# Hybrid Missing Value Imputation Algorithm- KLR

Deepti Sharma[1], Rajneesh Kumar[2], Anurag Jain[3]
[1]*Research Scholar, Department of Computer Engineering, MMEC, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, India, deeptisharma85@gmail.com*
[2]*Department of Computer Engineering, MMEC, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, India, drrajneeshgujral@mmumullana.org*
[3]*Virtualization department, School of Computer Science, University of Petroleum & Energy Studies, Dehradun, India, anurag.jain@ddn.upes.ac.in*

***Article Info***
***Page Number: 60 – 74***
***Publication Issue:***
***Vol 71 No. 2 (2022)***

***Abstract***
The data mining process mainly deals with the estimation, prediction, pattern extraction, and classification in big databases. The presence of missing values in the dataset decreases the accuracy of data mining classifiers. Therefore it is necessary to deal with missing values in the dataset to achieve accurate results. To improve the quality of data and prediction accuracy in the classification process, the authors have proposed a new hybrid missing value prediction algorithm, KLR, by combining the KNN and linear regression approach. The proposed KLR algorithm has been used for class validation and missing values imputation. Wisconsin Breast Cancer Diagnostic Dataset of 569 instances with 32 attributes from the machine learning repository of UCI, Irvinewasused to conduct the study. The Pearson Coefficient Correlation method is used for feature selection.Data normalization is performed using Min-max scaling technique. The Scikit-learn library for machine learning in python is used to complete all the experiments as the experimental framework. The mean square error method is used to evaluate the performance of the model. The proposed KLR algorithm with 450 nearest neighbors out of 569 gives the lowest MSE ie 0.00188 and more accurately predicts the missing values as compared to the classic models.

***Article History***
***Article Received****: 05 December 2021*
***Revised****: 12 January 2022*
***Accepted****: 02 February 2022*
***Publication****: 11 March 2022*

**Keywords**: Data Mining, Missing value Imputation, KNN (K Nearest Neighbor), Linear Regression, Adaboost, ANN (Artificial Neural Network).

## 1. INTRODUCTION

In recent times the amount of electronic data related to patients is constantly increasing. With the help of data mining techniques, this data is used to analyze and extract embedded hidden information.[1]. Data mining techniques are a useful and effective method that helps medical practitioners and radiologists handle many disease challenges [2]. Various standard and sophisticated tools for data mining are used for this. Machine learning and artificial neural network-based sophisticated tools demonstrate utility in disease

prediction compared to other traditional methods. Machine learning tools provide higher-order interaction among data and provide a flexible environment, which gives better and accurate predictions [3].

Data mining is a set of processes that involve the various steps to discover knowledge present in the big raw data. Data selection and integration, data cleaning and pre-processing, data transformation, pattern extraction, and knowledge discovery and use of that knowledge are the different sub-phases of the data mining process[4]. All these subphases are shown in topological order in Figure 1. Handling missing values as a part of pre-processing is a major task. Missing values are unwanted in data mining as it leads to many problems. Missing values in data can arise from many reasons like equipment malfunction, non-response questionnaire, image corruption, inaccurate measurements, incorrect data entered or experimental laboratory error, etc.[5].

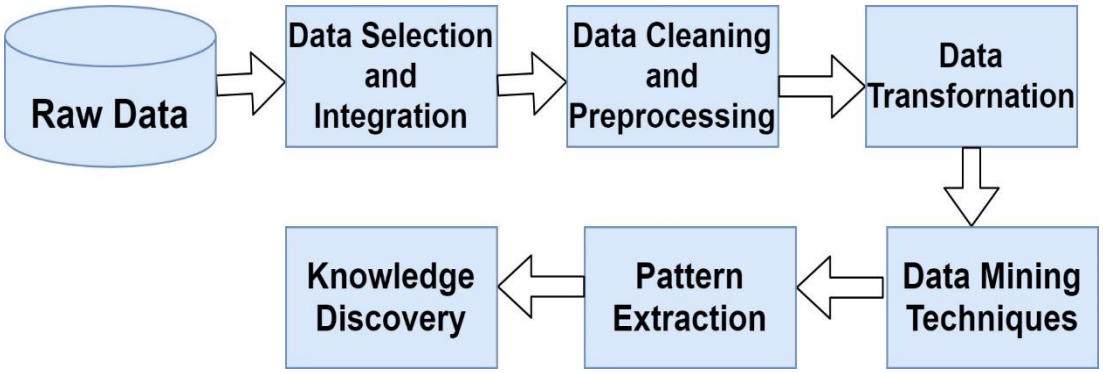


Figure 1: Data Mining Phases

### 1.1 Missing value concepts

In data mining and data warehousing, missing value handling is an important task. Incomplete values are generally present in the real-world data set, whether medical data, financial data, banking, or any other. Missing or not available values in the attributes are denoted by NA as shown in table 1 for attributes C, D, and E. Missing values in the data can arise from incorrect data entry, error in equipment, or sensor malfunctioning. The missing values present in data makes the analysis poor and inefficient prediction[6].

