

UCI Student Performance Dataset

Contains attributes such as study time, past failures, parental education, school support, absences, and grades for Portuguese secondary school students. It predicts final grade performance (pass/fail or grade level).

Link: <https://archive.ics.uci.edu/dataset/320/student+performance>

Target Variable: Final Grade (converted into high vs. low performance)

Features: Demographic, social, and academic variables (e.g., age, family support, study time, absences, alcohol consumption, etc.)

Lab Structure

1. Concept Demonstration Notebook

Students will explore a ready-to-run Jupyter Notebook that demonstrates:

- Overview of bagging, boosting, and stacking.
- Implementation of Random Forest, Gradient Boosting, and VotingClassifier.
- Comparison of ensemble vs. single model performance.

Each section includes explanatory markdowns and visualizations (confusion matrices, feature importance plots).

2. Implementation Steps Inside Notebook

Each student runs and reflects on:

1. Data Preprocessing: Handle missing data, categorical encoding, and feature scaling.
2. Baseline Models: Logistic Regression, Decision Tree.
3. Bagging Ensemble: Random Forest using `sklearn.ensemble.BaggingClassifier`.
4. Boosting Ensemble: Gradient Boosting and AdaBoost comparison.
5. Stacking Ensemble: Combine models (e.g., Logistic Regression + Random Forest + Gradient Boosting).

Evaluation metrics: Accuracy, Precision, Recall, F1 Score.

3. Assignments: Reflective Questions

Students will answer five critical reasoning questions:

1. Which ensemble method achieved the best accuracy? Why might that be so?
2. How does variance reduction differ between bagging and boosting?
3. In the stacking approach, which base model contributed most to performance?
4. Discuss feature importance and its implications for student success.
5. How could this framework be adapted to predict student dropout risk?

Responses can be written inside notebook markdown cells or as a short group discussion summary.