기계학습 – 중간고사 대체과제 (2 문제)
마감: 04.27 (일) 13:00
제출: 분류 과제 결과 압출 파일 + 회귀 과제 결과 압축 파일

I. 지도학습을 이용한 분류 모델 구현 및 분석

1. 과제목표
   A. 실제 분류 데이터를 활용해 머신러닝 모델을 설계하고 평가하는 경험을 갖는다.
   B. 분류 모델의 성능을 다양한 지표로 평가하고, 최적의 하이퍼파라미터를 찾아본다.
   C. scikit-learn 의 분류 모델 활용 능력과 데이터 전처리 이해도를 종합적으로 평가한다.

2. 과제 수행 내용
   A. 분류 데이터셋 선택
      i. ~~아래 제시된 예시 데이터셋을 사용하는데 자신의 학번을 3 으로 나누었을때 나머지가 0 이면 3 번을 1 이면 1 번, 2 면 2 번을 사용해야 합니다.~~
      아래 제시된 예시 데이터셋을 사용하는데 자신의 학번을 2 로 나누었을때 나머지가 0 이면 2 번을 1 이면 1 번을 사용해야 합니다.
   B. 데이터 전처리
      i. 결측치 처리
      ii. 범주형 변수 인코딩 (예: get_dummies, OneHotEncoder)
      iii. 스케일링 (StandardScaler 등)
      iv. 간단한 EDA 시각화
   C. 분류 모델 구현
      i. 선택 가능한 분류 모델: (두 개 이상을 선택)
         1. LogisticRegression
         2. KNeighborsClassifier
         3. DecisionTreeClassifier
         4. RandomForestClassifier
      ii. train_test_split 또는 KFold 사용
      iii. Pipeline 또는 GridSearchCV 활용
   D. **성능 평가**
      i. 평가 지표: (모두 사용해야 함)
         1. Accuracy
         2. Precision / Recall / F1-score
         3. Confusion Matrix
         4. ROC Curve & AUC (이진 분류일 경우)
      ii. 시각화 포함해야 함
   E. **하이퍼파라미터 튜닝**
      i. GridSearchCV 또는 RandomizedSearchCV 사용
      ii. 최적의 모델 파라미터 찾고 성능 향상 확인
   F. **최종 보고서 및 코드/데이터 제출**
      i. .ipynb 또는 .py 코드, 사용한 데이터 + PDF 보고서 제출
         1. 코드에는 충분한 comment 가 있어야 함
         2. 하나의 압출파일로 제출하는데 맥에서 압축하면 Windows PC 에서 제대로 열리는지 확인해야 함 (Windows PC 에서 열리지 않으면 0 점)
         3. 파일이름에는 한글이 없어야 함 (Windows PC 에서 파일이름이 깨져 보이면 0 점)
      ii. 데이터는 프로그램이 있는 디렉토리에 저장한 후 프로그램을 실행하면 처리될 수 있어야 함

iii. 보고서에는 문제 정의, 전처리, 모델 분석 및 시각화 결과 포함 (보고서 템플릿 참고)

평가 기준 (Rubric, 총 50 점)

항목	배점	평가 기준
데이터 선택 및 문제 정의	5	이진/다중 클래스 분류 문제의 적절성, 현실성 있는 목표 설정
데이터 전처리 및 EDA	7.5	결측치/인코딩/스케일링 수행 여부, 변수 분포 분석 및 시각화
분류 모델 구현 및 설명	10	코드 정확도, 모델 선택 이유, 파이프라인 구성 (가산점) 포함
성능 평가 지표 활용	10	정확도 외 추가 지표 포함 (정밀도, 재현율, F1, AUC 등), confusion matrix 포함
하이퍼파라미터 튜닝	7.5	튜닝 수행 여부, 성능 개선 확인, 적절한 하이퍼파라미터 조합 탐색
결과 해석 및 시각화	5	결과에 대한 분석력, 성능 시각화(ROC Curve, Bar plot 등) 포함
보고서 완성도 및 코드 품질	5	PDF 보고서 구성력, 코드 주석 및 가독성, 재현성 포함