Table1: Missing Values are denoted by "NA" in sample dataset

| ID | A | B | C | D | E | class |
|---|---|---|---|---|---|---|
| 101 | 6 | 146 | NA | NA | 11 | M |
| 102 | 4 | 122 | 102 | NA | 22 | B |
| 103 | 3 | 179 | 107 | 57 | NA | M |
| 104 | 2 | 148 | NA | 33 | NA | B |

### 1.2 Missing value handling methods

Different missing value handling techniques found in the literature[7] are shown in Figure 2.

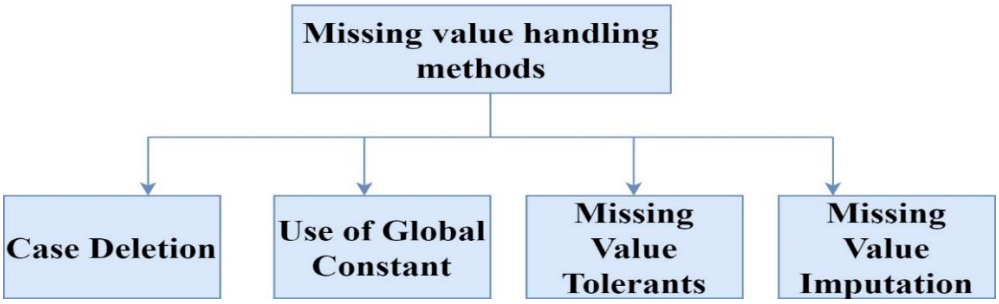


Figure 2: Missing Value Handling Methods

- **Casedeletion:**In this method, we delete the instances in which missing values are present in them.
- **Use of Global Constant**: It is the simplest method in which each missing value is replaced with some constant. Simplicity is the advantage of this, but the main disadvantage of this is that if the number of missing values is large, filling each missing value with a constant may lead to a new class that does not exist.
- **Missing value Tolerants**uses the only algorithm that consists of inherent capabilities of ignoring the missing value. Decision tree classifiers C4.5, Naïve Bayes are some algorithms.
- **Missing value imputation** is the technique in which some appropriate imputed missing value fills each missing value. Machine learning-based imputation approaches are used to evaluate missing values. The commonly used missing value imputation method is Mean, Mode, Nearest Neighbor, and Neural Network [8].

### 1.3 Types of Missing Data

Missingness Mechanism indicates the reason for the missingof values in data. Knowledge of the missingness mechanism helps decide the missing data imputation technique to find missing values. The three missing values mechanisms given by Rubin [9] are described below:

- **Missing at random (MAR)**:Horrendous missing values are not related to the missing data but related to other observed values. Other available features can guess this missing value. It may produce bias results and leads to unbalanced data.
- **Missing completely at random (MCAR)**: In MCAR, the missing data is not related to the studied variable; values in the data are said to be Missing Completely at Random if the missing data are independent of both observable data and non-observable variables and occur randomly.
- **Not Missing at Random (NMAR)**: When the missingness is neither MCAR nor MAR, it is NMAR. Missing values are dependent on the missing value itself. It is the most challenging missingness. It gives highly biased estimation results[10].

Thispaper aims to design analgorithmto handle missing values to improve the quality of data and prediction accuracy in classification. In this study, the authors analyze and use the KNN and Linear Regression algorithm to impute the missing value present in the data set. Wisconsin Breast cancer data set is used to conduct the study. No existing research has combined the K nearest neighbor and linear regression algorithm to handle missing value imputation and classification process to the best of our knowledge.

Therefore, it has motivated the authors to propose a KLR algorithm, a hybrid of KNN and linear regression algorithm.

This paper is organized as follows: Section 2 describes the related work about missing value imputation algorithms found in the literature. The proposed KLR algorithm, the dataset used, along pre-processing methods are explored in section 3. the experimental setup, approaches used, results and discussion are shown in section 4. At last, Section 5 signify the paper conclusion with future work.

## 2. RELATED WORK

As the data quantity related to the medicinal field is on the rise, the need for efficient and effective methods to deduct valuable knowledge from that data. Many researchers have been used data mining techniques on medical data for accurate predictions. This section reviews missing value imputation techniques and approaches found in the literature related to handlingmissing value imputations.

### 2.1 Missing data imputation technique

To handle the missing values, various missing value techniques have been analyzed by researchers. A few of them are given below :

- Case detection: The case Deletion method is used to remove the instances and examples having missing values from the data set.
- Mean imputation: In this approach, missing valuesarefilled by the mean of the known values of the same variation.
- Mode Imputation:In this approach, missing values are filled by the mode of the known values of the same variation
- Multiple imputations:Multiple complete data sets are created, and each data set is analyzed to impute missing values from the data set.
- KNN (K nearest neighbors): It discovers k most similar sample by non-missing values. The value of all k is aggregated by using some rules, and this value is filled.
- Hot direct imputation: It is the KNN imputation method in which k equal to 1 is considered always is called as hot back imputation
- Clustering-based KNN: An unsupervised clustering algorithm in which data gathers and divides among groups in k numbers of clusters.
- Matrix factorization:A matrix factorization is a method in which a matrix is reduced into constituent parts to perform complex matrix operations in an easier way. It is also known as the matrix decomposition method. Missing values for a given feature can be computed by the dot product of those vectors corresponding to a given feature and instance [11][12].

