Chapter 2

Literature review

2.1 Data Mining

Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "Knowledge mining," a shorter term may not reflect the emphasis on mining from large amounts of data. [1]

2. 1.1 Data Mining Model

Data mining model consist of following:

i. **Predictive Model (Supervised):** Using the known value to predictive unknown data value such as:

- Classification
- Regression
- Prediction
- Time Series Analysis

ii. **Descriptive Model (Unsupervised):** Identifies the patterns in data and explores the properties of the data such as:

- Clustering
- Association Rule
- Sequence discovery
- Summarization

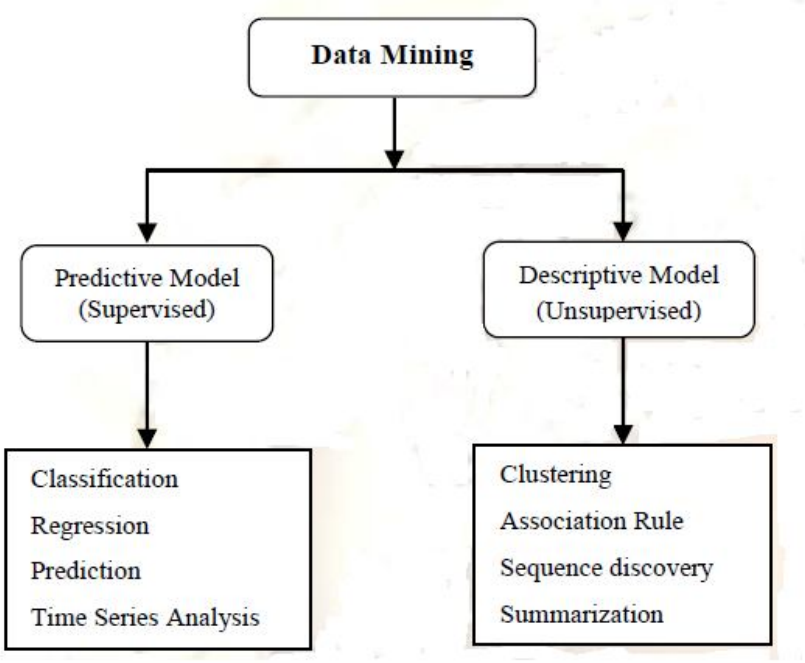

Figure 2.1: Data Mining Model [12]

2.1.2 Data Mining Tasks

There are different types of data mining tasks depending on result; these tasks are classified as follows:

i. Exploratory data Analysis:

It is simply and interactive techniques, exploring the data without any clear ideas of what we are looking for

ii. Descriptive data mining (unsupervised):

It describes the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

iii. Predictive data mining (Supervised):

It permits the value of one variable to be predicted from the known values of other variables.

2.1.3 The Component of Data Mining

i. Knowledge Discovery Process

Knowledge Discovery in Databases is the process comprises of few steps leading from raw data collections to some form of new knowledge. While data mining and KDD are often treated as equivalent terms but in the definition we consider here, data mining is an important step in the KDD process[12].

Data mining is the main step in the KDD process and involves automatically searching large volumes of data for patterns using algorithms such as classification, clustering, etc, data mining as the main analysis step in the KDD process needs high quality data. If the quality is not good, the results of the knowledge extraction will not be good [13], the following fig.2 show data mining steps in Knowledge Discovery process.

