ABSTRACT

In machine learning and other relevant areas such as database knowledge discovery and data mining, data quality is a crucial concern because most machine learning approaches acquire knowledge exclusively from the data. It is therefore important to determine the quality of the knowledge extracted; however, the actual data in the world today are particularly sensitive to noisy, losing and incomplete knowledge. This data can confuse the mining process and lead to incorrect results. This made missing values is the primary challenge for data quality; many machine learning algorithms are quite unsophisticated in handling missing values, so the handling of incomplete data should be carefully considered because otherwise bias might be injected further into discovery of knowledge. Here when attempts to pre-process the missing values we used imputation approach which is a concept that describes a strategy that fills incomplete values in a data set with constant values so we use the IBK classification algorithm to indicate the missing data and impute it based on KNN. This leads to develop a way to handle missing values, after imputing the incomplete data the user can examine and use the generated complete file and choose it as entry to data mining.

Key words: Machine Learning, Data Mining, Preprocessing, Missing Values, Classification

1. INTRODUCTION

In recent years our universe has witnessed a major revolution in communication and information technology and mobility, from the 60's on the field of artificial intelligence (AI) has begun to evolve by enabling devices doing a human activity, despite major progress, many labor-intended jobs are remained far away from being solved by AI systems but in recent years this has significantly changed, computerized systems have been adopted in multiple areas and the quantity of data created by multiple resources leads to an explosive growth, leading to an advanced machine learning revolution [1].

Machine Learning (ML) is the field in which expert systems can learn from generated data, so a mathematical model can be built to extract meaningful patterns from data, the predictive modelling is referred to as the process of constructing these models for prediction or forecasting, there are several couple extra study methods for this task, they have been categorized depending on the nature of the problem in hand into two main category classification or regressing [2].

Databases and servers need comprehensive cleansing help and assistance, they load and constantly explore large amounts of data from a wide variety of sources so that certain sources are ideal for storing null values, data warehouses are also utilized for decision-making, so it is sufficient to avoid erroneous conclusions to guarantee that their data is accurate, so data cleansing is taken into consideration to be among the major difficulties in database management due to the wide variety of prospective inconsistencies and the massive amount of data [3].

When evaluating actual data, the lack of data is the most frequent challenge in machine learning, the core idea of the lack of data concept is that it is often not a random lack of information, the strength of the association gets less if mean imputation is utilized while the imputation of regression strengthens the association and at the same time increases the danger of ambiguity in the results, generally existing methods for replacement of missing data are based on the mostly improper statement that called MCAR data increases the probability of incomplete and interpretative failure, in addition the application of traditional techniques like removal and single imputation could lead to massive biases and faults in datasets.

The K-Nearest Neighbors (KNN) algorithm was presumed the MCAR Method for Missing, this approach calculates null data from genes with identical expression pattern to the gene of interest, a weighted average of the k closest gene to the gene of interest is determined as the missing value.

At last it is required to describe clearly the structure of the missing and all the methods outlined above to understand why missing data can result in bias.
2. **BACKGROUND**

2.1. **DATA MINING**

In order to analyze particular mechanisms for this information or it is the method by which data from data centres are extracted, data mining technology is simply extracted significant information from enormous quantities of information, the data mining transfers, classifies and extracts the data from a variety of steps, beginning from data cleaning, standardizing and relevant test data [4].

All directions for discovering usable data knowledge are outlined in the KDD (knowledge Discovery from Databases), most people interpret data mining as synonymous with another popular phrase like data discovery, while others regard data mining as only an essential phase in the information extraction [5].

Data mining consists of the following steps in an iterative sequence:

- **Data cleaning** (to remove noise and inconsistent data).
- **Data integration** (where multiple data sources may be combined).
- **Data selection** (where data relevant to the analysis task are retrieved from the database).
- **Data mining** (an essential process where intelligent methods are applied in order to extract data patterns).
- **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on some interestingness measures).
- **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user). [6]
2.2. **Preprocessing**

Data preparation puts the data into a format which is easier and more efficient for the user, because most real world information can now be missing, unreliable or messy the preprocessing techniques are identified including data cleansing, data integration, transformation and data reduction.

