2021-1학기 생산시스템구축실무 기말과제

식품 살균 공정의 품질 예측

IT경영학과

김민경

목차

1. 프로젝트 개요

2. 탐색적 데이터 분석(EDA)

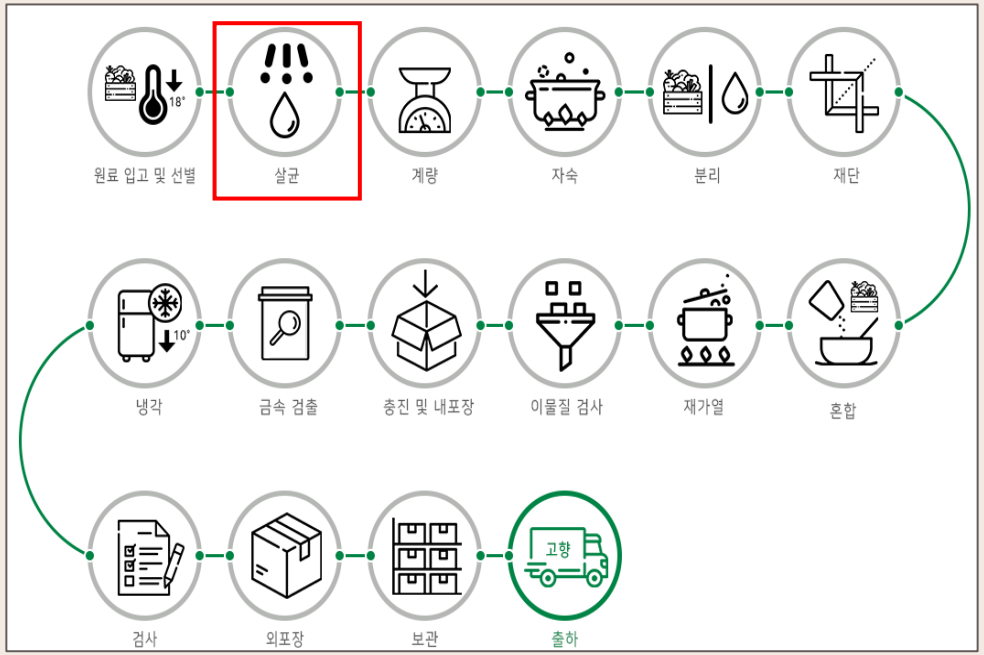
3. 데이터 전처리

4. 모델링

5. 결과 분석

1. 프로젝트 개요

1) 분석 배경 및 목적



<식품 제조 공정도>

- 식품 제조 공정의 두번째 단계인 살균공정(살균기)은 열처리를 통해 액상 식품의 미생물을 사멸 시켜 식품의 안전성과 보존성을 향상시키는 공정을 말한다.

- 식품을 생산하고 판매하기 위해서는 식품안전관리인증기준 HACCP등의 요건들을 만족시켜야 하는데 이에 있어 위생위협을 제거하는 살균공정은 제품의 품질을 판단하는 중요 공정이며 주요 관리 대상이다.

1. 프로젝트 개요

1) 분석 배경 및 목적

- 살균공정의 온도는 제품(식품)마다 다르다. 제품별로 목표 품질을 위한 공정 기준 값을 설정하여 공정을 수행하는데, 가열방식의 특성상 살균공정 동안 내용물의 온도는 계속 변동한다. 살균기의 내부 센서를 통해 가열 온도를 변경하지만, 온도가 조정되는 동안의 시차 때문에 어느 정도의 온도변화는 불가피 하고 이것은 제품의 품질에 영향을 준다.
- 온도 조절을 공정을 관리자의 주관적 판단으로 하고 있기 때문에 일관적이고 정확한 공정제어가 이루어지지 못하고 있다.

데이터를 기반으로 한 일관적이고 정확한 살균공정 품질관리 시스템이 필요하다!

