천재교육 빅데이터 8기

Machine Learning

팀프로젝트

빅데이터 8기 김영규, 김경수, 이현희

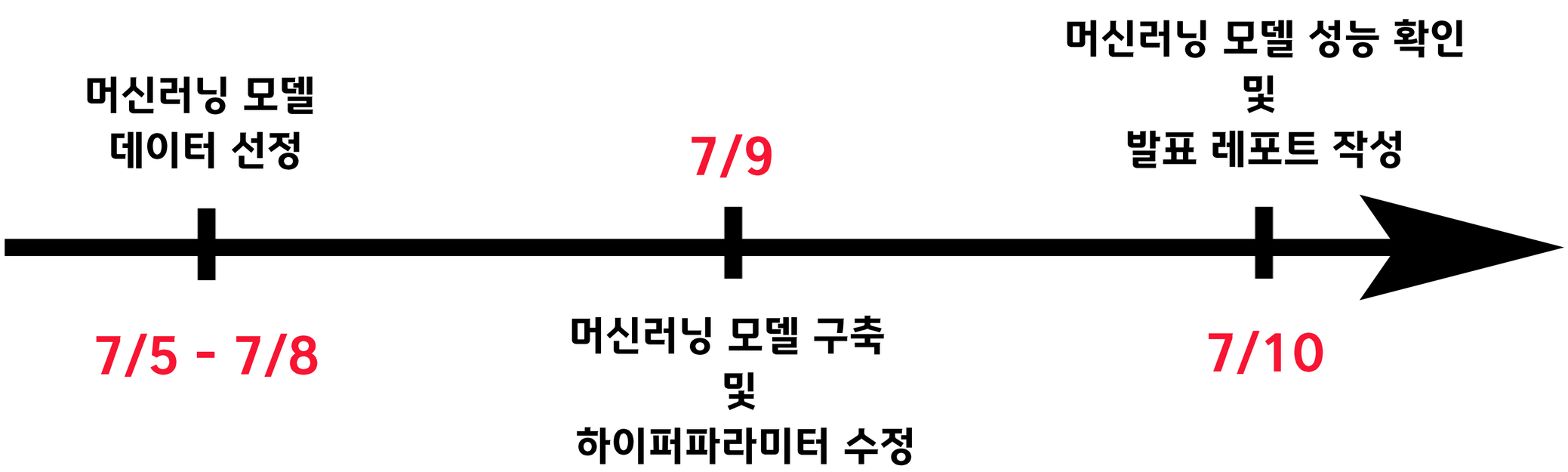
Contents.

01 역할분담 및 Timestamp

김경수	모델 성능 구축 및 하이퍼패러미터 수정
김영규	모델 성능 구축 및 데이터 전처리
이현희	모델 성능 구축 및 시각화, 발표 레포트 작성

01 역할분담 및 Timestamp

Timestamp



Student Stress Factors: A Comprehensive Analysis

< FEATURE는 크게 5개의 요인으로 구분 >

1. 심리적 요인 (4개)

anxiety_level (불안 수준)

self_esteem(자존감)

mental_health_history(정신 건강 병력)

depression(우울증)

2. 생리적 요인(4개)

headache(두통)

blood_pressure(혈압)

sleep_quality(수면의 질),

breathing_problem(호흡 문제)

Student Stress Factors: A Comprehensive Analysis

< FEATURE는 크게 5개의 요인으로 구분 >

3. 환경적 요인 (4개)

noise_level(소음 수준)

living_conditions(생활 조건)

safety(안전)

basic_needs(기본적 요구 사항)

4. 학업적 요인(4개)

academic_performance(학업 성취도)

study_load(공부 부담)

teacher_student_relationship(교사-학생 관계)

uture_career_concerns(미래 진로 고민)

Student Stress Factors: A Comprehensive Analysis

< FEATURE는 크게 5개의 요인으로 구분 >

5. 사회적 요인(4개)

social_support(사회적 지원)

peer_pressure(동료 압력)

extracurricular_activities(과외 활동)

bullying(괴롭힘)

