

1. 기본 정보 및 인구통계 (Basic Info & Demographics)

컬럼명	설명 (Description)	분석 활용
rid	Roster ID (환자 고유 번호, 숫자)	데이터 병합 Key
subject_id	환자 식별 문자열 (예: 027_S_6001)	시각화 라벨링
ptdobyy	출생 연도	나이 검증용
age	나이 (방문일 기준)	필수 공변량
gender	성별	1(Male), 2(Female)
enrlfg	등록 여부 (Enrollment Flag)	데이터 필터링
apoe4	APOE ε4 유전자 개수 (0, 1, 2)	알츠하이默 고위험 인자 / 대립유전자 개수 (0, 1, 2)

2. 연구 메타데이터 (Study Metadata)

컬럼명	설명	분석 활용
viscode2	방문 시점 코드 (bl, m06, m12 등)	시계열 시점 확인
examdate	검사 날짜	시간 간격 계산
image_id	매칭된 영상 이미지 ID	영상 데이터 로드 Key
origprot	최초 참여 프로토콜 (ADNI1/2/3)	코호트 구분
colprot	데이터 수집 프로토콜	데이터 출처 확인

3. 진단 및 질병 단계 (Diagnosis & Staging)

컬럼명	설명	분석 활용
diagnosis	최종 진단명 (CN, MCI, AD)	1(CN), 2(MCI), 3(AD)
mci_subtype	MCI 세부 타입 (Early/Late)	1(EMCI), 2(LMCI)
cdr_global	CDR (임상 치매 척도) 종합 점수	치매 심각도 라벨
cdr_sb	CDR-SB (Sum of Boxes)	진행도 예측 타겟 (연속형)
cdr_memory	CDR 항목 중 '기억력' 점수	기억력 특화 분석

```
# 조건 정의
# pMCI (2): 관찰 기간 상관없이 AD로 전환된 경우
cond_pmc = df_results['adConverted'] == True

# sMCI (1): AD로 전환되지 않았으면서, 24개월 이상 추적 관찰된 경우 (안정 or 회복)
cond_smci = (df_results['adConverted'] == False) & (df_results['maxFollowupMonth'] >= 24)
```

4. 인지 및 임상 척도 (Cognitive & Clinical Scores)

컬럼명	설명	분석 활용
mmse	MMSE 총점 (30점 만점)	인지 선별 검사
moca	MoCA 총점 (30점 만점)	MCI 민감도 높은 검사
adas_cog13	ADAS-Cog 13 문항 총점	임상시험 주요 지표
gds	노인 우울 척도	우울증 감별
faq	기능적 활동 설문 (IADL)	일상생활 수행 능력 평가
cci	찰슨 동반질환 지수	기저 질환 영향 보정

5.1 세부 신경심리 검사 (Neuropsychological Tests) - 언어 기억력 (AVLT / RAVLT)

컬럼명	설명 (Description)	분석적 의미
RAVLT_immediate	즉각 회상 점수	단어를 듣자마자 얼마나 기억하는지 (단기 기억)
RAVLT_learning	학습 곡선	반복해서 들려줄 때 기억량이 얼마나 늘어나는지 (학습 능력)
RAVLT_forgetting	망각률	학습 후 일정 시간이 지났을 때 잊어버리는 정도
RAVLT_pct_forgetting	망각 백분율 (%)	망각률을 백분율로 환산한 수치
AVTOT1 ~ AVTOT5	1~5회차 시도별 점수	회차별 기억력 상승 패턴 분석용
AVTOT6	간접 시행 점수	새로운 단어(List B)를 들려준 직후의 기존 단어 기억량
AVTOTB	간접 리스트(List B) 점수	방해 자극(새로운 단어)에 대한 기억력
AVDEL30MIN	30분 지연 회상	장기 기억력을 평가하는 핵심 지표 (가장 중요)
AVDELTOT	지연 회상 총점	지연 회상 능력의 종합 점수
AVERR1, AVERR2	오답 횟수	없는 단어를 말했다고 착각하는 횟수 (작화증 등 확인)
nxaudio	청각 언어 기억력 종합 점수	언어적 기억력을 종합한 표준화 점수 / 1(Absent), 2(Present)

5.2 세부 신경심리 검사 (Neuropsychological Tests) - 주의력 및 집행 기능 (Trail Making Test)

컬럼명	설명 (Description)	분석적 의미
TRAASCOR	TMT-A 소요 시간	숫자(1-2-3...)를 잊는 시간 → 단순 주의력 및 처리 속도
TRABSCOR	TMT-B 소요 시간	숫자와 문자(1-A-2-B...)를 번갈아 잊는 시간 → 집행 기능, 인지적 유연성

6. 바이오마커 (Biomarkers)

컬럼명	설명	분석적 의미
ABETA40	아밀로이드 베타 40 농도	병리 지표
ABETA42	아밀로이드 베타 42 농도	알츠하이머 핵심 병리
TAU	총 타우 (Total Tau)	신경세포 손상 지표
PTAU	인산화 타우 (p-Tau)	신경섬유 엉킴 지표
AB42/AB40	아밀로이드 42/40 비율	진단 정확도 보정 지표
ptau/AB42	p-Tau / A β 42 비율	강력한 알츠하이머 예측 인자
ttau/AB42	t-Tau / A β 42 비율	강력한 알츠하이머 예측 인자

Summary

Dataset Information:

- Total Subjects: 235
- Total Variables: 54
- Numeric Variables: 41
- Categorical Variables: 11
- Missing Values: 785 (6.19% of total data)

Demographics:

- Age: 72.4 ± 7.4 years (range: 56-91)
- Gender: Male: 129 (54.9%), Female: 106 (45.1%)
- APOE4 Status:
 - * 0 allele(s): 136 (57.9%)
 - * 1 allele(s): 77 (32.8%)
 - * 2 allele(s): 20 (8.5%)
 - * Total carriers: 97 (41.3%)
- Diagnosis: MCI: 235 (100.0%)
- MCI Subtype: LMCI: 130 (55.3%), EMCI: 105 (44.7%)
- Neurological Exam (Auditory): Absent: 211 (89.8%), Present: 24 (10.2%)

[목차]

1.일반 EDA

2.분류 모델 EDA

1. 일반 EDA

컬럼명과 데이터 타입

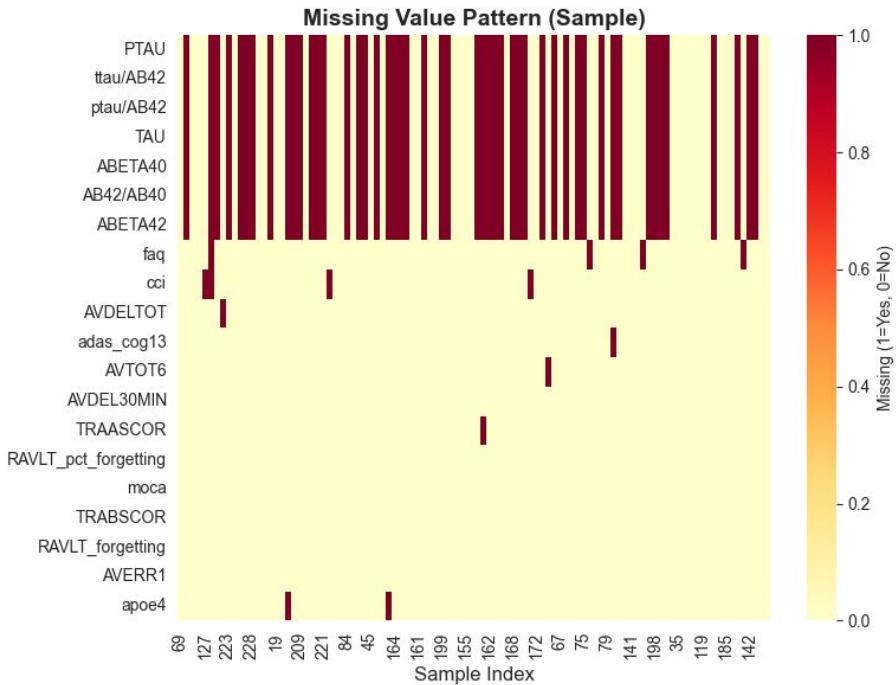
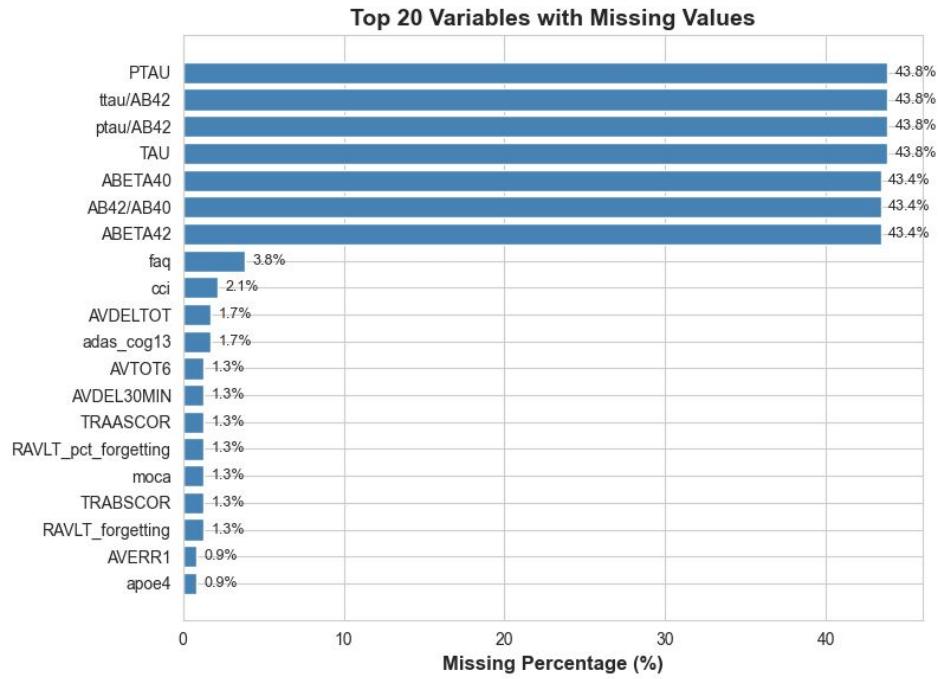
rid	int64
subject_id	object
ptdobyy	object
age	int64
gender	int64
apoe4	float64
image_id	float64
viscode2	object
examdate	object
origprot	object
colprot	object
enrlfg	object
diagnosis	int64
cdr_global	float64
cdr_memory	float64
cdr_sb	float64
mci_subtype	int64
mmse	int64
adas_cog13	float64
cci	float64
gds	float64
faq	float64
nxaudito	int64

