ISSN: 3022-5388
JKAIA website: https://accesson.kr/jkaia
DOI: http://doi.org/10.24225/jkaia.2024.2.2.31

심장마비 예측을 위한 머신러닝 알고리즘의 성능 비교 및 주요 변수 분석

A Comparative Study on the Performance of Machine Learning Algorithms and Key Feature Analysis for Predicting Heart Attack

JunSu KOH[1], Min Soo KANG[2], Dong Hoon HAN[3]

Received: November 13, 2024. Revised: December 12, 2024. Accepted: December 14, 2024.

Abstract

In this study, we compare the performance of various machine learning algorithms for predicting heart attacks, a major cause of mortality globally, with a focus on identifying key predictive features. Using a dataset of 918 records, the research evaluates models such as Random Forest, Logistic Regression, XGBoost, SVM, KNN, and Decision Tree to enhance prediction accuracy for heart attack risks. The methodology emphasizes robust preprocessing techniques, including feature scaling and handling class imbalances through Stratified K-Fold cross-validation, to improve model reliability. Results reveal that ensemble models, particularly Random Forest, achieve the highest ROC AUC score of 0.9301, significantly outperforming traditional algorithms. Key predictors, such as ST_Slope, were identified as critical variables in determining heart attack risks, while less influential features, such as RestingECG, had minimal impact. The findings underscore the efficacy of ensemble learning in predicting heart attacks and highlight the importance of feature importance analysis in enhancing model interpretability. This study provides valuable insights into the integration of machine learning in personalized healthcare, offering a foundation for future research to refine predictive models and improve early detection and prevention strategies for cardiovascular diseases.

Keywords: Heart Attack, Machine learning, Decision Tree, Random Forest, Recall, ROC AUC

Major Classification Code: Artificial Intelligence, etc

1. Introduction

심장질환은 우리나라 사망원인 중 암에 이어 두 번째로 높은 비중을 차지하고 있다. 통계청의 2021 년 사망원인 통계에 따르면, 심장질환으로 인한 사망률은 인구 10 만 명당 61.5 명으로 나타났으며, 이는 지속적인 증가 추세를 보이고 있다. 그림 1 은 대한민국의 주요 사망원인 순위 변화를 보여준다. 요구되며, 이를 유지하지 못할 경우 환 자의 건강을 심각하게 위협할 수 있다.

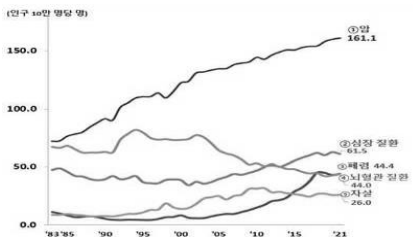

Figure 1: Trend of ranking of major causes of death in 1983-2021

1 Student. Department of Medical IT, Eulji University,Republic of Korea Email: rhwnstj00@naver.com
2 Professor. Department of Medical IT, Eulji University, Republic of Korea, Email: mskang@eulji.ac.kr
3 Corresponding Author. Researcher, MIIC, Republic of Korea, Email: d555v@naver.com

특히 급성 심근경색과 같은 심장마비는 발병 시 즉각적인 치료가 이루어지지 않으면 높은 사망률과 심각한 후유증을 초래한다. 이러한 심장마비의 조기 예측과 예방은 현대 의학의 중요한 과제 중 하나이다. 심장마비의 주요 위험인자로는 고혈압, 당뇨병, 고지혈증, 흡연, 비만 등이 있으며, 이들 위험인 자의 관리와 조기 발견을 통해 심장마비의 발생을 예방할 수 있다. 그러나 생활습관의 변화와 인구 고령화로 인해 이러한 위험 인자의 유병률이 높아지고 있으며, 이에 따라 심장마비 발생 위험도 증가하고 있다. 최근 머신러닝 기술의 발전으로 대규모 의료 데이터를 분석하여 질병 발생을 예측하는 연구가 활발히 진행되고 있다. 이러한 기술을 활용하면 심장마비의 발병 위험을 사전에 예측하고, 적절한 예방 및 치료 전략을 수 립할 수 있다. 따라서 본 연구에서는 다양한 머신러닝 알고리즘을 비교하여 심장마비 발병예측에 가장 적합한 모델을 선정하고자 한다.

