프로젝트명 : 사과 생산성 향상을 위한 품질 예측 모델 개발

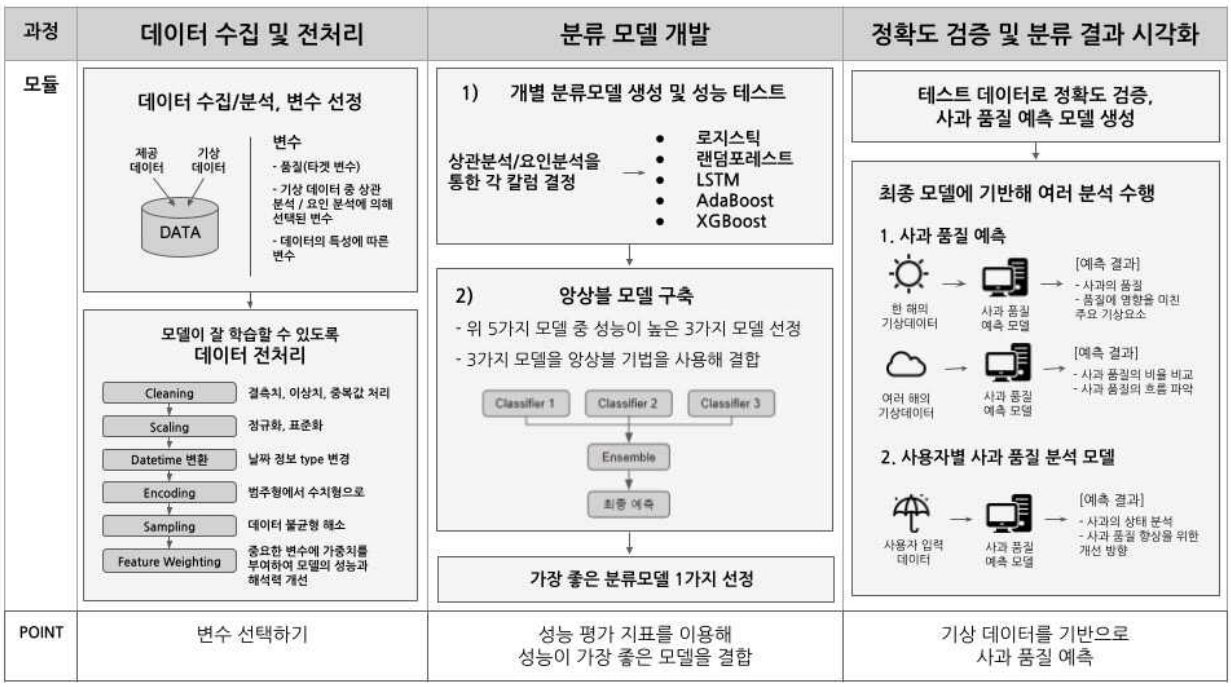
1. 문제 선택 이유

해당 문제를 선택한 이유는 사과가 세계적으로 많이 재배되고 소비되는 중요한 작물 중 하나이며, 기후 변화와 관련하여 사과의 품질에 큰 영향을 미치기 때문이다. 또한, [1]국립원예특작과학원 온난화대응농업연구소의 연구결과에 따르면 기후 변화의 영향을 가장 많이 받는 과일은 사과이며, 오는 2050년대에는 강원도 일부 산지에서만 사과를 재배할 수 있을 것이라고 예측되고 있다. 이러한 상황에서 사과 생산에 대한 대응 방안이 필요하다고 생각했으며, 사과 품질 예측을 통해 이를 해결하고자 한다.

사과의 품질과 기후 요소들 간의 상관관계를 파악한다면, 농가들은 더욱 효율적인 생산 방법을 도입하고 농작물 관리에 있어 더 나은 결정을 내릴 수 있을거라 생각한다. 또한, 소비자들은 예측 가능한 품질의 사과를 구매함으로써 만족도가 향상되며, 이는 사과 시장에 긍정적인 영향을 끼칠 것이다. 따라서 해당 문제는 농가와 소비자 모두에게 다양한 이점을 제공할 수 있어 의미있는 연구라고 생각하였고, 이를 주제로 선택하였다.