예시 분류 데이터셋

1. Mushroom Dataset (버섯의 독성 여부 예측)
    1. https://www.kaggle.com/datasets/rinichristy/uci-mushroom-dataset
2. Wine Variety Classification (포도주의 특징(산도, 당도, 알코올 등)으로 **품종(variety)** 예측)
    1. https://www.kaggle.com/datasets/samuelmcguire/wine-reviews-data
    2. **데이터 수가 많기 때문에 랜덤하게 1/10 만 선택해서 데이터로 사용하세요.**
3. ~~Painting Genre Classification (미술 작품의 다양한 장르를 예측 (예: Abstract, Impressionism, Realism 등)~~
    1. ~~https://huggingface.co/datasets/huggan/wikiart~~

II. 지도학습을 이용한 회귀 모델 구현 및 분석

1. 과제목표

A. 실제 데이터를 기반으로 연속형 값을 예측하는 회귀 모델을 구축한다.
B. 회귀 모델의 성능을 다양한 지표로 평가하고, 최적의 하이퍼파라미터를 찾아본다.
C. scikit-learn 라이브러리를 활용한 전처리, 모델링, 튜닝 능력을 통합적으로 평가한다.

2. 과제 수행 내용
   A. 회귀 데이터셋 선택
      i. 아래 제시된 예시 데이터셋을 사용하는데 자신의 학번을 3 으로 나누었을때 나머지가 0 이면 3 번을 1 이면 1 번, 2 면 2 번을 사용해야 합니다.
      ii. 목표(Target)는 반드시 **실수형 변수**이어야 함 (예: 가격, 점수, 수치 등)
   B. 데이터 전처리
      i. 결측치 처리 (삭제 또는 SimpleImputer)
      ii. 범주형 변수 인코딩 (get_dummies 또는 OneHotEncoder)
      iii. 스케일링 (StandardScaler)
      iv. 변수 간 상관관계 확인 및 시각화 포함 권장
   C. 회귀 모델 구현
      i. 선택 가능한 회귀 모델: (두 개 이상을 선택)
         1. LinearRegression
         2. Ridge
         3. Lasso
         4. DecisionTreeRegressor 또는 RandomForestRegressor
      ii. train_test_split 또는 KFold 사용
      iii. Pipeline 또는 cross_val_score 활용
   D. **성능 평가**
      i. 평가 지표: (모두 사용해야 함)
         1. **RMSE (Root Mean Squared Error)**
         2. **MAE (Mean Absolute Error)**
         3. **R^2 Score (결정계수)**
      ii. 시각화 포함해야 함 (예: 예측값 vs 실제값 산점도)
   E. **하이퍼파라미터 튜닝**
      i. GridSearchCV 또는 RandomizedSearchCV 사용
      ii. 성능 향상을 확인하고, 최적 하이퍼파라미터 명시
   F. **최종 보고서 및 코드/데이터 제출**
      i. .ipynb 또는 .py 코드, 사용한 데이터 + PDF 보고서 제출
         1. 코드에는 충분한 comment 가 있어야 함
         2. 하나의 압출파일로 제출하는데 맥에서 압축하면 Windows PC 에서 제대로 열리는지 확인해야 함 (Windows PC 에서 열리지 않으면 0 점)
         3. 파일이름에는 한글이 없어야 함 (Windows PC 에서 파일이름이 깨져 보이면 0 점)
      ii. 데이터는 프로그램이 있는 디렉토리에 저장한 후 프로그램을 실행하면 처리될 수 있어야 함
      iii. 보고서 구성: 문제 정의 → 전처리 → 모델링 → 평가 → 해석 (보고서 템플릿 참고)

평가 기준 (Rubric, 총 50 점)

