INDIVIDUAL CONTRIBUTION REPORT

ITAI 1371 - Machine Learning
Student Performance Prediction Using Classification
Alhassane Samassekou

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

CONTRIBUTION SUMMARY

I was responsible for implementing the data splitting strategy, training two of the best-performing models (Random Forest and Gradient Boosting), developing the Voting Ensemble, and creating all performance visualizations for the project.

1. DATA SPLIT IMPLEMENTATION (70-15-15)

I designed and implemented the stratified three-way data split that became the foundation for all model training and evaluation.

Implementation:

```
# First split: 70% train, 30% temporary
X_train_final, X_temp, y_train_final, y_temp = train_test_split(
    X, y, test_size=0.30, random_state=42, stratify=y
)
# Second split: 15% validation, 15% test
X_val_final, X_test_final, y_val_final, y_test_final = train_test_split(
    X_temp, y_temp, test_size=0.50, random_state=42, stratify=y_temp
)
```

Result: Created Training (700 samples), Validation (150 samples), and Test (150 samples) sets with proper stratification to prevent data leakage.

2. RANDOM FOREST MODEL

I trained and optimized the Random Forest Classifier, which achieved perfect 100% test accuracy.

Configuration:

- 300 estimators with max_depth=15

- Class-balanced weighting
- Integrated preprocessing pipeline

Performance:

- Validation Accuracy: 99.3%
- Test Accuracy: 100.0% (Perfect)
- Test ROC-AUC: 100.0%
- Generalization Gap: 0.67%

Outcome: One of two models achieving perfect classification, demonstrating excellent ensemble learning.

3. GRADIENT BOOSTING MODEL

I trained the Gradient Boosting Classifier, which also achieved perfect 100% test accuracy.

Configuration:

- 200 estimators with learning_rate=0.05
- Max_depth=6 for base learners
- Sequential error correction

Performance:

- Validation Accuracy: 98.0%
- Test Accuracy: 100.0% (Perfect)
- Test ROC-AUC: 100.0%
- Generalization Gap: -2.0% (improved on test)

Outcome: Validated Random Forest results through different ensemble methodology, confirming robustness.

4. VOTING ENSEMBLE DEVELOPMENT

I designed and implemented the Voting Classifier that combined our three best models.

Design:

- Automatic selection of top 3 models by validation accuracy
- Soft voting to average probability predictions

- Combined Random Forest, Gradient Boosting, and Decision Tree

Performance:

- Test Accuracy: 99.3%
- Test Recall: 100.0% (zero false negatives)
- Test ROC-AUC: 100.0%

Value: Provided production-ready ensemble with 99.3% accuracy and perfect recall, ensuring no at-risk students missed.

5. PERFORMANCE VISUALIZATIONS

I created six comprehensive visualizations documenting all model results:

1. Confusion Matrices (2×4 Grid): All 8 models on test set with accuracy labels
2. ROC Curves (Dual Plot): Validation and test sets with AUC values for all models
3. Dual Bar Chart: Validation vs test comparison across all metrics
4. Feature Importance: Top 25 features color-coded (engineered vs original) with cumulative curve
5. Ensemble Analysis (6-Panel): Comprehensive comparison including accuracy bars, metrics, gaps, and confusion matrices
6. Statistical Summary (6-Panel): Distribution analysis, rankings, scatter plots, heatmaps, and performance categories