데이터 분석 포트폴리오

1) image cluster 분석을 통한 PCR 검사품질관리 효율화

2) PCR 검사 품질관리 현황 시각화 대시보드 (PowerBI) 개발

3) DNA 서열 조합 및 길이 별 녹는점 예측

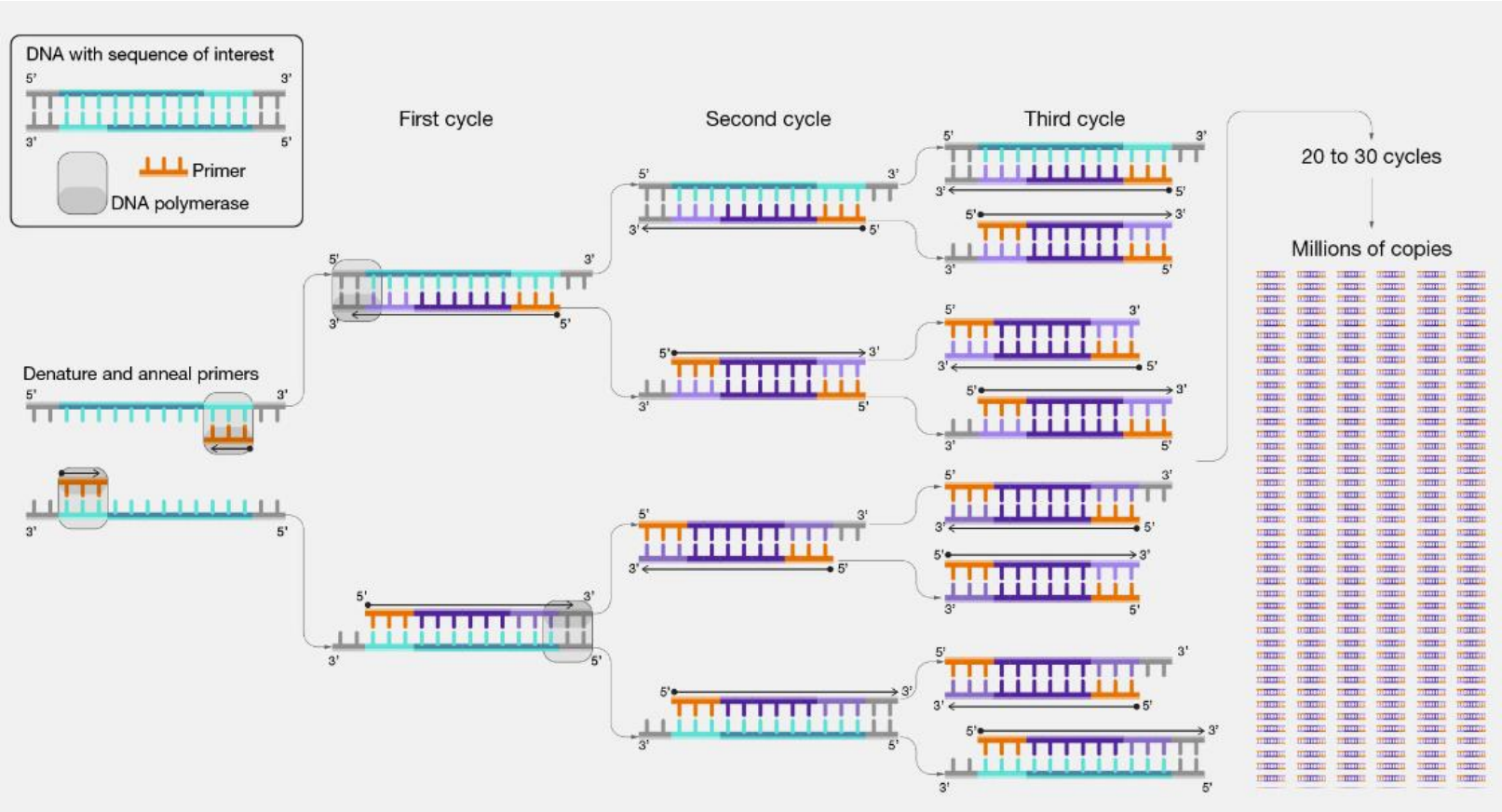
※ 본 발표자료에 기재된 정보는 실제 자료를 변형/가공한 것으로, 현 직장에서의 중요한 정보가 유출되지 않도록 하였습니다.

