***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 : - Nilratan 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

## 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 purpose of conducting this comparative study 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 implemented different imputation techniques .

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 easy understanding of the imputation results I have plotted images of of results of different evaluation metrics to show the differences in results , comparing with the original data-set .

*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 mostly results of : human mistake in entering the data, machine malfunctioning , participant refused to answer some personal questions, the raw data got destroyed due to neglect .

The missing data-point are a problem, common in all sector that works with data and results in various problems for example : degradation in performance of ML model , problems in analyzing data .

We can use two approach for handling the Missing value , those are complete deletion of all the rows with missing values and putting predicted values in place of missing values,this process is 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 three evaluation metrics,those are Root Mean Square Error(RMSE), Mean Absolute Error(MAE), Mean Absolute Percentage Error(MAPE) .

## 2. Missing data mechanisms : -

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

### *2.1 Missing-Completely-At-Random(MCAR) : -*

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

### *2.2 Missing-At-Random(MAR) : -*

Rate of missingness in the data is dependent on available data. MAR mostly occurs in medical data .

### *2.3 Missing-Not-At-Random(MNAR) : -*

This is when the missing data-points are 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 data-points are mostly impossible in this method, as it depends on the missing data-points also.

## 3. Missing values approaches : -

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

### *3.1 Deletion : -*

In this process all the rows with missing data-point in the data-set are deleted during analysis. Deletion is the simplest process, as it is not needed to estimate the values. The flaws of deletion process are, it gives biased outcomes in analysis .The deletion can be done in two ways, pairwise or list-wise deletion .

#### *3.1.1 List- wise or case deletion : -*

In list-wise deletion every row with missing data is deleted .

#### *3.1.2 Pairwise deletion : -*

To lower the information loss, one can use pairwise deletion rather than list-wise deletion .

### *3.2 Imputation : -*

Imputation involves the process of predicting values to put in place of missing values . The available data-points from the data-set is used to predict the values .Imputation methods can be categorized into single imputation and multiple imputation methods based on the number of values imputed .In accordance to the construction approach used for imputation, these methods can be classified also as statistics-based and machine learning-based (or model-based) methods.

## 4. The imputation techniques implemented : -

### *4.1 Simple imputer : -*

It is a statistic based approach in which a statistic(such as mean) is calculated from each column with missing values and after that calculated statistic are put in place of missing values . In simple imputation, missing data-points are imputed by three strategies those are : - mean,median, or mode from the available values.Mean imputation is the most used method in simple imputation; it puts the calculated mean of the available data-points in place of missing values . Medians is used instead of means for reliability in some cases . For cases where categorical variables are used , the missing data-points are commonly replaced with the most-frequent value of 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.

*4.2 Iterative-Imputer : -*

Iterative imputer is a multivariate imputer ,that means it estimates each features from all others. A imputation technique 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 is done for *max_iter* imputation rounds . The result of final imputation round is then returned .

*4.3 Multiple-Imputation-by-Chained-Equations(MICE) : -*

MICE works on assumption that the variables used in the imputation and the mechanism of the missing data-points is Missing At Random(MAR), which states that the possibility that a value is missing depends only on observed values and not on unobserved values. Implementing MICE when missingness of the data-points are not of MAR, it mostly gives in biased outcomes.

*4.4 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 missing data points. Each missing data-point are imputed from the mean of the ‘k’-neighbors found in the data-set using distance measures .Some of example the distance measures are : - Manhattan distance, Cosine distance, Hamming distance and Euclideandistances is used in KNN algorithm.

### *4.5 MissForest Imputer : -*

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

First , the missing data-points are filled using median/mode imputation method. Then we mark the imputed values as 'Predict' and the others as training rows, which later are fed into a Random Forest model trained for predicting the missing data-points .This process repeats itself several times , each iteration gives better results than the iterations done before . Iterations continue until maximum iterations criteria is met .

## 5. Performance Evaluation Metrics : -

For evaluation of performance of different imputation techniques there are various criteria . In this part we will discussthe criteria used in this project those are , Root Mean Squared Error(RMSE) , Mean Absolute Error(MAE), and Mean absolute percentage error(MAPE) .

