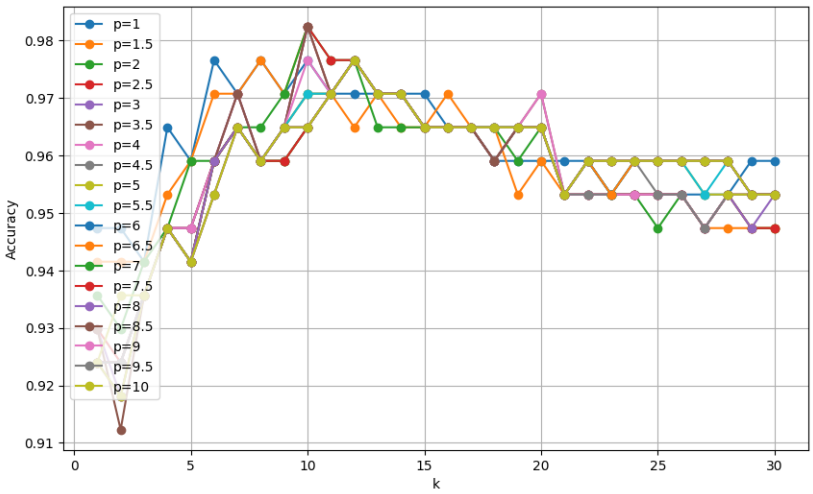
데이터마이닝 과제2 – 2022920024 박동찬

1. Breast Cancer Classification
   A. 실험

      Neighbor k는 1부터 30까지의 정수로, Minkowski distance p는 1부터 10까지 0.5씩 커지도록 정하여 Grid Search한다. 이때 결과를 보고 어떤 경향성을 보이는지 시각화하고 가장 높은 Accuracy를 도출하는 (k, p) 파라미터 값을 찾는다. 단순 Accuracy는 클래스 불균형 문제로 왜곡될 수 있으므로 가장 높은 결과를 보이는 파라미터 조합에서 F1 score를 추가로 도출한다. 또한 Breast Cancer Dataset에서 train set과 test set으로 분리할 때 stratify를 적용하지 않은 경우와 적용한 경우도 추가로 실험해본다.

   B. 실험 결과

      i. Stratify = False

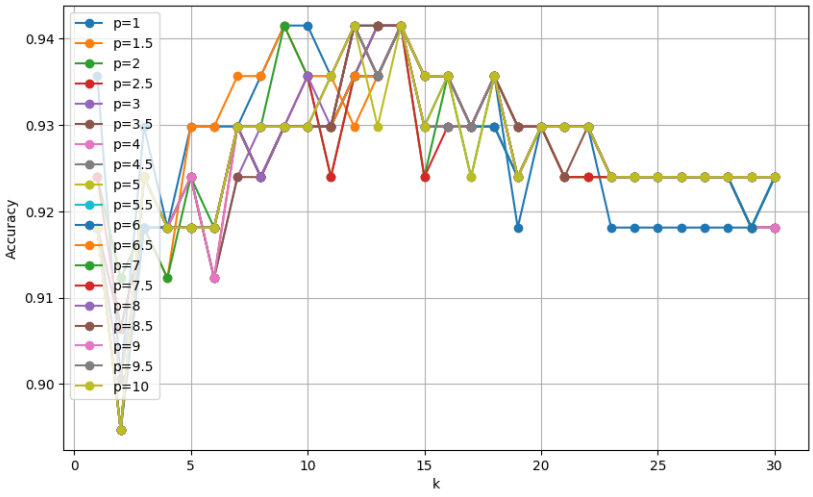


      Best Accuracy: 0.9825

      Best (k, p): (10, 1.5), (10, 2), (10, 2.5), (10, 3), (10, 3.5) – F1-score: 0.9862

      즉, k=10일 때, p가 1.5, 2, 2.5, 3, 3.5인 경우 최고 정확도를 얻었다.

