International Journal of Knowledge Based Computer Systems
6 (2), December 2018, 08-22
http://www.publishingindia.com/ijkbcs/

An Effective Method to Estimate Missing Value for Heterogonous Dataset

Priyakshi Mahanta[1*], Gulzar Ahmed Choudhury[2] and Tapash Dey[3]
[1]Assistant Professor, Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, India. Email: priyakshimahanta@dibru.ac.in
[2]Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, India. Email: gulzar5@gmail.com
[3]Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, India. Email: tapashdey1991@gmail.com
*Corresponding Author

Abstract: Missing value estimation renders a probable state in which it is required to predict a missing value in a data set. It gets induced due to various reasons but the deal is to find a value for that particular cell. To justify a conclusive result in contrast to the original value various techniques are used. It does not assure us to choose a specific method applying for a data set because different techniques will yield different result for different data set that may not trigger an authentic value. So we deployed a technique confined of various other methods in it by which we can choose any of the methods in it to acquire a missing values. The main goal of this paper is the technique we proposed by assembling different methods get an efficient value. We have applied the proposed method in 6 different data set and validate the result using 4 validation techniques that can accord us with a most precise value or result.

Keywords: Data set, Methods, Missing value, Techniques.

I. Introduction

DATA is said to be the accumulation of binary information that can be transformed into knowledge for some meaningful purpose and can be supervised by humans where one can store, modify or retrieve a set of information as required. Missing data is always been a concern and mostly we settle with an unknown value or neglect it as a traditional way. But today with the trend of computerization researchers brought up some of the efficient techniques that help us to determine a missing value nearly as possible.

It's very much classy to say the most expensive thing in today world is Data, as everyday we are entering into the pool of data featuring our daily living standard. Moreover, it is ubiquitous, lucrative as well as life depending because Data losses can be life threatening, huge losses of money, or loss of an identity of an individual. The need to estimate missing value in a data set is important due to the fact that it contains important fact and features that cannot be biased. Missing value cannot be obscure or unambiguous relating to the real value [7] that may shallow the result. Missing value can occur due to technical or manual actions- Files deleting accidently or mistakenly Power failures without a backup can wipe out the data viruses can corrupt the system leading to blackout. Mechanical damages of hardware can vanish the data Lack of adequate data while uploading into the system remains null.

II. Our Contribution

We claim the following contributions in this paper. We introduce a method effective for missing value estimation measurement. We propose an approach to estimate missing value estimation using FILTRATION method comparing other methods. We develop a spanning tree based method to extract the potential network modules.

III. Literature Review

Missing value is an ideal phenomenon occurs due to various reasons. The trends of digitalization suppress the need of pen and paper rather than everything get compressed to memories. The transformation from traditional to digitization allowed to upload data into computers storing in hard disk or other form of memory, perhaps there also a problem of missing data in records related to business, education, health, etc. for that there are ways to determine the missing value. In this paper we jolt different techniques and methods to deal with the missing values keeping in mind that accuracy to be maintained without much fluctuation which can be irrelevant to any source and cannot be accepted. The major problem is the missing [5] value and to understand its pattern for that we deal different error detection technique and we being following our work on different data set with all sort of possibility. In this paper we dealt with different

techniques to find out the optimal result. The process starts with the data set analysis, that can be said the pattern of missing value is to be determined to get the knowledge to produce a systematic work to be followed. We got to know that while calculating values for;

Row Average or Column Average the value varies in a wide range which does not fit for the result, as we don't know the best from it. We followed normalization which ranges the value from 0 to 1. Normalization of the data gives us a new dimension for applying other methods.

Distance Imputation [1]: Minimum distance imputation [6] is the method that followed by calculating different distances and then choosing the nearer distance from the missing cell of the test data. For doing so first we extract 80 percent data row wise from missing value data set and rest 20 percent value having the maximum value get discarded. We can take an example that if information needed of a student than father name or mother name is not sufficient to trace other data for the need of calculation percentage we can calculate as it is directly in the numerical form relating some other information. There a problem occurs after the evaluation variation occurs sharply regarding different data set that concludes that we are still not able to judge a value for a particular missing cell.

Fuzzy [3]: We move to using Fuzzy method as the process remains same extracting 80 percent data and 20 percent are left behind due to the wide variance of missing value. The speciality of Fuzzy is that it can take many rows while distance imputation can handle only one row. To calculate distance we use Pearson distance formula to find out the minimum distance. Moreover, after the distance calculation percentage of weighted is calculated considering the minimum distance and taking the sum of all minimum distance. Then we sort the weight and pick the min weight with the max value that means min distance will carry max value and vice versa. Then dividing with 100 provides us different values and when we add up the values we yield the value for that particular missing cell.

Knn: Knn [1] is the best method considered by many so we Knn to compare the values of Fuzzy and others. As we know Knn deals with K values that means it gives us boundaries to absorb the values falling under it. K=1 will be idle or zero cause it itself be the value to be considered. We pull out the Pearson distance formula again and calculate the min distance of the neighbours with K values ranging from 3 until a huge deviation does not appear. As we found out the efficient value with respect to K, we impute that for the missing cell. We observe that for some data set Fuzzy is far better than Knn and in some other data set Knn have the upper hand which again involve us in a uncertainties of determining the possible judged nearer value for the missing cell.

Filter: We finally used the filter method combining all the methods to comprise an ultimate result among all. Filtration method consists of two ways as described below. Filtration combined two steps as pre-imputation 1 and pre-imputation 2 and then a judged value is obtained. Basically in pre-imputation 1 we discard 20 percent with higher missing value and we create a test data set with rest 80 percent missing data set. We need to view the data set about the pattern of missing value as well as the form that is categorical or numerical. If we find the categorical value we apply mode operation or else we use median for the numerical value. Both are imputed in order to fit in the missing value and to form a complete data set without any missing value. In pre-imputation 2 the functions or methods we used such as Knn, Fuzzy, row average, column average, distance and other functions can be used. Converging all the values of functions we use mean or median and impute the value for that missing cell and forms a complete data set after comparison. Finally, combining both of them we filter out the best and accurate value that fits the value of that missing cell and thus finally having a solid and nearer output makes the incomplete data set to complete data set.

IV. Validation Measures

Validation measures for different techniques to obtain a judgemental treatment for the methods and techniques used in this paper for the estimation of missing values. We thoroughly have preceded over the processes but still for the reality we need structures that proofs our statement in a good formation. In this paper we used validation measures using four techniques as RMSE, MAE, SE, RAE. Other processes can be used depending upon own interest and satisfaction. We will discuss the variation developed in the process and also deploy a table for some better understanding and convincement.

RMSE (Root Mean Square Error) [3]: Is a validation process that abruptly helps in determining a result to compare the best. In this paper we applied RMSE among all 6 techniques used for estimating missing value and thereby we see fluctuation occurs in graphs row average was near but finally the filtration process makes the narrow path to prove that it's the accurate in acquiring a less error value and thus can be accepted for further processes. RMSE is a frequently used measure to calculate the difference between predicted values of a data set's result and the values actually observed that is being accepted for calculation.

$$\text{"RMSE"} = (((X \text{ "obs"}, i) - (X \text{ "model"}, i))^2)/n \quad (1)$$

MAE (Mean Absolute Error) [3]: Is a measure of difference between two continuous variables. Assume X and Y are variables of paired observations that express the observed and predicted variables. It is another error detection method which shows more concentrated values regarding the missing values, it means more minimal values are obtained and compromising less error with more accuracy as followed by RMSE. It also indicates that the process followed for missing value estimation is gaining weight in itself evolving a good result. The formula as mentioned below;

$$MAE = ((i = 1)^n |y_i - x_i|)/n \quad (2)$$

SE (Square Error) [3]: Here, X refers to a vector of n predictions, and Y refers to the vector of observed values of the variable

being predicted, then the within-sample or the data set will be; SE of the predictor is computed as:

$$SE = (i = 1)^{n} |x_i - y_i| \quad (3)$$

RAE (Relative Absolute Error) [3]: The relative absolute error is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total squared error. Thus, the relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor. V is the original value and V Approx is the predicted value than RAE can be calculated as:

$$RAE = (|V - VApprox|)/V \quad (4)$$

Based on the papers we reviewed that helped us and allowed us to perceive our own dimension to work through it. As it was accepted by many research scholars and other institution that Knn is found to be one of the best methods in dealing with missing value. Knn deals with the K value that is considering its number of neighbours and to retrieve the distance and the other way to calculate is to make cluster and to compare the values for the missing cell. Alireza Farhangfar, Lukasz Kurgan, Jennifer Dy had proposed a filtration technique comparing different methods and finding out the marginal value that showed accuracy among all. We need quality of values in a data set to deal, according to them in a data set they dealt with row column which showed different column and row shows a good result mostly because they together combine and give a result out of the two compare methods and obviously it will choose from the row column result. So considering their method we decided to add a pre-imputation 2 method in our proposed technique such that to get a solid quality of result. We applied many different methods to compare out of which a value is obtained to fulfil the missing value. The best part is that it only chooses a minimal value among all that is nearer to the missing value and that some methods may differ in result but it possesses the minimum value out of all methods to fulfil the best result. All the methods are described in Table I.

Table I: Methods Used

Methods	Normalization	Data Set Used	Types of Data	Validation Technique	Kind of Methods
Column Average	Yes	Balance scale	Numerical	RMSE MAE RAE SE	Average computing
Row Average	Yes	Flag	Numerical	RMSE MAE RAE SE	Average computing
Knn	Yes	Glass	Numerical	RMSE MAE RAE SE	Classification computing
Fuzzy	Yes	Hayes-roth	Numerical	RMSE MAE RAE SE	Soft computing
Filter	Yes	Machine	Categorical or Numerical	RMSE MAE RAE SE	Mix computing

V. Proposed Method

The proposed method of ours trails an extensive study over the research paper and super guidance that sums up in researching something new among those existing theories, methods and ideas. Our proposed method is genuinely a comparison based technique that means we carry out a lot of processes to find the missing value of a missing cell among that we choose the most narrowly converged value. We plot the idea of how our method works by flow chart in Fig. 1, 2 and 3.

