정신질환 관련 환자군 분류기법 개발

권병근[1], 신호섭[2], 이다인[2], 성다솜[2,*]

[1]부산대학교 수학과, 산업수학 소프트웨어 연계전공, 산업수학센터 학부연구생
[2]부산대학교 수학과, 빅데이터 연계전공, 산업수학센터 학부연구생
이메일 : *som0608@pusan.ac.kr

1. Abstract

이 프로젝트는 산업 수학센터 2023년 여름방학 학부 연구생 프로젝트로 진행하였다.

정신질환 환자군을 분류하는 새로운 모델링 접근법의 개발과 적용을 다루었다. 초기 단계에서는 알츠하이머의 특성 및 병인을 탐구하였고, 양전자 방출 단층 촬영과 신경 심리 검사 데이터에 대한 이해와 분석을 수행하였다.

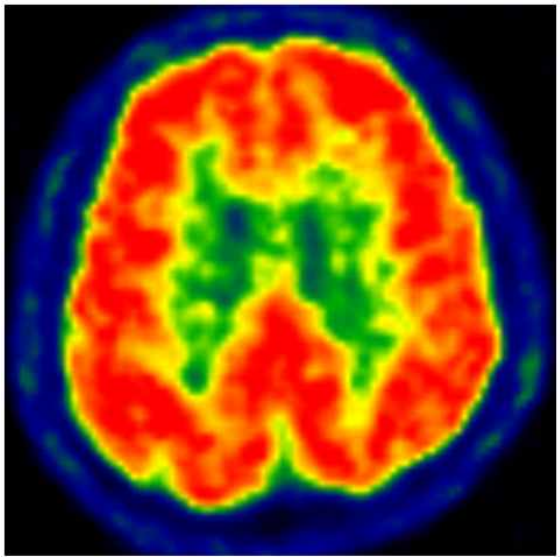
다양한 데이터 유형에 대한 학습 후, 각 데이터 유형별로 분류 모델을 개발하고 비교 분석하였다. 그러나 이 과정에서 발견된 일관성 부족으로 인해 비교 결과의 신뢰성이 떨어짐을 확인하였다.

이에 따라, 기존의 분류 모델 접근 방식에서 발생한 일관성 부족 문제를 해결하기 위해 새로운 모델링을 진행하였다. 이를 위해 여러 데이터 유형을 종합하고, target을 다양하게 정의하여 각각의 목표 변수에 대한 특화된 모델을 개발하였다. 이렇게 하여 여러 모델을 통해 도출된 결과를 비교하고 분석함으로써 보다 일관성이 있고 신뢰성 있는 분류 결과를 도출하였다.

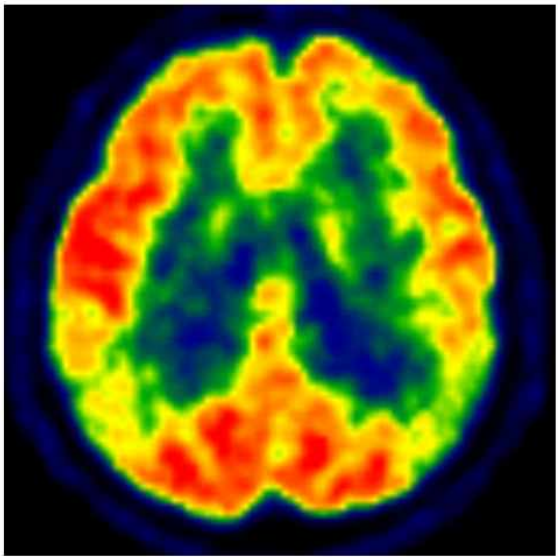
2. Introduction

알츠하이머란 65세 이상 노인에게 주로 생기는 퇴행성 뇌 질환이다. 기억력 장애를 시작으로 공간 지각력, 판단력이 떨어지며 악화 되며 일상생활 수행 능력이 상실된다. 병의 진단은 병력과 신경 심리 검사 그리고 뇌 영상정보를 참고하여 의사가 내리는 임상적 판단에 의존한다. 베타 아밀로이드(beta-amyloid)라는 단백질이 뇌에 축적되면서 병을 유발한다고 알려져 있다.

양전자 방출 단층 촬영(PET)은 핵의학 검사 방법 중 하나로, 양전자를 방출하는 방사성 의약품을 인체에 투여하여 대사 작용을 촬영하는 영상 검사 방법이다. Amyloid PET 영상은 베타-아밀로이드 침착 정도를 확인하는 PET 영상이다. SVUr은 표준 섭취 계수율로 특정 영역에 대한 아밀로이드 측정값을 정규화한 수치이다.



정상인



치매인

서울대학교 병원, https://www.snubh.org/dh/main/index.do?DP_CD=BCD7&MENU_ID=006005

신경 심리 검사(Neuropsychological test)는 뇌 손상이나 신경병리적 조건에 따른 인지기능 및 행동적 변화를 측정하여 뇌와 행동 간의 관계를 측정하여 뇌 손상 혹은 뇌 기능 장애를 진단하는 검사이다.

SGDepS (단축형 노인 우울 척도)	노인용 우울 검사로 짧은 시간에 실시 가능, 주관적인 평가나 스스로 자각한 삶의 질을 측정하는 검사
MMSE (간이 정신 상태 검사)	인지기능 장애의 정도를 반복적인 측정을 통해 정량적으로 평가하는 검사
GDS (전반적 퇴화 척도)	진단 의심 혹은 진단된 알츠하이머 환자의 진행 상태를 확인하기에 적합한 검사
CDR (임상 치매 척도)	전반적인 인지 및 사회 기능 정도를 측정하는 대표적인 등급 척도
CDR-SB (임상 치매 척도 박스 총점)	CDR 검사에 대한 점수 산출 방법을 다르게 사용한 결과

3. Main

(1) 사용한 데이터

1) FreeSurfer 기반 SUVr 데이터

- Clean version : 이상치, 결측치가 포함된 변수 제거 및 뇌 영역별로 묶음(신경 심리 결과 포함), 168 rows X 30 columns
- Raw version : SUVr 데이터 기준으로 정리(신경 심리 결과 포함), 168 rows X 64 columns

2) PMOD 기반 SUVr 데이터

- Clean version : 이상치, 결측치가 포함된 변수 제거 및 뇌 영역별 묶음(신경 심리 결과 포함), 168 rows X 30 columns
- Raw version : SUVr 데이터 기준으로 정리(신경 심리 결과 포함), 168 rows X254 columns

(2) 데이터 설명

- Tracer(=PET ligand type) : Amyloid-Pet 영상을 위해 사용된 방사성 의약품 ⇒ 3종류(FBB, FMM, FPN)
- Positivity : 의사 진단으로 결정된 질병 유무 ⇒ 환자군(Positive, BAPL2/3), 대조군(Negative, BAPL1)
- APOE 유전자 정보 : 6개의 유전형 다형성이 존재 ⇒ E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, E4/E4 (E4 단백의 경우, 발병 나이(Onset age)와 연관이 있음)
- 신경 심리 검사 결과 정보 : 우울증, 인지기능, 행동 변화 등을 바탕으로 6개의 검사 결과 정보가 존재
- Amyloid 영상정보 : Clean version ⇒ 해당 뇌 영역별 SUVr 값 표기,
  Raw version ⇒ Clean version으로의 가공 이전 데이터로 뇌 영역이 세분되어있고 부분별 volume 값이 포함

(3) EDA(Exploratory Data Analysis, 탐색적 데이터 분석)

바이올린 플롯(Violin plot)은 데이터 분포를 시각화하기 위한 통계 그래프 중 하나이며, 박스 플롯(Box plot)과 커널 밀도 추정(Kernel Density Estimation) 그래프의 조합으로 이루어져 있어서 데이터의 분포를 더 다양한 관점에 파악할 수 있다. 바이올린 플롯의 주요 의미와 그리는 이유는 다음과 같다.

① 데이터 분포의 시각화 : 데이터의 분포를 시각적으로 표현해준다. 평균, 중앙값, 분산 등 통계적 특성뿐만 아니라 데이터의 밀도를 보여줌으로써 데이터 세트의 특성을 빠르게 파악할 수 있다.

