Original Article

IMPUTATION BASED DATA PRE-PROCESSING IN MACHINE LEARNING FOR HEART DISEASE DATASET

R. Karthikeyan[1], B. Selvanandhini[2]

[1] Research Scholar, Department of Computer Science, Pollachi College of Arts and Science, Tamil Nadu India.
[2] AP, Research Supervisor, Department of Computer Science, Pollachi College of Arts and Science, Tamil Nadu India.

DOI: 10.47750/pnr.2022.13.S08.397

Abstract

New computational and mechanical standards that as of now guide advancements in the data society, i.e., the Internet of things, unavoidable technology favor the presence of new interruption vectors that can straightforwardly influence individuals' day-to-day routines. Heart disease diagnosis is a troublesome undertaking for example it ought to be performed exactly and proficiently. The exploration paper chiefly centers around which patient are more prone to have a coronary illness in view of different clinical characteristics of attributes. In this paper we proposed Imputation based preprocessing algorithm for preprocess the heart disease datasets. The consequences of the exploration show that the utilization of imputation based preprocessing algorithm had a job in working on the prescient exactness of inadequately effective classifiers, and shows agreeable execution in deciding the risk of coronary illness.

Keywords: Heart disease, imputation based preprocessing algorithm, heart disease datasets, classifiers.

Introduction

Heart disease (HD) incorporates a wide range of diseases that influence different parts of the heart. Diseases under the heart disease umbrella incorporate heart musicality issues (arrhythmias), as well as vein diseases like coronary conduit disease, congestive heart failure (CHF), and ischemic heart disease (IHD). Heart disease is quite possibly of the most common disease in the public arena and is considered among the main sources of death. As per the World Health Organization, an assessment of 17.7 million individuals kicked the bucket because of HD in 2016, addressing 31% of every worldwide passing; 7.4 million and 6.7 million of these passings were brought about by coronary heart disease and stroke separately. Thus, heart disease conclusion is an extremely significant clinical assignment that ought to be precisely and productively performed. Precise and early identification of cardiovascular diseases can save many lives by monitoring heart exercises.

Data preprocessing for machine learning techniques in heart disease

To examine and summarize the use of information preprocessing in medical informatics, Idri et al. did an efficient planning review, meaning to get and classify articles managing the use of information preprocessing in medical datasets. The goal of this SMS was to: (1) distinguish the amount of exploration, and (2) structure the kind of examination managing the use of information preprocessing in medical DM. They recognized 126 essential examinations distributed between January 2000 and December 2017, and they figured out that the analysts'

consciousness of utilizing information preprocessing in the medical spaces has expanded throughout the course of recent years.

Literature Survey

Dinesh, K. G., Arumugaraj, K., (2018) et.al proposed Prediction of Cardiovascular Disease Using Machine Learning Algorithms. Healthcare is an unavoidable errand to be finished in human life. Cardiovascular disease is a general class for a scope of diseases that are influencing heart and veins. The early methods of forecasting cardiovascular diseases helped in settling on conclusions about the progressions to have happened in high-risk patients which brought about the decrease of their dangers. The healthcare business contains heaps of medical information, in this way machine learning algorithms are required to go with choices really in the expectation of heart diseases. Late exploration has dug into joining these procedures to give mixture machine-learning algorithms. This nonethical concentrate on plans to use accessible machine learning procedures in R programming. Future work incorporates different troupe methods of these algorithms which can progress to better performance with more parameter settings for these algorithms.

Nair, P., & Kashyap, I. (2019) proposed Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for KNN Classifier. Information mining is a strategy of looking at tremendous quanta of previous information to discover new examples and connections among them, which will assist with pursuing better choices. Classification is an information mining method that puts together information into classifications. In this paper, to improve the performance of the k closest neighbor (kNN) classifier-a sort of classification strategy that is among the most generally used another information pre-handling method has been proposed, which can deal with some classification issues like imbalanced information and outliers. In an imbalanced dataset, the classification classifications are not similarly disseminated. Imbalanced datasets have an innate issue with regards to utilizing classifiers on them that have been created utilizing machine learning algorithms. The essential idea of these algorithms is to diminish blunders without depending on the balance of classes. One more issue tended to in this paper is the question of outliers or outrageous qualities. Exception or outrageous qualities are those values that are outside the normal scope of values. The nature of Classification modeling can be incredibly upgraded by recognizing and extraction these qualities. In this proposed strategy two information pre-handling procedures have been joined to shape a half breed pre-handling method. The two information pre-handling methods are resampled strategy and the interquartile range procedure (IQR).

