기말 프로젝트 보고서: 유전자 데이터를 활용한 선형 회귀와 딥러닝 모델 비교

20201811 차영원

---

1. 소개

유전자 데이터 분석은 현대 의학 및 생물학 연구에서 중요한 주제 중 하나이다. 개체의 유전체 정보를 분석함으로써 유전적 특성을 이해하고 질병 발생, 특정 특성, 그리고 환경과의 상호작용을 파악할 수 있다. 이러한 목적을 달성하기 위해서는 유전자 데이터의 특성을 정확하게 예측하는 모델이 필요하다. 본 프로젝트에서는 주어진 유전자 데이터를 활용하여 선형 회귀와 딥러닝 모델을 비교하여 연속형 변수와 이진형 변수를 예측하는 작업을 수행하였다.

---

2. 데이터

- training data와 testing data

training phenotype data: pheno_n300_conti.phe, pheno_n300_binary.phe
training genotype data: sim_n300_p1000.mldose, sim_n300_p1000.mlinfo
testing phenotype data: pheno_n150_conti.phe, pheno_n150_binary.phe
testing data: sim_n150_p1000.mldose, sim_n_150_p1000.mlinfo

---

3. 방법

1) 선형 회귀(Linear Regression):

- 연속형 변수를 예측하기 위해 표준 선형 회귀 모델을 사용하였습니다.
- 평가 지표로는 Mean Squared Error (MSE)를 사용하였습니다.

2) 로지스틱 회귀(Logistic Regression):

- 이진형 변수를 예측하기 위해 로지스틱 회귀 모델을 사용하였습니다.
- 평가 지표로는 Accuracy와 Area Under the ROC Curve (AUC)를 사용하였습니다.

3) MLP(Multi-layer Perceptron):

- 연속형 변수 및 이진형 변수를 예측하기 위해 MLP 모델을 사용하였습니다.
- 평가 지표로는 MSE, Accuracy, 그리고 AUC를 사용하였습니다.

4) CNN(Convolutional Neural Network):

- 연속형 변수 및 이진형 변수를를 예측하기 위해 CNN 모델을 사용하였습니다.
- 평가 지표로는 MSE, Accuracy, 그리고 AUC를 사용하였습니다.

5) RNN(Recurrent Neural Network):

- 연속형 변수 및 이진형 변수를를 예측하기 위해 RNN 모델을 사용하였습니다.
- 평가 지표로는 MSE, Accuracy, 그리고 AUC를 사용하였습니다.

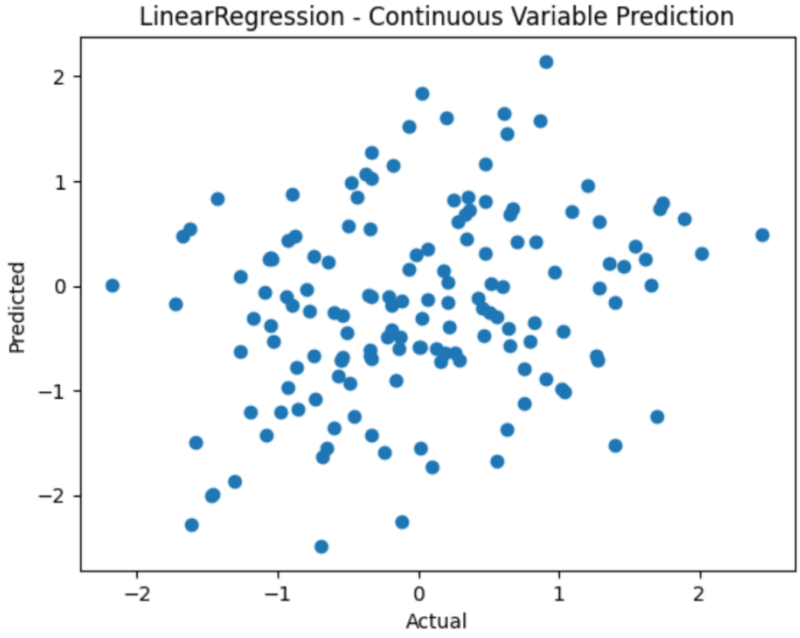
---

4. 결과

1) 선형 회귀:

연속형 변수에 대한 MSE: 1.2363

```
Mean Squared Error (Continuous): 1.2363372397496089
```



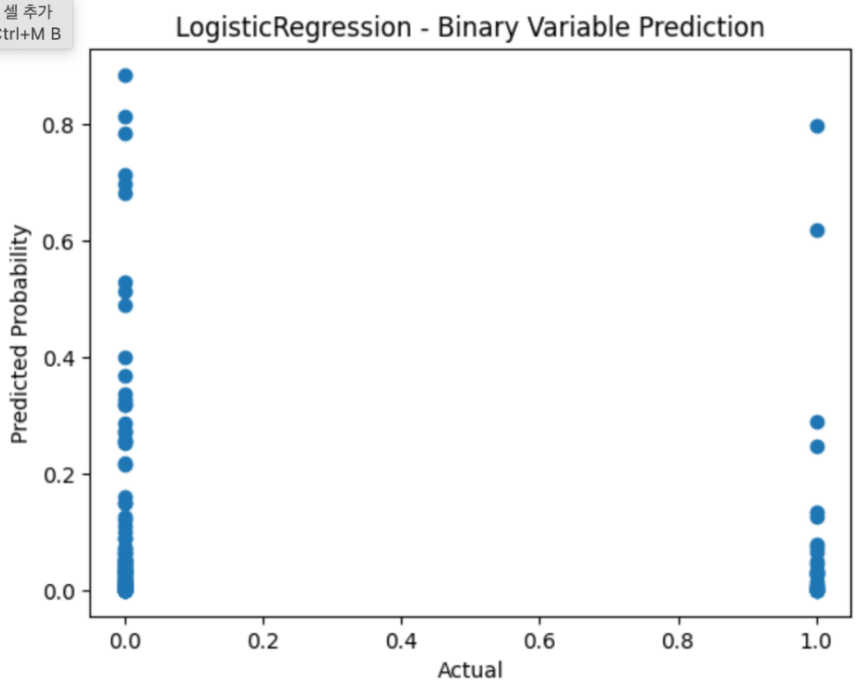
2) 로지스틱 회귀:

이진형 변수에 대한 Accuracy: 0.7867

이진형 변수에 대한 AUC: 0.5273

```
Accuracy (Binary): 0.7866666666666666
AUC (Binary): 0.5272952853598015
```

코드 셀 추가
⌘/Ctrl+M B



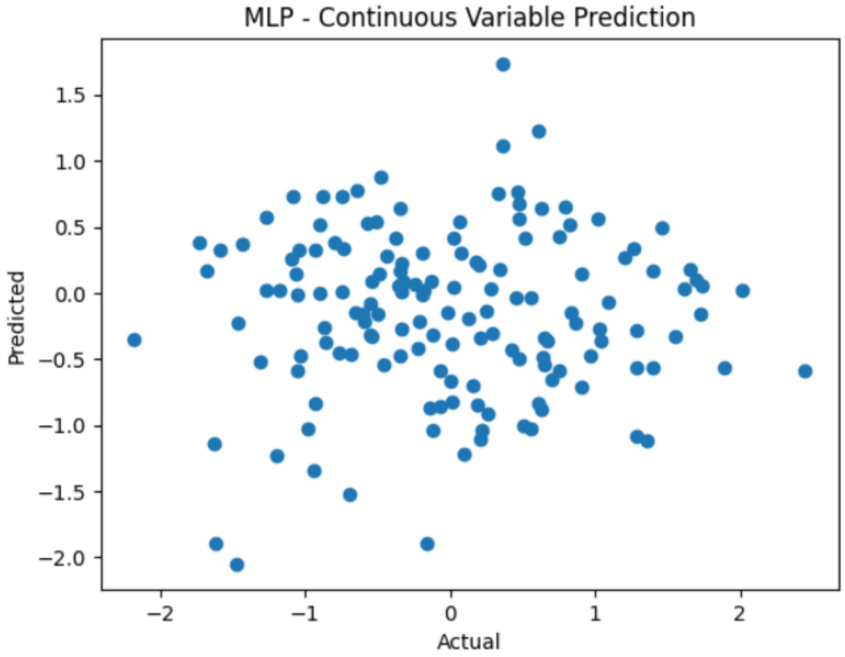
3) MLP:

연속형 변수에 대한 MSE: 1.1812

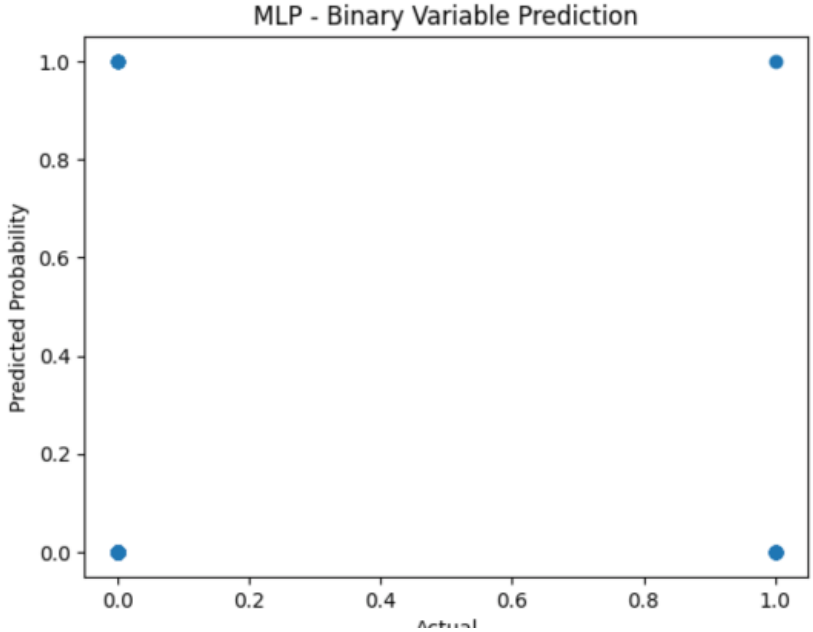
이진형 변수에 대한 Accuracy: 0.7867

이진형 변수에 대한 AUC: 0.4910

```
Epoch 16/100
8/8 [==============================] - 0s 8ms/step - loss: 7.1760e-04 - val_loss: 1.1125
5/5 [==============================] - 0s 4ms/step
Mean Squared Error (MLP - Continuous): 1.1811724758185516
```




```
Epoch 13/100
8/8 [==============================] - 0s 37ms/step - loss: 0.0088 - accuracy: 1.0000
5/5 [==============================] - 0s 7ms/step
Accuracy (MLP - Binary): 0.7866666666666666
AUC (MLP - Binary): 0.4910049627791564
```



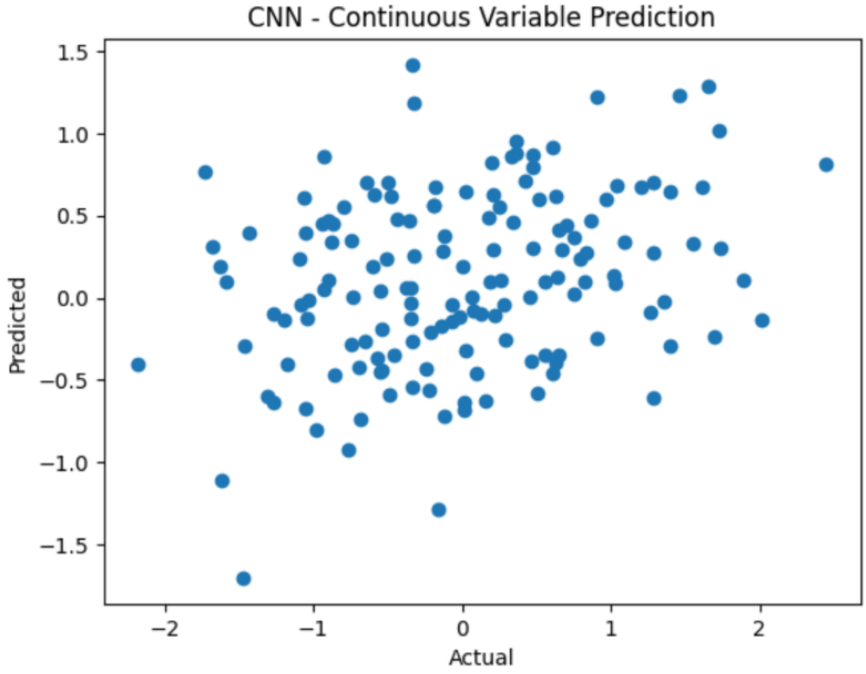
4) CNN:

연속형 변수에 대한 MSE: 0.8255

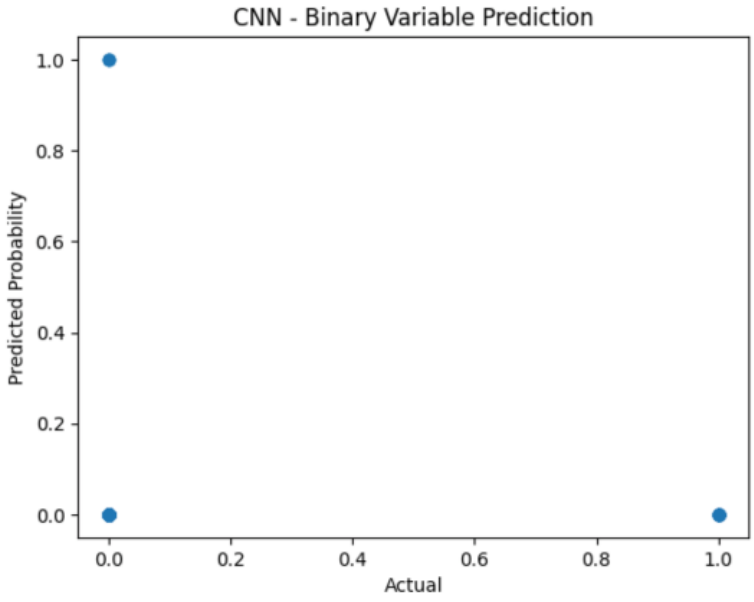
이진형 변수에 대한 Accuracy: 0.8067

이진형 변수에 대한 AUC: 0.4879

```
Epoch 17/100
8/8 [==============================] - 0s 29ms/step - loss: 0.0849 - val_loss: 1.0020
5/5 [==============================] - 0s 7ms/step
Mean Squared Error (CNN - Continuous): 0.8255049505526618
```




```
Epoch 16/100
8/8 [==============================] - 0s 29ms/step - loss: 0.0223 - accuracy: 1.0000 -
5/5 [==============================] - 0s 7ms/step
Accuracy (CNN - Binary): 0.8066666666666666
AUC (CNN - Binary): 0.4879032258064516
```



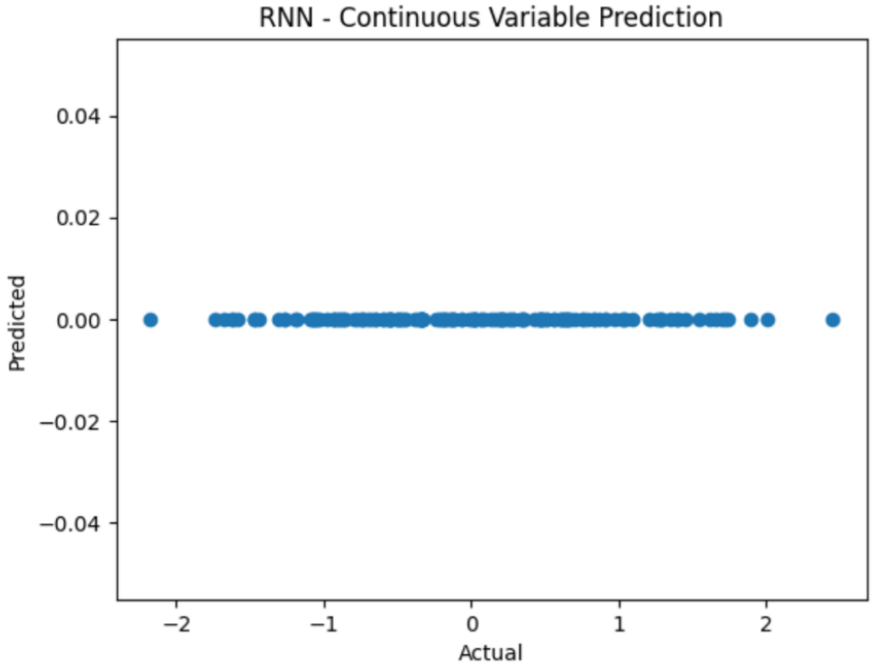
5) RNN:

연속형 변수에 대한 MSE: 0.8109

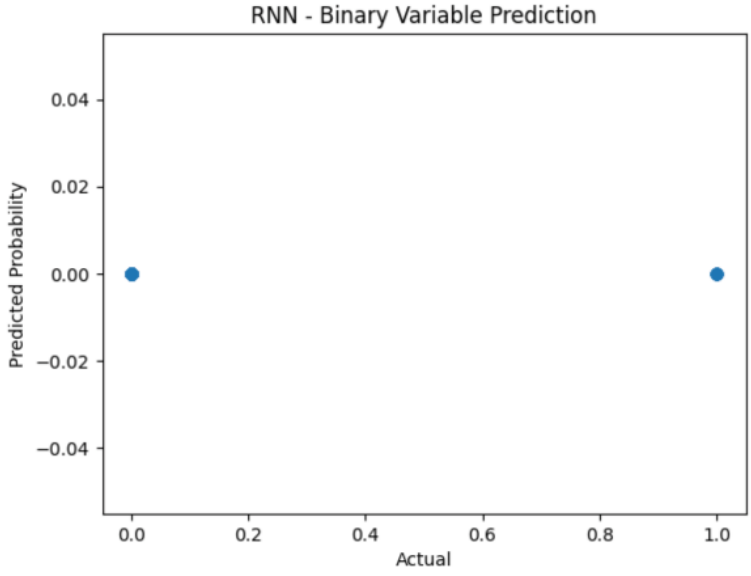
이진형 변수에 대한 Accuracy: 0.8267

이진형 변수에 대한 AUC: 0.5000

```
Epoch 11/100
8/8 [==============================] - 4s 495ms/step - loss: nan - val_loss: nan
5/5 [==============================] - 1s 95ms/step
Number of NaN values in predictions: 150
Mean Squared Error (RNN - Continuous): 0.8109445053228845
```




```
Epoch 12/100
8/8 [==============================] - 3s 367ms/step - loss: nan - accuracy: 0.8000
5/5 [==============================] - 1s 110ms/step
Accuracy (RNN - Binary): 0.8266666666666667
AUC (RNN - Binary): 0.5
```



6. 분석 및 결론

이번 프로젝트에서는 주어진 유전자 데이터를 활용하여 선형 회귀, 로지스틱 회귀, MLP, CNN, 그리고 RNN과 같은 다양한 딥러닝 모델을 적용하고 비교 분석하였다. 이진형 변수와 연속형 변수에 대한 예측 모델을 평가하고 가장 우수한 모델을 식별하는 데 중점을 두었다.

이진형 변수 예측에서 가장 좋은 성능을 보인 모델은 **RNN이**다. RNN의 이진형 변수에 대한 Accuracy는 0.8267로 가장 높았다. 또한, RNN의 이진형 변수에 대한 AUC는 0.5000으로 다른 모델보다 약간 더 높았다. 이러한 결과는 RNN 모델이 시계열 데이터에 대한 학습에 뛰어난 능력을 갖고 있음을 시사한다.

반면에 연속형 변수의 경우 연속형 변수 예측에서는 **CNN** 모델이 가장 좋은 성능을 보였다. CNN의 연속형 변수에 대한 MSE는 0.8255로 가장 낮았다. 이는 모델이 실제 값과 예측 값 간의 평균 제곱 오차가 작다는 것을 의미한다. CNN은 이미지 데이터에 대한 처리에 특화되어 있어서, 유전자 데이터와 같은 다차원적인 데이터에 효과적으로 적용될 수 있음을 시사한다.

따라서, 이러한 분석 결과를 토대로 하여 이진형 변수 예측에는 RNN 모델, 그리고 연속형 변수 예측에는 CNN 모델이 가장 적합하다고 할 수 있다. 그러나 더 정확한 예측을 위해서는 추가적인 데이터 탐색과 모델 튜닝이 필요할 것으로 보인다.