Filter method is a combination of multiple function that has been used in this project and it's also possible to add some other function if required that leads us to comprise an ultimate solid value among all. Filtration method consists of two very important stages as described below, Pre-imputation 1 and Pre-imputation 2 and then combining both us follows the last stage and obtaining a minimum value for the missing cell. In pre-imputation 1 we discard 20 percent rows possessing higher missing value because that may create a wide distribution of value that may biased the result. Thereafter, we create a test data set with remaining 80 percent rows of the data set. We need to visualize and study the pattern of the data as well as the missing value. We have to also determine the form of data exists, that is in categorical or numerical form. If we discover categorical value in contrast we apply mode operation or else we use median to convert it to the numerical value. In order to fill the missing value and to form a complete data set without much variation in result, we are able to predict a nearer value that may fill the blank space up to some possible extent that can help for analysis. Pre-imputation 2 follows pre-imputation 1 that compares the functions or methods we used such as Knn, Fuzzy, row average, column average, distance. Some other functions can also be added if necessary for concise matters. Calculating the values obtained from the functions using Pearson distance formula followed by using mean or median. The lowest possible value is secured by comparing among all methods and function that falls nearer to the missing value. The value approved for that missing cell and forms a complete data set. Finally, combining both pre-imputation 1 and pre-imputation 2 we summarize out the best and accurate value that fits the value of that missing cell. Filtration the final method begins with the missing value dataset followed by the pre-imputation measures that is pre-impuatation 1 with involving mean mode operation than the methods or functions applied. Then again pre-imputation draws in the mean calculation and link to the filter section so that a value can be retrieved. If it is not accepted with the justified value than it is rejected or else it furnishes the blank space with a value that can be accepted for further use and deliberation. The algorithm is described in Algorithm 1.

A. Flow Chart

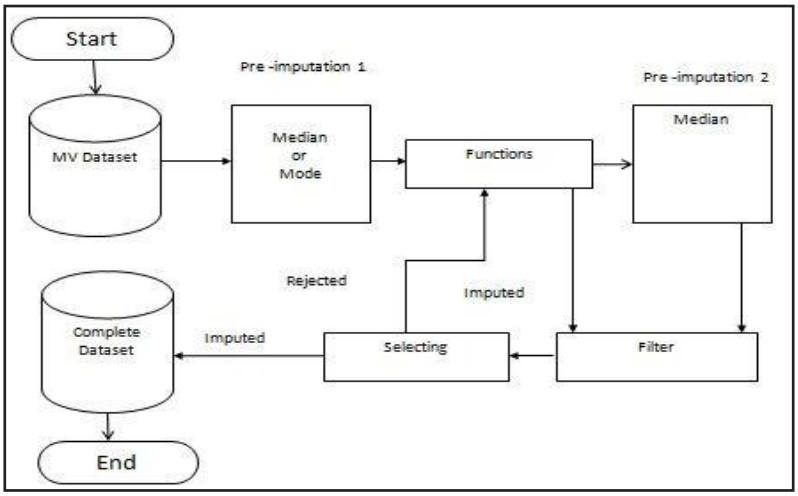


Fig. 1: Filteration

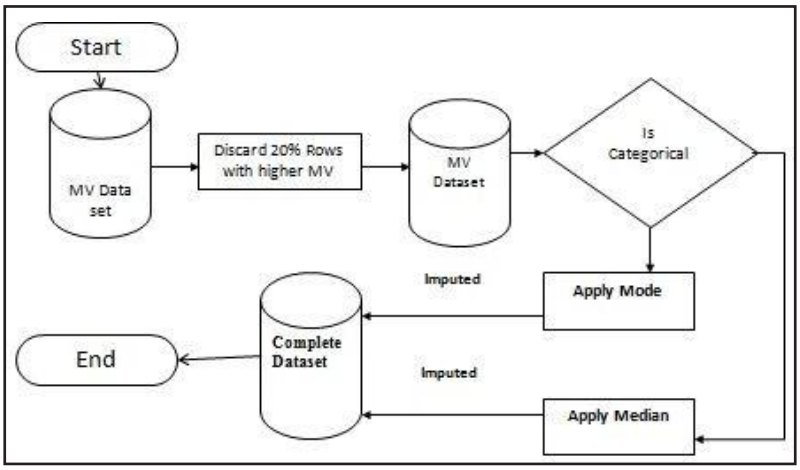


Fig. 2: Pre-imputation 1

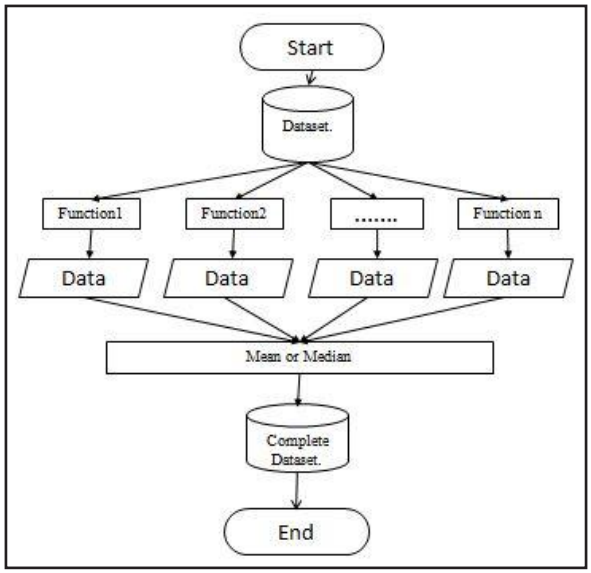
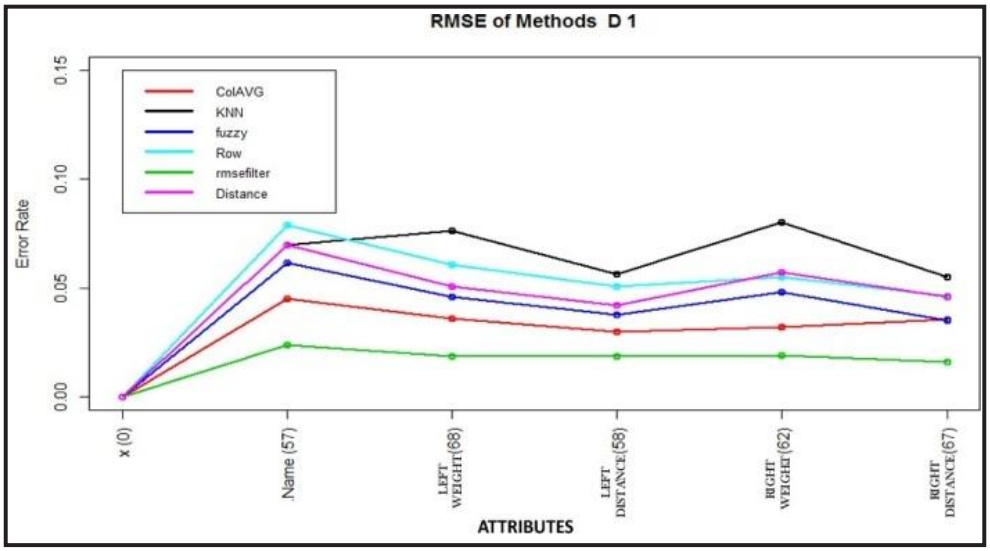
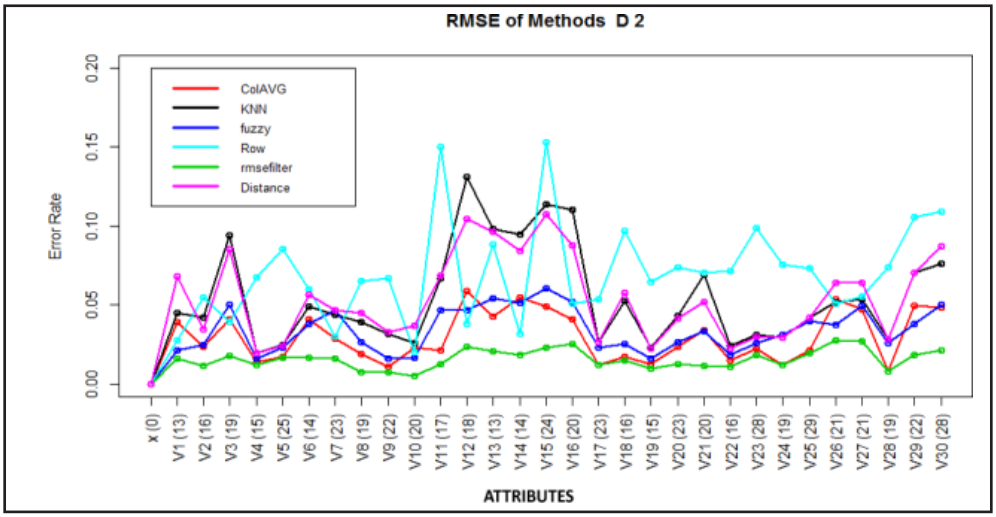
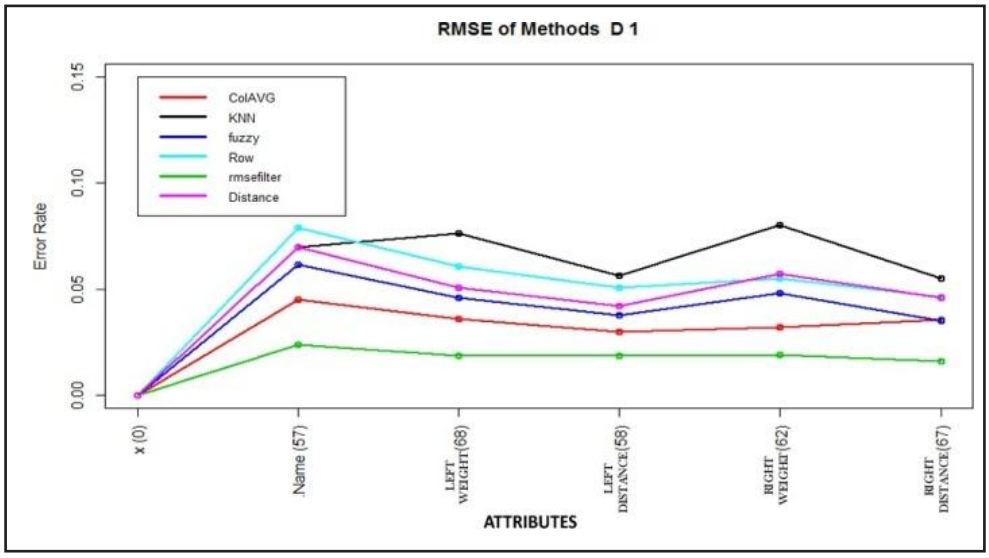


