**EEG 기반 알코올 중독 분류 결과 보고서**

**최도한, 전성현**

**요약**

본 팀프로젝트는 EEG 데이터를 통해 알코올 중독 여부를 정확히 분류하는 모델을 개발하기 위한 다양한 데이터 전처리 및 모델링 기법을 적용하였습니다. 목표는 EEG 신호의 특성을 최대한 활용하여 머신러닝 모델의 성능을 최적화하는 것이었습니다. 이를 위해 필터링, 기본선 보정, 데이터 세분화, 윈도우 함수 적용, 퓨리에 변환, 스케일링, 그리고 앙상블 기법을 포함하는 다양한 데이터 전처리 절차를 수행하였습니다. 각 단계는 EEG 데이터의 복잡성을 관리하고, 노이즈를 최소화하며, 신호의 중요 특성을 강조하는 데 중점을 두었습니다.

**1. 데이터**

팀프로젝트에서 사용된 EEG 데이터는 두피에 부착된 전극을 통해 측정된 뇌의 전기적 활동을 기록한 데이터로, 비침습적인 방법으로 뇌의 생리적 및 병리적 상태를 반영합니다. 이 데이터를 통해 뇌파의 패턴을 분석하고, 알코올 중독여부를 판단하는데 활용할 수 있습니다.

**2. 데이터 전처리**

데이터 전처리는 머신러닝 모델의 성능에 직접적인 영향을 미치는 중요한 단계입니다. EEG 데이터는 특히 노이즈가 많고, 신호 처리가 까다로운 특성을 지니고 있기 때문에, 효과적인 전처리가 필수적입니다. 다음은 저희 팀이 적용한 주요 전처리 단계입니다.

**2.1 필터링**

EEG 데이터는 자주 노이즈에 영향을 받기 때문에, 적절한 필터링 과정이 필수적입니다. 일반적으로 EEG 데이터에는 50Hz 또는 60Hz의 전력선 노이즈, 근전도(EMG) 노이즈 등이 포함될 수 있습니다. 이를 제거하기 위해 저주파 통과 필터(Low-pass filter)를 사용하여 고주파 노이즈를 제거하고, 고주파 통과 필터(High-pass filter)를 사용하여 저주파 드리프트를 제거합니다. 또한, 특정 주파수 대역(예: 알파파 8-12Hz, 베타파 12-30Hz)을 분리하여 분석하는 대역통과 필터(Band-pass filter)를 적용하였습니다.

**2.2 아티팩트 제거**

EEG 신호는 눈 깜박임, 눈 움직임, 심장 박동 등 인체의 다른 생리적 활동으로부터 아티팩트(예상치 못한 변동)가 발생할 수 있습니다. 이러한 아티팩트는 데이터의 해석을 왜곡할 수 있으므로, 독립 성분 분석(ICA)과 같은 기법을 사용하여 이를 식별하고 제거합니다. 하지만 ICA 적용후 데이터에 변화를 식별하지 못해 제공된 EEG 데이터는 아티팩트가 없다고 판단했습니다.

**2.3 재참조화**

EEG 데이터의 재참조화는 뇌파 신호의 기준을 일관되게 설정하여 데이터 간 비교를 용이하게 합니다. 공통 평균 참조(Common Average Referencing, CAR) 방법이나 특정 전극을 기준으로 하는 참조 방법이 사용됩니다.

**2.4 정규화**

데이터의 스케일링 차이를 보정하기 위해 정규화 과정을 거칩니다. 이는 모든 데이터가 비슷한 범위의 값을 갖도록 조정함으로써 모델의 학습 효율을 높이고, 다른 뇌파 데이터 간의 비교가 가능하도록 합니다.

