ADP 실기 시험 대비 학습 주제

- **개발 환경**: Jupyter Notebook 사용법 및 단축키

1. 데이터 전처리 및 탐색 (Data Preprocessing & EDA)

1.1. 데이터 핸들링

- **Python 기초**: List, Dict, Set, Tuple, 함수, 클래스
- **데이터 입출력**: `read_csv`, `read_excel`, `to_csv`
- **데이터 구조 확인**: `head`, `info`, `describe`, `shape`, `isnull.sum`
- **데이터 조작**: 행/열 선택 및 변경, 데이터 타입 변경, 필터링, `apply`, `map`
- **데이터 결합**: `concat`, `merge`

1.2. 데이터 정제 및 피처 엔지니어링

- **결측치 확인**: `isnull`, `isna`
- **결측치 처리**: `dropna`, `fillna`, `interpolate`
- **이상치 탐지 및 처리**: Box Plot, Scatter Plot, Z-score, IQR
- **변수 변환**: 로그 변환, 제곱근 변환
- **스케일링**: `StandardScaler`, `MinMaxScaler`, `RobustScaler`
- **범주형 변수 인코딩**: `LabelEncoder`, `OneHotEncoder`, `get_dummies`
- **파생 변수 생성**
- **차원 축소**: PCA, t-SNE
- **클래스 불균형 처리**: Undersampling, Oversampling (SMOTE), Class Weight

1.3. 데이터 시각화

- **주요 차트**: Line, Bar, Scatter, Histogram, Box Plot, Heatmap, Pair Plot
- **차트 옵션**: 축 레이블, 제목, 범례, 서브플롯(subplot), 여러 그래프 겹치기

1.4. 탐색적 데이터 분석 (EDA)

- **단변량 분석**:
  - **수치형**: 기초 통계량(평균, 중앙값, 분산 등), 분포 시각화(히스토그램, 커널 밀도 추정)
  - **범주형**: 빈도 분석(value_counts), 막대 그래프
- **이변량/다변량 분석**:
  - **수치형 vs 수치형**: 산점도(Scatter Plot), 상관계수(Correlation)
  - **범주형 vs 수치형**: 그룹별 통계량, Box Plot, Violin Plot
  - **범주형 vs 범주형**: 교차표(Crosstab), 누적/그룹 막대 그래프

2. 통계 분석 (Statistical Analysis)

2.1. 기술 통계 및 확률

- **확률**: 순열과 조합, 조건부 확률, 베이즈 정리
- **확률 분포**: 이항 분포, 포아송 분포, 정규 분포
- **기술 통계**: 중심 극한 정리, 표준 오차

- **평균 변화율**: 기하 평균, 로그 변화율
- **표본 크기 산출**: 신뢰수준, 오차한계 등 고려한 공식

2.2. 추론 통계 (가설 검정)

- **기본 개념**: 귀무/대립가설, 유의수준, p-value, 검정통계량
- **정규성 검정**: `Shapiro-Wilk`, `Kolmogorov-Smirnov`
- **등분산성 검정**: `Levene`, `Bartlett`
- **모수 검정**:
  - **T-검정**: 단일표본, 독립표본, 대응표본
  - **Z-검정**
  - **F-검정**
  - **분산분석(ANOVA)**: 일원배치, 이원배치, 사후분석 (`Tukey's HSD`, `Bonferroni`)
- **비모수 검정**:
  - `Wilcoxon signed-rank test`
  - `Mann-Whitney U test`
  - `Kruskal-Wallis H test`
- **범주형 자료 분석**:
  - **카이제곱 검정**: 적합도, 독립성, 동질성
  - **비율 검정**: 단일/두 집단 비율
  - `McNemar test`, `Cochran's Q test`

2.3. 상관/회귀 분석

- **상관 분석**: `Pearson`, `Spearman`, `Kendall`
- **편상관 분석**
- **선형 회귀**: `statsmodels.OLS`, `sklearn.LinearRegression`
- **고급 회귀**:
  - **다항 회귀** (`PolynomialFeatures`)
  - **강건 회귀** (`Robust Regression`)
  - **분위수 회귀** (`Quantile Regression`)
  - **베이지안 회귀**
  - **일반화 선형 모델 (GLM)**: 푸아송, 감마, 음이항 회귀
- **회귀 진단**:
  - 선형성, 잔차의 정규성/등분산성/독립성(Durbin-Watson)
  - **다중공선성**: VIF 확인 및 처리
- **로지스틱 회귀**: 이진/다중 분류

3. 지도 학습 (Supervised Learning)

