**건강검진 데이터를 활용한 음주 여부 예측**

작성자: [유성오]  
작성일: [2025-04-18]

# 1. 프로젝트 개요

본 프로젝트는 공공데이터 에서 주어진 국민건강 정보 데이터 건강검진 정보를 바탕으로 음주 여부를 예측하는 인공 신경망(ANN) 모델을 개발하는 프로젝트입니다. 목표는 모델을 통해 정확한 음주 예측을 하여 건강 관리에 도움이 되는 예측 도구를 만드는 것입니다. 사용된 데이터는 고객의 체중, 신장, 혈압, 나이 등 여러 변수를 포함하고 있으며, 예측 변수는 '음주 여부'입니다.

# 2. 데이터 탐색 및 전처리

"데이터 탐색을 통해 결측치와 이상치가 존재하는지를 확인하고 적절한 방법으로 처리하였습니다. 결측치는 총콜레스테롤 계산식을 참고하여 대체하였고, 이상치는 실제 있을 법한 데이터로 판단되어 제거하지 않았습니다."

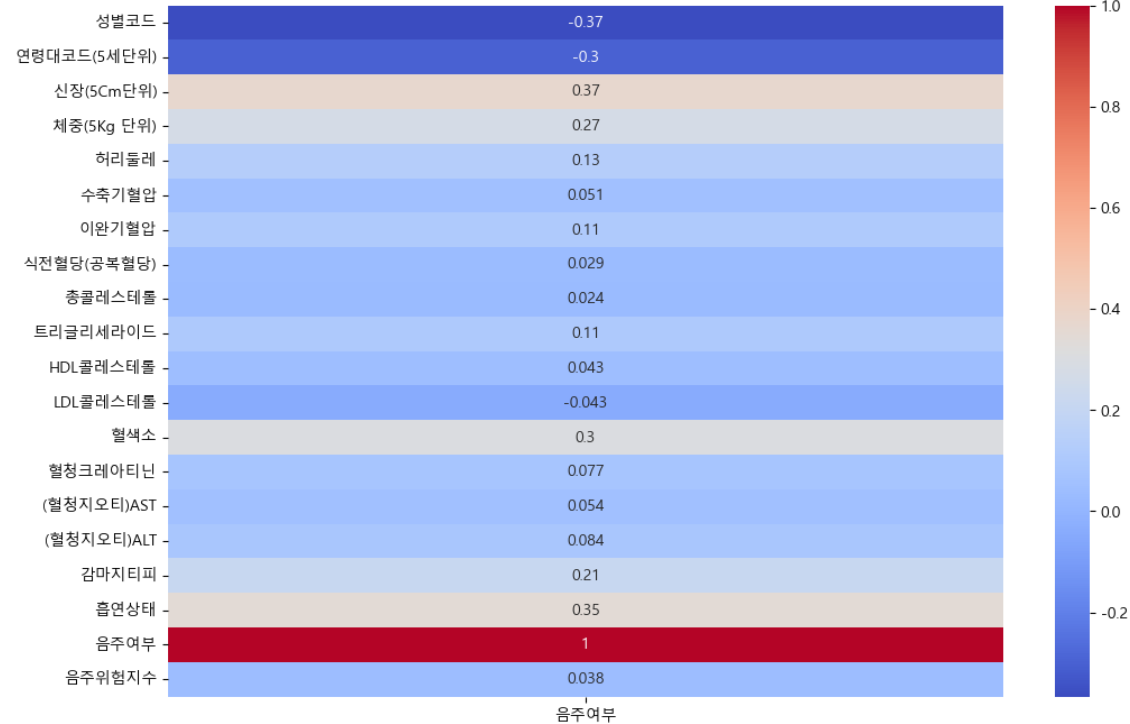
주요 데이터 전처리 과정은 다음과 같습니다:

|  |  |
| --- | --- |
| 전처리 항목 | 처리 방식 |
| 결측치 처리 | 계산식 참고  요단백 -> 상관 분석후 제거 |
| 이상치 처리 | 실제 있을 법한 데이터 제거 X |

