Project Report On

"Comparative Study of Techniques for Imputation of Missing Data in Datasets"

Submitted for partial fulfillment of

B.Tech in Computer Science and Engineering

Submitted by : -

Name : - Nilratam Sarkar

Roll no : - 1814017

Registration no : - (ADTU/L/2018-22/BCS/017)

8th semester ,Bachelor Of Technology,Computer Science & Technology

Under The Guidence of : -

Dr. Manoj Kumar Sarma, Program Co-ordinator

Faculty of Engineering and Technology

Computer Science and Engineering

Assam Down Town University

Sankar MadhabPath Gandhi Nagar, Panikhaiti,

781026, Guwahati, Assam,India

Assam Down Town University
Computer Science & Engineering
Faculty of Engineering & Technology
Sankar Madhab Path , Gandhi Nagar,
Panikhaiti ,781026 , Guwahati , Assam , India

CERTIFICATE

This is to certify that A Project titled **“Comparative Study of Techniques for Imputation of Missing Data in Datasets”** submitted by *Nilratan Sarkar bearing Registration no : -ADTU/L/2018-22/BCS/017 & Roll no : - 1814017* , students of 8th semester , B.Tech C. S . E , carried under my guidance for the Degree Bachelor of Technology in Computer Science & Engineering of *Assam Down Town University* and the work is original and not a copy of any other project.

Date : -

(*Signature of Dean*)

(*Signature of Supervisor)*

Assam Down Town University
Computer Science & Engineering
Faculty of Engineering & Technology
Sankar Madhab Path , Gandhi Nagar, Panikhaiti ,
781026 , Guwahati , Assam , India

CERTIFICATE

This is to certify that A Project titled **“Comparative Study of Techniques for Imputation of Missing Data in Datasets”** submitted by *Nilratan Sarkar bearing Registration no : -ADTU/L/2018-22/BCS/017 & Roll no : - 1814017* , students of 8th semester , B.Tech C. S . E , carried under my guidance for the Degree Bachelor of Technology in Computer Science & Engineering of *Assam Down Town University* and the work is original and not a copy of any other project.

Date : -

(*External Examiner*) (*Internal Examiner)*

DECLARATION

I hereby declare that the project named "**Comparative Study of Techniques for Imputation of Missing Data in Datasets"**, is on the basis of my own deeds , completed during the course under the guidance of Dr. Manoj Kumar Sarma .

I verify that the comments made and conclusions given are the result of our own work. I further declare that to the results given in the report have not been submitted to any other University or Institutions for the award of any other degree in this University or any other University.

Date : - (Signature of the Candidate)

Place : - Name : - Nilratan Sarkar

Roll no : - 1814017

Registration no : - (ADTU/L/2018-22/BCS/017)

ACKNOWLEDGEMENT

I would like to extend my gratitude and my sincere thanks to Assam Down Town University , for giving us such a great platform and I would like to convey my gratefulness towards Dr. Manoj Kumar Sarma, Program Co-originator , Faculty of Engineering and Technology , Computer Science and Engineering, Assam Down Town University for his support and guidance in accomplishment of this project on "***Comparative Study of Techniques for Imputation of Missing Data in Datasets***" **.**

Date : -

(Signature of the Candidate)

Place : -

Name : - Nilratan Sarkar

Roll no : - 1814017

Registration no : - (ADTU/L/2018-22/BCS/017)

CONTENTS

LIST OF CONTENTS PAGE NO

FIGURES

LIST OF FIGURES PAGE NO

FORMULAS

LIST OF FORMULAS PAGE NO

TABLES

LIST OF TABLES PAGE NO

❖ Abstract : -

Missing data (MD) is a common problem in data science job. When ignored or treated not appropriately, MD can lead to seriously biased results. The objective of this project is to discuss various methods to impute missing data. Here, in this project I have taken a data-set and removed some values randomly from one column then used different imputation techniques to impute the missing values.

