Chapter 2

Literature Review

2.1 Introduction

Various research techniques render knowledge about variables and their effect on the subjects. The present study focused on the concept of machine learning algorithms that are applied to the data set to bring out the outputs. One of these is the data mining technique (under which falls the classification technique) that helped in forming a different paradigm as concerns the students’ moral-social-psychological growth. These techniques further help in defining their academic career that relates to future prospects as well.

Data mining techniques perform their job to bridge knowledge gaps in educational settings. As a result of improved speed, efficiency and effectiveness of processes can be discovered using various data mining techniques. Therefore, this enhancement can take a lot of rewards in higher educational settings, such as, bringing most of the efficiency in educational settings, cutting down the drop-out rate in students, boosting promotion rate of students, improving retention of student’s and enhancing improvement of educational ratio, increasing student success rate, promoting learning outcomes of students, cutting down the cost of system processes. To accomplish the quality improvement for above-mentioned points, it is need of the hour to mine data sets that help in providing the necessary information and insights for the policy makers in the higher educational settings to maintain the quality.

2.2 Data Mining in Various Fields

Data mining is an interdisciplinary research area with developing techniques to extract interesting information from the voluminous data [35]. Presently DM techniques are rapidly growing in areas like neural network, pattern recognition, spatial data analysis, image databases, and many applications like education, retail, banking, business, economics and bioinformatics [26]. In banking, Data Mining is used to build risk prediction models in case of loan and mortgages. DM helps in fraud detection, finding defaulters in credit cards and detection of false
transactions. In marketing, it is used to boost up consumer satisfaction, to predict risky consumer activities etc.

DM facilitates in education by giving its hand in predicting the student’s performance so that the student can be saved being a dropout. Data mining is an emerging field in higher education, called Educational Data Mining (EDM). EDM is concerned with data set that is generated from diverse educational settings and developing methods to explore this unique data [80] a survey from.

Figure 2.2 shows the various fields where Data Mining is applicable in practical terms and users are getting most of out it.

![Data Mining in Various Fields](image)

**Figure 2.1: Data Mining in Various Fields**

One of the most essential parts of the educational setting is its quality. Its concern is to enhance education as well as workings of an organization related to these educational settings. Although various schemes by the government have provided a better platform to the students to grow the impending attributes concludes in demoralizing the cadets at various times.
2.3 Educational Data Mining (EDM)

As the growth of technology increases, accomplishing the highest level of quality in higher education settings has become easy. Finding the unknown variable has become easy and can be utilized for important and valuable information. Data mining technologies can help extract valuable information from educational data set.

Data mining techniques facilitated the researcher gather knowledge about various criterions, the ones that concern both the student and the teacher. Data mining has thus helped in various ways such as:

i. In discovering the retention rate of students in institutions.
ii. In evaluating student data based on different parameters, for example, attendance, assignment submission, study hours spent after college hours, previous semester grades.
iii. In acquiring academic behaviour of students, distance from the residence, distractions interfering in their academic performance, interference of the mentors in their daily routine and so on.
iv. In identifying the main hurdles found in student satisfaction that improve the overall performance of the students including interaction with teachers, problem discussion with classmates and so on.

The process of mining educational data to extract useful information from the voluminous data is generated in educational settings[21]. Its ultimate goal is to resolve issues associated with educational research after analyzing educational data. The large voluminous data generated in an educational setting is helpful in predicting the overall performance of the students and also helpful for mentors to take a timely intervention to improve their results [81]. To achieve better results with good performers it needs to keep track of students’ results so that the prediction of good scores can be done easily.

To better understand beginners and their learning issues, Educational Data Mining (EDM) extracts historical data to obtain useful results and extends computational approaches that combine data and theory together to benefit learners and updates in the existing system.
2.4 Educational Data Mining in Various Educational Systems

In recent times Educational Data Mining appeared as a challenging area in research, researchers from all over the world in various related fields are working on this. The increased demand for information technology in educational settings helped to store the voluminous data of students that helped to reflect on and improve the teaching-learning process. The educational data mining has become a vogue at various international conferences; to quote the establishment of the Annual Research Conference that was established in 2008 in Montreal, Quebec. The educational systems that involve educational data mining are:

- Traditional Classrooms
- Online Learning
- Learning Management Systems
- Intelligent Tutoring
- Adaptive Educational Hypermedia System

These educational systems are depicted in Figure 2.3.

Figure 2.2: Educational Data Mining in various Educational Systems
i. Traditional classrooms work on the criteria of face to face interaction between student and mentor. It helps obtain knowledge about students on various factors such as psychological, behavioural, and their performance in academics. Statistical and psychometric techniques are applied to get these results for the data generated in classrooms [17, 42, 49, 50, 91, 102].

ii. Online-learning and Learning Management System (LMS) works on the student database and log files. Instructions provided by online learning and communication, management, association and reporting tools are provided by LMS [82]. Web Mining (WM) techniques are applied to these databases to extract the results.

iii. Intelligent Tutoring (ITS) and Adaptive Educational Hypermedia System (AEHS) focuses on the special attention for assistance in teaching as required by a particular student [82]. Data mining techniques are applied to the student data collected from various platforms such as user models and log files.

