BAHRIA UNIVERSITY KARACHI CAMPUS

Department of Software Engineering

COURSE: MACHINE LEARNING

PROJECT PROPOSAL

CLASS: BSE – 5 B (SPRING - 2025)

Project Title:

Heart Disease Prediction System.

Group Members

S. No.	Name	Enrollment #
01	**Faiz ur Rehman**	02-131232-126
02	**Tahir Hameed**	02-131232-106

Submitted to:

Course Instructor: Engr. Aamana
Lab Instructor: Engr. Hamza

Date: 13/Nov/2025

1. Introduction & Background

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, accounting for approximately 17.9 million deaths annually according to the World Health Organization (WHO). Early detection and prevention play a vital role in reducing mortality and improving patient outcomes. However, traditional diagnosis often depends on manual interpretation of clinical data and physician expertise, which can be time-consuming and prone to error.

With the advancement of **Machine Learning (ML)** and **Data Science**, healthcare analytics has become increasingly data-driven. ML models trained on patient datasets can identify complex patterns in medical parameters such as age, blood pressure, cholesterol, and ECG readings—relationships that may not be easily recognized through conventional methods.

This project aims to develop a **Heart Disease Prediction System** using supervised ML algorithms to predict the likelihood of heart disease. Additionally, an **interactive web interface** (built with Streamlit) will enable users to enter patient information and receive real-time predictions with detailed explanations

---

2. Problem Statement

Heart disease diagnosis depends on multiple interrelated factors such as blood pressure, cholesterol level, and exercise tolerance. Manual assessment of these parameters may lead to inconsistent or delayed conclusions. Traditional diagnostic approaches also struggle to capture nonlinear relationships among variables.

Therefore, there is a need for an automated, data-driven system that can:

Learn from historical patient data.

Accurately classify patients as likely to have or not have heart disease.

Provide interpretable, explainable results for healthcare professionals.

Offer a simple, user-friendly interface for practical use.

Main Problem:

How can supervised ML models trained on clinical patient data accurately and reliably predict the likelihood of heart disease while maintaining interpretability and usability for end-users?

---

3. Proposed Solution

This project proposes an offline, dataset-based **Heart Disease Prediction System** using supervised learning. The system will analyze patient attributes such as age, sex, blood pressure, cholesterol, and ECG results to estimate heart-disease probability.

Algorithms including **Logistic Regression, SVM, Decision Tree, and Random Forest** will be implemented and compared. The best-performing model will be chosen based on accuracy and generalization performance. A **Streamlit GUI** will allow interactive prediction, and **SHAP plots** will be used for model explainability.

3.1 Features of the Project

Import and preprocess a labeled medical dataset.

Encode categorical and normalize continuous features.

Train multiple ML models and compare their metrics.

Evaluate using accuracy, precision, recall, and F1-score.

Apply cross-validation for reliability.

Visualize feature importance via correlation or SHAP plots.

Build an interactive Streamlit interface for predictions.

(Optional) Deploy online using Streamlit Cloud or Hugging Face Spaces.

3.2 Methodology

Data Understanding: Explore dataset attributes and relationships.

Data Preprocessing: Handle missing values, encode categorical data, and scale numeric features.

Feature Engineering: Derive new variables (e.g., Cholesterol/BP ratio, Risk Index).

Feature Selection: Use correlation or model-based importance scores.

Model Training: Implement Logistic Regression, Decision Tree, Random Forest, and SVM.

Model Evaluation: Assess accuracy, precision, recall, F1-score, and ROC-AUC.

Optimization: Apply GridSearchCV/RandomizedSearchCV for tuning.

Explainability: Use SHAP to visualize feature impact.

Interface Development: Build Streamlit GUI.

Deployment (Optional): Host online for demonstration.

3.3 Technologies to Be Used

Programming Language: Python 3

Libraries & Tools: pandas, NumPy, scikit-learn, matplotlib, seaborn, plotly, SHAP, Streamlit

Platform: Windows / macOS / Linux

IDE: Jupyter Notebook, VS Code, or PyCharm or Colab

---

4. Project Scope

Included

Offline model training on a labeled dataset.

Data cleaning, encoding, and scaling.

Implementation of Logistic Regression, SVM, Decision Tree, and Random Forest.

Cross-validation and hyperparameter tuning.

Visualization of results and feature importance.

Streamlit-based user interface for predictions.

Excluded

Real-time integration with hospital systems.

Deep-learning or neural-network architectures.

Live ECG signal or IoT data processing.

Mobile-app implementation.

Project Timeline

Week	Task
Week 1	Dataset exploration, cleaning, preprocessing
Week 2	Feature engineering and model training
Week 3	Model evaluation, tuning, and visualization
Week 4	Streamlit UI development, documentation, final testing

---

5. Project Abstract

This project applies machine learning to predict heart disease likelihood from clinical and demographic data. Supervised models—Logistic Regression, SVM, Decision Tree, and Random Forest—will be trained on a labeled dataset. Preprocessing steps such as encoding, normalization, and feature engineering will improve performance. Models will be compared using accuracy, precision, recall, and F1-score.

The best-performing model will be integrated into a **Streamlit web application** for real-time prediction and interpretation. The goal is to demonstrate how ML can assist in early detection of cardiovascular risk and support data-driven medical decisions.

---

6. Module Distribution

1. Data Collection & Preprocessing – Load dataset, handle missing data, encode and scale features.

2. Feature Engineering – Create derived variables; identify significant features.

3. Model Training & Validation – Implement algorithms, apply 80/20 split and cross-validation, tune hyperparameters.

4. Evaluation & Visualization – Measure metrics, plot confusion matrices, interpret with SHAP.

5. User Interface – Build Streamlit app for input and predictions; optional online deployment.

---

7. References

Bishop, C. M. *Pattern Recognition and Machine Learning.* Springer, 2006.

Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed., Springer, 2009.

Detrano, R. et al. *Cleveland Heart Disease Dataset.* UCI Machine Learning Repository, 1988.

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions (SHAP).* NIPS.

Goldstein, B. A., Navar, A. M., & Carter, R. E. (2017). *Moving Beyond Regression Techniques in Cardiovascular Risk Prediction: Applying Machine Learning to the Framingham Heart Study. European Heart Journal.*