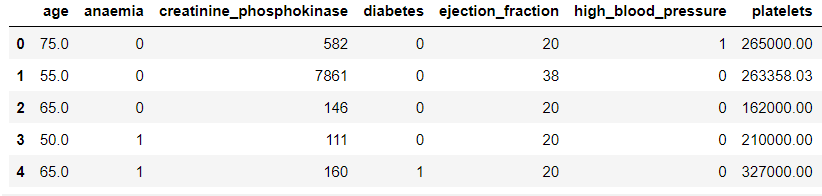
**보건의료 빅데이터 분석 보고서**

1. 데이터 살펴보기

HF\_data 데이터는 299개의 행과 13개의 열로 이루어져 있으며, NA값은 존재하지 않는다. 본 데이터를 사용해서 DEATH\_EVENT를 예측하고자 한다. 설명변수로는 'age', 'anaemia', 'creatinine\_phosphokinase', 'diabetes', 'ejection\_fraction', 'high\_blood\_pressure', 'platelets',

'serum\_creatinine', 'serum\_sodium', 'sex', 'smoking', 'time'을 사용하였고, 반응변수로는 ‘DEATH\_EVENT’를 사용하였다.



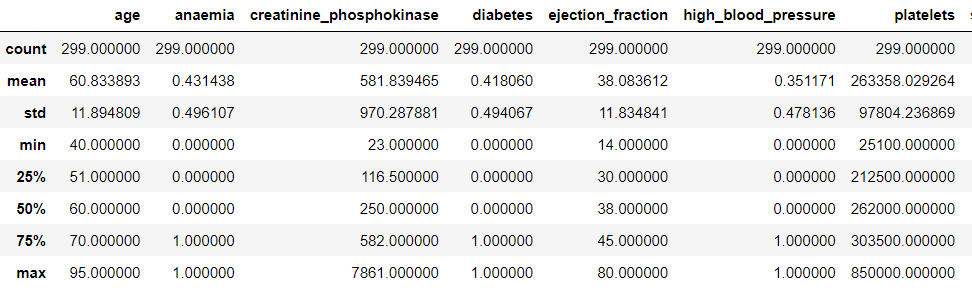
테이블이(가) 표시된 사진

자동 생성된 설명

[그림 1. HF\_data 데이터]

1. 데이터 기술통계량 확인

데이터의 개수, 평균, 표준편차, 최솟값, 최댓값, 4분위수에 대해 확인하였다.