1. 프로젝트 개요

2) 분석 과정

01

탐색적
자료 분석

02

데이터
전처리

독립변수(X변수) 설정
및 데이터 정제

03

모델링

의사결정 나무 모델을
사용한 불량률 예측

04

결과 분석

2. 탐색적 자료 분석(EDA)

1) 데이터셋 구성

	STD_DT	MIXA_PASTEUR_STATE	MIXB_PASTEUR_STATE	MIXA_PASTEUR_TEMP	MIXB_PASTEUR_TEMP	INSP
0	2020-03-04 6:00	1.0	1.0	551.0	524.0	OK
1	2020-03-04 6:30	1.0	1.0	584.0	536.0	OK
2	2020-03-04 7:00	1.0	1.0	584.0	536.0	OK
3	2020-03-04 7:30	1.0	1.0	585.0	536.0	OK
4	2020-03-04 8:00	1.0	1.0	585.0	536.0	OK

변수명	설명	값의 종류
STD_DT0	날짜, 시간(YYYY-MM-DD HH:MM:SS)	해당 값의 날짜와 시간
MIXA_PASTEUR_STATE (독립변수)	살균기A 가동상태	0 or 1
MIXB_PASTEUR_STATE (독립변수)	살균기B 가동상태	0 or 1
MIXA_PASTEUR_TEMP (독립변수)	살균기A 살균온도	날짜와 시간에 따른 온도
MIXB_PASTEUR_TEMP (독립변수)	살균기B 살균온도	날짜와 시간 따른 온도
INSP(종속변수)	불량여부	OK / NG

2. 탐색적 자료 분석(EDA)

2) 각 변수의 기술통계표

	MIXA_PASTEUR_STATE	MIXB_PASTEUR_STATE	MIXA_PASTEUR_TEMP	MIXB_PASTEUR_TEMP
count	1.113500e+04	10255.000000	201423.000000	1.988020e+05
mean	5.032693e+04	633.200390	566.867528	1.862568e+04
std	5.286901e+05	6408.270847	69.061703	8.111731e+05
min	0.000000e+00	0.000000	0.000000	0.000000e+00
25%	0.000000e+00	0.000000	543.000000	5.420000e+02
50%	1.000000e+00	1.000000	570.000000	5.690000e+02
75%	1.000000e+00	1.000000	596.000000	5.950000e+02
max	5.603841e+06	65536.000000	772.000000	4.279501e+07

- 변수의 기술통계표를 통해 각 변수에 대한 정보를 파악할 수 있었다.

- 살균기 가동 상태는 0과 1의 값만 가질 수 있는데 그 이외의 값이 나왔기 때문에 해당 컬럼의 데이터 정제의 필요성을 느꼈다.

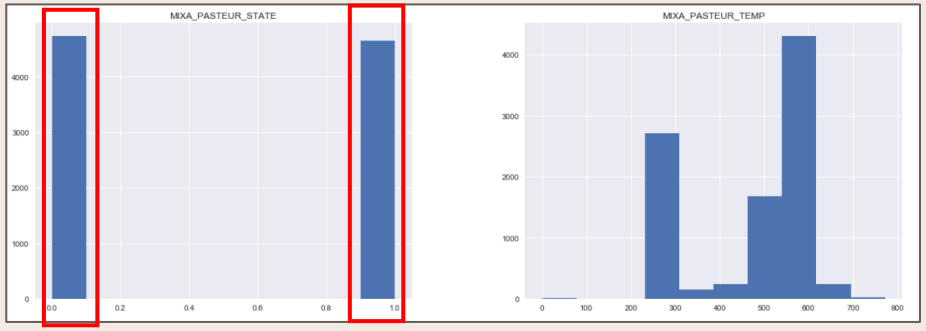
3) 변수 간의 상관계수 값

- 값이 1에 가까울 수록 상관관계가 높다고 볼 수 있다. 따라서 A,B 살균온도 변수는 상관계수 값이 0.92로 상관관계가 높다고 볼 수 있다.

