Table: Review of application of ML in education

Studies	The technique(s) selected and Analysis	Dataset details
[10]	▪ Used ML to enhance prediction of student academic performance ▪ The ML methods used are Artificial Neural Network, Logistic Regression, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, K-Means, and Decision Tree	▪ Dataset used using pre-academic program data of the students e.g. school system information, gender, scholarship, ethnicity, Math and English level, and result.
[11]	▪ Applied ML clustering and classification algorithms to predict student progress in HE ▪ Used K-mean and Random Forest to apply classification algorithm on each cluster identified by clustering algorithm	▪ Dataset contains student attributes from admission to graduation e.g. High school type, high school result, admission date, Time left to Degree, major, GPA, project marks, awards, entrance test grades.
[26]	▪ Used supervised classification ML algorithms to predict student retention rate at three levels of degree i.e. Decision trees, K-Nearest Neighbors, Random Forest, Naive Bayes, Logistic Regression , Support Vector Machines ▪ Result shows that high school result and student Socioeconomic indicator are important features to predict student dropout	▪ Chilean Educational dataset is used e.g. Student age, gender Socioeconomic index, high school performance and grades, health insurance, university grades and financial attributes.
[12]	▪ Used Classification ML algorithms to predict the student academic status and identify dropouts in coming semesters ▪ Applied three models i.e. Random Forest, Gradient Boosting and Decision tree techniques to identify student at-risk ▪ Result showed that academic year and high school GPA is directly related to the student dropout.	▪ Dataset is collected from KMUTT from 2012 to 2020 ▪ Data set consists of 81 student attributes like personal family information, high school information and GPA, admission information, faculty name and student academic year and status
[13]	▪ Detected student behaviour of interacting with the LMS to achieve better learning outcomes ▪ An experiment was done by Spearman Correlation analysis, K-means, and Network graph ▪ Found LMS modules that were either not accessed or least accessed by the students	▪ LMS dataset ▪ Static data - gender, region, age, credits transfer details, and GPA score. ▪ Dynamic data (LMS) - count of hits, forum posts details, and count of assessments viewed and submitted.
[14]	▪ Predict the progress of the student based on the academic record ▪ Employed supervised learning models of ML like logistic regression method ▪ The three factors that impacts the academic grades are student punctuality, student satisfaction level, and details about system interaction	▪ 16 distinct attributes. ▪ Dataset consists of demographical attributes like student gender, place of birth, parent academic background, residing region, and grade level, student engagement ratio, participation log data, and submitting assignments.
[15]	▪ Analyzed student examination data to predict student dropout ▪ logistic regressions and decision trees	▪ Dataset consists of data about 487 exams and student attributes like count and date of attempted exams, results, and the enrolment date
[16]	▪ dropout warning system to pre-emptively identifies K-12 at-risk students in the online courses.	▪ Data set collected from online classes consists of speech-related features, language-

	▪ The prediction model employed was Logistic Regression, Decision Tree and Random Forest ▪ Within two months, the dropout warning system has detected more than 70% dropped out students accurately.	related features, Interaction features, Pre-class and Post-class features like discount amount or change requests to their course schedules, and Time-variant like courses enrolled recently.
[17]	▪ Blended Learning environment ▪ Analyse homework submission behaviours to forecast student academic progress. ▪ Ten classification algorithms are used, e.g. Random Forest, ZeroR, J48, Prism, JRip, ID3, OneR, NBTree, PART, and decision stump to categorized the students as procrastinators or non-procrastinators	▪ LMS assessment dataset in the blended learning environment ▪ Three homework variables start and end date and submission date are used
[6]	▪ Predicted the students' performance by using four classification techniques from tree-based and ML boosting classifiers e.g. random forest, and extreme gradient boosting classifier	▪ UCI Publicly available data ▪ Dataset consists of 32 student attributes e.g. gender, age, school, address, parent's individual education and job, former grades, free time, frequency of going out with friends, weekday and weekend alcohol consumption, and other social, demographic attributes
[5]	▪ Forecasted the number of students who will be a pass/fail for necessary arrangements. ▪ Three supervised classification methods J48, NNge and MLP were employed ▪ J48 achieved the best accuracy of more than 95%	▪ Publicly available dataset - UCI Machine Learning Repository containing 33 attributes consists of academic grades, demographic and social features like [6]
[18]	▪ Predicted student dropout by Early detection system (EDS) by using administrative data of university students of private and public sectors. ▪ integrated decision trees classifiers, regression analysis techniques, and neural networks models.	▪ Student Administrative data and historical student dropout data ▪ Dataset consists of features a birth year, gender, nationality, name, previous education data like unity entrance qualification, number of a student previously enrolled in, number of previous courses studies at the same university, type of study program, study mode, academic performance data name of the exam, exam result, Graduate or drops out.
[19]	▪ Predicted the successful secondary students to improve student pass rates and reduce dropout rates ▪ Optimized ML support vector machine model and decision tree classifier. ▪ Optimized SVM performed best with 92% accuracy	The publicly available dataset comprises 32 student attributes similar to [6]
[20]	▪ Analysed the student performance and identify the key factors that limit their abilities or performance to get improve their potential. ▪ Used student performance estimator, student progress indicators and student attribute descriptors. ▪ Student classification results are correct and meaningful.	▪ Dataset consists of academic data of 2 years
[21]	▪ Predicted students' future performance by using their current and past performance.	▪ The dataset includes the high school GPA and SAT scores of the students, the lectures and lab

	▪ A bilayer structure of clustering consists of ensemble and base predictor is used.	scores in each academic quarter, the course credits, and the obtained grades.
[22]	▪ Analysing the students' graduation performance at the end of the degree to improve the program. ▪ Classification is done by using decision trees and other classifiers to categorise low and high achieving students	▪ Dataset consists of students' pre-admission marks for university and the marks for all the courses that are taught in the four years of the degree programme,
[23]	▪ Prediction academic performance of the engineering students by classification ▪ integrated classifier consists of three complementary algorithms, namely Decision Tree, K-Nearest Neighbour, and Aggregating One-Dependence Estimators (AODE) ▪ The integrated model results reflected consistent accuracy.	▪ 18 attributes based on academic information and demographic information are used
[24]	▪ J48, PART, RIPOR ▪ decision trees, artificial neural networks, and other classification techniques ▪ J48 performed the best with an accuracy of 86.4%	▪ Academic data like average test scores, first and second semester grades, and Demographic data like Age, gender, origin, social class of 932 Engineering student
[25]	▪ Analysed admission and placement data to understand student progress and improve behaviour. ▪ K-mean Clustering is implemented in RapidMiner studio and scatter, bars, or histogram charts	▪ Dataset consists of Grade 10 and Grade 12 qualification year and marks, placements, age, Gender, skill sets, and backlog to which weights are assigned
[27]	▪ Analysed student academical and social dataset to improve the didactic contents design and the progress of the student ▪ C5.0 Decision tree induction and Naïve Baye are used as ML classifiers, where primary model demonstration 100% accuracy	▪ Dataset consists of 12 attributes including name, student number, former grades, attendance, assessment, laboratory work, living location, family size, parents' qualification, and gender
[28]	▪ Blended learning environment ▪ Analysed student access to different modules and events of LMS to predict students' achievement ▪ Random Forests and Support Vector Machines are used as ML classifiers.	▪ Dataset consists of log data based on the frequency of interaction with each event in each module in LMS i.e. Mean, Standard deviation and variance of each module and event combination