항목	배점	평가 기준
데이터 선택 및 문제 정의	5	예측 대상이 회귀 문제로 적절하며, 데이터의 구조와 목적이 잘 설명됨
데이터 전처리 및 EDA	7.5	결측치, 인코딩, 스케일링 등 전처리 적절성 및 기본 분석 수행

회귀 모델 구현 및 설명	10	모델 선택 근거, 파이프라인 구성, 코드 구현 정확성 포함
성능 평가 지표 활용	10	RMSE, MAE, R^2 등 다양한 지표 활용과 해석 포함
하이퍼파라미터 튜닝	7.5	최적 파라미터 탐색 및 성능 개선 여부 설명
결과 해석 및 시각화	5	예측결과 시각화 포함 여부, 시각적 해석 및 정확도에 대한 논리적 설명
보고서 완성도 및 코드 품질	5	코드 가독성, 주석, 보고서 구성력, 결과 해석의 논리성 포함

예시 회귀 데이터셋

1. Spotify Track Popularity Dataset
   A. **타겟**: popularity (0~100)
   B. **피처**: danceability, energy, tempo, acousticness, genre 등
   C. **출처**: https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

2. Students' Mental Health & Productivity
   A. **타겟: stress level 또는 productivity**
   B. **피처: 학습 시간, 수면, 성별, 식사 여부 등**
   C. **출처:** https://www.kaggle.com/datasets/shariful07/student-mental-health

3. Energy Efficiency Dataset
   A. **타겟: 냉난방 부하 (연속형)**
   B. **피처: 건물 면적, 방향, 창 크기 등 건축 특성**
   C. **출처:** https://www.kaggle.com/datasets/ujjwalchowdhury/energy-efficiency-data-set
   D.

[분류 모델 개발 보고서 템플릿]

과목: 기계학습
과제명: 분류 예측 모델 구현 및 분석
학번 / 이름: [학번] / [이름]
제출일: [제출일]

1. 프로젝트 개요

- **문제 정의**: 어떤 클래스를 예측하는 문제인지 명확히 기술
  (예: 버섯의 독성 여부, 음악 장르 분류, 와인 품종 예측 등)
- **데이터셋 설명**:
  - **출처**:
  - **데이터셋 크기**: (샘플 수 × 변수 수)
  - **종속변수(target)** 및 주요 **독립변수(features)**:

2. 데이터 전처리 및 탐색적 분석 (EDA)

- **결측치 처리**:
  - 사용한 방법: (삭제, 평균/최빈값 대체, Imputer 등)
- **범주형 변수 처리**:
  - 인코딩 방식: (예: get_dummies, OneHotEncoder 등)
- **스케일링**:
  - 사용한 기법: (StandardScaler, MinMaxScaler 등)
- **EDA 시각화**:
  - 클래스 분포 시각화
  - 주요 변수 분포 / 상관관계 분석 등

3. 모델 구축 및 학습

- **사용한 알고리즘**:
  - (예: LogisticRegression, KNeighborsClassifier, DecisionTreeClassifier 등 **2 개 이상**)
- **데이터 분할 방식**:
  - 예: train_test_split, KFold, StratifiedKFold
- **파이프라인 사용 여부**:
  - 사용 시 구조 설명
- **학습 코드 요약**:
  - 핵심 코드 또는 pseudocode 간단히 첨부 가능

4. 성능 평가

- **사용한 지표**:
  - Accuracy, Precision, Recall, F1-score, ROC-AUC (이진 분류일 경우)
- **예측 결과 시각화**:
  - Confusion Matrix 시각화
  - ROC Curve, Precision-Recall Curve 등
- **해석**:
  - 어떤 모델이 더 성능이 좋은가?
  - 클래스 간 성능 차이는 존재하는가?

