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

Prediction and Analysis of Crime against Women Data using Machine Learning and Statistical Imputation Techniques

Tamilarasi P.[1], Uma Rani R.[2]

Submitted: 26/11/2023 **Revised**: 06/01/2024 **Accepted**: 16/01/2024

Abstract: Data analytics is a vast research field all around the world. Missing data is a significant difficulty in this discipline, as missing values decrease the algorithm's performance. In research, imputations of missing values are critical; incorrect imputation of absent values leads to erroneous predictions.These types of missing data must be handled efficiently.In this article, statistical-based and Machine Learning based imputation methods are proposed. These values are applied in two different crimes against women data set.There are thirteen types of machine learning algorithms implemented by using these proposed values. Simple linear, multiple linear, KNNr, SVR. Polynomial regression, Decision Tree Regression, and Random forest regressions are used to predict the crime rate against women in India and Salem District in Tamil Nadu. The Novel algorithm KNN_ET has proposed for missing data imputation whose value is compared with other statistical and machine learning-based imputations such as mean, median, mode, K-Means cluster, fuzzy C-Means cluster, K-Medoids and K-nearest neighbor method. The main aim of this work is to reduce errors and improve machine learning accuracy. Finally, the algorithm efficiency is compared based on MSE, MAE, RMSE for Prediction. Proposed algorithm KNN_ET has reduced the maximum number of errors and gives accuracy 94.78 % for Salem District Crime Data-set and 98.7 % accuracy for India level Crime dataset .This is better accuracy compared with all other imputation techniques. In the future, this work will be very helpful to the police department for control the crimes against women in India and Salem.

Keywords: *Statistics, Machine Learning, Missing Values Imputation, Fuzzy C-means, K-means, KNN*

1. Introduction

Data analytics is a vast area of research. Analytics are mainly used to analyze unrefined data to create a conclusion concerning that data.Data analytics is a highly developed category of data mining, a grouping of statistical and mathematical systems. The different types of fields used in data analysis, such as agriculture, health, and crime are some important areas. There are different types of statistical analysis methods available, including descriptive, predictive, positive, and diagnostic. This article focuses on predictive and descriptive analytics on crime against women data. Predictive Analytics is an environment-friendly technique. It is similar to Data analytics and machine learning techniques which help to predict possible conclusions, Primarily based on time-series statistics. Analytics of Descriptive is an algebraic approach used to discover and evaluate ancient statistics to capture patterns and meanings. This article focuses on predictive and descriptive analytics on crime against women data. Predictive Analytics is an environment-friendly technique. It is similar to Data analytics and machine learning techniques which help to predict possible conclusions, Primarily based on time-series statistics. Analytics of Descriptive is an algebraic approach used to discover and evaluate ancient statistics to capture patterns and meanings. Machine Learning (ML) is the most important sub-area of Artificial Intelligence; it is uncomplicated to realize. ML procedures are revised from familiarity like a person does the work without the coding.

When delivering the data to ML, those models read, build up and change them. These ML procedures independently adapt the new data through iteration. ML algorithms produced reliable results based on the previous computation and operation. Machine Learning models are primarily used for prediction and are strongly connected to statistics and mathematics functions.ML is categorized into two major parts such as supervised and unsupervised. Supervised handles the labeled data but unsupervised works with the unlabeled data. In this work, the supervised algorithms of regression and classification techniques are implemented to predict the crime rate against women. Various types of regression models have been proposed for prediction.

Crime against women has the potential to be a very serious and dangerous national problem. Recently it's increasing frequently. Various styles of crimes are registered against women is now on a daily basis. This could be controlled and avoided otherwise it will stop our nation's growth. The govt considers this problem and takes steps for controlling the crimes against women. In India, ladies' defence cannot be definite by just having a policy and a group of laws. Those

[1]*Research Scholar, Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem-16, Tamil Nadu. India. Email: tamilinresearch@gmail.com*

[2] *Principal, Sri Sarada College for Women (Autonomous), Salem-16, Tamil Nadu. India. Email: principalsscsalem@gmail.com*

days' the government framed many laws for growing our society with nil percent persistence of offenses registered against women, but the offenses are endlessly rising. Many researchers attempt to solve this problem with the assistance of information analytics. During this work, the crime data are collected from 2003 to 2012 in India. There are more than one billion records for various types of crimes. Cruelty and kidnapping are the highest recorded categories of crime against women in India. So these styles of crimes are used for prediction in this work. Misplaced data imputation is a method used to replace the missing data with some alternative values to retain most of the information. The researchers are mostly using these techniques because every time data removal is not feasible and data deletion leads to errors and inaccurate analysis. Missing data handling is important since it causes some issues. When using the Machine Learning packages it cannot handle the missing values automatically. A vast quantity of missing information is able to cause deformation in the label. So this is a boost or reduces the value of the specific class in the dataset. One more main reason is the need to store the complete information when all the values are more significant in the dataset. If the data volume is small, removing the information causes an important collision in the final model.

2. Related Work

Chia-Hui Liu et.al [1] proposed genetic and information gain algorithms for imputing the missing values using low-dimensionality medical datasets and decision tree algorithms are applied for high-dimensional data sets. Waseem Shahzad et.al [2] compared single and multiple value imputation with proposed techniques for genetic algorithm performance using various types of classification methods. Tlamelo Emmanuel et.al [3] used K Nearest Neighbor and Random forest algorithms for missing values imputation in the iris dataset and novel ID dataset. Finally, the author concluded that KNN performed better than Random forest. Dimitris Bertsimas et.al [4] applied K-NN, SVM, Decision Trees, and opt. impute the method for missing data and experiment with 84 datasets. The author has given the result with better Mean Squared Errors. KohbalanMoorthy [5] focus on the current techniques for the imputation of misplaced values in gene expression records.

Huimin Wang [6] implemented machine learning and traditional algorithms for handling the missing values and concluded the Ensemble Learning algorithms are well suited for missing values with a 2:1 missing values ratio. Michael B. Richman [7] proposed new methods for filling in the missing values and minimizing errors. C. G. Marcelino et.al [8] investigate the miss data effect on linear regression in eight various types of datasets and conclude that the K-NN has performed better than other algorithms. Ugochinyere I. Nwosu et al [9] compared the new proposed buys ballot table with the existing one by using time series data. Shin-mu Tseng et.al [10] applied a cluster and regression tree model for dealing with the missing values.

Ibrahim Gad [11] executed deep learning stochastic gradient decent optimization techniques for missing data and gives the result with high accuracy. Amjad Ali and Qamruz Zaman [12] reduced the out-of-bag errors by applying Conventional and IN/OUT Random forest algorithms for missing values. Kritbodin Phiwhorm et. al[13] execute adaptive multiple imputation techniques for missing values and compared the result with mean and median methods and state the result that the performance of the adaptive techniques is better than mean and median. Chung-yuan-cheng et al [14] applied deep learning methods like ANN and CNN techniques for missing values in the attention deficit hyperactivity disorder dataset and get the result with 89 percent accuracy. Huaxiong Li [15] utilized iterative techniques for filling in the missing data and the solution as this method gives better results.

