인공지능 데이터 전처리 - 데이터 분석 과정

데이터 수집

- 공공데이터 사이트에서 데이터 수집
- CSV 파일 / API를 사용하여 데이터 수집

데이터 확인

- 데이터의 내용 : .head()
- 데이터 수 / 컬럼 수 / 데이터 타입 / 결측치 : .info()
- 데이터 기초 통계 / 이상치 여부 : .describe() / boxplot
- 데이터의 분포 : histplot

데이터 전처리

- 결측치 처리 : 제거, 대치(평균값 / 중앙값 / 0 / ffill, bfill
- 데이터 타입 변경 : astype()
- 스케일링 : 표준화 / 정규화
- 범주형 데이터 수치화(encoding) : label / one-hot
- 신규변수 생성 : 집계 / 요약 / 기간별 구분

인공지능 데이터 전처리 - 데이터 분석 과정

데이터 시각화

- 수치데이터 상관관계 분석 : pairplot
- 수치 데이터 분포 : histplot
- 범주형 데이터 분포 : countplot
- Violinplot, swarmplot, boxplot

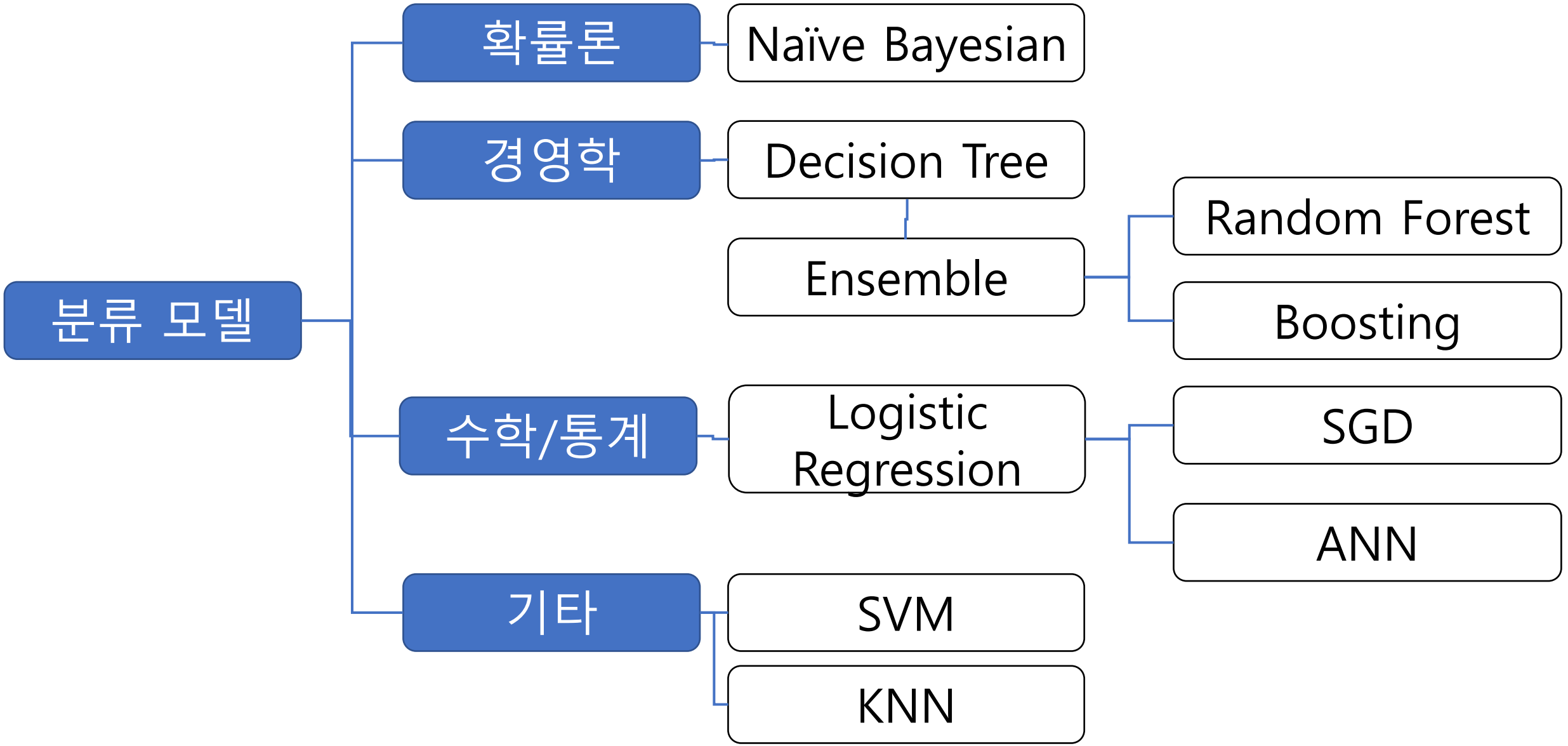
모델 선정

- 데이터 분리 : X, y 데이터 분리 / train, test data 분리 : train_test_split
- 알고리즘 (회귀 / 분류) 별 학습 및 테스트 데이터로 예측
- 회귀 / 분류 알고리즘에 따른 성능 지표를 기준으로 성능 비교

모델 최적화

- 과대적합 여부 확인 (학습 , 평가 결과 비교)
- 목표값의 클래스 편차가 심할 경우 Sampling 요소 고려
- 과대적합 해소를 위한 규제 적용
- 성능 향상을 위한 하이퍼파라미터 튜닝

분류 모델의 종류


분류 모델
확률론
Naïve Bayesian
경영학
Decision Tree
Ensemble
Random Forest
Boosting
수학/통계
Logistic Regression
SGD
ANN
기타
SVM
KNN