CHAPTER 2
REVIEW OF LITERATURE

2.1. INTRODUCTION

Wide literature review on academic performance of students and its prediction by using performance models was carried out. But, it was noticed that limited research investigations have been executed not only on the factors that are influencing the academic performance of the students at high school/higher secondary level but also on the prediction of the academic performance of the students using different classification algorithm in data mining. So, the specific review of literature pertaining to the present investigation have been listed in this chapter with the categorizations on reviews related to the factors that influence the academic performance and reviews with reference to performance models.

2.2. FACTORS THAT INFLUENCE THE ACADEMIC PERFORMANCE
Walters and Soyibo (2001) conducted a study to determine Jamaican high school students' (population n=305) level of performance on five integrated science process skills with performance linked to gender, grade level, school location, school type, student type, and socio-economic background (SEB). The results revealed that there was a positive significant relationship between academic performance of the student and the nature of the school.

Khan (2005) conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the predictive value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Sunitha and Khadi (2007) studied the influence of socio-economic factors on academic learning environment at home and school. The number of sample students of the study was 240 and the students were from 8th to 10th grade of eight co-educational schools with medium of instruction on both English and Kannada in Dharwar city of Karnataka state. Karl Pearson’s moment correlation test was used to measure the relationship among the factors like home environment, school environment, academic achievement and socio-economic status of the family of high school students. It was noticed that students of low socio-economic status studied in schools with Kannada as medium of instruction, whereas the students with relatively higher socio-economic status studied in schools with English as medium of instruction. One of the results of their study revealed that students from English medium schools were significantly higher on academic achievement than that of the students with Kannada as medium of instruction. It was concluded that medium of instruction would reflect in the academic performance of students in the region of study.

Hijazi and Naqvi (2006) conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab University of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance" was framed. By means of simple linear regression analysis, it was found that the
factors like mother’s education and student’s family income were highly correlated with the student academic performance.

De Lange and Mavondo (2004) analyzed the gender, motivational differences and students’ learning approaches. The study has been conducted on 246 business students studying in Australia through open learning and model has developed using a structural equation model. The model was subsequently exhibits that there are significant differences between males and females in learning strategies.

Tho (1994) studied the evidences on the determinants of the performance of the students in the introductory accounting course conducted by University of Malaya. It was reported that female students could outperform male students in accounting, as they are believed to have higher work needs. Also this study reveals that, high-school accounting, mathematics and grades obtained in high-school economics are important predictors of performance of the students.

Maric and Petro (2004) assessed the gender differences of secondary school in various cognitive-motivational variables and in performance attained in school subjects of Language Arts and Mathematics. A sample of 521(236 male sand 285 females) students between the ages of 14 and 18 was chosen from the second cycle of mandatory secondary education in the province of Jeon. Student’s ‘t’ test was used to study the academic difference between male and female of secondary students and it was reported that existence of gender difference in the academic performance of students.

Kristjansson, Sigfusdottir and Allegrante (2010) made a study to estimate the relationship among health behaviors, body mass index (BMI), and self-esteem and the academic achievement of adolescents. The authors analyzed a survey data of cross-sectional sample of 6,346 adolescents in Iceland. It was found that lower BMI, physical activity, and good dietary habits were all associated with higher academic achievement. Poor dietary habits negatively influenced self-esteem and academic achievement, and self-esteem was negatively influenced by increasing levels of BMI.

An entire volume of the journal Educational Leadership (Scherer, 2000) focused on children in public schools and the needed emphasis on the relationship between healthy bodies and minds. In this volume, Scherer (2000) offered the viewpoint that "the topic of health involves scientific challenges, ethical conflicts, (and) social dilemmas". In reality, providing students with quality instruction in healthy lifestyles has been and will continue to be a challenging task. It was
observed that students could be taught that their healthy development would be important when they could receive consistent health-promoting messages from all aspects of the school environment--in class; in the food choices offered; [and] in the opportunities for daily physical activity.