1) image cluster 분석을 통한 PCR 검사품질관리 효율화

※ 본 발표자료에 기재된 정보는 실제 자료를 변형/가공한 것으로, 현 직장에서의 중요한 정보가 유출되지 않도록 하였습니다.

Background

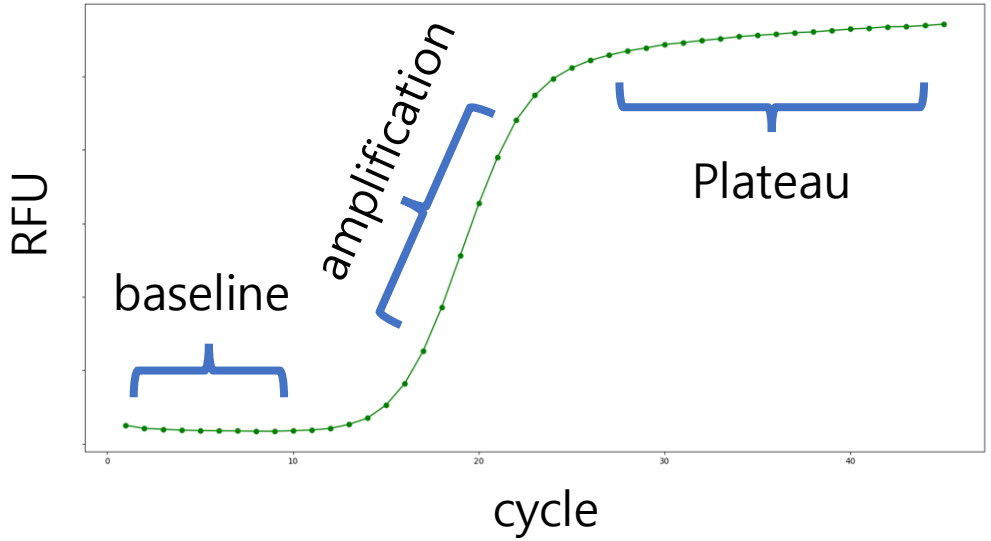
- PCR (Polymerase Chain Reaction)



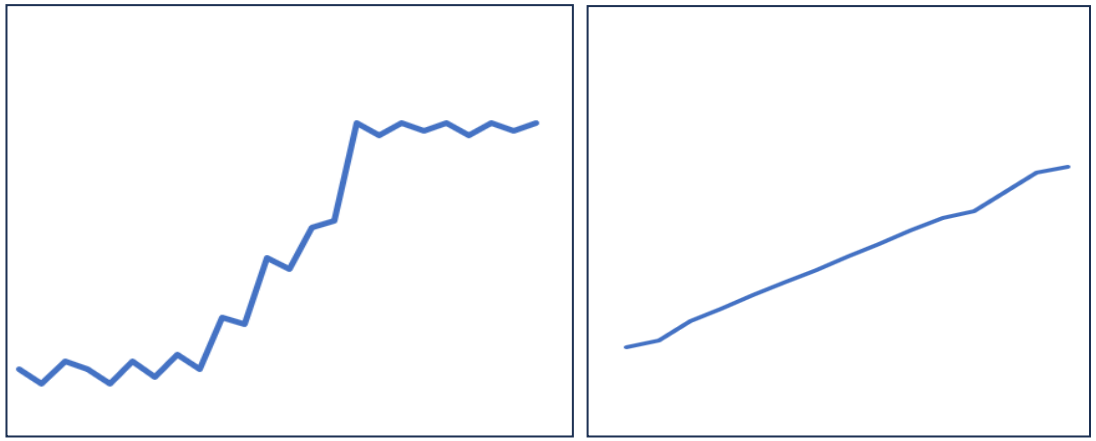
https://www.genome.gov/genetics-glossary/Polymerase-Chain-Reaction

Background

- Amplification curve 형태



정상증폭곡선



비정상증폭곡선

PCR 증폭곡선은 정상적인 경우 sigmoid 함수 형태를 보이지만 시약 또는 장비 품질 등 여러 요인에 의해 문제가 발생한 경우 sigmoid 함수 형태를 벗어난 다양한 형태를 보여준다.

