IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 6, February 2025 Impact Factor: 7.67

Student Performance Using Machine Learning Techniques

Siddhi Gharat[1] and Udit Patil[2]
Assistant Professor, Department of IT[1]
Student, P.G. Department of IT[2]
Veer Wajekar ASC College, Phunde, Uran

Abstract: *The ultimate aim of any educational institution is to deliver the best learning experience and knowledge to its students. Identifying students in need of academic support early and taking timely measures to enhance their performance is critical to achieving this goal. This research utilizes four machine learning techniques—Artificial Neural Network (ANN), Naïve Bayes, Decision Tree, and Logistic Regression—to develop a classifier that predicts student performance in a Computer Science course offered by Al-Muthanna University (MU), College of Humanities. Special emphasis is placed on the impact of internet usage for academic purposes and time spent on social networks on student performance. Performance is evaluated using ROC index, classification accuracy, error rate, precision, recall, and F-measure. The dataset, comprising 161 student records collected via surveys and gradebooks, indicates that ANN outperforms other models with a ROC index of 0.807 and an accuracy of 77.04%. Decision Tree analysis identifies five key predictors of performance: early computer grades, accommodation, interest in the subject, educational environment satisfaction, and residence.*

Keywords: Student Performance Prediction, Artificial Neural Network, Naïve Bayes, Decision Tree, Logistic Regression, Educational Data Mining

I. INTRODUCTION

Education plays a fundamental role in shaping the economic and social frameworks of any nation. In many countries, including Iraq, governments invest heavily in providing free or subsidized higher education. However, the failure of students to graduate on time incurs substantial additional costs to both governments and families. One effective way to mitigate this is by leveraging machine learning (ML) techniques to predict student performance and proactively identify at-risk students.

Students' academic success is influenced by numerous factors including GPA, psychological state, family background, learning habits, and social environment. Modern ML approaches offer powerful tools for processing such diverse data and drawing insights. In this study, we explore the effectiveness of four popular ML models—ANN, Naïve Bayes, Decision Tree, and Logistic Regression—to build predictive models based on a dataset from Al-Muthanna University. Our dataset incorporates novel attributes related to internet usage and time spent on social media.

II. LITERATURE REVIEW

Predicting student academic performance using data mining and ML is a widely researched domain. Shahiri et al. [3] provided a comprehensive review on prediction methods, emphasizing that GPA, demographic data, and psychological traits are strong indicators. Xu et al. [1] applied progressive prediction algorithms on large datasets, achieving high accuracy using Logistic Regression and Random Forests. Similarly, Guleria et al. [5] demonstrated the effectiveness of Decision Trees using attendance and sessional performance data.

Arsad et al. [6] developed a Neural Network model (NNSPPM) that achieved high prediction accuracy using academic scores. Gray et al. [8] explored the role of aptitude, personality, and learning strategies using models such as SVM and Naïve Bayes, finding that SVM performed best. Buniyamin et al. [9] and Alharbi et al. [10] implemented ensemble approaches, reflecting that no single model consistently outperforms others across datasets.

ISSN 2581-9429 IJARSCT

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 6, February 2025 Impact Factor: 7.67

Table 1 provides a summary of related studies, outlining dataset sizes, features, and best-performing algorithms.
Proposed System 3.1 System Components Figure 1 outlines the system components: data collection, preprocessing, model training, and evaluation. The input dataset, sourced from student surveys and gradebooks, is normalized and labeled. The ML algorithms train on this dataset, producing predictive models evaluated via various metrics.

II. METHODOLOGY

3.1 Artificial Neural Network (ANN)

ANNs mimic biological neural networks and are adept at handling non-linear relationships. A three-layer feedforward ANN was used in this research with 20 input neurons, two hidden layers (6 and 3 neurons respectively), and one output neuron. The Rectified Linear Unit (ReLU) was employed as the activation function [16]. The network was trained using backpropagation for 100 epochs, with hyperparameters optimized using grid search.

3.2 Logistic Regression (LR)

LR models the probability of a binary outcome based on input features using the logistic function: P(D=1|X) = 1 / (1 + e^-(α + ∑βiXi))
Parameters α and β were optimized using gradient descent [17]. The model is particularly useful when interpretability and linear relationships are desired.

3.3 Naïve Bayes (NB)

NB is a probabilistic classifier based on Bayes' theorem, assuming independence among features: Vmax = argmax_vj P(vj) ∏_i P(ai|vj) Though simplistic, NB is computationally efficient and often performs well on high-dimensional data [18].

3.4 Decision Tree (DT)

DT models classify data through a series of splits based on feature values. The model built in this study used information gain and pruning techniques. Popular in domains requiring interpretability, DTs have been widely applied in education, healthcare, and finance [16].

IV. EXPERIMENT

4.1 Dataset

Data was collected from students of the Archeology and Sociology departments at Al-Muthanna University during 2015-2016. The dataset includes 161 records (76 males, 85 females), covering 20 attributes categorized into personal, academic, family, lifestyle, and environmental factors (Table 2).
Each student is labeled as 'Good' or 'Weak' based on their final grade in a Computer Science course. Students scoring below 60% were labeled as 'Weak'. Early indicators like average grades from the first two exams were included to identify at-risk students early. Additional factors such as employment, marriage, and use of social media were also considered.