Iliou, T., Anagnostopoulos, (**2015)** et.al proposed A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance. Information preprocessing depicts any kind of handling technique performed on crude information to set it up for another handling procedure. Normally used as a primer information mining practice, information preprocessing methods transform the information into a format that will be all the more effective and really handled for the classification algorithms. In this paper, a clever information preprocessing technique is proposed and assessed in three troublesome classification informational collections of the notable UCI Vault, in which different classifiers have a typical performance lower than 75%. The performance of their proposed information preprocessing strategy and Head Part Investigation preprocessing technique was assessed utilizing the 10-overlay cross-approval technique surveying five classification algorithms, Nearest-neighbor classifier (IB1), C4.5 algorithm implementation (J48), Arbitrary Woods, Multilayer Perceptron and Turn Woodland, individually. The classification results are introduced and analyzed logically.

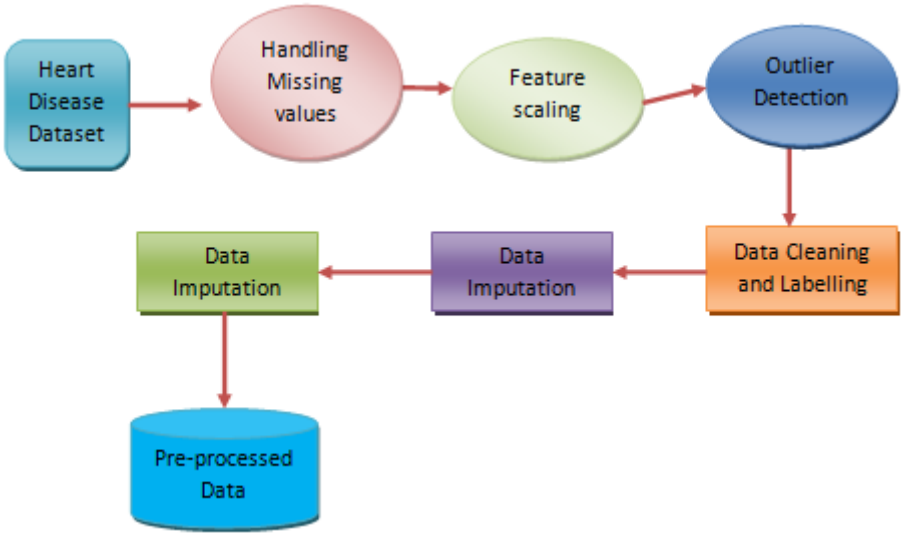
Proposed Methodology

Dataset Description

They used the UCI database of cardiology. It contains four datasets that have been recently used by ML specialists. The "target" property demonstrates the appearance or nonexistence of heart disease in the patient. This dataset

contains 76 features. These features are smoking, weight, active work, a healthy eating regimen, cholesterol levels, circulatory strain, fasting blood glucose, and so on. These properties are the very seven ideal estimates that the American Heart Affiliation has set to promote cardiovascular health and disease decrease.

Figure 1. Data preprocessing steps



The principal stage cleans the information by erasing copies, crediting missing information, and normalization called preprocessing. Second, we use algorithm 2 to perform information preprocessing, for example, information cleaning, information ascription, and element normalization. A few notable equations are used to perform highlight normalization preprocessing. These equations are considered as follows:

$$\boldsymbol{\mu_j = \frac{1}{n}\sum_{x=1}^{n} xj, xj \in D}$$

$$\boldsymbol{\sigma_j = \sqrt{\frac{1}{n}\sum_{x=1}^{n}(x_j - \mu_i)^2}, \quad x_j \in D}$$

$$\boldsymbol{x_i = \frac{x_i - \mu_i}{\sigma_j}}$$

Where μ = mean, σ = standard deviation, D = dataset, n = total number of values, x = single feature value. These data features are normalized using one unit mean and zero variance.

