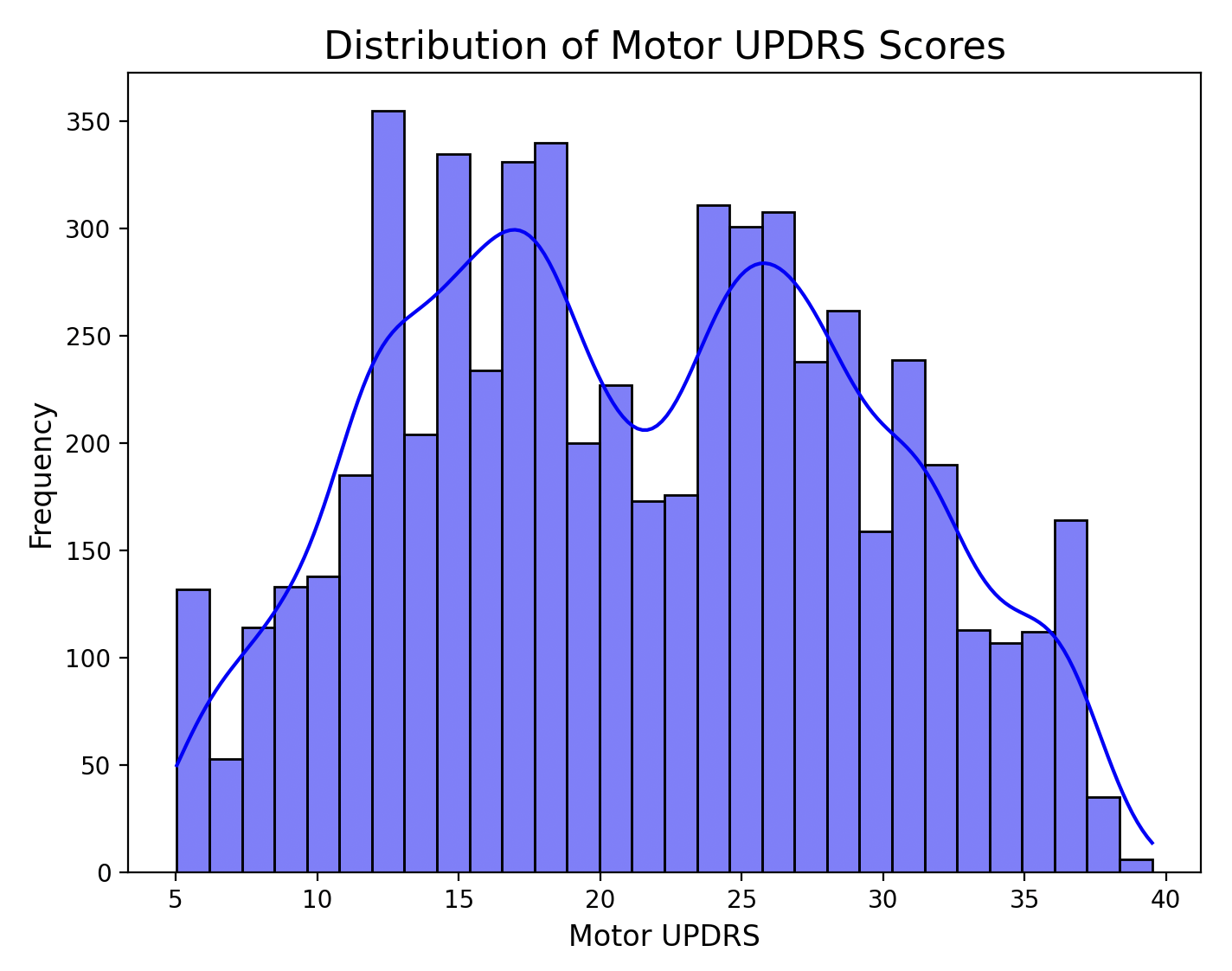
**[A2] Comparison of Supervised Learning Algorithms**

**2021313075 백경인**

1. **데이터셋**

<https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring> 에서 받아온 데이터로 원격 증상 진행 모니터링을 위한 원격 모니터링 장치의 6개월 시험에 모집된 초기 파킨슨병 환자 42명의 다양한 생체의학적 음성 측정으로 구성 되어있다.Data point 개수: 5875개/ Feature 개수: 16이며Label의 분포는 아래와 같다.



1. **과업의 특성에 따른 적절한 평가지표의 선정:**

아무래도 target이 실수형이라서 MSE와 MAE를 고민했으나, 보편적으로 MSE를 사용하며 이상치가 크게 발견되지 않아 이상치에 민감한MSE를 사용하더라도 큰 문제가 없다고 판단했다.

1. **3개 이상의 학습 알고리즘 비교**

LinearRegression, RandomForestRegressor, GradientBoostingRegressor 중에서 RandomForestRegressor, GradientBoostingRegressor, LinearRegression 순으로 성능이 좋았다.

1. **최적의 데이터 전처리 방법:**

애초에 비여있는 값이 없는 데이터를 받아와서 empty value를 처리할 필요는 없었다. 또한 regression 모델의 target 값이 실수라서 one-hot encoding도 필요없었다. 그렇기에 scaling에만 신경을 썼다. 이를 위해 min-max와 standard scaling을 했었는데 minmax가 더 좋은 성능을 보였다.

1. **최적의 학습 알고리즘:**

Random Forest가 최고의 성능을 보였다. 아무래도 과적합 위험 감소에 용이하고 중요한 feature를 잘 찾아내는 성향을 갖고 있어서 16개의 feature 중에서 가장 영향력이 있는 feature에 집중시킬 수 있게 모델을 짤 수 있었던 것 같습니다. 또한 여러 feature들이 target 값과 비선형적인 관계를 갖고 있어 이런면에서 linear model인 linear regression보다 우위에 있었던 것 같다. 또한 target이 정규분포와는 살짝 거리가 있는 느낌이라서 Linear Regression의 MLE 설정시 guassian distribution을 사용하는 것을 고려했을때에는 적절하지 않다고 판단했다.

1. **하이퍼파라미터 설정 제시:**

Random Forest 같은 경우에는 n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf를 hyperparmeter로 설정했으며, Gradient boosting은 추가적으로 learning rate또한 설정했다. 성능이 가장 좋았던 설정은 Random forest에서

1. **추가 성능 개선을 위한 방안 논의:**

Cost function 선택에는 자유가 있기에 상황별로 cost function또한을 학습하는 모델이 있다면 더 scable한 모델이 성능적인 측면에서 향상을 만들 수 있을 것 같다.

icampus 제출 (https://icampus.skku.edu/)

* 파일명: A2\_학번\_이름 (예: A2\_2020300000\_가나다)