📢 데이터마이닝과분석 텀프로젝트 공지 (회귀: 정규화)

🧭 개요

- **주제:** 수치형 변수를 이용한 **회귀 예측 모델 구축**
- **데이터:** 교수 제공 단일 파일
  - 팀이 자유롭게 train/validation을 분할
  - **비공개 Test set은 교수 측에서 최종 평가용으로만 사용**
- **팀 구성:** 3~4명
- **제출물:**
  - 분석 노트북(`.ipynb`)
  - 발표자료(`.pdf` 또는 `.pptx`)
- **발표:** 14주차 (팀당 15~20분: 10분 발표 + 5~10분 질의응답)

⚙️ 분석 조건

- **허용 알고리즘:**
  - **선형회귀 계열 모델만 허용**
  - (예: 기본 선형회귀 및 정규화 기법 적용 가능)
- **제외 알고리즘:**
  - 의사결정나무, 앙상블, 부스팅, 신경망 등 비선형 모델 일체 금지
- **전처리:**
  - 자유 (결측치 처리, 스케일링, 변수 선택, 이상치 제거 등)
- **Seed:**
  - 고정 `random_state=42` (모든 팀 동일)

📊 평가 지표 (가산점 전용)

- **사용 지표:** adjusted R^2 (수정된 결정계수), MAPE
- **최종 점수(M):**

$$M = \frac{\max(0, \text{adj-}R^2) + \max(0, 1 - \text{MAPE})}{2}$$

- **정의:**

- **adjusted R² (수정된 결정계수):** 변수 수(p)와 표본수(n)를 보정한 결정계수입니다.
  - 계산 방법: `sklearn.metrics.r2_score`로 R²를 구한 뒤 아래 공식을 사용해 adj-R²를 계산하세요:

$$\text{adj-}R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- **MAPE (Mean Absolute Percentage Error):** 평균 절대 백분율 오차로, 0에 가까울수록 좋습니다.
  - `sklearn.metrics.mean_absolute_percentage_error` 함수를 사용하세요.
  - **주의:** 소수 단위로 계산합니다. (예: 15% 오차 → 0.15)
- 최종 점수 `M`은 adjusted R²와 (1-MAPE)를 각각 0 이상으로 클리핑한 후 평균을 냅니다.

🧾 점수 구조 (총 100점 + 가산 가능)

구분	핵심 평가 관점	세부내용	배점	채점방식
A. 코드 품질 (EDA+모델링)	**"Why?"** (논리) (예: 이 방법을 "왜" 썼는가?)	전처리, 변수 선택, 회귀 구조 논리 **(* 학생 제출 Validation Set M과 비공개 Test Set M 간 차이가 현저할 경우, '과적합'으로 간주되어 본 항목에서 감점)**	**35**	ipynb (보고서)
B. 리포트 완성도	**"How?"** (전달력) (예: "얼마나" 깔끔하고 재현 가능한가?)	구조, 재현 가능성, 해석력	**20**	코드 실행 및 설명 평가
C. 발표 및 질의응답	**"Really?"** (이해도) (예: "정말" 본인이 한 것인가?)	발표 명료성, 이해도, 방어 능력 등 **(* 질의응답은 Peer 점수와 관계없이 모든 학생에게 랜덤하게 주어질 수 있음) (* Peer 평균 3.0 미만 시, 본 질의응답을 IPM 페널티 확정/구제를 위한 '안전장치'로 활용. 상세 규정은 `IPM_평가방식_수정.pdf` 참조)**	**45**	발표 완성도, 논리 방어 등
D. (가산) 성능 경쟁	**"How Well?"** (성능)	`M` 점수 순위 (1위:+10 ~ 4위:+3)	**+10**	교수 실행 평가
E. 기여도·참여도	**(개인별 적용)**	IPM 가중치로만 반영 (별도 배점 없음)	-	IPM 가중치로만 반영

> **A+B 점수는 정당한 분석 근거, 재현 노력, 해석의 깊이 등 초과적 노력이 뚜렷한 팀을 기준** 으로 삼아 평가하며, **A+B+C의 최고점은 100점입니다.**