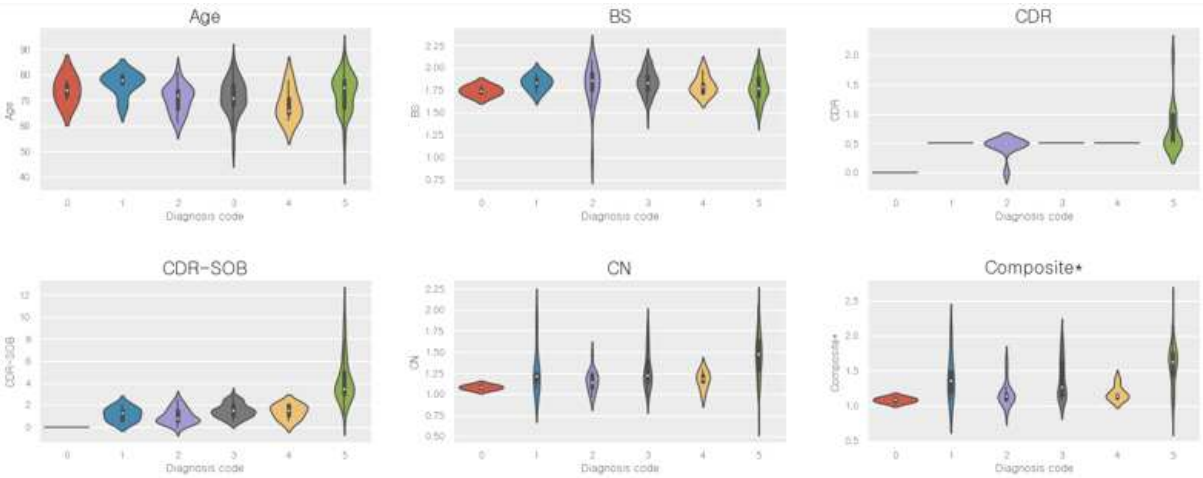
② 카테고리별 비교 : 여러 개의 카테고리의 데이터 분포를 한꺼번에 비교할 수 있는 장점이 있다. 각 카테고리의 분포를 나란히 배치하여 비교하면, 그룹 간의 차이나 유사성을 시각적으로 파악할 수 있다.

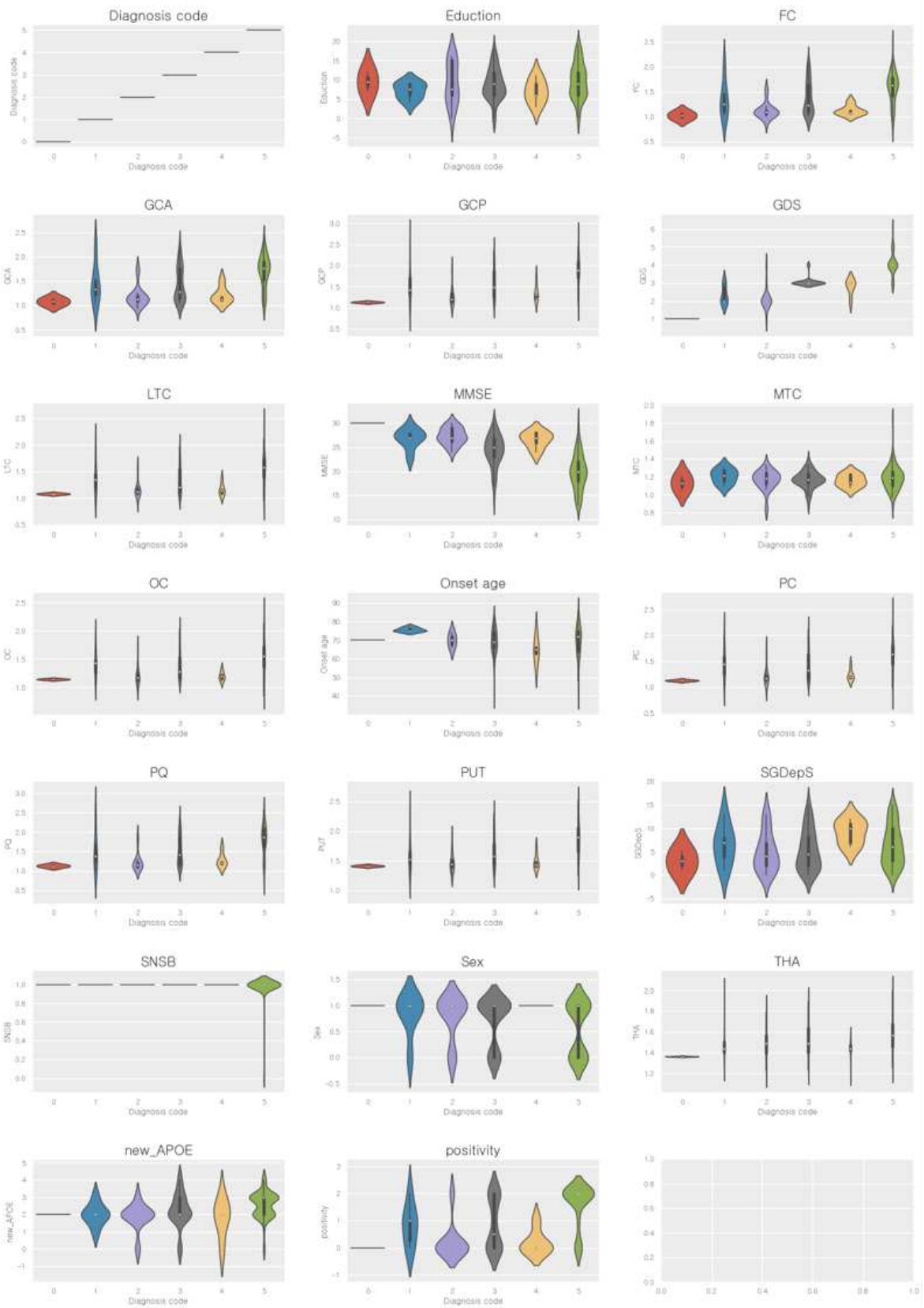
③ 커널 밀도 추정 : 커널 밀도 추정 그래프를 포함하고 있어, 실제 데이터 분포를 부드럽게 표현할 수 있다. 이는 박스 플롯만으로는 알기 어려운 분포의 세부 정보를 확인할 수 있다.

④ 이상치 감지 : 데이터의 분포를 보여주기 때문에 이상치(outlier)를 쉽게 발견할 수 있다. 박스 플롯의 "whisker" 부분에서 벗어나는 데이터 포인트는 이상치로 간주할 수 있다.

