DATA PRE-PROCESSING ON ELECTION VOTING

Affiliation
Mrs. K. MAHALAKSHMI 1, D.VITHYA 2,
[1] Mrs. K. MAHALAKSHMI,
[1]Assistant professor, [1]Department of B. Com (Business Analytics),
[1]PSGR krishnammal College for Women, Coimbatore, India.
muhilm9@gmail.com

[2]Ms. D.VITHYA,
[2]UG SCHOLAR,[2]Department of B. Com(Business Analytics),
[2]PSGR krishnammal College for Women, Coimbatore, India.
vithya.d2000@gmail.com

Abstract : Data preprocessing may be a crucial step within the info mainning process. The process of converting data into information is called data processing .after collecting data, the raw data is to be converted into meaningful form, the process of converting data into meaningful form is named processing or data reduction. that involves transforming data into a clear format. Data in the real would in dirty. the various steps involved in data preprocessing.(I)Data cleaning,(II)Data Integration,(III)Data Transformation,(IV)Data Reduction,(V)Data Discretization First the data is collected and the data is pre-processed by removing the duplicate votes, and removing null values.

***Index Terms* -** data preprocessing, datamining, null values

I. Introduction

Data preprocessing is an integral step in Machine Learning because the quality of knowledge and therefore the useful information which will be derived from it directly affects the power of our model to learn, therefore, it's extremely important that we preprocess our data before feeding it into our model. real-world dataset there are always few null values. It 's really matter whether it is a regression, classification or the opposite quite problem, no model can handle these NULL or NaN values on its own so we'd like to intervene. Preprocessing we need a dataset to preprocess the data. For that we are using INDIAN NATIONAL ELECTION dataset. In this some attributes have null values which are not mentioned as 0. To haldle the null-values in this dataset start with finding where the null values are taken place. Then start to preprocess each columns one by one.

II.OBJECTIVE

Identifying and pre-processing parties for customizing them and removing null values .A Null value in a table is a value in a field that appears to blank ,which means a field with a Null value is a field with no value. It is important to understand that a Null value is different than a zero value or a field that contains spaces. A field with a null value is one that has been left blank during record creation. Its remove Null values used for data pre-processing process.

.III. RELATED WORK

Data Preprocessing is that the method of simply transforming data into understandable format. Real world data is usually incomplete, inconsistent, redundant and noisy. Data preprocessing involves various steps .(I)Data cleaning,(II)Data Integration,(III)Data Transformation,(IV)Data Reduction,(V)Data Discretization[5]

Data cleaning, data cleansing, or data scrubbing is that the process of improving the standard of knowledge by correcting inaccurate records from a record set. The detecting and modifying, replacing, or deleting incomplete, incorrect, improperly formatted, duplicated, or irrelevant records, otherwise mentioned as “dirty data,” within a database. Data cleaning also includes removing duplicated data .[3]

Data Integration is that the process of transferring the data in source format into the destination format. warehousing has been supported to data migration and transportation by using Extract-Transform-Load (ETL) approach. These tools are widely fit handling large volumes of knowledge and not flexible to handle semi or unstructured data. Data Integration as a process is extremely cumbersome and iterative especially to feature new data sources. The process of adding these new data sources are time consuming which ends up in delay, loss of knowledge and irrelevance of the info and improper utilization of useful information.[1]

Data reduction has been used widely in data processing for convenient analysis. Principal component analysis (PCA) and correlation analysis (FA) methods are popular techniques. The PCA and FA reduce the amount of variables to avoid the curse of dimensionality. The curse of dimensionality is to increase the computing time exponentially in proportion to the quantity of variables. So, many methods are published for dimension reduction. Also, data augmentation is another approach to research data efficiently. Support vector machine (SVM) algorithm may be a representative technique for dimension augmentation. Both data reduction and augmentation are wont to solve diverse problems in data analysis. [2]

Discretization of numerical data . The purpose of attribute discretization is to seek out concise data representations as categories which are adequate for the training task retaining the maximum amount information within the original continuous attribute as possible. discretization as data preprocessing technique, developed within the literature for giant Data. [4]

IV. METHODOLOGY

4.1. DATAMINIG

It is the method of discovering or mining knowledge from an outsized amount of knowledge. Another term for datamining –KDD (knowledge discovery from data).attempts to extract hidden patterns and trends from large databases. Also support automatic exploration of data.

4.2. NEED OF DATA MINING

Needs comes evolution in size of database.
Db ↑[Big data]->manual analyze
Need of automatic analysis

4.3. EVOLUTION OF DATA MINING

4.3.1. STATISTICS:

Regression analysis, cluster analysis, standard deviation etc.