2.5 Stakeholders in Higher Education

Etymologically, the stakeholder is a party interested in a business deal/enterprise that prefers to take risks for a larger benefit. In the education field, stakeholders refer to people who invest in the welfare of students, teachers, staff members, school administration, parents, and thus the society at large. It highlights various preoccupations levelling from the least risk-takers to the highest. The role played by all of the stakeholders essentially forms a base for student-teacher growth. It is often perceived that the stakeholders affect or be affected by the organization’s actions, objectives and policies.

Higher education settings include such various user groups or members. They glance at educational data from diverse perspectives according to their individual task, vision and objectives. The settings also project the diverging motivations, aspirations, anxieties and frustrations of these participants in the educational process. In higher education, users can be broadly classified into various Users/Stakeholders [82] as illustrated in Figure 2.2. The Figure, however, does not highlight the stakeholders of an Indian education system but stakeholders that exist in any educational field at large.
Students/Scholars/Learners: Students are the major concern in this category. They need guidance at the right time to improve their grades. To improve their performance they usually need direct interaction with the teachers/mentors to discuss their problems – academic and personal. Further, the students prefer to have assignments and textbook discussions which is often seen as a major help as regards their performance. The teachers then personalise learning, generate hints, recommend course/courses and so on. Thus, the two-way process enhances the knowledge and academic requirements of the students.

Mentors/Guides/Advisers/Teacher: Their objective is to recommend relevant material to the students, modify teaching strategies according to the students’ requirements, provide remedial classes, and provide personal attention to the students. To get the objective response about their
teaching, the mentors observe/analyse students’ and their behaviour towards studies, detect students that need special attention, predict their academic performance, classify their learning groups, find repetition in their mistakes and suggest corrections. The teachers also act as a guiding source for setting future goals of the students.

iii. Educational Researchers/Course Planners: Course Planners plan and maintain course curriculum, progress student learning, make sure to maintain the effectiveness and efficiency of the learning process, create models for the tutor as well as students according to their needs. Best data mining techniques are recommended for teacher and student according to their task, based on these tasks specific data mining tools are developed to enhance educational purposes. Educational researchers and course planners also offer guiding techniques to the teachers as and when required.

iv. Universities/Colleges/Organizations: These educational settings improve the decision process in higher educational settings, to make a more efficient process to achieve the specified goal, propose valuable courses for learners of each phase, to improve retention and student grades, most cost-effective way to implement procedures, select the students with good grades to admit in the courses to bring laurels to the university.

v. School Administrators/School District Administrators/Network Administrators/System Administrators: To organise best way for educational promotion and institutional resources, develop educational program to present the effectiveness of the distance learning approach, make use of existing resources more effectively, evaluate teacher and curricula and set parameters for improving website efficiency and adapt it to users (optimal server size, network and traffic distribution). The administrators also facilitate new research methodologies in acknowledging the new/advanced study techniques.
vi. Parents: Any educational institution needs to understand the importance and the significant role the parents play in the academic growth of a child. It is important to mention that it is normally the parents who chose an educational institution that leaves an enormous influence on the child.

2.6 Data Mining Techniques

Data mining is a vast research field that provides various methods/techniques that are used in higher education settings. Being a vast canvas, voluminous databases from educational system face a variety of challenges and issues in routine. To achieve the target and to yield the best results, selection of appropriate data mining techniques, tools and algorithms are considered very important [20, 31, 49, 50, 54, 56, 60, 63, 69].

![Data Mining Techniques Diagram](image_url)

**Figure2.4: Data Mining Techniques in Education**

2.6.1 Supervised Learning

Supervised Learning (SL) is a machine learning task that works on the principle of trained data. The training data already have trained examples in it; it consists of the input value and output value. It works on the mechanism of leaning; the instructions fed into the task will give output
according to that training provided to the task. It also helps in the prediction of future actions based on the input value and output value. It is this process that renders the classification of low, medium and high-risk students.

2.6.1.1 Classification

A model can be developed which can infer a single feature of the data (predicted variable) and a mixture of other elements of the data (predictor variables) [5, 10, 51]. Classification techniques are used in educational domain, to withdraw qualitative patterns of previous data set such as classification of students in various categories based on their study pattern, record of teachers who are efficient as compared to other teachers, the different courses offered to students based on their calibre, patterns of students likelihood in courses, discovering student behaviour (fooling the system, missing classes, asking for hints), forecasting and appreciating student academic outcomes, predicting the accuracy and efficiency percentage in their academic performance, predicting the positive and negative influencing factors on academic performance of students, predicting remedial actions for students at-risk, predicting willingness to drop out before their failure.