Problem definition

- 검사량이 적을 때에는 임상검사자가 육안으로 쉽게 정상/비정상 여부를 파악하여 품질 관리를 할 수 있음.
- But, 코로나 팬데믹 등과 같이 검사량이 급증하는 경우 육안으로 확인하기에는 한계가 있음.
- 따라서 검사체계의 효율화를 위해 PCR 증폭곡선의 정상여부를 임상검사자에게 리포트하고, 나아가 비정상 증폭곡선으로 분류 된 형태를 다시 분류하여 그 형태에 맞는 이상패턴 원인에 대한 조치를 할 수 있게 할 필요가 있음.
- 현재까지 비정상 곡선의 패턴이 분류된 바가 없어 사전에 정의된 label이 없으며 비정상 곡선이 몇 개의 분류군으로 나뉠 지도 알려져 있지 않은 상태.

=> Cluster 분석 필요.

Objective

AS-IS

- 증폭곡선을 tracking 하는 임상검사자는 드묾
- 임상검사 현장으로부터 1달에 1~2건 정도 비정상증폭곡선에 대한 문의 받음.

TO-BE

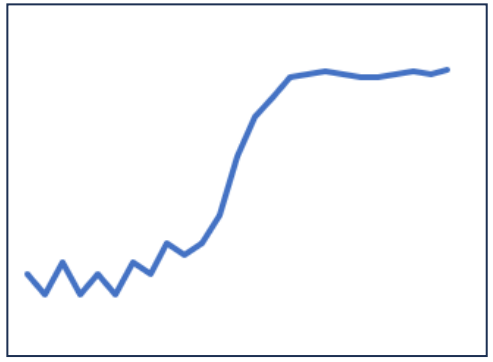
- 정상증폭곡선과 비정상증폭곡선을 자동으로 분류
- 비정상증폭곡선의 형태를 분류하여 각 형태에 맞는 적절한 조치가 취해지도록 함.

Expectation

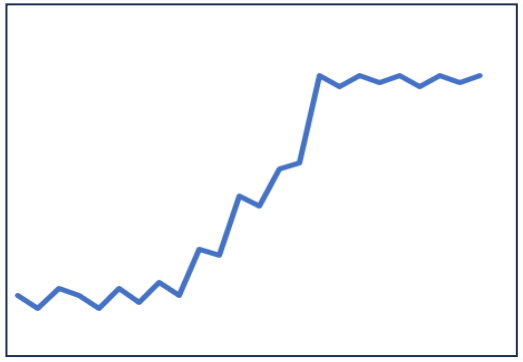
PCR 검사 품질 관리 효율화

- 검사자가 육안으로 증폭곡선 형태 확인 시 1개 검체당 10초 걸린다면, 1000건의 검체를 확인 시 10*1000초 (약 3시간) 걸릴 것으로 예상
- Cluster 분석으로 분류시 1분 내로 확인 가능 예상 => 99.4% 속도 향상
- Cluster 된 비정상 증폭 곡선의 타입에 따라 그에 맞는 조치를 취할 수 있음

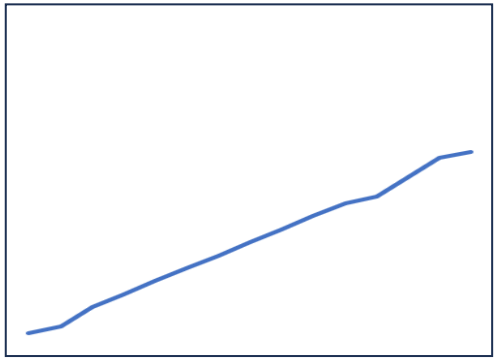
예)



타입 Baseline noise
조치 Template 희석



Entire noise
장비 수리/교체



Inefficient amplification
PCR cycle 온도 조절

How to cluster

- Step1: DBscan 을 이용하여 outlier detection

  (outlier 에 해당하는 검체의 증폭 곡선을 비정상증폭 곡선으로 간주)

- Step2: Outlier만을 대상으로 2가지 방법으로 cluster분석, 성능 비교.