AVDEL30MIN	float64
AVDELTOT	float64
AVERR1	float64
AVERR2	float64
AVTOT1	float64
AVTOT2	float64
AVTOT3	float64
AVTOT4	float64
AVTOT5	float64
AVTOT6	float64
AVTOTB	float64
RAVLT_forgetting	float64
RAVLT_immediate	float64
RAVLT_learning	float64
RAVLT_pct_forgetting	float64
TRAASCOR	float64
TRABSCOR	float64
ABETA40	float64
ABETA42	float64
PTAU	float64
TAU	float64
AB42/AB40	float64
ptau/AB42	float64
ttau/AB42	float64
moca	float64

결측치 요약

Column	Missing_Count	Missing_Percent
PTAU	103	43.829787
ttau/AB42	103	43.829787
ptau/AB42	103	43.829787
TAU	103	43.829787
ABETA40	102	43.404255
AB42/AB40	102	43.404255
ABETA42	102	43.404255
faq	9	3.829787
cci	5	2.127660
AVDELTOT	4	1.702128
adas_cog13	4	1.702128
AVTOT6	3	1.276596
AVDEL30MIN	3	1.276596
TRAASCOR	3	1.276596
RAVLT_pct_forgetting	3	1.276596
moca	3	1.276596
TRABSCOR	3	1.276596
RAVLT_forgetting	3	1.276596
AVERR1	2	0.851064
apoe4	2	0.851064
AVERR2	2	0.851064
AVTOT3	2	0.851064
AVTOT2	2	0.851064
AVTOT1	2	0.851064
RAVLT_learning	2	0.851064
AVTOT5	2	0.851064
AVTOT4	2	0.851064
AVTOTB	2	0.851064
RAVLT_immediate	2	0.851064
image_id	1	0.425532
gds	1	0.425532

결측치가 있는 컬럼 수: 31
총 결측치 수: 785



Demographic and Clinical Characteristics

--- 나이 분포 ---

평균 나이: 72.41 ± 7.42

중앙값: 73.00

범위: 56 - 91

--- 성별 분포 ---

gender_label

Male 129

Female 106

Name: count, dtype: int64

Male : Female = 129:106

--- APOE4 Status ---

apoe4

0.0 136

1.0 77

2.0 20

Name: count, dtype: int64

APOE4 carriers: 97 (41.3%)