Student Stress Factors: A Comprehensive Analysis

< 데이터 Train / Test 로 구분>

- StressLevelDataset.csv : 1100 rows
- train.csv : 880 rows
- test.csv : 220 rows

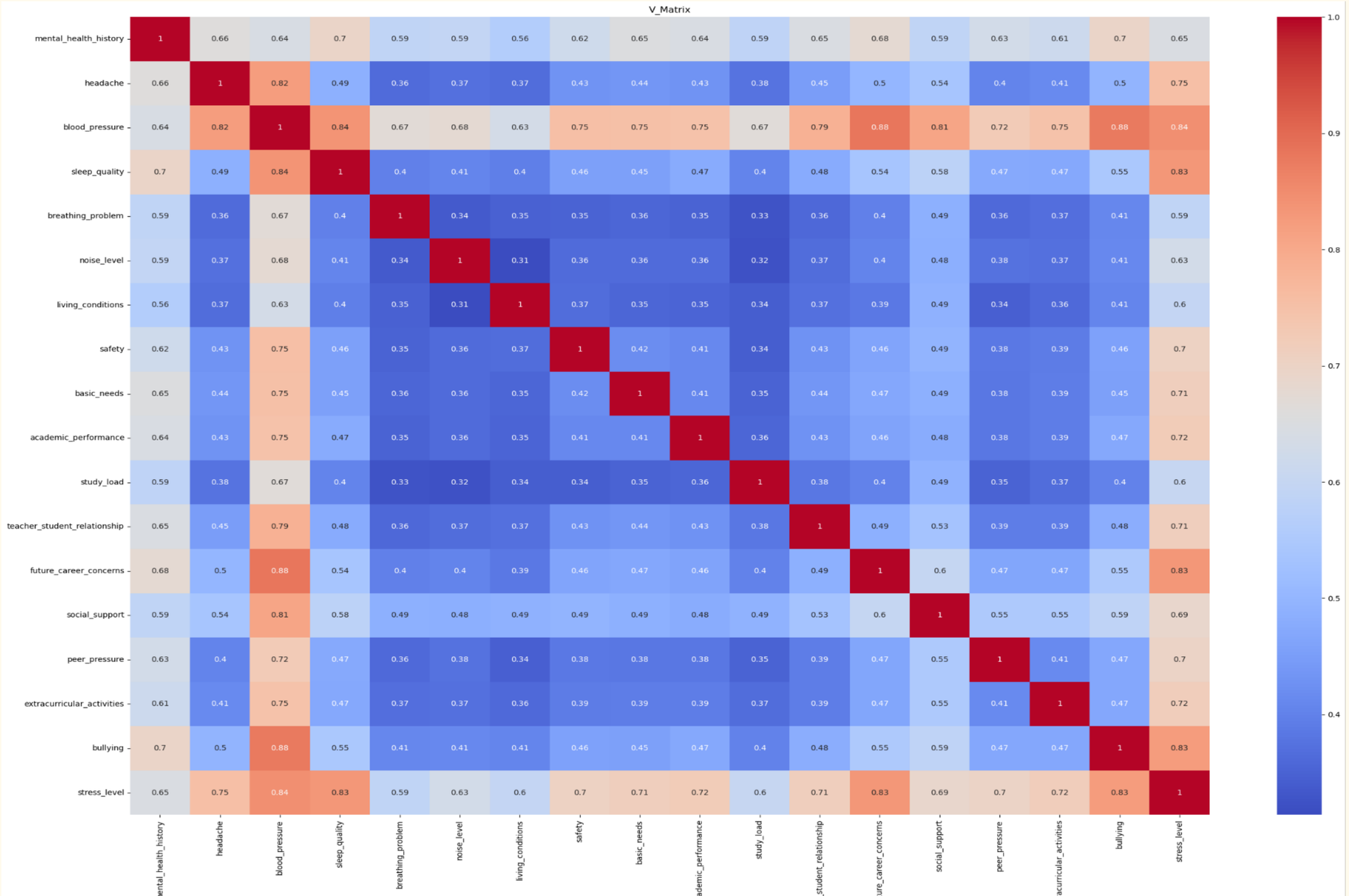
stress_level
0,1,2 클래스 예측

< Feature간의 아노바 검정>

Feature	stress_level
anxiety_level	5.296e-188
self_esteem	1.269e-210
depression	1.920e-187

