# CHAPTER 3

# LITERATURE REVIEW

## 3.1 Introduction

This chapter reviews the literature of previous work related to student potential and performance. A systematic review of all papers related to this field is needed to understand the how Machine Learning techniques can be used in education and how to predict the potential and performance of students. The reason behind the systematic review is to find suitable attributes and methods to fill the gaps in the existing research.

In this Chapter we discuss in detail the work done by researchers in predicting student performance in schools and colleges. Early prediction of potential and performance is beneficial for students and the education system as a whole.

There is a lot of work that has been done in this area, even from early times when Machine Learning techniques were not available. For example, the first study was related to student retention was conducted by Johnson in 1926 while the effect of student morality on academic behaviour was examined by McNeely in 1938 and how it affects the attrition of student.

While these were very interesting studies and laid the foundation for later work in this area, we will concentrate on recent studies. Here also, we shall first discuss those studies in which the emphasis has been on the type of attributes that were used, rather than on the techniques used, and which we consider as important to finalise them for our studies using ML techniques. After that, we will review research work that has used Machine Learning techniques, using some of these attributes.

## 3.2 Literature Review highlighting attributes

Kumar and Singh (2019) conducted a study to evaluate the performance of students to reduce human efforts. Data was taken from an inventory. It contains 34 attributes which include some school features and some subject-marks related attributes. It showed a strong relationship between the attributes and the performance of the students.

**Fernanderz et. al., (2019)** conducted the study to predict the weakness in the students' performance and support for the future. This study used student academic historic

records such as education information, students' behaviour, family behaviour and student's status.

**Ajibade et. al., (2019)** the study stated that features of learners was a very significant role in student performance. Feature selection was used to identify the optimal feature subset. The result showed a strong relationship between student behaviour and their performance.

**Kumar et. al., (2019)** used educational data like demographic, academic and behavioural features and analyse the impact on student's performance based on these attributes.

The work of **Sembiring et al. (2016)** used behavioural data for students' performance prediction in a Malaysia University. The data includes study behaviour, family support, their interest and study time. This study showed a strong relationship between a student's mental condition and his/her performance. In the paper, the author used student learning style with five different personality factors.

**Sungar et al. (2017)** conducted a study to examine and compare different classifiers for predicting student performance. The questionnaire for this study used demographic, personality, psychological, employment and institutional attributes. It showed a strong correlation between the attributes which were considered.

**Almasri et. al., (2019)** used the student's education information and behaviour in the collected dataset for predicting student performance. In another paper, (Zohair, 2019) data was collected from the administration department for the study held on graduate and postgraduate students. This dataset includes student age, bachelor degree grade, name of course taken in graduation with their grades and teachers name of each course. This showed that a student's grade in most courses was related to the student's dissertation course grade.

Another study **(Kumari et al., 2018)** found that prediction of student performance in the final year is possible, based on first- and second-year marks. They also considered higher and secondary school marks.

Another study mentioned that gender is the main factor for analysing the students' performance. **Suhaimi et al., (2019)** stated from their research, that when the behaviour

of female and male students is compared, it showed that female candidates have positive behaviour as compared to male.

One study proposed the PhD students' graduation status. Female candidates were found to have completed their graduation on time as compared to males. Apart from gender, age is also an important factor according to some studies in this field.

**Walter and Soyibo (2011)** examined high school student's performance on factors such as school location, gender, school type, grade and socio-economic attributes. The results showed a positive relationship between the nature of the school and students' academic performance.

**Sunita and Khadi (2007)** investigated the factors that influence the academic learning environment of home and school. The samples for this study were 240 and the selected students were from eight co-educational schools, with two different mediums of education i.e., English and Kannada in a Karnataka city. Measuring the relationship between factors like home environment, school environment, socio-economic attributes and academic achievement was calculated by Karl's Pearson correlation. It was shown that high socio-economic status students studied in English medium school and achieved high academic performance while students from low socio-economic status studied in Kannada medium and compared low academic performance. This showed that the medium of education would affect academic performance.