The data cleaning procedures try to complete null values, reduce noise while recognizing outliers and fix incoherences in the data.

**Figure 3: Preprocessing tasks**

Quality data is vital to get for the extraction of knowledge, since unclean data leads to poor interpretation of knowledge.

2.3. **Missing Values**

Missing Values (MVs) are an attribute value lost during recording, other reasons for data loss include data input failure, device failure and improper measurement.

The missing algorithm describes the link between a missing value's probability and the other data set variables, when Y represents all the data that may be segmented as (Yobs, Ymis) where Yobs is Y's observed part and Ymis represents missing in Y and R indicates random (or matrix) variable, whether or not Y's are observed or missing, so allow R= 1 to show the observed value and allow R = 0 to show the missing value, this will make the missing data model be P (R|y, Ø) in which Ø is the missing data process parameter, the missing mechanism is established by R's dependence on the data set variables.

The following are various missing mechanisms:

- **Missing Completely At Random (MCAR):**
  This missing mechanism is presented as:
  
  \[
  P (R|Y, Ø) = P (R, Ø).............(1)
  \]
  In other words, the probability of missing does not depend on any values observed or disregarded in Y, a computer fault that randomly deletes certain data values may be an example of MCAR.

- **Missing At Random (MAR):**
  This missing mechanism is presented as:
  
  \[
  P (R|Y, Ø) = P (R|Yobs, Ø).......(2)
  \]
  This means that the likelihood of missing is dependent on values observed in y and not on values not observed in Y, one simple example of MAR is an enquiry for which individuals beyond some age do not respond to a certain enquiry question and the covariate is noted.

- **Missing Not At Random (MNAR):**
  This missing mechanism is provided as:
  
  \[
  P (R|Y, Ø) = P (R|Yobs,Ymis, Ø) ........(3)
  \]
  This mechanism has been discovered when MAR conditions are broken so that the likelihood of lack depends on Ymis or any unexpected covariate, one example of MNAR may be those with an income over a specific value who do not declare an income, the lack here depends on the unobserved response, revenue.[7]

2.4. **Imputation**

Imputation (fill in) of missing data may just be the approach for handling any missing values, a variety of different algorithms are used for the imputation and the overall objective of the imputation is to replace the missing values carefully and avoid the data set distortions, because there are several techniques to solving the incomplete data problem, over the years such approaches were investigated and split into two categories: traditional and advanced approaches.

- **Traditional imputation methods:**
  Deleting missing values by lists and pair deletions is still one of the conventional techniques to missing data, even today this can be helpful if values are MCAR, so the basic strategies for imputations in this approach are Mean Imputation, Mode Imputation, Hot Deck
Imputation, Maximum Likelihood Imputation, Multiple Imputation.

- **Advanced imputation methods:** computational intelligence, newly designed technologies which led machine learning to treat complex issues so that proved very beneficial, we are familiar with advanced imputation procedures, such as: K-Nearest Neighbor Imputation, Random Forest Imputation, Decision Tree Imputation and other techniques.

2.5. **A Conceptual Explanation of K-Nearest Neighbors (KNN) Algorithm**

One of the monitored learning strategies widely used to classify the pattern recognition and also to regression is the k-nearest neighbor algorithm, in both cases the input includes the closest examples of training in datasets, and the output depends on whether k-NN is used for regression or classification.

- **KNN classification:** The K-Nearest Neighbor Algorithm is based on memory and does not require a particular model to match it, for a set of observations other than collecting vectors labeled with the class supplied, no special training is required. All intensive calculations involve two observations: the closest k in the training set and the most votes in iteration k and the labeling of the classes in each classification are found [8], the steps taken in the K-Nearest Neighbor categorization are:
  1. Identify k parameter.
  2. Determine using Euclidian approach the distance between each data to calculate the distance in Equation 1.
  3. Sorting data from small to big distances.
  4. Take the number given in the data which is k.
  5. Search for data with the most k number determined.
  6. Determine the most acquired data class obtained.