Hence, assists in early detection of at-risk students and providing them preventive action to take timely action. The data mining application can also help in academic progress, interference, reassign, and prediction. This will help higher authorities of the institution about the category of students who are at risk of not qualifying and not being eligible for higher level before they know it. These predictions help the institute management and mentors to pay more attention to at-risk students and need more academic assistance by providing them with extra classes.

i) Neural Network

It is a kind of network that resembles a human brain and works on the learning principle. There are multiple hidden layers that are responsible to
generate the desired output. Neural Network is not an algorithm but it provides a framework to the group of algorithms to work together to process complex data set. It provides the desired output with some labelling that are designed using machine learning, it recognises patterns and sense data through machine perception. It generates the best possible outputs by adapting varying inputs and there is no need to redesign the output criteria. In neural network, data can flow in multiple directions with greater learning abilities employed to solve complex problems such as learning handwriting and language recognition.

ii) Bayesian Network

It is also referred to as a belief network or decision network. It is a directed acyclic graph that represents the directed direction for the variables on their conditional dependencies. It works on the principle of the probability distribution on the set of variables. Bayesian networks combine principles from graph theory, probability theory, computer science and statistics. In many practical settings, the Bayesian Network is unknown and one needs to learn it from the data provided. The theory has been used for applications in various areas such as machine learning, text mining, signal processing, error-control codes, weather forecasting et al.

(iii) Decision Tree

A decision tree is a tree-like structure that works as a decision support tool and their possible consequences. It is a way to display an algorithm that has conditional control statements. It works on both numerical and categorical data. In Decision Tree each node symbolizes a characteristic (attribute), each connection (branch) stands for a decision (rule) and each leaf signifies a result (categorical or continues value). The goal of Decision Tree is to build noiseless or homogeneous data that classifies into pure nodes.
iv) Support Vector Machine

It operates on numerical attributes but if Categorical Attribute is encountered then it automatically changes categorical data into binary values 0(false) and 1 (true). So, based on SVMs processing, new important factors came to light that put a negative or positive impact on the academic performance of the student. A Support Vector Machine (SVM) is thus a discriminative classifier formally defined by a separating hyperplane.

2.6.1.2 Regression

Regression is a machine learning technique that works on the real and continuous variable such as weight and price. Linear regression is the simplest model that can be used; it fits the data with the best hyper-plane that crosses the points. When it is required to predict the price of the property, marketing and other fields of finance then regression does its job. As these fields have already had the previous data, so based on the previous data the prediction can be done based on this data. For example the number of sales in past, property hike etc.

2.6.2 Unsupervised Learning

Unsupervised learning is a machine learning task that works on the principal of unlabeled data. It draws conclusions and finds unknown patterns in the data set. It is a method of cluster analysis used to find hidden variables. For example dimensionality reduction, feature learning, density estimation, etc

2.6.2.1 Association

Association Mining (AM) helps to discover the relationship between the attributes in a data set with a variety of other attributes. AM is an attempt to find out strongly connected attributes with a single attribute of particular interest and tries to find out which attribute is strongest amongst them. In Association Rule Mining, the objective is to discover the if-then system of the form that explains some set of inconsistent values thus
found, rendering another variable that will definitely have an explicit value.

The AM techniques are used to mine samples of previous students’ historic data in the education domain, like attendance with respect to test score, the association between teacher guidance and student test score, the patterns of previous students’ that opted a number of courses, the association of students to the most suitable subject, the characteristic patterns of previous students who took a particular major, the relationship between exam level and student mark, the relationship between exam level and teacher-class performance, the relationship between academic student performance and teacher satisfaction, association between student course mark, and, time and venue of classes, association between teacher who cancels the class frequently and student test scores, association between teacher background and time and a students’ performance, association of training undertaken with various types of students and performance scores individually and in teams, association of students’ work profiles with the most appropriate project, association of students’ team building and leadership approaches, association of students’ attitude with performance, discovery of curricular associations in course sequences, and discovering pedagogical strategies which leads to more effective/robust learning.

2.6.2.2 Clustering

The purpose is to create clustering of students based on their adopted qualities, individual characteristics and so on [3]. Then, the clusters/groups of students acquired can be used by the instructor/developer to build a personalized learning system, to encourage efficient group learning, to offer adaptive contents, and so on. Cluster analysis or clustering is the distribution of a set of observations into subsets (called clusters) so that observations in the same cluster have some points in common.

In this regard, the following points can be taken care of: the application of clustering in educational settings, collection of similar pattern of student’s data in historic student database, pattern of previously dropped out students, pattern of students’ at-risk, discovery of student behavioural data,
clustering students not at-risk, likelihood towards special courses, popularity of institutes between students.

2.7 Approaches of Data Mining in Student Academic Prediction

As mentioned, Data mining techniques play a vital role in creating predictive models for educational settings. Various studies have been carried out to create models for the purpose of predicting unknown variables. Data mining techniques proved better as compared to mathematical and statistical techniques which used to create models in the educational domain as these are comprehensible and do not require much calculation. Additionally, they offer objective ideologies that form the basis of a teacher-student relationship. The data mining approaches further questions and highlights the risk factors that a high-risk student might encounter in the near future. Further, these techniques start working the moment a pupil enters an education field; incidentally, the work of a teacher/mentor/guide to starts at the same moment.