2. Literature Review

2.1. Risk Factor for Heart Attack

심장마비의 원인은 주요 위험요인과 기여요인으로 나뉜다. 심장마비의 주요 위험요인으로는 고혈압, 고콜레스테롤, 당뇨 병, 비만 및 과체중, 흡연, 신체 활동 부족, 성별 및 연령, 그리고 유전적 요인이 있다. 이러한 위험 요인들은 심혈관 시스템에 직접적인 영향을 미쳐 동맥경화, 관상동맥 질환, 심근경색 등 다양한 심혈관 질환을 유발한다. 특히, 고혈압과 콜레스테롤은 동맥벽에 플라크를 축적시켜 혈류를 막고, 이는 심장으로 가는 혈류를 제한해 심장마비로 이어질 수 있다. 비만과 당뇨병은 심혈관 질환의 발병 위험을 높이며, 흡연은 혈관을 손상시키고 혈액 내 응고를 증가시켜 심장마비의 가능성을 높인 다. 또한 나이가 들수록 동맥이 경직되고 심장의 기능이 저하되어 심장마비의 위험이 증가하며, 가족력이 있는 경우 심장 질환의 위험이 높아진다. 심장마비에 기여하는 요인으로는 스트레스, 알코올, 성 호르몬 등이 있다. 이러한 기여요인들은 주요 위험 요인과 결합하여 심장마비의 위험을 증가시킨다. 예를 들어, 스트레스는 혈압을 높이고 심장의 산소 요구량을 증가시키며, 알코올은 과도한 섭취 시 심혈관 시스템에 부정적인 영향을 미친다. 또한, 폐경 후 여성은 성 호르몬 변화로 인해 심장마비 위험이 증가할 수 있다.

2.2. Machine Learning Algorithm

머신러닝 문제를 다룰 때, 일반적으로 지도학습 데이터와 비지도 학습 데이터, 두 가지 유형의 데이터(및 머신 러닝 모 델)가 있다. 지도 학습 데이터 (Supervised Data)는 항상 하나 이상의 목표(target)가 연결되어 있고 비지도 학습 데이터 (Un supervised Data)는 목표 변수가 없다. 값을 예측해야 하는 문 제가 지도 학습 문제로 알려져 있다. 예를 들어, 과거 주택 가격을 바탕으로 주택 가격을 예측하는 문제가 있고, 병원, 학교 또는 슈퍼마켓의 존재, 가장 가까운 대중교통까지의 거리와 같은 특징(features)이 주어졌다면 이는 지도 학습 문제이다. 마 찬가지로, 고양이와 개의 이미지가 제공되고, 어떤 이미지가 고양이인지 개인지 사전에 알고 있으며, 주어진 이미지가 고양 이인지 개인지 예측하는 모델을 만드는 작업도 지도 학습 문제 로 간주된다. 이 데이터셋에서는 심부전 예측을 위한 지도 학 습 문제를 다루고 있다.

2.2.1. Random Forest

각 트리는 무작위로 선택된 데이터 샘플과 특성을 기반으로 학습되며, 최종 예측은 각 트리의 예측 결과를 평균내거나 다 수결 방식으로 종합하여 결정된다. 랜덤포레스트는 과적합(ov erfitting)을 방지하는데 효과적이며, 변수의 중요도(feature importance)를 계산할 수 있어 의료 데이터와 같은 고차원 데이터에서 널리 사용된다. 이러한 특성 덕분에 복잡한 변수 간의 관계를 효과적으로 처리할 수 있다.

2.2.2. Logistic Regression

로지스틱 회귀는 이진 분류 문제를 해결하기 위한 통계적 기법으로, 종속 변수와 독립 변수 간의 관계를 선형 회귀 모형으로 모델링하지만, 출력값이 0 과 1 사이의 확률로 제한되는 것 이 특징이다. 로지스틱 회귀는 예측된 확률값에 로지스틱 함수 (logistic function)를 적용하여 이진 분류를 수행한다. 의료 데이터에서 질병 유무와 같은 이진 결과를 예측할 때 자주 사용 되며, 독립 변수의 가중치 해석이 용이해 질병 위험 요인을 직 관적으로 이해할 수 있는 장점이 있다.

2.2.3. SVM, Support Vector Machine

서포트벡터머신(SVM)은 Vapnik 과 Cortes(1995)에 의해 개발된 분류 알고리즘으로, 데이터를 분류하는 초평면을 찾는 데 중점을 둔다. SVM 은 고차원 공간에서 데이터 포인트들 을 선형적으로 분리하기 위한 최적의 경계를 찾으며, 선형적으 로 분리되지 않는 데이터에 대해서는 커널 함수(kernel functio n)를 사용하여 비선형 문제를 해결할 수 있다. SVM 은 소규모 데이터 세트에서도 높은 성능을 발휘하며, 특히 의료 분야에서 종양이나 질병 진단과 같은 이진 분류 문제에 효과적으로 사용 된다.

