CHAPTER-2

BACKGROUND AND LITERATURE REVIEW

This chapter covers the brief introduction of the background and literature review, including the Decision Support System, Data Mining as well as Educational Data Mining. Additionally, it also discusses the incorporation of Data Mining techniques for the development of Decision Support System pertaining to the students’ performance prediction. Moreover, the role of Machine Learning techniques in the application of EDM is also discussed.

2.1 INTRODUCTION
Students' academic success is becoming a challenging issue in higher education and technical educational organizations and this requires an extensive knowledge related to the students’ performance and innovation of an effective model that forecast the students’ performance on the basis of previous academic records. During current years data mining has become more popular in the field of higher education. The classification and prediction methods are implemented at various universities for extracting knowledge from the students’ database that support various decision making situations to improve the success rate of students, additionally data mining methods also facilitates for the development of decision support system pertaining to student performance prediction. A number of decision support system using data mining has been implemented within last one decade.

2.2 DECISION SUPPORT SYSTEM
The notion of a decision support system (DSS) is extremely broad and the definitions of DSS vary depending on the author’s point of view (Druzdzel et al., 1999). (Finlay et al., 1994) define “a DSS broadly as a computer-based system that aids the process of decision making”. (Turban et al., 1995) defines “it as an interactive, flexible, and adaptable computer-based information system, especially developed for supporting the solution of a non-structured management problem for improved decision making. It utilizes data, provides an easy-to-use interface, and allows for the decision maker’s own insights”. According to (Keen et al., 1978), "DSS are computer-based support for management decision makers who are dealing with semi-structured problems."
(Sprague et al., 1982) stated “DSS are interactive computer based systems that help decision makers utilize data and models to solve unstructured problems “. DSS have been developed for many applicative domains such as, a DSS was proposed particularly for loan risk assessment using fuzzy theory, in this the authors defined a risk assessment analysis along with adoption strategies to represents the complexity of the system and MATLAB tool was used for the evaluation of different factors regarding new loan applicant such as age, assets, job stability, income etc. (Kumar et al., 2011). The DSS developed for to education system are discussed in detail in section 2.6.1.

However DSSs can build the decision making process more competent but the role of decision maker can’t keep away from the decision. Therefore, it is necessary to understand when to use a DSS, and to what extent it can be used? Thus DSS can be used as a precious tool in decision-making process instead of as a decision making mechanism.

2.1.1 Decision Making Process

A decision making process begin with the identification of a problem, once a problem is identified, a number of alternative solutions pertaining to the problem are generated (Basra et al., 2010). These alternative options are critically examined and evaluated, and the best alternate option is chosen for implementation of the DSS. The implemented solution is further evaluated over time to assure its immediate and continued effectiveness. If problem arise at any stage in the process, the entire process can be repeated for achieving the desired results. Thus, a decision making is a logical sequence of activities. A typical decision making process consist of three phases, namely intelligence, design and choice, respectively (Ismail et al., 2014) as shown in Figure 2.1.

- **Intelligence Phase:** The intelligence phase consists of identifying and formulating the problems. The end of this phase is the problem statement or decision statement.
- **Design Phase:** This phase starts with the formulation of a model and generation of set of alternatives based on problem statement.
- **Choice Phase:** In this phase all the generated options developed in design phase are evaluated and best option is selected for the decision.
2.1.2 Architecture of DSS

The architecture of DSS varies depending upon the authors’ viewpoint. According to (Sprague et al., 1982) “there are three elementary components of the DSS (i) The Database Management System (DBMS), (ii) The Model-base Management System (MBMS), and (iii) The Dialog Generation and Management System (DGMS) “. (Power et al., 2002) discussed four major components for development of a DSS: (i) “The user interface, (ii) The database, (iii) The model and analytical tools and (iv) The DSS architecture and network “. Building upon the various existing architectures (Marakas et al., 1999) proposed a generalized architecture for the DSS based on the existing structure of the DSS. This generalized architecture consists of five distinct parts: “(i) The Database Management System, (ii) The Model Management System, (iii) The Knowledge Engine, (iv) The User Interface, and (e) The users”. The five distinct components of DSS are shown in Figure 2.2.
- **Database Management System (DBMS):** A DBMS component of a DSS has a number of sub components for e.g. database, data dictionary, query tools etc. A DBMS is used to store, retrieve and organize data according to the specified domain of an application. It also provides security mechanism, integrity constraints, policies and general data administration duties.

- **Model management System (MMS):** The model management performs storage, retrieval and organizes activities for the quantitative models that offer the analyzing capability for DSS. The model based management system is used to create new models by using programming languages.

- **Knowledge Engine:** The Knowledge Engine plays the role of brain of the system. It provides knowledge in solving complex unstructured and semi structured problems. This is responsible for dealing reasoning, handling uncertainty and learning from historical data. The data and model come together in KE to offer the user with a significant help that supports in decision making context. This leads to intelligence DSS.

