Mohanraj Jayakumar
10577457

Data Mining - CA1
Mohanraj Jayakumar

Part–A (CRISP-DM Model)

ABSTRACT:

This document was based on three publications on "CRISP-DM Model". We explain about the data mining industry advantages and relationships across the circular economy which involves the process by using this model. This method was progressed in the year 1996 as an industrial tool and application neutral model. This model significantly plays a vital role in numerous industries in the coming generations with the new variants and methods, but the CRISP-DM Methodology is still unexplored in many ways, and this can be enhanced more in the future world. This paper is said about the advantages and benefits of CRISP-DM Model.

Keywords: *Process Methodology, Circular Economy and Business Analytics*.

INTRODUCTION:

CRISP-DM is an industry-independent process model for data mining, and we provided the concluded name of this model is "CRISP-DM" as many researchers mentioned in the literature review. And the important phases involved in CRISP-DM is based on the business structure and economic evaluation. The circular economy which involves CRISP-DM emerged as an umbrella concept in the year 2010 to grow and gain attention before developing any business progress. The arrow in diagram denotes the way to implement the business strategies to achieve the economic goals. This model is developed and refined in the several years for handling the data mining projects. In simple words the CRISP-DM Model is providing blueprints for conducting a data mining projects with the comprehensive understandings. In business terms if analyst could not recognize the model curricula this model helps to translate by redefining the problem at this point for data mining sectors.

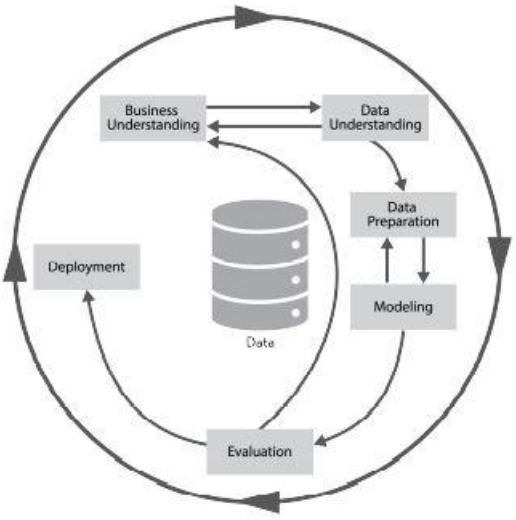

Figure 1: CRISP-DM PHASES

BUSINESS UNDERSTANDING:

Systems Analysis and Design (SA&D) protocols is followed in CRISP-DM business environment to balance the business needs and business objectives. The dimension of the workflow is blueprinted with the help of deploying and implementing the SA&D concepts. Therefore, the entire plan is predicted using this methodology.
The appropriate prediction techniques used for data mining is,

- Regression analysis
- Regression trees
- Neutral nets
- K nearest neighbor
- Box - Jenkins's methods
- Genetic algorithms

DATA UNDERSTANDING:

In this stage this model helps to collect the initial data and provide a detailed report by exploring the data with other data on the internet with the help looping concepts and finally it summarizes the data which is verified.

DATA PREPARATION:

In this stage it involves in selecting data for project after collection, it helps to clean the data which has null values, etc. by this, we get clarity of improvised data. The next stage is normalizing the data for an integrated data format

MODELLING:

In this level we check about a well-suited modelling technique and then we generate the techniques wisely to build the model for broad range of detailed sketches and this helps in creating a new data modelling tool with given the given data to assess the business profile

EVALUATION:

Crisp dm model is used by all the analyst and scientist to provide a better insight before the last stage -by exploring, evaluating, and reviewing to deploy the forecasted plan. Hence it is very convenient to summarize the output of the plan by evaluating the business needs.

DEPLOYMENT:

Crisp dm model is used by all the analyst and scientist to provide a better insight by deploying the business architecture, henceforth it generates the deployment schema, maintenance sketch and give the final review of the project. By this the entire business model is provided by documenting in a significant and peculiar way.

The CRISP-DM Model is basically a level of business approaches given by the help of data which is later implemented and deployed like a blueprint of project plan this is even called as CRISP eSNep phase. In this the below stages are followed to maintain a workflow from the early age of any business.