Fig. 3: Pre-imputation 2

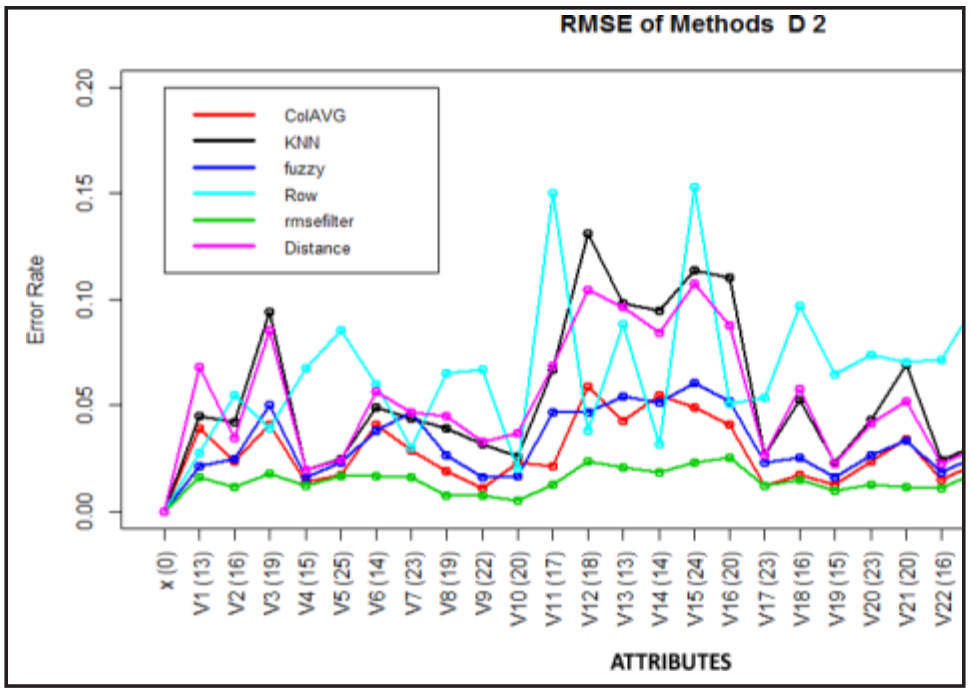
B. Graphs

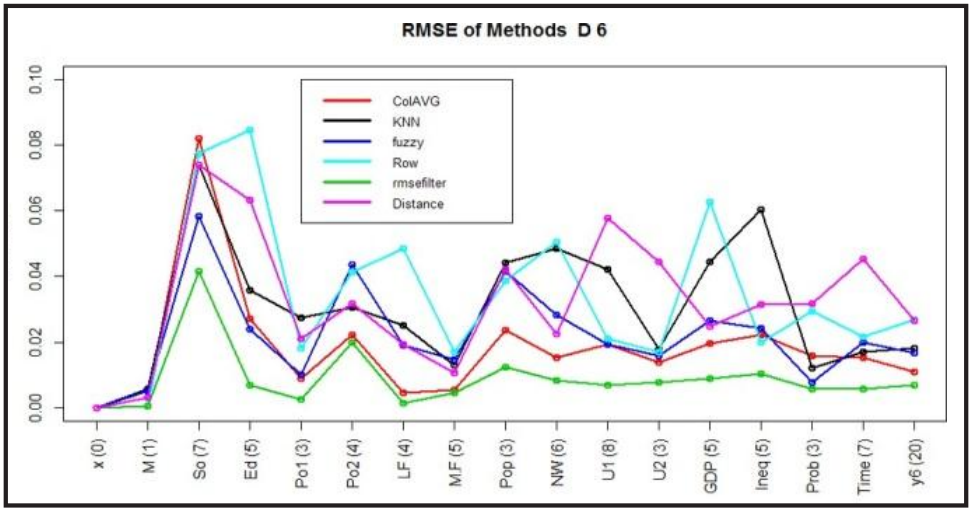


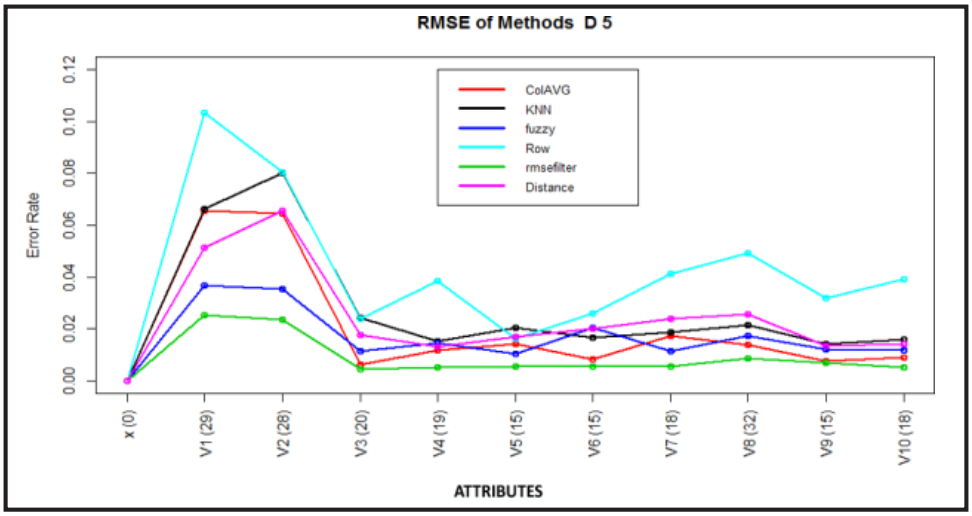




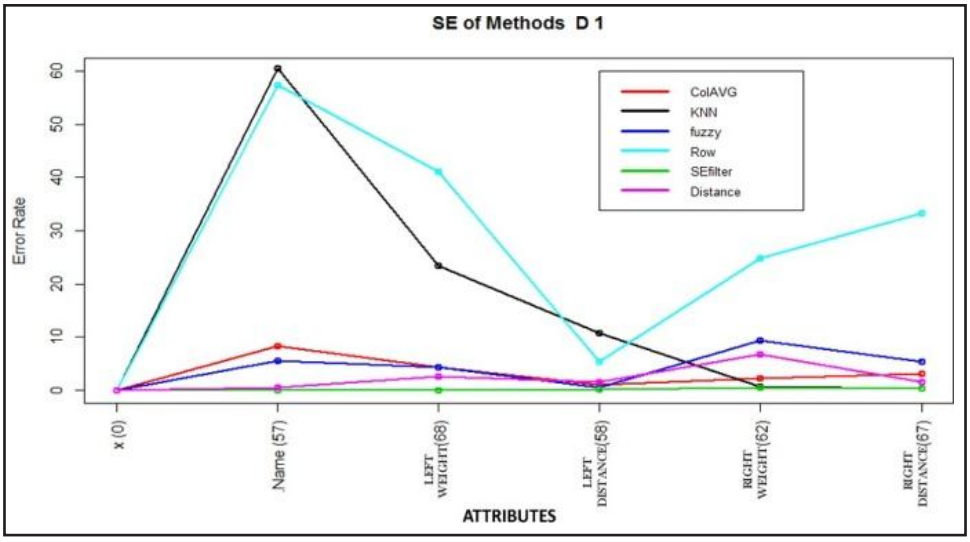
Contd.

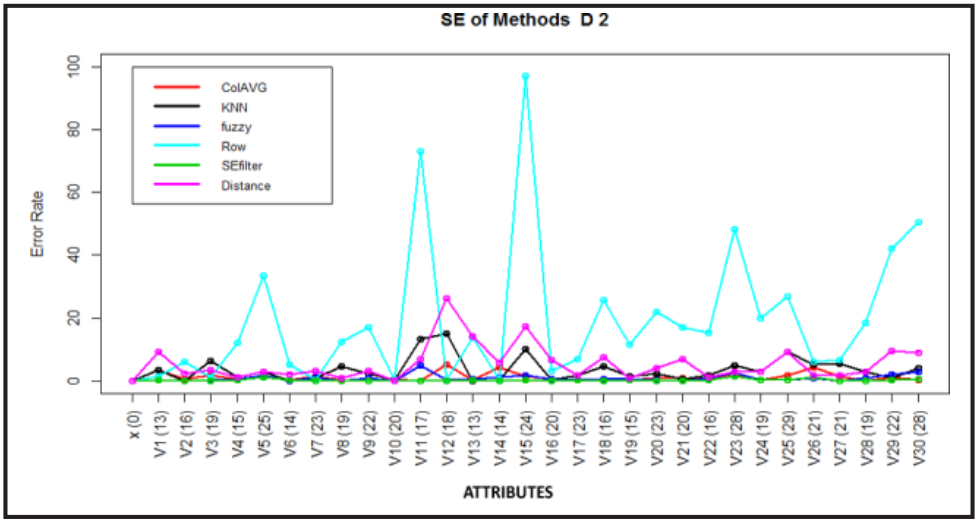

RMSE of Methods D 2
ColAVG
KNN
fuzzy
Row
rmsefilter
Distance
Error Rate
0.00
0.05
0.10
0.15
0.20
x (0)
V1 (13)
V2 (16)
V3 (19)
V4 (15)
V5 (25)
V6 (14)
V7 (23)
V8 (19)
V9 (22)
V10 (20)
V11 (17)
V12 (18)
V13 (13)
V14 (14)
V15 (24)
V16 (20)
V17 (23)
V18 (16)
V19 (15)
V20 (23)
V21 (20)
V22 (16)
ATTRIBUTES

