인공지능 데이터 전처리 결과서

프로젝트명: 통신사 고객 이탈 예측 프로젝트

# 1. 데이터 수집 개요

• 출처: https://www.kaggle.com/datasets/blastchar/telco-customer-churn(IBM)

• 데이터 건수: 총 7,043명 고객 정보

# 2. 원본 데이터 설명

• 주요 컬럼

|  |  |  |
| --- | --- | --- |
| 변수명 | 변수 설명 (description) | 변수 유형 |
| customerID | 고객의 고유 ID | object |
| gender | 고객의 성별 (남성/여성) | object |
| SeniorCitizen | 고령자 여부 (1 = 예, 0 = 아니오) | int64 |
| Partner | 배우자 유무 (예/아니오) | object |
| Dependents | 부양 가족 유무 (예/아니오) | object |
| tenure | 고객이 회사에 가입한 개월 수 | int64 |
| PhoneService | 전화 서비스 이용 여부 (예/아니오) | object |
| MultipleLines | 다중 회선 이용 여부 (예/아니오/전화 서비스 없음) | object |
| InternetService | 인터넷 서비스 종류 (DSL/광섬유/인터넷 서비스 없음) | object |
| OnlineSecurity | 온라인 보안 서비스 이용 여부 (예/아니오) | object |
| OnlineBackup | 온라인 백업 서비스 이용 여부 (예/아니오) | object |
| DeviceProtection | 기기 보호 서비스 이용 여부 (예/아니오) | object |
| TechSupport | 기술 지원 서비스 이용 여부 (예/아니오) | object |
| StreamingTV | 스트리밍 TV 이용 여부 (예/아니오) | object |
| StreamingMovies | 영화 스트리밍 서비스 이용 여부 (예/아니오) | object |
| Contract | 계약 유형 (월별 계약/1년 계약/2년 계약) | object |
| PaperlessBilling | 종이사용없는 청구서 여부 (예/아니오) | object |
| PaymentMethod | 결제 방식 (전자 수표/우편 수표/계좌 이체/신용카드) | object |
| MonthlyCharges | 월별 서비스 요금 | float64 |
| TotalCharges | 고객이 누적하여 지불한 총 요금 | object |
| Chrun | 이탈여부 (예/아니오) | object |

# 3. 전처리 목적

• 불필요한 변수 제거 및 데이터 정합성 확보

• 모델 학습 성능 향상을 위한 이상치/결측치 처리

• 범주형 변수의 인코딩 및 수치형 변수의 스케일링

# 4. 전처리 작업 내역

|  |  |  |
| --- | --- | --- |
| 1 | 결측치 처리 | 결측치 없음 확인 |
| 2 | 중복값 처리 | 중복값 없음 확인 |
| 3 | 텍스트 인코딩 | `TotalCharges` 에서 결측치가 아닌 빈 공간 값 제거 및 데이터 타입 변환 |
| 4 | 스케일링 | 인코딩 후 수치형 변수 스케일링 |
| 5 | 불필요한 컬럼 제거 | `Customer\_ID` 식별용 컬럼 제거 |
| 6 | 라벨 인코딩 | `churn` 컬럼 → Yes: 1, No: 0 |

# 5. 전처리 전/후 비교 예시

|  |  |  |
| --- | --- | --- |
| 고객 ID | 전처리 전: TotalCharges | 전처리 후 |
| 4472-LVYGI  3115-CZMZD | ‘ ’  ‘ ’ | 제거됨  제거됨 |

# 6. 사용 도구

• Python (pandas, numpy, scikit-learn)

• Matplotlib, Seaborn (EDA 시각화용)

# 7. 전처리 결과 요약

• 최종 학습 데이터 수: 7,032건

• 총 특성 수: 23개 컬럼 (라벨 포함)

• 이탈 고객 비율: 26.5%