< Feature간의 상관관계 분석>

데이터 소개



모델 손실 및 정확도

03

사용한 모델

- LogisticRegression
- SGD
- DecisionTree_RS(CV)
- DecisionTree_GS(CV)
- RandomForest_RS(CV)
- RandomForest_GS(CV)

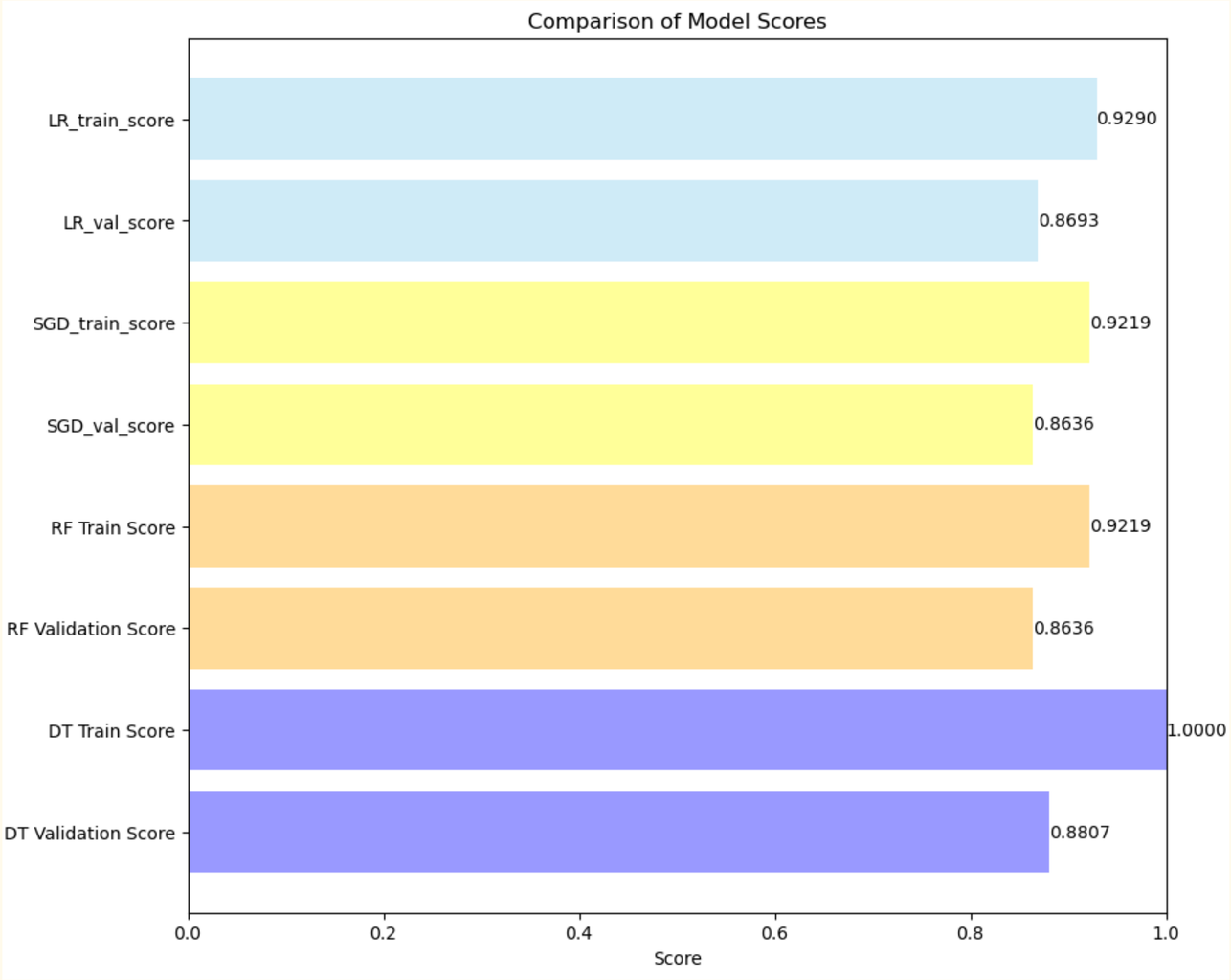
* 기본 모델

모델 손실 및 정확도

사용한 모델	정확도
LogisticRegression	train : 0.9290 test : 0.8693
SGD	train : 09219 test : 0.8636
RandomForest	train : 0.9815 test : 0.8593
DecisionTree	train : 1.0 test : 0.8807