### 2.2 Review of existing models

In recent times, predictive classification methods have been successfully used in medical diagnosis. Many researchers have used medical data to provide valuable decision support in disease prediction at an early stage. Various diseases such as Breast cancer, diabetes, lung cancer, and many more have been predicted by available prediction models in the literature.

Many hybrid methods have also been proposed and give better results, among these In 1996 Quinlan [13]used C4.5, In 1998 Carpenter, and Markuzon[14]used Adaptive Resonance Theory Map Instance

Count (ARTMAP-IC) in which there isatotal of 576 instances sample for the training dataset, and 192 instances are for testing purpose, the model produced the accuracy of 81%. In 2003 Meesad& Yen et al.[15] explains the accuracy of diagnosis of diabetes, breast cancer, and other diseases efficiently. In 2009 PedroJ.Garcı´a-Laencina et al. [16] et al. used the KNN imputation feature-weighted procedure using a distance metric based on mutual information. This missing data estimation method whileimproving the classification task is provided in this. In 2010 Hybrid prediction model by Patil et al.[17] given the accuracy of 84.5 % in the diabetes prediction. In 2010 Kahramanli&Allahverdiet al.[18] give the accuracy of 92.38%, while Illango&Ramaraj in 2010 given 98.84% with the same diabetes dataset.

In 2010 Jerez et al.[19]proved that all the imputation techniques instead of Hot-Deck could increase the accuracy of ANN based prediction. In this prediction of breast cancer, prognosis has been carried out by comparing statistical/ machine learning imputation techniques with deletion. KNN Imputation provided a 2.71% higher improvement in prediction over deletion.

In 2014 Seera et al.[20] given a hybrid prediction model FMM (Fuzzy Min-Max)-CART-RF and compare it with other existing models Lukka et al. (2011) and Orku& Bal et al. (2011). Seera et al. got an accuracy of 78.39%, while Lukka and Orkcu got 75.97% and 77.60% in their models, respectively.

In 2015, Archana Purwar et al.[3] given the HPM-MI (Hybrid Prediction Model- Missing Value Imputation) , K means clustering and multilayer perceptron are used in this. Three different datasets name PIMA, breast cancer, and Hepatitis we used and got the accuracy of 99.82%, 99.39%, and 99.08%.

Gracia-Laencina et al. [21]applied many different missing value techniques in the year 2015 on the dataset of breast cancer,missing data with high percentage was applied to predict the survival rate in patients. Institute Portuguese of Oncology of Porto dataset was used in the study by the authors, KNN imputation method and Mode imputation methods were compared, and in terms of accuracy, KNN performed best and mode was the worst.

In 2017,Jenghara et al.[22] used ensemble concept-based missing value imputation technique on eight different datasets. Evaluation of the model was done with both RMSE ( Root Mean Square Error) and NMAE (Normalized Mean Absolute Error) and gave better accuracy as compared to the other existing models.

In 2019, Ali Idri et al. [23] used three classifiers C4.5, SVM, and MLP on two different datasets to predict the missing values. 162 experiments using KNN for MCAR, MAR, NMAR were conducted. MLP achieved the lowest accuracy regardless of the Missing Data Mechanism percentage.

In 2019, Mohammad Faiz et al.[24] proposed a missing data imputation model with fuzzy feature selection. Fuzzy Principle Component Analysis (FPCA) was used to extract features. FPCA-SVM-FCM model with fuzzy Principle Component Analysis- Support Vector Machine- Fuzzy c-means has been proposed by the authors. The model gives better results while comparing with other corresponding models.

In 2021 Ching-Hsien Hsu et al. [25] gave a feature model based on machine learning to enhance predictive modeling. Breast cancer, cervical, and lung data sets were used to in the study. GA-CFS (Genetic Algorithm – Correlation-based Feature Selection) model used the classifiers Decision Tree, SVM, LDA, and MLP-NN on all three datasets. 10-Fold cross-validation method was used, the performance of the model was evaluated on the accuracy, f-score, precision, and recall. The study's outcome achieved the accuracy of 99.62%, 96.88%, and 98.21% on breast, cervical, and lung cancer datasets, respectively.

## 3. MATERIAL AND METHODS

In this section, the authors have described theproposedKLR missing value imputation method that is formed by ensembling of KNN and Linear Regression. Detail of the ensembling algorithms and the dataset is also given in this section.

### 3.1 K Nearest Neighbor

Many researchers have widely used KNN imputation, and it performs by selecting k nearest instances to the incomplete cases. K nearest neighbors can be selected for imputing missing values and the cases with incomplete patterns. The optimal value of k is usually chosen by cross-validation. When the k nearest neighbor has been found, the missing value attribute is estimated for its replacement. The new value for the missing value is calculated depending on the type of data. The replacement value calculation depends on the type of data; the mean is used for numerical or continuous data while the mode is for discrete data. The KNN algorithm is an instance-based learning method. In this, real-valued and discrete-valued functions are approximated. According to this, instances within a data set will generally exist near other cases that have the same properties [26].Calculation of k closest instances to the incomplete value depend on the distance metric. Several distance metrics used are Euclidean distance, Manhattan distance, and Hamming distance. In this paper, authors have used Euclidian distance, where the distance between the instances $x_i$ and $y_i$ is assessed by using the equation Eq. 1

$$D_E(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Here *n* describes the number of instances describing the instances $x_i$ and $y_i$ [27].

