National Taipei University of Technology

Introduction to Data Science
Fall 2019-2020

# Decision Tree Algorithm

Name: Li-Wei Yeh

StudentID: 108012047

Date: December 30th, 2019

# Contents

# 1. Data science

Data Science is a field that uses scientific methods to extract information from data. It is related to Statistics, Data Mining and Big Data. Data Science uses mathematical concepts to understand and analyze data, like; statistics, machine learning, computer science etc.

Data Science is primarily used to make decisions and predictions, making use of predictive analytics, prescriptive analytics, and machine learning.

# 2. Algorithms

An algorithm is a process or a set of rules to be followed in calculations or other problem-solving operations.

Typically, there are two kinds of algorithms, supervised and unsupervised algorithms

## 2.1 Supervised Algorithms

Supervised algorithms are algorithms that rely on labeled data, meaning that the outcome of the data is already known and thus, has the right “answer”. Supervised algorithms rely on labeled training data to predict outcome of unforeseen data.

## 2.2 Unsupervised Algorithms

Unsupervised algorithms are algorithms that don’t have labeled data. You need to make the model discover its own information. It is generally less accurate than supervised algorithms due to the trouble of having unlabeled data.

## 2.3 The Algorithms

There are numerous different algorithms to use for Data Science. Some of the more popular ones are: Linear Regression, K-means Clustering, Logistic Regression, Support Vector Machines, Feed Forward Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks and Decision Trees. And even those algorithms have different ways to implement them but come from the same idea.

This report will mainly focus on the Decision Trees algorithm.

# 3. Decision Trees

The Decision Tree algorithm is one of the predictive models used for statistics, data mining and machine learning. A decision tree can be used to visually represent decisions in decision making, thus making it easier for humans to understand the process it takes, to reach such conclusions.

Decision Tree learning is the construction of a decision tree from class-labeled training tuples. It has a flow-chart-like structure, where each node represents the outcome of the prediction. The trees are essentially a series of questions designed to assign a certain classification.

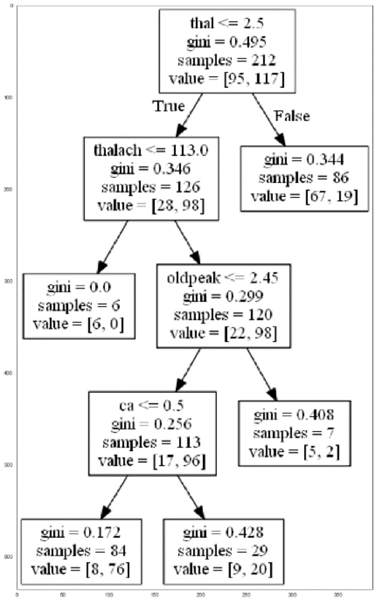
There are many different kinds of Decision Tree algorithms. Some working better than others. The more notable ones are:

- ID3
- C4.5 (successor of ID3)
- CART (Classification and Regression Trees)
- Chi-square automatic interaction detection (CHAID)

The most popular ones are ID3/C4.5 and CART. They were both created at roughly the same time independently from each other, whilst still following a similar approach for learning a decision tree from training datasets.

## 3.1 CART

With cart, we basically take the same steps. Instead of the measure being entropy, this time the unit of measure will be Gini values. The Gini value will decide which feature will be chosen first (comparable to the most information gain). A Gini value gives an indication of how good a split is by how mixed the classes are.



$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

The steps we have to take:

1. Compute the Gini index for every feature
2. Calculate Gini gain
3. Pick the best Gini gain attribute
4. Repeat

The result will be a decision tree, which is reliant on the previous choices in the flowchart-like graph.

# 4. The project

## 4.1 Introduction

### 4.1.1 Motivation

I wanted to use my programming knowledge to help other people. I can do this by analyzing data of a heart disease dataset and predict when a person may or may not have heart disease. Heart disease will thus, be identified easier, depending on the accuracy of my project.

### 4.1.2 Objectives

The goal of this project is to analyze the dataset. I will be using the heart disease dataset (.csv) and by using the data of the dataset, predict if other people will have heart disease or not.

## 4.2 Project Plan and Deadlines

### 4.2.1 Related work and Resources

I will be using my old project, which is a project using the Decision Tree Algorithm, to predict Spotify data. Further references and resources will be projects of other people, library documentation, etc.

### 4.2.2 Methodology and Tools

I will be working in sprints of 1 week, every week finishing certain tasks.

The tools I will be using:

- Python, as the programming language
- Anaconda, to handle environments
- Jupyter Notebook, to write code and see the results
- pandas library, to read (and transform) datasets
- numpy library, for containing data in easy to use multi-dimensional arrays
- matplotlib library, to plot the data
- sklearn library, for predictive analysis
- graphviz library. for graphing visualization of the tree

### 4.2.3 Expected Results

People will be more likely to have heart disease if the chest pain type is high, sex is male and thal is high.