*5.1 Root-Mean-Square-Error(RMSE) : -*

Root-Mean-Square-Error calculates mean in the difference of the predicted data-points and of the original data-points .

**Frm .1** Formula of RMSE

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

$\mathrm{RMSD}$ = root-mean-square deviation\error
$i$ = variable i
$N$ = number of non-missing data points
$x_i$ = actual values
$\hat{x}_i$ = imputed values

*5.2 Mean-Absolute-Error(MAE) : -*

Mean-Absolute-Error calculates the average of the absolute difference of the predicted data-points and the original data-points .

**Frm .2** Formula of MAE

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

$\mathrm{MAE}$ = mean absolute error
$y_i$ = imputed value
$x_i$ = true value
$n$ = total number of data points

## *5.3 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

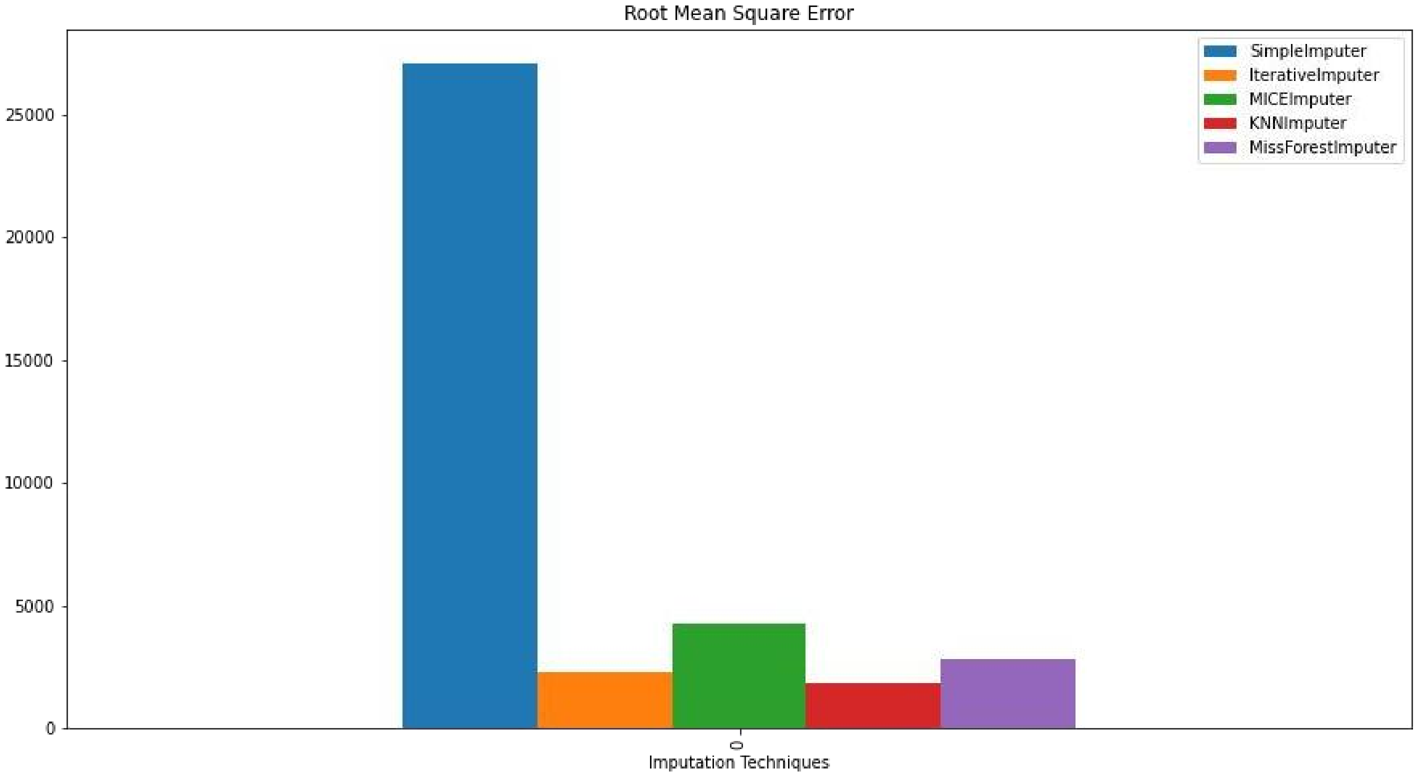
$F_t$ = imputed value

## 6. Results : -

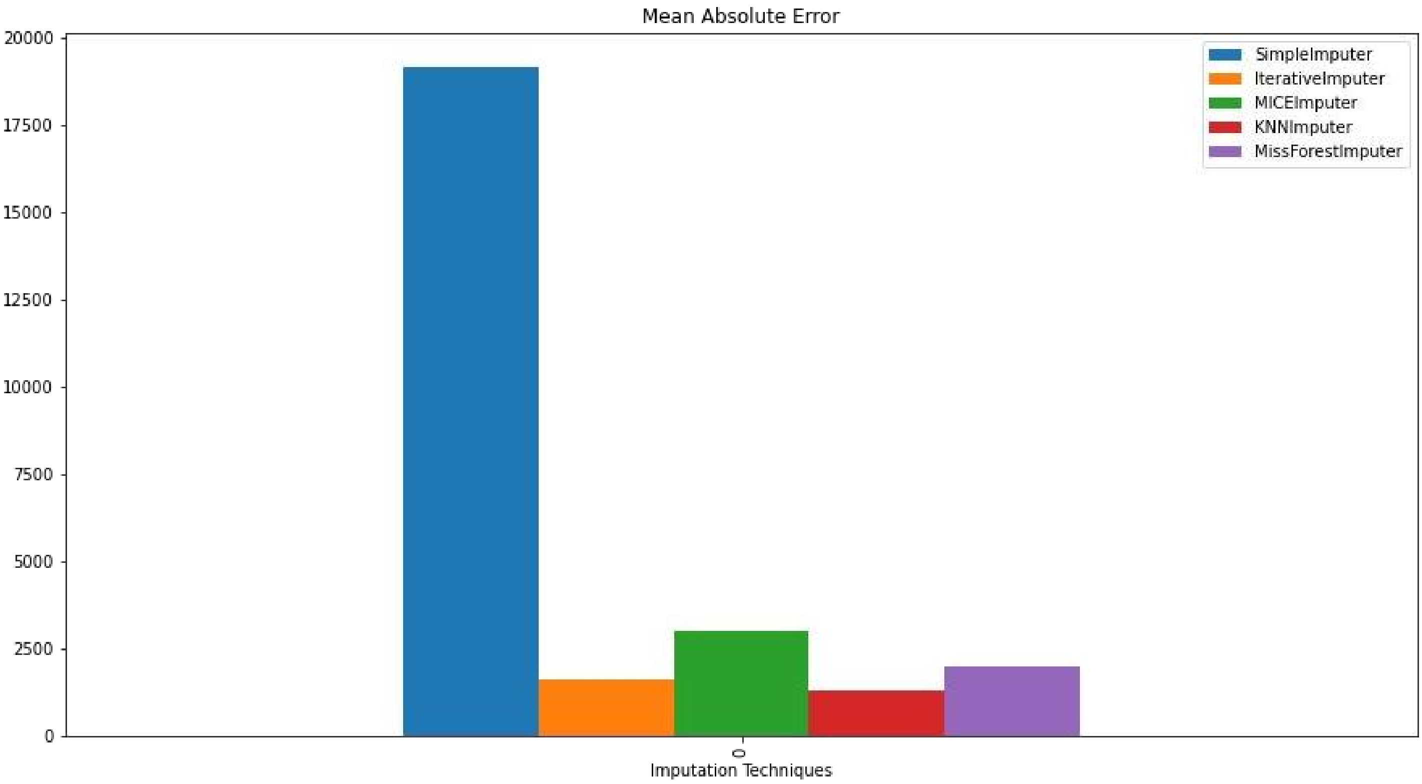
### *6.1 Table 1 : - Results of the imputation techniques .*

| | | 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 | 0.045955934540470623 |

