DATA MINING:

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.[1] Data mining is a subfield of computer science with an overall goal to extract information from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.[1] The difference between data analysis and data mining is that data analysis is to summarize the history such as analysing the effectiveness of a marketing campaign, in contrast, data mining focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data

The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases," the term "data mining" wasn’t coined until the 1990s. But its foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions). What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power.

Over the last decade, advances in processing power and speed have enabled us to move beyond manual, tedious and time-consuming practices to quick, easy and automated data analysis. The more complex the data sets collected, the more potential there is to uncover relevant insights. Retailers, banks, manufacturers, telecommunications providers and insurers, among others, are using data mining to discover relationships among everything from pricing, promotions and demographics to how the economy, risk, competition and social media are affecting their business models, revenues, operations and customer relationships.

Why is data mining important?

So why is data mining important? You’ve seen the staggering numbers – the volume of data produced is doubling every two years. Unstructured data alone makes up 90 percent of the digital universe. But more information does not necessarily mean more knowledge.

Data mining allows you to:

Sift through all the chaotic and repetitive noise in your data.

Understand what is relevant and then make good use of that information to assess likely outcomes.

Accelerate the pace of making informed decisions.

Example Applied to Insurance Industry: Data mining tools could be utilize to predict probabilities of customer claims being incurred and premiums to be set accordingly to hedge against it.

How It Works

Data mining, as a composite discipline, represents a variety of methods or techniques used in different analytic capabilities that address a gamut of organizational needs, ask different types of questions and use varying levels of human input or rules to arrive at a decision.

Descriptive Modeling: It uncovers shared similarities or groupings in historical data to determine reasons behind success or failure, such as categorizing customers by product preferences or sentiment. Sample techniques include:

Clustering - Grouping similar records together.

Anomaly detection - Identifying multidimensional outliers.

Association rule learning - Detecting relationships between records.

Principal component analysis Detecting relationships between variables.

Affinity grouping Grouping people with common interests or similar goals (e.g., people who buy X often buy Y and possibly Z).

Predictive Modeling: This modeling goes deeper to classify events in the future or estimate unknown outcomes – for example, using credit scoring to determine an individual's likelihood of repaying a loan. Predictive modeling also helps uncover insights for things like customer churn, campaign response or credit defaults. Sample techniques include:

Regression A measure of the strength of the relationship between one dependent variable and a series of independent variables.

Neural networks Computer programs that detect patterns, make predictions and learn.

Decision trees Tree-shaped diagrams in which each branch represents a probable occurrence.

Support vector machines Supervised learning models with associated learning algorithms.

Prescriptive Modeling: With the growth in unstructured data from the web, comment fields, books, email, PDFs, audio and other text sources, the adoption of text mining as a related discipline to data mining has also grown significantly. You need the ability to successfully parse, filter and transform unstructured data in order to include it in predictive models for improved prediction accuracy.

For example, determining the best marketing offer to send to each customer. Sample techniques include:

Predictive analytics plus rules

Developing if/then rules from patterns and predicting outcomes.

Marketing optimization

Data mining

Data mining involves six common classes of tasks:

Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.

Summarization – providing a more compact representation of the data set, including visualization and report generation

**Data mining applications for Finance**

A **huge amount of data is generated in online transactions**, so the ability to **identify the right information** at the right time can mean the difference between gaining or losing millions of dollars:

* **Increase customer loyalty** by collecting and analyzing customer behavior data
* Help banks predict customer behavior and launch **relevant services and products**
* **Discover hidden correlations between various financial indicators** to detect suspicious activities with a high potential risk
* **Improve due diligence** to speed alerts and support real-time decision-making
* **Identify fraudulent or non-fraudulent actions** by collecting historical data and turning it into valid and useful information.

**Data mining applications for Healthcare**

The pharmaceutical industry produces a **large amount of documents that are often underutilized**. Data mining can improve health systems and reduce costs:

* **Provide** government, regulatory and competitor **information that can fuel competitive advantage**.
* **Support to the R&D process** and the go-to-market strategy with rapid access to information at every phase of the development process.
* **Discover the relationships between diseases and the effectiveness of treatments** to identify new drugs, or to ensure that patients receive appropriate, timely care.
* **Support healthcare** insurers in detecting fraud and abuse.

**Data mining applications for Intelligence**

**Data mining helps analyze data** and clearly identifies how to connect the dots among different data elements. This is an essential aspect for government agencies:

* **Reveal hidden data** related to money laundering, narcotics trafficking, corporate fraud, terrorism, etc.
* **Improve intrusion detection with a high focus** on anomaly detection and identify suspicious activity from a day one.
* **Convert text based crime reports** into word processing files that can be used to support the crime-matching process.

**Data mining applications for Telecommunication**

The large volumes of call, customer and network data generated and stored by telecommunications companies require **data mining to extract hidden knowledge and identify useful data** to better understand customers and detect fraud:

* **Gain a competitive advantage** and reduce customer churn by understanding demographic characteristics and predicting customer behavior.
* **Increase customer loyalty** and improve profitability by providing customized services.
* **Support customer segmentation strategy** by developing appropriate marketing campaigns and pricing strategies.

**Data mining applications for Energy**

In the Oil & Gas industry,**the large amount of unstructured information** integrated with traditional structured data offers a clear and full picture of the process. Data mining offers solid support for the upstream oil and gas industry:

* **Capture weak signals** of potentially threatening events and identify previously unidentified patterns, connections and relations to empower bidding, hedging and trading strategies.
* **Structure identification of important information**, and quickly distill it to boost technical problem-solving, empower more informed decision-making and enable immediate notification of prospective technical breakthroughs.
* **Improve core processes** in upstream, midstream and downstream with analysis and intelligence capabilities using a variety of sources.
* Extract—in real time—the **relevant knowledge**from an unlimited amount of information streams relative to your core business domains.