Abisuga AO et al (2015) conducted a study using CGPA of a student to measure their ability to perform in the programme opted. Secondary School Certificate Examination is considered as qualifying criteria at the entry level. A data of 61 students was collected who had completed their graduation with good CGPA during the time period (1998-2014). This study has inspected the connection between admission requirements and performance in a building technology program for graduate building students. Linear regression was applied to perform the operation [1].

Adejo OW and Connolly T (2018) empirically investigated and compared the utilization of multiple data sources to predict academic performance of students using diverse classifiers and ensembles of classifiers technique. The study compared the performance and effectiveness of ensemble techniques that make use of variety of combinations of data sources with that of base classifiers with single data source. Decision Tree Neural Networks and Support Vector Machine were discussed to build ensembles [2].
Ahmed ABED and Elaraby (2014) is used Decision Tree Method to predict the academic performance of student's based on student's database. They worked on several attributes to predict the final grade of student's. This study helped the student's to progress their academic performance, to discover those students who need special awareness to lessen failing ration and taking required action at right time [6].

Alban M and Mauricio D (2019) performed a systematic literature review on university student drop-out where the authors have planned their study based on few questions. They divided their study based on the questions such as: techniques used for data pre-processing-integration, cleaning and transformation of the information, the factors that affect dropout-personal, academic, social, economic and institutional the techniques used for factor selection-Principal Components Analysis, the techniques used for prediction-ID3, Neural network classifiers, Linear Regression and Logistic Regression, Naive Bayes and Support Vector Machine. The tools used for the study are: WEKA and SPSS Modeler. They concluded that total 16 factors can be the reason of university drop-out [7].

Alharbi Z et al (2016) concentrated on early detection of students that were at a high risk of not completing a good honours degree so that intervention could be proposed. Authors have achieved this goal with rational accuracy using classification models and highlighted the students who were predicted to score low percentage with high probability. Attributes from historic academic performance was considered sufficient to identify low percentage students found up to 46% [8].

Alsaffar (2017) considered two attributes to present past performers achievements in a course. The two attributes Global performance it is the aggregated performance achieved by all the students in the course and Local performance is a condition of a student to enrol in a course. Six algorithms were experimented C4.5, CART Decision Trees, Naive Bayes, k-neighboring, rule-based induction and Support Vector Machines and Rapid miner tool was used to process the data. The study concluded that synthetic attributes progress the performance of classification algorithms [10].
Aluko et al (2018) aimed to explore the worth of using data mining techniques to envisage the academic performance of architecture students based on their previous academic performance. It is the pre-screening of the students before they take admission in architecture course. The study created a SVM model that can be used as decision making tool to select new intakes in the course of architecture [11].

Asif R et al (2016) performed a study for predicting the performance of students’ academic performance of students’ at an early stage; they used high school marks to give admissions to students’ no demographic variable which was used for the prediction. Found indicators for at-risk students’ who needed special attention based on these indicators. Decision Trees were used to create the classifiers to create model. Student’ progression increased in their graduation, created clusters to divide the students’ into categories based on their level of performance. Data Mining Techniques were implemented using Rapid Miner tool[14].

Asif R et al (2017) The authors utilized most appropriate data mining techniques to predict the performance of undergraduate students. Features of their performance was categorised in two aspects: predicting their academic achievements at the end of their degree and other one is to see their development and merging them with the prediction results. Two groups of students were found based on their results that show high achieving students and low achieving students that show their performance in the course. Classification and clustering techniques were explored to achieve the hidden knowledge about student’s educational data. Early prediction can be performed through their result category i.e. low and high grades [16].

Bharadwaj and Pal (2014) used the historic student data to predict the division of student and made use of the classification technique. From the number of approaches available for data classification, the decision tree method was preferred. To predict the academic performance of students at the end of semester, few important attributes were considered such as marks scored in assignment, attendance, and seminar attended/completed, marks scored in class test et al. The survey proved beneficial for both the teachers and students to raise their level of
interaction. The students who needed special attention were identified so that the drop-out rate in student could be controlled and required steps/remedial actions could be taken [24].

Huang S et al (2013) compared mathematical techniques with data mining techniques to create model for the prediction of student’s academic performance in engineering. The models were built on the basis of: high-enrolment, high-impact, and core course that many engineering undergraduates are required to take. These models include multiple linear regression models, the radial basis function Network Model, the Multilayer Perception Network Model, and the Support Vector Machine model. The input for these models includes student's cumulative GPA, grades earned in four pre-requisite courses. The author concluded that Data Mining Techniques are better than mathematical models [45].

Kim KJ et al (2017) explored the Bayesian Network to prove that it can be the best prediction model in data mining. The authors compared three evolutionary algorithms GA, FSGA, and DCGA they used three ensemble search heuristics, i.e., Expert, Greedy and 50C3 to create the best ensemble out of it. Hamming distance was used to measure the similarity and distance between Bayesian Network structures [52].

