AI4PH Summer Institute Data Challenge

Federated Learning for Privacy-Preserving Predictive Modeling

Organizers: Lisa Lix, David Buckeridge, Hassan Maleki Golandouz, Jean-Paul R. Soucy, Megha Roshan

July 15, 2025

Overview

- During the week, you will work with an existing dataset to complete a series of tasks (i.e., challenges) with your 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

- You need to have R downloaded on your computer for all activities: https://www.r-project.org/
- Python can also be used for selected activities: https://www.python.org/downloads/

Objective A

To explore the challenges and strategies for preparing multi-site datasets for federated learning, with a focus on data quality, heterogeneity, and feature harmonization.

A1: Assess data quality and structure across site-specific datasets, including missingness, inconsistencies, and outliers.
A2: Identify and address heterogeneity in sample sizes, feature availability, and coding practices across sites.
A3: Standardize and harmonize a common set of features to enable consistent model development across distributed data sources.

Objective B

To compare the performance and fairness of machine-learning models applied to pooled line-level data with models applied to distributed data.

B1: Implement site-specific models and aggregate results using federated learning approaches.
B2: Train a model using pooled data.
B3: Compare the accuracy, sensitivity, specificity, and fairness metrics of both approaches.

Objective C

To support participant skill development in applying machine-learning models to distributed data, and explore their assumptions, strengths, and limitations

C1: Practice model training in R/Python for binary classification.

C2: Reflect on modeling assumptions and limitations per approach.

Data Challenge Activities: Tuesday Afternoon

Tutorial session

- Instructors: Hassan Maleki Golandouz, University of Manitoba; Jean-Paul R. Soucy, McGill University
- This tutorial will use a synthetic dataset created for illustration purposes
- You will be provided with R code to demonstrate the techniques that will be introduced in this session

Team Session & Nerd Night

- Meet members of your pre-assigned team
- Establish tasks for each team member
- Conduct descriptive analyses
- Develop summary slides in PPT that describe the datasets

Data Challenge Presentations: Thursday Afternoon

- Present a summary of your team's challenge activities

- Develop a 12-13 minute presentation that provides an overview of the analyses your team completed this week
- In your presentation you will describe:
    - the specific tasks completed by each member of your group, and the expertise that these group members contributed to your challenge activities,
    - insights about the dataset that your team produced,
    - results of the predictive modeling and federated learning,
    - a summary of what you learned from working with these data
- Be prepared for questions from the audience

Comparative Table of Federated vs. Distributed Methods

Feature	Federated Learning	Federated Analysis	Distributed Learning	Distributed Analysis
Goal	Train a global ML model across decentralized data sources without sharing data	Conduct statistical/epidemiological analysis using local data without sharing data	Accelerate large-scale ML training by distributing data or model across compute nodes	Conduct statistical analysis across large datasets by distributing computation workload
Data location	Remains local at clients/sites	Remains local at sites	Data may be shared or moved across nodes	Data may be partitioned and distributed across servers
Privacy concern	Yes — designed to preserve privacy	Yes — used when data sharing is restricted	No — assumes data can be shared across infrastructure	No — focuses on scaling analysis, not privacy
Model type	ML models (e.g., neural networks, logistic regression)	Statistical models (e.g., regression, survival analysis)	ML models (e.g., deep learning, gradient boosting)	Statistical models (e.g., GLM, Cox model)

Federated Learning

The focus of this data challenge is on Federated Learning

Federated Learning is used to develop predictive models across multiple sites without transferring individual-level data, thereby preserving data privacy while enabling collaborative machine learning

About the Dataset

- 📍 UCI Heart Disease dataset
- 4 sources: **Cleveland**, **Hungary**, **Switzerland**, **Long Beach V**
- Total: 920 records (processed subset ~740)
- Features: 14 tabular features (e.g., age, sex, chest pain type, resting BP)
- Task: **Binary classification** (heart disease: presence vs. absence)

For more information:

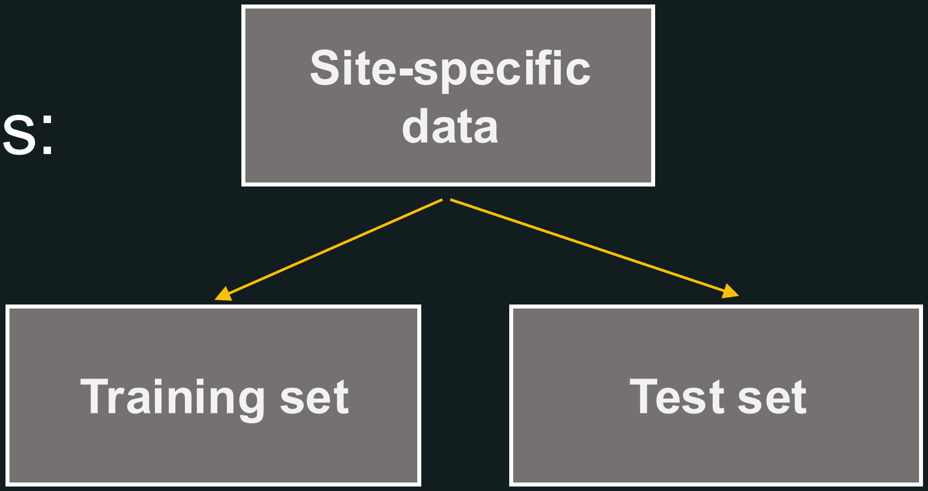
- visit: https://archive.ics.uci.edu/dataset/45/heart+disease
- Introductory Paper: *International application of a new probability algorithm for the diagnosis of coronary artery disease.* By R. Detrano, A. Jánosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, V. Froelicher. 1989, Published in the *American Journal of Cardiology.*

Train/Test Split per Site

- For each site, the data is split into two sets: a training set and a test set

- For example, for Cleveland:
  - cleveland_train_raw
  - cleveland_test_raw

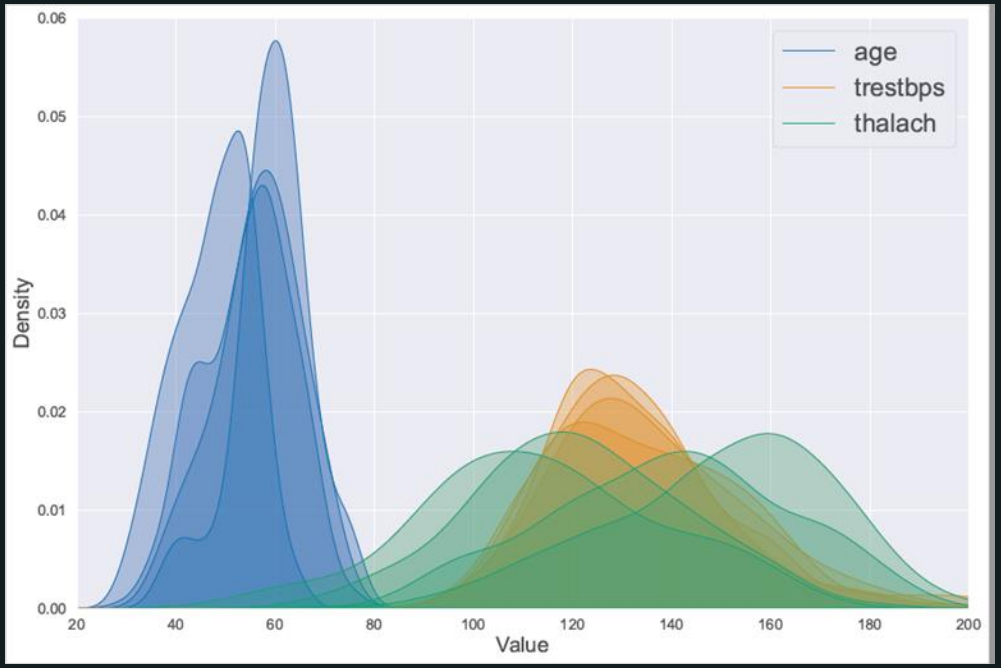
- Use the training sets to build your models, and the test sets to validate them



Data Heterogeneity

- Feature distributions vary across sites
    - For example, patient age and heart rate differ significantly amongst locations.

- Sample sizes also differ, which may affect model performance and generalizability.
    - For example, Switzerland has fewer cases compared to other sites.
-



Data Quality

- The datasets differ on:
  - Feature names and availability (some features may not be available for certain sites)
  - Missing values and data quality issues (e.g., inconsistent labels or formats)
  - Outliers (e.g., blood pressure of 10 or 400?)
  - Class imbalance (patients with and without disease)

Harmonizing data is essential for consistent modeling

Data Dictionaries

A proportion of Cleveland site's data dictionary is shown below:

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

Challenge Activities

- Explore site-specific data
- Document inconsistencies and differences across sites
- Harmonize data features across sites
- Train models: pooled vs. federated
- Evaluate model performance & fairness

Pooled data:

- Site-specific training and test datasets can be combined to form pooled training and test datasets.

Predictive Modeling Options

- Statistical Model:
    - Logistic Regression

- Machine-Learning/Non-Linear Models:
    - Random Forest
    - Neural Network
    - XGBoost

R code templates will be provided to help you get started!