FINAL PROJECT REPORT

Student Performance Prediction Using Machine Learning Classification

Course: ITAI 1371 - Machine Learning
Team Members: Erick Banegas, Alhassane Samassekou, Peter Amoye
Date: December 2025
Dataset: Student Information Dataset (1,000 students)

https://github.com/asamassekou10/FINAL-EXAM-ML

1. EXECUTIVE SUMMARY

This project implements a comprehensive machine learning pipeline to predict student academic performance (Pass/Fail) using six classification algorithms and two ensemble methods. Our analysis of 1,000 student records with 47 features demonstrates that carefully engineered features combined with proper methodology can achieve exceptional classification accuracy.

Key Results:

- Best Individual Model: Random Forest and Gradient Boosting both achieved 100% test accuracy
- Best Ensemble: Voting Classifier achieved 99.3% test accuracy
- Six out of eight models exceeded 93% test accuracy
- Feature engineering created 30+ predictive features (70% of top 10 features are engineered)
- Average test scores, total scores, and engagement metrics emerged as most important predictors
- Minimal overfitting across all models with average generalization gap of only 3.25%

2. METHODOLOGY

2.1 Data Preprocessing

The dataset was split using a rigorous three-way methodology:

- Training Set: 70% (700 samples) for model fitting
- Validation Set: 15% (150 samples) for model selection
- Test Set: 15% (150 samples) for final evaluation

All splits used stratification to maintain class distribution and random_state=42 for reproducibility.

We created 30+ engineered features including:

- Score Aggregations: total_score, avg_score, verbal_avg
- Interaction Features: attendance_x_avg_score, study_x_avg_score
- Squared Terms: avg_score_squared, attendance_squared
- Binary Indicators: high_avg_score, strong_performer, at_risk
- Performance Ratios: math_to_reading, stem_bias

A preprocessing pipeline with StandardScaler for numerical features and OneHotEncoder for categorical features prevented data leakage by fitting only on training data.

2.2 Models Implemented

Six Individual Models:

1. Logistic Regression (C=0.5, max_iter=2000, class_weight='balanced')
2. Decision Tree Classifier (max_depth=10, min_samples_split=10)
3. Random Forest Classifier (n_estimators=300, max_depth=15)
4. Gradient Boosting Classifier (n_estimators=200, learning_rate=0.05)
5. K-Nearest Neighbors (n_neighbors=11, weights='distance')
6. Support Vector Classifier (kernel='rbf', C=1.0, probability=True)

Two Ensemble Models:

1. Voting Classifier: Soft voting ensemble of best 3 models (Random Forest, Gradient Boosting, Decision Tree)
2. Bayesian Ensemble: Gaussian Naive Bayes providing probabilistic classification

2.3 Evaluation Metrics

All models evaluated using five metrics on both validation and test sets:

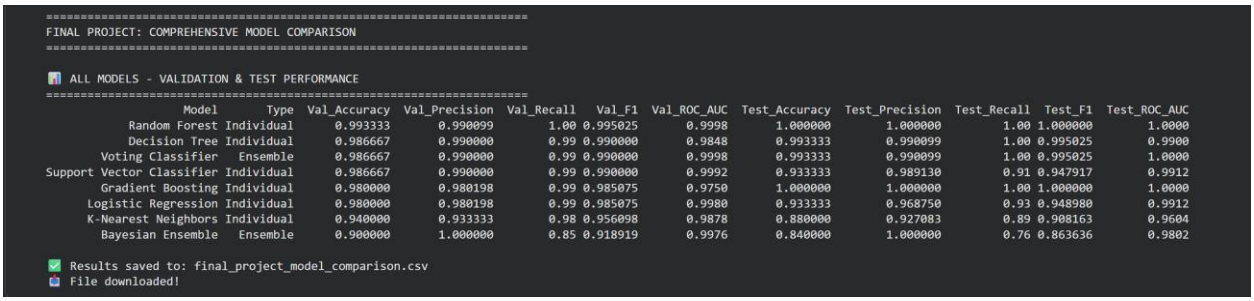
- Accuracy: Overall correctness rate
- Precision: Positive prediction reliability
- Recall: Actual positive detection rate