모델 손실 및 정확도 시각화

03

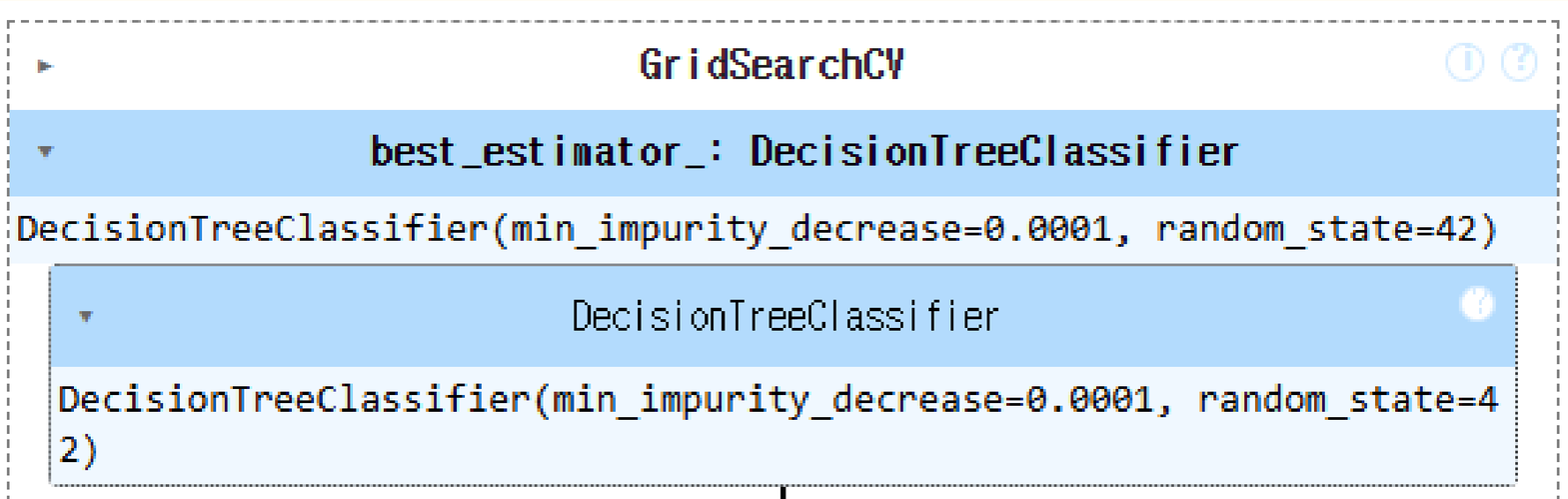


하이퍼패리미터 사용 전 train/test결과

- Logistic Regression
- SGD
- RandomForest
- DecisionTree