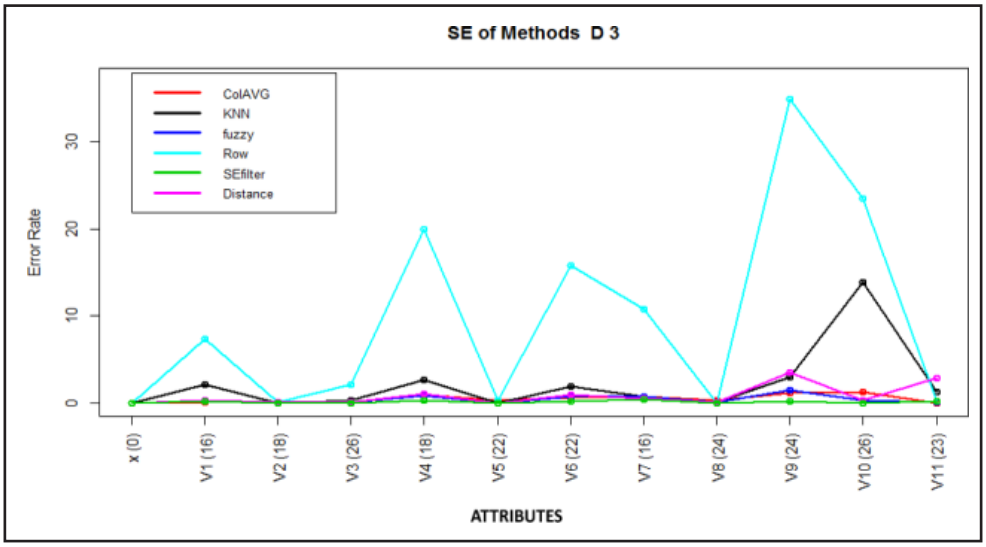

RMSE of Methods D 6
ColAVG
KNN
fuzzy
Row
rmsefilter
Distance
0.00
0.02
0.04
0.06
0.08
0.10
x (0)
M (1)
So (7)
Ed (5)
Po1 (3)
Po2 (4)
LF (4)
M.F (5)
Pop (3)
NW (6)
U1 (8)
U2 (3)
GDP (5)
Ineq (5)
Prob (3)
Time (7)
y6 (20)


RMSE of Methods D 5
ColAVG
KNN
fuzzy
Row
rmsefilter
Distance
Error Rate
0.00
0.02
0.04
0.06
0.08
0.10
0.12
x (0)
V1 (29)
V2 (28)
V3 (20)
V4 (19)
V5 (15)
V6 (15)
V7 (18)
V8 (32)
V9 (15)
V10 (18)
ATTRIBUTES

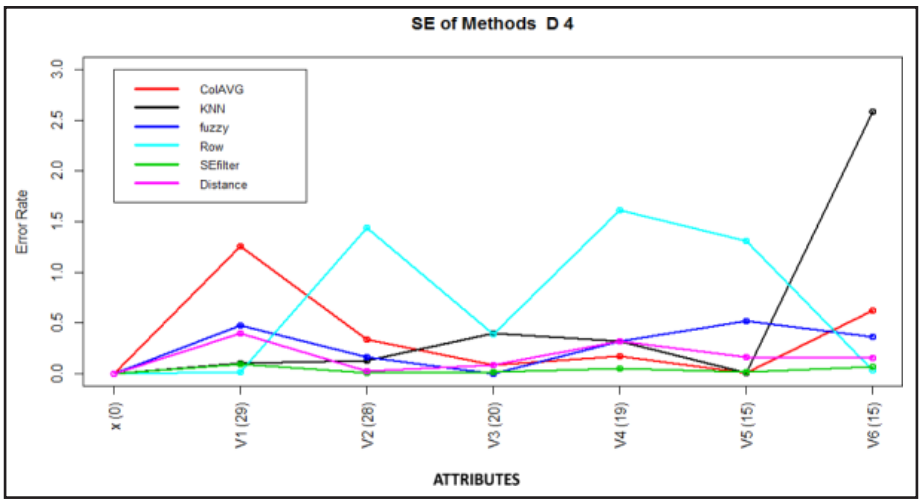
Contd.

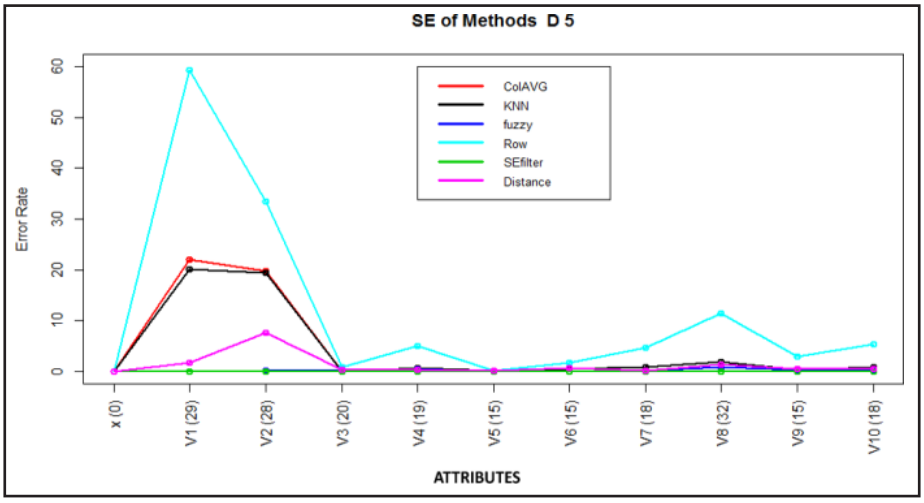

SE of Methods D 1
Error Rate
0
10
20
30
40
50
60
ColAVG
KNN
fuzzy
Row
SEfilter
Distance
x (0)
Name (57)
LEFT WEIGHT(68)
LEFT DISTANCE(58)
RIGHT WEIGHT(62)
RIGHT DISTANCE(67)
ATTRIBUTES

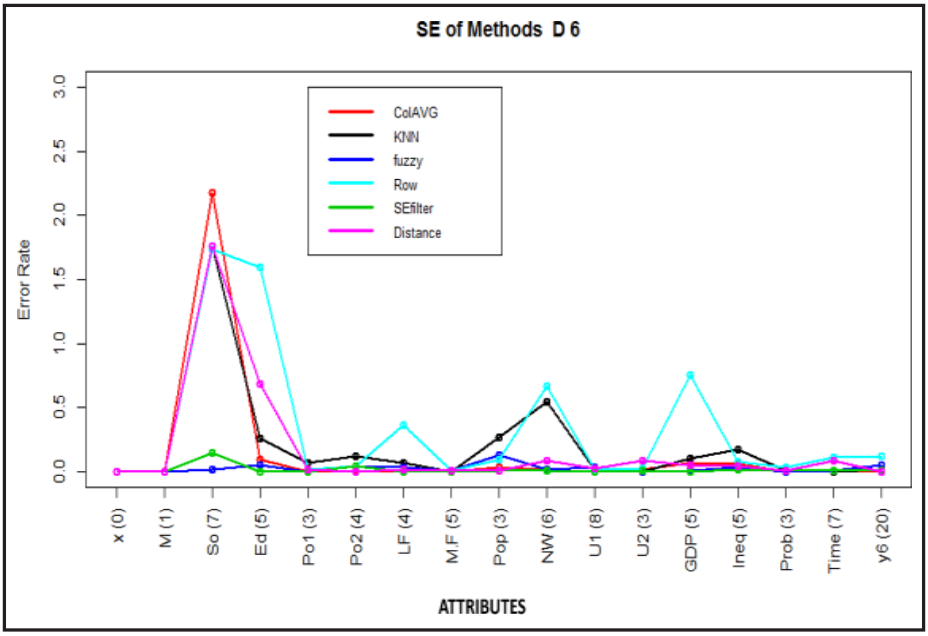

SE of Methods D 2
Error Rate
0
20
40
60
80
100
ColAVG
KNN
fuzzy
Row
SEfilter
Distance
x (0)
V1 (13)
V2 (16)
V3 (19)
V4 (15)
V5 (25)
V6 (14)
V7 (23)
V8 (19)
V9 (22)
V10 (20)
V11 (17)
V12 (18)
V13 (13)
V14 (14)
V15 (24)
V16 (20)
V17 (23)
V18 (16)
V19 (15)
V20 (23)
V21 (20)
V22 (16)
V23 (28)
V24 (19)
V25 (29)
V26 (21)
V27 (21)
V28 (19)
V29 (22)
V30 (28)
ATTRIBUTES


SE of Methods D 3
Error Rate
0
10
20
30
ColAVG
KNN
fuzzy
Row
SEfilter
Distance
x (0)
V1 (16)
V2 (18)
V3 (26)
V4 (18)
V5 (22)
V6 (22)
V7 (16)
V8 (24)
V9 (24)
V10 (26)
V11 (23)
ATTRIBUTES