- **User Interface:** It provides a communication (dialog) between user and DSS. It is used to access the data, model and the processing components of DSS. This hides the technical complexity of the DSS from the user and provides ease of use.

- **User:** The role of user in DSS cannot be avoided. User control is a primary characteristic of the DSS. A user can be a manager, staff specialists or intermediary user. Users’ skill set, knowledge domain and role within the organization are the essential element in successful application of DSS.

2.1.3 **Classification of DSS**

(Kacprzyk et al., 2007) classify the DSS in following seven categories:

1. **Data Driven DSS (DDDSS)**
   DDDSS emphasize an access and manipulation of internal business data as well as external data, it may access data from lower level to higher level for e.g. simple file systems, data warehouses, On-line Analytical Processing (OLAP) tool etc.

2. **Communication Driven DSS (CDDSS)**
   CDDSS uses the network and communications technologies to facilitate collaborations and communications (Kacprzyk et al., 2007). It is useful where more than one person involve in a task.
3. **Group DSS (GDSS)**
GDSS are interactive, computer-based systems that facilitate solution of unstructured problems by a set of decision-makers working together as a group (Kacprzyk *et al.*, 2007).

4. **Document Driven DSS (DDDSS)**
Integrate a variety of storage and processing technologies for a complete document retrieval and analysis. Documents may contain numbers, text, and multimedia (Kacprzyk *et al.*, 2007).

5. **Model Driven DSS (MDDSS)**
MDDSS call attention to access and manipulate a model, such as financial statistical, simulation models etc. It uses data and parameters given by users for analyzing a situation.

6. **Knowledge Driven DSS (KDDSS)**
KDDSS are interactive systems with a particular problem solving expertise consisting of knowledge about a particular domain, understanding of problems within that domain, and “skill” for solving some of these problems (Kacprzyk *et al.*, 2007)

7. **Web Based DSS (WBDSS)**
WBDSS are computerized system that delivers decision support related information and/or tools to a manager/analyst using a “thin-client” Web browser (Explorer), TCP/IP protocol, etc. (Kacprzyk *et al.*, 2007).

### 2.1.4 Advantages of DSS
- DSS enhance the decision making capabilities of decision makers by extending the capability to practice information and facts to handle large, complex and time consuming problems (Holsapple *et al.*, 2001).
- Enhance the reliability of a decision making process.
- Supports investigation and innovation ability of the decision makers.
- Expose new methods for the consideration of a decision context.
- Produce new facts to maintain the authentication of obtainable hypothesis.
- Generate a tactical or aggressive benefit for the competing business.
- Reduces the time related to making a decision. It also reduces the cost associated with wrong decisions.
- DSS facilitates interpersonal communication.
2.1.5 Limitations

- DSS can’t replace human decision making aptitudes such as creativity, imaginations or intuitions.
- DSS are usually intended for a particular application. This thwarts their generalized use of various decision making contexts.
- Languages and command interfaces are not sophisticated of user directives and queries.
- The influence of DSS is bounded by the computer system upon which it is running.

2.1.6 Problems /Challenges in existing DSS Approaches

Traditional techniques of data analysis are not efficient enough to handle all kinds of data analysis problems. Traditional DSS generally make use of OLAP tools for an advance data analysis and OLAP tools pursue a deductive approach (JSR-73 Expert Group, 2004). The disadvantage of this approach is that, it depends on luck or coincidence, if the right dimensions gets selected during the drilling down process to retrieve the information and patterns then results might be correct. This approach is also lacking to follow an algorithmic method; it also depends on the analysts inside and chance for obtaining the most valuable information, fashions and patterns. Moreover, there is a limitation for attributes selection and only few attributes can be selected simultaneously (Goebel et al., 1999). Therefore the notion of Data Mining comes into the picture. Data mining is an interdisciplinary field consisting of standard statistical and current machine learning techniques for supporting the data analysis. The data mining performance analysis follows an inductive approach of analyzing data where machine learning algorithms are employed to obtain knowledge from data (JSR-73 Expert Group, 2004). Data mining eliminates the above mentioned drawback.

2.2 DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASES (KDD)

Data Mining is one of the important and young features in the area of computer science. It generally refers to or means the extraction of hidden information from large amount of databases. Data Mining assists business to discover and identify the hidden patterns in databases. (Mitchell et al., 1997) wrote “Machine Learning is a mature and well-recognized research area of computer science, mainly concerned
with the discovery of models, patterns, and other regularities in data”. “Data mining generally refers to the extraction of hidden information from the large amount of databases “. (Dunham et al., 2003) cite “Data mining helps organizations to discover and identify the hidden patterns in databases. The extracted patterns are then made in use to build data mining models and hence can be used to predict performance and behavior with high accuracy. “Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic”. A frequently cited definition of data mining is “Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data”. According to (Janecek, 2009) the two basic objectives of data mining are:

i) To identify regularities in data.

ii) To discover relationships among data and predict the unknown values.