# 3. 상관 분석

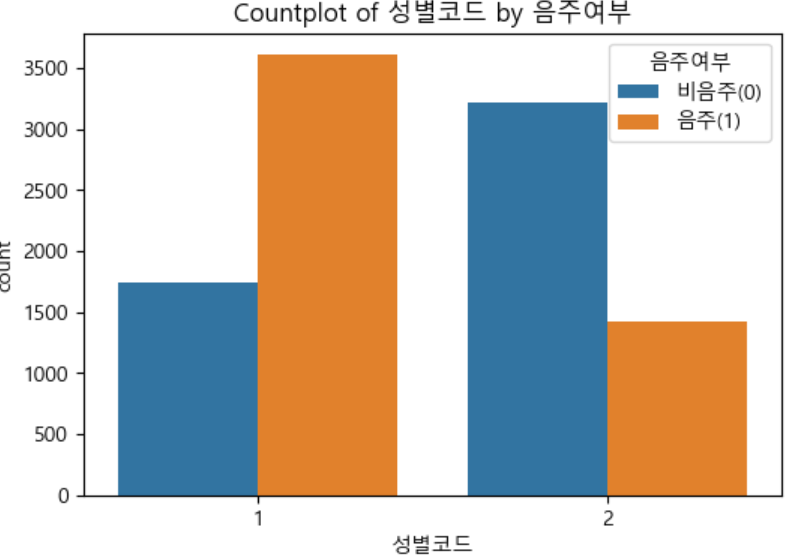
모델링에 들어가기 전에, 데이터 내 변수들 간의 관계를 파악하기 위해 상관 분석을 수행했습니다. 상관 분석은 각 피처들이 목표 변수와 어떻게 연관되어 있는지, 또한 피처들 간의 관계가 어떻게 형성되어 있는지를 이해하는 데 중요한 역할을 합니다. 이를 통해 모델에 포함시킬 변수들을 선정하거나, 다중 공선성 문제를 피하기 위해 상관 관계가 높은 변수들을 제거할 수 있습니다.

본 분석에서는 \*\*피어슨 상관 계수(Pearson correlation coefficient)\*\*를 사용하여 변수들 간의 상관 관계를 측정했습니다. 상관 계수 값은 -1에서 1 사이의 값을 가지며, 값이 1에 가까울수록 두 변수 간에 강한 양의 상관 관계가 있음을 나타내고, -1에 가까울수록 강한 음의 상관 관계가 있음을 의미합니다. 상관 계수가 0에 가까울수록 변수들 간에 관계가 적음을 나타냅니다.

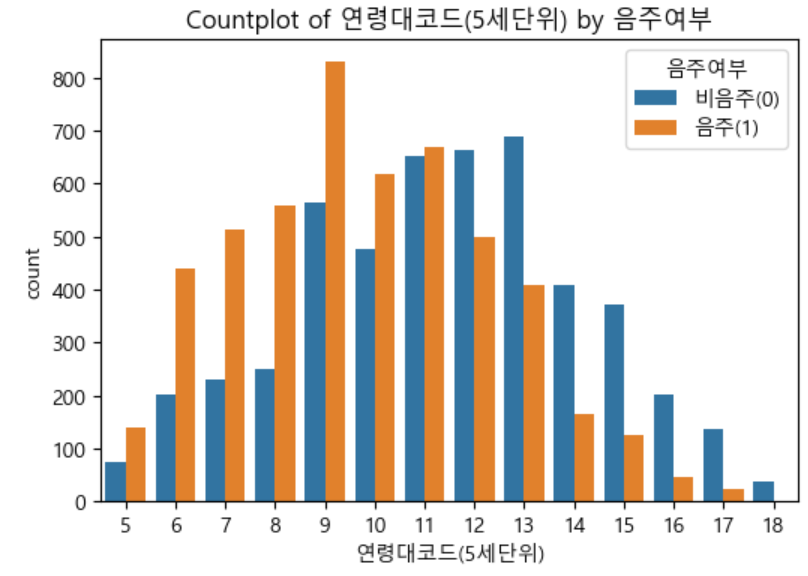


성별,연령,신장,혈색소,흡연상태의 경우에는 음주여부와 상관관계가 있는 것으로 판단이 되며 혈압, 혈당, 콜레스테롤, 혈청 같은 경우에는 음주여부와는 약한관계가 있는 것으로 판단이 됩니다.