Pre-processing is the process of converting crude data into a deliberate and significant format. The real data for the most part comprises conflicting, superfluous, surplus data containing countless invalid qualities. It is vital to pre-process the dataset prior to preparing it on a classifier to further develop its prediction capacity. Prior to applying the classification model, datasets are first pre-processed and afterward exposed to dimensionality decrease strategies. Figure 1 show the preprocessing steps used in this method.

Medical data is generally incomplete with missing data, noisy with errors or outliers, and inconsistent containing discrepancies in names of features. The missing values can be handled using some imputation techniques. It places the features in a similar range or a similar scale so no factor is overwhelmed by the other. On the off chance that it isn't applied, then the learning algorithm will in general weigh more prominent qualities as higher and think about more modest advantages as lower, no matter what the unit of the benefits. There are in a general sense various sorts of component scaling. In any case, the most generally used strategies are normalization and normalization. In normalization, we register the transformed qualities by calculating the distinction of each component esteem from the mean of every one of the expenses of that element and afterward isolating by the

standard deviation for that component. This transforms the data between the range of - 1 and +1. The processed data has a mean of 0 and a standard deviation of 1. In normalization, we register the transformed qualities by calculating the distinction of each component esteem from the base of the relative multitude of expenses of that element and afterward splitting by the contrast between the base and most extreme incentive for that component. This transforms the data between the range of 0 and 1. Normalization is by and large not a decent choice, basically when the data contains a ton of noise and outliers. Specifically, when there are outliers, normalization will transform the typical data, i.e., the data without outliers into a negligible range of values, which isn't alluring for machine learning models. Thus, normalization is used in this review for the scaling of features.

Proposed model

A machine learning algorithm finds natural patterns in data that generate insight and assist with pursuing better choices and predictions. They are used to going with basic choices in medical conclusion, stock exchanging, and energy load forecasting, and that's only the tip of the iceberg. The data pre-processing phase handles missing invalid qualities and anomaly data. Then dimensionality decrease has been applied to the data. Normalizing the information values for each property estimated in the training occasions will help to upgrade the proficiency of the learning stage. Particularly for the distance-based methods normalization upholds ascribes with huge ranges from out weighting credits with at first more modest ranges. The normalization score is thought of and it is called as z-score. The missing qualities have been taken out then the z-score esteem is calculated. The score computation technique thinks about significance standard deviation as an enhanced strategy. Thus it is more hearty to find outliers than the standard deviation from the mean which was not a squared worth.

To alleviate the asset utilization of the preprocessing task, we characterized an architecture following the circulated registering model. The process was finished by partitioning the capabilities into a grid of relations which is fit for interpreting the conditions between the capabilities, thinking about certain factors, i.e., coupling, union, recursion, idempotency, and so on. From that point onward, the code was parted into various pieces of code; these pieces of code were isolated into various daemons that are executed from a distance by various machines. These pieces of code (units) will be units that were isolated from a successive disposition to an equal disposition. While doing this, the code should be adjusted to accurately aggregate the preprocessing capability equally, while protecting basic composition/perusing operations to give dependable operations of software.

Algorithm: Imputation based pre-processing algorithm

Input*: Heart Disease Dataset (DS)*

Output*: Pre-Processed Data (PDS)*

Step 1: Start the procedure

Step 2: Import full dataset file

Step 3: Drop incomplete rows

Step 4: For $i = 1$*: length (ds)*

Step 5: $PDS(i) \leftarrow removeduplication$

Step 6: If $DS(i)$ *==missing value*

Step 7: $PDS(i) \leftarrow meanfield\ of\ that\ field$

Step 8: Start Data normalization

Step 9: For j=1: length (DS)

