Educational Administration: Theory and Practice
2024, 30(5), 10844-10854
ISSN: 2148-2403
https://kuey.net/
Research Article

Data Pre-Processing And Its Implications In Data Mining

Dr Preeti Bala[1*], Himani Tyagi[2], Ms. Rashmi Vaishnav[3], Shikha Tiwari[4], Sunil Kumar[5]

[1*]Assistant Professor Institute of Management studies, Ghaziabad (University Courses Campus) Preeti.bala@imsuc.ac.in
[2]Assistant Professor Sharda University (School of Computer science) himani.tyagi@sharda.ac.in
[3]Assistant Professor Institute of Management studies, Ghaziabad (University Courses Campus) rashmi.vaishnav@imsuc.ac.in
[4]Assistant Professor Institute of Management studies, Ghaziabad (University Courses Campus) shikha.tiwari@imsuc.ac.in
[5]Assistant Professor Institute of Management studies, Ghaziabad (University Courses Campus) Sunilkumar.sharma @imsuc. ac.in

Citation: Dr Preeti Bala *(2024),* Data Pre-Processing And Its Implications In Data Mining *Educational Administration: Theory and Practice, 30(5),* 10844-10854
Doi: 10.53555/kuey.v30i5.4848

ARTICLE INFO **ABSTRACT**

This study on data preparation and its techniques demonstrates the necessity of data pre-processing. How does it raise data quality, as good data produces results that are appropriate? This Paper also discusses the central role of data processing in data analysis. What different facets make up the KDD process? Without data pretreatment, we cannot perform KDD. Along with data visualization, this work also discusses data transformation and discretization. This paper also describes the methods and tools of data preprocessing as well as focused data preprocessing and its methods. Apart from that, it also discusses data visualization. The handling of the missing data was another major topic of the article. In this research Paper also delves a small project and describes a case study.

Keywords: *Data mining, Data pre-processing, Tools, Techniques, Quality of data*

I. INTRODUCTION

Data preprocessing is a method in which we process a Set of data with the help of some procedures, these procedures are Applied on raw data to getit prepared for data mining. This raw data is the primary data collected directly from source for use which has to go under many Phases like extraction, organization etc.to become useful information.

Data preparation makes the data effective. By turning raw data into information that may be used in a variety of data-related applications. Data pre-processing is commonly used as a preceding in data mining like a Neural network.[2]

For pre-processing, there are several different tools and methods which include sampling (from large population of data, it selects a representative subset), Transformations (Single input is produced by the manipulation of raw data), de-noising (removing elements from data that make data noisy) normalization and many more.

According to a customer relationship management context, data pre-processing is accomplished by web mining. In this User transactions are observed by web page logs, from there meaningful of data is gathered. User sessions multiple things are identified like the websites used regularly, their order accessing these sites, the time spent on each website.

When the raw data got gathered from these observations, more useful information is also collected for user's purpose like consumer search, interested marketing area. The raw data collectively make a database. This data is now available for more processing.

Now consider the question why there is a need to -data pre-processing. The following techniques help the data to get analyzed by data mining technique as follows:

- **Incomplete-** Incomplete data means there are missing some attribute values. In tuples having some missing values or some attributes, may need to be inferred.
- **Noisy** –Data is said to be noisy if there are some errors or over evaluated values that are derived from the expected values. Noise modifies the original values.

With the help of some techniques noisy data can be corrected.

- **Inconsistent**-It is when the data contains discrepancies between different items. For proper processing, we need to remove inconsistency in data.
- **Aggregate Information:** Aggregate Information means the actual desired data items in the data warehouse. It will be of more use if there is a provision to use aggregate information such as sales per customer region.
- **Enhancing mining process**-The data mining process become slow because of large no of data sets. Reduce number of data sets play a vital role in data pre - processing.
- **Improve data quality**– By following some techniques, the quality of data can be enhanced. In data processing. These techniques will make the data more the accurate and efficient. If we remove data anomalies, by identifying them earlier data can be analyzed.

KDD

KDD, or knowledge discovery in databases, the abbreviation, defines the general process of finding knowledge and it impacts the advanced uses of some Data Mining methods. Researchers of several areas like artificial intelligence, machine learning, databases, statistics, knowledge acquisition for expert systems, and data visualization, are interested in this area.

In the setting of huge databases, the primary goal of KDD method is to gain information from data.This is accomplished through the use of data mining algorithms to identify what is deemed to be knowledge.

Database knowledge discovery is discovered as planned elaborated investigation and modeling of huge data stores. KDD is a methodical implementation for extracting meaningful patterns from vast and complicated data sets.