**Khan (2005)** examined the academic performance at higher secondary level in the science stream. The study was conducted on 400 students, which included 200 boys and 200 girls selected from Aligarh Muslim University. Factors used were demographic and socio-economic attributes. The cluster sampling technique was selected in which the population related to this study was divided into clusters and a random sample of the cluster used for further analysis. The result found that the girls with high socioeconomic status performed better and boys with low socio-economic performed well.

**Hijazi and Nagi (2006)** investigated that the students' attitude towards attendance, study time on a daily basis, family income, mother's age and education have influenced the performance of the student. 300 students (225 males and 75 females) were selected from Punjab university of Pakistan. Linear regression analysis was done on this dataset.

It was shown that two factors, mother education and family income, were strongly correlated with the student academic performance.

**Kristjansoon et. al., (2010)** studied to estimate the relationship among different factors, such as body mass index (BMI), health and self-esteem and academic achievement. This study used 6346 records of adolescents in Iceland for analysis of survey data. They found attributes that were associated with higher academic achievement. Increasing levels of BMI was a negative impact on self-esteem.

**Maric and Petro (2004)** studied the gender differences of secondary school by using cognitive and motivational variables. This study used 521 samples (236 male and 285 females) of students aged between 14 and 18. T-test was used in this study to evaluate the difference between male and female secondary school students. It showed that there exists a gender difference in the students' academic performance.

**Grissom (2005)** examined the connection between student's fitness level and academic achievement in test scores. This was study conducted on fifth, seventh and ninth-grade students in a California school, during the year 2002. The dataset had 84715 records. It was found that there was a strong relationship between students' fitness level and achievement. It was found better in females than in males.

**Valerie and Dawn (2009)** did the study based on various attributes such as birth order, parental attention and academic performance. This study used 20 individuals (9 males and 11 females) age from 13 to 39 years. It was shown that parental attention decreases with the order of birth, i.e., parents give the most attention to the first child and this decreases with the next child, and so on. This impacted the student's academic performance. Intellectual competition between siblings may lead due to parental favouritism and expectations.

**Taiwo and Tyolo (2002)** studied the effect of early childhood education on performance. The dataset included three subject areas. The data was gathered from four selected primary schools of 120 students in Botswana. The result showed that pupils with preschool education experience gave better performance than those without such experience in all three school subject areas.

The above comprehensive literature review was very useful in finding out the attributes which should be considered in predicting a student's potential and performance. Some

attributes which are frequently used are demographic, internal assessment, cumulative grade point average (CGPA), External assessments and psychometric factor. The internal assessment was classified as quizzes, lab marks, attendance, and assignment. Demographic attributes consist of age, gender, and family background. While external assessments include marks obtained in exams for a particular stream or subject. A psychometric factor is used to study behaviour, engagement time, student Interest and family support.

On the basis of the above study the common factors have been seen to have been employed for student potential and performance prediction, are shown in Figure 3.1.

