|  |
| --- |
| **보고서** |
| 머신러닝(1) 팀 프로젝트  (810007/21001) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 제출일 | 2024.05.19 |  | 학과 | 인공지능응용학과 |
| 과목 | 머신러닝 (1) |  | 학번 | 21102345, 21OOOOOO |
| 담당교수 | 김OO 교수 |  | 이름 | 권도윤, 하OO |
| 권도윤 | -데이터 전처리  -모델 제작  -데이터 작성  -학습과정 작성  -개선안 및 결론 작성  -발표 |  | 하OO | -데이터 전처리  -모델 제작  -서론 작성  -성능 평가 지표 및 성과 작성  -개선안 및 결론 작성  -발표 자료 제작 |

**|목차|**

**Ⅰ. 서론……………………………………………………………………..……………2**

**Ⅱ. 데이터….………………………………………..…………………….…………..2**

**Ⅲ. 학습과정……………………………………………….………………………….3**

1. XGBoost……………………………………………………………………………………………………………….…………….3
2. 하이퍼 파라미터 최적화…………………………………………………………………………………………………..3
3. K겹 교차검증…………….………………………………………………………………………………………………….….4

**Ⅳ. 성능 평가 지표 및 성과……………………………………..………………4**

1. 오차 행렬과 ROC-곡선……………………………………………………………………………………………………….4
2. 모델 성과……………………………………………………………………………………………………….…………………..5

**V. 개선안 및 결론……………………………………………………………………5**

**1. 서론**

알코올 중독은 전 세계적으로 심각한 공중 보건 문제를 야기하고 있으며, 이로 인한 사회적 및 경제적 비용은 매우 크다. 알코올 중독은 개인의 건강은 물론 가족과 사회에도 부정적인 영향을 끼치는 만큼, 효과적인 진단과 조기 개입이 중하다. 현재 알코올 중독의 진단은 주로 자가 보고식 설문, 임상 평가에 의존하고 있으나, 이러한 방법들은 종종 편향되거나 정확도가 떨어질 수 있다.

이러한 문제를 해결하기 위해 뇌파 검사가 중요한 대안으로 떠오르고 있다. 뇌파 검사는 뇌의 전기적 활동을 측정하여 뇌의 기능적 상태에 대한 심층적인 이해를 제공한다. 특히, 알코올 중독자의 뇌파 패턴은 비중독자와 다르게 나타나는 특징이 있어, 이를 통해 알코올 중독의 생물학적 마커를 정확히 식별할 수 있다. 뇌파 검사를 이용하면 객관적이고 정량적인 데이터에 기반하여 알코올 중독을 조기에 발견하고, 개인 맞춤형 치료 계획을 수립하는 데 크게 기여할 수 있다.

뇌파 데이터를 활용한 연구는 알코올 중독으로 인한 뇌 기능의 변화를 명확히 파악할 수 있을 뿐만 아니라, 중독의 정도를 객관적으로 평가할 수 있는 기회를 제공한다. 이는 향후 알코올 중독 치료 방법의 효율성을 극대화하고, 재발 방지 및 관리에 있어 보다 체계적인 접근을 가능하게 한다.

본 프로젝트는 뇌파 데이터 전처리 과정과 모델링 방식을 소개하고, 결과를 분석하여 모델의 성능을 평가하는 것을 목표로 한다.

**2. 데이터**

본 프로젝트에는 알코올 중독자와 비중독자를 구분해 놓은 뇌파(EEG) 데이터가 사용되었다. 데이터셋 안에는 64개의 전극을 통해 측정한 뇌파 데이터가 1080개 존재한다. 모델의 성능을 평가하기위해 1080개의 데이터에서 10%를 테스트 데이터로 활용했고 하이퍼 파라미터를 최적화 하기위해 나눠진 학습 데이터의 12.5%를 검증 데이터로 활용했다.

뇌파 데이터는 고차원 특성을 갖고 있어 데이터의 원시 형태를 직접 사용할 경우 학습 모델의 복잡도가 증가하고 이는 과적합으로 이어질 수 있다. 따라서 뇌파 데이터로부터 학습 모델의 복잡도를 효과적으로 관리하면서도, 중독 상태를 정확하게 구분할 수 있는 유의미한 특징을 추출하기 위해 데이터 전처리 과정을 거쳤다.

