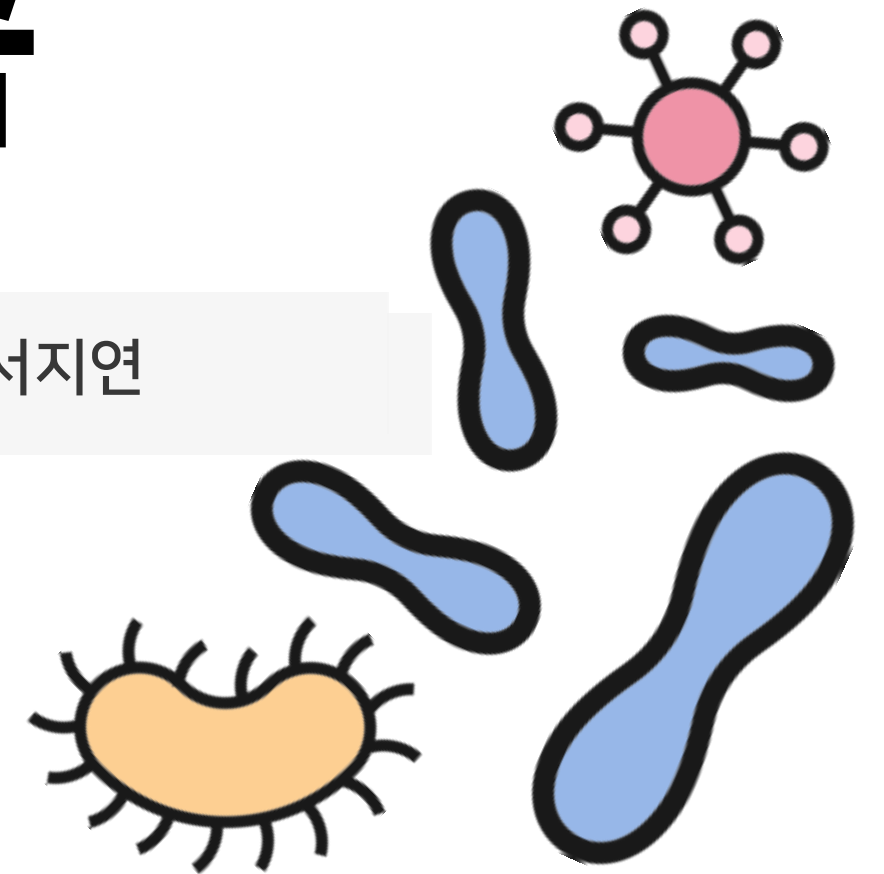


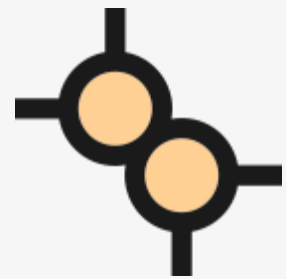
# 마이크로바이옴 기반

## 노화성 질병 예측

지도교수 : 서민석 교수님 | 컴퓨터융합소프트웨어 학과 | 2022271323 | 서지연



# CONTENTS



## 1. Introduction

Background  
Project Goal



## 3. Result

Evaluation  
DashBoard & Analysis



## 2. Method

Datasets  
Data Preprocessing  
Model



## 4. Conclusion

Summary  
Future Work

CHAPTER.

