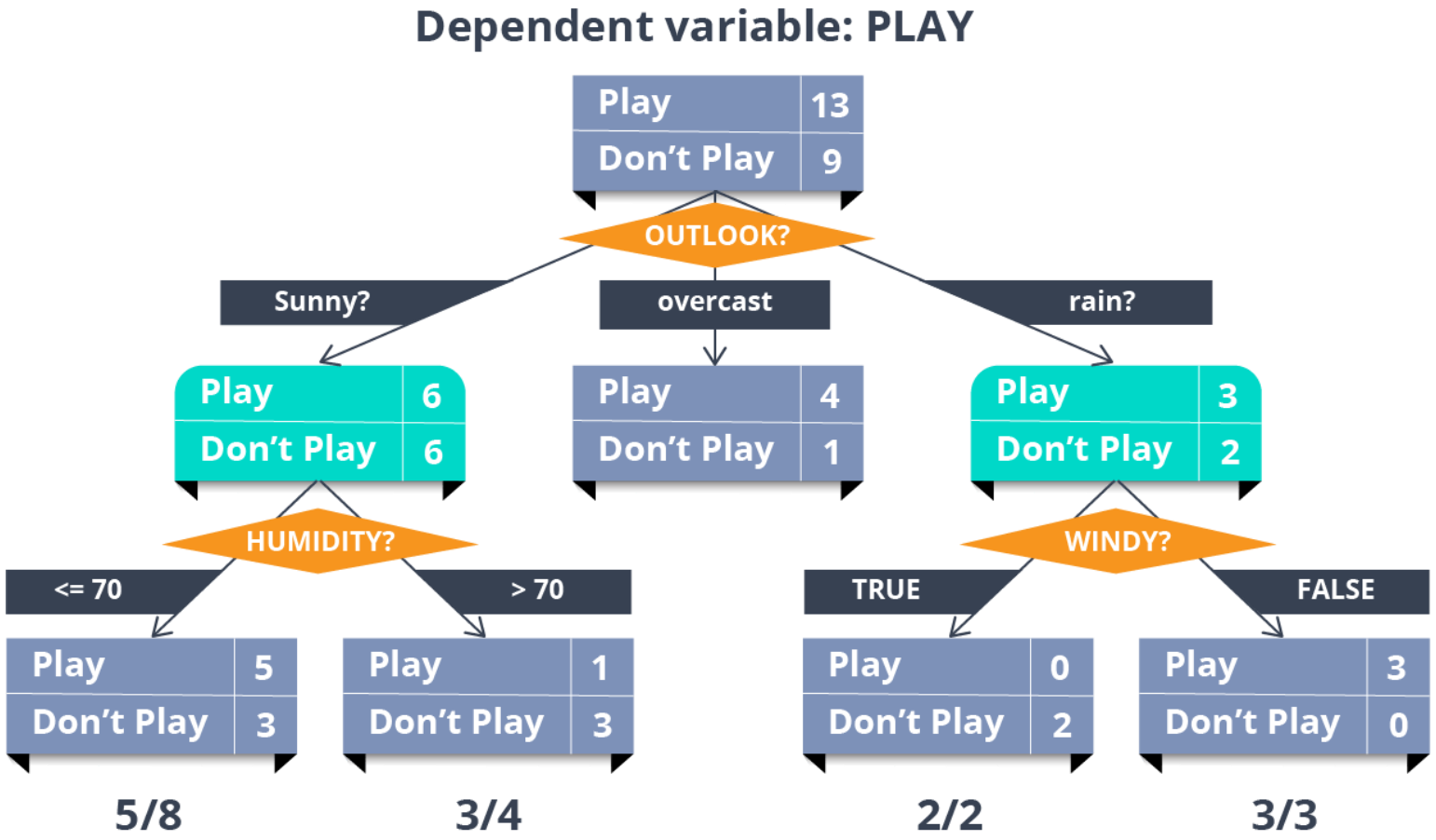
DECISION TREE
PHƯƠNG PHÁP CÂY QUYẾT ĐỊNH

- Nguyễn Hoàng Yến Như
- Nguyễn Trần Phúc Nghi
- Nguyễn Trần Phúc An
- Nguyễn Đức Anh Phúc
- Trịnh Thị Thanh Trúc
- KS. Cao Bá Kiệt
- KS. Quan Chí Khánh An
- KS. Lê Ngọc Huy
- CN. Bùi Cao Doanh
- CN. Nguyễn Trọng Thuận
- KS. Phan Vĩnh Long
- KS. Nguyễn Cường Phát
- ThS. Nguyễn Hoàng Ngân
- KS. Hồ Thái Ngọc
- ThS. Đỗ Văn Tiến
- ThS. Nguyễn Hoàn Mỹ
- ThS. Dương Phi Long
- ThS. Trương Quốc Dũng
- ThS. Nguyễn Thành Hiệp
- ThS. Nguyễn Võ Đăng Khoa
- ThS. Võ Duy Nguyên
- TS. Nguyễn Văn Tâm
- ThS. Trần Việt Thu Phương
- TS. Nguyễn Tấn Trần Minh Khang

Thuật toán Cây quyết định



- https://www.edureka.co/blog/decision-trees/

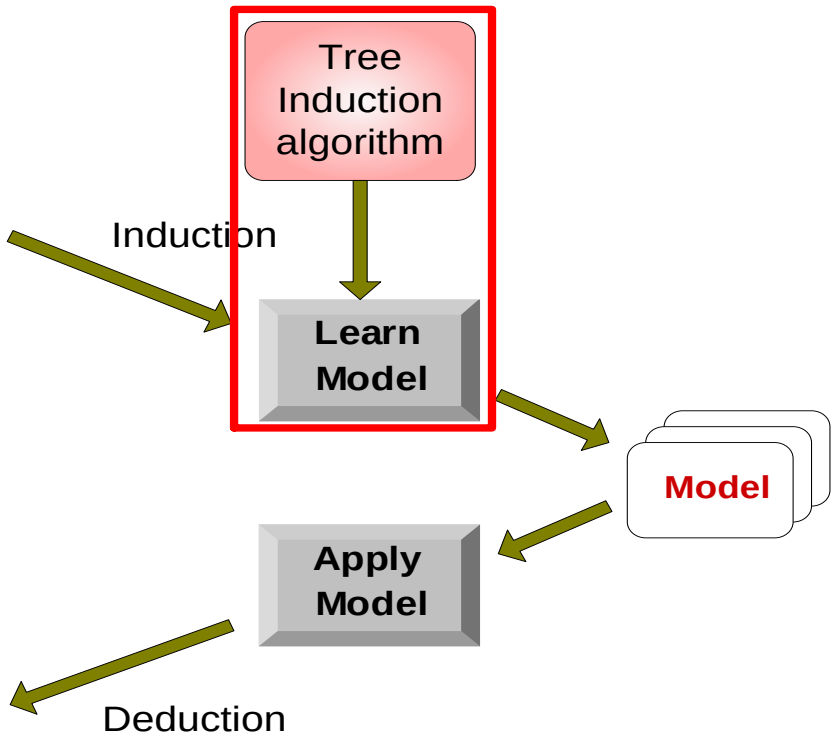
Thuật toán Cây quyết định

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Thuật toán Cây quyết định

- In computer science, Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).
- It is one of the predictive modeling approaches used in statistics, data mining and machine learning.

- https://en.wikipedia.org/wiki/Decision_tree_learning

Thuật toán Cây quyết định

- Tree models where the target variable can take a discrete set of values are called classification trees;
- In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.
- Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

- https://en.wikipedia.org/wiki/Decision_tree_learning

ĐỊNH NGHĨA CÂY QUYẾT ĐỊNH

Định nghĩa Cây quyết định

- Decision tree is a classifier in the form of a tree structure.
  + Decision node: specifies a test on a single attribute.
  + Leaf node: indicates the value of the target attribute.
  + Arc/edge: split of one attribute.
  + Path: a disjunction of test to make the final decisionan.
- Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node.

PHÂN LOẠI CÂY QUYẾT ĐỊNH

Phân loại cây quyết định

- Classification tree analysis is when the predicted outcome is the class (discrete) to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

- https://en.wikipedia.org/wiki/Decision_tree_learning

CÁC THUẬT TOÁN CÂY QUYẾT ĐỊNH

Các thuật toán xây dựng cây quyết định

- ID3 (Iterative Dichotomiser 3) – J. Ross Quinlan.
- C4.5 (successor of ID3) – J. Ross Quinlan.
- CART (Classification And Regression Tree) – L. Breiman, J. Friedman, R. Olshen, and C. Stone.
- Chi-square automatic interaction detection (CHAID). Performs multi-level splits when computing classification trees.
- MARS: extends decision trees to handle numerical data better.
- Conditional Inference Trees.

- https://en.wikipedia.org/wiki/Decision_tree_learning

SO SÁNH CÁC THUẬT TOÁN

Các thuật toán xây dựng cây quyết định

Methods	CART	C4. 5	CHAID	QUEST
Measure used to select input variable	Gini index; Twoing criteria	Entropy info-gain	Chi-square	Chi-square for categorical variables; J-way ANOVA for continuous/ordinal variables
Pruning	Pre-pruning using a single-pass algorithm	Pre-pruning using a single-pass algorithm	Pre-pruning using Chi-square test for independence	Post-pruning
Dependent variable	Categorical/ Continuous	Categorical/ Continuous	Categorica	Categorical
Input variables	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous
Split at each node	Binary; Split on linear combinations	Multiple	Multiple	Binary; Split on linear combinations

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

ỨNG DỤNG CÂY QUYẾT ĐỊNH

Ứng dụng thuật toán cây quyết định

Ứng dụng thuật toán cây quyết định

- Hotel.
  + Predicting high occupancy dates for hotels
  + Dự báo ngày đặt chỗ nhiều nhất.

  + Identifying factors leading to better gross margins on a retail chain.
  + Xác định các yếu tố chủ chốt mang đến lợi nhuận tốt nhất.

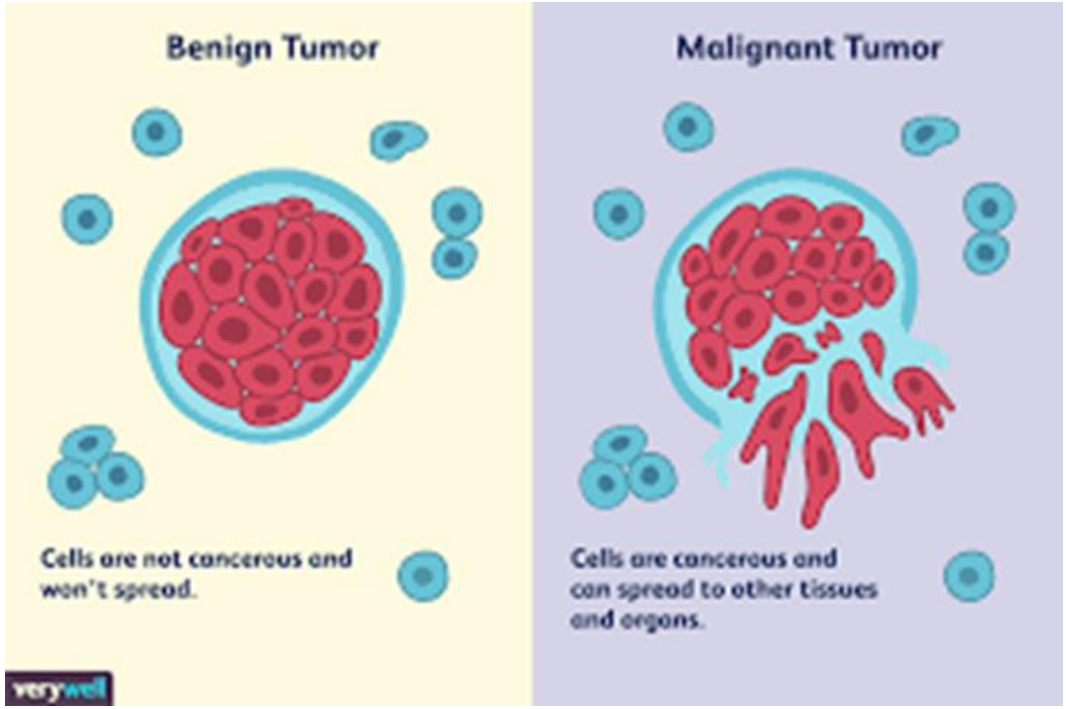
- https://www.quora.com/In-what-real-world-applications-is-the-decision-tree-classifier-used

Ứng dụng thuật toán cây quyết định

- Hotel.
  + Identifying correlates to high average checks for a global quick-service restaurant chain.
  + Xác định tương quan giữa check in và dịch vụ nhà hàng.