2.2.4. KNN, K-Nearest Neighbors

K-최근접이웃은 비모수적 학습방법으로, 새로운 데이터 포인트를 분류할 때 가장 가까 운 K 개의 이웃 데이터 포인트의 레이블을 참고하여 결과를 예측하는 알고리즘이다. KNN 은 학습 과정이 필요 없다는 장점이 있지만, 예측 과정에서 모든 데이터 포인트 간의 거리를 계산해야 하므로 대규모 데이터 세트에서는 계산 비용이 증가할 수 있다. 그러나 소규모 데이터 세트에서는 비교적 간단하고 직관 적인 방법으로 예측을 수행할 수 있어 의료 데이터에서 유용하게 사용될 수 있다.

2.2.5. Decision Tree

결정 트리는 분류와 회귀 문제를 해결하는 데 사용되는 트리 기반 모델로, 각 내부 노드는 특성에 대한 조건을 나타내고, 조 건에 따라 데이터를 분기하여 예측을 수행한다. 결정 트리는 직관적이고 해석이 쉬워 의료 데이터에서 변수 간의 관계를 시각화 하는데 유용하다. 그러나 트리의 깊이가 깊어질수록 과 적합 될 위험이 있으며, 작은 변화에도 민감할 수 있다는 단점 이 있다. 그럼에도 불구하고, 데이터의 전처리과정이 적고 범주형, 연속형 데이터를 모두 처리할 수 있어 다양한 의료 예측 모델에서 널리 사용된다.

2.2.6. Gaussian Naive Bayes

Gaussian Naive Bayes (GNB)는 통계적 방법에 기반한 분류 알고리즘으로, 각 특성이 서로 독립임을 가정하는 나이브베이즈 모델의 한 변형이다. 가우시안 분포를 사용하여 연속형 데이터를 처리하며, 이를 통해 특성의 분포가 가우시안 형태를 따를 경우 효율적인 성능을 보여준다. GNB 는 계산이 매우 빠르고, 고차원 데이터에서 좋은 성능을 발휘할 수 있어 특히 실시간 분류 작업에 유리하다.

2.2.7. XGBoost

XGBoost 는 데이터 과학과 머신 러닝에서 주로 사용되는 트리 부스팅 알고리즘으로, 그레이디언트 부스팅 기술을 기반으로 한다. 이를 통해 결정 트리를 순차적으로 조합하여 강력한 예측 모델을 형성하며, 규제 기법을 활용하여 모델의 복잡성을 줄이고 과적합을 방지한다. XGBoost 는 각 특징의 중요도를 계 산하여 데이터의 해석을 용이하게 하고, 효과적인 병렬 처리를 통해 대용량 및 고차원 데이터 셋에서 뛰어난 성능을 보장한다.

2.3. Model Evaluation Indicators

모델 평가 지표는 머신러닝 모델의 성능을 정량적으로 평가하는 데 사용되는 핵심 요소로, 모델의 예측이 실제 결과와 얼 마나 일치하는지에 대한 정보를 제공한다. 특히, 이진 분류 문 제에서는 여러 평가 지표를 종합적으로 활용하여 모델의 전반 적인 성능을 평가하고, 다양한 상황에서의 효율성을 파악하는 것이 중요하다. 본 논문에서는 분류 모델의 성능 평가를 위해 주요 지표로 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-스코어(F1-Score), ROC AUC(Receiver Operating Characteristic Area Under the Curve)를 사용하였다. 정확도는 가장 직관적인 성능 지표 중 하나이지만, 클래스의 불균형이 심한 데이터셋에서는 모델의 성능을 왜곡할 수 있다. 예를 들어, 전체 데이터 중 95%가 음성 클래스인 경우, 모든 샘플을 음성으로 예측하는 모델이 95%의 정확도를 얻을 수 있 지만 이는 유용한 모델이라고 할 수 없다. 따라서 정확도는 다른 지표와 함께 사용하여 모델의 성능을 종합적으로 평가해야 한다.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Samples} \quad (1)$$

정밀도는 거짓 양성(False Positive)의 수가 중요한 상황에서 유용한 지표다. 예를 들어, 의학 진단 모델에서 정밀도는 모델이 질병을 양성으로 예측했을 때 실제로 질병이 존재할 확률을 나타내므로, 오진으로 인한 불필요한 치료나 검사를 피할 수 있다.

