Data Mining

☙ *What is Data Mining?*

Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue. In a nutshell, *data mining* is the process of discovering useful patterns and trends in large data sets. During such analyses we often come across the term *predictive analysis.* Predictive analysis is nothing but the process of extracting information from large data sets in order to make predictions and estimates about future outcomes. So, we clearly see the association between the two approaches.

☙ *What Tasks can Data Mining Accomplish?*

The common goals data miners try to achieve are,

- ***Description:*** Which is common in any sort of data analysis. Researchers and analysts are always keen to find ways to *describe* patterns and trends lying within the data.
- ***Estimation:*** In estimation, we approximate the value of a numeric target variable using a set of numeric and/or categorical predictor variables. For example, estimating the grade point average (GPA) of a graduate student, based on that student's undergraduate GPA.
- ***Prediction:*** Prediction is the task of forecasting values of any variable based on observation from past or present of explanatory variables. For example, predicting the price of a stock 3 months into the future.
- ***Classification:*** Classification is similar to estimation, except that the target variable in this case is categorical in nature and we are to classify future observations based on observed ones. For example, diagnosing whether a particular disease is present.
- ***Clustering:*** Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A *cluster* is a collection of records that are similar to one another, and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering.
- ***Association:*** The association task for data mining is the task to uncover rules for quantifying the relationship between two or more attributes. For example, predicting degradation in telecommunications networks.

☙ *Stages of a Data Mining procedure:*

According to The Cross-Industry Standard Process for Data Mining (CRISP-DM), a given data mining project has a life cycle consisting of six phases,

- ***Business understanding:*** This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.
- ***Data understanding:*** The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
- ***Data preparation:*** The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data.
- ***Modeling:*** In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.
- ***Evaluation:*** At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives.
- ***Deployment:*** Model creation does not signify the completion of the project. Need to make use of created models. Example of a simple deployment: Generate a report.

Applications of Data Mining:

Some of the major areas in which data mining plays an important role are,

- Healthcare
- Market Basket Analysis
- Education
- Manufacturing Engineering
- Customer Relationship Management
- Fraud Detection
- Intrusion Detection
- Customer Segmentation
- Financial Banking
- Corporate Surveillance