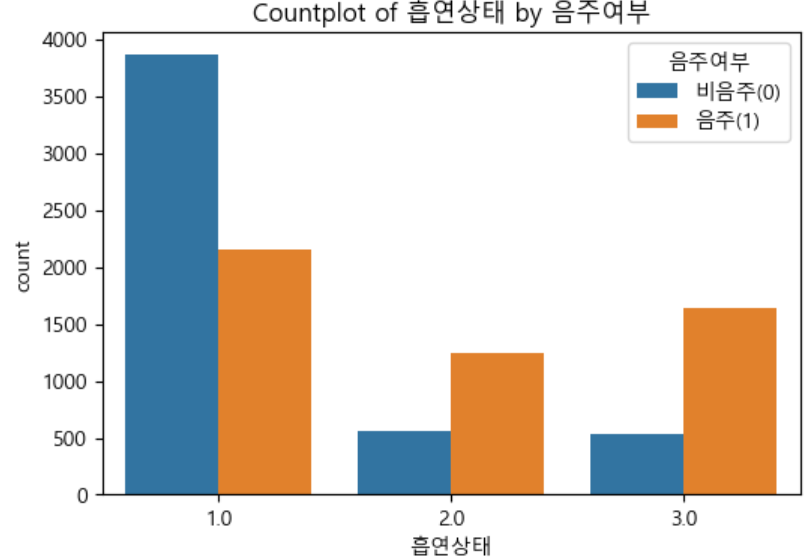
아래의 그래프 처럼 성별의 경우에는 **남성이 여성보다** 음주를 하는 수치가 **높은** 것을 보실 수 있습니다.



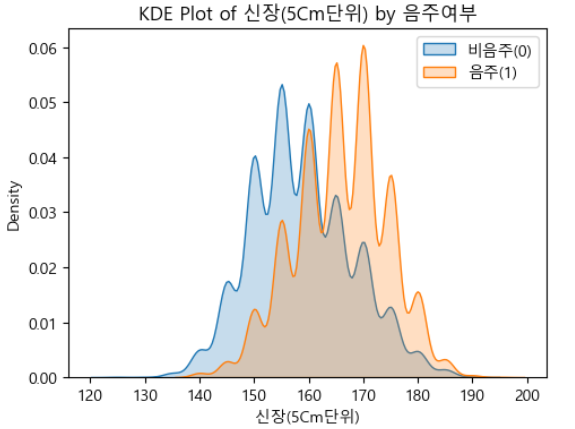
나이의 경우에는 7(30세 ~ 34세) ~ 11(50세 ~ 54세) 단위의 경우가 많으며 직장을 다니는 사람들이 음주를 하는 비율이 높은 것을 볼 수 있습니다.



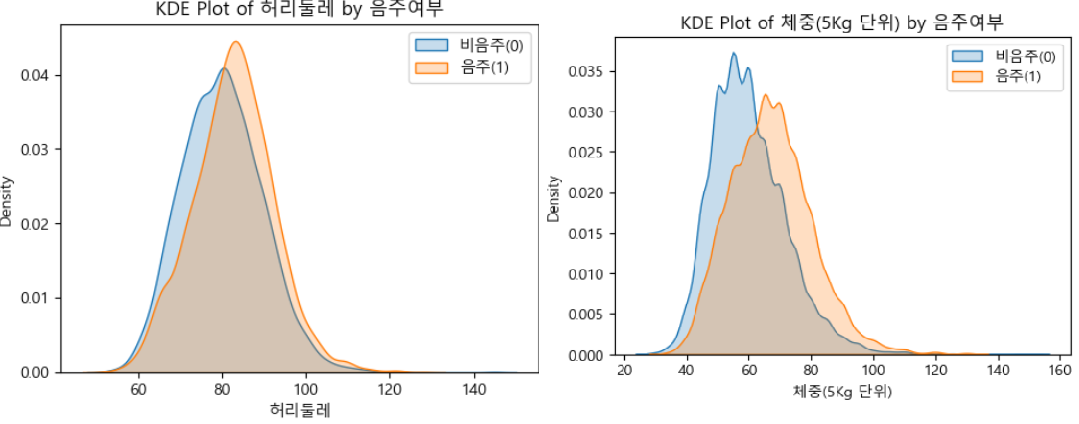
흡연의 경우에도 비흡연(1) 자는 음주를 안 하는 비율이 높으나 이제 금연(2) 자 나 흡연(3) 자의 경우에는 음주를 안 하는 비율보다 음주를 하는 비율이 높은 것을 볼 수 있습니다.



