Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 10, No. 2, June 2022, pp. 375~384
ISSN: 2089-3272, DOI: 10.52549/ijeei.v10i2.3730

The Effect of Using Data Pre-Processing by Imputations in Handling Missing Values

Abdelrahman Elsharif Karrar
College of Computer Science and Engineering, Taibah University, Saudi Arabia

Article Info

Article history:

Received Feb 19, 2022
Revised Apr 4, 2022
Accepted Apr 8, 2022

Keyword:

Data Pre-Processing
Imputation Model
Machine Learning
k-Nearest Neighbor Algorithm
Missing Values

ABSTRACT

The evolution of big data analytics through machine learning and artificial intelligence techniques has caused organizations in a wide range of sectors including health, manufacturing, e-commerce, governance, and social welfare to realize the value of massive volumes of data accumulating on web-based repositories daily. This has led to the adoption of data-driven decision models; for example, through sentiment analysis in marketing where produces leverage customer feedback and reviews to develop customer-oriented products. However, the data generated in real-world activities is subject to errors resulting from inaccurate measurements or fault input devices, which may result in the loss of some values. Missing attribute/variable values make data unsuitable for decision analytics due to noises and inconsistencies that create bias. The objective of this paper was to explore the problem of missing data and develop an advanced imputation model based on Machine Learning and implemented on K-Nearest Neighbor (KNN) algorithm in R programming language as an approach to handle missing values. The methodology used in this paper relied on the applying advanced machine learning algorithms with high-level accuracy in pattern detection and predictive analytics on the existing imputation techniques, which handle missing values by random replacement or deletion.. According to the results, advanced imputation technique based on machine learning models replaced missing values from a dataset with 89.5% accuracy. The experimental results showed that pre-processing by imputation delivers high-level performance efficiency in handling missing data values. These findings are consistent with the key idea of paper, which is to explore alternative imputation techniques for handling missing values to improve the accuracy and reliability of decision insights extracted from datasets.

Corresponding Author:

Abdelrahman Elsharif Karrar
College of Computer Science and Engineering, Taibah University, Saudi Arabia
Email: akarrar@taibahu.edu.sa

1. INTRODUCTION

Technological growth over the recent years has caused unprecedented growth in information and communication capabilities and mobility, especially in the field of Artificial Intelligence, which enables smart devices to perform automated functions even beyond human ability. Computer systems embedded with advanced capabilities such as automation and predictive analytics have been adopted in a wide range of areas such as transportation, manufacturing, healthcare, marketing, finance, and robotics. These applications generate huge quantities of data, which presents significant growth opportunities in the field of Machine Learning [1]. Machine learning is a subset of artificial intelligence in which specialized algorithms are designed to analyze and extract insights from unstructured datasets. Machine learning utilizes mathematical models to discover trends and make data-driven predictions concerning the variables under study. According to [2], predictive modelling refers to the formulation of computerized algorithms to perform data analytics operations and make forecasts about a problem by regression or classification.

The data generated from the devices and applications with AI and ML capabilities are stored in dynamic servers and databases, which require routine updating and upgrading due to the complex nature of the data and sources. Unstructured data from certain sources are subject to errors due to the possibility of storing null values, which are also analyzed during decision-making processes. Data may also contain inconsistencies and discrepancies, and noises, which may lead to unreliable output. Therefore, data cleansing is an essential aspect of database management to minimize the risks of bias resulting from the storage inconsistencies, especially for large data sets [3], [4]. In the process of data analytics by machine learning, shortage of data is the most prevalent challenge faced in developing predictive models.

Mean imputation is one of the most efficient methods for addressing the problem of missing data values in advanced analytics systems by reducing the strength of association while the imputation of regression can increase the strength of association although it poses ambiguity risks to the data values. Missing Completely at Random (MCAR) technique based on the K-Nearest Neighbors (K-NN) algorithm [5] is generally applied to the replacement of missing data values by increasing the likelihood of interpretative and incomplete failure. Imputation techniques correct data inconsistencies by replacing the missing values with mean observed values or the last observed value. The relative advantage of imputation based on ML algorithms over other existing methods for handling missing values is [6] that the replacement values rely on a combination of computational, mathematical, and statistical models rather than random parameters from the dataset. This makes it comparatively a more reliable, accurate, and applicable approach for data-driven decision-making, especially in high-precision fields such as manufacturing and healthcare.

1.1. Problem Statement

Structured and unstructured data collected from digital devices and computer systems in their real-world applications may have missing values, noise, and other inconsistencies. Subjecting such datasets to analytics increases the risks of inaccurate output leading to biased decisions and erroneous insights, which may have adverse negative implications on the enterprise. The existing techniques of handling missing values by imputation rely on random sampling technique to estimate and replace the missing values; making then potentially inaccurate and unreliable, especially when applied to data-driven decision-making in high-precision fields such as medical surgeries and manufacturing. Therefore, there is a need to adopt effective approaches based on computational, mathematical, and statistical models to address the problem of missing values in large datasets. This paper sought to apply the imputation technique to pre-process and classify data with missing values to reduce the risks of inconstancy or errors in decision-making processes.