CRISP-eSNeP Phase for revising the deployment process and stages.

- Data Acquisition
- Data Cleaning
- Data Formatting
- Data Validation
- Data Analysis

DATA MINING PROBLEM TYPES:

Data description and summarization provides an elementary and formatted form in data mining divisions to give an overview of all data structure. In every data mining projects the sub goal of every data is to find the hypothesis of hidden information's

For instance:

"A retailer might be interested in the turnover of all outlets, broken down by categories, summarizing changes and differences as compared to a previous data period". If data problem is stand -alone problem, no further modelling is required to carry out data mining engagements.

SEGMENTATION:

In data mining they are problem called segmenting the data wisely for but unfortunately it does not find a loop to interact with the missing data hence there will be major missing of fact tables and the clustering process is collapsed due to insufficient data.

Appropriate techniques to handle segmentation:

- Clustering
- Neural nets
- Visualization

SYSTEMATIC REVIEW:

It is an industry independent model for using data mining projects and the important result of this model is known as de-facto standard. By using CRISP-DM Model there are many challenges are there said in my studies because there no foresee a deployment phase. Due to the own stage of deployment in the CRISP-DM Model we want to answer only on how this phase conduct through the process model is research methodology.

The systematic literature is structed and assembled as follows according to "kitchenham"

- RQ1: It is known said to be the popular model in data mining
- RQ2: In this stage they identify what is suitable model and how to conduct this model.
- RQ3: The final process is to develop the project by using the given data which are collected for answering the business requirements.

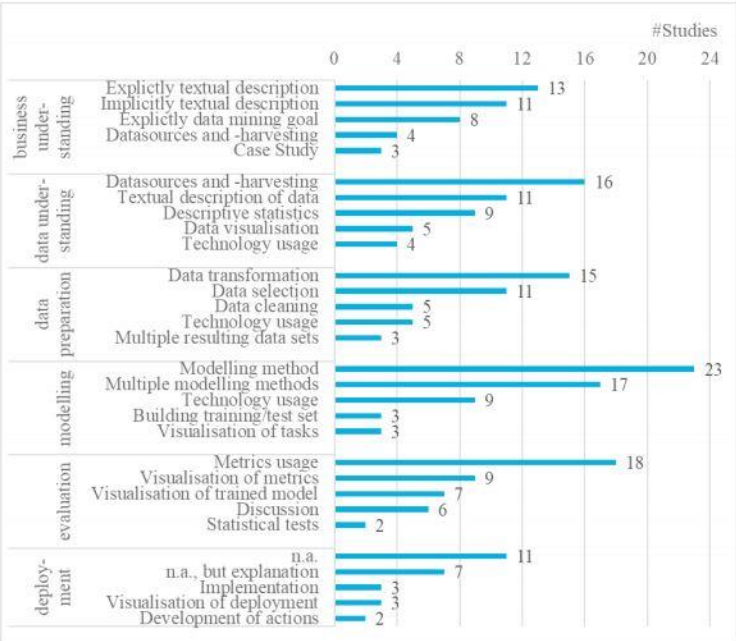

Figure 3: Systematic Review

The above recommendations from the CRISP-DM Model depicts that all the users form business, understand that it is more comfortable to evaluate the business plans or projects. In evaluation phase like data selection, data transformation, and data cleaning which basically depends on data mining firms.

CIRCULAR ECONOMY:

Circular economy was emerged as umbrella concept in the year 2010 to complete the business sustainability's by the action of structural waste. Therefore, it helps to gain and grow attention in early stage of business development, but even though there was not proper literature, however Circular economy was clearly stated by Ellen MacArthur Foundation,

Where CE is defined as a system

"*That provides multiple value creation mechanisms, which are decoupled from the consumption of finite resources*."