The implemented imputation techniques are , Simple Imputation, Iterative Imputation(II), Multiple Imputation by Chained Equations(MICE), k-nearest neighbors(KNN) , MissForest(MF) .

After implementing the imputation techniques I have evaluated the techniques , evaluation metrics used are, Root Mean Square Error(RMSE), Mean Absolute Error(MEA), Mean Absolute Percentage Error(MAPE) .

For better understanding of the results I have plotted an image to show the difference between original data-set and the imputed data-sets and also plotted Root Mean Square Error(RMSE) of the implemented imputation techniques.

Keywords : - Missing data (MD), Simple Imputation, Iterative Imputation(II), Multiple Imputation by Chained Equations(MICE), k-nearest neighbors(KNN) , MissForest(MF) .

1. Introduction : -

Missing values are usually attributed to : human error when entering the data, machine malfunctioning , participant refused to answer some personal questions, the raw data got destroyed due to neglect .

The missing data problem is common in all fields that uses data and causes different issues like performance degradation, data analysis problems and biased results caused by the differences in missing and complete values .

How much serious the missing data issue is that depends in part on how much data is missing, the pattern of missing data, and what mechanism the missingness of the data follows .

Missing values can be handled by certain techniques including, deletion of instances and replacement with estimated values , known as imputation .

In this study I use a data-set of Missing Completely At Random(MCAR) mechanism to perform different imputation techniques and evaluate and compare the techniques using different evaluation metrics like RMSE, MAE, MAPE .

2. Missing data patterns : -

Missing data patterns explains which values are missing and observed in a data set. There is no standard list of missing data patterns. We will discuss three missing data patterns that appears the most in the literature which are Univariate, Monotone and Non-Monotone.

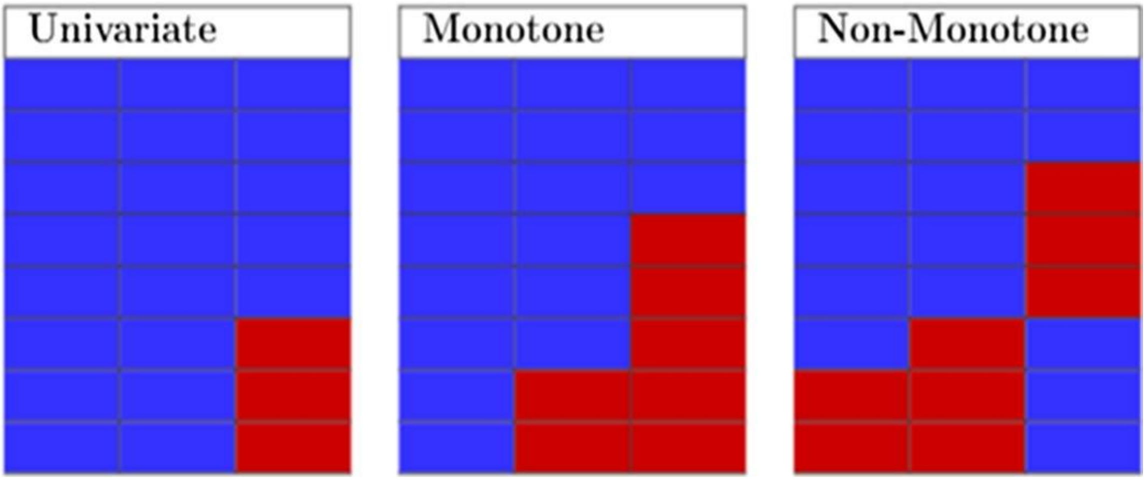


Fig. 1 : - Representation of missing data patterns data. Blue represents observed values; red is missing values

2.1 Univariate : -

When a data-set has missing data in only one variable then the missing data pattern is called univariate .

2.2 Monotone : -

Missing data pattern can be said to be as monotone if the variable values of the data set can be arranged .

2.3 Non-Monotone : - When the missingness of one variable does not affect the missingness of other variable then it can called Non-Monotone .