2.2.1 KDD process

By definition, data mining is a part of a (KDD) knowledge discovery in databases process, the word “database" refers to any kind of data storage and does not solely comprise data stored in database management systems. KDD can be defined as “Knowledge discovery in databases which describes non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (Fayyad, 1996). Moreover KDD address the problem of mapping low-level data (too large and complex to understand) into other structures that can be more conceptual and useful for future cases. Data mining is integrated as a single step of the KDD, usually between model selection and interpretation. A typical KDD process is depicted in Figure 2.3. It consists of following six steps:
1. **Data Collection**

The first step in KDD process is data extraction from internal as well as external sources. Data can be in any format for e.g. relational database system, a typical flat file format, XML or from a data warehouse. There might not be any extracted information about the data, but instead samples are stored in their original format. Moreover, the features and properties of the data need to be computed explicitly. If the data are not already available, the data extraction phase might take a larger part of the time and efforts of the entire knowledge discovery process (Janecek, 2009).

2. **Data Preprocessing**

This step consists of data cleaning, data integration and data transformation steps. As high-quality models require good quality of data obtained after a thorough cleansing of the data. An appropriate model also demands for the correctness and consistency of data to improve the quality of data mining methods. The data obtained from various sources are integrated into a single format and transformed into appropriate forms for mining by performing summary or aggregation operations. The data preprocessing step also deal with the problem of missing data, which is also a major problem, particularly when the dataset is small or the number of missing fields is large. A widely applied approach is to calculate a substitute value for missing fields is to calculate the median or mean of a variable (Hailu et al., 2015; Janecek, 2009).

3. **Feature Reduction**

In many application areas, datasets can have a very large number of features. (Michael et al., 2008) cited “as the dimensionality of the feature space increases, many types of data analysis and classification become significantly harder.
Additionally, the data becomes increasingly sparse in the space it occupies which can lead to big difficulties for both supervised and unsupervised learning. In the literature, this phenomenon is referred to as the curse of dimensionality” and was first mentioned in 1961 by Bellman. As a result, a higher number of features can lead to lower classification accuracy (Michael et al., 2008). Therefore a feature selection method can be used to select appropriate attributes for e.g. Chi-squared methods, information gain method, gain ratio method etc.

4. Data Mining

Data Mining is the core part of the knowledge discovery process. Typically, data mining methods has two high-level goals: prediction and description. The goal of predictive data mining tasks is to build a model that predicts unknown values of a particular attribute of interest based on known values of some other attributes. “The job of constructing a model for the class variable as function of the explanatory variables is called predictive modeling” (Janecek, 2009). Predictive modeling consists of classification and regression techniques. The purpose of classification as well as regression is to develop models that minimize the errors between the actual values and predicted values of the class variable. In almost any data mining application, the choice of exact learning algorithm is also crucial step, since there is no generally “best” learning algorithm. One algorithm may be more accurate in the sense of classifying unlabeled objects, but very slow in performance. Other algorithms may be very fast but not as accurate as others. The process of building predictive models, requires a well-defined training and test set in order to ensure that the data is trained on one dataset and tested on another dataset with different objects in order to measure the expected generalized errors. This results in an estimate of how the model will perform on future data that are similar to the training and test data. Sometimes a third set, called a validation dataset, is used to act as a further independent measure of the model's accuracy. There are several methods that are commonly used to divide the samples into two datasets namely training data set and testing dataset (Janecek, 2009).

5. Evaluation and Interpretation

The next step in KDD process is to evaluate the model proposed in the previous step. A number of evaluation measures are available in the literature including True Positive Rate (TPR), False Positive Rate (FPR), Recall, Precision, Confusion Matrix,
Cost Matrix ROC Curves and Traditional Performance Accuracy etc. Once a classification model gets constructed, it can be employed to forecast the class of unseen objects. These measures help to determine the interpretation of the discovered patterns obtained from the previous steps. In this step, these patterns are represented in possible visualization form, for example, graphs, charts etc.

6. Knowledge Representation

The discovered knowledge and information obtained from data mining can be applicable in various decision making situations, information management, and query processing. Therefore, data mining can be considered as one of the most important tools in information systems and information technology.

2.3 INTEGRATION OF DATA MINING AND DECISION SUPPORT

In general, a typical decision support system is constructed for a specific purpose based on a particular decision process, a set of methods and procedures. The DSS can be classified as simulation (Chen et al., 2004; Kuan et al., 2004) based DSS, analysis based DSS (Bohanec, 2001; Heinrichs et al., 2003; Ward, 2000) prediction based DSS (Zhong et al., 2005; Patelis, 2003) and optimization based DSS (Heinrichs et al., 2003). Most of the business applications use traditional OLAP based DSS for analysis purpose. Data analysis through OLAP follows a deductive approach for analyzing data (JSR-73 Expert Group, 2004). OLAP can answer the question like “What has been going on?” but “OLAP is not able to answer for the question like What are the characteristics of the best customers?”