### 3.2 Linear Regression

Linear regression is a linear modelin which input variables x and the single output variable y is assumed to be in a linear relationship. In simple linear regression, single input variable is used in linear combination to calculate y and for multiple inputs, variables are known as multiple linear regression. Linear regression is the next step after correlation. To predict the value of a variable based on the value of another variable linear regression is used. The variable which is predict is called the dependent variable, and the variable used to predict the other variable is the independent variable [28].
A linear regression line is calculated by equation(2) of the form

$$Y = a + bX \quad (2)$$

where *X* ; independent variable and *Y* ; dependent variable. The slope of the line is *b*, and *a* is the intercept. In higher dimensions, when we have more than one input x, the line is called a plane or a hyper-plane.

#### 3.2.1 Predictions Using Linear Regression

Linear regression defines the relationship between two variables by fitting a linear equation to observed data. To make a linear model to observed data, a relationship between the considered variable must be

there. A significant association between the variables must be exist but not necessary to have causal relationship between the variables. To determine the strength between the two variables scatterplot is an effective tool. If there is non existence of association among the proposed explanatory and dependent variables, application linear regression model to the data may not be helpful in developing a useful model. Correlation coefficient is valuable numerical measure of association between two variables,the strength of the association of the observed data for the two variables is indicated by the values between -1 and 1.

### 3.2.2 Least-Squares Regression

A regression line is the standard fitting method of least-squares. The best-fitting line of the observed data points is calculated in this method by minimizing the sum of the squares of the vertical deviations from each data point to the line. There are no cancellations between positive and negative values because the deviations are first squared and then summed.

### 3.2.3 Outliers and Influential Observations

Outliers are the data points that lie far away from the line when the regression line is computed for a group of data points. These outliers show the overfitting or underfitting regression line. It represents erroneous data. A point when lies far from the other data in the horizontal direction is called as an influential observation. The significant impact on the slope of the regression line is the reason for this distinction in data points[29].

## 3.3 Data Set Description

A variety of screening methods are available to predict the disease in various phases. Human errors while entering data or misinterpretation of data may cause adverse effects inprediction. This study aims to design a model to handle missing values to improve the quality of data and prediction accuracy in classification for computer-aided diagnostic methods.
The proposed model is validated through the Wisconsin Breast Cancer Diagnostic Data set from the University of California, Irvine. The data set is publicly available and contains a total of 569 instances described by 32 attributes. The class attribute has two classes Benign and Malignant[30].

## 3.4 Data Preprocessing

A breast cancer dataset is used to conduct the study. There are no missing values present in the original dataset. Therefore we have to induce the missing values in the dataset to validate the proposed study. Following pre-processing steps are implemented before algorithm testing.

### 3.4.1 Feature Selection

In machine learning, feature selection is an important phase before the training of the model. Feature selection is a pre-processing technique in which attributes that are not relevant are removed. It deals with finding the most useful and informative features that be used to increase the performance of prediction

training and evaluation[31]. Feature selection is a method to find the best subset of features from the entire big set of features. The three effective feature selection methods are

- Filter method
- Embedded method
- Wrapper method

Feature Selection methods like complete, heuristics and random using different evaluation functions like distance, information, classification error rate, and dependency between attributes are used to make a subset from the original dataset. In the diagnosis of disease in the patients, it is necessary to follow pre-processing steps to attain accuracy in classification problems[32].

### 3.4.2 Data Normalization

After feature selection, to avoid the harmful effects of large-scale attributes on evaluation criteria (mean squared error), all the attributes are normalized. The normalization of data is done by using Min-max scaling normalization.

### 3.5 Proposed KLR Algorithm

In this paper, a hybrid algorithm for Missing Value Imputation is developed, shown in Figure 3. The entire process is divided into three parts; First Missing value introduction, Second KNN Imputation, and third model performance evaluation and conclusion. Table 2 shows the step-by-step procedure of the proposed work, and all the details of the algorithmaregiven after this.