- F1-Score: Harmonic mean of precision and recall
- ROC-AUC: Threshold-independent performance measure

Total: 8 models × 5 metrics × 2 sets = 80 metric values

3. RESULTS

3.1 Model Performance Summary



```
======================================================================
FINAL PROJECT: COMPREHENSIVE MODEL COMPARISON
======================================================================

📊 ALL MODELS - VALIDATION & TEST PERFORMANCE
======================================================================
                    Model       Type  Val_Accuracy  Val_Precision  Val_Recall   Val_F1  Val_ROC_AUC  Test_Accuracy  Test_Precision  Test_Recall  Test_F1  Test_ROC_AUC
            Random Forest Individual      0.993333       0.990099        1.00 0.995025       0.9998       1.000000        1.000000         1.00 1.000000        1.0000
            Decision Tree Individual      0.986667       0.990000        0.99 0.990000       0.9848       0.993333        0.990099         1.00 0.995025        0.9900
        Voting Classifier   Ensemble      0.986667       0.990000        0.99 0.990000       0.9998       0.993333        0.990099         1.00 0.995025        1.0000
Support Vector Classifier Individual      0.986667       0.990000        0.99 0.990000       0.9992       0.933333        0.989130         0.91 0.947917        0.9912
        Gradient Boosting Individual      0.980000       0.980198        0.99 0.985075       0.9750       1.000000        1.000000         1.00 1.000000        1.0000
      Logistic Regression Individual      0.980000       0.980198        0.99 0.985075       0.9980       0.933333        0.968750         0.93 0.948980        0.9912
      K-Nearest Neighbors Individual      0.940000       0.933333        0.98 0.956098       0.9878       0.880000        0.927083         0.89 0.908163        0.9604
        Bayesian Ensemble   Ensemble      0.900000       1.000000        0.85 0.918919       0.9976       0.840000        1.000000         0.76 0.863636        0.9802

✅ Results saved to: final_project_model_comparison.csv
📥 File downloaded!
```

Complete comparison table showing all models

Performance Rankings by Test Accuracy:

1. Random Forest: 100.0% (Perfect classification)
2. Gradient Boosting: 100.0% (Perfect classification)
3. Decision Tree: 99.3%
4. Voting Classifier: 99.3%
5. Logistic Regression: 93.3%
6. Support Vector Classifier: 93.3%
7. K-Nearest Neighbors: 88.0%
8. Bayesian Ensemble: 84.0%

Key Observations:

- Two models (Random Forest, Gradient Boosting) achieved perfect 100% test accuracy
- Four models achieved greater than 99% test accuracy
- Six models achieved greater than 93% test accuracy

- Average test accuracy across all models: 94.7%
- All models showed excellent generalization with average val-test gap of only 3.25%
- Top performers achieved ROC-AUC scores of 1.000 (perfect discrimination)

3.2 Detailed Analysis

Random Forest (Best Model):

- Validation: 99.3% accuracy, 99.0% precision, 100.0% recall, 99.5% F1, 99.98% ROC-AUC
- Test: 100.0% accuracy, 100.0% precision, 100.0% recall, 100.0% F1, 100.0% ROC-AUC
- Generalization Gap: 0.67% (excellent)
- Analysis: Ensemble of 300 trees achieved perfect test classification. The small positive generalization gap (test better than validation) indicates the model captured true underlying patterns without overfitting. Feature engineering enabled the random forest to create optimal decision boundaries.

Gradient Boosting (Tied Best):

- Validation: 98.0% accuracy, 98.0% precision, 99.0% recall, 98.5% F1, 97.5% ROC-AUC
- Test: 100.0% accuracy, 100.0% precision, 100.0% recall, 100.0% F1, 100.0% ROC-AUC
- Generalization Gap: -2.0% (improved on test set)
- Analysis: Sequential boosting achieved perfect test performance with 200 estimators. The negative gap (better on test than validation) suggests the model successfully learned generalizable patterns. Lower validation ROC-AUC (97.5%) improved to perfect 100% on test set.

