1. Predicting Depression in Mental Health Data Using Supervised Learning

Work done by Group_A2_77:

- Francisco Miguel Pires Afonso (**up202208115**)
- Miguel Moita Caseira (**up202207678**)
- Pedro Trindade Gonçalves Cadilhe Santos (**up202205900**)

2. Problem Definition and Context

- This project addresses a **binary classification problem**, aiming to detect the presence of depressive symptoms in individuals based on various personal, academic, and lifestyle factors.

- The **target variable** is **Depression**:
  - **0** = No symptoms of depression
  - **1** = Signs of depression present

- We apply **Supervised Learning techniques** to learn patterns from labeled data and predict the mental health status of individuals.

- The problem is aligned with the **IART** assignment goals: to build and evaluate ML models following a complete pipeline - from exploratory data analysis to model selection, training, and performance comparison.

- Dataset source: **Kaggle Playground Series – S4E11** (train.csv and test.csv)

3. Dataset Overview and Preprocessing

- Dataset contains ~140k samples, with diverse features (personal, academic, lifestyle).

- The target variable **Depression** is binary and class-imbalanced: ~82% class 0 (no symptoms), ~18% class 1 (symptoms present).

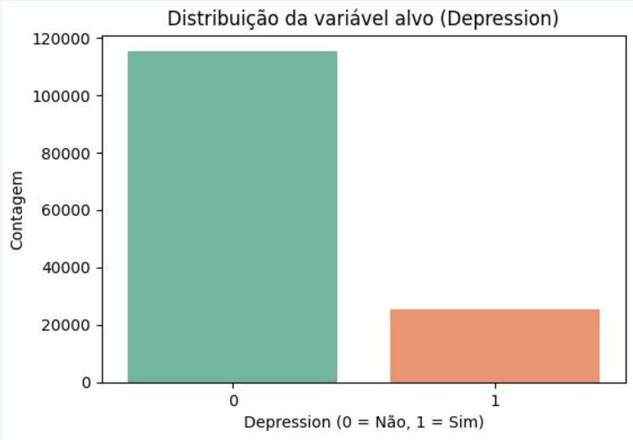
- Removed irrelevant columns: **id**, **Name**, etc.

- Imputed missing values:
  - **Numerical**: median
  - **Categorical**: most frequent

- Encoded categorical variables with LabelEncoder.

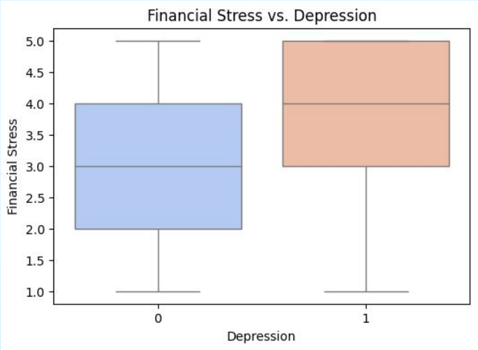
- Normalized numeric features using StandardScaler.

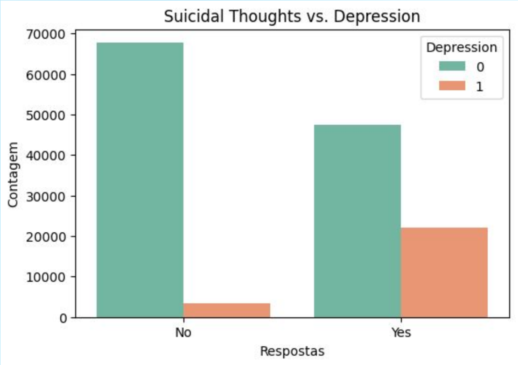
- Final dataset: cleaned, numeric and ready for model training.