**2.1데이터 전처리**

**2.1.1 PSD**

EEG데이터의 경우 대게 고속 푸리에 변환(FFT)을 하고 주파수 대역을 설정하여 그 대역에 해당하는 에너지 스펙트럼 밀도(PSD)의 산출값을 사용한다 [1]. 특징을 더 줄이기 위해 산출값의 중간값을 사용했다. 이때 평균값, 최소값, 최대값을 사용할 수 있지만 중간값을 사용했을 때 모델이 가장 높은 정확도를 기록했다. 계산에는 파이썬의 MNE라이브러를 사용했다. 변형 후 데이터는 행이 그 개수를 나타내고 열이 5개의 주파수 대역을 나타낸다. 대역 별 주파수의 범위는 <표1>과 같다. 기존에 널리 알려진 뇌파 주파수 대역(delta(0~4Hz), theta(4~8Hz), alpha(8~13Hz), beta(13~30Hz), gamma(30~50Hz))을 최대한 보존하면서 모델의 성능을 높이기 위해 주파수의 범위를 조금 조정했다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 대역 별 주파수 범위 (Hz) | | | | |
| a | B | c | d | e |
| 0.5 ~ 5 | 5 ~ 10 | 10 ~ 15 | 15~28 | 31~43 |

<표1> 대역 별 주파수 범위

**2.1.2 정규화**

데이터의 스케일이 다르면 파라미터 업데이트가 늦어질 수 있고 작은 스케일의 변수가 무시될 수 있다. 따라서 데이터를 정규화 함으로 이것을 예방할 수 있다. 정규화를 하는 방법에는 여러가지가 있지만 정해진, 가장 좋은 방법은 없다. 데이터와 알고리즘에 따라 성능이 제각각 이기 때문에 여러 개를 테스트해보고 선택해야 한다[2].

본 프로젝트에는 데이터를 평균이 0이고 분산이 1이되도록 만드는 standardization, 변수의 범위를 0과 1사이로 변환하는 min-max, 데이터의 중앙값을0, IQR이 1이되도록 만드는 robust 방법을 사용해 보았다. 학습데이터에서 분리한 테스트 데이터를 활용해 해당 정규화를 사용하고 optuna를 활용하여 하이퍼 파라미터를 최적화했을 때 <표2>와 같이 세 가지 방법 모두 최고91%의 정확도를 달성했다. 그러나 kaggle의 테스트 데이터에 대한 정확도는 min-max방법을 사용했을 때 95.833%로 가장 높았다[3]. 따라서 min-max 스케일링 기법을 정규화에 사용했다.



- Min-max정규화 식 –

|  |  |  |  |
| --- | --- | --- | --- |
| Scaling  method | Test data accuracy | Scaling  method | Kaggle accuracy |
| Standardization | 0.91 | Standardization | 0.94166 |
| Min\_max | 0.91 | Min\_max | 0.95833 |
| Robust | 0.91 | Robust | 0.91666 |

<표2> 데이터에 따른 scaling method 정확도

**3. 학습 과정**

**3.1 XGBoost**

데이터 학습에는 XGBoost 모델이 사용되었다. XGBoost란 그레디언트 부스팅 알고리즘의 변형중 하나로 병렬 학습이 이루어지는 알고리즘이다. 정형화된 데이터에 강하고 학습 속도가 빠르며 자체에 하이퍼 파라미터 형태로 여러 과적합 규제 기능이 있기 때문에 사용했다. <그림1>을 보면 분석 대회에서 많은 참가자 들이 XGBoost 사용함으로 그 능력이 입증되었다는 것을 알 수 있다. <표3>은 학습데이터에서 분리한 테스트 데이터로 5가지 성능 척도를 활용해 나타낸 표이다. 5개의 분류 알고리즘 중 XGBoost를 사용한 모델의 성능이 가장 좋은 것을 알 수 있다.

