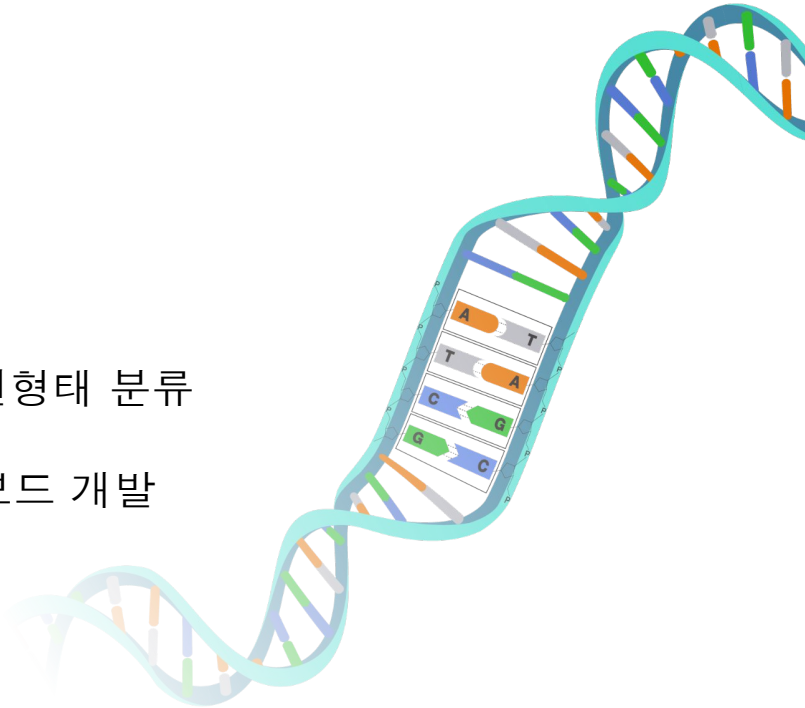
데이터 분석 포트폴리오



Project 1) Cluster analysis 를 통한 PCR 증폭곡선형태 분류

Project 2) PCR 검사 품질관리 현황 시각화 대시보드 개발

Project 3) DNA melting temperature 예측

작성자: 정화영

본 포트폴리오의 분석은 현업에서 실시했던 데이터 분석 결과 및 시각화 자료를 변형/blur 처리 하여 중요한 정보가 유출되지 않도록 하였습니다.

Project1) Cluster analysis 를 통한 PCR 증폭곡선형태 분류

데이터 소개 및 분석 목적

컬럼명	설명
sample_ID	각 sample 별 ID
device	PCR 장비의 시리얼 번호
amp_info_1	Amplification curve 파라미터1
amp_info_2	Amplification curve 파라미터2
amp_info_3	Amplification curve 파라미터3
……	……
amp_info_10	Amplification curve 파라미터10

분석 목적

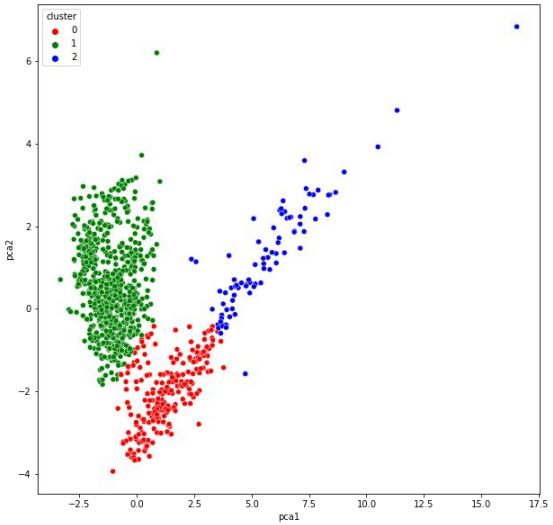
1) PCR 검사시, 양성 검체에서의 증폭 곡선 그래프는 sigmoid 형태. 하지만 생화학적, 혹은 장비 품질의 이슈로 인해 비정상적 증폭 패턴을 보이는 경우가 있음
2) 이 때 위양성등 판독의 오류 발생할 수 있음
3) 동일 장비에서 반복해서 비정상 증폭 패턴이 발생하는경우 장비 점검, 교체 필요
4) 육안으로 비정상 증폭 여부를 체크 하는 것은 비효율적
5) 비정상적 증폭 패턴을 효율적으로 detect 할 수 있도록 clustering 분석 실시

