1. What is Data Mining? What are the different tasks of Data Mining?

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

The different tasks of data mining are as follows:

A close-up of a paper

Description automatically generated

2.Discuss the Life cycle of Data Mining projects?

The life cycle of a Data Mining project involves several phases. These phases are described in detail in the Cross Industry Standard Process for Data Mining (CRISP-DM). The project life cycle of a data mining project according to CRISP-DM includes the following phases:

* + **Business understanding**: To identify the business goals and to determine how to measure success.
  + **Data understanding**: To select relevant data and to understand this data. This means to understand the semantics of tables and columns and to know the data distributions.
  + **Data preparation**: To cleanse the selected data and to transform it, for example, by joining and by aggregation so that it is suitable for data mining analysis.
  + **Modeling**: To run the data mining algorithms.
  + **Evaluation**: To look at mining models, understand influencing factors, and assess model accuracy.
  + **Deployment**: To score, this means to apply the data mining model to new data.

3. Name areas of applications of data mining?

Data Mining can be applied to various areas such as:

* + **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 analyzing these data.
  + **Intrusion Detection**: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources.
  + **Business**: 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.
  + **Telecommunications**
  + **Banking Sector**
  + **E-commerce**
  + **Finance**
  + **Medicine and Healthcare**
  + **Security (Information and Cybersecurity)**

4 What are the issues in data mining?

There are several issues in Data Mining such as:

* + **Mining Methodology and User Interaction Issues**: It refers to issues like mining different kinds of knowledge in databases, interactive mining of knowledge at multiple levels of abstraction, incorporation of background knowledge, data mining query languages and ad hoc data mining, presentation and visualization of data mining results, handling noisy or incomplete data, pattern evaluation.
  + **Performance Issues**: There can be performance-related issues such as efficiency and scalability of data mining algorithms. In order to effectively extract the information from a huge amount of data in databases, a data mining algorithm must be efficient and scalable.
  + **Diverse Data Types Issues**: Handling of relational and complex types of data. The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc.
  + **Data Security & Privacy**: Data privacy and security is another significant challenge in data mining. As more data is collected, stored, and analyzed, the risk of data breaches and cyber-attacks increases.

6 What is required technological drivers in data mining?

The required technological drivers in Data Mining include:

* + **Database size**: Since for maintaining and processing the huge amount of data, therefore powerful systems are required.
  + **Query Complexity**: In order to analyze the complex and large number of queries, therefore powerful system is required.

7 What is data mining? Is it a simple transformation or application of technology developed from databases, statistics, machine learning, and pattern recognition?

data mining is defined as the process of uncovering patterns and knowledge from large amounts of data. It’s not a simple transformation of technology developed from databases, statistics, machine learning, and pattern recognition. Instead, it’s an interdisciplinary subfield of computer science that involves methods at the intersection of artificial intelligence, machine learning, statistics, and database systems1. The overall goal of the data mining process is to extract information from a dataset and transform it into an understandable structure for further use1. It is also known as Knowledge Discovery in Databases (KDD)

8 How is a data warehouse different from a database? How are they similar?

A data warehouse and a database system are both used for storing data, but they serve different purposes and have distinct characteristics. A database system is designed for the creation, maintenance, and use of databases for organizations and end-users. It is well-known for its high scalability in processing very large, relatively structured data sets. On the other hand, a data warehouse integrates data originating from multiple sources and various timeframes. It consolidates data in multidimensional space to form partially materialized data cubes. This not only facilitates Online Analytical Processing (OLAP) in multidimensional databases but also promotes multidimensional data mining.

In terms of similarities, both database systems and data warehouses are essential components of an organization's data management infrastructure. They both handle large amounts of data and aim to provide useful insights from that data. Database systems can make good use of scalable database technologies to achieve high efficiency and scalability on large data sets. Similarly, many data mining tasks that handle large data sets or even real-time, fast streaming data can benefit from these scalable database technologies. Furthermore, recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities, extending the capability of existing database systems to satisfy advanced users’ sophisticated data analysis requirements.

8 Present an example where data mining is crucial to the success of a business.

What data mining functionalities does this business need?

An example where Data Mining is crucial to the success of a business is Amazon. Amazon uses both cross-market analysis (finding associations between product sales) and customer profiling (what types of customers buy what products). Based on the acquired profiles predictions can be made on what kind of marketing strategies would be most effective.

9 Is discrimination a classification task?

Yes, discrimination is a classification task. For the classification task, neural network-based approaches attempt to distinguish between two distributions by determining the joint distribution of input variables for each class.

10. Write the difference between prediction and classification in machine learning

with example.

In machine learning, classification is about predicting a categorial class for an element in a dataset whereas prediction is about predicting a missing/unknown element (continuous value) of a dataset. For example, in a hospital setting, grouping patients based on their medical records or treatment outcome is considered classification whereas using a classification model to predict the treatment outcome for a new patient is considered prediction.

11 Is noise same as outlier?

No, noise is not same as outlier. Noise is considered as a random error or variance in a measured variable whereas an outlier is something that is much different than other values.

12 What do you understand by Data Purging?

Data purging refers to the permanent deletion or removal of data from a system, database or storage medium. It involves elimination of unnecessary, outdated or redundant data to optimize system performance, ensure compliance with data protection regulations, safeguard sensitive information and free up storage space.

13 What are the different problems that "Data Mining" can solve?

* **Mining Frequent Patterns, Associations, and Correlations**
* **Classification and Prediction**
* **Cluster Analysis**
* **Outlier Detection**
* **Evolution Analysis**