Grissom (2005) investigated the link between fitness levels of students and their academic achievement in Stanford Achievement Test scores. The study was conducted on 884,715 fifth, seventh and ninth grade California school children enrolled in the year 2002. It was presented an evidence of a correlation between students’ fitness scores and their achievement. It was reported that physical fitness levels could impact the health of the child, in turn, promoting higher achievement. This study also indicates that the relationship between fitness and achievement appeared to be stronger for females than males and stronger for higher socio-economic status (SES) than lower SES students.

Dalton, Kathryn, and Melissa (2007) were studied the effect of sibling structure in the educational attainment in United States of America. They have collected data from the Panel Study of Income Dynamics (PSID) and Child Development Supplement (CDS) databases for the period of January 1996 to December 1997. The survey consists of 3563 black and white children in the age group of 0 to 12 from 2380 households. It was noted that children in two-parent households could perform more similarly on age-adjusted achievement tests than that of the children in single parent households. Results from within-family sibling comparisons revealed that first born generally could outperform their younger siblings on age-adjusted achievement tests.

Valerie and Dawn (2009) conducted a study that explores the relationship between birth order, parental attention, and academic achievement. The survey consists of 20 individuals (9 males, 11 females), ranging in age from 13 to 39 years and each one had parents with at least one sibling. Six of the 20 participants analyzed were firstborns, 4 were middleborns, and 10 were lastborns. Each participant completed the full academic school year of 2000–2001 and had a grade average established. Basic demographics were obtained along with information on birth order, parental attention, and academic achievement. It was reported that, parental attention decreases incrementally with birth order -- in other words, that parents give the most attention to their firstborn children, with less attention given to each subsequent child -- and that
as a result of this decreased parental attention, middle and lastborn children display diminished academic performances. More over, lack of parental attention may cause siblings to rely on one another; alternately, parental favoritism and expectations may lead to intellectual competition between siblings.

Taiwo and Tyolo (2002) were studied the effect of early childhood education experience on the academic performance of Botswana grade one pupils with pre-school education experience and their counterparts without such an experience on selected tasks in English language, Mathematics and Science. A total of 120 grade one pupils studied in year 2000 were randomly selected for the study from four selected primary schools in Botswana. The results of their survey study indicated that pupils with pre-school education experience significantly out-performed their counterparts without such experience in all the three school subject areas.

Johnson (1996) conducted a research in 1996 to study the effectiveness of preschool education on Academic Achievement of the 100 Third Grade students of May Community Academic Public School in Chicago. The results of research findings indicated that third graders who had preschool experience had significantly higher reading and mathematics achievement scores than third grades who had not had preschool education.

Tooley and Dixon (2005) conducted a study in 2005 at Delhi, found that children in low-budget unrecognized private schools did better than government school children students, with around 72% higher marks on average in mathematics than government students, 83% higher in Hindi and 246% higher in English.

Malathi and Malini(2006) found that there are significant differences in learning style of higher secondary students was found to be good and there was no significant difference in the learning style of higher secondary students in terms of their class and type of school. A study was conducted in Coimbatore district of TamilNadu in the year 2004 in selective aided, unaided schools. Results of their reveals that, there was significant difference in the learning style between boys and girls studying in higher secondary schools and the correlation was higher between learning style and achievement which could indicate that higher the achievement scores, the better the learning style among higher secondary students.
Another predominant factor as for academic performance of the students are concerned, is nature of schools i.e. whether co-educational or unisex schools. Most of the public schools are co-educational schools throughout India. There is a common phenomenon that, girls fare better in unisex schools than co-educational schools. The results may not apply to other countries given the cultural differences and geographical differences. The response from a few city-based teachers has been quite mixed. While some say that students studying in girls’ schools are less prone to ‘distraction’, others feel that the competition is higher in a co-ed school and so is the learning.

Kamalamani (2001) conducted a research study of the profiles of high and low achievers at Higher Secondary examination conducted in Coimbatore District of Tamil Nadu state. One of the interesting finding was that students with high academic achievement in higher secondary education belonging to Backward Class (BC) had the conducive atmosphere than those belonging to Most Backward Class (MBC), Scheduled Caste (SC) and Open Category (OC)

The impact of Television viewing on academic achievement and mathematical reasoning was examined in a group of 654 higher secondary students comprising 368 boy and 286 girl students from Maharashtra, India. Among the boys and girls, Television viewing had significant negative correlations with academic achievements. It was reported that heavy viewers of television were poor, compared to light viewers, in their academic achievement and mathematical reasoning. But, it was specifically found that Television viewing and mathematical reasoning were negatively correlated among boys only.