- **KNN Imputation:** The KNN technique was expanded in many datasets to the imputation of missing data, generally speaking the KNN imputation is a suitable alternative if we do not previously know the distribution of the data in a case where the system is incomplete, the approach selects a distance metric from its nearest k neighbors by estimating lack of data with the appropriate mean or mode, the mean rule is used to predict missing numerical parameters and the mode rule to predict missing categorical variables so KNN imputation is not explicitly creating prevision models as the training dataset is used as a lazy model this approach can also handle cases with several missing values easily.[9]

3. **Related Work**

Arkopal Choudhury & Michael R. Kosorok (2020) Propose the new kNN iterative gray class weighted imputation technique, based on the missing date and all the training data. In heterogeneous data with missing instances, gray distance works effectively. Mutual information (MI) weighs the distance, which is a measure of the importance between features and the class label. This guarantees that the data are imputed in order to improve classification performance. This weighted Gray kNN Algorithm shows better performance in imputation and classification challenges compared with other KNN Algorithms and typical imputation algorithms such as MICE and miss-Forest. These challenges are based on simulated circumstances using UCI data sets with different missing rates.[9]

Thomas F. Johnson & Nick J. B. Isaac & Agustin Paviolo & Manuela González-Suárez (2020) evaluate the performance of missing value management approaches in considering biased datasets by simulating continuous traits and separate response variables in a biased missing data scenario, for testing the performance of nine imputation processes, including complete case analysis (except missing data from the dataset). They characterized performance by assessing the error in imputed value parameters (deviation from true value) and the association of attribute and response (deviation from the true relationship between a trait and response).[7]

Christine R. Padgett & Clive E. Skilbeck & Mathew James Summers (2019) provide conceptual explanation of why newer missing data techniques are being employed, with a focus on multiple imputations (MI). 20 cases were selected randomly from a population study investigating cognitive sequelae of traumatic brain injury (TBI), and 8 of 20 cases were deleted on one variable to simulate a missing data set, to illustrate the technical effectiveness of deletion, single imputation, and multiple imputation techniques. When comparing each technique's parameter estimates with the known data set parameters, MI demonstrated that deletion and single imputing procedures were outperformed. More advanced procedures such as MI should consequently be kept in mind in clinical studies.[10]

SuvarJyoti Choudhury & Nikhil R. Pal (2019) Classifiers designed that differs from the
prediction of missing classification values. An autoencoder neural network is used for the imputation procedure. It uses training data innovatively to train the auto encoder without missing values in order to be more ready to predict missing values. It is a two-stage program of training. This research finally compares the suggested method with eight cutting-edge imputation technology utilizing 14 datasets and 8 classification methodologies.[6]

Dimitris Bertsimas & Colin Pawlowski & Ying Daisy Zhuo (2018) Offers a flexibly designed framework to impute missing information with mixed continuous and categorical variables based on formal optimisation. It allows for various model predictions such as closest K neighbours, vector support machines and decision tree methods and can be adapted to multiple imputations. This framework is easily obtainable. They produce almost first-order approaches that achieve high-quality solutions in seconds following an opt.impute generic algorithm. Authors show that their methodology provided improves the exactness of the sample in large-scale experiments through a selection of 84 data sets derived from the UCI Machine Learning Repository. Opt.impute produces the best total imputation in most sets of data with five different methods in all scenarios of missing random mechanisms and various missing percentages: mean impute, k-nearest neighbours, predictive knn, Bayesian PCA and predictive-mean matching, with an average reduction of 8.3 percent in the average absolute error compared with the best cross validated benchmark method.[11]

B. Mathura Bai, & N. Mangathayaru (2018) this article deal with mining medical datasets because these datasets contain various hidden obstacles compared to normal datasets. One of the key issues in mining medical datasets is handling null values which is part of preprocessing step. So, it covers the issue of handling null values in medical dataset consisting of categorical attribute values. The key aspect of this work is to apply the proposed imputation measure to estimate and correct the missing values.[12]

