213CSE2301 PREDICTIVE ANALYTICS

UNIT I INTRODUCTION TO DATA MINING - 9 HRS
Introduction, What is Data Mining?, Concepts of Data mining, Technologies Used, Data Mining Process, KDD Process Model, CRISP – DM, Mining on different kinds of data, Applications of Data Mining, Challenges of Data Mining.

Introduction:-

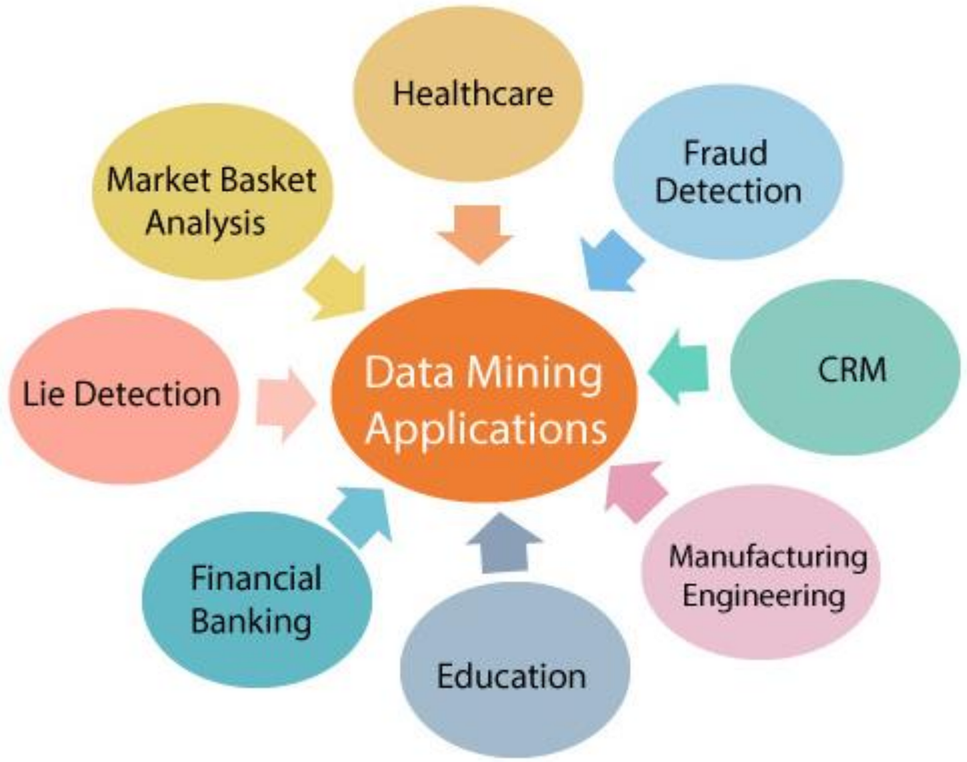
1.1 DATA MINING:

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data.

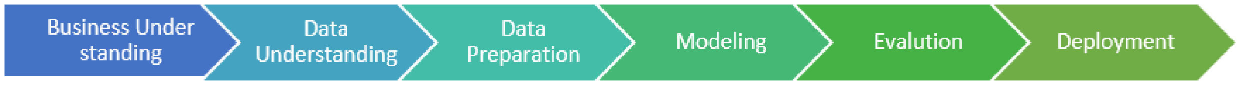
- Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.
- Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures.
- Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD).

Data Mining Applications

- Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits.



1.2 DATA MINING IMPLEMENTATION PROCESS:



Business understanding:

In this phase, business and data-mining goals are established.

- First, you need to understand business and client objectives. You need to define what your client wants (which many times even they do not know themselves)
- Take stock of the current data mining scenario. Factor in resources, assumption, constraints, and other significant factors into your assessment.
- Using business objectives and current scenario, define your data mining goals.
- A good data mining plan is very detailed and should be developed to accomplish both business and data mining goals.

Data understanding:

In this phase, sanity check on data is performed to check whether its appropriate for the data mining goals.

- First, data is collected from multiple data sources available in the organization.
- These data sources may include multiple databases, flat filer or data cubes. There are issues like object matching and schema integration which can arise during Data Integration process. It is a quite complex and tricky process as data from various sources unlikely to match easily. For example, table A contains an entity named cust_no whereas another table B contains an entity named cust-id.
- Therefore, it is quite difficult to ensure that both of these given objects refer to the same value or not. Here, Metadata should be used to reduce errors in the data integration process.
- Next, the step is to search for properties of acquired data. A good way to explore the data is to answer the data mining questions (decided in business phase) using the query, reporting, and visualization tools.
- Based on the results of query, the data quality should be ascertained. Missing data if any should be acquired.