Youngdoo Son and Wonjoon Kim [16] handled the different machine and deep learning procedures for filling null values by using non-linear data. C. G. Marcelino et al [17] used dissimilar categories of models like K-NN, SVM, AdaBoost, and Neural Networks for null values and conclude the output K-NN performed well compared to others.P.S.Raja and K.Thangavel proposed k-means centered based missing value imputation techniques based on roughest and compared the output with other cluster-based imputation techniques [18]. Gobal Krishna M et al used eight imputation techniques and compared the result based on the algorithm performance and decided the results of MICE and PMM efficiencies are well better than others [19]. Khaled M. Fouad et al executes K-NN and fuzzy c-mean techniques for missing data and proposed new techniques called FCKI. The FCKI work is superior to the others [20].

This paperwork intends to analyze the missing data using various machine learning and statistics strategies and broaden techniques to enhance the algorithm performance. Based on the analysis, we offered a soft computation primarily based on the imputation algorithms and verified with two various crimes against women datasets. The remaining part of the article gives details as follows. The workflow of the imputation strategies block diagram is defined in part 3. The statistical imputation strategies are mentioned in part 4. Existing machine learning models which can be used for imputation are mentioned in part 5. Results are discussed in part 6, imputation strategies performances and analysis of the offense records against women are in part 7. Part 8, concludes the result of the paperwork.

3. Work Flow of Imputation Strategies Block Diagram

Part 3 provides a brief explanation of the imputation process workflow in Figure 1. Initially, the crime datasets are partitioned with missing and without missing data. In this work, two types of imputation techniques are used for handling the Missing data sets which are the statistical-based approach and the machine learning-based approach.

Mean, median, mode, K-means clustering, and fuzzy c means are centric value-based approaches and Knn is the nearest value or attribute-based approach. Center values are applied for all missing fields in the datasets. In the K-mean cluster, the mean value from all the clusters is found and that value is applied to impute the missing values and the same process is applied for the fuzzy c-means cluster. Using Knn imputation techniques, nearest neighbor values are applied for all null field data.In this work, a novel method KNN_ET is used for imputing the missing data. This KNN_ET proposed method is used to predict the crime rate against women by using various categories of regression techniques. Algorithm 1, explained Imputations of missed value using mean.

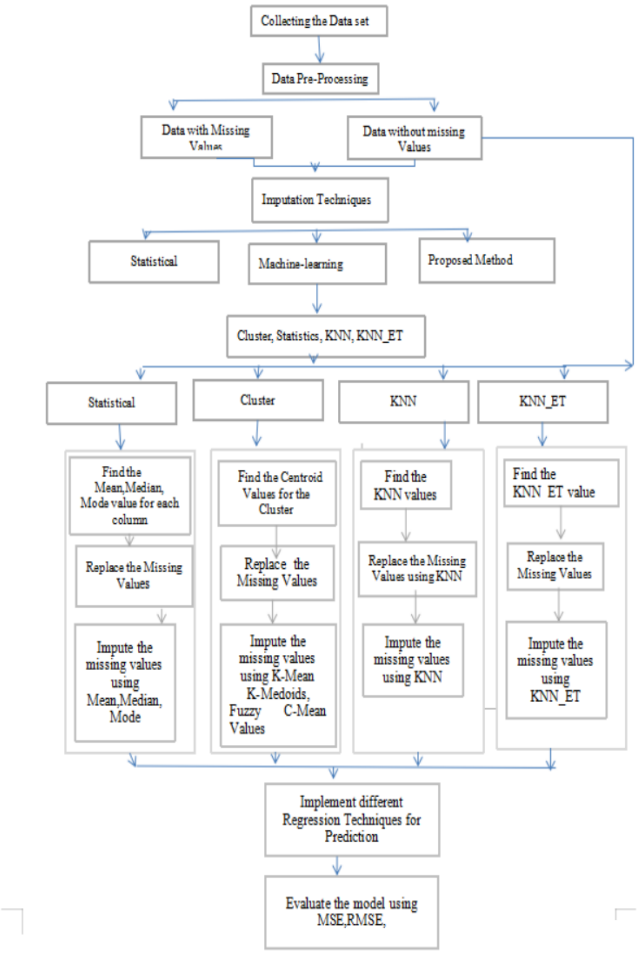


Fig 1: Work Flow of Imputation Process Block Diagram

4. Statistical Imputation Methods for Missing Values

The imputation modifies the missing data in the alternate evaluation in statistics. Data point replacement is called "unit mapping". When a substitution is applied to a part of a data item, it is called "element mapping". Most machine learning techniques require input in numeric format and all rows and columns of the data-set must hold numeric data. Therefore, missing values may cause problems in machine learning algorithms. Therefore, it is common to detect missing values in datasets and modify them numerically. This is known as data mapping or missing data mapping. A well-known approach to Data allocation is to use a statistical function to calculate a column approximation from an existing value and change the missing value in the column in the calculated statistics. Statistics are popular because they are quick and easy to calculate and are often very valuable. Missing attribute values are filled by means which are used for the method of single imputation and the resulting fulfilled data are used for implication. Missing information is replaced by estimated values which values are applied for named missing points in datasets. These imputed values are calculated from central tendency which are usually known as statistical strategies of the mean. This statistical approach is most continually used for missing data. The below formula 1 is used for imputing the mean value in missed place. In formula 1, xic is the missing value in the m field. The calculated value xic is imputed for nm missing place.Figure 2 shows the before and after imputation for Salem District Crime Data-Set

$$x^{-}_{ic} = \sum_{i:xic \in nm} \frac{x_{ic}}{n_m} \qquad (1)$$

Impute the Missing Values using Mean

Algorithm _1 : Impute the Missing values by Mean

Input : Crime Dataset (CD), Missing Attribute (A)

Output : Fill the missing data with Mean Value (μ) for each Observations(O)