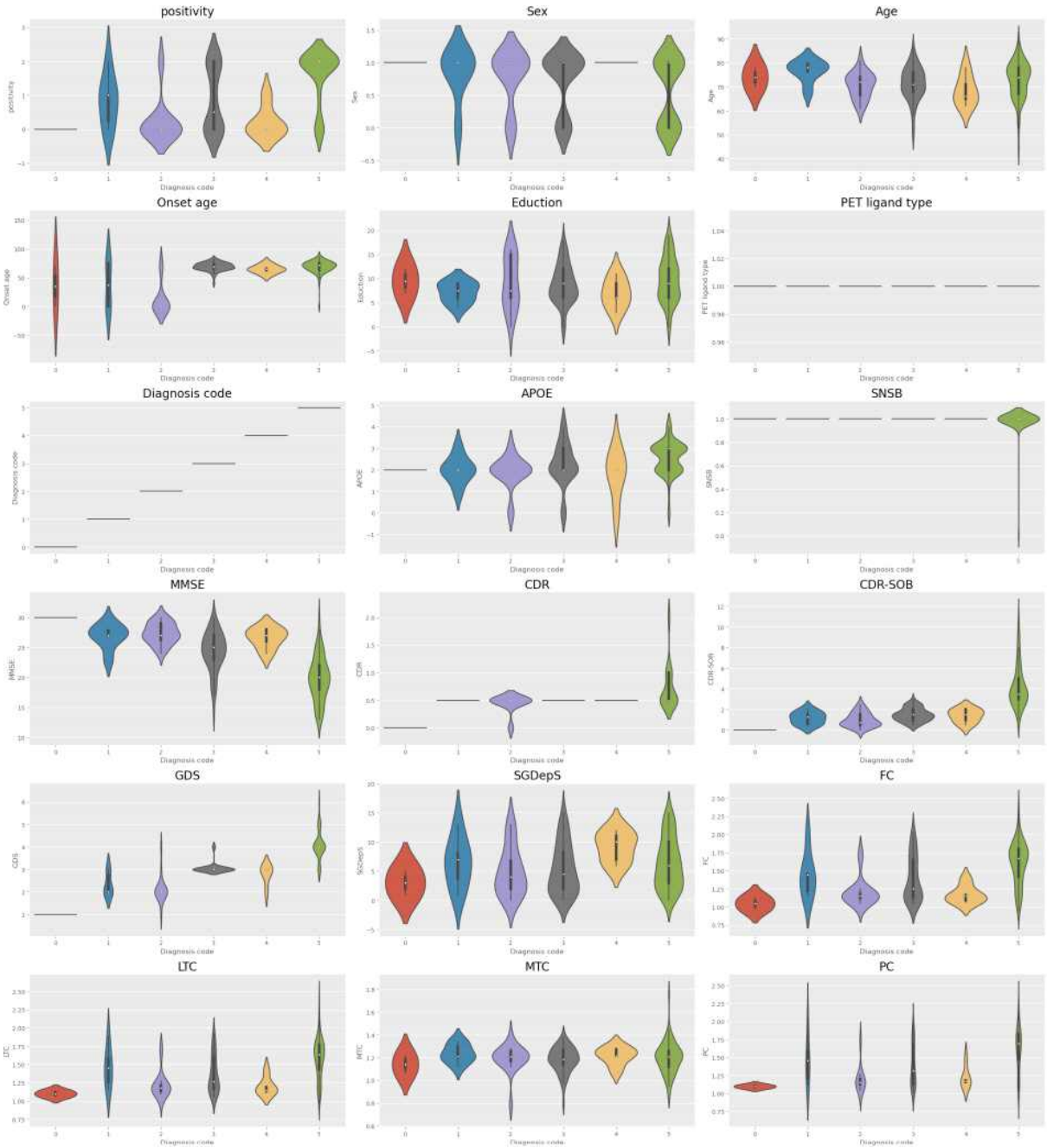
⑤ 다중 변수 분포 분석 : 바이올린 플롯은 한 변수의 분포를 다른 변수의 값에 따라 나누어 보여줄 수 있다. 이를 통해 두 변수 간의 관계를 탐색하고 상관성을 분석할 수 있다.

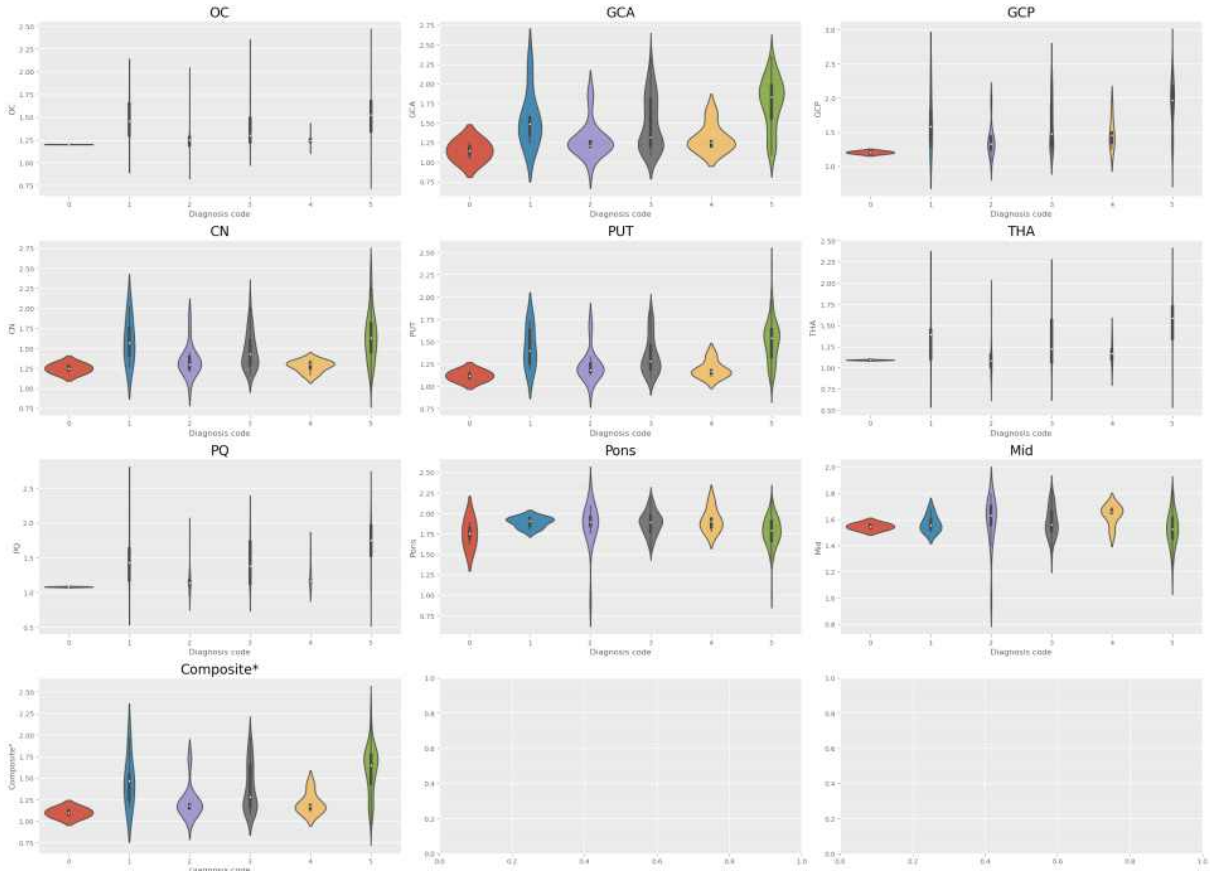
1) FreeSurfer Data




Diagnosis code
Eduction
FC
GCA
GCP
GDS
LTC
MMSE
MTC
OC
Onset age
PC
PQ
PUT
SGDepS
SNSB
Sex
THA
new_APOE
positivity
Diagnosis code

2) PMOD Data


positivity
Sex
Age
Onset age
Eduction
PET ligand type
Diagnosis code
APOE
SNSB
MMSE
CDR
CDR-SOB
GDS
SGDepS
FC
LTC
MTC
PC
Diagnosis code



(4) Before Networking Day

1) FreeSurfer Clean version 데이터

① objective 데이터 타입인 "positivity", "APOE"를 Category method Encoding을 통하여 수행

- positivity 인코딩(파생 변수 생성)

1. positivity : 환자군(BAPL2, BAPL3) ⇒ 1,2 / 대조군(BAPL1) ⇒ 0
2. new_positivity : 환자군(BAPL2, BAPL3) ⇒ 1 / 대조군(BAPL1) ⇒ 0

- APOE 인코딩(파생 변수 생성)

1. new_APOE : E2/E2, E2/E4, E3/E3, E3/E4, E4/E4 ⇒ 0, 1, 2, 3, 4
2. APOE_02/134 : (E2/E2, E3/E3) ⇒ 0 / (E2/E4, E3/E4, E4/E4) ⇒ 1
3. APOE_0123/4 : (E2/E2, E2/E4, E3/E3, E3/E4) ⇒ 0 / E4/E4 ⇒ 1

② 결측치 확인 => "Onset age" feature 결측치 20개 존재

③ 상관분석

Index	Diagnosis Code	Index	Diagnosis Code
GDS	0.7958	OC	0.3724
CDR-SOB	0.6670	new_APOE	0.2509
positivity	0.4618	THA	0.2359
PUT	0.4552	SGDepS	0.1165
GCA	0.4456	Eduction	0.1023
GCP	0.4372	MTC	0.0657
PQ	0.4240	Age	0.0394
FC	0.4105	Onset Age	-0.0086
Composite*	0.4090	SNSB	-0.0755
CDR	0.4058	BS	-0.1115
LTC	0.4056	Sex	-0.2043
CN	0.4023	MMSE	-0.6399
PC	0.3888		