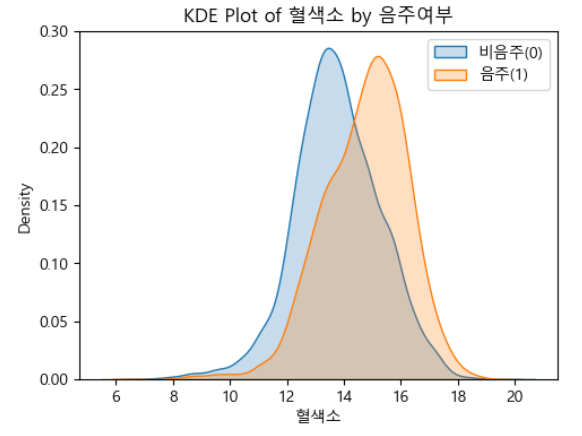
신장의 경우에도 이제 키가 크면 남성의 비율이 많이 분포해 있기 때문에 키가 클수록 음주하는 비율이 높은것을 볼 수 있습니다.



체중과 허리둘레의 경우에는 음주 때문에 체중과 허리둘레가 크다고 할 수는 없으나 음주를 하는 사람이 비음 주인 사람보다는 체중과 허리둘레가 클 것이라는 가정을 했습니다.



일반적으로 혈색소 수치가 높다고 해서 반드시 음주를 많이 하는 사람이라고 단정할 수는 없다. 하지만, 알코올은 혈색소 수치에 직간접적으로 영향을 줄 수 있다. 과도한 음주는 간 기능에 손상을 주어 간 질환을 유발할 수 있으며, 이는 혈색소 수치 변화로 이어질 수 있다. 또한, 음주 후에는 탈수 증상이 나타나 혈액의 농도가 일시적으로 높아져 혈색소 수치가 증가할 수 있다. 이를 바탕으로 음주하는 사람의 경우가 혈색소가 높게 분포 되어 있다.

