Welcome students to Unit 4 ”Basics of data mining”, the module name is “Technologies used in Data Mining : Statistics, Database and Datawarehouse systems, Information Retrieval, Machine Learning, Pattern Recognition. Data Mining System Architecture”. This is module number 14, I am Edwin D'Souza, Associate Professor, St. Xavier’s College.

In this module we are covering technologies used in data mining, that is, statistics, database systems and data warehouses, information retrieval, machine learning, pattern recognition and data mining system architecture.

After studying the contents of this module, the student will be able to explain the various technologies used in data mining. Explain the data mining system architecture.

The technologies used in data mining. The first one is **statistics**. Statistics is a science of learning from data and includes, Collecting of data, Drawing conclusions from numerical facts, Presentation of results. Data mining is an area of study that allows us to extract useful information. And predict from vast datasets with the help of statistics and its mathematical methods. Datamining has gradually become important and useful. Some of the main statistical metrics used to perform data analysis are mean, median, variance, and standard deviation.

The second technology is **Database Systems and Data Warehouses**. A database system application focuses on creation, maintenance and use of databases for organisations and its end users. Database systems are often well known for their high scalability in processing very large, relatively structured datasets. Many data mining tasks need to handle large datasets or even real time fast streaming data. Therefore, data mining can make good use of scalable database technologies to achieve high efficiency and scalability on large datasets. Recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities. A data warehouse integrates data originating from multiple sources and various timeframes. It consolidates data in multi-dimensional space to form partially materialized data cubes. The data cube model not only facilitates overlap in multi-dimensional databases. But also promotes multi-dimensional data mining.

Third Technology is **Information Retrieval**. Information retrieval IR is the science of searching for data or information in documents. Documents can be text or multimedia and web site on the web. The differences between traditional information retrieval and database systems are two-fold. Information retrieval assumes that the data under search is unstructured. The queries are formed mainly by keywords. Which do not have complex structure, unlike SQL queries in database systems. The typical approaches in information retrieval adopt probabilistic models. For example, a text document can be regarded as a bag of words, that is multi set of words appearing in the document. The similarity between any two documents can be measured by the similarity between the corresponding language models.

The Fourth technology is **Machine Learning**. Machine learning investigates how computers can learn. or improve the performance based on data. A main research area for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data. For example. A typical machine learning problem is to program a computer so that it can automatically recognize handwritten Postal codes on Mail after learning from a set of examples. Supervised learning is basically a synonym for classification. The supervision in the learning comes from the labelled examples in the training data set. For example, In the Postal code recognition problem. A set of handwritten Postal code images and the corresponding machine-readable translations are used as the training examples with supervised the learning of the classification model. Unsupervised learning is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labelled. Typically, we may use clustering to discover classes within the data.

Pattern recognition is the 6th technology. Pattern is a significant feature of a label in the data set or data sources under consideration. Data mining system has the potential to extract thousands or even millions of patterns or rules by applying different knowledge extraction algorithms. Example, the apriori algorithm. Pattern recognition is the ability to detect arrangements of characteristics or data that yield information about a given system or data set. In a technological context, a pattern might be particular configuration of features in images that identify objects, face recognition, handwriting recognition, iris recognition. Frequent combinations of words and phrases for natural language processing NLP or particular cluster of behaviour on the network that could indicate an attack. Recurring sequences of data. Overtime that can be used to predict trends. Frequent patterns underneath, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including the frequent item-sets. A frequent item-set typically refers to a set of items that often appear together in a transaction data set. For example, milk and bread are frequently bought together in grocery stores by many customers.

The following modules are part of the data mining system. Now this is the data mining system architecture. The first module is knowledge base. This module creates the knowledge base for data mining from multiple heterogeneous sources. The second module is data mining engine. The data mining engine has functional data mining modules for the implementation of data mining techniques. Namely, classification and classification analysis. Namely, classification analysis, Cluster Analysis Associate Rule Mining, Prediction analysis and outlier analysis. The third module is pattern evaluation module. This module evaluates the patterns thrown by the data mining engine module. The pattern evaluation module is integrated with the mining module. Example, the associative rules. Patterns thrown by apriori are filtered based on the thresholds for metrics, Confidence, lift, and support. The 4th module is a user interface. This module communicates between users and the data mining system, allowing the users to interact with the system by specifying a data mining query or task. Providing information to help focus the search. Performing exploratory data mining based on the intermediate data mining results. In addition this module allows the user to browse database and data warehouse schemas or data structures. Evaluate patterns and visualize the patterns in different forms.

The data mining system architecture is shown pictorially here. We have the user interface pattern evaluation data mining engine. Database or data warehouse server, that is the knowledge base. Their connections are shown. These are my references. Thank you.