Grenze International Journal of Engineering and Technology, Jan Issue

Importance of Data Features Selection: Analyze Students Performance using Optimal Features Logistic Regression Model

Javed Akhtar Khan
Indur Institute of Engineering & Technology, Computer Science & Engineering, Siddipet Telangana India
Email: er.javedkhan@gmail.com

***Abstract*—This is an informative article of Data Analytics for data .This article covers the various aspects of Data Analytics such as definition of Data analytics , various terminology used in Data Analytics related to data ,Types of Data and various methods of data collection. Motive of this paper is to describe the importance of data quality , describe the data processing and preprocessing task for structured and unstructured data . Perform the observation through student data to find the fact about the data design architecture in term of optimization of processing steps for quality data set a case study. A case student is based on the student performance analysis based on the selected feature among the various existing features in the Data set using the machine learning regression model .Implantation related work in this paper work to produce the student performance using Internal exam and attendance and try to predict the performance of student in External Result . For this I include the Logistic Regression model to produce the result.**

***Index Terms*— Data , Data Analytics , Data Management, Regression , Data Design architectures , Data preprocessing, Regression Model.**

I. INTRODUCTION

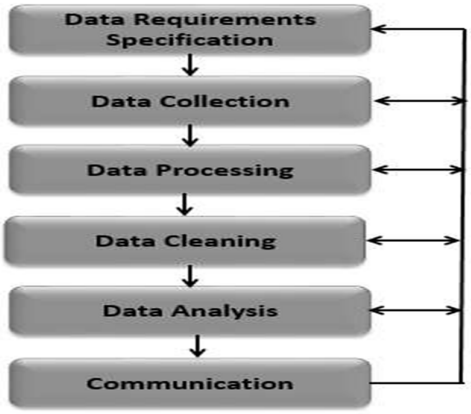
This introduction section covers the basic information about Data Analytics. The term Data Analytics includes the Two terms one is Data and second one is Analytics . Data Means a collection of things ,items , any relevant or irrelevant objects will be considered as Data sometimes it is also called Raw Data. Example of Data a Collection of Images so it may be a human images , animal images , food images , properties images, illusion images , Vehicles images ,home images , fruit images , vegetable images , furniture images ,clothes images, medicine images so many more there is no limit of images this is called the Data because this include the any images . This is one example of data. Many more real life examples are available in this world to understand the term Data . Data may be in digital form , physical form , statically form any other forms. Similarly student data , employee data Covid Patient is also the one kind of digital data , other patient data , etc Data is simple collection of various things ,Data does not generate any kinds of information .

Data Analysis is the step by step process of digging through the data for finding the suitable information , knowledge , and to make some predictions from the Data . Analysis means to filter the Data , Apply some conditions , some set of rules , remove the noise from Data , improve the Data Quality, and achieve the target in terms of prediction accuracy.

Grenze ID: 01.GIJET.9.1.307_1

Figuer.1 Data Analytics Steps [18]

Data analysis is complete one process for Targeted Data to find the hidden knowledge , improve the business , make good business policies , make a good set of rule in term of generate the maximum revenue , attract the max customer , solve the complex problem , forecast , make the relationship among the parameter , make decision making senses etc .



Figuer.2 Data Analysis process [18]

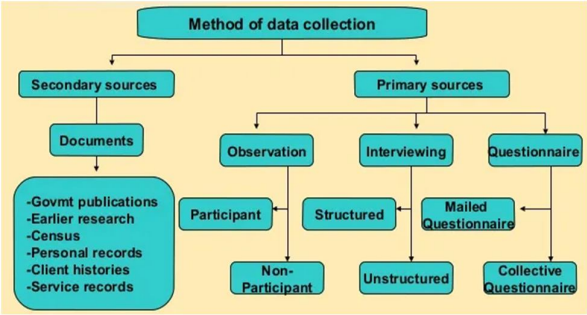
II. Type of Data and its Representation with Data Collection Methods

Collection of Data is the step by step process it includes the data acquiring, data collecting , data fetching ,and information about the data storing . Most of the time we just collected the data without worried about the data quality , data relevancy , and some other factor so these kinds of collected data consider as unstructured data .
Structured data means data should related with the particular aim so that data will be consider as structured data or it just follow some structured like email Some of the other Example are as various kinds of file like XML, image files is used in data analysis. Collected data is called the RAW data, in general term if you collect the data first time so that data will consider as the RAW data .This raw data sometimes not useful for analysis due the noise, so before using the raw data researcher applying the data cleaning process and data preprocessing task to make the data usable and improve the quality of data as well .