--- MCI Subtype Distribution ---

mci_subtype_label

LMCI 130

EMCI 105

Name: count, dtype: int64

--- Diagnosis Distribution ---

diagnosis_label

MCI 235

Name: count, dtype: int64

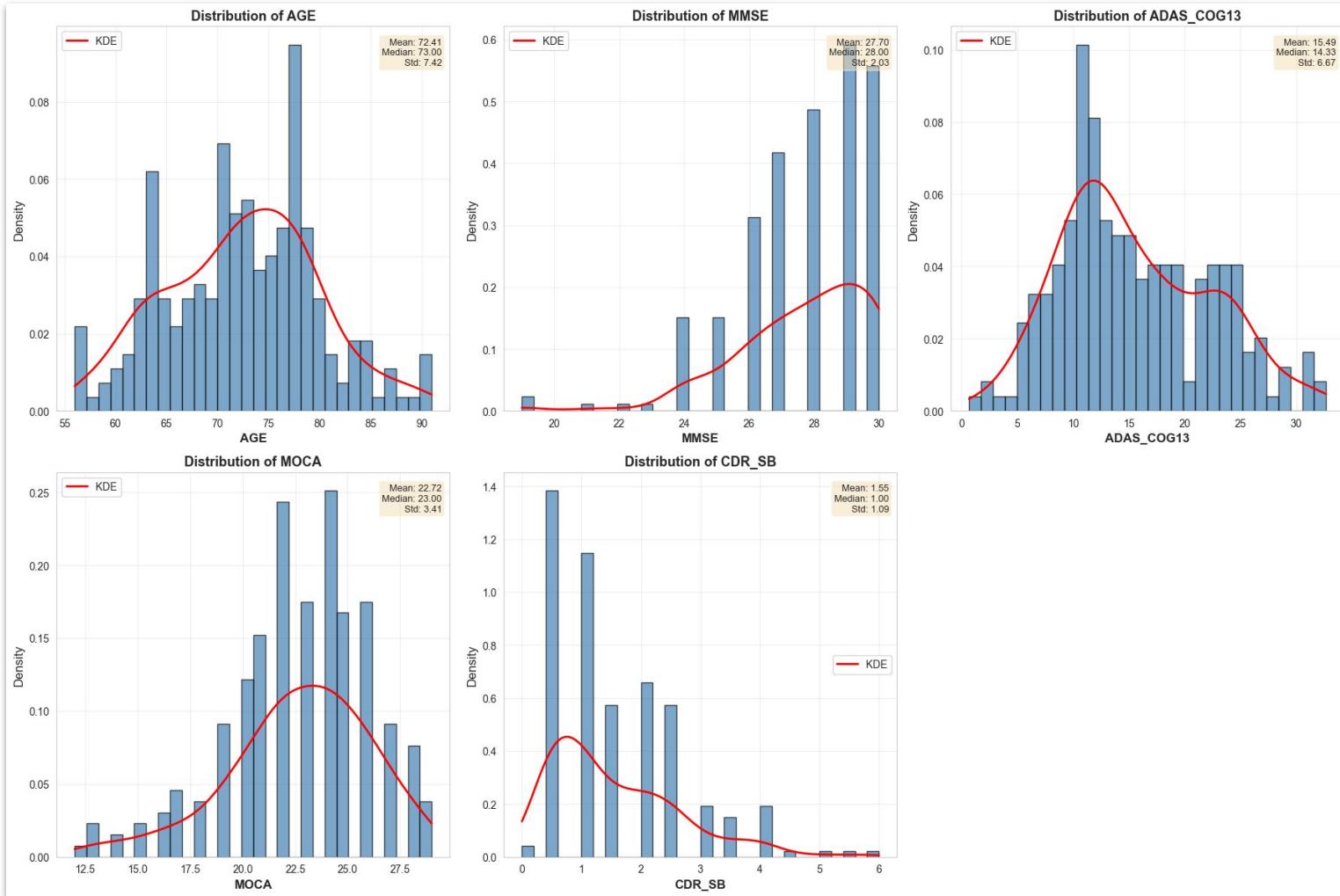
--- Neurological Exam (Auditory) ---

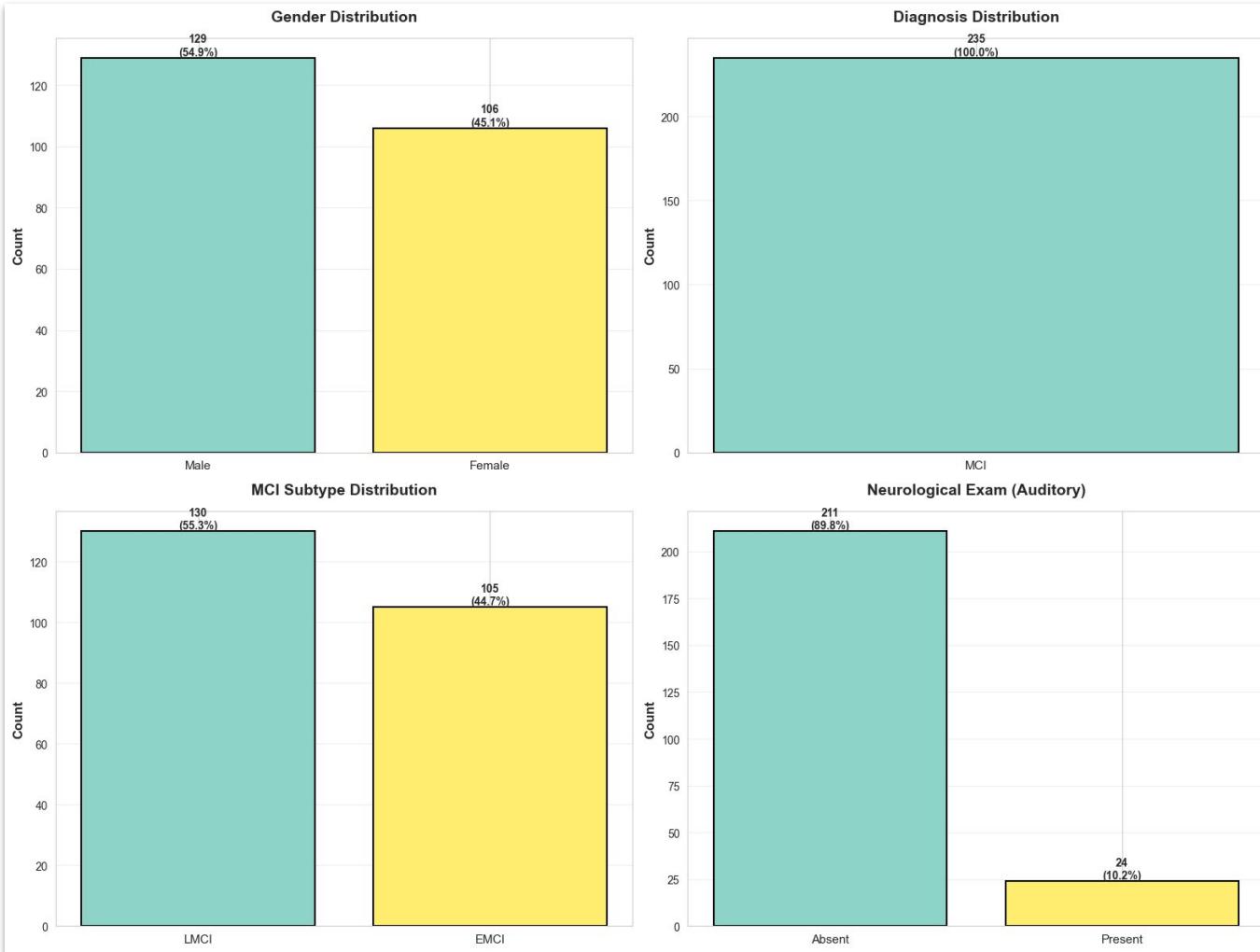
nxaudio_label

Absent 211

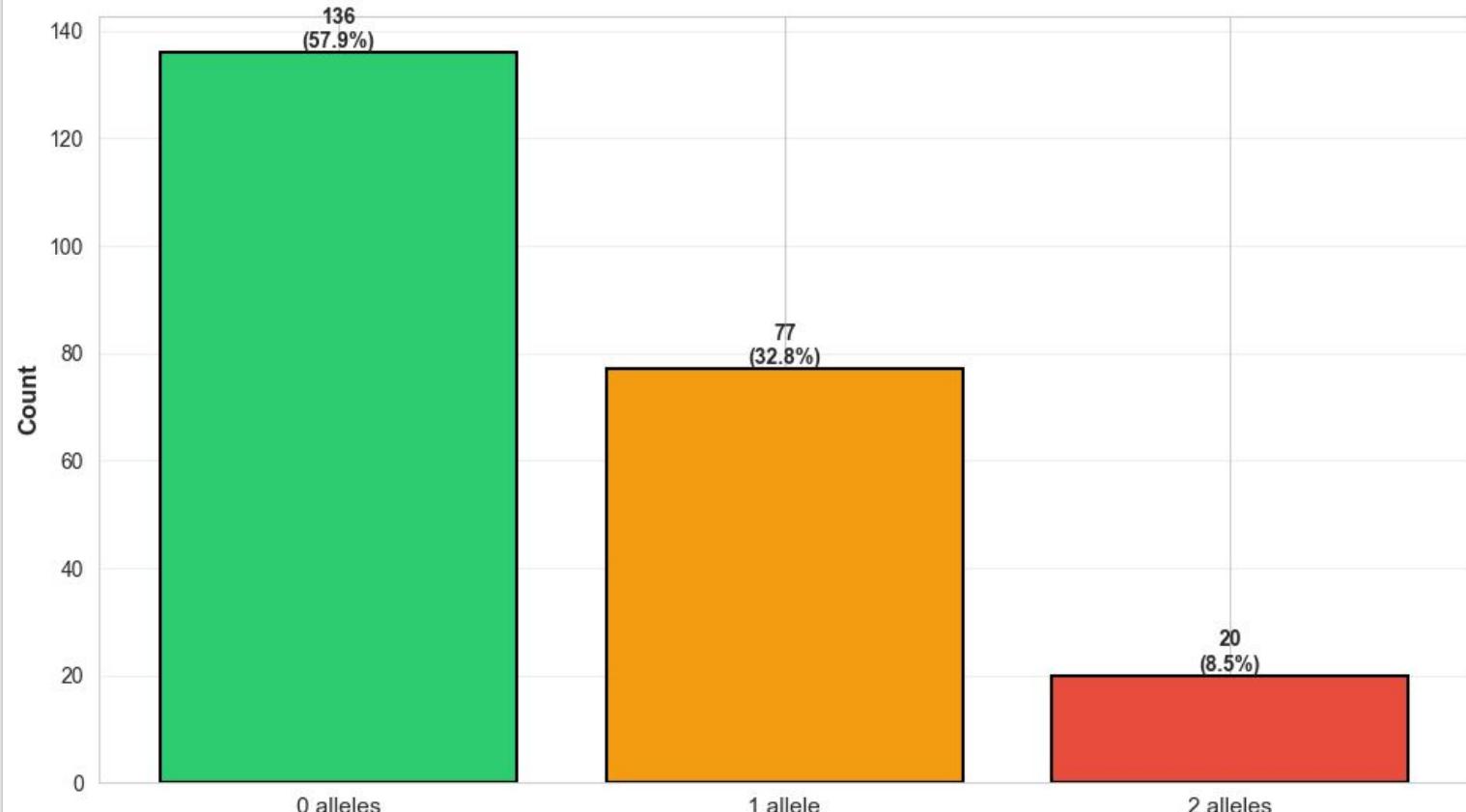
Present 24

Name: count, dtype: int64



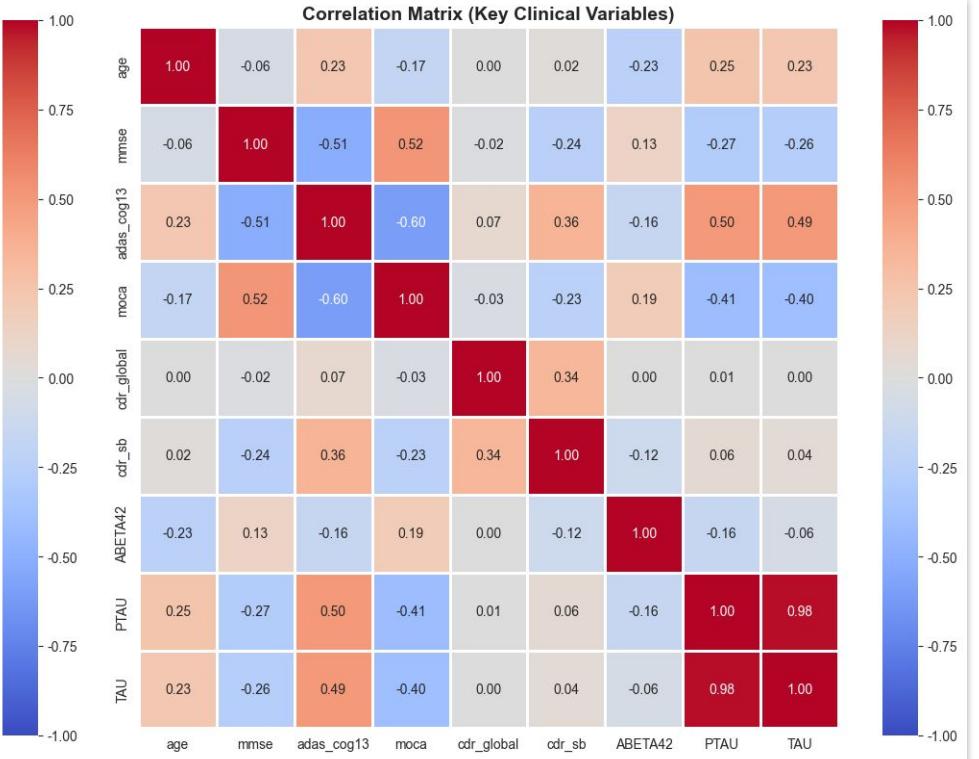
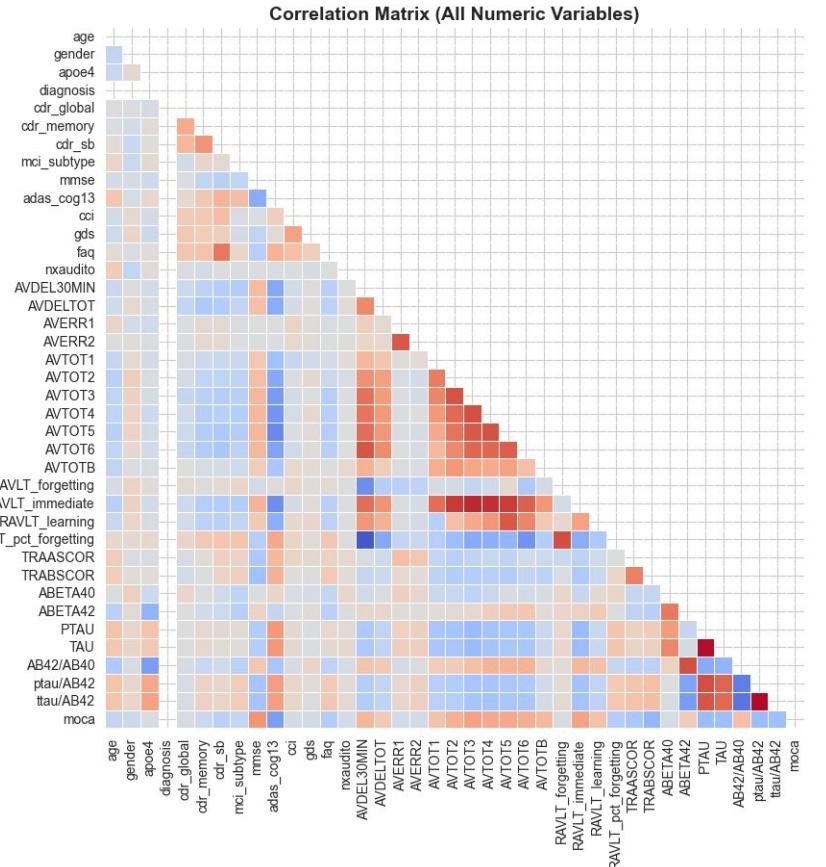


APOE4 Allele Distribution



Correlation Analysis

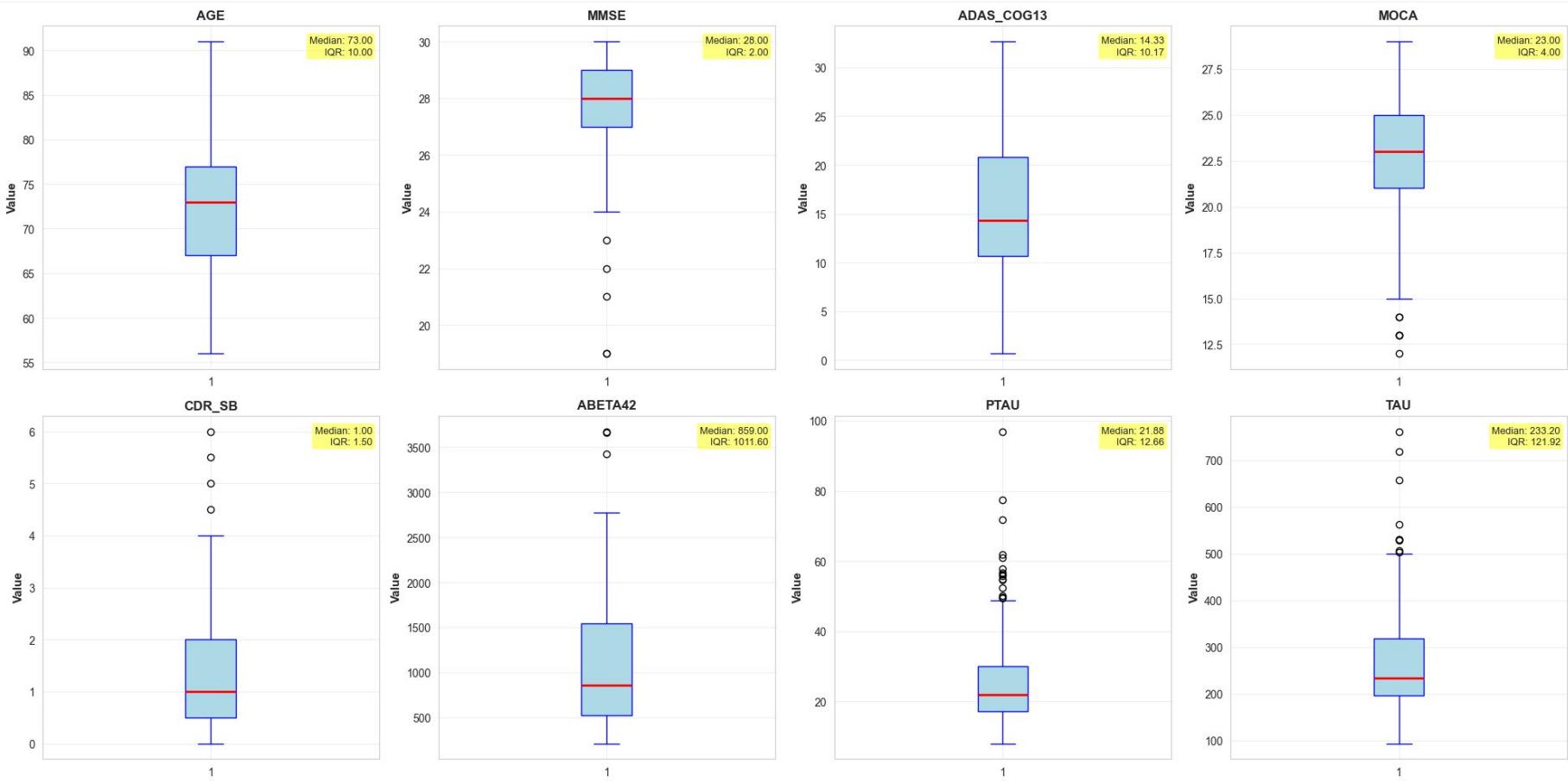
Variable 1	Variable 2	Correlation
ptau/AB42	ttau/AB42	0.989553
PTAU	TAU	0.979365
AVTOT3	RAVLT_immediate	0.931299
AVDEL30MIN	RAVLT_pct_forgetting	-0.931237
AVTOT4	RAVLT_immediate	0.907972
AVTOT2	RAVLT_immediate	0.877574
AVTOT5	RAVLT_immediate	0.874100
PTAU	ptau/AB42	0.816715
RAVLT_forgetting	RAVLT_pct_forgetting	0.815383
AVTOT3	AVTOT4	0.811913
ABETA42	AB42/AB40	0.809273
AVTOT4	AVTOT5	0.803674
AB42/AB40	ttau/AB42	-0.800342
AVTOT2	AVTOT3	0.795758
PTAU	ttau/AB42	0.784008
AVDEL30MIN	AVTOT6	0.783637
AVTOT5	RAVLT_learning	0.781177
AVERR1	AVERR2	0.774079
AVTOT3	AVTOT5	0.769993
AB42/AB40	ptau/AB42	-0.759153



