Student Performance Prediction Using Machine Learning

"A Data-Driven Approach to Academic Success Forecasting"

Author: Korada Vamsi[1] (MCA student), Dr. Bharati Bidikar [2] (Adjunct.Professor) 1,2

Department of Information Technology & Computer Applications, Andhra University

College of Engineering, Visakhapatnam, AP.

Corresponding Author: Korada Vamsi

(email-id: vamsi41202@gmail.com)

****

Abstract- *This study proposes a technically robust machine learning framework for predicting student academic performance, specifically targeting the final grade (G3), by leveraging the Portuguese student dataset. The approach incorporates comprehensive data preprocessing techniques, including categorical encoding, feature scaling, and dimensionality reduction via Principal Component Analysis (PCA), to enhance data quality and model performance. A Multi-Layer Perceptron Regressor (MLPRegressor) is employed to model the complex, non-linear relationships between academic, demographic, and behavioral input features and student outcomes. The system is deployed through a web-based interface developed using Flask, integrated with Gradio for interactive and real-time prediction capabilities. Additionally, an API layer is incorporated to support seamless integration with external platforms. This architecture supports data-driven educational interventions by enabling early identification of at-risk students and aiding institutions in implementing targeted academic support strategies.*

Index Terms: *Student Performance Prediction, Machine Learning, MLP Regressor, Educational Data Mining, Academic Analytics*

1.INTRODUCTION:

Utilizing their prior academic achievements and other pertinent traits, such as demographic and behavioral information, this study presents a machine learning-based approach for forecasting students' final grades. The system employs the MLP Regressor model, which is aided by preprocessing techniques such as scaling and PCA for dimensionality reduction. Flask and Gradio are used to create a user-friendly interface that allows for real-time, interactive predictions, making the application useful for teachers in identifying and helping pupils who are at risk of underperforming.

Statistical models like linear and logistic regression were used in early studies to forecast student achievement, but they were simple and unable to deal with complicated data patterns. Decision Trees, SVMs, and Random Forests are examples of models that have improved prediction accuracy thanks to improvements in machine learning, but they also face scalability issues and need to be carefully tuned. Although they frequently lacked interpretability and needed more resources, deep learning models like RNNs and MLPs made additional improvements by capturing non-linear and sequential patterns.

Compared to other models and methods like PCA, hybrid approaches offered superior performance and flexibility. Although predictive models have been used in real-world applications like online learning platforms and intelligent tutoring systems, they frequently operate within limited data ranges.

Using an MLP Regressor with PCA and a user-friendly interface, this study builds on these advancements to provide a reliable and scalable prediction of student grades.

RESEARCH OBJECTIVES

- Create a precise machine learning model for forecasting students' ultimate grades (G3).
- Use efficient data preprocessing and dimensionality reduction methods.

- Use serialization to guarantee that the model may be reused.

- Assist teachers in identifying vulnerable pupils early so that they may intervene in a timely manner.

2. METHODOLOGY:

Using machine learning, this study employs a methodical methodology to create a predictive system that predicts student academic achievement. To guarantee precise, efficient, and real-time predictions, the technique combines data preparation, model creation, and implementation tactics.

Data Collection and Understanding:

The Portuguese student performance dataset, which contains academic records, demographic data, and behavioral data, is the one utilized in this study. Age, gender, school kind, study time, and first and second period grades (G1 and G2) are some of the key variables used as input. We aim to forecast the student's final score, which is represented by G3.

Data Preprocessing:

A number of preprocessing procedures are used to guarantee the data's quality and appropriateness for machine learning:

- Treatment of Missing Data: To ensure data consistency, all missing data are treated using statistical imputation, such as the mean or mode.
- Categorical Encoding: Using Label Encoding and One-Hot Encoding, features like gender and school type are transformed into numerical form.
- Feature Scaling: To guarantee consistency in feature ranges and improve model performance, numerical features are normalized using standardization methods.
- Dimensionality Reduction: To reduce the number of input variables while retaining the most informative components, Principal Component Analysis (PCA) is utilized. This improves computational efficiency while reducing overfitting.