Many researchers argue the uses of data mining to facilitate decision support. According to Chen and Liu, the use of data mining in DSS enables the academic organizations to make critical decisions faster with a higher degree of confidence and reduces the uncertainty in decision process (Chen et al., 2004).

Mladenic stated that “the integration of data mining and decision support can enhance the performance of DSS and can help to handle new problems which were never tackled before”. They also advocate for the integration of data mining and decision support as it can significantly improve the existing methods and can create new methods for problem solving (Mladenic et al., 2003). Chen and Liu also agreed that data mining is a very helpful in the field of data analysis (Chen et al., 2005).

Data mining can be employed using two approaches (Rupnik et al., 2007)
i) **Data Mining Software Tool Approach.**

In this approach data mining projects make use of data mining software (Goebel et al., 1999; Holsheimer, 1999; Chen et al., 2004). However, data mining software tools require extensive knowledge in data mining methods, databases and statistics, they offer a variety of methods and parameters for complete analysis but they are rather complex and difficult for a user (Chen et al., 2004).

ii) **Data Mining Application System Approach.**

This approach requires the development of decision support system or an application software which is designed for the business users and other decision makers, that enables them to view data mining models presented in user friendly manner with an interactive GUI based graphical presentation (Aggarwal, 2002). This approach provides better integration of data mining technique with business environments and their decision processes in DSS (Goebel et al., 1999; Holsheimer, 1999; Chen et al., 2004; Bayardo et al., 2001).

### 2.4 EDUCATIONAL DATA MINING (EDM)

The application of data mining in higher education is a new promising interdisciplinary area. It is also referred as Educational Data Mining (Romero et al., 2007). It emerged as an independent research area in recent years, However, Educational data mining methods vary from the broader data mining due to the multiple levels of meaningful hierarchy in educational data. “Educational data mining is defined as the area of scientific inquiry centered on the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in”.

Educational Data Mining (EDM) is concerned with developing methods for educational data and uses these methods for understanding of students and their learning environments (Baker et al., 2009). The EDM process translates raw data educational systems into useful information that could potentially have a great impact on educational research and practice. This process is same as any other application of data mining because it also follows the same steps of the data mining process (Romero et al., 2004). EDM includes academic performance predictions of students,
clustering of students, and the association of students in suitable courses. *(Baker et al., 2009; Baker et al., 2010)* suggested four applications and five methods for EDM. The application of educational data mining includes “improving student models, improving domain models, studying the pedagogical support provided by learning software, and scientific research into learning and learners”. “The five approaches pertaining to educational data mining includes prediction, clustering, relationship mining, distillation of data for human judgment and discovery with models” *(Romero et al., 2010)*. The data mining in higher education is also useful to progress the procedure of decision making in the management system of university *(Luan, 2002)*. The objective of EDM is “to discover useful information from the large volume of electronic data collected from educational systems”. Figure 2.4 *(Romero et al., 2007)* depicts an “overview of EDM process” proposed by Romero and S. Ventura related to Educational Data Mining.

There are two popular taxonomies for classifying data mining techniques in educational systems proposed by *(Romero et al., 2007)* and *(Baker et al., 2009)*. Figure 2.5 obtained from *(Ribeiro, 2013)* shows the relationship between the two taxonomies.

It is clear from the Figure 2.5 that in Romero-Ventura’s taxonomy fourth group can be viewed as an extension of data mining for text data and it is related to web content mining *(Romero et al., 2007)*, while group one, group two and group three are the usual task of data mining.

Baker-Yassef’s taxonomy includes *(Maimon et al., 2002)* methods for data mining (second, third and fourth group) except the group one and group fifth, where group one corresponds to the Romero-Ventura’s "Statistical and visualization" group and the fifth group is considered as the most unusual classification.
2.5 MACHINE LEARNING

(Mitchell, 1997) wrote “Machine Learning is a mature and well-recognized research area of computer science, mainly concerned with the discovery of models, patterns, and other regularities in data” (Baker et al., 2009). “Machine learning can be defined as a set of methods that can automatically detect patterns in data, and then find out the uncovered patterns to predict future data”. Application of machine learning techniques to huge database is known as data mining. Machine Learning is not limited to database problems. It is also a part of artificial intelligence as an intelligent system has the ability to learn in changing environments. Machine Learning is also useful in building mathematical models. In this, computer science plays a vital role in twofold. i) During training, an efficient algorithm is responsible for solving the optimization problem as well as managing and processing of huge quantity of data. ii) Once a model is trained, it provides an efficient algorithmic solution and its representation (Alpaydin, 2004).

In recent research a heuristic based association rule hiding proposed by introducing the notion of the impact factor of the transaction on the rule (Pathak et al., 2012). According to them “higher the impact factor of a transaction, the higher is its sensitivity”. (Narender et al., 2008) also proposed an algorithm based on the semi supervised approach for classification using neural network pertaining to the concept of geometrical expansion. A classic machine learning process is shown in Figure 2.6 obtained from (Kilany, 2013).
2.5.1 Machine Learning Applications

- **Finance Sector**: Machine Learning techniques are widely used in finance and banking sectors. The bankers’ past data is employed to build the model. This model helps to identify whether a customer is genuine customer or not for new credit applications. It helps in fraud detection.