The methodology differs from native CRISP-DM in 3 ways:

1. Extending the life cycle
2. Increasing utilization
3. Looping the asset

These 3 phases in CE are looped to any chance to improve the insight and transparency into economic business assets conditions and usage history

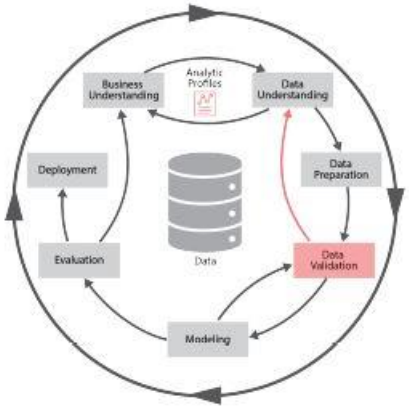

Figure 3: An enhanced CRISP-DM PHASE Model

This model is being embedded based on data understanding and data preparation phases in CRISP-DM version 1.0 for connecting with assets for easier manipulation and automation for the environment or enabling the highly rate in systemic resource efficiency and productivity of the CE.

The focus of CE is PdM which is known as pertinent strategy for OEMs which offers one highest potential for environment and economic impact of any business sector. Thus, following these norms help in identifying the failures and making wise decisions based on analytics which provide better view for management for keeping a complete track of their business and its impactdepending on analysis.

BUSINESS DOMAIN:

The data mining in business domain meets a lot of potential technical barriers when comes to business domain while processing the agile data like text mining, data preprocessing, inductive logic, featureextraction etc. and others like, project management, organization culture, project teams, project selection and initiation, etc. To evaluate scenario of the business understanding stage is enhanced by analyzing and designing by using modern CRISP-DM Technique-SA&D. The CRISP-DM is very sustainable and can be used on top projects having any level of complexity and this method is being tested for the field of robotics, automation and BI domains. The vital role played in Big Data is CRISP-DM methodology since it predicts the future analysis of the economy to make the data more powerful. And this model can even use in small scale industries to predict the business needs and the business requirements.

APPLICATION AND USE OF SA&D:

Many business models can bring forth the insight of business outcome, but the probability of high success rate of business success is given by this model.
The business understanding applications are used by many techniques they are:

- Systems Service Request – SSR
- Baseline Project Plan – BPP
- Statement of Work – SoW

The four major tasks that are undertaken in the CRISP-DM business understanding phases are:

- determining the business objectives
- assessing the situation
- describing the data mining goals
- producing the project plan

The goal is to demonstrate the planning stage of business of data mining process to entangle the problems of business.

CONCLUSION:

To the nutshell of this case the CRISP-DM Model is one of the most important models for data mining projects for all business understandings, without the process of this model a business entity cannot be focused or preplanned for running an organization. And there are many models for looping the business concepts, but CRISP-DM Model holds the highest priority for implementing any business.

Data Mining - CA1

Mohanraj Jayakumar

Part–B (Big Data Mining Issue)

ABSTRACT:

Big data is growing indestructible and problematic since ages when it started booming in the corporate sectors. Even though the most of the cooperate sectors faces many issues and challenges in visualizing the big data, they have more reasons to use the big data mining tools to make it simple and efficient. In simple terms big data is refer to the data which has a huge dataset with the structured and unstructured data. This paper provides the literature overview of the big data mining tool problems, issues and challenges facing in the real time scenarios.

Keywords: *Big data mining problems, Systematic literature, Tools and Techniques*

INTRODUCTION:

In the 21st century the volume of data's collection is in huge entity, and it has been increasing rapidly every upcoming year. Basically, data gathered, analyzed, and visualized for the purpose of future references and to make decisions for the present period. The data are not only collected for the purpose of visualization, but it is also implemented in business organization for changing condition in real time. Data mining is not only the tool to collect data from the resources it also to solve the systematical and technical problem of the dataset which are faced previous and further it gives a feedback or solution to encode the problems which they face every day. Data is the unique and primary functionality which have all detailed structure of any organization to run successfully, the data can be used by any promising tools but in our case, we are going to use mining tools to find the issues faced by them.
The common experience we go through the phases of big data mining are

Volume: The big data tools are so traditional, and it is complicated to sustain the high-level data like Terabytes and Peta bytes, so this makes the process of storing data more complex. Now a days the mining tools a have the highest scope in many sectors since it plays a vital role, But the specification provided by the tools are little slower to manage the huge amount of data or controlling, monitoring, and all the phases which a business organization goes through.

