Participant Guide & Data Dictionaries
AI4PH Summer Institute Data Challenge
July 2025

Overview

During the Summer Institute you will work with an existing dataset to complete a series of tasks (i.e., challenges) with your pre-assigned team members

These challenges will help you to build your skills in preparing distributed data for a research project and analyzing these data using federated learning approaches

Important Note:
You need to have R downloaded on your computer for all activities: https://www.r-project.org/
You will require the following R packages:
Please run this R script even if you have the required packages installed, as some functionality may depend on recent versions of the included packages.

```
install.packages(c(
 "rlang",
 "dplyr",
 "readr",
 "randomForest",
 "xgboost",
 "nnet",
 "yardstick",
 "ggplot2",
 "skimr"))
```

Python may also be used for selected activities and so you may wish to also download this software: https://www.python.org/downloads/

Dataset Characteristics

- UCI Heart Disease datasets
- Training and test datasets for 4 sites: Cleveland, Hungarian, Switzerland, Long Beach V
- Total: 920 records (processed subset ~740)
- Features: 14 features (e.g., age, sex, chest pain type, resting BP)
- Task: Binary classification (heart disease: presence vs. absence; use the “label” variable as the outcome)
- Data heterogeneity:
    - Feature distributions differ across sites (e.g., age, heart rate)
    - Label distributions are also imbalanced across sites
    - Site sample sizes (e.g., Switzerland has far fewer cases than other sites).

For more information: https://archive.ics.uci.edu/dataset/45/heart+disease

Introductory Paper to Learn More About the Data

International application of a new probability algorithm for the diagnosis of coronary artery disease. By R. Detrano, A. Jánosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, V. Froelicher. 1989. *American Journal of Cardiology.*

Site-Specific Data Dictionaries

Cleveland Site

Variable	Description	Type	Values / Units
age	Age of the patient in years	Integer	Example: 29, 45, 64
gender	Biological sex	Categorical	0 = Female, 1 = Male
cp	Chest pain type	Categorical	1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal pain, 4 = Asymptomatic
trestbps	Resting blood pressure (on hospital admission)	Float	mm Hg
chol	Serum cholesterol	Integer	mg/dL
fbs	Fasting blood sugar > 120 mg/dL	Binary	1 = True, 0 = False
restecg	Resting electrocardiographic results	Categorical	0 = Normal, 1 = ST-T abnormality, 2 = LVH
thalach	Maximum heart rate achieved	Integer	Beats per minute
exang	Exercise-induced angina	Binary	1 = Yes, 0 = No
oldpeak	ST depression induced by exercise relative to rest	Float	Measured in mm
slope	Slope of the peak exercise ST segment	Categorical	1 = Upsloping, 2 = Flat, 3 = Downsloping
ca	Number of major vessels (0–3) colored by fluoroscopy	Categorical	0, 1, 2, 3

thal	Thalassemia test result	Categorical	3 = Normal, 6 = Fixed defect, 7 = Reversible defect
label	Presence of heart disease	Binary	0 = No disease, 1 = Disease

Hungarian Site

Variable	Description	Type	Values / Units
age	Age of the patient in years	Integer	Example: 28, 30, 45
sex	Biological sex	Categorical	0 = Female, 1 = Male
cp	Chest pain type	Categorical	Typical angina, Atypical angina, Non-anginal pain, Asymptomatic
trestbps	Resting blood pressure (rounded to nearest 10)	Integer	mm Hg
chol	Serum cholesterol	Integer	mg/dL
fbs	Fasting blood sugar > 120 mg/dL	Binary	1 = True, 0 = False
restecg	Resting electrocardiographic results	Categorical	0 = Normal, 1 = ST-T abnormality, 2 = LVH
thalach	Maximum heart rate achieved	Integer	Beats per minute
exang	Exercise-induced angina	Binary	1 = Yes, 0 = No
oldpeak	ST depression induced by exercise	Float	Measured in mm
slope	Slope of the peak exercise ST segment	Categorical	1 = Upsloping, 2 = Flat, 3 = Downsloping
thal	Thalassemia test result	Categorical	3 = Normal, 6 = Fixed defect, 7 = Reversible defect

label	Presence of heart disease	Binary	0 = No disease, 1 = Disease

Switzerland Site

Variable	Description	Type	Values / Units
age	Age of the patient in years	Integer	Example: 34, 38, 56
sex	Biological sex	Categorical	0 = Female, 1 = Male
cp	Chest pain type	Categorical	1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal, 4 = Asymptomatic
trestbps	Resting blood pressure (on hospital admission)	Float	mm Hg
chol	Serum cholesterol	Integer	mg/dL
fbs	Fasting blood sugar > 120 mg/dL	Binary	1 = True, 0 = False
restecg	Resting electrocardiographic results	Categorical	0 = Normal, 1 = ST-T abnormality, 2 = LVH
max_hr	Maximum heart rate achieved	Integer	Beats per minute
exang	Exercise-induced angina	Binary	1 = Yes, 0 = No
oldpeak	ST depression induced by exercise	Float	Measured in mm
slope	Slope of the peak exercise ST segment	Categorical	1 = Upsloping, 2 = Flat, 3 = Downsloping
ca	Number of major vessels (0–3) colored by fluoroscopy	Categorical	0, 1, 2, 3
thal	Thalassemia test result	Categorical	3 = Normal, 6 = Fixed defect, 7 = Reversible defect

label	Presence of heart disease	Binary	0 = No disease, 1 = Disease

Long Beach V Site

Variable	Description	Type	Values / Units
age	Age of the patient in years	Integer	Example: 44, 60, 66
sex	Biological sex	Categorical	M = Male, F = Female
cp	Chest pain type	Categorical	1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal, 4 = Asymptomatic
resting_blood_pressure	Resting blood pressure (on hospital admission)	Float	mm Hg
chol	Serum cholesterol	Float	mg/dL
fbs	Fasting blood sugar > 120 mg/dL	Binary	1 = True, 0 = False
restecg	Resting electrocardiographic results	Categorical	0 = Normal, 1 = ST-T abnormality, 2 = LVH
thalach	Maximum heart rate achieved	Integer	Beats per minute
exang	Exercise-induced angina	Binary	1 = Yes, 0 = No
oldpeak	ST depression induced by exercise	Float	Measured in mm
slope	Slope of the peak exercise ST segment	Categorical	1 = Upsloping, 2 = Flat, 3 = Downsloping

label	Presence of heart disease	Binary	0 = No disease, 1 = Disease