1.2. Data Mining

Data mining [7] refers to a process in which valuable information is extracted from unstructured or structured datasets stored in databases or cloud platforms. According to [8]. data mining entails extraction, classification, and transfer of various data through a series of processes including data cleaning, standardization, and testing. The process of data mining entails a series of sequences, which include the following steps shown in Figure 1;

1. Cleaning the data to remove inconsistencies and noise.
2. Integration to separate data from overlapping sources.
3. Selecting and retrieving data sets that are appropriate for analysis from the database.
4. Extracting useful data patterns from the database.
5. Identifying and evaluating the variable patterns representing knowledge based on the set parameters.
6. Preprocessing and representing knowledge through visualization and presentation techniques such as schemas [9].

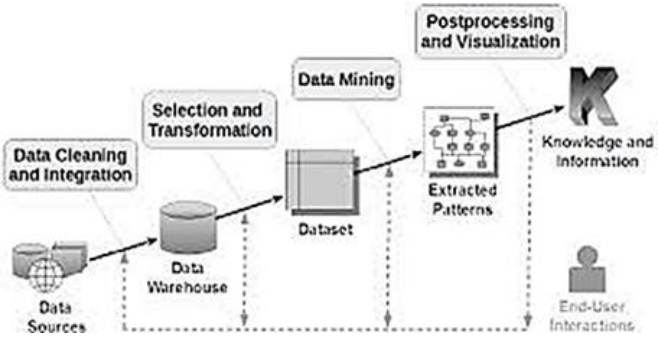

Figure 1. Steps Involved in the Process of Data Mining

1.3. Pre-processing

Data pre-processing [10] entails a series of preparations aimed at converting the data into a format that is easier to analyze since most of the information collected from day-to-say activities is largely unstructured and may be difficult to analyze due to missing values [11]. Pre-processing raw data involves various processes including cleansing, integration, transformation, and reduction as illustrated in Figure 2;

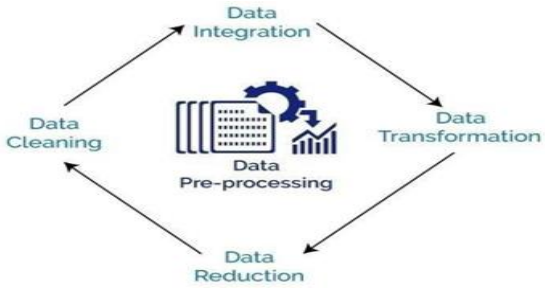

Figure 2. Stages of Data Pre-processing

The objective of data cleaning is to identify and complete null values, fixing incoherence, and standardizing outliers to reduce noise. When the data sets are imported from storage resources such as servers and databases, the first step of data cleaning [12] is merging the datasets with related information, which are then rebuilt to recreate missing values and de-duplicated to normalize the sets. The next step is data verification and enrichment, which entail confirming the validity and making further improvements to prepare the data for analysis and decision making.

1.4. Missing Values

Missing Values (MVs) in machine learning refers to the data attributes that may lack in a dataset due to errors that may arise in the input process due to improper measurements or device failure [13]. A missing value algorithm is used to determine whether a correlational link exists between the missing values and other variables in a dataset. For instance, if X represents a dataset (a, b) such that a is X's observed value while b is the missing value in X and Y is a random variable. Assuming that Y = 1 regardless of whether X's are observed or missing values, the observed value can be determined by letting X = 0 expressed in the form of a model P (Y/x, Ø) such that Ø represents the missing. The mechanism for fixing the missing values is based on Y's dependence on the variables contained in the dataset [14].

1.4.1. Mechanisms for Computing Missing Values

Missing Completely at Random (MCAR) Technique

MCAR technique determines the missing values through the (1);

$$P(Y|X,\emptyset) = (Y,\emptyset) \quad (1)$$

From (1), it is evident that the probability of missing values being proximate to the observed values is low and does not depend on the variables contained in the dataset X. Therefore, the most feasible option for dealing with missing values based on the MCAR approach is developing an algorithm that can randomly delete values to normalize the dataset.

Missing at Random (MAR)

The MAR technique is used to fix the problem of missing data based on the (2);

$$P(Y|X,\emptyset) = P(Ya|\emptyset) \quad (2)$$

From (2), the probability that the value of missing variables depends on the values noticed in a subset of the X and not entire dataset Y. This implies that the MAR approach conducts a systematic inquiry of the larger dataset to determine the values that do not correspond to the variable under inquiry hence determines the missing values based on covariance.

Missing Not at Random (MNAR)

The MNAR technique formulates a solution for missing values in a dataset based on the (3):

$$P(Y|X,\emptyset) = P(Y|a,b,\emptyset) \quad (3)$$

From (3), it is observed that the prerequisites for handling the missing values based on the MAR technique are violated such that the probability of missing values is dependent on b or other unexpected covariates within the dataset. For instance, when applied to tax computations, MNAR is able to determine that missing data values are dependent on unobserved revenue declarations by the tax payers [15]. The advantage of MNAR technique is that it separates data, which was never provided from that which was incorrectly input due to measurement error.