분석 설계

1) Deep learning-based clustering 모델인 autoencoder 생성
2) 적정 cluster 개수 파악
3) 단일 cluster 내 동일 장비 비율 파악

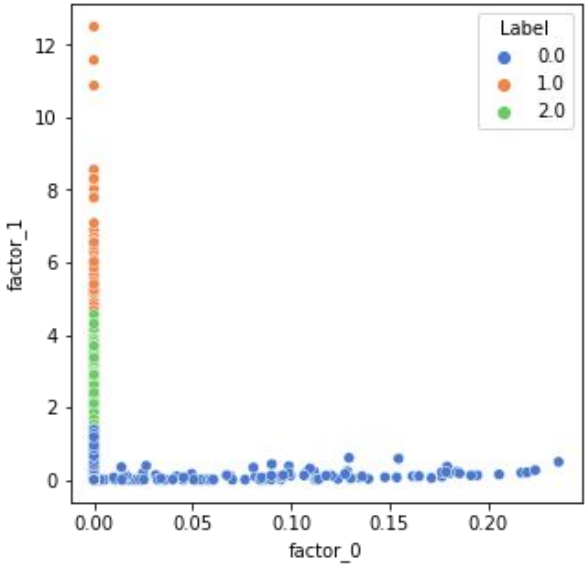
사용 tool: Python, tensorflow, Keras

Project1) Cluster analysis 를 통한 PCR 증폭곡선형태 분류



방법1: K-means & PCA

K-means로 cluster를 labeling 한 후
2차원 PCA 공간에서 시각화



방법2: Autoencoder & K-means

Autoencoder 로 latent vector 를 생성 후
K-means로 fitting

2가지 다른 방법으로 cluster하여
어떤 방법이 적절한지 비교,
결정.