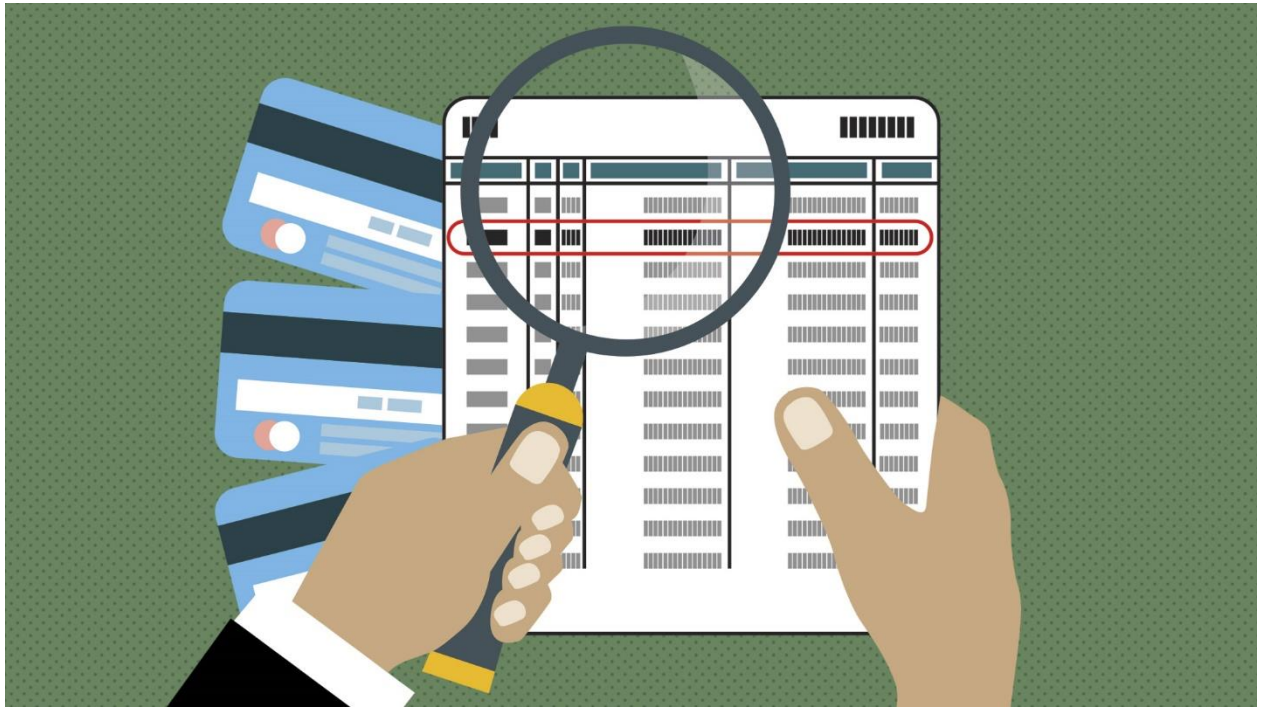
- https://www.quora.com/In-what-real-world-applications-is-the-decision-tree-classifier-used

Ứng dụng thuật toán cây quyết định



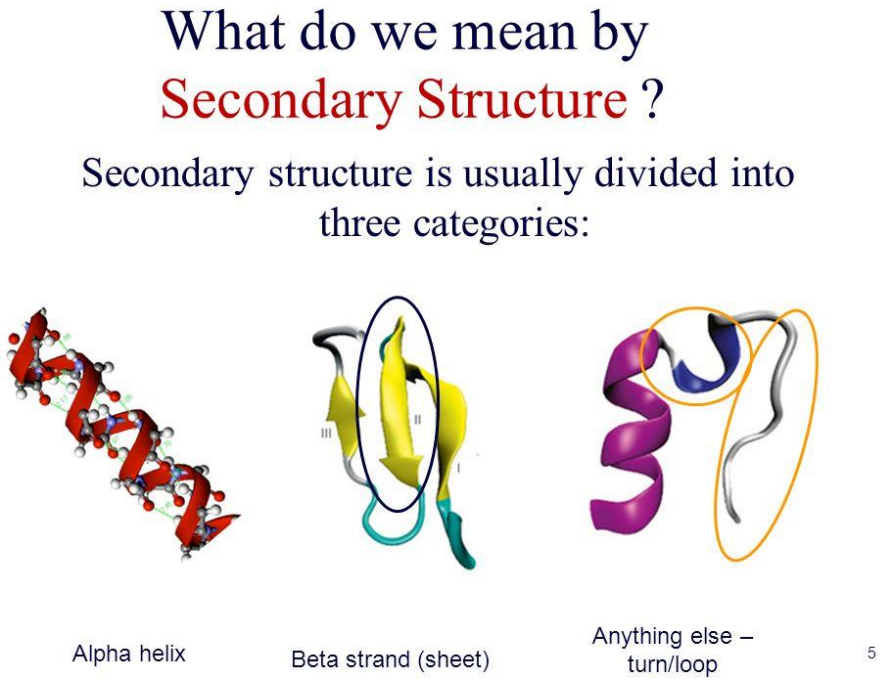
– Predicting tumor cells as benign or malignant

Ứng dụng thuật toán cây quyết định



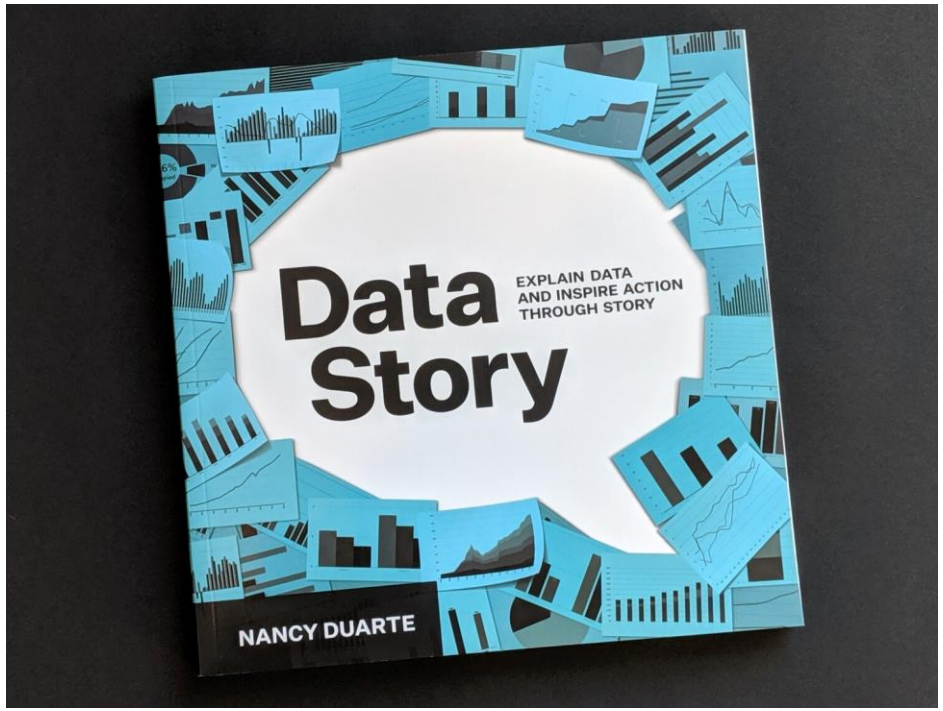
– Classifying credit card transactions as legitimate or fraudulent

Ứng dụng thuật toán cây quyết định



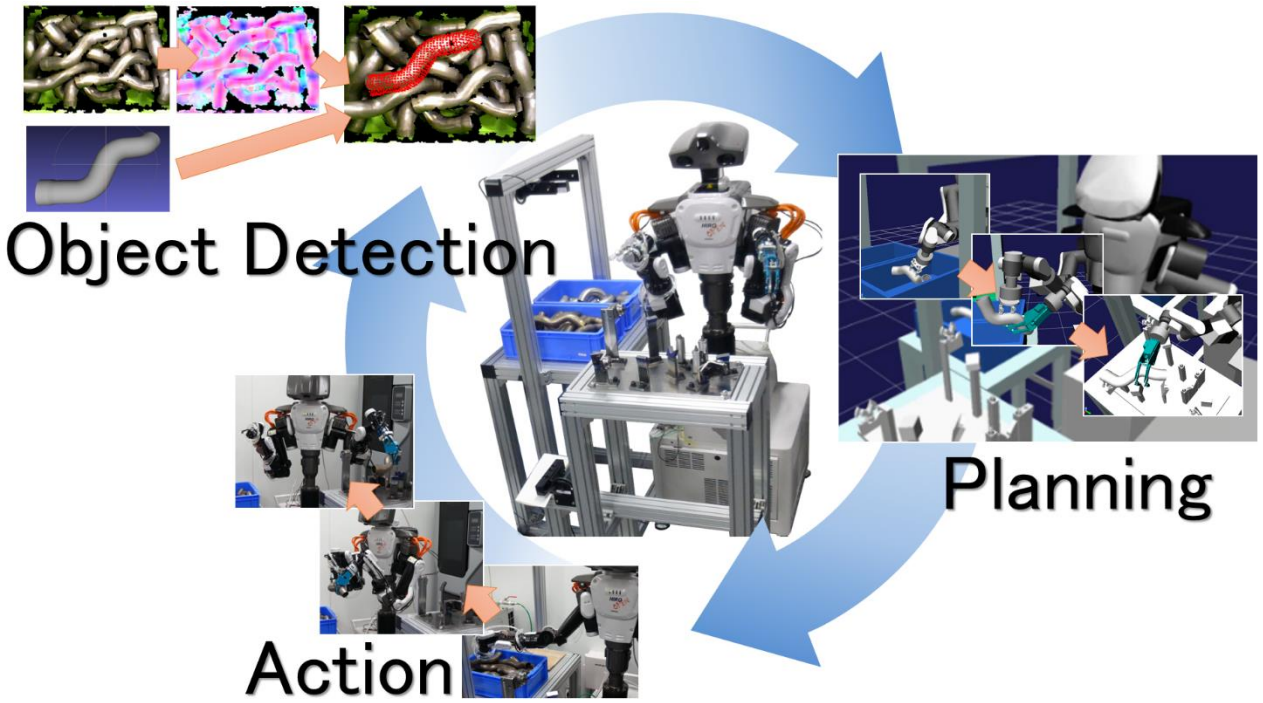
– Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil.

Ứng dụng thuật toán cây quyết định



- Categorizing news stories as finance, weather, entertainment, sports, etc.

Ứng dụng thuật toán cây quyết định



– Object classification for robot manipulator (Tan 1993)

ĐIỂM MẠNH CỦA CÂY QUYẾT ĐỊNH

Điểm mạnh của thuật toán

- Simple to understand and interpret.
- Cây quyết định dễ hiểu.

- Decision Trees are easy to explain. It results in a set of rules.
- Cây quyết định dễ giải thích. Kết quả của nó là tập các luật.

- https://en.wikipedia.org/wiki/Decision_tree_learning

Điểm mạnh của thuật toán

- Able to handle both numerical and categorical data.
- Cây quyết định có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại.

- Requires little data preparation.
- Việc chuẩn bị dữ liệu cho một cây quyết định là cơ bản hoặc không cần thiết.

- https://en.wikipedia.org/wiki/Decision_tree_learning

Điểm mạnh của thuật toán

- Uses a white box model.
- Cây quyết định là một mô hình hộp trắng.

- Possible to validate a model using statistical tests.
- Có thể thẩm định một mô hình bằng các kiểm tra thống kê.

Điểm mạnh của thuật toán

- Performs well with large datasets.
- Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn.

- It follows the same approach as humans generally follow while making decisions.
- Cây quyết định giống cách tiếp cận ra quyết định của con người.

- https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

Điểm mạnh của thuật toán

- Interpretation of a complex Decision Tree model can be simplified by its visualizations. Even a naive person can understand logic.
- Dễ trực quan hóa.

- https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

ĐIỂM YẾU CỦA CÂY QUYẾT ĐỊNH

Điểm yếu của thuật toán

- There is a high probability of overfitting in Decision Tree.
- Xác xuất xảy ra quá khớp cao.

- Generally, it gives low prediction accuracy for a dataset as compared to other machine learning algorithms.
- Độ chính xác của việc dự báo thấp so với các thuật toán học máy khác.

- https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

Điểm yếu của thuật toán

- Information gain in a decision tree with categorical variables gives a biased response for attributes with greater no. of categories.
- Độ đo Information gain gặp khó khăn với dữ liệu có miền giá trị là dữ liệu phân loại.

- Calculations can become complex when there are many class labels.
- Việc tính toán trở nên phức tạp nếu biến phụ thuộc có nhiều lớp (nhiều nhãn).
- https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

CÁC ĐỘ ĐO XÂY DỰNG CÂY

Các độ đo - Metrics

- Gini impurity – Gini index: CART.
- Information gain: ID3, C4.5, C5.0.
- Variance reduction: CART.

- https://en.wikipedia.org/wiki/Decision_tree_learning

Example

VÍ DỤ MINH HỌA

Cây quyết định – Ví dụ minh họa

– Cho tập dữ liệu các mặt hàng như sau. Xác định khi nào người dùng quyết định Mua và Không mua một mặt hàng.

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

Cây quyết định – Ví dụ minh họa

– Bài toán: sử dụng thuật toán Cây định danh để xác định khi nào người dùng quyết định Mua và Không mua một mặt hàng.

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 01: Có mấy thuộc tính.

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 01: Có mấy thuộc tính.

– Trả lời: 4 thuộc tính.

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 01: Có mấy thuộc tính.

– Trả lời: 4 thuộc tính.

– Câu hỏi 02: Thuộc tính thứ nhất tên gì?

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 01: Có mấy thuộc tính.

– Trả lời: 4 thuộc tính.

– Câu hỏi 02: Thuộc tính thứ nhất tên gì?

– Trả lời: Kích cỡ

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 01: Có mấy thuộc tính.

– Trả lời: 4 thuộc tính.

– Câu hỏi 02: Thuộc tính thứ nhất tên gì?

– Trả lời: Kích cỡ

– Câu hỏi 03: Thuộc tính thứ hai tên gì?

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 01: Có mấy thuộc tính.

– Trả lời: 4 thuộc tính.

– Câu hỏi 02: Thuộc tính thứ nhất tên gì?

– Trả lời: Kích cỡ

– Câu hỏi 03: Thuộc tính thứ hai tên gì?

– Trả lời: Màu.

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 04: Thuộc tính thứ ba tên gì?

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 04: Thuộc tính thứ ba tên gì?

– Trả lời: Hình dáng.

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 04: Thuộc tính thứ ba tên gì?

– Trả lời: Hình dáng.