Enhances decision-making: KDD offers insightful and useful information that can assist enterprises in making better decisions.

Enhanced efficiency: KDD automates time-consuming, repetitive operations and prepares the data for analysis, saving both resources.

Better customer service: KDD assists firms in better comprehending the wants and preferences of their clients, which can assist them in offering better customer service.

Fraud detection: KDD can be used to find patterns and anomalies in data that might point to fraudulent activity.

Forecasting future trends and patterns using predictive modeling: KDD can be used to create these models.

CRISP-DM: This one is the traditional methodology for data mining, which is abbreviated as Cross Industry Standard Process for Data Mining.

The given process includes six sequential steps required for the association and mining of certain data those steps are understanding of business, understanding of data, preparation of that data, modeling, evaluation and deployment having their explanation respectively:

- The first one explains the data analysis goals, means what are the requirements of that business for which model is going to be created or what are the metrics for verification of results whether the project is successful or not.
- Data understanding is the second step in this procedure that defines first the sources of data and type of data like the format, size, type of variables in it. This step is necessary for cleaning and exploration of data in any model.
- Preparation of data is the stage where data actually get ready for the analysis by cleaning, removing outliers, splitting and choosing important features etc.
- Modeling means selecting the appropriate technique that is as per the requirement of result, business objective and characteristics of that data.
- Evaluation part itself suggest that here model accuracy can be find, if using more than one technique the comparison for efficiency can also be done in this part.
- Deployment defines the integration of the model generated in the operational system, where results or performance of the model can be monitored.

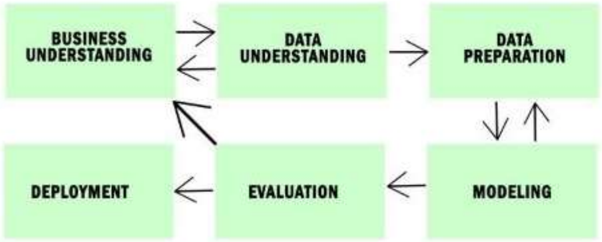


Figure 2- Crisp-DM methodology

Different keys form of data pre-processing

Sampling-sampling is a very important step of data collection, where we select a subset of the dataset that has similar properties to the original data sets. Statistician sample the whole set of interests and processing it is expensive and time-consuming too.

Data Cleaning- is defined as the process of cleaning the data. The cleaning process involves filling mission value, Making the noisy data smooth, removing over valued data etc. so that the user can trust the result of data mining applied on inappropriate data. The inappropriate data here means the Data that could create confusion in the procedure of mining, apart from that it ensures that the data used is accurate and suitable for further analysis.

Data Integration-is defined as the process of combining data from multiple databases. In a data mining environment, inconsistencies and redundancies are caused by attributes representing a given concept that has resulted from data integration.

Data Transformation- Data Transformation is the collection of data pre-processing techniques that include normalization and aggregation. Normalization involves scaling the analyzed data [1]. Data scaling yields aggregate information, which is then used in data analysis. It has to be calculated because it may not be included in any existing data collection or summary.

Data reduction ensures, we obtain set of data in a minimum representation, which produces same analytical results but is much smaller in volume.

- There is distinct variety of strategies in data reduction, listed as follows
- Data aggregation
- Attributes dependency analysis to select the most important deciding attributes for a target class (e.g.-Chi-Square Analysis)
- Attribute subset selection (e.g.-correlation analysis to remove irrelevant attributes)
- Dimension reduction

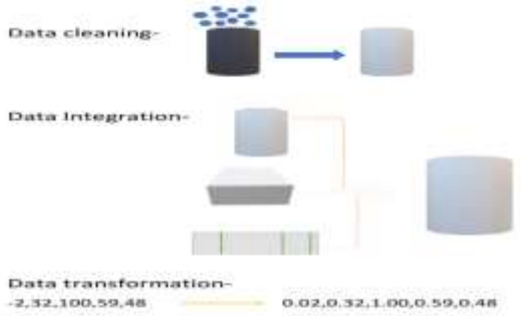


Figure 1-Forms Of Data Pre-Processing

- In this, we obtain a data set in a reduced representation, which produces the same analytical results but is much smaller in size. There is a distinct number of strategies in data reduction, listed as follows:

Data Aggregation

- Attributes dependency analysis to select the most important deciding attributes for a target class (chi square-Analysis)
- Attribute subset selection (e.g., correlation analysis to remove irrelevant attributes)

Dimension reduction (e.g., using statistical concepts such as PCA)

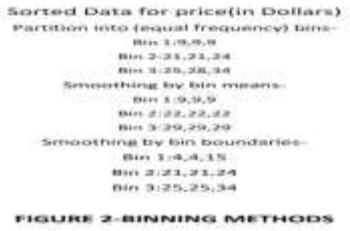
- Generalization with the use of concept hierarchies, It is done by organizing the concepts into different levels of abstraction
- Data discretization is vital for the automatic generation of concept hierarchies.
- Aggregation is a process in which there is a combination of two or more attributes into a single attribute. The main usage behind aggregation is data reduction change of scale and to get more balanced data.

