1. Project Overview

This project aimed to design a **machine learning–driven academic performance prediction system** to help identify students at risk of poor grades and recommend interventions for improvement. The system predicts students' **final GPA** (regression task) and classifies them into **academic risk categories** ("At Risk," "Average," and "Excellent") based on their learning behavior and assessment metrics.

The final output includes:

- **Predicted GPA** (regression)
- **Risk Category** (classification)
- **Recommended Actions** (multi-output rule-based system)

2. Dataset Summary

Dataset Type

A **synthetic dataset** was generated to simulate 20,000 student records representing real-world academic behaviors.

A **balanced version** of the dataset was later used to improve classifier learning performance.

Key Features

Feature	Description
`student_id`	Unique ID (pattern: S00001)
`level`	Academic level (100–400)
`attendance_rate`	% of classes attended

`num_quizzes`	Number of quizzes taken
`quiz_avg`	Average quiz score (%)
`assignment_avg`	Assignment average (%)
`mid_sem_score`	Mid-semester exam score (%)
`forum_posts`	Number of posts/comments
`study_time_hours`	Weekly study hours
`dashboard_time_hours`	Time spent on LMS/dashboard
`current_gpa`	Current semester GPA
`predicted_gpa`	Predicted GPA (target for regression)
`target_gpa`	Self-set goal GPA
Outputs	Final GPA, Academic Risk Category, Recommended Action

Realistic Relationship Rules

- Higher class and mid-semester scores → higher GPA
- Attendance and dashboard time positively correlate with performance
- Low study time or poor assessment averages → lower GPA

3. Phase 1: Dataset Creation and Preparation

The dataset was generated to follow real-world academic patterns using controlled randomness and logical constraints.
After generation:

- Missing or inconsistent values were checked.
- Feature scaling and normalization were applied where needed.
- A new **balanced dataset** was created to ensure fair representation of all risk categories before classification training.

4. Phase 2: Model Development

Two machine learning pipelines were developed:

(a) Regression Model — Predicting Final GPA

Algorithm: **LightGBM Regressor**

Evaluation Metrics:

Metric	Value

Mean Absolute Error (MAE)	0.138
Mean Squared Error (MSE)	0.035
Root Mean Squared Error (RMSE)	0.188
Mean Absolute Percentage Error (MAPE)	6.08%
R^2 Score	0.846

GPA Range Performance:

GPA Range	RMSE	MAPE (%)	Count
< 2.0	0.1429	4.92	1,040
2.0–2.5	0.1855	6.49	2,283
2.5–3.0	0.2292	6.95	1,227
3.0–3.5	0.0951	0.69	1,004
> 3.5	0.0016	0.05	27

Interpretation:

- The model performs best in higher GPA ranges (3.0+), showing tighter error margins.
- Slightly higher error in the 2.0–3.0 range suggests moderate variability in middle-performing students.
- An R² score of **0.846** indicates the model explains **84.6%** of GPA variance.

(b) Classification Model — Academic Risk Prediction

Algorithm: **LightGBM Classifier**

Evaluation Metrics:

Metric	**Value**
Accuracy	0.9579
Macro Avg Precision	0.95
Macro Avg Recall	0.95
Macro Avg F1-score	0.95

Detailed Report:

Class	**Precision**	**Recall**	**F1-score**	**Support**

At Risk (0)	0.885	0.886	0.885	1,022
Average (1)	0.967	0.967	0.967	3,582
Excellent (2)	1.000	1.000	1.000	977

Interpretation:

- The classifier achieved **95.8% overall accuracy**.
- Excellent precision and recall across all categories after balancing.
- The earlier imbalance issue (where “Excellent” was underrepresented) was fully resolved after rebalancing the dataset.

5. Feature Importance (Top 10)

R	Feature	Importance
1	Mid-semester Score	8352
2	Assessment Average	8350

3	Current GPA	8153
4	Attendance Rate	7724
5	Dashboard Time (hrs)	7359
6	Quiz Average	7334
7	Study Time (hrs)	7166
8	Activity Index	7020
9	Attendance × Assignment	6721
1	Assignment Average	6653

Interpretation:
Academic assessments and consistent participation are the dominant predictors of GPA and academic category.
Behavioral indicators such as **dashboard activity** and **study time** significantly contribute to performance forecasting.