01

Introduction



중앙SUNDAY : 오피니언

## 인체 내 39조 마리 미생물은 유익한 물질 만드는 '제2 장기'

중앙선데이 | 입력 2017.05.07 03:03 업데이트 2017.06.04 03:19

지면보기 ①

[조현욱의 빅 히스토리] 인간과 미생물의 공생

조현욱, "인체 내 39조 마리 미생물은 유익한 물질 만드는 '제2 장기'", 중앙일보, 2017.06.04

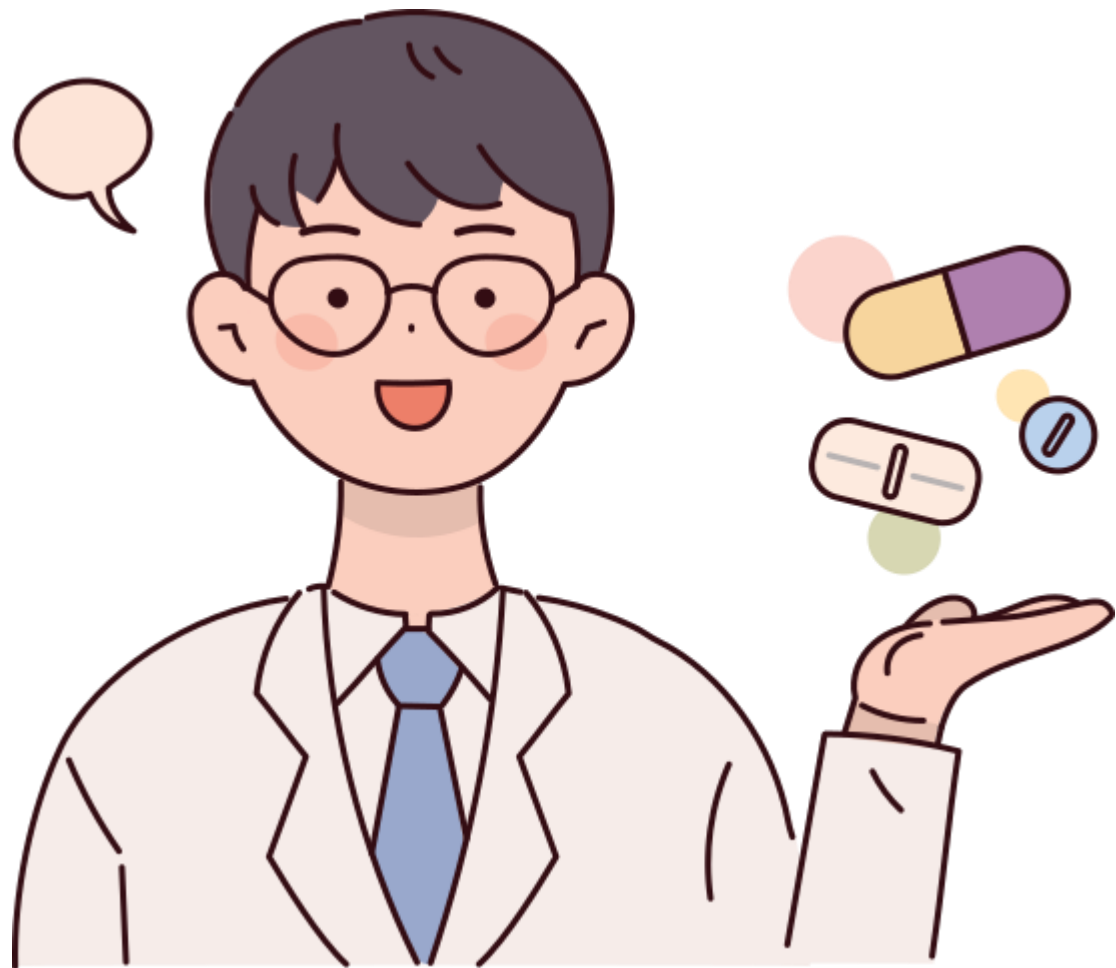
나이가 들면서 체내 독소가 축적돼 몸에 이로운 미생물이 점점 줄어드는 겁니다.

이렇게 미생물의 수가 크게 줄면 면역력과 체력이 현저하게 떨어집니다.

각종 질환의 원인 되는 것입니다.

이성규, "인간과 공생하는 미생물의 세계", YTN 사이언스, 2013.10.10

# Background



과학계는 장내 미생물이 만성 통증 환자에게 새로운 치료법이 될 수 있다고 기대하고 있다. 앞서 연구에서 장내 미생물은 소화 기능은 물론 뇌를 포함해 다양한 장기의 건강에 영향을 주는 것으로 밝혀졌다. 장내 세균은 소화기 질환은 물론, 관절염·비만·위염과 뇌질환까지 막아준다고 알려졌다.

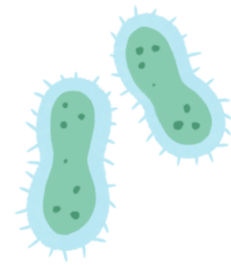
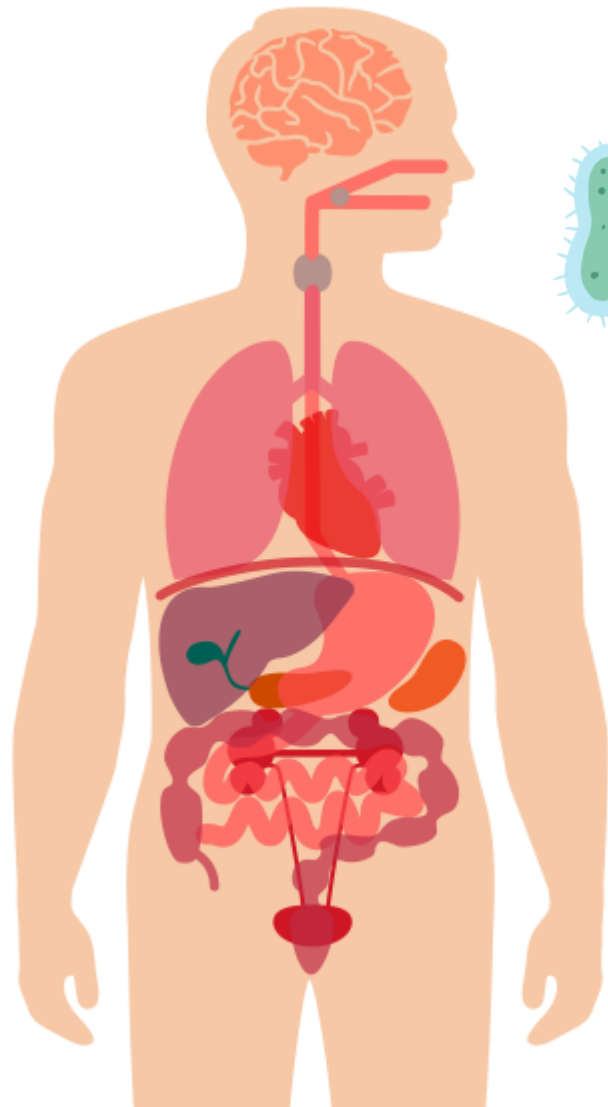
염현아, "장내 세균 바꿨더니..." 만성 통증, 치료 실마리 찾았다", 사이언스 조선, 2025.04.26

광주과학기술원(GIST)은 의생명공학과 류동렬 교수 연구팀이 충남대학교병원 내분비내과 이현승 교수팀, 고려대학교 생명공학부 최동욱 교수팀, 에이치이엠파마, 아모레퍼시픽 연구진과의 공동연구를 통해 장내 미생물이 노화를 늦추고 건강에 미치는 긍정적 효과를 규명했다고 9일 밝혔다.