Velocity: The habitual way of collecting information and details from the resource is slow enough since the model is too traditional. Now a days data mining tools are being automated by new technologies to build their system more powerful, but the problem is many reputed companies are following the pattern of CRISP-DM for big data

Variety: The data variants are in many formats but the critical part in big data mining is, it does not have the ability to choose the unstructured data to make it structure in a possible way, there are even audio, video, text, sensor data, graph and many more. Whatever formatted data can be recognized by this tool to understand data but now a days artificial intelligence data being more complicated with 3D, 4D,5D,6D effectual data so its being hard to tackle the crisis.

Variability: In this stage the questions from different variance have been insighted in different angle, but the interpretations are less frequent.

Value: The big data requirements and implantation process is much expensive in IT and in business organization, this causes the small-scale industry to invest much on data mining tools.

THE BIG DATA MINING TOOLS:

The tradition models of mining as been escalated and came up with new technologies to resolve the processing phases of any business organization. The different frameworks on big data has been emerged in the recent years with the great significance and specifications for user firendly usages.

	RDBMS	MPP	Hadoop Framework
Processing	Sequential processing	Some processing parallelism	Massively parallel processing. Grid processing
Scalability	Vertical	Limited horizontal	Massive horizontal
Storage	Relational / SQL database	Propriety data warehouse and data marts	No relational / NoSQL database
Data type	Structure	Structured	All types
Architecture	Shared disk and memory	Shared nothing	Shared nothing
Hardware	Single processor to multi core computing	Data warehouse appliances	Commodity hardware in a distributed grid or clusters
Analysis	Model-based	Model-based	Not model-based

Figure 1

BUSINESS CHALLENGE:

The Big Data consist of fundamental process in multiple ways for descriptive analysis, this helps the data for better insight to follow the organization challenges and drawbacks.

For instance: The below figure 2 depicts that there is a major downfall in SMB organization, moreover it affects the technical firms of the US service business with the percentage of 27% according to recent survey as shown. The major factor for the start-up failure is illustrated in figure 3.

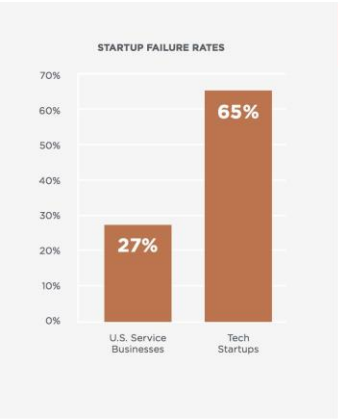

Figure 2

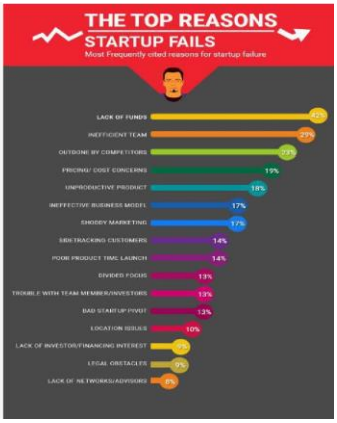

Figure 3

In these cases, the tech companies result in failure due to insufficient fund and lack of time constrains to build their idea or product. This are the major causes for the tackling the business and being a big milestone to be successful in business.

The complications and interruption faced during the business is illustrated below in following steps they are:

1. Heterogeneity and Incompleteness
2. Scale and complexity
3. Timeliness
4. Privacy

SECURITY CHALLENGES:

Big data effectively provides an enormous opportunity to the business organization in the major fields like security, marketing, credit risk, medical research, and urban planning. The focus of this is to establish a strong foundation in data privacy protection.

The data privacy policy involves of encryption and decryption of any data before implementing any dataset in organization, so the company protects their privacy by building a layout of user security authentication.

