's generalization performance was not significantly affected by the imputation process. For Naive Bayes, the training accuracy was slightly higher with the original dataset (89.98%) compared to the imputed dataset (83.58%). However, during testing, NB's accuracy dropped from 77.04% (original) to 75.41% (imputed). This indicates that NB's performance degraded slightly when applied to the imputed data, possibly due to the imputation process altering the data distribution in a way that affected NB's probabilistic assumptions.

In summary, these results suggest that SVM was more robust to the imputation process, maintaining similar testing accuracy on both the original and imputed datasets. On the other hand, Naive Bayes, while performing well on the original data, experienced a slight decrease in accuracy when applied to the imputed dataset, indicating a potential sensitivity to changes in data distribution caused by imputation. These findings underscore the importance of carefully considering the choice of classifier and the impact of data preprocessing, including imputation, when developing ML models.

5. Conclusion

In the landscape of data analysis, the challenge of missing data is ubiquitous, especially when dealing with heterogeneous datasets that encompass various data types, ranging from numerical to categorical and unstructured data. The critical need to address missing values arises from their potential to introduce biases and inaccuracies into analyses and modeling processes, ultimately undermining the reliability of data-driven decision-making. In response to this imperative, this study has introduced a novel and robust imputation method designed to tackle the intricate problem

of missing data fields within multivariate datasets. The proposed approach stands out as it boldly combines three distinct pillars: NLP-encoders for text data, ML-driven feature-extractors for structured data, and sequential-regression-imputation techniques. This amalgamation of methodologies offers a comprehensive and adaptable solution for handling missing values, ensuring that data completeness is achieved across diverse data types. It is this synergy that makes our approach particularly valuable, given the growing prevalence of mixed-data environments in contemporary data analytics. In addition, it is worth noting that this study has made a significant impact by thoroughly verifying the suggested approach using a well-established medical dataset of heart-disease retrieved from the repository from UCI. The findings presented in this paper affirm the significance of our work. The proposed method, by surpassing existing missing data imputation techniques in terms of accuracy, establishes itself as a practical, viable, and effective solution to a pressing concern in the realm of data preprocessing and analysis.

In conclusion, this study not only recognizes the intrinsic need for advanced imputation methods in heterogeneous datasets but also boldly presents an innovative approach that bridges the gap between diverse data types. With empirical validation and compelling results, our work paves the way for enhanced data quality, ultimately empowering data scientists and analysts to make more reliable and informed decisions based on comprehensive and imputed datasets.

References

[1] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data," *Soft Computing*, Feb. 2021, doi: 10.1007/s00500-021-05590-y.

[2] A. R. Ismail, N. Z. Abidin, and M. K. Maen, "Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare," *Journal of Robotics and Control (JRC)*, vol. 3, no. 2, pp. 143–152, Feb. 2022, doi: 10.18196/jrc.v3i2.13133.

[3] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation?," *Emerging Markets Finance and Trade*, pp. 1–11, Oct. 2020, doi: 10.1080/1540496x.2020.1825935.

[4] A. D. Woods et al., "Missing Data and Multiple Imputation Decision Tree," *PsyArXiv*, Aug. 2021, doi: 10.31234/osf.io/mdw5r.

[5] X. Miao, Y. Wu, L. Chen, Y. Gao, and J. Yin, "An Experimental Survey of Missing Data Imputation Algorithms," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–20, 2022, doi: 10.1109/tkde.2022.3186498.

[6] R. Pavithrakannan, N. B. Fenn, S. Raman, V. Kalyanaraman, V. K. Murugananthan and J. Janarthanan, "Imputation Analysis of Central Tendencies for Classification," *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Toronto, ON, Canada, 2021, pp. 1-7, doi: 10.1109/IEMTRONICS52119.2021.9422507.

[7] K. Slavakis, G. N. Shetty, L. Cannelli, G. Scutari, U. Nakarmi and L. Ying, "Kernel Regression Imputation in Manifolds Via Bi-Linear Modeling: The Dynamic-MRI Case," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 133-147, 2022, doi: 10.1109/TCI.2022.3148062.