3. Missing data mechanisms : -

Missing data mechanisms are defined based on the available and the missing data . Missing data mechanisms can be categorized into three main mechanisms these are discussed bellow ..

3.1 Missing completely at random (MCAR) : -

This is when the missing data is not dependent on any other available or missing values.The total percentage of missing values is not dependent on anything at all.

3.2 Missing at random (MAR) : -

Rate of missing data is dependent on available data.Missing at random (MAR) is mostly occur in medical science studies data sets. Under this mechanism, missing values can be handled by other variable in the data-set.

3.3 Missing not at random (MNAR) : -

When the missing data is neither MCAR or MAR ,then it refers to as MNAR .The rate of missing data equally depends on missing and available data.Handling the missing values is mostly impossible in this method, as it also depends on the unobserved data.

4. **Missing values approaches : -**

In this section we will discuss the approaches for handling of missing values in a data-set .

4.1 Deletion : -

In this approach all instances with missing values are deleted when doing analysis. Deletion is considered the simplest approach available as there is no need to try and estimate the values. The disadvantages of deletion, as it gives biased outcomes in analysis, mostly when the missing data is not randomly distributed.The deletion process can be carried out in two ways, pairwise or list-wise deletion .

4.1.1 List- wise or case deletion : -

In list-wise deletion every instance with missing values is deleted .

4.1.2 Pairwise deletion : -

To mitigate against information loss when doing deletion one can use pairwise deletion rather than list-wise deletion .

4.2 Imputation : -

Imputation involves the process of replacing missing values with some predicted values. The available values from the data set is usually mined to predict the values .Imputation methods can be divided into single and multiple imputation methods based on the number of values imputed .In accordance to the construction approach used for imputation of missing data, these methods can also be classified as statistics-based and machine learning-based (or model-based) methods.

5. The imputation techniques implemented : -

5.1 Simple imputer : -

It is a statistic based approach in which a statistic(such as mean) is calculated from each column with missing values and then all missing values are replaced with the calculated statistic . In simple imputation, missing data is imputed by different strategies such as, mean,median, or mode of the available values.Mean imputation is one of the most used methods in simple imputation; it fill missing values with the calculated means . Medians is used instead of means for reliability in some cases . For cases where categorical variables are used , the missing data are commonly replaced with the most frequent values in the data-set. Even if this method is simple and can be powerful, it has its limitations, simple imputation may produce biased or unrealistic outcome on a high-dimensional data-set.

5.2 Iterative Imputer : -

Iterative imputer is a multivariate imputer ,that means it estimates each features from all others. A technique for imputing missing data as a function of other features in a round-robin order .At each step a feature is designated as target value(y) and other features is designated as independent value(x). A regressor is fit on (x, y) for known y. Then the regressor is used to predict the missing values of y. This done for *max_iter* imputation rounds . The result of final imputation round is then returned .

5.3 Multiple Imputation by Chained Equations (MICE) : -

MICE works under the presumption that the variables used in the imputation technique, the missingness of the data are Missing At Random (MAR), which means that the possibility that a value is missing depends only on observed values and not on unobserved values. Implementing MICE when data are not Missing At Random (MAR) could result in biased outcomes.

5.4 K nearest neighbour(KNN) Imputer : -

The logic behind KNN methods is to identify ‘k’ samples in the data-set that are similar or close in the space. Then we use these ‘k’ samples to estimate the value of the missing data points. Each sample’s missing values are imputed using the mean value of the ‘k’-neighbors found in the data-set using a distance measure between instances .Some of the distance measures are such as the Minkowski distance, Manhattan Distance, Cosine Distance, Jaccard Distance, Hamming Distance and Euclidean distance can be used for KNN imputation. However the Euclidean distance is known to give more efficiency and productivity than other distance measure that is why it is the most widely used distance measure.

5.5 MissForest Imputer : -

MissForest is machine learning based missing data imputation algorithm that works on the basis of Random Forest algorithm .

