# DR B.R. AMBEDKAR NATIONAL INSTITUTE OF TECHNOLOGY JALANDHAR


Assignment-1

Data Mining and Data Warehousing

**SUBMITTED TO-**

Dr. Geeta Sikka

CSE Department

**SUBMITTED BY-**

Ankit Goyal (17103011)

Group - G-1

Branch - CSE

# Data Mining in Bio-Informatics


## Abstract:

Over recent years the studies in proteomic, genomics and various other biological researches has generated an increasingly large amount of biological data. Drawing conclusions from this data requires sophisticated computational analysis in order to interpret the data. One of the most active areas of inferring structure and principles of biological datasets is the use of data mining to solve biological problems. Some typical examples of biological analysis performed by data mining involve protein structure prediction, gene classification, analysis of mutations in cancer and gene expressions. As biological data and research become ever more vast, it is important that the application of data mining progresses in order to continue the development of an active area of research within bioinformatics. This essay aims to draw information from varied academic sources in order to discuss an overview of data mining, bioinformatics, the application of data mining in bioinformatics and a conclusive summary.


## Introduction:

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data Mining (DM) is the science of finding new interesting patterns and relationship in huge amount of data. It is defined as "the process of discovering meaningful new correlations, patterns, and trends by digging into large amounts of data stored in warehouses". Data mining is also sometimes called Knowledge Discovery in Databases (KDD). Data mining is not specific to any industry. It requires intelligent technologies and the willingness to explore the possibility of hidden knowledge that resides in the data. Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich, but

lacks a comprehensive theory of life's organization at the molecular level. The extensive databases of biological information create both challenges and opportunities for development of novel KDD methods. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience.

## Data mining tasks

The two "high-level" primary goals of data mining, in practice, are prediction and description. The main tasks well suited for data mining, all of which involves mining meaningful new patterns from the data, are:

*Classification:* Classification is learning a function that maps (classifies) a data item into one of several predefined classes.
*Estimation:* Given some input data, coming up with a value for some unknown continuous variable.
*Prediction:* Same as classification & estimation except that the records are classified according to some future behaviour or estimated future value).
*Association rules*: Determining which things go together, also called dependency modeling.
*Clustering:* Segmenting a population into a number of subgroups or clusters.
*Description & visualization:* Representing the data using visualization techniques.

Learning from data falls into two categories: directed ("supervised") and undirected ("unsupervised") learning. The first three tasks – classification, estimation and prediction – are examples of supervised learning. The next three tasks – association rules, clustering and description & visualization – are examples of unsupervised learning. In unsupervised learning, no variable is singled out as the target; the goal is to establish some relationship among all

the variables. Unsupervised learning attempts to find patterns without the use of a particular target field. The development of new data mining and knowledge discovery tools is a subject of active research. One motivation behind the development of these tools is their potential application in modern biology.

The discovery of intelligence or information gained from data mining has a number of objectives, including the popularity of forecasting, validation, diagnosis and imitation. Typically, the process of obtaining information includes data retention and processing, the use of algorithms, the perception / interpretation of results.

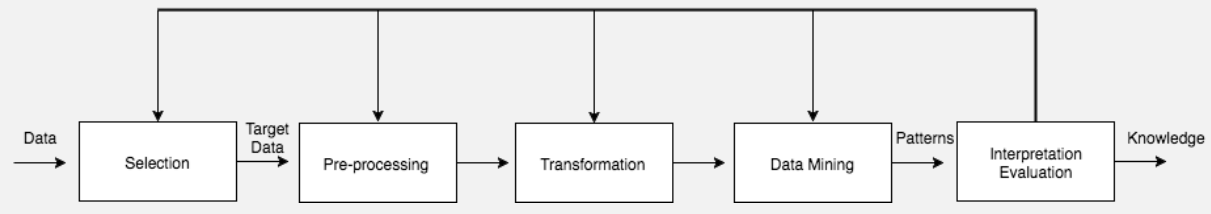


Figure: Process of Knowledge Discovery through Data Mining

It is important to note that the data mining process or KDD involves many techniques, such as machine learning. As a result, the data mining process involves many of the necessary steps to be repeated and sorted to provide accuracy and solutions in data analysis, which means that there is currently no standard framework for data mining.

However, CRISP-DM (Cross Industry Standard Process for Data Mining), defines one standard framework for the process of data mining across multiple industries containing phases, generic tasks, specialised tasks, and process instances Typically speaking, this process and the definition of Data Mining defines the extraction of knowledge. Where we define

machine learning within data mining is the automatic data mining methods used, Kononenko and Kukar state that

*"Machine Learning cannot be seen as a true subset of data mining, as it also compasses the other fields, not utilised for data mining"*

Following this, knowledge is gained through the use of differing machine learning methods used include: classification, regression, clustering, learning of associations, logical relations and equations.

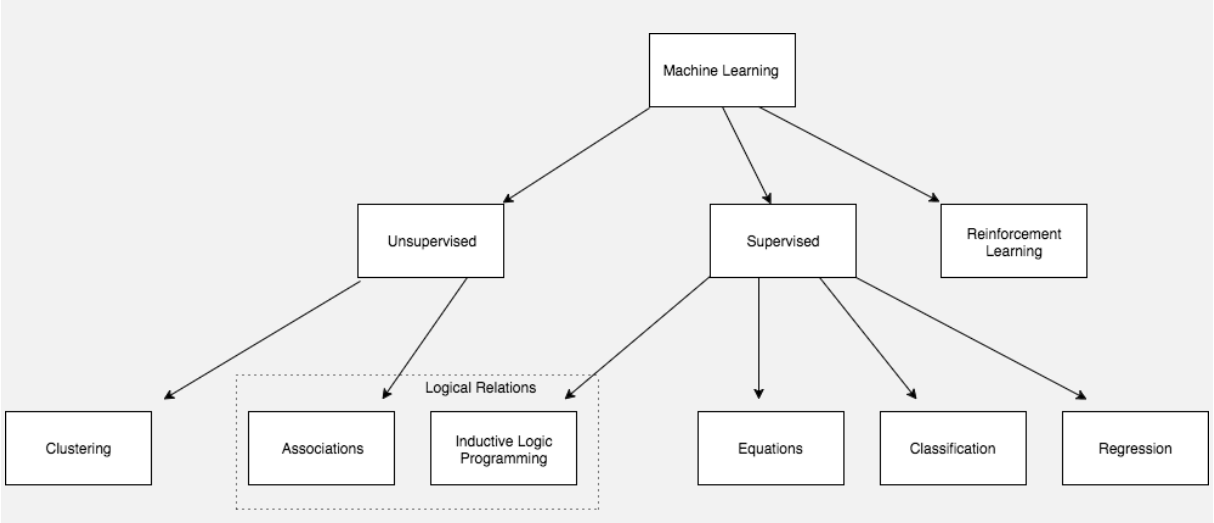


Figure : Methods of Machine Learning

As seen in Figure, Machine learning can be catergorised into unsupervised or supervised learning models. Unsupervised learning models involve data mining algorithms identifying patterns and structures within the variables of a data set, i.e clustering. Supervised learning defines where the variable is specified or provided in order for the algorithms to predict based off of these, i.e regression.

## Application of Data Mining in Bioinformatics:

Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

## References:

1. An introduction into Data Mining in Bioinformatics

   *An introduction into Data Mining in Bioinformatics. | by Littl3field | Littl3field (littlefield.co)*

2. Application of Data Mining in Bioinformatics

   *(PDF) Application Of Data Mining In Bioinformatics (researchgate.net)*