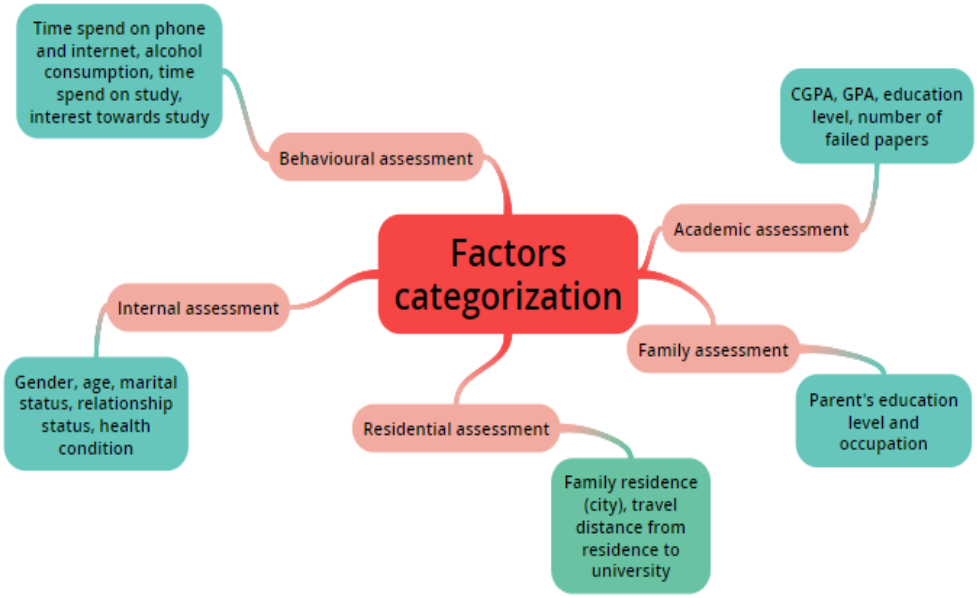


**Figure 3.1** Common Factors used for student performance prediction

Source: Suhaimi et al., 2019

### 3.3 Literature Review using Machine Learning techniques in Academics

In this section we review recent literature that has used Machine Learning techniques for academic potential and performance of students.

**Vaidu and Sornalakshmi (2016)** discussed ML techniques to evaluate students' performance for employability skills. They implemented two ML techniques, K-Nearest Neighbors (KNN) and Naïve Bayes to classify the students into several groups. The results showed that KNN achieved better performance at 95.33% accuracy and Naïve Bayes got 67.67% accuracy.

**Zafar, Junaid and Faisal (2017)** discussed different ML techniques to predict grades in different subjects. They used matrix Factorization, Classification models, like collective filtering, regression and Restricted Boltzmann Machine (RBM) techniques. These methods were applied to analyse data collected from Information Technology University (ITU), Pakistan. They evaluated the performance of the ITU students who were studying in the Bachelor's Program. The RBM technique was shown to achieve a better result and found to be the best among different applied ML techniques.

**Kim Byung-Hak et al. (2018)** have proposed a Deep Leaning technique for predicting the future performance of students. They collected the student data from Udacity Nanodegree Programs. This study implements the GritNet algorithm, based on the concept of deep learning. This method provides better results as compared to Logistic Regression.

**Jie, Kyeong and Schaar (2017)** proposed an ML technique for predicting the performance of students in Degree programs. The calculation is based on past and present performance. The planned system uses a bilayered structure, consisting of a data-driven approach based on a latent factor model to construct more than one base predictor and an efficient basis prediction. This paper shows that the proposed method achieves better performance as compared to benchmark approaches.

**Murat (2017)** examined the ML algorithms to predict the performance of the student. This study applied three different techniques viz.) Linear Regression (LR), Decision Trees (DT) and Naïve Bayes (NB) classification on two different datasets. The best technique is NB for the first dataset with 98% accuracy and DT for the second dataset with 78% accuracy.

**Singh and Singh (2013)** applied ML techniques on engineering students for predicting academic performance according to subjects. This study predicts subject scores in ongoing courses by analysing subjects based on the previous semester. This paper implemented two classification techniques, Naïve Bayes and C4.5 Decision Tree classifier for the proposed purpose. The result shows that C4.5 DT has achieved better accuracy than NB.

**BendengnuKsung and Prabhu (2018)** suggest a Deep Neural Network model for predicting the performance of the student and class category. The paper used a Deep Neural network with existing different ML techniques such as Decision Tree (J48),

Naïve Bayes and ANN and compared the accuracy between them. This model got 84.3% accuracy and is better than other applied ML Algorithms.

**Balachew and Gobena (2017)** proposed a model using ML techniques to predict students' performance. The dataset of the students used in this paper is taken from Wolkite University. They apply three ML Techniques Neural Network, Naïve Bayes and Support Vector Machine (SMO) to build the models based on each method. It evaluates the performance of the students and compares the results of each predictive model.

**Agrawal and Mavani (2015)** propose a Neural Network for the prediction of student performance. The algorithm is applied to a dataset of 60 engineering students. The performance of the students in academics is based on their past performance. This paper compares the neural network algorithm with the NB. The result shows that the Neural Network algorithm results improved with an increase in dataset size.

**Ali, Helali and Mohamad (2013)** find the factors affecting the course and academic performance of the students using Data Mining techniques. The dataset was got from the community college, computer science and business administration departments at Narjan University, Saudi Arabia. This paper implements a predictive model where both clustering and classification techniques are used to identify features affecting the performance of the students.