Data preparation:

- In this phase, data is made production ready.
- The data preparation process consumes about 90% of the time of the project.
- The data from different sources should be selected, cleaned, transformed, formatted, anonymized, and constructed (if required).
- Data cleaning is a process to “clean” the data by smoothing noisy data and filling in missing values.
- For example, for a customer demographics profile, age data is missing. The data is incomplete and should be filled. In some cases, there could be data outliers. For instance, age has a value 300. Data could be inconsistent. For instance, name of the customer is different in different tables.
- Data transformation operations change the data to make it useful in data mining. Following transformation can be applied

Data transformation:

Data transformation operations would contribute toward the success of the mining process.

Smoothing: It helps to remove noise from the data.

Aggregation: Summary or aggregation operations are applied to the data. I.e., the weekly sales data is aggregated to calculate the monthly and yearly total.

Generalization: In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.

Normalization: Normalization performed when the attribute data are scaled up o scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.

Attribute construction: these attributes are constructed and included the given set of attributes helpful for data mining.

The result of this process is a final data set that can be used in modeling.

Modeling

In this phase, mathematical models are used to determine data patterns.

- Based on the business objectives, suitable modeling techniques should be selected for the prepared dataset.
- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.
- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

Evaluation:

In this phase, patterns identified are evaluated against the business objectives.

- Results generated by the data mining model should be evaluated against the business objectives.
- Gaining business understanding is an iterative process. In fact, while understanding, new business requirements may be raised because of data mining.
- A go or no-go decision is taken to move the model in the deployment phase.

Deployment:

1. Classification:

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

2. Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. Regression:

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. Outer detection:

This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

6. Sequential Patterns:

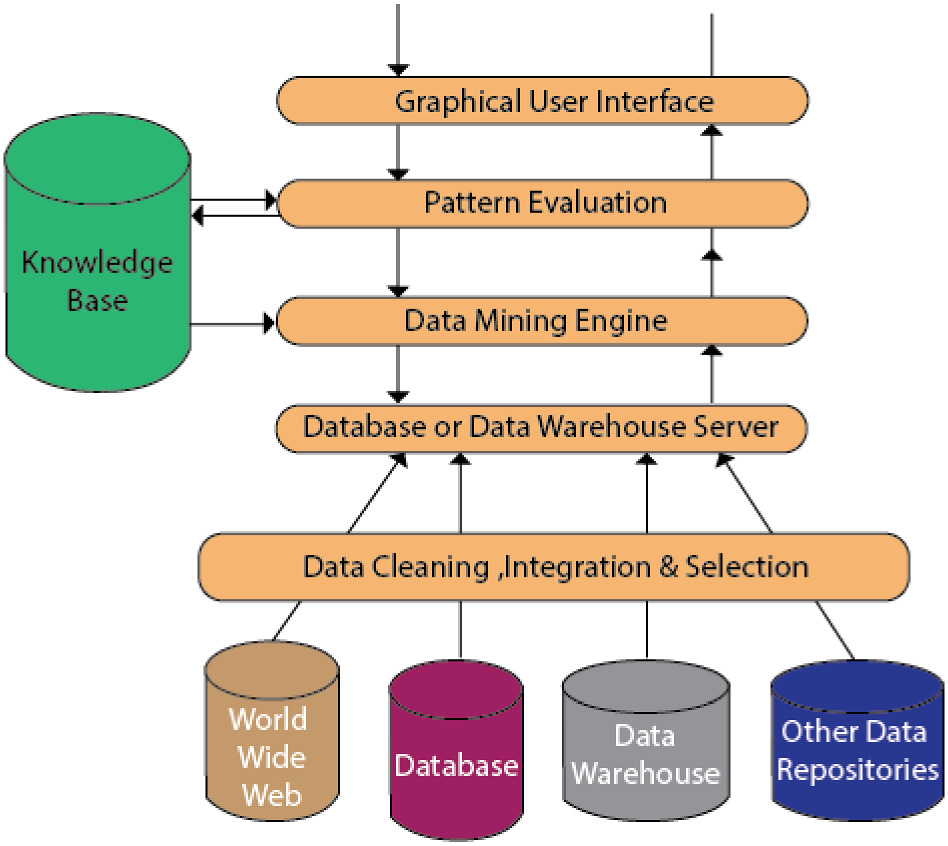
This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

7. Prediction:

Prediction has used a combination of the other techniques of data mining like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

1.3 DATA MINING ARCHITECTURE

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.