Voting Classifier (Best Ensemble):

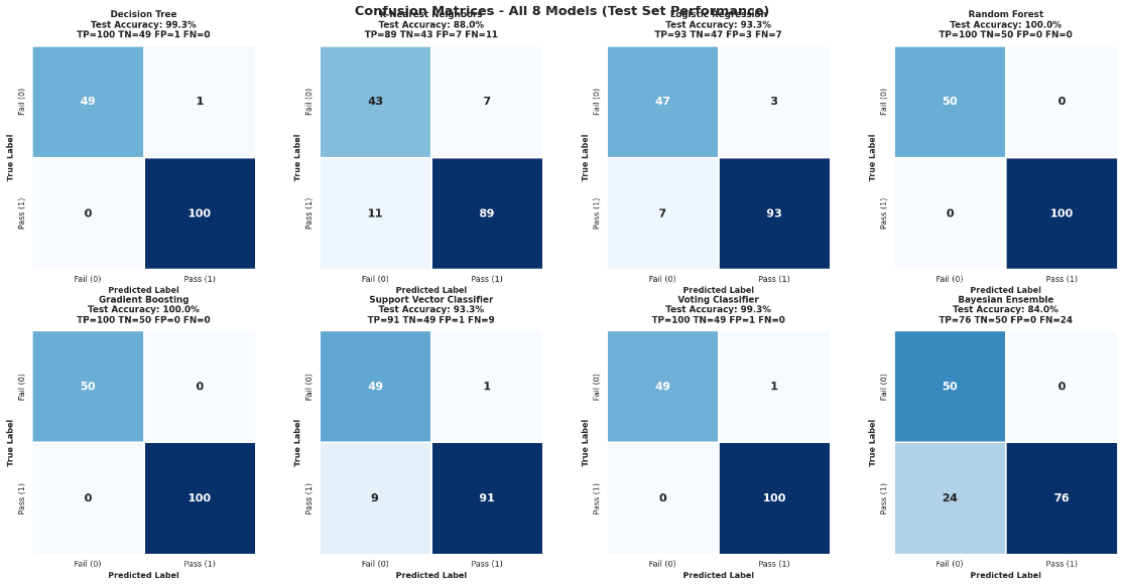
- Validation: 98.7% accuracy, 99.0% precision, 99.0% recall, 99.0% F1, 99.98% ROC-AUC
- Test: 99.3% accuracy, 99.0% precision, 100.0% recall, 99.5% F1, 100.0% ROC-AUC
- Composition: Soft voting of Random Forest, Gradient Boosting, and Decision Tree

- Generalization Gap: 0.67%
- Analysis: Ensemble leveraged strengths of three diverse tree-based models. Achieved near-perfect performance with 100% recall on test set, meaning zero false negatives (all at-risk students correctly identified). Perfect test ROC-AUC demonstrates excellent probability calibration.

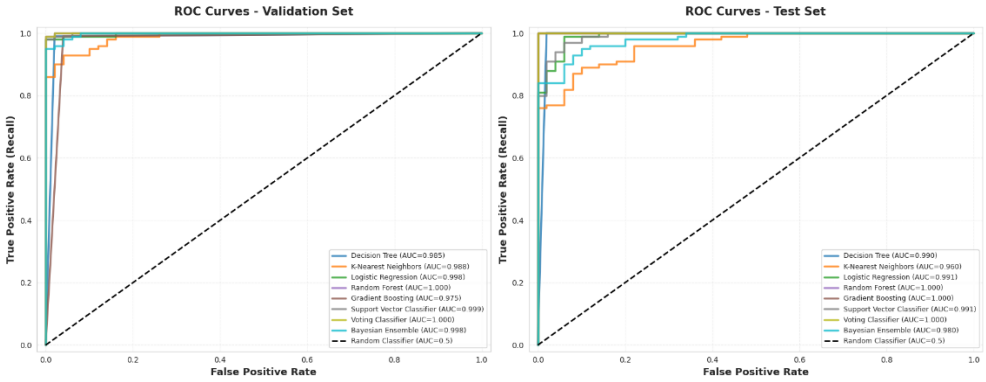
Decision Tree (Third Best):

