MDD Data Dictionary

This data dictionary describes all variables contained in **mdd_scores_baseline.csv**, which is used for the prediction of depressive symptom severity. Each entry includes the variable name, type and a brief description.

Variable Definitions

gender

Type: Categorical (0 or 1)
Description: Participant gender. Encoded numerically after preprocessing. Exact mapping depends on the original dataset structure.

Age group

Type: Categorical (integer codes)
Description: Age group classification for each participant. Represents grouped age categories rather than exact ages.

impatient

Type: Categorical (0 or 1)
Description: Indicates whether the participant was receiving inpatient care at the time of assessment.

marital

Type: Categorical (integer codes)
Description: Marital status of the participant. Encoded numerically after preprocessing. Categories may include single, married, separated, divorced or equivalent.

work

Type: Categorical (integer codes)
Description: Employment or work status. Original nonnumeric labels are converted to numeric values during preprocessing. Rows with unconvertible values (for example, old job codes with letters) were removed.

madrs1

Type: Continuous (numeric)
Description: Baseline **Montgomery–Åsberg Depression Rating Scale (MADRS)** score. Represents initial depression severity and serves as the **target variable** in the current modeling setup.

madrs2

Type: Continuous (numeric)
Description: Follow up MADRS score collected later point. Useful for understanding symptom change and often treated as an outcome in longitudinal studies. In the present project, it is used as a predictor.

age

Type: Continuous (numeric)
Description: Participant's chronological age in years.

education

Type: Continuous (numeric)
Description: Years of formal education or education level coded numerically.

condition

Type: Categorical (0 or 1)
Description: Experimental or clinical condition group. May represent study condition, treatment arm or control group.

control

Type: Categorical (0 or 1)
Description: Indicator for control group membership when applicable. May overlap with or complement the condition variable depending on study design.

Notes on preprocessing

The dataset originally contained mixed variable types and missing values. The following preprocessing steps were applied:

- Nonnumeric variables such as work were converted to numeric form.
- Remaining missing values in numeric variables were imputed using the mean.
- Rows with remaining incomplete information were removed after imputation.
- The final dataset used for modeling is a fully numeric subset of the original file.

Usage in modeling

- **Target variable: madrs1**

- **Predictor variables:** All other numeric columns
- **Model types applied:** Linear regression and Random Forest regression
- **File location in repo:** data/mdd_scores_baseline.csv