Step1: Initialize the CD

Step 2: Method for selecting each attribute from CD

Step3: Calculate μ for each A

$$\mu = \frac{\sum_{i=1}^{n} fixi}{\sum_{i=1}^{n} fi} \qquad (2)$$

end for

Step4: for find each attribute A

If the A is null then

Fill in the missing O by me

End if

end for

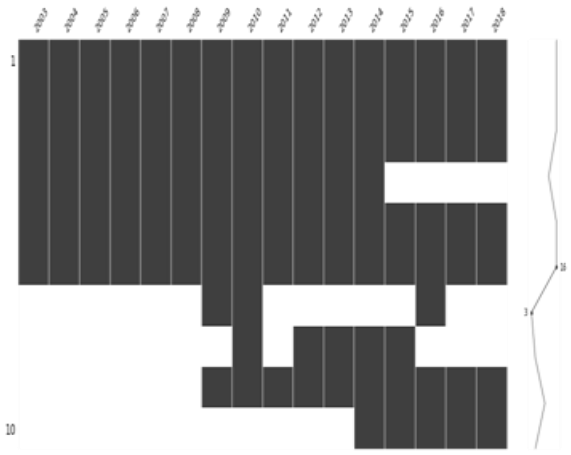


Fig: 2 (a) Before Imputation of Missing values

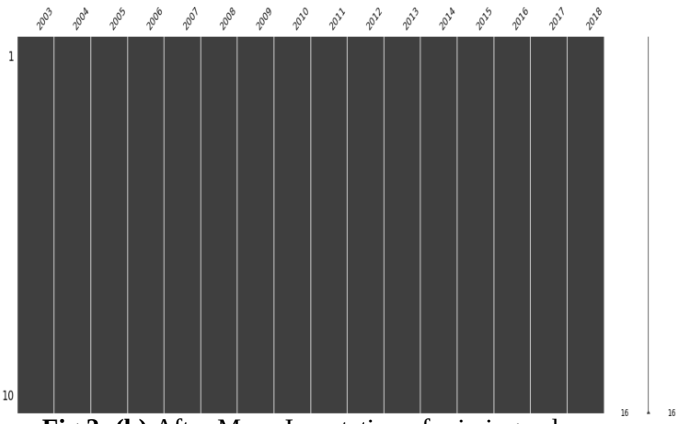


Fig 2: (b) After Mean Imputation of missing values

Impute the Missing Values using Median

Algorithm_2: Missing Data imputed by Median

Input : CAW Data with missing (Bl)

Output : Form the complete data by filling in the median (M) for each Attribute(A)

Step1: Load the B

Step 2: Function: choosing every field from B1

Step 3: find there is an odd or even no of A

Step 4: arrange all attributes from least to highest

Step 5: if the no of the attribute is odd then fix the median value using (n+1)/2

Step 6: if the attribute count is odd then

Compute the value(n/2)

Compute (n/2)+1

The Median value is the mean value of the above-computedvalues

end for

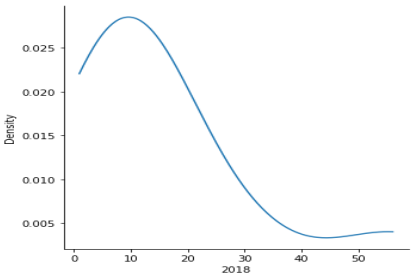
Step7: for calculate each attribute A

If the A is lost then

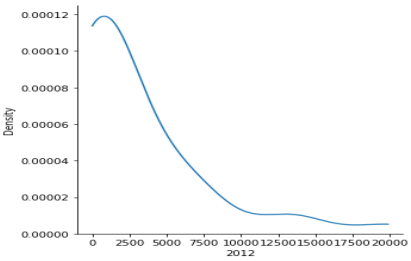
Fill the missing A by M

End if

The middle value for any group is represented by the median. At this stage, half of the data is more and half is less. The Median makes it possible to represent a lot of data points with just one. The median is the most straightforward statistical metric to compute.The number of data points also affects how to calculate the median.The median is the middlemost data when the data is odd number and for an even number of data,the average of the two middle numbers is known as the median.Figure 3 a and b shows skewness of Salem and India Crime against Women Dataset.



(a) skewness of Salem crime



(b) skewness of India crime dataset

Fig: 3

The optimal method for replacing a missing value in a dataset's done by median.When the information is skewed, it is good to use the median value for exchange the missing values.The skewness is used to calculate the symmetry of our data distribution and discover the shape of our data. The skewness value may be positive, zero, negative and undefined. The below formula 3 explains the skewness.

$$\mu_3^- = \frac{\sum_i^N (X_i - X^-)}{(N-1) * \sigma^3} \quad (3)$$

Here μ_3^- is used for skewness

N -> total number of distribution

$X_{i\ ->}$ This symbol is used for variables of random

x^- -> Mean distribution and σ denotes the standard deviation

Impute the Missing Values using Mode

Algorithm_3: Missing values handled by Mode

Input : Crime records with missed values (M1)

Output : complete the data set by using (M_0) for all Attribute (A) or Observations

Step1: Initialize M1

Step 2: Method3:

for choose each column from M1

Step3: Calculate M_0 for every A

$$M_0 = l + \left(\frac{f1-f0}{2f1-f0-f2}\right)h \qquad (4)$$

end for

Step4: for search each attribute A

If the A is empty then

Fill the missing O by M_0

End if

end for

Mode is the most frequently occurring value in a record. A dataset can have one mode, multiple modes, or no mode at all. The mean, or average of a set, and the median, or middle value in a set, are two more common measurements of central tendency.Bimodal sets of numbers have two modes, tri-modal sets of numbers have three modes, and multi-modal sets of numbers have more than one mode.Mode calculation is a fairly simple process. Arrange the numbers in the set in any order from smallest to largest or largest to smallest and count the number of times each number appears in the groups. Mode is most often displayed.The mode has some advantages

- ✓ The modes are easy to understand and easy to calculate.
- ✓ Extreme values have no effect on the mode.
- ✓ Modes are easily recognizable in both discrete frequency distributions and data sets.
- ✓ This method is useful for qualitative data.
- ✓ There is no limit to the number of frequency tables that can be used for mode calculations.

Graphical location of the mode is possible. From the formula 4

l = least value of the class

h = range of the class

f_0 = occurrences of the class priority

f_1 = occurrences of the class

f_2 = occurrences of the class ensuing

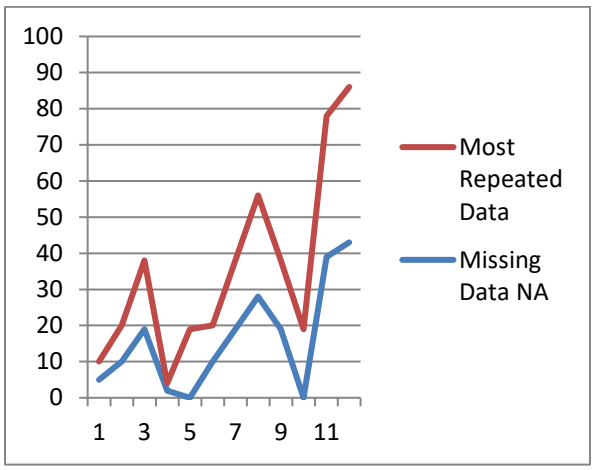
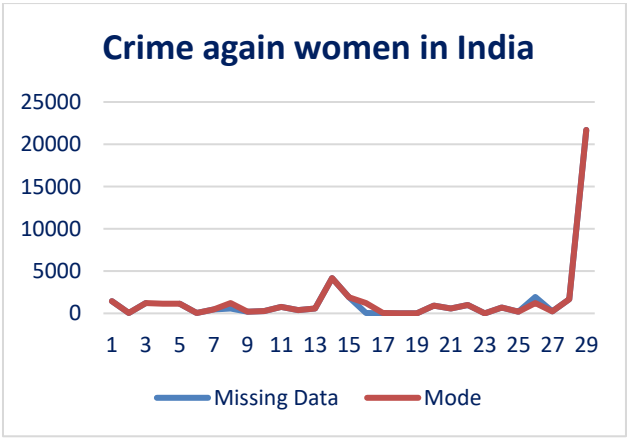


