# Project 3: Heart Disease Prediction

Problem Statement:

- How can we optimize the prediction accuracy of heart disease presence using feature selection techniques and machine learning algorithms?

Context:

- This capstone project involves exploring the Cleveland Heart Disease Database, which contains 76 attributes, with only a subset of 14 attributes commonly used in experiments. The goal is to predict the presence of heart disease based on these attributes. However, the dataset has undergone anonymization, with patient names and social security numbers replaced by dummy values. The task for the capstone would be to explore feature selection techniques and machine learning algorithms to optimize prediction accuracy while reducing the dimensionality of the dataset. Additionally, understanding the ethical implications of data anonymization and ensuring compliance with privacy regulations would be crucial aspects of the project.

Criteria for success:

- Achieve a high prediction accuracy for heart disease presence.
- Identify the most relevant features contributing to prediction.
- Implement efficient feature selection techniques to reduce dataset dimensionality.

Scope of solution:

- Exploration of feature selection techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and SelectKBest.
- Experimentation with various machine learning algorithms including logistic regression, random forests, support vector machines (SVM), and neural networks.
- Utilization of cross-validation and grid search for hyperparameter tuning.
- Evaluation of model performance using metrics such as accuracy, precision, recall, and F1-score.
- Presentation of findings through reports, visualizations, and presentations.

Constraints:

- Limited access to the Cleveland Heart Disease Database and its subset of 14 attributes.
- Constraints on computational resources for training and evaluating machine learning models.
- Adherence to ethical guidelines and privacy regulations regarding healthcare data.

Stakeholders:

- Healthcare professionals
    - That are interested in improving heart disease diagnosis and prediction.
- Data scientists and machine learning practitioners:
    - Seeking insights into feature selection and model optimization techniques.

Data Sources:

- Kaggle and UC Irvine Machine Learning Repository.