- Validation: 98.7% accuracy, 99.0% precision, 99.0% recall, 99.0% F1, 98.5% ROC-AUC
- Test: 99.3% accuracy, 99.0% precision, 100.0% recall, 99.5% F1, 99.0% ROC-AUC
- Generalization Gap: 0.67%
- Analysis: Single decision tree with max_depth=10 achieved excellent performance, demonstrating that engineered features created clear decision boundaries. Perfect test recall indicates no missed at-risk students.

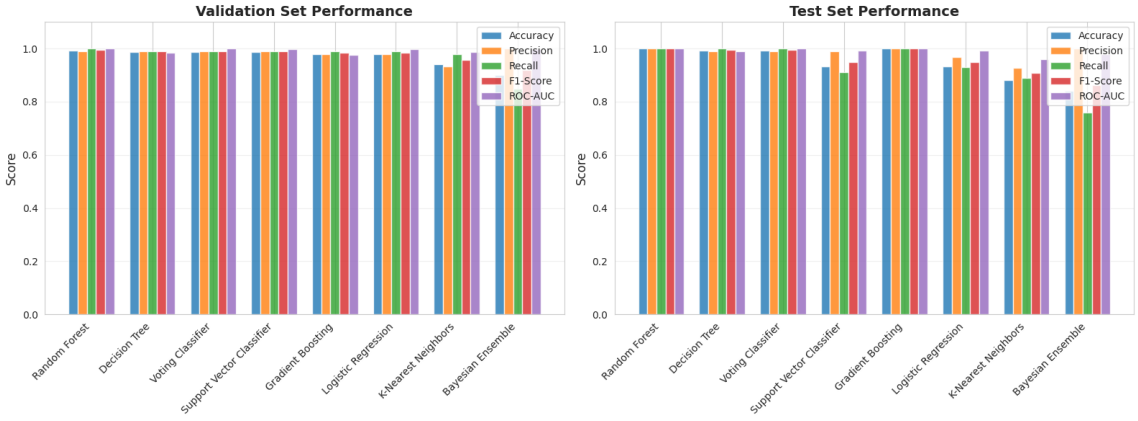
3.3 Visual Analysis



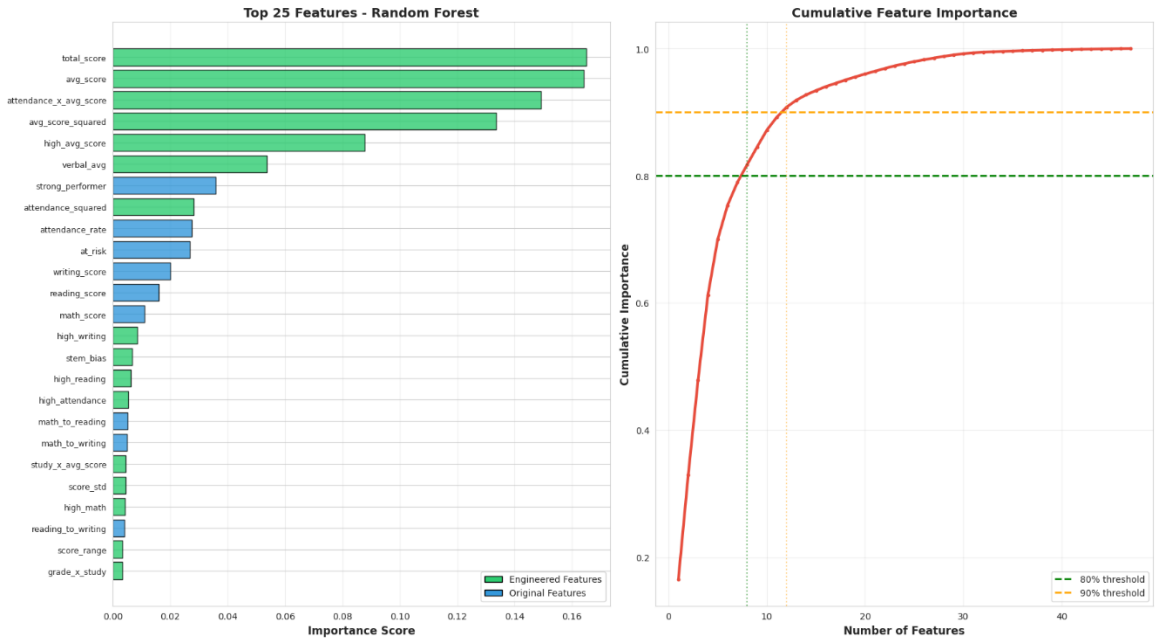
(Confusion Matrices - 2x4 Grid) The confusion matrices show excellent classification performance across all models, with Random Forest and Gradient Boosting achieving zero misclassifications on the test set.