Outlier Detection

--- 이상치 요약 (IQR Method) ---

Variable	Lower_Bound	Upper_Bound	N_Outliers	Outlier_Pct
age	52.0000	92.0000	0	0.000000
mmse	24.0000	32.0000	5	2.127660
adas_cog13	-4.5775	36.0825	0	0.000000
moca	15.0000	31.0000	6	2.586207
cdr_sb	-1.7500	4.2500	4	1.702128
ABETA42	-989.0000	3057.4000	3	2.255639
PTAU	-1.7250	48.9150	15	11.363636
TAU	13.2125	500.9125	9	6.818182



Comparative Analysis by Key Groups

--- 성별에 따른 인지검사 점수 ---

gender_label	mmse			adas_cog13			moca		
	mean	std	count	mean	std	count	mean	std	count
Female	27.603774	2.082486	106	15.131442	6.753662	104	22.221154	2.221154	104
Male	27.782946	1.988094	129	15.787402	6.621051	127	23.117188	2.117188	127

gender_label	cdr_global			cdr_sb			\\		
	std	count	mean	std	count	mean	std	count	\\
Female	3.649596	104	0.500000	0.097590	106	1.382075	0.916910	106	\\
Male	3.166312	128	0.503876	0.076448	129	1.693798	1.202715	129	\\

count

gender_label	count
Female	106
Male	129

mmse - T-test: t=-0.673, p=0.5017

adas_cog13 - T-test: t=-0.742, p=0.4586

moca - T-test: t=-2.001, p=0.0465 **

cdr_global - T-test: t=-0.341, p=0.7332

cdr_sb - T-test: t=-2.195, p=0.0292 **

--- APOE4 상태에 따른 인지검사 점수 ---

apoe4_carrier_label	mmse			adas_cog13			\\		
	mean	std	count	mean	std	count	mean	std	count
Carrier	27.371134	2.204728	97	16.645368	6.712127	95	27.371134	2.204728	95
Non-carrier	27.934783	1.868707	138	14.686471	6.552743	136	27.934783	1.868707	136

apoe4_carrier_label	moca			cdr_global			\\		
	mean	std	count	mean	std	count	mean	std	count
Carrier	22.406250	3.652189	96	0.494845	0.088236	97	22.406250	3.652189	97
Non-carrier	22.933824	3.229957	136	0.507246	0.085126	138	22.933824	3.229957	138

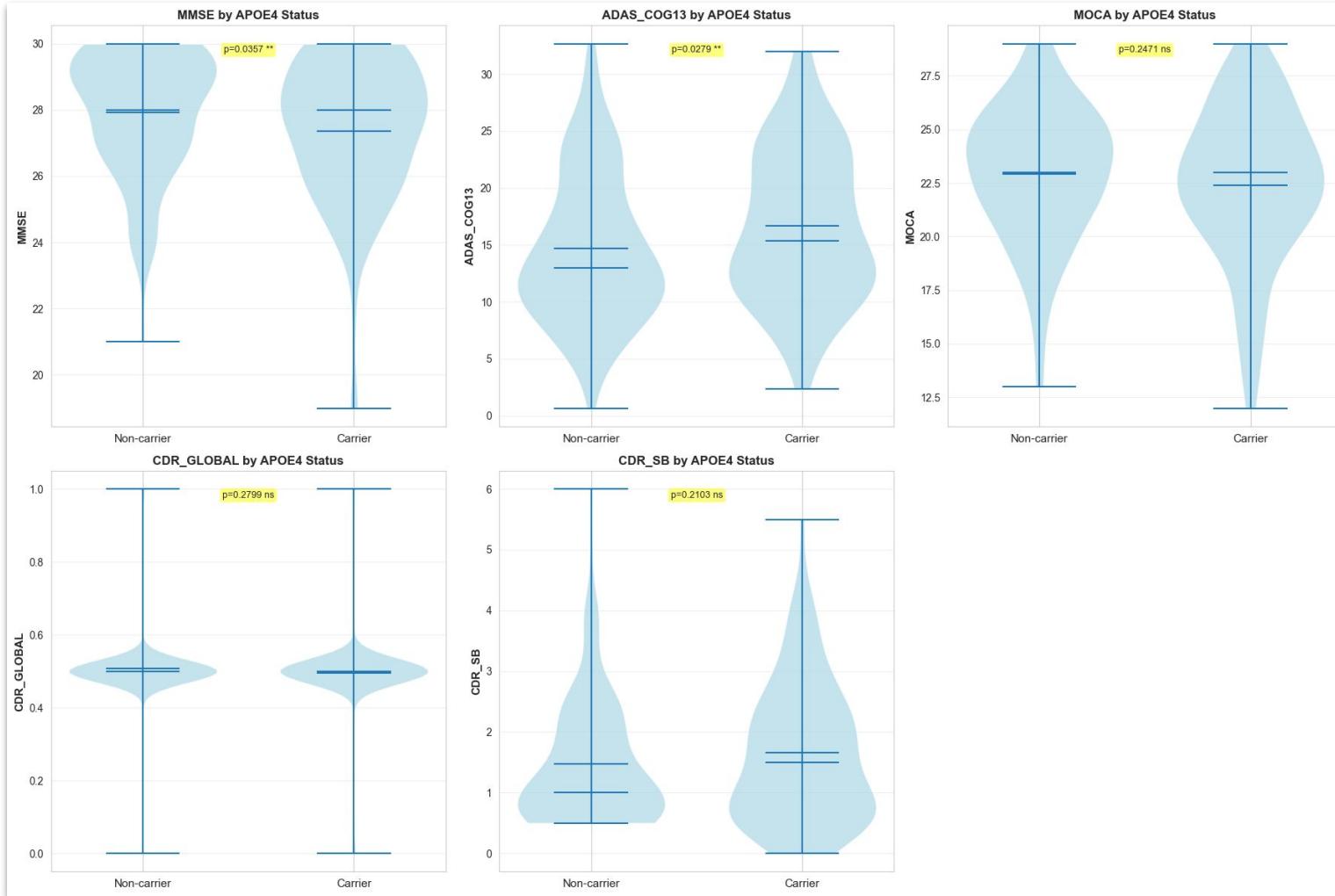
apoe4_carrier_label	cdr_sb			\\			\\		
	mean	std	count	mean	std	count	mean	std	
Carrier	1.659794	1.146885	97	1.659794	1.146885	97	1.659794	1.146885	97
Non-carrier	1.478261	1.049626	138	1.478261	1.049626	138	1.478261	1.049626	138