Fig: 4 (a) Comparisons between Missing Values

Comparisons between Missing Values with Most Repeated Values using Salem Crime Dataset



(b)Comparisons between Missing Values with Most Repeated Repeated Values using India Crime Dataset

Fig 4a and 4b shows ,comparison of missing values and most repeated values (Mode) of both dataset Crime against Women in India and Salem.From this fig ,the missing values counting is less compared with most repeated values. So we can use Mode values for filling the missing values.

5. Machine Learning Based Imputation techniques

Machine learning is one of the valuable techniques for handling the missing information in the record. Many algorithms are proposed to address the missing information to predict crime and estimate the algorithm based on that performance. The following algorithms are used for

imputation. This algorithm is further classified as center value and parametric values-based models

Impute the Missing Values using k-mean Cluster Algorithm _4 : Impute the Missing values by k-mean Cluster

Input : Crime Dataset (CD_1), Missing Attribute (A)

Output : Fill the missing data with centre Value (c) for each Observations(O)

{a1,a2,a3.....an } and group of centred values {c1,c2,c3.....}

Step 1: Load all Observations (O){a1,a2,a3.....}

Step 2: Determine the no of clusters K

Step 3: Initialize the cluster mid value.

Step 4: Measures the distance by using Euclidian distance formula.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(qi - pi)^2} \quad (5)$$

Step 5: Estimate mean value for every cluster

$$c_{i=\frac{1}{|si|}} \sum_{xi \in si} xi \quad (6)$$

Step 6: Repeat step- 3& 5 for assign new cluster

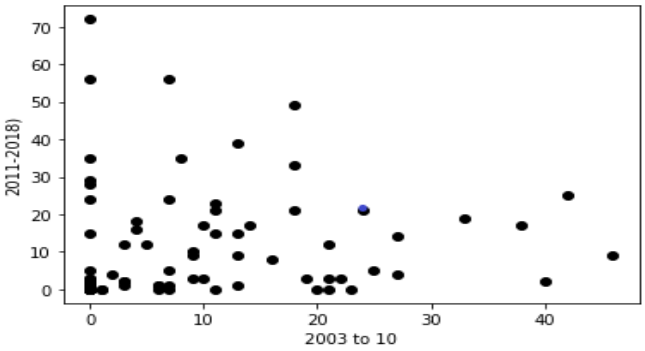
Step 7: Estimate the variance value for cluster

$$S^2 = \frac{\sum(x_i - x^-)^2}{n-1} \quad (7)$$

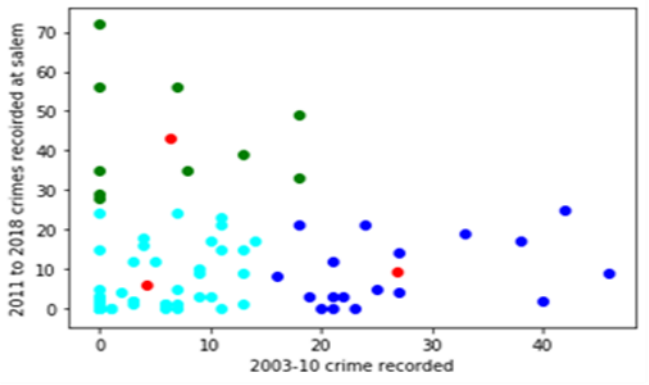
Step 8: Execute step 2& 7 until get lowest or no variance between the clusters

Step 9: Fill the missing values by cluster mean values

In formula 5, p and q are two data points in Euclidian. qi and pi is the initial point of Euclidian areas denoting the no of the space. In formula 6, si is the group of every point of the i^{th} cluster. In formula 7, s2 is the variance of all samples' This is the total number of attributes and xi denotes the single attribute. x- is the mean of all attributes. K-mean cluster is a universally recognized unsupervised machine learning technique that have a comprehensible idea in such a way it split the data position so that by way of the data point in the alike cluster are as related as feasible. Data values from different clusters are as unlike as feasible. In this algorithm, each and every cluster are symbolized by its center value. It matches up to the statistical mean of data value dispense to the cluster. Cluster center values are represented by their centroid value and this should not be a part of the data-set. This process is executed until all data values are closer to their cluster center value than another cluster-centric value, reducing the cluster distance in every process.Figure 5 a and b shows before and after Clustering using Salem District Crime Data.



(a) Before Cluster



(b) After cluster

Fig.5

Impute the Missing Values using Fuzzy c -mean

Algorithm _5:Impute the Missing values by fuzzy c-mean Cluster

Input : Crime Dataset (C_1D), Missing Attribute (A)

Output : Fill the missing data with centre Value (V) for each Observations(O)

{a1,a2,a3.....an } and group of centered values {c1,c2,c3.....}

Step 1: Execute all attributes (A) of C_1D {a1, a2,a3.....}

Step 2: Decide the no of clusters K

Step 3: Initialize the data values randomly for required no of clusters.

Step 4: Calculate the centroid value V using below formula 8.

$$V(i,j) = \sum_1^n(\gamma_{ik}^m * x_k) / \sum_1^n \gamma_{ik}^m \quad (8)$$

Step 5: Find space of all data points from centered value

Step 6: Revise the membership data by using the below formula 9.

$$\gamma = \sum_1^n(d_{ki}^2 / d_{kj}^2)^{1/1-m}]^{-1} \quad (9)$$

Step 7: Repeat step 2 to 4 until get the less variance

Step 8: De-fuzzify the received sponsorship data.

Step 9: Load the absent values by fuzzy cluster average.