방법1: K-means & PCA

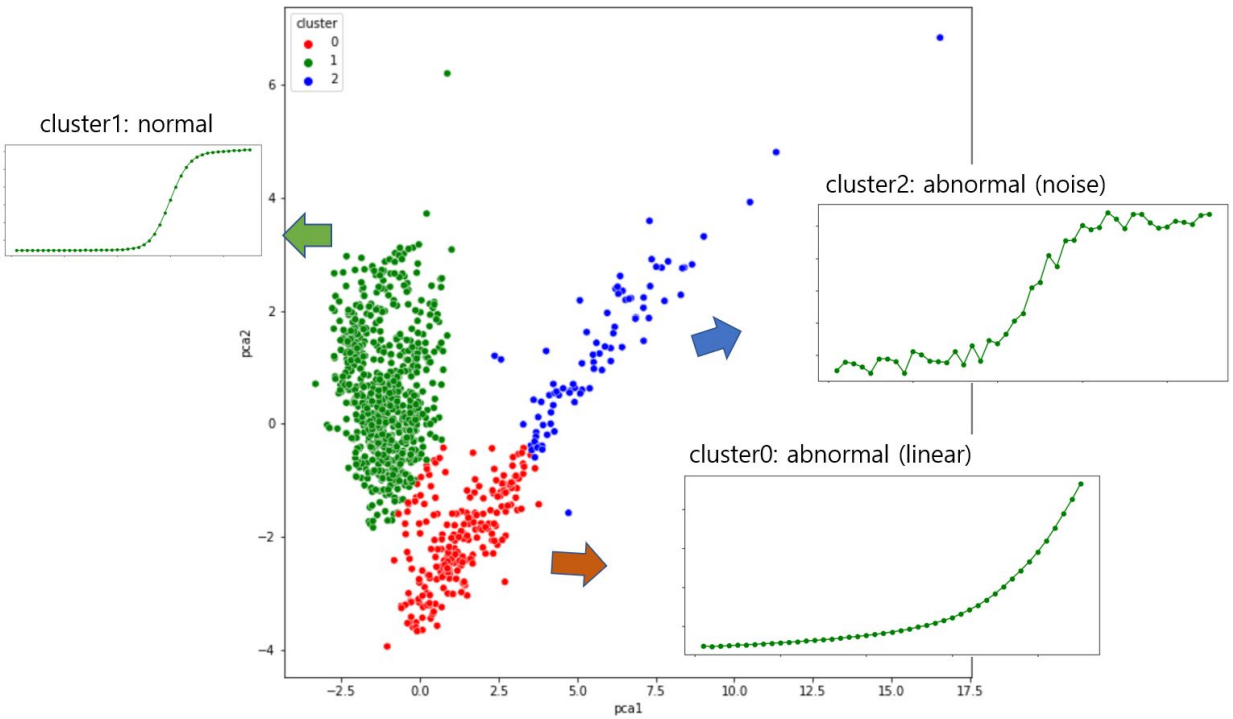
방법2: Autoencoder & K-means

결과
K-means & PCA 를 이용한 경우
Cluster 가 적절하게 분류 됨.

Autoencoder & K-means 이용
시 Cluster 가 x축과 y축에
몰려있음.

=> K-means & PCA 방법 선택

Project1) Cluster analysis 를 통한 PCR 증폭곡선형태 분류



결과

K-means & PCA 를 이용한 경우
Cluster 가 3개로 분류되며
Cluster 별로 다른 증폭곡선 형태

Cluster 1: 정상 증폭
Cluster 0: 비정상 증폭 (늦은 증폭)
Cluster 2: 비정상 증폭 (noise)

Cluster 0,2는 모두 비정상 증폭에 속하지만 형태가 다름.
Cluster 0- 증폭이 늦게 되는 형태
Cluster 2- noise가 심한 형태,
80%가 특정 장비에서만 발생.

각 비정상 증폭의 분류에 맞는 품질관리 조치가 실시 될 수 있는 방안 마련.

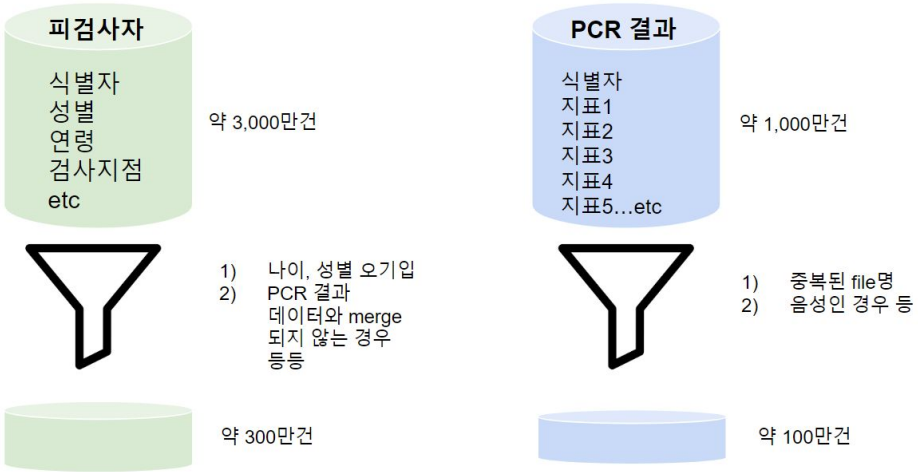
Project2) PCR 검사 품질관리 현황 시각화 대시보드 개발

연구 배경

- 코로나로 인해 PCR 검사량이 급증했으나 검사 현장에서는 전문적인 검사 보조 인력이 부족
- PCR 검사 결과에 이상이 있을 경우진단의가 신속하고 효율적으로 이에 대해 조치를 취해야 함
- PCR 검사의 체계적인 검사 품질 관리 시스템 필요하다는 고객사의 요청이 있었음