– Câu hỏi 05: Thuộc tính thứ tư tên gì?

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 04: Thuộc tính thứ ba tên gì?

– Trả lời: Hình dáng.

– Câu hỏi 05: Thuộc tính thứ tư tên gì?

– Trả lời: Quyết định.

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 04: Thuộc tính thứ ba tên gì?

– Trả lời: Hình dáng.

– Câu hỏi 05: Thuộc tính thứ tư tên gì?

– Trả lời: Quyết định.

– Câu hỏi 06: Miền giá trị của thuộc tính Kích cỡ?

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 04: Thuộc tính thứ ba tên gì?

– Trả lời: Hình dáng.

– Câu hỏi 05: Thuộc tính thứ tư tên gì?

– Trả lời: Quyết định.

– Câu hỏi 06: Miền giá trị của thuộc tính Kích cỡ?

– Trả lời: {Nhỏ, Trung Bình, Lớn}.

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 07: Miền giá trị của thuộc tính Màu?

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 07: Miền giá trị của thuộc tính Màu?

– Trả lời: {Đỏ, Vàng, Xanh}.

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 07: Miền giá trị của thuộc tính Màu?

– Trả lời: {Đỏ, Vàng, Xanh}.

– Câu hỏi 08: Miền giá trị của thuộc tính Hình dáng?

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 07: Miền giá trị của thuộc tính Màu?

– Trả lời: {Đỏ, Vàng, Xanh}.

– Câu hỏi 08: Miền giá trị của thuộc tính Hình dáng?

– Trả lời: {Cầu, Hộp, Trụ, Nón}.

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 07: Miền giá trị của thuộc tính Màu?

– Trả lời: {Đỏ, Vàng, Xanh}.

– Câu hỏi 08: Miền giá trị của thuộc tính Hình dáng?

– Trả lời: {Cầu, Hộp, Trụ, Nón}.

– Câu hỏi 09: Miền giá trị của thuộc tính Quyết định?

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Câu hỏi 07: Miền giá trị của thuộc tính Màu?

– Trả lời: {Đỏ, Vàng, Xanh}.

– Câu hỏi 08: Miền giá trị của thuộc tính Hình dáng?

– Trả lời: {Cầu, Hộp, Trụ, Nón}.

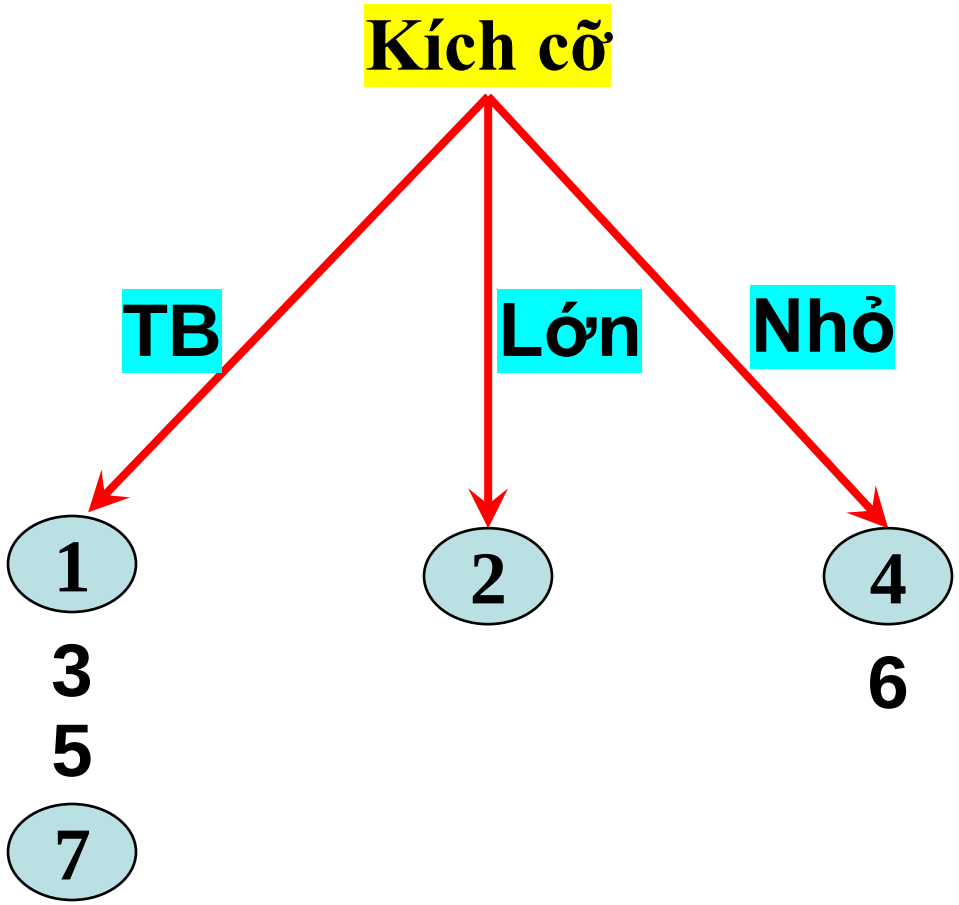
– Câu hỏi 09: Miền giá trị của thuộc tính Quyết định?

– Trả lời: {Mua, Không Mua}.

Cây quyết định – Ví dụ minh họa

Kích cỡ

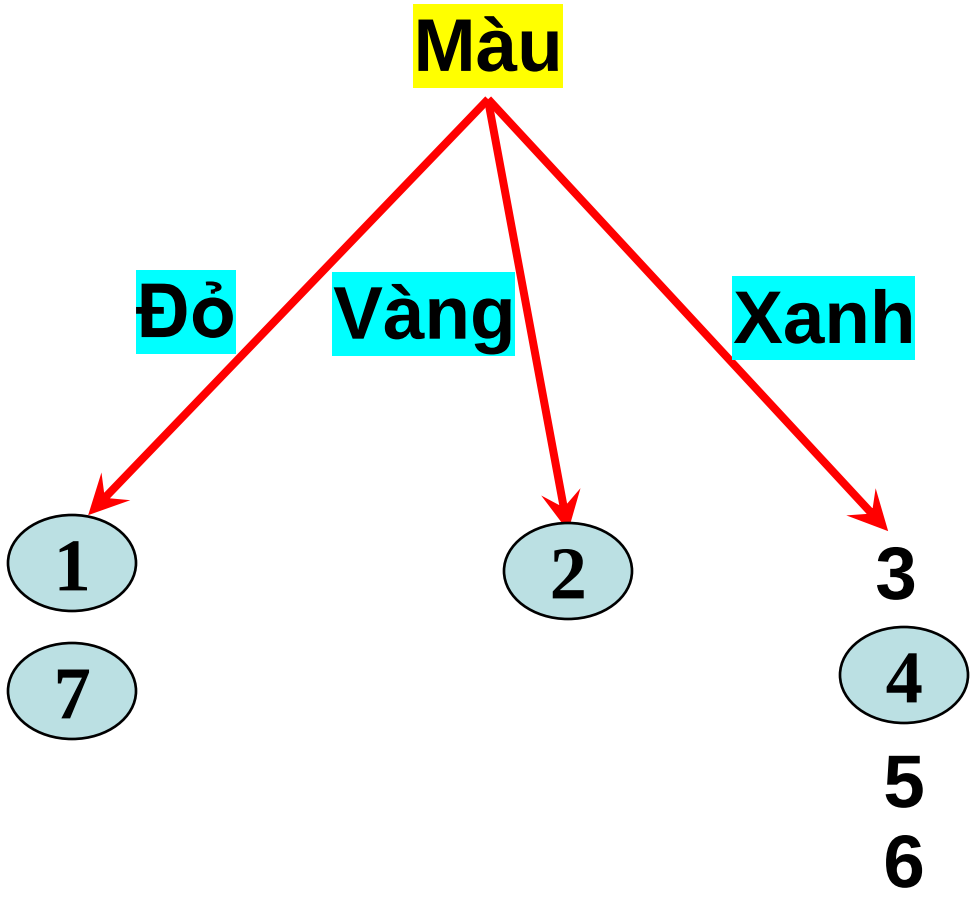
STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua



Cây quyết định – Ví dụ minh họa

Màu

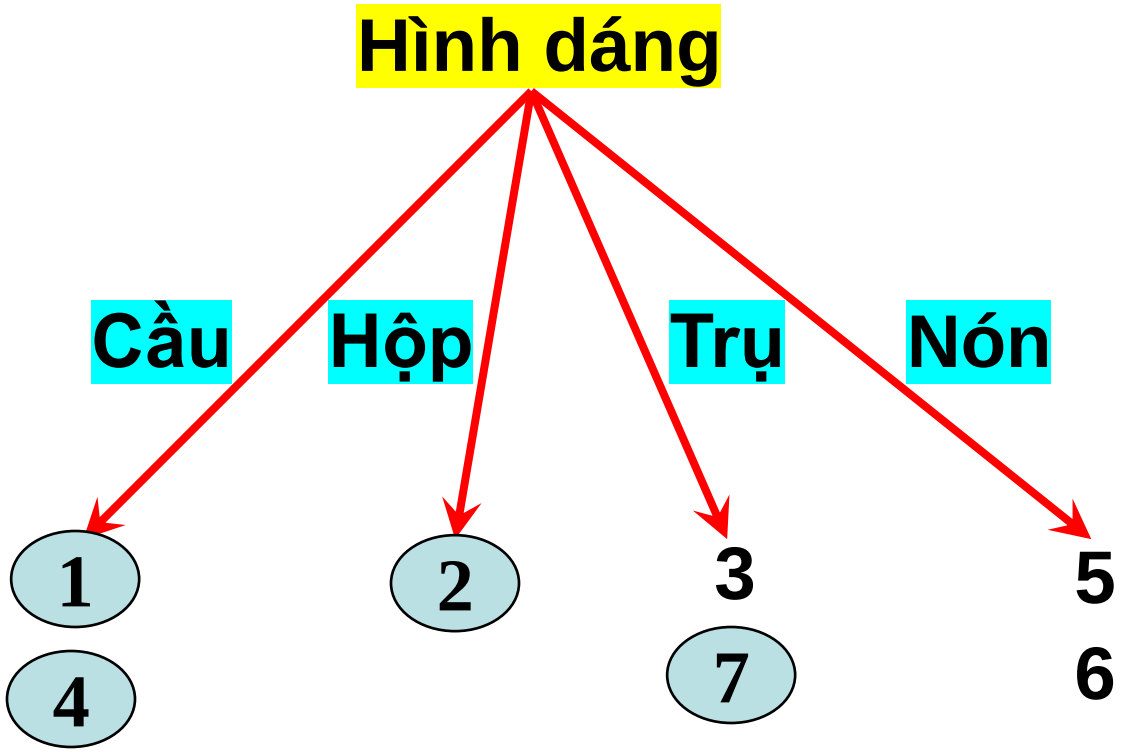
STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua



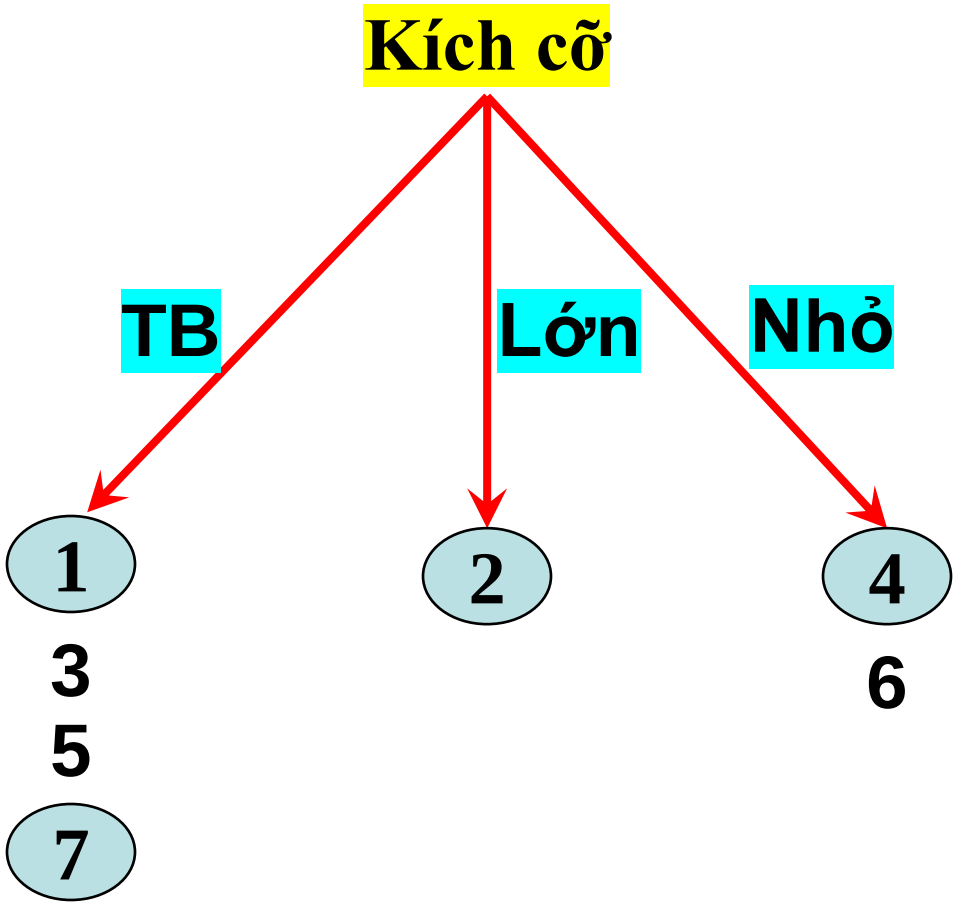
Cây quyết định – Ví dụ minh họa

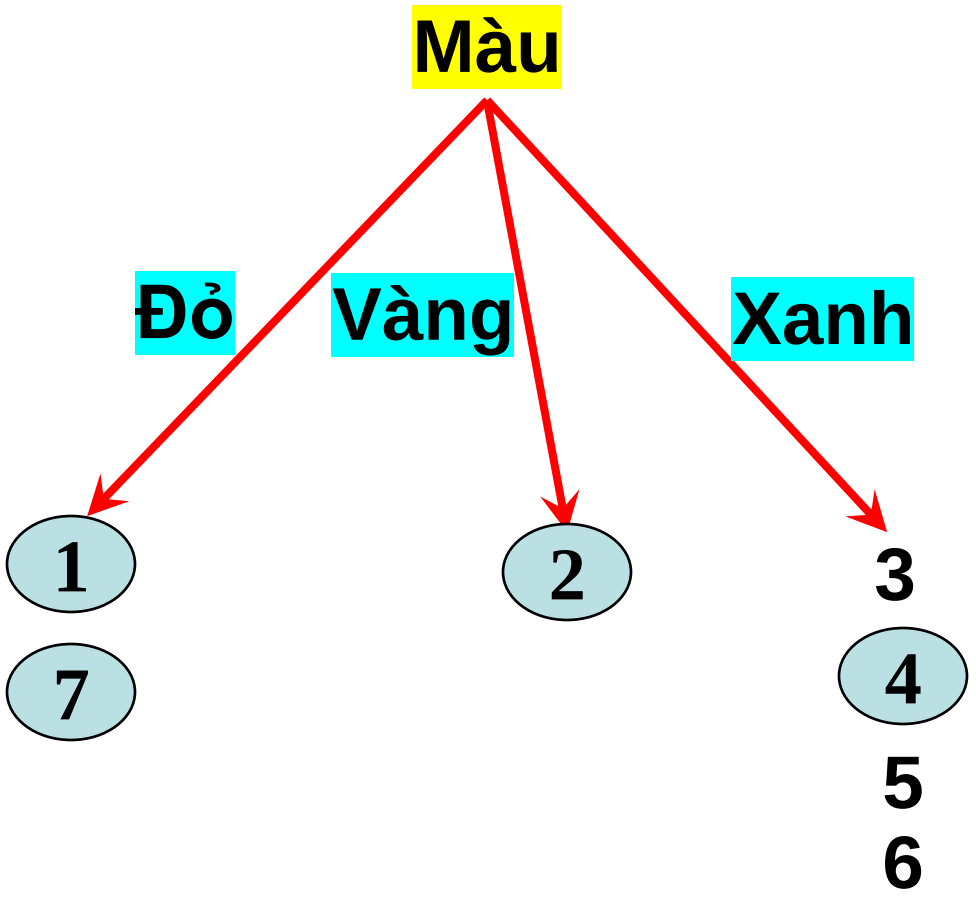
Hình dáng

STT	Kích cỡ	Màu	Hình dáng	Quyết định
1	Trung bình	Đỏ	Cầu	Mua
2	Lớn	Vàng	Hộp	Mua
3	Trung bình	Xanh	Trụ	Không Mua
4	Nhỏ	Xanh	Cầu	Mua
5	Trung bình	Xanh	Nón	Không Mua
6	Nhỏ	Xanh	Nón	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

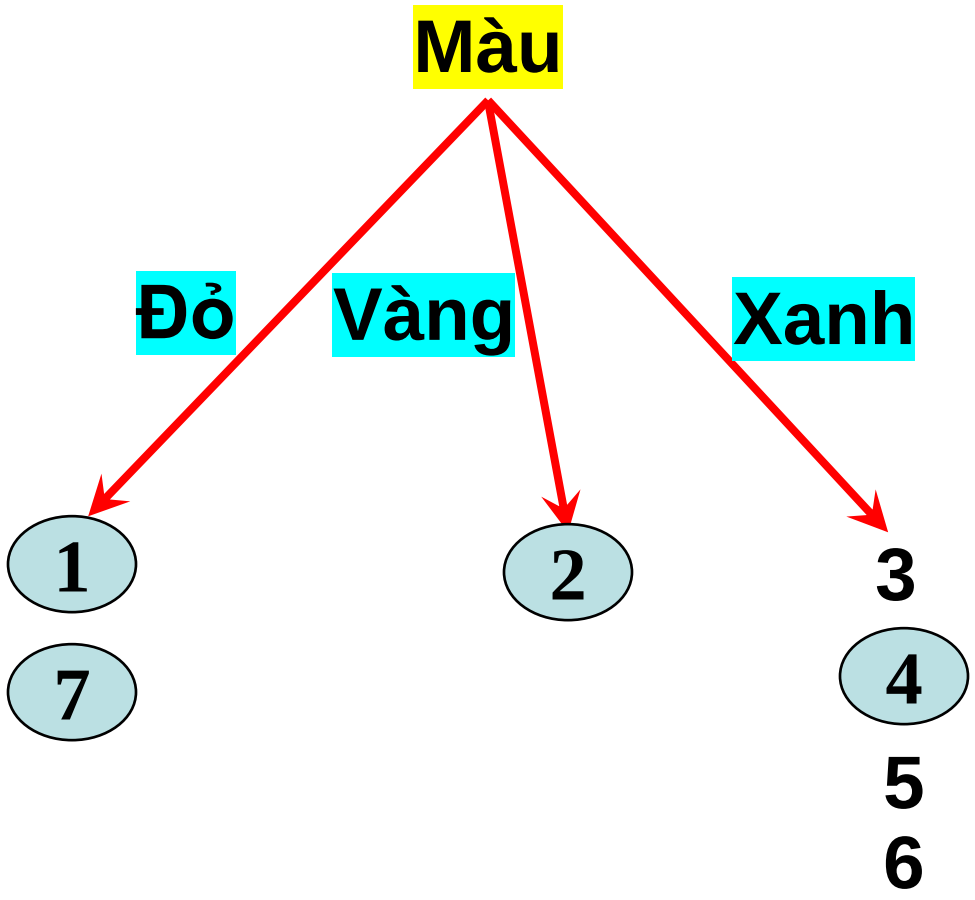


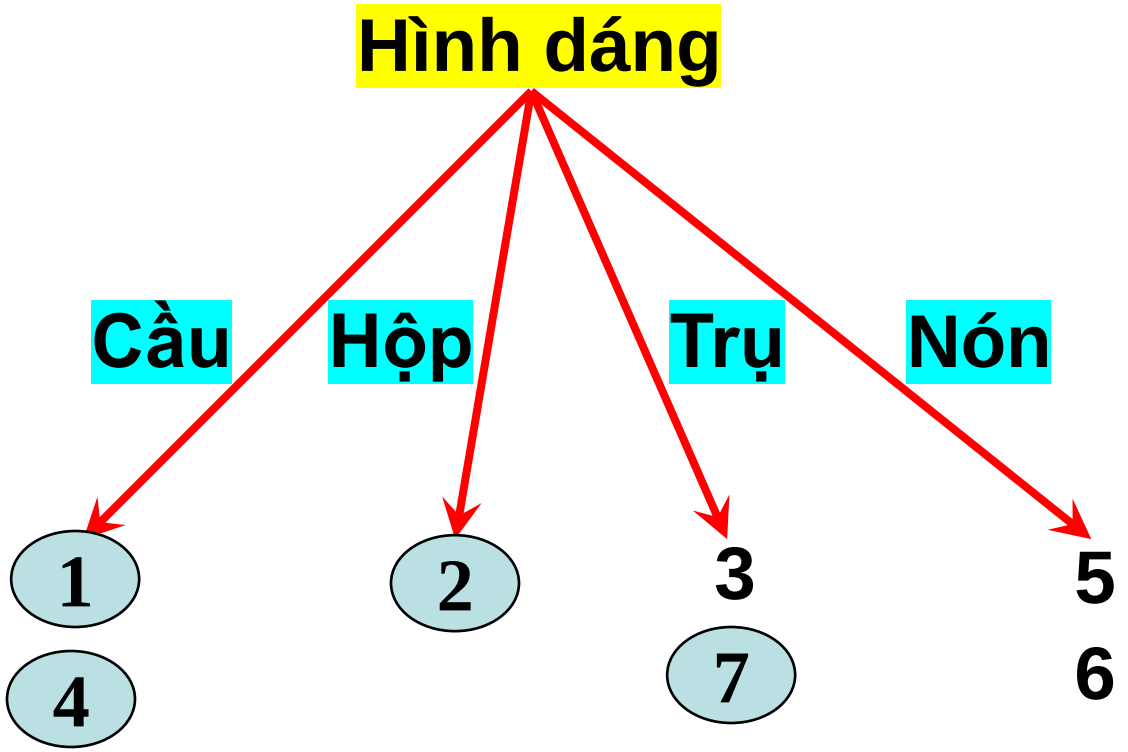
Cây quyết định – Ví dụ minh họa





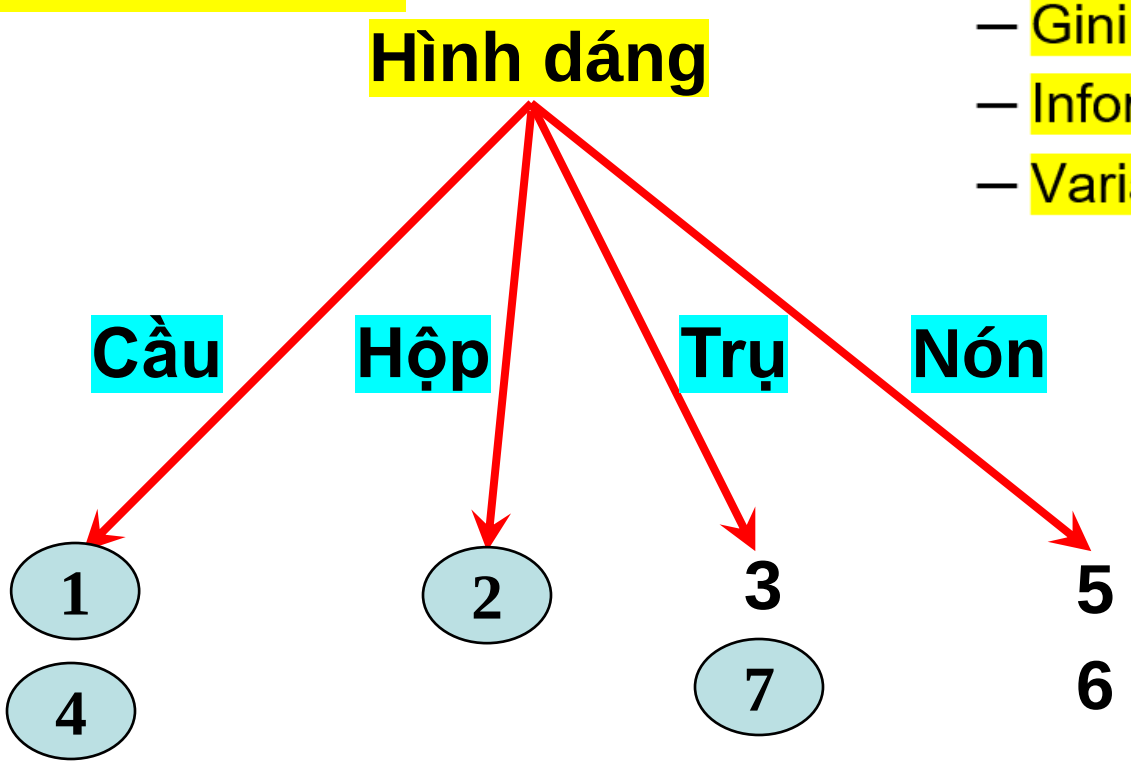
Cây quyết định – Ví dụ minh họa





Cây quyết định – Ví dụ minh họa

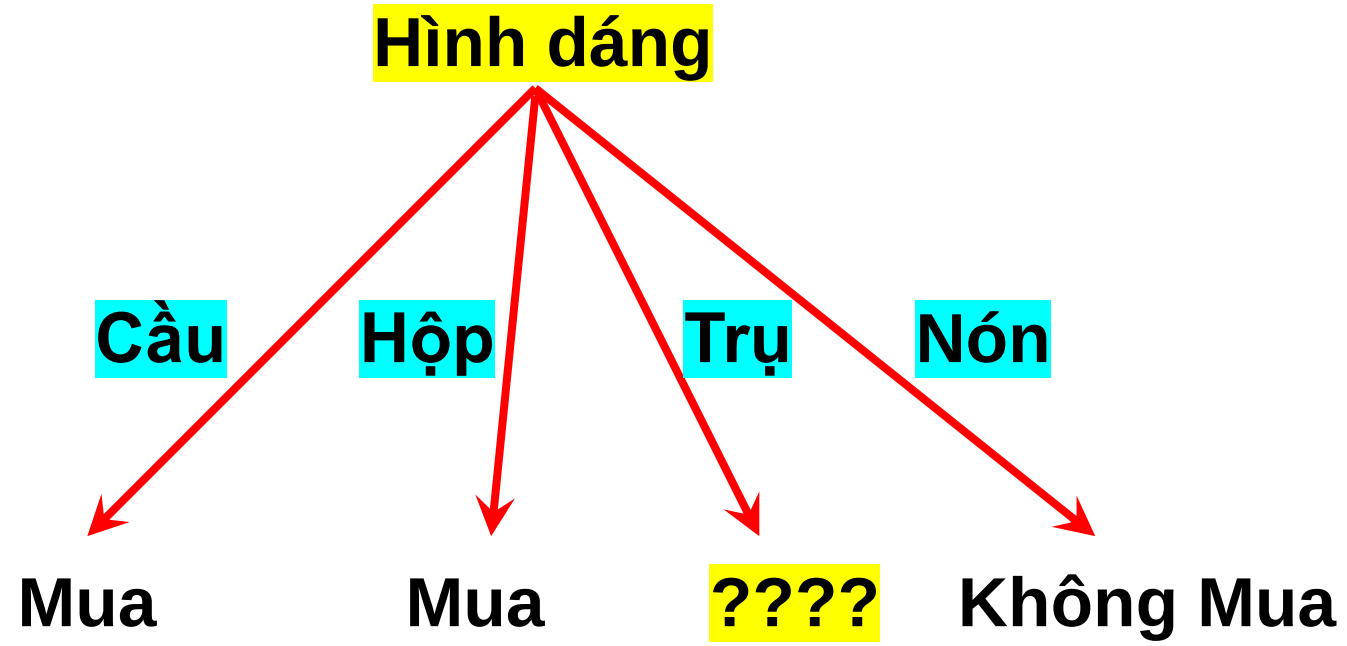
– Chọn thuộc tính Hình dáng làm thuộc tính phân loại vì thuộc tính này có xxx xxx xxx xxx xxx.