      ii. Stratify = True



Best Accuracy: 0.9415

Best (k, p): (9, 1), (10, 1), (12, 1), (13, 1), (14, 1), (9, 1.5), (14, 1.5), (9, 2), (13, 2), (14, 2), (13, 2.5), (14, 2.5), (13, 3), (14, 3), (12, 3.5), (13, 3.5), (14, 3.5), (12, 4), (14, 4), (14, 4.5), (14, 5), (14, 5.5), (14, 6), (14, 6.5), (12, 7), (14, 7), (12, 7.5), (14, 7.5), (12, 8), (14, 8), (12, 8.5), (14, 8.5), (12, 9), (14, 9), (12, 9.5), (14, 9.5), (12, 10), (14, 10) – F1-score: 0.9541

C. 결과 분석

i. k 값에 따른 변화

k 값이 작은 경우(ex. 1~5)는 이웃 수가 적어 overfitting이 일어나 정확도가 들쭉날쭉하고 비교적 낮은 것 볼 수 있다. k 값이 큰 경우(ex. 20 이상)에서는 많은 이웃을 고려하므로 구분해야 하는 패턴이 희석되어 정확도가 낮다. 그래프를 보았을 때 k=10 전후에 집중적으로 정확도가 가장 높게 나타나고 이 경우가 Breast Cancer dataset에서 가장 균형있게 경계를 형성하는 지점이라고 추정할 수 있다. 또한 F1 score도 0.9862, 0.9541로 정밀도와 재현율이 모두 균형 있게 높음을 알 수 있다.

ii. p 값에 따른 변화

p 값의 그래프를 보았을 때 대부분 비슷한 결과와 경향을 보인다. 즉, 정확도가 p 값에 민감하게 변하지 않는다고 생각할 수 있다. Breast Cancer dataset은 선형적으로 분리하기 쉽게 분포해있기 때문에 p 값을 바꾸어도 비슷한 성능이 보여진다.

iii. Stratify에 따른 변화

Stratify=True로 설정하면 전체 데이터 클래스 비율을 train set과 test set에 동일하게 유지한다. Stratify=False인 경우는 Best Accuracy: 0.9825, Stratify=True인 경우는 Best Accuracy: 0.9415으로 오히려 클래스 비율을 맞추지 않은 경우가 더 높았다. 이는 우연히 더 분류하기 쉬운 데이터가 test set에 들어갈 수 있고 이로 인해 인위적으로 성능이 좋게 나온 것이라 생각할 수 있다. Stratify=True인 경우, 정확도가 약 0.94로 낮아졌지만 이 결과가 신뢰할 수 있는 실제 성능에 가깝다고 생각할 수 있고 여전히 분류 모델로써 좋은 성능이다. 또한 더 많은 파라미터 조합에서 유사한 정확도를 보이는데 데이터 분포가 고르게 반영된 결과라고 볼 수 있다.