Step 10: $\mu_j = \frac{1}{n}\sum_{x=1}^{n} x_j, x_j \epsilon DS$

Step 11: $\sigma_j = \sqrt{\frac{1}{n}\sum_{x=1}^{n}(x_j - \mu_i)^2}, x_j \epsilon DS$

Step 12: End the process

Step 13: Return PDS

Step 14: End procedure

The proposed algorithm presents how the dataset is preprocessed. This undertaking is finished by columns because it is required to think about the various qualities in each component (each element corresponds with a column). This permits constructing a file cluster of the various components, which is used to compute the typical dispersion of the qualities that shape the element or just to encode them in the ideal manner as required for each situation. The entire process is finished in a light-footed manner, separating the various capabilities accurately and protecting basic perusing/composing operations to give solid activity of the software that will be executed in lined up in the previously mentioned machines. One of the accessible strings (sent off constantly) deals with the execution. This first string begins by opening the full dataset document and the various columns comparing to each component distinguished. Then, this first string begins assigning features for preprocessing to every one of the accessible strings in equal. When the errands finish, each string returns the outcome to the main string, which relegates more features in lined up until there are something else to process.

Experiment Result

Precision: Precision is the proportion between the quantity of right positives and the quantity of true positives in addition to the quantity of false positives.

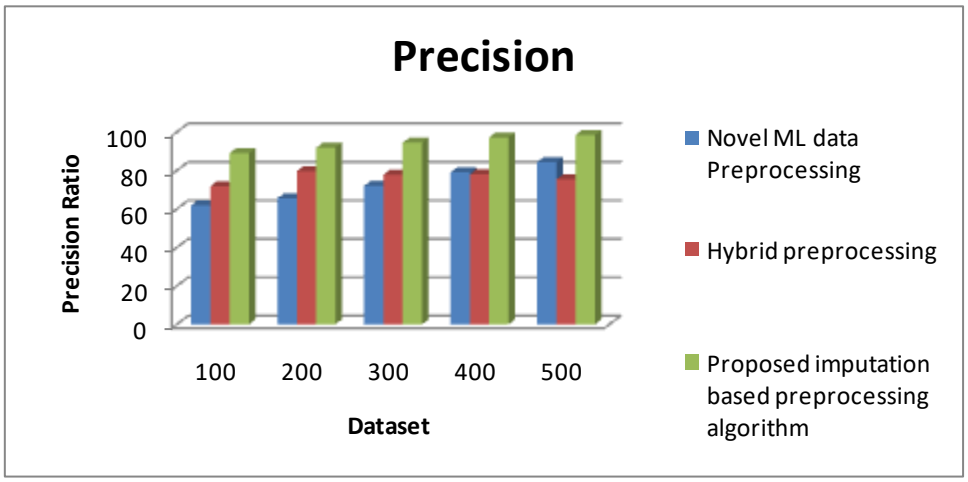
$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

Table 1. Comparison table of Precision

Dataset	Novel ML data Preprocessing	Hybrid preprocessing	Proposed imputation based preprocessing algorithm
100	61.94	71.91	89.01
200	65.66	79.77	91.87
300	72.12	77.93	94.48
400	79.09	78.05	97.23
500	84.38	75.39	98.52

The Comparison table 1 of Precision Values explains the different values of existing Novel ML data Preprocessing, Hybrid preprocessing and proposed imputation based preprocessing algorithm. While comparing the Existing algorithm and proposed imputation based preprocessing algorithm provides the better results. The existing algorithm values start from 61.94 to 84.38, 71.91.39 to 79.77 and proposed imputation based preprocessing algorithm values starts from 89.01 to 98.52. The proposed method provides the great results.

Figure 2 Comparison chart of Precision



The Figure 2 Shows the comparison chart of Precision demonstrates the existing Novel ML data Preprocessing, Hybrid preprocessing and proposed imputation based preprocessing algorithm. X axis denote the Dataset and y axis denotes the Precision ratio. The proposed imputation based preprocessing algorithm values are better than the existing algorithm. The existing algorithm values start from 61.94 to 84.38, 71.91.39 to 79.77 and proposed imputation based preprocessing algorithm values starts from 89.01 to 98.52. The proposed method provides the great results.

