INTRODUCTION

Education is a process of imparting or acquiring knowledge and habits through instruction or study and this process results in desirable changes in the behavior of human beings. It provides the skills to individuals to become self-confident, self-reliant and self-sustained and inculcates buoyancy to face challenges in all walks of life. It enhances the ability of individuals to manage health problems, improve nutrition and childcare, and prepare for the future. It sustains the human values which contribute to individual and collective well-being. It is the key which allows people to move up in the world, seek better jobs, and ultimately succeed in their lives. It is essential for eradicating poverty and it allows people to be more productive playing greater roles in economic life and earn a better living. It is worth mentioning that education forms the basis for lifelong learning in the context of human development and it is one of the fundamental requirements of democracy. It makes the people to aware of opportunities and rights that in turn result in more responsible and informed citizens. These citizens can have
a voice in politics and society, which is essential for sustaining democracy and so education is the only tool which takes the country to greater heights.

1.2. EDUCATIONAL RESEARCHES

As education provides multifaceted developments of human beings, it is imperative to conduct researches in education for its effective implementation for the benefits of end users. One of the major goals of educational research is to investigate behavioral patterns in pupils, students, teachers and other participants in schools and other educational institutions. In fact, educational researchers like other social science researchers use a variety of techniques which can be broadly summarized as well as categorized in to two forms of methods viz. qualitative and quantitative research methods. Both these research methods are used in the fields of natural science, social science and technology, though these methods are differ to each other in all aspects.

1.2.1. Quantitative Method of Research

Quantitative methods are used to gather large volume of data, information dealing with numbers and anything that is measurable and discovering underlying meanings and patterns of relationships, including classifications of types of phenomena and entities. The objective of quantitative research is to develop and employ mathematical models, theories and / or hypotheses pertaining to natural phenomena. The process of measurement is central to quantitative research, as it provides the fundamental connection between empirical observation and mathematical expression of quantitative relationships. In this connection, quantitative research can be viewed as the systematic scientific investigation of quantitative properties and phenomena and their relationships.

The rapid growth of databases in many disciplines has overwhelmed the traditional interactive approaches to data analysis and created a new generation of tools called data mining, critical to intelligent and automated data discovery. A data mining is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques. The most appealing aspect of data mining is nothing but the transformation of complex statistics into usable visualization charts that convey large amounts of information to a user in the most meaningful and transferable forms. This type of information
transfer helps to reduce time spent with data and increases how data can connect to models of service delivery.

Guttentag and Clark (1999) have documented the information overload in a variety of existing education settings all over the world. Their findings sound an alert that schools drown with the variety of daily transactions pertaining to students, teachers and administrators of educational institutions. Student and teacher information mainly includes his / her historical data, demographic data, unique family data, health records, extra curricular evidences, standardized test scores and grades. Administrator information mainly includes data histories with details such as the most successful students, the most successful teachers and the most successful classes, pedagogy, schedules, or other variables like salary, educational attainment, certification and professional development of their customers in addition to regional, school-based, and student domains. As the number of variables pertaining to students, teachers and administrators’ increases, the intricacies of determining the trio namely administration, teaching and learning becomes increasingly complex. Under such conditions, data mining can provide improved situation awareness for all that are involved in educational institutions.

The data mining is effectively utilized in business and medical sectors to accurately identify and predict the most effective way to deliver respective services. The researchers and administrators in education (Luan, 2001) can find larger and wider applications for data mining than their counterparts in the business and medical sectors. This is mainly due to the nature of duties of academic institutions that are intensively pertaining to data mining process. These duties include teaching that concerns with the transmission of knowledge, scientific research that relates to the creation of knowledge, and institutional research that pertains to the use of knowledge for decision making. So, the data mining techniques have great promise for providing anticipatory guidance in teaching and learning in the development of models critical for promoting student academic achievement and they can be effectively used for applying knowledge-discovery principles in the field of education. It is worth mentioning that there are additional legal and ethical requirements for teachers and support service personnel to address issues of universal access and individual accommodation, while data mining is applied at all levels of education.

Specifically, in the field of education, data mining can be used for forecasting and proactive decision making. Tsantis and Castellani (2001) advocate the use of data mining as a feedback mechanism for continuous improvement of the learning environment. Using data mining techniques, program administrators can identify and monitor categories of students in
need of special attention, predict possible outcomes, and make predictive decision before problems become critical.