A. Data Collection

Collection of data form the different – different sources for analysis work , we can start the data collection by asking simple questions to audience like group of students , group of school students , group of shopkeeper etc so in this based one our objective we can start collection of data .
Generally collected data divided in to two types one is called “Qualitative Data “ its example like Male and Female, Good and Bad Low and High , and another one is called “Quantitative Data” as its name gives hint it is

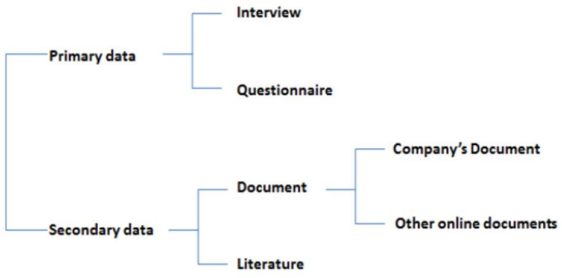
countable data, means it include number or numeric type value like total student in your class, Total number of member in your family .
So finally "Qualitative Data" is Non numerical type of data and "Quantitative Data" is numerical type of data. Two major parts of any Data mention above is again divided into two parts theses are : one is Primary Data and Secondary Data. Here primary data may be qualitative or quantitative similarly secondary data may be qualitative or quantitative.



Figuer.3. Data Collection Sources[17]

B. Primary Data

Primary data can be consider as Raw data or original data these data extracted from the official sources. Primary data is collected directly by applying some techniques like questionnaires, data may be collected through interviews, and data collected through surveys. Primary data collection methods, these methods are very useful for data analytics to collect the correct data for further work in the right direction. These methods are as:
Interview Method: Does everybody know very well what an interview is ? In this method the interviewer asks many questions to the interviewee. It means it is based on a simple Question and Answer pattern but important is that selection of Question selection of interviewee.



Figuer.4. Type of Data For Data Analytics[22]

Next Method for primary data collection is Survey method. In this method researchers prepare the proper list of relevant questions and collect the answers form the targeted audience ,answers may be in text format , audio, video , multiple choice question etc . This method can be online or offline mode after collection answer data analytics performs the analysis process.
Observation method is based on selected targeted audience like proper monitoring the behavior of Students in one Class room and monitoring the behavior of customers for the particular product. In this data collection methods researchers keenly observe the behavior of the selected audience using some data collection tools collected may be in the video format , or in some other format like text , picture , images etc
Experimental methods: As names suggested experiment so in this type of method data should be collected through by applying the some experiments these experiments related to what kind of data or resulted data researcher want to collected .Already existing some widely used experimental methods are as RBD, CRD,FD, LSD.

- **CRD- Completely Randomized Design** this method is widely applying for to make a comparing of the particular experiments.

- **RBD- Randomized Block Design** is comes under the an experimental design method for data collection this methods based on the fragmentation means divide the big unit in
- smaller units or blocks. Than after applying experiment on these smaller unit rather than one big unit . One more very important things this method originated form Agriculture land. By using this method research make a smaller unit of land and perform random the experiments and collect the resultant data .
- **LSD – Latin Square Design** is based of matrix structure and include the rows and columns .Using this method researcher can able to create the n number of row and n number of Column nxn here rows and column a unique letter at once only means no repetition required .This kind of arrangement is help to us to find error .one of the good example of this methods is Latin Square Design Concept .
- **FD-Factorial design** is based of the an experimental design in this every experiment must has two factors with possible values and on performing some trials other combinational factors are derived.