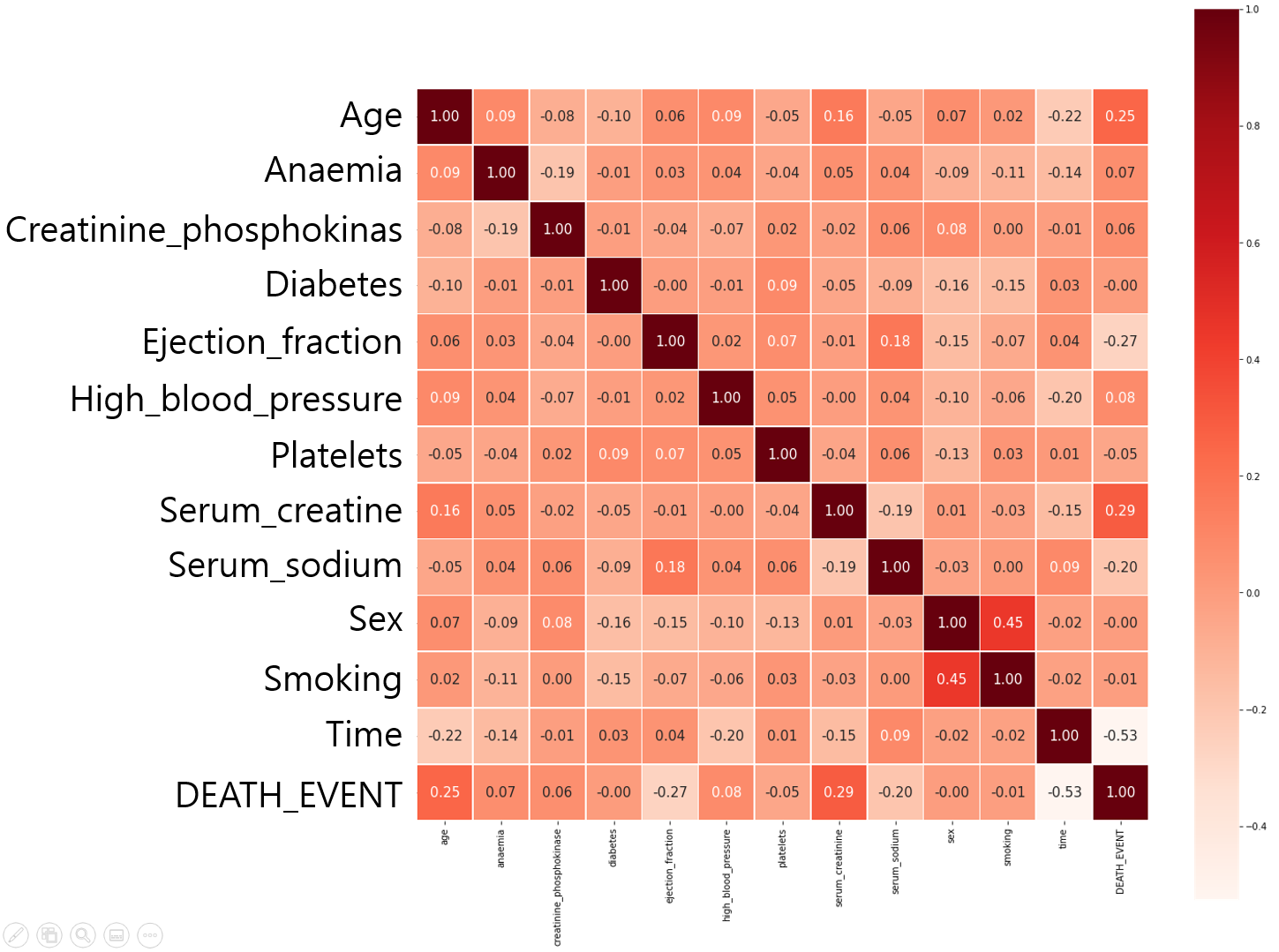
테이블이(가) 표시된 사진

자동 생성된 설명

[그림 2. HF\_data 기술통계량 결과]

1. 변수간 상관관계 확인

변수간 상관관계를 비교해보면, Age, Ejection\_fraction, Serum\_creatine, Time이 DEATH\_EVENT와 0.2이상으로 다른 변수들과 비교했을 때, 비교적 높은 상관관계를 보였으며, Diabetes와 Sex, Smoking은 DEATH\_EVENT와 적은 상관관계를 보였다.



[그림 3. HF\_data 변수간 상관관계 결과]

1. 전처리

Sklearn.model\_selection의 train\_test\_split을 사용해서 데이터를 분할하였으며, random\_state=18로 고정하여, 값이 변하지 않도록 고정하여 모델 학습에 사용하였다. 앞서 언급하였듯이, 설명변수로는 'age', 'anaemia', 'creatinine\_phosphokinase', 'diabetes', 'ejection\_fraction', 'high\_blood\_pressure', 'platelets', 'serum\_creatinine', 'serum\_sodium', 'sex', 'smoking', 'time'을 사용하였고, 반응변수로는 ‘DEATH\_EVENT’를 사용하였다. Train data와 Test data의 shape은 다음과 같다.



1. 모델 적용

각 설명변수들을 표준화 한 후, 공통적으로 모든 모델에 적용하였다.