Sang Kyu Kwak & Jong Hae Kim (2017) This overview includes the varieties of null data, strategies to identify outliers and deal with both because during data collection, missing values and outliers frequently arise. The occurrence of missing values limits the available data to be analyzed, reducing its statistical capacity and, ultimately, its dependability. Outliers have a substantial impact on the statistical estimation process leading to inflated or understated data. The results of the data analysis therefore depend substantially on the manner of processing of missing values and outliers.[13]

Zhun-ga Liu & Quan Pan & Jean Dezert & Arnaud Martin (2016) Present an incomplete pattern creedal classification approach based on belief function theory, which includes an adaptive imputation of missing values. At beginning, attempt classifying the object on the basis of the supplied values (incomplete pattern 23). As a principle, if a certain class for the object can only be discovered with the existing data, the missing information is not necessary for the classification. The object is committed to this specific class in this situation. However, it indicates the missing data have become a central factor in attaining accurate classification if an object cannot be classified without ambiguity. In this situation, the missing values are imputed using the self-organizing map (SOM) technology, which is subsequently classed as an altered pattern. The pattern (original or altered) is classified according to each training class and the outcomes of classification represented by basic tenets Which will be combined with appropriate criteria for the creedal categorization.[14]

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Year</th>
<th>Title</th>
<th>Section</th>
<th>Mechanism Used</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arkopal Choudhury &amp; Michael R. Kosorok</strong></td>
<td>2020</td>
<td>Missing Data Imputation for Classification Problems</td>
<td>lazy learning technique</td>
<td>novel iterative kNN technique</td>
</tr>
<tr>
<td><strong>Thomas F. Johnson &amp; Nick J. B. Isaac &amp; Agustin Pavilo &amp; Manuela González-Suárez</strong></td>
<td>2020</td>
<td>Handling missing values in trait data</td>
<td>trait datasets</td>
<td>Imputation</td>
</tr>
<tr>
<td><strong>Christine R. Padgett &amp; Clive E. Skilbeck &amp; Mathew James Summers</strong></td>
<td>2019</td>
<td>Missing Data: The Importance and Impact of Missing Data from Clinical Research</td>
<td>Missing Data in Clinical database</td>
<td>multiple imputation (MI)</td>
</tr>
<tr>
<td><strong>Suvra Jyoti Choudhury &amp; Nikhil R. Pal</strong></td>
<td>2019</td>
<td>Imputation of missing data with neural networks for classification</td>
<td>use data with missing values for designing classifiers</td>
<td>Imputation</td>
</tr>
</tbody>
</table>
4. **Experimental Procedures**

Here we worked to deal with the missing data problem by using the simplest way, and it is important to use the most accurate data collection, so here we used one of the data mining Algorithms (IBK algorithms) to address this important problem in preprocessing using an imputation technique and take raw data to create a standardized methodology that can enhance data accuracy.

### 4.1. Dataset

The dataset utilized here is the actual data from Juba Insurance & Reinsurance Company, massive data amounts are already in use for data mining however the paper employed basic arbitrary 100 samples and three missing mechanisms namely MCAR, MAR and NMAR, were considered to incorporate fictitious missing values into the insurance dataset, the original data sets are run on random data generating units to simulate missing value on the 4 classes from DocType attribute, which contains the same likelihood $\alpha$ (the missing rate), this generates 3 percentage of missing rate (3%, 6% and 10%), so it is possible to simulate missing values on the DocType attribute which contain 4 classes, this creates: (5 Third Part, 8 Complete, 3 Marine, 3 Fire+stole) after using one of the missing methods to enter artificial missing values in it.