1.5. Imputation as a Solution for Missing Data Values

Imputation is a computational technique that utilizes mathematical and statistical algorithms to resolve the problem of missing values in a dataset by replacement without interfering with the attributes and values of the entire dataset. Imputation approaches are broadly classified into traditional and advanced categories depending on the applied method for replacing the missing values. Under the traditional imputation technique [16], the problem is solved through pairwise and listwise deletion of the missing values, especially under the condition of Missing Completely at Random (MCAR). Traditional imputation entails computational procedures such as Multiple Imputation, Maximum Likelihood Imputation, Hot Deck Imputation, Mode Imputation, and Mean Imputation. However, the advanced imputation approach relies on computational intelligence to learn complex interdependencies in large data sets and determine the optimal method for handling missing values based on the observed characteristics. Computational models such as Decision Trees, Random Forest Algorithm, and k-Nearest Neighbor are implemented in advanced imputation.

The k-Nearest Neighbors (k-NN) algorithm, which is among the most efficient technique for advanced imputation conceptualizes missing data values as regression and pattern recognition problems. K-NN algorithms classify datasets based on the memory of the observed values and attributes rather than labeled vectors. The computational processes used to replace the missing values are based on the observed closest k-neighbor in the training sets and the highest number of k iterations. The study [17] suggest that the individual variable classifications define the procedures used to determine the K-Nearest Neighbor in each instance. K-NN Imputation [18]is applied in the case of incomplete system and unknown data distribution. The technique imputes missing data values by computing a metric/variable that is distant from the nearest k neighbors based on computed estimations of data sets that lack the appropriate mean and mode. The numerical value of the missing parameters is then predicted using the mean rule while the missing categorical variables are computed using the mode rule hence making it possible to efficiently handle the problem of missing values in large data sets.

This paper is organized as follows: Section 2 briefly introduces related work. Section 3 clarifies on the proposed methodology for implementing an imputation technique based on R-programming language as a solution for handling missing values in data sets. Section 4 of this research paper discusses results based on the experimental works. Then in the subsequent sections, the conclusion comes in Section 5 and Finally, Section 6 provides recommendations for future studies.

2. RELATED WORK

A study [19] proposed a novel K-NN model for handling missing values through two-stage training scheme based on the training data and missing date. This approach effectively handles missing instances in heterogeneous data sets by computing Mutual Information (MI) weights between class labels and attributes of the dataset. This ensures the imputation of missing data values to enhance classification performance, especially in UCI data samples with varying rates of missing values. The performance efficiency of handling missing values in biased datasets can be determined by successive simulations of continuous traits and segmenting response variables through imputation procedures such as complete case analysis [14]. The model performance is measured based on the degree of deviation/marginal error and the covariance between traits and responses.

Findings from a research study [20]suggests that the adoption of advanced techniques for handling missing values delivers high-level performance by allowing multiple imputations on a single dataset. In a study, 20 samples were randomly selected from a dataset of Traumatic Brain Injury data sets. In 8 of the samples, one variable was deleted to create missing data, which was used to determine the technical performance efficiency of multiple imputation and single imputation methods [21]. The Multiple Imputation approach demonstrated higher effectiveness in variable deletion compared to single imputation based on the estimated parameter comparison [21].

According to [9], auto-encoder neural networks can be applied to the imputation of missing data to innovatively predict missing values and automatically encode new files without missing values through a two stage model. A more advanced imputation framework capable of imputing categorical and mixed continuous variables through formal optimization, predicts missing values using mathematical models such as decision trees, support vector machines, and closest k-neighbors [22]. The implementation of opti-impute generic

algorithm in this framework produces high-quality solutions due to the improved sampling accuracy in multiple datasets obtained from the UCIML repository. This approach performs precise imputations of missing value sets through predictive K-NN , mean-matching, and Bayesian techniques with low average absolute error through cross-validated benchmarking [23].

According to [24], medical datasets with missing values may be difficult to impute as the null values are often contained in categorical attributes hence complicating pre-processing stages. However, advanced imputation techniques based on machine learning and decision-tree models are capable of effectively identifying outliers and replacing the missing values through K-NN computations [25]. Outliers pose significant risks of bias during statistical estimation procedures by increasing the likelihood of overstated or understated decision outcomes hence the dependability of imputations techniques is a critical consideration.

According to [26] pattern creedal classification models may be applied to the adaptive imputation of missing data based on the observed variables. This technique is founded on the belief function theory, which implies that missing data is fundamentally required for unambiguous and accurate classification of the observed datasets hence allowing the imputation though the self-organizing map (SOM) algorithms for pattern extraction. The algorithm classifies data patterns as either altered or original depending on the representation outcomes from the training classes.