1.2.2. Qualitative method of research

Qualitative research is a free-form technique that is used to gain insight into the underlying issues surrounding a research problem. It is based on the opinions and non-statistical feedback that are rooted in people's feelings, attitudes, motivations, values, and perceptions. It is an empirical research in which the researcher explores relationships using textual, rather than quantitative data. Observation, case study, and ethnography are considered as different forms of qualitative research.

. The qualitative method investigates the why and how of decision making, not just what, where, when. Hence, smaller but focused samples are more often needed, rather than large random samples. It is to be noted that the results are not usually considered generalizable, but are often transferable.

Qualitative approaches have the advantages of allowing for more diversity in responses as well as the capacity to adapt to new developments or issues during the research process itself. It is worth mentioning that qualitative research is expensive and time-consuming for its execution, but many fields of research employ qualitative techniques that have been specifically developed to provide more succinct, cost-efficient and timely results.

1.3. INDIAN EDUCATION SYSTEM

Indian Education System dates back to Ancient India imparting formal education under the Gurukul1 (In Sanskrit, guru means "teacher" or "master"; kul means domain, which is derived from the word kula, "extended family") system. In this system, young boys passing through the Brahmacharya stage (which is a period of spiritual education in the traditional scheme of life in Hinduism that takes place during the teenage years) of life had to stay at the Guru or the teacher's home and complete their education. Although the ancient system of education has produced many geniuses and still a major area of research, it was hardly egalitarian. At this juncture, the British administrators brought a revolution in the Indian education system. It is worth mentioning that the same education system is being adopted till now through out the country.

1.3.1. Levels of School Education in India

1 http://en.wikipedia.org/wiki/Gurukul
The educational structure in India is generally referred to as the **Ten + Two + Three** (10+2+3) pattern. School education is a two-tier system, the first ten years covering general education followed by two years of higher secondary education. The general education covers primary and secondary education and the former is divided into two stages. While the first five years constitute the primary stage (standards I to V), the next three years form the upper primary stage or middle school (standards VI to VIII). After finishing upper primary stage, there are two years of schooling to the pupils so as to complete ten years of education (standard X) for taking the Secondary School Certificate (SSC). They enter higher secondary schools or junior colleges at the minimum age of 16 or above and complete a further two more years of education (standards XI and XII). Public examinations are conducted at the end of standard XII either by the state or central board that leads to the award of the Higher Secondary School Certificate (HSC) and marks at the end of school education. In **+3 stage**, which involves college education, the student goes for higher studies in his chosen field of subject. Alternatively, the students can prefer and join professional courses which last four-year duration after their higher secondary education.

In India, free and compulsory education to all children up to the age fourteen is constitutional commitment. Through the ban on child labor, the constitution of India provides for absolute legal obligation on children to attend school or on their parents to send them to school.
1.3.2. Academic Performance of Students in School Education

According to Tuckman (1975), academic performance or excellence is used to label the observable manifestation of knowledge, skills, concepts, and understanding and ideas. In educational institutions, success of the student is measured by academic performance, or how well a student meets standards set out by an educators and the institution itself. As career competition grows ever fierce in the working world, the students are the focus of attention of parents, family members and teachers those who believe that good academic results will provide more career choices and job security. Although education is not the only road to success in the working world, much effort is made to identify, evaluate, track and encourage the progress of students in schools. By providing due considerations to these existing facts, it is felt that there is an urgent need for devising a method of measuring academic performance of students in order to create plans for improvement during their course of study.
In this study, the definition of the academic performance is restricted by the author and it is defined as the percentage of marks obtained by a student at his final examination of the higher secondary course that is conducted by the Government of TamilNadu. If the academic performance of the students is forecast during the higher secondary course itself, it will be helpful to identify slow learners to whom additional academic care is to be given. It will also be helpful to encourage moderate and effective learners for their better academic performance.

1.3.3. Factors that Influence Academic Performance of Students

The main goal of education has always been better student academic achievement. In past decades, researchers and educators have conducted many researches to determine the factors that influence (positively or negatively) the student achievement in their academic track. It was reported that measurement of academic performance of student would be challenging, since student performance would be the product of socio-economic, psychological and environmental factors.

Examinations occupy a unique position as a measure of assessing the academic performance of a student. In fact, the performances of students in the examination mainly rely on three factors namely demographic and academic-environment and socio-economic factors.

1.3.3.1. Demographic Factors

Demographic factors are those relating to personal characteristics such as age, gender, body mass index (BMI), food habits, disabilities, type of family system, living area, and sibling structure. Now, the demographic factors that influence the student academic performance are highlighted briefly.