The above formula 8 and 9, is a fuzzy sponsorship data and m is the membership parameter of fuzzy. is the data value. Fuzzy logic philosophy can be utilized to cluster multidimensional information, conveying every face a sponsorship in every cluster midpoint from 0 percent to 100 percent. It could be extremely influential evaluate to the usual solid threshold group where each tip is allocated a crusty, precise tag. These techniques works by handing over sponsorship to all point equivalents to each group midpoint on the source of detachment between the group middle and the data point. Additionally, the information is close to the group midpoint furthering its sponsorship on the way to the scrupulous group midpoints. The following are some advantages of using fuzzy c-mean. There are

1. It gives good accuracy for overlapped records compared to k-mean cluster algorithms

2. In this model, every data point is allocated sponsorship to every group mid-values.

Disadvantages of Fuzzy c-mean:

1. Apirior requirement of the number of clusters.

2. If the value is low, it will be good but a maximum number of iterations is needed for this.

3. Euclidean distance may calculate unequally.

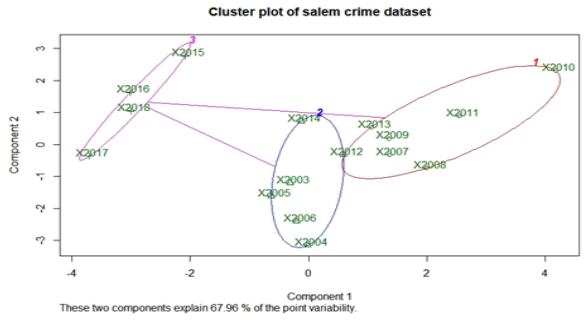


Fig 6. Fuzzy cluster plot for Salem District Crime Data-set

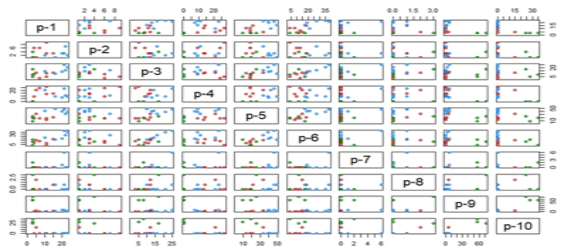


Fig 6 b shows Pearson plot for same data set.

The above figure 6 a shows fuzzy cluster plot for Salem District Crime Data-set

Algorithm _6 : Fill the absent values by K-medoids Cluster

Input : Crime _Dataset (CDS), Missing Attribute (MA)

Output : load the absent field with Mid Value (MV) for each Field (F)

{c1,c2,c3.....cn } and group of Mid values {m1,m2,m3.....mn}

Step 1: Read the Data_set (CDS) with all column {c1, c2,c3.....}

Step 2: Select the no of clusters K by using elbow formula 10

$$W_k = \sum_{r=1}^{k} \frac{1}{n_r} D_r \qquad (10)$$

Step 3: Initialize the data values randomly for required no of clusters.

Step 4: Repeat the process step2 and 3

Step 5: Assign all data value to its close centroids by using formula 11

$$c = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci| \qquad (11)$$

Step 6: Evaluate average value for all the centroid values

Step 7: Repeat the above process until the centroid place is not misshapen.

Step 8: Fill the missed values in the dataset by using Centroid mean value.

An elbow technique is implemented for finding the best possible K value. This scheme is used for changing the cluster count from 1 to ten. We calculate the WCSS value for every k.WCSS means within-cluster sum of the square. WCSS estimates the sum of squared distances between data points and cluster centroids. The WCSS value is elevated when the k value is 1. Figure 7(a) analyses, the data value certainly changed in one data point and create the elbow symbol which the value moves towards the x-axis This equivalent value of k is optimal for the cluster. Formula 10 is for the elbow which is executed for calculating the optimal k value. From this equation ,

- ✓ k symbolize the count of the clusters
- ✓ nr represent total data points,
- ✓ r &Dr signifies the cluster's distance.
- ✓ Figure 7(b) and 7(c) shows the earlier k-medoids and later
- ✓ k-medoids execution results.

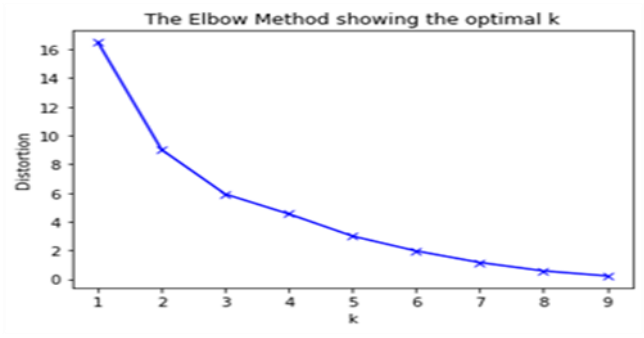
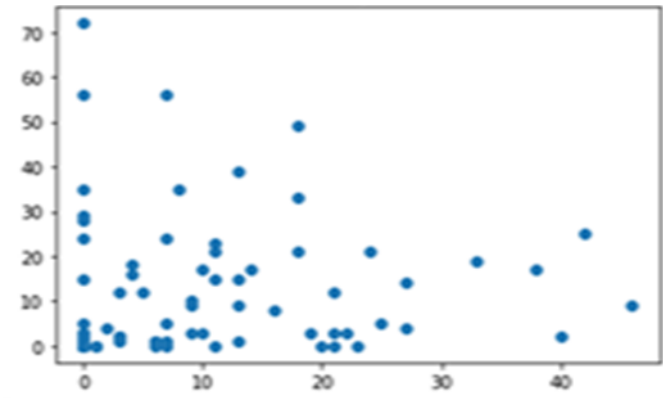
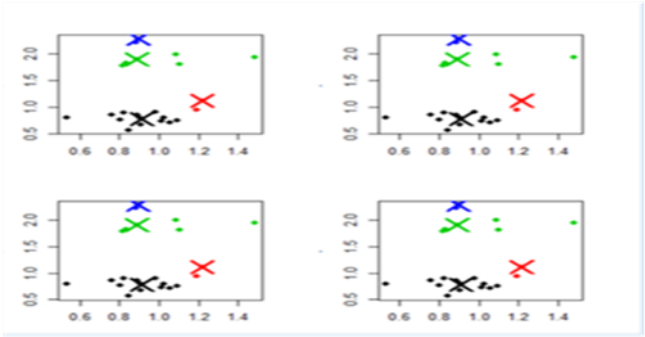


Fig 7 (a) Find the K optimized Value



(b) Before K-Medoids



(c) After K-Medoids

K-medoids is one more significant technique for clustering. This method is more related to K-mean. Each cluster is identified with a data point surrounded by the group. These data values are identified as k-medoids. The k-medoids have dissimilar optimization techniques and slightly vary from k-means. There are various types of procedures available for estimating the k-medoids but the PAM method is more capable than others. This method is as well known as PAM [Partitioning_Around_Medoids]. Formula 11 is used for calculating the center value. From this formula,it is possible to calculate the difference between object(Pi) and medoid(Ci) using the equation E=|Pi - Ci|.