텍스트, 스크린샷, 번호, 그래프이(가) 표시된 사진

자동 생성된 설명

<그림1> 2022년 캐글대회 사용 라이브러리와 알고리즘

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 정확도 | 민감도 | 특이도 | 정밀도 | F1-점수 |
| XGBoost | 0.90 | 0.91 | 0.89 | 0.89 | 0.90 |
| Random  Froest | 0.84 | 0.87 | 0.81 | 0.82 | 0.85 |
| SVM | 0.56 | 0.96 | 0.15 | 0.53 | 0.68 |
| LDA | 0.75 | 0.83 | 0.67 | 0.71 | 0.77 |
| KNN | 0.83 | 0.78 | 0.89 | 0.88 | 0.82 |

<표3> 알고리즘별 모델 성능 비교

**3.2 하이퍼 파라미터 최적화**

모델을 제작한 후 검증 데이터를 통해 하이퍼 파라미터를 최적화했다. 그 과정에서 하이퍼 파라미터 최적화 과정을 도와주는 파이썬의 프레임워크인 Optuna를 사용했다. Optuna는 하이퍼 파라미터의 최대값과 최소값, 간격을 정해주면 자동탐색을 통해 최적의 하이퍼 파라미터를 찾아준다. 더 좋은 결과를 얻기 위해 최적화 과정을 100번 반복한다. Optuna를 통해 최적화한 하이퍼 파라미터는 <표4>과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 하이퍼 파라미터 |  | 하이퍼 파라미터 |
| 1 | Max\_depth | 6 | colsample\_bynode |
| 2 | Learning\_rate | 7 | reg\_lambda |
| 3 | n\_estimators | 8 | subsample |
| 4 | colsample\_bytree | 9 | min\_child\_weight |
| 5 | colsample\_bylevel | 10 | gamma |

<표4> 최적화 대상 하이퍼 파라미터 종류

“Max\_depth”는 생성할 트리의 최대 깊이를 나타낸다. 이 값을 크게 할수록 모델이 더 복잡해지고 과적합 될 가능성이 높아진다. “Learning\_rate”는 각 트리가 기존 트리의 오차를 얼마나 강하게 보정할지를 결정하는 학습률이다. “n\_estimators”는 사용할 트리의 개수를 나타낸다. 이 값이 커질수록 모델이 과적합 될 가능성이 높아진다. ”colsample\_bytree”, “colsample \_bylevel”, “colsample\_bynode는 각각 트리를 구성하는 요소를 샘플링 하는 비율을 나타낸다. 이를 통해 트리 간의 상관관계를 줄이고 다양성을 증가시켜 과적합을 방지할 수 있다. “reg\_lambda”는 L2규제의 강도를 나타낸다. 트리 가중치를 제한하여 모델의 복잡성을 줄여 과적합을 방지할 수 있다. “subsample”은 각 트리를 학습할 때 사용하는 데이터 샘플의 비율을 나타낸다. 트리 성장전 훈련 데이터를 무작위로 샘플링 함으로 과적합을 방지할 수 있다. “min\_child\_weight”는 각각의 샘플이 가져야 하는 최소 가증치를 나타내는데 트리분할단계에서 인스턴스 가중치 합계가 이것보다 작은 리프 노드가 생성되면 빌드 프로세스에서 추가 분할을 포기한다. 이 값이 작을수록 과적합 가능성이 높아진다. “gamma”는 트리의 분할을 결정할 때 필요한 최소 손실 감소를 나타낸다. Gamma가 작을수록 과적합될 확률이 높아진다. 결론적으로 이 하이퍼 파라미터들은 과적합을 제어하는 역할을 가지고 있기 때문에 최적화 화면 그 정도를 줄일 수 있다.

**3.3 K-겹 교차 검증**

모든 데이터를 학습과 검증에 활용하여 과적합을 방지하고 더욱 일반화된 모델을 생성하기 위해K-겹 교차 검증 방법을 사용했다. K-겹 교차 검증은 데이터 셋을 K개의 fold로 나누어 각 학습마다 1~K번째의 fold중 하나를 검증 세트로 사용하고 나머지 데이터를 훈련 세트로 사용하여 모델을 k번 학습하고 평가하는 과정을 반복하는 방법이다. K번 fold내의 클래스 비율을 유지하기 위해 파이썬의 StratifiedKFold를 사용했다.