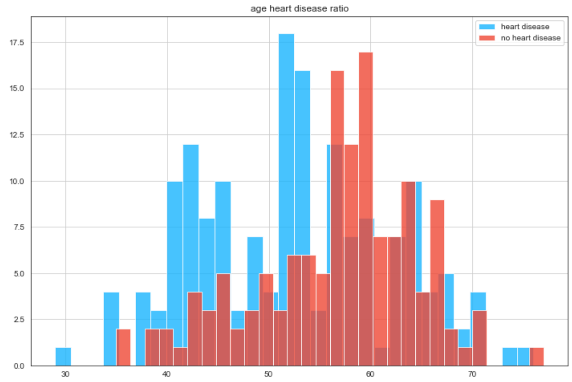
### 4.2.4 Timeline

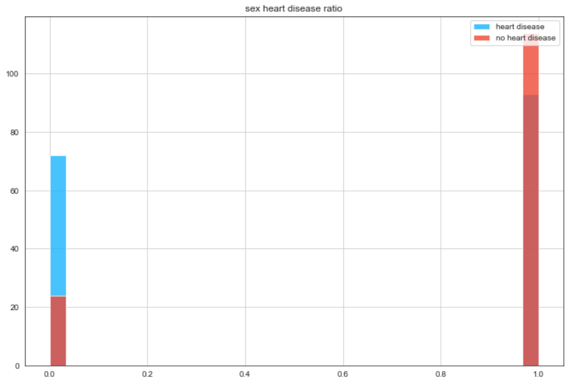
Plotting in the first week, predicting in the second week.

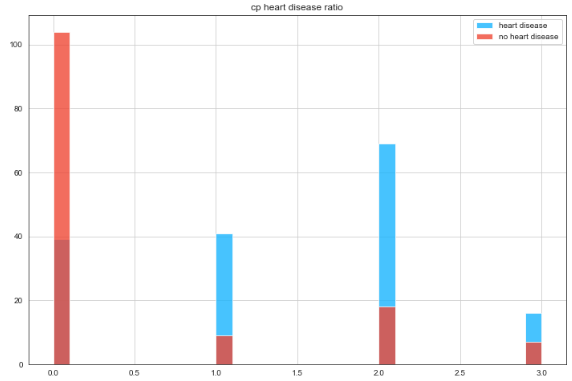
# 4.3 Results

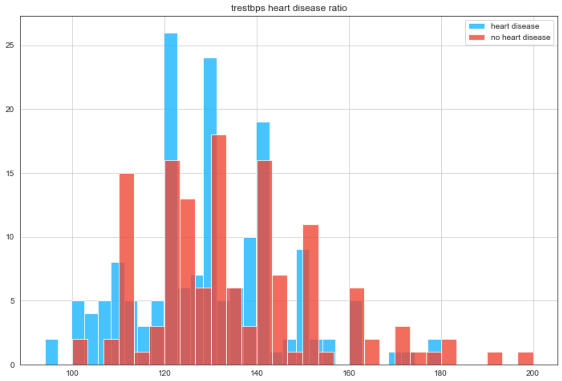
## 4.3.1 Graphs of the attributes

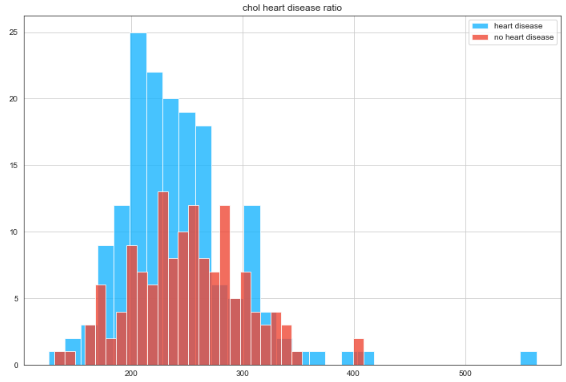
I used the pandas library to read the dataset. Then I used pandas as well to separate the attributes and afterwards plotted them in graphs, using matplotlib.