4. Exploratory Data Analysis (EDA)

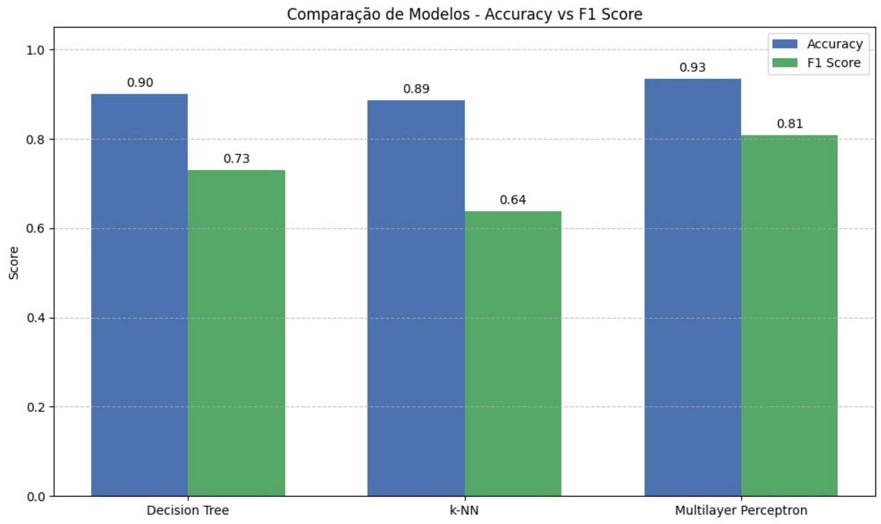
- Analyzed distribution of key features across the dataset:
- Observed **strong class imbalance** (confirmed previously).
- Correlation heatmap used to identify relationships between numeric features.
- Boxplots revealed potential associations between:
    - **Financial Stress, CGPA, Work/Study Hours and Depression.**
- Some features (e.g. Family History, Suicidal Thoughts) showed clear association with Depression.





5. Initial Models and Performance

- Tested three supervised learning models with default parameters:
  - **Decision Tree**
  - **k-Nearest Neighbors (k-NN)**
  - **Multilayer Perceptron (MLP)**
- Evaluated with **Accuracy** and **F1 Score** on test data.
- Performance varied, especially on class 1 (minority class).
- MLP showed best balance between accuracy and recall.



6. Hyperparameter Tuning

- Applied **GridSearchCV** to tune parameters for:
  - **Decision Tree**
  - **k-Nearest Neighbors (k-NN)**
  - **Multilayer Perceptron (MLP)**

- Used **F1 Score** as the scoring metric (focus on minority class performance).

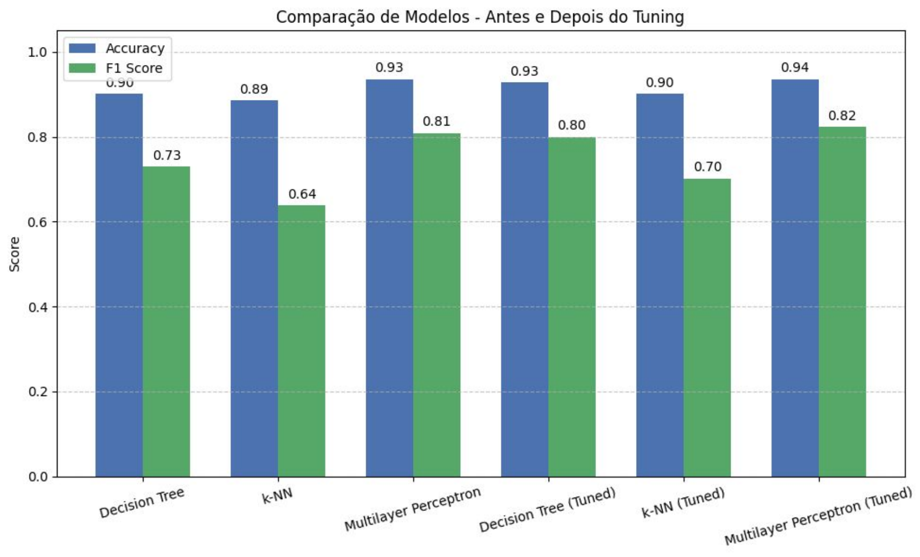
- **Cross-validation** (cv=3) used to ensure reliable evaluation.

- Tuned parameters included:
  - Tree depth, splitting criteria (for **Decision Tree**): {'class_weight': None, 'criterion': 'gini', 'max_depth': 10, 'min_samples_split': 10}
  - Number of neighbors, distance metric (for **k-NN**): {'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'distance'}
  - Hidden layer size, activation, learning rate (for **MLP**): {'activation': 'tanh', 'alpha': 0.0001, 'hidden_layer_sizes': (50,), 'learning_rate': 'constant'}

- Best parameters were selected and used for final evaluation.