– Gini impurity – Gini index: CART.
– Information gain: ID3, C4.5, C5.0.
– Variance reduction: CART.

Cây quyết định – Ví dụ minh họa

– Chọn thuộc tính Hình dáng làm thuộc tính phân loại vì thuộc tính này có xxx xxx xxx xxx xxx.



Cây quyết định – Ví dụ minh họa

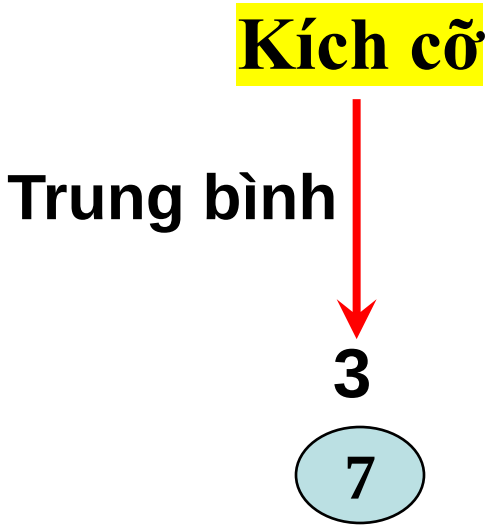
– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
3	Trung bình	Xanh	Trụ	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

Cây quyết định – Ví dụ minh họa

– Bảng quan sát

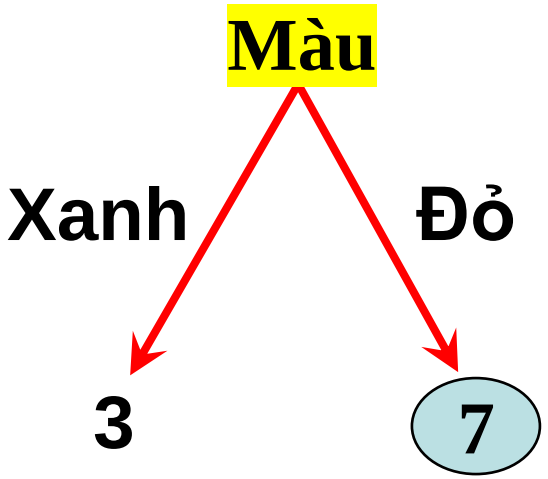
STT	Kích cỡ	Màu	Hình dáng	Quyết định
3	Trung bình	Xanh	Trụ	Không Mua
7	Trung bình	Đỏ	Trụ	Mua



Cây quyết định – Ví dụ minh họa

– Bảng quan sát

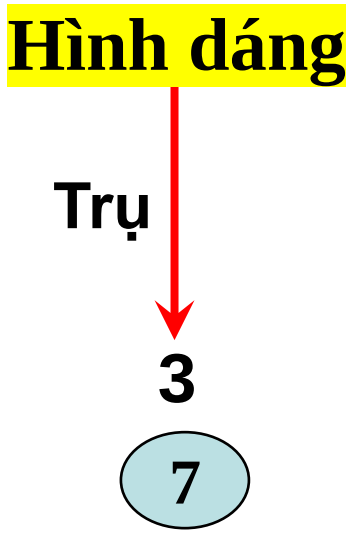
STT	Kích cỡ	Màu	Hình dáng	Quyết định
3	Trung bình	Xanh	Trụ	Không Mua
7	Trung bình	Đỏ	Trụ	Mua



Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
3	Trung bình	Xanh	Trụ	Không Mua
7	Trung bình	Đỏ	Trụ	Mua



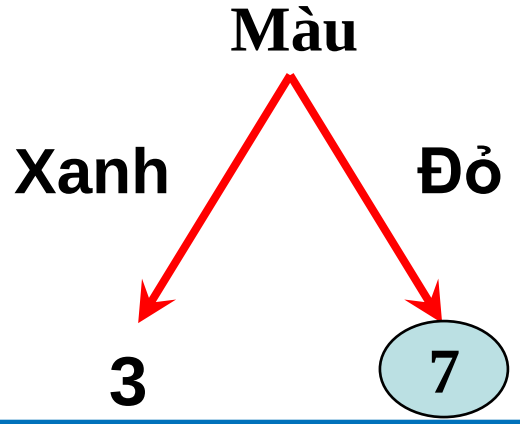
Cây quyết định – Ví dụ minh họa

– Bảng quan sát

STT	Kích cỡ	Màu	Hình dáng	Quyết định
3	Trung bình	Xanh	Trụ	Không Mua
7	Trung bình	Đỏ	Trụ	Mua

– Chọn thuộc tính Màu sắc làm thuộc tính phân loại vì thuộc tính này có xxx xxx xxx xxx xxx.

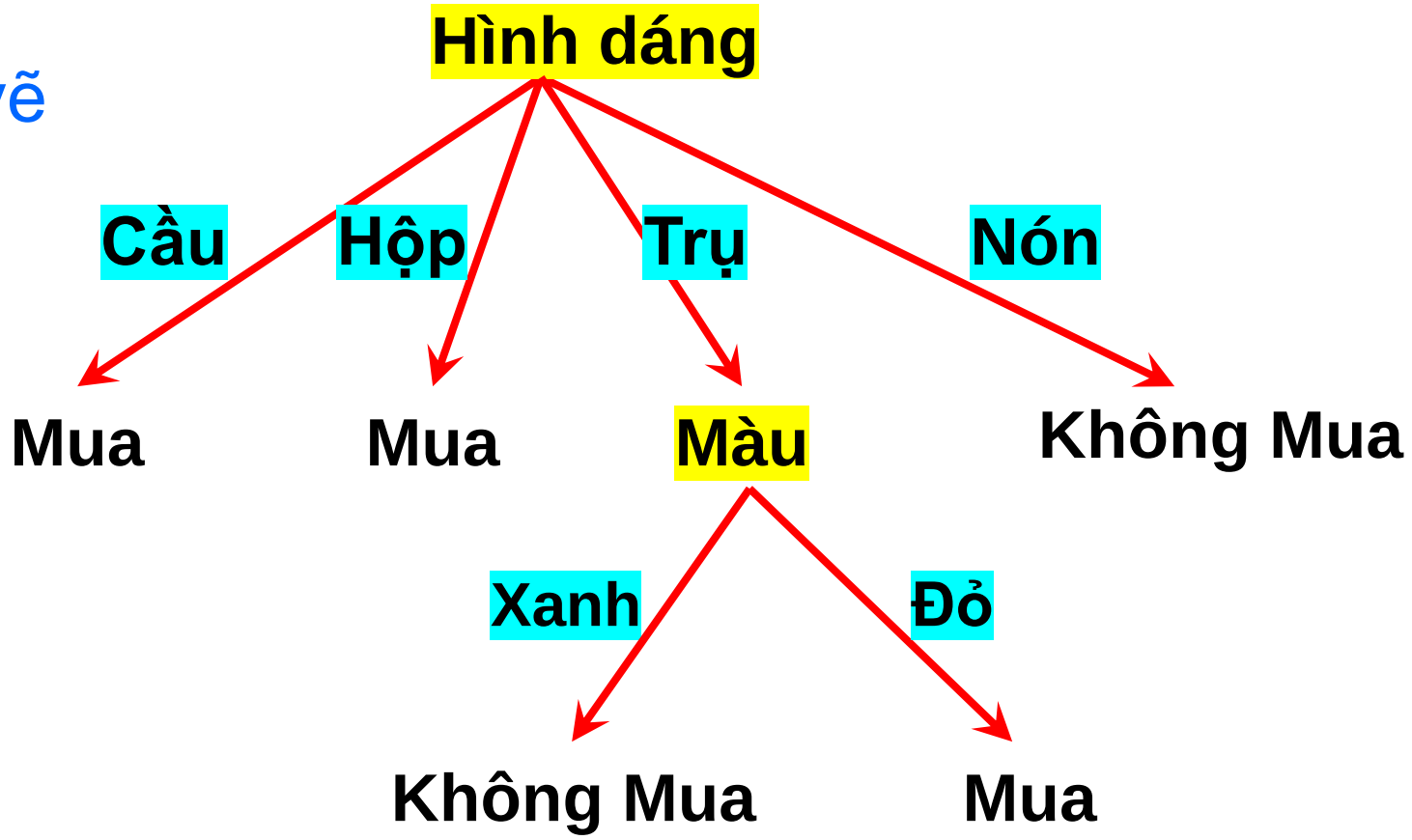
- Gini impurity – Gini index: CART.
- Information gain: ID3, C4.5, C5.0.
- Variance reduction: CART.



Cây quyết định – Ví dụ minh họa

– Kết luận

+ Hình vẽ



Cây quyết định – Ví dụ minh họa

+ Rút luật

- Luật 1: Nếu Hình dáng = Cầu thì Quyết định = Mua.
- Luật 2: Nếu Hình dáng = Hộp thì Quyết định = Mua.
- Luật 3: Nếu Hình dáng = Trụ và Màu = Xanh thì Quyết định=Không Mua.
- Luật 4: Nếu Hình dáng = Trụ và Màu = Đỏ thì Quyết định = Mua.
- Luật 5: Nếu Hình dáng = Nón thì Quyết định = Không Mua.

THỰC HÀNH
PHƯƠNG PHÁP CÂY QUYẾT ĐỊNH

1. TS. Nguyễn Tấn Trần Minh Khang
2. ThS. Võ Duy Nguyên
3. Cao học. Nguyễn Hoàn Mỹ
4. Tình nguyện viên. Lê Ngọc Huy
5. Tình nguyện viên. Cao Bá Kiệt

DATASET

Dataset

- Tên tập dữ liệu: Social Network Ads.
- Nguồn: https://www.superdatascience.com/pages/machine-learning.
- Tập dữ liệu cho biết các thông tin của khách hàng và họ có mua hàng hay không.

Dataset

- Tập dữ liệu chứa 400 điểm dữ liệu, mỗi điểm dữ liệu có 5 thuộc tính gồm:
  + UserID: Mã số định danh của người dùng.
  + Gender: Giới tính của người dùng.
  + Age: Độ tuổi người dùng.
  + Estimated Salary: Mức lương ước đoán của người dùng.
  + Purchased: Là một trong hai số 0 và 1. Số 0 cho biết khách hàng không mua hàng và số 1 cho biết khách hàng có mua hàng.

Dataset

– Dưới đây là 5 điểm dữ liệu ngẫu nhiên trong tập dữ liệu.

UserID	Gender	Age	Estimated Salary	Purchased
15624510	Male	19	19,000	0
15810944	Male	35	20,000	1
15668575	Female	26	43,000	0
15603246	Female	27	57,000	0
15804002	Male	19	76,000	1

Dataset

- Bài toán: Yêu cầu dựa vào 2 thuộc tính:
  - + Độ tuổi (Age).
  - + Mức lương ước đoán (Estimated Salary).

Dự đoán khách hàng sẽ mua hàng hay không?