A recent survey has been conducted by an Non-Government Organization (NGO) in India - Pratichi Trust showed that the expenses pertaining to education in government schools of Kolkata. It was estimated that an average additional cost of Rs.1000 per annum was spent for private tuition for school children even at the primary level. It was also estimated that the average annual expenditure per child on tuition was more than Rs. 850 for even in the Education Centres (Shishu Shikha Kendras) that cater to less privileged groups.

Moriana, Alos, Alcala, Pino, Herruzo, and Ruiz (2006) studied the possible influence of extra-curricular activities like study-related (tutoring or private classes, computers) and/or sports-related (indoor and outdoor games) on the academic performance of the secondary school students in Spain. A total number of 222 students from 12 different schools were the samples and they were categorized into two groups
as a function of student activities (both sports and academic) outside the school day. Analysis of variance (ANOVA) was used to verify the effect of extra curricular activities on the academic performance and it was observed that group involved in activities outside the school yielded better academic performance.

Bray (2007) in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Sri Lanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

Davis, Akers, Green, and Zartman (2007) conducted a study to identify student related factors that could influence student performance in an introductory class on soils (PSS 2432) at Southwestern Agriculture University in the year 2003. They evaluated the influences of SAT score, percentile rank in high school class, high school chemistry background, major, gender, and academic semester in which the class was taken, with student performance in introductory soil class. By means of correlation and stepwise multiple regression analysis, they identified that, SAT score, percentile rank in high school class, major, and gender were highly correlated to grade in PSS2432.

Most of the research work was carried out only in developed countries and limited research was conducted in developing countries like India and other South Asian countries. On the basis of the findings of the literature, it was found that the academic performance of students was mostly dependent on socio-economic, psychological and environmental factors.

### 2.3 STUDENT PERFORMANCE MODELING

A number of studies have been made on modeling of student performance at various levels of education. Some of the interesting studies related to the present study are listed in the following sections:

Fang and Lu (2009) developed a decision-tree model to predict student performance in engineering dynamics. A decision-tree model was constructed by using student’s score in dynamics as dependent variable and five variables including GPA score of four prerequisite courses namely Engineering Statistics, Calculus I, Calculus II, and Physics as independent
variables. It was found that there was a strong co-relationship between student prior achievements and course performance in Dynamics. The multivariate linear regression technique was employed for assessing the prediction accuracy of student performance of the decision tree approach. The statistical results revealed that decision tree based model predictions were more accurate than the predictions from the traditional multivariate linear regression technique.

Nghe, Janecek, and Haddawy (2007) discussed the accuracy of decision tree and bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two different academic institutions - Can Tho University (CTU), a large national university in Viet Nam, and the Asian Institute of Technology (AIT), a small international postgraduate institute in Thailand. The accuracy of the decision trees were compared and bayesian network algorithms for predicting student performance in the two different case studies with 2-class, 3-class and 4-class of values by using 14 different student related independent variables were studied. It was found that the outcomes of the decision tree model could provide 3 to 12% more accuracy consistently than the results of the Bayesian network model for all type of class values. It was specifically reported that the predictive accuracy of decision tree model was 73/71% for a 4-case class variable {fail, fair, good, very good} and 94/93% for a two-case class variable {fail, pass} at the CTU/AIT respectively.

Ma, Liu, Wong, Yu, and Lee (2000) applied a data mining technique on a real-life application for the Gifted Education Programme (GEP) of the Ministry of Education (MOE), Singapore. The students of GEP were selected when they were in 3rd primary school education and selection was done with a series of tests. The GEP was interested in improving the A-level exam results of their students by selecting the weak students to attend remedial classes based on cut-off marks. The effect of selecting weak students was studied in both traditional method and data mining method. In conventional approach, the cut-off marks of each subject of O-level examination were studied individually, and a subject cut-off mark for attending the remedial class of the subject was imposed. If a GEP student did not do better than a subject cutoff mark in his/her O-level exam, he/she would be recommended to attend the remedial class of the subject. They applied three types scoring function based on decision tree system C4.5 (C4.5-score), Naïve Bayesian technique (NB-score) and Scoring Based on Associations (SBA-score) for allocating students to three different groups using their scores. The SBA could identify weak students very well in the first 20% of top ranking students. The SBA could also identify a small number of weak students in the next 30% and it was found that all the students were good
students in the last 50%. It was concluded that SBA-based scoring method could outperform than the traditional techniques substantially.

