International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Missing Data Imputation Using Big Data Component Hive

M.Vennila[1], I.Jenifer[2]

Asst. Professor, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India [1, 2]

ABSTRACT: Now days a day we are using great deal of knowledge. Consistent with the organization's expansion the database becomes large. And it becomes so huge and sophisticated that traditional storage; analysis and processing won't work anymore. Taking care of such information is characterized as large information. Circumstances may happen where once in a while information is absent. But it becomes problematic if the info is said to medical Sector. And processing of such data becomes difficult. Hence during this paper we've proposed algorithm to seek out and predict Missing Value from calculating mean, median and mode values. We used hive to predict the missing values.

KEY WORDS: Big Data, Hive, Missing Value, Prediction, Mean, Median, Mode

I. INTRODUCTION

Big data is rapidly growing huge amount of data. The Big Data is so large and complex to handle using relational database management tools with analysis, capture, data curation, search, sharing, storage, transfer, visualization. Big data is portrayed by volume, velocity and variety. Such type of data is very difficult to process that contains billions of records that includes the web sales, social media, audio, images and so on. Example Yahoo, Google, Facebook, etc.., Big Data is problem and Hadoop is solution. Hadoop is an open source framework that provides solutions for handling big data. Apache's Pig and Hive are two major components of Hadoop to use MapReduce Library. Pig provides the scripting language to perform operations like reading, writing, filtering, joining and transforming. Instead of writing these operations in thousands of lines of java code which uses MapReduce. Pig is scripting and the Hive is SQL like inquiries both for Hadoop. Hive is a magnificent apparatus for examiners and business advancement types who are acclimated with the SQL-like inquiries and business insight frameworks. Often in official statistics we've large data sets with many variables and lots of missing data. Anyway we basically can't erase deficient records since this adds up to a generous loss of expensive gathered information. In some cases the loss is completely at random (MCAR), A more realistic hypothesis is to assume that the missing data are missing at random (MAR), that is, the likelihood that a perception is missing may rely upon watched information yet not on other missing qualities or non-watched factors. At times we can't accept MAR, we have the Missing Not at Random component (MNAR).

II. LITERATURE SURVEY

Farhangfar, Alireza, Lukasz A. Kurgan, and Witold Pedrycz. An epic structure for ascription of missing qualities in information bases." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 37.5 (2007): 692-709. The least demanding procedure to eliminate the examples containing missing qualities in their traits. Other way is by estimating the parameters such as variance and covariance depend on complete data using maximum likelihood methods. Or the missing values are calculated by analyzing the relationship within the attributes [1].

Bakshi, Kapil. "Considerations for big data: Architecture and approach" The enlightening assortment inspecting has gotten more genuine as the proportion of data has been growing. The big data aimed at improving the public healthcare system by clinical decision support system, Public sector, Retail industries, business modeling and marketing etc [4]. Now, the problem here comes to handling the issues in processing the datasets i.e. data cleaning which is a step of KDD knowledge discovery from data by performing some operations to extract the meaningful information. Out of these issues there is an important issue handling the missing values in a large database i.e. missing data imputation [2].

S.Sridevi, Dr. S.Rajaram, C.Parthiban, S.SibiArasan and C.Swadhikar [3] introduces missing value finding algorithm named autoregressive-model-based missing value estimation method (ARLSimpute). This calculation is for finding

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

missing an incentive for time arrangement informational index. So at nay point of time some data is missing or if the whole time point is missing so we can find using this algorithm. They had used linear prediction technique and quadratic prediction method to find this type of missing values. And using this technique they predict future data from historical data.

Thirumagal R, Deepali A Patil [4] develops two algorithms KNN impute and ARL impute. Both algorithms are utilized to ascribe and appraise missing qualities from informational collection. Based on imputation for finding missing data from data sets. Here KNN gives best result only when k value is between 10 to 20. This KNN attribute calculation is utilized for a specific section yet ARL ascription calculation is utilized for finding many missing an incentive from same segment.