Impute the Missing Values using KNN

Algorithm _7 : Missing data values filled using KNN

Input : Crime _Dataset1 (CAW_1), Missing value in Attribute (MAV)

Output : Execute missing field with KNN Value (KMV) for each column (C) {f1,f2,f3.....fn }

Step 1: Read the Data_set (CAW_1) along with attribute {c1, c2, c3.....}

Step 2: Decide the no K for the neighbour

Step3: Execute the above formula 5 for finding Euclidian distance for the neighbour

Step 4: Assign the k value, it must be estimated by Euclidian distance

Step 5: From the K neighbour, count each class data point

Step 6: Assign the new data values to that class which K_neighbour is high

Step 7: Fill that value, where the missing data occurs in the record.

Data in the record can be blank. One or more records or columns data are missing in a row. Sometimes data values are completely missing, and sometimes special characters like? are missing. Include values are missing for various reasons due to domain difficulty and reason unavailability. Machine learning procedures generally require numerical data, which can be present in every row and column of a data set. Missing values can cause problems with algorithm performance. It is very common to detect missing data and reconstruct it numerically. This is known as lacking data accusation. Using a model for accusation is a more expeditious accusation process. A model that fills in missing data with calculated values is estimated using availability data. KNNs are best known for their lack of data imputation. Missing values were filled by sampling the next data point in the training set and averaging the next value. Here the Euclidean value is estimated by calculating the distance between ignoring missing data and increasing the weight of present data. The below formula 12 and 13 is used to estimate the Euclidean region for KNN.

$$dxy = \sqrt{weight * squared\ distance\ from\ coordinates} \quad (12)$$

$$\text{Weight} = \frac{\text{Total Number of Coordinates}}{\text{Number of present coordinates}} \quad (13)$$

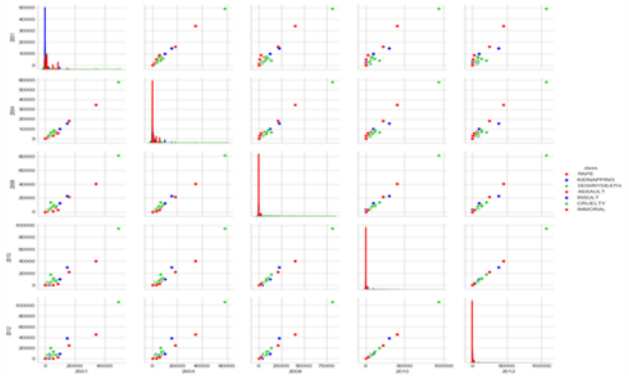


Fig 8 : (a) crime data Jitter plot for KNN

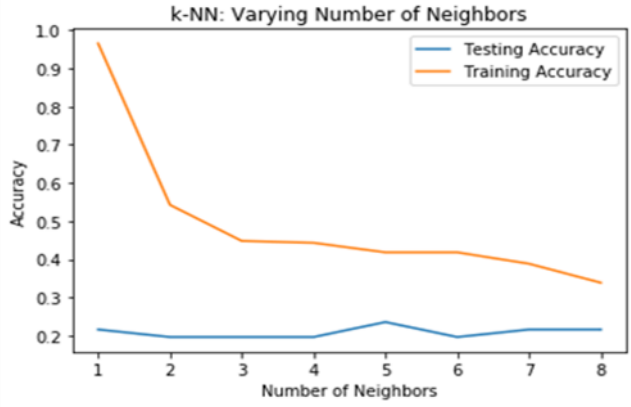


Fig 8 : (b) KNN Neighbours Comparisons in testin and training data

The figure 8 (a) shows the jitter plot for crime against women data in India and fig (b) shows comparisons of KNN neighbours in training and testing data.

Missing Data Imputation using KNN_ET values

Algorithm _8 : proposed values used to fill the missing space

Input : Crime _Dset (Cs_1), lacking field LF)

Output : perform LF with Proposed Value (PV) for ALL field (F)

{x1,x2,x3.....xn }

Step 1: Interpret the Cs_1 along with attribute {x1, x2, x3.....xn} Step 2: Find the missing place in the dataset

Step 3: Calculate Mean and standard Deviation for every field

Step 4: Find the Most Repeated value

Step 5: Enhanced step2 and step3 results with the help of logarithmic

Step 6: Finally add step 5 result with mean.

Step 7: Stuff the step6 result in the data lost place.

Imputation signifies the process of reinstate the value in missing area. When replace with a data value this is called as unit imputation. When replacing the piece of a data value that is called item imputation in statistics. In the records missing data may be ?,Null, NAN ,blank cell etc. This article proposed an enhanced statistical method for missing value.The combination of mean, standard deviation and mode are performed well in this method. There are various types of mean values available such as mean of arithmetic and geometric, median values, most repeated values are regularly used for calculating the statistical mean. Here the arithmetic mean is used for scientific field. The geometric mean is used in finance department for calculating the multifarious quantity. The median is executed in skewness data-set and mode is used in most frequent occurring dataset. Standard deviation is one of most usual techniques in statistics. It measures the diffusion records relative to that mean and evaluated as the variance of the Root Mean Squared. The Standard deviation mainly used for calculating the Root mean squared which is used to find the deviation of all data points and that data values are relative to its mean. Standard deviation is estimated by using Root mean Squared which is imitative from contrast to a accumulative mean of samples. The below formula is 14 used to calculate the standard deviation.

$$Standard\ Deviation = \sqrt{\frac{\sum_{i=1}^{n}(xi - x^{-})^2}{n-1}} \qquad (14)$$

Here, xi -> Denotes the ith position value in record

x- -> the average value for the record

n -> It mentions the number of entries in the dataset. The following steps are used to estimate the standard deviation.

Step 1: Find the average value of records. The value of the average is considered by adding all data values and that value is divided by the number of data entries

Step 2: Estimate the variance of all data. This value is considered for all data minus the average value from the number of data entries.

Step 3: From step2 get the square root for the variance.

Step 4: Using step 4 values compute the squared value for the

The degree of a data set's dispersion is measured by variance. According to mathematics, it is the average of the squared deviations from the mean.The following figure 9 shows the difference between the mean and standard deviation of the 2003 to 2018 Salem District crime data set.

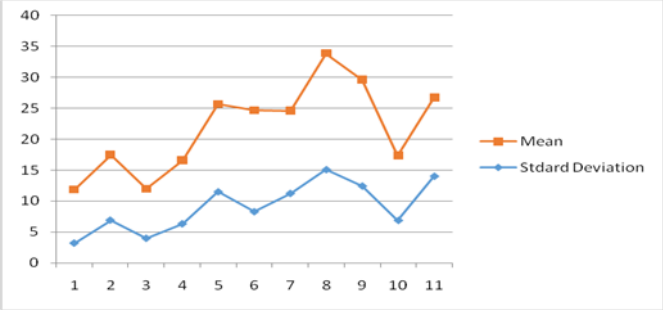


Fig 9 . Mean and standard deviation of Crime against women data at Salem

6.Experimental Results and discussions

The statistical and machine-learning-based imputation techniques are used in two different crime against women datasets among first data-set is collected from Data.gov.in and another one is collected from the Salem District commissioner office, Tamil Nadu, India.

