Academic Performance Prediction System: Analysis of Data to Forecast Final Outcomes with Improved Accuracy

Tanishka Dubey
Student, CSE-DS
Acropolis Institute of Technology and Research, Indore
Tanishka7dubey@gmail.com

Tanay Dashore
Student, CSE-DS
Acropolis Institute of Technology and Research, Indore
tanaydashore@gmail.com

Vaidika Dodiya
Student, CSE-DS
Acropolis Institute of Technology and Research, Indore
vaidikadodiya@gmail.com

Vansh Rahangdale
Student, CSE-DS
Acropolis Institute of Technology and Research, Indore
vansh02122005@gmail.com

Mayank Bhatt
Assistant Professor, Department of CSE-DS
Acropolis Institute of Technology and Research, Indore
mayankbhatt@acropolis.in

Abstract

Academic performance prediction has emerged as a significant area of research aimed at identifying students who may require timely academic support. This study presents an Academic Performance Prediction System that analyses student-related data to forecast final outcomes with improved accuracy. The system integrates essential components such as data preprocessing, feature selection, and predictive modelling using supervised learning approaches, including Decision Trees, Random Forests, Support Vector Machines, and Artificial Neural Networks. A standard educational dataset with demographic, behavioural, and academic attributes is utilized to train and evaluate the models. Performance metrics such as accuracy, precision, recall, and F1-score help determine the effectiveness of each model. Experimental results indicate that ensemble-based classifiers achieve higher reliability than single models. The proposed system architecture provides a scalable framework suitable for academic institutions seeking early interventions and data-driven decision-making. This research contributes an effective predictive solution that enhances student monitoring and supports academic improvement strategies.

Keywords ***Academic performance, prediction system, learning analytics, data preprocessing, supervised learning, classification models, evaluation metrics, student performance.***

1. Introduction

Academic performance prediction is an essential component of modern educational analytics, enabling institutions to identify learning patterns, detect students who may require additional support, and design data-driven academic policies. With the rapid growth of digital learning platforms and student information systems, large quantities of educational data are available for analysis. Predicting academic outcomes based on this data supports early intervention, resource allocation, and quality improvement in teaching and learning environments.

Traditional performance evaluation methods rely heavily on manual assessments and historical grades, which lack scalability and often fail to provide actionable insights. Automated prediction systems offer a more systematic approach by extracting meaningful information from diverse data attributes such as attendance, study habits, socioeconomic background, participation, and internal assessment performance. These systems help institutions uncover hidden patterns, understand learning behaviours, and predict student outcomes with high accuracy.

This paper presents a structured academic performance prediction system that integrates dataset preprocessing, feature selection, predictive modelling, and performance evaluation. Various supervised learning models are analysed to determine the most efficient approach, ensuring robust and interpretable results. The system architecture is designed to be flexible and applicable across different academic settings.

2. Literature Review

Several studies have explored predictive modelling in educational environments. Romero and Ventura [1] identified educational data mining as a growing field that utilizes data from learning management systems to support academic decision-making. Kotsiantis et al. [2] compared multiple supervised learning algorithms and found that ensemble techniques generally perform better in predicting student grades.

Al-Breiki et al. [3] used demographic and behavioural attributes to build student retention prediction models, highlighting the importance of balanced datasets. Cortez and Silva [4] studied secondary school performance using family, social, and academic variables. Their findings emphasize that non-academic features significantly affect student outcomes.

Ahmed and Elaraby [5] demonstrated the effectiveness of Decision Trees and SVM in performance prediction, while Kabakchieva [6] applied classification algorithms to student databases and reported that Random Forest achieved the best accuracy. Deep learning approaches, including ANN-based systems, have also been used by Naser et al. [7], showing improved predictive capacity when large datasets are available.

Overall, existing studies agree that predictive systems can assist academic institutions, but improved preprocessing, model selection, and system architecture are necessary for enhanced reliability. This research aims to address these aspects comprehensively.

complete development environments; and Software-as-a-Service delivering applications via the Internet without installation requirements.