TIỀN XỬ LÝ DỮ LIỆU

Tiền xử lý dữ liệu

– Ở bài này, ta chỉ quan tâm đến hai thuộc tính tuổi và mức lương ước đoán.

```
1. import pandas as pd
2. import numpy as np
3. dataset = pd.read_csv("Social_Network_Ads.csv")
4. X = dataset.iloc[:, [2, 3]].values
5. Y = dataset.iloc[:, 4].values
```

Tiền xử lý dữ liệu

- Để thuận tiện cho trực quan hóa kết quả sau khi huấn luyện, ta chuẩn hóa dữ liệu về dạng:
  + Kỳ vọng bằng 0.
  + Phương sai bằng 1
- Lớp `StandardScaler` trong module `sklearn.preprocessing` đã được xây dựng sẵn để chuẩn hóa dữ liệu.

```
7. from sklearn.preprocessing import StandardScaler
8. SC = StandardScaler()
9. X = SC.fit_transform(X)
```

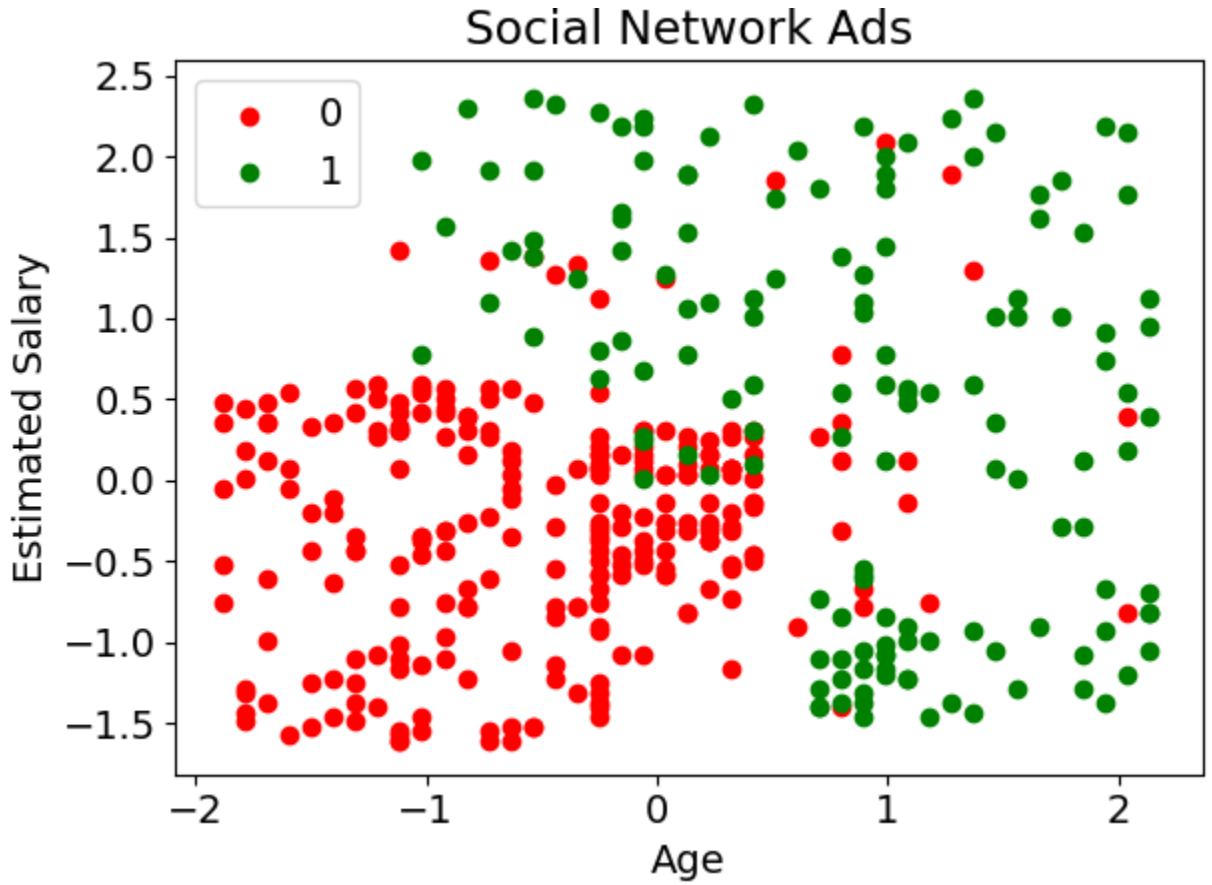
Tiền xử lý dữ liệu

- Chia dữ liệu thành hai tập training set và test set.
- Ta dùng hàm `train_test_split` được cung cấp trong module `sklearn.model_selection`.

```
10.from sklearn.model_selection import train_test_split
11.X_train, X_test, Y_train, Y_test = train_test_split(
   X, Y, train_size = 0.8, random_state = 0)
```

TRỰC QUAN HÓA DỮ LIỆU

Trực quan hóa dữ liệu



Trực quan hóa dữ liệu

– Xây dựng hàm trực quan hóa các điểm dữ liệu.

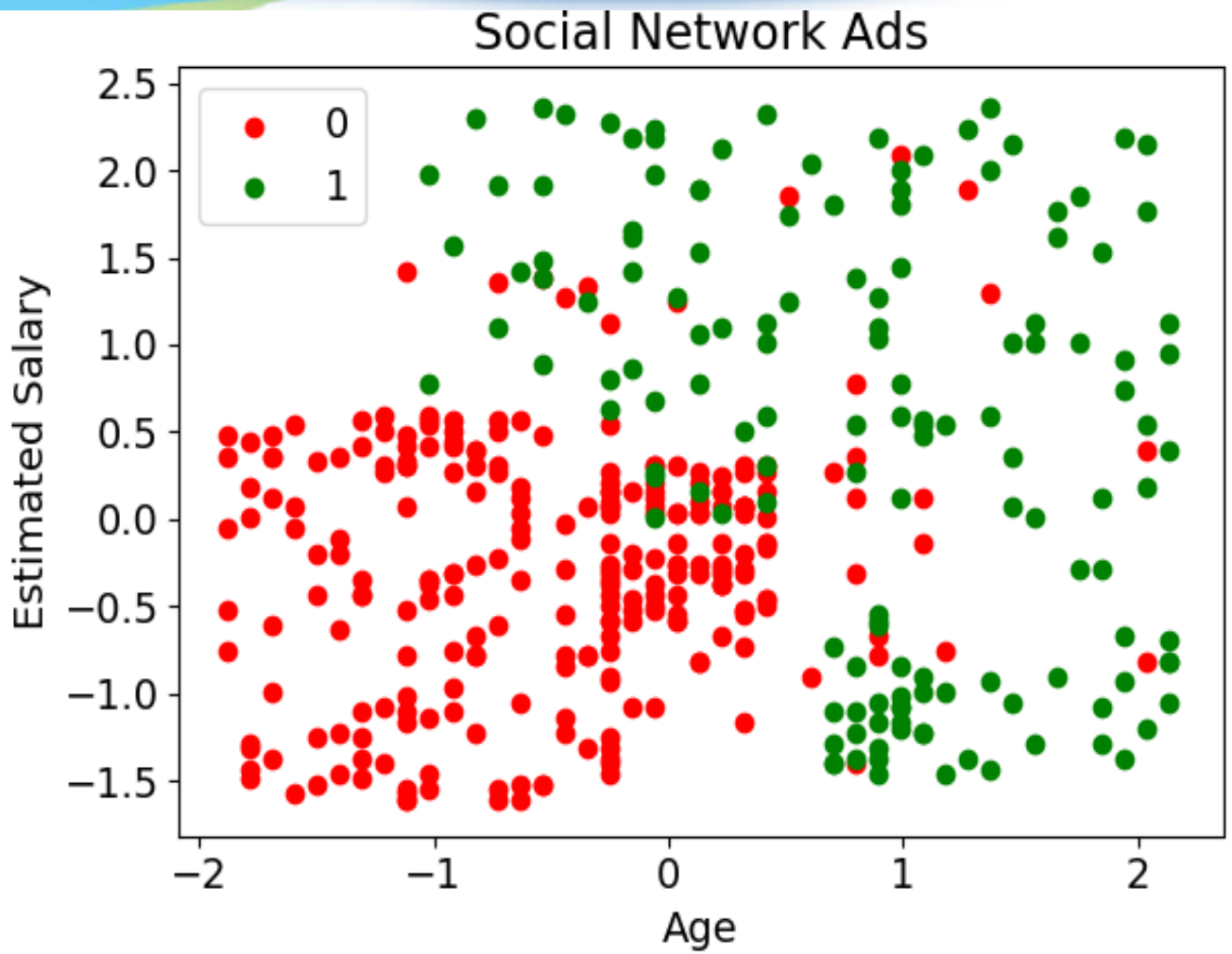
```
11.from matplotlib.colors import ListedColormap
12.import matplotlib.pyplot as plt
13.def VisualizingDataset(X_, Y_):
14.    X1 = X_[:, 0]
15.    X2 = X_[:, 1]
16.    for i, label in enumerate(np.unique(Y_)):
17.        plt.scatter(X1[Y_ == label], X2[Y_ == label]
    ,
```

Trực quan hóa dữ liệu

– Gọi hàm trực quan hóa dữ liệu.

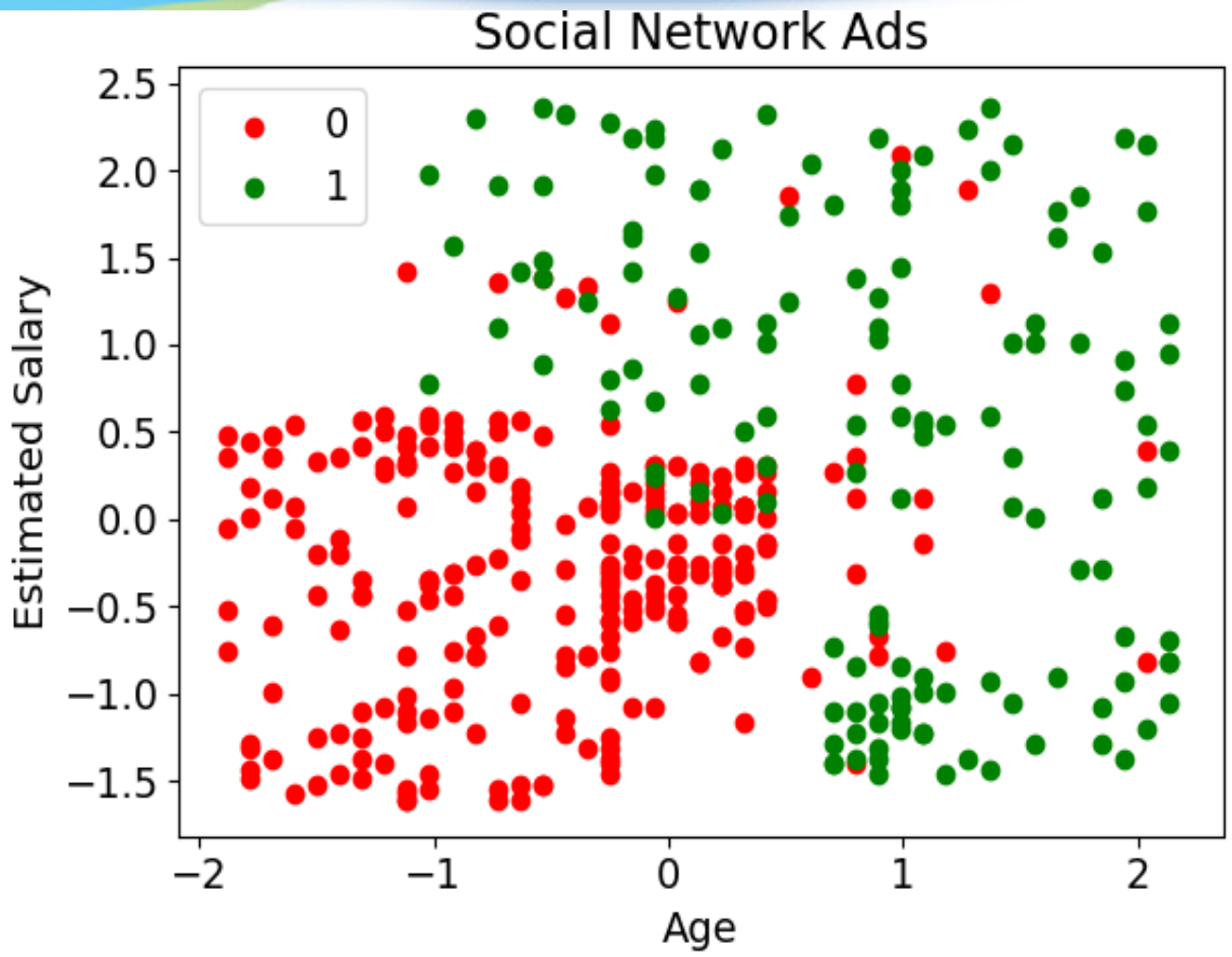
```
18.VisualizingDataset(X, Y)
19.plt.show()
```

Trực quan hóa dữ liệu



- Theo hình vẽ, ta thấy các điểm có sự phân bố thành 2 mảng.
  - + Mảng dưới trái phần lớn có màu đỏ, tức khách hàng không mua hàng.
  - + Mảng bên phải và mảng bên trên phần lớn có màu xanh, tức khách hàng có mua hàng.

Trực quan hóa dữ liệu



- Điều này là phù hợp vì các khách hàng trẻ và có mức lương thấp sẽ thường không mua hàng.
- Ngược lại, khách hàng cao tuổi hoặc có lương cao sẽ thường mua hàng nhiều hơn.