- **Telecommunications**: Telecommunication companies analyze the call patterns of the customers for network utilization and maximizing the quality of service (QoS). This is also helpful to retain customers by providing attractive offers for regular and loyal customers.

- **Manufacturing**: Machine Learning methods are used for optimization, control, and troubleshooting.

- **Medical**: Machine Learning methods are popular in medical diagnosis. Now a day’s it is widely used in clinical research.

- **Science**: A huge amounts of data in science such as astronomy, physics, and biology can only be examined and analyzed using machine learning techniques.

- **Research**: Machine Learning is also helpful in many research based problems for example image recognition, speech recognition, robotics etc.

2.5.2 Machine learning Methods in EDM

- **Predictions (Classification and Regression)**

  Prediction is the process of developing a model that determines the unknown class variable (predicted) on the basis of the combination of known values variable (predictor). It typically consists of two steps. First is model generation and second is model usage. During the first step an algorithm is used to induce
a mathematical model for the training dataset. In second step this model is used
to evaluate for other datasets to determine the predictive accuracy of this
model. Predictions typically involve classification and regression algorithms.
There are number of classification algorithms available in literature For e.g.
Decision Tree, Naïve Bayes, K-Nearest Neighbor etc. However classification
can be performed only for discrete valued variables (nominal) whereas
regression can be used only for continuous (numerical) variables. Regression
analysis also used to construct a model for determining the relationship
between one or more predictor variables and a dependent variable (which is
continuous-valued). A regression technique includes linear regression, logistic
regression and Poisson regression etc. *(Han et al., 2006)*.

Prediction has been extensively used to understand the behavior of the students
and causes of student attrition. Prediction can be used in two ways in EDM.
First, prediction methods can be used to identify the most important features of
a model for prediction about the underlying construct. This is a common
approach that allows predicting student educational outcomes *(Romero et al,
2008)* without predicting intermediate or mediating factors *(Baker, 2010)*.
Second, prediction methods can also be used to predict the output values where
it is not required to find a label for that construct. A prediction model was
developed *(Baker et al., 2008)* using the observational methods where data was
collected from interactions between students and software for predictor
variables and model accuracy was evaluated during the new examples

- **Clustering**

Clustering is the process of grouping a set of instances into classes of similar
instances. A cluster is a collection of data objects that are similar in a group
while dissimilar to the objects in other group *(Han et al., 2006)*. Clustering
can also be referred as data segmentation as it partitions large datasets into
groups according to their similarity. There are number of clustering
algorithms available in literature. Some algorithm starts with the specific
hypothesis for e.g. Gaussian Mixture Models *(Baker, 2010)* while few
clustering algorithms does not require any prior hypotheses for e.g. k-means
algorithm *(Steinbach, 2000)*. In EDM clustering can be used for grouping the
students based on educational background, age, areas of interest, specialization
and so on *(Ranjan et al., 2007)*. Similarly students’ behavior can be clustered
together to discover patterns of behavior of the students (Amershi et al., 2006; Beal et al., 2006). Clustering can also be performed to examine similarities and differences between the universities.

➢ **Relationship Mining**

The relationship mining determines the relationships between the large set of variables in a dataset. Relationship mining is classified in four categories such as association rule mining, correlation mining, sequential pattern mining and causal data mining. Association rule mining determines the relationship between the variable in terms of if then rules for e.g. a rule might be found of the form (If student attendance is more than 80%, then probability of success is 90%). A correlation mining discovers linear correlations between the variables. The correlation can be positive or negative. In sequential pattern, mining determines the temporal associations between events. The causal data mining is used to find that whether the cause of one event can be the cause of another event. A study related to relationship mining was conducted using market basket analysis to analyze “why students’ uses of practice tests decreases over a semester of study” (Bienkowski, 2012). The results of relationship mining are valuable only when they expose unexpected rules from dataset.

➢ **Outlier Detection**

An outlier is an instance that does not accomplish the common actions of the model of the data (Han & Kamber, 2006). Outliers are also known as exceptions or noise. However in many applications outliers are ignored but Outlier detection is an important activity for some data mining applications such as fraud detection. The objective of Outlier detection is to determine these instances and the cause to it. They can be generated due to errors during loading the data into databases, errors in measurement of data or there is no error in measured data but data is associated with a rare event (Ribeiro, 2013). In educational data mining, outlier detection get used to address the behavioral issues of students having learning problem (Ribeiro, 2013)

2.6 **LITERATURE REVIEW**
This section covers the work done so far in the field of students’ performance prediction. The Literature Review is presented in four subsections including Decision support System in education, followed by students’ performance prediction, data mining for students’ performance prediction and ends with the most popular current technique namely ensemble methods for education environments.