### 4.2. IBK Imputation

Now we can utilize the IBK algorithm (k-Nearest Neighbors) for imputing the data after the data has been prepared, thus the k-Nearest Neighbor’s algorithm (k-NN) as pattern recognition is a non-parametric method suggested for classification and regression, in all circumstances the entry consists of the closest examples of training in the functional area. The output is a membership of the class in k-NN, an object is classifiable by a vote of plurality of its neighbours which allocates the object to its closest neighbours, the class most common, oftentimes if distance metric is learnt with specialist algorithms like Large Margin Nearest Neighbor or Neighborhood Components Analysis, the classification [15] accuracy of k-NN can be greatly enhanced, so the ideal choice of k depends on the data in general, larger k values lower the noise influence on the classification, but provide less distinctive boundaries between classes, different heuristic strategies can identify a good k the particular case when the class of the nearest training set is estimated to be (i.e. when k=1) is the closest neighbor algorithm, the sensitivity of the k-NN algorithm can be seriously reduced due to noisy or unnecessary features if the feature scales do not correspond.
Handling Missing values are one of the key factors to analyze data further, which makes the imputation of null values a main step in data preparation so we construed R programming language to utilize the classification method for imputing the MVs in dataset by following next procedures:

1. Place the dataset in record and attributes (rows and columns).
2. Settle two sections of the dataset (Complete, with missing values).
3. Normalize the above dataset.
4. Impute the missing values for the records, each one at a time.
5. Test the similarity between new records and current records using the nearest neighbor (k-NN).
6. Fill in the missing attribute value with the nearest matching record attribute value to which it has the highest similarity.
7. Put the column with the full dataset part including its handled attribute.
8. Repeat all of the above to include the missing information in all records.

This proposed imputation of the IBK algorithm (k-NN) relies on the modification for parameters of the gower distance, and in order to use all the dataset you have to provide the name of the dataset and the parameters that indicate null values and the number of nearest neighbors.

4.3. LANGUAGES

To implement the proposed algorithm we used two languages, first one is Java and the second is R language which is a language for data analysis and graphics, it was created in 1991 by Ross Ihaka and Robert Gentleman in the department of statistics at the university of auckland, in 1993 the world received the first announcement of R, and to apply the R language we used RStudio tool which is one of the Integrated Development Environments (IDE) available for R language that is built by RStudio.

5. IMPLEMENTATION

The first thing to do is to set the work directory of the session to the file pane position we have used here:

```r
setwd("~/R implementation").
```

Once we read the data set file, we save it in certain objects for analysis and utilize it by:

```r
ins <- read.csv("missing enc.csv")
```

Which means read the whole dataset in another object called (ins).

The original data set was saved as a csv file prior before loaded in the RStudio and it’s a real data containing 7 attributes and 100 instances and null data.

We noted DocType is the most incomplete data element parameter and the data is MCAR, too much missing data can be a problem too so we can therefore observe the essential structure of the data by:

```r
> str(ins)
'data.frame': 100 obs. of 7 variables:
  $ Id     : num 91 96 97 98 99 100 101 102 103 104 ...  
  $ DocType: chr "check" "check" "check" "check" "check" ...
  $ TotalMon : num 7.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 ...
  $ Payment : chr "19/04/2013" "19/04/2013" "19/04/2013" "19/04/2013" ...
  $ TotalPay : num 55455 44300 14476 20658 12357 13483 10001 15254 105555 11436 ...
  $ Payment : chr "check" "check" "check" "check" ...
  $ DocType : chr "check" "check" "check" "check" ...
```

Figure 8: Dataset structure

Then we will read the entire sample data and try to set the NA value in the dataset for the DocType parameter, this means all "" in the dataset goes to "NA" which allows the suggested algorithm to detect all the null values.

![Figure 7: Proposed algorithm flowchart](image-url)
Furthermore, to analyze incomplete data, we examine the missing values by utilizing a user-defined function that uses variables (columns) and samples (Rows).

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

Figure 10: Missing values user defined function

![Figure 10: Missing values user defined function](image)

Then we utilized a few R packages to analyze the missing data pattern, which provide the `md.pattern()` function that gives better comprehension to the lack of data pattern.

```r
> md.pattern(ins)
```

![Figure 13: Row corresponds to missing data pattern](image)

The prior output shows 81 clean samples (no missing values) but for the DocType variable, 19 values are missing.