3. METHODOLOGY

The objective of this section is to handle missing data values through the implementation of imputation techniques based on IBK algorithms. The implementation entails manipulating unstructured datasets and testing the performance efficiency and accuracy of imputation techniques in replacing the missing values.

3.1. Dataset

A sample dataset obtained from Juba Insurance & Reinsurance Company was used in this project. An arbitrary selection of 100 samples were selected and prepared for imputation using MAR, NMAR, and MCAR techniques to insert fictitious missing values as the table shown in Figure 4. The original dataset was then implemented on a random data generation algorithm to model missing values into 4 categories with a DocType attributes having an equal probability of missing rates. The datasets were classified into 3 categories of percentage missing rates (3%, 6%, and 10%) hence allowing the simulation of missing values from attribute instances containing 4 classes as the table shown in Figure 3.

	A	B	C	D	E	F	G
1	Id	DocumentClass	PayMoney	PayDate	IsalNo	PayType	DocType
2	95	2	700	19/04/2013	55455	check	Third part
3	96	2	1000	19/04/2013	44300	check	Third part
4	97	2	1000	19/04/2013	14476	check	Third part
5	98	2	900	19/04/2013	20058	check	Third part
6	99	2	1240	19/04/2013	12397	check	Third part
7	100	2	1140	19/04/2013	12483	check	Third part
8	101	2	1140	19/04/2013	10001	check	Third part
9	102	2	2180	19/04/2013	25254	check	Third part
10	103	2	940	19/04/2013	55555	check	Third part
11	104	2	840	19/04/2013	11426	check	Third part
12	105	2	780	19/04/2013	55367	check	Third part
13	106	2	680	19/04/2013	22299	check	Third part
14	107	2	580	19/04/2013	44000	check	Third part
15	108	1	623	19/04/2013	33200	check+cash	comprehensive
16	109	2	1000	19/04/2013	12224	check	comprehensive
17	110	1	1000	19/04/2013	12154	check+cash	Third part
18	111	1	900	19/04/2013	23782	check+cash	Third part
19	112	1	800	19/04/2013	22665	check+cash	Third part
20	113	1	1240	19/04/2013	40201	check+cash	Third part
21	114	1	1140	19/04/2013	11001	check+cash	Third part
22	115	1	735	19/04/2013	11520	check+cash	marine
23	116	1	500	19/04/2013	15211	check	comprehensive
24	117	1	100	19/04/2013	55000	check	comprehensive
25	118	1	100	19/04/2013	15270	check	comprehensive
26	119	1	1237	19/04/2013	11411	check	comprehensive
27	120	1	100	19/04/2013	54228	check+cash	marine
28	121	1	200	19/04/2013	77711	check+cash	marine
29	122	1	1837	19/04/2013	44789	check+cash	comprehensive
30	123	1	500	19/04/2013	12582	check	comprehensive

all_enc100

Figure 3. The Original Dataset

	A	B	C	D	E	F	G
1	Id	Documentclass	PayMoney	PayDate	IsalNo	PayType	DocType
2	93	2	700	19/04/2013	55455	check+cash	Third part
3	96	2	1000	19/04/2013	44300	check	Third part
4	97	2	1000	19/04/2013	14476	check	Third part
5	98	2	900	19/04/2013	20058	check	NA
6	99	2	1240	19/04/2013	12397	check	Third part
7	100	2	1140	19/04/2013	12483	check	Third part
8	101	2	1140	19/04/2013	10001	check	Third part
9	102	2	2180	19/04/2013	25254	check	Third part
10	103	2	940	19/04/2013	55555	check	NA
11	104	2	840	19/04/2013	11426	check	Third part
12	105	2	780	19/04/2013	55367	check	Third part
13	106	2	680	19/04/2013	22299	check	Third part
14	107	2	580	19/04/2013	44000	check	Third part
15	108	1	623	19/04/2013	33200	check+cash	comprehensive
16	109	2	1000	19/04/2013	12224	check	NA
17	110	1	1000	19/04/2013	12154	check+cash	Third part
18	111	1	900	19/04/2013	23782	check+cash	Third part
19	112	1	800	19/04/2013	22665	check+cash	NA
20	113	1	1240	19/04/2013	40201	check+cash	Third part
21	114	1	1140	19/04/2013	11001	check+cash	Third part
22	115	1	735	19/04/2013	11520	check+cash	marine
23	116	1	500	19/04/2013	15211	check	comprehensive
24	117	1	100	19/04/2013	55000	check	comprehensive
25	118	1	100	19/04/2013	15270	check	NA
26	119	1	1237	19/04/2013	11411	check	comprehensive
27	120	1	100	19/04/2013	54228	check+cash	NA
28	121	1	200	19/04/2013	77711	check+cash	marine
29	122	1	1837	19/04/2013	44789	check+cash	comprehensive
30	123	1	500	19/04/2013	12582	check	comprehensive