④ 오토 머신러닝(Auto machine learning) 실시

데이터 안의 모든 feature들을 넣고 Pycaret을 사용하여 오토 머신러닝을 실시했다.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.7906	0.1881	0.7906	0.7385	0.7519	0.6581	0.6715	2.2680
lightgbm	Light Gradient Boosting Machine	0.7829	0.1910	0.7829	0.7690	0.7650	0.6571	0.6673	3.1120
xgboost	Extreme Gradient Boosting	0.7755	0.1906	0.7755	0.7434	0.7520	0.6430	0.6511	1.9580
et	Extra Trees Classifier	0.7684	0.1830	0.7684	0.7436	0.7391	0.6300	0.6464	2.1700
gbc	Gradient Boosting Classifier	0.7527	0.1816	0.7527	0.7249	0.7309	0.6084	0.6170	2.7460

상위 5개의 모델일 Random Forest, Light GBM, XGBoost, Extra Tree, Gradient Boosting의 정확도가 0.75 이상의 성능을 보여주었다.

⑤ 결과 중 절댓값이 0.4 이상인 feature만 활용하여 Pycaret을 사용하여 오토 머신러닝을 실시하였다.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.8356	0.1782	0.8356	0.7801	0.8032	0.7387	0.7462	3.2180
lda	Linear Discriminant Analysis	0.8211	0.1806	0.8211	0.7960	0.8043	0.7201	0.7253	2.7220
et	Extra Trees Classifier	0.8211	0.1783	0.8211	0.7883	0.7983	0.7179	0.7268	3.6420
gbc	Gradient Boosting Classifier	0.7903	0.1746	0.7903	0.7839	0.7841	0.6790	0.6844	2.8360
lightgbm	Light Gradient Boosting Machine	0.7838	0.1832	0.7838	0.7756	0.7707	0.6652	0.6733	3.5120
xgboost	Extreme Gradient Boosting	0.7835	0.1863	0.7835	0.7629	0.7687	0.6633	0.6698	3.3940

상위 5개 모델인 Random Forest, Linear Discriminant Analysis, Extra Tree, Gradient Boosting, LightGBM의 정확도가 0.78 이상의 성능을 보여 주었다.

⑥ 앞으로의 분류 모델을 만들 때는 "Diagnosis code"와의 상관관계의 절댓값이 0.4 이상인 Feature들만 활용하기로 했다.

⑦ 0, 1, 2, 3, 4, 5 다중 분류

Method	Accuracy
Random Forest	0.7941
Random Forest, Hyperparameter tuning 실시	0.8235
LDA(선형 판별 분석)	0.7647
LDA, Hyperparameter tuning 실시	0.8213
XGBoost	0.8529
PCA 차원 축소 이후 XGBoost	0.6764

Diagnosis code를 기반으로 다중 분류를 시행했을 때 XGBoost가 Accuracy : 0.8529로 가장 좋은 성능을 보여 주었다.

⑧ 0, 1, 2, 3, 4 ⇒ 0 // 5 ⇒ 1

Method	Accuracy
Random Forest	0.9117
Random Forest, Hyperparameter tuning 실시	0.9117
XGBoost	0.9411
XGBoost, Hyperparameter tuning 실시	0.9411
Random Forest, XGBoost Ensemble	0.9411

Diagnosis code에서 알츠하이머병으로 진단하는 5와 나머지로 인코딩을 진행하여 분류를 시행했을 때 XGBoost가 Accuracy : 0.9411로 가장 좋은 성능을 보여주었다.

⑨ 0, 1, 2 ⇒ 0 // 3, 4 ⇒ 1 // 5 ⇒ 2

Method	Accuracy
Random Forest	0.8529
Random Forest, Hyperparameter tuning 실시	0.8886
XGBoost	0.8235
XGBoost, Hyperparameter tuning 실시	0.8823
Random Forest, XGBoost Ensemble	0.8823

Diagnosis code에서 normal과 SGD(주관적 인지 감퇴), MIC(경도 인지장애), AD(알츠하이머병) 세 가지로 나누어 인코딩을 진행하여 분류를 시행했을 때 Random Forest를 Hyperparameter tuning을 실시했을 때가 Accuracy : 0.8886으로 가장 좋은 성능을 보여주었다.

⑩ Result

FreeSurfer Clean version 데이터를 활용하여 학습한 결과, Diagnosis code에서 알츠하이머병으로 진단하는 5와 나머지로 인코딩을 진행하여 (0, 1, 2, 3, 4 ⇒ 0 // 5 ⇒ 1) 분류를 시행했을 때 XGBoost가 Accuracy : 0.9411로 가장 좋은 성능을 보여주었다.

2) FreeSurfer Raw 데이터

① objective 데이터 타입인 "positivity", "APOE"를 Category method Encoding을 통하여 수행

1. positivity : 환자군(BAPL2, BAPL3) ⇒ 1,2 / 대조군(BAPL1) => 0
2. APOE : (E2/E2, E2/E4, E3/E3) ⇒ 0 // (E3/E4, E4/E4) ⇒ 1

② PET ligand type은 검사약으로 모든 환자에게 같은 약을 썼다. 따라서 제거하였다.

③ 상관분석

Index	Diagnosis Code	Index	Diagnosis Code
GDS	0.7958	Age	0.0394
CDR-SOB	0.6670	(Vol)GCP-r	0.0206
(SUV)PUT-l	0.4596	(Vol)CBL-r	0.0076
(SUV)PUT-r	0.4457	(Vol)GCA-r	-0.0202
(SUV)GCA-l	0.4429	(Vol)GCP-l	-0.0216
(SUV)GCA-r	0.4407	(Vol)CBL-r	-0.0293
(SUV)GCP-l	0.4392	(Vol)CN-r	-0.0441
(SUV)GCP-r	0.4296	(Vol)OC-l	-0.0560
(SUV)PQ-r	0.4274	(Vol)OC-r	-0.0560
(SUV)PQ-l	0.4156	(Vol)CN-l	-0.0597
(SUV)FC-r	0.4146	(Vol)FC-r	-0.0665
(SUV)LTC-r	0.4104	(Vol)GCA-l	-0.0690
positivity	0.4091	SNSB	-0.0755

CDR	0.4058	(Vol)FC-l	-0.0795
(SUV)FC-l	0.4026	(SUV)CBL-r	-0.1102
(SUV)CN-r	0.3963	(Vol)THA-r	-0.1306
(SUV)PC-r	0.3948	(Vol)LTC-r	-0.1136
(SUV)LTC-l	0.3946	(Vol)PC-r	-0.1608
(SUV)CN-l	0.3844	(Vol)PUT-l	-0.1612
(SUV)PC-l	0.3765	(Vol)PUT-r	-0.1666
(SUV)OC-l	0.9724	(Vol)THA-l	-0.1698
(SUV)OC-r	0.3724	(Vol)LTC-l	-0.1891
APOE	0.3214	Sex	-0.2043
(SUV)THA-l	0.2327	(Vol)PQ-r	-0.2051
(SUV)THA-r	0.2280	(Vol)PQ-l	-0.2123
SGDepS	0.1165	(Vol)PQ-l	-0.2615
(SUV)CBL-l	0.1087	(Vol)MTC-l	-0.3585
Eduction	0.1023	(Vol)MTC-r	-0.3832
(SUV)MTC-l	0.0736	MMSE	-0.6399
(SUV)MTC-r	0.0523		

④ H0(base) : 0, 1, 2, 3, 4, 5총 6개의 Diagnosis Code로 분류

Method	Accuracy
Random Forest Classifier	0.8235
Random Forest Classifier with Grid Search(Hyperparameter tuning)	0.7385

