A screen shot of a computer code

Description automatically generated

Chúng ta tạo một biểu đồ, trong đó trục tung là các biến quan trọng trong quá trình dự đoán còn trục hoành là thang đo độ quan trọng của chúng. Có thể thấy, chất lượng chung của căn nhà (OverallQual) là biến quan trọng nhất; gara ô tô, chất lượng ngoại thất hay hàng xóm là nhóm biến quan trọng tiếp theo; còn diện tích đất, diện tích gara, chất lượng móng căn nhà, thời gian xây dựng nhà, chất lượng căn bếp hay tổng diện tích dưới lòng đất không quá quan trọng.

A graph with numbers and a bar

Description automatically generated

\* Tải dữ liệu kiểm tra:

A computer screen with text and images

Description automatically generated

Mô hình RandomForest (rf) được sử dụng để dự đoán giá nhà trên dữ liệu kiểm tra. Một dataframe mới được tạo ra với 2 cột là ‘Id’ là các giá trị đã loại bỏ từ dữ liệu kiểm tra và 'SalePrice' là các giá trị dự đoán được tính từ mô hình RandomForest. Sau khoảng hơn 1 giây, 5 dòng đầu của dataframe đó được thể hiện, cho thấy mức giá dự đoán của đối với từng căn nhà với những biến số khác nhau.

A screenshot of a computer

Description automatically generated

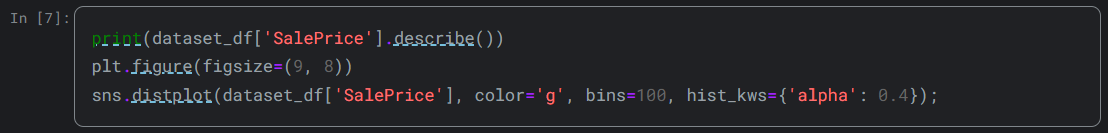
CÁC THAM SỐ VÀ MÔI TRƯỜNG CÀI ĐẶT

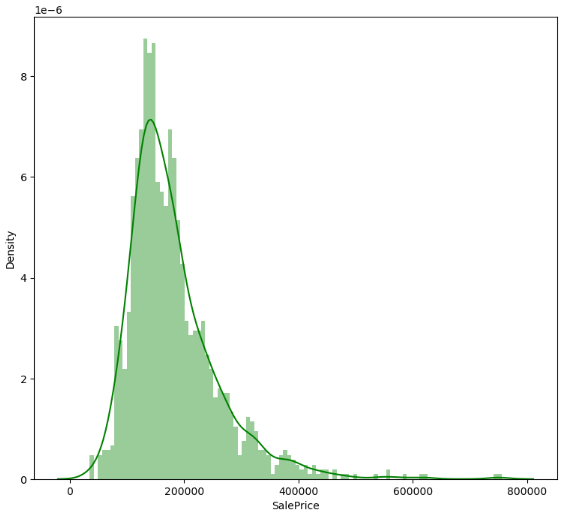
Môi trường cài đặt cho bài toán "House price prediction" thường bao gồm các công cụ và thư viện phổ biến được sử dụng cho xử lý dữ liệu và khai thác quy tắc kết hợp từ tập dữ liệu mua sắm. Dưới đây là các thành phần cần thiết cho bài toán này:

Ngôn ngữ lập trình: Nhóm sử dụng ngôn ngữ Python vì ngôn ngữ này cung cấp các thư viện mạnh mẽ cho xử lý dữ liệu và phân tích.

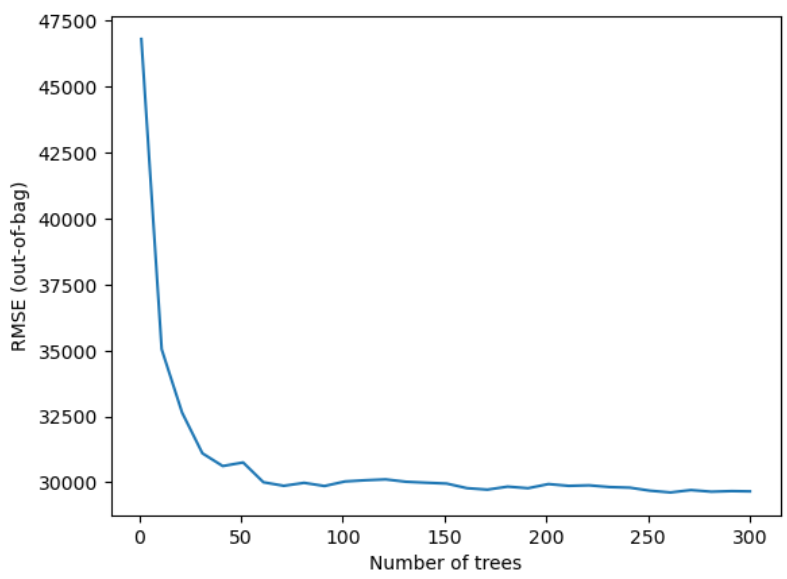
Nhóm tải các thư viện cần thiết cho việc phân tích dữ liệu và trực quan hóa dữ liệu bằng các biểu đồ như: pandas, seaborn, matplotlib, tensorflow

Thư viện Seaborn là một thư viện Python dùng để trực quan hóa dữ liệu dựa trên Matplotlib. Nó cung cấp một giao diện cấp cao để tạo ra các biểu đồ thống kê và trực quan hóa dữ liệu một cách dễ dàng và hiệu quả, bao gồm biểu đồ đường, biểu đồ cột, biểu đồ tròn,… Trong bài này, thư viện seaborn được cung cấp để vẽ biểu đồ displot.





Thư viện Matplotlib được cung cấp để tạo biểu đồ đường:



Các tham số được sử dụng trong bài toán “House price prediction”

Tập dữ liệu (Dataset): Đây là file dữ liệu gồm 80 trường dữ liệu khác nhau, mỗi trường thể hiện một thông số của căn nhà như: chất lượng căn nhà, diện tích nhà, phố, số lượng phòng,...

Biến quan trọng: Thể hiện các yếu tố khác nhau với mức độ quan trọng khác nhau, biến càng quan trọng thì mức độ tác động đến mô hình của nó càng lớn.

PHƯƠNG PHÁP CƠ SỞ

Đoạn code sử dụng quyết định rừng (Random Forest), một thuật toán học máy để dự đoán giá nhà. Random Forest  hoạt động dựa trên việc kết hợp nhiều cây quyết định (decision tree) đơn giản.

Mỗi cây quyết định được huấn luyện trên một tập hợp con (subset) của dữ liệu và đưa ra dự đoán độc lập. Dự đoán cuối cùng của mô hình là tổng hợp (aggregation) của các dự đoán từ các cây quyết định thành phần.

Random Forest có thể được coi là một phương pháp học tập tổng hợp (ensemble learning method) vì nó kết hợp nhiều mô hình yếu (individual weak models) để tạo ra một mô hình mạnh hơn.

Một số kỹ thuật để cải thiện hiệu quả của mô hình Random Forest:

* Sử dụng các kỹ thuật chọn lọc đặc trưng (feature selection) để chọn ra các đặc trưng quan trọng nhất cho mô hình.
* Sử dụng các kỹ thuật tăng cường dữ liệu (data augmentation) để tăng kích thước tập dữ liệu và giảm thiểu nguy cơ quá khớp (overfitting)
* Sử dụng các kỹ thuật ensemble để kết hợp nhiều mô hình Random Forest lại với nhau để tạo ra một mô hình mạnh mẽ hơn.