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

재현율은 거짓 음성(False Negative)의 수가 중요한 상황에서 특히 중요하다. 예를 들어, 질병 진단 모델에서 재현율이 높을수록 실제로 질병이 있는 사람들을 더 많이 찾아낼 수 있기 때문에, 초기 선별 검사와 같은 응용 분야에서 높은 재현율이 요구된다

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

F1-스코어는 정밀도와 재현율 사이의 균형을 중요시하는 경우 유용하다. 특히 클래스의 불균형이 심하거나 거짓 양성과 거짓 음성의 비용이 다를 때 유용한 평가 지표다.

$$F1 - Score = \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

ROC 곡선은 모델의 참 양성 비율(True Positive Rate)과 거짓 양성 비율(False Positive Rate) 간의 관계를 시각화한 곡선이다. 이 곡선의 아래 면적을 AUC(Area Under the Curve)라고 하며, 모델의 분류 성능을 나타낸다. AUC 값은 0 과 1 사이에 위치하며, 1 에 가까울수록 모델의 분류 성능이 우수함을 의미한다.

ROC AUC 는 특히 클래스의 불균형이 있는 경우에도 모델의 성능을 평가하는 데 효과적이다. AUC 값이 높을수록 모델이 양성 샘플과 음성 샘플을 잘 구분할 수 있음을 나타낸다. 이는 다양한 임곗값(threshold)에서 모델의 성능을 한 번에 평가할 수 있는 이점을 제공한다.

3. Research Design

3.1. Research Environment

본 연구는 고성능 컴퓨터 하드웨어와 Python 프로그래밍 언어를 기반으로 다양한 데이터 분석 및 머신러닝 라이브러리를 사용하여 수행되었다. 실험의 재현성을 확보하기 위해 사용된 하드웨어 및 소프트웨어 환경을 구체적으로 기술한다.

3.1.1. Hardware Environment

Table 1: 학습에 사용된 장비 성능

Type	Content
CPU	Intel i7-8700K 3.70 GHz
GPU	NVIDIA GeForce GTX 1660 SUPER
Memory	16GB

OS	Windows 11 Pro

3.2. Analysis of Research Data

본 연구에서는 심부전 예측을 위한 데이터셋을 이용하였다. 이 데이터셋은 심혈관 질환 예측에 사용될 수 있는 다양한 환자 특성 및 의료 정보를 포함하고 있으며, 총 11 개의 속성과 918 개의 관측치를 포함하고 있다. 데이터는 심장병을 앓고 있는 환자와 정상인 환자의 정보를 제공하며, 이러한 정보를 바탕으로 심부전 예측 모델을 구축하고 평가했다.

3.2.1. Configuring Datasets

데이터셋은 총 918 개의 샘플로 구성되어 있으며, 데이터는 다음 그림 2 와 같이 11 개의 속성으로 이루어져 있다.

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0
37	F	NAP	130	211	0	Normal	142	N	0	Up	0
58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M	ATA	120	204	0	Normal	145	N	0	Up	0
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
42	F	NAP	115	211	0	ST	137	N	0	Up	0
54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0
38	M	ASY	110	196	0	Normal	166	N	0	Flat	1
43	F	ATA	120	201	0	Normal	165	N	0	Up	0
60	M	ASY	100	248	0	Normal	125	N	1	Flat	1

Figure 2: DataSet

3.2.2. Dataset Source

이 데이터셋은 5 개의 서로 다른 심장 질환 데이터셋을 결합하여 만들어졌다. 원본 데이터셋은 Cleveland, Hungarian, Switzerland, Long Beach VA, Stalog(Heart) 데이터셋이며, 각각 303, 294, 123, 200, 270 개의 관측치를 포함한다. 총 1190 개의 관측치 중 272 개의 중복 데이터가 제거되었고, 최종적으로 918 개의 관측치가 남았다. 이 데이터는 UCI 머신러닝 저장소에서 제공된 데이터셋을 기반으로 하였다. 각각의 데이터셋에서 공통적으로 나타나는 11 개의 특성만을 사용하였으며, 데이터 통합 과정에서 중복 및 결측 데이터는 처리되었다. 특히 중복된 관측치는 제거되었고, 결측값이 존재하는 경우 해당 특성을 제거하거나 평균값으로 대체하는 방식을 사용하였다.