First , the missing values are filled using median/mode imputation method. Then we mark the imputed values as 'Predict' and the others as training rows, which are then put into a Random Forest model trained to predict the missing values .This process repeats itself several times , each iteration is giving better and better results . Iterations continue until maximum iterations criteria is met .

6. Performance Evaluation Metrics : -

Te performance evaluation for different missing value imputation techniques can be done using different criteria, on this section we discuss the most used which are , Root Mean Squared Error (RMSE) , Mean Absolute Error (MAE), and Mean absolute percentage error(MAPE) .

6.1 Root Mean Square Error (RMSE) : -

Root Mean Square Error calculates mean in the difference between the imputed values and the actual values .

Frm .1 Formula of RMSE

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \hat{x}_i\right)^2}{N}}$$

RMSD = root-mean-square deviation\error

i = variable i

N = number of non-missing data points

x_i = actual values

$\hat{x}_i$ = imputed values

6.2 Mean Absolute Error (MAE) : -

Mean Absolute Error calculates the mean of the absolute difference between the imputed values and the actual values.

Frm .2 Formula of MAE

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left|y_i - x_i\right|}{n}$$

MAE = mean absolute error

y_i = imputed value

x_i = true value

n = total number of data points

6.3 Mean Absolute Percentage Error(MAPE) : -

Mean Absolute Percentage Error calculates the total percentage of Mean Absolute Error .

Frm .3 Formula of MAPE

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

M = mean absolute percentage error

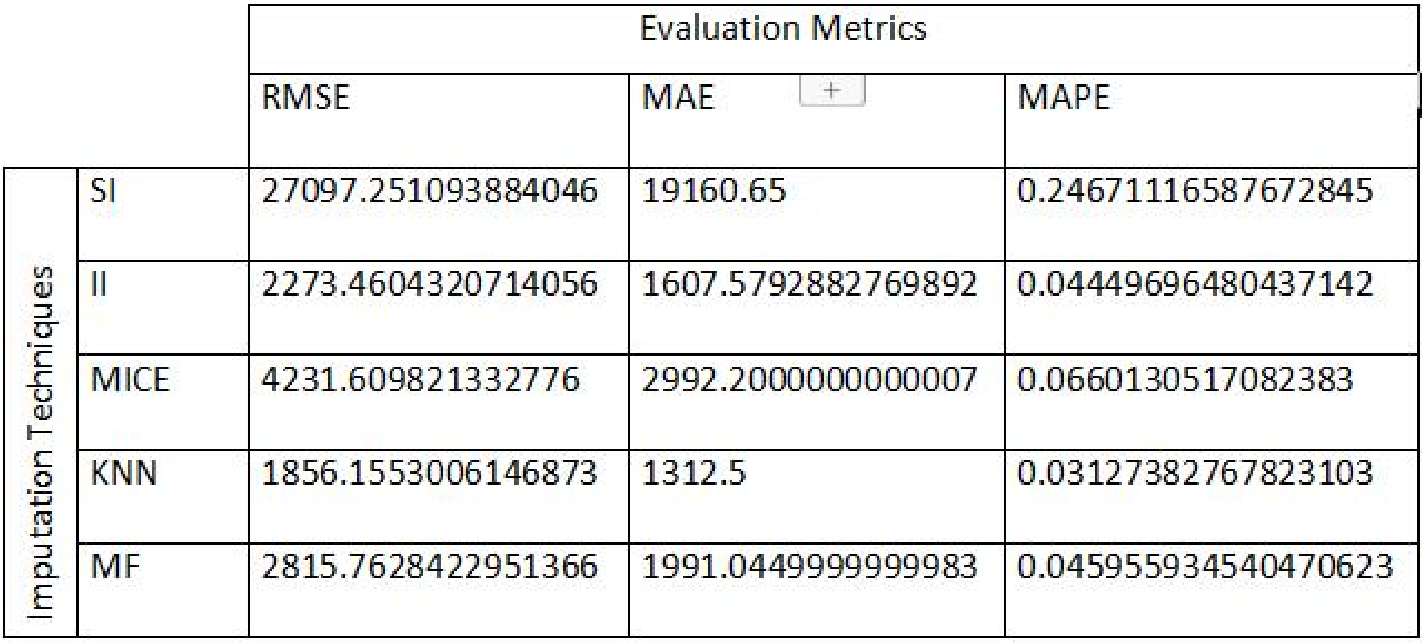
n = number of times the summation iteration happens

