Data Collection and Preprocessing Techniques for Business Data Using Data Mining

Suman Khurana
Associate Professor
Dept. of Computer Science and Applications
K.M.V., Jalandhar
Punjab, India
kmvskhurana@gmail.com

Nidhi Kochhar
Student
Dept. of Computer Science and Applications
K.M.V., Jalandhar
Punjab, India
nidhikochhar280@gmail.com

Abstract: *Business sector produces enormous quantity of data that clutches complex information relating to customers and their expectations. Correct information can help businesses to make better decision making. This huge amount of data cannot be processed using traditional methods because of the complexity and volume of the data, that's why Data collection and data preprocessing methods are used to collect data in the correct manner. Business data comprises multimedia content as well. Data cleaning methods help to clean the data. Integration and transformation help to transform the heterogeneous data into single format. This paper features various data collection methods and data pre-processing techniques along with their pros and cons.*

Keywords: **Data Collection, Data Cleaning, Data mining, Data preprocessing, Business data, Knowledge Discovery in Databases.**

1. INTRODUCTION

Data can be defined as raw facts and figures. Today, Data is everywhere but this data is of no use if it is not converted into the information that we require. So, it is necessary to analyze the data and extract the useful information from it. Data plays a vital role in businesses. Data are essentially the plain facts which are collected during the day to day operations of a business. It is rightly said that the success of any business lies in its ability to search and analyze the data. In the world of business, it is the most important task because without processing the data, useful information can't be extracted. With the right information in hand, businesses can analyze the market trends and can take decisions according to it.

To convert raw data into useful information a process called Data Mining is used. Data Mining is the process of extracting hidden information from the large databases for analysis purpose. Data Mining is also known as Knowledge Discovery in Database (KDD), data mining has a great importance in businesses. It helps businesses in Sales Forecasting - can help owners to show how many customers in the market are buying, what is the latest trend.

The Data Mining process breaks down into five steps. First, Organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams and information technology professionals access the data and determine how they want to organize it. Then, application software sorts the data based on the user's results and finally, the end user presents the data in an easy-to-share format such as a graph or table [1]. Data mining process includes data preprocessing which is an important procedure that includes data cleaning, data transformation and feature selection. Feature selection is a data preprocessing step in the data mining process, which can be employed to reduce storage requirements while also maintaining the minimum quality. This process aims to filter out outliers (or noisy data) from a given dataset. When the dataset is very large in size, more time is required to accomplish the selection task [2]. Therefore, it is necessary to understand the data deeply so that useful information can be extracted from it. This paper mainly focuses on the different ways of collecting data and the various data preprocessing techniques.

2. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

Knowledge Discovery in Databases (KDD) is defined as a method of finding, transforming, and refining meaningful data and patterns from a raw database in order to be utilized in different domains and applications. KDD is a lengthy and complex process which involves many steps and iterations [3]. KDD (Figure1) is an iterative and interactive method consisting of seven steps. The procedure starts with determining the KDD goals and ends with the execution of the discovered knowledge [4].

A. Goal-Setting and Application Understanding

This is the first step in the process and requires prior understanding and knowledge of the field to be applied in. This is where we decide how the transformed data and the patterns arrived at by data mining will be used to extract knowledge. This premise is extremely important which, if set wrong can lead to false interpretations and negative impacts on the end user.

B. Data Selection and Integration

After setting the goals and objectives, the data collected needs to be selected and segregated into meaningful sets based on availability, accessibility importance and quality. These parameters are critical or data mining because they make the base for it and will affect what kinds of data models are formed.

C. Data Cleaning and Preprocessing

This step involves searching for missing data and removing noisy, redundant and low quality data from the data set in order to improve the reliability of the data and its effectiveness. Certain algorithms are used for searching and eliminating unwanted data based on attributes specific to the application.

D. Data Transformation

This step prepares the data to be fed to the data mining algorithms. Hence, the data needs to be in consolidated and aggregate forms. The data is consolidated on the basis of functions, attributes, features etc.

E. Data Mining

This is the root or backbone process of the whole KDD. This is where algorithms are used to extract meaningful patterns from the transformed data, which help in prediction models. It is an analytical tool which helps in discovering trends from a data set using techniques such as artificial intelligence, advanced numerical and statistical methods and specialized algorithms.

