ASIAN SCHOOL OF MEDIA STUDIES
SCHOOL OF DATA SCIENCE

Predicting Student Performance

Diploma in
Data Science

By
Samriddhi Negi

Under the supervision of
Ms. Neema Jha

ASIAN SCHOOL OF MEDIA STUDIES

2025

Abstract:

This project aims to develop a machine learning model that predicts student performance based on various academic and socio-demographic features such as attendance, previous grades, family background, and study hours. The goal is to help educational institutions identify at-risk students early and provide timely support to improve outcomes.

Motivation:

In today's competitive academic environment, student success is a top priority for educational institutions. Despite efforts by teachers and administrators, many students struggle or drop out due to issues that go undetected until it's too late. Early identification of at-risk students can prevent academic failure, improve retention rates, and provide timely interventions.

With the increasing availability of student-related data, there is a huge opportunity to apply machine learning to analyze trends and predict performance. This can help educators make data-driven decisions, provide personalized support, and ultimately improve educational outcomes.

Literature Review:

In recent years, there has been growing interest in applying machine learning (ML) techniques to educational data to predict student outcomes. Several studies have demonstrated that predictive analytics can significantly improve student support and academic planning.

1. Romero & Ventura (2007) conducted a comprehensive review of educational data mining (EDM) techniques. They emphasized that classification algorithms like Decision Trees and Naïve Bayes are commonly used to predict student performance. Their findings showed that data mining can uncover patterns not easily visible through traditional analysis.

2. Cortez & Silva (2008) used the UCI Student Performance Dataset to predict students' final grades using Decision Trees, Random Forests, and Support Vector Machines. Their research indicated that parental education level, study time, and past grades were significant predictors.

3. Al-Barrak & Al-Razgan (2016) applied classification algorithms like Naïve Bayes, K-NN, and SVM to predict student performance. The study concluded that Naïve Bayes gave better accuracy when dealing with smaller datasets and fewer attributes.

4. Jayaprakash et al. (2014) developed an early warning system using logistic regression and decision trees. Their system helped identify students at risk of dropping out, enabling timely intervention.

5. Kotsiantis et al. (2004) found that the combination of demographic data, previous academic records, and behavioral attributes provided higher accuracy in predicting performance when used with ensemble learning models.

Traditional data analysis and visualization tools:

Popular tools like **Excel, Tableau, and Power BI** allow users to create visual dashboards, charts, and tables to analyze data. However, these tools require **manual effort and domain expertise** to interpret insights, making them inaccessible to non-technical users.

Rule-Based Data Storytelling Systems:

Early automated storytelling solutions used **predefined templates and rule-based logic** to generate textual insights from structured data. Systems like **Automated Insights and**

Narrative Science provided automated reporting but lacked **adaptability for diverse datasets** and struggled with **unstructured data analysis**.

Machine Learning and NLP-Based Storytelling

Recent advancements in **machine learning (ML) and NLP** have led to the development of intelligent data storytelling models. Large Language Models (LLMs) like **GPT-4 and BERT** can generate human-like narratives, explaining insights dynamically. However, challenges such as **bias, computational costs, and hallucinations (incorrect insights)** remain key limitations.

Comparison of existing techniques

Technique	Pros	Cons
Logistic Regression	Easy to interpret, fast, works well with linear relationships.	Limited in capturing complex patterns.
Decision tree	Easy to visualize, good for categorical data.	Prone to overfitting without pruning.
Random Forest	High accuracy, handles missing values well.	Less interpretable, longer training time.
Support vector machine (SVM)	Works well in high-dimensional spaces.	Difficult to tune, slower on large datasets.
K- Nearest Neighbors(KNN)	Simple to implement, non-parametric.	Computationally expensive, sensitive to noise.
Naïve Bayes	Fast, works well with small datasets.	Assumes independence, which may not be true.
Gradient Boosting (XGBoost, LightGBM)	High accuracy, handles imbalanced data.	Requires tuning, less interpretable
Neural Networks	Captures complex relationships.	Requires large data, less transparent.

Methodology:

To design, implement, and evaluate a system that leverages artificial intelligence to automatically:

1. Process and analyze datasets to extract actionable insights.

2. Generate human-like textual narratives explaining the insights.

3. Combine these narratives with interactive visualizations to create comprehensive, user-friendly data stories.

Planning of work:

1. Data Collection

- Source: Public datasets like the UCI Student Performance dataset or academic records from an institution (if available).
- Features: Includes academic scores, attendance, demographic data, parental education, study hours, etc.

2. Data Preprocessing

- **Handling Missing Values:** Imputation or removal of null entries.
- **Feature Encoding:** Converting categorical variables (e.g., gender, school type) using Label Encoding or One-Hot Encoding.
- **Normalization/Standardization:** Ensures all features are on the same scale, especially important for models like KNN and SVM.
- **Splitting the Dataset:** Dividing data into training (70%) and testing (30%) sets using `train_test_split()`.

3. Exploratory Data Analysis (EDA)

- Use of graphs and charts to analyze feature distributions and correlations.
- Visualization tools: **Matplotlib**, **Seaborn**, or **Power BI**.
- Identify key variables that affect student performance.

4. Model Selection and Training

Multiple algorithms will be applied for comparison:

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **Support Vector Machine (SVM)**
- **Gradient Boosting (XGBoost/LightGBM)**

These models will be trained using the training set.

5. Model Evaluation

Models will be evaluated based on the test data using:

- **Accuracy**
- **Precision**
- **Recall**

- **F1-Score**
- **Confusion Matrix**

The best-performing model will be selected based on a balance between these metrics.

6. Model Optimization

- Hyperparameter tuning using **GridSearchCV** or **RandomizedSearchCV**.
- Feature selection to remove less important or redundant variables.

7. Deployment (Optional / If Applicable)

- If part of the project scope, results may be visualized using a **dashboard** built in **Power BI** or **Streamlit** (for Python-based UI).
- A user-friendly report will be created to present results to academic stakeholders.

Bibliography

1. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135–146. https://doi.org/10.1016/j.eswa.2006.04.005

2. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th FUture Business Technology Conference (FUBUTEC)*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Student+Performance
3. Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1), 6–47. https://doi.org/10.18608/jla.2014.11.3
4. Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: A case study. *International Journal of Information and Education Technology*, 6(7), 528–533. https://doi.org/10.7763/IJIET.2016.V6.745
5. Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5), 411–426. https://doi.org/10.1080/08839510490442058
6. UCI Machine Learning Repository. (n.d.). Student Performance Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Student+Performance
7. Scikit-learn: Machine Learning in Python. (n.d.). Retrieved from https://scikit-learn.org/