Diego Armando Salinas Lugo ds24353 2401168

README - ADHD Data Exploration (Stage 1)

Project Overview

This project stage focuses on data exploration and preprocessing for ADHD datasets, so that later a model for diagnosis can be created. The dataset consists of behavioral, demographic, and medical information. Stage 1 involves cleaning, handling categorical and numerical features, and analyzing correlations to prepare for model training in Stage 2.

How to use the code

Requirements:

To run this notebook, the following dependencies are needed:

- **Python version**: 3.11 (or compatible version)
- **Required libraries**:

  pip install numpy pandas matplotlib seaborn scikit-learn

Running the Notebook:

1. Open the Jupyter Notebook.
2. Ensure the dataset is correctly loaded in the expected format.
3. Execute the cells step by step to perform:
   - Data loading & exploration
   - Data cleaning & preprocessing
   - Handling categorical & numerical variables
   - Feature correlation analysis
   - Saving the cleaned dataset for Stage 2
4. The final cleaned dataset will be saved as cleaned_train_data.csv.

Assumptions & Preprocessing Decisions

- **Categorical Features**:
  - Maintained as categorical where appropriate.
  - Temporarily encoded to numerical for correlation analysis.

- **Numerical Features**:
    - Not normalized in Stage 1 (Normalization will be done in Stage 2 for modeling). Since Stage 1 focuses on exploration and preprocessing, ensuring data integrity before applying transformations that could affect its interpretation.
- **Missing Data**:
    - For metadataset_a where features are numerical, the media was applied for the NaN values.
    - For metadataset_b where features are categorical, the mode was applied for the NaN values.
- **Feature Importance**:
    - Exploratory analysis performed to assess feature importance using correlation.

Expected Output

After executing all steps in the notebook, the user should have:

- A cleaned and structured dataset (cleaned_train_data.csv).
- Insights into feature relationships through correlation matrices & visualizations.
- A clear plan for Stage 2, where feature scaling and model training will take place.

Next Steps (Stage 2)

- Implement feature scaling (normalization/standardization) for numerical features.
- Apply machine learning models for ADHD prediction.
    - Potential Models:

- **Logistic Regression**: Since it is a linear model which can estimate the probability of ADHD presence based on the features. It can also provide easily interpretable coefficients, allowing to understand which features contribute the most to ADHD classification.
- **k-Nearest Neighbors (k-NN)**: ADHD classification could benefit from k-NN due to similar behavioral or demographic patterns among participants.
- **Random Forest:** ADHD diagnosis is influenced by different behavioral, demographic, and cognitive features. Random Forest’s capability to handle mixed data types makes it robust for this dataset. Besides, it automatically captures feature importance, which can help to get insights into the features influence.
- **Support Vector Machine (SVM):** ADHD diagnosis may not be linearly separable, meaning a model that finds non-linear decision boundaries could be very helpful. SVMs can identify complex patterns, improving classification accuracy.

- **Neural Networks**: As ADHD diagnosis is a multi-faceted problem with diverse variables interacting in non-trivial ways. NN can automatically detect these interactions, making them ideal for capturing complex relationships.

- **Evaluate the models with metrics**

    - Accuracy as a general performance measure.
    - Precision & Recall to address potential class imbalance issues.
    - AUC-ROC To evaluate model discrimination ability.