⑤ 가설 1 : Diagnosis Code [5]는 1로, [0, 1, 2, 3, 4]는 다 0으로 인코딩

Method	Accuracy
Random Forest Classifier	0.9705
Random Forest Classifier with Grid Search(Hyperparameter tuning)	0.8943

⑥ 가설 2 : Diagnosis Code [4, 5]는 1로, [0, 1, 2, 3]은 0으로 인코딩

Method	Accuracy
Random Forest Classifier	0.8529
Random Forest Classifier with Grid Search(Hyperparameter tuning)	0.9028

⑦ 가설 3 : Diagnosis Code [3, 4, 5]는 1로, [0, 1, 2]는 0으로 인코딩

Method	Accuracy
Random Forest Classifier	0.9411
Random Forest Classifier with Grid Search(Hyperparameter tuning)	0.9475

⑧ 가설 4 : Diagnosis Code [5]는 2로, [3, 4]는 1로, [0, 1, 2]는 0으로 인코딩

Method	Accuracy
Random Forest Classifier	0.9117
Random Forest Classifier with Grid Search(Hyperparameter tuning)	0.8732

⑨ Result

FreeSurfer Raw version 데이터를 활용하여 학습한 결과, 진단 코드 [5]는 1로, [0, 1, 2, 3, 4]로 인코딩한 가설 1 상황에서 Random Forest Classifier로 분류를 시행했을 때, Accuracy : 0.9705로 가장 좋은 성능을 보여주었다.

3) PMOD Clean 데이터

① objective 데이터 타입인 "positivity", "APOE"를 Category method Encoding을 통하여 수행

1. positivity : BAPL1 ⇒ 0 // BAPL2 ⇒ 1 // BAPL3 ⇒ 2
2. APOE : E2/E3 ⇒ 0 // E2/E4 ⇒ 1 // E3/E3 ⇒ 2 // E3/E4 ⇒ 3, E4/E4 ⇒ 4

② 상관분석

Index	Diagnosis Code	Index	Diagnosis Code
positivity	0.46	LTC	0.41
Sex	-0.2	MTC	0.02
Age	0.038	PC	0.41
Onset age	0.49	OC	0.33
Eduction	0.1	GCA	0.43
APOE	0.25	GCP	0.43
SNSB	-0.076	CN	0.33
MMSE	-0.64	PUT	0.36
CDR	0.41	THA	0.39

CDR-SOB	0.67	PQ	0.42
GDS	0.8	Pons	-0.21
SGDepS	0.12	Mid	-0.18
FC	0.41	Composite*	0.41

③ Feature Importance

- XGBoost

Index	F Score	Index	F Score
CDR-SOB	102.0	Pons	46.0
MMSE	102.0	GDS	46.0
Mid	97.0	GCP	44.0
Onset age	77.0	GCA	40.0
Age	77.0	PUT	38.0
MTC	72.0	THA	31.0
CN	64.0	PC	26.0
Eduction	57.0	LTC	23.0
PQ	51.0	OC	18.0
SGDepS	48.0	APOE	17.0

-Random Forest

Index	importance	Index	importance
CDR-SOB	0.1885	PC	0.0256
GDS	0.1854	Age	0.0525
Onset age	0.0873	Composite*	0.0230
MMSE	0.0794	THA	0.0229
LTC	0.0462	CN	0.0217
PUT	0.0345	MTC	0.0217
GCP	0.0291	PQ	0.0199
Pons	0.0271	Mid	0.0197
FC	0.0263	SGDeps	0.0196
GCA	0.0261	OC	0.0185

XGBoost와 Random Forest를 활용하여 각 feature의 중요도를 계산해서 상윗값들을 출력하였다. CDR-SOB가 두 모델 모두 가장 높은 중요도를 보여주었다.

④ 다중 분류 1 : target ⇒ "Diagnosis Code"

- 데이터를 훈련 데이터와 테스트 데이터로 나누어 사용하였다. train : test = 7 : 3 비율로 train_test_split을 진행하였다.

train			test		
Diagnosis Code	Data #	rate(%)	Diagnosis Code	Data #	rate(%)
0	1	0.8547	0	1	1.96.7
1	4	3.4188	1	2	3.9215
2	13	11.1111	2	5	9.8039
3	42	35.8974	3	18	35.2941
4	3	2.5641	4	2	3.9215
5	57	46.1538	5	23	45.0980
Total	117	100	Total	51	100

- 이진 분류 중첩 사용

"Diagnosis Code"의 데이터 불균형 해결을 위해, 가장 빈번한 target 데이터부터 분류 작업을 시작하고, 남은 데이터로 다음으로 빈번한 target을 분류하였다. 이 과정을 반복하여 0부터 5까지의 "Diagnosis Code"를 분류하였다. 순서는 5, 3, 2, 1, 4, 0 순으로 진행하였다.

Classification report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.00	0.00	0.00	2
2	0.57	0.80	0.67	5
3	0.78	0.78	0.78	18
4	0.00	0.00	0.00	2
5	0.88	0.91	0.89	23
accuracy			0.76	51
marco avg	0.37	0.42	0.39	51
weighted avg	0.73	0.76	0.74	51

Confusion matrix

```
[[0,  0,  1,  0,  0,  0]
 [0,  0,  2,  0,  0,  0]
 [0,  0,  5,  0,  0,  0]
 [0,  0,  0, 16,  1,  0]
 [0,  0,  1,  1,  0,  1]
 [0,  0,  0,  2,  0, 21]]
```

정확도만 확인하면 0.76으로 괜찮은 수치이지만, Diagnosis Code가 5, 3을 분류하고 난 후, 2를 분류하는 과정에서 모든 데이터를 2로 판단해버려 1, 4를 모두 맞추지 못하는 문제가 발생하였다.

- XGBoost 다중 분류

분할한 데이터셋을 활용하여 XGBoost로 objective = 'multi:softmax'를 활용하여 다중 분류를 실행하였다.

Classification report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	1.00	0.50	0.67	2
2	0.71	1.00	0.83	5
3	0.84	0.89	0.86	18
4	0.00	0.00	0.00	2
5	0.92	0.96	0.94	23
accuracy			0.86	51
marco avg	0.56	0.56	0.55	51
weighted avg	0.82	0.86	0.84	51

Confusion matrix

```
[[0,  0,  1,  0,  0,  0]
 [0,  0,  2,  0,  0,  0]
 [0,  0,  5,  0,  0,  0]
 [0,  0,  0, 16,  1,  1]
 [0,  0,  1,  1,  0,  0]
 [0,  0,  0,  2,  0, 21]]
```

정확도 0.86으로 준수한 성능의 모델이 만들어졌다. 하지만 0 target, 1 target이 모두 1개씩 존재하여 정확하게 분류하지 못한다고 판단했다.

⑤ 다중 분류 2 : target ⇒ 인지장애 정도

- 주어진 데이터 설명에 따른 인지장애 정도를 기반으로 target 데이터를 새롭게 설정하였다.

Diagnosis Code	인지장애 정도	New Target
0~2	normal SGD(주관적 인지 감퇴)	0
3~4	MIC(경도 인지장애)	1
5	알츠하이머병	2

- 새롭게 데이터를 훈련 데이터와 테스트 데이터로 나누어 사용하였다. 이전과 마찬가지로 train : test = 7 : 3 비율로 train_test_split 했다.

train			test		
Diagnosis Code	Data #	rate(%)	Diagnosis Code	Data #	rate(%)
0	18	15.3846	0	8	15.6862
1	45	38.4615	1	20	39.2156

2	54	46.1538	2	23	45.0980
Total	117	100	Total	51	100

- LightGBM 다중 분류

Classification report

	precision	recall	f1-score	support
0	0.89	1.00	0.94	8
1	0.95	0.90	0.92	20
2	0.96	0.96	0.96	23
accuracy			0.94	51
marco avg	0.93	0.95	0.94	51
weighted avg	0.94	0.94	0.94	51

- XGBoost 다중 분류

Classification report

	precision	recall	f1-score	support
0	0.88	0.88	0.88	8
1	0.95	0.95	0.95	20
2	0.96	0.96	0.96	23
accuracy			0.94	51
maroc avg	0.93	0.93	0.93	51
weighted avg	0.94	0.94	0.94	51

- Random Forest 다중 분류

Classification report

	precision	recall	f1-score	support
0	0.88	0.88	0.88	8
1	0.90	0.90	0.90	20
2	0.96	0.96	0.96	23
accuracy			0.92	51
marco avg	0.91	0.91	0.91	51
weighted avg	0.92	0.92	0.92	0.92

⑥ Result

PMOD Clean version 데이터를 활용하여 학습한 결과, target을 인지장애를 기반으로 0, 1, 2는 0으로 3, 4는 1로 5는 2로 데이터를 설정한 이후 LightGBM으로 분류를 시행했을 때, Accuracy : 0.94로 가장 좋은 성능을 보

여주었다.

