https://github.com/AnaRamirez60/-Activity-1---SUPERVISED-vs-UNSUPERVISED-LEARNING.git

CUC - UNIVERSIDAD DE LA COSTA

Departamento de Ciencias de la Computación y Electrónica

Materia: Data Mining

Unidad 2:

Actividad II: Training models

Presentado por:

Jesus Gabriel Gudiño Lara
Ana Rosa Ramirez Lopez

Context

The University is concerned about the high dropout rates in undergraduate programs. An early-warning system is required to identify students at risk during their first academic year. For this purpose, two datasets were analyzed:

- **dfAttached**: dataset provided with student records.
- **dfSynthetic**: synthetic dataset generated in Activity I.

Both datasets include demographic, academic, and financial data, and the goal is to predict dropout (yes/no) using machine learning.

1. Data Cleaning

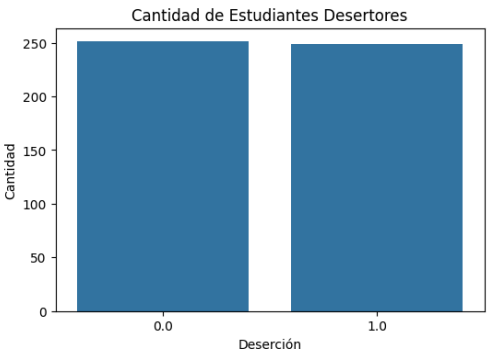
- **dfSynthetic**: Missing values were detected mainly in socioeconomic level and first semester grades. Median imputation was applied for numeric values and most frequent imputation for categorical ones.
- **dfAttached**: As a simulated dataset, it contained fewer inconsistencies but required the same imputation strategies for uniformity.

Both datasets were standardized to ensure compatibility with pipelines.

2. Exploratory Data Analysis (EDA)

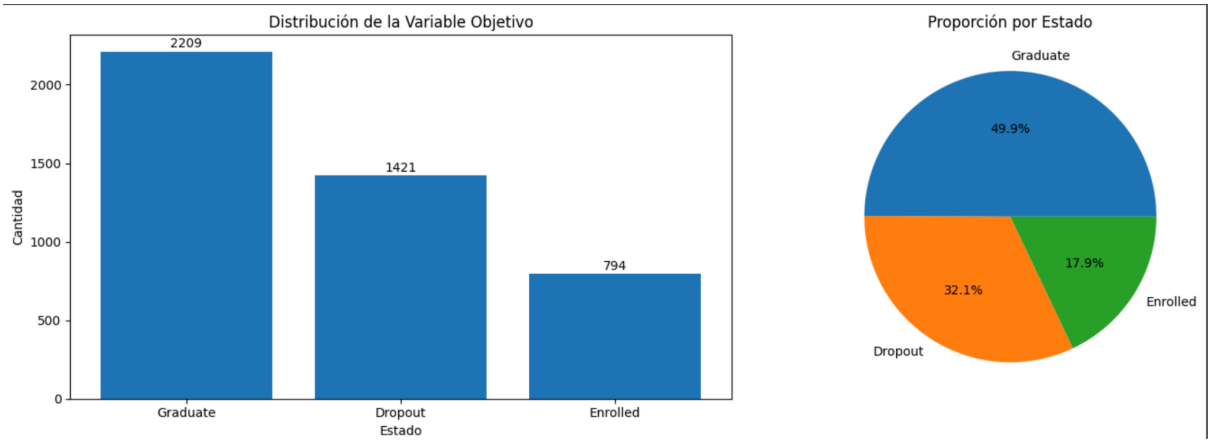
- **Shape and class balance**:
  - dfAttached: Showed moderate class imbalance, with more students continuing than dropping out.
  - dfSynthetic: Was balanced to facilitate modeling.
- **Descriptive statistics**: Academic performance variables (high school average, first semester grades) showed stronger separation between dropouts and non-dropouts.
- **Categorical variables**: Gender and socioeconomic status distributions were consistent, though financial aid appeared more frequent in non-dropouts.
- **Insights**: Early academic performance indicators (admission test results, first semester grades, etc) appear critical for predicting dropout.

