JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

STUDENT PERFORMANCE PREDICTION USING MACHINE LEARNING

A Comparative Study and Proposed Framework for Early Academic Intervention

Aditi Sachin Patil , Rutuja Kiran Mangale
Department of CSE (AIML),
New Institute of Technology, Kolhapur, Maharashtra, India
Mentor: Mrs. Neha Nazare (Assistant Professor)

Abstract : Various factors that impact student academic achievement are connected to each other, such as prior academic performance, attendance, evaluation activity, attitude to studies and the education environment. Getting rid of students who may have difficulties in studying can provide teachers with the opportunity to deliver their help on time. The paper aims at presenting the resume of the review carried out so far and a suggested model of the machine learning algorithm designed for forecasting student performance. The role of educational data mining, the characteristics of students commonly used in work, datasets available to researchers, supervised learning methods employed, as well as other measures is discussed in the paper. The review also deals with Decision Tree (DTree), Random Forest, Logistic Regression (LR), K-Nearest Neighbor (K-Nearest Neighbor) and regression methods, and offers a description of the work algorithm that includes data gathering, processing, feature selection, model training, validation, prediction and interpretation. Rather, it outlines the methodological criteria for ensuring an effective implementation, which comprises leakage prevention, dealing with class imbalance, fair evaluation, explain ability and privacy of students' data. The framework proposed aims to supplement educators' professional judgment, rather than replace it, and to promote early, evidence-based help of students.

IndexTerms - Student performance prediction, machine learning, educational data mining, learning analytics, classification, academic performance.

I. INTRODUCTION

Educational institutions gather academic documents, and increasingly, the data on students' digital learning activities. This data is used when analyzing the information to identify patterns in learning and understand what determines success in studies. Under Student Performance Prediction (SPP), statistical and machine learning methods are applied to predict the outcomes such as grades, pass/fail status, and risk category in the future. The predictions are valuable when they are made in advance and allow teachers to help students on time. Traditional assessment relies on exam scores that students get after finishing the course or semester. Traditional assessment can explain what status is reached by students but cannot let them know in time if they are lagging behind. The predictive assessment can utilize the information obtained during the learning process and include attendance, marks for assignments, previous grades of students, and learning platform activity. The task is not to label students but to identify what kind of help they will need in order to make better decisions in educational management.

Educational Data Mining (EDM) is the technique of utilizing data analysis and computational methods for handling educational data. The learning analytics employs learner data collection and analysis methods for better understanding and improvement of the learning process. According to Zhang et al., SPP is referred to as processes consisting of data collection, problem framing, model building, forecasting and application [1]. Staged view is useful as the success of a model depends not only on the algorithm but on factors like input data relevance, timing and quality. In this paper, the most relevant concepts are discussed for the implementation of the model of prediction of the performance of students, including the proposal of a reproducible framework. The paper includes the following sections: Section II discusses related works; Section III outlines theoretical framework; Section IV describes the research methodology and data; Section V outlines possible algorithms; Section VI discusses evaluation and validation; Section VII presents criteria for publication; Section VIII presents discussions of interpretation, ethics, and limitations; Section IX offers conclusion.

A. Problem Description

The key issue is predicting a student's expected performance based on data that is available ready before the actual performance is known. An effective model should generalize well to new students or future academic years, measure performance properly, and not use information that was unavailable during prediction.

B. Purpose

1) Discern academic and behavioral factors that can influence student performance.

2) Examine ML techniques for classification and regression.

3) Evaluate preparation and model evaluation stages.

4) Explore public datasets that help in repeatable experiments.

II. LITERATURE REVIEW

The study of predicting students' academic success has involved several factors such as demographic factors, academic performance, and behavioral characteristics using a range of classification and regression methods. There are variations in terms of student populations, education, definition of outcome, size of dataset, validation process in different studies and hence results cannot be compared directly without taking into consideration the similarity of datasets, targets and evaluation methods. Zhang et al. have reviewed the literature and analyzed methods of prediction of academic success and have categorized the process into data collection, problem formulation, learning, predicting and applying. They explain that prediction of students' performance can be treated as classification when the response is a category, regression when it is a number, or clustering when the intention is to find groups of students with common features [1].

A systematic review of literature produced between 2015 and 2021 was undertaken by Roslan and Chen. Their research followed the PRISMA protocol and involved examining 219 articles and analyzing 58. The findings suggest that academic records and demographic factors are usually taken into account in the various studies, supervised learning is often used to analyze the data and Decision Trees are one of the methods widely used in the area of study [2]. As for Cortez and Silva, they presented a novel study based on social, demographic, and school-connected factors affecting student performance in two secondary schools in Portugal [3]. The researchers used publicly available UCI Student Performance data set in their work, which can be a reliable data source since UCI is largely recognized in educational studies [7].