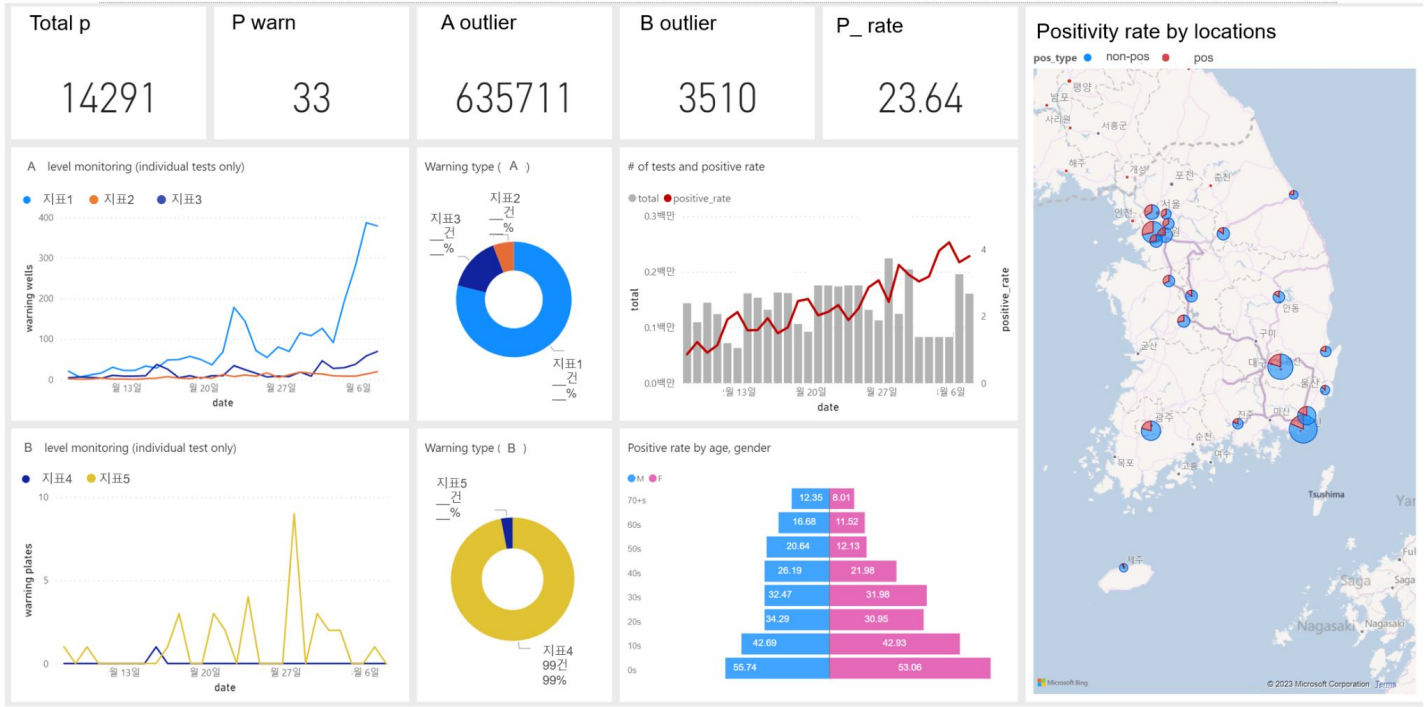
연구 방법: 데이터 전처리

2종류의 데이터 (피검사자, PCR 결과) 전달 및 클리닝



연구 결과

PowerBI 로 구축한 대시보드 메인화면



품질관리의 정도를 나타내는 5가지 지표 선정

각 지표별 분포에 기반하여 이상치 기준 선정 (Q1 − 1.5*IQR 미만, Q3+1.5*IQR 초과)

이상치 여부 판별하여 일별 지표별 이상치 개수 count, 비율 계산

연구 결과

대시보드 상세화면– 피검사자 정보 Part

날짜 구간, 양/음성 판독의 기준 값을 사용자가 선택하면 이에 맞는 결과가 나타남

연구 결과

대시보드 상세 화면– 품질관리 지표별 분포 및 이상치 기준 시각화

날짜 구간, 이상치 기준 값을 사용자가 직접 선택하면 이에 맞는 결과가 나타남.

