2022-2 머신러닝 및 딥러닝 프로젝트 발표

유방암 데이터셋 활용방안

2016110838
행정학전공
홍서이

Contents

1.프로젝트 세부 내용

데이터 전처리, 데이터 탐색, 모델 선택 등의 프로젝트 진행 과정에 대한 구체적인 설명

2. 프로젝트를 진행하면서 배운 점

프로젝트 세부 내용
: 데이터프레임 생성

BC-TCGA-Normal.txt
BC-TCGA-Tumor.txt

Hybridization REF		TCGA-BH-A0AY-11A-23R-A089-07	TCGA-A7-A0DB-11A-33R-A089-07
0	ELMO2	0.204333	0.869417
1	CREB3L1	-0.242000	0.878250
2	RPS11	0.591875	-0.024625
3	PNMA1	0.538500	0.819500
4	MMP2	0.707667	1.932333

전치 수행

label 추가

cancer_classification.csv

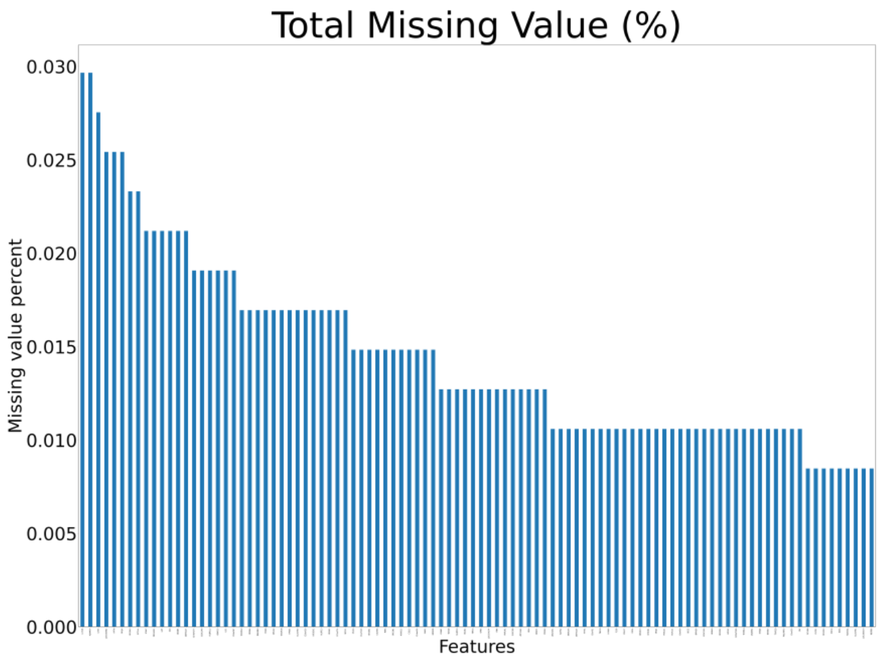
	ELMO2	CREB3L1	RPS11
0	0.610333	1.7550	0.403875
1	0.055917	0.2450	0.337125
2	0.785583	1.1935	0.314375
3	0.232667	0.0055	0.745750
4	0.286917	1.1100	0.209750

shape: (590, 17815)

프로젝트 세부내용: 결측치 처리

결측치 비율 0.03% 이하

중간값으로 대치하여 결측치 처리



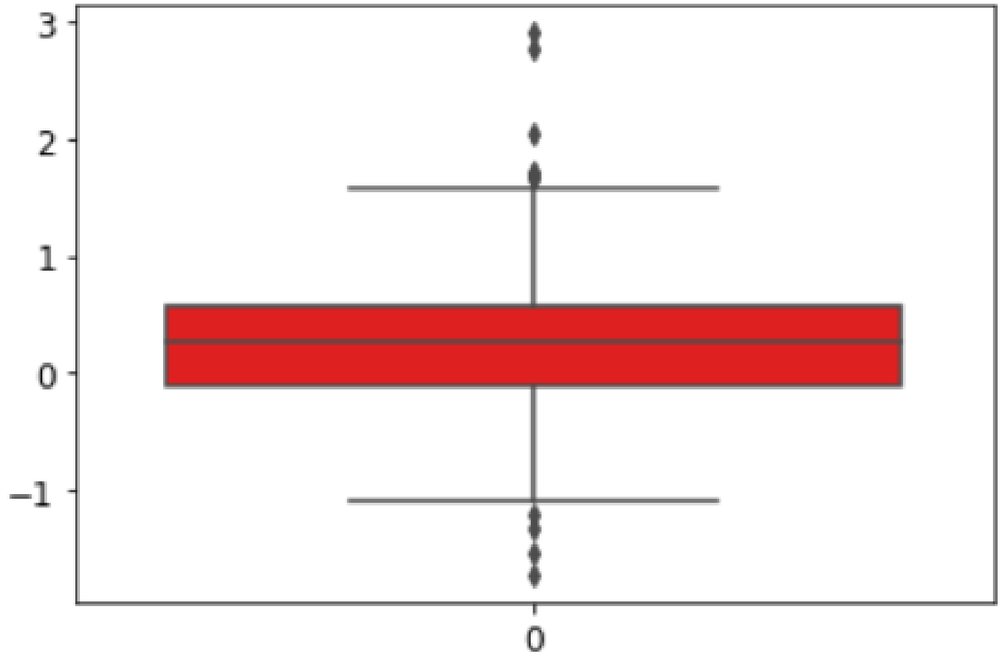
결측치 비율	처리 방법
10% 미만	제거 or 어떠한 방법이든지 상관없이 Imputation
10% 이상 20% 미만	hot deck , regression , model based method
20% 이상	model based method , regression