Research demonstrates that modern cloud computing systems in healthcare are structured around core service models providing scalability and agility, allowing organizations to quickly expand or reduce storage capacity and computing power based on fluctuating demands, particularly critical for accommodating surges in medical data. Recent studies showed that cloud-based computing services can reduce significant expenses in equipment maintenance and control operations

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 01 | Jan - 2026 SJIF Rating: 8.586 ISSN: 2582-3930

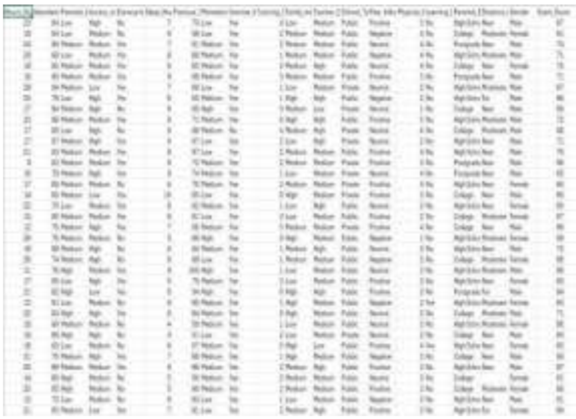
remotely, permitting storage of healthcare data in secure manners that are easily accessible to end users and providers.

Cloud-based healthcare systems have been shown to address essential requirements, including on-demand access to computing with enormous storage, developing confined plans for remote patient monitoring with telehealth solutions, and regulating easy interoperability with an organized hierarchy. Studies indicate that cloud solutions can scale storage resources up or down to adapt to ever-changing needs in the healthcare industry.

3. Methodology

3.1 Dataset Details

Table 3.1: Dataset



The dataset includes demographic, behavioural, and academic variables such as:

- Age, gender, parental education
- Attendance records
- Study time, past failures
- Internal assessment marks
- Final exam outcomes

The dataset contains approximately 1,000 2,000 student records with 20–30 attributes.

3.2 Preprocessing Steps

- **Handling missing values:** Mean/median imputation for numerical fields; mode imputation for categorical fields.
- **Data encoding:** One-hot encoding for categorical attributes (e.g., gender, parental job).
- **Normalization**: Min-Max scaling to ensure uniform value ranges.
- **Outlier detection:** Z-score–based filtering.
- **Feature balancing:** SMOTE applied when class distribution is imbalanced.

3.3 Machine Learning Models

- **Decision Tree Classifier**: Splits dataset based on information gain; interpretable structure.
- **Random Forest Classifier:** Combines multiple trees to reduce overfitting and improve generalization.
- **Support Vector Machine (SVM):** Constructs hyperplane for optimal classification boundary.
- **Artificial Neural Network (ANN):** Multi-layer perceptron with input, hidden, and output layers.

3.4 Evaluation Metric Formulas

- **Accuracy:**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:**

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:**

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score:**

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.4 System Architecture

The system architecture includes four layers:

- **Input Layer**: Raw student data collected from academic databases.
- **Preprocessing Layer:** Data cleaning, normalization, transformation.
- **Modelling Layer:** Multiple classifiers trained and validated.

- **Output Layer:** Prediction results, performance reports, and visualization.