이종훈, "장내 미생물과 노화, 직접적인 연결고리 확인했다", 헬스라이프헤럴드, 2025.04.07

01

# Project Goal



방대한 마이크로바이옴 데이터를 효율적으로 구조화  
다양한 머신러닝 모델 평가 및 시각화  
-> 실시간 예측 결과 확인 가능한 반응형 웹을 만들자!

CHAPTER.

02

Method

## 1. Metadata

- 다양한 질병 정보, 성별, 나이 등의 인간 미생물 정보가 담긴 데이터 (45,242 샘플)
- 총 53가지 질병 그룹으로 분류

출처) <https://www.ebi.ac.uk/metagenomics/search/studies?query=microbiome>

+ 생물학적 분류 체계 (taxonomy)

: 생물학에서 생물들을 체계적으로 분류

Phylum(문) -> Class(강) -> Order(목) -> Family(과) -> Genus(속)

## 2. 미생물 데이터

- Phylum, Class, Order, Family, Genus 수준의 각 미생물 데이터 (by 생물학적 분류체계)
- 각 샘플은 미생물 군집의 abundance 값을 포함

ex. 개의 분류: 척삭동물, 포유류, 육식목, 개과, 개속

**metadata (45242, 95)**

	OTUid	Analysis_accession	Run	Group
1	ERR589383	MGYA00003425	ERR589383	Healthy
2	ERR589384	MGYA00003426	ERR589384	Healthy
3	ERR589686	MGYA00003630	ERR589686	Rheumatoid arthritis
4	ERR589702	MGYA00003646	ERR589702	Rheumatoid arthritis
5	ERR589704	MGYA00003648	ERR589704	Healthy
6	ERR589712	MGYA00003656	ERR589712	Healthy

	OTUid	disease	o_0319-7L14	o_11-24	o_258ds10	o_32-20
1	ERR1912958	Parkinson disease	4.289887	4.289887	4.289887	4.289887
2	ERR1912959	Parkinson disease	4.289887	4.289887	4.289887	4.289887
3	ERR1912960	Parkinson disease	4.289887	4.289887	4.289887	4.289887
4	ERR1912964	Parkinson disease	4.289887	4.289887	4.289887	4.289887
5	ERR1912965	Parkinson disease	4.289887	4.289887	4.289887	4.289887

**Genus(17111,2243)**

**Family(17111,670)**

**Order (17111, 519)**

**Class (17111, 291)**

**Phylum(17111,168)**

02

# Data Preprocessing

01

데이터프레임 병합

02

NA/결측치 처리 및 데이터 클렌징

03

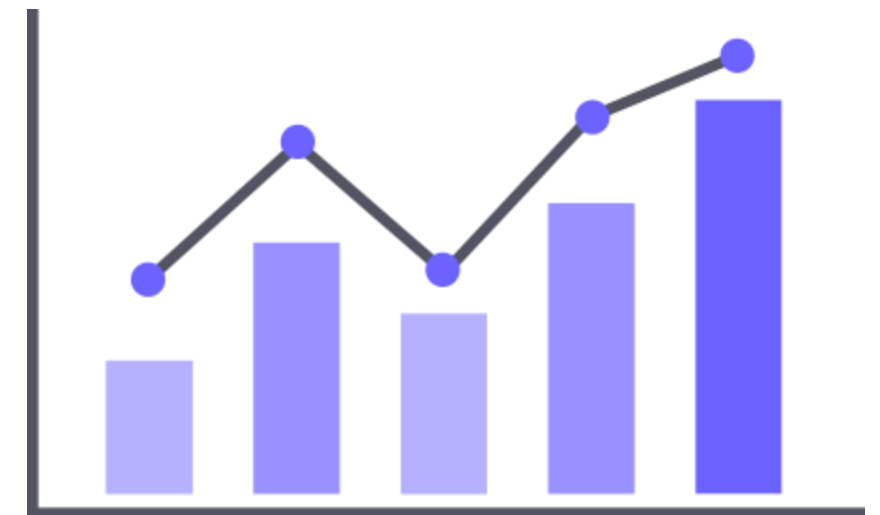
taxonomy 수준 분리

04

Quality Control

05

데이터 정규화

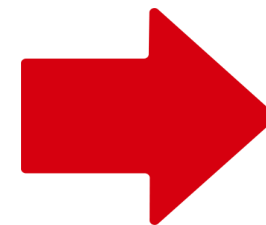


02

# Data Preprocessing

기존 )

	S1	S2	S3
미생물1			
미생물2			
미생물3			



변경 )

	미생물1	미생물2	미생물3
S1			
S2			
S3			

k\_Bacteria;p\_Actinobacteria;c\_Actinobacteria;o\_Actinomycetales;f\_Actinomycetaceae; ... | 풍부도

데이터 테이블 전치 및 Taxonomy 별 파일 분리

## 02

# Data Preprocessing

이름

- Phylum\_NxP\_OTUtable\_total\_20230929\_1.tsv
- Order\_NxP\_OTUtable\_total\_20230929\_1.tsv
- Genus\_NxP\_OTUtable\_total\_20230929\_1.tsv
- Family\_NxP\_OTUtable\_total\_20230929\_1.tsv
- Class\_NxP\_OTUtable\_total\_20230929\_1.tsv