  방법2) Image 분석: VGG16 이용하여 feature extraction 한 것을 standardize 하여 K-means 로 cluster.

About Dataset

- Data collection: 고객사와의 NDA, IRB 연구를 통해 임상데이터를 전달 받음. (2021년)
- Data storage: Azure cloud에 PostgreSQL DB형태로 데이터 적재
- 분석 대상: 2022-02-10 부터 2022-02-17 까지 양성 검체 70,000건
- Outlier detection 에서 사용된 Data structure

  Amplification curve 의 형태를 결정짓는 parameter 10가지
- Image cluster 에 사용된 Data

  Outlier 로 detect 된 검체의 amplification curve 의 이미지 파일

How to cluster

Why DBscan?

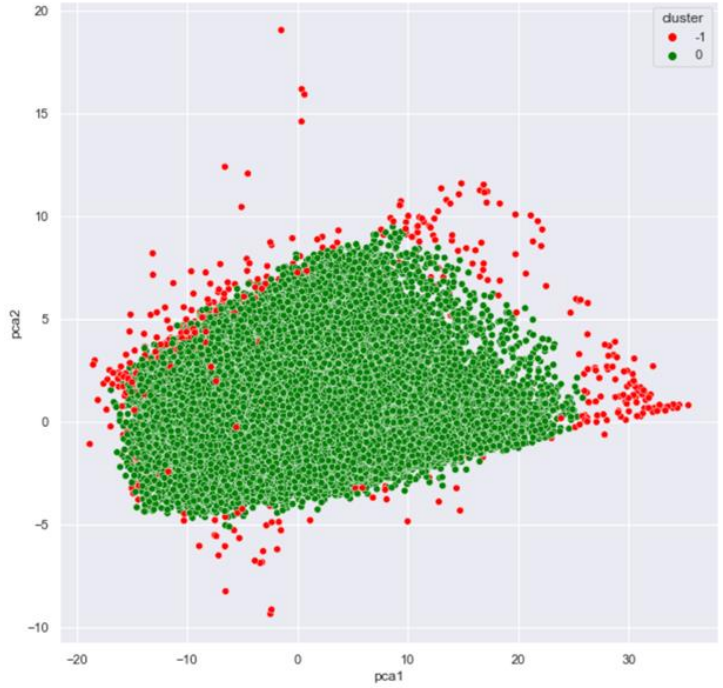
- 초기 조건에 제약이 덜함
- Cluster 개수를 지정하지 않아도 됨
- Outlier 에 대한 정보가 부족하며 각기 다른 cluster 에 속해 있을 수도 있는 dataset에 적합

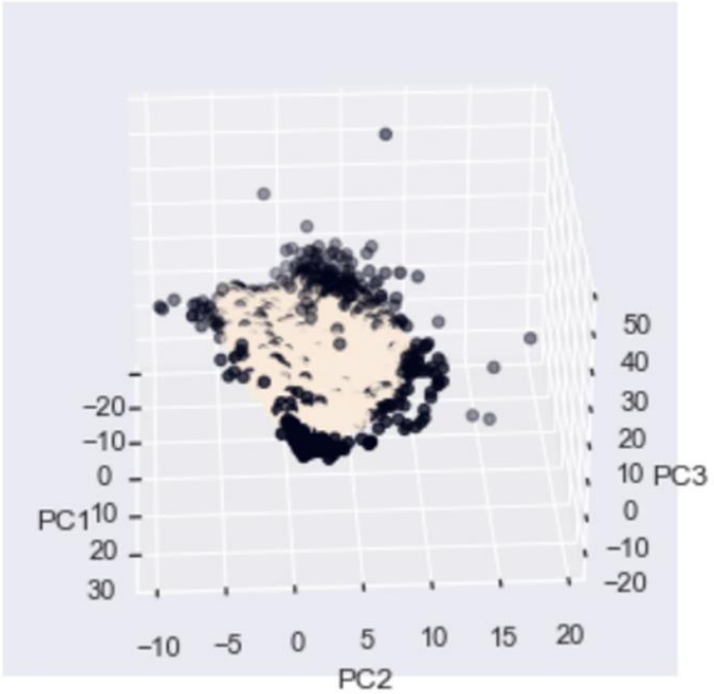
Why K-means?