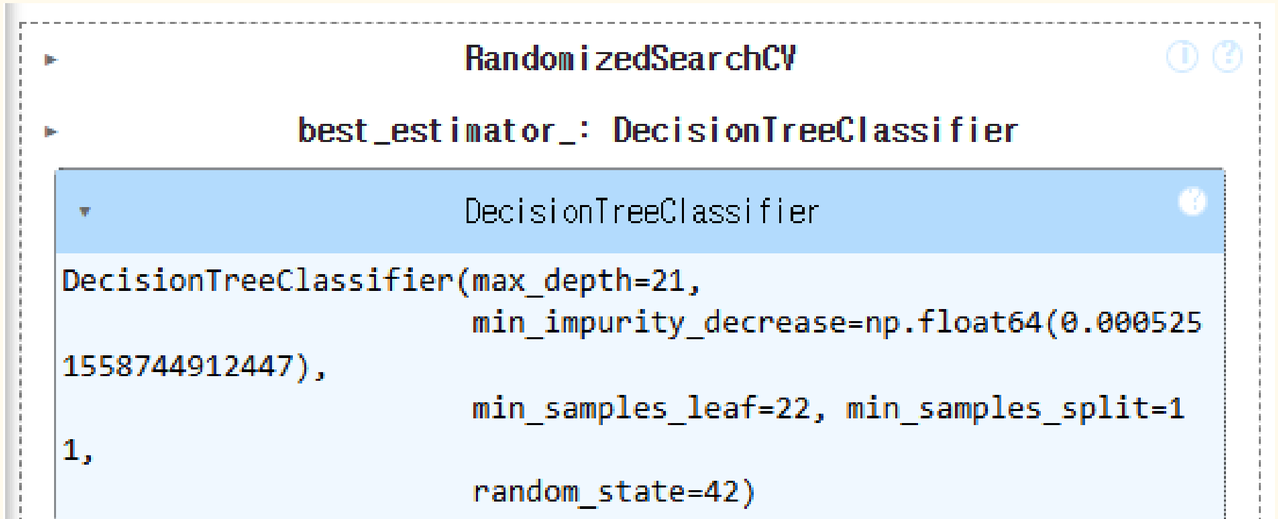
모델 손실 및 정확도

사용한 모델	정확도
DecisionTree	train : 1.0 test : 0.88
DecisionTree_GS(CV)	train : 1.0 test : 0.88