[표1] 결측치의 비율에 따른 처리 방법

프로젝트 세부내용: 이상치 처리

Boxplot 확인시 이상치 존재

standard scaler 이용해 이상치 변환처리




```
# 데이터 변환: scaling
standard_scaler = StandardScaler()
standard_scaler.fit(X_train)
X_scaled_train = standard_scaler.transform(X_train)
X_scaled_vaild  = standard_scaler.transform(X_valid)
X_scaled_test  = standard_scaler.transform(X_test)
```

프로젝트 세부내용:
Dimension reduction-PCA

```
from sklearn.decomposition import PCA
pca = PCA()
pca.fit(X_scaled_train)
cumsum = np.cumsum(pca.explained_variance_ratio_)
d = np.argmax(cumsum>=0.95)+1
```

```
[164] d

      337
```

```
[168] pca = PCA(n_components=337)
      pca.fit(X_scaled_train)
      X_PCA_train = pca.transform(X_scaled_train)
      X_PCA_valid = pca.transform(X_scaled_valid)
      X_PCA_test  = pca.transform(X_scaled_test)
```

train dataset의 분산을 95% 유지하는데 필요한

최소한의 차원 수 계산

프로젝트 세부내용: 모델 훈련

Classifiers

- Logistic Regression
- Decision Tree
- Support Vector Machine
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Random Forest
- K-Nearest Neighbors
- Naive Bayes

Scoring

- precision score
- recall score
- F1 score
- support score
- accuracy score
- AUC/ROC

프로젝트 세부내용:train Score

	Model	Fitting time	Scoring time	Accuracy	Precision	Recall	F1_score	AUC_ROC
0	Logistic Regression	0.066499	0.006739	0.997826	0.991667	0.998810	0.998017	1.000000
6	K-Nearest Neighbors	0.001572	0.011218	0.997826	0.991667	0.998810	0.998017	0.998810
2	Support Vector Machine	0.058943	0.006316	0.995743	0.990530	0.990476	0.995750	1.000000
1	Decision Tree	0.036790	0.004411	0.981069	0.979644	0.939286	0.979817	0.939286
3	Linear Discriminant Analysis	0.132410	0.008877	0.955707	0.878333	0.975000	0.960112	0.986905
5	Random Forest	0.413417	0.025026	0.934692	0.866218	0.708333	0.916373	0.999206
4	Quadratic Discriminant Analysis	0.061007	0.008879	0.890217	0.445109	0.500000	0.838617	0.984921
7	Bayes	0.003430	0.005258	0.834420	0.694764	0.823214	0.858704	0.883532

프로젝트 세부내용: Test Score

Testset Score

	Model	Accuracy	Precision	Recall	F1_score	AUC_ROC
6	K-Nearest Neighbors	0.966102	1.000000	0.962963	0.981132	0.981481
0	Logistic Regression	0.949153	0.963636	0.981481	0.972477	0.790741
2	Support Vector Machine	0.949153	0.963636	0.981481	0.972477	0.790741
3	Linear Discriminant Analysis	0.949153	0.963636	0.981481	0.972477	0.790741
5	Random Forest	0.949153	0.947368	1.000000	0.972973	0.700000
1	Decision Tree	0.932203	0.946429	0.981481	0.963636	0.690741
4	Quadratic Discriminant Analysis	0.915254	0.915254	1.000000	0.955752	0.500000
7	Bayes	0.355932	0.944444	0.314815	0.472222	0.557407

프로젝트를 진행하면서 배운점

감사합니다!