Qianya Zhang, Ashfaqur Rahman, and Claire D'Este [5] said, because of sensor faults and sensor communication error some time some readings are missing. So they assessed attribution procedure execution in sensor field. They considered two scenarios for this techniques (I) the value missing only in prediction phase (II) Value missing in introduction phase as well as Prediction phase. They concentrated on two main words "Impute" and "Ignore" for missing value. Impute means missing value replace using imputation to existing value in data set through Prediction. Ignore means ignore the missing value if it is not useful and no need to imputation and no need change with random or artificial value.

III. METHODS OF MISSING DATA IMPUTATION:

Missing attribute values one or more of the attribute values could also be missing both for examples within the training set and for objects which are to be classified Missing data might occur because the worth isn't relevant to a specific case, couldn't be recorded when the info was collected, or is ignored by users due to privacy concerns. I f properties are absent in any preparation set, the framework may either disregard this item thoroughly; attempt to consider by, for example, finding what is the missing quality's most likely worth, or utilize the worth "missing", "obscure", or "Invalid" as a different incentive for the property.

The dealing with and overseeing of missing information system is grouped in three classes:- **Missing completely at random:** In this type of randomness the missing data occurs completely at the random, it occurs when the probability of a case having missing value for an attribute does not depend on the known values or the missing data.

Missing at Random: In this sort of irregularity the missing worth isn't arbitrarily dispersed among all the perceptions yet are haphazardly circulated inside at least one. The likelihood of a record containing missing an incentive for a property relies upon the known worth yet not on the benefits of missing information itself.

Not missing at random: This kind of component is known as not missing aimlessly. It is additionally considered as Non-insignificant. It happens when the likelihood of an example containing the missing worth doesn't rely upon the estimation of the quality. We can understand this by returning to the source information and afterward investigate more data about the component.

Case deletion: - In this strategy, the whole line or section is erased having missing worth traits.

Treating missing attribute as special value: - This is very surprising methodology. Rather than discovering some new worth, the missing worth is viewed as itself a unique incentive for the occurrence that containing missing worth.

Closest fit:- This is very surprising methodology. Rather than discovering some new worth, the missing worth is viewed as itself a unique incentive for the occurrence that containing missing worth.

K-nearest neighbor: - This technique finds the K-nearest neighbors, and among all neighbors the most widely recognized worth is considered for ostensible characteristics.

Weighted imputation with K-nearest neighbor: -In this the separation of each missing worth occurrences from its neighbors is determined. Here, the separation utilized for figuring weight. Missing qualities are determined by weighted mean for mathematical traits.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

K-means clustering imputation: -This strategy is like "K-Nearest Neighbor" the cases are bunched by utilizing K-implies grouping. Also, the occasions in each bunch are considered closest neighbors of one another.

Fuzzy k-means clustering imputation: - This strategy is better than K implies attribution. Here the information item can be a piece of more than one bunch community. It is best technique for covering information. In this unreformed properties for each - uncompleted information are subbed. Missing qualities are determined by weighted total of all entroids based on enrollment degrees.

Support vector machines imputation: - This technique is productive in memory utilization and huge dimensional spaces. This model is utilized to foresee missing traits from the total examples which don't have missing qualities. Here the condition properties and choice qualities are thought of.

IV. IMPLEMENTATION OF MISSING DATA IMPUTATION

In our implementation we find out missing values, we computed

1. **Mean**
2. **Median**
3. **Standard deviation**
4. **Variance**

Mean Substitution

The most usually rehearsed approach is mean replacement single attribution procedures. Mean substitution replaces missing values on a variable with the mean of the observed values. The imputed missing values are contingent upon one and only variable subjects mean for the variable supported the available data. Mean substitution involves imputing the mean of a variable within the place of any case which is missing a worth for that very same variable.

The example mean is that the ordinary and is processed in light of the fact that the total of the apparent multitude of watched results from the example isolated by the entire number of occasions. We use x as the image for the example mean. In math terms, where n is the example size and the x compare to the watched esteemed.

The formula for mean is as stated below

$$x = \frac{\sum x}{N}$$

Here,

$\sum$ represents the summation

X represents the Scores

N represents number of scores

Median substitution

Mean or median substitution of covariance and outcome variable is still frequently used. This method is slightly improved by first stratifying the data into subgroups and using subgroups average. Median imputation results in the median of the entire dataset being the same at it would be with case detection, but the variability between individuals' responses is decreased, biasing variance and covariance toward zero.