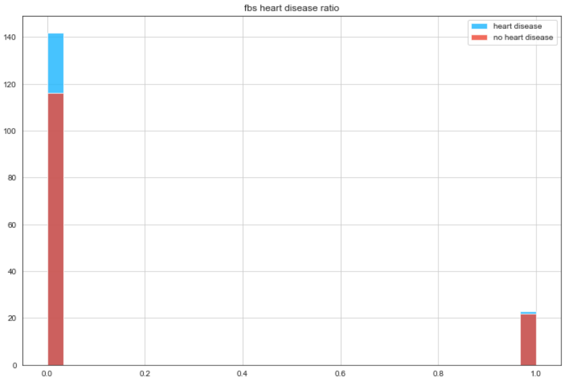


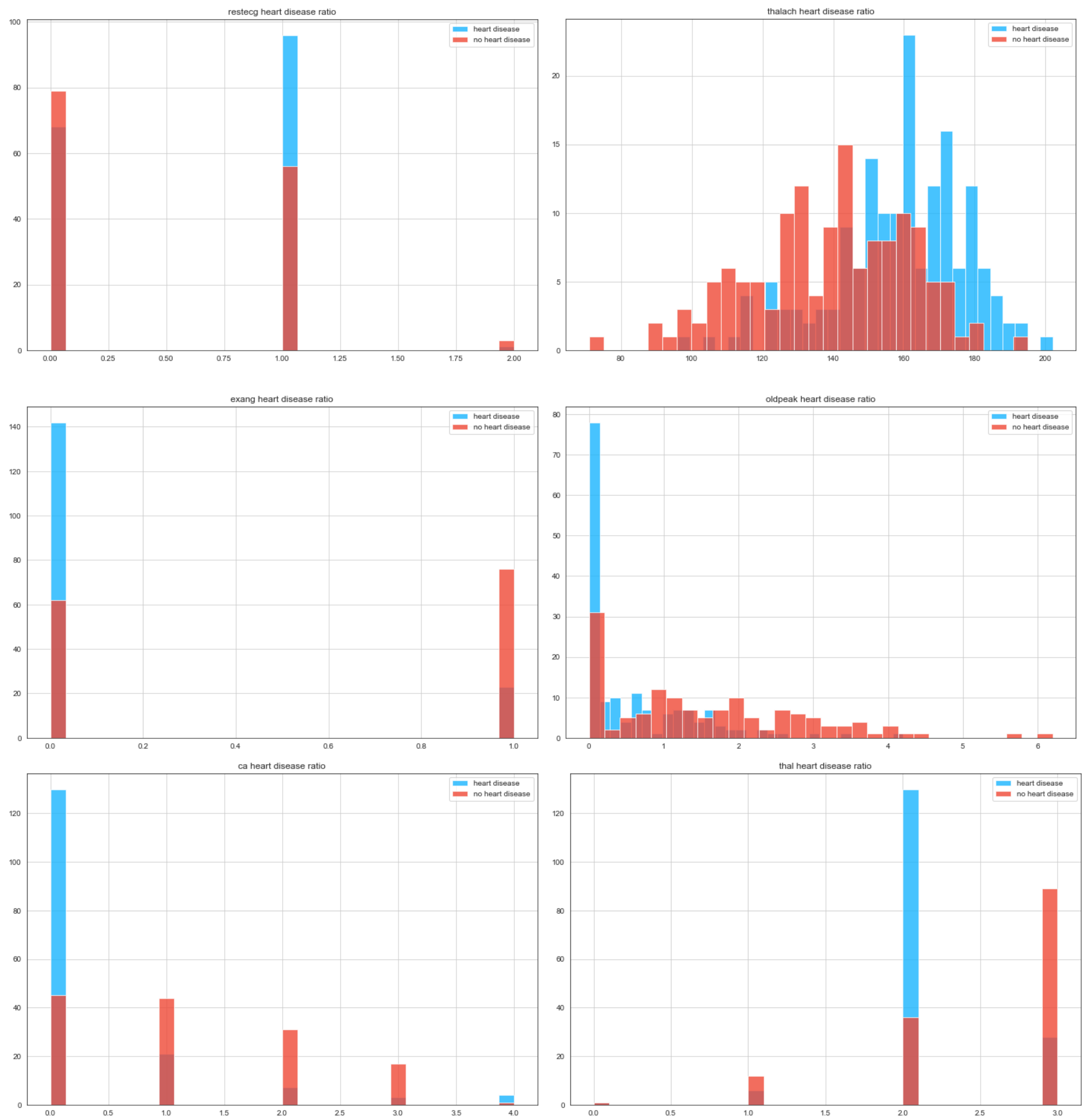











restecg heart disease ratio

thalach heart disease ratio

exang heart disease ratio

oldpeak heart disease ratio

ca heart disease ratio

thal heart disease ratio

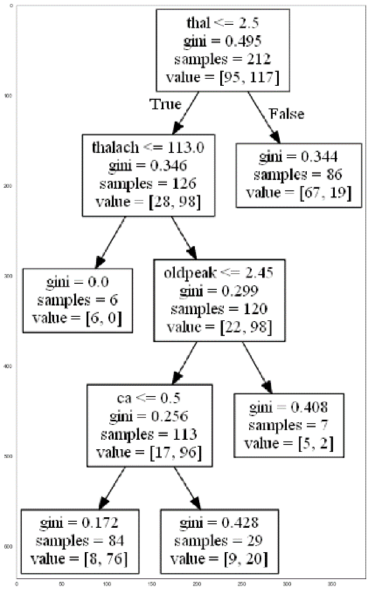
### 4.3.2 Classifier

For the classifier, I used the sklearn library to use the DecisionTreeClassifier (which uses the CART method with Gini values). In section 3.1, the algorithm is explained.

Using this classifier, I've trained the dataset to predict the outcome.

### 4.3.3 The Tree

To make the tree, I've used graphviz to create a tree in .png file. The tree shows which questions the program gets asked, coupled with the Gini value is has calculated. It also has the amount of samples that's left.



### 4.3.4 Outcome

The accuracy of this Decision Tree is 73.63%. Looking at the tree, it gets the most information gain from the attribute thal, thalach and oldpeak.