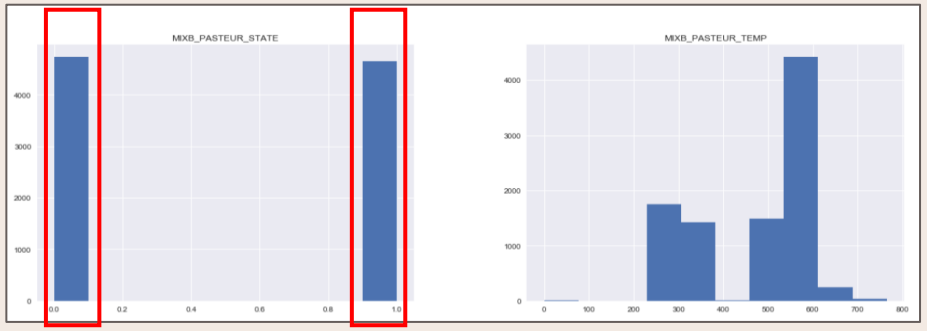
3. 데이터 전처리

1) 결측치 제거

<독립변수 값 분포 확인>



<살균기A 가동상태> <살균기A 살균온도>



<살균기B 가동상태> <살균기B 살균온도>

- 값이 없는 결측치가 있는 행을 모두 삭제하고 각 컬럼의 값이 될 수 있는 범위에 없는 값들을 제거 했다.
- 각 컬럼의 히스토그램을 통해 변수의 값의 분포를 확인할 수 있었다.
- 최종적으로 6개의 컬럼과 9383개 행으로 이루어진 데이터셋을 사용하여 모델링을 진행하였다.

4. 데이터 모델링

1) 사용 모델 선정

▸모델명: 의사결정 나무 모델

▸모델 설명: 머신러닝에서 분류 또는 예측 모형으로 사용되는 지도학습 방법론으로 나무처럼 가지를 뻗어가며 조건에 따라 값을 분류하는 알고리즘이다.

▸고려 사항: 학습데이터(train), 시험데이터(test) 비율 설정(test_size), 트리의 최대 깊이 설정(max.depth)에 따라 모델링 결과와 모델 성능이 달라진다.

▸모델 평가 지표: 오차 행렬을 바탕으로 한 정확도, 정밀도 , 재현율, F1-score, AUC 값

4. 데이터 모델링

2) 모델링

Model 1

- 테스트 데이터 비율 30%
  -> test_size = 0.3
- 모델 깊이 3
  -> max.depth = 3

Model 2

- 테스트 데이터 비율 40%
  -> test_size = 0.4
- 모델 깊이 3
  -> max.depth = 3

Model 3

- 테스트 데이터 비율 30%
  -> test_size = 0.3
- 모델 깊이 5
  -> test_size = 5

5. 결과분석

1) 모델 평가 지표

Model 1

- 정확도: 0.9185
- 정밀도: 0.9294
- 재현율: 0.9788
- F1: 0.9535

Model 2

- 정확도: 0.9300
- 정밀도: 0.9239
- 재현율: 0.9996
- F1: 0.9603

Model 3

- 정확도: 0.9897
- 정밀도: 0.9958
- 재현율: 0.9920
- F1: 0.9939

5. 결과분석

2) 모델 비교 평가

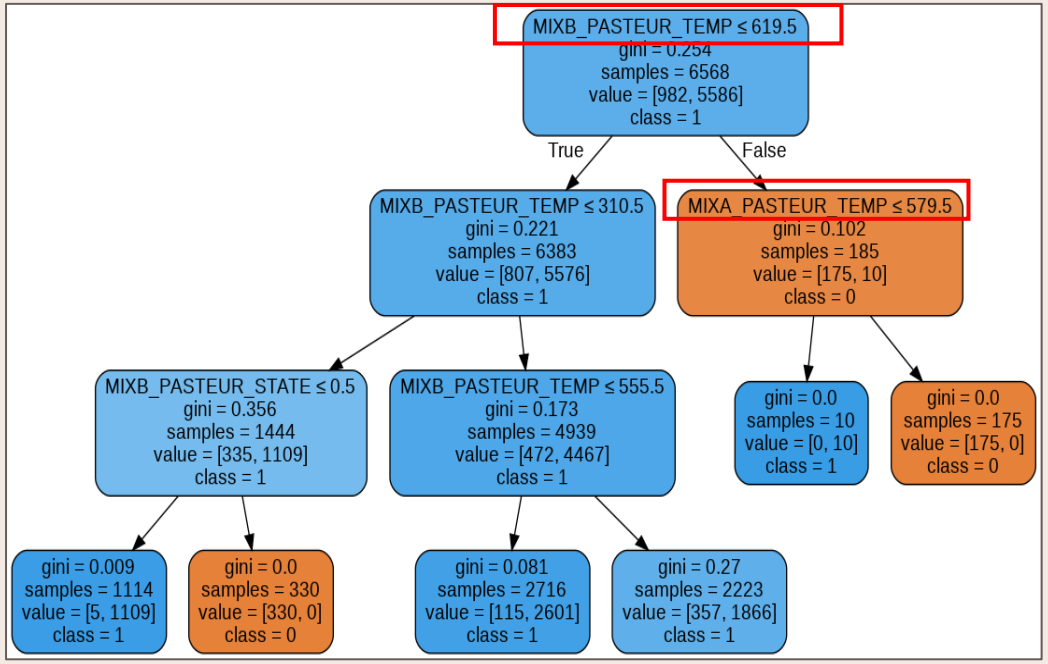
- 학습, 시험 데이터의 비중은 모델의 정확도, 정밀도, 재현율 등의 모델 성능에 영향을 준다. Test 데이터의 비율이 높은 모델2의 정확도와 재현율의 값은 커졌지만 정밀도의 값은 낮아 졌다. F1-score의 값 또한 모델2가 더 높은 것으로 보아 분류기의 예측율이 더 좋아 졌다고 볼 수 있다.