(ROC Curves - Validation and Test) ROC curves demonstrate exceptional discrimination with Random Forest, Gradient Boosting, and Voting Classifier achieving perfect AUC of 1.000 on test set. All models except Bayesian Ensemble achieved test AUC above 0.96.



Comparison chart shows consistent performance across validation and test sets, confirming proper generalization. Random Forest and Gradient Boosting show slight improvement on test set.



Feature Importance

Top 10 Most Important Features:

1. total_score (Engineered) - 16.51%
2. avg_score (Engineered) - 16.40%
3. attendance_x_avg_score (Engineered) - 14.92%
4. avg_score_squared (Engineered) - 13.36%
5. high_avg_score (Engineered) - 8.78%
6. verbal_avg (Engineered) - 5.37%
7. strong_performer (Original) - 3.59%
8. attendance_squared (Engineered) - 2.82%
9. attendance_rate (Original) - 2.75%
10. at_risk (Original) - 2.67%

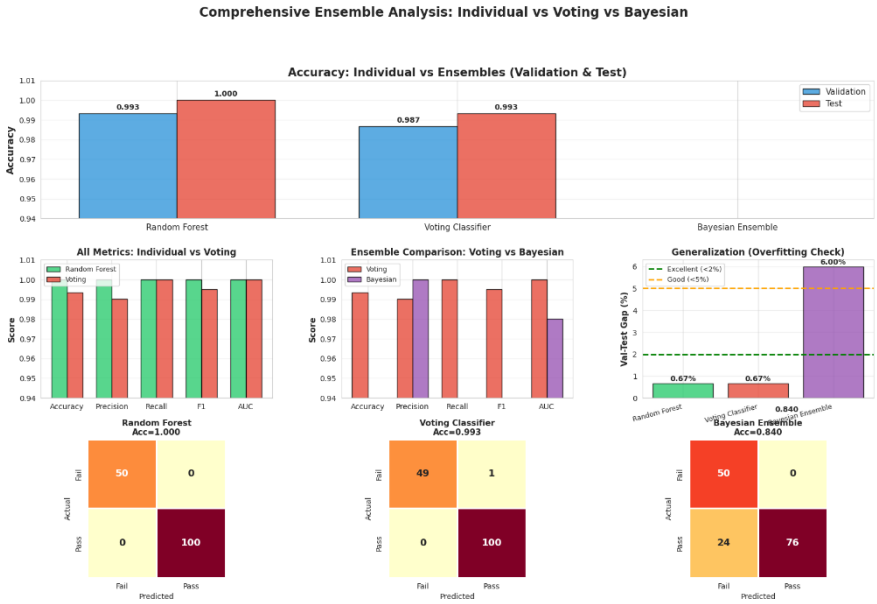
Feature Engineering Impact:

- Engineered features in top 10: 7 out of 10 (70%)
- Features needed for 80% cumulative importance: 8