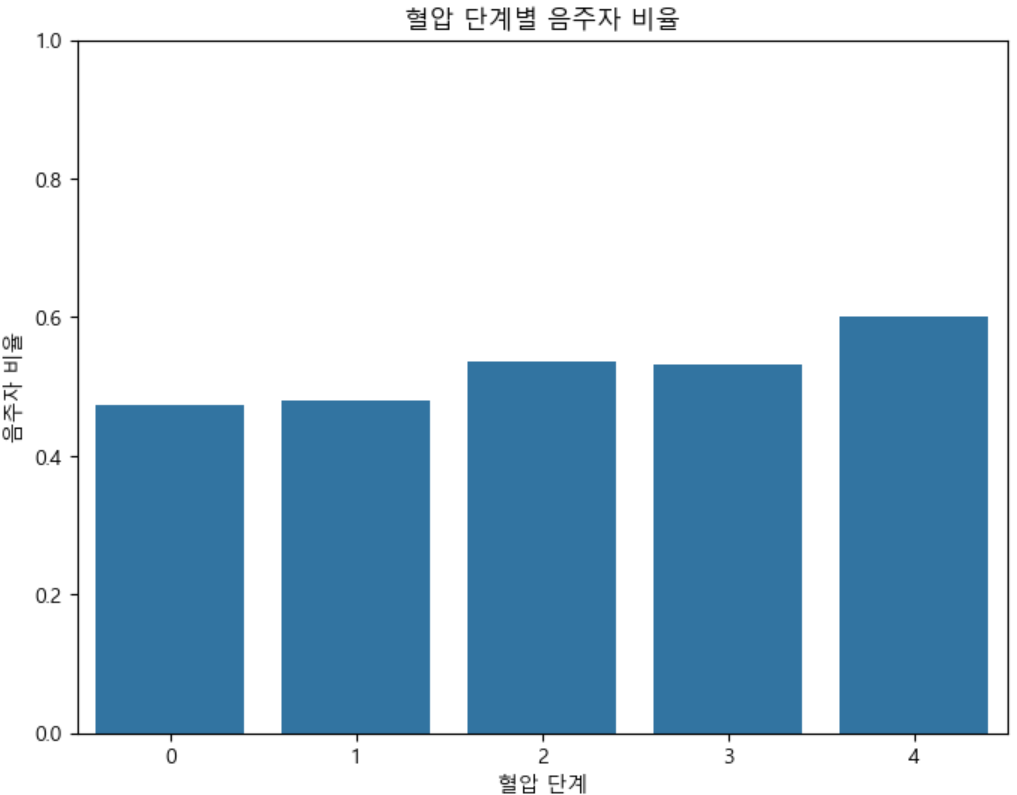


# 4. 파생변수 생성

**1. 혈압단계**

수축기 혈압과 이완기 혈압 데이터를 바탕으로 파생변수 생성

|  |  |  |
| --- | --- | --- |
| 혈압분류 | 수축기혈압(mmHg) | 이완기 혈압(mmHg) |
| 정상혈압 | 120미만 | 그리고 80 미만 |
| 주의혈압 | 120~129 | 그리고 80 미만 |
| 고혈압 전단계 | 130~139 | 또는 80~89 |
| 1기 고혈압 | 140~159 | 또는 90~99 |
| 2기 고혈압 | ≥ 160 | 또는≥ 100 |



음주와 고혈압의 관계는 복잡하나, 음주량, 개인의 건강 상태, 유전적 요인등 다양한 요소에 의해 영향을 받는다. 일반적으로 과도한 음주는 고혈압 발병 위험을 높이는 것으로 알려져 있기에 이렇게 파생변수로 만들었습니다.

**2. 심혈관 위험인자**

심혈관 위험인자 계산식을 참고하여 음주위험지수 파생 변수를 생성했습니다. 음주와 심혈관 건강에 복잡한 영향을 미치며, 섭취량과 음주 습관에 따라 긍정적 또는 부정적인 결과가 나타날 수 있습니다. 일반적으로 적당량의 음주는 일부 심혈관 질환에 긍정적인 영향을 줄 수 있지만, 과도한 음주는 심혈관 질황의 위험을 증가합니다.

# 5. 데이터 전처리 완료 데이터

# 

# 6. 모델 설계 및 하이퍼파라미터 튜닝 전략

모델 설계는 기본적인 인공 신경망(ANN)을 기반으로 하였습니다. 하이퍼파라미터 튜닝을 통해 최적의 성능을 이끌어내기 위해 여러가지 옵티마이저와 은닉층 크기를 실험하였습니다.

모델의 구조는 다음과 같습니다:

입력층: 32개의 뉴런, 활성화 함수 'ReLU'  
은닉층: 8개의 뉴런, 활성화 함수 'ReLU'  
출력층: 1개의 뉴런, 활성화 함수 'Sigmoid' (이진 분류용)