A_t = actual value

F_t = imputed value

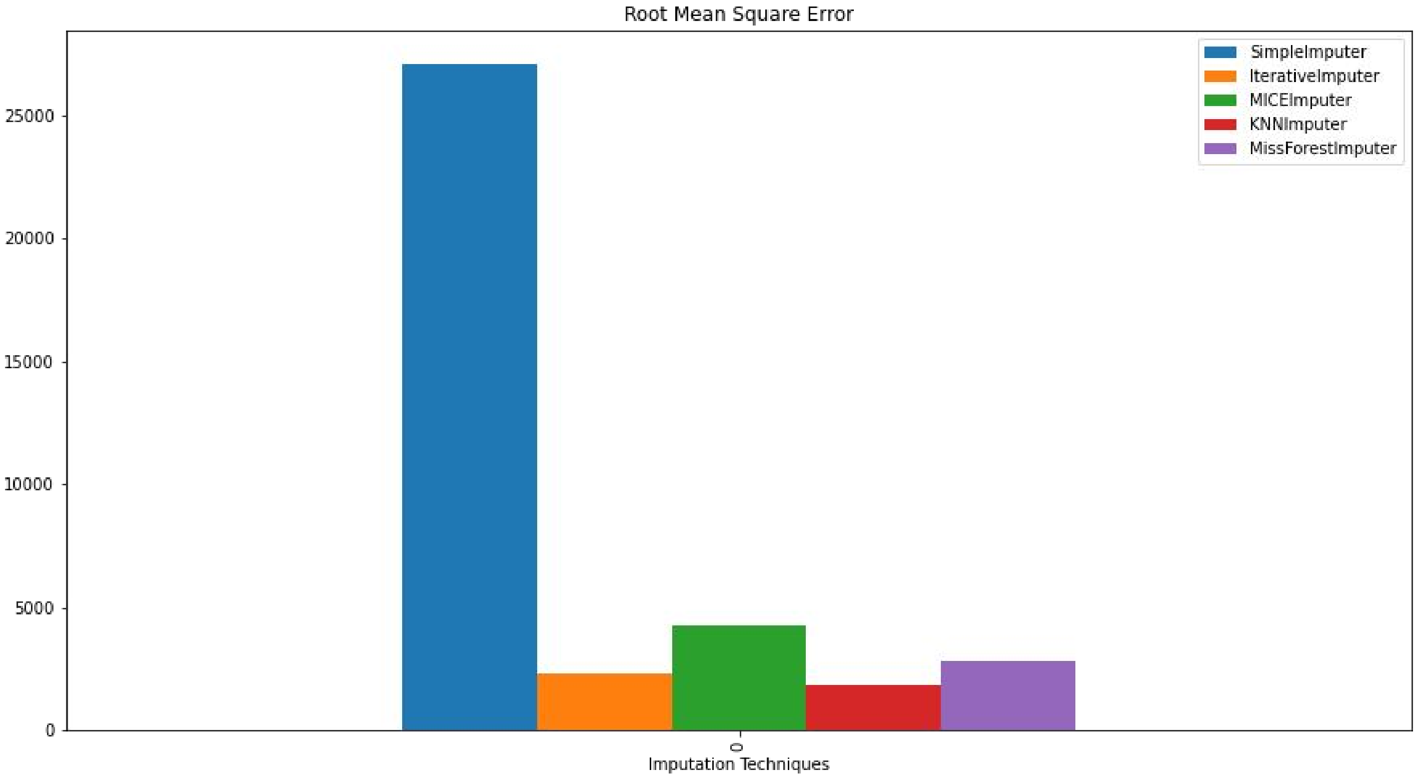
7. **Results : -**

7.1 Table 1 : - Results of the imputation techniques .

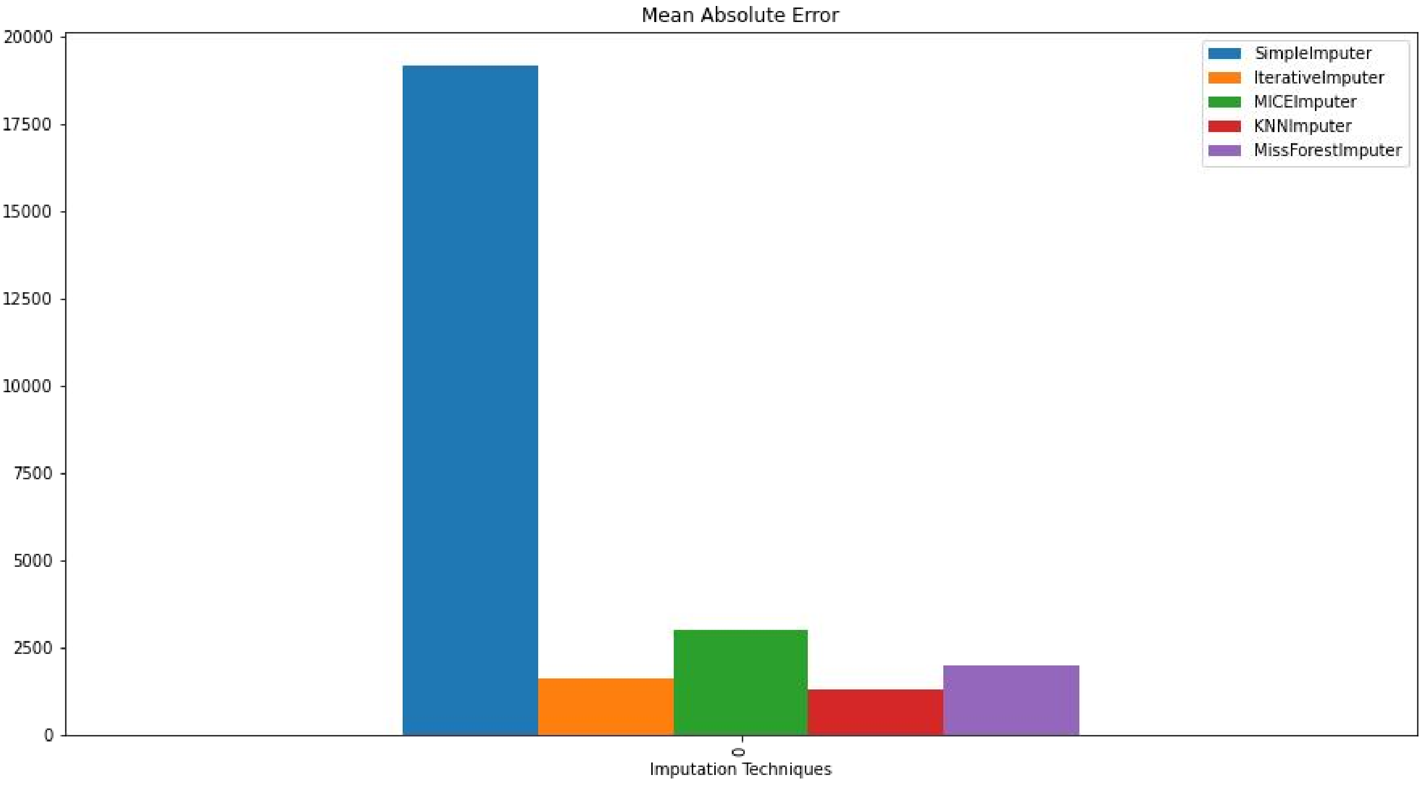


		Evaluation Metrics		
		RMSE	MAE	MAPE
Imputation Techniques	SI	27097.251093884046	19160.65	0.24671116587672845
	II	2273.4604320714056	1607.5792882769892	0.04449696480437142
	MICE	4231.609821332776	2992.2000000000007	0.0660130517082383
	KNN	1856.1553006146873	1312.5	0.03127382767823103
	MF	2815.7628422951366	1991.0449999999983	0.04595593454070623