**Fok et al. (2018)** proposed a prediction model using deep learning and Tensorflow Artificial Intelligence Engine, to predict a student's future development. They used a dataset of 2000 students, 75% of the data was used as training data and 25% was used as testing data. The paper adjusted and compared the number of hidden layers, hidden nodes, number of iterations and the learning rate. This study demonstrated that deep learning could be an effective tool for predicting the performance of students. The accuracy ranged from 80% to 91%.

**Aly, Hegazy and Heba (2013)** have developed a student advisory framework based on classification and clustering. The paper used the C4.5 algorithm to predict the likely department for first-year university students. K-means clustering was also used to divide the students into a number of groups and find the success in each cluster for each department. The dataset of students was taken from Cairo Higher Institute for engineering, department of Computer science and Management. The study concluded

that the existence of different proportions of successes in each department in each cluster.

**Guo et al. (2015)** developed a classification model to predict student performance using Deep Learning. This model used pretrained hidden layers of features using an Unsupervised Learning algorithm and then used Supervised training for fine-tuning the parameters. A Students' performance prediction model (SPPN) was trained on about a 120,000 labelled student dataset with training parameters. The paper implemented three different classification algorithms: Naïve Bayes, Multilayer Perception (MLP) and SVM to compare results with SPPN on the same dataset. The result showed that SPPN has a higher average accuracy of 77.2% than other algorithms.

**Singh (2013)** applied ML techniques on engineering students for predicting academic performance according to subjects. This study predicts subject scores in ongoing courses by analysing subjects based on the previous semester. This paper implemented two classification techniques, Naïve Bayes and C4.5 Decision Tree classifier for the purpose. The result found that C4.5 DT achieved better accuracy than NB.

**Ermiyas and Gobena (201**7) proposed a predictive model using Machine Learning techniques to predict students' performance. The dataset of the students used in this paper was taken from Wolkite University. They applied three Machine Learning Techniques: Neural Network, Naïve Bayes and Support Vector Machine to build the models based on each method. It evaluated the performance of the students and compared the results of each predictive model.

**Gorikhan and Abdullah (2016)** evaluated various classification methods to predict the students' performance in the computer science course at a vocational institute. The objective of the work was to analyse data on student records. He used different Machine Learning Algorithms Decision Tree, KNN, Logistic Regression, Support Vector Machine and Neural Network. Rapid Miner was used to pre-processing data. Two datasets were used, dataset1 considering science & math marks and dataset2 did not consider science and math marks. After comparison, it was found that dataset1 produced more accurate prediction. In their thesis, the authors also compare different ML techniques. The result showed that the Decision Tree had the highest accuracy (89%) for predicting the performance of students.

**Bergin and Reilly (2006)** studied factors that influence introductory programming success. They predicted whether students will be weak or strong programmers. The authors investigated 25 factors at four different institutions. They also investigated the effectiveness of various Machine Learning algorithms to predict introductory programming performance. Six different Machine Learning Techniques Naïve Bayes, Logistic Regression, Support Vector Machine, C4.5, Neural Network. PCA was implemented to reduce the dimensionality of the dataset. The result showed that NB and Support Vector Machine had more accuracy in predicting the students' performance.

**Mishra, Kumar and Gupta (2014)** built a performance prediction model by applying different classification methods, based on students' social integration, academic integration and various emotional skills. This study implemented two algorithms: J48 (implementation of C4.5) and Random Forest. These algorithms were applied to a dataset of MCA students of colleges affiliated with Guru Govind Singh Indraprastha University to predict third-semester performance. The result showed that Random Forest was more accurate than the J48 algorithm.

**Han et al. (2017)** developed a model based on the ensemble algorithm, AdaBoost, to predict the classes of students. This model was compared with other algorithms, such as Decision Tree, Neural Network, Random Forest and Support Vector Machine. This paper used 123 undergraduate students' dataset in the school of Urban and Geographical in a university. This model achieved an accuracy of 91.67% and was better than other Machine Learning algorithms.