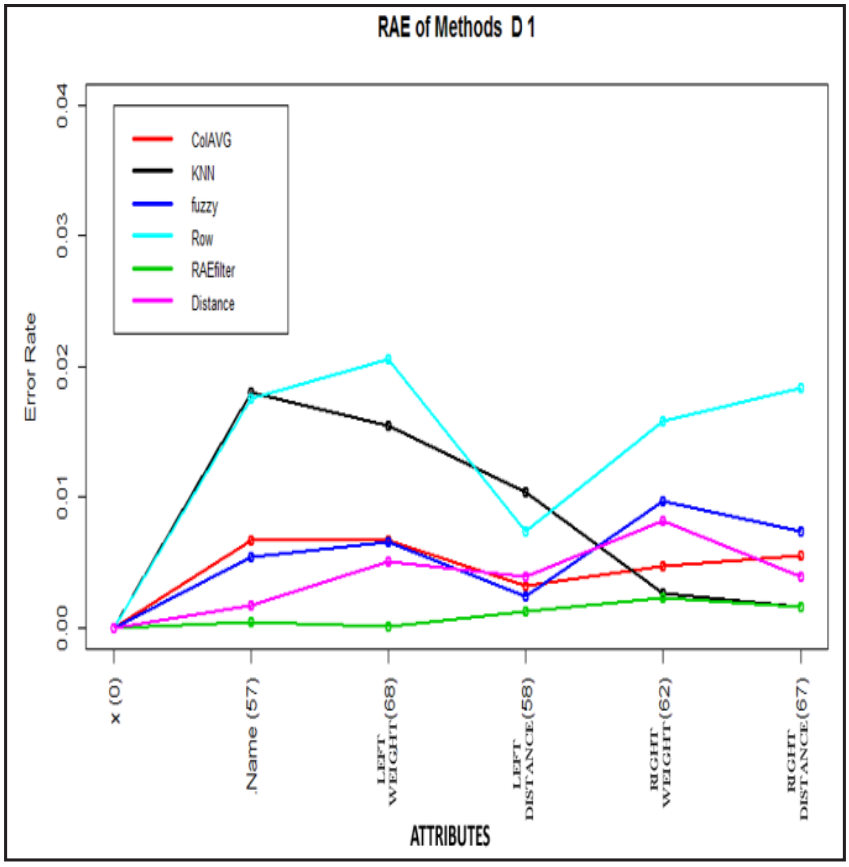
Contd.

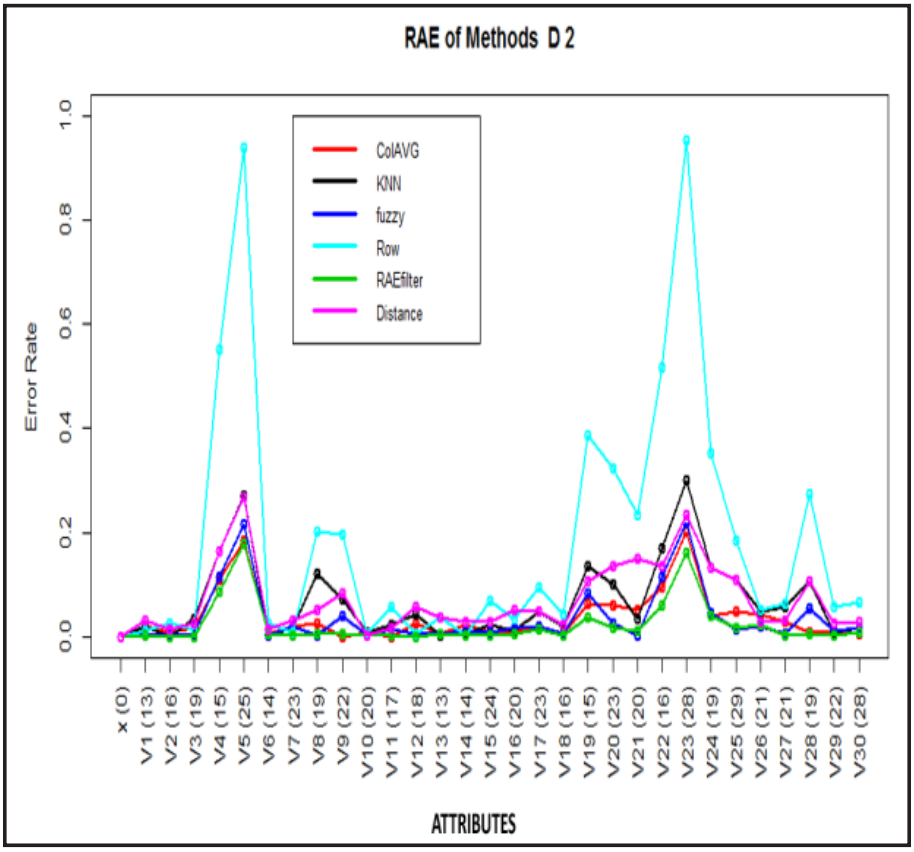

SE of Methods D 4
ColAVG
KNN
fuzzy
Row
SEfilter
Distance
Error Rate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x (0)
V1 (29)
V2 (28)
V3 (20)
V4 (19)
V5 (15)
V6 (15)
ATTRIBUTES


SE of Methods D 5
ColAVG
KNN
fuzzy
Row
SEfilter
Distance
Error Rate
0
10
20
30
40
50
60
x (0)
V1 (29)
V2 (28)
V3 (20)
V4 (19)
V5 (15)
V6 (15)
V7 (18)
V8 (32)
V9 (15)
V10 (18)
ATTRIBUTES

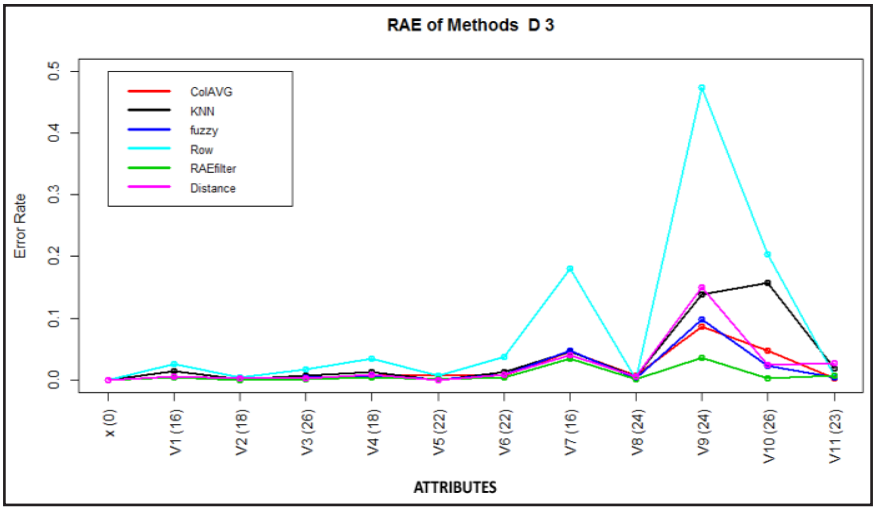

SE of Methods D 6
ColAVG
KNN
fuzzy
Row
SEfilter
Distance
Error Rate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x (0)
M (1)
So (7)
Ed (5)
Po1 (3)
Po2 (4)
LF (4)
M.F (5)
Pop (3)
NW (6)
U1 (8)
U2 (3)
GDP (5)
Ineq (5)
Prob (3)
Time (7)
y6 (20)
ATTRIBUTES

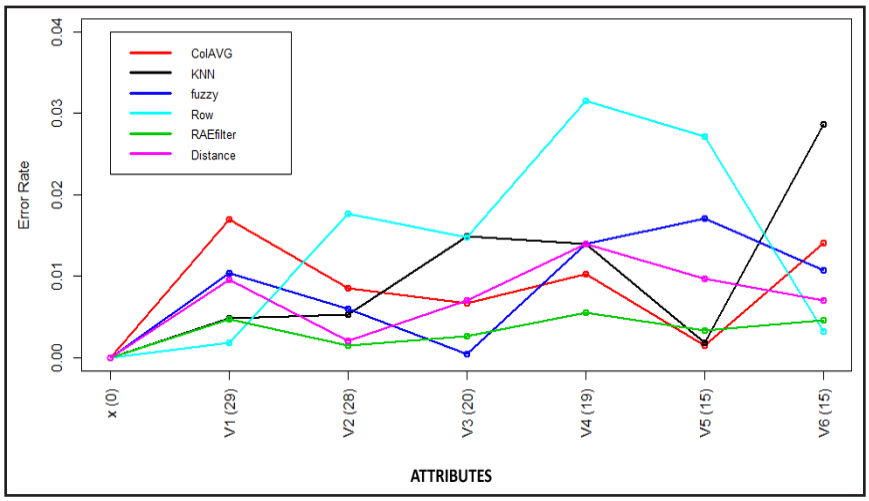
Contd.