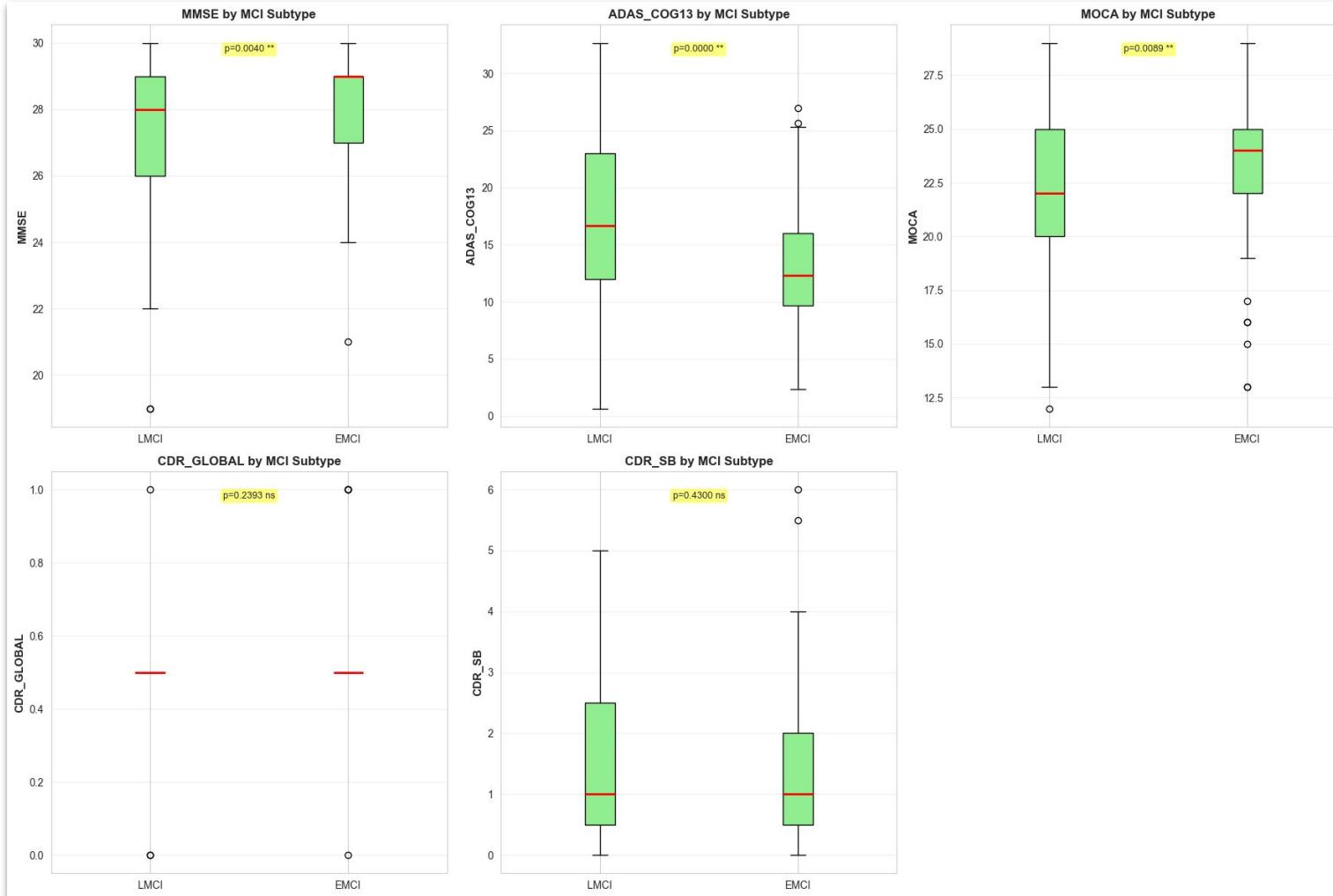
Comparative Analysis by Key Groups

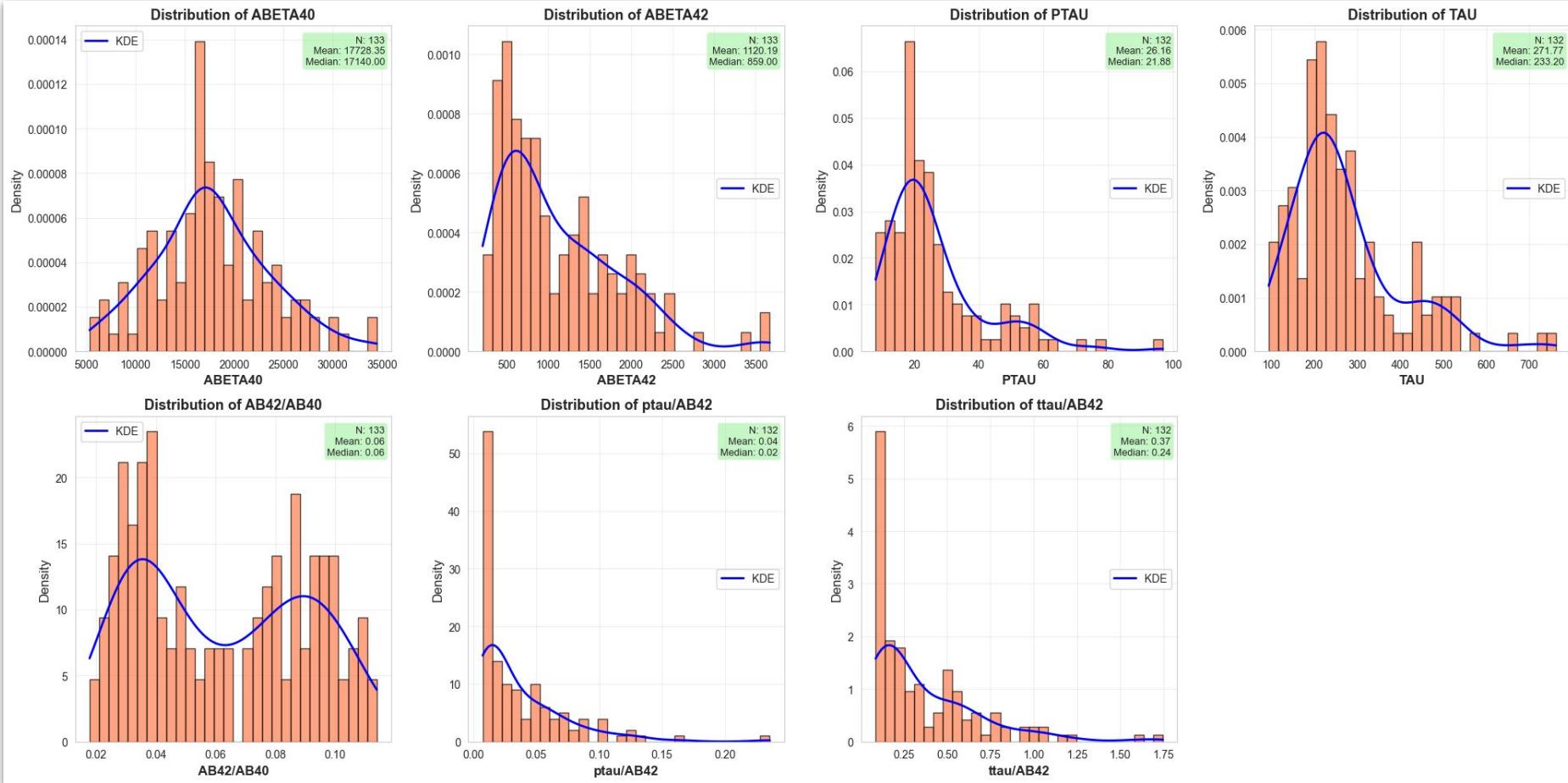
--- MCI Subtype에 따른 인지검사 점수 ---						
mci_subtype_label	mmse			adas_cog13		
	mean	std	count	mean	std	count
EMCI	28.123810	1.679602	105	13.336505	5.595113	103
LMCI	27.361538	2.220426	130	17.226641	6.981087	128

--- MCI Subtype에 따른 인지검사 점수 ---						
mci_subtype_label	moca			cdr_global		
	mean	std	count	mean	std	count
EMCI	23.368932	3.134221	103	0.509524	0.097590	105
LMCI	22.193798	3.546736	129	0.496154	0.076151	130

--- MCI Subtype에 따른 인지검사 점수 ---						
mci_subtype_label	cdr_sb			cdr_global		
	mean	std	count	mean	std	count
EMCI	1.490476	1.111524	105	0.509524	0.097590	105
LMCI	1.603846	1.077782	130	0.496154	0.076151	130







2.분류 모델 EDA

Summary

===== EMCI vs LMCI 분류 모델을 위한 EDA 요약 리포트 =====

1. 데이터 개요

- 총 샘플 수: 235
- EMCI: 105명 (44.7%)
- LMCI: 130명 (55.3%)
- 클래스 불균형 비율: 1.24:1

2. Feature 개요

- 총 Features: 37개
- 수치형: 37개
- 범주형: 3개

3. 결측치 현황

- 결측치 있는 Features: 30개
- Drop 권장 (>50%): 0개
- Impute 필요 (<=50%): 30개

4. 통계적 유의성

- 검정 Features: 40개
- 유의한 차이 ($p<0.05$): 20개

5. Feature Importance (Top 10)

adas_cog13	: 0.0579
TRABSCOR	: 0.0458
ABETA42	: 0.0438
TRAASCOR	: 0.0437
RAVLT_pct_forgetting	: 0.0433
cci	: 0.0382
AVDEL30MIN	: 0.0379
RAVLT_immediate	: 0.0374
ABETA40	: 0.0353
AVTOT6	: 0.0322

6. 다중공선성 (VIF)

- VIF ≥ 10: 10개 (심각)
- VIF 5-10: 2개 (주의)

7. 추천 Features

총 31개 (Importance > 0.01)

8. 전처리 권장사항

- 결측치: Median/Mode imputation
- 이상치: Winsorization (1-99 percentile)
- Scaling: StandardScaler
- Feature selection: Top 20 features

9. 모델링 권장

- 교차검증: Stratified K-Fold (k=5)
- 모델: Random Forest, XGBoost, SVM
- 평가지표: Accuracy, F1-score, ROC-AUC

```
=====
EMCI vs LMCI 분류 모델을 위한 EDA
=====
```

```
1. 데이터 로드
```

```
데이터 크기: 235 rows x 48 columns
```

```
Target 변수 분포:
```

```
target
```

```
LMCI    130
```

```
EMCI    105
```

```
Name: count, dtype: int64
```

```
클래스 비율:
```

```
target
```

```
LMCI    0.553191
```

```
EMCI    0.446809
```

```
Name: proportion, dtype: float64
```

```
=====
3. 클래스 불균형 분석
=====
```

```
EMCI: 105명 (44.7%)
```

```
LMCI: 130명 (55.3%)
```

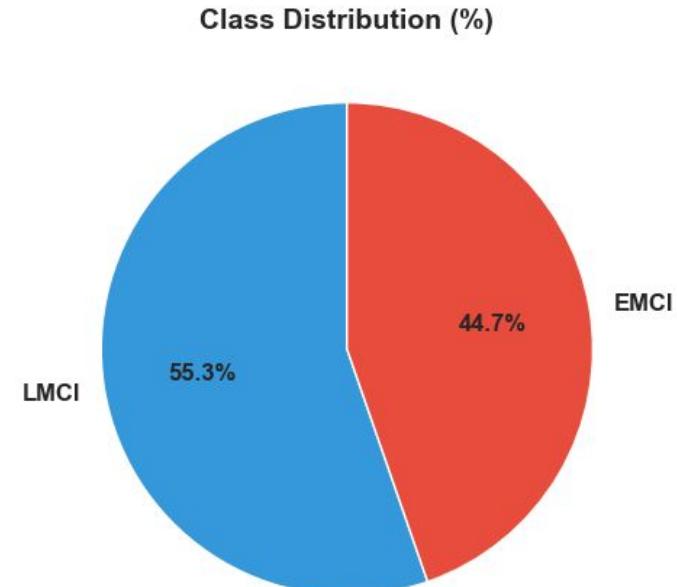
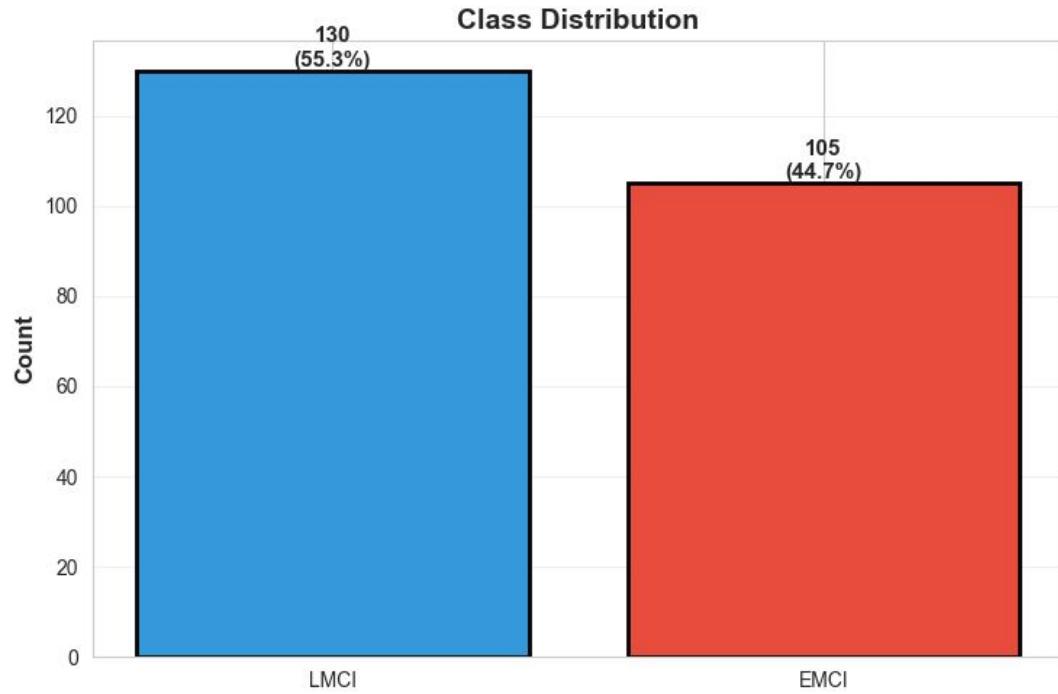
```
불균형 비율: 1.24:1
```