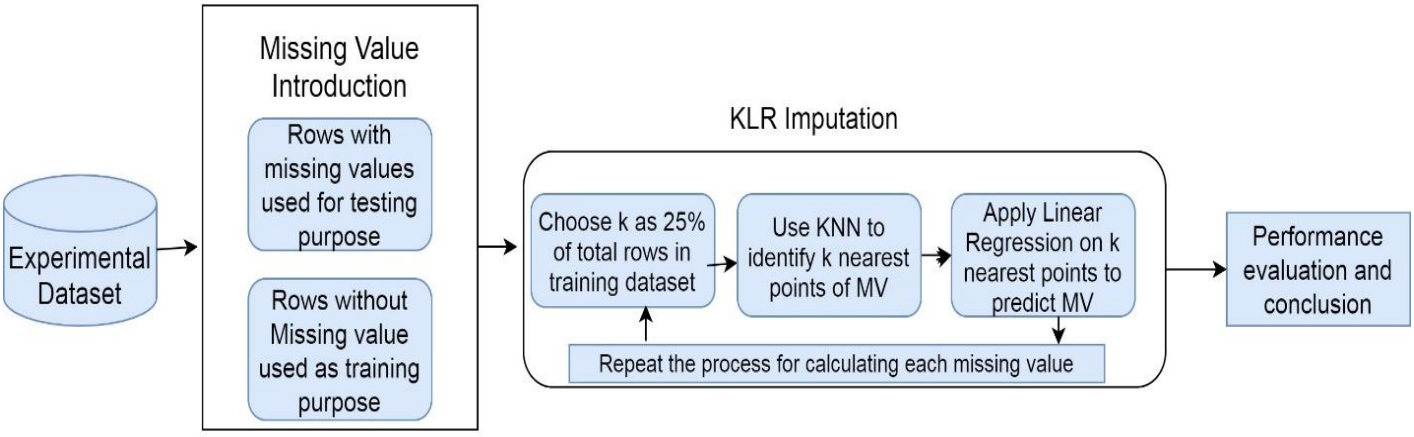


Figure 3. Proposed KLR Model

Table 2: Procedure for developing KLR model

| |
|---|
| Step 1: Get the dataset<br>Step 2: Segregate the rows having missing values and use those rows for testing purposes. At the same time,other rows which don't have any missing value use them for training purposes.<br>Step 3: Repeat the following process for each missing value:<br>Step 3.1: Choose the value of k as 25% of the number of rows in the training dataset.<br>Step 3.2: Use the KNN algorithm to identify the K nearest point of the missing value.<br>Step 3.3: Apply linear regression on the k nearest points found in step 3.2 to predict the missing value.<br>Step 4: Evaluate the performance of the model. |

### 3.5.1 Missing Value Introduction

According to the proposed algorithm,first, segregate the rows having missing values from the dataset. All the rows with missing values will be used for the testing of the model. Rows that don't have the missing values have to be used for the training of the model. Itis the first step of the proposed model in which training and testing datasets are decided.

### 3.5.2 KLR Imputation Phase

This proposed model deals with the missing value imputation and gives predictions about the classification of patients. This K Nearest Neighbor and Linear Regression phase of the model have further distributed in three subparts, as shown in Figure 3.

In step 1,Choose the value of k as 25% of the number of rows in the training dataset.

In step 2 of this phase, K nearest neighbor algorithm is used to identify the k nearest points of the missing value.

In step 3,applytheLinear Regression algorithm on the k nearest points found in step 2 to predict the missing value. All three steps are repeated and processed to impute all the missing values.

Missing value imputation is a critical phase in this proposed model. The main objective of the model is to provide an accurate classification of patients while dealing with missing values present in the data. For developing our proposed model, K Nearest Neighbor and Linear Regression Algorithm are used to attain accurate results. After evaluating the missing values, a performance evaluation of the model is carried out.

### 3.5.3 Performance Evaluation

The performance of the model is evaluated using Mean squared error;it is the average of the mean squared error or deviation. The average square difference between the estimated values and the actual values is the MSE(Mean Squared Error). MSE is a risk function related to the expected value of the squared error loss.

It is an essential method to determine the estimator's performance. It is denoted as MSE; an average of squares of errors. The error is the difference between the estimated attributes and the estimator is given by Equation (4);

$$\text{MSE}\ (\theta_{obs,i}) = \text{E}\left[\left(\theta_{obs,i} - \theta\right)^2\right] \qquad (3)$$

Mean Squared Error is not a random variable as it is an expectation. $\theta$is the unknown parameter, and the estimator is $\theta_{obs,i.}$[33].

## 4. RESULTS AND ANALYSIS

The main objective of the experiment conducted in this work is to handle missing values and to improve the prediction accuracy. Details of the simulation environment, results, and their analysis are given in this section.

### 4.1 Simulation Environment

The data set used for the study is Breast Cancer Wisconsin Diagnostic Dataset of 569 instances with 32 attributes from UCI, Irvine Machine learning repository. The Scikit-learn library for machine learning in python is used to perform all the experiments as the experimental framework. The feature selection of the relevant attributes is processed using Pearson's Coefficient Correlation method, selecting those features whose absolute value of correlation coefficient with the dependent feature is more than 0.5. After feature selection for missing value, the estimation dataset is normalized to remove the effect of large-scale attributes. Data normalization using Min-Max scaling is performed to analyze the results effectively. In Wisconsin Diagnostic dataset, there are no missing values, so we approached in the following way:

1. Iterate through every column of the dataset
2. In the selected column, introduce 10% missing values randomly.
3. Now, we will impute these introduced missing values with the below proposed algorithm which has 3 methods

a. KNN
b. KNN + Adaboost Regression (KADR)
c. KNN + Linear Regression (KLR)