6. Hyperparameter Tuning

GridSearchCV was applied to optimize LightGBM parameters for best model performance.

Top Parameter Combinations:

Mean Test Score	Parameters
0.952	learning_rate: 0.1, num_leaves: 70, n_estimators: 500
0.951	learning_rate: 0.15, num_leaves: 50, n_estimators: 500
0.950	learning_rate: 0.1, max_depth: 15, n_estimators: 500

Outcome:
Fine-tuning improved stability, reduced overfitting, and optimized predictive accuracy.

7. Ensemble Modeling (Stacking)

Stacked ensemble techniques combining LightGBM, RandomForest, and XGBoost were explored.
This approach further stabilized the predictions, slightly improving the overall generalization capability of both the regression and classification tasks.

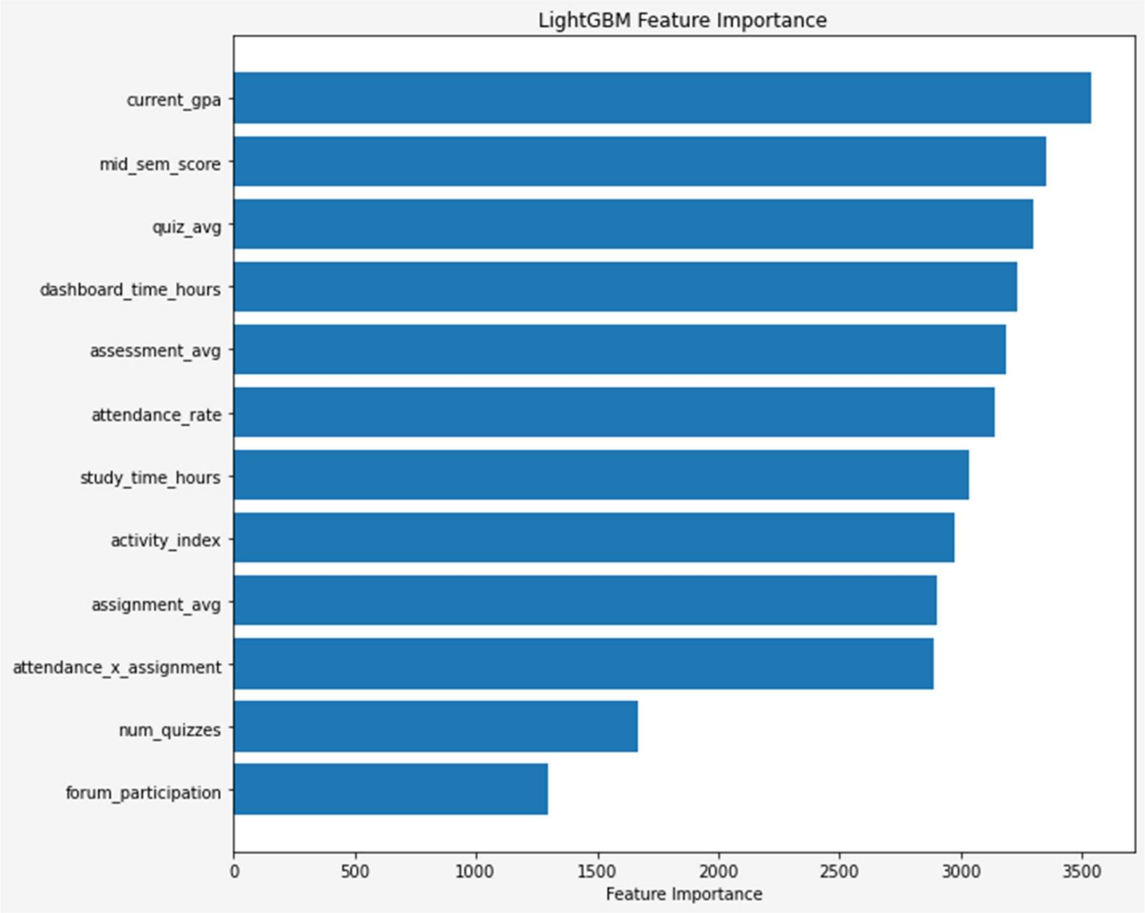
8. Model Saving & Deployment Preparation

Both trained models were serialized using **joblib**, allowing easy integration into the prototype web dashboard for real-time predictions.