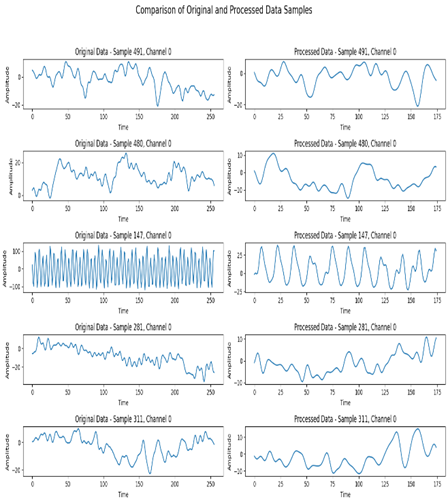
**2.5 기본선 제거 및 세그먼트화**

기본선 제거는 신호에서 일반적인 노이즈나 오프셋을 제거하는 과정입니다. 이후 신호를 더 명확하게 분석하기 위해 수행했습니다. EEG 데이터는 연속적인 신호로 이루어져 있기 때문에 분석을 위해 특정 시간 간격으로 데이터를 나누는 세그먼트화 작업이 필요합니다. 이러한 전처리 과정을 거친 후에, EEG 데이터는 특징 추출 과정을 통해 학습 가능한 형태로 변환됩니다. 이는 모델링 과정에서 입력 데이터로 사용됩니다.

**2.6 FFT 변환과 윈도우 함수 적용**

시간 영역의 데이터를 주파수 영역으로 변환하여 주요 주파수 구성 요소의 파워 스펙트럼을 분석하였습니다. 이 과정에서 파워스펙트럼 평균값을 기준으로 임계값 처리를 수행해 주요 신호만을 추출하여 노이즈의 영향을 최소화하였습니다. 추가적으로 한넬 윈도우 함수를 적용하여 주파수 해상도를 최적화하였습니다.

**2.7 데이터 전처리 결과**



노이즈 감소: 고주파 및 저주파 필터링을 통해 불필요한 신호가 제거되어, 분석에 필요한 신호만을 남겼습니다.

신호의 명확성 향상: 기본선 보정 및 세분화 과정을 통해 신호의 주요 특성이 강조되며, 이로 인해 데이터 분석의 정확성이 높아졌습니다.

분석 용이성 증대: 윈도우 함수 적용 및 퓨리에 변환을 통해 시간 영역 데이터가 주파수 영역으로 변환되면서, 데이터의 스펙트럼 분석이 용이해졌습니다.

**3. 머신러닝 모델 선택 및 모델링**

머신러닝 모델로 랜덤포레스트(RF), 서포트 벡터 머신(SVM), 그리고 XGBoost 모델을 선택하여 앙상블하는 방법을 선택하였습니다.

**3.1 모델 학습**

세 가지 모델을 앙상블 기법 사용하여 모델 학습을 진행하였습니다. 각 모델의 성능을 개별적으로 평가하고, 소프트 보팅을 통해 여러 모델의 예측 결과를 결합하여 최종 예측의 정확도를 높였습니다.

**3.2 시행착오**

한넬, 해밍, 블랙맨, 카이저 윈도우: 각각의 윈도우 함수를 실험적으로 적용하여 스펙트럼 누출을 줄이고, 데이터의 주파수 해상도를 최적화하였습니다. 이러한 접근은 스펙트럼 분석의 정확도를 높이는 데 기여하였습니다**.** 최종적으로 한넬 위도우 함수를 선택하였습니다.

스태킹 앙상블 기법을 적용하였으나 오히려 예측 성능이 감소하였습니다.

**4. 하이퍼파라미터 최적화**

하이퍼파라미터 탐색에는 Optuna 라이브러리를 활용하였습니다. 하이퍼파라미터 탐색은 머신러닝 모델의 성능을 최적화하는 데 중요한 역할을 합니다. 하이퍼파라미터 최적화를 위한 효율적인 도구를 사용하여 EEG 기반 알코올 중독 분류 모델의 성능을 극대화하였습니다. Optuna는 베이지안 최적화 기법을 기반으로 하는 자동화된 하이퍼파라미터 최적화 프레임워크로, 광범위한 실험과 복잡한 조건들을 효율적으로 관리할 수 있는 기능을 제공합니다.