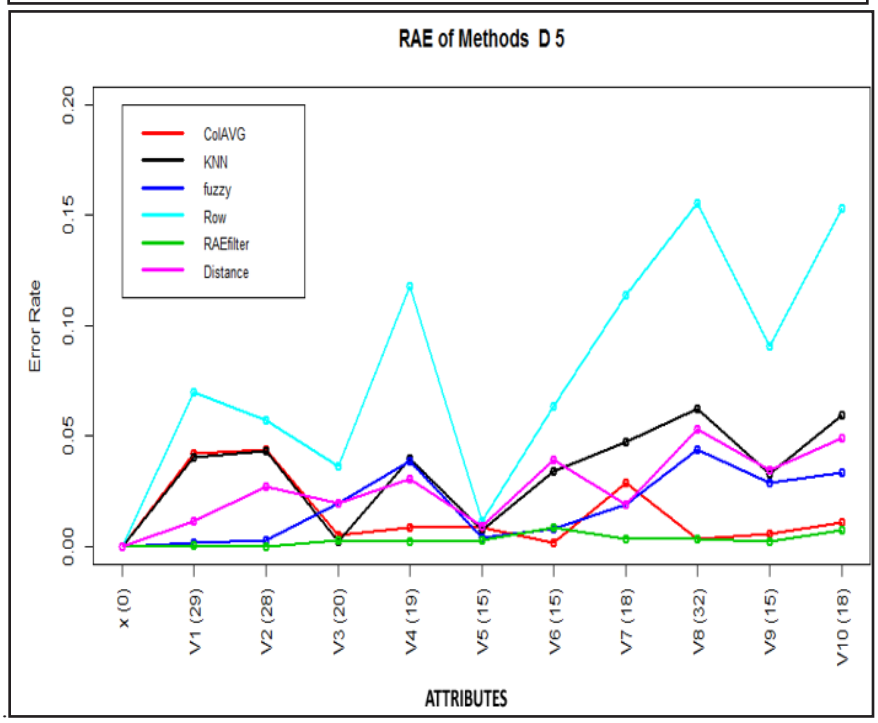

RAE of Methods D 1
ColAVG
KNN
fuzzy
Row
RAEfilter
Distance
Error Rate
0.00
0.01
0.02
0.03
0.04
x (0)
Name (57)
LEFT WEIGHT(68)
LEFT DISTANCE(58)
RIGHT WEIGHT(62)
RIGHT DISTANCE(67)
ATTRIBUTES


RAE of Methods D 2
ColAVG
KNN
fuzzy
Row
RAEfilter
Distance
Error Rate
0.0
0.2
0.4
0.6
0.8
1.0
x (0)
V1 (13)
V2 (16)
V3 (19)
V4 (15)
V5 (25)
V6 (14)
V7 (23)
V8 (19)
V9 (22)
V10 (20)
V11 (17)
V12 (18)
V13 (13)
V14 (14)
V15 (24)
V16 (20)
V17 (23)
V18 (16)
V19 (15)
V20 (23)
V21 (20)
V22 (16)
V23 (28)
V24 (19)
V25 (29)
V26 (21)
V27 (21)
V28 (19)
V29 (22)
V30 (28)
ATTRIBUTES

Contd.


RAE of Methods D 3
ColAVG
KNN
fuzzy
Row
RAEfilter
Distance
Error Rate
0.0
0.1
0.2
0.3
0.4
0.5
x (0)
V1 (16)
V2 (18)
V3 (26)
V4 (18)
V5 (22)
V6 (22)
V7 (16)
V8 (24)
V9 (24)
V10 (26)
V11 (23)
ATTRIBUTES


ColAVG
KNN
fuzzy
Row
RAEfilter
Distance
Error Rate
0.00
0.01
0.02
0.03
0.04
x (0)
V1 (29)
V2 (28)
V3 (20)
V4 (19)
V5 (15)
V6 (15)
ATTRIBUTES


RAE of Methods D 5
ColAVG
KNN
fuzzy
Row
RAEfilter
Distance
Error Rate
0.00
0.05
0.10
0.15
0.20
x (0)
V1 (29)
V2 (28)
V3 (20)
V4 (19)
V5 (15)
V6 (15)
V7 (18)
V8 (32)
V9 (15)
V10 (18)
ATTRIBUTES

Contd.

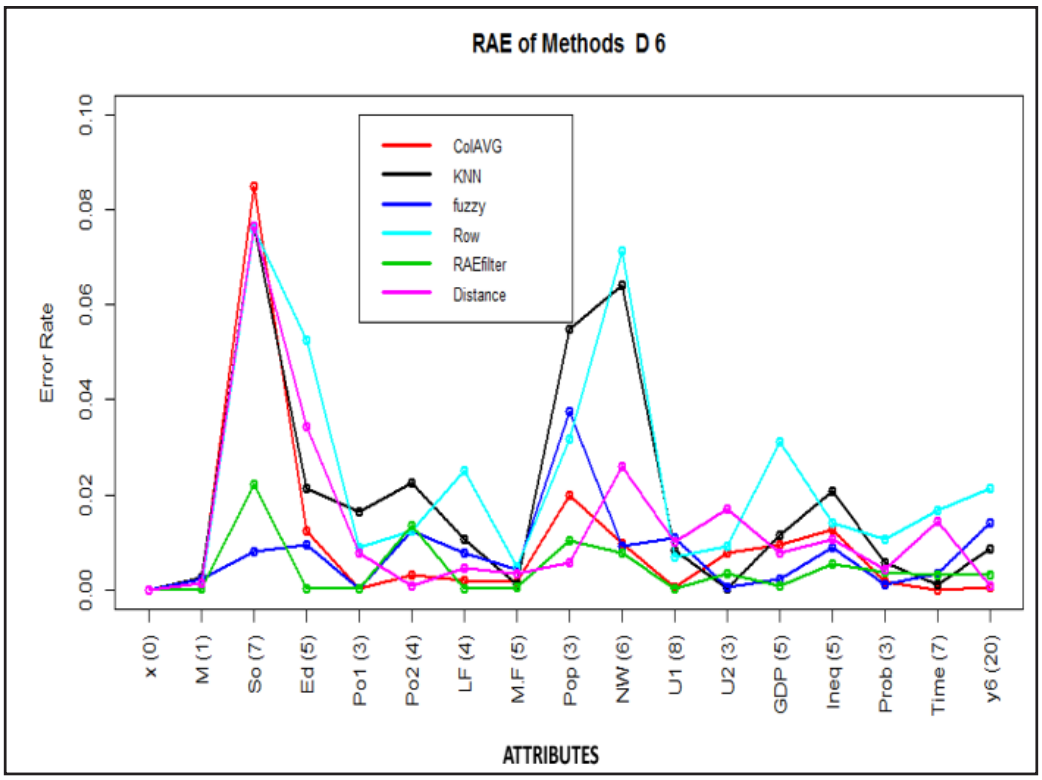
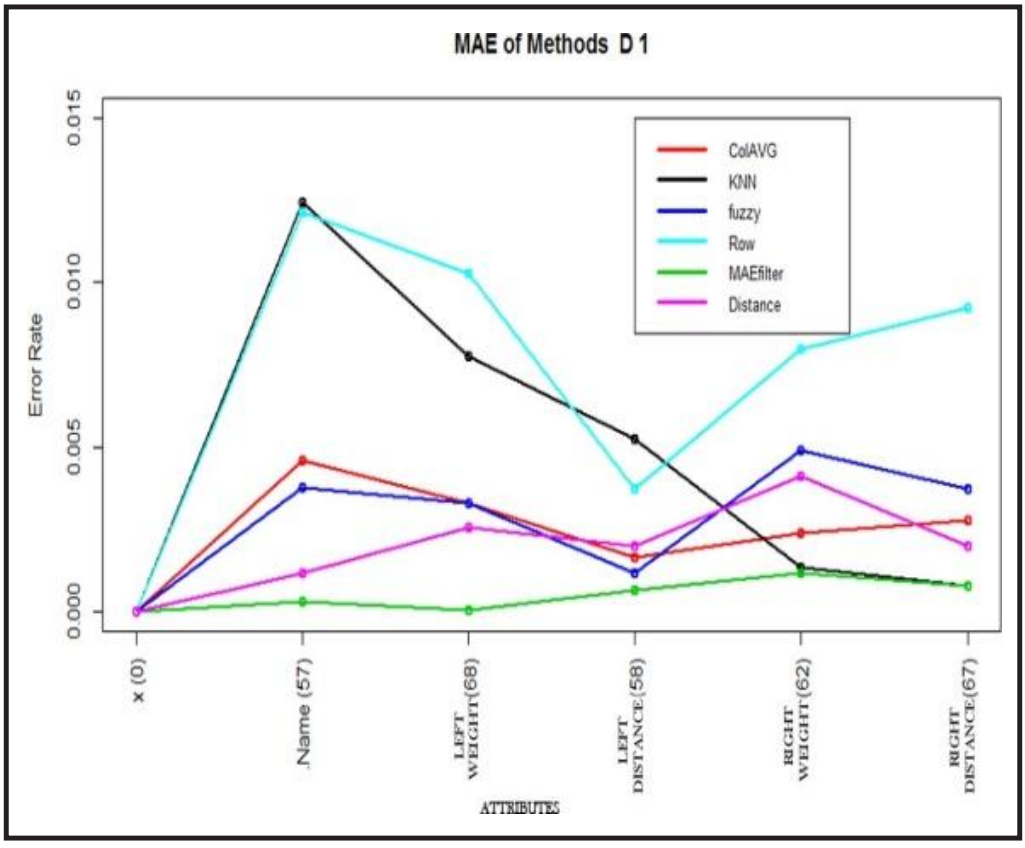
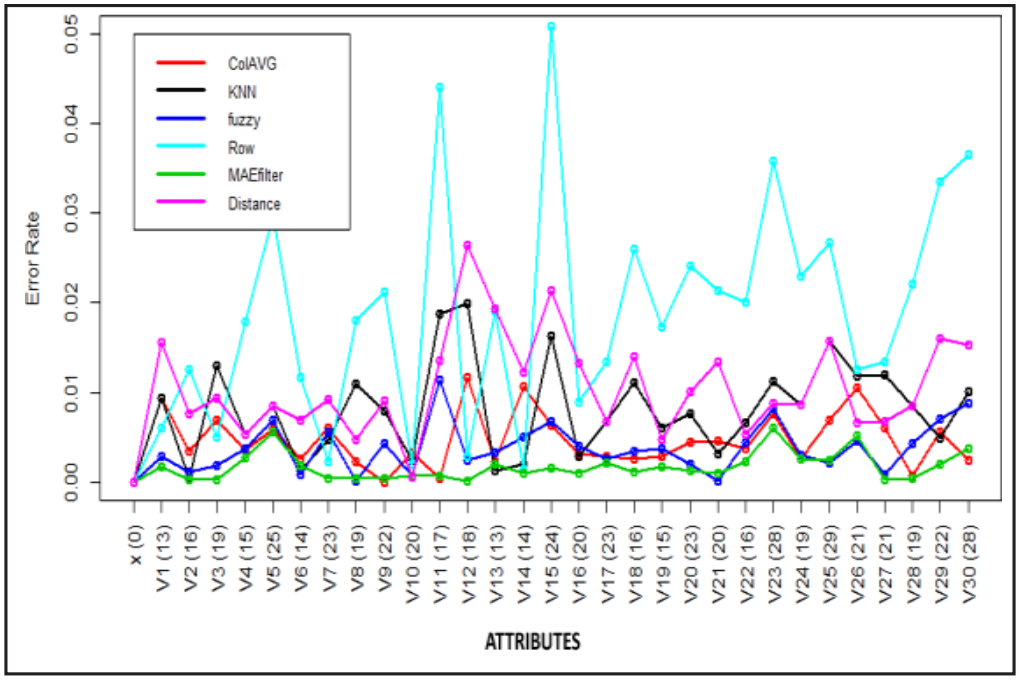
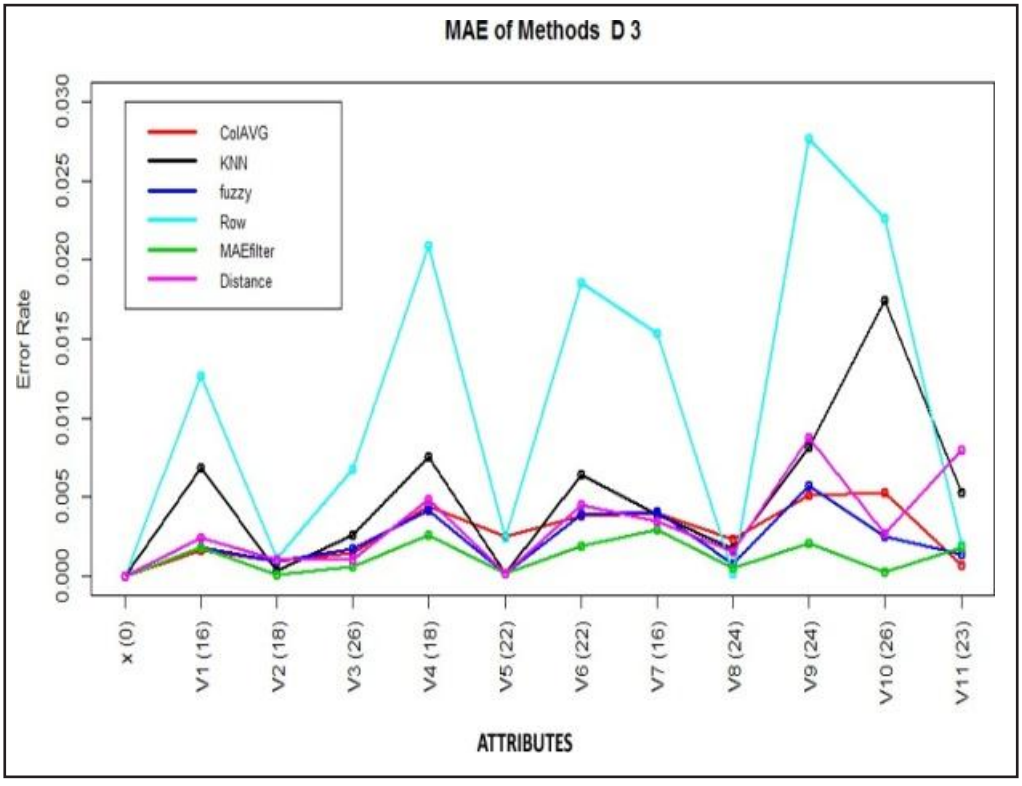


