**문제 유형1**

\* 다음은 음식 조리 시간 관련 데이터셋이다. 이 데이터셋을 활용하여 음식 조리 시간에 대한 예측을 수행하시오.

파일명: indian\_food2.csv (구분자: 쉼표(,), 255 행, 10 컬럼, UTF-8 Encoding, 결측치: Empty string(nan, 빈칸(‘ ‘), 문자열(‘’) )

|  |  |  |
| --- | --- | --- |
| **변수명** | **설명** | **데이터 타입** |
| id | 음식 ID | Double |
| name | 음식 이름 | String |
| ingredients | 재료 목록(“, ”로 구분, 예: 고추장, 고추 가루, 된장) | String |
| diet | 고기 포함 여부(vegetarian/non vegetarian) | String |
| prep\_time | 재료 준비 시간 | Double |
| cook\_time | 음식 조리 시간(Target 변수) | Double |
| flavor\_profile | 음식 맛 구분(“-1”, “spicy”, “sweet”, “bitter”, “sour”) | String |
| course | 음식 종류(dessert/main course/snack/starter) | String |
| state | 지역(Andhra Pradesh ~ West Bengal) | String |
| region | 음식 지역(“-1”, “Central”, “East”, “North”, “North East”, “South”, “West”, “nan”) | String |

**\* 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

**단계 1-1**. 지역(state) 컬럼을 제거하시오.

**단계 1-2**. 조리 시간(cook\_time)과 재료 준비 시간(prep\_time)이 양수(>0)인 데이터만 추출하시오.

**단계 1-3**. region의 값이 “-1”인 경우와 빈문자(empty string)는 제거한 후 ‘North East’ 지역을 ‘North’로 변경하시오(결측치는 제거하시오).  
**단계 1-4**. 음식 맛 구분(flavor\_profile)을 다음과 같이 2가지로 정리한 변수(변수명: ed\_flavor\_profile)를 추가하시오.

IF flavor\_profile= ‘sweet’ then ‘sweet’ else ‘not sweet’

**단계 1-5**. 음식 종류(course)를 2가지로 정리한 변수(변수명: ed\_course)를 추가하시오.

IF course = ‘main course ‘ then ‘main course’ else ‘not main course’

**단계 1-6**. 재료 목록(ingredients)를 이용하여 재료 수 변수(변수명: ingredients\_no)를 추가하시오.

예: 고추장, 마늘 🡪 ingredients\_no=2

**\* 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 212 Rows)으로 다음 분석(문제 1~3) 수행**

1. (basetable1을 이용) 2범주로 정리된 음식 맛(ed\_flavor\_profile)이 5개로 정리된 음식 지역(region) 간에 통계적으로 유의미한 차이가 있는지 적절한 검정을 수행하고, 검정 결과 중 검정 통계량을 기술하시오.

- 검정 통계량은 소수점 넷째 자리 이하는 버리고 소수점 셋째 자리까지 기술. (답안 예시) 0.123

|  |  |
| --- | --- |
| 정답 | 22.689 (19.540) |

2. (basetable1을 이용) 5개 지역(region)에 따라 재료 준비 시간(prep\_time)의 평균이 통계적으로 유의미한 차이가 있는지 적절한 검정을 수행하고, 검정 결과 중 유의 확률(P-Value) 값을 기술하시오.

- 주어진 데이터는 정규성과 등분산성을 가정하며, 유의 확률(P-Value) 값은 소수점 넷째 자리 이하는 버리고 소수점 셋째 자리까지 기술. (답안 예시) 0.123

|  |  |
| --- | --- |
| 정답 | 0.0644 (0.069) |

3. (basetable1을 이용) 음식 조리 시간(cook\_time)을 예측하는 모델을 생성하고자 한다. 이를 위하여 basetable1에 다음의 전처리를 추가로 수행하시오.

**단계 1-7**. 다음 변수들을 추가하시오(총 2개).

- ed\_course으로부터 생성한 가변수(1개), ed\_flavor\_profile로부터 생성한 가변수(1개)

- 가변수화 Hint

|  |  |
| --- | --- |
| Brightics Studio | One Hot Encoder 사용, Drop Last=True |

**단계 1-8**. 음식 조리 시간(cook\_time) 예측을 위한 선형 회귀 모형을 생성하고자 한다. 다음 가이드에 따라 Train Data로 학습시킨 모형을 Test Data에 적용하여 음식 조리 시간(cook\_time)을 예측하시오.

- Train Data: id가 3의 배수가 아닌 데이터, Test Data: id가 3의 배수인 데이터

- 독립변수(3개): **단계 1-7**의 가변수들(2개), prep\_time

- 선형 회귀 가이드

|  |  |
| --- | --- |
| Brightics Studio | Default 사용 |

생성한 회귀 분석 모델을 Test Data에 적용하여 음식 조리 시간(cook\_time)를 예측한 후 예측값들에 대한 평균제곱근오차(Root Mean Square Error, RMSE)를 기술하시오.

- RMSE는 소수점 둘째 자리 이하는 버리고 소수점 첫째 자리까지 기술. (답안 예시) 1.2

|  |  |
| --- | --- |
| 정답 | 83.2 |