**Mô hình dự đoán chất lượng nước sử dụng thuật toán cây quyết định (Decision tree algorithm)**

Thuật toán cây quyết định là một phướng pháp học có giám sát được sử dụng trong thống kê, khai thác dữ liệu và học máy. Có thể được dùng để giải quyết bài toán hồi quy (Regression) và phân loại (Classification)

Một thuật toán học máy thường sẽ có 2 bước:

1. Huấn luyện: Từ dữ liệu huấn luyện thuật toán sẽ học và tạo ra mô hình.
2. Dự đoán: Dùng mô hình học được từ bước trên dự đoán ra các giá trị mới.

Thuật toán cho cây quyết định:

- Cây quyết định được xây dựng từ tập dữ liệu huấn luyện bằng cách phân chia và chinh phục đệ quy từ trên xuống.

- Khi bắt đầu, tất cả các ví dụ đào tạo đều ở gốc.

- Các thuộc tính được phân loại (nếu có giá trị liên tục, chúng dược phân chia rời rạc trước).

- Các ví dụ được phân vùng đệ quy dựa trên các tuộc tính đã chọn.

- Các thuộc tính kiểm tra được chọn trên cơ sở đo lường thống kê hoặc kinh nghiệm (thu thập thông tin - information gain)

- Dừng phân vùng khi tất cả các mẫu cho một nút nhất định thuộc cùng một lớp hoặc không còn thuộc tính nào để phân vùng hoặc không còn mẫu nào

- Để lựa chọn tuộc tính phân vùng tốt cho mô hình cần sử dụng các kỹ thuật tính toán lựa chọn thuộc tính. VD như Information Gain, Gini Index.

Ưu điểm:

- Mô hìn đơn giản và dễ hiệu, tạo ra bộ luật với mỗi nhánh là là một luật của cây.

- Dữ liệu đầu vào có thể là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả.

- Có thể làm việc với dữ liệu số và dữ liệu phân loại.

- Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê.

- Có khả năng làm việc với dữ liệu lớn.

Nhược điểm:

- Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu huấn luyện. Với một sự thay đổi nhỏ trong bộ dữ liệu cấu trúc cây có thể thay đổi hoàn toàn.

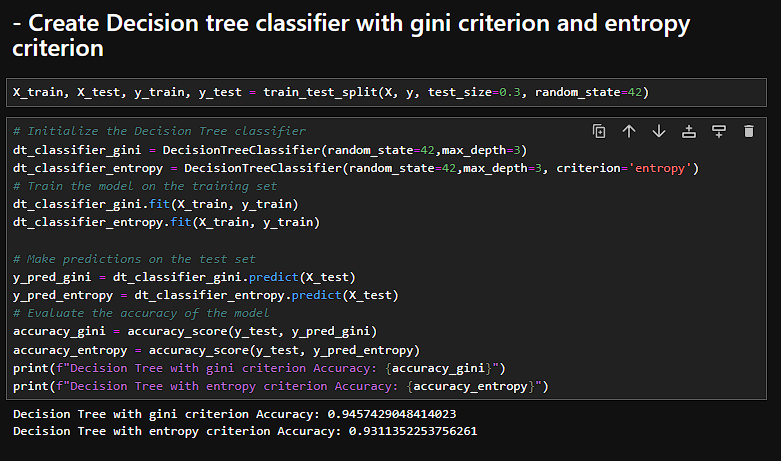
- Cây quyết định hay gặp ván đề overfitting.

**Thực hiện với bài toán dự đoán chất lượng nước**

1. Khởi tạo và huấn luyện:

Khởi tạo mô hình dự đoán với thuật toán Cây quyết định với kỹ thuật lựa chọn thuộc tính Information gain tiêu chuẩn entropy và kỹ thuật Gini impurity tiêu chuẩn gini.

Huấn luyện với tập dữ liệu huấn luyện và dự đoán với tập dữ liệu kiểm thử từ đó cho ra được độ chính xác.



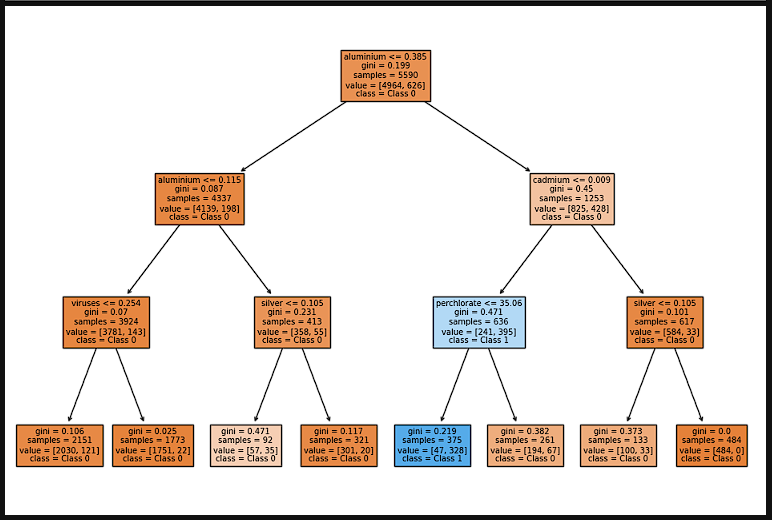
Với câu hình độ sâu của cây = 3:

- Độ chính xác của mô hình cây quyết định (Information gain) = 94%

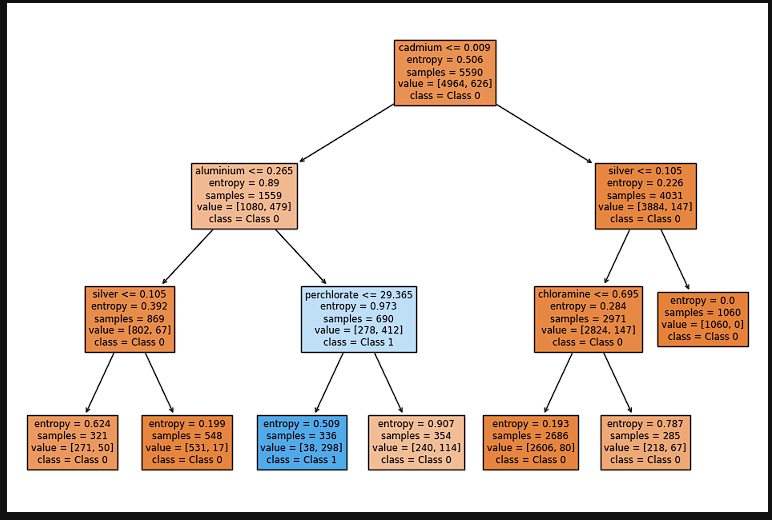
- Độ chính xác của mô hình cây quyết định (Gini impurity) = 93%

Sau khi khởi chạy với các độ sâu khác nhau từ khoảng 3 đến 20 thu được kêt quả với độ sâu phù hợp nhất là 3 cho cả 2 mô hình trên. Độ sâu không quá sâu nhưng độ chính xác vẫn rất tốt.

Hình ảnh cây quyết định (Information gain):



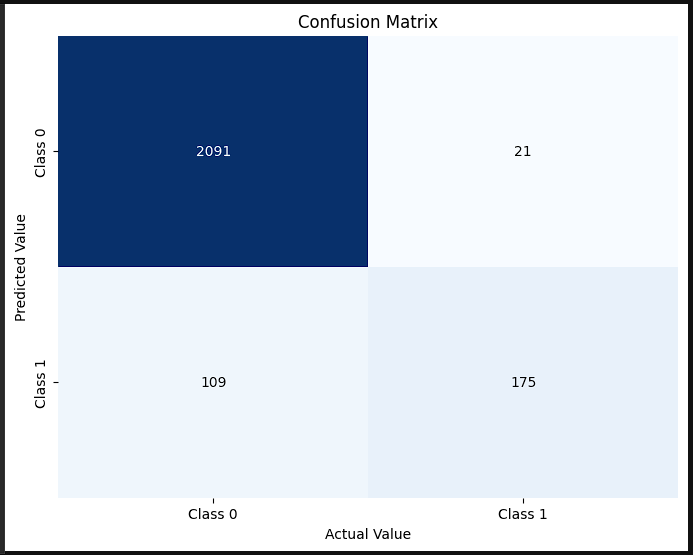
Hình ảnh cây quyết định (Gini impurity):