missing_enc_new100

Figure 4. Dataset with Missing Values

3.2. Imputation of the k-Nearest Neighbors (k-NN)

The next step is implementing the k-Nearest Neighbors algorithm to impute the missing data values through pattern recognition after which non-parametric regression and classification are performed to pre-process the dataset. The output contains data values classified into k-NN classes based on the plurality of the neighbors in which objects allocated to the closest neighbors. Implementing algorithms such as the Neighborhood Components and Large Margin k-NNs allocates missing data values based on the closest neighbor classes [27]. Since imputation methods can be used to improve the classification performance of k-Nearest Neighbors, the ideal value (k) of missing data is dependent on the larger values of k, which minimize the impact of noises on classification accuracy by creating less distinctive class boundaries. Therefore, the optimal value of k in each case can be predicted based on the estimates from the nearest training sets (i.e. assuming k = 1). After the imputation of missing values, a R-programming model was developed to perform further classification of the missing values through the following steps as illustrated in Figure 5;

Step 1: Organize the data into rows and columns representing records and attributes respectively
Step 2: Segment the dataset into two classes; 'complete' and 'with MVs'
Step 3: Carry out normalization on the dataset
Step 4: Iteratively perform the imputation on the missing values independently
Step 5: Use K-NN algorithm to test variations between the new and original records
Step 6: Replace the missing values with the nearest attribute having the highest similarity
Step 7: Create a column with complete data values
Step 8: Apply the above procedure to impute missing values for all the records.

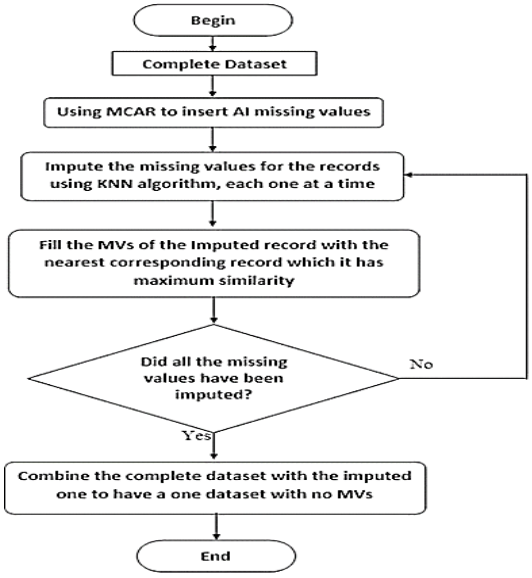

Figure 5. A Flowchart of the k-NN Imputation Procedure

The above proposed imputation technique utilizing k-NN algorithm is based on the modification of power distance parameters to determine the appropriate values of the missing data hence the need to specify the attributes with null values and their nearest neighbors.

3.3. Programming Languages

The implementation of imputation technique for missing values in data sets requires Java and R programming languages, which are used for statistical modelling and mathematical computation of relational patterns among the variables.

3.4. Experimental Procedures

The first step of implementing the R model for imputation is preparing the work directory to read and save the data file through the following commands;

```
setwd("~/R implementation")
ins <- read.csv("missing enc.csv")
```

The original dataset contains 100 instances, 7 attributes, and the null values shown in Figure 6. Since DocType is recorded as an incomplete parameter with randomly missing values, the following r-code is run to restructure the dataset;

```
> str(ins)
'data.frame':   100 obs. of  7 variables:
 $ Id           : int  93 96 97 98 99 100 101 102 103 104 ...
 $ Documentclass: int  2 2 2 2 2 2 2 2 2 2 ...
 $ PayMoney     : int  700 1000 1000 900 1240 1140 1140 2180 940 840 ...
 $ PayDate      : chr  "19/04/2013" "19/04/2013" "19/04/2013" "19/04/2013" ...
 $ IsalNo       : int  55455 44300 14476 20058 12397 12483 10001 25254 55555 11426 ...
 $ PayType      : chr  "check+cash" "check" "check" "check" ...
 $ DocType      : chr  "Third part" "Third part" "Third part" "" ...
```

Figure 6. Original Dataset Structure

The original dataset is then read and the NA value of DocType parameter is set to allow for the detection of null values using the R-code as the table shown in Figure 7;

```
> ins %>% replace_with_na(replace = list(DocType = ""))
    Id Documentclass PayMoney    PayDate IsalNo       PayType       DocType
1   93             2      700 19/04/2013  55455   check+cash    Third part
2   96             2     1000 19/04/2013  44300         check    Third part
3   97             2     1000 19/04/2013  14476         check    Third part
4   98             2      900 19/04/2013  20058         check          <NA>
5   99             2     1240 19/04/2013  12397         check    Third part
6  100             2     1140 19/04/2013  12483         check    Third part
7  101             2     1140 19/04/2013  10001         check    Third part
8  102             2     2180 19/04/2013  25254         check    Third part
9  103             2      940 19/04/2013  55555         check          <NA>
10 104             2      840 19/04/2013  11426         check    Third part
11 105             2      780 19/04/2013  55367         check    Third part
12 106             2      680 19/04/2013  22299         check    Third part
13 107             2      580 19/04/2013  44000         check    Third part
14 108             1      623 19/04/2013  33200   check+cash comprehensive
15 109             2     1000 19/04/2013  12224         check          <NA>
16 110             1     1000 19/04/2013  12154   check+cash    Third part
```

