MEDI: Challenge Lab – Framingham Heart Study

Objective

In this assignment, medical students will utilize the **Framingham Heart Study** (https://www.framinghamheartstudy.org) dataset to develop and validate predictive models for estimating the 10-year risk of cardiovascular disease (CVD). This assignment will provide medical students with practical experience in applying engineering principles to medical data, enhancing their skills in data analysis, predictive modeling, and effective communication of complex findings.

Learning Objectives

1. Data Preprocessing and Feature Engineering:
    a. Apply techniques to handle missing data, outliers, and normalize features.
    b. Select and create features to enhance predictive modeling.
2. Predictive Model Development:
    a. Implement and compare different predictive models, optimizing their performance through hyperparameter tuning.
3. Model Evaluation and Validation:
    a. Assess model performance using key metrics and cross-validation to ensure reliability and generalizability.
4. Interpretation and Visualization:
    a. Analyze feature importance and visualize model outputs to derive and communicate key insights.
5. Reporting and Presentation:
    a. Compile findings into a structured report and effectively present the project's results and implications.

Specific Learning Objectives

1. Identify and address missing data, outliers, and normalization requirements within a clinical dataset.
2. Select relevant features based on statistical analysis and domain knowledge.
3. Implement various predictive modeling techniques and use appropriate metrics (accuracy, precision, ROC-AUC) to evaluate model performance.
4. Document the methodology, results, and implications of the analysis in a structured report.
5. Effectively communicate the project’s findings and respond to questions related to the analysis and its implications for cardiovascular risk prediction.

Introduction

Cardiovascular disease (CVD) remains a leading cause of morbidity and mortality worldwide. Accurate prediction of CVD risk is crucial for effective preventive measures and personalized treatment. The Framingham Heart Study dataset is a well-known longitudinal dataset that has been widely used in cardiovascular research. It originates from the Framingham Heart Study, which began in 1948 in the town of Framingham, Massachusetts, and was designed to identify common factors or characteristics that contribute to cardiovascular disease (CVD). This study is one of the most influential and longest-running epidemiological studies in the world.

The Framingham Heart Study dataset provides an invaluable opportunity to apply engineering and data science techniques to medical data, aiming to improve risk prediction models and ultimately contribute to better health outcomes.

Framingham Heart Study Dataset Overview

The Framingham Heart Study dataset includes data collected from multiple generations of participants over several decades. It contains a wealth of information about participants' health, lifestyle, and medical history, making it a valuable resource for studying cardiovascular risk factors and outcomes.

The dataset can be broken down into several key components:

1. **Participant Demographics**:
    a. Age: The age of participants at the time of data collection.
    b. Sex: Gender of the participants (male/female).
    c. Education: The highest level of education attained by the participant.
    d. Ethnicity: Racial and ethnic background, though early cohorts were primarily of European descent.
2. **Medical History**:
    a. Personal Medical History: Information about pre-existing conditions, such as hypertension, diabetes, and hyperlipidemia.
    b. Family History: Data regarding the history of cardiovascular diseases in family members.
    c. Smoking Status: Information about whether the participant smokes, including current and past smoking behavior.
    d. Alcohol Consumption: Details on the frequency and amount of alcohol consumption.
    e. Physical Activity: Data on the level of physical activity, both occupational and recreational.
3. **Physical Examinations and Measurements**:
    a. Blood Pressure: Systolic and diastolic blood pressure measurements taken during medical exams.
    b. Cholesterol Levels: Total cholesterol, HDL, LDL, and triglyceride levels.
    c. Blood Glucose Levels: Fasting glucose levels, important for assessing diabetes risk.
    d. Body Mass Index (BMI): Calculated from height and weight measurements to assess obesity.
    e. Electrocardiogram (ECG) Data: Includes information on heart rate, rhythm, and evidence of previous myocardial infarction or other cardiac abnormalities.
4. **Laboratory Tests**:
    a. Blood Chemistry: Levels of various blood biomarkers, including hemoglobin, hematocrit, creatinine, and others.
    b. Lipid Panels: Detailed breakdown of lipid profiles, which are crucial for assessing cardiovascular risk.
5. **Outcome Variables**:
    a. Cardiovascular Events: Data on occurrences of myocardial infarction (heart attack), stroke, heart failure, and other cardiovascular events.
    b. Mortality Data: Information on deaths, including cause of death and survival time from the baseline examination.
    c. Time-to-Event Data: For survival analysis, time-to-event data (e.g., time to first cardiovascular event or death) is often included.

Activity Outline

Students will be divided into several groups and provided with a dataset (comma-separated-values, CSV file: *frmgham2.csv*) containing limited (for teaching purposes) data obtained during the Framingham Heart Study and a PDF file (*Framingham Longitudinal Data Documentation.pdf*) containing detailed information on the dataset.

The **goal of this activity** is to determine the epidemiology and risk factors associated with heart disease, including hypertension, cholesterol levels, smoking, obesity, and diabetes, based on statistical analysis and domain knowledge. Students can employ any available tools (Python, MATLAB, R, etc.) and statistical tests to reach conclusions, which will be presented in the form of a scientific report and a video presentation of the findings.