	OTUId	p_AC1	p_Acidobacteria	p_Actinobacteria	p_AD3	p_Annelida	p
1	ERR589383	0	10	1830	0	0	
2	ERR589384	0	17	4174	0	0	
3	ERR589686	0	36	12378	0	0	
4	ERR589702	0	39	10657	0	0	
5	ERR589704	0	57	6025	0	0	
6	ERR589712	0	114	8516	0	0	
7	ERR589714	0	32	1880	0	0	

Taxonomy 별 tsv 파일로 저장

문제 1. 0 너무 많이 존재  
문제 2. 피처 간 발현량 범위 차이가 큼

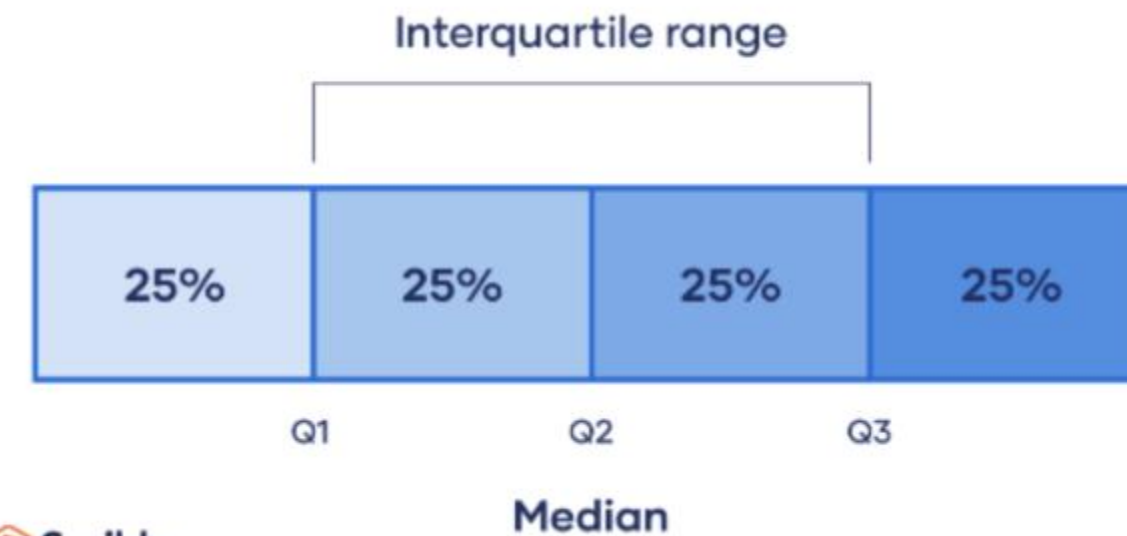
=> QC 진행

## ① 0값 제거

- 모든 샘플에서 검출되지 않은 taxon 삭제
- 모든 taxon 카운트 합이 0인 샘플 삭제

## ② IQR기반 필터링

- IQR이 낮다 -> 관측 오류/ 변별력 x
- IQR이 크다 -> 발현량 차이, 유용한 정보



# Data Preprocessing (Normalization)

## POINT. 01

### 상대풍부도 구하기

각 샘플의 taxon 카운트를  
그 샘플 전체 카운트로 나눈 백분율

$$\text{RelativeAbundance}_{i,j} = \frac{c_{i,j}}{\sum_k c_{i,k}}$$

샘플 간 총 카운트 **차이를 보정**하고,  
각 taxon이 전체에서 차지하는 **비율**을 구함

## POINT. 02

### CpmLog 변환

$$\text{CPM}_{i,j} = \frac{c_{i,j}}{\text{LibSize}_i} \times 10^6$$

$$\log\text{CPM}_{i,j} = \log_2(\text{CPM}_{i,j} + 1)$$

스케일을 맞추기 위해

라이브러리 크기(총 리드 수) 차이를 보정  
로그 변환을 통해 값의 **분포를 안정화**

## POINT. 03

### TMM 정규화

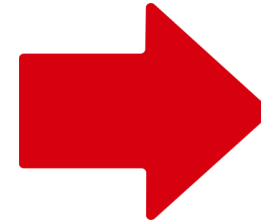
샘플 간 로그 비율을 보고  
극단치를 걸러낸 뒤 평균을 내어  
정규화 계수를 산출하는 방법

**샘플 간** compositional bias **보정**

# Data Preprocessing

전)

	OTUid	p_AC1	p_Acidobacteria	p_Actinobacteria	p_AD3	p_Annelida	p
1	ERR589383	0	10	1830	0	0	
2	ERR589384	0	17	4174	0	0	
3	ERR589686	0	36	12378	0	0	
4	ERR589702	0	39	10657	0	0	
5	ERR589704	0	57	6025	0	0	
6	ERR589712	0	114	8516	0	0	
7	ERR589714	0	32	1880	0	0	



후)

	OTUid	disease	p_AC1	p_Acidobacteria	p_Actinobacteria
1	ERR1912958	Parkinson disease	4.635856	4.635856	14.38579
2	ERR1912959	Parkinson disease	4.635856	4.635856	14.43668
3	ERR1912960	Parkinson disease	4.635856	4.635856	14.39164
4	ERR1912964	Parkinson disease	4.635856	7.013757	14.82567
5	ERR1912965	Parkinson disease	4.635856	4.635856	15.04731
6	ERR1912966	Parkinson disease	4.635856	6.182961	14.98281
7	ERR1912976	Parkinson disease	4.635856	4.635856	12.40160

"총 14가지 머신러닝 기반 분류기를 사용하여 노화성 질병 예측 성능을 비교 수행"