The key roles of securing the data mining projects are

- Using authentication methods: The verification of any user will be monitoring in this phase
- Use file encryption: The file is encrypted with the high-level security platforms like operating system is managed by server management
- Implementing access controls: The security level is always in the control privileges for user or system
- Use key management: This service is used to distribute keys, certificates and manage different application from different end
- Logging: It is safely secured from hackers with the proper authentication procedures
- Use secure communication: The data are pre recorded and post recorded for the purpose of future visions.

As big data emerges in every domain, but the important factor about the system is about the security and authentication of user and the customer, the highly security portal for big data mining is always being functionalized with the data block.

The following are some of the security threats:

- An unauthorized person can use or hack the system
- An unauthorized person may use any data
- An unauthorized client my use any data
- An unauthorized person can go to any extend by assigning queries automatically

SOLUTION:

The Big Data is necessary in many sectors which help them in handling the huge collection of data in a variant circumstance. The various sectors are Public, Financial service, Healthcare, Manufacturing, Telecommunications, Retails, Other industries. The following sectors where we need the Big Data tools should be effective by implementing the algorithms which is based on the concepts of evolution. The needs of the sectors are making the domains go with the many methodologies with the right workflow. Every traditional tool should be upgraded with the new entities, specifications and feature every year by the firm which owns the mining tools so that it helps many sectors in functioning the data's

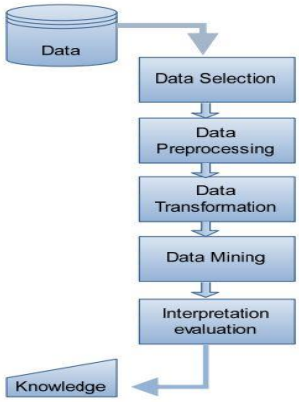

Figure 4: Evolution Process for sorting Big Data

Knowledge discovery is extraction of minimum data recovery from large dataset which the first core step of this model, which help the analyst to declare the final data. In this, it also involves data normalization, data cleansing, data classification, data clustering and data imputation. The main purpose of this mining is data dredging which is known as analyzing data without a prior hypothesis.

TECHNIQUES AND TOOLS FOR BIG DATA MINING:

Big Data has effective measures to organize a company successfully without the escalation and complex challenges. According to data mining tools to handle big data there are multitude open-source tools. The emerging tools and technologies are so handy for the data analyst for descriptive analysis, it helps the analyst as well the clients who is visualizing any kind of data. Therefore, these spectacular tools are automated discovering of problematic data which helps them to recognize the hidden patterns. The tools which are predominantly used for Big Data analysis are

Hadoop:

This tool is an open source which can be used by any firms without any expenses to handle big data for better visualization and helps to store data and functions the data. This tool is run by commodity hardware, with the feature called fault tolerant. This also process with data like structured and unstructured data for handling the velocity, heterogeneity of data, Hadoop is also called as Hadoop Distributed file System by Apache is widely used for storing and managing big data

Rapid Miner:

Rapid Miner is also open source which analyze the data automatically, this tool is useful in research, education, training, rapid prototyping, and application development for better understanding of data in visualization.

CONCLUSION:

Big data is one of the powerful emerging tools in 21st century, even though it must improve in many sections like heterogeneity, speed, accuracy, scalability, trust, provenance, privacy, and instructiveness. If this happens it going to be the best in solving Big Data problems.

References

Binita Kumari, R. R. K., 2015. Visualizing Big Data Mining:Challenges,Problem and Opportunities. p. 5.

Christoph Schröerab, F. K. J. M. G., 2020. A system Litrerature Review on Applying CRISP-DM Proces Model. p. 9.

Eivind Kristoffersen, O. O. B. M. L., 2019. Exploring the Relationship Between Data Science and Circular Economy: an Enhanced CRISP-DM Process Model. p. 14.

James J. Pomykalski, J. B., 2017. Using Systems Analysis and Design to Enhance the Business Understanding Stage in CRISP-DM. p. 7.

Jaseena K.U., J. M. D., 2014. ISSUES, CHALLENGES, AND SOLUTIONS: Big Data Mining. Volume 2, p. 10.

Shearer, C., 2000. The RISP-DM Model The new blueprint for data mining. *Journal of data warehousing,* Volume 5, p. 10.