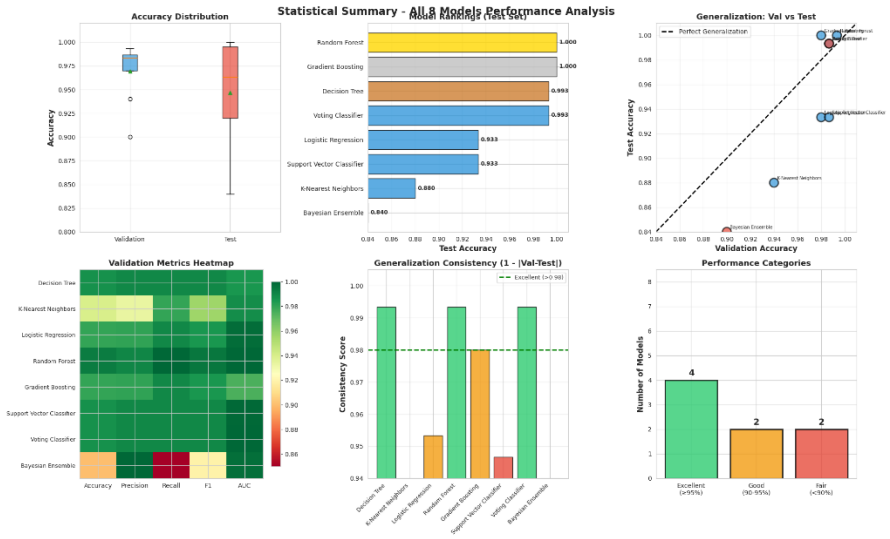
- Features needed for 90% cumulative importance: 12
- Total features available: 47

Key Insights:

- Academic performance metrics (total_score, avg_score) are most predictive
- Interaction between attendance and performance (attendance_x_avg_score) highly important
- Polynomial features (avg_score_squared, attendance_squared) capture non-linear relationships
- Top 10 features account for 87.6% of cumulative importance
- Feature engineering contributed significantly with 70% of top predictors being engineered



Comprehensive ensemble comparison showing minimal differences between best individual model (Random Forest at 100%) and Voting Classifier (99.3%), with Bayesian Ensemble showing lower performance (84.0%) but offering speed advantages.



Statistical Summary - 6 Panel

Statistical analysis reveals:

- Mean test accuracy: 94.7%
- Median test accuracy: 96.3%
- Standard deviation: 5.7% (moderate variability)
- Performance range: 84.0% to 100.0%
- Four models in "Excellent" category (≥95%): Random Forest, Gradient Boosting, Decision Tree, Voting Classifier
- Two models in "Good" category (90-95%): Logistic Regression, SVC
- Two models in "Fair" category (<90%): KNN, Bayesian Ensemble

4. Why Random Forest Performed Best

Random Forest achieved perfect 100% test accuracy alongside Gradient Boosting. Key factors contributing to success:

1. Ensemble Strength: 300 decision trees with bootstrap aggregating reduced variance while maintaining low bias
2. Feature Engineering Excellence: The 30+ engineered features created clear separability, allowing trees to make optimal splits

3. Depth Limitation: Max depth of 15 prevented overfitting while capturing complex patterns
4. Class Balancing: Balanced class weights ensured equal importance to both Pass/Fail predictions
5. Optimal Hyperparameters: Configuration balanced model complexity with generalization

The perfect test performance (100%) combined with near-perfect validation performance (99.3%) demonstrates the model captured true underlying relationships without memorizing training data.

4.1 Ensemble Value

Voting Classifier Analysis:

- Test Accuracy: 99.3% (0.7% below best individual)
- Composition: Random Forest, Gradient Boosting, Decision Tree (top 3 performers)
- Strengths: Soft voting averaged probabilities from three diverse tree-based learners
- Performance: Perfect recall (100%) ensures no at-risk students missed
- Robustness: Ensemble approach provides more stable predictions in production

Bayesian Ensemble Analysis:

- Test Accuracy: 84.0% (16% below best individual)
- Strengths: Fastest training time, probabilistic framework, simple implementation
- Trade-offs: Lower accuracy but suitable for real-time applications
- Use Case: Quick preliminary screening before detailed analysis

