PS3192 Coursework dataset information

Matteo Lisi

Table of contents

Data Dictionaries

Below is a collection of data dictionaries for the main datasets referenced in the coursework. Each section highlights the most important variables, their general meaning, and (where applicable) the target or outcome variable for analyses.

`college_recent_grads`

Below is a summary of the commonly referenced variables in the `college_recent_grads` dataset from the `fivethirtyeight` package. This dataset contains information on recent college graduates, including median salaries, major categories, gender composition, etc.

- **rank**
  Integer rank based on median earnings (1 = highest median earnings).
- **major_code**
  Numeric code identifying each major.
- **major**
  Name of the major (e.g., “Petroleum Engineering”).
- **major_category**
  Broad category to which the major belongs (e.g., “Engineering”).
- **total**
  Total number of people with that major.
- **sample_size**
  Number of survey respondents from that major (used to estimate statistics).
- **men**
  Number of men with that major.
- **women**
  Number of women with that major.
- **sharewomen**
  Proportion of women among graduates with that major.
- **employed**
  Number of individuals employed (full- or part-time).
- **employed_fulltime**
  Number of individuals employed full-time.
- **employed_parttime**
  Number of individuals employed part-time.
- **employed_fulltime_yearround**
  Number of individuals employed full-time, year-round.
- **unemployed**
  Number of individuals unemployed.
- **unemployment_rate**
  Proportion of the labor force (employed + unemployed) that is unemployed.
- **p25th**
  25th percentile of earnings.
- **median**
  Median earnings.
- **p75th**
  75th percentile of earnings.
- **college_jobs**
  Number of jobs requiring a college degree.
- **non_college_jobs**
  Number of jobs not requiring a college degree.
- **low_wage_jobs**
  Number of low-wage service jobs.

`titanic`

The `titanic` dataset includes information on Titanic passengers, including whether they survived or not.

- **name**
  Passenger's name (string).
- **survived**
  Integer indicating survival status (1 = survived, 0 = did not survive).
- **sex**
  Passenger's gender (e.g., "male", "female").
- **age**
  Passenger's age (in years). May contain missing or approximate values.
- **passengerClass**
  Ticket class (e.g., "1st", "2nd", "3rd").

(Target variable: `survived`)

`heartdisease`

The heartdisease dataset contains medical and demographic information about individuals, used for predicting heart disease risk.

- **Age**
  Age of the individual (numeric).
- **Sex**
  Biological sex of the individual ("M" = Male, "F" = Female).
- **RestingBP**
  Resting blood pressure in mmHg (numeric).
- **Cholesterol**
  Serum cholesterol level in mg/dL (numeric).
- **FastingBS**
  Binary indicator for fasting blood sugar > 120 mg/dL (1 = true, 0 = false).
- **RestingECG**
  Results of resting electrocardiogram test ("Normal", "ST", etc.).
- **MaxHR**
  Maximum heart rate achieved during exercise (numeric).
- **Angina**
  Presence of exercise-induced angina ("Y" = Yes, "N" = No).
- **HeartPeakReading**
  Peak exercise ST segment reading (numeric).
- **HeartDisease**
  Indicator of heart disease presence (1 = has heart disease, 0 = no heart disease).

(Target variable: `HeartDisease`)

`oiltype`

This dataset shows the fatty acid concentrations of commercial oils measured via gas chromatography. It is used to predict the **type of oil** (A = pumpkin, B = sunflower, C = peanut, D = olive, E = soybean, F = rapeseed, G = corn).

- **Palmitic**
  Palmitic acid concentration (percentage).
- **Stearic**
  Stearic acid concentration (percentage).
- **Oleic**
  Oleic acid concentration (percentage).
- **Linoleic**
  Linoleic acid concentration (percentage).
- **Linolenic**
  Linolenic acid concentration (percentage).
- **Eicosanoic**
  Eicosanoic acid concentration (percentage).
- **Eicosenoic**
  Eicosenoic acid concentration (percentage).
- **oilType**
  Factor indicating the oil type (A through G). *(Target variable for classification.)*

`forest_mapping`

Multi-temporal remote sensing data of a forested area in Japan, used to classify different forest types based on spectral information.