Data Source:

- The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses.
- Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Different processes:

- Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate.
- So, the first data requires to be cleaning and unifying. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server.
- 4These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:

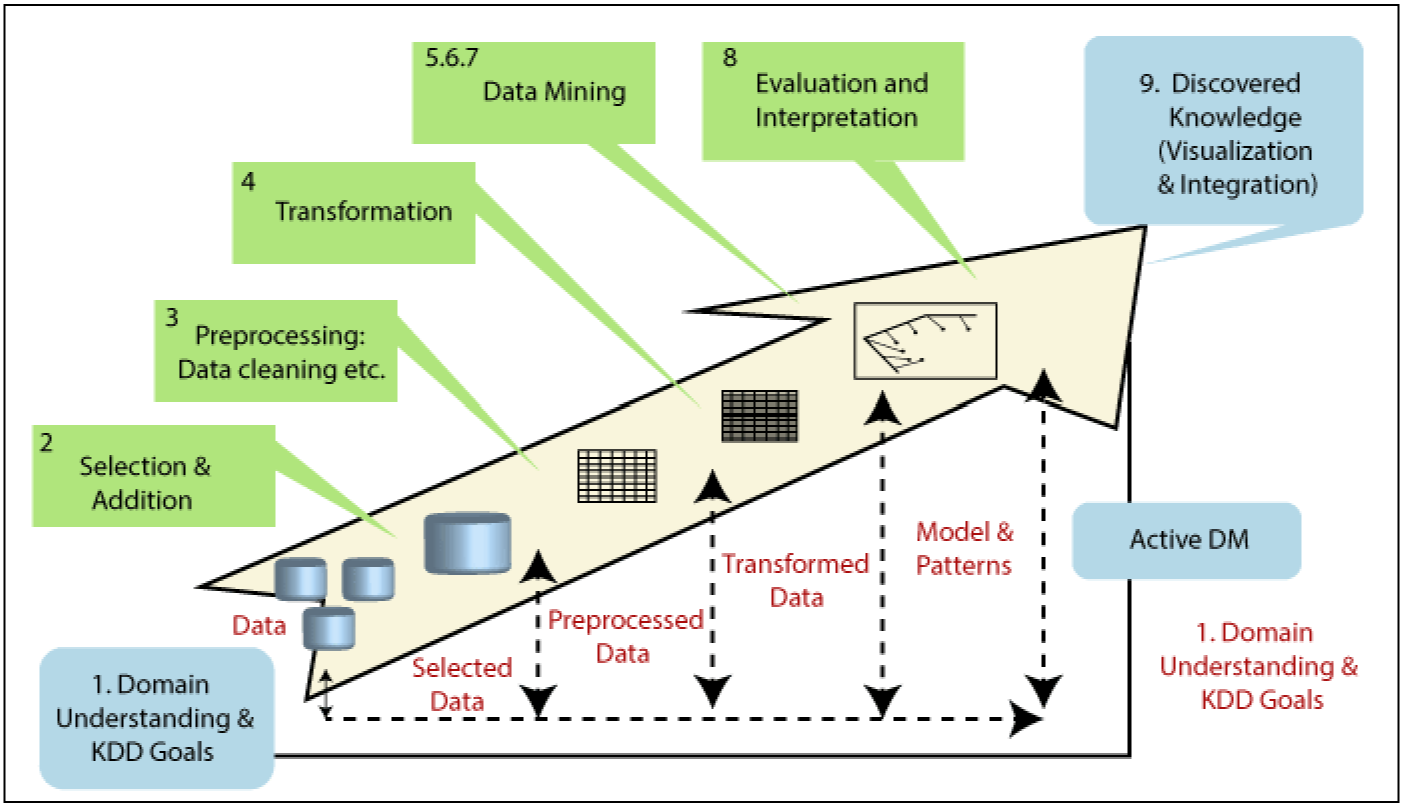
- The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.
- In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module:

- The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

1.4 THE KDD PROCESS

- The knowledge discovery process (illustrates in the given figure) is iterative and interactive, comprises of nine steps. The process is iterative at each stage, implying that moving back to the previous actions might be required.
- The process has many imaginative aspects in the sense that one cant presents one formula or makes a complete scientific categorization for the correct decisions for each step and application type. Thus, it is needed to understand the process and the different requirements and possibilities in each stage.
- The process begins with determining the KDD objectives and ends with the implementation of the discovered knowledge.
- At that point, the loop is closed, and the Active Data Mining starts. Subsequently, changes would need to be made in the application domain. For example, offering various features to cell phone users in order to reduce churn.
- This closes the loop, and the impacts are then measured on the new data repositories and the KDD process again. Following is a concise description of the nine-step KDD process, Beginning with a managerial step:


5.6.7
Data Mining
8
Evaluation and Interpretation
9. Discovered Knowledge (Visualization & Integration)
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Data
Selected Data
Preprocessed Data
Transformed Data
Model & Patterns
Active DM
1. Domain Understanding & KDD Goals
1. Domain Understanding & KDD Goals

1. Building up an understanding of the application domain

This is the initial preliminary step. It develops the scene for understanding what should be done with the various decisions like transformation, algorithms, representation, etc. The individuals who are in charge of a KDD venture need to understand and characterize the objectives of the end-user and the environment in which the knowledge discovery process will occur (involves relevant prior knowledge).