Figure 7. The Dataset After Replacing Missing values with NA

A user-defined function is implemented to analyze missing values in the incomplete dataset using the R-code shown in Figure 8;

```
> percm <- function(x)
+ {
+   sum(is.na(x))/length(x)*100
+ }
```

Figure 8. The R-code

The R output for this command is shown in Figure 9;

```
> apply(ins,2,percm)
         Id Documentclass     PayMoney      PayDate      IsalNo      PayType      DocType
          0             0            0            0           0            0           19
```

Figure 9. The R Output

Further analysis of the incomplete dataset using R-packages for pattern identification produced the following output shown in Figure 10;

```
> md.pattern(ins, plot = TRUE)
   Id Documentclass PayMoney PayDate IsalNo PayType DocType
81  1             1        1       1      1       1       1  0
19  1             1        1       1      1       1       0  1
    0             0        0       0      0       0      19 19
```

Figure 10. Further Analysis Using R-packages for Pattern Identification

From the Figure 10, it is observed that the observed and missing values are represented as binary values 1 and 0 respectively. The number of observations made from the data file is observed in the first column while the total of variables with incomplete data is shown in the last column. A plot of the missing and complete data values is shown in Figure 11;

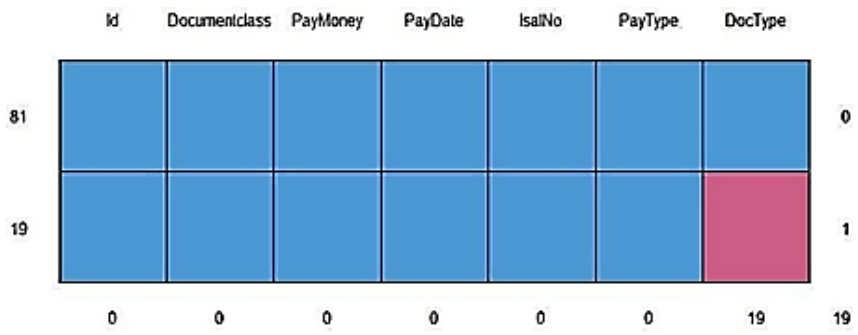

Figure 11. Complete and Incomplete Datasets

From the graph, it is observed that the initial output shows 81 samples with no missing values and 19 samples with missing values, which are further analyzed using the aggregate plot function in R as shown in the Histogram (Figure 12).

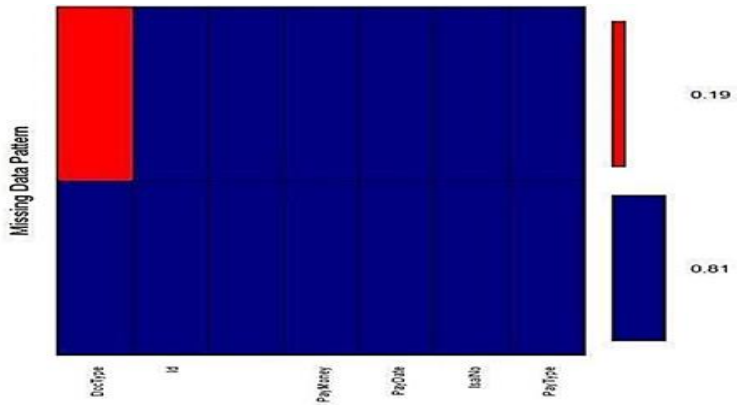

Figure 12. Complete and Incomplete Datasets

From the histogram, it is observed that the dataset contained 19% missing values (shown in Red) and 81% complete values (shown in Blue) from the DocType attribute.

4. RESULTS AND DISCUSSION

Imputation technique based on the k-NN algorithm and IBK implementation on R programming language can effectively compute missing values in a dataset. The proposed technique applied multiple computational and statistical methods on the imputation approach for handling missing values that had been artificially created in a sample dataset. An insurance dataset was obtained and input into R-studio with variables (Third-party, comprehensive, marine, and Fire + stolen) clearly defined. Running a series of pattern extraction and analytical functions in R-studio detected missing values from the dataset with 89.5% classification accuracy as summarized in Figure 13;

```
> summary(ins1)
       Id          Documentclass      PayMoney        PayDate             IsalNo          PayType             DocType          DocType_imp
 Min.   : 93.0   Min.   :1.00   Min.   :    5   Length:100         Min.   :10001   Length:100         Length:100         Mode :logical
 1st Qu.:119.8   1st Qu.:1.00   1st Qu.:  200   Class :character   1st Qu.:13164   Class :character   Class :character   FALSE:81
 Median :144.5   Median :1.00   Median : 1086   Mode  :character   Median :32809   Mode  :character   Mode  :character   TRUE :19
 Mean   :144.5   Mean   :1.74   Mean   : 3183                      Mean   :36685
 3rd Qu.:169.2   3rd Qu.:2.00   3rd Qu.: 4368                      3rd Qu.:55273
 Max.   :194.0   Max.   :4.00   Max.   :14901                      Max.   :96423
```