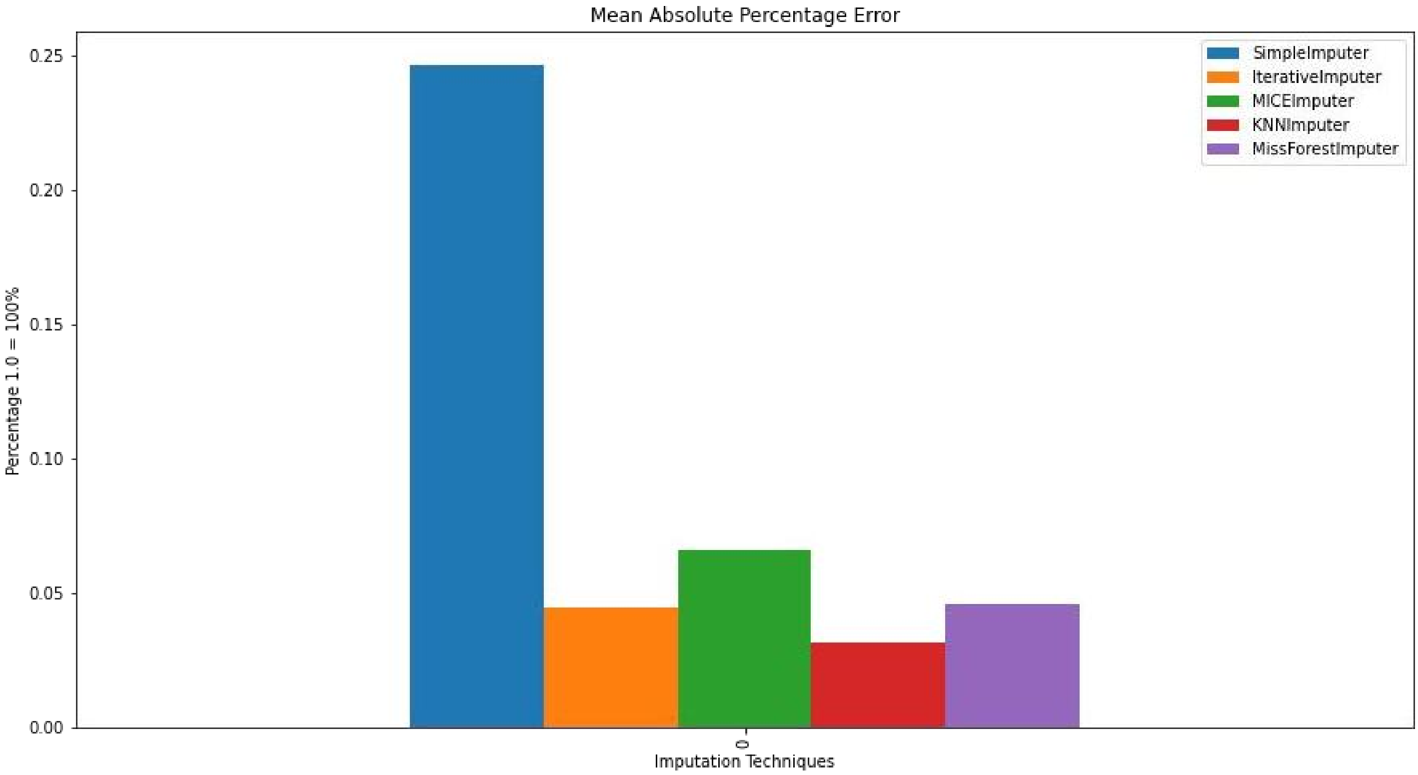
*6.2 Fig .1 : - Graphical comparison of Root Mean Square Error(RMSE) of the imputation techniques .*




*6.3 Fig .2 :-Graphical comparison of Mean Absolute Error(MAE) of the imputation techniques.*



*6.4 Fig .3 : - Graphical comparison of Mean Absolute Percentage Error(MAPE) of the imputation techniques .*




- ✧ Figure 1 shows the simple imputer has highest RMSE of 27097.25 and knn imputer lowest RMSE of 1856.18 .
- ✧ Figure 2 shows the simple imputer has highest MAE of 19160.65 and knn imputer got lowest MAE of 1312.5 .
- ✧ Figure 3 shows the simple imputer has highest MAPE with 0.2467(or 24%) and knn has lowest MAPE with 0.0312(or 3%) .

## 7. 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 .