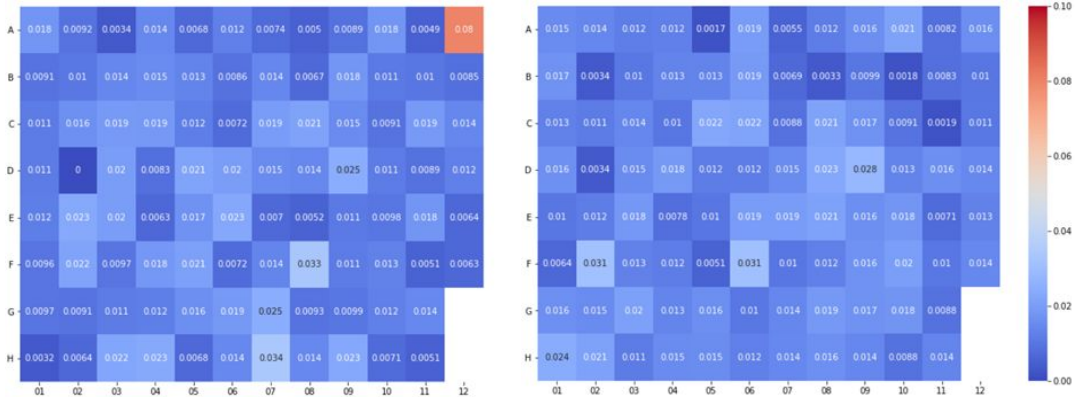
연구 결과

대시보드 상세 화면– 5가지 지표별 분포 시각화, 이상치 기준 선정 =>장비별 이상치 개수 count

장비별 이상치 건수 및 비율

장비n 에서 높은 비율로 이상치 발생– 장비 점검 필요성 알려줌

장비 내 특정 파트 이상치 건수 및 비율
시각화 heatmap



왼쪽 장비 내 A12 part (붉은 색) 에서 상대적으로 높은 비율로 이상치 발생– 장비 점검 필요성 알려줌

성과 및 결론

- PCR 검사 품질 개선을 위한 5가지 지표를 고객사와 공동 개발하여 이상치 발생 현황 시각화 대시보드 개발 => PCR 양음석 판독의 잠재적 오류 23% 감소
- 지점별 Quality Control 현황 비교하여 QC에 문제가 있는 지점 관리 강화
- 지점별 양성률 비교시, 특정 지점에서 양성률이 큰 곳은 검체 수집과정에서 오염 있었을 가능성 제시
- 장비 노후화 또는 고장 tracking: 특정 장비, 또는 장비 내의 특정 파트에서 이상치 반복적으로 발생할 경우 장비 교체 제안.
- 고객사 2곳을 대상으로 QC 모니터링 대시보드 프로토타입 시연회 개최 이 대시보드의 필요성에 대한 동의 이끌어냄
- 본사 연구과제 연구종료심의에서 대시보드 과제가 실제로 고객사에 사용될 수 있도록 후속과제화 의견
- 해외 고객사를 타겟으로 한 대시보드 영문 매뉴얼 영상 제작

Project3) DNA melting temperature 예측

데이터 소개 및 분석 목적

컬럼명	설명
DNA_ID	각 DNA별 서열 ID
length	DNA별 서열의 길이
Seq_info_1	DNA의 특정 파트의 서열 정보
Seq_info_2	DNA의 특정 파트의 서열 정보
Seq_info_3	DNA의 특정 파트의 서열 정보
……	……
Seq_info_n	DNA의 특정 파트의 서열 정보
DNA_FE	feature engineering 시 추가 된 변수
DNA melting temperature	DNA 가 녹는 온도

예측 목적

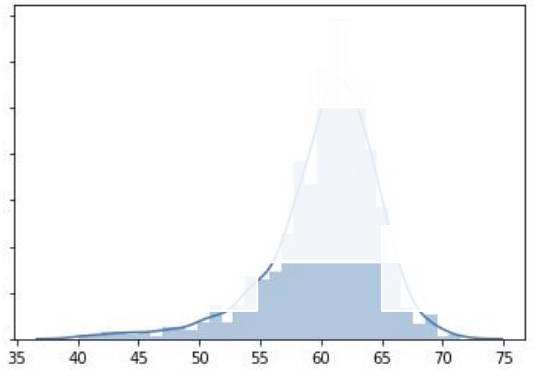
1) DNA의 녹는 온도를 예측 하는 것은 분자진단 제품 개발을 위한 기초단계
2) 기존에 타부서에서 수리적 최적화를 이용하여 예측
3) 더 다양한 방면에서 예측하여 어떤 방법이 제일 최적의 예측 방법일지 비교하기 위하여 머신러닝으로 접근