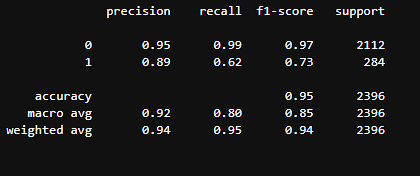


1. Đánh giá hiệu suất của mô hình

1. Ma trận nhận lẫn

Ma trận của cây quyết định (Information Gain):

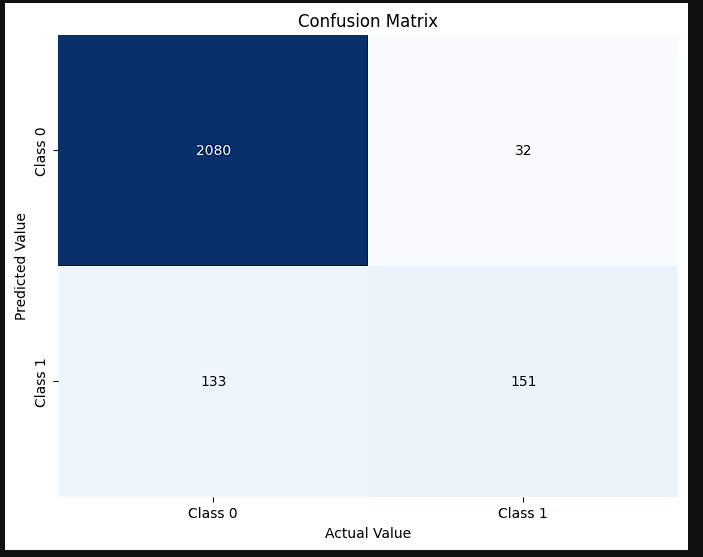


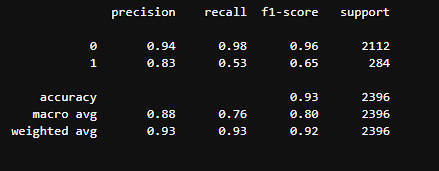


- Độ chính xác: 95%

- Điểm f1(đánh giá tính đầy đủ (recall) và độ chính xác gán nhãn tích cực (precision) cùng lúc): nhãn 0 = 0.97; nhãn 1 = 0.73

Ma trận của cây quyết định (Gini impurity)





- Độ chính xác: 93%

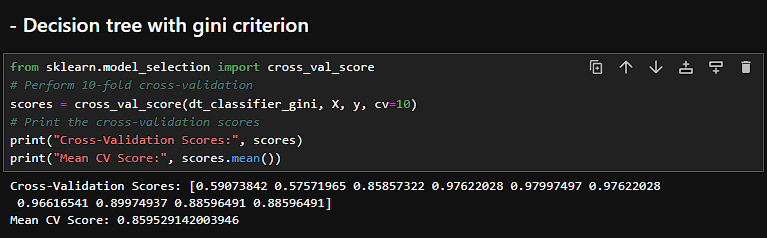
- Điểm f1(đánh giá tính đầy đủ (recall) và độ chính xác gán nhãn tích cực (precision) cùng lúc): nhãn 0 = 0.96; nhãn 1 = 0.65

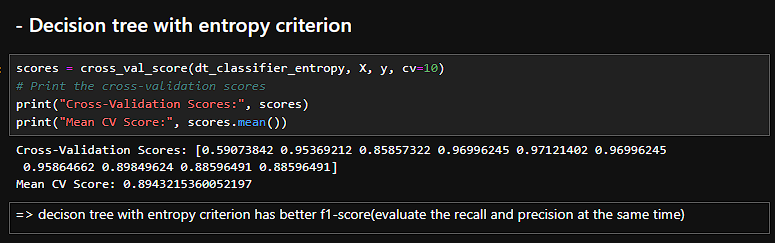
Đánh giá chung:

- 2 mộ hình đều cho ra hiệu suất khá tốt. Với tỷ lệ % bộ giá trị mà bộ phân loại có nhãn tích cực cao

- tính đầy đủ của nhãn 0 ở 2 mô hình đều cao với tỷ lệ hơn 90%. Với nhãn 1 trên 50%

1. Thẩm địch chéo





- Với kỹ thuật thẩm địch chéo kết quả kiểm tra hiệu suất 2 mô hình không chênh lệch nhau nhiều.

- Với Decision tree (Information gain) = 86% và Decision tree (Gini inpurity) = 89%