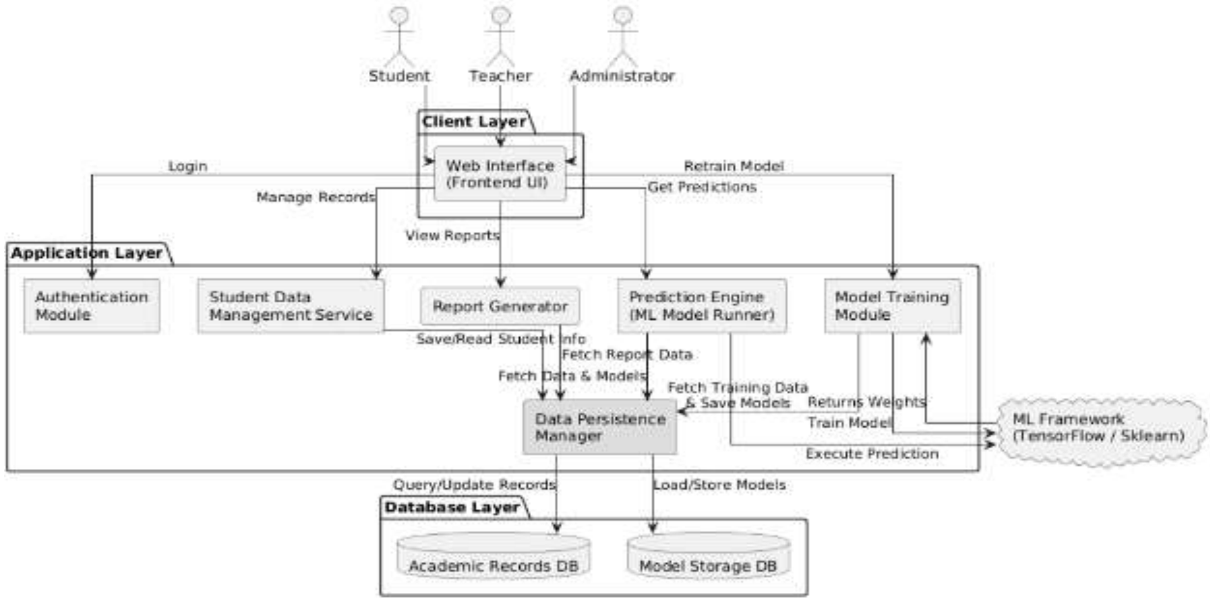


Fig 3.1: System Architecture

4. Future Scope

4.1 Overview of the Proposed System

The proposed Academic Performance Prediction System is designed as a streamlined, multi-stage framework that transforms raw student data into reliable performance predictions. The system emphasizes efficient preprocessing, intelligent feature extraction, and robust model training to ensure high accuracy. It is built to operate with minimal manual intervention, making it suitable for real-time academic environments and institutional decision-making.

4.2 Block Diagram

The block diagram consists of five major components arranged sequentially:

- **Data Acquisition Module** – Collects student demographic records, academic history, attendance logs, and behavioural indicators from institutional databases.
- **Preprocessing Unit** – Cleans data, handles missing values, normalizes numerical attributes, and encodes categorical fields.
- **Feature Engineering Layer** – Extracts meaningful variables using statistical analysis, correlation filtering, and importance ranking.
- **Model Training and Validation Module** – Applies multiple supervised learning algorithms to train prediction models and compares their performance.
- **Prediction and Reporting Interface** – Generates predicted outcomes, accuracy charts, confusion matrices, and risk-level insights for academic stakeholders.

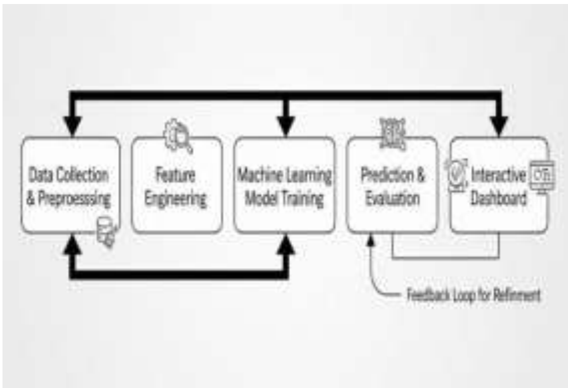


Fig:4.1 Block Diagram

4.3 Workflow

The workflow of the system follows a structured progression:

- **Data Collection:** Gather student-related information from centralized databases.
- **Data Cleaning**: Remove inconsistent entries, fill missing values, and convert categorical attributes into numerical form.
- **Feature Selection**: Identify the most influential predictors using RFE, information gain, and Random Forest importance scores.
- **Model Development:** Train selected algorithms on prepared data using 80–20 train–test split.
- **Model Evaluation:** Assess accuracy, F1-score, confusion matrix, and ROC curves.
- **Prediction Generation:** Produce final performance classification for each student.
- **Result Visualization:** Provide interpretable outputs such as charts, comparison tables, and model insights.