2. 문제해결 모델 개발 계획

- 최종 요약 자료



1) "'기후변화 영향' 가장 많이 받는 과일은?", news1, 2022년 11월 24일, https://www.news1.kr/articles/4875362

<데이터 분석 및 AI 모델 개발 과정>

사과 종류(후지/홍로)별로 사과 품질 예측 모델과 사용자별 사과 품질 분석 AI 모델을 개발한다. 사과 품질 예측 모델은 데이터 분석을 통해 사과의 품질을 예측하고 사과의 품질에 영향을 미친 요인 등을 알려주는 모델이다. 사용자별 사과 품질 분석 AI 모델은 사용자가 사과를 기르고 있을 때, 현재까지의 기상 데이터를 기입하면 해당 사과의 품질을 결정하는 주요 기상 데이터를 자동으로 분석하여 현재 사과의 상태를 예측하고 개선 방향을 제공한다. 해당 모델을 개발하기 위해 필요한 전처리와 성능 개선 단계를 거쳐 진행한다. 모델 개발 전 사전 단계는 다음과 같다.

1. 사과 종류(후지/홍로)별로 나눠 각각 모델을 생성
2. 논문 및 자료를 토대로 (후지/홍로) 사과의 품질(특/상/보통) 기준을 선정
3. 해당 기준에 따라 제공 받은 사과 데이터의 품질을 결정하여 타켓 변수를 생성
   (출처: 중부매일 - 충청권 대표 뉴스 플랫폼(http://www.jbnews.com)) 기사에 따르면 현재 국내에서 재배되는 사과 품종은 후지 68%, 홍로 15%이다. 후지는 소비자의 수요가 많기 때문에 소비자의 선호도를 고려한 사과 품질 등급 기준을 사용한다. 본 자료는 '소비자의 선호를 반영한 등급 표준화' 논문을 참고하였다.)

〈표 5〉 소비자 선호를 반영한 사과(후지) 품질의 결정경계

	당도 (°Bx)	당산비 (%)	무게 (g)	경도 (kgf)	색깔 (Hunter a)	모양 (%)	WTP[15] (원/kg)
1과 2+3등급 경계	11.736	35.523	166	5.631	29.164	75.135	5,200
2+3과 4등급 경계	10.943	35.653	166	5.629	15.900	75.135	3,577

반면에 홍로 사과는 소비자의 수요가 비교적 적어 소비자 수요 기준에 따른 조사가 불충분하다. 따라서 농촌진흥청 자료를 활용하여 품질 기준을 결정한다.

표 1. 품질 등급 기준표

구분	농산물 품질 관리원		농촌진흥청 'Top fruit'
항목	특	상	
고르기	무게 구분 표에 따라 다른 것이 섞이지 않음		
무게	후지 : 250g이상 홍로 : 250g이상	후지 : 215g이상 홍로 : 215g이상	홍로 : 320g±10%
색택	품종고유색택 후지 : 60%이상 홍로 : 70%이상	품종고유색택 후지 : 40%이상 홍로 : 50%이상	품종고유색택 후지 : 70%이상 홍로 : 80%이상
당도	후지 : 14°Bx이상 홍로 : 14°Bx이상	후지 : 12°Bx이상 홍로 : 12°Bx이상	후지 : 14°Bx이상 홍로 : 14°Bx이상
신선도	육기가 다고 과피 수축현상이 없음	육기는 다소 약하나 과피 수축현상이 없음	육기가 다고 과피 수축현상이 없음
중 결점과	없음	없음	없음
경 결점과	없음	10%이하	없음

4. 8개 지역 기상 데이터 수집 (군위, 청송, 영주, 장수, 거창, 충주, 포천, 화성)
   - 월별 최저기온, 최고기온, 평균기온, 평균풍속, 최대풍속, 일조시간, 평균 강수량, 최대 강수량, 습도, 일사량, 일조율, 풍향, 평균 지면 온도
5. 상관 분석 방법과 요인 분석 방법을 통해 중요 변수 5개 추출하여 모델 생성

   1) 상관분석에 따른 모델
   • corr 메서드를 사용해 사과 품질 칼럼과 다른 칼럼과의 상관계수를 추출한다. 이때

피어슨 상관계수는 두 변수가 모두 연속형 변수일 때만 사용 가능하므로, 순위형 변수의 상관계수를 구하는 스피어만 상관계수 혹은 켄달의 타우를 이용하여 구한다.

