UNIVERSIDAD
PANAMERICANA®

SG1_Team#2

Course:

Simulation & Data Visualization

Due date:

Thursday, May 29, 2025

Professor:

Gabriel Castillo Cortés

Team members:

Renata Calderón Mercado
Regina Ochoa Gispert
Miguel Angle Velez Olide

1. Dataset Search and Understanding

Challenge: At the beginning of the project, one of the main challenges was finding a database that truly fit the project objective: predicting students' academic performance. Many available datasets were incomplete, irrelevant, poorly structured, or lacked useful variables for predictive analysis.

Solution: After exploring repositories like Kaggle, UCI Machine Learning Repository, and GitHub, I found a dataset that included academic, personal, and study habit characteristics. Even so, extensive cleaning was necessary to make it usable.

Sources explored:

- Kaggle.com
- UCI Machine Learning Repository
- GitHub (open educational repositories)

Selection criteria applied:

- Presence of relevant academic performance indicators
- Inclusion of lifestyle and behavioral variables
- Sufficient sample size for statistical significance
- Data quality and completeness assessment

2. Data Cleaning and Preprocessing

Challenge: The dataset contained null values, inconsistent entries (for example, different ways of writing gender), and variables with little impact on prediction.

Solution:

- Standardized categories using pandas for consistent data representation
- Handled missing values by removing or imputing them with mean/mode as appropriate
- Applied encoding to categorical variables with LabelEncoder and OneHotEncoder
- Removed non-predictive columns such as student IDs that added no analytical value

Specific preprocessing steps:

- Gender standardization: "male", "Male", "M" → unified format
- Missing value strategy: median for numerical, mode for categorical variables
- Outlier detection using IQR method and clipping extreme values
- Feature engineering: created composite variables like wellness_index and study_ratio

3. Visualization and Exploratory Data Analysis

Challenge: Finding meaningful relationships between variables and academic performance.

Solution:

- Used matplotlib and seaborn to visualize:
    - Grade distributions
    - Variable correlations
    - Study habit vs. performance comparisons
- Dictionary implementation advantage: Using dictionaries made it possible to group and classify variable values more clearly, facilitating the generation of comprehensible and aesthetically appropriate graphs. This tool helped us assign readable and ordered labels to categorical data, improving visual interpretation.

Key stories discovered in the data:

1. Students who dedicate more than 2 hours to daily study tend to have better performance.
2. Variables related to family support and mental health showed significant impact.
3. Sleep patterns and exercise frequency correlated with academic success.
4. Screen time showed negative correlation with study effectiveness.

4. Model Training

Challenge: Selecting an effective model that was easy to interpret and performed well.

Solution:

- Compared K-Nearest Neighbors and Random Forest algorithms
- Random Forest proved superior in accuracy and noise tolerance
- Data split: 70% training and 30% testing for robust evaluation

Model configuration:

- Random Forest with 100 estimators
- Maximum depth of 10 to prevent overfitting
- Feature importance analysis for interpretability
- Standard scaling applied to numerical features

5. Model Evaluation

Challenge: Correctly evaluating the model, especially with imbalanced classes.

Solution:

- Used comprehensive metrics: accuracy_score, confusion_matrix, and classification_report
- Identified model difficulty in predicting "medium" performance students
- Considered weight adjustment but didn't implement due to time constraints

Evaluation results:

- Overall accuracy: acceptable performance level
- Precision and recall analysis revealed class-specific strengths
- Feature importance highlighted study habits as primary predictors
- Model successfully identified at-risk student patterns

6. Unresolved Problems and Future Improvements

Current limitations:

- Missing emotional/psychological factors that could be key predictors
- No hyperparameter optimization techniques like GridSearchCV applied
- Lack of cross-validation for more robust performance assessment
- Static model that doesn't retrain with new data

Proposed future enhancements:

- Expand dataset to include psychological and socioeconomic variables
- Implement automated hyperparameter tuning for optimal performance
- Add cross-validation for better generalization assessment

- Develop dynamic retraining pipeline for continuous model improvement
- Include ensemble methods to improve prediction accuracy
- Implement explainable AI techniques for better result interpretation

Conclusion

This academic performance prediction project represents a comprehensive demonstration of how machine learning can be successfully applied to educational data analysis, despite the inherent challenges of working with real-world datasets. The journey from data discovery to model implementation revealed both the potential and limitations of predictive analytics in educational contexts.

The project successfully established a complete machine learning pipeline that transforms raw educational data into actionable insights. By identifying study habits, wellness factors, and family support as primary predictors of academic success, the model provides evidence-based foundations for educational interventions. The Random Forest algorithm's ability to handle mixed data types while maintaining interpretability proved particularly valuable for this domain, where understanding *why* predictions are made is as important as the predictions themselves.

The technical implementation showcased essential data science skills, from data preprocessing and feature engineering to model evaluation and interpretation. The creation of composite variables like the wellness_index and study_ratio demonstrates sophisticated feature engineering that goes beyond simple data cleaning to extract meaningful insights from multiple related variables.