4) Before Networking Day result

최종적으로 FreeSurfer Raw version 데이터를 활용하여, 진단 코드 [5]는 1로, [0, 1, 2, 3, 4]는 0으로 인코딩한 가설 1 상황에서 Random Forest Classifier 모델로 환자 분류를 시행했을 때, Accuracy : 0.9705로 가장 좋은 성능을 보여주었다.

하지만 데이터별로 다른 전처리 방법과 다른 학습을 사용하여 정확도에서 신뢰성이 떨어진다. 따라서 다음 연구를 진행하였다.

(5) After Networking Day

1) Introduction

국가수리과학연구소 부산 의료수학센터에서 개최한 산학연 협력 기반 산업수학 문제해결 네트워킹 데이에 갔다 온 이후 새로운 방향으로 연구를 진행하였다.

PMOD 데이터와 FreeSurfer 데이터를 합쳐서 진행하였고, 다양하게 target 데이터를 정하여 학습을 진행하였다.

2) 데이터 전처리

① PMOD Clean 데이터와 FreeSurfer Clean 데이터 병합

공통 Feature이며 데이터도 같은 개인정보 및 신경 심리 검사는 하나만 사용하였다.

['ID', 'positivity', 'Sex', 'Age', 'Onset age', 'Eduction', 'Diagnosis code', 'APOE', 'SNSB', 'MMSE', 'CDR', 'CDR-SOB', 'GDS', 'SGDepS']

② 두 데이터의 PET 영상을 Feature들 중 중복되는 것들은 각각의 평균으로 처리하였다.

['FC', 'LTC', 'PC', 'OC', 'GCA', 'GCP', 'PQ','MTC', 'CN', 'PUT', 'THA', 'Composite*']

③ 두 데이터에 공통되지 않고 각각 있는 데이터들은 그대로 사용하였다.

- FreeSurfer → ['BS', 'CBL']
- PMOD → ['CBLCTX', 'Mid', 'Pons']

④ object 데이터 타입인 "positivity", "APOE"를 Category method Encoding

을 통하여 수행하였다.

- 'positivity': BALP1 → 0 / BALP2/3 →1
- 'APOE': E2/E3, E2/E4, E3/E3 → 0 / E3/E4, E4/E4 → 1

⑤ 단일데이터 Column 제거

'PET ligand type', 'CBL', 'CBLCTX' column은 모두 값이 1인 데이터만 존재해서 삭제하였다.

⑥ 결측치 처리

'Onset Age' feature에 20개의 결측치가 존재하였다. 평균값 대체, 중앙값 대체와 선형회귀, XGBRegresoor, RandomForestRegressor의 방법을 사용해 보았고, 'Onset Age'와 선형 관계를 가지며 상관관계가 매우 높은 'Age'와 나머지 변수 중에 상관관계가 상위 2개인 'BS', 'SNSB'를 사용하여 모델의 학습을 진행했다.

Method	positivity	Diagnosis Code
평균값 대체	0.146151	-0.007137
중앙값 대체	0.133587	-0.028932
선형회귀	0.187087	0.012075
XGB	0.184602	0.006441
RF	0.185550	0.011959

결측치 처리 이후의 'Onset age'와 target 데이터로 사용할 'positivity', 'Diagnosis cod'의 상관관계가 가장 높은 선형회귀 방법을 채택하여 결측치를 처리했다.

3) 1차 모델링

① target 데이터 : 'Diagnosis code' : 5 ⇒ 1 // [0, 1, 2, 3, 4] ⇒ 0

Diagnosis code는 신경심리검사 결과로 알츠하이머를 진단할 때 사용하는 방법 중 하나이다. 0에서 5까지 총 6가지로 분류될 수 있지만 기준에 따라 더 포괄적인 분류도 가능하다. 이 모델링을 진행하는 목적인 알츠하이머는 경미한 인지장애와는 차이가 있다고 생각하여 값이 5인 Diagnosis code만을 알츠하이머라고 분류 하였다.

또한 이전에 실행했던 모델링에서 FreeSurfer Raw 데이터를 사용하여

target 데이터를 'Diagnosis code'로 설정하고, 값이 5일 때는 알츠하이머, 나머지 값은 정상으로 분류하였을 때 성능이 제일 좋았으므로 새로 생성한 데이터에 동일하게 시도하였다.

② 상관 분석

Diagnosis code와의 상관관계			
GDS	0.727094	OC	0.358633
CDR-SOB	0.701848	APOE	0.300501
GCA	0.443342	Eduction	0.118357
PUT	0.438358	Age	0.098601
GCP	0.434930	SGDepS	0.091770
PQ	0.429478	ID	0.075908
LTC	0.419607	Onset age	0.063636
positivity	0.417806	MTC	0.054784
Composite*	0.415314	SNSB	-0.084123
PC	0.414880	BS	-0.147222
FC	0.408992	Sex	-0.211183
CDR	0.396167	Mid	-0.220719
THA	0.394502	Pons	-0.261733
CN	0.394501	MMSE	-0.623521

※ 빨간색: 매우 높은 상관관계, 파란색: 다소 높은 상관관계

③ 모든 feature 사용

			XGBClassifier		RandomForestClassifier	
accuracy score			0.941176		0.941176	
recall score			0.937728		0.952380	
precision score			0.937728		0.933333	
f1 score			0.937728		0.939285	
confusion matrix	TN	FP	20	1	19	2
	FN	TP	1	12	0	3

④ 상관계수가 –0.4 이하 +0.4 이상인 feature들만 사용

['GDS','CDR-SOB','MMSE','GCA','PUT','GCP','PQ','LTC','positivity','Composite*','PC','FC ']

			XGBClassifier		RandomForestClassifier	
accuracy score			0.976190		0.952380	
recall score			0.978260		0.956521	
precision score			0.975		0.952380	
f1 score			0.976068		0.952272	
confusion matrix	TN	FP	22	1	21	2
	FN	TP	0	19	0	19

⑤ 상관계수가 –0.7 이하 +0.7 이상인 feature들만 사용
['GDS','CDR- SOB']

			XGBClassifier		RandomForestClassifier	
accuracy score			0.880952		0.904761	
recall score			0.882151		0.894736	
precision score			0.879545		0.925925	
f1 score			0.880341		0.901176	
confusion matrix	TN	FP	20	3	23	0
	FN	TP	2	17	4	15

⑥ Result

1차 모델링을 통해 Diagnosis code가 신경 심리 검사 결과인 만큼 신경 심리 검사들인 'GDS', 'CDR-SOB', 'MMSE' 들과의 상관관계가 매우 높다는 것을 알 수 있다. 반면에 PET 영상 데이터와는 상관관계가 높지 않다는 것을 알 수 있다. 따라서 신경 심리 검사들만으로 알츠하이머를 진단하기에는 다소 부족하다고 판단하였다.

4) 2차 모델링

① target 데이터 : 'positivity'

positivity는 의사의 진단 데이터로 BAPL2, BAPL3은 환자군, BAPL1은 대조군을 의미하고 앞서 진행한 전처리를 통해 BAPL2, BAPL3는 1로, BAPL1은 0으로 인코딩되었다.

1차 모델링에서 target 데이터로 사용한 Diagnosis code로도 알츠하이머를 진단할 수 있지만 알츠하이머는 의사의 임상적 진단에 의존하는 질병이므로 결론적으로 병의 유무를 판단하는 의사의 결정이 중요하다고 할 수 있다. 따

라서 신경 심리 검사로 나온 Diagnosis code가 아닌 의사의 진단 positivity로 모델링을 진행하였다.