2. Choosing and creating a data set on which discovery will be performed

Once defined the objectives, the data that will be utilized for the knowledge discovery process should be determined. This incorporates discovering what data is accessible, obtaining important data, and afterward integrating all the data for knowledge discovery onto one set involves the qualities that will be considered for the process. This process is important because of Data Mining learns and discovers from the accessible data. This is the evidence base for building the models. If some significant attributes are missing, at that point, then the entire study may be unsuccessful from this respect, the more attributes are considered. On the other hand, to organize, collect, and operate advanced data repositories is expensive, and there is an arrangement with the opportunity for best understanding the phenomena. This arrangement refers to an aspect where the interactive and iterative aspect of the KDD is taking place. This begins with the best available data sets and later expands and observes the impact in terms of knowledge discovery and modeling.

3. Preprocessing and cleansing

In this step, data reliability is improved. It incorporates data clearing, for example, Handling the missing quantities and removal of noise or outliers. It might include complex statistical techniques or use a Data Mining algorithm in this context. For example, when one suspects that a specific attribute of lacking reliability or has many missing data, at this point, this attribute could turn into the objective of the Data Mining supervised algorithm. A prediction model for these attributes will be created, and after that, missing data can be predicted. The expansion to which one pays attention to this level relies upon numerous factors. Regardless, studying the aspects is significant and regularly revealing by itself, to enterprise data frameworks.

4. Data Transformation

In this stage, the creation of appropriate data for Data Mining is prepared and developed. Techniques here incorporate dimension reduction(for example, feature selection and extraction and record sampling), also attribute transformation(for example, discretization of numerical attributes and functional transformation). This step can be essential for the success of the entire KDD project, and it is typically very project-specific. For example, in medical assessments, the quotient of attributes may often be the most significant factor and not each one by itself. In business, we may need to think about impacts beyond our control as well as efforts and transient issues. For example, studying the impact of advertising accumulation. However, if we do not utilize the right transformation at the starting, then we may acquire an amazing effect that insights to us about the transformation required in the next iteration. Thus, the KDD process follows upon itself and prompts an understanding of the transformation required.

5. Prediction and description

We are now prepared to decide on which kind of Data Mining to use, for example, classification, regression, clustering, etc. This mainly relies on the KDD objectives, and also on the previous steps. There are two significant objectives in Data Mining, the first one is a prediction, and the second one is the description. Prediction is usually referred to as supervised Data Mining, while descriptive Data Mining incorporates the unsupervised and visualization aspects of Data Mining. Most Data Mining techniques depend on inductive learning, where a model is built explicitly or implicitly by generalizing from an adequate number of preparing models. The fundamental assumption of the inductive approach is that the prepared model applies to future cases. The technique also takes into account the level of meta-learning for the specific set of accessible data.

6. Selecting the Data Mining algorithm

Having the technique, we now decide on the strategies. This stage incorporates choosing a particular technique to be used for searching patterns that include multiple inducers. For example, considering precision versus understandability, the previous is better with neural networks, while the latter is better with decision trees. For each system of meta-learning, there are several possibilities of how it can be succeeded. Meta-learning focuses on clarifying what causes a Data Mining algorithm to be fruitful or not in a specific issue. Thus, this methodology attempts to understand the situation under which a Data Mining algorithm is most suitable. Each algorithm has parameters and strategies of leaning, such as ten folds cross-validation or another division for training and testing.

7. Utilizing the Data Mining algorithm

At last, the implementation of the Data Mining algorithm is reached. In this stage, we may need to utilize the algorithm several times until a satisfying outcome is obtained. For example, by turning the algorithms control parameters, such as the minimum number of instances in a single leaf of a decision tree.