- Autoencoder로 차원 축소 시 bottle neck 의 dimension을 3개로 적게 지정.
- 결과를 이해하기 쉽고 속도가 빠름.

Results: outlier detection by DBscan

- Outlier detection: 70,000 건 중 622건이 outlier 로 분류.





Results

Outlier 에 해당하는 amplification image feature를 VGG16 으로 extract
K-means 로 3 clusters 로 분류

Cluster 1
초반 cycle 급격 증가 패턴

Cluster 2
Baseline, plateau noisy 패턴

Cluster 3
늦은 증폭 패턴

Conclusion & Further plan

- VGG16 으로 image feature 추출, cluster 하여 실제로 이상 증폭 곡선의 형태가 유사도에 따라 다른 cluster 로 분류되는 것을 확인
- 임상검사 시스템에 적용시 효율적으로 단시간내에 비정상증폭곡선을 detect 하고 비정상증폭곡선의 형태 별 원인 파악을 추정 할 수 있는 단서 마련. (e.g. cluster1 의 경우 특정 장비에서만 발생. 장비교체 등 고려 가능)
- 보완할 점: 같은 cluster 내에 완벽하게 분리가 안되는 amplification curve 가 아직 남아있음. MobilenetV2, DEC 등 다른 방법 등 사용하여 성능비교 필요.

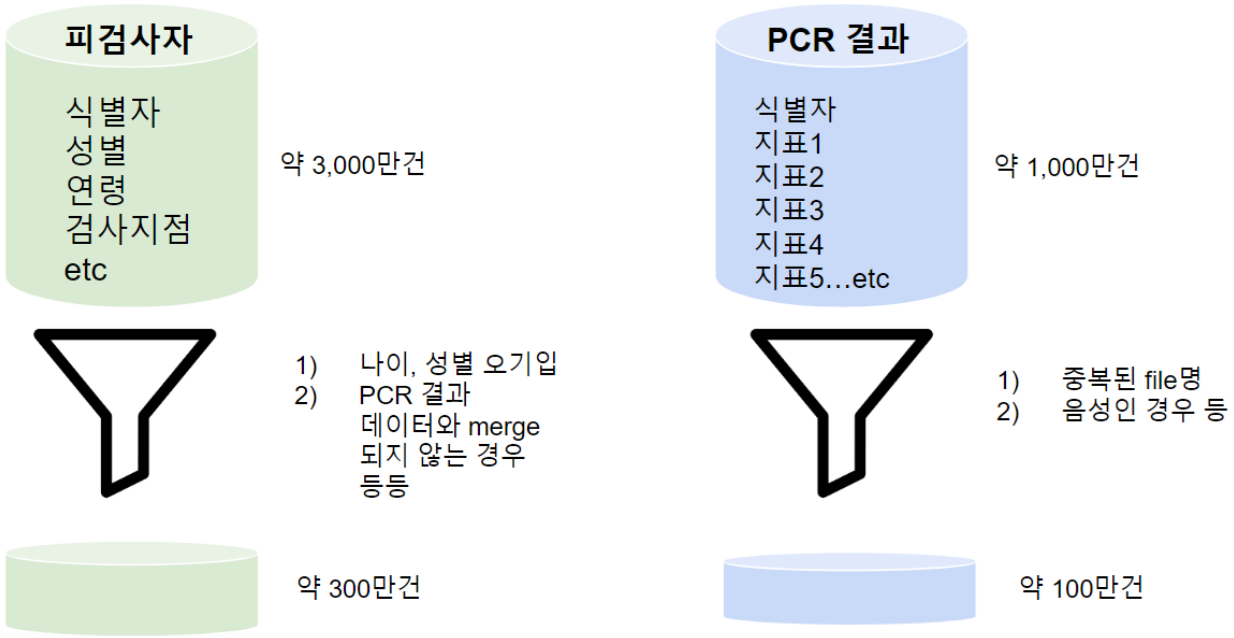
2) PCR 검사 품질관리 현황 시각화 대시보드 (PowerBI) 개발

※ 본 발표자료에 기재된 정보는 실제 자료를 변형/가공한 것으로, 현 직장에서의 중요한 정보가 유출되지 않도록 하였습니다.