Data Cleaning - Data cleaning is a method for enhancing the accuracy of the data by spotting and eliminating mistakes and inconsistencies. Data scrubbing or cleansing are other terms for data cleaning.

- In the context of data warehouses, the requirement of data cleaning is increases significantly and other types of database systems where there is integration of various other data sources. To access the consistent data, valid data, the Solidification of different data representations and removal of duplicate data becomes a mandate.
- The warehouses of data need to avoid wrong conclusions as they are used in decision making.
- The chances of the data being dirty is very much in data warehouses because in the process called ETL (extraction, Transform, load), they collect data from files and database.

- Data cleaning is known to be one of the hardest problem of data warehousing because there is wide variety of possible data inconsistencies and the absolute data volume [1,3].
- Furthermore, it is done with schema-related data transformations. It should not be carried out in separation. Data transformations and cleaning is mapping functions to be specified in such a way that data sources as well as query processing phases may reuse these data transformations.
- In consideration of reliability, the execution of all data transformation steps is efficient and thus, should be treated in a uniform manner. Any changes in the data, its structure, representation, or content are supported by data transformations.
- Noise defines the random error that can occur in data. variance in a variable for measurement. Following are some methods to manage such noisy data.
- Binning techniques arrange and smooth the neighborhood data values.
- This approach distributes sorted values across many buckets or bins. Binnig techniques can only execute the operation with neighboring values, hence they are considered to do local smoothing.
- The median is utilized to replace each bin value in the smoothing by binning technique, resulting in improved smoothing impact. In general, the higher the smoothing effect, the broader the breadth.

Alternatively, bins might be equal in width. In width the interval range of values for each bin is constant. Binning is also serving as a discretization technique.



Sorted Data for price(in Dollars)
Partition into (equal frequency) bins-
Bin 1:8,4,8
Bin 2:21,21,24
Bin 3:25,28,34
Smoothing by bin means-
Bin 1:9,9,9
Bin 2:22,22,22
Bin 3:29,29,29
Smoothing by bin boundaries-
Bin 1:4,4,15
Bin 2:21,21,24
Bin 3:25,25,34

FIGURE 2-BINNING METHODS

Figure 3-Binning Methods

Data cleaning defines a process consisting of many types of operations-

Discrepancy Detection: - Data Entry errors, the data entry forms which are designed with various optional fields, deliberate errors, decay, and inconsistency in data representations are the metadata – that we can consider for detecting discrepancy. Domain, data trends are grasped, and anomalies are identified by descriptive data summaries. When developers finalize new attribute to be set in unused portion of existing and defined attributes, It leads to a new source of error, called attribute overloading unique rules, sequencial rules, and null are used to examine the data. Simple domain knowledge is used to detect errors and correct them in the data by scrubbing tools.

When trying to get the clean data we need to discover some rules and relationships for detecting data that like conditions. For example, they are the different data mining tool variants. For example, they may make statistical analysis to find correlations or to identify the clustering to outliers.

External references are used to find out some data and its inconsistencies that may be corrected manually.

- **Data Transformation**- Data Transformation is required by most data mining applications, especially neural network-based applications. It is the second step of data cleaning procedure. A sequence of data transformation is required to be defined and applied to the data correctly continuously to check whether any new discrepancies have been introduced by mistake after the completion of the data cleaning procedure. The new data cleaning strategies signifies enhanced interactivity in the process. Metadata updation to reflect this knowledge is also an important thing. Future versions of the same data store will get speeded up by such data cleaning procedures.
- **Data Analysis:** A detailed detection is required to find out the Inconsistencies in dataset. To develop insight in values involved in data we are using, detection anomalies in the quality of data, analysis programs may be used.
- The data transformation workflow and mapping -A large number of data transformation and multiple cleaning steps are required as per the the number of data sources, the diversity of these sources and the impurities involved in the data. A schemas translation is considered to map the source to a common data model. Earlier in cleaning steps, a relational representation has been taken by the data-warehouse such that single-source instance problems were rectified to prepare data integration.
- **For data warehousing**-the work flow should be specified by the control and flow of data in the transformation and cleaning steps. Declarative query and mapping language will specify the schema

associated with data transformations along with the the cleaning steps. Automatic generation of transformation code is enabled from this.