### 트리 기반

extraTrees  
ranger  
randomForest  
ordinalRF

### 부스팅

xgbTree  
gbm\_h2o  
LogitBoost

### 선형 모델

regLogistic  
slda

### 서포트벡터머신

svmLinear  
svmPoly

### 기타 앙상블

RRF  
wsrF  
LMT

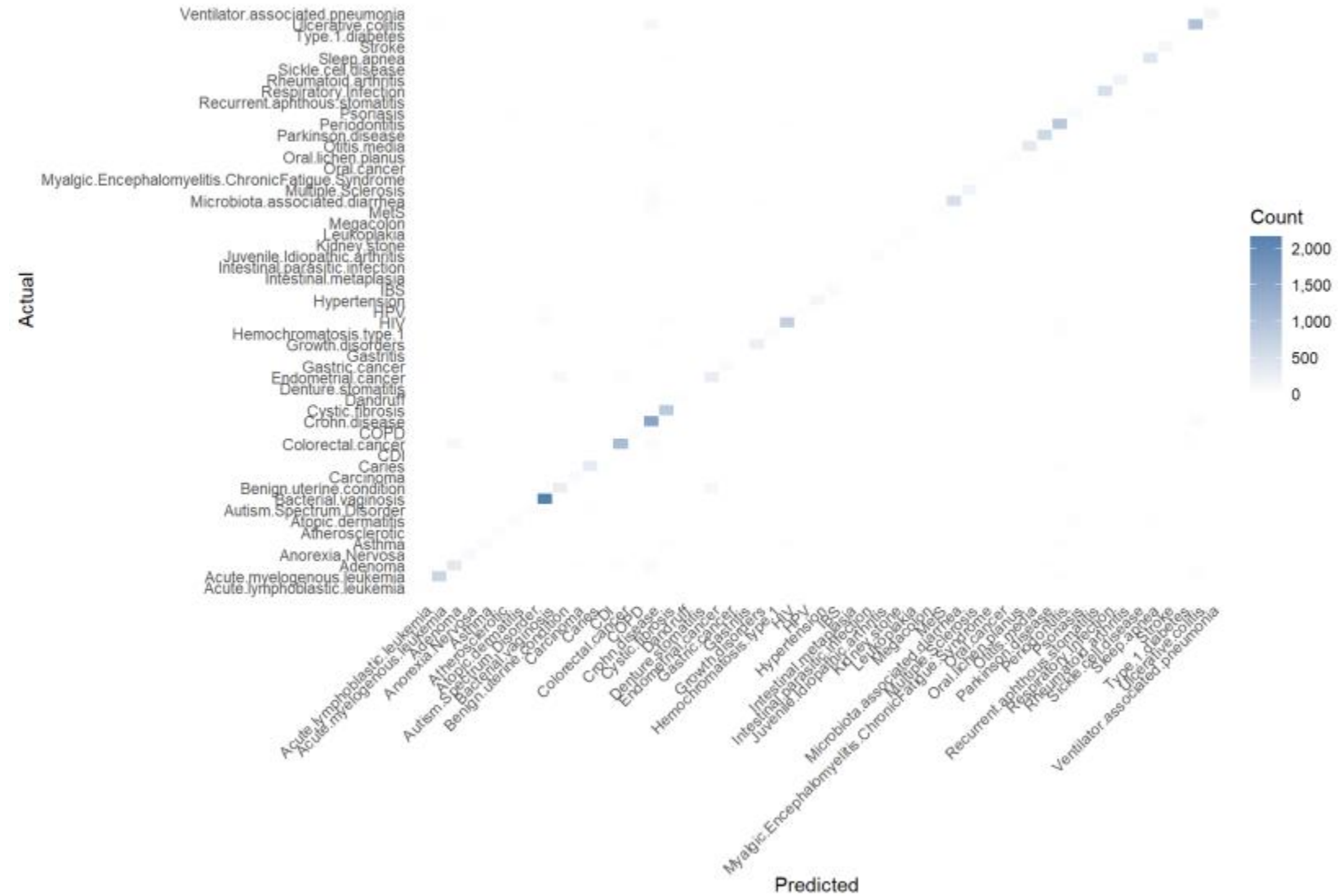
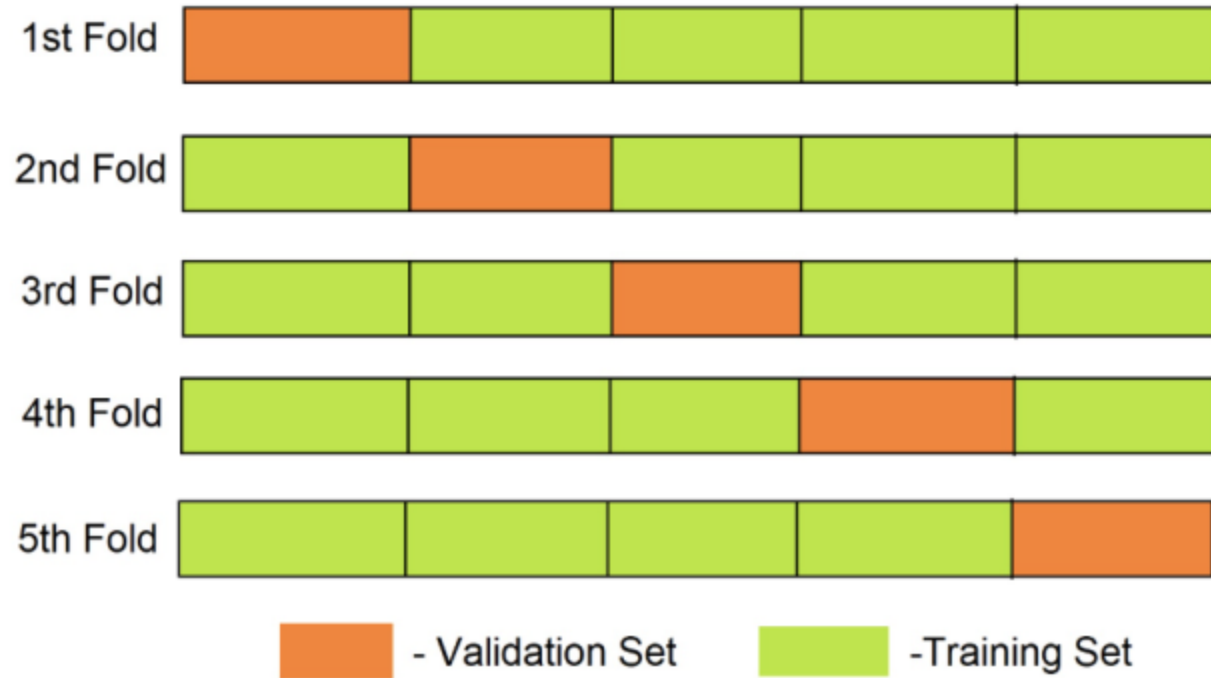
CHAPTER.