Please note that the recommended workflow, including task breakdown, is presented below only as an example. Based on the group discussion, the students can modify this workflow or use another one – you can be creative!

Task Breakdown (an example)

1. **Data Preprocessing**
   Objective: Prepare the dataset for analysis by addressing data quality issues and ensuring that it is ready for modeling.
   a. Steps:
      i. Handling Missing Data:
         - Begin by exploring the dataset to identify any missing or incomplete values. For instance, if some records are missing blood pressure measurements, these need to be addressed.
         - Choose an appropriate method for handling missing data. For example, if only a small percentage of records are incomplete, you might decide to impute missing values using the mean or median. For more extensive missing data, consider advanced imputation techniques such as k-Nearest Neighbors (k-NN) imputation.
      ii. Outlier Detection:
         - Detect outliers in features such as cholesterol levels or blood pressure. Outliers might indicate measurement errors or rare conditions that could affect model performance.
         - Decide whether to transform or exclude outliers based on their impact on the analysis.
      iii. Data Normalization:
         - Normalize features to bring them onto a comparable scale. For instance, cholesterol levels and blood pressure measurements should be scaled so that they contribute proportionately to the model's performance.
      iv. Data Splitting:
         - Divide the dataset into training (e.g., 70%), validation (e.g., 15%), and test sets (e.g., 15%). This ensures that the model is trained on one portion of the data and validated and tested on separate, unseen data to assess its generalizability.

2. **Feature Engineering**
   Objective: Enhance the dataset by creating and selecting features that improve the predictive power of the model.
   a. Steps:
      i. Feature Selection:

- Use statistical techniques and domain knowledge to select the most relevant features. For example, features like age, blood pressure, and cholesterol levels are known to be strong predictors of CVD risk.

ii. Feature Creation:
- Create new features if they might provide additional predictive power. For example, you could calculate the ratio of HDL to total cholesterol as an additional feature.

iii. Feature Transformation:
- Apply transformations to improve model performance. For instance, if cholesterol levels have a skewed distribution, applying a logarithmic transformation might normalize the distribution.

3. Model Development

Objective: Build and evaluate predictive models to estimate the 10-year risk of cardiovascular disease.

a. Steps:

i. Model Selection:

Implement and compare several predictive models:
- Logistic Regression: Serves as a baseline model for binary classification.
- Decision Trees: Useful for understanding feature importance and capturing complex relationships.
- Random Forests: An ensemble method that aggregates multiple decision trees to improve predictive accuracy and robustness.
- Gradient Boosting Machines (GBM): A powerful technique that improves prediction accuracy by combining weak learners into a strong model.

ii. Hyperparameter Tuning:
- Optimize model performance by tuning hyperparameters. Use techniques like grid search or random search to find the best parameters for each model.

4. Model Evaluation and Validation

Objective: Assess the performance of the predictive models and ensure their reliability.

a. Steps:

i. Evaluation Metrics:
- Evaluate model performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. These metrics help in understanding how well the model predicts CVD risk and the balance between sensitivity and specificity.

ii. Cross-Validation:
- Implement k-fold cross-validation to verify that the model performs consistently across different subsets of the data. This helps in mitigating overfitting and ensuring that the model generalizes well.

iii. Model Comparison:
- Compare the performance of different models to identify the best one. For instance, a Random Forest might perform better than a Logistic Regression model due to its ability to capture complex interactions between features.

5. Interpretation and Visualization

Objective: Analyze and present the results in a way that highlights key insights and model performance.

a. Steps:
    i. Feature Importance:
        - Analyze and visualize which features are most influential in predicting cardiovascular risk. For example, Random Forests and Gradient Boosting Machines provide feature importance scores that help in understanding which variables contribute most to the model's predictions.
    ii. Visualization:
        - Create visualizations to communicate your findings:
            a. ROC Curves: Show the trade-off between sensitivity and specificity across different models.
            b. Feature Importance Plots: Highlight which features are most predictive of CVD risk.
            c. Scatter Plots and Heatmaps: Illustrate relationships between features and the predicted risk.

6. Report Preparation

Objective: Compile a detailed report that documents the methodology, results, and implications of the project.

a. Steps:
    i. Methodology:
        - Describe the data preprocessing steps, feature engineering processes, and modeling techniques used. Explain why specific methods were chosen and how they contribute to the overall analysis.
    ii. Results:
        - Present the performance metrics of the models, highlighting which model achieved the best results. Include visualizations to support your findings.
    iii. Discussion:
        - Interpret the results in the context of cardiovascular risk prediction. Discuss how the findings could impact clinical practice and suggest potential improvements or future research directions.
    iv. Visualizations:
        - Include charts, plots, and tables that enhance the understanding of your findings.