Knowledge management is a broad concept that encompasses various multidisciplinary approaches, including content management, collaboration, organizational behavior science, trends and anomalies analysis, clustering, classification, summarization, taxonomy building, and more.

Although brief and to the point, this definition (Davenport, 1994) is perhaps one of the most commonly cited: Knowledge management is the process of acquiring, sharing, and effectively using knowledge.

To achieve corporate goals and support business decision-making based on business analytics, KM refers to a set of approaches used to collect, share, and utilize the information that is currently available.

Over the past ten years, the region of knowledge management has experienced tremendous expansion, and new tools and applications that promote information sharing have emerged.

- **Verification** -The testing and evaluation of correct and effective transformation of workflow and definition is required. The steps iterated multiple times are Analysis, Design, and verification, if needed, e.g. after applying some transformations, some errors only become apparent.
- **Transformation –** This happens as a part of the ETL phase or during answer of queries from multiple sources.
- **Backflow in cleaned data**: • After data transformation, replace dirty data with original sources to prevent repeating cleaning operations in the future. In the context of data warehousing, a data staging area stores cleaned or improved data.

Large volumes of metadata, or data about data, necessitate fundamental data structure, instance-level data properties, transformation mappings, and workflow requirements for the transformation process. This metadata is kept in a DBMS central repository, allowing for numerous characteristics like as consistency, isolation, and facilitation of future revisions. To achieve the best results, the specifics of the transformations made to the original objects should be thoroughly documented and stored in a data warehouse.

Methods to Handle Missing Data-

A common problem is handling the missing data in statistical analysis. Missing data rates which are less the 1% are considered trivial, those of 1% -5% are Considered manageable, those between 5% and 15% need simple methods to handle , and those who have more than 15% might seriously impact some kind of interpretation .

In the literature to treat missing data, there are several methods. They will treat missing data and several methods have been proposed for this.

There are four approaches for dealing with missing values in supervised classification tasks. They include the case deletion (CD) approach, mean imputation, median imputation, and k-nearest neighbour (KNN) imputation. The criterion for comparing them is the influence of these approaches on two classifiers: linear discriminant analysis (LDA) and K-Nearest Neighbour (KNN), where LDA is a parametric classifier and KNN is a nonparametric classifier.

The following four approaches are available in a supervised classification environment.

Case Deletion

CD is also called complete case analysis. Many software packages have module that implements this technique. This method includes all instances with missing values for at least one feature discarded Determining.

Mean imputation

MI. This is the most popular choice for dealing with missing values in data. In this case, the mean of all current Attribute values in that specific class is the instance that corresponds to missing attributes and will be computed and replaced for missing data features. Let's attempt to define it formally. Assume that the value Xij of the kth Class, Ck, is missing. Then it may be substituted by Xij=∑xij/nk.

The value nk represents the number of non-missing values in the jth feature of the kth class. We thought that the overall mean does not account for the sample size of the class that contains instances with mission values**.** Some of the drawbacks of mean imputation are overestimation of sample size, under-estimation of covariance, and negativity biased correlation. Data sets are supervised classification purposes give good experimental results for mean imputation.

Mode imputation- As the mean get impacts if there is some outliers, the mode may be the next choice. The Mode can be assumed as the most repeated value in a data sequence. As an example, consider the following scenario. Suppose we want to replace the missing value of the sex attribute of a record in a woman college. Naturally it has to be female. This is the notation behind the mode of certain data sequence. Another possibility can be replacement with median which is called median imputation. This one is also widely adopted technique for missing value substitution.

The correlation structure of the current data values must also be evaluated. Replacing missing data may become useless because of the existence of the other features with similar information.

KNN imputation (KNNI)- In this method, a number of instances that is mostly similar to the instance of interest are used to impute the missing value of an instance. A distance function is there to determine the similarity in two instances, The algorithm of which is as follows:

1-Data set D is divided into two parts. Instances in which at least one of the feature is missing is contained in Dm. Dc is the set of instance which will have complete feature information.
2- For each Vector x in Dm:
a) Instance vector divided into two parts that are observed and missing parts as x=[xo, xm]
b) distance between xo and all the instance vectors from the set Dc is calculated. Only feature that are observed in vector x can be used from the set Dc.
c) A maximum voting estimate is performed for the missing values for categorical attributes by using the K closet instances vectors (K Nearest neighbors)
d) The average value of the attribute in the K-nearest neighborhood is used to replace the missing value attribute. The median can be used instead of the mean.
The advantages of KNN Imputation are :
1-Qualitative attributes and quantitative attributes are predicted by the k nearest neighbor.
2 Creating a Predictive Model for each characteristic with incomplete data is not essential. The K-nearest neighbour technique does not produce an explicit model.
3- Multiple missing values can be treated.