3.2.3. Data Correlation Matrix Analysis

본 연구에서는 변수 간의 관계를 분석하고, 다중공선성 문제를 피하기 위해 상관 행렬을 작성하였다. 상관 행렬을 통해 각 변수 간의 상관 계수를 시각화함으로써, 특정 변수들이 높은 상관관계를 가지는지 확인하였다. 일반적으로 상관 계수가 0.8 이상일 경우 다중공선성 문제를 일으킬 가능성이 있어 변수 제거를 고려해야 한다. 이번 상관 행렬 분석 결과, 변수들 간의 상관 계수가 0.8 을 초과하는 경우는 발견되지 않았으며, 다중공선성 문제가 발생할 우려가 없는 것으로 확인되었다. 특히, "Heart Disease"는 "MaxHR"과 음의 상관관계를, "Oldpeak" 및 "RestingBP"와 양의 상관관계를 가지지만, 이러한 관계는 다중공선성을 초래할 정도로 강하지 않았다. 따라서, 본 연구에서는 모든 변수를 유지하는 방향으로 분석을 진행하였으며, 변수 제거는 필요하지 않다고 판단하였다. 그림 3 은 데이터의 상관행렬을 보여준다.

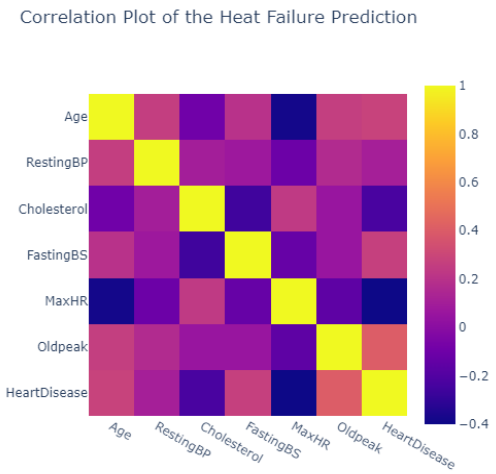

Figure 3: Correlation Matrix of Variables for Heart Failure Prediction

3.2.4. Data preprocessing and cross-validation

본 연구의 데이터 전처리 과정에서는 모델 학습의 안정성 및 성능 향상을 목표로 Standard Scaling 을 선택하여 모든 모델에 적용하였다. Standard Scaling 은 각 피처를 평균 0, 표준편차 1 로 변환하는 방식으로, 데이터의 스케일 차이를 줄여 모델 학습이 특정 피처에 치우치지 않도록 돕는다. 이는 특히 비트리 기반 모델(예: SVM, KNN, 로지스틱 회귀)에서 중요한 역할을 한다. 이들 모델은 피처 간의 상대적인 크기를 고려하여 학습하기 때문에, 스케일 차이가 크면 학습에 부정적인 영향을 미칠 수 있다. 따라서 Standard Scaling 을 통해 각 피처의 중요도가 고르게 반영될 수 있도록 하였다. 반면, 트리 기반 모델(Decision Tree, Random Forest, XGBoost)은 스케일링의 영향을 크게 받지 않지만, 모델 간 공정한 비교를 위해 모든 모델에 일관되게 적용하였다. 트리 기반 모델은 데이터의 분포나 범위에 관계없이 각 노드에서 최적의 분할을 찾기 때문에 스케일링이 직접적인 성능 향상에 기여하지 않을 수 있다. 그러나 일관된 데이터 전처리는 성능 비교 시 스케일 차이로 인한 영향을 최소화하여 보다 객관적인 평가를 가능하게 하였다. 그림 11 은 데이터의 스케일링 전후 분포 변화를 나타낸 것으로, 다양한 스케일링 기법이 적용된 후 각 피처가 어떻게 변형되는지를 비교한 그래프이다. 그래프는 왼쪽부터 순서대로 스케일링 전, Robust Scaling, Standard Scaling, Min-Max Scaling 이 적용된 후의 데이터 분포를 보여준다. 또한, 범주형 변수에 대해서는 Label Encoding 과 One-Hot Encoding 을 적절히 활용하였다. Label Encoding 은 트리 기반 모델에서 유용하게 사용되며, 각 범주형 변수를 고유한 정수로

변환하여 메모리 사용 효율성을 높였다. 반면, One-Hot Encoding 은 비트리 기반 모델에서 범주 간 순서나 관계를 고려하지 않도록 이진 벡터로 변환하여 모델의 학습을 돕는 방식이었다. 모델 평가 과정에서는 Stratified K-Fold 교차 검증을 사용하여 클래스 불균형 문제를 완화하고, 각 폴드에 클래스 비율이 균등하게 분포되도록 하여 성능 평가의 왜곡을 방지했다. 이 방식은 모델이 다양한 클래스 분포에 대해 안정적으로 평가될 수 있도록 하였고, 성능의 변동성을 최소화할 수 있었다. 결론적으로, Standard Scaling 의 적용은 모델 간 공정한 성능 비교를 가능하게 하고, 트리 기반 모델을 포함한 다양한 모델에서 일관된 성능 평가를 도출할 수 있도록 하였다. 또한, 범주형 변수 처리와 Stratified K-Fold 교차 검증을 통해 데이터의 분포 특성을 잘 반영한 모델 성능 평가가 가능하게 되었으며, 연구 결과의 신뢰성을 높이는 데 기여하였다.