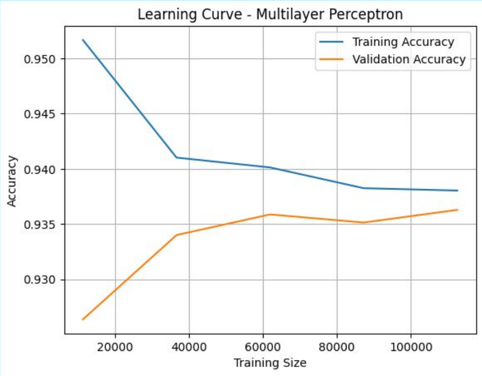
7. Performance after Tuning

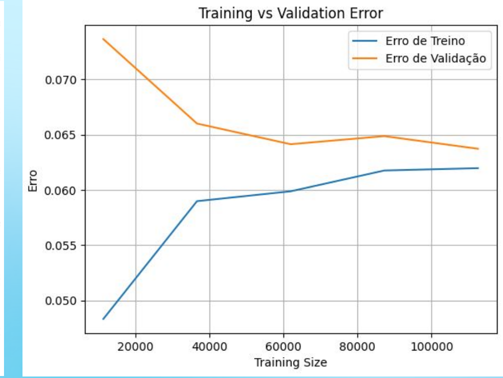
- All three models were re-evaluated using their best parameters from tuning.
- Performance was compared using **Accuracy** and **F1 Score**.
- Improvements observed, especially in **k-NN** and **Decision Tree**.
- **MLPClassifier** remained the best overall, showing both high accuracy and good sensitivity to the minority class.



8. Learning Curve and Generalization

- Learning curve shows model performance on both training and validation sets as training size increases.
- The **validation accuracy increases** and **approaches training accuracy**, indicating good generalization.
- **Small gap between curves** suggests no overfitting.
- Model performance stabilizes after ~80,000 samples — additional data brings diminishing returns.
- Complementary error plot confirms **low and converging error rates**.





9. Class-by-Class Performance

- Used classification_report to assess precision, recall and F1 Score per class.
- Focus on class 1 (**Depression = Yes**), which is the minority and most critical.
- **MLPClassifier (Tuned) achieved:**
    - **Precision: 0.94**
    - **Recall: 0.94**
    - **F1 Score: 0.94**
- **Decision Tree** and **k-NN** showed lower recall and F1 for class 1.
- MLP generalizes best without sacrificing sensitivity to positive cases.

Relatório de Classificação - Decision Tree (Tuned)

	precision	recall	f1-score	support
0	0.95	0.96	0.96	23027
1	0.81	0.79	0.80	5113
accuracy			0.93	28140
macro avg	0.88	0.87	0.88	28140
weighted avg	0.93	0.93	0.93	28140

Relatório de Classificação - k-NN (Tuned)

	precision	recall	f1-score	support
0	0.92	0.96	0.94	23027
1	0.78	0.64	0.70	5113
accuracy			0.90	28140
macro avg	0.85	0.80	0.82	28140
weighted avg	0.90	0.90	0.90	28140

Relatório de Classificação - Multilayer Perceptron (Tuned)

	precision	recall	f1-score	support
0	0.96	0.96	0.96	23027
1	0.83	0.82	0.82	5113
accuracy			0.94	28140
macro avg	0.89	0.89	0.89	28140
weighted avg	0.94	0.94	0.94	28140

10. Conclusions and Results

- **MLPClassifier (Tuned)** was the best-performing model:
  - **Accuracy**: 0.94 | **F1 Score**: 0.94
- Parameter tuning improved performance, especially for **k-NN** and **Decision Tree**.
- Learning curves showed **no overfitting** and good generalization.
- Classification reports confirmed strong performance for **class 1 (Depression = Yes)**.
- The final model was successfully applied to the **unseen test set** (test.csv), using the same pipeline for preprocessing and encoding (including handling of previously unseen categorical values).