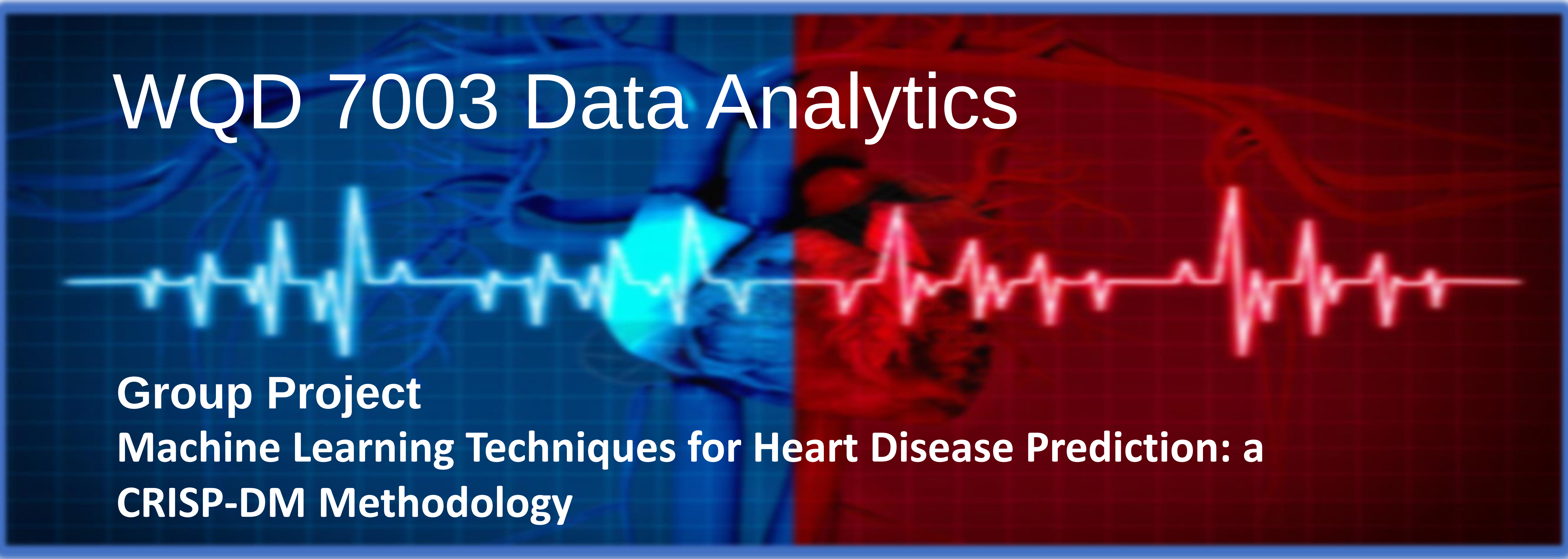


WQD 7003 Data Analytics

Group Project

Machine Learning Techniques for Heart Disease Prediction: a CRISP-DM Methodology

Group #8 | Prepared by	Matric No.
Kristian Surya Dinata	S2043845
Xin Dong	22060696
Jiang Jiajia	22069349
Zhu Mei	22060214
Yuejing Huang	S2158553

Machine Learning Techniques for Heart Disease Prediction: a CRISP-DM Methodology

WQD7003 DATA ANALYTICS | Group Project

A. Introduction:

Heart disease is one of the leading causes of death worldwide. According to the World Health Organization (2020), cardiovascular diseases (CVDs), including heart disease, account for approximately 17.9 million deaths each year, making it the leading cause of death globally. Early detection and prevention can help reduce the risk of CVDs, particularly heart disease and ultimately lead to saving lives.

Machine learning algorithms can assist in predicting the likelihood of heart disease in patients based on their health data. The algorithms can be trained to predict based on various risk factors such as age, gender, blood pressure, cholesterol levels, and smoking habits.

In this project, we will use the CRISP-DM methodology to develop heart disease prediction models using a publicly available dataset from UCI.

Reference: *World Health Organization. (2020). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)*

B. Problem Statement:

Heart disease is a leading cause of death worldwide, and early identification of individuals at risk can help prevent the development of heart disease and improve patient outcomes. Despite advances in medical technology, traditional risk assessment methods have limitations, and there is a need for more accurate and efficient methods for identifying individuals at risk. The unknown and important aspect of this problem is the potential of machine learning techniques to improve heart disease prediction accuracy and how to implement these techniques in real-world clinical practice.