The fraction of incomplete data by variable will be viewed and analyzed using the aggregate plot, the `aggr()` function allows the combination of histogram and pattern chart to create an aggregate graphic.

```r
> aggr.plot <- function(x, y, xlab="", ylab="", main="Missing data by variable")
```

![Figure 15: Visualize the proportion of missing data by variable](image)

Each row reflects a lost pattern of data (1=observed, 0=missing), increasing numbers of incomplete details organized by rows and columns, the first line represents the number of observations while the last line shows the number of variables with incomplete data, and the plot value is the incomplete pattern which is converted into a pattern plot.
Only the percentage of null data is displayed in the above charts, the graph demonstrates the fraction of the values of Non-Missing (Blue) and Missing (Red), and from these plots we can now see that 81% of the observations do not lack data in the dataset, 19% of observations on DocType variable, so we might consider the missing values for this variable to be imputed.

6. **RESULTS AND DISCUSSION**

The imputed incomplete data used the K-Nearest Neighbor imputation approach with the IBK classification function by the R language package, the proposed algorithm we have offered with the name of the object dataset, the names of the variables that includes the missing values, and at last we used 4 number of Nearest Neighbors because we have (Third part, Comprehensive, Marine, Fire+stole), this utilized by:

```r
ins1 <- KNN(ins, variable = c("DocType"), k=4)
```

So we export the data from RStudio after the value has been imputed and stored as imputed.csv by:

```r
write.table(ins1,file = "imputed.csv",row.names = FALSE,sep =",")
```

The ins1 is the name of my RStudio data, the file is our new export data name, the row.names means don’t show all the data in one column, the sep is the csv separator that is used.

Now we can compare the summary of the imputed data set with the initial data set, to see that each variable really does have a substantial difference.

```
> summary(ins1)
  31 DocumentClass PayAmount PayDate  Value PayType DocType DocType_imp node
  Min. :13.00 Min. :1.00 Min. :1.00 Min. : 5 Min. :1000 Min. :1000 Min. :1000
  1st Qu.:12.00 1st Qu.: 2.00 1st Qu.: 2.00 1st Qu.: 5000 1st Qu.:10000 1st Qu.:10000
  Median :14.50 Median : 2.74 Median : 3.00 Median :10000 Median :20000 Median :20000
  Mean :14.50 Mean : 2.74 Mean : 3.00 Mean :10000 Mean :20000 Mean :20000
  3rd Qu.:19.00 3rd Qu.: 3.00 3rd Qu.: 3.00 3rd Qu.:15000 3rd Qu.:25000 3rd Qu.:25000
  Max. :24.00 Max. : 4.00 Max. : 4.00 Max. :60000 Max. :60000 Max. :60000
```

**Figure 18: Original data summary**

Shown above figure represents that there is column called DocType_imp that has a 81 variables considered false because there is no MVs on it and the 19 others are true because they’re being imputed MVs on it, and the last accuracy was 89.5% as 17 variables of the total 19 were the true imputed variables.

### Table 2: Missing vs imputed values

<table>
<thead>
<tr>
<th>Missing Values</th>
<th>Imputed Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Third part</td>
<td>6 Third part</td>
</tr>
<tr>
<td>8 Comprehensive</td>
<td>9 Comprehensive</td>
</tr>
<tr>
<td>3 Marine</td>
<td>2 Marine</td>
</tr>
<tr>
<td>3 Fire+stole</td>
<td>2 Fire+stole</td>
</tr>
</tbody>
</table>

7. **CONCLUSION**

In summary, the main objective of this paper is to handle the missing data, so this fundamental problem is solved completely when we utilized the k-NN imputation technique by the proposed approach algorithm that used the IBK classification and formed it to help providing a complete dataset which can present a direct motivation to the research being carried out.

In future, massive data amounts could be used, and the missingness mechanism can be investigated automatically to drive consistency for the data.

### References