② 상관분석

Positivity와의 상관관계			
PQ	0.867611	CDR-SOB	0.358633
GCP	0.861187	MTC	0.300501
PC	0.848809	CDR	0.118357
GCA	0.848655	Age	0.098601
Composite*	0.844886	Onset age	0.091770
FC	0.840276	Eduction	0.075908
LTC	0.829676	Onset age	0.063636
CN	0.796282	Sex	0.054784
PUT	0.777077	SNSB	-0.084123
THA	0.769440	SGDepS	-0.147222
OC	0.763990	BS	-0.211183
Diagnosis code	0.409148	Pons	-0.220719
APOE	0.392635	Mid	-0.261733
GDS	0.366383	MMSE	-0.623521

※ 빨간색: 매우 높은 상관관계, 파란색: 다소 높은 상관관계

③ 모든 feature 사용

			XGBClassifier		RandomForestClassifier	
accuracy score			1.0		1.0	
recall score			1.0		1.0	
precision score			1.0		1.0	
f1 score			1.0		1.0	
confusion matrix	TN	FP	17	0	17	0
	FN	TP	0	25	0	25

④ 상관 계수가 –0.4 이하 +0.4 이상인 feature들만 사용

['Diagnosis code','FC','LTC','PC','OC','GCA','GCP','PQ','CN','PUT','THA','Composite*']

	XGBClassifier	RandomForestClassifier
accuracy score	1.0	0.970588
recall score	1.0	0.975
precision score	1.0	0.966666

f1 score			1.0		0.969938	
confusion matrix	TN	FP	14	0	14	0
	FN	TP	0	20	1	19

⑤ 상관 계수가 –0.7 이하 +0.7 이상인 feature들만 사용
['FC','LTC','PC','OC','GCA','GCP','PQ','CN','PUT','THA','Composite*']

			XGBClassifier		RandomForestClassifier	
accuracy score			1.0		1.0	
recall score			1.0		1.0	
precision score			1.0		1.0	
f1 score			1.0		1.0	
confusion matrix	TN	FP	14	0	14	0
	FN	TP	0	20	0	20

⑥ Result

positivity를 target 데이터로 하여 모델링을 진행한 결과 Diagnosis code를 target 데이터로 하였을 때보다 모델들의 성능이 더 좋아진 것을 확인할 수 있다. 하지만 positivity와의 상관관계를 보았을 때, 매우 높은 상관 계수를 가진 feature들은 모두 PET 영상 데이터인 것을 알 수 있다. 이는 의사의 진단이 신경 심리 검사가 아닌 PET 영상 데이터에 높은 의존도를 가진다는 것을 의미한다. 즉 의사가 PET 영상 데이터만으로 진단한다는 것이다.

앞서 진행한 모델링에서 언급한 바와 같이 하나의 검사만으로 알츠하이머를 진단할 수 없다. 따라서 target 데이터로 Diagnosis code와 positivity 모두 고려해 봐야 할 필요가 있다.

5) 3차 모델링

① target 데이터 : Diagnosis code + positivity

알츠하이머는 하나의 검사로 판단할 수 있는 질병이 아닐뿐더러 의사의 임상적 판단에 의존하는 신경 심리 검사 결과와 의사의 판단 모두 고려해야 할 필요가 있다. 하지만 주어진 데이터에서 신경 심리 검사 결과와 의사의 판단이 상반되는 경우들이 발생하여 두 데이터를 이용한 새로운 답안 column을 생성하여 알츠하이머 환자를 분류하는 모델링을 진행하였다. 또한 알츠하이머가 가지는 여러 특성을 고려하여 target 데이터를 3종류로 만들어 3가지의

모델을 개발하였다.

② Case 1
- 신경 심리 검사 결과와 의사 진단 모두 알츠하이머인 경우 ⇒ 2
- 신경 심리 검사 결과와 의사 진단 둘 중 하나만 알츠하이머인 경우 ⇒ 1
- 신경 심리 검사 결과와 의사 진단 모두 정상인 경우로 분류 ⇒ 0

알츠하이머를 3개의 클래스로 분류함으로써 정상, 초기 알츠하이머, 알츠하이머를 구별하여 진단해주는 모델링을 진행해 보았다.

● 상관분석

Case 1 answer와의 상관관계			
PQ	0.768317	APOE	0.411204
GCP	0.767792	CDR	0.367450
GCA	0.765438	MTC	0.199919
PC	0.748501	Age	0.186690
Composite*	0.746448	Onset age	0.148340
LTC	0.740053	Eduction	0.147539
FC	0.739949	SGDepS	-0.015219
PUT	0.720270	SNSB	-0.087940
CN	0.705349	Sex	-0.136458
THA	0.689530	BS	-0.198852
OC	0.664857	Mid	-0.295166
GDS	0.650959	Pons	-0.316207
CDR-SOB	0.596691	MMSE	-0.582262

● 모든 feature 사용

	XGBClassifier			RandomForestClassifier		
accuracy score	0.823529			0.882352		
recall score	0.818840			0.882051		
precision score	0.823076			0.877777		
f1 score	0.816666			0.878840		
confusion matrix	11	0	0	11	0	0
	1	7	2	1	8	1
	0	3	10	0	2	11

● 상관 계수가 –0.4 이하 +0.4 이상인 feature들만 사용

['APOE','MMSE','CDR-SOB','GDS','FC','LTC','PC','OC','GCA','GCP','PQ','CN','PUT','THA','Composite*']

	XGBClassifier			RandomForestClassifier		
accuracy score	0.823529			0.911764		
recall score	0.823076			0.907692		
precision score	0.816666			0.915343		
f1 score	0.818840			0.910331		
confusion matrix	11	0	0	11	0	0
	1	7	2	0	8	2
	0	3	10	0	1	12

● 상관 계수가 -0.7 이하 +0.7 이상인 feature들만 사용

['FC', 'LTC', 'PC', 'GCA', 'GCP', 'PQ', 'CN', 'PUT', 'Composite*']

	XGBClassifier			RandomForestClassifier		
accuracy score	0.735294			0.676470		
recall score	0.689855			0.650116		
precision score	0.710722			0.644817		
f1 score	0.713071			0.638281		
confusion matrix	10	1	0	8	3	0
	2	3	5	2	3	5
	0	1	12	0	1	12

③ Case 2

- 신경 심리 검사 결과와 의사 진단 모두 알츠하이머인 경우 ⇒ 1
- 신경 심리 검사 결과와 의사 진단 둘 중 하나라도 알츠하이머가 아닌 경우 ⇒ 0

알츠하이머라는 질병이 우울증이나 다른 정신 질환 증상과 비슷하여 오진할 가능성도 있으므로 신경 심리 검사 결과와 의사의 진단 모두 알츠하이머라 판단한 경우만 알츠하이머로 분류하는 모델링을 진행해 보았다.

● 상관분석

Positivity와의 상관관계			
PQ	0.642173	APOE	0.402997
GCA	0.636639	CDR	0.357398
GCP	0.636622	MTC	0.221848
PC	0.623585	Education	0.191153
LTC	0.619473	Age	0.089440
Composite*	0.617046	Onset age	0.056545
PUT	0.607060	SGDepS	-0.070927
FC	0.603318	SNSB	-0.099900
THA	0.588992	Sex	-0.104328
CN	0.586593	BS	-0.203144
GDS	0.586094	Mid	-0.244004
CDR-SOB	0.582376	Pons	-0.315194
OC	0.550451	MMSE	-0.512052

● 모든 feature 사용

	XGBClassifier		RandomForestClassifier	
accuracy score	0.928571		0.904761	
recall score	0.942307		0.899038	
precision score	0.921052		0.899038	
f1 score	0.926530		0.899038	
confusion matrix	23	3	24	2
	0	16	2	14

● 상관 계수가 –0.4 이하 +0.4 이상인 feature 들만 사용

['APOE','MMSE','CDR-SOB','GDS','FC','LTC','PC','OC','GCA','GCP','PQ','CN','PUT','THA','Composite*']

	XGBClassifier		RandomForestClassifier	
accuracy score	0.928571		0.904761	
recall score	0.918269		0.887019	
precision score	0.929629		0.929629	
f1 score	0.923311		0.896296	
confusion matrix	25	1	25	1
	2	14	3	13

● 상관 계수가 –0.7 이하 +0.7 이상인 feature 들만 사용
=> 존재하지 않아서 진행 X

④ Case 3
- Diagnosis code가 3~5이면서 positivity가 1이거나
Diagnosis code가 5이면서 positivity가 0인 경우 ⇒ 1
- 그 외의 경우 ⇒ 0

알츠하이머를 진단할 때 사용하는 PET 영상의 베타 아밀로이드 침착 정도도 중요하다. 하지만 겉으로 발현되는 행동이나 심리 등의 증상 또한 무시할 수 없다. 따라서 알츠하이머의 가능성이 보이는 사람들도 미리 알츠하이머로 진단하여 적절한 치료 및 추적 관찰을 할 수 있도록 증상과 의사의 진단을 모두 고려하는 모델링을 진행하였다.