2.6.1 Decision Support System in Education
(Luan, 2001) developed an influential decision support tool with the help of data mining techniques.

(Deniz et al., 2002) designed a decision-support system for academic administrators in which data is made available through the use of standard user interfaces, displays and graphs, as well as in tabular form. This minimizes the time and effort needed to analyze pages of data.

(Krisper, 2006) defined a business system called “DMDSS (Data Mining Decision Support System)”. The designed model was very much inspired from the traditional model provided by oracle called Oracle Data Mining (ODM) model. The presented model incorporates the complex tools and demands that expertise the data mining algorithms and the parameters. The presented data mining model is based on CRISP-DM (Cross Industry Standard Process for Data Mining). A DSS was implemented for university enrollment management by (Maltz et al., 2007).

(Nguyen et al., 2007) defined a Bayes Network based prediction model for the prediction of students’ performance. The input given to the system is in the form of grade point analysis and based on the study the prediction process is performed. The author has defined a case based mechanism from the Bayesian network to perform the prediction. The main classification attribute considered in this work is the student CGPA. According to this the classification of students is done in the form of grades such as B Grade, C+ Grade etc. The obtained results from the system are satisfactory enough to predict the student grade.

(Bresfelean et al., 2009) discussed several issues for higher education environments such as decisions, decision process, Decision Support System, role of Data Mining and Decision Support System (DSS) in decision making. They also proposed the architecture of a DSS in higher education for the Faculty of Economics and Business Administration at Babes-Bolyai University of Cluj-Napoca.

(Elayidom et al., 2009) designed a rule based hierarchical system to obtain the information gain and to take the predictive decision based on current and the historical
dataset. Author has designed tool in weka for public use, so that students can take the appropriate decision regarding the course selection.

(Changjiang Li et al., 2010) performed an analytical study on hybrid model based on Data Warehouse and Online Analysis Processing that collectively defines the Decision Support System. The author has defined a multi dimensional model called Star data model and the Snow flake model. The author basically used the oracle data mining model in an integrated form to derive the optimum results from the system.

(Nripendra et al., 2011) proposed architecture of Decision Support System for the decision making processes in higher educational institutions and also demonstrate the use of DSS in admission process.

(S. B. Kotsiantis, 2011) described the Machine Learning techniques for forecasting the student’s grades. The author employed regression techniques for the development of a DSS for the prediction of students’ academic performance.

(Salma et al., 2012) performed a factor based analysis for the teaching system to answer the classroom questions and presented an interactive case based learning system for the business intelligence.

(Saxena et al. 2012) proposed a simulation tool for the classification and analysis of students’ performance. The authors used Fuzzy System, Neural Network, and Genetic Algorithm for their study to evaluate and to predict students’ performance.

(Livieris et al., 2012) employed neural network to develop a student performance prediction tool particularly for mathematics course that enables educators to identify weak students.

2.6.2 Student Performance Prediction

Student’s success is of great importance for the progress and development of the country. It helps in producing the skilled and educated persons required for good country growth.

Higher percentage of failure of students reflects the University reputation. Therefore, Student’s performance is a major area of concern in Universities for attaining good academic achievements (Fennolar et al., 2007). Universities can plan several strategic programs for the improvement and maintenance of academic performance of students. (Rusli et al., 2008).

Predicting the student’s performance and improving the teaching learning process by using academic evaluation are vital. In higher education, the study involved in
identifying the key indicators that lead to academic success is an important area of concern. (Ervin et al., 2005). “The performance indicators related to education aims at measuring the access, retention and achievement of the students” (Ho et al., 2001). “The evaluation can be carried out in many ways such as examination, projects, homework, assignments etc”. (Cardenas, 2000).

According to (Kinnunen et al., 2007) CGPA acts as a main indicator for the student’s academic performance prediction for evaluating the examination results. Student’s performance can be evaluated based on their academic achievements (Baird, 1982; Hilton, 1982; Ditcher et al., 1999). (Baird, 1982) stated that “academic achievements is the major consideration in the award of scholarship”. According to (Hilton, 1982) “previous educated status especially in high school caused the retention in higher education”. (Ditcher et al., 1999) mentioned that the academic achievement will always be described in terms of grades or degree completion.

(Toh, 2010) cited “high self esteem and determination of students results in achievement of good academic performance. Therefore an interest has been developed in students as well as in instructors for identifying the factors that lead to the prediction of academic performance which also facilitates to take corrective actions within the time”.

In modern technical era data mining became more popular in the field of classification and prediction. Data mining techniques are very powerful tool as it is used to extract and describe the database and helps in the prediction of new data based on a set of models or rules (Pimentel et al., 2005). Data mining also has been successfully applied in students’ academic performance prediction. The following section presents a literature review of data mining for student performance prediction.

2.6.3 Data Mining for Students’ Performance Prediction

According to (Haddawy et al., 2007) “Data mining is the convergence of multiple disciplines such as database technology, machine learning, information science, statistics, visualizations and other disciplines”.

