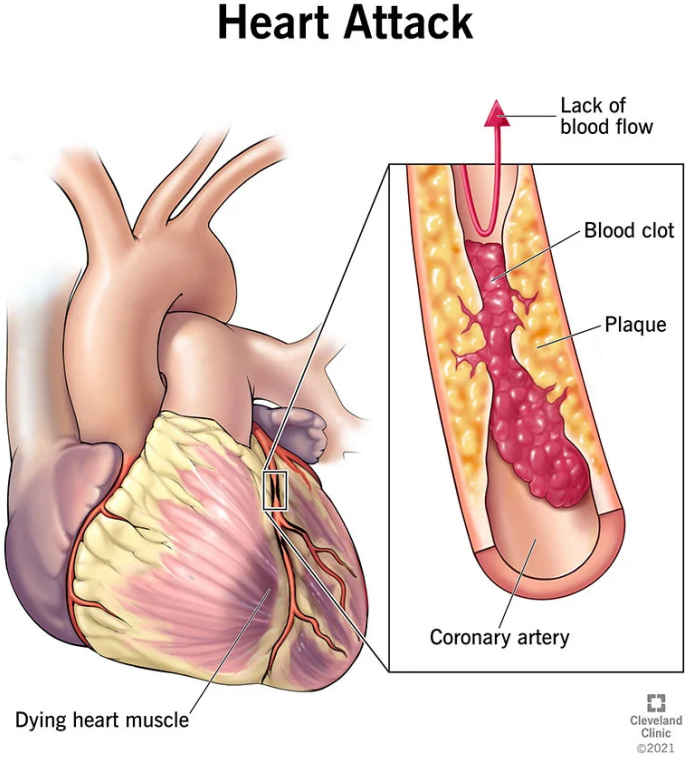
Laboratory Activities for Week 5: Supervised Learning (Part I)

SC310005 Artificial Intelligence
Khon Kaen Business School

(10 Points) Predicting Heart Attack Risk Using Decision Tree

Task: You are given a dataset containing various patient attributes. Your task is to create a machine learning model using a Decision Tree Classifier to predict the likelihood of a heart attack for a patient based on their medical attributes.

Dataset:
https://raw.githubusercontent.com/kaopanboonyuen/SC310005_ArtificialIntelligence_2023s1/main/dataset/heart_attack_dataset.csv



Dataset Description:

1. Age: Age of the patient.
2. Sex: Gender of the patient. (0 = female, 1 = male)
3. Chest Pain Type (CP): The type of chest pain experienced by the patient. It's categorized into four types: typical angina, atypical angina, non-anginal pain, and asymptomatic.
4. Resting Blood Pressure (trtbps): The patient's resting blood pressure measured in mm Hg.
5. Cholesterol Level (chol): Serum cholesterol level in mg/dl.
6. Fasting Blood Sugar (fbs): Whether the fasting blood sugar is greater than 120 mg/dl. (1 = true, 0 = false)
7. Resting Electrocardiographic Results (restecg): Results of the resting electrocardiogram, indicating normalcy or abnormalities in heart rhythm.
8. Maximum Heart Rate Achieved (thalachh): Maximum heart rate achieved during the cardiac stress test.
9. Exercise Induced Angina (exng): Presence of exercise-induced angina. (1 = yes, 0 = no)
10. ST Depression (oldpeak): ST depression induced by exercise relative to rest.
11. Slope of the Peak Exercise ST Segment (slp): The slope of the peak exercise ST segment categorized as upsloping, flat, or downsloping.
12. Number of Major Vessels (caa): Number of major vessels colored by fluoroscopy (ranging from 0 to 3).
13. Thalassemia (thall): A blood disorder type categorized as normal, fixed defect, or reversible defect.
14. **Output (Target Variable):** Presence of heart disease. (0 = no, 1 = yes)

Task Requirements:

- [ ] Load the dataset and preprocess it (handle missing values, encode categorical variables if any).
- [ ] Split the data into training (80%) and testing (20%) sets using a random_state of 2023.
- [ ] Train a Decision Tree Classifier using the training data.
- [ ] Make predictions on the test set and evaluate the model's performance.
- [ ] Calculate and report the accuracy score of the model on the test set.

Submission:

Submit your Colab code along with comments explaining each step, and a brief summary or report mentioning the accuracy achieved by your model on the test set.

Note: Ensure to handle any preprocessing steps required for the dataset and properly comment on your code to explain each part of the process.