Project Update 2 – Machine Learning Project

Project Title: Predicting Student Academic Performance Using Machine Learning
Group Members: Thanh Nguyen, Santiago Ponce, Josh Rubino
Date: July 28, 2025

1. Significant Progress:

   Since the last update, our group has made strong progress in both implementation and evaluation of our models:

   - We have completed full data loading and preprocessing, including handling categorical features and scaling.
   - Implemented both regression models (Linear Regression, SVR) and classification models (Logistic Regression, Naive Bayes, SVM).
   - Added advanced modeling techniques, including:
     - Support Vector Regression (SVR) for better non-linear prediction of final grades.
     - Support Vector Machine (SVM) for classification of performance groups.
     - PCA (Principal Component Analysis) for dimensionality reduction and explained variance analysis.
     - GridSearchCV for tuning hyperparameters (SVM).
   - Included confusion matrix visualization, feature importance analysis, and model performance comparisons.
   - Used class binning and distribution visualization to assess potential imbalance in classification labels.

2. Revisions:

   We revised our plan in the following ways:

   - Focus Shift: We clarified that our primary focus is regression, specifically predicting G3 final grade using SVR. Classification models are used to supplement this with practical groupings for intervention.

- Model Expansion: Based on instructor feedback, we added additional classification models such as SVM, and will soon include Decision Tree and Random Forest to strengthen performance comparison.
- Class Imbalance Awareness: We added class distribution checks and discussed potential remedies like resampling or using class weights.

3. Tasks Achieved (from Implementation Plan):

Task from Proposal	Status	Contributor
Data Collection	Completed	Josh
Exploratory Data Analysis (EDA)	Completed	Thanh
Preprocessing (encoding, scaling)	Completed	Santiago
Regression Models (Linear, SVR)	Completed	Thanh
Classification Models (LogReg, NB, SVM)	Completed	Josh and Santiago
PCA Analysis	Completed	Thanh
Model Tuning (GridSearchCV)	Completed	Josh
Visualization & Plots	Completed	Thanh

4. Challenges and Solutions:

Challenge	Solution
● Class imbalance warnings in classification	● Used zero_division=0 to suppress warnings and analyzed distribution
● Low recall in certain categories (e.g. “High”)	● Added SVM tuning and plan to explore Random Forest
● Complex model comparison	● Created unified table and plotted accuracy/F1 for all models

5. Preliminary Results:

Model	Type	Metric	Result
Linear Regression	Regression	MSE	12.9
SVR (RBF kernel)	Regression	MSE	11.7
Logistic Regression	Classification	F1 (macro avg)	0.70
Naive Bayes	Classification	F1 (macro avg)	0.65
SVM	Classification	F1 (macro avg)	0.74

- SVR shows better performance for grade prediction than Linear Regression.
- SVM (RBF) currently performs best for classifying students into Low/Medium/High.

6. Next Steps:
    - Finalize and evaluate Decision Tree and Random Forest classifiers.
    - Visualize feature importances and decision boundaries (if time allows).
    - Integrate model saving for potential deployment.
    - Polish final report and prepare presentation slides.

7. Conclusion:

At this stage in our project, we have successfully transformed our initial proposal into a fully operational machine learning pipeline. We have conducted extensive data preprocessing, implemented multiple models, and evaluated them using industry-standard metrics. The use of both regression and classification techniques has allowed us to approach the problem from complementary perspectives predicting exact student grades using SVR and identifying performance categories using classifiers like Logistic Regression, Naive Bayes, and SVM.

In response to instructor feedback, we clarified our project's primary focus on regression while strengthening our classification track through class balance analysis and hyperparameter tuning. This dual-track approach provides flexibility for real-world use cases, where educators may require both fine-grained grade predictions and broader performance groupings to intervene early.

Our initial results are promising, particularly the improved performance of SVR over linear regression in capturing non-linear relationships, and the strength of SVM in classification tasks. We've also deepened our analysis through feature importance visualization and confusion matrix evaluation, which enhances the interpretability and transparency of our findings.

Looking ahead, we plan to expand our classification models to include tree-based approaches such as Decision Trees and Random Forests. These models will be evaluated not just for accuracy but for their ability to provide explainable results that educators can act upon. We also aim to prepare our final report, polish visualizations, and potentially build an interface or demo that allows predictions on new student data.

Overall, we are confident in the direction and quality of our work. We believe our project not only fulfills academic objectives but also presents practical applications for improving educational outcomes through data-driven decision-making.