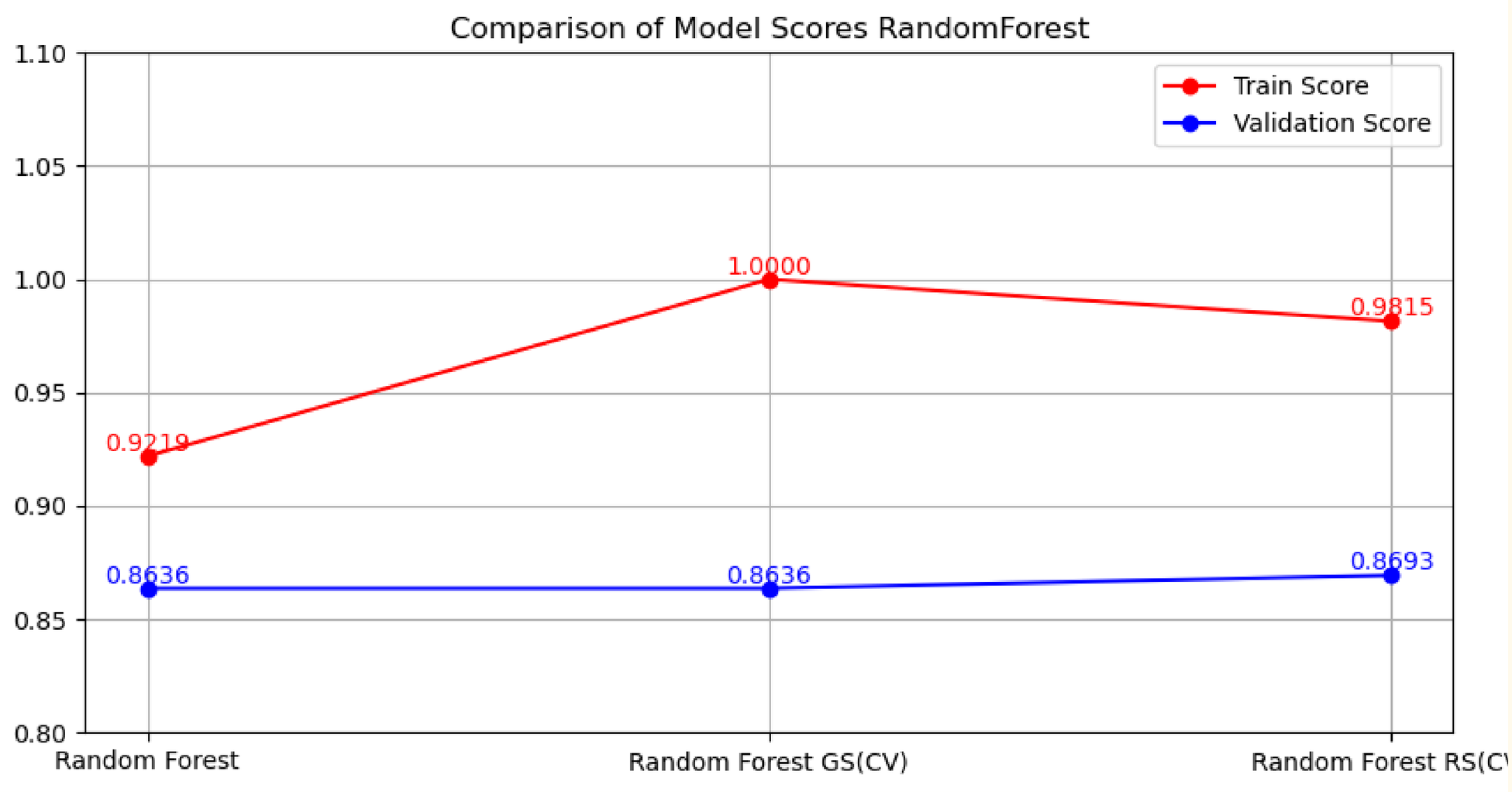
모델 손실 및 정확도

사용한 모델	정확도
DecisionTree	train : 1.0 test : 0.88
DecisionTree_RS(CV)	train : 0.95 test : 0.86



모델 손실 및 정확도(하이퍼패리미터 시각화)

하이퍼패리미터 사용 후 RandomTree train/test결과



모델 손실 및 정확도

03

사용한 모델	정확도
RandomForest	train : 0.92 test : 0.86
RandomForest_GS(CV)	train : 1.0 test : 0.86

GridSearchCV

best_estimator_: RandomForestClassifier

RandomForestClassifier

```
RandomForestClassifier(max_depth=7,
                       min_impurity_decrease=np.float64(0.005623413
251903491),
                       min_samples_split=4, random_state=42)
```

모델 손실 및 정확도

사용한 모델	정확도
RandomForest	train : 0.92 test : 0.86
RandomForest_RS(CV)	train : 0.98 test : 0.86