HUẤN LUYỆN MÔ HÌNH

Huấn luyện mô hình

- Ta sử dụng lớp `DecisionTreeClassifier` trong module `sklearn.tree` để huấn luyện mô hình.
- Đặt tham số `criterion` (tham số mô tả thuật toán tạo cây quyết định) là `"entropy"`.

```
20.from sklearn.tree import DecisionTreeClassifier
21.classifier = DecisionTreeClassifier(criterion = "ent
   ropy")
22.classifier.fit(X_train, Y_train)
```

TRỰC QUAN HÓA KẾT QUẢ MÔ HÌNH

Trực quan hóa kết quả mô hình

- Ta tạo một *confusion matrix*. Đây là một ma trận có kích thước là $p \times p$ với p là số phân lớp trong bài toán đang xét, ở đây là 2.
- Phần tử ở dòng thứ i, cột thứ j của confusion matrix biểu thị số lượng phần tử có loại là i và được phân vào loại j.
- Hàm `confusion_matrix` trong module `sklearn.metrics` sẽ hỗ trợ ta xây dựng confusion matrix.

```
23.from sklearn.metrics import confusion_matrix
24.cm = confusion_matrix(Y_train, classifier.predict(X_
   train))
25.print(cm)
```

Trực quan hóa kết quả mô hình

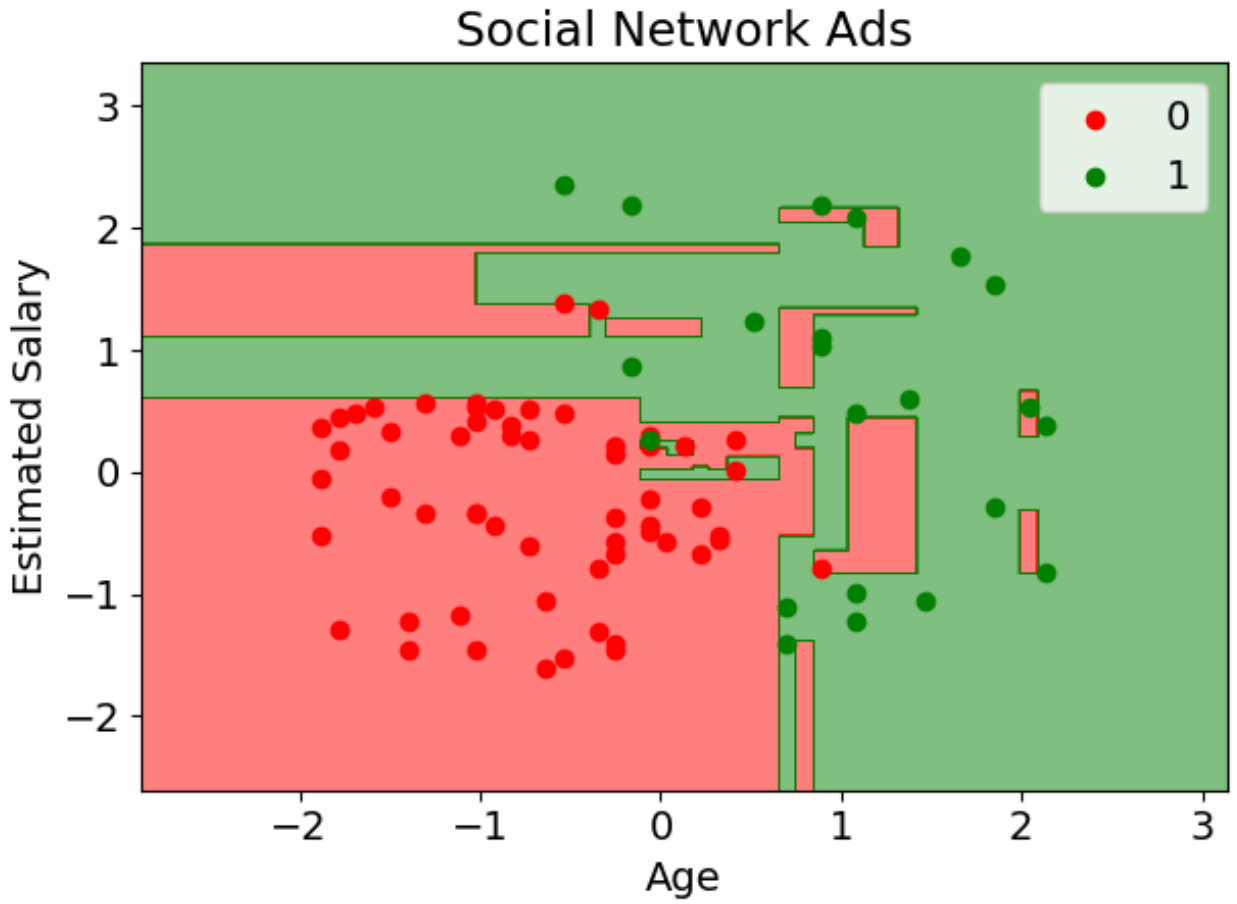
- Confusion Matrix được in ra là:

	0	1
0	199	0
1	1	120

- Theo ma trận trên, số lượng dữ liệu được phân loại đúng là $199 + 120 = 319$ điểm dữ liệu.
- Số lượng dữ liệu phân loại sai là 1 điểm dữ liệu.
- Tỉ lệ phân loại sai là $1/320 \approx 0.003$.

Trực quan hóa kết quả mô hình

- Ta trực quan hóa kết quả mô hình trên mặt phẳng tọa độ bằng cách vẽ các vùng phân chia mà mô hình thu được sau quá trình huấn luyện.



Trực quan hóa kết quả mô hình

– Xây dựng hàm trực quan hóa kết quả bằng cách tạo 2 vùng phân chia mà mô hình đạt được.

```
26.def VisualizingResult(model, X_):
27.    X1 = X_[:, 0]
28.    X2 = X_[:, 1]
29.    X1_range = np.arange(start= X1.min()-1, stop= X1.max()+1,
   step = 0.01)
30.    X2_range = np.arange(start= X2.min()-1, stop= X2.max()+1,
   step = 0.01)
31.    X1_matrix, X2_matrix = np.meshgrid(X1_range, X2_range)
```

Trực quan hóa kết quả mô hình

– Xây dựng hàm trực quan hóa kết quả bằng cách tạo 2 vùng phân chia mà mô hình đạt được.

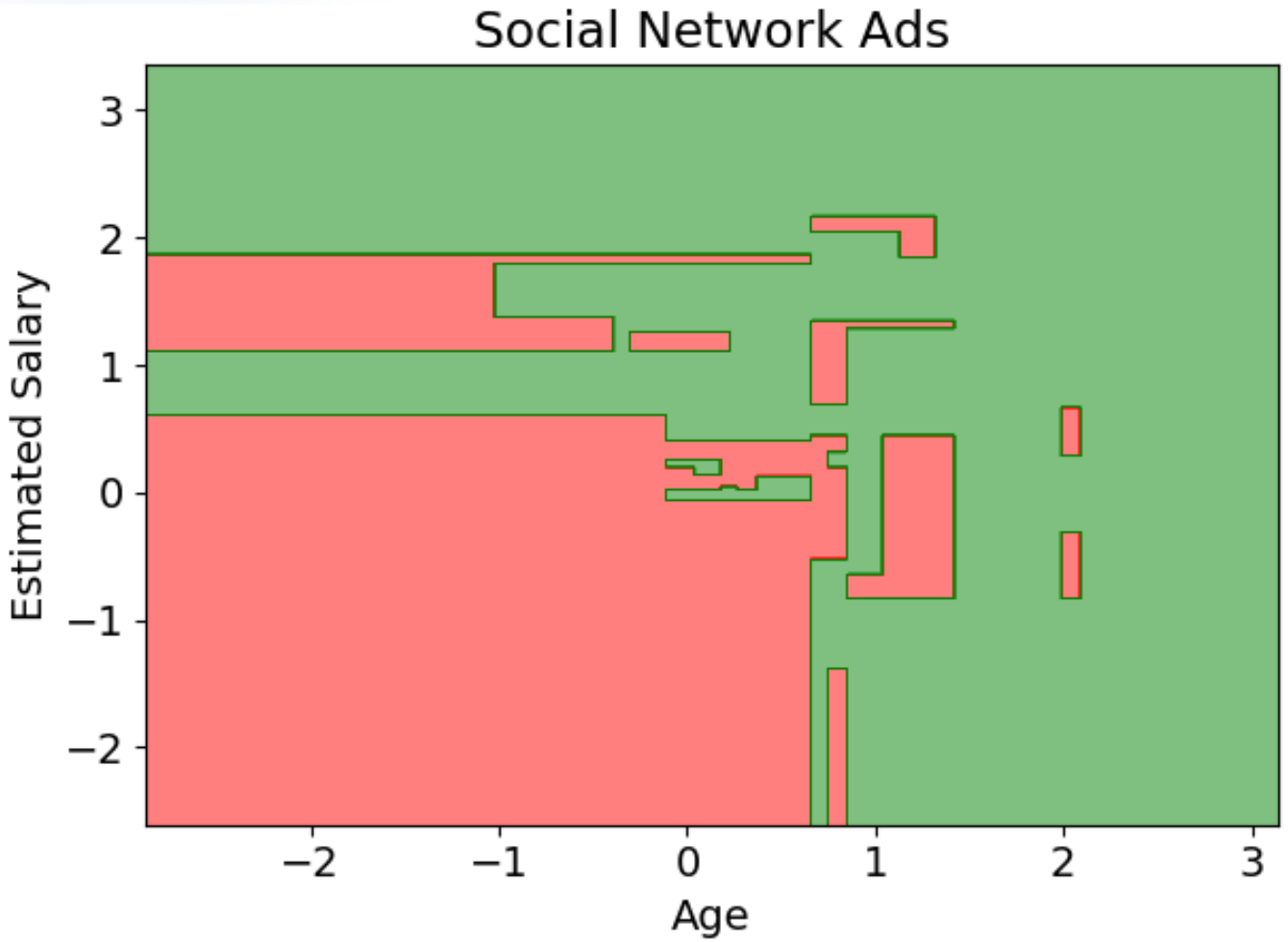
```
26.def VisualizingResult(model, X_):
31.     ...
32.     X_grid= np.array([X1_matrix.ravel(), X2_matrix.ravel()]).T
33.     Y_grid= model.predict(X_grid).reshape(X1_matrix.shape)
34.     plt.contourf(X1_matrix, X2_matrix, Y_grid, alpha = 0.5,
    cmap = ListedColormap(("red", "green")))
```

Trực quan hóa kết quả mô hình

– Trực quan hóa kết quả mô hình.

```
35.VisualizingResult(classifier, X_train)
36.plt.show()
```