C. Objectives:

1. To evaluate the effectiveness of current machine learning techniques in predicting heart disease

2. To improve a practical model for implementing these techniques in clinical practice

Metadata	
Source	https://archive.ics.uci.edu/ml/datasets/heart%2Bdisease
Area	Life/Health
Attributes	14
Observations	303
Associated task	Classification

Machine Learning Techniques for Heart Disease Prediction: a CRISP-DM Methodology

WQD7003 DATA ANALYTICS | Group Project

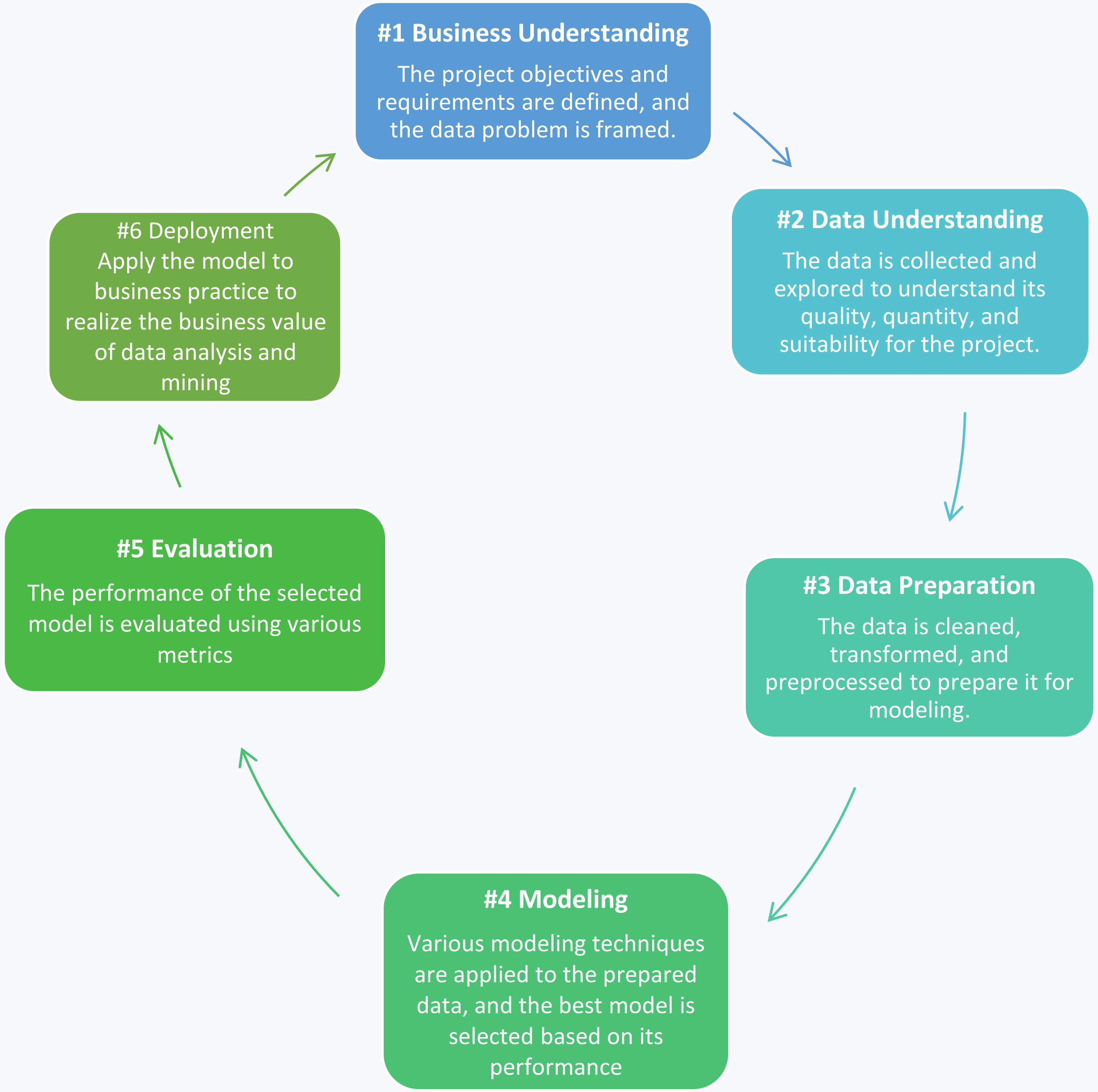

#1 Business Understanding
The project objectives and requirements are defined, and the data problem is framed.
#2 Data Understanding
The data is collected and explored to understand its quality, quantity, and suitability for the project.
#3 Data Preparation
The data is cleaned, transformed, and preprocessed to prepare it for modeling.
#4 Modeling
Various modeling techniques are applied to the prepared data, and the best model is selected based on its performance
#5 Evaluation
The performance of the selected model is evaluated using various metrics
#6 Deployment
Apply the model to business practice to realize the business value of data analysis and mining

Figure 1. CRISP-DM framework

Step #1: Business Understanding

- Objective: To develop a machine learning model that predicts the likelihood of a person having heart disease based on various risk factors, for early intervention and prevention.
- Business question: Can we build a predictive model that classifies patients as having or not having heart disease based on medical history and demographic info for clinical use?