[8] N. Karmitsa, S. Taheri, A. Bagirov and P. Mäkinen, "Missing Value Imputation via Clusterwise Linear Regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1889-1901, 1 April 2022, doi: 10.1109/TKDE.2020.3001694.

[9] M. Chen, H. Zhu, Y. Chen, and Y. Wang, "A Novel Missing Data Imputation Approach for Time Series Air Quality Data Based on Logistic Regression," *Atmosphere*, vol. 13, no. 7, pp. 1044–1044, Jun. 2022, doi: 10.3390/atmos13071044.

[10] D. M. P. Murti, U. Pujianto, A. P. Wibawa and M. I. Akbar, "K-Nearest Neighbor (K-NN) based Missing Data Imputation," *2019 5th International Conference on Science in Information Technology (ICSITech)*, Yogyakarta, Indonesia, 2019, pp. 83-88, doi: 10.1109/ICSITech46713.2019.8987530.

[11] B. N. Vi, D. Tan Nguyen, C. T. Tran, H. Phuc Ngo, C. C. Nguyen and H. -H. Phan, "Multiple Imputation by Generative Adversarial Networks for Classification with Incomplete Data," *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, Hanoi, Vietnam, 2021, pp. 1-6, doi: 10.1109/RIVF51545.2021.9642138.

[12] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang, "Deep learning versus conventional methods for missing data imputation: A review and comparative study," *Expert Systems with Applications*, vol. 227, p. 120201, Oct. 2023, doi: 10.1016/j.eswa.2023.120201.

[13] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities," *Neural Computing and Applications*, vol. 33, no. 22, pp.

15091–15118, Aug. 2021, doi: 10.1007/s00521-021-06406-8.

[14] M. I. Gabr, Y. M. Helmy, and D. S. Elzanfaly, "Effect of Missing Data Types and Imputation Methods on Supervised Classifiers: An Evaluation Study," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 55, Mar. 2023, doi: 10.3390/bdcc7010055.

[15] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, "Machine Learning and Integrative Analysis of Biomedical Big Data," *Genes*, vol. 10, no. 2, Jan. 2019, doi: 10.3390/genes10020087.

[16] M. Yumuş, M. Apaydın, A. Değirmenci and Ö. Karal, "Missing data imputation using machine learning based methods to improve HCC survival prediction," *2020 28th Signal Processing and Communications Applications Conference (SIU)*, Gaziantep, Turkey, 2020, pp. 1-4, doi: 10.1109/SIU49456.2020.9302222.

[17] S. A. Ansari, C. Sharma and T. Agarwal, "Mean and Prediction Imputation-Based Approach for Predicting Water Potability Using Machine Learning," *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2022, pp. 1-6, doi: 10.1109/ICRITO56286.2022.9964809.

[18] S. Tabassum, N. Abedin, R. I. Maruf, M. Taufiq Ahmed and A. Ahmed, "Improving Health Status Prediction by Applying Appropriate Missing Value Imputation Technique," *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, Osaka, Japan, 2022, pp. 345-348, doi: 10.1109/LifeTech53646.2022.9754794.

[19] A. Deshmukh, J. Choudhary and D. P. Singh, "Multi Kernel Scaled Deep Time Series Imputation," *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2022, pp. 829-834, doi: 10.1109/ICACCS54159.2022.9784998.

[20] A. Hassan and N. Yousaf, "Bankruptcy Prediction using Diverse Machine Learning Algorithms," *2022 International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, 2022, pp. 106-111, doi: 10.1109/FIT57066.2022.00029.

[21] V. Peter and Ma. Sheila, "Cardiovascular disease prediction with imputation techniques and recursive feature elimination," *Nucleation and Atmospheric Aerosols*, Jan. 2023, doi: 10.1063/5.0124079.

[22] "UCI Machine Learning Repository," archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/45/heart+disease