Welcome to Unit 4 “Basics of Data Mining”. The module name is “Data Mining Techniques”. This is module number 15. I am Edwin D'Souza, Associate Professor, St. Xavier’s College Mapusa, Goa.

In this module, we are covering data mining techniques: Classification analysis, Clustering Analysis, Association Learning, Prediction, or Regression analysis. A comparison of some of these techniques will also be done. After studying the contents of this module, the student will be able to describe the data mining techniques. Identify the data mining technique for a given problem. Compare the available data mining techniques.

The list of data mining techniques: Classification analysis, Clustering Analysis. Association rule learning, Anomaly or outlier detection, Prediction analysis, or regression analysis.

Given a data mining problem, we need to choose the right technique to solve the problem.

The first technique, Classification analysis.
Classification is a classic data mining technique that is used to classify an item in a set of data into one predefined set of classes or groups.

Example,
Classifying an email as legitimate or spam. Classifying a bank loan application as either safe application or risky application. The categories for classifying emails and the categories to classify a bank loan application are already known before the data is analysed.

Next, we are looking at the technique Clustering Analysis. Cluster analysis is a class of techniques that are used to categorize objects or cases into relative groups, called clusters. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects. Grouping similar objects together gives us insight into underlying patterns of different groups. Cluster analysis involves formulating a problem. Selecting and applying a clustering procedure. Interpreting the profile clusters. Finally, assessing the validity of clustering. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Clustering is the technique of grouping a set of objects in such a way that objects in the same group are most similar to each other, then to those in other groups. Clustering is an unsupervised learning technique, which means that it finds natural grouping of objects. From the given unlabelled data.

Example, in the field of psychiatry. The characterization of patients on the basis of clusters of symptoms can be useful in the identification of an appropriate form of therapy. In online shop based on purchasing patterns of customers, clusters of customers are created. Customers with similar purchases are put in a cluster. The purchases of customers in a cluster are different from purchases of customers in other clusters. Customers within a cluster can be targeted with a customized advertisement campaign. A good clustering method will produce high quality clusters in which the intra cluster similarity is high. And the interclass similarity is low.

The third data mining technique is Association Rule Learning. This technique helps us to identify some interesting relations that is the dependency modelling between different variables in large databases. The technique can help to unpack some hidden patterns in the data that can be used to identify variables within the data. And the concurrence of different variables that apply very frequently in the data set. Examples. Association rules are useful for examining and forecasting customer behaviour. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering catalog design and store layouts.

The fourth data mining technique is Anomaly or Outlier Detection. This refers to the observation for data items in the data set that do not match an expected pattern or an expected behaviour. Anomalies are also known as outliers, novelties, noise or exceptions. An anomaly is an item that deviates considerably from the common average within a data set or a combination of data. These types of items are statistically aloof as compared to the rest of the data and hence it indicates that something out of the ordinary has happened and requires additional attention.
Example, anomaly or outlier detection technique can be used in a variety of domains such as intrusion detection, system health monitoring fraud detection, detecting ecosystem disturbances.

The fifth data mining technique is Prediction analysis or Regression analysis. Prediction or regression is a data mining technique used to predict a numerical value. Given a particular data set, the regression functions are used to determine the relationship between the dependent variable that is the target Field and one or more independent variables. The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on.

Example, Regression might be used to predict the cost of a product or service given other variables. Regression would also be used to predict a home's value based on its location square feet price, when it was last sold, the price of similar homes and other factors.

Comparison of classification, analysis and clustering analysis. Clustering technique is used when we are not aware of the categories and what the problem is?. Problem in hand is to discover the categories. Classification technique is used when the categories are clearly known and the problem is to classify a new data item into one of the known categories.

Comparison of Classification analysis and clustering.
The data is unlabelled in clustering technique. In classification technique, the data is labelled. Clustering technique is unsupervised learning. Clustering is un-guided learning as a problem is from unknown area. The categories of data are not known and the problem is to discover the data categories. Classification technique is supervised learning, guided learning as the problem is from known area. The categories of data are already known and the problem is to classify any new data into in to one of the known categories.

Comparison of classification analysis and prediction analysis

In classification technique. A classification model predicts categorical class labels. Whereas in prediction regression technique, prediction model predicts continuous valued functions. For example. We can build a classification model to categorise bank loan application as either safe or risky whereas prediction model to predict the expenditure in dollars of potential customers for computer equipment given the income and occupation. My references are here. Thank you.