기초인공지능

HW02 20201036 김도훈

1. 분류 성능 결과표

- 결과 표(소수점 넷째 자리까지 표시)

	Decision Tree	Random Forest	Naive Bayes	XGBoost	MLP	My Model
Accuracy	0.7468	0.7338	0.7662	0.7078	0.7013	0.8000

2. 5가지 모델

1. Decision Tree

- 알고리즘 개념

  Decision Tree Algorithm은 데이터를 분류하거나 회귀 분석을 수행하기 위해 트리 구조를 사용하는 지도 학습 알고리즘이다. 각 노드는 특정 특징(feature)을 기준으로 데이터를 분할하며, 엔트로피를 계산하고, 리프 노드에서 최종 예측 값을 제공한다. 이해와 해석이 쉬운 장점이 있지만, 과적합을 방지하기 위해 트리 깊이 제한, 가지치기(pruning) 등의 기법이 필요하다.

- 파라미터 값 및 간략한 설명

```
decision_tree_classifier.fit(X_train, y_train)#train set 의 input 과 output 값을 이용하여 학습 진행
y_prediction = decision_tree_classifier.predict(X_test)#test set 의 input 으로 예측 진행. 이후 모든 모델들에 대하여 동일하게 진행됨.
```

2. Random Forest

- 알고리즘 개념

  Random Forest Algorithm은 다수의 결정 트리(Decision Tree)를 생성하고, 각 트리의 예측 결과를 앙상블하여 최종 예측을 수행하는 지도 학습 알고리즘이다. 트리 생성 시 데이터와 특징을 무작위로 샘플링하여 다양성을 높이고, 과적합을 방지한다. 높은 예측 정확도와 안정성을 가지며, 분류(Classification)와 회귀(Regression) 문제에 모두 사용할 수 있다.

3. Naive Bayes

- 알고리즘 개념

Naive Bayes Algorithm은 베이즈 정리를 기반으로 특정 클래스가 주어졌을 때 데이터가 관찰될 확률을 계산하여 분류를 행한다. 각 특징을 독립적으로 계산하여 곱셈을 한다. 빠르고 계산 효율성이 높아 텍스트 분류, 스팸 필터링 등에서 널리 사용된다. 특징 간의 상관관계가 강한 경우 성능이 저하될 수 있지만, 간단하고 실용적인 모델로 유용하다.

기초인공지능

HW02 20201036 김도훈

4. XGBoost

■ 알고리즘 개념

XGBoost Algorithm은 Gradient Boosting을 개선한 앙상블 학습 알고리즘으로, 각 단계에서 오류를 최소화하는 새로운 결정 트리를 추가하며 예측 성능을 점진적으로 향상시킨다. 모델의 목적 함수를 <손실 함수+정규화 항> 형태로 정의하여 과적합을 방지하고, 계산 속도를 최적화하기 위해 병렬 처리와 조기 종료(Early Stopping)를 지원한다. 높은 예측 정확도와 효율성을 제공하며, 분류(Classification)와 회귀(Regression) 문제에서 모두 사용될 수 있다.

5. MLP

■ 알고리즘 개념

MLP (Multi-Layer Perceptron) 알고리즘은 입력층, 하나 이상의 은닉층, 출력층으로 구성된 완전 연결 신경망(fully connected neural network)이다. 각 뉴런은 활성화 함수(예: ReLU, sigmoid)를 통해 비선형 변환을 수행하며, 가중치와 편향을 학습하여 입력 데이터와 출력 간의 복잡한 관계를 모델링한다. 역전파(Backpropagation) 알고리즘을 사용해 가중치를 최적화하며, 분류(Classification)와 회귀(Regression) 문제에서 유연하고 강력한 성능을 제공한다.

3. My Method

1. 데이터 전처리

1. 데이터가 누락되어 0인 경우가 있다. **'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'Age', 'BMI'** 와 같은 파라미터의 값은 0이 나오는 것이 거의 불가능한 값이므로, 이 부분에서 나온 0은 KNN 결과값으로 대체하였다.

2. 현재 8열 데이터 구조인데, 특징을 늘리고자 4열을 추가하였다. 추가한 방식은 관련이 있을 법한 두 열의 값을 곱한 것이다.

3. StandardScaler 클래스를 이용하여 분산 범위가 다른 각 열의 값들을 같은 범위로 통일시켰다.

4. 데이터 부족 문제를 해결하기 위하여 Gaussian Noise를 추가하였고, SMOTE를 통해 데이터 양을 늘렸다.

2. 앙상블 기법

1. 위에서 다룬 여러 모델들을 통하여 각 항목 별로 가장 좋은 예측 결과를 택하는 앙상블 기법을 수행하였다.

기초인공지능

HW02 20201036 김도훈

2. 이때 각 모델에 대한 파라미터를 수정해가면서 성능 평가를 지속하였다.

4. 성능 평가

실험 결과 다음과 같이 성능이 소폭 상승하였음을 확인했다.

```
1/1 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 115ms/step - accuracy: 0.7719 - loss: 0.1584 - val_a
7/7 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 13ms/step
Accuracy: 0.8000
```