The dataset of Open University Learning Analytics (OULAD) was created by Kuzilek, Hlosta and Zdrahal, and it is based on 22 courses and 32,593 students, including demographic data and assessment results together with virtual learning environment usage summaries [4]. OULAD is useful for higher education learning analytics and opens opportunities to research not only academic achievements but also behavior in online learning. Chaka conducted an overview of reviews in the sphere of educational data mining and prediction tools [5].

Table I: Selected research studies

Study	Focus and contribution
Zhang et al. (2021)	Discusses the SPP problem formulation as well as its data, models, and applications.
Roslan & Chen (2022)	Carries out a review of 58 studies with an emphasis on the common nature of the usage of academic records, classification and Decision Trees.
Cortez & Silva (2008)	A study on data mining and a dataset from two secondary schools in Portugal.
Kuzilek et al. (2017)	Utilizes the OULAD dataset which presents information on students' characteristics, assessments, and activity at the VLE.
Chaka (2022)	Considering the publications in the sphere of educational data mining and prediction.

III. THEORETICAL BACKGROUND

Machine learning enables a computer system to learn a relationship between input variables and an outcome from examples. In student performance prediction, input variables describe a student's academic history or learning activity, while the target variable represents a grade, score, pass/fail result or risk category. The model learns a mapping from the input feature vector to the target.

A. Classification

Classification is used when the target contains a finite set of labels. For example, students may be grouped into pass and fail categories, or into low, medium and high performance categories. The model estimates a class or class probability. The class thresholds must be defined before evaluation and should have an educational justification.

B. Regression

The method for determining both the target and predictor variables is called regression. In this case, continuous predictive values are produced. An observation made through regression analysis is that the estimates are much more precise than in case of simple score estimation but should still be taken cautiously.

C. Types of Features

There are four broad types of features: underlying academic features (including scores and grades); engagement features such as attendance record, online date visits, and submission timing; background data whenever available, including age and history of individuals or organizations; and lastly, features related to the course variables including its difficulty and ways of assessment. The types of available feature depend on the initial dataset type and predicting time period.

D. Association Between Constructs

The theory supposes that educational results can be linked to prior performance, learning engagement and other conditions of the course besides which these links may be expected rather than automatically establish causation of one variable over another.

For instance, attending classes less frequently may correlate with lower performance, but the correlation does not mean that not attending classes leads to a specific outcome.

$$y = f(X) + \varepsilon$$

In this case, X denotes the selected inputs regarding student and course, y refers to the expected output, f means the prediction function while ε indicates unexplained variability. If it is a classification case, f finds a class or probability of a class; if it is a regression case, f gives an estimate.

E. Anticipating Early

Predicting at the right time is a significant theoretical and practical problem. It is impossible to issue early-term alerts based on a model created using final grades because final exam scores are not available. The prediction moment must be defined, and features that can be used at this moment should be taken into consideration. This avoids target leakage and allows for relevant outcomes for practical educational intervention.

IV. RESEARCH METHODOLOGY

This paper proposes a student performance prediction system and describes how this system can be developed. This is a framework to enable future development, not a description of an already completed work. The given datasets would have to be worked on to perform various operations such as model training and hyperparameter tuning before results could be posted.

A. Research Design

This study will employ a quantitative, supervised-learning research design. The data used in this study will be cleaned and labeled. From the cleaned and labeled data, a subset will be created for training and a different subset will be created for testing. The algorithms will be trained on the training data. The models will be evaluated on the testing data. The data will be ordered in the chronological order of the academic year. Splitting the data randomly would test if the model is able to predict the performance of students in a different academic time period.

B. Dataset Selection

We have two primary publicly available datasets from which to work, dependent upon the research question. The UCI Student Performance dataset contains records from two Portuguese secondary schools, including various student and family background and school and academic history variables, and may include the final grade of the students, or be derived from grades [3], [7]. OULAD contains 22 course presentations and 32,593 students. Some of its tables contain personal information of students, registrations, assessments and a daily record of a student's virtual learning environment interactions [4]. The two datasets, however, represent different populations and may not be merged without an integrating rationale.