- 사과의 품질과 상관계수가 가장 큰 기상 요인 변수 5개만 추출하여 이용한다.

2) 요인 분석에 따른 모델

- 요인 또는 잠재 변수는 공통적인 반응 패턴을 갖는 여러 관측 변수와 관련이 있다. 분석을 통해 관측 변수 간의 분산을 설명하고, 변수의 수를 줄여서 데이터 해석에 도움이 된다.
- KMO 검정을 이용하여 요인 분석에 대한 데이터의 적합성을 측정한다. 각 관측 변수와 전체 모형에 대한 적합성을 결정한다. 모든 관측 변수 간의 분산 비율을 추정하여 0에서 1사이의 출력값을 얻는다. 이때 0.6 미만의 KMO 값은 부적합으로 간주한다.

6. 데이터 특성에 따라 결정한 새로운 변수 생성

1) 수확 '월' 별로 결정된 성장 시기 분류 변수

EX) 화아분화/성숙착색/자발휴면 (출처 :사과 | 농사로 (nongsaro.go.kr))

2) 이상 기후, 자연재해 유무 변수

EX) 홍수, 장마 기간, 태풍, 폭염 횟수

3) 지역마다 수확 시기에 영향이 있기 때문에 지역&수확 '월'을 합친 변수 생성

<EDA 및 가설>

0. 제공받은 사과 데이터 중 '지역' 변수와 본 팀이 생성한 품질 타켓 데이터 시각화를 통해 지역별 품질 차이를 확인한다.
1. 봄철(3,4,5월) 기온이 낮으면 착과 수와 과실의 무게에 부정적 영향을 미친다.
2. 과실 수확기의 많은 강수량은 과실 비대에 긍정적 영향을 미친다.
3. 수확기 최저 기온이 높으면 병충해로 인해 과실의 품질이 저하될 것이다.
4. 과실 비대기의 고온은 과실 비대에 부정적 영향을 미친다.
5. 여름철 (6,7,8월) 일조 시간은 과실 비대에 좋은 영향을 주며, 당도가 높아질 것이다.

<전처리>

상관분석에 따른 방법과 요인 분석에 따른 방법을 활용하여 변수 상관관계 분석 후 최종적으로 사용할 변수를 선택한다. 데이터의 결측값, 이상치, 중복값을 처리하고 정규화 및 표준화를 진행한다. 날짜 데이터를 datetime 타입으로 변경하고 인코딩도 함께 진행한다. 그리고 자료의 불균형 해소를 위한 오버샘플링 또는 언더샘플링을 하고 사과의 품질을 결정하는 데에 중요한 변수는 가중치를 부여하는 단계를 거친다. 이 외의 추가적인 전처리는 데이터 분포를

확인하며 진행할 예정이다.

<예측 모델>

데이터 분석을 통한 사과 품질 예측 모형은 다음 세 가지 모델을 사용한다. 앞서 설명한 요인 분석/상관분석 기법을 통해 사과 품종 별로 각 5개의 모델을 생성한다. 가장 성능이 좋은 3가지 모델을 앙상블하여 최종 모델로 결정한다.

- LSTM : 시계열 데이터에 대한 예측 성능이 우수하기 때문에, 시계열 기상 데이터에 적합할 것으로 예측하여 선정
- AdaBoost : 과적합 발생 확률이 낮고, 모델 수행 시간이 빠르며, 우수한 성능을 나타내기 때문에 선정
- XGboost : 높은 예측 성능과 중요한 피처 특징을 감지하여 작동하기 때문에 선정
- Logistic : 모델 특성상 해석에 용이하여 선정
- Random Forest : 과적합을 방지하고, 높은 성능을 나타기 때문에 선정

<개발 환경>

운영체제는Windows 11이며, 패키지는NumPy, Pandas, Matplotlib, Scikit-learn, seaborn, factor_analyzer 등을 사용할 예정이다. 프로그래밍 언어는 Python을 사용한다.