수치형 Features:

1. age
2. gender
3. apoe4
4. cdr_global
5. cdr_memory
6. cdr_sb
7. mmse
8. adas_cog13
9. cci
10. gds
11. faq
12. nxaudito
13. AVDEL30MIN
14. AVDELTOT
15. AVERR1
16. AVERR2
17. AVTOT1
18. AVTOT2
19. AVTOT3
20. AVTOT4
21. AVTOT5
22. AVTOT6
23. AVTOTB
24. RAVLT_forgetting
25. RAVLT_immediate
26. RAVLT_learning
27. RAVLT_pct_forgetting
28. TRAASCOR
29. TRABSCOR
30. ABETA40
31. ABETA42
32. PTAU
33. TAU
34. AB42/AB40
35. ptau/AB42
36. ttau/AB42
37. moca

범주형 Features:

1. gender
2. apoe4
3. nxaudito



=====
4. 결측치 분석 (모델링 관점)
=====

결측치가 있는 Features: 30개

결측치 요약:

	Feature	Missing_Count	Missing_Pct	Strategy
	ttau/AB42	103	43.829787	Impute
	ptau/AB42	103	43.829787	Impute
	TAU	103	43.829787	Impute
	PTAU	103	43.829787	Impute
	AB42/AB40	102	43.404255	Impute
	ABETA42	102	43.404255	Impute
	ABETA40	102	43.404255	Impute
	moca	11	4.680851	Impute
	faq	9	3.829787	Impute
	cci	5	2.127660	Impute
	adas_cog13	4	1.702128	Impute
	AVDELTOT	4	1.702128	Impute
	AVDEL30MIN	3	1.276596	Impute
	RAVLT_pct_forgetting	3	1.276596	Impute
	AVTOT6	3	1.276596	Impute
	TRAASCOR	3	1.276596	Impute
	RAVLT_forgetting	3	1.276596	Impute
	TRABSCOR	3	1.276596	Impute
	apoE4	2	0.851064	Impute
	AVERR2	2	0.851064	Impute
	AVERR1	2	0.851064	Impute
	AVTOT2	2	0.851064	Impute
	AVTOT1	2	0.851064	Impute
	AVTOT5	2	0.851064	Impute
	AVTOT3	2	0.851064	Impute
	AVTOT4	2	0.851064	Impute
	AVTOTB	2	0.851064	Impute
	RAVLT_learning	2	0.851064	Impute
	RAVLT_immediate	2	0.851064	Impute
	gds	1	0.425532	Impute

✓ 전처리 권장사항:

- Impute (<=50% 결측): 30개

권장: Median imputation 또는 KNN imputation

=====
5. EMCI vs LMCI 그룹 간 통계적 차이 검정
=====

수치형 변수 (Independent t-test):

범주형 변수 (Chi-square test):

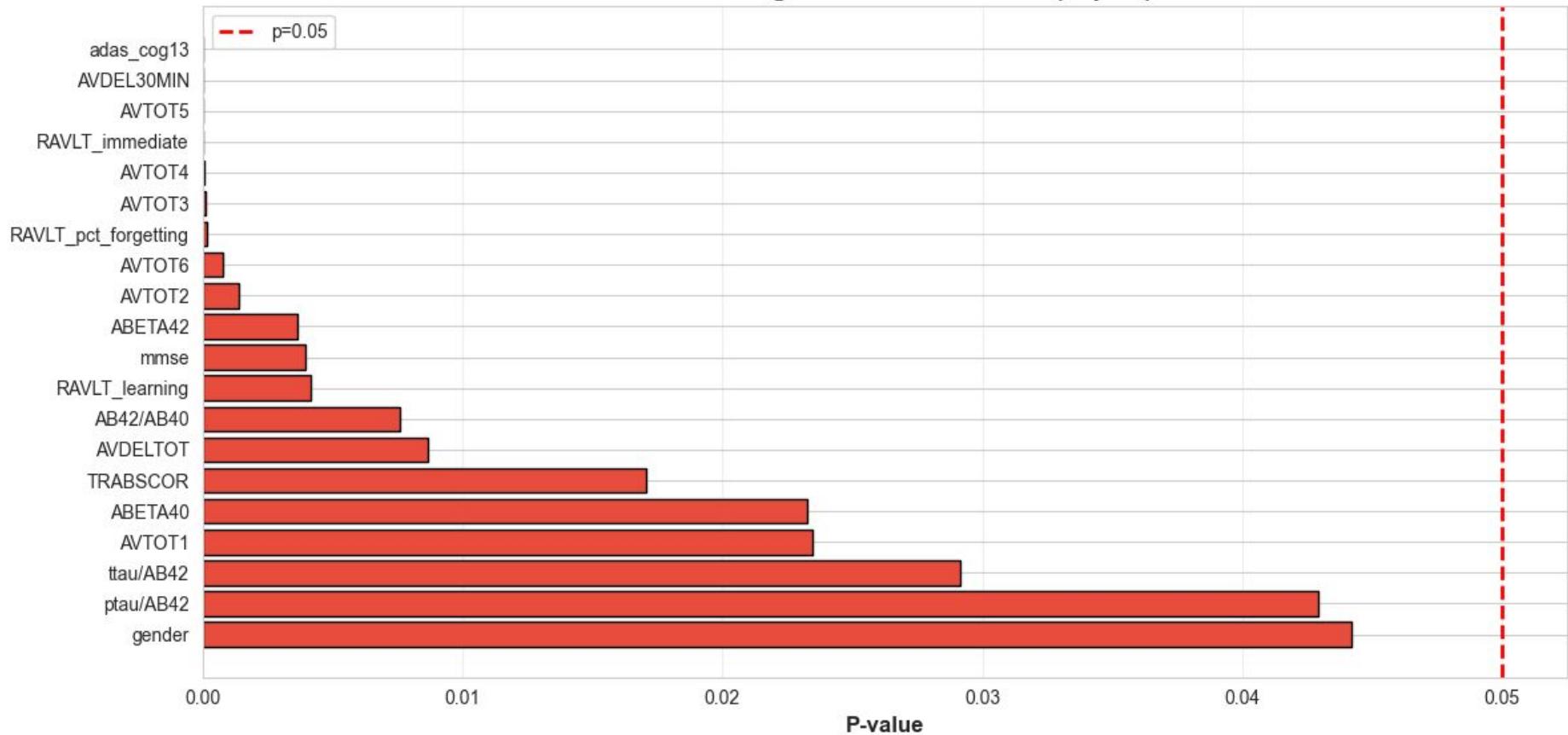
총 40개 변수 검정

유의한 변수 ($p < 0.05$): 20개

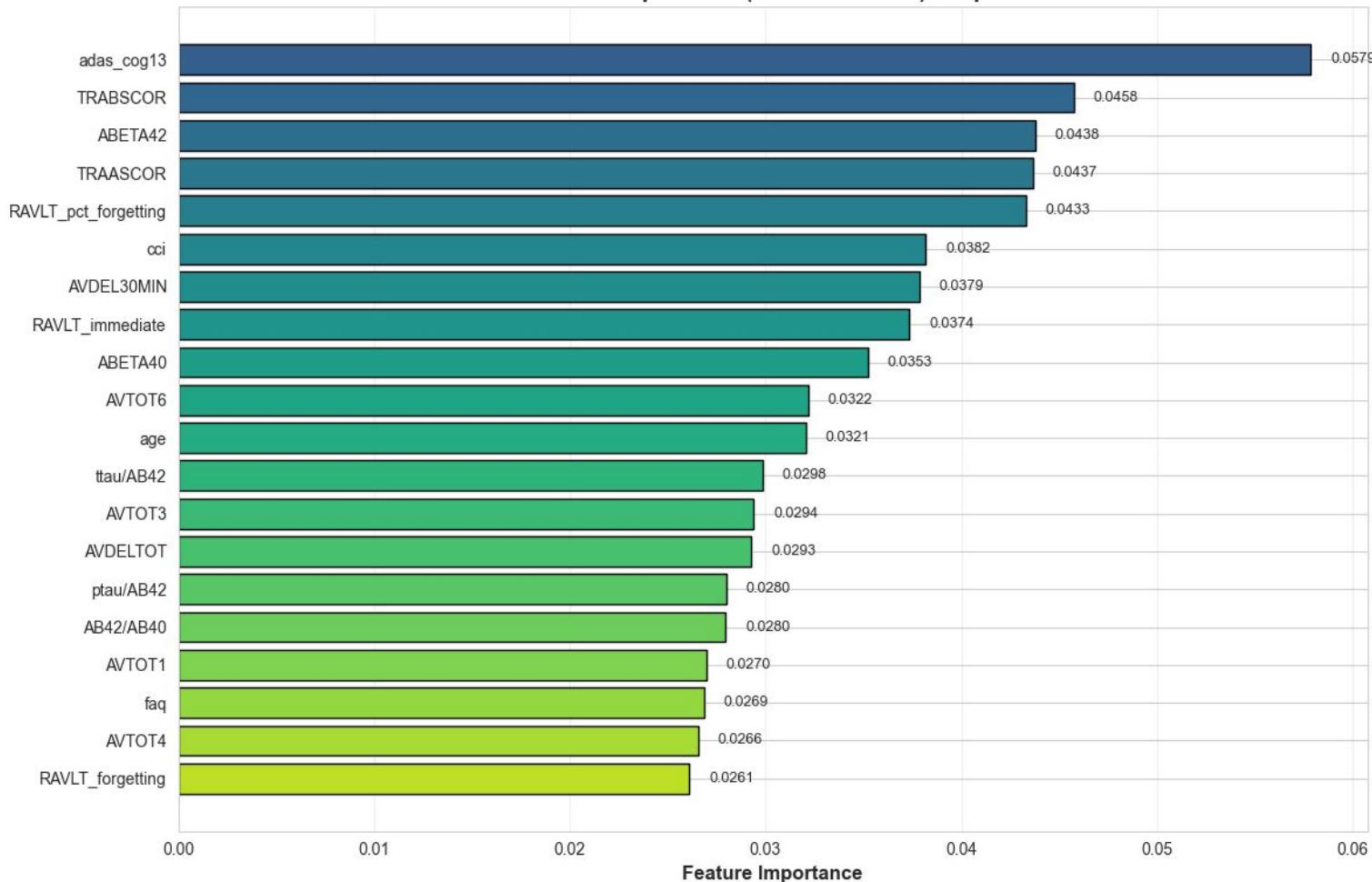
✓ 유의한 차이를 보이는 상위 변수:

	Feature	Type	P_value	Mean_Diff
	adas_cog13	Numeric	0.000007	3.890136
	AVDEL30MIN	Numeric	0.000021	1.993990
	AVTOT5	Numeric	0.000027	1.429613
	RAVLT_immediate	Numeric	0.000029	5.435714
	AVTOT4	Numeric	0.000049	1.346949
	AVTOT3	Numeric	0.000134	1.267857
	RAVLT_pct_forgetting	Numeric	0.000151	15.761374
	AVTOT6	Numeric	0.000803	1.538433
	AVTOT2	Numeric	0.001426	0.887426
	ABETA42	Numeric	0.003653	360.764719
	mmse	Numeric	0.003979	0.762271
	RAVLT_learning	Numeric	0.004159	0.925744
	AB42/AB40	Numeric	0.007605	0.012984
	AVDELTOT	Numeric	0.008680	1.215021
	TRABSCOR	Numeric	0.017082	19.265617

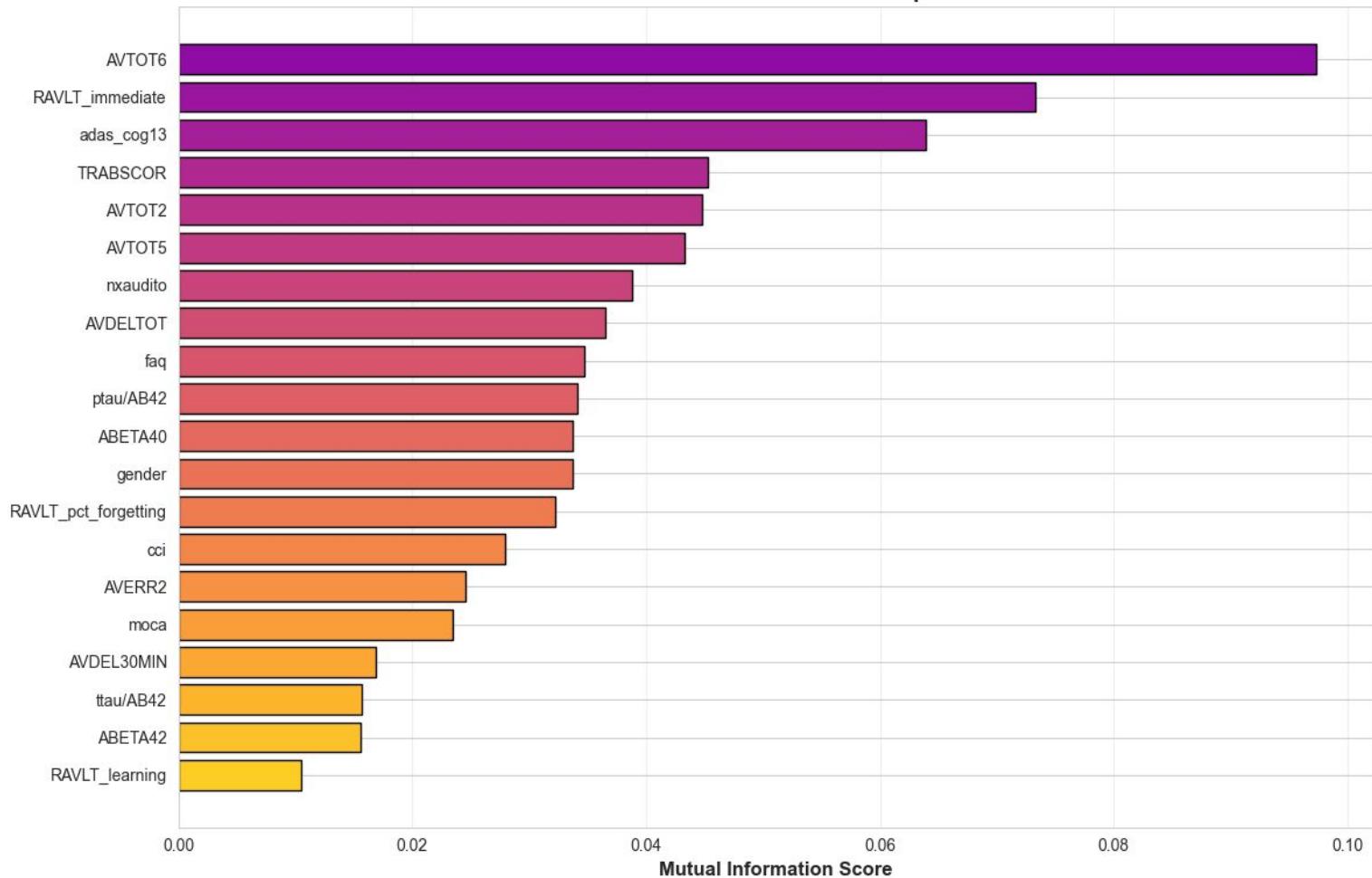
Statistical Significance of Features (Top 20)

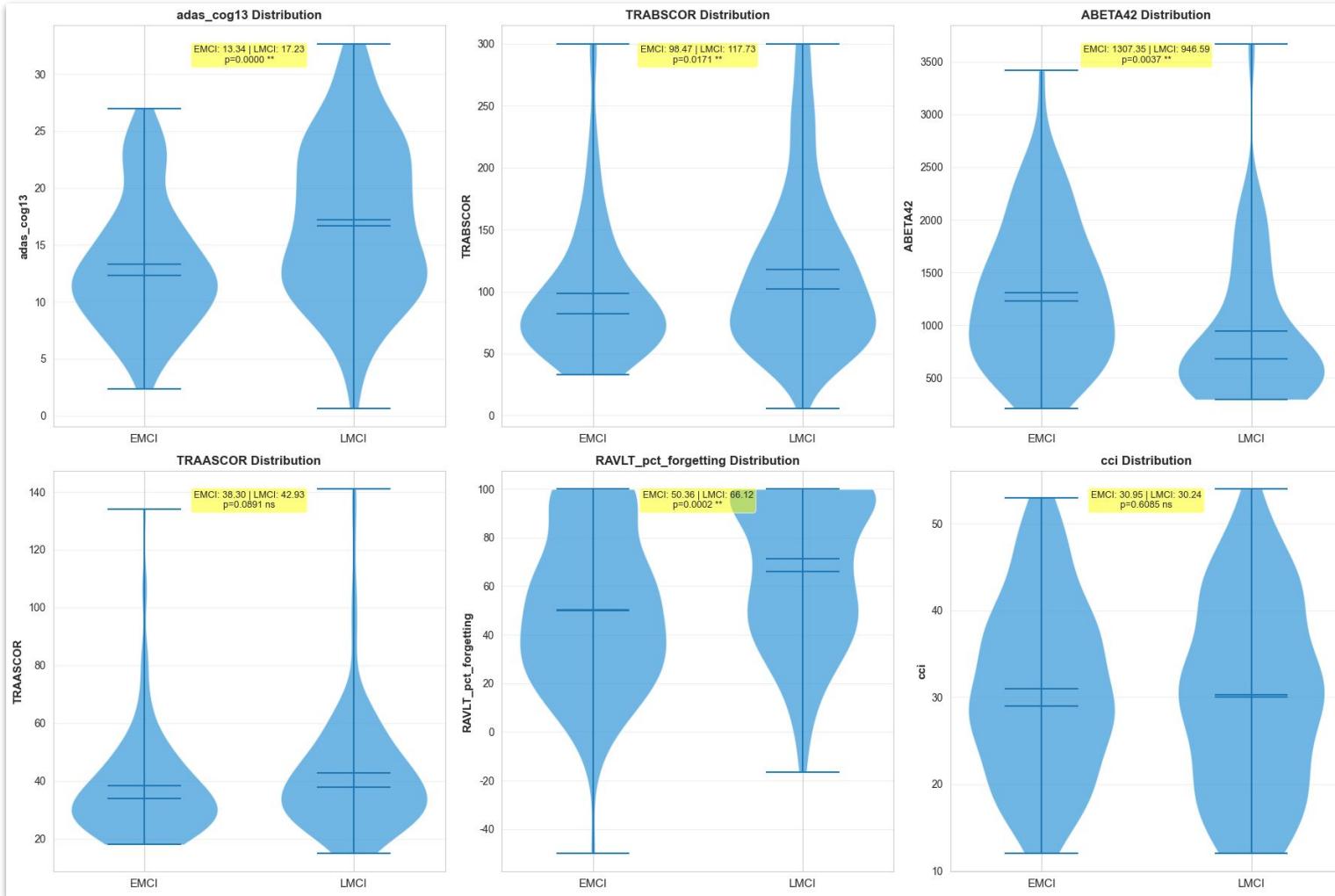


Feature Importance (Random Forest) - Top 20



Mutual Information Score - Top 20





=====
9. 다중공선성 분석 (VIF)
=====

✓ VIF 값 (Top 20 features):

	Feature	VIF
	RAVLT_immediate	78.757346
	ttau/AB42	75.767087
	ptau/AB42	65.850428
	AVDEL30MIN	27.716752
	ABETA42	16.074967
	RAVLT_pct_forgetting	15.188521
	AVTOT3	14.269643
	AB42/AB40	13.543834
	RAVLT_forgetting	12.613482
	AVTOT4	10.952830
	ABETA40	5.916569
	AVTOT1	5.521544
	AVTOT6	3.890484
	adas_cog13	2.775244
	AVDELTOT	1.813299
	TRAASCOR	1.801812
	TRABSCOR	1.783924
	faq	1.311470
	cci	1.190118
	age	1.171947

다중공선성 해석:

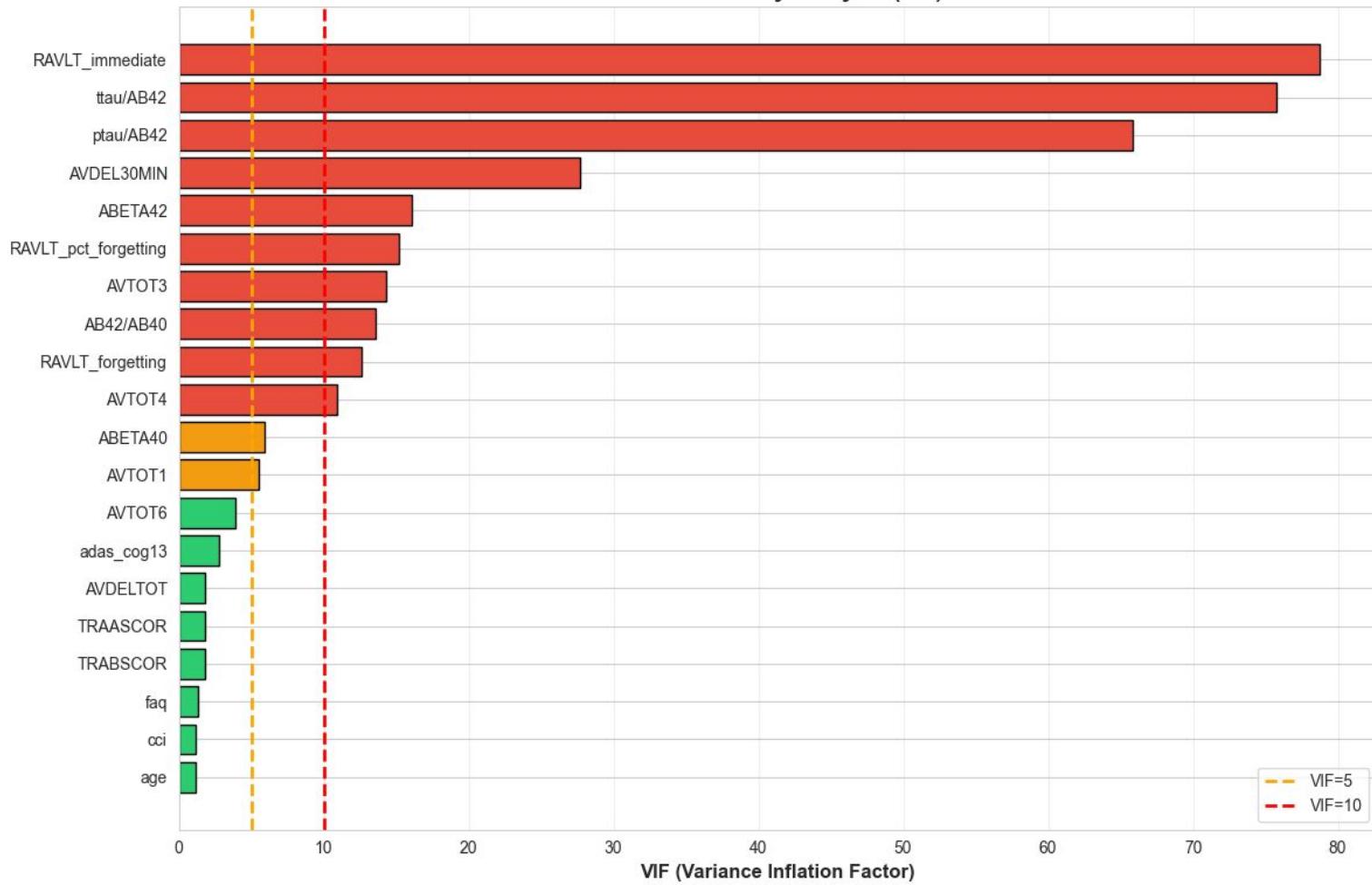
- VIF < 5: 문제 없음
- 5 ≤ VIF < 10: 주의 필요
- VIF ≥ 10: 심각한 다중공선성

▲ VIF ≥ 10인 변수: 10개

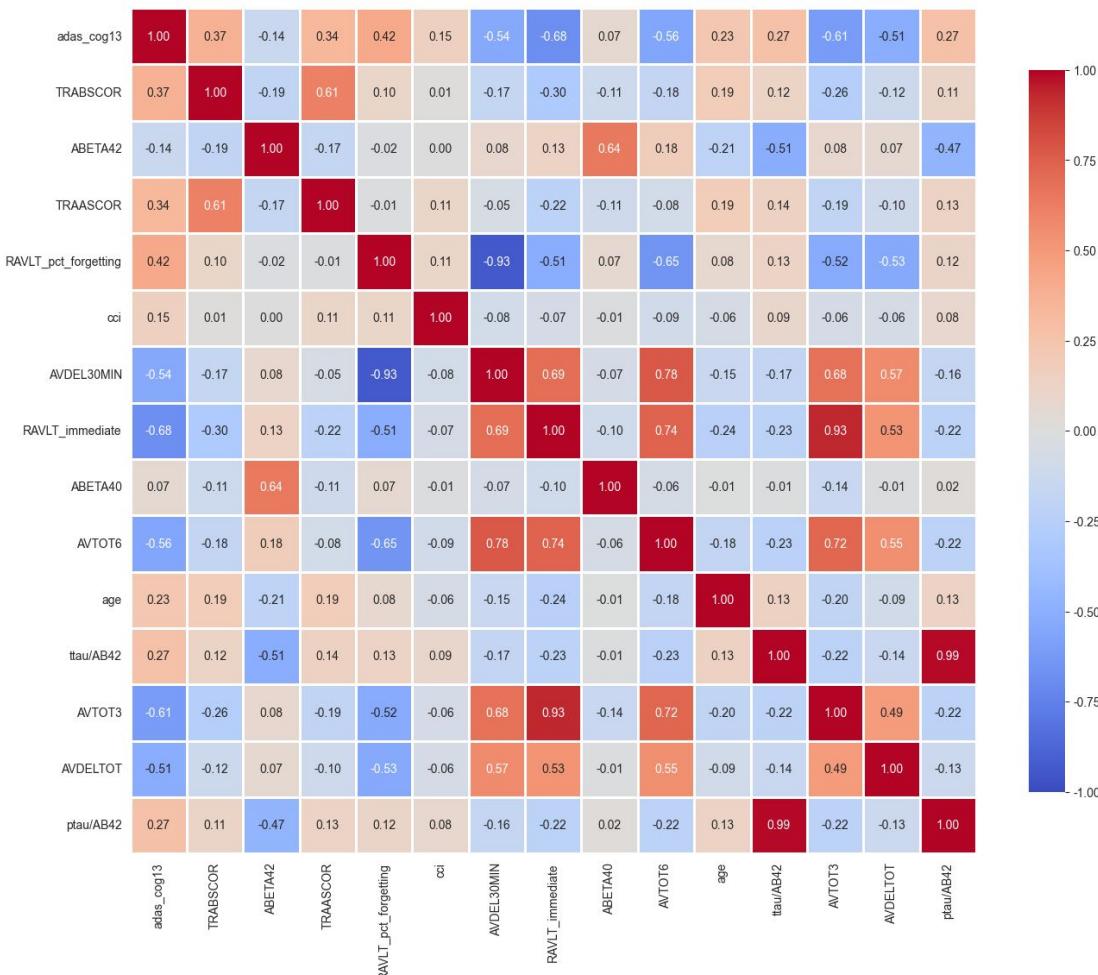
다음 변수 제거 고려:

- RAVLT_immediate
- ttau/AB42
- ptau/AB42
- AVDEL30MIN
- ABETA42
- RAVLT_pct_forgetting
- AVTOT3
- AB42/AB40
- RAVLT_forgetting
- AVTOT4

Multicollinearity Analysis (VIF)



Correlation Matrix (Top 15 Features)



⚠️ 높은 상관관계 ($|r| > 0.7$): 6쌍

Feature_1	Feature_2	Correlation
ttau/AB42	ptau/AB42	0.990169
RAVLT_immediate	AVTOT3	0.931262
RAVLT_pct_forgetting	AVDEL30MIN	-0.931221
AVDEL30MIN	AVTOT6	0.779376
RAVLT_immediate	AVTOT6	0.740825
AVTOT6	AVTOT3	0.719260

권장: 각 쌍에서 하나씩 제거 고려

===== 11. 이상치 분석 (상위 Features) =====

✓ 이상치 요약:

	Feature	N_Outliers	Outlier_Pct	Lower_Bound	Upper_Bound
	adas_cog13	0	0.000000	-4.577500	36.082500
	TRABSCOR	16	6.896552	-33.000000	231.000000
	ABETA42	3	2.255639	-989.000000	3057.400000
	TRAASCOR	13	5.603448	-0.500000	75.500000
	RAVLT_pct_forgetting	1	0.431034	-43.973214	166.939484
	cc1	0	0.000000	-2.875000	64.125000

이상치 처리 옵션:

1. 제거 (Removal)
2. Winsorization (상/하위 1-5%)
3. 변환 (Log, Box-Cox)
4. 유지 (도메인 지식 기반)

=====

12. 모델링 권장사항

=====

✓ 추천 Features: 31개

상위 20개:

1. adas_cog13 (Importance: 0.0579)
2. TRABSCOR (Importance: 0.0458)
3. ABETA42 (Importance: 0.0438)
4. TRAASCOR (Importance: 0.0437)
5. RAVLT_pct_forgetting (Importance: 0.0433)
6. cci (Importance: 0.0382)
7. AVDEL30MIN (Importance: 0.0379)
8. RAVLT_immediate (Importance: 0.0374)
9. ABETA40 (Importance: 0.0353)
10. AVTOT6 (Importance: 0.0322)
11. age (Importance: 0.0321)
12. ttau/AB42 (Importance: 0.0298)
13. AVTOT3 (Importance: 0.0294)
14. AVDELTOT (Importance: 0.0293)
15. ptau/AB42 (Importance: 0.0280)
16. AB42/AB40 (Importance: 0.0280)
17. AVTOT1 (Importance: 0.0270)
18. faq (Importance: 0.0269)
19. AVTOT4 (Importance: 0.0266)
20. RAVLT_forgetting (Importance: 0.0261)

✓ 전처리 파이프라인 권장:

1. 결측치 처리:

- Drop: 0개 features (>50% 결측)
- Impute: 30개 features (\leq 50% 결측)

2. 이상치 처리: Winsorization (1-99 percentile)

3. Feature scaling: StandardScaler

4. 다중공선성 처리:

- 6쌍의 높은 상관관계 확인
- 각 쌍에서 덜 중요한 feature 제거

✓ 모델 선택 권장:

- Random Forest
- XGBoost
- SVM
- Logistic Regression

✓ 교차검증 전략:

- Stratified K-Fold (k=5 or 10)
- 클래스 비율 유지 필수

✓ 평가 지표:

- Accuracy
- Precision, Recall, F1-score
- ROC-AUC
- Confusion Matrix