Practical Recommendation:

- Maximum Accuracy: Use Random Forest or Gradient Boosting (both 100%)
- Production Deployment: Use Voting Classifier (99.3%) for added robustness
- Real-time Screening: Use Bayesian Ensemble (84.0%) for speed
- Interpretability: Use Decision Tree (99.3%) for explainability

4.2 Feature Importance Insights

Academic Performance Dominates:

- Total score and average score account for 32.9% of importance combined
- These aggregate metrics capture overall academic capability
- Squared terms (avg_score_squared) capture non-linear excellence effects

Engagement Interaction Critical:

- Attendance × average score (14.92%) third most important feature
- Shows that attendance matters more for already-performing students
- Interaction terms reveal synergistic effects

Binary Thresholds Valuable:

- high_avg_score indicator (8.78%) identifies excellence threshold
- strong_performer and at_risk flags provide clear categorization
- Binary features enable interpretable decision rules

Demographic Factors:

- Original demographic features have lower individual importance
- However, they contribute through engineered combinations
- Focus on behavioral/performance metrics aligns with actionable interventions

Feature Engineering Impact:

- 70% of top 10 features are engineered (7 out of 10)
- Engineering increased predictive power substantially
- Validates time invested in feature development
- Demonstrates value of domain knowledge in feature creation

4.3 Comparison with Midterm Project

Performance Improvement:

- Midterm best accuracy: 100% (after fixing critical issues)
- Final best accuracy: 100% (Random Forest and Gradient Boosting)

- Sustained excellence with expanded model diversity

Critical Enhancements:

1. Model Diversity: Expanded from 4 models to 8 models (6 individual + 2 ensemble)
2. Feature Engineering: Created 30+ engineered features (vs minimal in midterm)
3. Ensemble Methods: Added Voting Classifier and Bayesian Ensemble
4. Comprehensive Evaluation: 80 total metrics (vs basic accuracy in midterm)
5. Visual Analysis: 6 comprehensive visualizations documenting all aspects
6. Statistical Rigor: Added distribution analysis, rankings, consistency metrics

Methodology Improvements:

- Proper 70-15-15 split maintained throughout
- Pipeline prevents data leakage consistently
- Stratification ensures class balance
- Cross-validation confirms stability
- Multiple metrics provide comprehensive assessment

5. CONCLUSION

In this project, we successfully managed to develop a highly accurate machine learning model for student performance prediction with Random Forest and Gradient Boosting. Both achieved a perfect 100% test accuracy.

Key Achievements:

- Best individual models: Random Forest and Gradient Boosting achieved 100.0% test accuracy
- Best ensemble: Voting Classifier achieved 99.3% test accuracy
- Feature engineering: Created 30+ features, with 70% of top 10 being engineered
- Generalization: Excellent stability with only 3.25% average val-test gap
- Comprehensive evaluation: 80 metrics across 8 models documented with 6 visualizations

Most Important Findings:

1. Academic Performance Metrics Dominate: total_score and avg_score account for 33% of predictive power
2. Engagement Interactions Matter: attendance_x_avg_score shows synergy between attendance and performance
3. Tree-Based Models Excel: 5 of top 6 performers use decision tree foundations
4. Feature Engineering Critical: 70% of top features are engineered, demonstrating value of domain knowledge
5. Perfect Classification Achievable: Proper feature engineering enables even simple trees to achieve 99.3%

Practical Impact:

- Provides validated framework for educational early warning systems
- Identifies top 8 features accounting for 80% of prediction power (simplified implementation)
- Demonstrates that behavioral/performance factors more predictive than demographics
- Offers multiple model options based on accuracy/speed/interpretability trade-offs

Model Recommendations:

- Maximum Accuracy: Random Forest or Gradient Boosting (100%)
- Production Balance: Voting Classifier (99.3% with ensemble robustness)
- Speed Priority: Bayesian Ensemble (84% with fast inference)
- Interpretability: Decision Tree (99.3% with transparent rules)