8. Evaluation

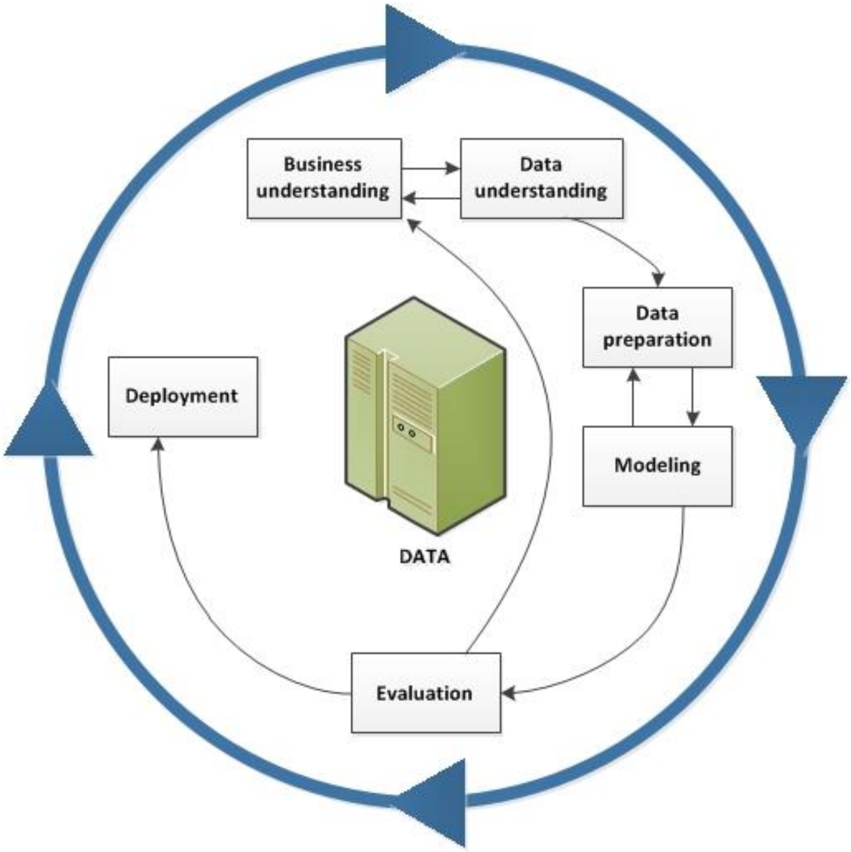
In this step, we assess and interpret the mined patterns, rules, and reliability to the objective characterized in the first step. Here we consider the preprocessing steps as for their impact on the Data Mining algorithm results. For example, including a feature in step 4, and repeat from there. This step focuses on the comprehensibility and utility of the induced model. In this step, the identified knowledge is also recorded for further use. The last step is the use, and overall feedback and discovery results acquire by Data Mining.

9. Using the discovered knowledge

Now, we are prepared to include the knowledge into another system for further activity. The knowledge becomes effective in the sense that we may make changes to the system and measure the impacts. The accomplishment of this step decides the effectiveness of the whole KDD process. There are numerous challenges in this step, such as losing the "laboratory conditions" under which we have worked. For example, the knowledge was discovered from a certain static depiction, it is usually a set of data, but now the data becomes dynamic. Data structures may change certain quantities that become unavailable, and the data domain might be modified, such as an attribute that may have a value that was not expected previously

1.5 CRISP –DM:

CRISP-DM, which stands for **Cross-Industry Standard Process** for Data Mining, is an industry-proven way to guide your data mining efforts.



The life cycle model consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary.

The CRISP-DM model is flexible and can be customized easily. For example, if your organization aims to detect money laundering, it is likely that you will sift through large amounts of data without a specific modeling goal. Instead of modeling, your work will focus on data exploration and visualization to uncover suspicious patterns in financial data. CRISP-DM allows you to create a data mining model that fits your particular needs.

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process. It has six sequential phases:

1. Business understanding – What does the business need?

2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

1. Business Understanding

The *Business Understanding* phase is to understand what the business wants to solve. Important task within this phase according to the Data Science Project Management including:

1. **Determine the business question and objective**: What to solve from the business perspective, what the customer wants, and define the business success criteria (Key Performance Indicator or KPI). For fresher, research what kind of the situation company would face and try to build your project on top of it.

2. **Situation Assessment**: You need to assess the resources availability, project requirements, risks, and cost-benefit from this project. While you might not know the situation within the company if you are not hired yet, you could assess it based on your research and explain what your assessment is based on.

3. **Determine the project goals**: What the technical data mining perspective success criteria. You could set it based on model metrics or availability time or anything as long as you could explain it — what is important is that it logically sounded.

4. **Project plan**: Try to create a detailed plan for each project phase and what kind of tools you would use.

You might assume that when you are hired as a Data Scientist, you will work with the model all the time — alas, it is not true. This business understanding phase is the phase you would meet the most and arguably the most important one; you are hired to solve the business problem, after all.

For you **who want to stand out with the hiring**, make sure you have a complete pack of the business understanding phase. This phase is the one that serves as the project backbone and the one that everyone could understand.

One problem for the Data Science project is about the data — you already know what you want to solve, but where is the data? Let me answer it in the next phase.