Table II: Dataset comparison

Dataset	Population and contents	Suitable use
UCI Student Performance [7]	Secondary-school students from two Portuguese schools; background and academic attributes.	Grade prediction and school-level classification.
OULAD [4]	Presentations of Open University courses; demographics, assessments and VLE activity.	Higher education and online learning prediction.

C. Data Preprocessing

During data preprocessing, we identify inconsistencies with respect to types of variables, missing values, duplicates, out-of-range values and categories. For instance, we can scale numerical variables for algorithms like KNN and SVM where the algorithms are affected by feature magnitude. For categorical variables, we have one-hot and ordinal encodings to name a few. Imputation and scaling must be performed based on the training data, and the same parameters must be applied on validation and testing data. The target variable is defined explicitly. Class labels must also be defined with respect to classification problem. If classes are imbalanced during training, class weights must be defined, or dataset must be resampled. The test data must be defined in a way that it retains the natural classes in the dataset.

D. Feature Selection

Feature selection can reduce noise and enhance model interpretability and efficiency. It can also limit the collection of personally identifiable information. Feature selection can use domain knowledge, analysis of correlations, statistical tests or model-based importance. Feature selection should take place in the training pipeline to avoid the test set information leaking and affecting the model.

V. MACHINE LEARNING ALGORITHMS

A good comparison should examine different algorithms, assumptions and levels of complexity. The 'best' model may not be identified in advance in every dataset. The impacts of sample size, missingness, class balance, features, definition and validation of targets should be considered.

A. Decision Tree

Decision Tree predicts an outcome by following a chain of rules based on features. It is easy to understand and can handle relationships that are not straight lines. If a tree becomes too deep it may learn the training data too well and not work well on new data. You can limit the depth, set a minimum number of samples for each leaf and prune the tree to keep its complexity in check.

B. Random Forest

Random Forest builds Decision Trees using different subsets of data and features. By combining the predictions of these trees, Random Forest can capture complex nonlinear patterns. It is often a strong starting point for tabular data. Its decision process is less clear than that of one tree, so you may need methods that show which features are important or explain why a decision was made.

C. Logistic Regression

Logistic Regression estimates the chance that a categorical outcome will happen, usually a simple yes or no label. It is quick to compute and gives coefficient values that help explain how each predictor is linked to the result. It assumes a straightline relationship between the predictors and the log odds unless you add interactions or transform the variables.

D. Support Vector Machine (SVM)

Support Vector Machine finds a boundary that separates classes with the largest possible margin. Kernel functions let SVM draw boundaries in high-dimensional space. SVM can perform well on complex data, but it often needs the features to be scaled and careful choice of kernel and regularization settings.

E. K-Nearest Neighbor (KNN)

K-Nearest Neighbor assigns a label based on the labels of the training examples closest to it. It is simple and can capture local patterns, but it is very sensitive to how the features are scaled, to irrelevant variables, and to the choice of the number k. As the training set grows, predictions can become slower.

F. Linear Regression

Linear Regression is used when the target is a numerical score. It offers a baseline and shows how input variables relate to the numeric outcome. If the pattern is not a straight line, you may need to transform the data or use a more flexible model.

Table III: Candidate algorithms and issues

Algorithm	Primary function	Issue with the algorithm
Decision tree	Classification/Regression	Easy to interpret; tends to overfit.
Random forest	Classification/Regression	Robust ensembles; less interpretability.
Logistic regression	Classification Simple method	Assumes linear relationship of the log odds.
Support vector machine	Classification/Regression	Scales and parameters matter.
KNN	Classification/Regression	Dependent on distance measures and feature scaling.
Linear regression	Regression Interpretable	Assumes linear relationship.

A starting model needs to be created before adjusting algorithms. When choosing a model it is important to look at how it predicts, how easy it is to understand, how well it matches real data, how much computing power it uses and what happens if it makes wrong predictions. A model that is just a little more accurate might not be good enough if teachers can't figure out what it is saying or use its predictions properly.

VI. MODEL TRAINING AND EVALUATION

A. Training and Validation

The dataset needs to be divided into training data, validation data and test data or evaluated using cross-validation when the dataset is small. The training data is used to adjust model parameters. The validation data helps with choosing the model and setting the right hyperparameters. The test data is kept for a fair estimate and should only be used after all decisions about model development are made. For data that comes from the student more than once, the way the data is split should make sure that data from the same person does not end up in both the training data and the test data when the goal is to predict results for new students. If the goal is to predict results for students who are already known, a split based on time might be better to match how the model will be used. The way the data is split must match the intended use.