Model Development:

The Multi-Layer Perceptron Regressor (MLP Regressor), a neural network architecture that is ideal for regression applications, is used to construct the prediction model.

- Input Layer: Receives the preprocessed and dimensionally-reduced feature set
- Hidden Layers: There may be one or more hidden layers that use activation functions like ReLU to learn complicated patterns.

A subset of the data is used to train the model, and cross-validation methods are used to optimize hyperparameters like the learning rate, number of neurons, and number of iterations.

Model Evaluation:

The trained model is evaluated using a test dataset to assess its accuracy and generalization capability. Evaluation metrics include:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R-squared (R^2) Score

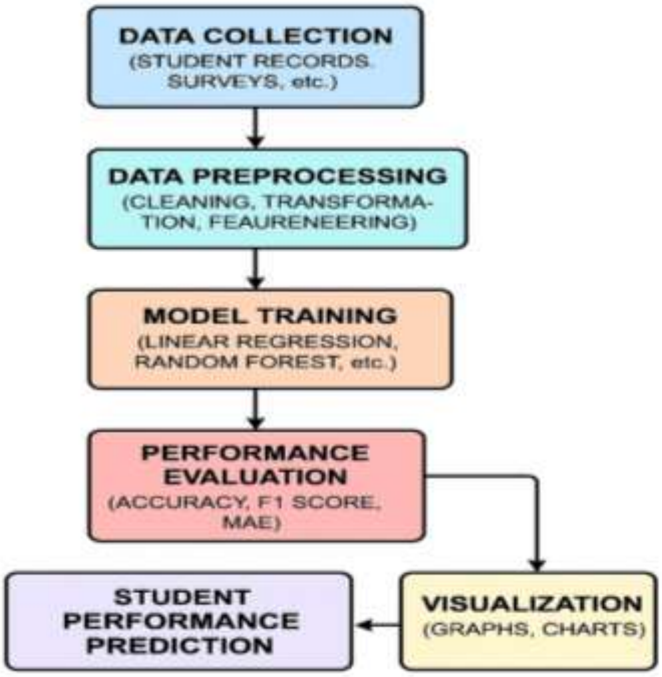

DATA COLLECTION
(STUDENT RECORDS SURVEYS, etc.)
DATA PREPROCESSING
(CLEANING, TRANSFORMA-TION, FEAURENEERING)
MODEL TRAINING
(LINEAR REGRESSION, RANDOM FOREST, etc.)
PERFORMANCE EVALUATION
(ACCURACY, F1 SCORE, MAE)
STUDENT PERFORMANCE PREDICTION
VISUALIZATION
(GRAPHS, CHARTS)

Fig 1: Architecture of Student Performance Prediction

System Implementation:

The prediction system includes the following components to improve usability:

- **Flask Web Interface:** Enables users to enter student information using a straightforward web form and get dynamic predictions.
- **Gradio Interface:** Provides a live, intuitive interface for experimenting with different input combinations in the model.

- **Model Persistence:** To guarantee reusability without retraining, the trained model, the scaler, and the PCA transformer are saved using joblib.
- **Prediction API:** This API offers programmatic access to the prediction system, making it ideal for integration with other software or instructional frameworks.

Summary of the Workflow:

- Raw student data is obtained via forms or APIs.
- Preprocessing: The data is reduced using PCA after it has been cleaned, encoded, and scaled.
- Model Prediction: The MLP Regressor model receives the processed data.
- Model Reuse: Saved model components enable quick future predictions without retraining.

With this approach, the resulting system is guaranteed to be accurate, usable, and expandable, as well as capable of making predictions. A solid framework for forecasting student performance in real-world educational settings is created by combining preprocessing, dimensionality reduction, and a regressor based on a neural network.

3.RESULTS AND DISCUSSION:

The efficacy of the predictive system was assessed by its capacity to predict students' final grades (G3) using their demographic and academic data. The model's performance and its contribution to the objective of early academic intervention and decision-making support are summarized in the following results.