Recall: Recall is the proportion between the quantity of right positives and the quantity of true positives in addition to the quantity of false negatives.

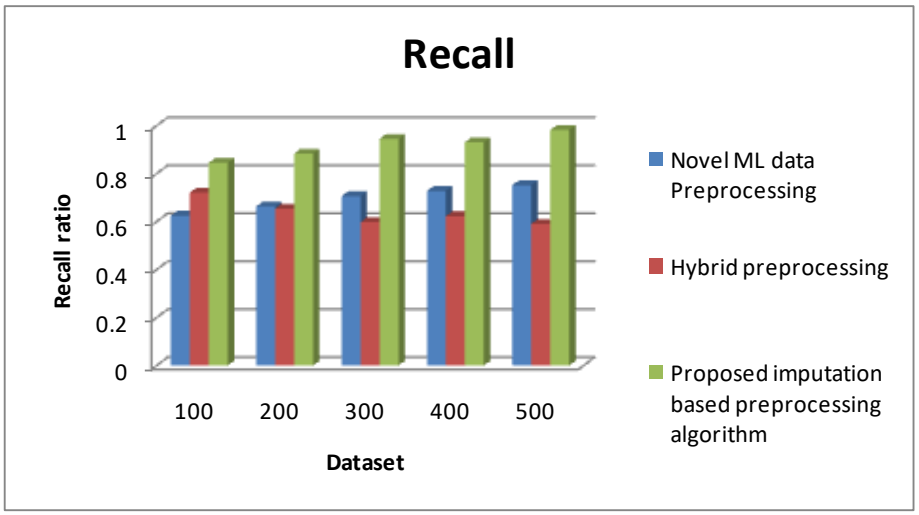
$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

Table 2. Comparison table of Recall

Dataset	Novel ML data Preprocessing	Hybrid preprocessing	Proposed imputation based preprocessing algorithm
100	0.625	0.721	0.846
200	0.663	0.654	0.884
300	0.706	0.598	0.945
400	0.728	0.623	0.931
500	0.752	0.591	0.982

The Comparison table 2 of Recall Values explains the different values of existing Novel ML data Preprocessing, Hybrid preprocessing and proposed imputation based preprocessing algorithm. While comparing the Existing algorithm and proposed imputation based preprocessing algorithm provides the better results. The existing algorithm values start from 0.625 to 0.752, 0.591 to 0.721 and proposed imputation based preprocessing algorithm values starts from 0.846 to 0.982. The proposed method provides the great results.

Figure 3 Comparison chart of Recall



The Figure 3 Shows the comparison chart of Recall demonstrates the existing Novel ML data Preprocessing, Hybrid preprocessing and proposed imputation based preprocessing algorithm. X axis denote the Dataset and y axis denotes the Recall ratio. The imputation based preprocessing algorithm values are better than the existing algorithm. The existing algorithm values start from 0.625 to 0.752, 0.591 to 0.721 and proposed imputation based preprocessing algorithm values starts from 0.846 to 0.982. The proposed method provides the great results.

Accuracy: Accuracy is the proportion between the quantity of right predictions and complete number of predications.

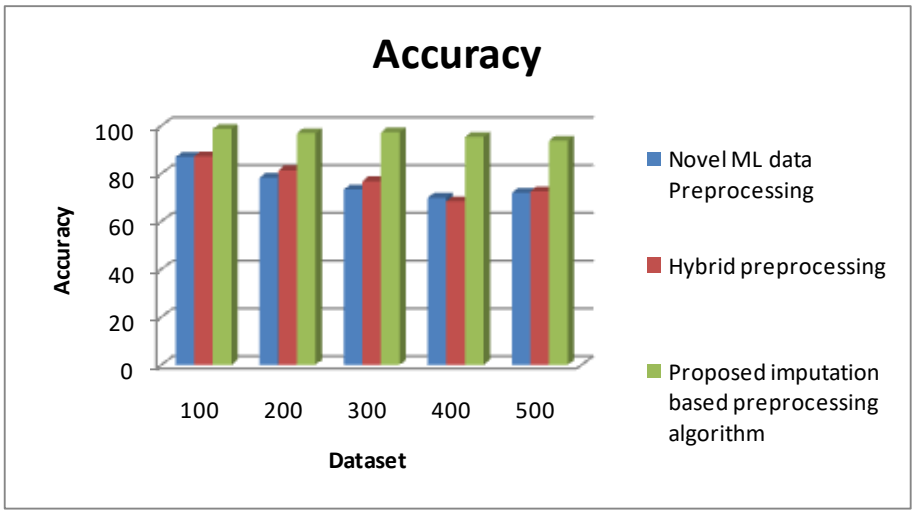
$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}}$$