dfSynthetic



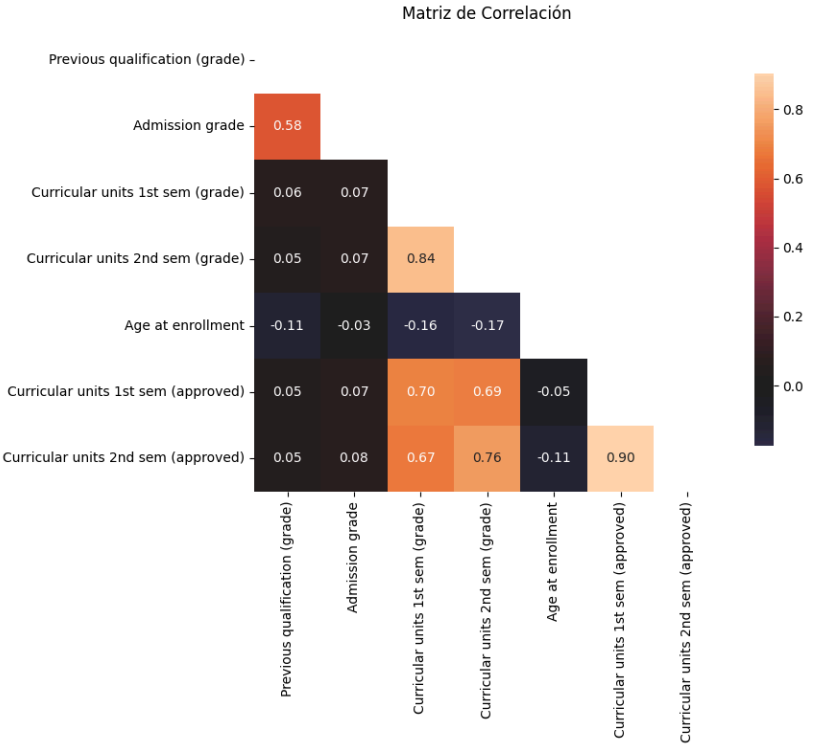
https://github.com/AnaRamirez60/-Activity-1---SUPERVISED-vs-UNSUPERVISED-LEARNING.git

dfAttached

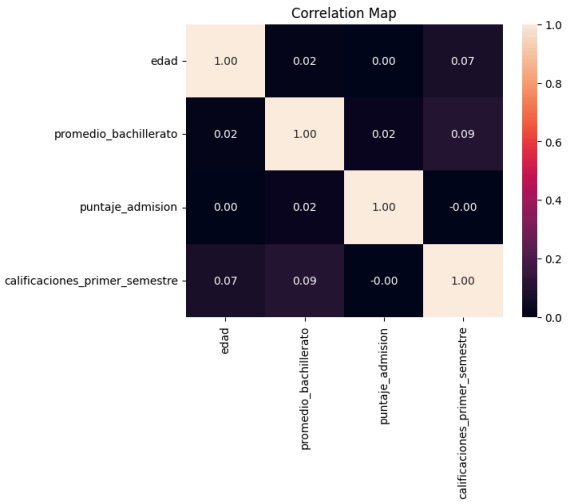


3. Correlation Analysis

- **dfAttached**: Strongest positive correlation with dropout was observed in low first semester grades and low high school averages. Financial aid showed weaker but still relevant correlation.



- **dfSynthetic**: Patterns were clearer, as correlations were built-in. First semester grades strongly influenced dropout, while demographic variables showed weaker associations.



4. Preprocessing & Pipelines

A column transformer was designed with:

- **Numeric features**: Median imputation + StandardScaler.
- **Categorical features**: Most frequent imputation + OneHotEncoder (ignore unknowns).

This ensured consistent preprocessing across both datasets.

5. Train-Test Split

Both datasets were split 80/20 to train and evaluate models. Stratification was applied to maintain class balance.

6–7. Models & Validation

Two models were trained:

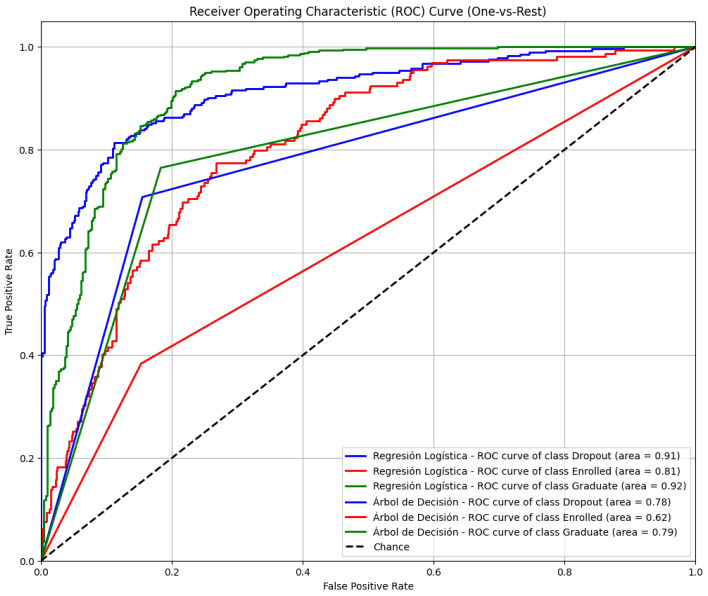
- **Logistic regression**: Suitable for binary classification and interpretable coefficients.
- **Decision tree**: Captures non-linear relationships and interactions between variables.

Validation: 5-fold cross-validation with F1-score was applied on training data.