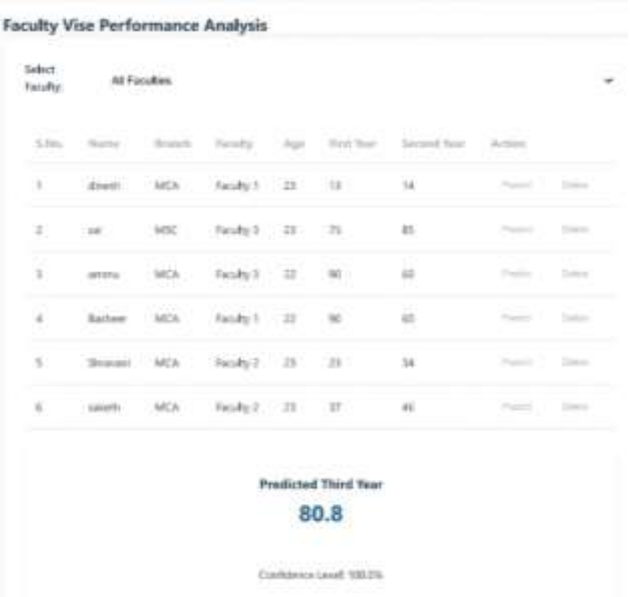
Model Performance

The model was trained using the Multi-Layer Perceptron Regressor (MLPRegressor) after preprocessing the student data and performing Principal Component Analysis (PCA). The model was assessed using the following parameters:

- Mean Absolute Error (MAE): Calculated the average of the absolute differences between predicted and actual values.
- Mean Squared Error (MSE): Emphasized the average squared differences, giving more importance to larger mistakes.

Real-Time Prediction Interface:

A Gradio interface and a Flask-based web application were created to make the model available to educators. With these tools, users could enter student information and get instant forecasts without requiring any prior technical knowledge. Particularly the Gradio interface offered a fluid and visual user experience, enabling users to test various input values and see how they affected the overall grade forecast.



Faculty Vise Performance Analysis

Select Faculty: All Faculties

S.No.	Name	Branch	Faculty	Age	First Year	Second Year	Actions
1	[illegible]	MCA	Faculty 1	23	13	14	Predict Delete
2	sai	MSC	Faculty 3	23	75	85	Predict Delete
3	amma	MCA	Faculty 3	22	90	60	Predict Delete
4	Basheer	MCA	Faculty 1	22	90	65	Predict Delete
5	Shravani	MCA	Faculty 2	23	23	34	Predict Delete
6	[illegible]	MCA	Faculty 2	23	37	46	Predict Delete

Predicted Third Year
80.8
Confidence Level: 100.0%

Fig 2: Performance Analysis

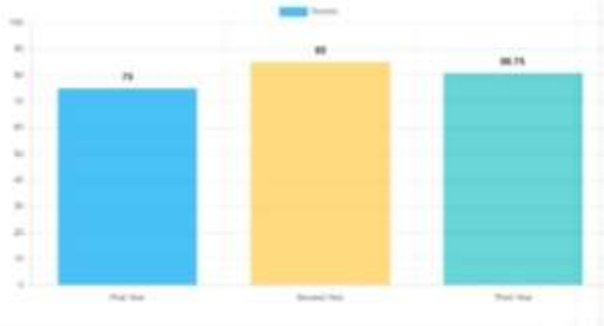


Fig 3: Predicted values of a student Performance

Model Reusability and Integration

The joblib package was used to store the trained MLP model as well as the PCA and scaler components, guaranteeing that the system will be able to generate future predictions without needing to be retrained. As a result, the system's run time was significantly shortened, making it appropriate for use in actual classroom environments. Additionally, an API endpoint was established to facilitate integration with outside systems like Learning Management Systems (LMS), which will help the tool become more widely used.

Interpretation and Educational Implications

The hypothesis that machine learning can be used to accurately predict student performance when sufficient preprocessing and feature selection are used is supported by the data. Prior grades (G1 and G2) are strongly correlated with the end result (G3), which is consistent with earlier studies in educational analytics. The real-time tool also gives teachers the ability to identify students who might require more assistance and tailor interventions to their specific needs.