Liew J et al (2018) they studied sensitive issues related to students data such as interpersonal skills between student and teacher. However it helps in early detection of their temperament towards learning skills. The students go through number of mood changes and behaviour since their childhood, they attain ego in some cases and when they have social and peer pressure. Every person is unique adaptability according to their relicense towards increasing level of competence [59].

Lu OH et al (2018) focused on blended learning that combines traditional classroom and on-line learning performance and facilitates students to accomplish higher learning performance all the way through well-defined and associated approaches that involved online and traditional learning activities. The authors worked upon mixture of online behaviour and traditional factors to predict the early academic performance of students. The authors concluded that the
prediction can be performed within one to six weeks of the starting of the course. They applied MOOCS and OAS enabled calculus course [62].

Pal S et al (2017) the authors worked upon the students’ data collected through reports and questionnaire that helped the authors to analyse the students’ behaviour. From the data authors generalised their view towards students as the alcohol affect the learning ability and memory. The activities after having excessive drinking can harm the students as well as the surroundings in which they are living. Activities for example are: injuries, blackouts, violence, STDs, rapes etc can be the reason of their failure. They predict the affect and harm done to students and used four classifiers to perform this analysis such as BFTree, J48, RepTree and Simple Cart [71].

Superby JF et al (2006) classified the students into three groups based on the personal history of the student, behaviour of the student and different perceptions of the students. They were categorised into: “Low-Risk”: that have High probability of succeeding. “Medium-Risk”: that may succeed in their examination. “High-Risk”: that has High probability of failing or drop-out. Techniques applied on the student data included Discriminant Analysis, Neural Networks, Random Forests and Decision Trees [91].

Tair MM et al (2012) worked on a case study of the undergraduate students to improve their performance in academics. The educational data mining techniques were used to discover knowledge about students’ historic data such as Association Rules, Classification Methods, Rule Induction and Naïve Bayesian Classifier to predict the grade of the graduate student. The students were clustered into groups using K-Means clustering algorithm. The outlier detection was used to detect all outliers in the data. Distance-based Approach and Density-Based Approach were used for the same [93].

**2.8 Approaches of Data Mining in Various Learning Environments**

A student’s moral and social growth takes place in various learning environments– primary being the household and primary educational institutes. Since times immemorial, students received their education in various institutes.
There have been the *gurukul* and western oriented educational institutes; with the passage of time and changes in technology, online learning, commonly known as e-learning, established itself as an important platform of learning.

Feng M et al (2006) promoted the use of ASSISTments (break problem into steps), different aspects of student like performance (Speed, Accuracy, and Help Seeking Behaviour) considered being influential. E-learning system (Integrate Assistance and Assessment), Instructional Support provides good test-beds. Different Tests were performed (MCAS, MAD, R Square, BIC*) on the student’s data. The concept of Scaffolding Questions (breaks down problem into steps to eventually get the problem correct) embedded with the system. Different models were created to implement the data extracted from the students, models for example: Lean Model, Item Response Theory Style, Rasch model were used [36].

Antunes C (2008) pointed out that to build an intelligent tutoring system for student prediction the main constraint is acquiring the background knowledge of teaching strategy and behaviour of the student. Acquiring background knowledge can be solved with sequential pattern mining. In conjunction with constraints and constraint relaxations, it can be used to discover the usual misconceptions or the bug library. The author concluded that context free languages are more expressive than regular. Statistical methods like: T-Test, ANOVA and SPSS were used to process the data [13].

Amershi S et al (2009) discussed about the exploratory learning environments. They presented a data-based modelling framework using both unsupervised and supervised classification methods to build student models. Two different data sources (logged interface and eye-tracking data) were used to analyse the behaviour of the students’. Experiments were carried out by applying framework to build user models for two such exploratory environments: the CSP Applet for helping students to understand an algorithm for constraint satisfaction, and the ACE environment for the exploration of mathematical functions [12].

Romero C et al (2008) investigated the different data mining approaches that can be used to improve the performance of first-year students on the basis of participation in an On-Line forum. It was also discussed how the student performed in the traditional classroom. Classification algorithms compared to
clustering approaches proved better in predicting students’ performance. It was also proposed to discover class association rules on each cluster; advantage using this representation to create rule-based models, these models proved to be robust and reliable[81].

Ramaswami and Bhaskaran (2010) discussed the type of classroom the student is attending i.e. traditional classroom and on-line education. Traditional classrooms are considered better as compared to online classroom, as students have direct interaction with the teacher. On the other hand, online classrooms lack face to face interaction. Students feel lost as they cannot interact with teachers directly. Demographic factors are also discussed that affect the performance of the student such as: qualification of parents, factors influencing the student’s performance such as environment in which the student is living, their awareness about opportunities provided by the institutes nearby [75].

Kuzilek J et al (2015) analysed the objective carried out in an Open University project that was proposed to predict 'at-risk' students in the early years of their graduation. The techniques they used for prediction included Bayesian Classifier, Classification and Regression Tree (CART), k-Nearest Neighbours (k-NN). To identify the students at-risk they created predictive models using machine learning techniques. These models provided the information to the teachers for the performance of students. The authors concluded a number of challenges, such as different course prediction needs, previous presentations and different ordering of assessments which still exist. These challenges often stem from changes in course design and need to be addressed on a course opted by the students [57].