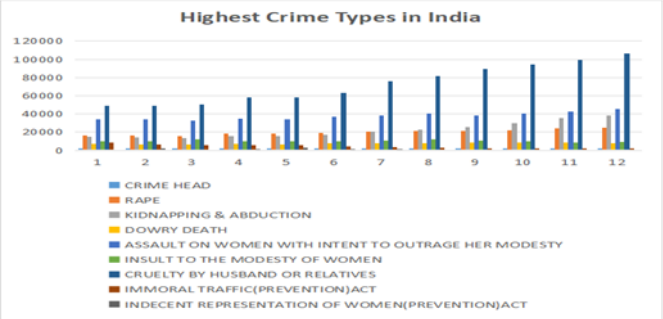


Fig. 10 .Highest crime categories in India

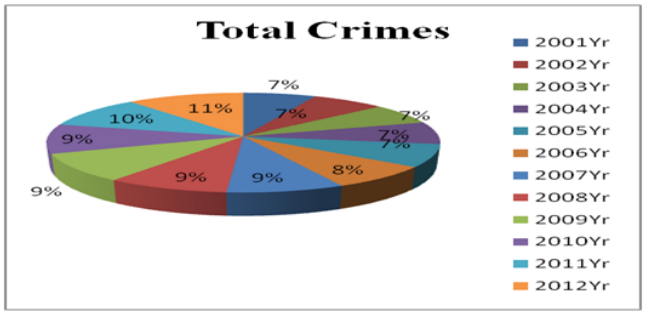


Fig 11. Highest crime recorded year wise in India

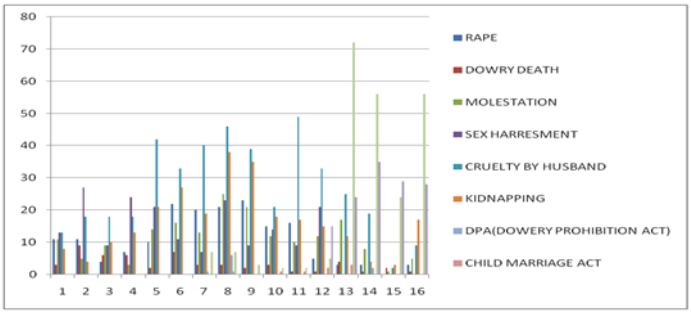


Fig 13.Year wise Highest crime recorded in Salem District

The above Fig 10,11,12,13 are visualize the Total crimes and high level registered crimes against women in India,Salem.

Table:1 comparison of year-wise Actual and Predicted crime against women data Values using KNN_ET

Data Set	S.No	Year	Actual Value	Predicted Value
Crime against Women in India	1	2003	1211	1042
	2	2007	2097	1916
	3	2010	2395	2172
	4	2012	2591	2467
Crime against Women in Salem	5	2003	13	9
	6	2008	27	17
	7	2010	46	39
	8	2015	25	19
	9	2018	56	47

The above table1 shows the comparison of actual and predicted crime rate values.

Table:2 Algorithms Prediction Evaluation for Salem District Crime data-set

Name of the Dataset used for Predictions	S.No	Imputed values for Missing Values	Imputed values	MAE	MSE	RMSE	Accuracy
Crime data in Salem district,India	1	Simple linear	Mean	5.23	41.31	6.2	66.52
			median	5.23	41.31	6.2	66.52

			Mode	5.11	40.79	6.38	68.22
			K-mean	8.68	127.23	11.62	50.81
			Fuzzy C-Mean	4.86	40.76	6.38	69.34
			K-medoids	4.05	23.29	4.82	72.79
			KNN	6.22	46.70	6..83	80.23
			KNN_ET	**2.3**	**8.60**	**2.93**	**94.78**
	2	Multiple Linear	Mean	26.2	968.40	31.43	50.21
			Median	25.2	1230.4	35.07	59.23
			Mode	24.6	1037.8	32.29	59.21
			K-mean	22.54	1256.8	30.13	40.89
			Fuzzy C-Mean	25.2	774.5	27.82	74.20
			K-medoids	26.4	1031.8	31.84	55.82
			KNN	24.2	1035.08	32.17	59.72
			KNN_ET	**22.4**	**665.6**	**25.79**	**89.12**
	3	SVR	Mean	30.1	874	30.2	62.45
			Median	34.21	823.36	31.56	62.41
			Mode	31.34	735.56	34.52	59.12
			K-mean	29.34	845.23	32.56	63.54
			Fuzzy C-Mean	29.42	745.87	28.94	64.23
			K-medoids	27.74	839.45	30.26	69.28
			KNN	30.12	783.72	35.67	70.12
			KNN_ET	**23.12**	**639.45**	**28.12**	**83.16**
	4	Decision Tree Regression	Mean	5	25	5	67
			Median	6	36	6	62
			Mode	9	81	9	60
			K-mean	19	361	19	54
			Fuzzy C-Mean	8	64	8	61
			K-medoids	5	54	9	58
			KNN	4	16	4	69
			KNN_ET	**3**	**9**	**12**	**72**
	5	KNNr	Mean	8.93	129.46	11.37	50.52
			Median	8.73	128.44	11.37	50.52
			Mode	8.73	128.44	11.33	50.34
			K-mean	8.93	129.46	11.37	53.56
			Fuzzy C-Mean	8.53	128.13	11.31	40
			K-medoids	10.46	121.10	11.00	50.96
			KNN	8.13	107.81	10.38	54.1
			KNN_ET	**6.60**	**71**	**8.42**	**84.92**
	6	Polinomial Regression	Mean	6	32	25	65
			Median	6	34	26	63

			Mode	5	25	45	62
			K-mean	5.8	23	32	43
			Fuzzy C-Mean	5.2	21	34	50
			K-medoids	4.3	43	42	53
			KNN	6.2	32	24	58
			KNN_ET	**4**	**12**	**18**	**69**
	7	Randomforest Regressor	Mean	7.6	80.4	8.96	68.23
			Median	7.26	76.6	8.75	69.56
			Mode	7.23	76.8	8.75	70.23
			K-mean	7.6	8.4	8.96	75.23
			Fuzzy C-Mean	6.93	73.23	8.57	72.90
			K-medoids	5.6	43.30	6.58	72.86
			KNN	8.13	87.86	9.37	73.93
			KNN_ET	6.6	71	8.42	69

The table 2 shows the comparisons of machine Learning algorithms performance using various regression Techniques