● 상관분석

Positivity와의 상관관계			
GCP	0.664150	APOE	0.319246
GCA	0.658137	CDR	0.299344
PQ	0.653714	Age	0.179990
FC	0.652239	Onset age	0.151019
Composite*	0.641270	MTC	0.098847
GDS	0.639166	Education	0.097784
PC	0.631852	SGDepS	0.063815
LTC	0.629375	SNSB	-0.057677
PUT	0.613983	Sex	-0.131762
CN	0.600768	BS	-0.161264
THA	0.569481	Pons	-0.240374
OC	0.553114	Mid	-0.268018
CDR-SOB	0.499855	MMSE	-0.551808

● 모든 feature 사용

	XGBClassifier	RandomForestClassifier
accuracy score	0.970588	0.970588
recall score	0.977272	0.958333
precision score	0.961538	0.978260

f1 score	0.968372		0.967149	
confusion matrix	12	0	11	1
	1	21	0	21

● 상관 계수가 –0.4 이하 +0.4 이상인 feature 들만 사용

['MMSE','CDR-SOB','GDS','FC','LTC','PC','OC','GCA','GCP','PQ','CN','PUT','THA','Composite*']

	XGBClassifier		RandomForestClassifier	
accuracy score	1.0		1.0	
recall score	1.0		1.0	
precision score	1.0		1.0	
f1 score	1.0		1.0	
confusion matrix	12	0	12	0
	0	22	0	22

● 상관 계수가 –0.7 이하 +0.7 이상인 feature 들만 사용

⇒ 존재하지 않아서 진행 X

⑤ Result

positivity와 Diagnosis code를 모두 고려해 새로 생성한 3가지의 target 데이터를 사용해 모델링을 진행해 본 결과, 3가지 경우 모두 PET 영상과 신경 심리 검사 데이터와의 상관관계가 괜찮은 상관관계를 가지는 것을 알 수 있다. 또한 더 많고 다양한 feature 들이 다소 높은 상관관계 또는 높은 상관관계를 가지고 있어 모델 학습에 더 많은 변수들이 영향을 주었다. 결론적으로 여러 Case 중, Case 3의 모델들이 가장 좋은 성능을 보였다.

4. Summary & Conclusion

FreeSurfer Clean version 데이터를 활용하여 학습한 결과, Diagnosis code에서 알츠하이머병으로 진단하는 5와 나머지로 인코딩을 진행하여(0, 1, 2, 3, 4 => 0 // 5 => 1) 분류를 시행했을 때 XGBoost가 Accuracy : 0.9411로 가장 좋은 성능을 보여주었다.

FreeSurfer Raw version 데이터를 활용하여 학습한 결과, 진단 코드 [5]는 1로, [0, 1, 2, 3, 4]로 인코딩한 가설 1 상황에서 Random Forest Classifier로 분류를

시행했을 때, Accuracy : 0.9705로 가장 좋은 성능을 보여주었다.

PMOD Clean version 데이터를 활용하여 학습한 결과, target을 인지장애를 기반으로 0, 1, 2는 0으로 3, 4는 1로 5는 2로 데이터를 설정한 이후 LightGBM으로 분류를 시행했을 때, Accuracy : 0.94로 가장 좋은 성능을 보여주었다.

하지만 데이터별로 다른 전처리 방법과 다른 학습을 사용하여 각각의 모델들을 정확도로만 판단하는 것이 신뢰성이 떨어져 추가 연구를 진행하였다.

PMOD Clean 데이터와 FreeSurfer 데이터를 합쳐서 진행하였고, target 데이터를 'Diagnosis code', 'positivity'로 각각 진행 해 보았다. target 데이터가 'Diagnosis code'인 경우에는 신경 심리 검사 feature 들과의 상관관계는 높았지만, PET 영상 데이터와는 상관관계가 높지 않아서 알츠하이머를 진단하기에는 다소 부족하다고 판단하였다. 반대로 target 데이터가 'positivity'인 경우에는 신경 심리 검사 feature 들의 상관관계는 낮고, PET 영상 데이터와의 상관관계는 높아 알츠하이머를 진단하기에는 다소 부족하다고 판단하였다. 따라서 하나의 검사만으로 알츠하이머를 진단할 수 없다고 판단하여, target 데이터로 Diagnosis code와 positivity 모두 고려해 보았다.

positivity와 Diagnosis code를 모두 고려해 새로운 target 데이터를 사용해 모델링한 결과, Diagnosis code가 3~5이면서 positivity 가 1이거나 Diagnosis code 가 5이면서 positivity가 0인 경우를 1로 인코딩하고, 그 외의 경우를 0으로 했던 경우에 가장 좋은 성능을 보여주었다.

5. Utilization Plan & Expected Effect

프로젝트에서 PMOD 데이터와 FreeSurfer 데이터를 융합하여 모델링한 결과 더 높은 성능을 얻을 수 있었다. 이러한 결과는 다양한 데이터 소스를 활용하여 알츠하이머병 진단 모델을 구축하는 것의 중요성을 강조한다. 의료 영상 데이터와 신경 심리 검사 데이터를 조합함으로써 종합적인 분석이 가능해지며, 이는 진단의 정확성을 높이는 데 큰 도움이 될 것이다.

또한 분석 결과에서 하나의 검사만으로 알츠하이머병을 정확하게 진단하기는 어려움을 확인할 수 있다. 이에 따라 더 정확한 진단을 위해서는 다양한 변수를 고려하고 개인의 특성을 반영하는 맞춤형 접근이 필요하다. 의료 분야에서는 환자 개인의 변화와 특성을 고려하여 개인 맞춤형 치료와 관리를 제공하는 접근법이 강조될 것이다.

마지막으로 이 프로젝트는 의료 분야에서 인공지능과 데이터 분석기술을 접목하는 하나의 예시이다. 이러한 연구는 의료 분야의 인공지능 응용 확대 및 의료 기술의 혁신을 촉진할 수 있다. 더 나아가서는 이러한 모델들이 실제 임상 환경에서 활용되어 의사와 환자들에게 실질적인 도움을 제공하는데 기여할 수 있을 것이다.

6. Future Work

168명의 환자 데이터가 제공되었는데 데이터의 양이 적어서 분류가 어려웠고, Diagnosis code와 의사의 진단인 positivity가 상반된 경우가 꽤 많아 알츠하이머 진단 모델을 개발하는 데 어려움을 겪었다. 의사들이 어떤 과정과 근거에 의해 알츠하이머를 진단하는 지 좀 더 자세히 알 수 있으면 모델 개발에 더 도움이 될 것이다.

7. Reference

* 최용인(2023), 정신질환 관련 환자군 분류기법 개발(동아대학교병원, 핵의학과), 국가수리과학연구소
* 윤현진, 정용진, 박경원, 강도영(2023), classification and clustering with machine learning for alzheimer's dementia, 동아대학교의료원
* 서은현(2018), 치매 및 인지장애에서의 신경심리평가(조선대학교 의과대학 의예과), JKNA(Journal of Korean Neuropsychiatric Association)