Quality education is a big priority for every academic organization. Data mining helps in improving the quality level. (Vranic et al., 2007) used the Data Mining techniques in their research to improve the level of quality in education. The objective of this research was to study the behavior of the students’, their learning ability in engineering discipline and to predict the success rate of the students.
(Baker et al., 2009) “Machine Learning techniques include classification and regression algorithms, association rules, sequential pattern analysis, as well as clustering and web mining”.

(Hsia et al., 2008) conducted a study for a university in Taiwan. In this research three Machine Learning techniques including Forest tree, Decision Tree and Link Analysis etc. were used to analyze the course preferences and course completion rates of the students.

(Sacín et al., 2009) described a different Machine Learning method based on classification for predicting an appropriate course is for a particular student. (Jantan et al., 2010) determined the latent classification techniques for academic capacity predicting in higher education organizations. (Lykourentzou et al., 2009) employed three Machine Learning techniques, namely Support Vector machines, Feed-Forward Neural Networks, and Probabilistic Ensemble Simplified Fuzzy ARTMAP to predict the dropouts in e-learning courses.

(Campbell, 2007) conducted a study based on regression analysis for prediction of student’s academic performance and conclude that LMS variable (predictor) is three times better than the SAT scores (predictor) of the students.

(Tseng et al., 2007) proposed a “Two-Phase Concept Map Construction (TP-CMC) algorithm based on Fuzzy Set Theory, Education Theory and a Machine Learning techniques to find grade fuzzy association rules. (García et al., 2010) used association rule mining to develop a data mining tool for the improvement of e-learning courses.

(Anaya et al., 2011) proposed an approach to examine collaboration using two Machine Learning techniques namely decision tree and clustering, so that this can be used by assigning a collaborative value to each student for the comparative study of students’ collaborative behavior.

(Lin et al., 2011) conducted a study for Automobile Corporation in Taiwan and China. They used two-stage clustering techniques (SOM and K-means) to collect personnel training data for data mining and analysis.

(Luan, 2004) used data mining as a prediction tool, to predict whether students will be able to successfully complete their graduation or not and the information obtained from the prediction can help to lower performance students to improve their performance.

(Lee et al., 2000) conducted a study to identify weak students. (Luan, 2002) also discussed the role of data mining to maximize efficiency in academics. Another
The application of data mining in educational systems was shown by (Erdogan et al., 2005) to improve the making process.

(Rajagopalan et al., 2006) developed a classifier for predicting the students’ performance on the basis of similar patterns of the available data. A model was proposed by (Sargenti et al., 2006) for diffusion of knowledge within a small business university. A case study was discussed by (Chang et al., 2006) using predictive modeling to predict the college admission for a state university admitted applicants at a large state university.

A prediction has been made about the results of students to warn them before the final exams in (Galit et al., 2007). A study (Shannaq et al., 2010) of the existing students’ behavior has been done to manage the relation with future students, so that loyalty of student can be enhanced. In another research the academic performance record of the students were used for the enrollment process (Vialardi et al., 2011).

A research has been conducted by (Delavari et al., 2005) to prove the importance of quality education. The model includes main processes of educational system namely evaluation, computing, marketing, performance examination etc. The decision tree and association rule machine learning techniques were used to improve these processes.

(Nguyen et al., 2007) conducted a research in Vietnam at CTU and AIT and used the Bayesian Network and Decision Tree to predict the academic performance of undergraduate and postgraduate students. This prediction was made for selecting best students for scholarship at the AIT and to assist failing students at CTU.

(Kabra et al., 2011) developed a system for student as well as for teachers. Students can enhance their learning process with the help of the system, while teachers can identify weak students and can provide the necessary guidance for their success.

(Lobo et al., 2011) conducted their research on the application of EDM, which deals with the development of novel process to determine the information from educational databases. They have studied and applied data mining to the educational system and have shown that how data mining in education can help in improving the students’ performance.

(Bichkar et al., 2011) proposed a classification model using decision tree for the prediction of students’ academic performance particularly for engineering students.

(Qasem et al., 2006) used CRISP methodology and proposed a decision tree based classification model for predicting students’ performance grade.
(Vijayalakshmi et al., 2011) conducted a study on internal marks of the MCA students and used a decision tree C4.5 classifier to predict the results. The predicted results were compared with the original results using two versions of decision tree classifiers namely C4.5 and ID3.

(Bharadwaj et al., 2011) employed ID3 decision tree algorithm to analyze the students’ performance. They collected the semester information such as attendance marks, class test marks, assignment marks etc. from the student’s database and predict the performance at the end of the semester.

(Diego Garc et al., 2011) defined a comparative analysis of different classification approaches for prediction of student performance. Two basic machine learning techniques called Bayesian network and the Decision Tree approach have been used. The obtained results from the system shows that Meta algorithm gives more significant results then these two algorithms.