Al-Radaideh, Al-Shawakfa, and Al-Najjar (2006) applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University in the year 2005, Jordan. They used 12 predictive variables and a 4-class response variable for the model construction. Three different classification methods namely ID3, C4.5, and the NaïveBayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models with the predictive accuracy of 38.33% for four-class response variable.

Kotsiantis, Pierrakeas, and Pintelas (2004) applied several data mining algorithms to predict the performance of computer science students from distance learning stream of Hellenic Open University (HOU), Greece. A representative algorithm was used for each one of the most common machine learning techniques, decision-trees, bayesian-nets, perceptron-based learning, instance-based learning and rule-learning. It was found that the efficiency of machine-learning algorithms could predict student performance with satisfying accuracy before the final examination. For model construction, they used two groups attribute sets viz demographic attributes and performance attributes. The demographic attribute set consisted of seven independent variables students' sex, age, marital status, number of children, occupation, computer literacy and the association between students' jobs and computers. The performance attribute set comprised four pair of attributes- students’ marks on the written assignments and their presence or absence in face-to-face meetings. Representative of each of the techniques C4.5, NaïveBayes, 3-NN, RIBBER and WINNOW were used to predict class (pass or fail) of each student. The predictive accuracy was calculated for each subset of attributes starting with demographic attributes and only one of the attribute of performance attribute set was added at a time in incremental manner and predictive accuracy was estimated against each subset of attributes. It was reported that high predictive accuracy was achieved by using the Naive Bayes algorithm and the RIPPER algorithm.

Cortez and Silva (2008) attempted to predict secondary school student performance in two core classes of Mathematics and the Portuguese language who appeared in 2006 examination at Portugal, Europe. The prediction models were developed for these two core classes using four techniques namely decision trees (DT), random forest (RF), neural networks (NN) and support vector machines (SVM). The database was constructed from two main sources of data viz. school reports related to paper sheets and attributes like three period grades and number of school absences and questionnaires with closed end questions related to
several demographic features like mother’s education, family income, social/emotional (e.g. alcohol consumption) and other school related variables that were expected to affect student performance. These two separate databases were integrated into two datasets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes with final-grade (G3) as class variable. All four types of models were evaluated against class variable G3 with three goals such as (i) binary classification (pass/fail); ii) five-level classification (I very good or excellent, II – good, III - satisfactory, IV – sufficient and V – insufficient/ fail); and iii) regression, the G3 value (numeric value that ranges between zero and twenty). They used Percentage of Correct Classifications (PCC) was used as a measure for the evaluation of DT, RF, NN and SVM based prediction models, while in regression the Root Mean Squared (RMSE) was used. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

In 2003 (Minaei-Bidgoli, Kashy, Kortemeyer, and Punch, 2003) online student grades from the Michigan State University were modeled using three classification approaches (i.e. binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score). The database included 227 samples with online features (e.g. number of corrected answers or tries for homework) and the best results were obtained by a classifier ensemble (e.g. Decision Tree and Neural Network) with accuracy rates of 94% (binary), 72% (3-classes) and 62% (9- classes).