4.3.2. ARTIFICIAL INTELLIGENCE:

Applying of human-thought like processing.

4.3.3. MACHINE LEARNING

Union of statistics and AI about learning by software about data

4.4. PRE-PROCESSING

- It is done-to improve the quality of data in data warehouse.
- Increase of mining process.
- Removes noisy data,incrsistent data and incomplete data.

the various steps involved in data preprocessing.(I)Datacleaning,(II)Data Integration,(III)Data Transformation,(IV)Data Reduction,(V)Data Discretization

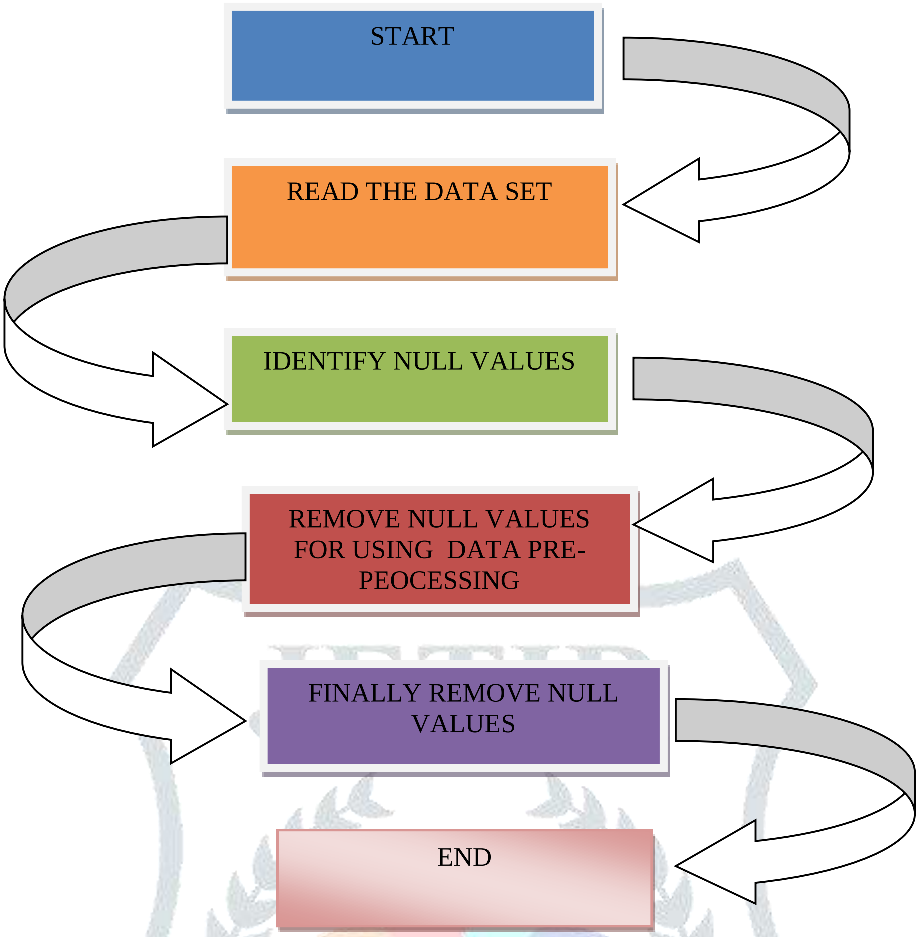
4.5. DATA CLEANING

It cleans the data by filling in the missing values, smoothing noisy data, resolving the inconsistency and removing the outlines

4.6. WAYS TO HANDEL MISSING DATA DURING CLEANING

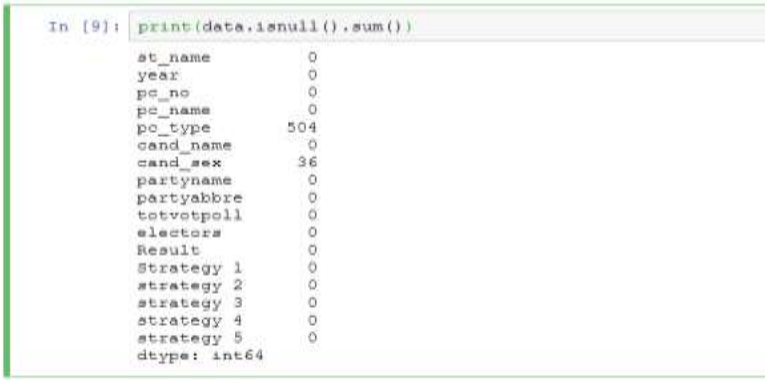
- Mannual entry of missing data
- Using attribute mean
- Using most probable value
- Using global constant(NA)
- Ignore the tuple

FLOWCHART OF THE PROCESS


START
READ THE DATA SET
IDENTIFY NULL VALUES
REMOVE NULL VALUES FOR USING DATA PRE-PEOCESSING
FINALLY REMOVE NULL VALUES
END

First,we have to check the null values in the dataset to clean the dataset.Python uses the keyword None to define null objects and variables. As the null in Python, None isn't defined to be 0 or the other value. In Python, None is an object and a first-class citizen!.The null values in the dataset is displayed using the below code. Print (data.isnull.sum ()).