Background

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

2종류의 데이터 (피검사자, PCR 결과) 전달 및 클리닝



Results: Dashboard development

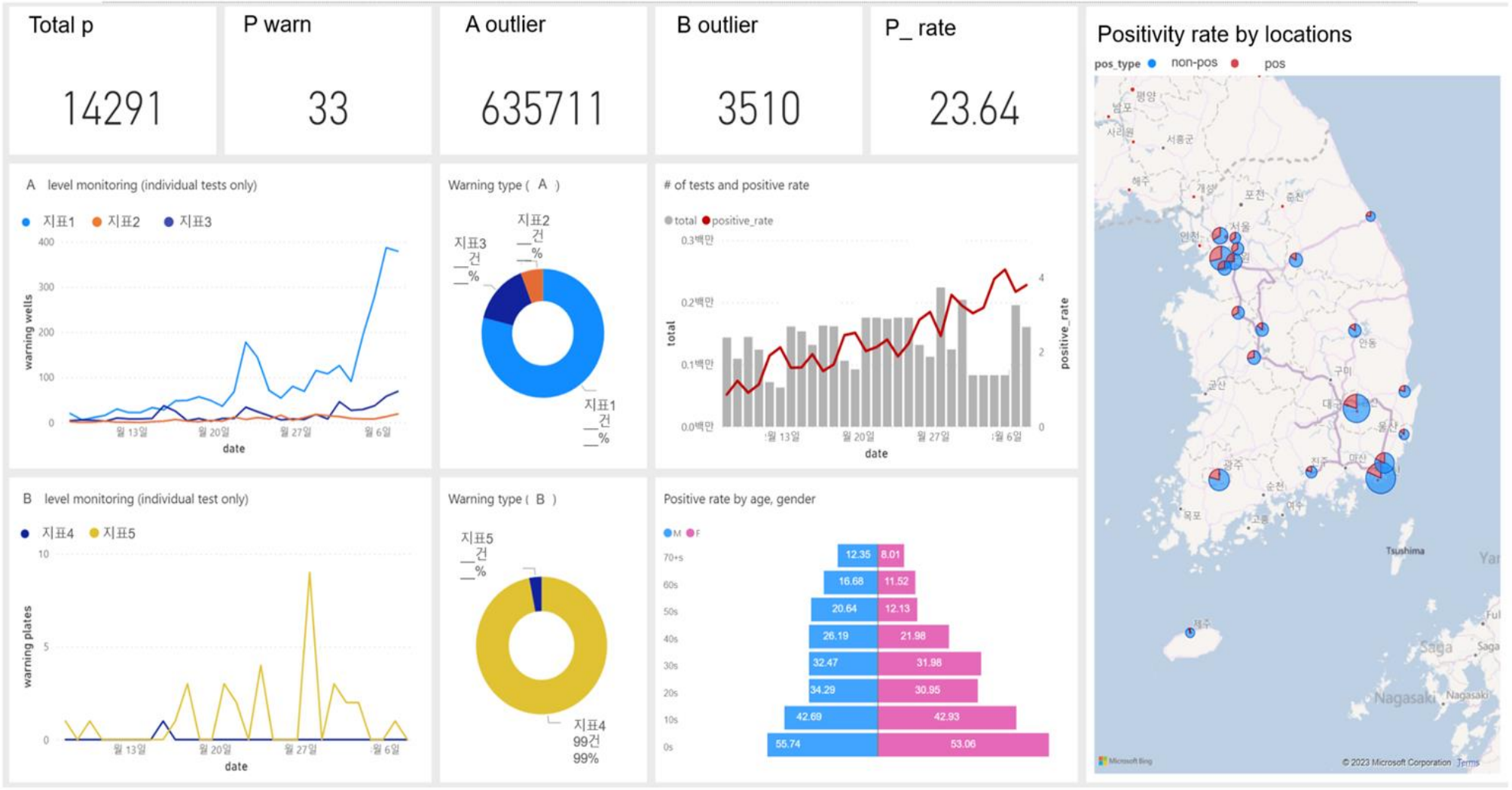
PowerBI 로 구축한 대시보드 메인화면

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



Results: Dashboard development

대시보드 상세화면- 피검사자 정보 Part

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

Results: Dashboard development

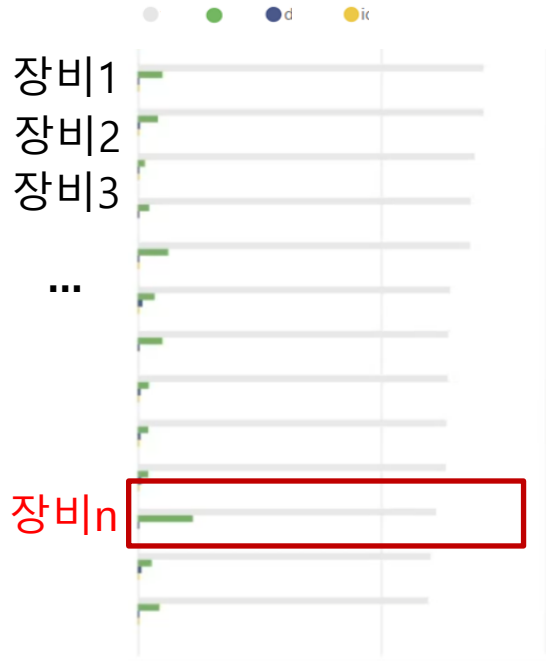
대시보드 상세화면- 품질관리 지표별 분포 및 이상치 기준 시각화