B. Classification Metrics

Accuracy is the proportion of all predictions that are correct. Precision measures the proportion of predicted cases that are truly positive. Recall measures the proportion of positive cases correctly identified. The confusion matrix shows true positives, true negatives, false positives and false negatives.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

In an early-warning system, recall for the at-risk class is especially important because failing to identify a student who needs help can have consequences. However, many false alarms can create unnecessary stress for both educators and students. That is why it is essential to report precision, recall, F1-score and the confusion matrix together. Relying on accuracy alone can be misleading and does not give a complete picture of how well the system works.

C. Regression Metrics

Mean Absolute Error (MAE) tells how far the predicted scores are from the actual scores on average. Root Mean Squared Error (RMSE) gives weight to big mistakes. The coefficient of determination (R^2) shows how much of the change in the scores the model can explain compared to a baseline guess. All of these metrics should be written in the same units as the scores when that makes sense.

$$MAE = (1/n)\ \Sigma\ |y_i - \hat{y}_i|$$
$$RMSE = \sqrt{[(1/n)\ \Sigma\ (y_i - \hat{y}_i)^2]}$$

D. Model Comparison and Reproducibility

All candidate models must use the same data split and target definition. Hyperparameter search should be limited to the training and validation data. A final comparison table should include the chosen values, confidence intervals or variation across folds where possible, and the size of the test set. The software environment, random seed, feature list and preprocessing pipeline should be documented to ensure reproducibility. No numerical results are included in this review because no dataset was trained or evaluated for this paper. Any future implementation should only add measured results after the full experiment has been completed. Accuracy values taken from datasets or papers must not be presented as results of the proposed project.

VII. RESEARCH ARTICLE QUALITY AND CONTRIBUTION

A. Relevance of Title and Research Article

The title clearly shows the research task and method area. The article stays focused on predicting student outcomes with machine learning, not on general artificial intelligence in education. The scope lists the prediction target, the data needed, possible algorithms and the evaluation process.

B. Scholarly Quality

The discussion uses peer-reviewed research and recognized public dataset documentation. The literature review tells the difference between a review, a dataset descriptor and an empirical data-mining study. Claims about dataset size and population link back to the sources. This clear distinction is needed for accuracy.

C. Continuity in the Article

The paper follows a logical sequence: the introduction defines the problem; the literature review shows prior work; the theoretical part explains classification and regression; the methodology describes data and preprocessing; the algorithm part lists candidate models; and the evaluation part shows how performance should be measured. The conclusion naturally follows from these sections.

D. Use of Theory

The paper uses supervised learning concepts, classification, regression, feature representation and generalization. The paper also separates prediction from explanation. These concepts give a base for choosing the target features, model and evaluation protocol.

E. Novelty and Originality

The contribution of this paper is a consolidated implementation-oriented framework that links dataset choice, prediction timing, leakage prevention, evaluation and responsible use. The contribution is not claimed as a novel algorithm or a finished empirical discovery. A stronger original contribution would need a documented experiment, a defined student population, a justified feature set and a comparison against established baselines.

F. Technical Contents and Correctness

Technical correctness requires the algorithm to match the target type, preprocessing to be fitted on training data, leakage to be prevented, metrics to fit the class distribution, and the test protocol to be fully reported. The paper separates the UCI secondary-school dataset from OULAD's higher-education context and does not mix their statistics.

G. Understanding and Illustrations

The proposed workflow can be shown as: records → data validation → preprocessing → feature selection → train/validation split → model training → test evaluation → explanation → educator-supported intervention. The described sequence shows the path from raw data to responsible use. A figure or implementation diagram should use these stages and clearly label the prediction point.

H. Critical Qualities

The paper acknowledges that model accuracy alone does not establish value. Data imbalance, dataset representativeness, changes in course practices, proxy variables, privacy, and the consequences of predictions must be considered. Predictive associations should not be interpreted as causal effects.

I. Quality of References

References include research articles, a systematic literature review, and official dataset documentation. Each dataset description is attributed to its source. Bibliographic details and DOI links are supplied where available.

J. Clarity of Conclusions

The conclusion summarizes the proposed framework and states its limitations. It does not claim that one algorithm is universally best or that a particular accuracy has been achieved. The next step is an empirical evaluation on a suitable dataset.

VIII. DISCUSSION, INTERPRETABILITY AND ETHICAL CONSIDERATIONS

A. Interpretation of Predictions

A prediction only becomes helpful when its meaning fits the surrounding situation. Global feature importance can show which variables sway model predictions across a dataset, while local explanations can highlight factors linked to a single prediction. These explanations serve as descriptions of model behavior; they are not guarantees that altering one feature will automatically change a student's outcome. For example, a model might give a high-risk score to a record that shows low assessment completion and low attendance. Educators should double-check the information and talk with the student about potential barriers before taking action. The system should suggest review and support rather than immediate punishment, exclusion, or cutting back on opportunities.