In simple words, Median is the center an incentive in the rundown of numbers. To calculate Median value, the list should be arranged in the ascending order first, and then the formula is stated as

If the total number of numbers (n) is odd number, then the formula is given below

Median = (n+1/2)

If the total number (n) is an even number, then the formula is given

Median = (n/2) + (n/2+1) /2

Standard deviation and Variance substitution

The mean, mode, median, and tended to mean does a pleasant work in telling where the focal point of the informational collection is, yet frequently we are keen on additional. This is the thing that the fluctuation and standard deviation do. First we show the equations for these estimations.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

We define the variance to be

$$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

and the standard deviation to be

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Variance and Standard Deviation: Step by Step

1. Figure the mean, x.
2. Compose a table that takes away the mean from each watched esteem.
3. Square every one of the distinctions.
4. Include this segment.
5. Gap by n - 1 where n is the quantity of things in the example this is the fluctuation.
6. To get the standard deviation we take the square foundation of the fluctuation.

Data Set

The database used for our experiment is or Diabetes database take from Diabetes Open Directory (DOD) with 9 variables.

Patient	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree function	Age
1	6	148	72	35	82	33.6	0.627	50
2	1	85	66	29	90	26.6	0.251	31
3	8	183	64	34	0	23.3	0.673	32
4	1	89	66	23	94	28.1	0.167	21
5	0	137	40	35	168	43.1	2.288	33
6	5	116	74	0	0	25.6	0.201	30
7	3	78	50	32	88	31	0.248	26
8	10	115	0	35	102	35.3	0.134	29
9	2	197	70	45	543	30.5	0.158	53
10	125	96	0	30	120	0	232	54

Table 1 Sample Dataset

Steps in the Imputation Algorithm

Step 1:

Spot the dataset in record and characteristic lattice design with records as lines and qualities as segments. Consider the Table 5.2 which has ten medical records and four records with missing values as shown below.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
R_1	6	148	72	35	82	33.6	0.627	50
R_2	1	85	66	29	90	26.6	0.251	31
R_3	8	183	64	34	0	23.3	0.673	32
R_4	1	89	66	23	94	28.1	0.167	21
R_5	0	137	40	35	168	43.1	2.288	33
R_6	5	116	74	0	0	25.6	0.201	30
R_7	3	78	50	32	88	31	0.248	26
R_8	10	115	0	35	102	35.3	0.134	29
R_9	2	197	70	45	543	30.5	0.158	53
R_{10}	125	96	0	30	120	0	232	54

Table 2 Sample Dataset with attributes

Step 2:

Deteriorate the clinical dataset into two sections. The initial segment comprises of records with no missing qualities and second part comprises of records with missing qualities. Table.3 shows the records R1, R2, R4, R5, R7, and R9 which have all the property estimations characterized. We can see that there are no missing qualities.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
R_1	6	148	72	35	82	33.6	0.627	50
R_2	1	85	66	29	90	26.6	0.251	31
R_4	1	89	66	23	94	28.1	0.167	21
R_5	1	137	40	35	168	43.1	2.288	33
R_7	3	78	50	32	88	31	0.248	26
R_9	2	197	70	45	543	30.5	0.158	53

Table.3 Records with no missing values

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
R_3	8	183	64	34	0	23.3	0.673	32
R_6	5	116	74	0	78	25.6	0.201	30
R_8	10	115	0	35	102	35.3	0.134	29
R_{10}	125	96	68	30	120	0	232	54

Table.4 Records with missing values

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Step 3:
Standardize the datasets got in sync 2 appropriately. This includes planning the credit esteems to suit the requirement for applying the proposed measure to discover missing qualities for example we fix the area of property estimations. In Table.5 the standard deviation of all characteristics in the records of Table 2 are figured and recorded. The traits esteems for A3, A4, A5, and A6 are planned to an alternate area to make it appropriate for handling by the calculation and proposed measure.