3. 예상하는 결과물 형태와 내용

1. 사과 품질 예측 모델

1) 한 해의 기상 데이터를 입력한 경우

=> 사과의 품질(특, 상, 보통), 사과 품질의 비율 그래프 출력, 사과의 품질에 영향을 미친 주요 요인을 알려주는 모델이다. 농가들은 해당 모델을 이용하여 사과 품질에 영향을 미쳤던 최적의 기상 요소와 시기를 제공받아 내년 농업에 반영할 수 있도록 돕는다.

2) 여러 해의 기상 데이터를 입력한 경우

=> 사과의 품질 비율을 비교해주며 사과의 품질이 향상되고 있는지 떨어지고 있는지를 알려준다. 또한, 한 해의 기상 데이터를 입력했을 경우와 마찬가지로 사과 품질에 영향을 미친 주요 요인도 함께 알려주는 모델이다. 이와 같은 모델을 통해 해당 연도의 사과 품질에 대한 예측 뿐만 아니라 사과 품질의 흐름을 파악할 수 있도록 한다.

위에서 언급한 어떤 기상요인이 사과의 품질에 결정적인 역할을 하고, 주요 요인인지에 대한 분석은 XAI알고리즘을 활용한다. 이 알고리즘은 인공지능 모델이 특정 결론을 내리기까지 어떤 근거로 의사결정을 내렸는지를 알 수 있게 설명 가능성을 추가하는 기능이 있다. 모델의 feature importance에서 발생하는 변수 간의 의존성을 간과하는 문제를 해결하기 위해 XAI 알

고리즘의 SHAP 기법을 사용한다. 이는 의존성까지 고려해 모델에 미치는 영향력을 계산한다. 또한 음의 영향력도 계산할 수 있어 넓은 범위를 볼 수 있다. 현재 독립변수들은 기상데이터이기 때문에 변수 간의 의존성이 존재할 것이다. 따라서 SHAP을 사용하여 변수의 영향력을 살펴보고, 사과 품질에 영향을 미친 feature를 제공받는다.

2. 사용자별 사과 품질 분석 AI 모델

사과의 발아기부터 수확기까지 기상 데이터가 누적되어갈 때 품질에 크게 영향을 미쳤던 변수의 데이터가 있음에도 불구하고, 품질이 좋았던 과거 데이터를 토대로 앞으로의 대처 방안을 계획할 수 있도록 한다. 해당 모델은 1번 모델에서 성능이 제일 좋았던 모델을 이용하여 다음과 같은 단계를 거쳐 진행된다

1) 상관분석을 통해 사과 품질에 영향을 미치는 각 월의 기상 요인 분석

   먼저 주요 기상 요인을 추출하기 이전의 기상 데이터 프레임에서 상관분석을 진행하여 각 '월'마다 주요 기상 요인을 분석한다.
   => EX) 1월 습도, 2월 최저온도, 3월 최고온도, 4월 강수량

2) 주어진 데이터에서 사과의 품질(특/상/보통)별로 그룹을 나눈 뒤, 1)의 기상 요인이 어떻게 분포되어 있는지 확인
   => EX) 특/상.보통 집단의 기상요인(습도)에 대한 시각화 진행

3) 사과 품질이 구분되는 월별 기상요소(ex.습도)의 기준치를 각각 정한다.

4) 사용자가 입력한 데이터에서 중요한 기상요인이 기준치 이하인 것이 있는지 확인한다.
5) 있을 경우, 이전 시기 데이터에서 해당 기상요인이 비슷한 양상을 보이는 데이터 중 가장 품질이 좋은 데이터(EX. 특/상 이라면 품질이 '특'인 데이터선택)만 추출한다.