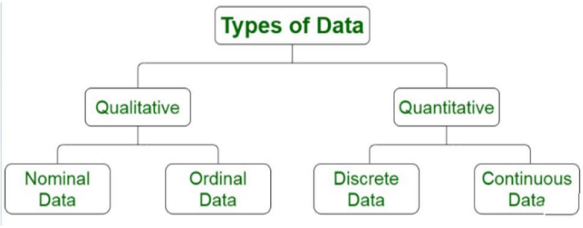
C. Secondary Data

What ever the data you collected first time if you used same again for performing the new task so in this case data will be consider as Secondary data , here related task means find the some relation between the collected data , to predict some result from the data it means researcher further analyze the collected data for some valid purpose . Secondary data sources are internal sources and external sources.

- **Internal source of Data:** if we found the data within the particular organization that data consider as Internal Source of Data such as data from sale , data form market ,education hub many more . This type of data collection is less Time consumption , we can easily collect the data from internal sources .
- **External source:** As name mention external means data collection through external sources like others third party so these collected data called external source data. Due to high volume of these kind of data cost and time consumption also high. Some examples related with external sources like , international labor bureau ,published news , planning commission ,RGI that is Registrar General of India, , Government publications and non-governmental publications, syndicate services, Data related to treatment of particular diseases in various hospitals etc .External data Collection is littlie bit difficult because to find the particular resources or data provider is complex task , collected data is in very large . This is some basic information about the Primary data secondary data with its suitable example and sources .

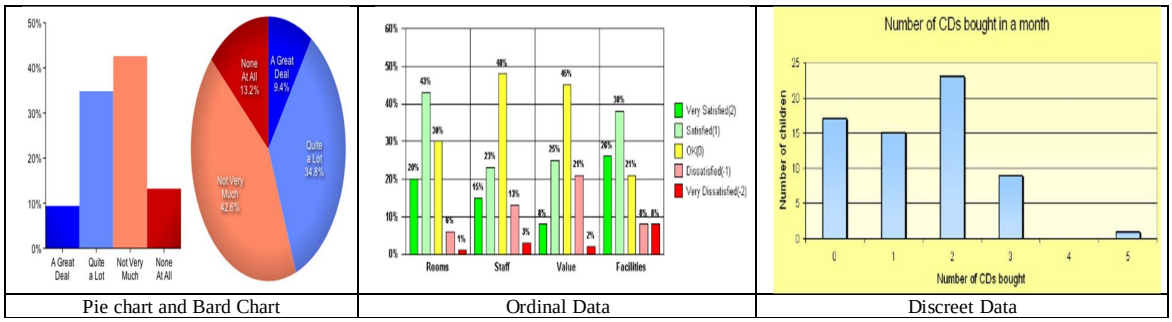
III. Data Type for Data Analytics

Above section shows collective information about two main categories of Data and introduction about the techniques that are used for data collection . Now this part of the paper includes information about data type of Data , that can be used by data analytics for the various research oriented work .Data selection is important for the researcher to produce a good output. Now let's take some more information about the Data type for Data Analytics. Starting with Quantitative Data it means that “How many ?” “How Much?” If we understand these questions so very easily we understand Quantitative Data. This category of Data is reflected as a number or count of things in integer or numeric type data or numeric variable .Quantitative data is
mostly include the statistical type of data these data can be represented in the form of graph , chart bar graph , Some related Example for Quantitative Data like Student exam marks . Temperature ,number of team participants in last Olympic Games , the number of yours family members , number of student in you college etc so all these are related with the numeric value . Main thing This Quantitative data is further fragmented into Discrete data and Continuous data .
Next Type of Data is Qualitative data. This type data only deal with the words, picture , images and symbol rather than numeric value . The second name of this data is categorical data because this type of data deals with non numeric data values. Ordinal Data deal with the number which is in some order or in a sequence; here it is not a matter of sequence order like descending or ascending .. So also say that nominal cant not be in a sequence . The value of Ordinal Data is comes under the Qualitative data because it is deal with the numeric type of data in proper order. One very important point here ordinal data can not be represented in numerical type value because ordinal data comes under the categories of Qualitative data . Example of ordinal data type is First , Second , …., A, B, C ….. low , medium high , etc.
Discrete Data This is represented only in terms of numeric value means some count we can represent like total number of female students in your class , total count of students in school , etc . Continuous data means height of child , sped of your vehicle ,average strike rate of your best cricket player etc.