Missing values are not present in the dataset used in the study. Therefore artificial missing values of 10% rate according to the Missing at Random mechanism were generated. With the artificial missing values, we control the missing value, which helps in the evaluation process. The total instances in the data set are 569, so we decided to run the experiment for ten iterations. In each iteration, every selected column will be introduced with 10% missing values randomly. For each column now, we will impute the missing value with the proposed KLR algorithm. As we observe by introducing 10% missing values in a column, we are introducing 59 missing values in each iteration. In this way, after all the iterations, we have analyzed all the values of the complete dataset. As in our dataset 10% MAR mechanism works well, the percentage of missing mechanisms can be increased according to the size of the dataset. At the end of all the iterations, the true error rate can be estimated by accumulating the mean error rate for each iteration. The detailed missing value imputation evaluation algorithm performance procedure is stated below:

**pseudo code for the algorithm:**

iteration_wise_mse_list = []
iterations = 10
1. for each iterations:
column_wise_mse_list = []

for each column in data:
a. introduce 10% missing values at random.
b. impute these introduced values using below methods
    1. K-Nearest Neighbors (KNN)
    2. K-Nearest Neighbors + AdaboostReggression (KADR)
    3. K-Nearest Neighbors + Linear Reggression (KLR)
c. calculate Mean Squared Error (MSE) for the selected column by
comparing actual values and predicted missing values
d. store/append MSE for selected column in a list(column_wise_mse_list)

calculate average of MSE stored in column_wise_mse_list
store/append the average MSE in iteration_wise_mse_list.

2. calculate average of MSE stored in iteration_wise_mse_list

In our case we ran the above algorithm on below number of neighbors:
n_neighbors = [270, 300, 330, 360, 390, 420, 450, 480]

While Applying adaboost we used DecisionTreeRegressor with max depth equals to 5

```
dt = DecisionTreeRegressor(max_depth=5)
```

Graph in the figure 5 will show the comparison of all the 3 algorithms being used KNN, KADR and KLR.

### 4.2 Discussion of Results

Statistical and Machine learning-based missing value imputation techniques were used to predict the missing values in the breast cancer dataset. The main aim of the paper is to impute the missing value and to find the accuracy. For this, we have analyzed K Nearest Neighbor, K Nearest Neighbor with Adaboost and proposed a hybrid KLR(K Nearest Neighbor with Linear Regression) model. To validate the choice of selection of neighbors in the proposed algorithm, the value of Nearest Neighbors is analyzed with different nearest neighbors as shown in Table 3.
We have also analyzed KNN,KNN_Adaboost and, proposed KLR algorithms through min-max scaling normalizations and compared them for better dealing with the large-scale attributes.

Table 3: Average Mean Squared Error Values for KNN, KNN_AdaBoost and KNN_Linear Regression with different Nearest Neighbors

| Algorithm | Nearest Neighbor= 270 | Nearest Neighbor= 300 | Nearest Neighbor=330 | Nearest Neighbor 360 |
|---|---|---|---|---|
| KNN | 0.00597 | 0.00573 | 0.00592 | 0.00594 |
| KNN_AdaBoost | 0.00299 | 0.00305 | 0.00297 | 0.00305 |
| KLR(Proposed) | 0.00202 | 0.00207 | 0.00215 | 0.00201 |

| Algorithm | Nearest Neighbor= 390 | Nearest Neighbor= 420 | Nearest Neighbor= 450 | Nearest Neighbor= 480 |
|---|---|---|---|---|
| KNN | 0.00598 | 0.00606 | 0.00608 | 0.00628 |
| KNN_AdaBoost | 0.00295 | 0.00319 | 0.00309 | 0.00320 |
| KLR(Proposed) | 0.00192 | 0.00195 | 0.00188 | 0.00195 |