V. IMPLEMENTATION AND RESULT

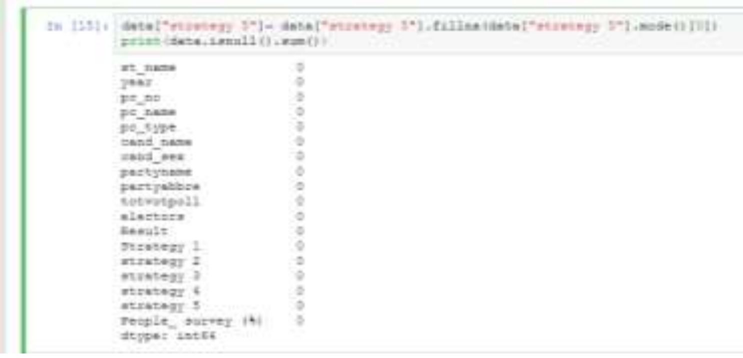


```
In [9]: print(data.isnull().sum())

        st_name         0
        year            0
        pc_no           0
        pc_name         0
        pc_type       504
        cand_name       0
        cand_sex       36
        partyname       0
        partyabbre      0
        totvotpoll      0
        electors        0
        Result          0
        Strategy 1      0
        strategy 2      0
        strategy 3      0
        strategy 4      0
        strategy 5      0
        dtype: int64
```

FIG 4.1 Display How Many Null Values

In the above figure all the attributes with null values are Displayed.We have null values in the column pc_type,cand_sex, strategy 2,strategy 3,strategy 4 and strategy 5.The null values in each column have to be removed to get a clear dataset.



```
In [15]: data["strategy 5"]= data["strategy 5"].fillna(data["strategy 5"].mode()[0])
         print(data.isnull().sum())

         st_name              0
         year                 0
         pc_no                0
         pc_name              0
         pc_type              0
         cand_name            0
         cand_sex             0
         partyname            0
         partyabbre           0
         totvotpoll           0
         electors             0
         Result               0
         Strategy 1           0
         strategy 2           0
         strategy 3           0
         strategy 4           0
         strategy 5           0
         People_ survey (%)   0
         dtype: int64
```

FIG 4.2 Remove Null Value

In fig 4.2 all the null values in the dataset is removed using the fillna ("unknown") command. This replaces the null values with a 0 and the data is preprocessed

VI. CONCLUSION

In this paper, Jupyter Notebook is the tool to analyze the status of Data pre-processing is an important step in preparing raw data for statistical analysis and to get accurate results. Throughout the process it is important to understand the choices made in pre-processing steps and how different methods may impact the dataset validity and applicability of study results. We have null values in the column pc_type, cand_sex, strategy 2,strategy 3,strategy 4 and strategy 5.The null values in each column have to be removed to get a clear dataset. all the null values in the dataset is removed using the fillna ("unknown") command. This replaces the null values with a 0 and the data is preprocessed.

FURTHER WORK:

It is suggested that this method of solution can further exten by using more tools.data preprocessing is necessary step before building a model with these fetures. In any tool preprocessing in must not only in python also in R, Weka,RapidMiner, etc

REFERENCES:

[1] Arputhamary Bonfring .pInternational Journal of Data Mining, Vol.5 ,No.1,Feb 2015

[2] Daiho Uhm International Journal of Fuzzy Logic and Intelligent Systems 12(1)DOI: 10.5391/IJFIS.2012.12.1.1 March 2012

[3] Mike AllenPublished: 2017 https://dx.doi.org/10.4135/9781483381411.n126

[4]Sergio Ramírez-Gallegoovember 2015Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 6(1):n/a-n/a DOI: 10.1002/widm.1173

[5] Vivek AgarwalInternationa l Journal of Computer Applications (0975 – 8887)
Volume 131 – No.4, December2015

[6] Dhairya kumar, By Towards Data Science.: https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

[7] Son NH (2006) Data mining course—data cleaning and data preprocessing. Warsaw University. Available at URL http://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf