Onior: có thứ bậc feature

Nomino: ko có thứ bậc feature

Unique(): lấy giá trị ko bị trùng

Phần nếu dùng của imputer và standardscaler thì sẽ dùng pipeline(steps=[(“imputer”,simpleimputer(missing\_values=-1,strategy=”median”)),(“scaler”,Standardscaler())])

Education=[‘some high school’,’high school’,’some college’,”asocite degree’,’bachelor degree’]

Pipeline(steps=[(“imputer”,simpleimputer(strategy=”most\_frequent”)),(“encode”,Ordi hoac onhotEncoder(sparse\_ouput=False True thi la vtri cua 1)(categories=[Education]))])

Dùng columnTransfomer để dùng tất cả xử lí phần feature

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên** | **Dữ liệu áp dụng** | **Mã hóa/chuẩn hóa thành** | **Dùng cho** | **Thứ tự có quan trọng?** |
| **LabelEncoder** | Cột nhãn (y) | Số nguyên (0, 1, 2, ...) | Nhãn đầu ra (target) | ❌ Không |
| **OneHotEncoder** | Cột feature dạng chuỗi | Dãy nhị phân (0 hoặc 1) | Feature phân loại **không có thứ tự** | ❌ Không |
| **OrdinalEncoder** | Cột feature dạng chuỗi | Số nguyên theo thứ tự | Feature phân loại **có thứ tự** | ✅ Có |
| **StandardScaler** | Dữ liệu số (liên tục) | Phân phối chuẩn (mean=0) | Mọi mô hình cần dữ liệu chuẩn hóa | ❌ Không |
| **MinMaxScaler** | Dữ liệu số (liên tục) | Trong khoảng [0, 1] | Khi cần giữ tỷ lệ gốc (vd: ảnh) | ❌ Không |
| **RobustScaler** | Dữ liệu số có **outlier** | Dựa trên median & IQR | Khi có dữ liệu **nhiễu / outlier** | ❌ Không |
| **MaxAbsScaler** | Dữ liệu số có dấu âm | Trong [-1, 1] theo max abs | Sparse data (vd: NLP, TF-IDF) | ❌ Không |
| **QuantileTransformer** | Dữ liệu số | Phân phối chuẩn hoặc đều | Cần **loại bỏ outlier & làm trơn** | ❌ Không |
| **PowerTransformer** | Dữ liệu số lệch/skewed | Gần chuẩn | Dữ liệu lệch (skewed), log-like | ❌ Không |