🧮 (가산) Test Set 성능 경쟁

- **비공개 Test set**을 교수 측에서 동일 조건(`random_state=42`)으로 실행
- 평가 기준은 `📊 평가 지표` 섹션의 M 점수 순위로 함
- **가산 규칙 (팀 순위 기준):**
  - 1위: +10점
  - 2위: +7점
  - 3위: +5점
  - 4위: +3점
  - 5위 이하: +0점
- 팀 수가 4개 이하인 경우: 1~4위까지만 가산(1위:+10, 2위:+7, 3위:+5, 4위:+3)
- 팀 수가 5개 이상인 경우: 1~4위까지만 가산(1위:+10, 2위:+7, 3위:+5, 4위:+3), 5위 이하: +0점
- Test M 동점 시: C(발표/질의) → A+B(코드·리포트) 순으로 tie-break
- **총점은 100점을 초과할 수 있음**

🗣️ 발표 및 질의응답

- **진행:**
  - 제출된 노트북 이후 **코드 수정 불가**, 발표 중에는 교수 지시에 따라 **지정 셀 실행만 허용**
  - 10분 발표(핵심 요약 중심)
    - 초과 시 감점
  - 5~10분 질의
    - 팀마다 질의 내용과 대상은 다를 수 있음

📈 IPM 계산 및 예외 규정

- **모든 IPM 관련 규칙(4-Tier), 예외 케이스(담합, 미제출), 3/4인팀 공정성, 만회 기회, 점수 계산 예시 등은 아래 링크된 문서를 기준으로 합니다.**
- **필독:** `IPM_평가방식_수정.pdf`

⚖️ 공정성 규정

상황	조치	비고
코드 복제·유사	0점	전체 배점 박탈
제출 지연	하루 -15점	3일 이상 0점

상황	조치	비고
무임승차자	패널티 적용	상세 규정은 `IPM_평가방식_수정.pdf` 참조

🧮 채점 및 계산 요약

1. 발표 후 즉시 C항목 평가 (45점)
2. `.ipynb` 코드 및 리포트 검토 (A+B항목, 55점)
3. **Test set 실행 후 M 계산** → `D. (가산) 성능 경쟁` 점수 부여
4. Peer 참여·팀 내 평가 자동 반영 (IPM 가중치 적용)

- 팀 점수는 **A+B+C (총 100점)** 에 순위 **가산점(D항목)** 을 더해 계산하며,
- **E항목(기여도)은 팀 점수에 직접 점수를 더하지 않고 IPM 가중치로만 개인 점수에 반영**됩니다.
- Peer Review 미제출 시 페널티 등 모든 IPM 관련 사항은 `IPM_평가방식_수정.pdf` 를 따릅니다.
- 최종 성적 산출 시 소수점은 반올림하여 정수로 표기합니다.

✅ 채점표 (요약형)

항목	배점	기준
EDA/모델링 품질	**35**	전처리·변수선택·선형회귀 적용 논리 (과적합 포함)
리포트 완성도	**20**	구조·재현 가능성·해석력
발표/질의응답	**45**	명료성·대응력·모델 이해 (IPM 안전장치 / 랜덤 질의)
가산점 (성능)	**+10**	M 순위 가산(1:+10, 2:+7, 3:+5, 4:+3, 기타:0)
참여도/기여도	IPM 가중치	상세 규정은 `IPM_평가방식_수정.pdf` 참조

🎯 요약

- **모델 제약:** 선형회귀 계열만 허용
- **분석 목표:** 해석 가능한 회귀모델 구축
- **성능 평가: 가산점으로만 반영** (adjusted R^2와 (1-MAPE)의 평균)
- **공정성 확보:** 동일 조건, 동일 Seed, 동일 지표
- **최대 점수(팀 기준): A+B+C=100점**
  - **성능 가산(D)**으로 최대 **110점** (팀 기준)
  - IPM 반영시 110점 초과 가능