# 7. 하이퍼파라미터 실험 결과

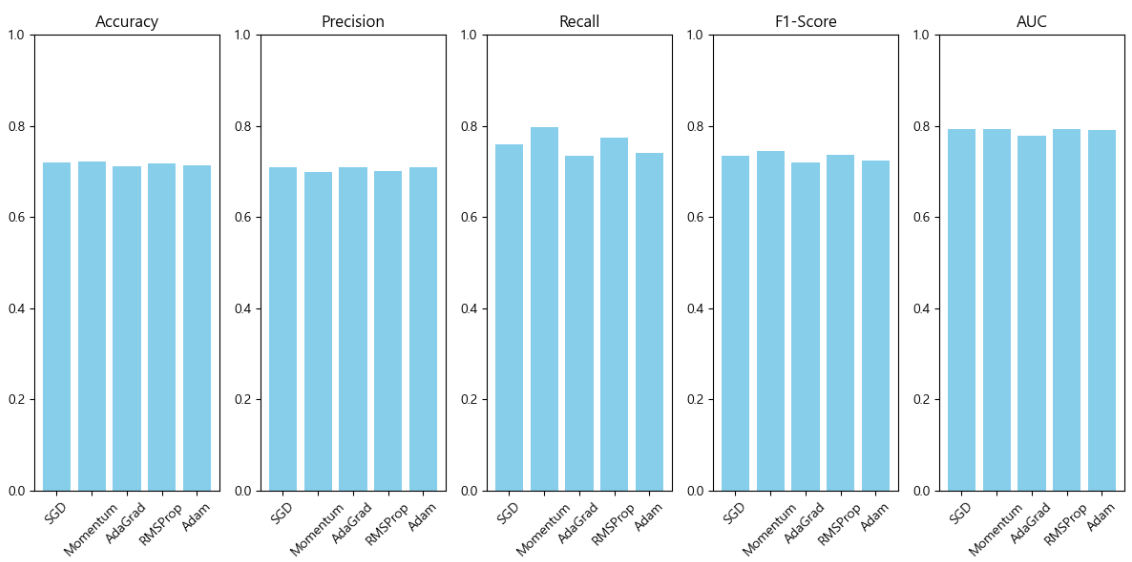
다양한 옵티마이저와 하이퍼파라미터를 실험하여 최적의 조합을 찾았습니다. 각 옵티마이저 및 하이퍼파라미터에 대한 실험 결과를 표로 정리합니다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Optimizer | Precision | Recall | F1-Score | Accuracy | AUC |
| Momentum | 0.6924 | 0.8227 | 0.7519 | 0.7245 | 0.7959 |
| Adam | 0.7075 | 0.7675 | 0.7363 | 0.7210 | 0.7942 |
| SGD | 0.7126 | 0.7645 | 0.7376 | 0.7240 | 0.7941 |
| RMSProp | 0.7037 | 0.7793 | 0.7396 | 0.7215 | 0.7939 |
| AdaGrad | 0.7147 | 0.7034 | 0.7090 | 0.7070 | 0.7796 |