03

Result

# Evaluation

- K-fold Cross-Validation  
5-fold CV 를 사용.



## 3가지 파일 제작

## 1. Confusion.matrix.14.models.tsv

모델×Taxon별로 실제 클래스 대비 **예측 횟수**를 기록

	Model	Taxon	pred	ref	value
1	extraTrees	Class	Acute.lymphoblastic.leukemia	Acute.lymphoblastic.leukemia	10
2	extraTrees	Class	Acute.myelogenous.leukemia	Acute.lymphoblastic.leukemia	2
3	extraTrees	Class	Adenoma	Acute.lymphoblastic.leukemia	0

## 2. bestModel.byClass.tsv

모델×Taxon별 각 클래스에 대한 **민감도** 정리

	Classifiers	Sensitivity	Class	Taxon
1	extraTrees	0.24242424	Acute.lymphoblastic.leukemia	Phylum
2	extraTrees	0.77135348	Acute.myelogenous.leukemia	Phylum
3	extraTrees	0.52804642	Adenoma	Phylum

## 3. bestModel.resamples.tsv

모델×Taxon별 10-fold 재샘플링 과정의 **Accuracy** 값 기록

	Classifiers	Accuracy	Taxon
1	extraTrees	0.7407085	Phylum
2	extraTrees	0.7494138	Phylum
3	extraTrees	0.7454599	Phylum

**민감도와 정확도에 집중 => 정확도 "전반적 맞춘 비율" + 민감도 "놓치는 환자 비율"**

# DashBoard & Analysis

"웹 기반 인터랙티브 마이크로바이옴 분석 및 예측 플랫폼 개발"



① 전체 모델 성능 비교 탭  
: 모든 모델·Taxon 조합의 Accuracy 분포(박스플롯)  
및 평균·표준편차 테이블 제공

=> 어떤 모델이 정확도 분포가 가장 높고 안정적인지 확인

**extraTrees**가 정확도가 가장 높았음

## "웹 기반 인터랙티브 마이크로바이옴 분석 및 예측 플랫폼 개발"



## ② 질병별 민감도 탭

: 선택한 모델·Taxon에서 상위 N개(기본 20개) 질병의 Sensitivity 막대그래프

모델이 특히 잘 찾아내는 질병은 무엇인지,  
Taxon 레벨에 따른 질병별 탐지 강점 차이 확인 가능

## "웹 기반 인터랙티브 마이크로바이옴 분석 및 예측 플랫폼 개발"



## ③ 혼동 행렬

: 특정 모델·질병(ref)에 대한 예측(pred) 분포 상위 10개를 집계한 막대그래프

해당 질병을 예측할 때 주로 어떤 다른 질병과 혼동하는지를 통해 모델의 혼동 패턴, 질병 간 유사성 단서, 데이터·모델 개선 포인트 등을 분석 가능

# 03

# DashBoard & Analysis

## "웹 기반 인터랙티브 마이크로바이옴 분석 및 예측 플랫폼 개발"

미생물 데이터 기반 Dashboard

① 전체 모델 성능 비교   ② 질병별 민감도   ③ 분류 순위 선택 (단일 클래스)   ④ 미생물 분포   ⑤ 전체 클래스 정확도 요약   ⑥ 샘플 메타데이터   ⑦ 질병 예측

Taxon별 미생물 상대 abundance  
 Max 노미 레벨 선택  
 Phylum

Copy   CSV   Excel   PDF   Print   Search:

OTUId	disease	p_AC1	p_Acidobacteria	p_Actinobacteria	p_AD3	p_Annelida	p_Apicomplexa	p_Aquificae	p_Armatimonadetes
1	ERR1912958	Parkinson disease	4.63585625206165	4.63585625206165	14.3857864999451	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
2	ERR1912959	Parkinson disease	4.63585625206165	4.63585625206165	14.4366811332641	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
3	ERR1912960	Parkinson disease	4.63585625206165	4.63585625206165	14.3916376130152	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
4	ERR1912964	Parkinson disease	4.63585625206165	7.01375717610193	14.8256662364395	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
5	ERR1912965	Parkinson disease	4.63585625206165	4.63585625206165	15.0473083017399	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
6	ERR1912966	Parkinson disease	4.63585625206165	6.18296147403405	14.9828113429628	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
7	ERR1912976	Parkinson disease	4.63585625206165	4.63585625206165	12.4016047956789	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
8	ERR1912977	Parkinson disease	4.63585625206165	4.63585625206165	12.7835371228609	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
9	ERR1912978	Parkinson disease	4.63585625206165	4.63585625206165	12.3867691610068	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165
10	ERR1912983	Parkinson disease	4.63585625206165	4.63585625206165	13.0951309738106	4.63585625206165	4.63585625206165	4.63585625206165	4.63585625206165

Showing 1 to 10 of 17,111 entries   Previous   1   2   3   4   5   ...   1,712   Next

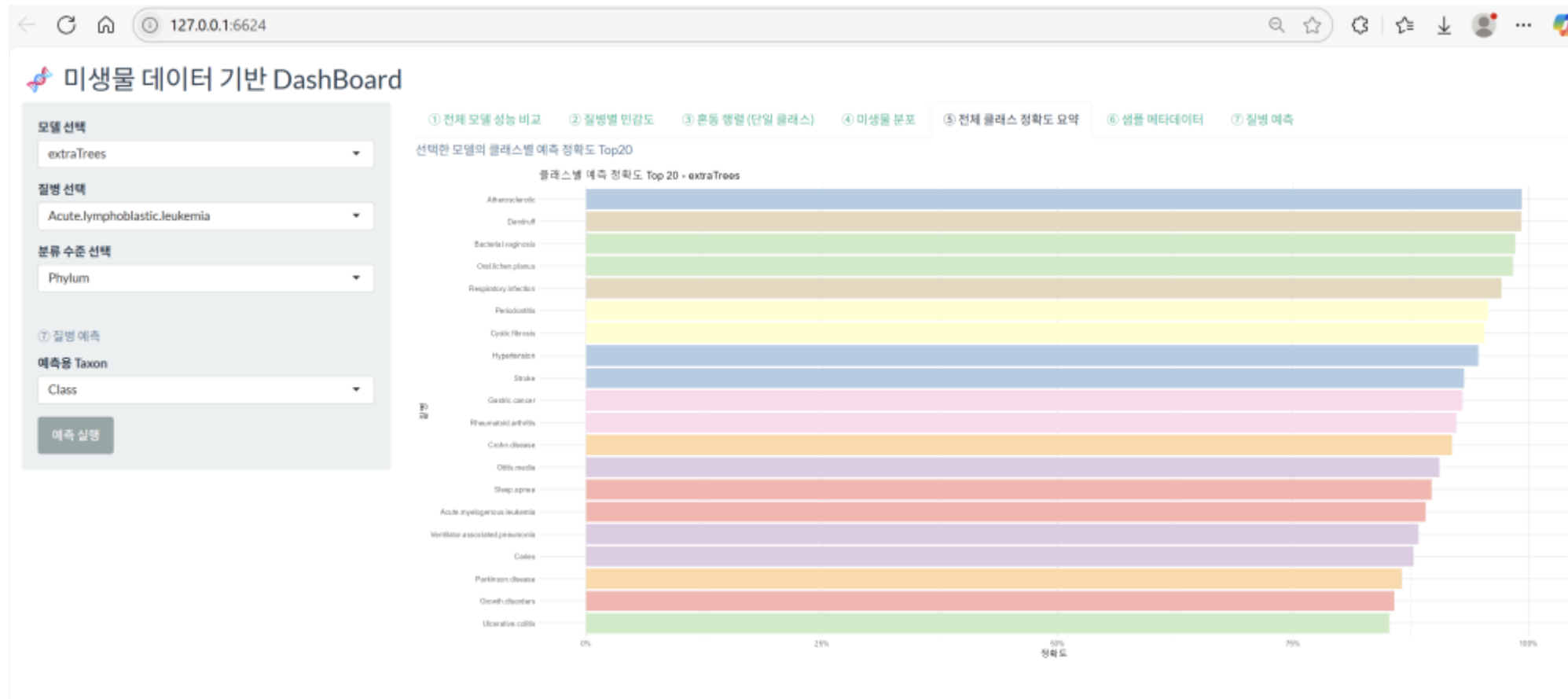
### ④ 미생물 분포

: 선택한 Taxonomy 레벨의 원시/정규화 OTU 테이블을 인터랙티브하게 탐색

민감도·정확도 높은 질환에서 특징적으로 높거나 낮은 분류군 찾기

Ex) Phylum Bacteroidetes가 Alzheimer's·Parkinson's에서 높다면, 이 분류군이 바이오마커 후보

## "웹 기반 인터랙티브 마이크로바이옴 분석 및 예측 플랫폼 개발"



## ⑤ 전체 클래스 정확도 요약

: 선택 모델의 클래스별 정확도를 Top 20까지  
바 차트로 시각화

모든 질병에 대한 "정확도 순위"를 확인

모델이 약한 질병을 빠르게 식별 및 개선 포인트 도출

# 03

# DashBoard & Analysis

"웹 기반 인터랙티브 마이크로바이옴 분석 및 예측 플랫폼 개발"

미생물 데이터 기반 DashBoard

모형 선택: extraTrees  
질병 선택: Acute lymphoblastic leukemia  
분류 수준 선택: Phylum  
예측용 Taxon: Class