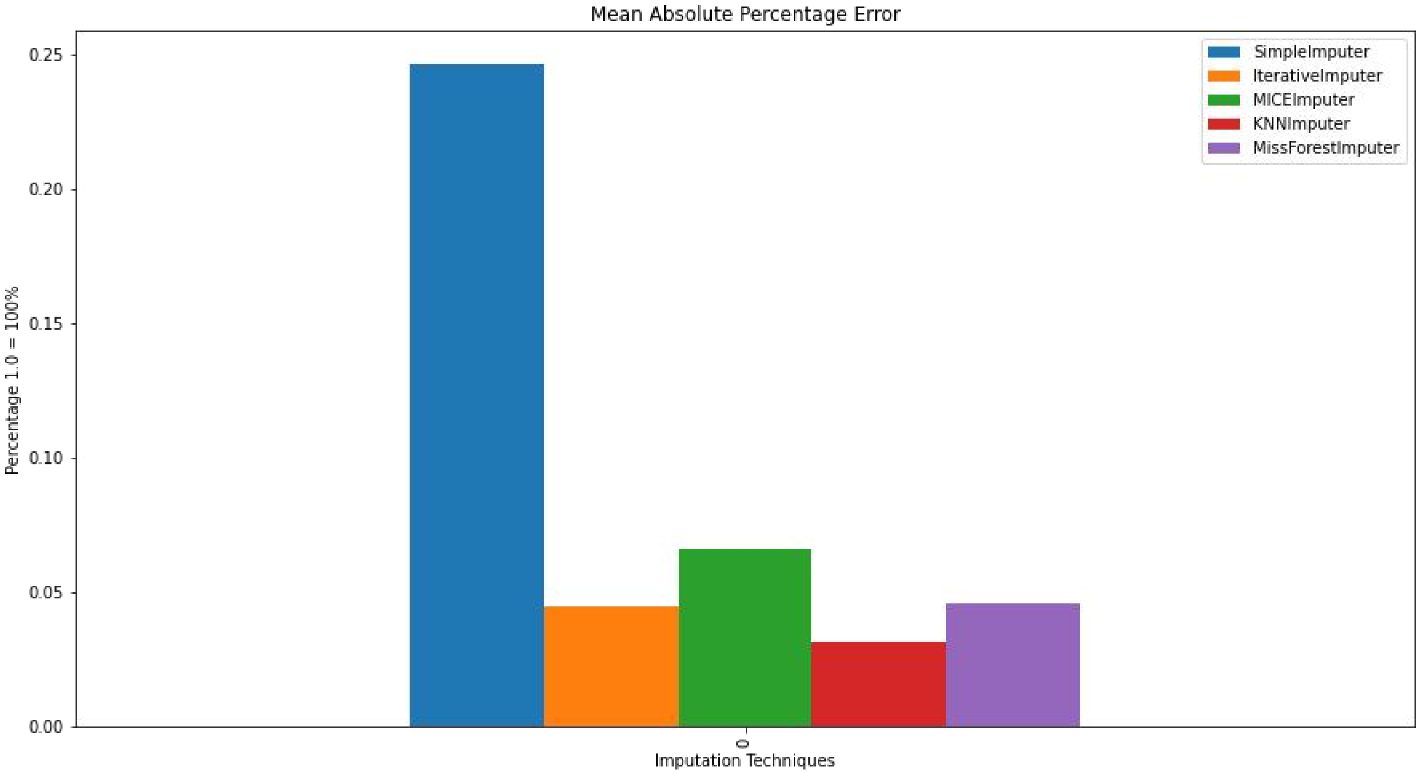
*7.2 **Fig .2** : - Graphical comparison of Root Mean Square Error(RMSE) of the imputation techniques .*



*7.3 **Fig .3** :-Graphical comparison of Mean Absolute Error(MAE) of the imputation techniques.*



7.4 **Fig .4** : - Graphical comparison of Mean Absolute Percentage Error(MAPE) of the imputation techniques .



