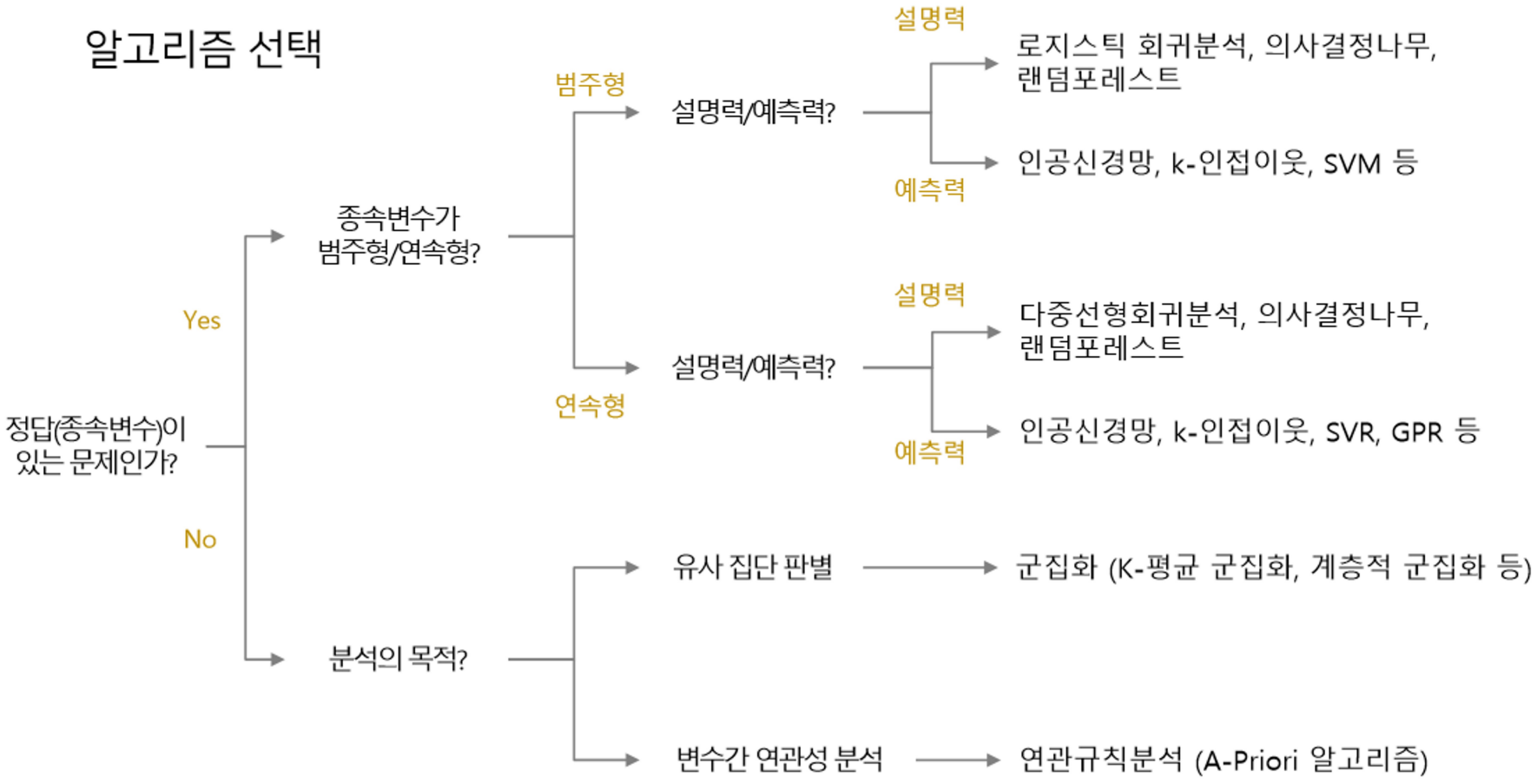
Machine Learning FLOW

2019년 3월 20일 수요일 오후 4:37

- 질문의 속성, 데이터의 특징, 결과의 설명력 포함 유무 등을 고려하여 적합한 분석 알고리즘 선택



❖ 각 단계별 주요 과업 및 산출물

	목적 및 문제 정의	데이터 수집/검증/수정	데이터 전처리	모형 구축	평가 및 해석
주요 활동	• 데이터 분석을 통해 달성하고자 하는 목표 구체화	• 데이터 원천 확인 • 독립변수/종속변수 정의 • 변수별 이상치/결측치 탐지 및 제거	• 불필요한 변수 삭제 • 변수 변환 • 비지도 방식의 변수 선택 및 추출 • 데이터 분할	• 모델 학습 • 최적 파라미터 선택	• 모델링 결과 평가 • 개선안 수립
주 사용 기법			• 기초통계분석을 포함한 EDA • 주성분분석	• 분류 알고리즘 • 회귀 알고리즘 • 군집화 알고리즘 • 이상치 탐지 알고리즘	
산출물	• 문제 기술서 • 모형의 유형(분류/회귀 등)	• 행렬 형태의 모델링 기초 데이터 (행: 레코드, 열: 변수)	• 정제된 모델링용 데이터	• 구축된 모형 • 성능 평가 결과	• 모델 결과 평가표 • 개선 아이디어 리스트
고려 사항	• 현재 보유 데이터로 달성 가능한 목적인가?	• 최대한 많은 레코드와 변수를 이 단계에서 수집	• 사용 모형에 따른 데이터 분할 비율 • 문제에 따른 적절한 변수 수	• 다양한 알고리즘 시도 • 최적 파라미터 선택시 충분한 영역 탐색	• 모델의 결과가 현장에서 수용 가능한 수준인가?

1. **Defining the problem statement**
2. **Collecting the data**
3. **Exploratory data analysis**
4. **Feature engineering**
5. **Modelling**
6. **Testing**