Fig. 4: RAE Graphs of all Datasets

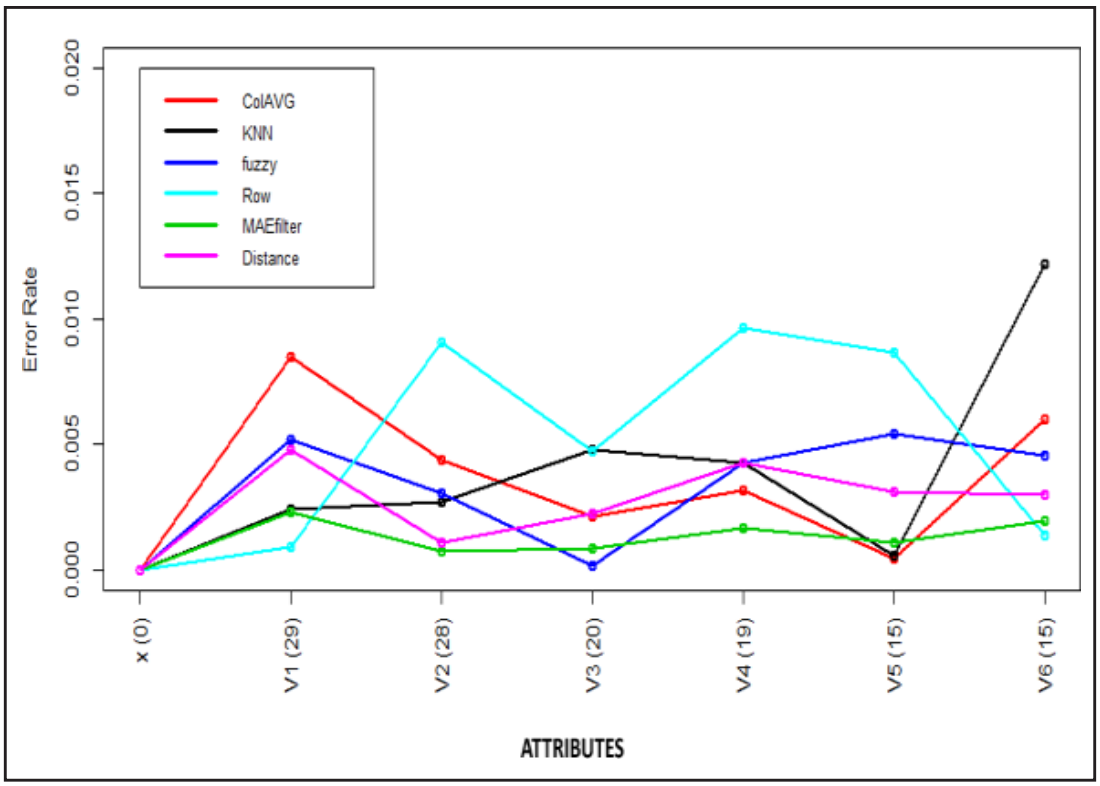


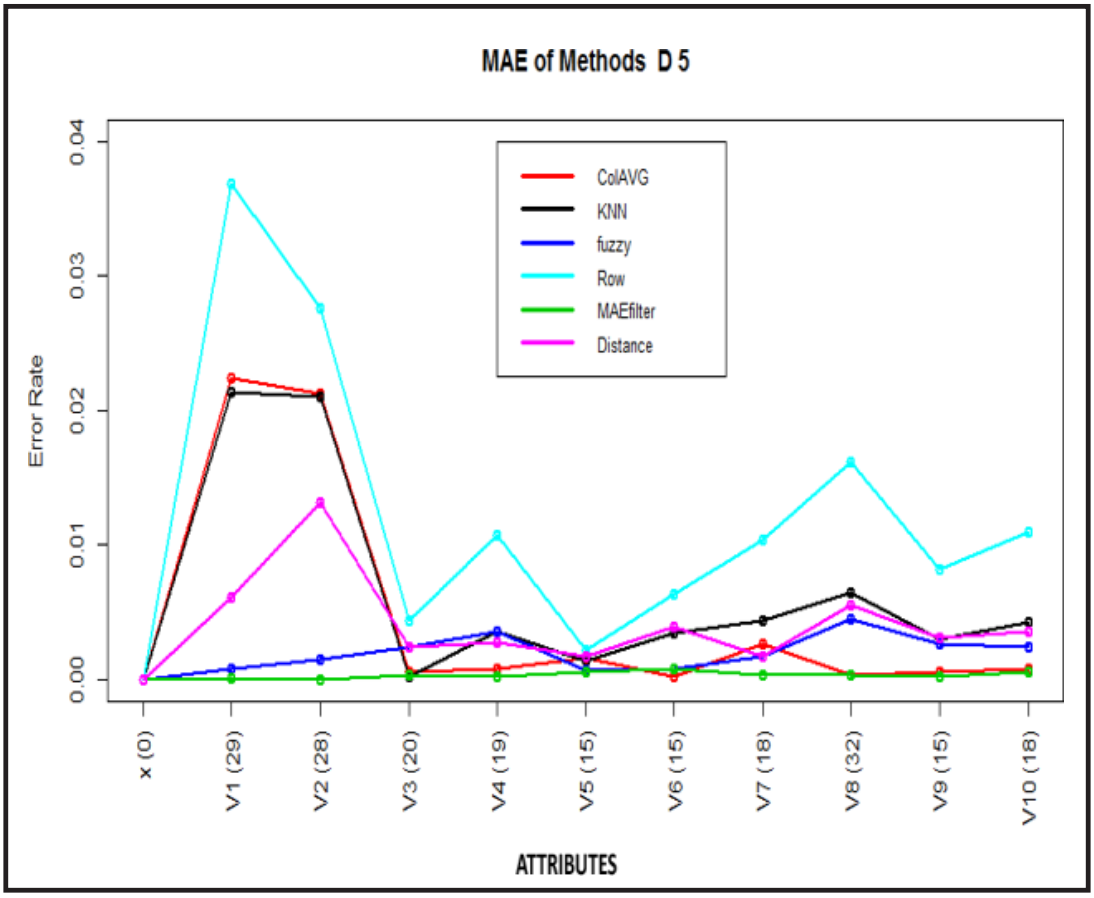
Contd.