F. Patterns Evaluation /Interpretation

Once the trend and patterns have been obtained from various data mining methods and iterations, these patterns need to be represented in discrete forms such as bar graphs, pie charts, histograms etc to study the impact of data collected and transformed during previous steps. This also helps in evaluating the effectiveness of a particular data model in view of the domain.

G. Knowledge Discovery and Use

This is the final step in the KDD process and requires the knowledge extracted from the previous step to be applied to the specific application or domain in a visualized format such as tables, reports etc. This step drives the decision-making process for the said application [3].

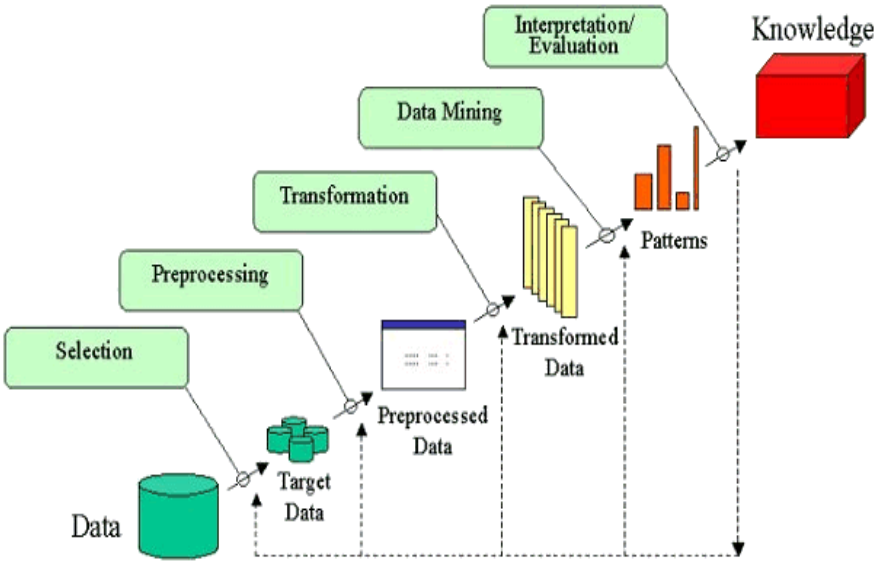

Figure.1- KDD Process

3. DATA COLLECTION METHODS

Data can be collected in two data sources, which are categorized as primary data sources and secondary data sources.

1. ***Primary data:*** Data that are collected freshly for the first time. Examples- Observation, survey, interview and focus groups.
2. ***Secondary data*:** Data collected from a source that has already been published in any form is called secondary data. e.g. Research articles, Internet and library [19].

Sources of Primary Data: Sources for primary data are limited and at times it becomes difficult to obtain data from primary source because of either scarcity of population or lack of cooperation. Following are the sources of primary data:

- ***a. Experiments:*** Experiments require an artificial or natural setting in which to perform logical study to collect data. Experiments are more suitable for medicine, psychological studies, nutrition and for other scientific studies. In experiments the experimenter has to keep control over the influence of any extraneous variable on the results.
- ***b. Survey:*** Survey is most commonly used method in social sciences, management, marketing and psychology to some extent. Surveys can be conducted in different methods.
- ***c. Questionnaire:*** It is the commonly used method in survey. Questionnaires are a list of questions either open-ended or close-ended for which the respondents give answers. Questionnaire can be conducted via telephone, mail, live in a public area or in an institute through electronic mail or through fax and other methods.
- ***d. Interview:*** Interview is a face-to face conversation with the respondent. In interview the main problem arises when the respondent deliberately hides information otherwise it is an in depth source of information. The interviewer can not only record the statements the interviewee speaks but he can observe the body language, expressions and other reactions to the questions too [19].