Wolff A et al (2014) carried the latest work the clubbed demographic data with VLE data to predict the performance of student, whether the student is a failure or dropout. The first assessment marks has been considered to be a good predictor for final performance of students. They pointed out the right time of intervention, so that one could find out at-risk at an early stage and offer help to students at the right time. These interventions can improve student’s chances of success. Authors created predictive models using Machine Learning, Bayesian models, k Nearest Neighbours (k-NN), Classification and Regression Tree (CART). The four
sources used to collect the information about students were: resource; books and handouts for the students, forum; is a platform provided to the students to communicate with the tutors, subpage; means of navigating in the VLE environment, content; specification of TMAs and the guidelines to their elaboration [101].

Hlosta M et al (2014) focused on problems faced by students in an open university’s online course. They discussed problems of student retention by exploring two methods in Virtual Learning Environment: GUHA and Markov Chain-Based Graphical Models. Both methods provided useful insights into the students' behaviour during their studies. On the basis of these outcomes the teachers and supporting staff could plan the interventions and establish steps to help them improve their academic performance. They pointed out towards the validity of these models. These models proved good predictors for students who were at-risk of failure. Markov method proved to be good as compared to Graphical Output Provider [43].

Romero C et al (2010) reviewed the Educational Data Mining and it was found that the objective was to analyse the data mining techniques and the type of data used in order to resolve educational research issues. The different research related areas discussed offline education, e-learning and Learning Management System (LMS), Adaptive Educational Hypermedia System (AEHS) and Intelligent Tutoring (ITS). Different mining models were created such as: Association Rules, Clustering, Classification, Sequential Pattern Analysis, Dependency modelling and Prediction have been used to enhance web-based learning environments to improve the degree to which the educator can evaluate the learning process. Decision Trees, Bayesian Models and other prediction techniques have been proposed to address the admission and counselling process in order to assist in improving the quality of education and student performance [82].

Wolff A et al (2013) discussed how the predictive capacity of different sources of data changes as the course progresses (a case study performed on OU). Data Mining Techniques like Decision-Tree, SVM, GUHA, Bayesian Networks, Linear
and Logistic Regression were explored to build predictive models using data from several OU modules. OU offers a good test-bed for the prediction of students in various phases of their academic career. It also highlights the importance of understanding how a student’s pattern of behaviour changes during the course [99].

Wolff A et al (2014) conducted their study using data from VLE combined with student data such as grades in modules, students’ activity on the VLE (logged in response). GUHA (using LIStp Miner) with the VLE were used to generate hypotheses in the form of rules about the factors that contribute in student’s failure. These rules could be applied to new data sets and could help predict accurately which students will fail in their course. Possible factors that affect students’ performance include demographics lower such as economic backgrounds, less formal qualifications and a higher proportion of disabled students. Also, due to lack of face to face contact, students could feel isolated and unsupported by their tutors. Decision trees have been demonstrated to be suitable for prediction, particularly at the start of a module, when there is commonly a high attrition rate. The authors concluded the best predictor was the behaviour of the student that changed during the course [100].

Kabakchieva D (2012) discussed several well-known data mining classification algorithms: a Rule Learner, a Decision Tree Classifier, a Neural Network, Nearest Neighbour, Random Forest Method and Linear Discriminant Analysis. Prediction of student’s at-risk and drop-out rate, by determining factors influencing the achievement of the first-year university students. They categorized students into three classes: Low-risk:- the students with the probability of scoring good marks, Medium-risk:- the students with the probability of scoring average marks but with the help of extra efforts like providing assignments and class test. High-risk: the students with high probability of failure. Implementing predictive models maximized students’ selection and retention rate in education domain [48].

Kamal P and Ahuja S (2017) reviewed the literature and found influential variables that affect the academic performance of the students. The prediction was
based on various influential factors. The authors explored various data mining techniques to build the classification model for student performance prediction in their early years of graduation. Decision Tree was found to be most appropriate for the study [49].

Nasiri M et al (2012) illustrated EDM applications. In this case study they used e-learning and online learning. The data of first year students is collected from LMS used by online education centre of Iran University of Science and Technology (IUST). CRISP has been used to predict GPA of students. The author has discussed various major problems in student’s graduation such as probability of dismissal, capable tools to predict GPA, drop out and other challenges. Regression and Classification (C5.0 algorithm) is used by the author to perform the prediction. SPSS tool used to analyse the results collected from the students [66].

Yadav S et al (2012) used the approaches of DM: Classification, Decision Tree and Cluster Sampling Technique to create a predictive model. The historic data of students was used to assess the performance of the students. The factors included attendance, class test, seminar and assignment marks scored in semester examination. Student's attitude towards attendance in class, hours spent for self-study after college, family income, students' mother's age and mother's education proved important factors influencing student performance [105].

