# Project Deliverable 4: Final Insights, Recommendations, and Presentation for Heart Disease Analysis Using Machine Learning

## Sahaj Shrestha

## MSCS-634 Advanced Big Data and Data Mining

**Introduction**

In this project, we analyzed the UCI Heart Disease dataset using a variety of machine learning techniques to predict the presence or absence of heart disease. The dataset comprises 1,035 records with 14 attributes, including age, sex, chest pain type, cholesterol levels, and other medically relevant indicators. Our goal was to preprocess the data, apply predictive models, evaluate their performance, and extract actionable insights.

**Data Preparation Steps**

To ensure high-quality inputs for our models, we carried out several data preparation and preprocessing steps:

1. Handling Missing Values  
   We identified missing values in the chol, thalach, and oldpeak columns. These missing values were addressed using median imputation, ensuring minimal distortion of the data distribution.
2. Removing Duplicate Rows  
   We detected and removed duplicate records using the drop\_duplicates() method. This step helped prevent data leakage and ensured unbiased model evaluation.
3. Outlier Detection and Handling  
   We found outliers in the chol, trestbps, thalach, and oldpeak attributes. These outliers were handled using the Interquartile Range (IQR) method, which allowed us to cap extreme values while retaining data integrity.
4. Encoding Categorical Variables  
   To handle categorical features effectively, we applied one-hot encoding. This transformation allowed our machine learning algorithms to process non-numeric data such as chest pain type (cp) and thalassemia (thal) accurately.

**Modeling Details**

After preparing the dataset, we applied multiple machine learning models for both regression and classification tasks:

* Linear Regression  
  We used linear regression to predict cholesterol levels, establishing a baseline performance for numeric prediction.
* Ridge Regression  
  We implemented ridge regression with a regularization parameter of alpha = 1.0 to improve generalization and reduce overfitting. Ridge regression outperformed linear regression in terms of error metrics, validating the use of regularization.
* Decision Tree Classifier  
  To predict the presence or absence of heart disease, we built a decision tree classifier. This model provided transparent, interpretable rules for classification based on attributes like cp, thalach, and oldpeak.
* K-Nearest Neighbors (KNN)  
  We also used KNN classification with k = 5 to evaluate how well instance-based learning performs on this dataset. The model provided competitive results and offered a complementary approach to tree-based models.

**Evaluation**

We assessed the performance of each model using appropriate metrics:

* For Regression Models (Linear & Ridge):
  + Mean Squared Error (MSE)
  + Root Mean Squared Error (RMSE)
  + R² Score
* For Classification Models (Decision Tree & KNN):
  + Accuracy

These metrics helped us compare the models and identify which approaches were most suitable for different predictive tasks within the dataset.

**Insights**

Through our analysis, we uncovered several meaningful insights:

* The dataset, after cleaning and transformation, was ready for robust modeling and yielded consistent performance across multiple algorithms.
* Key predictive features for heart disease classification included chest pain type (cp), maximum heart rate achieved (thalach), ST depression (oldpeak), and exercise-induced angina (exang).
* Ridge regression showed lower error than linear regression, emphasizing the benefit of regularization in regression tasks.
* Both the decision tree and KNN classifiers performed well in classifying heart disease presence, suggesting that the dataset is well-suited to supervised learning techniques.

**Ethical Considerations**

While working with medical datasets, we acknowledged the ethical responsibilities involved. Potential biases in the dataset such as demographic imbalances could impact model fairness. Moreover, protecting data privacy is essential, especially in healthcare applications. Misclassification of heart disease could lead to severe consequences, making model interpretability and validation critical in real-world deployment.

**Recommendations**

For future work, we recommend:

* Exploring additional models like Random Forest, Logistic Regression, and Support Vector Machines.
* Applying feature selection techniques to reduce dimensionality and improve performance.
* Evaluating models with cross-validation for more reliable performance estimates.
* Testing different target variables, such as cholesterol levels or maximum heart rate, to expand insights.

**References**

* Kaggle. (n.d.). Heart Disease UCI. Kaggle. https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uciScikit-learn Documentation
* Scikit-learn developers. (n.d.). Scikit-learn: Machine learning in Python. <https://scikit-learn.org/>