2. Data Understanding

The *Data Understanding* phase is where we focus on understanding the data we had to support the Business Understanding and solve the business problem. This phase, according to the Data Science Project Management, could consist of:

1. **Collect Data**: Acquire the data. As I mentioned previously — For people outside the company, the data we want would not be available; in this case, you could either try to collect the data or using the available data from free resources and build your project based on it. You need to work with what you have.
2. **Describe data**: Examine the data format, number of rows and columns, field identities, and available features. Try to describe the data you have at a glance.
3. **Explore data**: Exploring the data. Describe the relationship between data, visualize the data, and be creative. What is important is your data exploration could verify the business question.
4. **Verify data quality**: How is your data quality? Many missing values? Is the data collection appropriate enough?. Make sure that the data is past your quality threshold.

Data Understanding is where you show everything you could understand about the data and relate it with the business question. My advice here is to try to answer the question of business without over-exploration — many people are stuck here because of the freedom. However, you must remember to focus on the business.

3. Data Preparation

After you understand the data you have, it is time for the *Data Preparation*. This phase is what we did to prepare the data for the modeling phase. The phase according to the Data Science Project Management including:

1. **Data Selection**: Selecting the dataset, columns, and/or rows you would use. When you exclude data, make sure you have a valid explanation. The way you filter data should reflect the business question as well.

2. **Data Cleaning**: Garbage-in, garbage-out — what happens if you did not clean the data properly. This is the task when you need to make sure your data is right. Cleaning data takes a lot of preparation and data understanding — you need a reason to correct data or impute new data. Your data science project should explain the steps you take to clean the data in detail. For me personally, I would take a closer look at this task when I am looking at the candidate.

3. **Feature Engineering**: Feature engineering you might think interesting or helpful. Try to be creative when creating new data from existing data.

4. **Data Integration**: New data set from combining two or more data sets. It might be a lot harder to acquire the dataset from free resources, but it is possible.

5. **Data Formatting**: Formatting data when you need it. For example, convert the categorical value into numerical value or vice versa. Don’t forget to state your reasoning.

Data preparation is the key to a great modeling process. If you already messed up in this phase — the next phase would not produce any viable result. That is why this is the phase that you need to focus on the most and describe as much as possible to make your data science project stand out even more.

4. Modeling

For many data enthusiasts out there, the *Modeling* phase might be the most exciting — well, it is. However, this phase is shorter compared to the other phase we just passed previously. In this phase, we would develop our machine learning model/product to answer the business question. The task, according to the Data Science Project Management, could include:

1. **Model Selection**: Using which machine learning algorithms to try. You might want to experiment with many models. It would be desirable if you could explain why you select a certain algorithm.

2. **Test design**: Design your modeling test design by splitting the data into training, test, validation sets, or cross-validation. Justify the reason for how you design the test.

3. **Model development:** Fit your model using the data you have prepared. Manage your resources well here — it might take a long time, depend on your data and your experiment design. The development should consider how the business question and industrial scene

would be as well, such as “Would the model I develop is possible in the business,” “Is the resources I need to develop this model is costly?” etc.

4. **Model Assessment**: Set your success technical metrics and choose the best model(s) viable for solving the business question. Try to explain why you decide on certain metrics and why not the other.

In a real-world working environment, we don't try to achieve perfection. What we want is a “good enough” model — CRISP-DM lifecycle would improve the model in future iterations. If you feel you haven’t achieved that “99% Accuracy” model, it is fine. What is important is you able to explain the process.

5. Evaluation

The *Evaluation* phase is different from the Modeling technical evaluation. This phase **evaluates the model concerning the business indicator and what to do next**. This phase task, according to the Data Science Project Management includes:

1. **Evaluate results**: Would the business success criteria be met using your model? which model(s) would you choose?. This is the phase when you should explain how your model would help the business. Explain it as realistic as possible, and don’t use too much technical jargon that people outside of the data world could understand.

2. **Review process**: Review your work process. Was anything missing? Need more time? Were all phases executed? Try to Summarize your findings and correct anything if required. Your first data science project iteration did not need to be the perfect one. Learning from the mistake is part of the process and what the company wants to see as well.

3. **Determine next steps**: Based on the previous tasks, decide if the model is ready for deployment, needs more iteration, or just creates a new project.

For many data enthusiasts, this is the step that they overlooked the most. This is because the study guideline or data Bootcamp would rarely be taught any of the business evaluations. This is a shame because I would argue that this part would make you different compared to any other candidates.

Imagine the situation when you present your data science project, and you show the model to the business user. The most valid question would be, “Is this model could help my business? If yes, how?” — Try to answer this question when working on your project.

6.Deployment

You might have the best model in the world, but if **your user/customer could not access the model result, it is useless.** This phase might seem too much for a fresher, but if you could build a realistic enough concept, it would score you many points with the company. This phase task, according to the Data Science Project Management, includes:

1. **Deployment**: How would you like to deploy the model? How is the result presented or delivered? Try to plan and document the process.