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
2.1370	46.7319	12.7541	7.3325	181.901	5.8865	0.8337	12.9872

Table.5 Standard deviation of attributes

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
R_1	6	148	72	35	82	33.6	0.627	50
R_2	1	85	66	29	90	26.6	0.251	31
R_4	1	89	66	23	94	28.1	0.167	21
R_5	1	137	40	35	168	43.1	2.288	33
R_7	3	78	50	32	88	31	0.248	26
R_9	2	197	70	45	543	30.5	0.158	53
Std.Dev	2.1370	46.7319	12.7541	7.3325	181.901	5.8865	0.8337	12.9872

Table.6 Mapping A_3, A_4, A_5, and A_6 attribute values

Step 4:
Choose one record with missing attribute value or missing values, each time to fix the missing values of the record.

Step-5:
From the record considered in above advance, dispose of the quality section comprising missing qualities (Second piece of dataset). Likewise, dispose of this characteristic section when thought about residual records with no missing qualities (First aspect of the dataset).

Step-6:
Apply the proposed measure to discover the comparability between the new record and existing records without missing qualities.

	Similarity with R_3
R_1	78.505
R_2	83.827
R_4	84.124
R_5	143.921
R_7	80.003
R_9	421.351

Table 7 Similarity of records R_3 with recorded in Table 2

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

	Similarity with R_6
R_1	29.210
R_2	18.145
R_4	15.821
R_5	29.210
R_7	27.541
R_9	33.819

Table 8 Similarity of records R_6 with recorded in Table.2

Step-7:

The class label of new test record is the label of the record to which the new test record shows maximum similarity.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
R_3	8	183	64	34	?	23.3	0.673	32
R_6	5	116	74	?	78	25.6	0.201	30
R_8	10	115	?	35	102	35.3	0.134	29
R_{10}	125	96	68	30	120	?	232	54

Table 9 mapping the records with missing values

Step-8:

Fill the missing quality estimation of the test record with the comparing trait estimation of the record to which it has most extreme similitude. We may likewise think about the recurrence of event if there should arise an occurrence of vagueness.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
R_3	8	183	64	34	84	23.3	0.673	32
R_6	5	116	74	29	78	25.6	0.201	30
R_8	10	115	69	35	102	35.3	0.134	29
R_{10}	125	96	68	30	120	29.5	232	54

Table 10 Records with missing values fixed

Step-9:

Rehash the above strides for each record having missing qualities as needs be. The records R3 and R6 show greatest closeness with the records of Table 7, and Table8 separately. Consequently we fix the missing trait estimations of these records R.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

V. CONCLUSION

Missing data place major role in data pre-processing process. Missing qualities are exceptionally normal in clinical information. Handling missing values is challenging and also an interesting area of research in perspective of data mining, information retrieval. In this work we propose imputation measure which is used to fix missing values of attributes for the records of dataset. This is done by finding similarity values between records with missing values and records without missing values. In our proposed approach we proposed mean substitution, median substitution, variance and standard deviation substitution. From the experiments on diabetes dataset, it is observed that standard deviation is the best missing data imputation technique to find the similarity attributes in missing data imputation.

Our proposed algorithm is implemented in the HADOOP tool. In hadoop hive is used in our expriment. The database used for our experiment is our diabetes database taken from (DOD) Diabetes Open Directory with 9 variables. Hadoop tool is used for the implementation process.

In future this dissertation can be enhancing by using various other clustering techniques and various imputation techniques so that accuracy extends up to 80%.

REFERENCES

[1] Farhangfar, Alireza, Lukasz A. Kurgan, and Witold Pedrycz. "A novel framework for imputation of missing values in databases." *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 37.5 (2007): 692-709.

[2] Bakshi, Kapil. "Considerations for big data: Architecture and approach."*Aerospace Conference, 2012 IEEE*. IEEE, 2012.

[3] Sridevi, S., et al. "Imputation for the analysis of missing values and prediction of time series data."Recent Trends in Information Technology (ICRTIT), 2011 International Conference on. IEEE, 2011.

[4] Thirumahal, R., and Deepali A. Patil. "KNN and ARL Based Imputation to Estimate Missing Values." Indonesian Journal of Electrical Engineering and Informatics (IJEEI) 2.3 (2014): 119-124