- **Gender** – Gender is one of the personal variables that have been related to differences found in motivational functioning and self regulated learning. Different researches (Khan, 2005., Lange and Mavondo, 2004., Tho, 1994) demonstrate the existence of different attributional patterns in boys and girls. Many studies (Robinson and Nichole,2003; Maric and Petro, 2003) show that girls tend to give more emphasis to effort when explaining their performance and boys appeal more to ability and luck for their academic achievement. The presence of gender differences with respect to performance may have implications on students’ career choices, course enrolment, and the use of knowledge, etc. Student gender can affect degree of motivation and thus academic performance.
Body Mass Index - The body mass index (BMI), or Quetelet index (Keys, Fidanza, Kimura and Tailor, 1972) is a statistical measurement which compares a person's weight and height. Though it does not actually measure the percentage of body fat, it is a useful tool to estimate a healthy body weight based on how tall a person is. Due to its ease of measurement and calculation, it is the most widely used diagnostic tool to identify weight problem within a population including: underweight, overweight and obesity. BMI is defined as the individual's body weight divided by the square of his height. The formulae universally used in medicine produce a unit of measure of kg/m$^2$.

Overweight during childhood and adolescence increases the risk of developing high cholesterol, hypertension, respiratory ailments, orthopedic problems, depression and type-two diabetes as a youth. American Association of School Administrators (Arlington, 2006) indicates that “healthful living is as important to quality of life as academic achievement”. Hence, it is evident that health index can be used as a matrix for measuring academic performance of students.

Sibling Structure. There has been a long-standing fascination with exploring associations between sibling variables, such as the number of children in a family and a child's birth-order position in the family, and children's academic achievement. Studies (Dalton, Kathryn, and Melissa, 2007 and Valerie, and Dawn, 2009) reveal that these sibling variables have small but significant inverse associations with academic measures of achievement.

Student Community/Caste. Adolescence in India is shaped by various contextual elements like religion, caste, gender, class etc. Even though, constitutionally, India defined itself as a secular state, religion and caste are deeply entrenched in the identity of Indians across ages and play a direct or indirect role in the daily lives of young people. In terms of social status, the Indian population can be grouped into five categories: Scheduled Castes (SC) or Dalits, Scheduled Tribes (ST) or Adivasis, Most Backward Classes (MBC), Backward Classes (BC) and Others (OC). This caste system with these five categories is rooted in educational system.

1.3.3.2. Academic Environment factors

---

Academic environmental factors are those directly influencing academic performance of the student at higher secondary level which includes primary education, type of elementary education, type of secondary education, type of school, location of school, medium of instruction, private tutoring, marks obtained at secondary level and type of group opted at higher secondary level and extra-curricular activities.

- **Preschool education:** Preschool education is the provision of education for children before the commencement of statutory education, usually between the ages of two and five, dependent on the jurisdiction. Preschool is also known as nursery school, day care or kindergarten. The first five years of life are a time of enormous growth for children. The cognitive, social, and emotional skills that children develop during their early years are essential building blocks for their entire educational lives. The pace of learning, however, depends on whether and to what extent they encounter and engage in supporting environments. Taiwo and Tyolo (2002) show that, there is a statistically significant difference between the achievements of the children who had attended preschool and the children who had not.

- **Type of School:** In India, there are two types of schools namely government schools and private schools offering higher secondary education. Most of the government schools are located in rural and semi-urban areas, whereas the non-government schools are running in semi-urban and urban areas. Among the non-government schools, majority of the schools are run by private managements without state government fund. Private schools are more effective in imparting learning and private school students generally outperform their public school counterparts in learning achievement. The pupil teacher ratios are much better in private schools and more teachers in private schools are female. Studies carried out in India all share the common conclusion that private-school students outperform their government-school counterparts.

- **Private Tuition.** Private tuition, or "shadow education", has become a prominent feature in many countries internationally. A number of studies reveal that such tuition is associated with improved examination performance and enhanced access to tertiary education. Elsewhere, in all cities and towns across India, children of school-going age regularly attend private tuition for their academic improvement. Even among poorer families, there are tremendous pressures on parents to engage in private tuition once the child starts doing not so well in school. The dominance of private tuition may reflect a peculiar academic culture, whereby competitive pressure and high aspirations combine
to create a milieu in which it is seen as not only the norm, but even as a minimal requirement for any kind of academic achievement.

- **Nature of Institution.** Another predominant factor for academic performance of the students is the nature of schools i.e. whether co-educational or unisex schools. Most of the government schools are co-educational schools throughout India. There is a common phenomenon that, girls faring 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.