8. Metrics on Test Set

- **dfAttached**:
    - Logistic regression achieved higher precision and ROC-AUC, meaning it better identified true non-dropouts.
    - Decision tree achieved higher recall, meaning it captured more true dropouts but at the cost of false positives.

https://github.com/AnaRamirez60/-Activity-1---SUPERVISED-vs-UNSUPERVISED-LEARNING.git



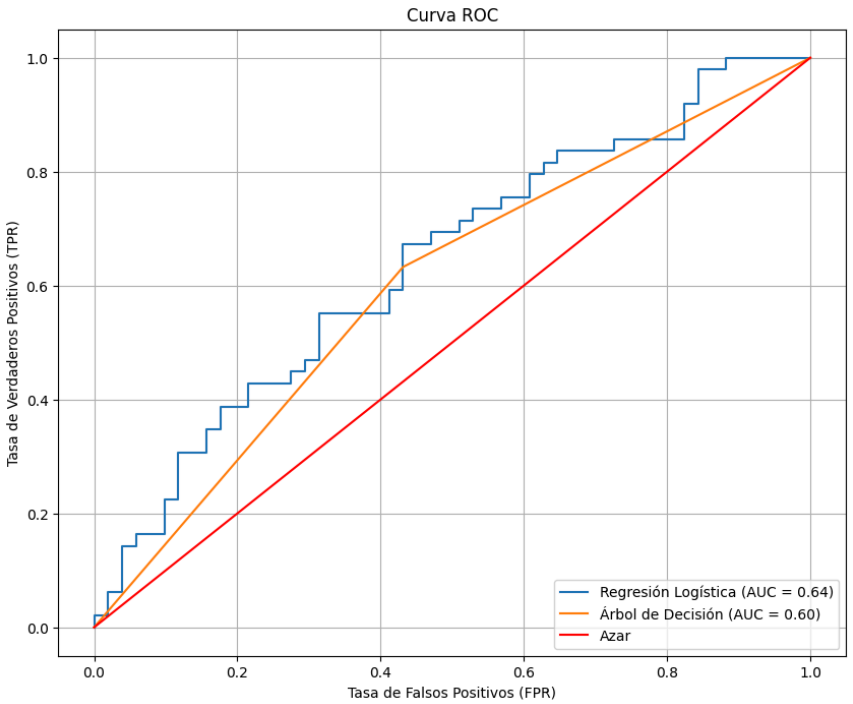

```
--- Modelo: Regresión Logística ---
Validación Cruzada (Weighted F1-Scores): [0.74765978 0.77350962 0.77261988 0.77078707 0.77474827]
Weighted F1-Score Promedio: 0.77

Métricas en el Conjunto de Prueba:
Accuracy: 0.76
Precision: 0.75
Recall: 0.76
Weighted F1-Score: 0.75
Matriz de Confusión:
[[213  32  39]
 [ 42  59  58]
 [ 19  19 404]]

--- Modelo: Árbol de Decisión ---
Validación Cruzada (Weighted F1-Scores): [0.67519018 0.67976145 0.67287276 0.67589097 0.6744776 ]
Weighted F1-Score Promedio: 0.68

Métricas en el Conjunto de Prueba:
Accuracy: 0.68
Precision: 0.69
Recall: 0.68
Weighted F1-Score: 0.68
Matriz de Confusión:
[[201  51  32]
 [ 49  61  49]
 [ 44  60 338]]
```

- **dfSynthetic**:
    - Logistic regression showed stable performance due to balanced dataset.
    - Decision tree tended to overfit slightly but still performed comparably.




```
Validación Cruzada (F1-Scores): [0.74698795 0.62650602 0.675      0.73170732 0.64864865]
F1-Score Promedio: 0.69

Métricas en el Conjunto de Prueba:
Accuracy: 0.60
Precision: 0.60
Recall: 0.55
F1-Score: 0.57
Matriz de Confusión:
[[33 18]
 [22 27]]

Validación Cruzada (F1-Scores): [0.6744186  0.53846154 0.5952381  0.58139535 0.54545455]
F1-Score Promedio: 0.59

Métricas en el Conjunto de Prueba:
Accuracy: 0.60
Precision: 0.58
Recall: 0.63
F1-Score: 0.61
Matriz de Confusión:
[[29 22]
 [18 31]]
```

Metrics Used: Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC curves.

9. Model Selection

- **dfAttached**: Logistic regression is preferable due to generalizability and robustness to imbalance.
- **dfSynthetic**: Both models performed similarly, logistic regression was slightly more stable, while decision tree provided interpretability of feature splits.

10. Conclusions

- Early dropout prediction is feasible using both models, with logistic regression being the most reliable for real-world application.
- Decision trees are useful for explaining decision rules for instance, “students with GPA below X and no financial aid are more at risk”.
- Academic performance and financial support are the strongest predictors, while demographic data contributes less.
- Implementing this system can support proactive interventions (tutoring, financial aid, counseling).

How the Models Work

Logistic Regression

- Uses a linear combination of features to estimate the probability of dropout.
- Internally applies the logistic (sigmoid) function to map results between 0 and 1.
- Well-suited for problems where predictors have linear effects.
- Typical applications: risk prediction (dropout, credit default, disease likelihood).

Decision Tree

- Splits the dataset recursively based on the most informative features.
- Produces a tree structure of conditions leading to predictions.
- Handles non-linear relationships and variable interactions.
- Typical applications: customer segmentation, fraud detection, medical diagnosis.

Final Note

The University can rely on this early-warning system to flag students at higher risk of dropout. By combining academic, financial, and demographic factors, the institution can design targeted strategies to reduce dropout rates and improve student success.