(Garcia et al., 2011) compared the results of different classifiers applied to the educational datasets and proposed a Meta Algorithm to enhance the accuracy of the classifiers.

2.6.4 Ensemble Methods in Education

Ensemble is an effective technique to combine multiple learning algorithms and improves the overall prediction accuracy (Bahador et al., 2011).

(Sharaf et al., 2013) proposed ensemble based algorithm by combining Adaboost with the Genetic Algorithm for the students’ performance prediction in near the beginning stages to minimize the failure rate by counseling the risk associating students.

(Essa et al., 2012) developed an ensemble-based generalized system to keep track of the student academic success (S3). This system used flexible predictive modeling engine to identify students that are at danger zone, a set of advanced data visualizations and a case management tool for applying management intervention.

(Maclin et al., 1999) compared several different models for tracking students’ knowledge within intelligent tutoring systems. They also performed some ensemble techniques to ensemble multiple student models at the action level and evaluate the predictions in terms of future performance.

(Kotsiantis et al., 2010) proposed online ensembles of a classifier for prediction of students’ performance in distance education. They proposed an online ensemble of
classifiers which integrates three online algorithms namely Naive Bayes, 1-NN and
WINNOW incrementally using the voting methodology and concluded that this
approach is better than batch line mode of ensemble classifiers for developing
decision support tool for distance education.

(Paris et al., 2010) combined the HNB and Decision stumps (DS) algorithms using
voting technique for the students’ performance prediction. They used a number of
classifiers for comparison with the proposed voting technique for 2 class prediction, 3
class prediction and 4 class prediction. (Kotsiantis et al., 2005) conducted a study of
the local voting scheme of classifiers using three weak learning algorithms namely
OneR, Decision Stump and Naive Bayes. Authors also carried out a comparison with
well-known ensemble methods on benchmark datasets and find that the proposed
technique is more accurate.

2.7 CHALLENGES IN EXISTING STUDENTS’ PERFORMANCE PREDICTION
AND OPPORTUNITIES

A students’ performance prediction helps to determine how students may perform in
the class. The purpose of the students’ performance system is to evaluate the teaching
learning process as well as to find the students’ who are at risk, so that an innovative
and effective teaching learning plan can be developed and implemented for achieving
the success particularly for weak students. Many educational researchers proposed
many efficient models to predict students’ academic performance in a class (Emerson
et al., 2004; Holland, et al., 1966; Kotsiantis et al., 2003; Lowis et al., 2008;
Pittman, 2008).

However, identification and selection of effective modeling techniques is also
essential step in predictive modeling. Various mathematical techniques, namely
regression and neural networks, were applied for developing the predictive models.
These techniques has some advantages and some disadvantages. Regression is one of
the most frequently used approaches for developing predictive models but regression
cannot determine complex relationships. Moreover, it is also prone to outliers due to
the mean that is included in regression formulas. Neural Network is another popular
classification technique used for prediction, and this technique can be used for any
linear or nonlinear function. The limitation of this technique is the interpretation of
the results.

In another work (Fang et al., 2010) used a decision-tree approach for the prediction of
students’ academic achievement in an engineering discipline. The model proposed by
them was able to generate a set of “if-then” rules only regarding a student’s overall performance.

A core problem in current approaches is that a single hypothesis/model is selected from the space of all possible hypotheses, and then applied to make predictions. Moreover it cannot be assumed that a predictive model developed for a particular learning context is applicable to another learning context (Essa et al., 2012). A particular learning algorithm can perform efficiently for a particular learning context while in another context some other algorithm can be performing better. This limits the single classifiers’ approach to provide a scalable generalized solution. The second limitation in current approaches is the ability of practitioners to interpret the results of predictions.

In this research a generalized solution is proposed for developing a modelling strategy that is appropriate and supports a wide range of requirements for educational institutions and can take full advantage of predictive analytics using Ensemble Technique. “Ensemble methods are designed to boost the predictive generalized ability by blending the predictions of multiple models” (Huang, 2011). Mathematically, an ensemble classifier offers an extra degree of freedom in the classical bias/variance trade off, and provides a solution that would be difficult to achieve with only a single classifier (Huang, 2011).

2.8 CONCLUSION

In this chapter, the taxonomy of DSS, Data mining, Educational Data Mining and Machine Learning were explored in detailed. Initially the Decision Support System along with their types, advantages, disadvantages and traditional approaches for DSS were discussed. Thereafter Data Mining & KDD, Data Mining with DSS, Educational Data Mining and introduction to machine learning with their application were also covered. Moreover the machine learning techniques particularly for students’ performance prediction were also presented in this chapter. Additionally, the literature review section presented a number of frameworks and methods which were used to develop DSS related to students’ performance prediction. In addition to all above, the problems in existing DSS approach as well as problems in existing student’s performance modeling were also addressed. The subsequent chapters discuss the efforts that have been made to overcome these limitations during the entire research. The detailed methodology for this research is presented in the chapter 3.