B. Fairness and Bias

Historical educational data often mirror access to resources, varied teaching conditions, or structural disadvantages. A model can copy these patterns if they exist in the training data. Therefore, performance should be checked across student groups in ways that are lawful and ethical. Differences in false-negative rates deserve thorough investigation, and institutions should document how sensitive attributes are managed.

C. Privacy and Data Governance

Student records are sensitive educational information. Data collection must stay limited to variables that are truly needed for an educational purpose. Access controls, secure storage, de-identification, retention limits, and clear governance procedures should all be put in place. Institutions must follow all privacy and education policies and clearly explain to students how their data are used.

D. Limitations

There are several limitations in the proposed framework. First, it does not include experimental results. Second, public datasets may not match the specific student population, curriculum, or assessment system of an institution. Third, relationships that a model learns from data can shift over time. Fourth, some datasets contain variables that will not be available at the exact moment when predictions are needed. Finally, even a model that performs well cannot capture every social or institutional factor that affects student learning.

E. Future Scope

Future work should build the pipeline with a chosen dataset and explicit prediction target. It should compare ensemble models, use cross-validation or time-based testing, report classwise measures, and check calibration. Explainability techniques should be tested with educators to verify if the explanations are clear and actionable. A pilot deployment must monitor model drift and educational outcomes while maintaining human oversight.

F. Practical Application

A student-support dashboard should display a risk category, the date of the prediction, factors shaping the estimate, and a suggested review action. It must not present the prediction as a fixed judgment. Educators must be able to correct data, add contextual notes, and override the model if needed. The system should ultimately be judged by whether it delivers effective support, not just by its statistical score.

IX. CONCLUSION

Student performance prediction using machine learning is a vital application of educational data mining and learning analytics. By analyzing academic records and suitable learning-behavior indicators, predictive models can help identify students who may benefit from early academic support. The literature shows the use of classification and regression approaches across a range of algorithms including Decision Trees, Random Forests, Logistic Regression, SVM, KNN, and neural network methods. This paper has presented a comprehensive review and proposed framework covering problem definition, dataset selection, preprocessing, feature selection, model development, validation, and evaluation.

The UCI Student Performance dataset and OULAD provide valuable research settings and should be selected according to the intended student population and prediction objective. A reliable implementation must define when the prediction is made and prevent the use of information that would not be available at that time. Accuracy should be supplemented by precision, recall, F1-score, or regression error measures depending on the target. Finally, privacy, fairness, explainability, and educator oversight are essential for practical deployment. Machine learning should support informed decisions and timely intervention while final decisions remain grounded in human understanding of each student's circumstances.

ACKNOWLEDGMENT

The authors thank the researchers and organizations that have provided data-mining studies and public datasets for academic research.

REFERENCES

[1] Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai and X. Shang, "Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis," Frontiers in Psychology, vol. 12, Art. No. 698490, 2021. Doi: 10.3389/fpsyg.2021.698490.

[2] M. H. Bin Roslan and C. J. Chen, "Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015–2021)," International Journal of Emerging Technologies in Learning, vol. 17, No. 5, pp. 147–179, 2022. Doi: 10.3991/ijet.v17i05.27685.

[3] P. Cortez and A. M. G. Silva, "Using Data Mining to Predict Secondary School Student Performance," in Proc. 5th Annual Future Business Technology Conference, 2008.

[4] J. Kuzilek, M. Hlosta and Z. Zdrahal, "Open University Learning Analytics Dataset," Scientific Data, vol. 4, Art. No. 170171, 2017. Doi: 10.1038/sdata.2017.171.

[5] C. Chaka, "Educational Data Mining, Student Academic Performance Prediction, Prediction Methods, Algorithms and Tools: An Overview of Reviews," Journal of E-Learning and Knowledge Society, 2022. Doi: 10.20368/1971-8829/1135578.

[6] "Academic Performance Prediction Using Machine Learning Approaches: A Survey," IEEE Transactions on Learning Technologies, vol. 18, pp. 351–368, 2025. Doi: 10.1109/TLT.2025.3554174.

[7] P. Cortez, "Student Performance [Dataset]," UCI Machine Learning Repository, 2008. Doi: 10.24432/C5TG7T.