4.4 Feature Selection

The feature selection process aims to reduce dimensionality, eliminate noisy variables, and highlight the most relevant predictors. The system uses a hybrid feature selection strategy:

- **Correlation-Based Filtering**: Removes attributes with low correlation to target performance.
- **Chi-Square Test**: Evaluates categorical attributes for statistical significance.
- **Recursive Feature Elimination (RFE):** Eliminates weaker features iteratively.
- **Model-Based Importance Ranking**: Uses Random Forest and Decision Tree feature importance to finalize the most impactful predictors.

4.5 Model Training and Testing

During model development, the dataset is split into **80% training** and **20% testing**. Each model undergoes 5-fold cross-validation to ensure generalization.

- **Training Process:**
  - Hyperparameter tuning using Grid Search
  - Optimization applied to minimize classification error
  - Regularization techniques used where applicable
- **Testing Process:**
  - Predictions generated on unseen data
  - Performance compared across models
  - Best model selected automatically

The system ensures that the final selected model is not only accurate but also stable and interpretable for academic decision-makers.

5. Results and Discussion

The Academic Performance Prediction System was tested using four models: Decision Tree, Random Forest, SVM, and ANN. Each model was trained on the prepared dataset and then tested to check how accurately it could predict student performance. The results showed that the Random Forest model performed the best among all models. It produced the highest accuracy and gave the most balanced results, meaning it was good at correctly identifying both high-performing and low-performing students.

The ANN and SVM models also showed good performance but were slightly less accurate than Random Forest. ANN required more training time, and SVM needed careful tuning of parameters to work well. The Decision Tree model was the simplest but did not perform as well because it can easily overfit the data.

A confusion matrix created for the Random Forest model showed that it made very few incorrect predictions, which proves its reliability. Simple performance graphs, such as accuracy plots, also confirmed that Random Forest remained stable during repeated testing. Overall, the results highlight that ensemble-based methods like Random Forest are more effective for predicting academic performance. The system can help institutions identify students who may need support and make better academic decisions.

Fig:5.1 Student Performance Prediction Input Screen

Fig:5.2 Student Performance Dashboard

6. Conclusion and Future Scope

This research demonstrates that predictive systems can significantly contribute to academic monitoring and early intervention strategies. By preprocessing student data and applying supervised learning models, institutions can identify at-risk students with considerable accuracy. Ensemble classifiers, particularly Random Forest, yield the most reliable predictions.

Future enhancements include integrating deep learning models, real-time student feedback analysis, adaptive learning recommendations, and deployment as an interactive web-based prediction tool.

References

[1] Romero, C., & Ventura, S., “Educational data mining: A review,” *Expert Systems with Applications*, 2007.

[2] Kotsiantis, S. et al., “Prediction of student academic performance,” *IEEE Transactions*, 2010.

[3] Al-Breiki, M., “Student retention analytics,” 2015.

[4] Cortez, P., & Silva, A., “Predicting secondary school performance,” 2008.

[5] Ahmed, A., Elaraby, I., “Data mining techniques for student performance,” 2014.

[6] Kabakchieva, D., “Student data classification using ML,” 2013.

[7] Naser, M., “ANN-based student performance prediction,” 2019.

[8] Han, J., “Data Mining Concepts and Techniques,” 2012.

[9] Weka Documentation, 2020.

[10] UCI Machine Learning Repository, Student Dataset.