3.1. 모델링 준비

- **데이터 분할**: `train_test_split`
- **교차 검증**: K-Fold, Stratified K-Fold

3.2. 회귀 모델

- 선형 회귀: `LinearRegression`

- 규제 모델: `Ridge`, `Lasso`, `ElasticNet`
- 트리 기반: `DecisionTreeRegressor`, `RandomForestRegressor`
  - 모델 시각화
- 서포트 벡터 머신: `SVR`
- 부스팅 계열: `GradientBoostingRegressor`, `XGBRegressor`, `LGBMRegressor`

3.3. 분류 모델

- 로지스틱 회귀: `LogisticRegression`
- K-최근접 이웃: `KNeighborsClassifier`
- 트리 기반: `DecisionTreeClassifier`, `RandomForestClassifier`
  - 모델 시각화
- 서포트 벡터 머신: `SVC`
- 부스팅 계열: `GradientBoostingClassifier`, `XGBClassifier`, `LGBMClassifier`
- 나이브 베이즈: `GaussianNB`

3.4. 앙상블 모델

- **보팅 (Voting)**: Hard/Soft Voting
- **배깅 (Bagging)**: `RandomForest`
- **부스팅 (Boosting)**: `AdaBoost`, `GradientBoosting`, `XGBoost`, `LightGBM`
- **스태킹 (Stacking)**

4. 비지도 학습 (Unsupervised Learning)

- **군집 분석**:
  - `K-Means` (분할 군집)
  - `Hierarchical Clustering` (계층적 군집)
  - `DBSCAN` (밀도 기반 군집)
- **연관 규칙 분석**: `Apriori` (지지도, 신뢰도, 향상도)
- **이상 탐지**: `Isolation Forest`, `Local Outlier Factor (LOF)`

5. 시계열 분석 (Time Series Analysis)

- **시계열 기본 개념**: 데이터 변환, 정상성(Stationarity), 자기상관(Autocorrelation)
- **정상성 검정**: `ADF` (Augmented Dickey-Fuller) Test
- **시계열 데이터 처리**
  - 이동, 이동 통계량, 차분, 집계, 평활법(이동평균, 지수평활법 (Holt-Winters))
- **시계열 분해 및 파악**: 추세, 계절성, 주기, 불규칙 요소 / ACF, PACF 플롯
- **시계열 모델**: `AR`, `MA`, `ARMA`, `ARIMA`, `SARIMA`, `MARIMA (미작성)`
- **고급 시계열 모델**: `VAR`, `Prophet`, `LSTM`, 상태 공간 모델 (State Space Models)

6. 텍스트 마이닝 (Text Mining)

- **텍스트 전처리**: 토큰화, 불용어 제거, 형태소 분석, 표제어/어간 추출
- **텍스트 벡터화**: `CountVectorizer`, `TfidfVectorizer`
- **분석 기법**: 감성 분석, 토픽 모델링(LDA)
  - `pytorch` 및 `keras`에서 제공하는 텍스트 모델 종류 및 활용

7. 모델 평가 및 최적화

7.1. 모델 평가지표

- **회귀**: `MSE`, `RMSE`, `MAE`, `MAPE`, `R²`, `Adjusted R²`
- **분류**: 혼동 행렬, `Accuracy`, `Precision`, `Recall`, `F1-Score`, `ROC Curve & AUC`
- **군집**: 실루엣 계수

7.2. 모델 최적화

- **하이퍼파라미터 튜닝**: `GridSearchCV`, `RandomizedSearchCV`, 베이지안 최적화
- **변수 중요도**: 트리 기반 모델의 `feature_importances_`
- **변수 선택법**: 후진 제거법, 전진 선택법

8. 고급 주제 (Advanced Topics)

- **생존 분석**: `Kaplan-Meier` 생존 곡선, `Log-rank` test
- **최적화**: 선형 계획법 (Linear Programming), 정수 선형 계획법, 혼합 정수 선형 계획법
- **이미지 분석 (미완)**
    - 이미지 데이터 전처리 및 증강 방식 정리
    - CNN을 활용한 이미지 분류
        - `pytorch` 및 `keras`에서 제공하는 이미지 모델 종류 및 활용
- **딥러닝**:
    - **모델**: DNN, CNN, RNN, LSTM
        - 단일 구축, Sequential, Functional 활용 등 가능한 구축 방식 전부 작성
        - 문제에 따른 최종 레이어 구성 (회귀, 이중 분류, 다중 분류) 각각 보이기
    - **프레임워크**: PyTorch, Keras/TensorFlow 전부 사용해서 작성
    - **주요 개념**: 활성화 함수, 손실 함수, 옵티마이저, 규제, Dropout
    - 구축한 모델 시각화하는 방법 작성
    - **모델 학습 및 평가** 예제 코드 작성
        - 가중치 업데이트 등 상세하기 작성 필요
        - 성능 평가 코드 작성
        - 추론 코드 작성
        - 모델 저장 방법
        - 학습 과정 시각화
            - loss나 정확도 등