Table 3. Comparison table of Accuracy

Dataset	Novel ML data Preprocessing	Hybrid preprocessing	Proposed imputation based preprocessing algorithm
100	86.85	87.12	98.63
200	78.25	81.29	96.81
300	73.21	76.73	97.25
400	69.82	68.35	95.31
500	71.86	72.54	93.72

The Comparison table 3 of Accuracy Values explains the different values of existing Novel ML data Preprocessing, Hybrid preprocessing and proposed imputation based preprocessing algorithm. While comparing the Existing algorithm and proposed imputation based preprocessing algorithm provides the better results. The existing algorithm values start from 69.32 to 86.85, 68.35 to 87.12 and proposed imputation based preprocessing algorithm starts from 93.72 to 98.63. The proposed method provides the great results.

Figure 4 Comparison chart of Accuracy



The Figure 4 Shows the comparison chart of Accuracy demonstrates the existing Novel ML data Preprocessing, Hybrid preprocessing and proposed imputation based preprocessing algorithm. X axis denote the Dataset and y axis denotes the Accuracy ratio. The proposed imputation based preprocessing algorithm values are better than the existing algorithm. The existing algorithm values start from 69.32 to 86.85, 68.35 to 87.12 and proposed imputation based preprocessing algorithm values starts from 93.72 to 98.63. The proposed method provides the great results.

F-measure: F-measure is known as the consonant mean of precision and recall.

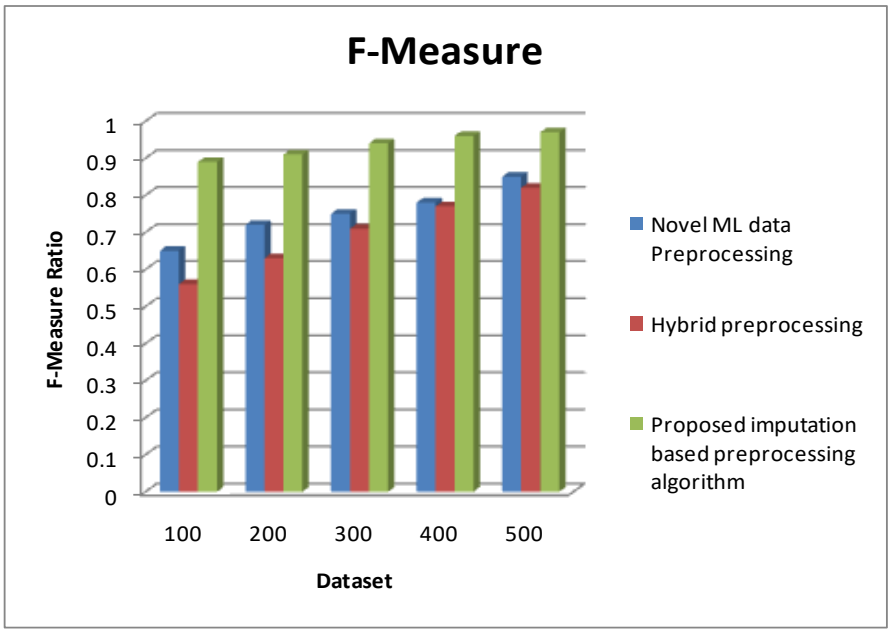
$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table 4. Comparison table of F-measure

