# 모델 선택 가이드라인 QA

Q1. 분류 문제에서 로지스틱 회귀 vs 결정트리 중 어떤 모델을 선택해야 하나요?
A1.
1. 로지스틱 회귀: 해석이 중요하고, 입력 변수와 출력 간 선형적인 경향이 있을 때 적합.
2. 결정트리: 변수 간 비선형 관계나 복잡한 조건 분기가 존재할 때 유리.
3. 성능보다 이해와 해석이 중요한 경우 로지스틱, 비정형 패턴 인식이 중요한 경우 트리를 선택합니다.

Q2. 분류 문제에서 샘플 수가 적을 때는 어떤 모델이 적합한가요?
A2.
1. 로지스틱 회귀, 나이브 베이즈 등 파라미터 수가 적고 과적합에 덜 민감한 모델이 적합.
2. 트리 기반 모델은 샘플 수가 적으면 오히려 과적합 우려가 높습니다.
3. 단순하고 일반화 가능한 모델부터 시도하는 것이 안전합니다.

Q3. 다중회귀 vs Lasso/Ridge는 어떤 기준으로 선택하나요?
A3.
1. 다중회귀: 변수 수가 적고 다중공선성이 크지 않은 경우
2. Ridge: 다중공선성 완화 목적 (계수를 0에 가깝게 수축)
3. Lasso: 변수 선택 기능이 필요한 경우 (불필요 변수 계수를 0으로 만듦)
4. 예측 정확도와 변수 해석력 사이에서 균형 필요 시 ElasticNet을 고려합니다.

Q4. 범주형 변수만 있을 때 적합한 모델은?
A4.
1. 나이브 베이즈, 로지스틱 회귀가 효과적.
2. 트리 기반 모델(XGBoost, LightGBM)은 범주형을 잘 다루나, 원-핫 인코딩 또는 라벨 인코딩이 필요.
3. 특히 변수 간 독립성이 높은 경우 나이브 베이즈가 빠르고 정확도도 우수합니다.

Q5. 비선형 관계가 강할 때 선형 모델을 쓰면 안 되나요?
A5.
1. 선형 모델은 해석이 용이하지만 예측력이 떨어질 수 있음.
2. 잔차 플롯 분석으로 패턴이 남아 있다면 비선형 모델(트리, SVM, 신경망 등) 고려.
3. 또는 다항 회귀, 스플라인 회귀, 로그변환 등 선형 모델 내 확장 기법을 사용할 수도 있습니다.

Q6. 시계열 예측에서 단순 회귀와 ARIMA는 어떻게 구분해서 쓰나요?
A6.
1. 단순 회귀: 시간 변수 외의 예측 변수가 있는 경우 (ex: 온도에 따른 판매량 예측)
2. ARIMA: 시계열의 **자체 패턴(계절성, 추세)**을 기반으로 예측할 때 적합
3. 예측 정확도와 패턴 복잡도에 따라 선택하며, 시계열 분해 후 특성 파악이 선행되어야 합니다.

Q7. 예측 정확도가 가장 중요한 경우 어떤 모델을 선택해야 하나요?
A7.
1. 트리 기반 앙상블 모델 (XGBoost, LightGBM, Random Forest 등)이 일반적으로 성능이 뛰어납니다.
2. 특히 변수 수가 많고 상호작용이 복잡한 경우 강력한 예측력을 가집니다.
3. 단, 해석력은 떨어지므로 설명 가능한 AI(XAI) 기법과 병행 사용하는 것이 좋습니다.

Q8. 모델 학습 시간이 중요할 때는 어떤 모델을 쓰는 게 좋나요?
A8.
1. 로지스틱 회귀, 나이브 베이즈, KNN은 학습 시간이 짧습니다.
2. 트리 기반 앙상블, 딥러닝은 상대적으로 오래 걸립니다.
3. 빠른 프로토타이핑에는 선형 모델부터 시작하고, 성능 향상이 필요하면 복잡한 모델로 전환합니다.

Q9. 다중 분류 문제는 어떤 모델이 적합한가요?
A9.
1. 로지스틱 회귀(Softmax), 랜덤포레스트, XGBoost, SVM 등이 가능
2. XGBoost와 LightGBM은 내부적으로 다중분류를 지원하며 성능이 좋음
3. 클래스 불균형 문제 시, 샘플링 기법(SMOTE 등) 병행 필수

Q10. 클래스 불균형 문제가 있는 경우 어떤 모델이 유리한가요?
A10.
1. 앙상블 모델(XGBoost, LightGBM)은 클래스 가중치 조정 옵션 제공
2. 로지스틱 회귀 + 가중치 부여, 리샘플링(SMOTE, 언더샘플링) 기법도 효과적
3. 평가 지표는 정확도 대신 F1, ROC AUC로 확인

Q11. 고객 세분화에 적합한 모델은 무엇인가요?
A11.
1. K-평균(K-Means): 대표적이고 빠른 군집화 알고리즘, 수치형 데이터에 적합
2. DBSCAN, GMM, 계층적 군집화: K-Means의 한계를 보완
3. 예: 비선형 분포, 밀도 기반 세분화가 필요할 때 GMM이나 DBSCAN 고려

Q12. 이상치 감지에 적합한 모델은?
A12.
1. Isolation Forest, One-Class SVM, LOF(Local Outlier Factor) 등 비지도 학습 기반
2. 회귀/분류 기반 모델에서 예측 오차 기반 감지도 가능
3. 데이터 양과 이상치 비율에 따라 적절한 모델을 선택합니다

Q13. 변수 선택이 중요한 경우 적합한 모델은?
A13.
1. Lasso 회귀: 계수를 0으로 만들어 변수 제거
2. 트리 기반 모델: 변수 중요도(Feature Importance) 제공
3. 변수 수가 많고 해석력이 중요한 경우 Lasso → 트리 기반으로 확장하는 전략 유효

Q14. 해석 가능한 모델이 필요한 경우 어떤 모델을 선택하나요?
A14.
1. 로지스틱 회귀, 선형 회귀, 의사결정나무는 직관적 해석이 용이
2. 복잡 모델은 SHAP, LIME 같은 설명 가능한 AI 도구 병행 필요
3. 규제 환경, 의료 등 해석력이 중요할 땐 단순 모델 우선 사용

Q15. 텍스트 데이터를 다룰 때 적합한 모델은?
A15.
1. TF-IDF + 로지스틱 회귀 / 나이브 베이즈: 빠르고 해석 가능
2. 워드 임베딩(BERT, FastText) + 딥러닝: 의미 기반 고성능 모델
3. 데이터 규모와 목표에 따라 전처리 수준과 모델 복잡도 조절