8. Conclusions

Data Analytics is one of the most familiar research areas. In this field, Missing data handling is a very challenging one. This type of data will reduce the Efficiencies of Machine Learning algorithms. To overcome this problem, various kinds of statistical and Machine Learning Imputation methods such as Mean, Median, Mode, K-Mean, Fuzzy C-means, K-Medoids, KNN, and newly proposed KNN_ET values are used for filling the missing data in crime against women dataset. Crime against women is one of the largest difficulties in our nation. To control this offense, Prediction is essential. In this work, two different crime-against-women-datasets are used. From this, the first dataset is collected from the Commissioner's office, Salem District, Tamil Nadu, India, and the second dataset is collected from Data.gov.in.This both Datasets have various crime types. To predict, various types of regression techniques are implemented such as Simple linear, Multiple Linear, SVR, Decision Tree Regression, Polynomial Regression, Random forest Regression, and KNNr. KNN_ET is a proposed method, It is used to fill the missing values.This imputed value along with Simple linear regression has given 94.78% accuracy for the prediction of crime against dataset in Salem District and 98.7 % accuracy for predict the Crime against women in India Dataset with fewer errors. The Algorithm performances are measured based on the statistical methods MAE, MSE, and RMSE.This gives better accuracy compared with all other Machine-learning algorithms. In future, this result will be very helpful to the crime department for control the crime against women in India.

References

[1] Liu, C.-H., Tsai, C.-F., Sue, K.-L., & Huang, M.-W. (2020). The Feature Selection Effect on Missing Value Imputation of Medical Datasets. Applied Sciences, 10(7), 2344. doi:10.3390/app10072344.

[2] Waseem Shahzad, Qamar Rehman, Ejaz Ahmed (2017) Missing Data Imputation using Genetic Algorithm for Supervised Learning, International Journal of Advanced Computer Science and Applications,Vol. 8, No. 3, 2017,PP 438-445.

[3] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big Data, 8(1), 1-37. https://doi.org/10.1186/s40537-021-00516-9

[4] Dimitris Bertsimas,Colin Pawlowski,Ying Daisy Zhuo(2018)From Predictive Methods to Missing Data Imputation: An Optimization Approach,Journal of Machine Learning Research,pp 1-39.

[5] KohbalanMoorthy, Mohammed HasanAli, MohdArfianIsmail, Chan Weng Howe, MohdSaber Mohamad, SafaaiDeris(2019),An Evaluation of Machine Learning Algorithms for Missing Values Imputation,Internationa Journal of Innovative Technology and Exploring Engineering (IJITEE) ,Volume-8 Issue-12S2,PP 415-420

[6] Wang, H., Tang, J., Wu, M. *et al.* Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Med Inform Decis Mak* **22**, 13

(2022). https://doi.org/10.1186/s12911-022-01752-6.

[7] Richman, M.B., Trafalis, T.B., Adrianto, I. (2009).Missing Data Imputation Through Machine Learning Algorithms.In: Haupt, S.E., Pasini, A., Marzban, C. (eds) Artificial Intelligence Methods in the Environmental Sciences. Springer,Dordrecht. https://doi.org/10.1007/978-1-4020-9119-3_7.

[8] C.G Marcelino, G. M. C. Leite, P. Celes & C. E. Pedreira(2022)Missing Data Analysis in egression,AppliedArtificial Intelligence, 36:1, D.

[9] DOI: 10.1080/08839514.2022.2032925

[10] Iwueze, E. Nwogu, O. Johnson and J. Ajaraogu, "Uses of the Buys-Ballot Table in Time Series Analysis," Applied Mathematics, Vol. 2 No. 5, 2011, pp. 633-645. doi: 10.4236/am.2011.25084.

[11] Tseng, S., Wang, K., & Lee, C. (2003). A pre-processing method to deal with missing values by integrating clustering and regression techniques. Applied Artificial Intelligence, 17(5-6), 535–544. doi:10.1080/713827170

[12] Gad, I., Hosahalli, D., Manjunatha, B. R., & Ghoneim, O. A. (2020). A robust deep learning model for missing value imputation in big NCDC dataset. Iran Journal of Computer Science. doi:10.1007/s42044-020-00065-z

[13]]Nwosu, Ugochinyere & Obite, Chukwudi. (2020). Methods for Estimating Missing Values in Descriptive Time Series Statistics: Novelty and Efficiency under Buys-Ballot. 0.18488/journal.24.2020.91.72.80.

[14] Amjad Ali* and Qamruz Zaman(2019)A New Method of Imputation for the Missing Value in the IN/OUT Procedure of the Random Forest (RF),Indian Journal o Science and Technolog,Year: 2019, Volume: 12, Issue: 14, Pages: 1-11,DOI: 10.17485/ijst/2019/v12i14/141616.

[15] Kritbodin Phiwhorm1 , Charnnarong Saikaew2 , Carson K. Leung3 , Pattarawit Polpinit1 and Kanda Runapongsa Saikaew1* (2022),Adaptive multiple imputations of missing values using the class center, Journal of Big Data,https://doi.org/10.1186/s40537-022-00608-0.

[16] Cheng, C.-Y., Tseng, W.-L., Chang, C.-F., Chang, C.-H., & Gau, S.S.-F. (2020). *A Deep Learning* Approach for Missing Data Imputation of Rating Scales Assessing Attention-Deficit Hyperactivity Disorder. Frontiers in Psychiatry, 11. doi:10.3389/fpsyt.2020.00673

[17] Li, Huaxiong. (2013). Missing Values Imputation Based on Iterative Learning. International Journal of Intelligence Science. 03. 50-55. 10.4236/ijis.2013.31A006.

[18] Cheng, C.-Y., Tseng, W.-L., Chang, C.-F., Chang, C.-H., & Gau,S.S.-F (2020). A Deep Learning Approach for Missing Data Imputation of Rating Scales AssessinAttention-Deficit Hyperactivity Disorder. Frontiers in Psychiatry, 11. doi:10.3389/fpsyt.2020.00673.

[19] Raja, P.S., Thangavel, K. Missing value imputation using unsupervised machine learning techniques. Soft Comput 24, 4361–4392 (2020). https://doi.org/10.1007/s00500-019-04199-6.

[20] Gopal Krishna M, Durgaprasad N, Deepa Kanmani S, Sravan Reddy G, Revanth Reddy D(2019),Comparative Analysis Of Different Imputation Techniques For Handling Missing Dataset,International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-7, May, 2019,PP347-351.

[21] Fouad KM, Ismail MM, Azar AT, Arafa MM. Advanced methods for missing values imputation based on similarity learning. PeerJ Comput Sci. 2021 Jul 21;7:e619. doi: 10.7717/peerj-cs.619. PMID: 34395861; PMCID: PMC8323724.