Educators may better comprehend the relative importance of each feature by visualizing and testing inputs through the Gradio interface. By offering specific advice, this knowledge helps pupils make better decisions and improve their learning experience.

4.CONCLUSION:

By analyzing a combination of academic, demographic, and behavioral characteristics, this study proposes a machine learning-based method for precisely forecasting student academic achievement. The Multi-Layer Perceptron Regressor (MLP Regressor) model, which successfully captures intricate patterns in the data to predict ultimate student grades (G3), serves as the foundation for the system. By employing meticulous data preparation techniques, such as encoding, scaling, and dimensionality reduction using PCA, the system guarantees greater performance and lower model complexity.

The system combines an interactive Gradio interface with a user-friendly online application made with Flask to encourage usability and real-world use. These platforms allow teachers to interact with the model naturally, even if they don't have any technical knowledge, and provide real-time predictions. Furthermore, the system's reusability and scalability for broader use in educational institutions are guaranteed by model persistence using joblib and the availability of an API.

This project demonstrates that predictive analytics may offer significant insights into student achievement if used carefully. Teachers can provide timely assistance and use individualized learning techniques by identifying at-risk students early. Consequently, this system improves the whole educational experience and results in addition to monitoring academic performance.

5.REFERENCES:

1. Yadav, S. K., & Pal, S. (2012). Machine learning models were used to predict how students perform academically, focusing on decision tree techniques. International Journal of Computer Science and Engineering, 4(6), 723–729.

2. Cortez, P., & Silva, A. (2008). The study applied neural networks and support vector machines to predict secondary school grades, utilizing demographic and performance-related data. Proceedings of the European Conference on Artificial Intelligence in Education, 1–6.

3. Al-Barrak, M. A., & Al-Razgan, M. (2016). Compared multiple classification techniques like k-NN and Naive Bayes to estimate student academic results.Appeared in the International Journal of Advanced Computer Science and Applications, Volume 7, Issue 5, on pages 72 through 79.

4. Kabakchieva, D. (2013). Evaluated various classifiers to determine academic risk in students, focusing on models like decision trees, rule learners, and Bayesian networks. Cybernetics and Information Technologies, 13(1), 61–72.

5. Asif, R., Merceron, A., & Pathan, M.-K. (2015). Utilized academic history features and ML classifiers to predict final semester performance. Computers & Education, 83, 104–117.

6. Kotsiantis, S. B. (2012). Developed predictive models to assist educators in identifying students likely to fail, applying decision trees and ensemble learning. Applied Artificial Intelligence, 26(3), 235–244.

7. Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. (2008). Created an artificial neural network-based model for forecasting students' success based on entrance examination results. Expert Systems with Applications, 34(1), 207–214.

8. Musso, M. F., et al. (2013). Combined machine learning and cognitive assessment data to detect at-risk students in early semesters. Computers & Education, 63, 104–112.

9. Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Presented an early warning system using logistic regression to identify students likely to underperform. Journal of Learning Analytics, 1(3), 6–47.

10. Nghe, N. T., Janecek, P., & Haddawy, P. (2007). Implemented machine learning algorithms like decision

trees and Bayesian classifiers to predict outcomes in education. International Conference on Knowledge and Systems Engineering, 149–155.

11. N. Ali, M. Hatala, D. Gašević, and J. Jovanović (2012). This study investigated the effectiveness of Bayesian networks and support vector machines in forecasting student achievement within online learning environments. Educational Technology & Society, 15(1), 147–157.

12. Meier, Y., Xu, B., & Szu, H. H. (2016). Utilized deep learning to estimate academic performance using course behavior data. Proceedings of the IEEE SMC Conference, 1723–1728.

13. Herodotou, C., et al. (2020). Used large-scale data and machine learning to forecast dropout and student success in online courses. British Journal of Educational Technology, 51(5), 1224–1241.

14. Bunkar, K., Sharma, C., & Umesh, S. (2012). Applied classification algorithms for academic failure prediction using historical data. International Journal of Computer Applications, 41(5), 1–5.