Shih B et al (2011) conducted the study to create the model that distinguishes between good student and bad student who made use of Bottom-Out-Hint. The different steps have different types of hints; students could explore the hints till the last hint available. Logged-in response time: predicts learning as well as captures behaviours related to self-explanation. There was presence of distractions such as low attendance, fooling the system by asking the hints again and again. WEKA tool was used to find appropriate classifier for a given dataset. Bayesian Network was used to predict the necessity of help-requests [86].

Kumar V et al (2011) reviewed the educational data mining algorithms that are becoming popular in research area. Different tasks, tools and data mining techniques were discussed. These tools and techniques could be applied on the data collected from higher education. To identify different techniques which are useful and could suit the available data. Different data mining algorithms were
discussed with their percentage of usage in data mining. Important algorithms of data mining like: Classification Algorithms, Regression Algorithms, Segmentation Algorithms, Association Algorithms and Sequence Analysis were discussed. Applications of different algorithms were discussed for different types of data [56].

Baker RS et al (2009) discussed the importance of educational data mining methods in EDM like Statistics, Visualization, Classification, Clustering, Web Mining, Association Rule Mining, Sequential Pattern Mining and Text Mining. The different techniques of data mining are discussed; one key area of application has been the improvement of student models. Also the most frequently cited and popular paper references are discussed [21].

Vandamme JP et al (2007) categorized students into three groups: the ‘low-risk’ students have a high probability of succeeding; the ‘medium-risk’ students, who may succeed; and the ‘high-risk’ students, who have a high probability of failing (or dropping out). The techniques they used to extract the results are: Discriminant Analysis, Neural Networks, Random Forests and Decision Trees to predict the academic success of students'. However, they found none of the three models provide high rate of correct classification. They found the sample size was small to implement the models [96].

2.9 Various Tools Used in Prediction

Prediction in layman terms is a projection of the unknown but which can be considered to have a value in the future. In this study, prediction was used to render meaning to the unknown variables. However, it is assumed to be an important technique as it helps in predicting the risk factors that eventually offer solutions to the risk factors. Earlier YALE AND ORANGE were considered as relevant tools of prediction but with the ever increasing research, Rapid Miner and WEKA are acknowledged as important prediction tools.

WEKA (Waikato Environment for Knowledge Analysis). WEKA is open source software issued under the GNU General Public License WEKA is a collection of machine learning algorithms for data mining tasks. WEKA contains
tools for data pre-processing, Classification, Regression, Clustering, Association Rules and Visualization. It is also well-suited for developing new machine learning schemes. Its portable as whole implementation is done in JAVA and thus runs on almost any modern computing platform [72].

KEEL: KEEL (Knowledge Extraction based on Evolutionary Learning) is an open source (GPLv3) Java software tool that can be used for a large number of different knowledge data discovery tasks. KEEL provides a simple GUI based on data flow to design experiments with different datasets and computational intelligence algorithms (paying special attention to evolutionary algorithms) in order to assess the behaviour of the algorithms. It contains a wide variety of classical knowledge extraction algorithms, pre-processing techniques (training set selection, feature selection, discretization and imputation methods for missing values, among others), computational intelligence based learning algorithms, hybrid models, statistical methodologies for contrasting experiments and so forth. Moreover, KEEL has been designed with a two-fold goal research and educational [67].

RapidMiner: Formerly known as YALE (Yet Another Learning Environment) Rapid Miner is written in Java. It is used for both research and real-world data mining tasks. RapidMiner provides data mining and machine learning procedures including: data loading and transformation (Extract, transform and load (ETL)), data pre-processing and Visualization, Predictive Analytics and Statistical Modelling, Evaluation, and Deployment. RapidMiner is written in the Java programming language. RapidMiner provides a GUI to design and execute analytical workflows. Those workflows are called “Process” in RapidMiner and they consist of multiple “Operators”. Each operator is performing a single task within the process and the output of each operator forms the input of the next one [3].

Adhatrao K et al (2013) compared clustering and classification techniques to predict the performance of the students. The authors preferred classification algorithms over clustering based on their data set, classification was most suitable. ID3 and C4.5 was used to process the data using Rapid Miner. The prediction
accuracy was achieved approximately 75% for both the techniques on data set of 173 students. These techniques mapped with PHP as well to perform the prediction [3].

Nghe NT et al (2007) compared the various data mining tools and techniques to conduct this study. On the basis of accuracy, they found Decision Tree to be well performing as compared to Bayesian Network Algorithms. In comparison of tools like WEKA, ORANAGE and YALE they found WEKA and YALE better data preparation tools. Prediction of 3rd year undergraduate students done on the basis of their performance in exams and they divided the students into different categories: fail, fair, good, very good [67].

Kumar S et al (2011) author proposed comparison of C4.5 algorithm with ID3 in terms of the efficiency in building the tree. Predicting the grades of student’s at an early stage, helps improve studies of the student and help them to pass the examinations. C4.5 (J48 in WEKA) is used to predict the performance of the student. Classification: Decision Tree, Support Vector Machines, Bayesian Network (Predict behaviour in education setting). Regression: Linear Regression, Neural n/w, Support Vector, Machine Regression, Density Estimation [55].

