# Heart Disease Diagnosis

1. Project Summary:
   - Develop a machine learning model that can accurately predict the presence of heart disease based on patient information such as age, sex, blood pressure, cholesterol levels, and other relevant risk factors.

2. Project Objectives:
   - The project aims to develop a predictive model for early detection of heart disease using machine learning algorithms. By analyzing various medical and lifestyle factors, the model will provide accurate risk assessments for individuals, allowing for timely interventions and improved patient outcomes.

3. Data Requirements and Sources:
   - The dataset used for this project is the Cleveland Heart Disease dataset taken from the UCI repository.
   - It covers 13 features which are listed below and the diagnosis.
        - Age
        - Sex
            - 1 = male
            - 0 = female
        - Chest-pain type
            - 1 = typical angina

        - 2 = atypical angina
        - 3 = non — anginal pain
        - 4 = asymptotic
    - Resting Blood Pressure (in mmHg)
    - Serum Cholesterol (in mg/dl)
    - Fasting Blood Sugar (in mg/dl)
    - Resting ECG
        - 0 = normal
        - 1 = having ST-T wave abnormality
        - 2 = left ventricular hypertrophy
    - Max heart rate achieved
    - Exercise induced angina
        - 1 = yes
        - 0 = no
    - ST depression induced by exercise relative to rest
    - Peak exercise ST segment
        - 1 = upsloping
        - 2 = flat
        - 3 = downsloping
    - Number of major vessels (0–3) colored by fluoroscopy
    - Thalassemia (Thal)
        - 3 = normal
        - 6 = fixed defect
        - 7 = reversible defect
    - Diagnosis (Whether the individual is suffering from heart disease or not)
        - 0 = absence
        - 1, 2, 3, 4 = present.

4. Data Analysis and Modeling:

Step 1: Read in parquet dataset about heart disease.
Refers to the process of loading the dataset into Spark to make it accessible for further analysis and modeling. This step is essential because the dataset contains the necessary information for training a supervised machine learning model for heart disease prediction.

Step 2: EDA (Exploratory Data Analysis)
EDA is an essential step to understand the structure, patterns, and characteristics of our dataset. It helps you gain insights into the variables, identify relationships, detect outliers, and prepare the data for modeling.

- Correlation Analysis: Calculate and visualize the correlation between numerical features using techniques such as correlation matrices or heatmaps. This helps identify relationships and dependencies among variables.

- Data Visualization: Utilize various data visualization libraries, such as Matplotlib or Seaborn, to create clear and informative visualizations of the dataset.

Step 3: Scaling the data using StandardScaler
Improve the performance of machine learning models, particularly those that are sensitive to the scale of the input features. The StandardScaler is a common scaling technique that standardizes the features by subtracting the mean and dividing by the standard deviation. Using StandardScaler to standardize or normalize the features of a dataset.

Step 4: Split the data into features and target (X and y) and then into testing and training sets.
Refers to the process of separating the dataset into input features (X) and the corresponding target variable (y). Afterward, the data is further divided into training and testing sets. This step is crucial for supervised machine learning tasks.

Step 5: Fit using SVM, Linear Regression, Logistic Regression, Random Forest, and KNN.
Refers to training a supervised machine learning model using the training data. SVM, Linear Regression, Logistic Regression, Random Forest, and KNN are commonly used algorithms for binary classification problems, making them suitable for heart disease prediction tasks. Using confusion metrics and classification reports, will find the best algorithm for our scenario.

Step 6: Create the predicted values for the testing and the training data.
After fitting the model, we can create predicted values for both the testing and training data. This allows you to evaluate the model's performance and compare its predictions with the actual target values.

Step 7: Print a confusion matrix for the training data.
You can print a confusion matrix to evaluate the performance of the model on the training data. The confusion matrix provides a detailed breakdown of the model's predictions and their agreement with the actual target values.

Step 8: Print a confusion matrix for the testing data.
The confusion matrix provides insights into the model's predictions and their agreement with the actual target values for the testing set.

Step 9: Print the training classification report.
You can print a classification report to get a comprehensive evaluation of the model performance on the training data. The classification report provides metrics such as precision, recall, F1-score, and support for each class, allowing you to assess the model's performance in more detail.

Step 10: Print the testing classification report.
You can print a classification report to evaluate the performance of the model on the testing data. The classification report provides metrics such as precision, recall, F1-score, and support for each class, allowing you to assess the model's performance in more detail.

Step 11: Create web application that makes prediction with the data model that we trained

- Use HTML, CSS, Javascript and D3 to create a web-page that access web services created in Flask.
- Flask application will use the model that was trained and saved using Pickle.

Step 12: Presentation

- Summarize the key points covered in the project.

5. Project Timeline:

- Project Initiation and Planning: May 30, 2023:
  Define project goals, objectives and deliverables
  Identify audience / stakeholders
  Brainstorming project plan
  Allocate resources and establish communication channel
- Data collection and project plan finalization: June 1,2023:
  Identify relevant heart disease datasets and sources
  Obtain necessary permission and access to the data
  Complete the project plan in README file and submit to Github
- Data preparation and EDA: June 5, 2023:
  Clean the data and perform data processing tasks such as feature engineering or normalization
  Conduct basic data summary and statistical analysis
  Create visualizations to explore patterns, distribution and relationships
  Assess data quality and make necessary corrections
  Generate hypotheses and refine the questions if needed

-Algorithm selection and model development: June 6, 2023:
  Select appropriate algorithms
  Develop and train the chosen algorithms using the preprocessed data
  Evaluate the models based on performance metrics

-Document and reporting: June 8, 2023:
  Prepare the report by summarizing the project and findings
  Document the entire process including : data sources, preprocessing steps, algorithms used and validation techniques.
  Create a dashboard with Tableau
  Review and finalize the report

-Create a web page for self-evaluated of heart disease risk: June 9, 2023:

Using HTML files and other libraries such as Bootstrap to create a single web page about predicting heart disease risk based on users' input.

-Testing, revising and practicing for the presentation: June 12, 2023:

Test all the codes to ensure there are not technical issues take place

Revise the final report before submitting to Canva

Create PowerPoint file for the presentation

-Presentation day: June 13, 2023:

6. Team members:

- Lynn Hoang
- John James
- Gavin Wan
- Isaac Rodriguez