Sources of secondary Data: Following are some ways of collecting secondary data. It may be either published data or unpublished data. Published data are available in various resources including:

a. Government Publications
b. Public records
c. Historical and statistical documents
d. Business document

Unpublished data includes:

a. Diaries
b. Letters
c. Unpublished biographies etc [14]

4. BUSINESS DATA COLLECTION METHODS

Today, Data is very important for businesses. Right information can help to know more about the customers. Businesses should understand their customer's needs carefully. The more they understand the more they can fulfill their customer's expectations. Most frequently used data collection methods which businesses use are:

a. Online Tracking: Company's Website or mobile app is one of the most important tools for collecting client data. When users visit Company's website, he/she spends some time on certain pages. From this one can easily find out what number of individuals visited the website, to what extent they were on it, what they engaged with and much more. Google Analytics is the best example to gather this kind of information.

b. Social Media Monitoring: This is another great data collection technique. From social media, one can easily know about the market trends. Using the social media businesses can easily make good reputation by offering good quality products and services at reasonable prices.

c. In-Store Traffic Monitoring: If it is a physical store, owner can collect data from monitoring in-store activity. The best approach to do this is by keeping track of the number of individuals that come into the store each day. This can help in business growth by knowing about the customers choices and what kind of improvement they want in store and what they love. In-store Traffic Monitoring helps retailers optimize and plan their product offerings [15].

d. ***Use Online Marketing Analytics:*** Businesses can gather essential data through marketing efforts. They can even import information from offline marketing efforts that they run. Through marketing campaigns, they can collect data about who clicked on their advertisements, what time of day they were clicked, what device they were using, and much more.

e. ***Transactional Data Tracing:*** Regardless of whether you sell products in-store, online, or both, transactional data can give significant insights about the customers and the business. This data may originate from online store or in-store record keeping framework. This information can give you insights about what number of items sold, how frequently certain individuals buy and even what days and times are most profitable for the business [15].

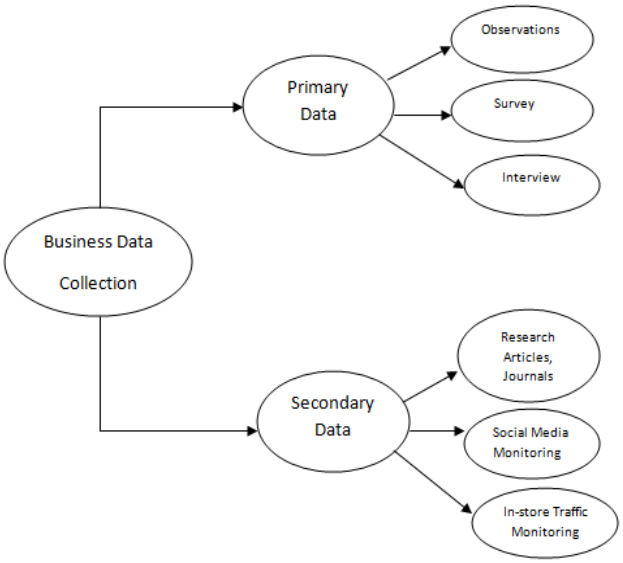

Figure.6 Business Data Collection Methods

5. DATA PREPROCESSING

Data Preprocessing is an important step in data mining process. Data Preprocessing is a Data Mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent and/or lacking in certain behaviors or trends and is likely to contain many errors. Data Preprocessing is a proven method of resolving such issues. [5]. Data preprocessing can turn the raw data into an organized form, it can sort it and merge it.

6. NEED OF DATA PREPROCESSING

Real World data is usually incomplete, noisy and inconsistent. Raw data can have missing values as well as present a lot of redundant information. The most common problems with raw data can be divided into 3 groups:

a. ***Missing data:*** It is inaccurate data because some data is missing from the data. Filling the missing values is a very difficult task. Missing data often appears when there is a problem in the collection phase, such as a glitch that caused a system's downtime, mistakes in data entry or issues with biometrics use, among others.

b. ***Noisy data:*** This group encompasses erroneous data and outliers that one can find in the data set but that is just meaningless information. Here one can see noise made of human mistakes, rare exceptions, mislabels and other issues during data gathering.

c. ***Inconsistent data:*** Inconsistencies happen when you keep files with similar data in different formats and files. Duplicates in different formats mistakes in codes of names, or the absence of data constraints often load to inconsistent data that introduce deviations that you have to deal with before analysis. If these issues are not handled carefully then final output would be plugged with faulty insights. That's especially true for more sensitive analysis that can be more affected by small mistakes, like when it's used in new fields where minimal variations in raw data can lead to wrong assumptions [6].