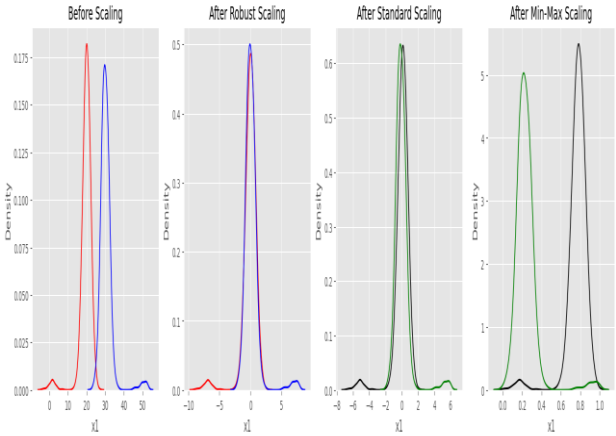

Figure 4: Comparison of Scaling Techniques

4. Research Results

본 연구에서는 각 모델에 대해 최적의 성능을 도출하기 위해 GridSearchCV 를 사용하여 하이퍼파라미터 튜닝을 진행하였다. 각 모델의 특성에 맞는 하이퍼파라미터를 설정하여 최적화하였으며, 이 과정에서 사용된 주요 하이퍼파라미터는 모델마다 달랐다. 이러한 하이퍼파라미터 최적화를 통해 각 모델의 성능을 극대화하고, 공정한 비교가 가능하도록 하였다. 첫 번째 실험은 로지스틱 회귀(Logistic Regression) 모델로 5 개의 폴드 결과 값을 평균내어 정확도: 0.8627 정밀도(precision): 0.8641 재현율(recall): 0.8627 F1 스코어: 0.8623 ROC AUC 스코어: 0.9262 을 얻었다. 두 번째 실험은 Gaussian Naive Bayes 모델로 5 개의 폴드 결과 값을 평균내어 정확도: 0.8617 정밀도(precision): 0.8632 재현율(recall): 0.8617 F1 스코어: 0.8524 ROC AUC 스코어: 0.9225 를 얻었다. 그림 8 은 Gaussian Naive Bayes 모델을 시각화한 것이다.

세 번째 실험은 선형 커널을 적용한 서포트 벡터 머신(Support Vector Machine, SVM)을 사용하여 5 개의 폴드 결과 값을 평균내어 정확도: 0.8595 정밀도(precision): 0.8638 재현율(recall): 0.8595 F1 스코어: 0.8580 ROC AUC 스코어: 0.9250 다. 네 번째 실험은 K-최근접 이웃(K-Nearest Neighbors, KNN)을 사용하여 5 개의 폴드 결과값을 평균내어 정확도: 0.8671 정밀도(precision): 0.8683 재현율(recall): 0.8671 F1 스코어: 0.8668 ROC AUC 스코어: 0.9294 를 얻었다. 다섯 번째 실험은 결정 트리(Decision Tree) 모델을 사용하여 5 개의 폴드 결과 값을 평균내어 정확도: 0.8148 정밀도(precision): 0.8163 재현율(recall): 0.8148 F1 스코어: 0.8147 ROC AUC 스코어: 0.8772 을 얻었다. 그림 9 는 Decision Tree 모델을 시각화한 것이다.

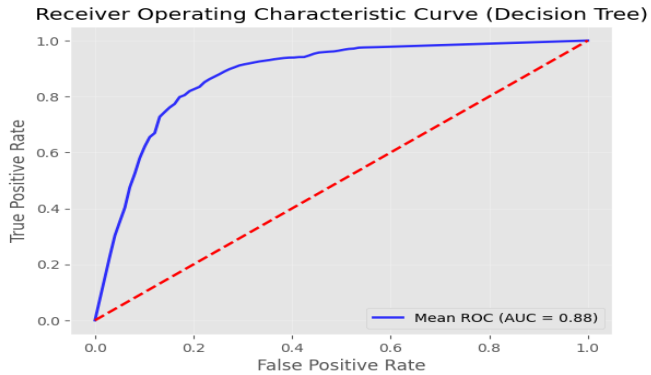

Figure 5: Decision Tree

여섯 번째 실험은 랜덤 포레스트(Random Forest) 모델을 사용하여 5 개의 폴드 결과 값을 평균내어 정확도: 0.8660 정밀도(precision): 0.8678 재현율(recall): 0.8660 F1 스코어: 0.8654 ROC AUC 스코어: 0.9301 을 얻었다. 그림 7 은 Random Forest 모델을 시각화한 것이다.