Trực quan hóa kết quả mô hình

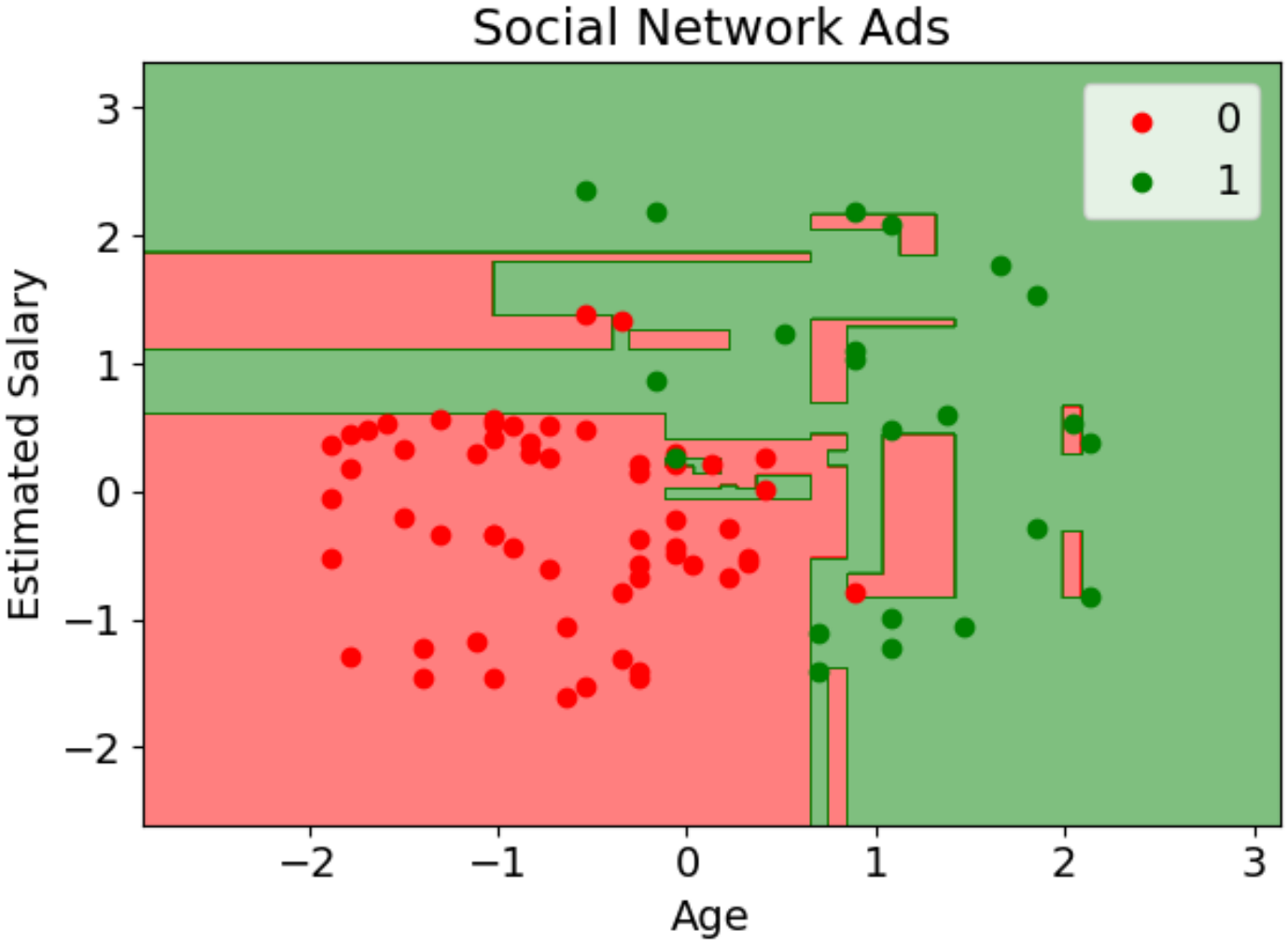


Trực quan hóa kết quả mô hình

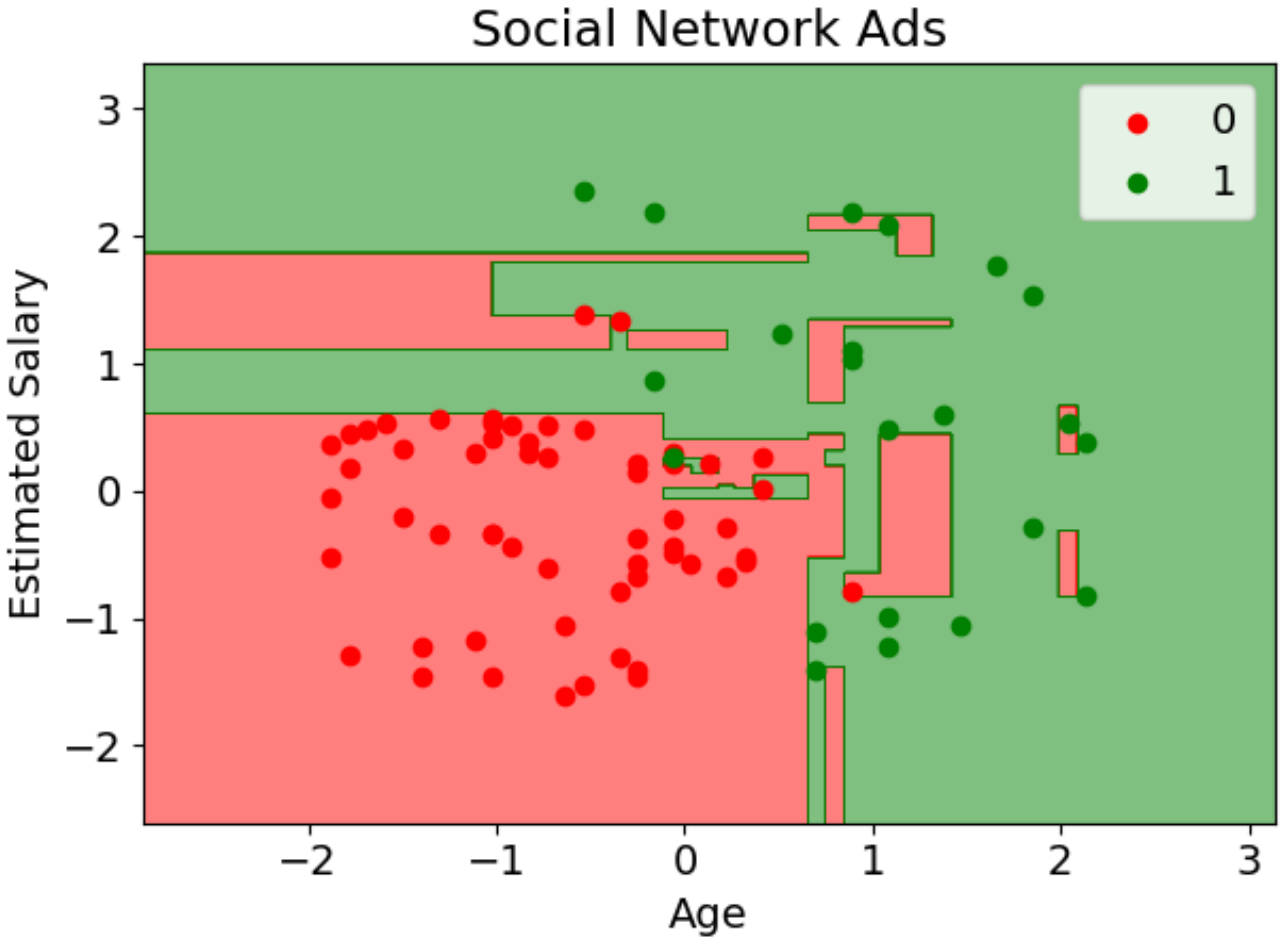
- Hoàn thiện quá trình trực quan bằng cách vẽ thêm các điểm dữ liệu huấn luyện lên mặt phẳng tọa độ.

```
37.VisualizingResult(classifier, X_train)
38.VisualizingDataset(X_train, Y_train)
39.plt.show()
```

Trực quan hóa kết quả mô hình



Trực quan hóa kết quả mô hình



- Nhận xét:
  - + Mô hình có độ chính xác cao.
  - + Tuy nhiên nó đang cố gắng mô tả các điểm dữ liệu nhiễu.

KIỂM TRA KẾT QUẢ TRÊN TẬP TEST

Kiểm tra kết quả trên tập test

– Tạo *confusion matrix* trên tập test.

```
40.cm = confusion_matrix(Y_test, classifier.predict(X_t
   est))
41.print(cm)
```

Kiểm tra kết quả trên tập test

- Confusion Matrix được in ra là:

	0	1
0	53	5
1	3	19

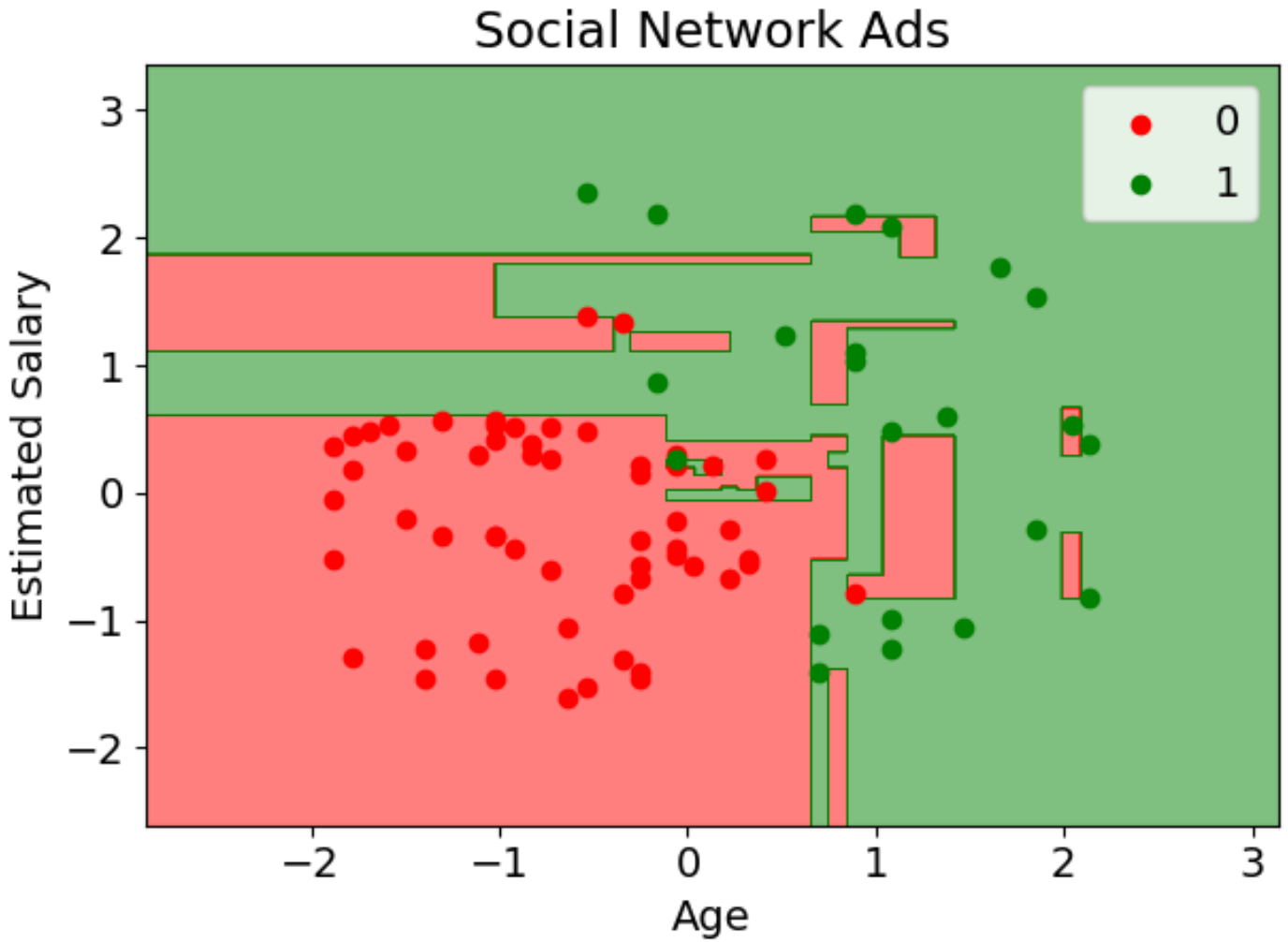
- Theo ma trận trên, số lượng dữ liệu được phân loại đúng là $53 + 19 = 72$ điểm dữ liệu.
- Số lượng dữ liệu phân loại sai là $3 + 5 = 8$ điểm dữ liệu.
- Tỉ lệ điểm dữ liệu phân loại sai là $8/80 = 0.1$.

Kiểm tra kết quả trên tập test

– Thực hiện tương tự trực quan hóa kết quả mô hình trên tập traning.

```
42.VisualizingResult(classifier, X_test)
43.VisualizingDataset(X_test, Y_test)
44.plt.show()
```

Kiểm tra kết quả trên tập test



	0	1
0	53	5
1	3	19

Kiểm tra kết quả trên tập test

- Xây dựng hàm so sánh kết quả trên một điểm dữ liệu trong tập test.

```
45.def compare(i_example):
46.     x = X_test[i_example : i_example + 1]
47.     y = Y_test[i_example]
48.     y_pred = classifier.predict(x)
49.     x_inv = SC.inverse_transform(x)
50.     print(x_inv, y, y_pred)
```

Kiểm tra kết quả trên tập test

– Gọi thực hiện hàm so sánh trên 5 điểm dữ liệu, có chỉ mục từ thứ 7 đến 11 trong tập kiểm thử.

```
51.for i in range(7, 12):
52.     compare(i)
```

Kiểm tra kết quả trên tập test

Age	Estimated Salary	Purchased	Predicted Purchased
36	144,000	1	1
18	68,000	0	0
47	43,000	0	0
30	49,000	0	0
28	53,000	0	0

Chúc các bạn học tốt

Thân ái chào tạm biệt các bạn

ĐẠI HỌC QUỐC GIA TP.HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN TP.HCM

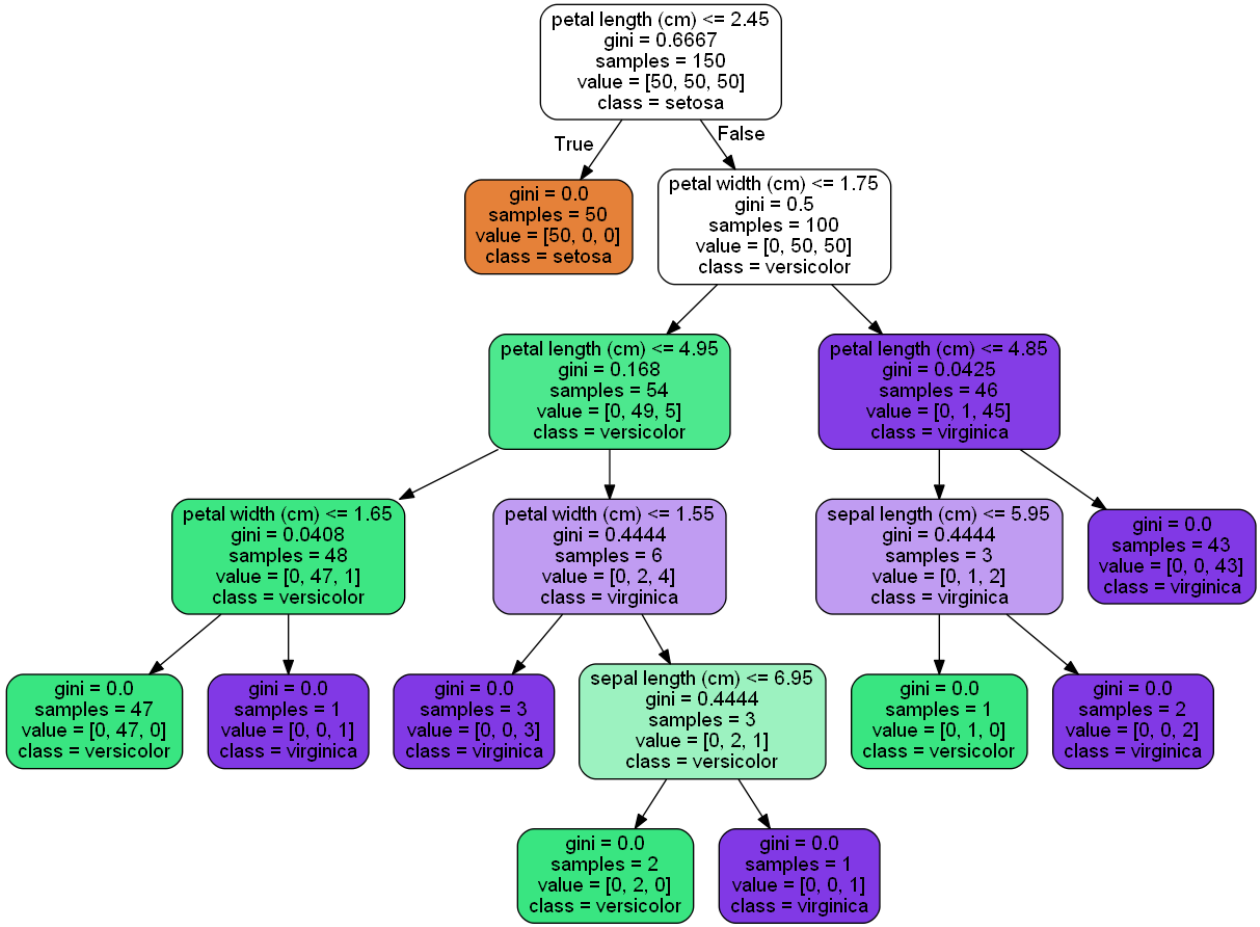
TOÀN DIỆN – SÁNG TẠO – PHỤNG SỰ

Programming Exercise

BÀI TẬP THỰC HÀNH

Bài tập thực hành trên Python

- Bài 01. Iris Data Set
  - + Xây dựng cây quyết định cho tập dữ liệu: Iris Data Set.
  - + Nguồn dữ liệu: https://archive.ics.uci.edu/ml/datasets/iris.
  - + Hình ảnh minh họa: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html



Chúc các bạn học tốt

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN TP.HCM

Nhóm UIT-Together

Nguyễn Tấn Trần Minh Khang