Few Other Imputations-

Hot deck Imputation- In this approach, the value of a missing attribute is replaced with a value derived from an estimated distribution of the existing data. In the fandom hot deck, the missing value attribute is replaced with the observed value of the (donor) of the randomly picked attribute.
Imputation with prediction Model- In this model, missing data is replaced with approximated values anticipated by the model. The attributes other than the missing data values are utilized as input for the prediction model, while the attributes with missing data will be used as the response attribute.
The model-estimated values frequently perform better than the genuine ones here. If there are no links between the characteristics in the data set and the attribute with missing data, the model will be inaccurate when predicting missing values.

We must construct a huge number of models to anticipate missing values based on computational cost.
Imputation uses decision tree techniques. All decision tree classifiers employ built-in ways to handle missing data. The missing value of a particular attribute is replaced by the value of the surrogate attribute having the highest connection to the original attribute. This technique use the CART algorithm to manage missing data in both the training and testing samples.

Multiple imputations: - In this technique, missing values in a feature are filled up with values taken randomly from a fitted distribution for the treatment values and their influence on the classifier accuracy. For precision, the same process is done multiple times (say, M=15). After that, the misclassification error for each data set is calculated by applying the classifier to the entire dataset. A single estimation is obtained by averaging the misclassification error rates .The Variance of the error rate is also estimated.
Data transformations and discretization- This section will go over some of the most popular forms of transformation and normalization. In attribute transformation, for example, 0 can be used for male and 1 for female.
To apply decimal scaling to the range [0,1]/[0,1,0.4,0.5,0.65,0.85,1.0}, divide by the maximum value 20. However, the most often used data normalization approach is min-max normalization, which converts data to a new range [0,1] by applying a single formula to all data[4]. It is done by translating their origin, adding/subtracting a constant, and scaling down the range to another value. The data values are adjusted to another value, as in ceil (2.5)= 3, i.e. rounding value transformations; in Scrubbing transformations, text strings are located and changed whenever they are required. For example, if all names "Williams "were entered as "bill" owing to a data entry error, a find-replace function may be used to repair the mistakes.

Data Visualization – Information that has been in the schematic form includes variables for the units of information are called data visualization. This will represent the data visually. The structure of data is to be mined is analyzed by data visualization. So, Data visualization is a very important task in data -mining. The similarities, dissimilarities, clusters, and dependencies that exist in data are visualized by this technique.
Weka is one of the most popular free and open-source software that is used nowadays for academic and research activities in data mining. It was developed by a team at the university of Waikato, New Zeeland. It was developed as a collection of Java classes that are reusable, modifiable and distributable. Almost all data mining algorithms and classifiers are available as Java classes, and they may be developed as a collection of Java classes, as well as new algorithms by editing the source code. WEKA expertise may be used to graphical user interfaces that contain source code. WEKA knowledge explorer has a user-friendly interface that boosts the capabilities of the WEKA program. WEKA's Explorer tool allows you to manage major data mining tasks like classification and clustering graphically. The visualization tool assists in obtaining a visual interpretation of the outcomes of these exercises.

Along with the software package Weka, there are many standard data sets available, Such as IRIS, CAR, and DIABATES, which are maintained by UCI. These are the standard data sets that are used as the benchmark for testing the majority of new data mining algorithms in data mining research.

Data visualization takes advantage of people's innate need to visualize their surroundings. When attempting to comprehend images or lengthy texts, the human brain responds differently. When reading, the information must first be processed by the brain before ideas can be formed in response to it. In contrast, the mind instantly processes information presented in a visual format as if it were a scene or a new location.

Simple data visualization features like making objects appear bigger, smaller, shorter, or taller have a significant psychological effect. Effective data visualization harnesses our psychological reactions to pictures in combination with a dataset to give a succinct summary. The goal of visualization, according to eminent American computer scientist Ben Schneiderman, is insight rather than pretty graphics.

Data visualization has many aspects that help us make our business strategy stronger and gain deep insights through different tools that help us evaluate our business strategy, like Organizations may assess the success of their strategies and have clear indications of any emerging concerns with the aid of data visualization, which they can then readily communicate to other team members.