Pandey M et al (2013) worked on data set of engineering students enrolled in an undergraduate course. Four Decision Tree algorithms were compared and J48 decision tree algorithm was selected for model construction, where J48 is a java version of C 4.5. The model obtained accuracy of 80.15% and 82.58% in 10 fold cross validation method respectively. The authors proposed a classification model, particularly decision tree algorithm, to predict the future grades of the students in their final examinations. Classification rules were used to identify the weak students, so that appropriate actions could be taken, and success rate of students could be increased sufficiently. The WEKA tool was used for analysis and prediction [72].

Romero C et al (2008) focused to discover potential student groups with similar characteristics and how the students responded and reacted to a particular pedagogical strategy to detect misuse of the platform provided to the student or
game-playing the system to fool the mentor. To find out the students category students were grouped in two groups: Hint-Driven or Failure-Driven. The authors also tried to find common misconception in students about their mentors and course they opted for. This helped in identifying the low-motivated students and providing them the remedial actions so that they could cut down the drop-out rates. By Predicting/Classifying students using Intelligent Tutoring System, the author has used different classification methods to perform the analysis: Statistical Classification, Decision Tree and Rule Induction. Machine learning techniques such as Fuzzy Logic, Neural Network, and Discriminant Function Analysis is used for the purpose. The author concluded with the tools and methods that proved better for the study i.e. KEEL framework using JAVA, Pattern Mining, Black Box Mechanism. A Classification Model was created to classify the students into two categories: FAIL (If have learning problems) GOOD/EXCELLENT (if they don’t have any problem in course content) [81].

Chirumamilla Vet al (2014) used Rapid Miner to perform their student prediction. According to them, Rapid Miner tool provides different types of classifiers for classification. They also offered various test options such as: Use Training Test, Supplied Test Set, Cross Validation and Percentage-split. They proposed their future work with Rapid Miner by creating dynamic tress to predict the performance of the students’. They also intended to consider various other attributes they left in the present study such as: extra-curricular activities and other vocational courses thus opted for[27].

Oskouei RJ et al (2014) analysed student’s prediction using two different case studies they collected from the data set of schools as well as college students to perform their study on stress level considering the academic and economic attributes of the students’. Classification and prediction techniques were used to build a predictive model. The data was collected from proxy server to check their log-in in various activities such as internet access per day, types of sites opened per day and the number of papers visited in a day and number of hits recorded per day. All these activities helped in predicting the academic performance of the
students’. The authors used Rapid Miner to perform the analysis of the student data. [69].

Puyalnithi T et al (2016) performed comparison of various classification algorithms with ensemble method using Rapid Miner tool, as it analyses the data in two phases: Testing and Training sections. The designing of the model also consisted of two parts: Pre-processing and Applying model using the performance measures. Naïve Bayes performs best in smaller data sets, Decision Tree performs better with larger data sets and Random Forest performed average in both the cases [74].

2.10 Identification of Research Gaps

Research gap can be defined as the insufficient information that puts barrier in decision making for policy makers, practitioners and patients. From the exhaustive literature review following are derived as the concluded research gaps:

i. Most of the research work is done in the field of educational data mining that is essentially based on online learning. The techniques and methods applied also focus on the online education. No method is applied on traditional learning environment as followed by majority of the Indian institutes (Nasiri, M., et al., 2012) (Hlosta, M., et al., 2014).

ii. Experiments are carried out on the open universities in other countries (outside India) and on selected courses i.e. psychology, mathematics, history and home science. Very few focused on professional/engineering courses (Kuzilek, J., et al., 2015) (Wolff, A., et al., 2014).

iii. The data mining models proved to be better than the mathematical and statistical models (Antunes, C. 2008).

iv. Majority of the studies conducted predict the performance of the students on the basis of the demographic, academic and social factors outside India. Since Indian culture and living style is distinguished and different, less study has been done which relates to the Indian educational system. So the focus will be to study the basis of Indian education system.
v. Online study was considered as an important variable but the researcher was left with no idea than to accept the loopholes.

2.11 Objectives

Most challenging task in educational settings is to predict the accurate academic performance of the students. An attempt is made to predict future academic performance of students in higher educational settings. The main objective of this work is to develop a prediction model using data mining techniques that can predict the academic performance of students on the basis of historic and behavioural data of the students.

To achieve the goal the study has been broadly divided into the following objectives:

i. To identify the most influential features in terms of predicting student academic performance.

ii. To collect the data on the basis of identified features to generate the training data set.

iii. To propose a classification model based on the data mining techniques for predicting student performance.

iv. To deploy and validate the proposed model with respect to Indian education system.

It is relevant to mention that the researcher did not only focus on the academic performance of the students but the background of the families as well. This facilitates the researcher to analyze the causes that lead to poor performance of the students, and acknowledges the factors that might lead to further loss, be it academic or personal. The study was done with the help of various data mining techniques.