**Sungar, Shinde and Rupnar (2017)** discussed different Machine Learning Algorithms for predicting student's performance. Students' marks, aptitude, family background, educational environment were the main factors while selecting a career path and these factors acted as a training set to the Learning System for classification. This paper used four pairs of classifiers, namely, MLP-CART, MLP-SVM, SVM-CART and SVM-SVM for the experiments. The model showed comparative results of MLP and SVM mapping functions with SVM and CART as base classifiers. MLP mapping function gave good results.

**Mahboob, Irfan and Karamat (2016)** analysed the chances to predict the success rate of students enrolled in a course using Machine Learning algorithms. This paper used

60 students' dataset. Weka tool was used to apply different algorithms to the student dataset. Four algorithms, Quinlan's C4.5, J48, Naïve Bayes and Random Forest were applied to the dataset. The result showed that Random Forest was more accurate than other Machine Learning algorithms.

**Tanuar, Heryadi and Gaol (2018)** predicted the final year GPA of students, based on first- semester results, using machine learning algorithms. The data used in this paper were from the computer science subjects, laboratory results and the GPA on their graduation year in Bina Nusantara University. Rapid Miner was used as the platform. This research used three models, viz., General Linear Model, Deep Learning and Decision Tree. The result showed that the important factors that had an impact on the result can be extracted. This will enable the students to prepare themselves for the exams well in advance.

**Martín et al. (2018)** evaluated the performance of four ML algorithms to predict dropout rate with a different perspective in university students. Four algorithms were used in this research, Neural Networks, Support Vector Machines, Random Forest and Logistics Regression. The dataset was collected from Instituto Tecnológico de Costa Rica (ITCR) whose students enrolled in a degree program between the years 2011 and 2016. The Random Forest algorithm with two variables was found the best for predicting the dropout rate.

**Pushpa et al. (2017)** predicted final year results, based on the performance of the students in the previous semesters, using Machine Learning Algorithms, to know whether a student would pass or fail. Four algorithms were used in this paper: Support Vector Machine, Naïve Bayes, Random Forest and Gradient Boosting. The result showed that the accuracy of Random Forest 89.06% was higher than other algorithms.

**Gerritsen and Conijn (2017)** developed a model using a Neural Network for predicting the student performance from Learning Management System data, in the context of educational data mining. This paper collected data from a Moodle log file, containing information about 4601 students. This research compared the performance of Neural Network with other classifiers. These algorithms were NB, kNN, DT, Random Forest, SVM and Logistic Regression. This study concluded that the accuracy of the Neural network was better than other classifiers.

**Kumar and Singh (2019)** proposed a classification model to evaluate student performance using ML algorithms. This paper used a combination of k-means clustering with SVM and ANN to evaluate student performance. The online dataset used contained 34 different attributes. The evaluation of student performance was done based on mean square error and effort estimation. The result shows that the performance of ANN is better than SVM. The mean square error is 5-20% better.

**Fernandez D.B. et al. (2019)** implemented machine learning techniques to predict the final grades to evaluate student performance based on their historical academic information. The dataset used 335 students' academic records. This dataset was taken from engineering students of Ecuador university. Data collection and preprocessing was done in the initial step and then a grouping of students, based on the similar pattern carried was out. Machine learning effectiveness shows in the prediction of student performance.

**Almasri Ammar et al. (2019)** constructed an ensemble meta based tree model (EMT) classifier method for predicting the performance of students. 400 student's records with 13 attributes were used in this study. The experimental result shows that EMT is more accurate than other ML techniques.

**Zohair and Mahmoud (2019)** developed a classification model to predict student performance using clustering algorithms. This paper tried to identify the key indicators in the small datasets which were used in creating the prediction model. 50 graduate students' record was used for this study. This paper used MS excel and python 3.6.2 for the analysis of collected data. It also used R studio for data visualization. Among the implemented algorithms, the SVM algorithm for small dataset size shows better accuracy than other ML algorithms.

**Lau E.T et al. (2019)** examined both conventional statistical analysis and the Neural Network modelling approach for the prediction of students' performance. The data was collected for about 1000 undergraduate students consisting of 275 female and 810 male students and was taken from a Chinese university. The neural network was implemented with 11 input variables, two layers of hidden neurons and one output layer. This model achieved 84.8% accuracy.

**Sudani and Palaniappan (2019)** proposed a multi-layered neural network to predict students' performance. The objective was to classify students' degrees into either good

or basic class. A feed-forward network with 100 nodes and a trained layer using the levenberg – Marquardt learning algorithm was used. The data used was of 481 students and divided into three sets: 70% of data used in training, 25% of data used as test data and 5% used for validation. This paper used four algorithms K- Nearest Neighbor, Decision Tree, Support Vector Machine and Neural Network. The Neural Network model was compared with other classifiers on the same dataset. The result showed that Neural Network performed better than other algorithms in terms of accuracy.

**Pal and Bhatt (2019)** studied prediction accuracy rate using R programming. This paper evaluated student performance for postgraduate students. The objective of the study was to analyse the factors that affect the academic performance of students. The dataset consisted of 395 students with 30 attributes. This dataset was gathered from the UCI repository. The paper used Deep Learning with other methods like Linear Regression and Random Forest. They used accuracy, Recall and F-measure to compare total data. The outcome of Deep Learning is better than other ML techniques.

**Suhsimi N.M. et al. (2019)** discussed the factors used to predict student's performance. Their paper studied the impact of different types of factors using different ML techniques. It used a Neural Network, Support Vector Machine and Decision Tree. Neural Network has the highest accuracy of 95% compared to other ML algorithms. This paper shows that when predicting a student's graduation time, the academic assessment was a prominent factor.

**Roy and Garg (2017)** implemented different ML techniques to predict student academic performance. This study helps us to identify the interests and weaknesses of students. Different types of attributes like social, demographic and those related to the school, influenced the performance of the student. This paper used different classification algorithms Naïve Bayes, J48 Decision Tree and MLP. Naïve Bayes has higher accuracy of 68.60%.

**Hassan H.M. et al. (2019)** studied different ML techniques for predicting student performance. This paper tried to predict future results based on the current status of the students. Data of 1170 students of three subjects were used in this paper. It used K-nearest Neighbor and Decision Tree. The Decision Tree has the highest accuracy (94.88%).

**Heryadi and Gaol (2018)** predicted final year GPA based on first-year semester results using machine learning algorithms. The data used in this paper were from the computer science subjects, laboratory results and the GPA on their graduation year in Bina Nusantara University. Rapid Miner was used as the platform. This research used three models viz., General Linear Model, Deep Learning and Decision Tree. The result concluded that the important factors that had an impact on the result can be extracted. This will enable the students to prepare themselves for the exams well in advance.

**Islam (2017)** predicted students' academic performance based on two categories: behavioural and student absence in class, using some data mining techniques. They used four classification algorithms; KNN, Naïve Bayes, Decision Tree and Artificial Neural Networks. They also used ensemble methods bagging, Adaboosting and Random Forest to get more accuracy. The student Dataset was collected from an e-learning system called kalboard 360, which has 480 instances and 16 features. The ensemble filtering technique has higher accuracy of 84.3%.

**Turabieh H. (2019)** evaluated ML algorithms for predicting student marks. This paper focused on developing a new approach to finding meaningful knowledge from collected data. The dataset was taken from students' e-learning log files in a virtual course. The paper used a hybrid feature selection algorithm with different classification algorithms, ie., KNN, Convolutional Neural network, Naïve Bayes and Decision Tree (C4.5). The binary genetic algorithm as a wrapper feature selection was used on collected data. The results show that BGA increases the performance of all classifiers.

**Patil A.P. et al. (2017)** predicted student grade point averages using an effective deep learning model as compared to ML techniques. Feedforward Neural Network and Recurrent Neural Network for predicting the student's GPA. They used various recurrent Neural architectures and compared their results. The paper compared the proposed model with ML Techniques. Accuracy of the Bi-directional long term memory networks (92.6%) were higher than other algorithms.

**Soni A. et al. (2018)** examined different classification algorithms for analysing pupil performance. This paper used graduate and undergraduate student's data collected from different universities during the period from 2017 through 2018. The data was collected with the help of a questionnaire survey. Three classifiers, viz., Naïve Bayes, Support

Vector Machine and Decision Tree were used for evaluation of students' performance. The accuracy of SVM is 83.33%, which is higher than the other algorithms.