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

Results: Dashboard development

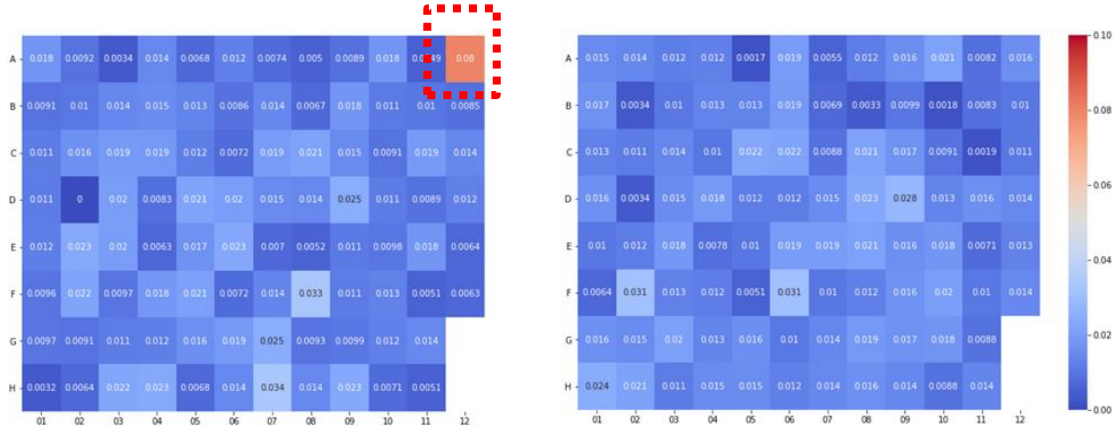
대시보드 상세화면- 품질관리 지표별 분포 및 이상치 기준 시각화

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 점선) 에서 상대적으로 높은 비율로 이상치 발생- 장비 점검 필요성 알려줌

Conclusion & Further plan

- 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

3) DNA 서열 조합 및 길이 별 녹는점 예측- 문헌조사 기반

※ 본 발표자료에 기재된 정보는 실제 자료를 변형/가공한 것으로, 현 직장에서의 중요한 정보가 유출되지 않도록 하였습니다.