Figuer.5 Types of Data [16]

Visualization of various Data Types in Data Analytics :



Pie chart and Bard Chart	Ordinal Data	Discreet Data

Figuer.5 Data Visualization [19]

IV. Data Quality and Data Preprocessing

Data Quality is very important and known in domain Data analytics DBMS , data quality is effect the resultant value for the research work[1][2]. Data Quality is faced many issue and challenging to improve data quality in domain of BigData its is very time consuming and costly task generally done by Data preprocessing. In The Data Preprocessing process we can achieve the data quality .For improving the data Quality collected data must be noise free for this data processing is required , in this process data cleansing, filtering of data , de duplication of data , compression , conversion of unique format, done . Data Quality plays the important role to achieve the target . To ensure the data quality data analytics want to remove the irrelevant data from the data set . These data may be noise ,outliers, missing values , duplicate data in general etc . Good Quality data should have the good accuracy, more reliable , consistent and up-to-date .Data quality is important for good quality analysis;

Figure 9. Data Quality Various Dimensions[13]

V. Literature Survey

This part of the article covers the importance and introduction about the Data Quality of collective data . This is again very important for the researcher or data analytics to get a good result. Good Quality of data gives an excellent result for the selected problem if the researcher deals with unprocessed data it reflects the result. Data Quality means to remove the unnecessary data , unrelated data, noise , outliers , missing values , duplicate data etc from the collected data set in this way researchers will improve the Data quality before performing the operation on it .

As per the record of IBM[3]McKinse[6][7][9]and Gartner[4][5]per day very high amounts of various kinds of data are generated; this data represents2.5quintillion bytes {Exabyte {EB} =1018 bytes}. In the year of 2000,

total 800,000 Petabyte {1PB is it equal to 1015 Bytes} this huge amount of data was stored , form the various sources of Point . And after the 20 years this will touch the 35 Zettabytes {1ZB= 1021 bytes}[10]. Through literature surveys I cover the basic overview and information about the collected data. Many researchers work on different types of data for various applications . What data they collected for which application? In which way they used the collected data ? many more queries answer is cover in this literature survey :[11]Ali works on Time series data. They used time series data for visual analysis In his survey paper ,they focus on the clustering or classification algorithm for analyzing the selected data. This is one kind of survey paper. They explore the sixty research article to make an excellent survey of Time series data for clustering and classification related work .In this paper they cover the various approaches to understand the data , its characteristics to find some meaningful statistics, trends, for making decisions and predicting the future .[12]Dabeeruddin and Ameema et. al they wok on the Smart grid big data analytics this article cover the survey of various technologies different type of techniques and application , various kinds of sources of data that relevant with the smart grid as they mention in this article big data analysis include the semi structured data , unstructured data and quasi structured data in this article they include the proper explanation of these types of data . They conduct literature review for the Big data Sources and Data the related to Smart grid , They also highlights the Big Data technologies with its latest development , good coverage of big data processes , techniques ,well designed commercial solutions, and discuss the various applications in domain of Smart grid . this work covers information about the smart grid , big data analysis, comparative review , technologies used for big data and its literature , They also include the information about related technologies of Smart Grid .This article is based on the data Quality[13] in this article the author focuses on quality evaluation. They work on the quality management framework to describe the proper step step by evaluation for big data So the main focus of this work is provide the good framework that able to to provide good quality management and a simple path for the Data Scientists to manage the quality of various data .This paper covers the introduction of Big Data and Data Quality , Quality management model for big data etc .In this paper authors go through the various research articles they work on and produce the data quality classification .This paper includes the various different ways of data collection [14] in the field of smart transportation . so it is very clear in this paper the author collected the various different ways of data collection for the smart city . They collected data from Pneumatic tubes for traffic related data , Video vehicle detection is one of the techniques that is used for identifying the vehicle here data is collected using CCTV camera, GPS based data collection many more different ways are introduced in this paper .This is a survey paper based on the preprocessing methods [15] that can be apply for mining in Big Data . in this paper authors work on the data process to improve the data quality they also discuss about the data cleaning process and its related issue .in this article they introduce the Filter based method is used for cleaning the raw or noise data as well as some other methods for performing cleaning the data are as imputation method, and hybrid method and wrapper method for getting the good data in term of quality these method is widely used .