Dataset	Novel ML data Preprocessing	Hybrid preprocessing	Proposed imputation based preprocessing algorithm
100	0.65	0.56	0.89
200	0.72	0.63	0.91
300	0.75	0.71	0.94
400	0.78	0.77	0.96
500	0.85	0.82	0.97

The Comparison table 4 of F-measure Values explains the different values of existing Novel ML data Preprocessing, Hybrid preprocessing and proposed imputation based preprocessing algorithm. While comparing the Existing algorithm and proposed imputation based preprocessing algorithm provides the better results. The existing algorithm values start from 0.65 to 0.85, 0.56 to 0.82 and proposed imputation based preprocessing algorithm values starts from 0.89 to 0.97. The proposed method provides the great results.

Figure 5 Comparison chart of F-measure



The Figure 5 Shows the comparison chart of F-measure demonstrates the existing Novel ML data Preprocessing, Hybrid preprocessing and proposed imputation based preprocessing algorithm. X axis denote the Dataset and y axis denotes the F-measure ratio. The proposed imputation based preprocessing algorithm values are better than the existing algorithm. The existing algorithm values start from 0.65 to 0.85, 0.56 to 0.82 and proposed imputation based preprocessing algorithm values starts from 0.89 to 0.97. The proposed method provides the great results.

Conclusion

In this paper we proposed Imputation based pre-processing algorithm for pre process the heart disease dataset. It was possible to use a great portion of the dataset while reducing the time taken to preprocess the data by more than 50% when compared with its local execution. The experiment method showed that the proposed Imputation based preprocessing algorithm method provides the great results.

References

1. Alrawashdeh, K.; Purdy, C. Toward an online anomaly intrusion detection system based on deep learning. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 195–200.
2. Barolli, L., Takizawa, M., Xhafa, F., Enokido, T., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 498–511.
3. Belouch, M.; El Hadaj, S.; Idhammad, M. Performance evaluation of intrusion detection based on machine learning using Apache Spark. Procedia Comput. Sci. 2018, 127, 1–6.
4. Brownlee, J. Supervised and unsupervised machine learning algorithms. Mach. Learn. Mastery 2016, 16.
5. Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018). *Prediction of Cardiovascular Disease Using Machine Learning Algorithms. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT).*
6. Heaton, J. AIFH, Volume 3: Deep Learning and Neural Networks; Heaton Research, Inc, 2015; ISBN 1-5057-1434-6.
7. Iliou, T., Anagnostopoulos, C.-N., Nerantzaki, M., & Anastassopoulos, G. (2015). A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance. (INNS) - EANN ’15.

8. Learning, D. Ian Goodfellow, Yoshua Bengio, Aaron Courville; MIT Press: Cambridge, MA, USA, 2016.

9. Nair, P., & Kashyap, I. (2019). Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for KNN Classifier. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

10. Rajagopal, S.; Kundapur, P.P.; Hareesha, K.S. A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets. Secur. Commun. Netw. 2020, 2020, 4586875.

11. Revathi, S.; Malathi, D.A. A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. Int. J. Eng. Res. Technol. 2013, 2, 1848–1853.

12. Strigl, D.; Kofler, K.; Podlipnig, S. Performance and Scalability of GPU-Based Convolutional Neural Networks. In Proceedings of the 2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing; IEEE: Piscataway, NJ, USA, 2010; pp. 317–324.

13. Tu, D.Q.; Kayes, A.S.M.; Rahayu, W.; Nguyen, K. ISDI: A New Window-Based Framework for Integrating IoT Streaming Data from Multiple Sources. In Proceedings of the Advanced Information Networking and Applications;

14. Vinayakumar, R.; Soman, K.P.; Poornachandran, P. Applying convolutional neural network for network intrusion detection. (ICACCI), Udupi, India, 13–16 September 2017; pp. 1222–1228. 30. Haykin, S. Neural Networks: A Comprehensive Foundation, 1st ed.; Prentice Hall PTR: Upper Saddle River, NJ, USA, 1994; ISBN 978-0-02-352761-6.