**Livieris I.E. (2018)** predicted secondary school students' performance using a semi-supervised learning approach, using two wrapper methods. They examined and evaluated the effectiveness of algorithms for predicting the performance in the final exam. The dataset of 3716 students in mathematics courses was taken from Microsoft showcase school "Avgoulea- Linardatou" during the years 2007 through 2016. The result shows that classification accuracy can be increased by using a semi-supervised learning algorithm.

**Adekitan A.I. and Salau D. (2019)** implemented different data mining algorithms for predicting the student graduation result by using engineering students' performance in the first three years. The dataset of 1841 students from 2002 to 2014 across seven engineering departments was taken from Covenant university in Nigeria. Six data mining algorithms, viz., probabilistic Neural Network, Random Forest, Decision Tree, Naïve Bayes, Tree Ensemble and Logistic Regression were used. Logistics Regression achieves higher accuracy (89.15%).

**Mduma N. et al. (2019)** studied different ML approaches for the prediction of student dropout. This paper surveyed literature in books, journals and case studies. Several work has been done using supervised and unsupervised learning algorithms. This paper concluded that several techniques have been used for this problem in developed countries, but there is a lack of research in developing countries.

**Kumar M. and Salal K. (2019)** analysed the performance of students using different data mining algorithms. The authors reviewed 46 papers and discussed the different classification algorithms with their maximum and minimum accuracy. They also concluded that there are lots of software tools like WEKA, Rapidminer, MATLAB, KNINE, GUI, R and Python available for this work. They found that academic performance of students is a good topic of research which helps students, academician and management for improving academic performance.

**Bunker and Thabah (2019)** studied a machine learning framework for sports result prediction using Artificial Neural Networks. Data was taken online from publicly available sources. This paper critically analysed some research papers on sports

prediction. It proposed the SRP- CRISP-DM framework for the sports results prediction.

**Atujjar et al. (2019)** implemented the ID3 Decision Tree induction algorithm for predicting academic performance. The dataset of female students' record in the bachelor program at the IT department, King Saud University, Saudi Arabia was collected. They developed classification models for each year of the program. The result shows that the classification model based on year performance is more accurate.

**Mienye et al. (2019)** reviewed different Decision Tree algorithms from several fields. ML Decision Tree algorithms which included ID3, C4.5, C5.0 and CART were discussed. The paper compared these algorithms and C4.5 is an improved version of ID3. It studied the performance of different Decision Trees.

**Fernandes and Holanda M. (2017)** presented a predictive analysis of students' performance. This study was done on public school students in the Federal district of Brazil. The paper used two datasets. The first dataset contained features taken before the start of the school year and the second included academic information features collected after two months. The result showed that the attributes like grades and absences were the most relevant features for academic performance. They also concluded that neighbourhood school and age were also important factors of students' success or failure.

**Verma and Mehta (2017)** proposed a novel approach "BBS Method" for the classification of the five UCI datasets taken from the field of bioinformatics. BBS is an ensemble method and stands for bagging, boosting and stacking. This paper showed that the proposed method gave better accuracy as compared to individual algorithms. It also showed that the ensemble approach minimized root mean square error.

**Kumari P. et al. (2018)** implemented machine learning techniques and ensemble methods to predict student performance. The objective of the paper was to show the importance of student's behaviour features. Data was collected from Learning Management System (LMS). This paper implemented some classifiers, namely: ID3, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine. The accuracy of the proposed model was achieved by using ensemble methods.

**Ajibade S. M. et al. (2019)** constructed a new performance model using Ensemble methods for predicting the performance of students. The proposed model compared some Machine Learning Classifiers like Decision Tree, KNN, and Support Vector Machine. This paper showed that 91.5% accuracy was achieved when ensemble methods were used.

**Adejo O.W. and Comolly T. (2019)** investigated different Machine Learning Classifiers and Ensemble Methods to predict student academic performance. The data was collected through a survey questionnaire. They developed a hybrid model that was high in accuracy and efficiency in performance. The findings showed that the accuracy, precision and recall were better in the ensemble classifier, rather than the base classifiers.

**Kumar A. D. et al. (2019)** used a hybrid classification model for predicting student performance. Two popular classification algorithms, ID3 and J48, were applied in this paper. The dataset was collected from various departments of UG students. Weka tool was used to implement different algorithms. The accuracy of hybrid classification was 62.66%.

