# Initial Proposal — Predicting Heart-Disease Risk with Machine-Learning

*Cognitive Computing (STA 141C) • Spring 2025 • Team Cardio-ML (5 students)*

## 1. Project Overview & Motivation

Cardiovascular disease (CVD) is the world’s leading cause of death, yet many events are preventable with timely risk stratification. We propose to build an interpretable machine-learning pipeline that **predicts the presence of heart disease** using a composite Kaggle dataset that merges five classic clinical cohorts (918 observations, 11 features). Our goals are to:

1. **Answer the core question:** *Given easily obtained clinical measurements, can we accurately flag patients who already have heart disease?*
2. **Explore two sub-questions (added value / impact):**
   * *Which features—age, cholesterol, chest-pain type, etc.—carry the most predictive weight?*
   * *Do model errors differ by sex or age, raising fairness concerns?*

Early identification will help clinicians route high-risk patients toward confirmatory testing and lifestyle or pharmacological interventions, potentially reducing premature CVD mortality.

## 2. Dataset & Predicted Variable

| Aspect | Details |
| --- | --- |
| **Source** | Kaggle “Heart Disease” compilation (Cleveland, Hungarian, Switzerland, Long Beach VA, Statlog) |
| **Size** | 918 unique patient records |
| **Features** | 7 numeric (Age, RestingBP, Cholesterol, MaxHR, FastingBS, Oldpeak, …) and 4 categorical (Sex, ChestPainType, RestingECG, ST\_Slope, ExerciseAngina) |
| **Target** | **HeartDisease** (1 = disease present, 0 = normal) |

The dataset’s moderate size lets us iterate rapidly while still testing models prone to over-fitting. Because the outcome is binary, evaluation will focus on ROC-AUC, PR-AUC, sensitivity, and specificity.

## 3. Proposed Methods & Rationale

| Stage | Method | Why Appropriate |
| --- | --- | --- |
| **Data prep** | Imputation (median / mode); one-hot encoding; stratified train/validation/test split; SMOTE if imbalance > 1.5 : 1 | Ensures clean inputs and balanced classes |
| **Baseline** | Logistic Regression | Clinically interpretable odds ratios; benchmark for complex models |
| **Non-linear models** | Random Forest, XGBoost | Capture feature interactions; proven tabular-data accuracy |
| **Hyper-parameter tuning** | Bayesian Opt / GridSearch with cross-validation | Robust model selection under limited data |
| **Interpretability** | SHAP values, Partial-Dependence Plots | Identify top risk factors; support clinical acceptance |
| **Fairness audit** | Grouped ROC by Sex & Age-quartile | Detect disparate error rates |

We expect tree-based ensembles to outperform linear baselines, while SHAP values will translate findings into clinician-friendly language.

## 4. Work Plan & Team Responsibilities

| Member | Main Tasks | Week 1-2 | Week 3-4 | Week 5-6 | Week 7-8 |
| --- | --- | --- | --- | --- | --- |
| **A. Research Lead (Rayan)** | Scope refinement, literature survey, progress checks | ✓ | ✓ | ✓ | Final synthesis |
| **B. Data Engineer** | Cleaning, feature engineering, pipeline automation | ✓ | ✓ |  |  |
| **C. ML Engineer** | Model training & tuning scripts |  | ✓ | ✓ | ✓ |
| **D. Analyst** | SHAP / fairness analysis, result visualization |  |  | ✓ | ✓ |
| **E. Writer & Presenter** | Slide deck, report drafting, oral demo rehearsal |  |  | ✓ | ✓ |

### Milestones

* **Week 2:** cleaned dataset & baseline metric
* **Week 4:** tuned models & cross-validated scores ≥ 0.85 ROC-AUC
* **Week 6:** interpretability & fairness results
* **Week 8:** final paper (< 10 pp) & presentation

## 5. Expected Deliverables

1. Reproducible Jupyter notebook with data pipeline and trained models
2. 10-page final report detailing methodology, results, and ethical considerations
3. 10-minute conference-style presentation + 5-minute Q&A
4. Public GitHub repo under MIT license

## 6. Significance

A portable, transparent heart-disease classifier contributes to the broader push for preventative cardiology, especially in low-resource settings where advanced imaging is scarce. Our fairness audit aims to ensure that the model benefits all demographic groups equally, aligning with modern standards for trustworthy AI in healthcare.