Figure 13. Summary results of the incomplete dataset after imputation

The findings from this study are consistent with the results from similar studies [28], [29] which showed that K-Nearest Neighbor (KNN) technique is a superior classification algorithm with high accuracy when applied to the imputation of missing values in a datasets. The relative accuracy of this technique is based on factors such as weighted estimation, feature relevance to specific datasets under study, predictive detection of missing values based on a series of statistical analyses. Findings from a related study show that KNN-based imputation of missing values in a dataset delivers 90% accuracy in the classification of missing values and 86.27% performance in the replacement of numerical values in relatively less computational time and error compared to other techniques [30]. Imputation of missing values using KNN algorithm attains superior

performance in the experimental replacement of numerical values due to its unique ability to classify the missing parameters and assign cluster ratios for each type unlike other techniques that perform replacement in whole datasets based on the normalized computation of mean absolute errors and root mean square error. A study [31] observes that imputation based on computational and statistical models is recommended by scientists due to its unique ability to determine the missing values by averaging a summarized likelihood function of the entire dataset over a mathematically defined predictive distribution with considerably high precision.

The implementation of reverse data mining using the IBK classification algorithm has been effectively demonstrated as a reliable solution for handling missing data values. The pre-processing implementation functions replace missing values by computing the nearest neighbors with the highest similarity index. This approach recreates the dataset with missing values into a complete with accurately imputed variables and attributes for use in data analytics and decision support.

R-studio increases the performance accuracy of the imputation technique by replacing the missing values in a dataset with several independently imputed values rather than a single randomly imputed value [32] . This reflects possible uncertainties that may arise due to technical errors with the imputation model. For instance, if a regression model is applied to imputing the missing values, it is desired that imputations reflect both sampling variability and uncertainties regarding the regression coefficients utilized in the model. Independent modelling of the coefficients makes it possible to create a new set of imputed values for each instance based on the coefficient distribution through multiple imputations. Therefore, R-studio made it possible to run the standard analysis of the datasets with missing values to generate inferences, which are then combined to determine the most appropriate value of the missing data.

5. CONCLUSION

In conclusion, the objective of this study was to discuss and implement the imputation technique based on R-programming language as a solution for handling missing values in data sets. The problem of missing data may arise due to input or measurement errors, especially in our daily interactions with technology hence impacting the quality of analytical insights from such data. The implementation of k-NN imputation method based on the IBK classification algorithm proved a reliable approach to replacing missing data values with the value of nearest attributes showing the highest similarity.

The experimental results also showed that pre-processing by imputation delivers high-level performance efficiency in handling missing data values. Where these findings are consistent with the key idea and objective of paper, which is to explore alternative imputation techniques for handling missing values to improve the accuracy and reliability of decision insights extracted from datasets.

6. FUTURE WORK

While this paper presents an important knowledge framework for the future of data analytics and data-driven decision-making, more research is needed to refine the imputation mechanisms for replacing missing values to eliminate noise and inconsistencies, especially in the massive data generated from the Internet of Things. Future work should focus on the development of automated imputation techniques based on machine learning and artificial intelligence for improved efficiency in data pre-processing and analytics.

Since the dataset used in this study is very small, choosing the most common dataset would be useful for an extended study, especially for comparing results with other methods.

REFERENCES

[1] A. Nikitas, K. Michalakopoulou, E. T. Njoya and D. Karampatzakis, "Artificial Intelligence, Transport and the Smart City: Definitions and Dimensions of a New Mobility Era," *Sustainability*, vol. 12, no. 7, p. 2789, 2020.

[2] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, with Applications in R, vol. 2, New York, NY: Springer, 2021.

[3] Z. ÇETİNKAYA and F. HORASAN, "Decision Trees in Large Data Sets," *International Journal of Engineering Research and Development*, vol. 13, no. 1, pp. 140-151, 2021.

[4] S. Chuprov, I. Viksnin, I. Kim, T. Melnikov, L. Reznik and I. Khokhlov, "Improving Knowledge Based Detection of Soft Attacks Against Autonomous Vehicles with Reputation, Trust and Data Quality Service Models," in *2021 IEEE International Conference on Smart Data Services*, Chicago, IL, USA, 2021.

[5] S. Pei, H. Chen, F. Nie, R. Wang and X. Li, "Centerless Clustering: An Efficient Variant of K-means Based on K-NN Graph," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[6] ChaoFu, C. Xu, M. Xue, W. Liu and S. Yang, "Data-driven decision making based on evidential reasoning approach and machine learning algorithms," *Applied Soft Computing*, vol. 110, no. 15, 2021.