각 모델의 하이퍼파라미터 설정과 최적화 결과는 다음과 같습니다.

**4.1 하이퍼파라미터 설정 및 탐색 범위**

랜덤 포레스트(Random Forest)

- n\_estimators: 트리의 개수 (100 ~ 200)

- max\_depth: 트리의 최대 깊이 (3 ~ 10)

- min\_samples\_split: 노드를 분할하기 위한 최소 샘플 수 (2 ~ 20)

- min\_samples\_leaf: 리프 노드가 가져야 할 최소 샘플 수 (1 ~ 10)

SVM (Support Vector Machine)

- C: 오류 항의 페널티 (0.1 ~ 10, 로그 스케일)

- gamma: 커널의 계수 ('scale', 'auto')

XGBoost

- n\_estimators: 부스팅 라운드 수 (50 ~ 200)

- max\_depth: 각 트리의 최대 깊이 (3 ~ 10)

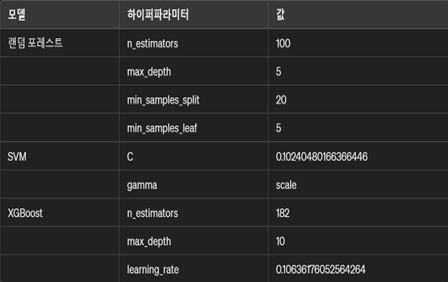
- learning\_rate: 학습률 (0.01 ~ 0.5)

**4.2 최적화 방법**

Optuna의 `study.optimize` 함수를 사용하여 위에서 정의한 범위 내에서 하이퍼파라미터를 탐색했습니다. 탐색은 총 30회의 시도(trials)를 통해 수행되었으며, 각 시도는 5-fold 교차 검증을 통해 평가되었습니다. 평가 기준은 ` accuracy` 스코어였습니다.

**4.3 최적의 하이퍼파라미터 결과**

최적화 과정을 통해 얻은 최적의 하이퍼파라미터는 다음과 같습니다:



**5. 모델 성능 검증**

모델의 학습 성능 평가를 위해 sklearn.metrics의 accuracy\_score, recall\_score, precison\_score 라이브러리를 사용하였습니다.

또한 데이터 불균형을 측정하기 위해 f1\_score를 사용하였습니다.

**5.1 Ensemble 구성 모델 별 Confusion Metrics 및 ROC Curve**

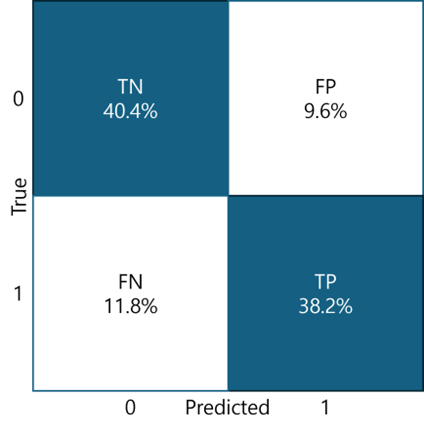
랜덤 포레스트(Random Forest)

- accuracy : 0.7861111111111111

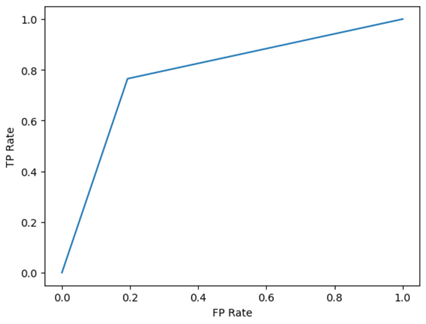
- recall : 0.7648148148148148

- precision : 0.7988394584139265

- f1 score : 0.7814569536423842



<랜덤 포레스트 Confusion Metrics>



<랜덤 포레스트 ROC Curve>

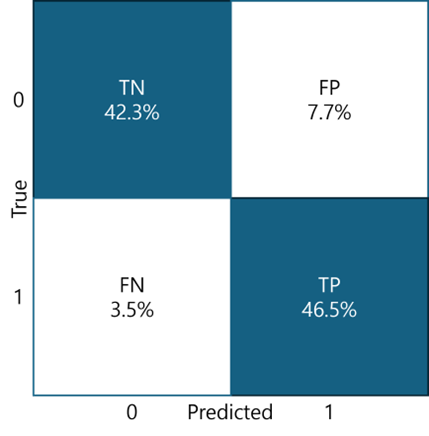
SVM (Support Vector Machine)