6) 추출한 데이터를 예측 모델을 돌리고, SHAP 기법을 사용하여 영향을 미친 feature를 도출한다.

7) 해당 분석을 농가에 제공함으로써 사과의 상태와 사과 품질 향상을 위한 개선 방향성을 제공한다.

Ex) '월'마다 품질에 영향을 미치는 요소가 무엇인지 상관분석을 통해 사전 분석 후, 농민들이 사과를 기른 현재까지의 기상 데이터를 넣었다고 하자. 3월의 평균 기온이 기준치보다 낮았다면, 기존에 학습한 모델에서 3월의 평균 기온이 낮았던 데이터들 중 품질이 비교적 높은 사과가 있다면 어떤 기상 요소가 품질을 향상시킬 수 있었는지를 분석해 대처 방안으로 사용한다.

<모델 기대 수준>

예측 결과를 바탕으로 지역별 농업 생산성 향상 방안을 도출한다. 어떤 지역에서 어떤 시기에 특정 기후 조건이 가장 적합한 사과를 재배해야 하는지를 제안한다. 농업 생산 뿐만 아니라 소비자 입장에서도 예측 결과를 활용하여 사과 구매에 도움을 주는 정보를 제공할 수 있다.

수치적인 수준으로는 데이터 규모, 불균형 정도를 반영하여 accuracy, recall, precision, f1score 등을 사용할 예정이다. 해당 모델의 성능 평가 지표로는 정확도(accuracy)를 사용할 때

모델 개발을 통해 기대하는 수준은 정확도 0.7~0.9사이로 설정하였다. 별도의 test data가 제공되지 않는다면, 주어진 데이터를 이용하여 validation data를 설정해 학습 과정 중 성능을 측정할 것이다. 해당 모델을 이용하여 높은 정확도를 통해 소비자와 공급자의 신뢰를 높일 수 있을 것이다.

4. 활용 방안

<활용 방안>

1. 농업 생산성 향상
   AI 모델을 농가들에게 제공하여 사과 품질에 영향을 미치는 최적의 기상 요소와 시기를 알려줌으로써 생산성을 향상시킬 수 있다. 이를 통해 농가들은 더 나은 사과를 생산하고 농작물 관리를 최적화함으로써 수익을 증대시킬 수 있다.

2. 자원 효율성 증대
   AI 모델을 활용하여 효과적인 농작물 관리 방안을 제시함으로써 물, 비료 등의 자원을 효율적으로 사용할 수 있다. 이는 환경 보호와 더 지속가능한 농업을 실현하는 데 도움이 될 것 이다.

3. 소비자 만족도 향상
   AI 모델을 사용하여 사과 품질을 예측하고 향상시킬 수 있으므로, 소비자들이 더 맛있고 신선한 사과를 즐길 수 있다. 이는 소비자 만족도를 높이고 사과 시장에서의 경쟁력을 강화하는 데 도움이 될 것이다.
4. 농작물 관리와 위기 대응
   AI 모델을 활용하여 사과 품질 예측과 기후 변화에 대응하는 방안을 제시함으로써 농가들은 예상치 못한 기후 변화에 대응할 수 있다. 이는 농작물 관리와 생산 안정성을 향상시키는 데에 도움이 된다.

<향후 연구 방향>

1. 기후 예측 데이터 포함
   기후 예측 데이터를 모델에 포함하여 더 정확한 사과 품질 예측을 가능하게 할 수 있다. 과거의 기상 데이터 뿐만 아니라 미래의 기후 예측 정보를 활용함으로써 모델의 성능을 향상시킨다.

2. 개별 농가에 맞는 맞춤형 조언 제공
   AI 시스템은 개별 농가의 생산 환경과 조건을 고려하여 맞춤형 조언을 제공해야 한다. 각 농가의 특성에 따라 최적의 기상 요소와 관리 방안을 추천함으로써 개별 농가의 생산성을 최대화할 수 있다.

이러한 방향으로 연구를 진행함으로써 사과 산업과 농가, 소비자들에게 더 큰 이익과 가치를 제공하는 AI 모델을 개발할 수 있을 것입니다.