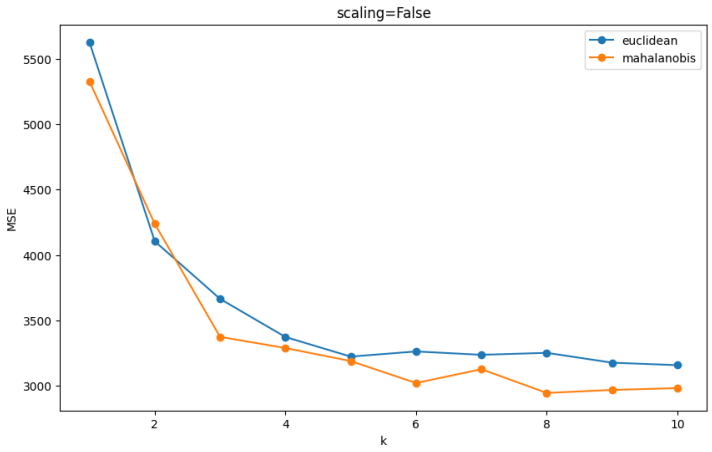
2. Diabetes Regression

A. 실험

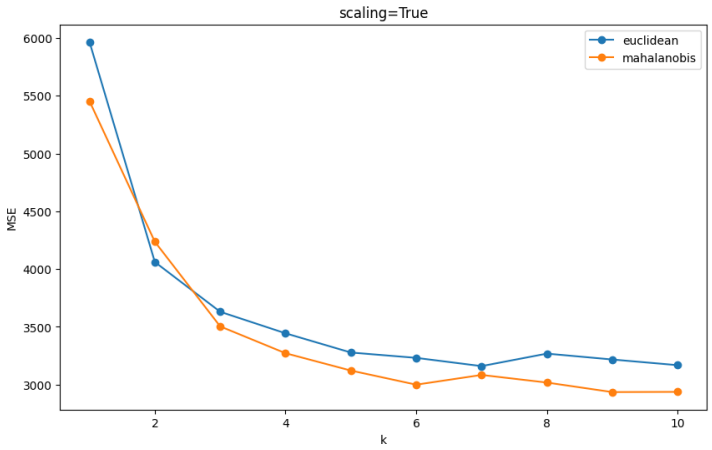
실험 결과에서 안정성과 일반화 성능을 확인하기 위해 neighbor k를 1부터 10까지의 정수로 정하여 테스트한다. Scaling 방식은 StandardScaler를 사용한다.

B. 실험 결과

i. Scaling=False



ii. Scaling=True



C. 결과 분석

i. Scaler 미적용 - Mahalanobis vs Euclidean

K=2인 경우를 제외하고 나머지 경우에서 모두 Manalanobis distance metric이 Euclidean distance보다 낮은 MSE 성능을 보였다. Mahalanobis가 평균적으로 약 160 정도 낮은 MSE 값을 기록했다. 스케일링 없이 원본 데이터로 학습하면 각 특성의 단위와 분산 차이가 그대로 반영된다. Euclidean은 단순히 특성의 차이를 제곱하여 더하므로 단위 차이가 결과에 영향을 미친다.

Mahalanobis는 특성들 간 상관관계를 고려하므로 단위 차이와 분산의 불균형을 보정할 수 있다. 따라서 스케일링이 안된 상태에서는 Mahalonobis가 관계를 더 잘 반영하여 좋은 결과가 도출된다.

ii. Scaler 적용 - Mahalanobis vs Euclidean

K=2인 경우를 제외하고 나머지 경우에서 모두 Manalanobis distance metric이 Euclidean distance보다 낮은 MSE 성능을 보였다. Scaling을 적용하면 각 특성의 단위 차이가 제거되어 모든 특성이 같은 스케일로 균형 잡히게 된다. Scaling 하면 두 metric 모두 안정적이지만 Mahalanobis가 상관관계를 잘 반영하여 여전히 더 낮은 MSE 결과가 도출된다. 평균적으로 Mahalanobis가 약 190 정도 낮은 MSE 값을 기록했다.

iii. Scaler 미적용 vs 적용 – Mahalanobis

Mahalanobis의 경우, Scaling 적용 여부와 관계없이 거의 비슷한 경향을 보인다. Scaling 적용하면 평균적으로 MSE가 약 10 정도 증가했다.

iv. Scaler 미적용 vs 적용 – Euclidean

Euclidean의 경우, 전반적으로 Scaling 적용 여부와 관계없이 거의 비슷하게 나타났다. Scaling 적용하면 평균적으로 MSE가 약 35 정도 증가했다. Diabetes dataset은 원래 각 특성이 적절한 단위와 분포를 갖도록 pre-processing 되어 있다. 이미 잘 처리되어 있는 데이터에 추가적인 Scaling을 가하면 원래의 정보 차이 값이 감소할 수 있다. 또한 각 특성이 가지고 있는 단위와 분산 차이가 유의미할 수 있고 이를 Scaling 하면 중요도가 희석되어 정보를 오히려 잃을 수도 있다. 이러한 이유들로 인해 Scaling 했을 때 MSE의 변화가 크게 없거나 오히려 증가하는 경향을 보인다.