- accuracy : 0.887962962962963

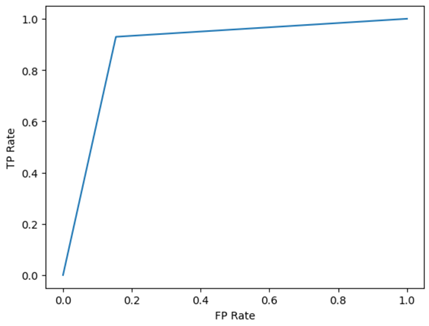
- recall : 0.9296296296296296

- precision : 0.8581196581196581

- f1 score : 0.8924444444444444



<SVM Confusion Metrics>



<SVM ROC Curve>

XGBoost

- accuracy : 1.0

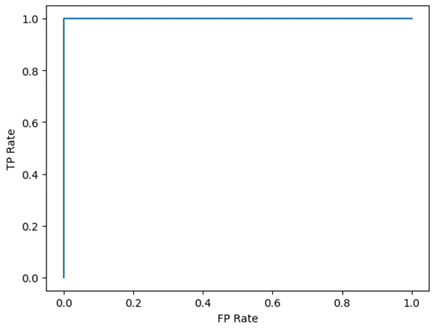
- recall : 1.0

- precision : 1.0

- f1 score : 1.0



<XGBoost Confusion Metrics>



<XGBoost ROC Curve>

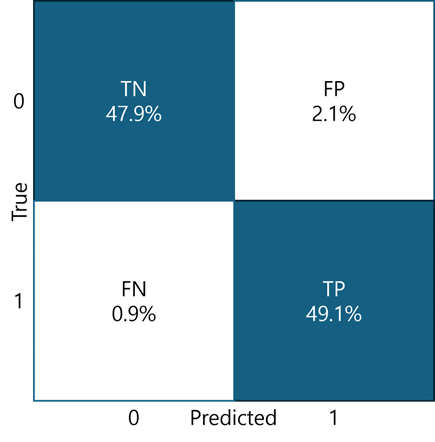
**5.2 Ensemble 모델의 Confusion Metrics 및 ROC Curve**

- accuracy : 0.9694444444444444

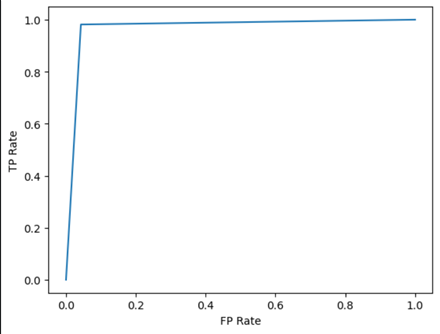
- recall : 0.9814814814814815

- precision : 0.9584086799276673

- f1 score : 0.969807868252516



<Ensemble Confusion Metrics>



<Ensemble ROC Curve>

**참고문헌**

EEG object recognition: Studies for criminal investigation and neuro-applications in social care - Constantin 저

Investigating the role of data preprocessing, hyperparameters tuning, and type of machine learning algorithm in the improvement of drowsy EEG signal modeling - Farbod Farhangi 저 (2022)

**역할분담**

최도한 :

* 데이터 전처리
* 머신러닝 모델링
* 보고서 작성

전성현 :

* 모델별 학습률 측정
* 모델별 학습률 시각화
* 발표자료 제작
* 보고서 작성