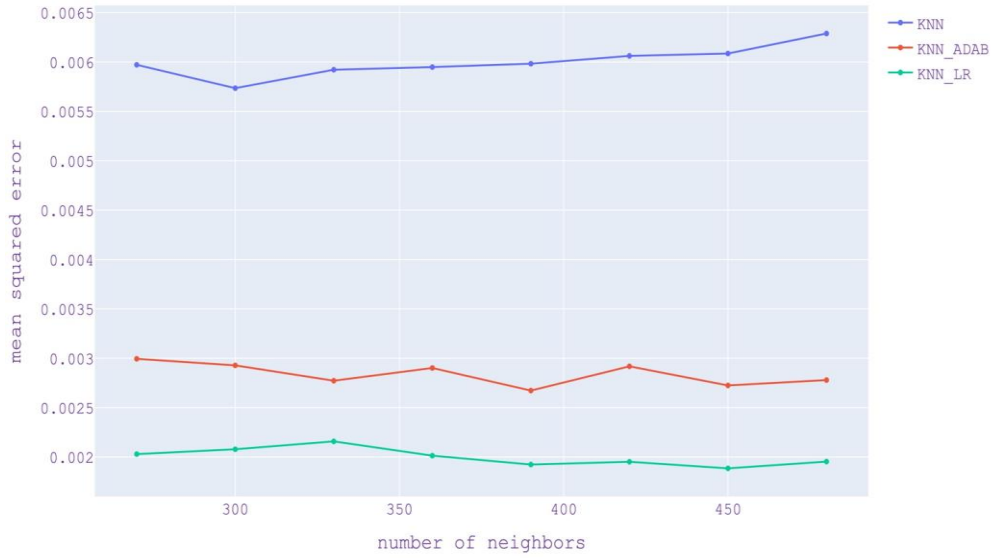


Figure 4: A plot of the Mean Squared Error with different numbers of neighbors

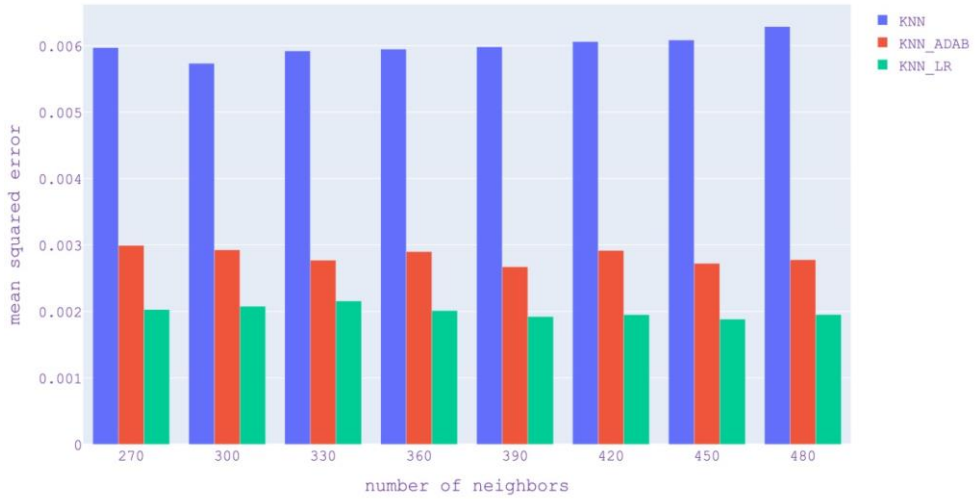


Figure 5 : Graph for KNN, KNN_ADABOOST, and KNN_LR

Figure 4 shows the values plotted for the different nearest neighbors and their corresponding Mean Squared Error as shown in table Table 3, and Figure 5 shows the graph-based upon the values of table 3. The performance of KNN,KNN_Adaboost, and KLR algorithms are analyzed and compared. It has been concluded that the proposed KLR algorithm works better as compared to the KNN and KNN_AdaBoost algorithms..The performance of the algorithm is evaluated on the mean squared error method. The value of MSE shows the error in the prediction, and for better classification prediction, it should be less/low. Higher the value of mean squared error; lower the accuracy and lower the value of mean squared error; higher the accuracy.

## 5. CONCLUSION AND FUTURE SCOPE

Missing values are an obstacle in the area of data mining. Presence of missing values in the data may mislead the outputs of data mining applications in an incorrect way. A hybrid KLR algorithm for missing value imputation for accurate classification and improving data quality has been presented in this paper. KNN and linear regression algorithms are merged for designing the new algorithm. Experimental results of the model have shown that the proposed model helps effectively in dealing with missing values and gives accurate results. The approach used in the proposed algorithm for estimating the missing value has an extremely significant effect on the performance of the model. The proposed algorithmevaluated using the Mean Square Error method.In the future, authors have planned to test the proposed algorithm on other datasets containing some missing values, with different values of K. Moreover, different imputation methods with the proposed algorithm could be analyzed with different datasets (other than medical datasets) could be analyzed to explore which kind of data suits it more.

## References

1. Lin, Wei-Chao, and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)." *Artificial Intelligence Review* 53, no. 2 (2020): 1487-1509.
2. Jerez, José M., Ignacio Molina, Pedro J. García-Laencina, Emilio Alba, NuriaRibelles, Miguel Martín, and Leonardo Franco. "Missing data imputation using statistical and machine learning methods in a real breast cancer problem." *Artificial intelligence in medicine* 50, no. 2 (2010): 105-115.
3. Purwar, Archana, and Sandeep Kumar Singh. "Hybrid prediction model with missing value imputation for medical data." *Expert Systems with Applications* 42, no. 13 (2015): 5621-5631.
4. Bertsimas, Dimitris, Colin Pawlowski, and Ying Daisy Zhuo. "From predictive methods to missing data imputation: an optimization approach." *The Journal of Machine Learning Research* 18, no. 1 (2017): 7133-7171.
5. Xia, Jing, Shengyu Zhang, GuolongCai, Li Li, Qing Pan, Jing Yan, and Gangmin Ning. "Adjusted weight voting algorithm for random forests in handling missing values." *Pattern Recognition* 69 (2017): 52-60.
6. Bertsimas, Dimitris, Colin Pawlowski, and Ying Daisy Zhuo. "From predictive methods to missing data imputation: an optimization approach." *The Journal of Machine Learning Research* 18, no. 1 (2017): 7133-7171.
7. Purwar, Archana, and Sandeep Kumar Singh. "Empirical evaluation of algorithms to impute missing values for financial dataset." In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, pp. 652-656. IEEE, 2014.
8. Peng, Liu, and Lei Lei. "A review of missing data treatment methods." *Intell. Inf. Manag. Syst. Technol* 1 (2005): 412-419.
9. Rubin, Donald B. "Inference and missing data." *Biometrika* 63, no. 3 (1976): 581-592.