4.2 Data Preprocessing and Tools

All features were normalized using min-max scaling. RapidMiner Studio was used for model training and evaluation. A 3-fold cross-validation approach ensured that results were not biased by specific data partitions.

4.3 Evaluation Metrics

Models were evaluated using:

- Accuracy
- Classification Error

Precision, Recall, F-Measure: F = 2 * (Precision * Recall) / (Precision + Recall)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 6, February 2025 Impact Factor: 7.67

ROC Index: AUC = Σ_i (FPR[i] - FPR[i-1]) * (TPR[i] + TPR[i-1])/2
A ROC index above 0.7 indicates a strong classifier [21].

4.4 Model Implementation

Each model was tuned using grid search:

- ANN used ReLU activation, 100 epochs, learning rate and L2 regularization optimized.
- DT tuned for splitting criterion, minimum node size, and pruning.
- LR optimized for solver method and regularization.
- NB tuned for Laplace correction, kernel functions, and grid size.

4.5 Results

Table 3 summarizes the performance metrics:

Model	TP	FP	TN	FN	Precision	Recall	F1	Accuracy	Error	ROC Index
ANN	67	18	57	19	79.17%	77.92%	78.47%	77.04%	22.96%	0.807
Decision Tree	67	19	56	19	77.96%	77.83%	77.88%	76.93%	23.61%	0.762
Logistic Reg.	62	17	58	24	79.23%	71.91%	74.87%	74.53%	25.47%	0.767
Naïve Bayes	55	23	52	31	70.51%	64.27%	67.21%	66.52%	33.48%	0.697

ANN outperformed other models in all evaluation metrics, followed closely by Decision Tree and Logistic Regression. Naïve Bayes showed the weakest performance.
The Decision Tree identified five key attributes impacting performance:

- Computer Grade (Course 1)
- Accommodation
- Interest in studying Computer
- Educational Environment Satisfaction
- Residency

These can inform early interventions by faculty to improve student outcomes.
Conclusion Predicting student performance allows academic institutions to implement early interventions for at-risk students. This study compared four ML models using real-world data from Al-Muthanna University. ANN provided the most accurate predictions with a 77.04% accuracy and a ROC index of 0.807. Decision Trees revealed the most influential factors, providing actionable insights for faculty and administrators. Future research may involve ensemble learning or deeper neural networks to further improve accuracy.
Conflict of Interest: None declared.

REFERENCES

[1] Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. IEEE Journal of Selected Topics in Signal Processing, 11(5), 742-753.
[2] Shaleena, K. P., & Paul, S. (2015). Data mining techniques for predicting student performance. ICETECH 2015.
[3] Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science.
[4] Meier, Y., Xu, J., Atan, O., & Van Der Schaar, M. (2016). Predicting grades. IEEE Transactions on Signal Processing, 64(4), 959-972.
[5] Guleria, P., Thakur, N., & Sood, M. (2015). Predicting student performance using decision tree classifiers and information gain. PDGC 2014.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 6, February 2025 Impact Factor: 7.67

[6] Arsad, P. M., Buniyamin, N., & Manan, J. L. A. (2013). A neural network students' performance prediction model. ICSIMA 2013.
[7] Li, K. F., Rusk, D., & Song, F. (2013). Predicting student academic performance. CISIS 2013.
[8] Gray, G., McGuinness, C., & Owende, P. (2014). An application of classification models to predict learner progression in tertiary education. IACC 2014.
[9] Buniyamin, N., Bin Mat, U., & Arshad, P. M. (2016). Educational data mining for prediction and classification. ICEED 2015.
[10] Alharbi, Z., Cornford, J., Dolder, L., & De La Iglesia, B. (2016). Using data mining techniques to predict students at risk. SAI Computing Conference.
[11] Livieris, I. E., Drakopoulou, K., & Pintelas, P. (2012). Predicting students' performance using artificial neural networks.
[12] Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student academic performance prediction using Decision Tree and Fuzzy Genetic Algorithm. Procedia Technology.
[13] Arsad, P. M., Buniyamin, N., & Manan, J. L. A. (2014). Neural Network and Linear Regression methods for prediction of students' academic achievement. EDUCON.
[14] Sarker, F., Tiropanis, T., & Davis, H. C. (2014). Linked data, data mining and external open data for better prediction of at-risk students. CoDIT 2014.
[15] Huang, S., & Fang, N. (2012). Early prediction of students' academic performance. FIE.
[16] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
[17] Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text (3rd ed.). Springer.
[18] Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques.
[19] Muñoz-Bullón, F., Sanchez-Bueno, M. J., & Vos-Saz, A. (2017). The influence of sports participation on academic performance. Sport Management Review.
[20] CDC (2010). The Association Between School-Based Physical Activity and Academic Performance. Atlanta, GA.
[21] Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press.