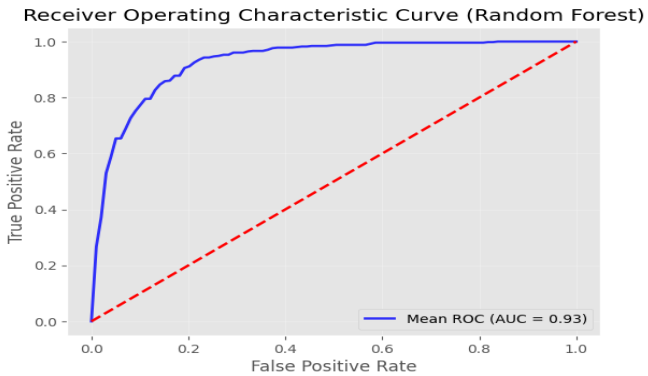

Figure 6: Random Forest

일곱 번째 실험은 XGBoost 모델을 사용하여 5 개의 폴드 결과 값을 평균내어 정확도: 0.8725 정밀도(precision): 0.8744 재현율(recall): 0.8275 F1 스코어: 0.8718 ROC AUC 스코어: 0.9293 을 얻었다. 그림 11 은 모델 성능 비교 요약이다.

모델	정확도	정밀도 (Precision)	재현율 (Recall)	F1 스코어	ROC AUC 스코어
Logistic Regression	0.8627	0.8641	0.8627	0.8623	0.9262
Gaussian Naive Bayes	0.8617	0.8632	0.8617	0.8524	0.9225
SVM (선형 커널)	0.8595	0.8638	0.8595	0.8580	0.9250
K-최근접 이웃 (KNN)	0.8671	0.8683	0.8671	0.8668	0.9294
결정 트리 (Decision Tree)	0.8148	0.8163	0.8148	0.8147	0.8772
랜덤 포레스트 (Random Forest)	0.8660	0.8678	0.8660	0.8654	0.9301
XGBoost	0.8725	0.8744	0.8275	0.8718	0.9293

Figure 7: Model Performance Comparison Summary

또한, 본 연구에서는 각 피처의 중요도를 평가하여 모델의 성능에 미치는 영향을 분석하였다. 랜덤 포레스트 모델을 사용한 피처 중요도 분석 결과, 'ST_Slope'가 가장 중요한 피처로 도출되었으며, 그 중요도 값은 0.264 로 나타났다. 반면, 'RestingECG'는 중요도가 0.021 로 가장 낮은 피처로 평가되었다. 이를 바탕으로 주요 피처의 제거가 모델 성능에 미치는 영향을 분석하기 위해, 가장 중요한 피처(ST_Slope)와 가장 중요도가 낮은 피처(RestingECG)를 각각 제거하여 추가 실험을 수행하였다. ST_Slope 를 제거한 결과, 모델의 ROC AUC 스코어는 0.918 로 눈에 띄게 하락하였다. 이는 ST_Slope 가 모델의 성능에 중요한 영향을 미치는 피처임을 나타낸다. 반면, RestingECG 를 제거했을 때는 모델의 ROC AUC 스코어가 0.928 로 소폭 하락하여, 모델 성능에 거의 영향을 미치지 않는 것으로 나타났다. 이러한 결과는 ST_Slope 가 모델 성능에 중요한 기여를 하는 반면, RestingECG 는 상대적으로 영향력이 낮음을 시사한다.

