Đề thi:

DATA PRE-PROCESSING AND DATA ANALYSIS

Ngày thi : 3/1/2026

**** Học viên tạo 1 thư mục là **DL04_HoVaTen**, lưu tất cả bài làm vào để nộp chấm điểm ****

**** Học viên được sử dụng tài liệu ****

**** Với mỗi câu, sử dụng Markdown để mô tả yêu cầu ****

Phần 1 : Đọc tập tin dữ liệu dự báo giá nhà (1đ)

1. Đọc tập tin dữ liệu housing-prices-dataset.csv
2. Xem thông tin sơ bộ : shape/head/tail/info …
3. Kiểm tra dữ liệu bị trùng và xử lý (nếu có)

Phần 2 : EDA (4.5đ)

1. Chọn các biến sau đây để phân tích : 'LotShape', 'Street', 'HouseStyle', 'LotArea', 'YearBuilt', '1stFlrSF', '2ndFlrSF', 'FullBath', 'BedroomAbvGr', 'TotRmsAbvGrd', 'SalePrice' (**Chú ý : biến SalePrice là biến output**)
2. Trong các biến trên hãy xác định các biến định tính và các biến định lượng. In thông tin các biến này như số lượng giá trị duy nhất và các giá trị duy nhất
3. Kiểm tra dữ liệu bị thiếu và xử lý (nếu có)
4. Thực hiện thống kê mô tả (describe) cho các biến trên và nêu nhận xét

Phân tích tất cả các biến được chọn trong câu 1

5. Phân tích 1 biến (cho nhận xét)
6. Phân tích 2 biến (cho nhận xét)
7. Kiểm tra và xóa các outlier (nếu có)

Phần 3 : Feature Engineering (2đ)

1. Chọn ra các biến định lượng input có tương quan với biến ouput (xét hệ số tương quan >=0.3 hoặc hệ số tương quan <=-0.3)
2. Chọn ra các biến định tính input có tương quan với biến ouput (xét p-value<=0.05)
3. Chuẩn hóa các biến định tính input bằng one-hot encoder
4. Chuẩn hóa các biến định lượng input bằng StandardScaler

Phần 4 : Tạo mô hình Linear Regression và đánh giá (1.5đ)

Với tập dữ liệu đã chuẩn hóa ở phần 3. Hãy :

1. Xác định các tập X và y
2. Chia tập dữ liệu thành 2 tập train và test (test size : 0.2)
3. Tạo mô hình Linear Regression và huấn luyện với tập train
4. Đánh giá mô hình (score) trong 2 trường hợp : Train, Test. Cho nhận xét

Phần 5 : Cải tiến hiệu suất mô hình (1đ)

1. Với **tất cả các biến định lượng**, sử dụng SelectKBest (sklearn) để chọn ra các feature có score cao nhất. Thực hiện phần 3 và phần 4
2. Giống câu trên nhưng có **chọn thêm các biến định tính nào có mối tương quan với biến SalePrice**

HẾT