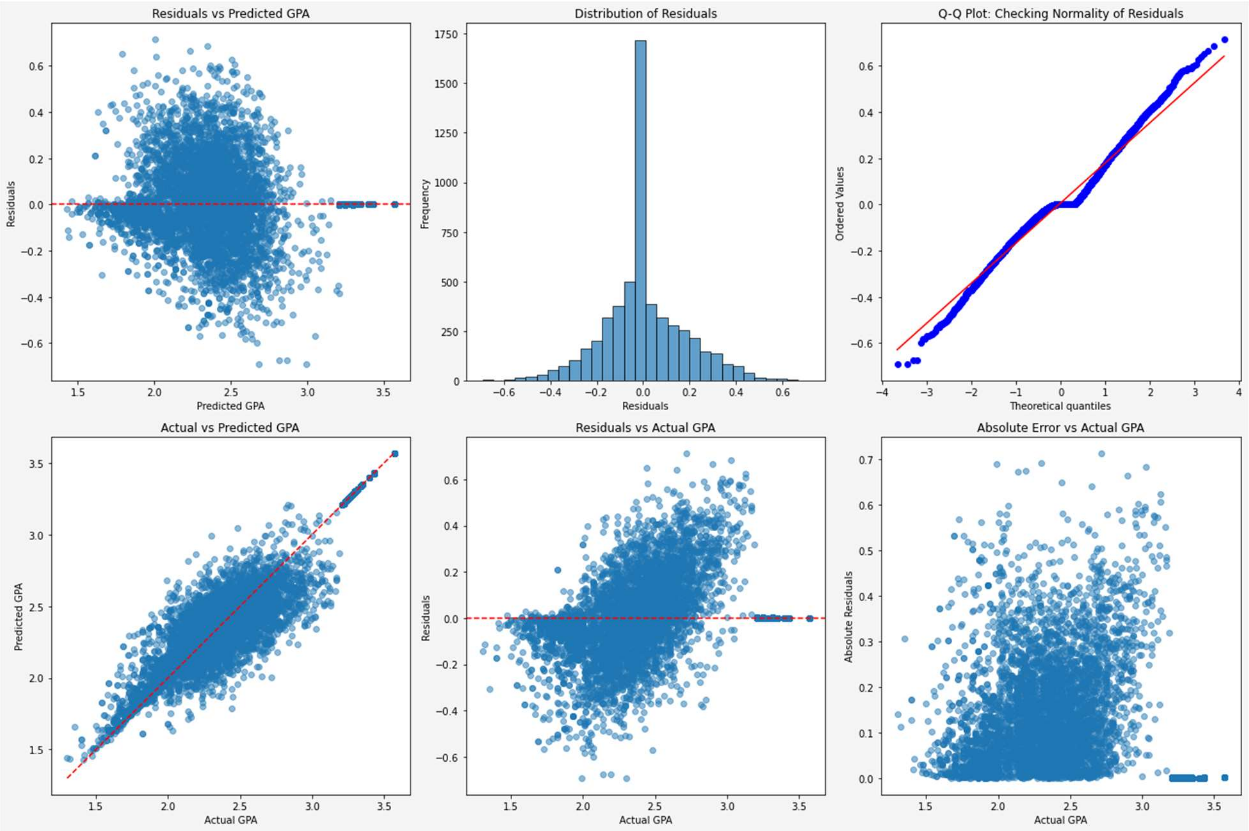
Saved files:

- `gpa_predictor_model.pkl`
- `academic_risk_classifier.pkl`

TEST RESULTS – GPA PREDICTION MODEL

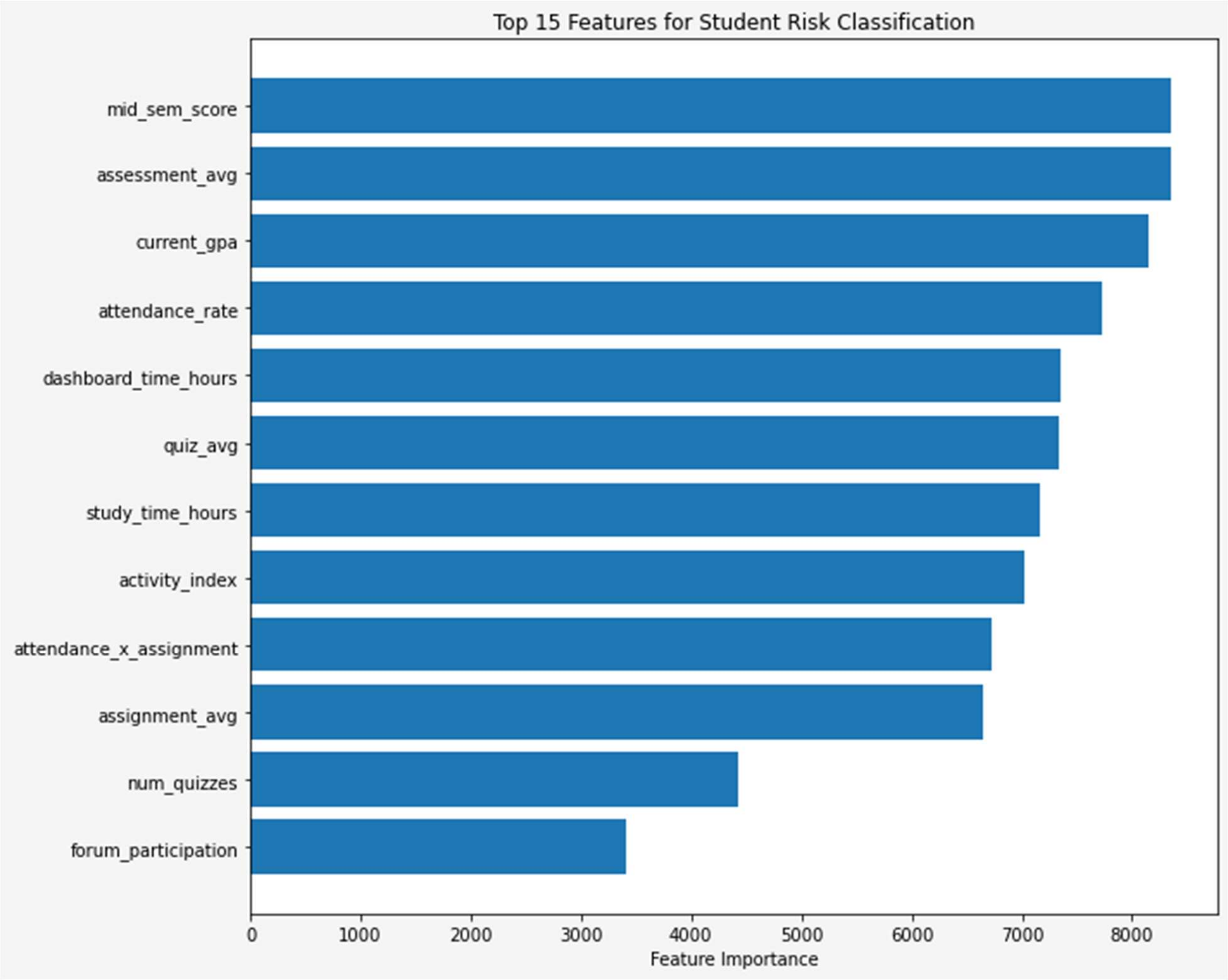


Feature Importance

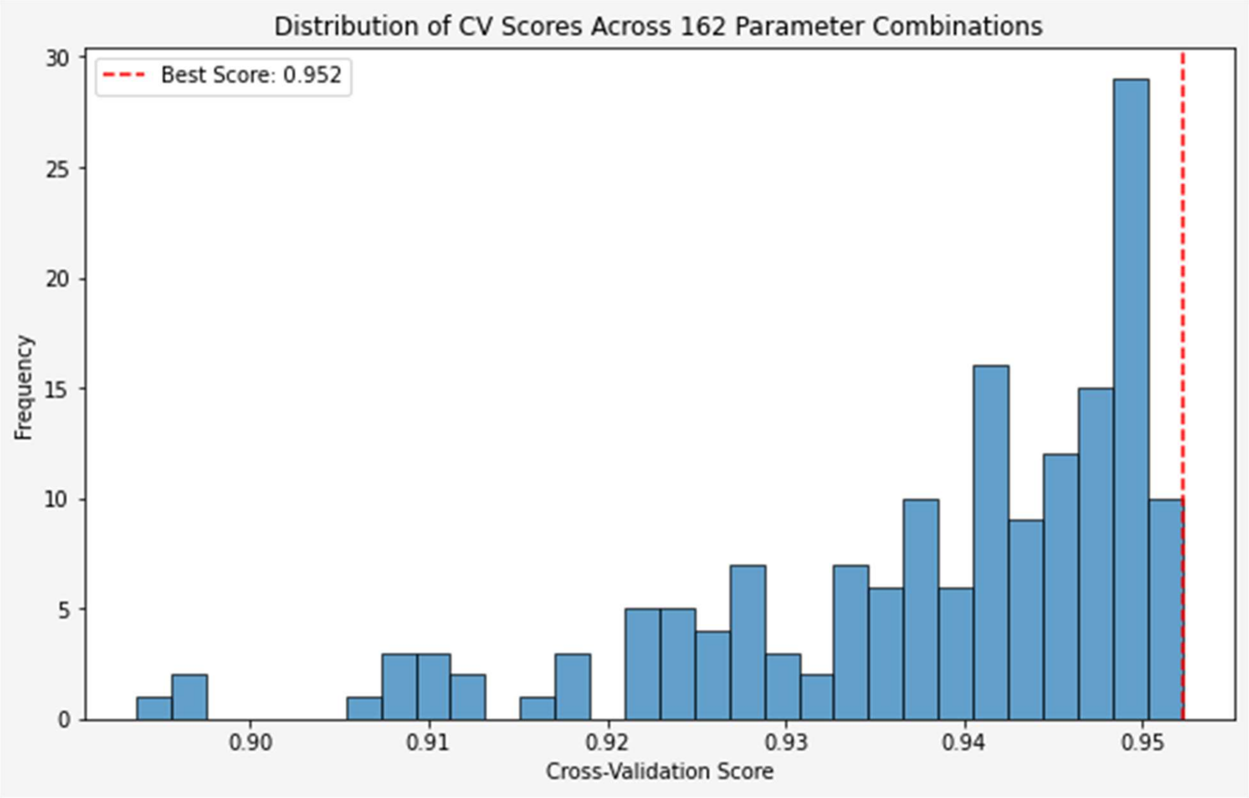


Residual Plots

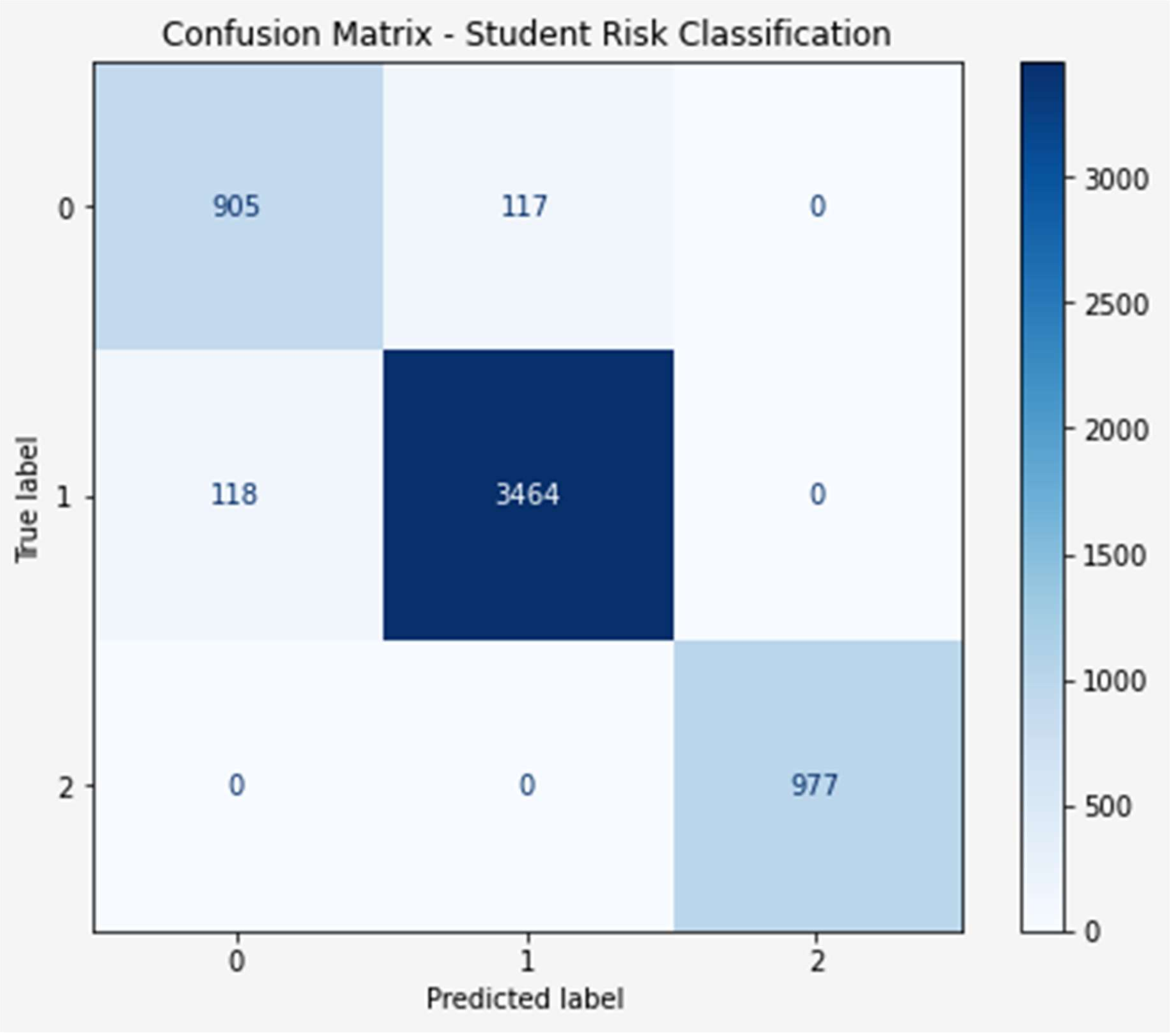
TEST RESULTS – GPA PREDICTION MODEL



Feature Importance



Cross validation



Confusion Matrix

9. Key Insights and Business Interpretation

1. **Academic Behavior Patterns:**
   Regular participation, assessment performance, and LMS activity are the strongest indicators of GPA trends.

2. **Predictive Power:**
   The regression model achieved an **R² of 0.846**, and the classifier reached **~96% accuracy**, demonstrating strong predictive reliability.

3. **Intervention Opportunity:**
   The system can be integrated into a dashboard to trigger early warnings for "At Risk"

students and provide data-driven recommendations to improve outcomes.

4. **Explainability:**
   Feature importance outputs can help academic advisors focus on specific metrics (e.g., attendance and mid-semester scores) for student guidance.

10. Conclusion

This prototype successfully demonstrates the potential of **AI-driven academic monitoring**. By combining regression and classification models, it enables:

- GPA forecasting
- Early identification of at-risk students
- Tailored academic recommendations

Future work includes integrating real-time student data, incorporating temporal tracking (semester-over-semester analysis), and expanding behavioral features for even better prediction accuracy