**Almasri A. et al. (2019)** predicted student performance for finding the weakness and providing support for future students. The goal of the study was to analyse the factors that affect the academic performance of students. It also proposed an Ensemble Meta based Tree model (EMT) for predicting student performance. The result showed that the EMT as an Ensemble technique has a high accuracy of 98.5%.

**Sultana J. et al. (2019)** studied prediction accuracy rate using Weka Tool. They implemented some classifiers to predict student performance. The dataset was collected from a Saudi University database. This paper used Deep Neural Network – MLP with other classifiers. They used accuracy, kappa statistics and ROC curve to compare total data. The outcome of Deep NN-MLP was better than other ML techniques.

**Krishna M. et al. (2019)** aimed to study and test the ability of CART analysis to predict course performance. Data was extracted from a Moodle-based blended learning course and built a student performance model. In this paper, the CART technique achieved a very high accuracy (99.1%).

**Dietterich T. G. (2017)** studied Ensemble Methods that construct a set of classifiers. This paper reviewed these methods and explained why ensembles can often perform better than other classifiers. They compared some previous studies based on ensemble methods and provided some experimental results which showed that Adaboost performs well.

**Aziz S.M. et al. (2019)** focused on constructing a classification model for predicting student performance. The dataset was collected from three different educational institutes using a questionnaire. This paper has shown that numerous attributes may have a high impact on students' performance. It used the Decision Tree algorithm to build the classification model.

**Kaur A. et al. (2018)** implemented different Machine Learning techniques and Hybrid ML to predict student academic performance. This paper also discussed how these Machine Learning models can help to improve the education system. They used accuracy, recall, precision and F-measure to predict the performance of students. The dataset consists of 1735 instances and 37 attributes of BTech second-year students. The results showed that hybrid Machine Learning techniques gave better performance.

**Rahman M.H. and Islam M.R. (2017)** predicted students' academic performance based on two categories: behavioural and student absence in class. This paper measured the effect of two categories of features on performance, using some data mining techniques. They used four classification algorithms; KNN, Naïve Bayes, Decision Tree and Artificial Neural Networks. They also used ensemble methods bagging, Ada boosting and Random Forest to get more accuracy. The dataset was collected from an e-learning system, which has 480 instances and 16 features. The ensemble filtering technique has higher accuracy of 84.3%.

### 3.4 Summary of Literature Review

The above systematic and comprehensive review, as has been done, helps to map out related research work and identify the gaps of knowledge. It defines the areas where further research is required. All related studies have been classified into parts such as objectives, the different applied techniques, and the data used for each study. The above studies reiterate that although Machine Learning algorithms have been used to study and predict student performance, no such study has been conducted on students passing $12^{th}$ standard. Also, the study has not been done to find a student's eligibility for

technical programs. This is very important in the context of India, where more than 15 lakh students appear for such programs.

The literature reviewed above focused on academic and behavioural attributes of students, but not on their financial background. The factors that influence the student potential and performance are not solitary but are interconnected, interrelated, and interdependent. So, there is a requirement for research to create a new prediction model that is comprehensive and general in its style.

### 3.5 Research Gaps identified

Based on the above literature review, we have identified the following research gaps:

1) ML techniques have not been used on Rajasthan students.
2) Even outside Rajasthan, no study has been done on $12^{th}$ standard students.
3) No one has predicted the potential of $12^{th}$ class students for admission into a particular technical stream.
4) ML techniques have been used for evaluating student performance and comparing results with conventional methods. But no studies have been done for actually grading students using clustering techniques.

### 3.6 Chapter Summary

This chapter reviewed the different studies that discussed and investigated the various attributes that have been used to study and predict student performance. It also analysed different factors like demographic, socio-economic, academic etc., that influence student potential and performance. Further, it also reviewed the ML classification techniques which are used for the analysis of student performance prediction. Based on the review, the Research Gaps were identified in this chapter.

In Chapter 4, we explain the methodology of research to fulfil the objectives of this research. After that, the data collection, and data pre-processing techniques to be used in this research will also be discussed.