- **Prior-Academic performance.** Previous academic performance is identified as a predictor of higher secondary school performance. It is to be noted that two types of secondary education are offered in the schools, which are known as *state-board* pattern or *matriculation* pattern. Most of the unaided private schools are offering matriculation type of secondary education, whereas government, government aided and private institutions are following state board secondary education. As the matriculation pattern is better in many academic aspects than state-board pattern, the students' hail from private institutions with matric pattern at tenth level may perform well at higher secondary examination also. Hence, academic performance of the student at secondary level is one of the indicators for the academic achievement at higher secondary level.

1.3.3.3. Socio-Economic factors

A family's socioeconomic status is based on family income, parental education level, parental occupation, and social status in the community. Families with high socioeconomic status often have more success in preparing their young children for school because they typically have access to a wide range of resources to promote and support young children's development. They are able to provide their young children with high-quality child care, books, and toys to encourage children in various learning activities at home. Also, they have easy access to information regarding their children's health, as well as social, emotional, and cognitive development. In addition, families with high socioeconomic status often seek out information to help them better prepare their young children for school.

Thus, Students’ academic performance hinges on diverse factors like personal, socioeconomic, psychological and other environmental variables rendering the prediction of students’ performance uncertain. It is worth mentioning that every year several thousands of students
come out of the higher secondary education generating intense competition in getting the desired courses and only a few are able to go to for further courses of their choice in institutions of their liking. This necessitates prediction models to forecast the student performance and also assist the students who lack the desired requirements to perform better.

1.4. DATAMINING

Data Mining refers to the extraction of implicit, previously unknown, and potentially useful knowledge from data (Witten and Frank, 2005). Data mining is an iterative process that typically involves the following phases:

- **Problem definition:** A data mining system starts with the understanding of the problem. Data mining experts and domain experts work closely together to define the objectives and the requirements of the system.

- **Data exploration:** Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In this phase, traditional data analysis tools, for example, statistics, are used to explore the data.

- **Data preparation:** Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In this process, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase.

- **Modeling:** Data mining experts select and apply various mining functions because we can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.
Evaluation: Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

Deployment: Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

1.5. MODELLING

A model is a pattern, plan, representation, or description designed to show the main object or workings of an object, system, or concept. It also in some cases used as prediction of future outcomes. Models are often used in both quantitative analysis and technical analysis. The formal idea of a model to represent the relationship between the variable to be predicted (dependent variable) and the relevant predictor information is simple, which is represented below:

\[ \text{Dependent Variable} = \text{Function of Predictor Information} + \text{Error} \]

where Dependent Variable represents the response being measured, Function of Predictor Information is the function of predictor variables, and Error is an error variable with zero mean and constant variance.
Modelling is an essential and inseparable part of all scientific activity, and many scientific disciplines have their own applications in specific types of models. Scientific modelling is the process of generating abstract, conceptual, graphical and/or mathematical models. Science offers a growing collection of methods, techniques and theory about all kinds of specialized scientific modelling. The major models that are used for scientific modeling are briefed in the following sections:

**Predictive models** analyze past performance to assess how likely a customer or a student is to exhibit a specific behavior in the future in order to improve the performance. This category also encompasses models that seek-out subtle data patterns to answer questions about customer or student's behavior. Predictive models often perform calculations during live transactions. For example, predictive analytics encompasses a variety of techniques from statistics and data mining that analyze current and historical data to make predictions about future events.

**Descriptive models** quantify relationships in data in a way that is often used to classify customers or students or prospects into groups. Unlike predictive models that focus on predicting a single customer or student behavior, descriptive models identify many different relationships between students or customers. Descriptive models do not rank-order customers or students likelihood of taking a particular action the way predictive models do. Descriptive models can be used, for example, to categorize students by their inherent characteristics. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.

**Decision models** describe the relationship between all the elements of a decision like the known data (including results of predictive models), the decision and the forecast results of the decision in order to predict the results of decisions involving many variables. These models can be used not only in optimization, but also in maximizing certain outcomes while minimizing others. Decision models are generally used to develop decision logic or a set of business rules that will produce the desired action for every customer or circumstance.

Model selection is a problem in statistics, machine learning, and data mining. Given training data consisting of input-output pairs, a model is built to predict the output from the input, usually by fitting adjustable parameters. Many predictive models of data mining have been proposed to perform such tasks, including linear regression analyses, neural networks,
decision-trees, and kernel methods. Finding methods to optimally select models, which will perform best on new test data, is the main object of all researches. In fact, the performance prediction challenge is connected to model selection because accurate performance predictions are good model ranking criteria and predict how well it will perform on a large test set.