5. 하이퍼파라미터 튜닝

- **튜닝 방법**:
  - GridSearchCV, RandomizedSearchCV 등
- **튜닝한 하이퍼파라미터**:
  - (예: C, max_depth, n_neighbors, criterion 등)
- **튜닝 결과 분석**:
  - 성능 변화 비교, 선택한 최적 파라미터 해석

6. 결론 및 고찰

- 최종 모델 성능 종합 평가
- 데이터 또는 모델의 한계
- 실생활 응용 가능성 또는 확장 방향
- 다음 단계에서 고려할 점 (예: 클래스 불균형 처리, 이상치 처리, 앙상블 등)

7. 참고자료

- 데이터셋 출처 URL
- 참고한 논문, 블로그, scikit-learn 문서 등
- 사용한 주요 라이브러리 버전 (선택)

부록

- 중요 코드 스니펫 (예: 파이프라인 정의, 성능 평가 루프 등)
- confusion matrix 및 ROC curve 시각화 추가

제출 유의사항

- PDF 형식으로 제출
- 코드 파일은 .ipynb 또는 .py 형태로 별도 제출
- **표절 시 0점 처리**

[회귀 모델 개발 보고서 템플릿]

과목: 기계학습
과제명: 회귀 예측 모델 구현 및 분석

학번 / 이름: [학번] / [이름]
제출일: [제출일]

1. 프로젝트 개요

- **문제 정의**: 어떤 연속형 값을 예측하는 문제인지 명확히 기술
  (예: 음악의 인기 점수, 주택 가격, 시험 점수 등)
- **데이터셋 설명**:
  - 출처:
  - 데이터셋 크기 (샘플 수 × 변수 수):
  - 종속변수(target) 및 주요 독립변수(features):

2. 데이터 전처리 및 탐색적 분석 (EDA)

- **결측치 처리**:
  - 사용한 방법: (삭제, 평균/중앙값 대체, Imputer 등)
- **범주형 변수 처리**:
  - (예: get_dummies, OneHotEncoder 활용 여부)
- **스케일링**:
  - 사용한 기법: (StandardScaler, MinMaxScaler 등)
- **EDA 시각화**:
  - 변수 간 상관관계 분석
  - 종속변수 분포 시각화 (히스토그램, 박스플롯 등)

3. 모델 구축 및 학습

- **사용한 알고리즘**:
  - (예: Linear Regression, Ridge, Lasso, DecisionTreeRegressor 등)
- **데이터 분할 방식**:
  - 예: train_test_split, KFold
- **파이프라인 사용 여부**:
  - 사용 시 구조 설명
- **학습 코드 요약**:
  - 핵심 코드 또는 pseudocode 간단히 첨부 가능

4. 성능 평가

- **사용한 지표**:
  - RMSE, MAE, R^2 등
- **예측값 vs 실제값 시각화**:
  - 산점도, 잔차(residuals) 플롯 등 포함
- **해석**:
  - 모델이 잘 작동하는가? 과소/과대 예측 패턴이 있는가?

5. 하이퍼파라미터 튜닝

- **튜닝 방법**:
  - GridSearchCV, RandomizedSearchCV 등
- **튜닝한 하이퍼파라미터**:
  - (예: alpha, max_depth 등)
- **튜닝 결과 분석**:
  - 성능 변화 비교, 선택한 최적 파라미터 해석

6. 결론 및 고찰

- 최종 모델 성능 종합 평가
- 데이터 또는 모델의 한계
- 실생활 응용 가능성 또는 확장 방향

- 다음 단계에서 고려할 점 (특성 선택, 앙상블, 이상치 제거 등)

7. 참고자료
- 데이터셋 출처 URL
- 참고한 논문, 블로그, scikit-learn 문서 등
- 사용한 주요 라이브러리 버전(optional)

부록
- 중요 코드 스니펫 (예: 파이프라인 정의, 모델 평가 루프 등)

제출 유의사항
- PDF 형식으로 제출
- 코드 파일은 .ipynb 또는 .py 형태로 별도 제출
- **표절 시 0점 처리**