Data visualization helps us to identifying error. Data visualization is useful for identifying and fixing problems since the psychological impact of numbers on a spreadsheet differs from a graphical depiction of what those numbers signify.

Data visualization helps us make decisions. In the modern organization era, we are continuously dealing large amount of data that can play a crucial role in taking decisions and making good strategies. To deal with such an amount of data, we require technological help. Data visualization tools help us provide the output in a minute, which we can find manually in a month.

Businesses may expedite the decision-making process while also having the data to support their decisions by integrating data visualizations tools into their processes.

Data Visualization performs sentiment analysis –

Sentiment analysis is generally performed by companies and brands to check customer satisfaction. Data visualization helps in sentiment Analysis Huge datasets are used in sentiment analysis, and unless an algorithm runs through them and parses the data, they are completely unintelligible. If the end user wants to understand this catalogued data correctly, data visualizations tools must be used after that.

Figure 4: shows the visualization of the standard data set named as IRIS that is available with the WEKA Package.

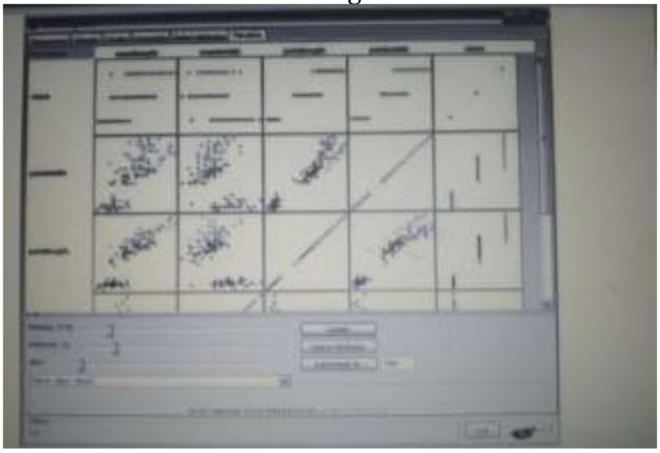


Figure 5: shows the data distribution with respect to classes along with the class distribution too. The visualize pane of the WEKA package is the exclusive pane in WEKA for data visualization. It could be used to analyzing various similarities and dependencies that are inherent in the data

Attributes dependency Analysis selects those attributes that are decisive for a target class. Chi-square analysis is a popular tool for this purpose. Finally, dimensionality reduction is a step in data preprocessing that minimizes the amount of characteristics in a data source. Principal component analysis (PDA) and linear discriminant analysis (LDA) are common dimensionality reduction techniques.

Summary statistics are numbers that summarize multiple statistical aspects, providing insight into the nature of data. Some examples include frequency, location, average, and dispersion, such as standard deviation. The percentage of times a value occurs in a data set is termed as its frequency. Consider the attribute "gender" and a sample of people. Suppose the gender female occurs about 50 % of the time. Here female can be taken as mode. Categorical data uses notations of frequency and mode.

Project Example-

In a related research work, the training of neural network is done by data transformation. The inputs to the problem are particular rank sex, reservation, sector, and branch of a student. The output is to predict the placement chances for that student. Inputs are fixed as categorical data and output has to be one from the set {Excellent, Good, Average , Poor}

Usually, transformation operations involve converting categorical data to numeric data. Normalization is the process of distributing the data evenly and scaling down into an acceptable range for the model. The details of the needed transformations are shown in Table 1

Attribute	Range	Mapped to
Rank	1to 4000	0 to 1
SeX	1 to 2	0 to 1
Category	1to 4	0 to 1
Sector	1to 2	0 to 1
Branch	A to J	0 to 1

Table 1: Attributes values mapped to 0 to 1 scale

Activity1 to 4 One of the four values:E, G, A and P

For data normalization,(1) was used for transforming each data value D to I

I=Imin+(Imax-Imin)x(D-Dmin)/(Dmax-Dmin)

Linear scaling requires minimum and maximum values for each data input. The values are Dmin and Dmax, respectively.Imin-Imax is the input range required by the network. Dmin and Dmax are computed using attribute values. Because the neural network takes input in the range of -1 to 1 or 0 to 1, all input and output data are translated to values between 0 and 1. For example, the number 500 in the range {1-5000} will be translated into 0+(1-0)x(500-1)/(5000-1) =0.099 in the range 0 to 1.