7. DATA PREPROCESSING TECHNIQUES

Raw data tends to be incomplete noisy and inconsistent so it important to covert this data and Data preprocessing Techniques can help in converting it into useful information. Data preprocessing is one of the most critical steps in a data mining process which deals with, the processing and transformation, of the initial dataset. Data preprocessing methods are divided into following

categories [8]:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

A. Data Cleaning

Real world tend to be incomplete, noisy and inconsistent. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. "Dirty" data can cause confusion for the mining procedure. Although most mining routines have some procedures, they deal incomplete or noisy data which are not always robust. Therefore, a useful data preprocessing stop is to run the data through some data cleaning routines. Table1 shows an example of dirty data [9].

Table 1: Dirty data

Dirty data	Problems
Gender = S	Wrong value
Address = 00	Incomplete record
C1_name = Rose M	Duplicate record
C2_name = R. Mohan	
Name = Rose 15-10-2015	Multiple values in single column

Following are the basic methods of data cleaning:

1. ***Missing Values:*** If there are records with unrecorded values for its records then these values may be filled using the following ways:
   - ***a. Ignore the tuple:*** This choice is selected when the value of class label does not exist. This method is not effective but it is used when the tuple have several attributes with empty values.
   - ***b. Enter the Missing value manually:*** When the database contains large missing value, then filling manually method is not feasible. Meanwhile, this method is time consuming. Another method to fill it global constant.
   - ***c. Filling the missing value with attribute mean:*** Filling the missing value with attribute value can be the other option. Filling with the most probable value uses regression, Bayesian formulation or decision tree [6].

2. ***Noisy Data:*** Noise refers to any error in a measured variable. If a numerical attribute is given you need to smooth out the data by eliminating noise. Binning method is the best method that can remove noisy data. Some of the binning techniques are as follows:
   - ***a. Smoothing by bin means:*** In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
   - ***b. Smoothing by bin median:*** In this method, each bin value is replaced by its bin median value.
   - ***c. Smoothing by bin boundary:*** In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Every value of bin is then replaced with the closest boundary value [23].

Let us take an example that shows how data smoothing works:

Sorted data for price (in dollars):

2, 10, 18, 18, 19, 20, 22, 25, 28

Partition into (equal-depth) bins:

- Bin 1: 2, 10, 18
- Bin 2: 18, 19, 20
- Bin 3: 22, 25, 28

Smoothing by bin means:

- Bin 1: 10, 10, 10
- Bin 2: 19, 19, 19
- Bin 3: 25, 25, 25

Smoothing by bin medium:

- Bin 1: 10, 10, 10
- Bin 2: 19, 19, 19
- Bin 3: 25, 25, 25

Smoothing by bin boundaries:

- Bin 1: 2, 2, 18
- Bin 2: 18, 18, 20

-Bin 3: 22, 22, 28

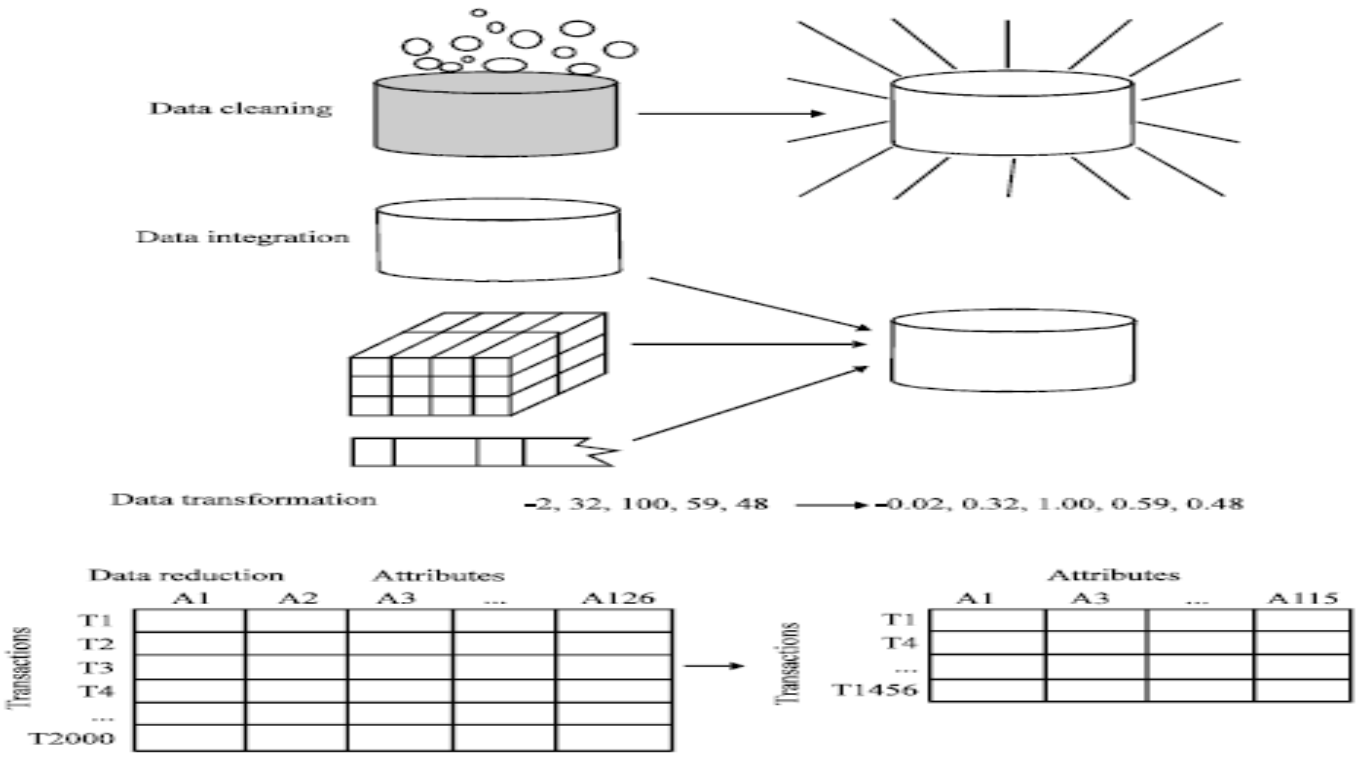

Figure.3 Forms of Data Preprocessing

3. Regression:

Regression is used to predict the value. Linear regression uses the formula of a straight line which predicts the value of y on the specified value of x whereas multiple linear regression is used to predict the value of a variable is predicted by using given values of two or more variables [10].

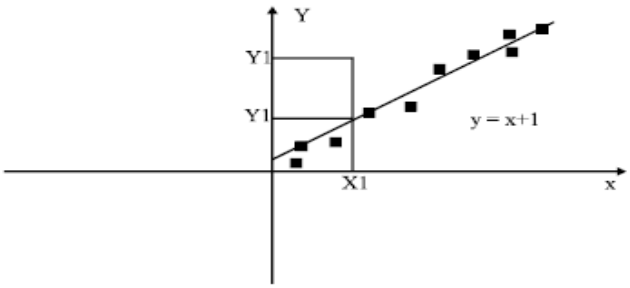

Figure.4 Linear Regression

4. Clustering:

Clustering is defined as grouping set of points into clusters according to a distance measured. It splits data into several subsets. Each subset contains data that is similar to each other and these subsets are called clusters. The result of clustering is a set of clusters each cluster will have set of points with small distance from one another and with large distance from other clusters. This technique can detect outliers, since it grouping similar points into a cluster while points that fall out the clusters are considered as outlier points. As shown in this figure5, there are 3 clusters and the points which do not belong to any clusters are outliers [10].

B. Data Integration

Data Integration merges data from several heterogeneous sources to attain meaningful data. The sources involve several databases, multiple files or data cubes. The integrated data must exempt inconsistencies, discrepancies, redundancies and disparity. Data integration is important as it provides a unified view of the scattered data not only this it also maintains the accuracy of data. This helps the data mining program in mining the useful information which in turn helps the executive and managers in taking the strategic decisions for the betterment of the enterprise [12].

- ***a. Tight Coupling:*** Here, a data warehouse is treated as an information retrieval component. In this coupling, data is combined from different sources into a single physical location through the process of ETL- Extraction, Transformation and Loading.
- ***b. Loose Coupling:*** Here an interface is provided that takes the query from the user, transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result and the data only remains in the actual source databases [11].