7. Presentation

Objective: Summarize and present the project findings clearly and effectively.

a. Steps:
    i. Slide Deck. Prepare a presentation with clear and concise slides. Key components should include:
        - Introduction: Objectives and significance of the project.
        - Data: Description and preprocessing of the dataset.
        - Models: Overview of the models developed and their performance.
        - Results: Key findings and visualizations.
        - Conclusions: Summary of insights and implications for cardiovascular risk prediction.
    ii. Delivery:

- Practice delivering the presentation to ensure clarity and confidence. Be prepared to answer questions and discuss the implications of your work.

Deliverables

1. **Code and Documentation**
   a. Submit clean, well-documented code for all stages of the project, including data preprocessing, feature engineering, model development, and evaluation. Documentation should explain the functionality of the code and the rationale behind the chosen methods.
2. **Report**
   a. Provide a detailed report that includes the methodology, results, visualizations, and interpretations. The report should be well-organized and include references to relevant literature where applicable (see details of presentation - Point 7 above, and the rubric below).
3. **Presentation**
   a. Deliver a video presentation summarizing the project, with a focus on key findings and implications. Ensure that the presentation is clear, engaging, and accessible to both technical and non-technical audiences.

Final Presentation and Rubric

The final graded assignment is the group presentation of findings from the MEDI Challenge lab activity. The video presentation format should be like a clinical case report. **The duration of the video presentation should not exceed 10 minutes**. The presentation should include:

- Introduction: Objectives and significance of the project.
- Data: Description and preprocessing of the dataset.
- Models: Overview of the models developed and their performance.
- Results: Key findings and visualizations.
- Conclusions: Summary of insights and implications for cardiovascular risk prediction.

Final group assignments (videos and reports) are due on Monday, **September 29, 2025, at 10 AM**. Videos will be watched together in class and graded by your instructor.

The following rubric will be used to grade the MEDI Challenge Lab video presentations:

Category	Great	Good	Min Grade
Presentation Organization, Professionalism, and Format	(7) Extremely clear, concise, and professional. Engaged and easily understood. Innovative and creative use of video format. Adhered to timing and guidelines.	(4) Mostly clear, concise, and professional. Somewhat engaging and easily understood. Acceptable use of slides to explain research and results. Mostly adhered to timing and guidelines.	(1) Unclear and unprofessional. Not easily understood and not engaging. Poor use of slides to explain the research. Did not adhere to timing and guidelines.
Technical Appeal of the Video	(7) Outstanding technical appeal. Visuals are clear and appropriate for the topic. The text and	(4) Good technical appeal. Visuals, text, and narration are mostly	(1) Inadequate technical appeal. Visuals, text, and narration are unclear

	narration are clear and appropriate for the topic.	clear and appropriate for the topic.	and inappropriate for the topic.
Content, Methodology, and Technical Description	(7) Detailed descriptions of all aspects of the analysis are given, including methodology and analysis process. Understandable to an educated non-expert.	(4) Some details are missing or quickly dismissed. Mostly addresses the research objective. Understandable by an expert but not an educated non-expert	(1) Many details are missing. An inadequate description of the analysis and methodology is given. Did not align with the research objective. Confusing to all.
Explanation of Results and Conclusions	(7) Results are well-organized and clearly and precisely explained. Conclusions are well articulated and based on results.	(4) Some results are omitted or with inadequate commentary and summary of the investigations.	(1) Many results are missing, and there is a lack of visuals such as images and graphs. Conclusions do not relate to results.

The following assessment criteria will be used to evaluate the scientific reports:

1. **Data Handling**: Effectiveness in addressing missing values, outliers, and normalization. Quality of data preparation and cleaning.
2. **Model Performance**: Accuracy and reliability of predictive models. Evaluation using appropriate metrics and cross-validation.
3. **Insight and Interpretation**: Depth of analysis and clarity of interpretations. Ability to draw meaningful conclusions from the results.
4. **Report Quality**: Organization, completeness, and clarity of the written report. Quality of visualizations and explanations.

References

1. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet. 2014 Mar 15;383(9921):999-1008. doi: 10.1016/S0140-6736(13)61752-3. Epub 2013 Sep 29. PMID: 24084292; PMCID: PMC4159698.
2. Andersson C, Nayor M, Tsao CW, Levy D, Vasan RS. Framingham Heart Study: JACC Focus Seminar, 1/8. J Am Coll Cardiol. 2021 Jun 1;77(21):2680-2692. doi: 10.1016/j.jacc.2021.01.059. PMID: 34045026.
3. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. Diabetes, blood lipids, and the role of obesity in coronary heart disease risk for women. The Framingham study. Ann Intern Med. 1977 Oct;87(4):393-7. doi: 10.7326/0003-4819-87-4-393. PMID: 199096.
4. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. Predicting coronary heart disease in middle-aged and older persons. The Framington study. JAMA. 1977 Aug 8;238(6):497-9. PMID: 577575.
5. Draper, Norman Richard, and Harry Smith. Applied Regression Analysis. Third edition. New York: Wiley, 1998. Print. (available online at library.illinois.edu)