**[문제 1] K-NN 알고리즘을 활용한 결측치 대치**

다음은 제품의 정상/불량을 분류하는 인공지능 모델을 구축하고자 한다. 입력변수는 반도체 공정에서 모니터링된 센서 값으로 결측치가 많은 상황이다.

(a) 결측치를 모두 0으로 대치하고 검증 데이터에 대해 성능평가 지표를 작성하시오. (단, 검증 데이터에 대한 평가는 **Decision Tree 알고리즘**으로 진행하시오.)

|  |  |  |
| --- | --- | --- |
| F1-score | G-mean | Accuracy |
| 0.939 | 0.941 | 0.94 |

(b) K-NN 기반의 결측치 대치 방법을 통해 결측치 보완을 진행하시오. 검증 데이터에 F1-Score 기준으로 우수한 성능을 보이는 최적의 K와 표의 성능평가 지표를 작성하시오. **(**단, 검증 데이터의 성능 평가는 **Decision Tree 알고리즘**으로 진행하며**, train 및 test 데이터**에 대한 **scaling**을 진행하지 않는다.**)**

|  |  |  |  |
| --- | --- | --- | --- |
| 최적의 K | F1-score | G-mean | Accuracy |
| 6 | 0.981 | 0.979 | 0.98 |

(c) Min-Max 스케일링을 적용한 후 (b)를 다시 수행하고 결과의 차이가 있는지 답하고, 그 이유를 간단히 설명하시오.

|  |  |  |  |
| --- | --- | --- | --- |
| 최적의 K | F1-score | G-mean | Accuracy |
| 6 | 0.981 | 0.979 | 0.98 |

Decision Tree에서는 특성의 값이 다른 범위에 있어도 각 특성의 경계를 찾아내고 분할하기 때문에 특성의 스케일이 달라도 잘 동작합니다.

반면 회귀모델에서는 입력 변수에 곱을 통해 예측하기 때문에 큰 값을 가진 변수에 더 많은 중요성을 부여하기 때문에 Scaling 작업을 진행해야 합니다.

* 문제 1은 k-fold 교차검증 및 하이퍼파라미터 최적화를 진행할 필요 없음

**[문제 2] KNN, DT, 비선형 SVM 모델 활용 문제**

2번 데이터는 은행의 고객 데이터이다. 총 12개의 입력변수(predictors)와 1개의 출력변수(response)로 구성되어 있다. 출력 변수는 고객이 계좌를 닫았는 지의 여부(닫았으면 1, 아니면 0)이다.

첨부된 학습 데이터를 이용해 k-nearest neighbors, decision tree, 비선형 SVM 모델(**‘rbf커널’**로 한정함)을 만들고 테스트 데이터의 response를 예측하시오.

* 성능향상을 위해 추가적인 Feature Engineering을 진행하여도 무방하며, 성능평가 기준은 f1 score로 한다.

|  |  |  |
| --- | --- | --- |
| KNN 예측 | DT 예측 | SVM 예측 |
| 0 | 0 | 0 |
| … | … | … |
| 1 | 0 | 1 |

테스트 데이터에 대한 예측 파일 (csv 파일로 아래 양식에 맞춰 제출)

<sample\_submission.csv 파일 생성 예시>

[제출양식]

제출파일 1: 문제 1에대한 답변을 포함한 word파일

제출파일 2: 문제 1, 문제 2에 대한 ipynb 파일(개별파일로 제출하거나 묶어서 제출하거나 상관없습니다.)

제출파일 3: 문제 2번에 대한 csv파일