VI. Methodologies Logistic Regression Model

Regression is a statistical method widely used in various sector such as investing, finance related work, and other disciplines that attempts to find the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Here this method I used for to compute the student performance by using the optimal attribute . This method will produce the some hidden fact about the students performance using the dependent variable and independent variable .I consider selected attribute as independent variable to evaluate the student external performance in various possible relation. I used here the Logistic Regression model to evaluate the student performance .

Logistic Regression Model – The main advantage of this model is to deal with the Integer value as well qualitative data type as I already mention in above section . So here I consider the Paper complexity , attendance , from the student data set . Next section is cover the case study about the list of attribute and it evaluation process using the regression model .

VII. Case Study

This part include cases study of student include the various attributes for analysis purpose to predict the student performance . To indentify the Impact of attribute and its combination to predict the student performance. Challenging task for this case study work is highlights the importance of relevant data vs irrelevant data to achieve the productive result . Some selected attribute are used for evaluate the student performance. Sample data 2550 . List of All Attribute in Data Set :

TABLE I. LIST OF ATTRIBUTE FOR STUDENTS DATA SET

Student Count	Internal Test 2	Student Name	Student 10th %	Student Gender
Subject Name	Number of Classes Delivered	Student Hall ticket Number	Student 12th %	Student Father Name
Student Attendance	Lab Attendance	Execution of Practical (Y/N)	Student Address	Student Coding Skills
Internal Test 1	Assignment Work Theory	Assignment Work Practical (Y/N)	Student Contact Details	Students Email
Exam Paper Complexity (H/L/Avg)	Syllabus Cover (Y/N)	Faculty Experience (G/A/P)	Lecture Session Interaction (Y/N)	Infrastructure (G/A/P)

List of various attribute are mention in table, but these all attribute are not contribute for evaluating the student performance for external exam. Only few selected attribute are sufficient for to compute or predict the student's performance. So selection of attribute is also play the important role for fact finding.

TABLE II. STUDENT PERFORMANCE ANALYSIS SELECTED ATTRIBUTE ATTENDANCE

Selected Attribute	80--100	60-80	40-60	30-40	0-20
Attendance	*1780*	*345*	*210*	*147*	*68*
Internal Test1 Eval	*67*	*73*	*48*	*21*	*17*
Internal Test2 Eval	*42*	*42*	*37*	*11*	*11*
External	*11*	*17*	*12*	*8*	*4*

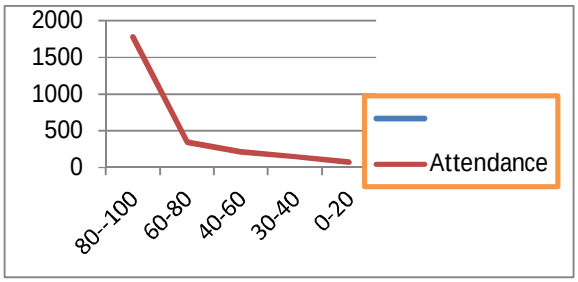


Fig 9. Student Attendance Cluster Analysis Representation

TABLE III. STUDENT PERFORMANCE ANALYSIS SELECTED ATTRIBUTE ATTENDANCE AND INTERNAL TEST EVALUATION 1

Selected Attribute	80--100	60-80	40-60	30-40	0-20
Attendance	*1780*	*345*	*210*	*147*	*68*
Internal Test1 Eval	*67*	*73*	*48*	*21*	*17*
Internal Test2 Eval	*42*	*42*	*37*	*11*	*11*
External	*11*	*17*	*12*	*8*	*4*