10. Moons, Karel GM, Rogier ART Donders, Theo Stijnen, and Frank E. Harrell Jr. "Using the outcome for imputation of missing predictor values was preferred." *Journal of clinical epidemiology* 59, no. 10 (2006): 1092-1101.
11. Cheng, C., and H. Huang. "A Distance-threshold K-NN Method for Imputing Medical Data Missing Values." *Journal of Advances in Computer Networks* 7, no. 1 (2019): 13-17.
12. Vazifehdan, Mahin, Mohammad Hossein Moattar, and MehrdadJalali. "A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction." *Journal of King Saud University-Computer and Information Sciences* 31, no. 2 (2019): 175-184.
13. Quinlan, J. Ross. "Improved use of continuous attributes in C4. 5." *Journal of artificial intelligence research* 4 (1996): 77-90.
14. Carpenter, Gail A., and Natalya Markuzon. "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases." *Neural Networks* 11, no. 2 (1998): 323-336.
15. Meesad, Phayung, and Gary G. Yen. "Combined numerical and linguistic knowledge representation and its application to medical diagnosis." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 33, no. 2 (2003): 206-222.
16. García-Laencina, Pedro J., José-Luis Sancho-Gómez, Aníbal R. Figueiras-Vidal, and Michel Verleysen. "K nearest neighbours with mutual information for simultaneous classification and missing data imputation." *Neurocomputing* 72, no. 7-9 (2009): 1483-1493.
17. Lobo, Vijaya, AvinashPatil, A. Phatak, and Naresh Chandra. "Free radicals, antioxidants and functional foods: Impact on human health." *Pharmacognosy reviews* 4, no. 8 (2010): 118.
18. Kahramanli, Humar, and NovruzAllahverdi. "Design of a hybrid system for the diabetes and heart diseases." *Expert systems with applications* 35, no. 1-2 (2008): 82-89.
19. Jerez, José M., Ignacio Molina, Pedro J. García-Laencina, Emilio Alba, NuriaRibelles, Miguel Martín, and Leonardo Franco. "Missing data imputation using statistical and machine learning methods in a real breast cancer problem." *Artificial intelligence in medicine* 50, no. 2 (2010): 105-115.
20. Seera, Manjeevan, and Chee Peng Lim. "A hybrid intelligent system for medical data classification." *Expert Systems with Applications* 41, no. 5 (2014): 2239-2249.
21. García-Laencina, Pedro J., Pedro Henriques Abreu, Miguel Henriques Abreu, and NoémiaAfonoso. "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values." *Computers in biology and medicine* 59 (2015): 125-133.
22. Jenghara, Moslem Mohammadi, Hossein Ebrahimpour-Komleh, VahidehRezaie, SamadNejatian, Hamid Parvin, and Sharifah Kamilah Syed Yusof. "Imputing missing value through ensemble concept based on statistical measures." *Knowledge and Information Systems* 56, no. 1 (2018): 123-139.
23. Chlioui, Imane, Ali Idri, IbtissamAbnane, Juan Manuel Carillo de Gea, and Jose Luis Fernández-Alemán. "Breast cancer classification with missing data imputation." In *World Conference on Information Systems and Technologies*, pp. 13-23. Springer, Cham, 2019
24. Dzulkalnine, Mohamad Faiz, and RoselinaSallehuddin. "Missing data imputation with fuzzy feature selection for diabetes dataset." *SN Applied Sciences* 1, no. 4 (2019): 1-12.
25. Hsu, Ching-Hsien, Xing Chen, Weiwei Lin, Chuntao Jiang, Youhong Zhang, ZhifengHao, and Yeh-Ching Chung. "Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning." *Measurement* 175 (2021): 109145.
26. Acurna, E., and C. Rodriguez. "The treatment of missing values and its effect in the classifier accuracy, classification, clustering, and data mining applications." In *Proceedings of the meeting of the International Federation of Classification Societies (IFCS)*, pp. 639-647. 2004.
27. Zhang, Shichao, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng. "Learning k for knn classification." *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, no. 3 (2017): 1-19.

28. Khashei, Mehdi, Ali ZeinalHamadani, and Mehdi Bijari. "A novel hybrid classification model of artificial neural networks and multiple linear regression models." *Expert Systems with Applications* 39, no. 3 (2012): 2606-2620.
29. Weisberg, Sanford. *Applied linear regression*. Vol. 528. John Wiley & Sons, 2005.
30. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
31. Sharma, Deepti, Rajneesh Kumar, and Anurag Jain. "A Systematic Review of Risk Factors and Risk Assessment Models for Breast Cancer." *Mobile Radio Communications and 5G Networks* (2021): 509-519.
32. Poolsawad, N., C. Kambhampati, and J. G. F. Cleland. "Feature selection approaches with missing values handling for data mining-a case study of heart failure dataset." *World Academy of Science, Engineering and Technology* 60 (2011): 828-837.
33. Willmott, Cort J., and Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." *Climate research* 30, no. 1 (2005): 79-82.