Pardos, Heffernan, Anderson, and Heffernan (2006) employed the use of Bayesian Networks not only to model user knowledge but also to predict student performance in 8th grade Mathematics in e-learning and e-assessing system. The Massachusetts Comprehensive Assessment System (MCAS), which was a state administered standardized test that covered English, Math, Science and Social studies for grades 3 through 10, test would be of particular importance to students and teachers in Massachusetts because students must pass the test in order to graduate high school. Four different skill models were considered; one that had unidimensional, WPI-1, one that had five skills, WPI-5, one that had 39 skills, WPI-39, and our most fine-grained model that had 106 skills, WPI-106. A separate network was created for each skill model. The top layer nodes represented knowledge of a skill that was set to a prior probability of 0.50. This model was simple and assumed all skills were as equally likely to be known prior to being given any evidence of student responses. Once the network was presented with evidence, it could quickly infer probabilities about what the student would know.
The bottom layer nodes were the question nodes with conditional probabilities set ad-hoc to 0.10 for the probability of answering correctly without knowing the skill, guess, and 0.05 for the probability of answering incorrectly if the skill would be known, slip. The intermediary secondary layer consisted of AND3 gates that, in part, allowed only specifying a guess and slipping parameter for the question nodes regardless of how many skills were tagged to them. The results showed that the WPI-39 had the best accuracy with an error of 12.86% that translated to a raw score error of 3.73. The finest-grain model, the WPI-106, came in second followed by the WPI-5 and finally the WPI-1 and it was concluded that the fine grain models were the best for predicting the external test.

Bekele and Menzel (2005) used Bayesian networks to predict performance of high school students of Addis Ababa University. Their model categorized students into three categories namely below satisfactory, satisfactory, and above satisfactory. The researchers studied the automatic prediction of performance by using personal, social and cultural features that might be influenced. The goal of predicting the student performance study was three fold: (i) to identify major attributes which help to effectively predict performance; (ii) to develop an instrument for measuring those attributes; (iii) to investigate the application of belief network tools in classifying students into categories of performance. Practically, it was difficult to model the exact values of attributes of a student with respect to some attributes when those attributes are uncertain. The uncertainty would transfer to the prediction accuracy, which might also result in poorly adapted classification of performance. The Bayesian approach was employed to tackle this problem since it was found to be a clear and manageable language for expressing what was certain and certain about. Empirical experiments were conducted on the prediction of performance with a data set of high school students containing 8 attributes. The names and possible categories of each attributes were gender (male, female); group work attitude (positive, indifferent, negative); interest for math(interested, indifferent, uninterested); achievement motivation (high, medium, low); self confidence (high, medium, low); shyness (extrovert, medium, introvert); English performance (above satisfactory, satisfactory, below satisfactory) and mathematics performance (above satisfactory, satisfactory, below satisfactory). The results showed that Bayesian model could predict that about 64% of the records were classified correctly out of the 172 total records. Students’ negligence to show their true feelings for some of the items in the questionnaire and the sample size used might be attributed to the prediction errors. A closer look showed that 78% of the below satisfactory performance category was correctly classified; only 34% of the satisfactory performance
category was correctly classified and 54.2% of the above satisfactory performance category was correctly classified.

A study was made by Ramasubramanian, Iyakutti and Thangavelu (2009) to analyze student information system (SIS) database using rough set theory to predict the future of students. They used three kinds of attributes of a student namely academic (A), non-academic (NA) and human behavior relationships (HBR). The value of the academic attribute was calculated on the basis of student performances in theory, practical, attendance, participating seminars, paper presentations, interactions, reading books and participating department activities. The value of the non-academic attribute was calculated on the basis of student performances in Sports, NSS, NCC, YRC and Social activities. The value of the Human behavior relationships was also calculated on the basis of students’ relationship with their teachers, fellow students, public, family members and their performances in hobbies and entertainment. They performed basic operations of rough set theory like lower-approximation, upper-approximation of all boundary decisions on a hypothetical SIS dataset.

Kalles and Pierrakeas (2006) constructed student performance prediction model based on a combination of machine learning approaches, namely decision trees and genetic algorithms. They evolved populations of tree representations according to a fitness function that allowed for fine-tuning decision tree size vs. accuracy on the training set. At each time-point a certain number of decision trees (population) was generated and sorted according to fitness value. Based on that ordering, certain transformations (genetic operators) were performed on some members of the population to produce a new population that would be subject to the same cycle until a predefined number of generations was reached (or no further improvement was detected). They built GATREE system by using the GALIB library. They compared the accuracy of GATREE induced trees with conventional classifiers already made by Kotsiantis et al., (2004) and it could provide good accuracy estimation than conventional classifiers.

To sum up, many studies have identified different factors that influence the academic achievement of students and various prediction models were proposed in different context.