- Figure 2 shows the simple imputer has highest RMSE of 27097.25 and knn imputer lowest RMSE of 1856.18 .
- Figure 3 shows the simple imputer has highest MAE of 19160.65 and knn imputer got lowest MAE of 1312.5 .
- Figure 4 shows the simple imputer has highest MAPE with 0.2467(or 24%) and knn has lowest MAPE with 0.0312(or 3%) .

8. Conclusion : -

After comparing the results of all the implemented imputation techniques we can say for this data-set, knn imputer is the best and simple imputer is the worst performing imputation techniques among all the techniques implemented in this project .

References : -

1. Tlamelo Emmanuel , Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago and Oteng Tabona(2021) .-- “A survey on missing data in machine learning”,Department of Computer Science and Information Systems, Botswana International University of Science and Technology, Palapye, Botswana .

2. Hatice UENAL, Benjamin MAYER ,Jean-Baptist DU PREL . -- “CHOOSING APPROPRIATE METHODS FOR MISSING DATA IN MEDICAL RESEARCH: A DECISION ALGORITHM ON METHODS FOR MISSING DATA”, - JOURNAL OF APLIED QUANTATIVE METHODS .

3. Yingpeng Fu , Hongjian Liao and Longlong Lv (2021) . -- “A Comparative Study of Various Methods for Handling Missing Data in UNSODA”, School of Human Settlements and Civil Engineering, Xi’an Jiaotong University, Xi’an 710049, China .

4. Youngdoo Son and Wonjoon Kim (2020) . -- “Missing Value Imputation in Stature Estimation by Learning Algorithms Using Anthropometric Data: A Comparative Study ”, Department of Industrial and Systems Engineering, Dongguk University—Seoul, Seoul 04620, Korea; Department of Industrial & Management Engineering, Sungkyul University, Anyang 14907, Korea .

5. CONCEPCIÓN CRESPO-TURRADO* , University of Oviedo, Maintenance Department, San Francisco 3, Oviedo 33007, Spain. - JOSÉ LUIS CASTELEIRO-ROCA, University of A Coruña, Departamento de Ingeniería Industrial, A Coruña 15405, Spain . - FERNANDO SÁNCHEZ-LASHERAS, University of Oviedo, Department of Mathematics, Facultad de Ciencias, Oviedo 33007, Spain. - JOSÉ ANTONIO LÓPEZ-VÁZQUEZ, University of A Coruña, Departamento de Ingeniería Industrial, A Coruña 15405, Spain. - FRANCISCO JAVIER DE COS JUEZ, University of Oviedo, Department of Mathematics, Facultad de Ciencias, Oviedo 33007, Spain. - FRANCISCO JAVIER PÉREZ CASTELO, University of A Coruña, Departamento de Ingeniería Industrial, A Coruña 15405, Spain. - JOSÉ LUIS CALVO-ROLLE, University of A Coruña, Departamento de Ingeniería Industrial, A Coruña 15405, Spain. - EMILIO CORCHADO, University of Salamanca, Departamento de Informática y Automática, Plaza de la Merced s/n, 37.008, Salamanca, Salamanca, Spain(2020). -- “Comparative Study of Imputation Algorithms Applied to the Prediction of Student Performance” .

6. (Howell, D.C. (2008) The analysis of missing data. In Outhwaite, W. & Turner, S. Handbook of Social Science Methodology. London: Sage.). -- “The Treatment of Missing Data” .

7. Therese D. Pigott , Loyola University Chicago, Wilmette, IL, USA . -- “A Review of Methods for Missing Data” - Educational Research and Evaluation,2001, Vol. 7, No. 4, pp. 353±383 .