Background

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

배경 및 목적

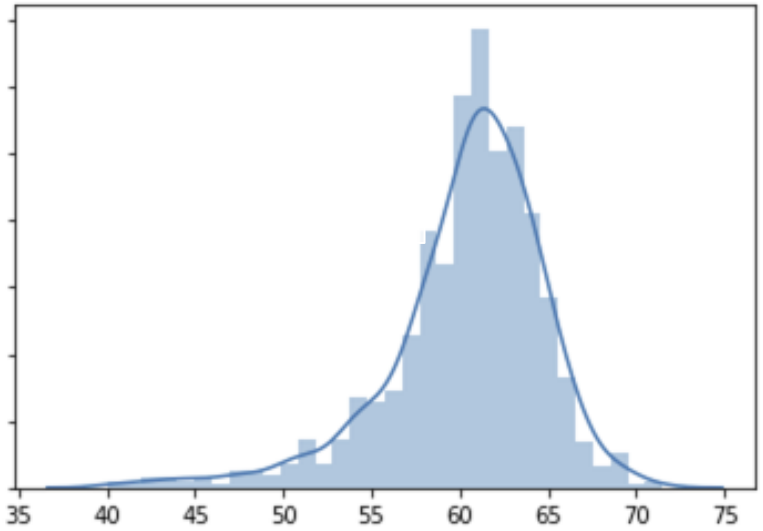
1) DNA의 녹는 온도를 예측 하는 것은 분자진단 제품 개발을 위한 기초단계
2) 기존에 타부서에서 수리적 최적화를 이용하여 예측
3) 더 다양한 방면에서 예측하여 어떤 방법이 제일 최적의 예측 방법일지 비교하기 위하여 머신러닝으로 접근

분석 설계

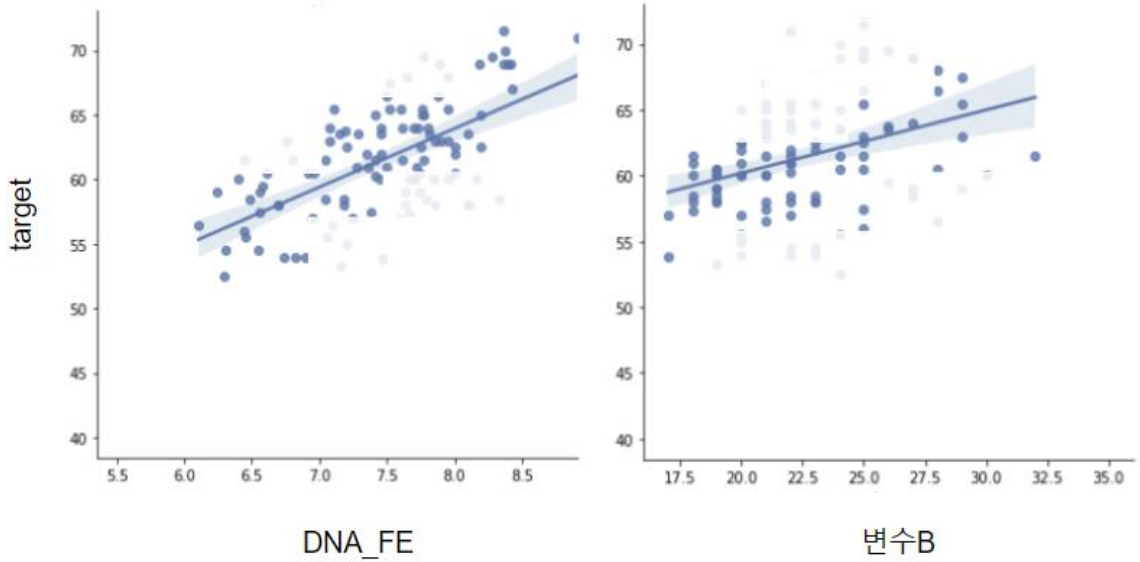
1) EDA를 통한 Feature, target 변수 특성 파악
2) Random forest 등의 모델 생성
3) Feature engineering: **논문 약 20편을 참고**하여 머신러닝으로 DNA melting temperature 예측할 시 추가하면 좋을 feature (**DNA_FE**)를 추가
4) target 예측에 가장 크게 기여하는 변수 파악

사용 언어: Python

Major EDA part



DNA 녹는점 (C°)



DNA_FE

변수B

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.

Conclusions

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.