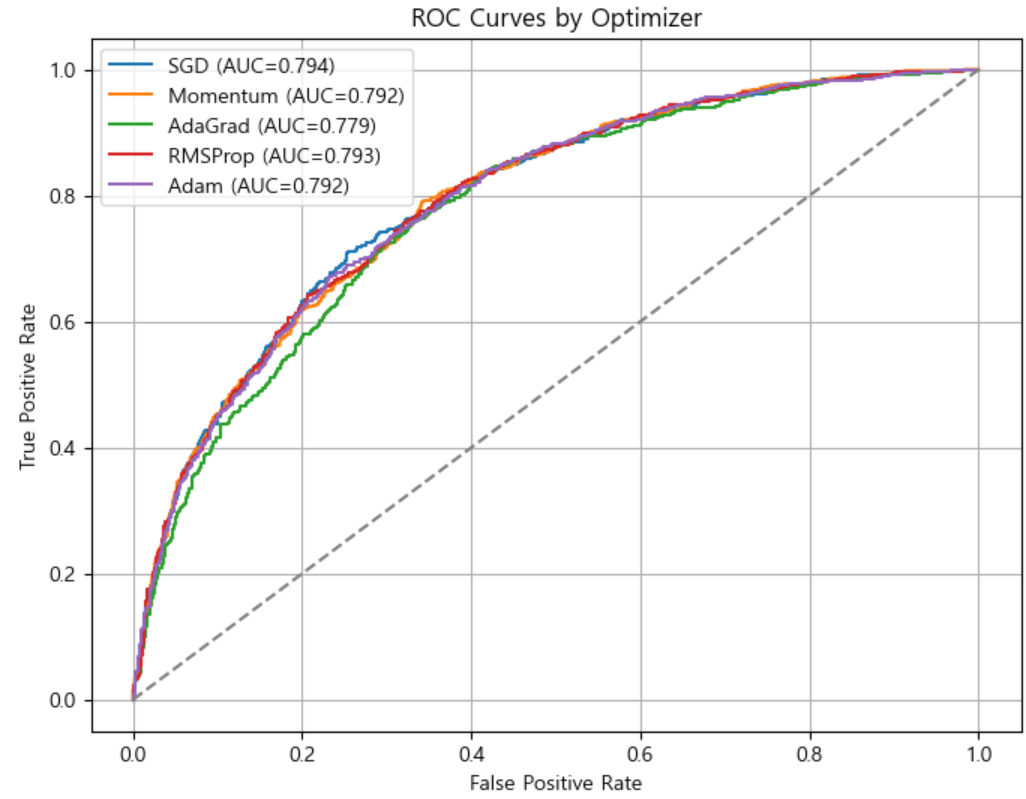
# 8. 예측 성능 비교 및 시각화

모델 성능을 비교하기 위해 ROC Curve, 예측 결과 분포, 학습 과정 추이 등의 시각화를 진행하였습니다.다양한 모델 성능 지표(AUC, 정확도 등)를 비교하여 최적 모델을 선택하였습니다. ROC Curve 및 학습 추이 그래프는 다음과 같습니다:

**8\_1 각 Optimizer 별 성능 비교**



**8-2 ROC Curve**



**8-3 Accuracy/Loss 곡선 그래프**



# 6. 결론 및 개선 방향

본 프로젝트에서는 건강검진 데이터를 바탕으로 음주 여부를 예측하는 이진 분류 모델을 개발하였습니다. 데이터 전처리 과정에서는 결측치를 총콜레스테롤 계산식 등을 참고하여 보완하였고, 혈압을 포함한 파생 변수를 생성함으로써 예측 성능을 향상시키는 데 기여 하였습니다.

여러 최적화 기법을 적용한 인공신경망(ANN) 모델을 구성하고, 다양한 하이퍼파라미터를 실험한 결과, Momentum 옵티마이저를 사용하고 은닉층 구성과 학습 조건을 적절히 설정한 모델이 가장 우수한 성능(AUC 기준)을 보였습니다. 본 모델은 Precision, Recall, F1-score 등 여러 평가지표에서도 비교적 균형 잡힌 성능을 보여주었습니다.

**향후 개선 방향**

- **데이터 다양성 확보 :** 현재 사용한 데이터는 특정 시점의 건강검진 결과에 국한되어 있어, 장기적인 건강 변화나 음주 습관의 변화를 반영하기 어렵습니다. 향후 시계열 데이터나 다양한 연령·지역의 데이터를 추가하여 모델의 일반화를 모도할 수 있습니다.

**- 피처 엔지니어링 고도화**: 본 프로젝트에서는 도메인 지식이 부족한 상태에서 진행되었기 때문에 현재 생성한 파생 변수(혈압 등)에 한정하여 분석을 수행했습니다. 향후에 는 의료 및 건강 데이터에 대한 이해를 바탕으로, 변수 간의 비율, 변화량 등의 고급 파생 변수를 추가하여 모델 성능을 더욱 개선할 수 있습니다.

**- 하이퍼파라미터 최적화:** 현재는 여러 하이퍼파라미터(batch size, epoch, 은닉층 노드 수 등)에 대해 실험은 수행했으나, 최적화 알고리즘(GridSearch RandomSearch, Bayesian Optimization 등) 을 활용한 체계적인 최적화 과정은 진행하지 않았습니다. 향후 하이퍼파라미터 튜닝을 통해 모델의 예측 성능을 극대화할 수 있습니다.

# 7. 사용 도구 및 개발 환경

본 프로젝트에서는 Python과 다음과 같은 주요 라이브러리를 사용하였습니다:

|  |  |
| --- | --- |
| 도구/라이브러리 | 역할 |
| Python | 개발 언어 |
| TensorFlow / Keras | 모델 설계 및 학습 |
| pandas, numpy | 데이터 전처리 및 분석 |
| sklearn | 성능 평가 및 분류 지표 계산 |