[7] V. Singh and V. D. Kaushik, "Concepts of Data Mining and Process Mining," in *Process Mining Techniques for Pattern Recognition*, CRC Press, 2022.

[8] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226-235, 2019.

[9] S. J. Choudhury and N. R. Pal, "Imputation of missing data with neural networks for classification," *Knowledge-Based Systems*, vol. 182, 2019.
[10] B. C. Wesolowski, "Data Preprocessing and Data Manipulation," in *From Data to Decisions in Music Education Research*, 1 ed., 2022.
[11] M. Umair, F. Majeed, M. Shoaib, M. Q. Saleem, M. S. Adrees, A. E. Karrar, S. Khurram, M. Shafiq and J.-G. Choi, "Main Path Analysis to Filter Unbiased Literature," *Intelligent Automation and Soft Computing*, vol. 32, no. 2, pp. 1179-1194, 2022.
[12] N. Whitmore, "Data cleaning," in *R for Conservation and Development Projects*, Chapman and Hall/CRC, 2020.
[13] E. W. Steyerberg, "Missing Values," in *Clinical Prediction Models. Statistics for Biology and Health*, Cham, Springer, 2019.
[14] T. F. Johnson, N. J. B. Isaac, A. Paviolo and M. González-Suárez, "Handling Missing Values in Trait Data," *Global Ecology and Biogeography*, vol. 30, no. 1, pp. 51-62, 2020.
[15] C. Bonander and U. Strömberg, "Methods to handle missing values and missing individuals," *European Journal of Epidemiology*, vol. 34, no. 1, 2019.
[16] A. E. Karrar, "Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 181-188, 2022
[17] Z. Hu and D. Du, "A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction," *PLoS ONE*, vol. 15, no. 9, pp. 1-15, 2020.
[18] A. K.S., R. Ramanathan and M. Jayakumar, "Impact of K-NN imputation Technique on Performance of Deep Learning based DFL Algorithm," in *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking*, Chennai, India, 2021.
[19] C. Arkopal and M. R. Kosorok, "Missing Data Imputation for Classification Problems," arXiv, 2020
[20] A. Yadav, A. Dubey, A. Rasool and N. Khare, "Data Mining Based Imputation Techniques to Handle Missing Values in Gene Expressed Dataset," *International Journal of Engineering Trends and Technology*, vol. 69, no. 9, pp. 242-250, 2021.
[21] A. E. Karrar, "A Novel Approach for Semi Supervised Clustering Algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 6, no. 1, pp. 1-7, 2017.
[22] A. Orfanoudaki, A. Giannoutsou, S. Hashim, D. Bertsimas and R. C. Hagberg, "Machine learning models for mitral valve replacement: A comparative analysis with the Society of Thoracic Surgeons risk score," *Journal of Cardiac Surgery*, vol. 37, no. 1, pp. 18-28, 2022.
[23] D. Bertsimas, A. Orfanoudaki and C. Pawlowski, "Imputation of clinical covariates in time series," *Machine Learning*, vol. 110, no. 1, pp. 185-248, 2021.
[24] B. M. Bai, N. Mangathayaru, B. P. Rani and S. Aljawarneh, "Mathura (MBI) - A Novel Imputation Measure for Imputation of Missing Values in Medical Datasets," *Recent Advances in Computer Science and Communications*, vol. 14, no. 5, pp. 1358-1369, 2021.
[25] D. Lee and K. Shin, "Robust Factorization of Real-world Tensor Streams with Patterns, Missing Values, and Outliers," in *2021 IEEE 37th International Conference on Data Engineering*, Chania, Greece, 2021.
[26] T. Siswantining, T. Anwar, D. Sarwinda and H. S. Al-Ash, "A Novel Centroid Initialization in Missing Value Imputation towards Mixed Datasets.," *Communications in Mathematical Biology and Neuroscience*, vol. 2021, 2021.
[27] F. Yin and F. Shi, "A Comparative Survey of Big Data Computing and HPC: From a Parallel Programming Model to a Cluster Architecture," *International Journal of Parallel Programming*, vol. 50, no. 11, pp. 27-64, 2022.
[28] R. Pan, T. Yang, J. Cao, K. Lu and Z. Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information," *Applied Intelligence*, vol. 43, no. 3, pp. 614-632, 2015
[29] P. Keerin and T. Boongoen, "Improved KNN Imputation for Missing Values in Gene Expression Data," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 4009-4025, 2022.
[30] K. M. Fouad, M. M. Ismail, A. T. Azar and M. M. Arafa, "Advanced methods for missing values imputation based on similarity learning," *PeerJ Computer Science*, vol. 7, 2021.
[31] M. Pampaka, G. Hutcheson and J. Williams, "Handling missing data: analysis of a challenging data set using multiple imputation," *International Journal of Research & Method in Education*, vol. 39, no. 1, pp. 19-37, 2014
[32] J. C. Jakobsen, C. Gluud, J. Wetterslev and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts," *BMC Medical Research Methodology*, vol. 17, pp. 162-171, 2017.