K를 선택할 때 그 값이 작으면 모델이 각 폴드에서 학습하는 데이터 양이 적어져 모델이 편향되기 쉽고 k값이 크면 분산이 증가한다. 따라서 이러한 편향과 분산 사이의 균형을 유지하기 위해 적당한 값을 선택해야 한다. 또한 K값의 변화와 머신 러닝 모델 알고리즘의 정확도 변화와는 직접적인 연결성이 없기 때문에 값을 바꿔보며 정확도를 개선시키는 최적의 k를 실험적으로 찾아내야 한다[4]. <표5>는 학습데이터에서 분리한 테스트 데이터를 활용하여 k값을 5~10으로 변경했을 경우 모델의 정확도를 나타낸 표이다. K=8일때의 모델 정확도가 가장 높은 것을 볼 수 있다. 따라서 교차 검증에서의 k값을 8로 설정했다.

|  |  |
| --- | --- |
| 교차검증 k값 | 모델 정확도 |
| K = 5 | 0.88 |
| K = 6 | 0.83 |
| K = 7 | 0.88 |
| K = 8 | 0.90 |
| K = 9 | 0.85 |
| K = 10 | 0.89 |

<표5> k값에 따른 모델 정확도

**4. 성능 평가 지표 및 성과**

**4.1 오차 행렬과 ROC-곡선**

모델의 예측 성능을 평가하기 위해 모델이 예측한 클래스와 실제 클래스 간의 관계를 보여주는 오차행렬을 사용했다. 성능 평가 지표들의 수치는 학습 데이터에서 분리한 테스트 데이터의 정확도가 아닌 kaggle에 있는 테스트 데이터의 정확도가 가장 컸을 때의 optuna로 얻은 하이퍼 파라미터를 사용한 결과이다[4]. <그림2> 가 모델에 대한 오차행렬을 나타낸다.

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

<그림2> 오차행렬

모델의 성능 평가 지표에는 정확도(accuracy), 민감도(sensitivity or recall), 특이도(specificity), 정밀도(precision), F-1점수(F1-score)가 있으며 해당 값이 클수록 모델의 성능이 좋다는 것을 의미한다.

ROC(Receiver Operating Characteristic) 곡선은 모델의 민감도와 거짓 양성 비율(False Positive Rate, FPR)간의 관계를 시각화 한 곡선이다. ROC-AUC-점수는 ROC곡선 아래 영역을 나타내는 값으로 모델의 성능을 평가하는 데에 사용된다. ROC-AUC점수는 0과1사이의 값을 가지며 1에 가까울수록 모델의 성능이 우수하다. <그림3> 는 ROC곡선을 나타내고 <표6>는 정확도와 5가지의 평가 결과, ROC의 AUC점수를 나타낸다.

|  |  |
| --- | --- |
| 성능 평가 지표 | 성능 평가 백분율 |
| 정확도(Accuracy) | 90% |
| 민감도(Sensitivity or Recall) | 91% |
| 특이도(Specificity) | 89% |
| 정밀도(Precision) | 89% |
| F1-점수(F1-score) | 90% |
| ROC-AUC-점수 | 97% |

<표6> 성능 평가 지표

텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

<그림3> ROC-곡선

**4.2 모델 성과**

주어진 학습데이터셋에서 레이블의 비율을 동일하게 하고 10퍼센트를 분리한 108개의 임의의 테스트 데이터셋을 활용해 6가지 방법으로 모델의 성능을 평가했다. <표2>는 그 결과를 보여준다. 전체 표본 중 예측을 성공한 비율을 나타내는 정확도는 90%, 알코올 중독자를 중독자로 예측한 비율인 민감도는 91%를 달성했다. 비중독자를 비중독자로 예측한 비율인 특이도는 89%, 알코올 중독자라고 예측한 결과가 맞았을 확률인 정밀도는 89%를 달성했다. 민감도와 정밀도를 종합한 성능을 수치화한 F1-점수는 90%, 대한 거짓 양성 비율(FPR)에 대한 진짜 양성 비율(TPR)을 측정한 ROC-AUC점수는 97%를 달성했다. 이 6가지 척도의 결과들을 종합적으로 보았을 때 해당 모델은 충분히 신뢰성이 있다고 할 수 있다.