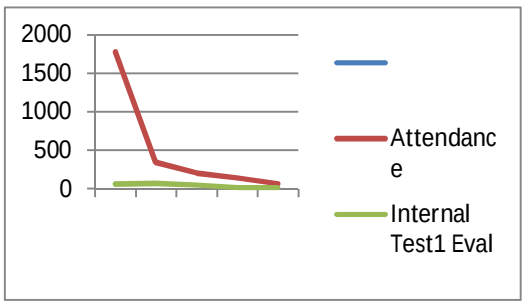


Fig.10 Student Attendance Vs Internal Assessment Cluster Analysis Representation

Here Internal Test 1 evaluation analysis is represent the statics of fail students.

TABLE IV. STUDENT PERFORMANCE ANALYSIS SELECTED ATTRIBUTE ATTENDANCE AND INTERNAL TEST EVALUATION 2

Selected Attribute	80--100	60-80	40-60	30-40	0-20
Attendance	*1780*	*345*	*210*	*147*	*68*
Internal Test1 Evaluation	*67*	*73*	*48*	*21*	*17*
Internal Test2 Evaluation	*42*	*42*	*37*	*11*	*11*

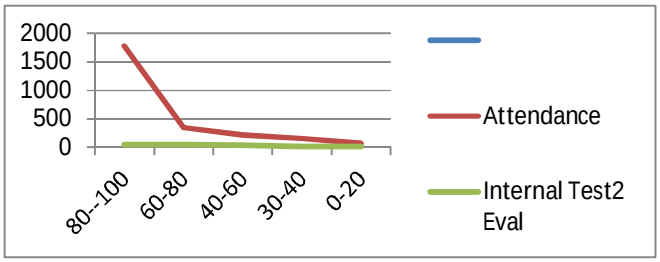


Fig 11. Student Attendance Vs Internal Assessment Cluster Analysis Representation

Here Internal Test 2 evaluation analysis is represent the statics of fail students.

TABLE V. STUDENT PERFORMANCE ANALYSIS SELECTED ATTRIBUTE ATTENDANCE AND EXTERNAL EVALUATION

Selected Attribute	80--100	60-80	40-60	30-40	0-20
Attendance	*1780*	*345*	*210*	*147*	*68*
Internal Test1 Evaluation	*67*	*73*	*48*	*21*	*17*
Internal Test2 Evaluation	*42*	*42*	*37*	*11*	*11*
External	*11*	*17*	*12*	*8*	*4*

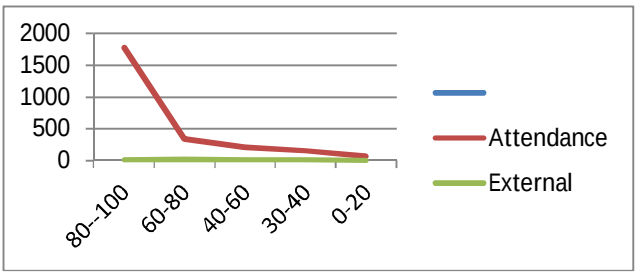


Figure 12. Student Attendance Vs External Assessment Cluster Analysis Representation

Here External evaluation analysis is represent the statics of fail students.

TABLE VI. STUDENT PERFORMANCE ANALYSIS SELECTED ALL ATTRIBUTE

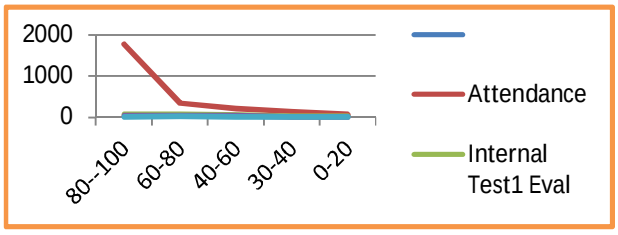


Figure 13.Student Attendance Vs Internal Test 1,2 ,External Assessment Cluster Analysis Representation

Here External evaluation analysis is represent the statics of fail students. The observation of above analysis is relevant attribute sufficient for compute the student performance as well as also increases the evaluation processes efficiency .

