HỒI QUY TUYẾN TÍNH ĐƠN

1. Dự đoán điểm thi cuối kỳ
Dataset: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

Input: midterm
Output: final
Tasks:
- Phân tích mối quan hệ giữa 2 điểm
- Xây dựng mô hình dự đoán
- Đánh giá độ chính xác

2. Dự đoán chiều cao dựa vào cân nặng
Dataset: https://www.kaggle.com/datasets/burnoutminer/heights-and-weights-dataset

Input: cân nặng
Output: chiều cao

Tasks:
- Visualize dữ liệu
- Xử lý outliers
- Xây dựng mô hình dự đoán
- Đánh giá độ chính xác

3. Phân tích tiêu thụ nhiên liệu
Dataset: https://archive.ics.uci.edu/dataset/9/auto+mpg
Input: trọng lượng xe
Output: mức tiêu thụ nhiên liệu

4. Dự đoán doanh số bán hàng
Dataset: https://www.kaggle.com/datasets/purbar/advertising-data
Input: chi phí quảng cáo
Output: doanh số
Tasks:
- Phân tích trend
- Xây dựng mô hình
- Cross-validation

5. Dự đoán nhiệt độ
Dataset: https://www.kaggle.com/datasets/muthuj7/weather-dataset

Input: độ ẩm (Humidity)
Output: nhiệt độ (temperature)
Tasks:
- Phân tích mùa vụ
- Xử lý missing values
- So sánh metrics
- Xây dựng mô hình

HỒI QUY TUYẾN TÍNH ĐA

6. Dự đoán giá nhà
Dataset: https://www.kaggle.com/code/ahmedmahmoud16/california-housing-prices

Features: diện tích, số phòng, tuổi nhà, thu nhập khu vực

7. Dự đoán lương
Dataset: https://www.kaggle.com/datasets/rohankayan/years-of-experience-and-salary-dataset

Input features: years_experience (số năm kinh nghiệm), education (trình độ học vấn), job_role (vị trí công việc), location (địa điểm làm việc)
Output: salary
Tasks:

1. Tiền xử lý dữ liệu:
- Label Encoding cho biến categorical
- StandardScaler cho biến numeric
- Chia train/test
2. Phân tích dữ liệu:
- Correlation matrix
- Scatter plot kinh nghiệm vs lương
- Box plot học vấn vs lương
3. Xây dựng mô hình:
- Sử dụng Linear Regression
- Đánh giá bằng R2 và RMSE
- Phân tích Feature Importance
4. Phân tích dự đoán:
- So sánh giá trị thực tế vs dự đoán
- Phân tích phân phối residuals

8. Dự đoán hiệu suất xe
Dataset: https://www.kaggle.com/datasets/uciml/autompg-dataset

Features gốc:

- mpg (miles per gallon - mức tiêu thụ nhiên liệu)
- cylinders (số xi-lanh)
- displacement (dung tích động cơ)
- horsepower (mã lực)
- weight (trọng lượng)
- acceleration (gia tốc)
- year (năm sản xuất)
- origin (xuất xứ)

Features tạo thêm:

- power_to_weight (tỷ lệ mã lực / trọng lượng)
- displacement_per_cylinder (dung tích trên mỗi xi-lanh)

Tasks:

1. Xử lý dữ liệu:
- Loại bỏ missing values
- Chuyển đổi kiểu dữ liệu
- Tạo features mới
2. Phân tích features:
- Vẽ correlation matrix
- Pair plots cho features chính
- Box plots cho biến categorical
3. So sánh 3 mô hình:
- Linear Regression (tuyến tính)
- Polynomial Regression (bậc 2)
- Ridge Regression (có regularization)

9. Dự đoán giá cổ phiếu
Dataset: https://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset

features: open (giá mở cửa), high(giá cao nhất), low(giá thấp nhất), volume (khối lượng giao dịch)

features mới

- MA5, MA20 (trung bình động 5 và 20 ngày)
- Price_change (thay đổi giá)
- Volume_change (thay đổi khối lượng)
- Volumn_MA5 (trung bình khối lượng 5 ngày)

10. dự đoán chi phí y tế
https://www.kaggle.com/code/mragpavank/medical-cost-personal-datasets

features: tuổi, BMI, số con, hút thuốc