ColAVG
KNN
fuzzy
Row
MAEfilter
Distance
Error Rate
0.00
0.01
0.02
0.03
0.04
0.05
x (0)
V1 (13)
V2 (16)
V3 (19)
V4 (15)
V5 (25)
V6 (14)
V7 (23)
V8 (19)
V9 (22)
V10 (20)
V11 (17)
V12 (18)
V13 (13)
V14 (14)
V15 (24)
V16 (20)
V17 (23)
V18 (16)
V19 (15)
V20 (23)
V21 (20)
V22 (16)
V23 (28)
V24 (19)
V25 (29)
V26 (21)
V27 (21)
V28 (19)
V29 (22)
V30 (28)
ATTRIBUTES


MAE of Methods D 3
ColAVG
KNN
fuzzy
Row
MAEfilter
Distance
Error Rate
0.000
0.005
0.010
0.015
0.020
0.025
0.030
x (0)
V1 (16)
V2 (18)
V3 (26)
V4 (18)
V5 (22)
V6 (22)
V7 (16)
V8 (24)
V9 (24)
V10 (26)
V11 (23)
ATTRIBUTES

Contd.


ColAVG
KNN
fuzzy
Row
MAEfilter
Distance
Error Rate
0.000
0.005
0.010
0.015
0.020
x (0)
V1 (29)
V2 (28)
V3 (20)
V4 (19)
V5 (15)
V6 (15)
ATTRIBUTES


MAE of Methods D 5
ColAVG
KNN
fuzzy
Row
MAEfilter
Distance
Error Rate
0.00
0.01
0.02
0.03
0.04
x (0)
V1 (29)
V2 (28)
V3 (20)
V4 (19)
V5 (15)
V6 (15)
V7 (18)
V8 (32)
V9 (15)
V10 (18)
ATTRIBUTES

Contd.

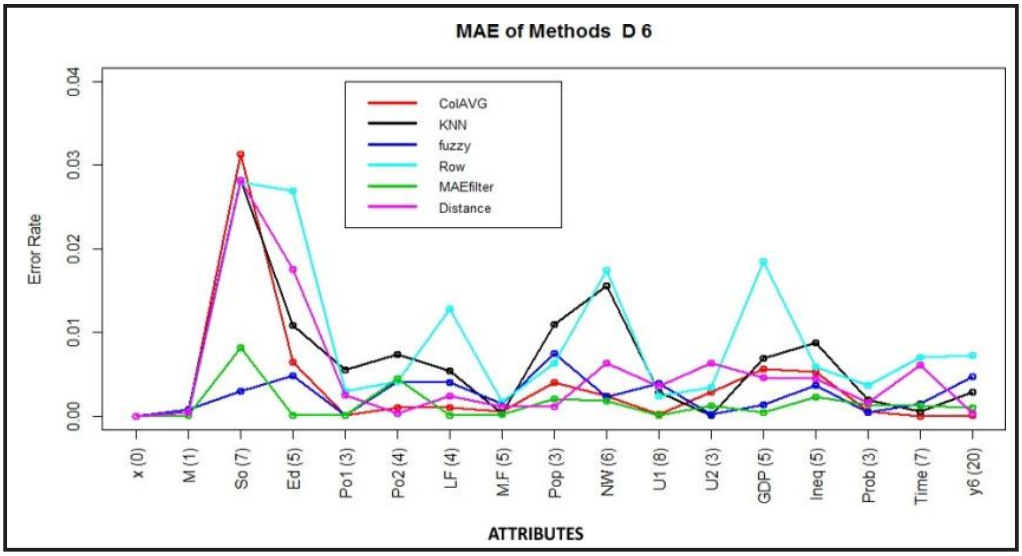


Fig. 5: MAE Graphs of all Datasets

VI. Experimental Results

In this paper, 4 validation technique has been used. These validation techniques applied on the six dataset of different variations and dimension. The graph procured showed a strong effect for Filter method gaining a much nearer value for the missing cell. We will discuss each dataset regarding the variation of the graph in Table III; Dataset that consists of values ranging from 0 to 1 has been normalized. There is also a need to observe the missing pattern by which we can apply some particular function or method for a beneficiary result. As the validation processed we found that Filter method is a delicate method that excels among all methods and overshadowed the rest. Validation techniques such as RMSE (D1-D6), RAE (D1-D6), MAE (D1-D6), SE (D1-D6) shows a positive graph line for making the Filter method as the best for heterogeneous dataset having different number of variables. Column average and fuzzy is taken in an average count for determining a missing value that varies from the original value. Knn, minimum distance, row average shows some flaws that cannot be considered as acceptable measures, but Knn sometimes follows nearer to fuzzy with number of variables less. Finally, we do consider Filter method as one of the best method in retaining a judgemental amount of value that can be taken in account for a profitable cause.

Table II: Symbolic Representation

Symbols	Meaning
Preimput	Preimputation 2
MV Dataset	Missing value dataset
For each Value[ij]	Row column belong to MV dataset
ImpData	Imputed dataset
Dist	MV dataset
ImpData[ij]	One cell value of imputed dataset
MV Dataset[ij]	One cell of misssing value dataset

ALGORITHM 1: Filteration.

Input: MV Dataset
Output: PDataset

PDataset ← [*preimput*(*MV Dataset*)];
/* Preimputed Dataset Using Preimputation 2 */
Imputed dataset ← [1, 2, . . . , N];
/* Imputed datasets Using Various Methods */
MinDist ← ; ∞
/* Assign minimum distance to infinity */
foreach $Value_{[ij]} \in (MV\ Dataset)$ **do**
  Value ← ∅
  if $[NA] presentin Value_{[ij]}$ **then**
    foreach $Imputed\ dataset_{[k]}$ **do**
      ImpData← $Imputeddataset_{[k]}$;
      Dist← $abs(ImpData_{ij} - PDataset_{ij})$;
      if $MinDist > Dist$ **then**
        MinDist ← *Dist*;
        Value ← $ImpData_{ij}$;
    $PDataset_{ij} \leftarrow Value$;

Table III: Data Set Used

Dataset Naming	Name	Attributes	Instances
D1	Balance Scale Weight and Distance Database	625	4
D2	Flag database	194	30
D3	Glass Identification Database	214	10
D4	Hayes-Roth and Hayes-Roth (1977) Database	160	5
D5	Relative CPU Performance Data	209	9
D6	The Effect of Punishment Regimes on Crime	48	16

VII. Future Direction

This paper has presented a new Filtration algorithm based on the comparison of multiple imputations techniques. Existence of missing values in the data set can sometime be really misleading. It becomes important to predict a value and to eliminate the presence of blank value so as to get the right results. We proposed a simple technique called Filtration technique to impute a judgemental value for the missing values from different data sets and also applied some other methods such as Knn, Distance Imputation, Fuzzy, etc. on the same. Substantial experiments are needed so as to check the correctness of the algorithm by taking a different set of data sets and then analyzing the output. We also dealt with the validation techniques as RMSE, MAE, SE, RAE. This will help in identifying any faults in the existing algorithm and can also help in predicting the accuracy of the methods used. In the paper we proposed a heuristic approach to identify missing values and then develop an approach to eliminate them. These experiments were performed on categorical data sets as well as numerical data set and the algorithm developed was successful in cleaning the data sets with missing values. It is also applied to use in large data set and it maintained well in establishing a value that can fulfill the empty cell. This new Technique will help in assisting others in handling missing values in large data sets. Finally, we can contribute our idea and work that will be helpful at some extent to be believed and motivate others to develop some more beneficial and develop a method of more accuracy adding to the world of knowledge.

VIII. Conclusion

In this paper, we have discussed the different data set with different patterns having missing values and moreover estimated a value for the missing cell, in the context of estimating missing value data quality completeness has to be maintained up to certain extent as possible. Completeness of data is one of the important attributes of data quality. The ultimate objective of our method is to fully justify the judged value as to ascertain a good quality of Data. As we know features of the data set is dependent upon the methods that support data to hold its integrity as much as possible. Knn or other methods casted for missing value estimation regarded as an important method in data mining and its determination is not always simple or totally efficient in some different kind of data set. Different techniques must be used dependent upon the nature of the data. Missing data is a common problem [4] in data mining. A blank entry in the database sometimes taken for granted as zero or, in some cases, cannot possibly exist (e.g. an entry for an individual in kind of salary or in any banking data base). However, it is not possible to get the exact value, but using the Filtration method a blank field can be filled with a judged that differ narrowly with the original value adding with some more calculation that to be compared represents an unknown quantity and techniques based on correlation can then be used to complete that entry. This should enable us for the replacement of missing values to be achieved more effectively.

References

1. A. Farhangfara, L. Kurganb, and J. Dy, “Impact of imputation of missing values on classification error for discrete data,” *Pattern Recognition*, vol. 41, no. 12, pp. 3692-3705, December 2008.
2. S. Singh, and J. Prasad, “Estimation of missing values in the data mining and comparison of imputation methods,” *Mathematical Journal of Interdisciplinary Sciences*, vol. 1, no. 2, pp. 75-90, March 2013.
3. Madhu G., and Nagachandrika G., “A new paradigm for development of data imputation approach for missing value estimation,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 6, pp. 3222-3228, 2016.
4. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, ……., and R. B. Altman, “Missing value estimation methods for DNA microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520-525, June 2001.
5. P. Schmitt, J. Mandel, and M. Guedj, “A comparison of six methods for missing data imputation,” *Journal of Biometrics and Biostatistics*, vol. 6, no. 1, article 224, May 2015. DOI: 10.4172/2155-6180.1000224
6. S. Bhushan, and A. P. Pandey, “Optimal imputation of missing data for estimation of population mean,” *Journal of Statistics and Management Systems*, vol. 19, no. 6, pp. 755-769, 2016.
7. N. J. Perkins, S. R. Cole, O. Harel, E. J. Tchetgen Tchetgen, B. Sun, E. M. Mitchell, and E. F. Schisterman, “Principled approaches to missing data in epidemiologic studies,” *American Journal of Epidemiology*, vol. 187, no. 3, pp. 568-575, 2017.