VIII. CONCLUSION

Through this article I try to give the basics introduction about the Data , its definition with suitable our daily life example ,this article covers the information about the various types Data and its various point of sources form where data will be collected . This paper includes the introduction of Types of data that are widely used by data analytics for various purposes .Then I cover some information about Data Quality , it's important , data preprocessing ,flowed by literature survey. Case study of this article is indicate the a optimal attribute is also generate the result for fact finding . One more important things is that if we analysis the data set with relevancy of fact ,so it will optimize the data processing steps.

REFERENCES

[1] P.Z.Yeh and CA.Puri,"An Efficient and Robust Approach for Discovering Data Quality Rules," In 22nd International Conference Proceedings on Tools with Artificial Intelligence IEEE (ICTAI) vol. 1,pp. 248–255 2010 .
[2] F.Chiang and RJMiller, "Discovering data quality rules," Proc. VLDB Endow., vol. 1, no. 1, pp. 1166–1177, 2008. Check
[3] IBM http://www- 01.ibm.com last Accessed:30-May-2016 check
[4] Gartner IT Glossary http://www.gartner.com/it-glossary/big-data/. Accessed: 30-May-2016 . check
[5] D.Laney, "The importance of'Big Data': A definition," Gart. Retrieved, vol. 21, pp. 2014–2018, 2012. check
[6] J.Manyikaet al.,"Big data:Thenextfrontier forinnovation,competition, and productivity," McKinsey Glob. Inst., pp. 1–137, 2011. Check
[7] A.DeMauro, MGreco, and MGrimaldi, "What is big data? consensual definition and a review of key research topics" In: International Conference on Integrated Information (IC-ININFO)pp 97 -104 American Institute of Physics AIP , 2015
[8] I.Emmanuel and C. Stanier, "Defining Big Data," In proceedings of the International conference on big data and advanced Wireless technologies pp 1-6 ACM Digital Library , Blagoevgrad Bulgaria 2016
[9] J.S Ward and A. Barker, "Undefined by data: a survey of big data definitions," ArXiv Prepr. ArXiv13095821, 2013.
[10] P Zikopoulos and C. Eaton, "Understanding bigdata: Analytics for Enterprise Class hadoop and streaming Data," 2011.
[11] Mohammed Ali and Ali Alqahtani "Clustering and Classification for Time Series Data in visual Analytics: A Survey" IEEE Access pp 181314-181338 2019
[12] Dabeeruddin Syed and Ameema Zainb "Smart Grid Big Data Analytics :Survey of Technologies , Techniques and Applications " IEEE Access DOI 10.1109/Access 2020.3041178. Year 2020
[13] Ikbal taleb and Mohamed Adel Serhani "Big Data Quality: Survey" In: IEEE International Congress in Big Data pp 166 -173 IEEE Honolulu, HI, USA 2018
[14] Sumit N Shuklaet "Surveyof VariousData CollectionWays for SmartTransportation Domain of Smart City" International conference on Io Tin Social Mobile (ISMAC) Analytics and Cloud (I-SMAC 2017) pp 681-685 IEEE Chennai, India (2017)
[15] Jayaram Harihara Krishnan "Survey of pre-processing techniques for mining big data"In: International Conference on Computer, Communication and Signal Processing (ICCCSP) Publisher IEEE Chennai, India (2017)
[16] https://www.google.com/imgres?imgurl.
[17] https://theintactone.com/2018/02/27/br-u2-topic-6-methods-of-data-collection-primary-and-secondary-sources/
[18] https://www.projectpro.io/article/data-analysis-process/612
[19] https://www.google.com/imgres
[20] https://www.analyticsvidhya.com/blog/2021/
[21] https://blogs.sas.com/content/sgf/2019
[22] https://www.researchgate.net/figure/Data-collection-sources_fig1_290989398