RandomizedSearchCV

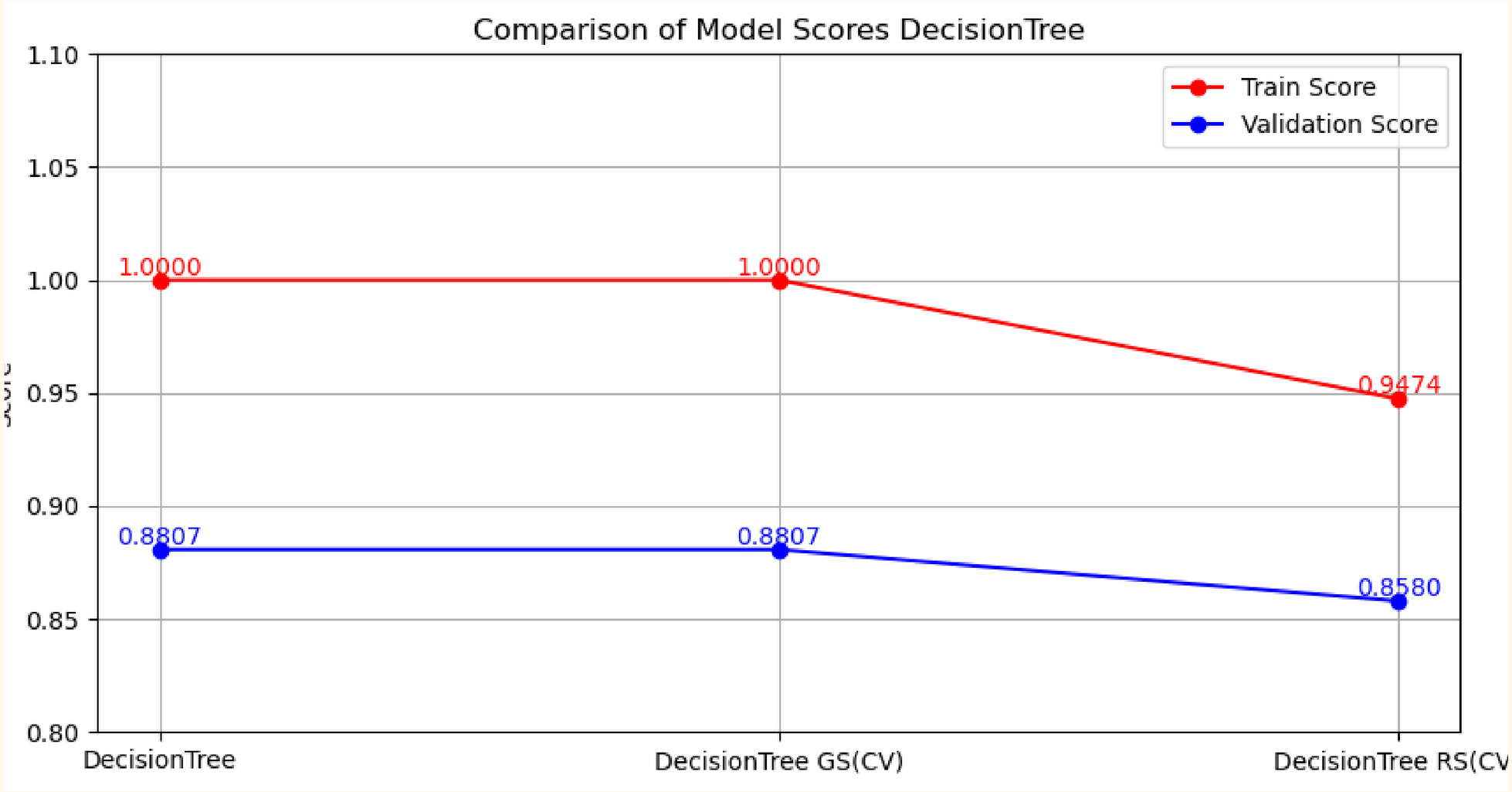
best_estimator_: RandomForestClassifier

RandomForestClassifier

```
RandomForestClassifier(max_depth=45,
                       min_impurity_decrease=np.float64(0.000612093
0582992811),
                       min_samples_leaf=19, min_samples_split=21,
                       random_state=42)
```

모델 손실 및 정확도(하이퍼패리미터 시각화)

하이퍼패리미터 사용 후 DecisionTree train/test결과



RandomTree_RS(CV)

test data

정확도 : 0.85

개선점

모델별로 다양한 하이퍼 파라미터가 있지만,

시간 여건상 제대로 알지 못해서,

random forest와 decision tree 이 두가지의

모델의 하이퍼 파라미터만 적용한 것이 아쉽습니다.

추가적인 학습을 통해서 다양한 모델의 하이퍼 파라미터를 적용시켜

모델의 성능을 더욱더 향상 시키고 싶습니다.

감사합니다 !