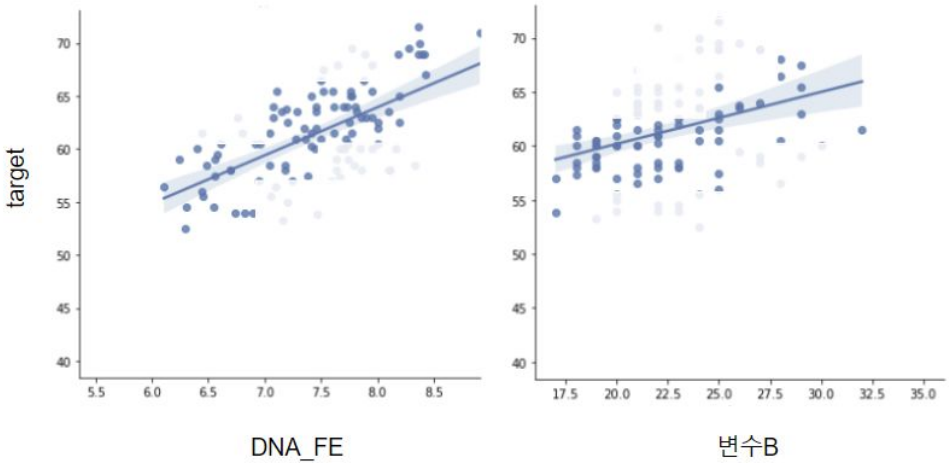
분석 설계

1) EDA를 통한 Feature, target 변수 특성 파악
2) Random forest 등의 모델 생성
3) Feature engineering: 논문 약 20편을 참고하여 머신러닝으로 DNA melting temperature 예측할 시 추가하면 좋을 feature (DNA_FE)를 추가
4) target 예측에 가장 크게 기여하는 변수 파악

사용 언어: Python

Project3) DNA melting temperature 예측
target 및 feature 변수 특징





target
DNA melting temperature 의 분포는 오른쪽으로 치우치고 꼬리가 왼쪽으로 길게 뻗어있는 형태.

feature
DNA_FE, 변수B 가 증가할 수록 DNA의 melting temperature 가 증가하는 것으로 나타남.

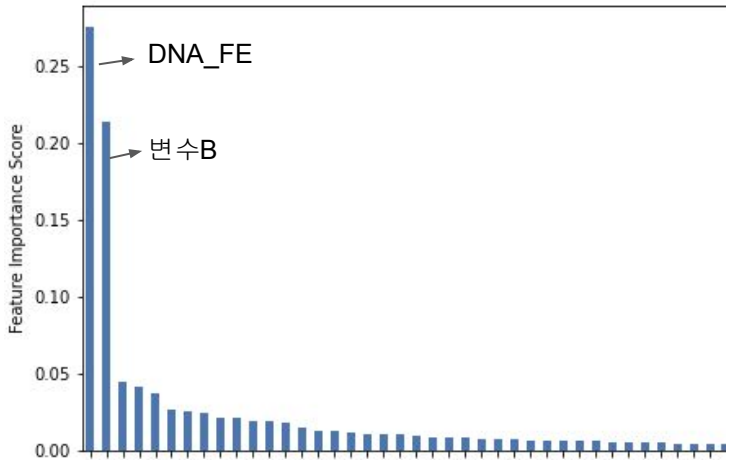
Project3) DNA melting temperature 예측
Model 분석 및 결론

사용 model: Random forest (with DNA_FE) vs Random forest (wo DNA/FE)

RF with DNA_FE MSE: 26.171

RF wo DNA_FE MSE: 41.308 => DNA_FE 변수를 포함한 경우 MSE 57.8% 감소

타부서 수리적 최적화에서의 MSE는 1.64



결론

- DNA_FE 와 변수 B 가 가장 중요한 변수. But, 수리적 최적화 방법에 비해 머신러닝 방법은 MSE가 높게 나타났음
- 분석 방향 제시
  - 향후 DNA의 melting temperature는 수리적 최적화 방법으로 하는 것을 권장.
  - 혹은 데이터를 더 많이 모아서 추후에 다시 머신러닝을 적용해 보는 것을 고려.
  - target 변수의 distribution을 고려하여 log 또는 sqrt transformation을 해 볼 것.
  - 머신러닝으로 분석 시 DNA_FE 변수는 반드시 넣을 것.

인사이트

- 머신러닝이 모든 예측 문제를 해결 할 수는 없음. 상황에 따라 최적화가 유리할 때, 머신러닝이 유리할 때가 존재 하며 이에 따라 적절한 분석 방법을 택해야 함.
- 머신러닝으로 분석 시 문헌 search 에 기반한 feature engineering 이 중요함.