**5. 개선안 및 결론**

**5.1 개선안**

데이터 전처리 부분에서 개선할 부분이 보인다. 본 프로젝트는 64개의 전극으로 수집한 EEG데이터를 사용한다. EEG데이터는 신호 안에 눈, 근육의 움직임, 심장 박동등의 잡음이 포함되어 있다. 그러나 데이터 전처리 단계에서 잡음을 제거하는 과정이 없다. 해당전극의 이름을 알 수 있었다면 그것을 통해 뇌전지도를 그려 잡음 유무를 확인하고 독립 성분 분석(independent component analysis, ICA)방법을 통해 중간에 튀는 경향을 가지는 요소를 배제함으로써 잡음을 제거할 수 있었을 것이다. 또한 필요한 전극에 해당하는 데이터만 선택하여 모델의 성능을 더 높일 수 있었을 것이다. 마지막으로 데이터를 더 잘 분석해서 모델이 레이블을 더 잘 구분할 수 있도록 주파수 대역을 설정한다면 모델의 성능을 더 높일 수 있을 것이다

**5.2 결론**

본 프로젝트는 XGBoost 알고리즘을 활용하고 정규화, 교차검증 등 여러 방법을 사용하여 알코올 중독자를 분류하는 모델을 만들어 신뢰도를 평가했다. 프로젝트 초반에 로지스틱 회귀 모델과 SVM 모델을 활용하여 train\_accuracy는 90퍼센트 이상인 것에 반해 test\_accuracy는 60퍼센트 정도를 달성하는 것을 보고 과적합이 일어난 것을 확인했다. 그것을 막기 위해 교차검증과 L2규제 등 여러 방법을 사용했다. 그럼에도 과적합이 해결되지 않자 데이터 자체의 문제라 판단해 전처리 과정을 거쳤다. 날것의 EEG데이터에서 의미 있는 특징을 추출하기 위해 고속푸리에 변환을 사용하고 주파수별 파워 스펙트럼 밀도의 중간값을 학습을 위한 데이터로 변형하여 사용했다. 그 후 모델 또한 과적합을 방지하고 예측 성능을 높이기 위해 XGBoost로 변경했다.

모델의 성능을 요약해보면 가장 대표적으로 사용되는 정확도는 90%임으로 모델이 데이터를 잘 예측하고 있다는 것을 알 수 있다. 민감도, 특이도, 정밀도, F1-점수 또한 대략90%를 달성하며 모델의 신뢰도를 증명하고 있다. 특히 ROC-AUC 점수가 97%로 매우 높은 수치를 기록함으로 모델이 데이터를 매우 잘 분류하고 있다는 것을 알 수 있다.

신뢰할 만한 모델을 제작했지만 kaggle의 테스트 데이터로 100%의 정확도를 달성하지 못했다. 같은 데이터로 완벽하게 분류해낸 팀이 존재하는 것을 보면 본 프로젝트에서 보완해야 할 부분이 존재한다는 것을 알 수 있다. 뇌파데이터 신호처리에 대한 부분을 좀 더 잘 분석해서 데이터의 경향을 더 잘 나타내는 특징을 추출한다면 모델의 성능을 더 높일 수 있을 것으로 예상된다.

**\* 전체 코드 확인 주소:**

https://github.com/Hagyeongjun/machine-learning-1/commit/be38af31626092ce0c438eb24239001cf04fad6e

**참 고 문 헌**

[1] 서정렬, 손경아, “EEG 데이터를 활용한 알코올 중독자 식별 모델 연구”, 2018 한국소프트웨어 종합 학술 대회 논문집, 한국정보과학회, 2018, 705쪽

[2] Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, Siddique Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies. 2021; 9(3):52.

[3] <https://www.kaggle.com/competitions/seoultech-applied->ai-machine-learning1/leaderboard

[4] Isaac Kofi Nti, Owusu Nyarko-Boateng, Justice Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation", International Journal of Information Technology and Computer Science(IJITCS), Vol.13, No.6, pp.69, 2021. DOI: 10.5815/ijitcs.2021.06.05