1. Mô tả cách thực hiện classification sử dụng decision trees.

Decision trees thực hiện bài toán classification bằng cách xây dựng ra một tập dữ liệu, model được tổ chức dạng cấu trúc dữ liệu cây nhị phân. Trong cây nhị phân đó có nhiều node, mỗi node có các dữ liệu như tổng số sample ở node đó, có điều kiện của threshold cứ mỗi sample qua 1 node là sẽ kiểm tra có thỏa điều kiện của threshold đó không nếu có thì đi theo chiều đúng, sai thì đi theo chiều sai. Quá trình cứ lặp lại đến khi gặp nút lá với label nào thì label đó chính là kết quả prediction của mình.

1. Mô tả cách thực hiện regression sử dụng decision trees.

Decision trees thực hiện bài toán regression bằng cách xây dựng ra một tập dữ liệu, model được tổ chức dạng cấu trúc dữ liệu cây nhị phân. Trong cây nhị phân đó có nhiều node, mỗi node có các dữ liệu như tổng số sample có trong khoảng nhỏ hơn mốc threshold đó, giá trị trung bình của tất cả sample trong khoảng đó, có điều kiện của threshold cứ mỗi sample qua 1 node là sẽ kiểm tra có thỏa điều kiện của threshold đó không nếu có thì đi theo chiều đúng, sai thì đi theo chiều sai. Lặp lại đến khi gặp nút lá và giá trị trung bình ở nút đó sẽ là giá trị prediction của mình.

1. Nêu cách thực hiện regularization với decision tree.

Cách regurlarization decision tree bằng cách giảm độ sâu cây, khống chế số lượng sample leaf,…

1. Nếu một decision tree bị overfitting thì ta nên tăng hay giảm max\_depth của nó? Tại sao?

Nếu một decision tree bị overfitting thì nên giảm độ sâu cây vì độ sâu cây càng lớn thì càng dễ gây overfitting với mỗi sample trong tập là một nút trong cây. Còn giảm độ sâu thì đồng nghĩa với việc giảm số nút trong cây thì model sẽ đơn giản hơn sẽ ngăn được overfitting.

1. Nếu một decision tree bị underfitting, thì ta có nên thực hiện feature scaling? Tại sao?

Nếu decision tree bị underfitting thì ta chỉ nên điều chỉnh các tham số như tăng độ sâu cây, … Và feature scaling không ảnh hưởng hay liên quan đến việc chạy thuật toán. Vì decision tree chỉ kiểm tra giá trị đó có lớn hơn hay bé hơn điều kiện của threshold thì feature scaling không gây ảnh hưởng dù có hay không.