Table 2 contains sample data used to train neural networks. Many data mining applications employ neural network methods. All of them are provided categorical or numeric data as inputs they require data transformations. Many built-in data preprocessing utilities are provided with many sophisticated data mining packages such as WEKA .one of the most popular tools is SPSS whose latest version is IBM SPSS statistics 21.0 as on August, 2012. There are many add-ons such as SPSS missing values and SPSS data preparation, which is exclusive for data preprocessing .Another alternative is R Commander programming language.

SEX	Reservation	Location	Rank	Branch
0	0	1	0.72	0.47
1	0	1	0.72	0.47
0	1	0	0.59	0.33
0	0	1	0.4	0.66
0	1	0	0.72	0.47
1	1	1	0.27	0.38

Table 2- Snippet of the sample data used to train the neural network

CASE STUDY

Data mining with data of Super Market :

Supermarket loyalty card programmers typically collect substantial consumer data for data mining purposes. This happened lately, particularly with the American store Target. As part of its data mining program, the company developed criteria to predict whether its clients would become pregnant. They were able to target marketing for nappies (diapers), cotton wool, and other things by identifying consumers who seemed to be expecting based on the items in their shopping carts. Because the forecast was so accurate, Target made news by providing discounts to those who had not yet found (or announced) what they were expecting!

Performance Evaluation: Comparing KDD and CRISP-DM for Data mining for given case study.

We presented the results of our comparative analysis of two Data mining methodologies- Knowledge discovery database and the traditional one CRISP-DM for mining a Supermarket loyalty card data. We evaluated the methods on the basis of visual interpretation of the graph created, and the results illustrate the distinct performance characteristics of each approach.

The result in our visualization is revealing the intriguing insights and elucidating the performance disparities between the two techniques of data mining. While both models were able to capture feature importance analysis trends, the Linear Regression model demonstrated a smoother and more consistent fit to the observed data points, indicating almost same result in features analysis. However, when we visualize through bar chart KDD is coming more efficient in giving result in comparison to CRISP-DM. here are the graphical presentation of results we can include

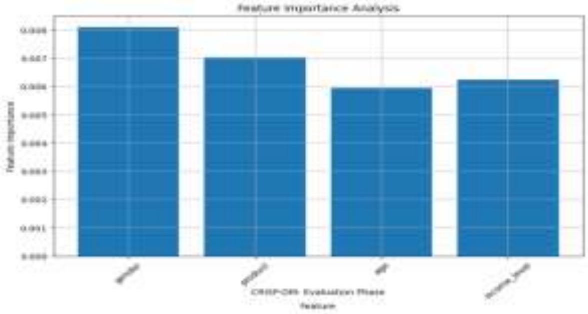


Figure 6- CRISP-DM Feature importance analysis

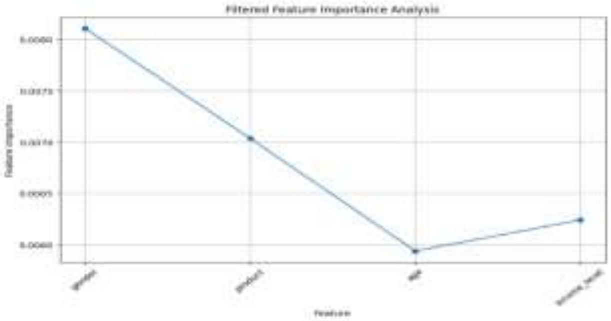


Figure 7: Feature analysis through regression method for the dataset of supermarket loyalty card.

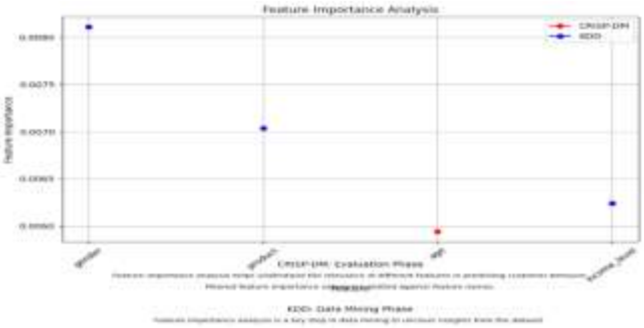


Figure 8: Feature importance analysis for both mining techniques

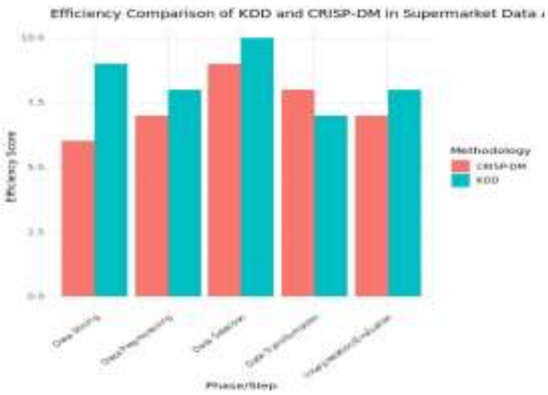


Figure 9: Showing efficiency comparison for both mining methods

Conclusion:

Overall, in our comparative analysis it can be seen that in the case of particular dataset that is of supermarket loyalty card dataset KDD, knowledge discovery database is performing better instead of traditional CRISP-DM methodology of Data mining. In the given Bar Graph it is clearly showing the better efficiency of KDD.

Discussion Questions :

1. Explain the various missing values in handling data. Discuss different methods through which we can handle missing data?
2. What is data visualization and how does it help to make an intelligent business decision?
3. What is Dimensionality Reduction? How can we minimize it using other methods? How may adding dimensions to our datasets make things more complicated?

References

[1] Bifet A.,"Mining Big Data in Real Time",Informatica journal,vol. 37,Number 1,2013

[2] Oracle : Big Data for Enterprise ,An Oracle White paper ,June 2013

[3] Peglar R ," Introduction to Analytics and Big Data Hadoop" ,EMCIsilon,Storage and network Industry Assciation, 2012.

[4] Labrinidis A, Jagadish H.V.,"challenges and opportunities with Big Data,procedding of the VLDB Endowment" ,Vol 5,Issue 12,August 2012.

[5] Han J, kamber M ,Pei J, Data,"Mining Concepts Technique", 3rd Edition , MK Publisher.

[6] Guptha G. K. , "Introduction to Data mining with Case studies" ,PHI Publishers ,2006.

[7] Tan P. N.,Streinbach M., Kumar V., "Introduction to Data mining Techniques", Pearson, 2006

[8] Elayidom S., Idicula S.M. and Alexandar J ,"Comparison of Data Mining Techniques Based on Decision Trees and Nureal Networks for placement Chance prediction", Proceddings of National Conference ICONCEPT , Kerla , India

[9] https://www.integrate.io/blog/real-life-applications-of-data-mining-and-business-intelligence/

[10] Kotsiantis S. B., Kanellopoulos, D. N.,"Survey of data preprocessing techniques in data mining", Journal of Computer Science, 1(1), 27-36, 2006. (https://acadpubl.eu/jsi/2017-117-20-22/articles/20/68.pdf)

[11] Guyon I., & Elisseeff, A."An introduction to variable and feature selection", Journal of Machine Learning Research, 3, 1157-1182, 2003. (https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf)

[12] Chawla N. V., Bowyer, K. W., Kegelmeyer, W. P., & Smola, A. J.," Smote: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, 16, 321-357, 2002. (https://www.jair.org/index.php/jair/article/view/10302)

[13] Cortes C., & Vapnik, V. N.,"Support-vector networks. Machine learning", 20(3), 273- 297,1995.
(https://link.springer.com/article/10.1007/BF00994018)

[14] James G., Witten D., Hastie T., & Tibshirani, R.,"An introduction to statistical learning: with applications in R", 2021, Springer.

[15] Brownlee, J.," Machine learning mastery with Python", Scarcity and variety edition, Machine Learning Mastery, 2020.

[16] Pedregosa F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Blondel, V., & Courty, P. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2861-2878, 2011. (https://scikit-learn.org/)

[17] Han J., Kamber, M., & Pei, J.,"Data mining: Concepts and techniques" (3rd ed.), Morgan Kaufmann, 2011.

[18] Witten I. H., & Frank, E.,"Data mining: Practical machine learning tools and techniques", (2nd ed.), 2005, Morgan Kaufmann.

[19] Zhang W.,"Learning from imbalanced data: A review of progress and issues in data mining field", International Journal of Data Mining & Knowledge Management Process, 7(6), 126-140, 2016. (https://link.springer.com/chapter/10.1007/0-387-25465-X_40)

[20] Liu H., Zhou J., & Li H., "Improving text classification using active learning with expected error reduction as stopping criterion", Information Sciences, 190, 192-201, 2012. (https://www.sciencedirect.com/science/article/abs/pii/S0306457313000964)

[21] Kim Y., & Provost, F., "Cost-sensitive active learning for classification", In Proceedings of the KDD-2003 workshop on mining for data streams (pp. 1-8), 2003. (http://fansmale.com/downloadRAR/publicationPdf/CATS.pdf)