① 전체 모델 성능 비교   ② 질병별 민감도   ③ 혼동 행렬 (단일 클래스)   ④ 미생물 분포   ⑤ 전체 클래스 정확도 요약   ⑥ 샘플 메타데이터   ⑦ 질병 예측

인상 및 실행 정보

Copy CSV Excel PDF Print

OTUId	Analysis_accession	Run	Group	Assay Type	BioSample	Center Name	Consent	Experiment	Instrument	LibraryLayout	LibrarySelection	LibrarySource	
1	ERR589383	MGYA00003425	ERR589383	Healthy	WGS	SAMEA2738026	BGI	public	ERX547378	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
2	ERR589384	MGYA00003426	ERR589384	Healthy	WGS	SAMEA2738029	BGI	public	ERX547379	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
3	ERR589686	MGYA00003630	ERR589686	Rheumatoid arthritis	WGS	SAMEA2738133	BGI	public	ERX547681	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
4	ERR589702	MGYA00003646	ERR589702	Rheumatoid arthritis	WGS	SAMEA2738149	BGI	public	ERX547697	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
5	ERR589704	MGYA00003648	ERR589704	Healthy	WGS	SAMEA2738151	BGI	public	ERX547699	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
6	ERR589712	MGYA00003656	ERR589712	Healthy	WGS	SAMEA2738159	BGI	public	ERX547707	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
7	ERR589714	MGYA00003658	ERR589714	Rheumatoid arthritis	WGS	SAMEA2738161	BGI	public	ERX547709	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
8	ERR589715	MGYA00003659	ERR589715	Healthy	WGS	SAMEA2738162	BGI	public	ERX547710	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
9	ERR589721	MGYA00003665	ERR589721	Rheumatoid arthritis	WGS	SAMEA2738168	BGI	public	ERX547716	Illumina HiSeq 2000	PAIRED	other	METAGENOMI
10	ERR589722	MGYA00003666	ERR589722	Rheumatoid arthritis	WGS	SAMEA2738169	BGI	public	ERX547717	Illumina HiSeq 2000	PAIRED	other	METAGENOMI

Showing 1 to 10 of 45,242 entries

Previous 2 3 4 5 ... 4,525 Next

⑥ 샘플 메타데이터 : 원본 메타데이터를 DataTable로 제공

## "웹 기반 인터랙티브 마이크로바이옴 분석 및 예측 플랫폼 개발"

미생물 데이터 기반 DashBoard

① 전체 모델 성능 비교 ② 질병별 민감도 ③ 혼동 행렬 (단일 클래스) ④ 미생물 분포 ⑤ 전체 클래스 정확도 요약 ⑥ 샘플 메타데이터 ⑦ 질병 예측

선택 Taxon 모델로 테스트셋 랜덤 3개 샘플 예측 결과

	실제 질병	예측 질병	맞았는지
159	Ventilator-associated pneumonia	Cystic fibrosis	false
65	Cystic fibrosis	Cystic fibrosis	true
236	Periodontitis	Periodontitis	true

⑦ 질병 예측

예측용 Taxon: Genus

예측 실행

## ⑦ 질병 예측

: 사용자가 보유한 신규 데이터를 입력하면 학습된 Taxon 모델이 질병을 예측해 주는 기능

임시로, 선택된 Taxon 모델에 대해 테스트셋에서 랜덤 3개 샘플을 추출해서 결과 확인하는 형식으로 구현

	실제 질병	예측 질병	맞았는지
43	Crohn disease	Crohn disease	true
24	Ulcerative colitis	Ulcerative colitis	true
253	Bacterial vaginosis	Bacterial vaginosis	true

CHAPTER.

04

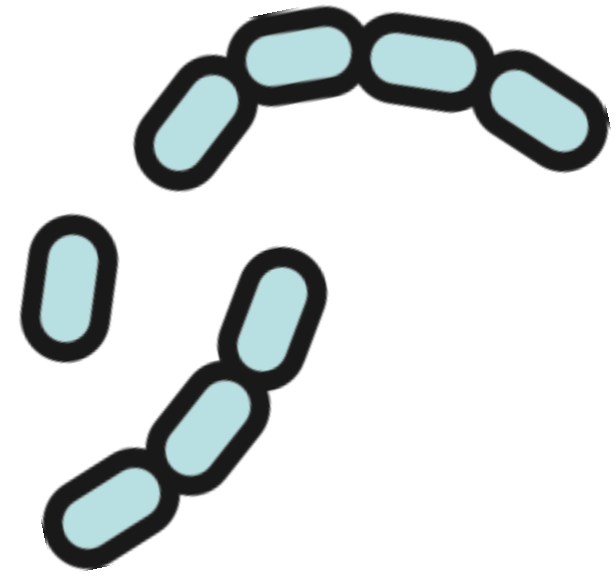
Conclusion

## SUMMARY

1. 마이크로바이옴 데이터를 활용해 **노화성 질환 예측**
2. 다양한 머신러닝 모델을 비교·평가해 **최적의 예측 모델 도출**
3. **직관적 Shiny 대시보드**로 의료진·연구자가 실시간으로 결과를 시각화·분석 가능

## FUTURE WORK

1. 머신러닝을 넘어 **딥러닝 아키텍처를 도입**해 미생물 데이터의 비선형성·상관관계를 심층 학습
2. 분류학 계층을 그래프 구조로 모델링하여, Graph Neural Network 기반의 **multi-view learning**으로 확장
3. 미생물 관련 연구 확장



Thank you