- 트리의 최대 깊이(분류 가지수)도 모델 성능에 영향을 준다는 것을 알 수 있었다. 하지만 너무 많은 나무 가지 수는 모델학습에 과적합 문제가 발생할 수 있다는 것을 알 수 있었다.

- 최대 모델 깊이를 5로 설정한 모델3의 정확도, 정밀도, 재현율, F1-score값이 모델1, 2 보다 높은 값을 가진다. 정밀도와 재현율의 경우는 서로 반비례하는 관계를 가져야 하는데 전체적인 모든 값이 높아진 것으로 보아 과적합 문제가 발생할 수 있는 모델로 볼 수 있다.

5. 결과분석

3) 모델 결과와 의미



<의사결정 나무 모델링 결과 – Model1>

- B 살균온도가 61.95도를 초과하고, A 살균온도가 57.95도 이하인 경우에 양품으로 분류했다.
- B 살균온도가 61.95를 초과하고, A 살균온도가 57.95 도를 초과하는 경우에 불량으로 분류했다.
- 살균공정에서 살균기B의 살균온도는 61.95℃ 이상, 살균기A의 살균온도는 57.95℃ 이하로 운영하는 경우에 양품을 생산할 수 있다고 예측할 수 있다.

5. 결과분석

4) 향후 응용 방향

▶ 의사결정나무모델을 사용해서 만든 AI모델을 사용하여 살균 공정에서 제품에 따라 온도조절을 하지 않아 발생하는 불량률을 줄일 수 있을 것이다.

▶ 제품에 따라 살균공정에 적용되는 온도가 다르기 때문에 각 제품에 따른 모델을 사용하여 공정을 운영한다면 공정은 다양한 제품을 효율적으로 생산할 수 있을 것이다.

▶ 센서 등을 사용하여 실시간으로 모니터링 하면서 더 정확한 온도 범위내에서 제조 공정이 운영된다면 불량품 처리 비용 등의 생산 운영비용을 줄일 수 있을 것이라 생각된다.

5. 결과분석

5) 느낀점

▶ 머신러닝의 분류 모델을 실무에 어떻게 적용하는지 알고 싶어서 본 프로젝트를 선택하였고, 프로젝트를 통해 의사결정 나무 모델을 사용하는 방법과 모델을 평가하는 방법에 대해 이해할 수 있었다. 어떤 모델을 사용하는지도 중요하지만 모델의 성능을 평가하는 방법을 이해하고 있어야 분석 결과를 가지고 산업에 잘 적용할 수 있다는 생각이 들었다.

▶ 데이터 정제를 하기 전에 데이터를 탐색해보는 단계가 중요하다는 것을 느꼈다. 어떤 데이터를 사용할 것인지 잘 이해하지 못한다면 데이터 전처리, 모델링 과정에 문제가 발생할 것이고 잘못된 결과를 야기 할 수 있겠다는 생각을 했다.