C. Data Transformation

In data transformation process data are transformed from one format to another format that is more appropriate for data mining. Data are transformed into appropriate forms of mining. Data Transformation involves the following:

- ***a. Smoothing:*** It is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.
- **b.** ***Aggregation:*** Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. For example Sales, data may be aggregated to compute monthly and annual total amounts.
- ***c. Discretization:*** It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes. Most Data Mining activities in the real world require continuous attributes. If a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete value.
  For example (1-10, 11-20) (age: young, middle age, senior).
- **d.** ***Generalization:*** It converts low-level data attributes to high-level data attributes using concept hierarchy.
  For example Age initially in Numerical form (20, 25) is converted into categorical value (young, old).
- **e.** ***Normalization:*** Data normalization involves converting all data variable into a given range. In Normalization, where the attribute data are scaled to fall within a small specified range.
  For example -1.0 to 1.0 or 0 to 1.0 [20].

D. Data Reduction

Data Reduction is an important step of data preprocessing. The data reduction technique helps to minimize the size of the dataset without affecting the result. The following are the methods that are commonly used for data reduction:

- ***a. Data cube aggregation*****:** It refers to a method where aggregation operations are performed on data to create a data cube, which helps to analyze business trends and performance.
- ***b. Attribute subset selection*****:** It refers to a method where redundant attributes or dimensions or irrelevant data may be identified and removed.
- ***c. Dimensionality reduction*****:** It refers to a method where encoding techniques are used to minimize the size of the dataset.
- ***d. Numerosity reduction*****:** It refers to a method where smaller data representation replaces the data [21].

8. CONCLUSION

Data Collection helps to collect large amount of data. Through various data collection methods huge amount of data is generated in Business sector which is lacking useful information to make decision and analysis. The data mining has played an important role in businesses. It helps to discover patterns and relationships in the data so that better decisions can be made. Rich data exists in businesses but researchers finds difficulty to collect data through primary and secondary techniques because real world data contains inconsistent and noisy data that leads to wrong decision making. After data collection, data preprocessing techniques are used to clean the data. Data cleaning help to fill up the missing values and data integration combines all the data from the multiple sources and provides a unified view. Then, data is transformed in appropriate forms. Finally, Data Reduction technique is used to reduce data storage and analysis costs. After all the steps are successfully completed finally we get the information that can help businesses to solve their problems and make decisions according to it.

REFERENCES

[1]https://www.investopedia.com/terms/d/datamining.asp#:~:text=Data%20mining%20is%20a%20process,increase%20sales%20and%20decrease%20costs

[2] https://pubmed.ncbi.nlm.nih.gov/25515050/

[3] https://www.upgrad.com/blog/kdd-process-data-mining/

[4] Jaiwei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, ISBN 13:978-1-55860-901-3.

[5]https://www.techopedia.com/definition/14650/data-preprocessing

[6]https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-is-important/

[7]https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[8]http://apps.iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf

[9]https://medium.com/easyread/basics-of-data-preprocessing-71c314bc7188

[10]Saud A. Alasadi and Wesam S. Bhaya,"Review of Data Preprocessing Techniques in Data Mining" Medwell Journals 2017

[11]https://www.geeksforgeeks.org/data-integration-in-data-mining/

[12] https://binaryterms.com/data-integration.html

[13]https://medium.com/easyread/basics-of-data-preprocessing-71c314bc7188

[14] https://byjus.com/maths/data-collection-methods/

[15]https://lionbridge.ai/articles/8-simple-data-collection-techniques-for-businesses/

[16]https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.skedsoft.com%2Fbooks%2Fdata-mining-data-warehousing%2Fintroduction-to-data-preprocess&psig=AOvVaw04OeVGp_VAV_T5N7D1DTwh&ust=1610445972251000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCMCe4_rQk-4CFQAAAAAdAAAAABAZ

[17]https://www.google.com/url?sa=i&url=http%3A%2F%2Fwww2.cs.uregina.ca%2F~dbd%2Fcs831%2Fnotes%2Fkdd%2F1_kdd.html&psig=AOvVaw2lEw-ETTnRuW71VUG2VLkF&ust=1610446337407000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCIi8-qHSk-4CFQAAAAAdAAAAABAD

[18]https://www.researchgate.net/publication/325846997_METHODS_OF_DATA_COLLECTION

[19]https://www.geeksforgeeks.org/data-transformation-in-data-mining/

[20]https://www.geeksforgeeks.org/data-reduction-in-data-mining/

[21]http://www.lastnightstudy.com/Show?id=45/Data-Discretization-and-Concept-Hierarchy-Generation

[22]https://www.includehelp.com/basics/data-preprocessing-in-data-mining.aspx