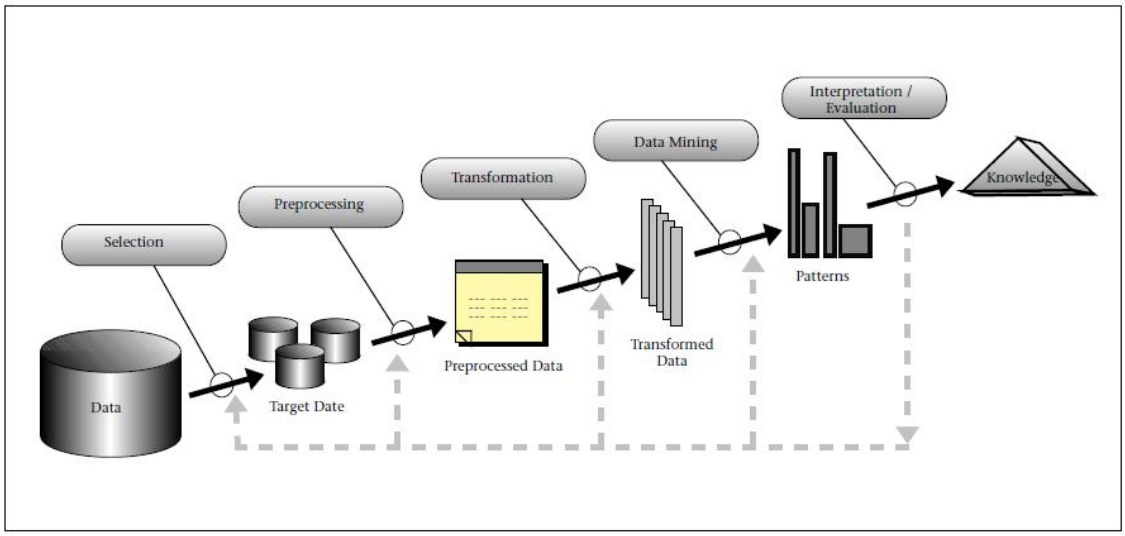

Figure 2.2: Steps of KDD [12]

The process of Knowledge Discovery KDD in database consists of the following steps [12]:

a. Data Selection:

In this step decide on the data relevance to the data mining goals to be used for analysis, its covers selection of attributes (columns) as well as selection of records (rows) in a table.

b. Data preparation:

This step covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times. Cleaning the data is one of the main objectives of this step. This means for example the filling of missing values.

c. Data Integration:
   In this step, multiple data sources and heterogeneous combined in a common source.

d. Data Transformation:
   It is also known as data consolidation, it is the stage in which the selected data is transformed into forms appropriate for the mining procedure.

e. Data Mining:
   In this step of the process clever techniques are applied to extract patterns potentially useful.

f. Interpretation and Evaluation:
   The result has to be interpreted and evaluated to come up with suitable actions.

g. Knowledge representation:
   In the final step after discovered knowledge is represented to the user, visualization techniques are used to help users understand and interpret the data mining results.

2.1.4 Data Mining Process

In the Knowledge Discovery process, the data mining step for extracting patterns from data this Knowledge or pattern depends in data mining tasks applied, data mining process involves the following phases [12]:

i. Problem definition:
   In the first phase identify the goals, based on the defined goal; the tools can be applied to the data to build the behavioral model.

ii. Data exploration:

In this phase if the quality of data is not suitable for an accurate model then recommendations on future, all data needs to be consolidated so that it can be treated consistently.

iii. Data preparation:

The purpose of this phase is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for analysis.

iv. Modeling:

Is this phase based on the data and the desired outcomes, data mining algorithm include classification techniques and clustering this algorithm is selected based on the particular objective to be achieved.

v. Evaluation and Deployment:

Based on the results and analysis of the data mining algorithms, determine the conclusions from the analysis and create recommendations for future work.

2.1.5 Data Mining Techniques

There are many data mining techniques and methods use (classification, clustering, association rules) to extract the pattern, and there are many algorithms for each techniques, these techniques are:

i. Classification:

Maps (or classifies) a data item into one of several predefined categorical classes, and extracts models to describe important data classes, the table below show the different classification algorithms[14]:

Approach	**Algorithm**
Statistical	Regression Bayesian
Distance	Simple distance K-nearest neighbors

Decision Tree	ID3 C4.5 CART SPRINT
Neural network	Propagation NN Supervised learning Radial base function network
Rule based	Genetic rules from DT Genetic rules from NN Genetic rules without DT and NN

Table 2.1: Classification Algorithms

ii. Clustering:

Technique use to partition asset of data object into group or subset each subset is cluster, the object in a cluster is similar to another; the table below shows the different clustering algorithms [14]:

Approach	**Algorithm**
Similarity and distance measure	Similarity & distance measure
Outlier	Outlier 3333
Hierarchical	Divisive, agglomerative
Partitional	Minimum spanning tree Squared matrix K-means Nearest Neighbor PAM Bond Energy Clustering with Neural Network

Clustering large database	BIRCH DB Scan CURE
Categorical	ROCK

Table 2.2: Clustering Algorithms

iii. Association rule:

Technique use to discovering associations between items, Association rules mining is a two step process, in the first step frequent item sets are generated and the second step association rules are derived from the frequent item-sets obtained in the first step, association rule has the several algorithms such as (Apriori, CDA, DDA, etc).

2.1.6 Challenges in Data Mining

Data mining systems face a lot of problems. A system which is quick and correct on some small training sets, could behave completely different when applied to a larger database. A data mining system may work perfect for consistent data and perform significant worse when a little noise is added to the training set. Problems and challenges of data mining systems [15]:

- Larger databases
- Changing data and knowledge
- Missing and noisy data.

2.2 Clustering

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the likeness (or homogeneity) within a group, and the greater the disparity between groups, the better or more distinct the clustering. [3]

Also clustering means the process of organizing objects into groups whose members are similar in some way. The initial data in the dataset are separated but when make cluster on it will be collect the near points together, show it in the figure.3 .below:

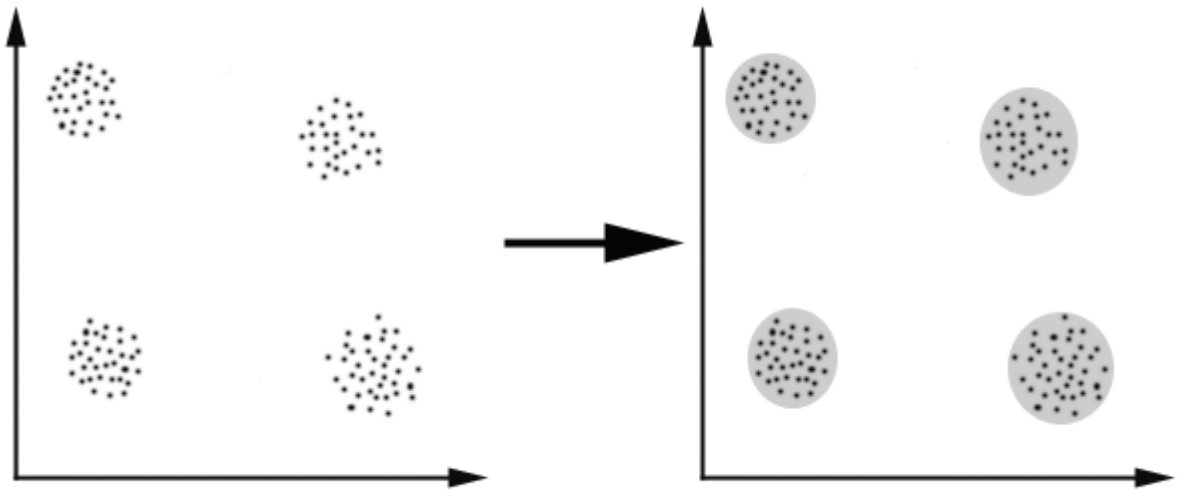

Figure 2.3: Cluster analysis [3]

2.2.1 The Goal of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unordered and clear data. In other words the process of identifying the collected data compilation.

2.2.2 Possible Applications of Clustering

Clustering algorithms can be applied in many fields, for instance: [3]

- **Marketing:** finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- **Biology:** classification of plants and animals given their features;
- **Libraries:** book ordering;
- **Insurance:** identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- **City-planning:** identifying groups of houses according to their house type, value and geographical location;
- **Earthquake studies:** clustering observed earthquake epicenters to identify dangerous zones;
- **WWW:** document classification; clustering weblog data to discover groups of similar access patterns.

2.2.3Requirement of Clustering

The main requirements that a clustering algorithm should satisfy are: [10]

- **Scalability:**

  The ability of the algorithm to perform well with large number of data objects (tuples).

- **Dealing with different types of attributes:**

  The ability to analyze single as well as mixtures of attribute types.

- **Discovering clusters with arbitrary shape:**

  The shape usually corresponds to the *kinds* of clusters an algorithm can find and we should consider this as a very important thing when choosing a method, since we want to be as general as possible. Different types of algorithms will be biased towards finding different types of cluster structures/shapes and it is not always an easy task to determine the shape or the corresponding bias. Especially when categorical attributes are present we may not be able to talk about cluster structures.

- **Minimal requirements for domain knowledge to determine input parameters:**
  Many clustering algorithms require some user-defined parameters, such as the number of clusters, in order to analyze the data. However, with large data sets and higher dimensionalities, it is desirable that a method require only limited guidance from the user, in order to avoid bias over the result.

- **Ability to deal with noise and outliers:**

  Clustering algorithms should be able to handle deviations, in order to improve cluster quality. Deviations are defined as data objects that depart from generally accepted norms of behavior and are also referred to as outliers. Deviation detection is considered as a separate problem.

- **Insensitivity to order of input records:**

  The same data set, when presented to certain algorithms in different orders, may produce dramatically different results. The order of input mostly affects algorithms that require a single scan over the data set, leading to locally optimal solutions at every step. Thus, it is crucial that algorithms be insensitive to the order of input.

- **High dimensionality:**

  The number of attributes/dimensions in many data sets is large, and many clustering algorithms cannot handle more than a small number (eight to ten) of dimensions. It is a challenge to cluster high dimensional data sets

- **Interpretability and usability:**

  Most of the times, it is expected that clustering algorithms produce usable and interpretable results. But when it comes to comparing the results with preconceived ideas or constraints, some techniques fail to be satisfactory. Therefore, easy to understand results are highly desirable.

2.2.4 Major Clustering Methods

Many algorithms exist for clustering, but there are three major clustering methods for clustering which are: [1]

i. K-means Clustering:

The term "k-means" was first used by James MacQueen in 1967, though the idea goes back to 1957. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982. K-means is a widely used partitional clustering method in the industries. The K-means algorithm is the most commonly used partitional clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time. [1]

- **K-means Clustering Algorithm:**

Using this algorithm to assemble multiple data (examples) depending on the characteristics of the K gathering, and the process of the assembly by reducing the distances between the data center and assembly (cluster centroid). [1]

The following figure shows the algorithm K-Means Clustering:

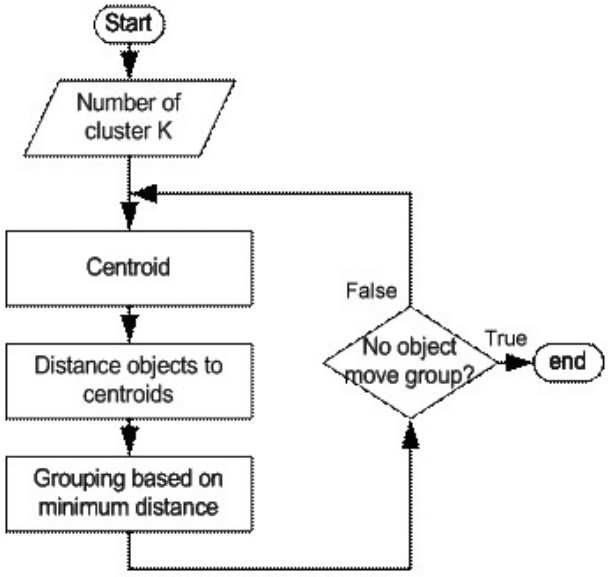

Figure 2.4: k-means lifecycle [1]

The steps of this algorithm are: [1]

1. Determine the number of clusters K, which is a step Preconditioning.
2. Determine the coordinates of the population centers Centroid randomly.
3. Calculate the distance between each instance and between all centers, and is used Euclidean dimension. Given the Euclidean distance between the two examples i, j the following relationship:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Where:

n: number of properties example.

X_{ik}: the property k coordinates for example i.

X_{jk} : k coordinates for examples j property (which is usually the center) coordinates.

4. Data Collection (examples) with its nearest center.
5. Repeat steps 2 through 4 until you get stability (lack of moving objects within communities), or even repeating a certain number of times.

The performance of this algorithm is based on the initial sites for the centers of clusters (Centroid), it is advisable to implement this algorithm several times with different centers at all times previous times. [1]

- **Advantages of algorithm K-means Clustering: [1]**

1. Highly effective.
2. Easy to implement.
3. Dealing with continuous values and discrete values (nominal).

- **Disadvantages algorithm K-Means Clustering: [1]**

1. Determine the number of clusters K is randomly before processing examples.
2. Sensitive to initial condition. Selecting multiple initial cases to produce different results gatherings, as a result, the algorithm may be located in the local end of the problem.
3. Assembly circular shape because it is based on calculating the distance.

ii. Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram [5].

I. Agglomerative (bottom up)

1. Start with 1 point (singleton).
2. Recursively adds two or more appropriate clusters.
3. Stop when k number of clusters is achieved.

II. Divisive (top down)

1. Start with a big cluster.
2. Recursively divides into smaller clusters.
3. Stop when k number of clusters is achieved.

iii. Density Based Clustering

Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts). One of the most well known density-based clustering algorithms is the DBSCAN [6].

2.2.5 Problems in the Clustering

There are a number of problems with clustering. Among them: **[7]**

- Current clustering techniques do not address all the requirements adequately (and concurrently).
- Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.
- The effectiveness of the method depends on the definition of “distance” (for distance-based clustering).
- If an obvious distance measure doesn’t exist we must “define” it, which is not always easy, especially in multi-dimensional spaces.
- The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.
- Outlier Analysis: the outliers can show many times when use clustering because in part of the data does not belong to any cluster. If the cluster was more than 15% makes for them cluster alone, or treatment them like outlier and use for them this kind of the data mining who called Outlier Analysis.

2.3 Related Works

2.3.1 Revealing Online Learning Behaviors and Activity Patterns and Making Predictions with Data Mining Techniques in Online Teaching

The study was conducted in an undergraduate course on Business Software Applications in a four-year vocational-track university in Taiwan. The course was delivered completely online via Wisdom Mater v.2.4, a widely used course management system in Taiwan. In a project-based learning (PBL) approach, the online learning experience required active collaboration among students. Ninety-eight students' online learning behaviors were recorded in server logs for six consecutive weeks. A total of 17,934 sever logs were retrieved from the LMS and analyzed in this study. [8]

Clustering algorithms were used to categorize students into homogeneous groups. K-means clustering techniques were applied to group students based on their shared characteristics: learning preference, time, duration, frequency, and learning performance. This method was based on distance concepts among individual participants and was intended to gather individuals who were close into the same group for further analysis. [8]

The results revealed that a small group of the students had a low frequency of accessing course materials, a low number of messages posted, and a low number of messages read. However, most of the students' patterns were scattered in the graphic. The results revealed again that students would rather read messages than post messages, as most of the plots gathered to the side of total frequency of accessing course materials (FAC) and total number of messages read (NR).[8]

2.3.2 Proposed Framework for Data Mining in E-Learning: The Case of Open E-Class

This paper proposes a platform dependant framework for recording, processing and analyzing data from Learning Management Systems (LMS). Its purpose and functionality is to facilitate instructors towards achievement of course efficiency. In depth, the improvements of course content are based on feedback information from data mining techniques over LMS data. In addition, a case study is discussed and an automated data mining tool, under development, is presented. [9]

The clustering was performed using the open source data mining tool Weka. The metrics and index described in the previous step were used with the SimpleKmeans for clustering

platform courses. The properties of SimpleKmeans were Euclidean distance with 2 predefined clusters. The produced results show that 22 (28%) of the courses had high activity and 56 (72%) of the courses had low activity. [9]

This paper proposes a framework for analyzing data from LMS. The main advantages of the framework are that: i) It uses data mining techniques for user and course evaluation; ii) it proposes new indexes and metrics to be used with data mining algorithms; iii) it can be easily adapted to any LMS. [9]

2.4.3 Mining Educational Data to Improve Students' Performance: A Case Study

In this paper, we gave a case study in the educational data mining. It showed how useful data mining can be used in higher education particularly to improve graduate students' performance. We used graduate students data collected from the college of Science and Technology in Khanyounis. The data include fifteen years period [1993-2007]. We applied data mining techniques to discover knowledge. Particularly we discovered association rules and we sorted the rules using lift metric. Then we used two classification methods which are Rule Induction and Naïve Bayesian classifier to predict the Grade of the graduate student. Also we clustered the students into groups using K-Means clustering algorithm. Finally, we used outlier detection to detect all outliers in the data, two outlier methods are used which are Distance-based Approach and Density-Based Approach. Each one of these tasks can be used to improve the performance of graduate student. [11]