Step #2: Data Understanding

- Data source: Kaggle's Heart Disease UCI dataset, with 303 rows and 14 columns of preprocessed medical and demographic info.
- Exploratory data analysis: Use summary statistics, histograms, box plots, and correlation matrices to better understand the data.

Step #3: Data Preparation

- Data cleaning
- Feature scaling: Use MinMaxScaler to ensure all features are on the same scale.
- Data splitting: Split the data into training and testing sets with a 70/30 ratio.

Step #4: Modeling

- Model selection: Try logistic regression, decision tree, random forest, SVM, and ANN.
- Evaluate the models: Use k-fold cross-validation to tune hyperparameters and evaluate accuracy, precision, recall, and F1-score.

Step #5: Evaluation

- Use k-fold cross-validation to evaluate model performance.
- Metrics: Accuracy, precision, recall, and F1-score.

Step #6: Deployment

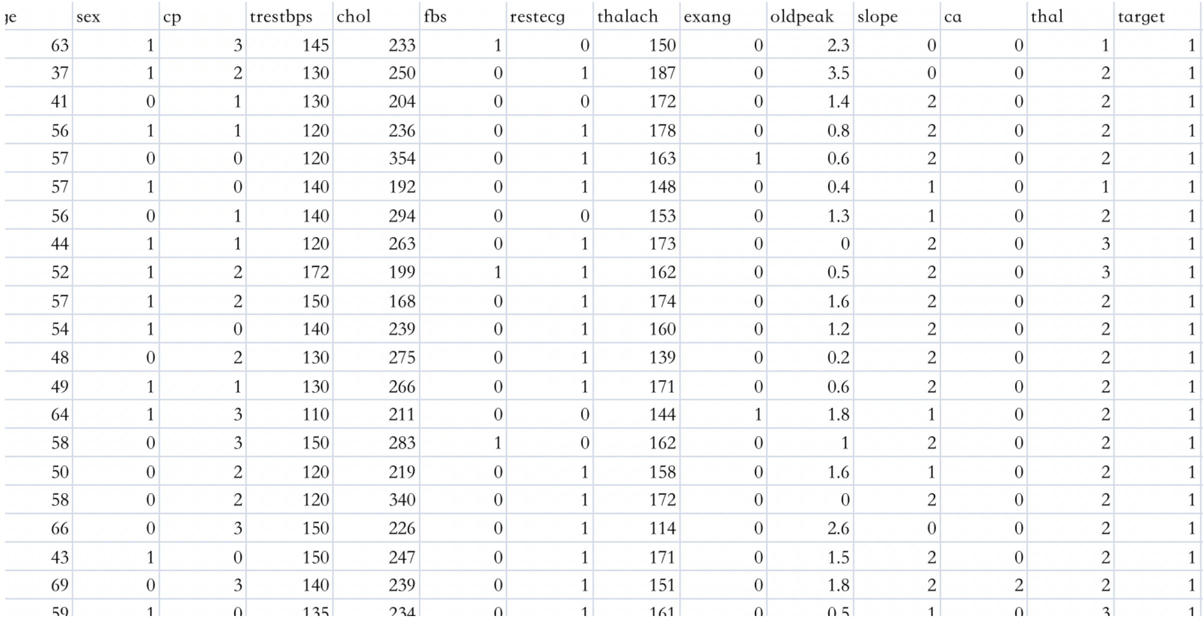
- Benefits: A machine learning model for predicting heart disease risk could aid clinical practice by identifying high-risk patients and enabling early intervention.
- Use cases: Healthcare providers could use the model to provide targeted interventions, while insurance companies could assess client risk and adjust premiums accordingly.
- Impact: Encourage individuals to adopt healthier habits and reduce the burden of disease on healthcare systems.

Machine Learning Techniques for Heart Disease Prediction: a CRISP-DM Methodology

WQD7003 DATA ANALYTICS | Group Project

Appendix
Display of the dataset in use

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type
   -- Value 0: typical angina
   -- Value 1: atypical angina
   -- Value 2: non-anginal pain
   -- Value 3: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
   -- Value 0: normal
   -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    -- Value 0: upsloping
    -- Value 1: flat
    -- Value 2: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
    and the label
14. condition: 0 = no disease, 1 = disease



ge	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1