* 1. Logistic Regression (로지스틱 회귀분석)

Logistic Regression의 max\_iter=1000으로 지정하여 모델을 구축하였다. Train data의 Accuracy는 0.831, Test data의 Accuracy는 0.827의 결과로 비교적 우수한 결과를 보였다.

|  |  |  |
| --- | --- | --- |
| Logistic Regression | Train | Test |
| Accuracy | 0.831 | 0.827 |

[표 1. Logistic Regression 결과]

5.2 Cross-Validation (교차검증)

cv를 5로 설정하여, 5-fold로 검증하였다. 0.756, 0.8, 0.778, 0.845, 0.841의 결과가 나왔

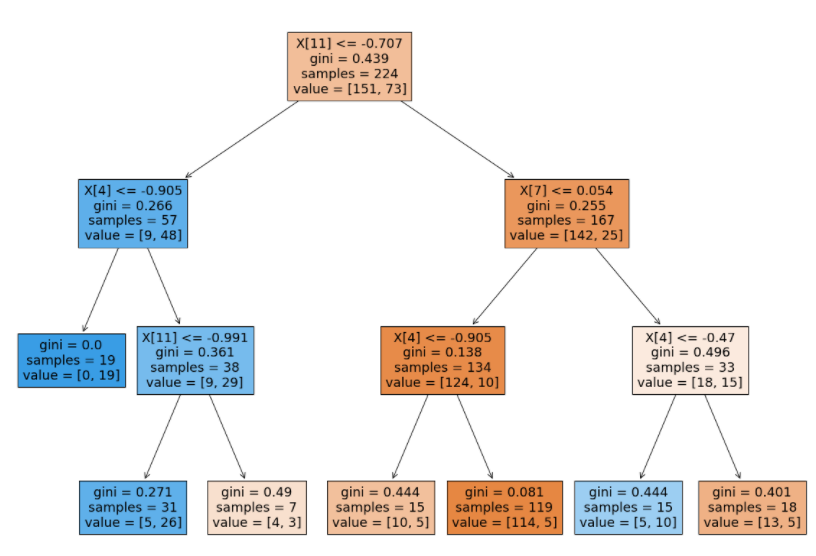
으며, 결과적으로는 0.804의 Acuuracy를 보였다.

5.3 결정트리모형

Radom\_state를 42로 고정하여 결과 값이 변하지 않도록 설정하였으며, max\_depth는3으로 설정하였다. 모델의 결과는 다음과 같다. Train data의 Accuracy는 0.875, Test data의 Accuracy는 0.893로 Underffiting이 의심된다.

|  |  |  |
| --- | --- | --- |
| 결정트리모형 | train | test |
| Accuracy | 0.875 | 0.893 |

[표 2. 결정트리모형 결과]



[그림 4. 결정트리모형 구조]

결정트리모형에 가장 큰 영향을 주는 인자로는 'ejection\_fraction', 'serum\_creatinine',

'time'으로 예상된다.

5.4 KNN

n\_neighbors=3으로 설정하여 모델을 설계하였다. Train Data의 Accuracy로는

0.799 Test data의 Accuracy로는 0.733으로 결과가 나왔다.

|  |  |  |
| --- | --- | --- |
| KNN | Train | Test |
| Accuracy | 0.799 | 0.733 |

[표 3. KNN 결과]

1. 결과

적용한 모델 Logistic Regression, 결정트리 모형, KNN의 결과를 비교하면 다음과 같다.

Test Accuracy를 비교하면 결정트리모형이 가장 우수하지만, underfitting된 결과임을 고려하였을 경우, Logistic Regression이 0.827로 가장 우수한 성능을 보였으므로 Logistic Regression을 최종 모형으로 채택하고자 한다.

|  |  |  |
| --- | --- | --- |
| Model | Train Accuracy | Test Accuray |
| Logistic Regression | 0.831 | 0.827 |
| 결정트리모형 | 0.875 | 0.893 |
| KNN | 0.799 | 0.733 |

[표 4. 전체 모델 결과 비교]