1.6. STUDENT PERFORMANCE MODELING

The performance model computes the competence of students based on their academic performance. The outcome of the performance model will be useful to upgrade the academic achievement of students through individual care, adequate assistance and additional concern.

Many studies have been performed to identify the factors that influence the student performance, albeit only a very few studies have come out with prediction models. Especially, accurate prediction of academic performance of student at higher secondary level is very much desirable to provide a student with the necessary assistance in the learning process. In this connection, a performance model, which has twin stages of processes namely identification of the best predictive variables among the set of variables used in our study by means of applying different variable selection techniques and prediction of the academic performance of the higher secondary students using decision tree, bayesian net, neural network and rough set theory, is developed in the present investigation. This model is validated and its accuracy is found in the present investigation.

1.7. LIMITATIONS OF THE STUDY

The present investigation has some limitations and they have been listed here for reference.

The geographical area of coverage of the investigation is limited to five educational districts of the state of Tamilnadu in India. It is to be noted that nine different types of schools were chosen on the basis of their existence (urban, rural, semi-urban), nature of financial assistance (aided and private schools) and medium of instruction (Tamil (regional language) and English (national language) along with the sex of students (single-sex and co-educational) in each educational district. This is mainly due to the manual collection of data, which is restricted only in working hours of schools, and hence the limited number of samples in the
present investigation. So, the student performance data is limited to 30 schools and the generalization is made on the basis of collected database from these schools.

The collected dataset is limited to 6000, but only 5784 datasets have been used in the construction of model. This is mainly due to filling up of missing values and transformation of values in one form into another during the course of data preprocess from the raw data. So, the datasets are limited to 8893 and the generalization is made on the basis of used database for model construction and its subsequent performance.

The number of variables utilized for model process is 35. The nature of variables includes gender, age, height, weight, eye-sight, community, physical growth, disabilities, food habit, living area, family size, sibling structure, type of family system, mode of transportation, availability of vehicle at home, pre primary education, type of elementary education, type of secondary education, marks obtained at secondary level, medium of instruction, private tuition, type of school at HSC level, location of school, sports personal, care of study at home, parents education, profession and their income. All the generalizations are made on the basis of these input variables and outcome of the model and this can be considered as the last limitation of the present investigation.

1.8. OBJECTIVES OF THE STUDY

The academic achievement of higher secondary school education in India is a turning point in the life of any student, as it serves as a very important link between the higher and higher secondary education of students. But, there are determinants like demographic, academic and socio-economic factors of students that restrict the students' performance. This necessitates the need for some forecasting systems to predict the academic performance of students at plus two examinations. This is an attempt made first time in this aspect, which is mainly devoted to design and develop a prediction model by taking into account variables pertaining to the Indian society, for Indian educational system.

In this connection, the following objectives are framed for the present investigation:

- Identification and collection of the variables that are related to the academic performance of higher secondary students in Indian context
Investigation on the highly correlated variables that are influencing student academic performance predominately by means of different feature selection of data mining and proposition of suitable performance prediction model with reasonable predictive accuracy

As the present investigation pertains to the model prediction of the performance of students in higher secondary education, it will be useful to parents, educators, academics, and policy makers. It has high social relevance due to its linkage with academic achievement of students and hence the development of the society.

1.9. THESIS STRUCTURE

The present thesis consists of five chapters and the structure of the thesis is briefed here for reference.

Chapter 2 provides an overview on highly potential variables which are influencing the academic performance of the student both at school level and college level and show how these variables are correlated both negatively and positively. Moreover, an overview on various student performance prediction models proposed by different authors using different classification algorithms namely Decision Trees, Induction rules, Naïve Bayes and Rough set algorithms is presented.

Chapter 3 presents various feature selection techniques as one of the most important and frequently used in data preprocessing for data mining. The general overview on Feature Selection in terms of Filter, Wrapper and Hybrid methods is followed with the effect of feature selection techniques on higher secondary student data. Effectiveness of these algorithms are presented in terms different measures.

Chapter 4 deals with various classifier algorithms that are responsible for the predictability of various performance prediction models. The performance of predication models are investigated in terms of predictive accuracy, Precision, Recall, ROC values and other measures for different variable subset generated in different Feature Selection methods.

Chapter 5 draws the main conclusions that are drawn from the predictions of student performance model and discusses the scope of future work.