5. Conclusion

본 연구는 심장마비 발병 예측에 가장 적합한 머신러닝 알고리즘을 선정하기 위해 로지스틱 회귀, 가우시안 나이브 베이즈, 서포트 벡터 머신(SVM), K-최근접 이웃(KNN), 결정 트리, 랜덤 포레스트, XGBoost 등 총 7 가지 모델을 비교하였다. 각 모델의 성능은 정확도, 정밀도, 재현율, ROC AUC 등 다양한 평가 지표를 통해 종합적으로 평가되었다. 실험 결과, 랜덤 포레스트 모델이 가장 높은 ROC AUC 스코어인 0.9301 을 기록하며 심장마비 예측에서 가장 우수한 성능을 보였다. 이 모델은 정확도(0.8660), 정밀도(0.8678), 재현율(0.8660)에서도 뛰어난 성과를 나타냈다. 반면, 결정 트리 모델은 ROC AUC 스코어가 0.8772 로 가장 낮았으며, 이는 단일 트리 구조로 인해 과적합의 영향을 받기 쉽고, 복잡한 데이터 구조를 처리하는 데 한계가 있음을 보여준다. 랜덤 포레스트 모델을 사용한 피처 중요도 분석 결과, 'ST_Slope'가 가장 중요한 변수로 평가되었으며, 이 변수는 심장 기능에 대한 중요한 정보를 제공하는 것으로 나타났다. 반면, 'RestingECG'는 중요도가 낮아 모델 성능에 거의 영향을 미치지 않는 것으로 평가되었다. ST_Slope 를 제거한 추가 실험에서는 모델의 ROC AUC 스코어가 0.918 로 감소하여 이 변수가 모델의 예측 성능에 중요한 기여를 하고 있음을 확인하였다. 반면, RestingECG 를 제거했을 때는 성능에 큰 변화가 없었다. 이러한 분석 결과는 모델을 단순화하고 해석 가능성을 높이는 데 유용한 정보를 제공한다. KNN 모델은 재현율 측면에서 우수한 성과를 보였으며, 환자를 놓치지 않고 예측하는 데 강점을 나타냈다. 심장마비와 같은 응급 상황에서 재현율이 중요한 의미를 가지므로, KNN 모델의 높은 재현율은 임상적 의의가 크다. 대부분의 모델에서 ROC AUC 값이 0.9 이상으로 나타나, 다양한 머신러닝 알고리즘이 심장마비 예측에 효과적으로 활용될 수 있음을 확인할 수 있었다. 본 연구에서는 정확도뿐만 아니라 재현율을 주요 평가 지표로 사용하여, 심장마비와 같은 응급 상황에서 환자를 놓치지 않는 것이 중요함을 강조하였다. 데이터 전처리 과정에서는 중복된 관측치를 제거하고, 결측값은 평균값으로 대체하여 데이터 품질을 향상시켰다. 또한, Stratified K-Fold 교차 검증을 통해 모델의 일반화 성능을 높이고, 클래스 비율을 균등하게 분배하여 평가의 일관성을 확보하였다. 향후 연구에서는 종단적 데이터를 활용하여 시간에 따른 위험 요인의 변화와 심장마비 발병 간의 인과관계를 심층적으로 분석하고, 더 다양한 머신러닝 알고리즘과 딥러닝 기법을 적용하여 예측 모델의 성능을 한층 더 향상시킬 계획이다. 또한, 실제 임상 환경에서 활용할 수 있는 신뢰성 있는 모델을 개발하여 의료 현장에 기여하고자 한다. 본 연구는 심장마비 예측에 있어서 가장 효과적인 머신러닝 기법을 선정하고, 주요 피처들의 중요도를 분석함으로써 예측 모델의 성능을 최적화하는 방법을 제시하였다. 랜덤 포레스트 모델은 높은 예측 성능을 통해 심장마비 예측에 적합한 모델임을 입증하였으며, KNN 모델은 높은 재현율을 통해 응급 상황에서 환자를 조기에 발견하는 데 중요한 기여를 할 수 있음을 보여주었다. 이러한 결과는 의료 현장에서 심장질환 관리와 예방 전략 수립에 중요한 인사이트를 제공하며, 향후 개인 맞춤형 의료 서비스와 공중보건 정책 개발에도 기여할 수 있을 것으로 기대된다.

References

Statistics Korea. (2021). Results of the 2021 cause of death statistics. Retrieved from https://kostat.go.kr/board.es?mid=a10301060200&bid=218

Korea Centers for Disease Control and Prevention (KCDC). National health information portal. Retrieved from https://health.kdca.go.kr/healthinfo/biz/health/gnrlzHealthInfo/gnrlzHealthInfo/gnrlzHealthInfoView.do?cntnts_sn=5243

Ministry of Health and Welfare. Comprehensive plan for the prevention and management of cardiovascular and cerebrovascular diseases. Retrieved from https://www.kdca.go.kr/contents.es?mid=a20303020300

Krittanawong, C., et al. (2020). Machine learning prediction in cardiovascular diseases: A meta-analysis. *Scientific Reports, 10*(1), 16057. https://doi.org/10.1038/s41598-020-72951-z

Texas Heart Institute. Heart disease risk factors. Retrieved from https://ko.texasheart.org/heart-health/heart-information-center/topics/heart-disease-risk-factors/

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. https://doi.org/10.1007/BF00994018

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010