- **Class**: One of ‘s’ (Sugi), ‘h’ (Hinoki), ‘d’ (Mixed deciduous), ‘o’ (Other land).
- **b1 - b9**: ASTER image bands (spectral data in green, red, near-infrared) from three dates.
- **pred_minus_obs_S_b1 - pred_minus_obs_S_b9**: Difference between predicted vs. observed spectral values for the ‘S’ class.
- **pred_minus_obs_H_b1 - pred_minus_obs_H_b9**: Difference between predicted vs. observed spectral values for the ‘h’ class.

`autism`

Dataset related to screening for Autism Spectrum Disorder (ASD) in adults. It contains 10 behavioral features (AQ-10-Child) plus additional demographic characteristics. The outcome is often whether an individual is diagnosed with ASD or not.

- **A1_Score - A10_Score**: Behavioral screening questions (0 or 1).

- **age**: Age (int).
- **sex**: Sex (e.g., “m” or “f”) (chr).
- **ethnicity**: Participant’s ethnicity (chr).
- **jundice**: History of jaundice (yes/no) (chr).
- **austim**: Family history of autism (yes/no) (chr).
- **contry_of_res**: Country of residence (chr).
- **ASD_diagnosis**: Actual diagnosis label (yes/no) (chr).

`banknote_authentication`

Data extracted from images (400×400 pixels) of genuine and forged banknotes. Wavelet transforms used to compute features.

- **variance**: Variance of Wavelet Transformed image (num).
- **skewness**: Skewness of Wavelet Transformed image (num).
- **kurtosis**: Kurtosis of Wavelet Transformed image (num).
- **entropy**: Entropy of the image (num).
- **class**: 0 or 1, indicating genuine or forged banknote (int).

`iris`

Measurements of sepal length, sepal width, petal length, and petal width for 50 flowers from each of 3 species of iris. The **Species** column typically has 3 levels: *setosa*, *versicolor*, and *virginica*.

- **Sepal.Length**: Sepal length in cm (num).
- **Sepal.Width**: Sepal width in cm (num).
- **Petal.Length**: Petal length in cm (num).
- **Petal.Width**: Petal width in cm (num).
- **Species**: Factor of three iris species (factor).

For unsupervised learning, you can remove `Species` to treat the data as unlabeled as follow:

```r
iris <- iris %>% dplyr::select(-Species)
```

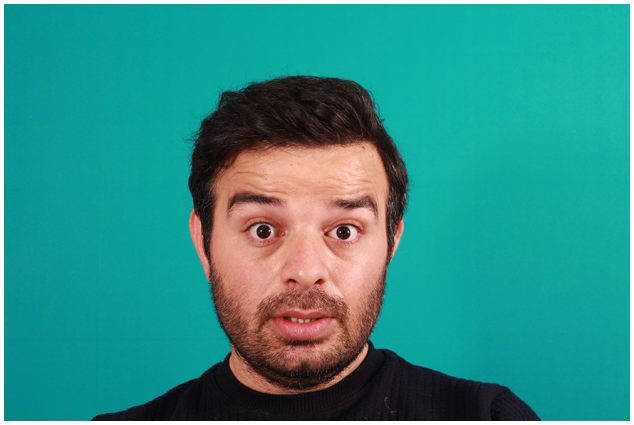
`faithful`

The **Old Faithful** geyser dataset, containing:

- **eruptions**: Eruption time in minutes.
- **waiting**: Waiting time to next eruption (in minutes).

`gazedata`

The gazedata dataset contains gaze fixation points collected from a sample of chidren participants as they viewed the below image.



The fixations represent pauses in visual exploration, measured with an eye tracker. The dataset includes the following variables:

- `x` Normalized horizontal position of gaze fixation on the image (range: 0 to 1).
- `y` Normalized vertical position of gaze fixation on the image (range: 0 to 1).

`wine`

Chemical analysis of wines grown in the same Italian region from three different cultivars. Thirteen constituents (originally ~30) are measured.

- **alcohol**: Alcohol content (num).

- **malic_acid**: Malic acid concentration (num).
- **magnesium**: Magnesium level (int).
- **tot_phenols**: Total phenols (num).
- **flavanoids**: Flavanoid content (num).
- **nonflavanoid_phenols**: Non-flavanoid phenols (num).
- **proanthocyanins**: Proanthocyanin content (num).
- **color_intensity**: Color intensity (num).
- **hue**: Hue (num).
- **proline**: Proline concentration (int).