Final Project Guide

Joshua Llano
Universitat Politècnica de Catalunya
joshua.llano@upc.edu

Dr. Alexandre Perera
Universitat Politècnica de Catalunya
alexandre.perera@upc.edu

Description

The main objective of this project is to develop a **robust and precise predictive model** capable of estimating the **probability** that a patient will develop a specific disease based on a set of **clinical variables and biomarkers**. The model will be built using **heterogeneous data** from various sources, such as electronic medical records, genomic databases, and population health registries, ensuring a comprehensive view of the patient's clinical profile.

To facilitate its integration into real-world workflows, the model will be encapsulated in an **API** that allows users to obtain predictions easily through requests. Subsequently, an **interactive graphical interface** will be developed using Shiny, enabling users to access the model's functionalities without requiring programming knowledge.

The ultimate aim of the project is to provide an **innovative tool** that contributes to improved **prediction, early diagnosis, and personalized clinical assessment**, identifying the most relevant risk factors associated with the studied disease.

Development

The project will be organized into three distinct phases, each designed to support the systematic design, implementation, and integration of the predictive model and its associated components.

Phase 1: Model Development and Data Preparation

In this phase, each group will select a complex disease to study and will carry out the following steps:

1. **Literature Review:** A comprehensive review of the scientific literature will be conducted to identify and understand the risk factors associated with the selected disease.
2. **Data Collection:** Relevant datasets will be identified from public repositories, such as PhysioNet, Gene Expression Omnibus, and MetaboLights, or extracted from studies reviewed during the literature review, providing a solid foundation for model development.
3. **Model Training:** The collected data will be used to train a predictive model using logistic regression techniques covered in class, with particular attention to interpreting the coefficients to understand each variable's contribution to disease risk.
4. **API Development:** The predictive model will be encapsulated within an API developed using *plumber*. The API will:
   - Be documented using *Swagger*.
   - Be hosted in a public *GitHub* repository to enable collaboration and transparency.

Phase 2: Collaborative Work and Integration

In the second phase, projects will be randomly exchanged between groups to encourage collaborative work and peer evaluation. The activities in this phase will include:

1. **Issue Management:** Each group will maintain their own *GitHub* repository, addressing and resolving issues reported by peers.
2. **User Interface Development:** Groups will develop a user interface for the assigned disease using a *Shiny* application, ensuring an intuitive and functional experience for users.
3. **Deployment:** The API and *Shiny* application will be deployed in two interconnected *Docker* containers, ensuring seamless integration and smooth operation between both components.

Phase Three: Documentation and Presentation

In the final phase, groups will concentrate on the completion, documentation, and presentation of their project. The activities in this phase will include:

1. **Final Documentation:** Each group will prepare comprehensive documentation of their project, including:
   - A detailed description of the methodology employed.
   - Instructions for deploying and using both the API and the user interface.
   - A summary of the challenges encountered and the solutions implemented.
   - A complete bibliography of all references used during the literature review and data collection, including references to the R software and packages utilized in the project.
2. **Project Submission:** The complete project must be submitted through Campus UB as a **single ZIP file** containing all documentation, supplementary materials, and the presentation. The source code for model training, the API, and the user interface must be provided through the group's *GitHub* repository and **must not** be included in the ZIP file.
3. **Presentation:** Each group will deliver a presentation of their project, which should:
   - Explain the selected disease and its clinical relevance.
   - Demonstrate the functionality of the API and the user interface application.
   - Highlight the importance of the project and its potential applications.
   - Last 25 minutes, followed by 5 minutes for questions. The presentation should be clearly structured and supported by slides or visual materials to enhance understanding.

Timeline and Deadlines

The project will follow the schedule outlined below:

1. **Phase 1: Model Development and Data Preparation**
   - Submission deadline: 17 December 2025
   - To submit: a Public GitHub repository with the API and the predicted model.
2. **Phase 2: Collaborative Work and Integration**
   - Work period: 18 December 2025 – 18 January 2026
3. **Phase 3: Final Documentation and Presentation**
   - ZIP Submission (documentation, presentation & materials): 18 January 2026
   - GitHub Submission (all source code): 18 January 2026
   - Presentations: 21 January 2026 (25 minutes + 5 minutes Q&A per group)