2. **Monitoring and maintenance**: What is the monitoring and maintenance plan going to be like? Monitoring the result and maintaining the model quality is as important as any other phase. Businesses don’t want the model to become a burden.

3. **Final report**: Conclude the project by creating the summary report, create the presentation, and try to present it to someone you know.

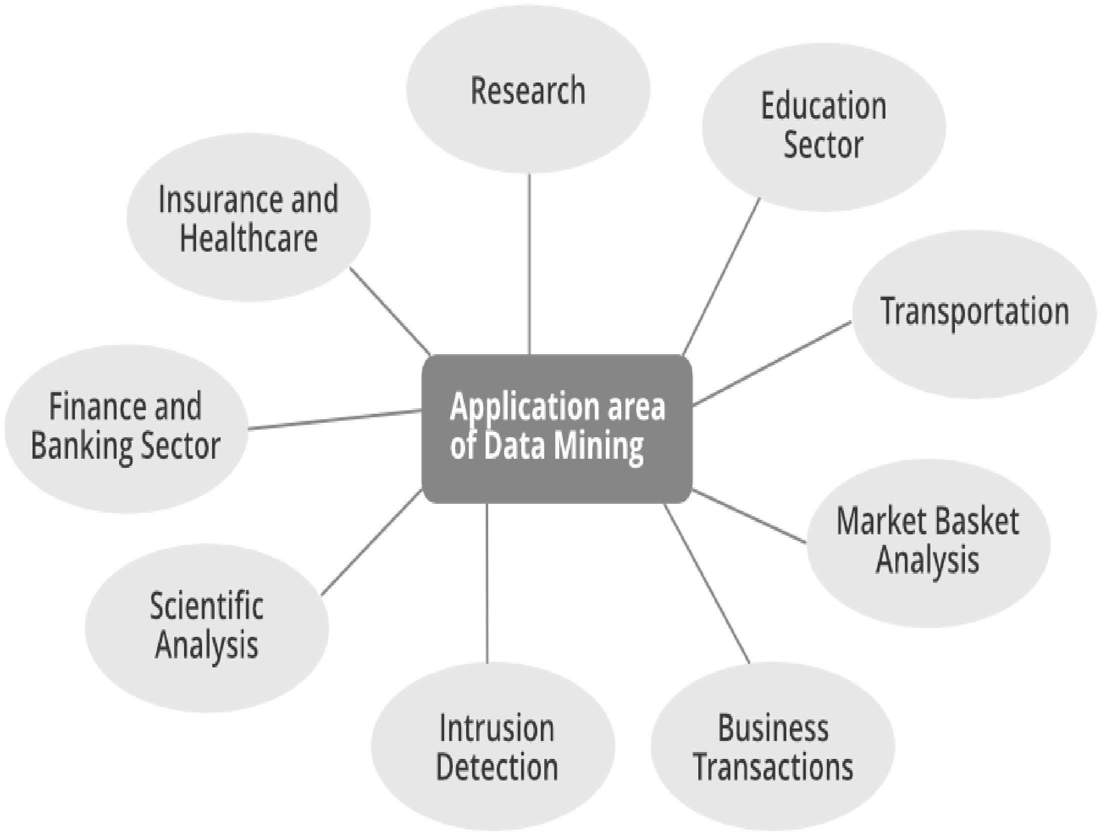
4. **Review**: Review the project by thinking about what is good, what you could improve, and what you think is lacking. Determine if the project needs more iteration or only needs frequent maintenance.

The project might end with the deployment, but it is a continuous cycle. When you develop your data science project, make sure you think for a long-term and not a one-time project (except if that is what you want).

1.6 DATA MINING APPLICATIONS

Data mining provides competitive advantages in the knowledge economy. It does this by providing the maximum knowledge needed to rapidly make valuable business decisions despite the enormous amounts of available data.

There are many measurable benefits that have been achieved in different application areas from data mining. So, let’s discuss different applications of Data Mining:



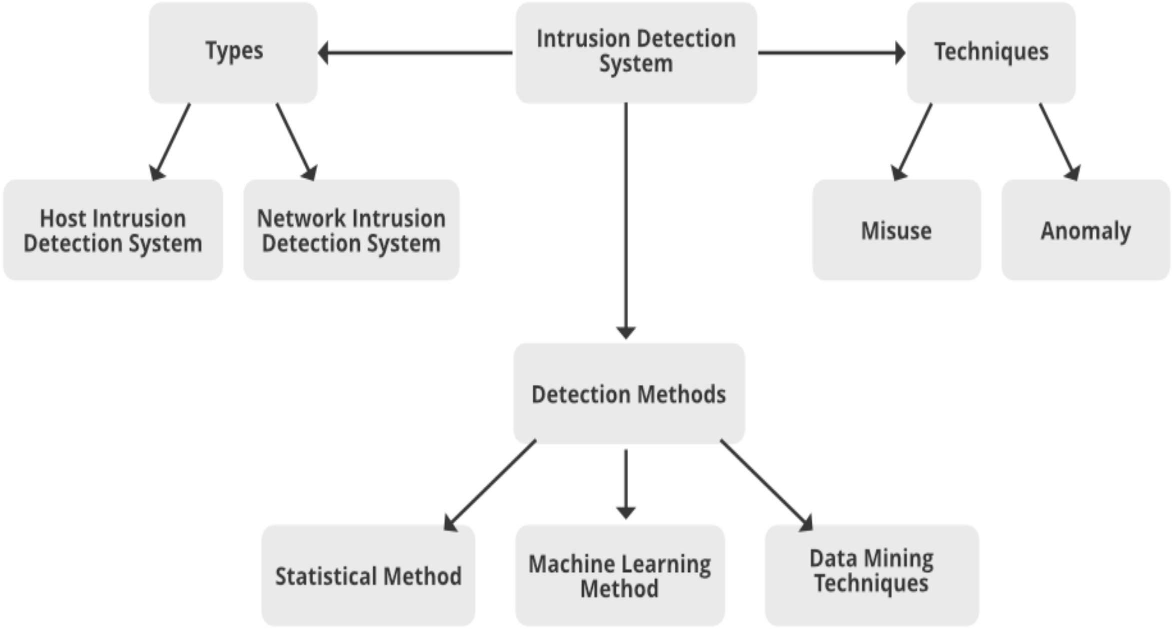
Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyse the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

Intrusion Detection: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources. Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

- Detect security violations
- Misuse Detection

- Anomaly Detection



Business Transactions: Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making. Example :

- Direct mail targeting
- Stock trading
- Customer segmentation
- Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things)and Cybersecurity
- Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

- Claims analysis i.e which medical procedures are claimed together.
- Identify successful medical therapies for different illnesses.
- Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

- Credit card fraud detection.

- Identify ‘Loyal’ customers.
- Extraction of information related to customers.
- Determine credit card spending by customer groups.

Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

1.7 CHALLENGES OF DATA MINING

Challenges of Data Mining Nowadays Data Mining and knowledge discovery are evolving a crucial technology for business and researchers in many domains. Data Mining is developing into established and trusted discipline, many still pending challenges have to be solved. Some of these challenges are given below.

Security and Social Challenges:

Decision-Making strategies are done through data collection-sharing, so it requires considerable security. Private information about individuals and sensitive information are collected for customers profiles, user behaviour pattern understanding. Illegal access to information and the confidential nature of information becoming an important issue.

User Interface: The knowledge discovered is discovered using data mining tools is useful only if it is interesting and above all understandable by the user. From good visualization interpretation of data, mining results can be eased and helps better understand their requirements. To obtain good visualization many research is carried out for big data sets that display and manipulate mined knowledge.

(i) **Mining based on Level of Abstraction**: Data Mining process needs to be collaborative because it allows users to concentrate on pattern finding, presenting and optimizing requests for data mining based on returned results.

(ii) **Integration of Background Knowledge**: Previous information may be used to express discovered patterns to direct the exploration processes and to express discovered patterns.

(iii) **Mining Methodology Challenges**: These challenges are related to data mining approaches and their limitations. Mining approaches that cause the problem are:

(i) Versatility of the mining approaches,

(ii) Diversity of data available,

(iii) Dimensionality of the domain,

(iv) Control and handling of noise in data, etc.

Different approaches may implement differently based upon data consideration. Some algorithms require noise-free data. Most data sets contain exceptions, invalid or incomplete information lead to complication in the analysis process and some cases compromise the precision of the results.

Complex Data: Real-world data is heterogeneous and it could be multimedia data containing images, audio and video, complex data, temporal data, spatial data, time series, natural language text etc. It is difficult to handle these various kinds of data and extract the required information.

New tools and methodologies are developing to extract relevant information.

(i) **Complex data types**: The database can include complex data elements, objects with graphical data, spatial data, and temporal data. Mining all these kinds of data is not practical to be done one device.

(ii) **Mining from Varied Sources**: The data is gathered from different sources on Network. The data source may be of different kinds depending on how they are stored such as structured, semi-structured or unstructured.

Performance: The performance of the data mining system depends on the efficiency of algorithms and techniques are using. The algorithms and techniques designed are not up to the mark lead to affect the performance of the data mining process.

(i) **Efficiency and Scalability of the Algorithms**: The data mining algorithm must be efficient and scalable to extract information from huge amounts of data in the database.

(ii) **Improvement of Mining Algorithms:** Factors such as the enormous size of the database, the entire data flow and the difficulty of data mining approaches inspire the creation of parallel & distributed data mining algorithms.