제 2차 데이터시각화 과제

20191007 이희원

<목차>

1. 문제 정의
2. 데이터 탐색 및 전처리
3. 예측 분석
4. 테스트 데이터로 예측한 결과 군집화
5. 결과 해석
6. references
7. 문제 정의

아동과 모성 사망을 예방하기 위해 태아 건강을 분류하는 데이터인 fetal\_health를 가지고 Decision Tree, Random Forest, Support Vector, Machine, Logistic Regression, DNN알고리즘을 가지고 태아 건강 정보를 분류/예측과 성능평가, ROC-AUC 등의 과정을 진행하고, K-means, K-medoids로 군집화를 진행하고 PCA(주성분 분석)을 통해 데이터의 정확한 군집 시각화를 표현한다.

1. 데이터 탐색 및 전처리

데이터 탐색 및 전처리 전 데이터 정보를 통해 타입과 결측치를 확인했다. 타입은 전부 float64이고, 결측치는 없는 것을 확인했다.

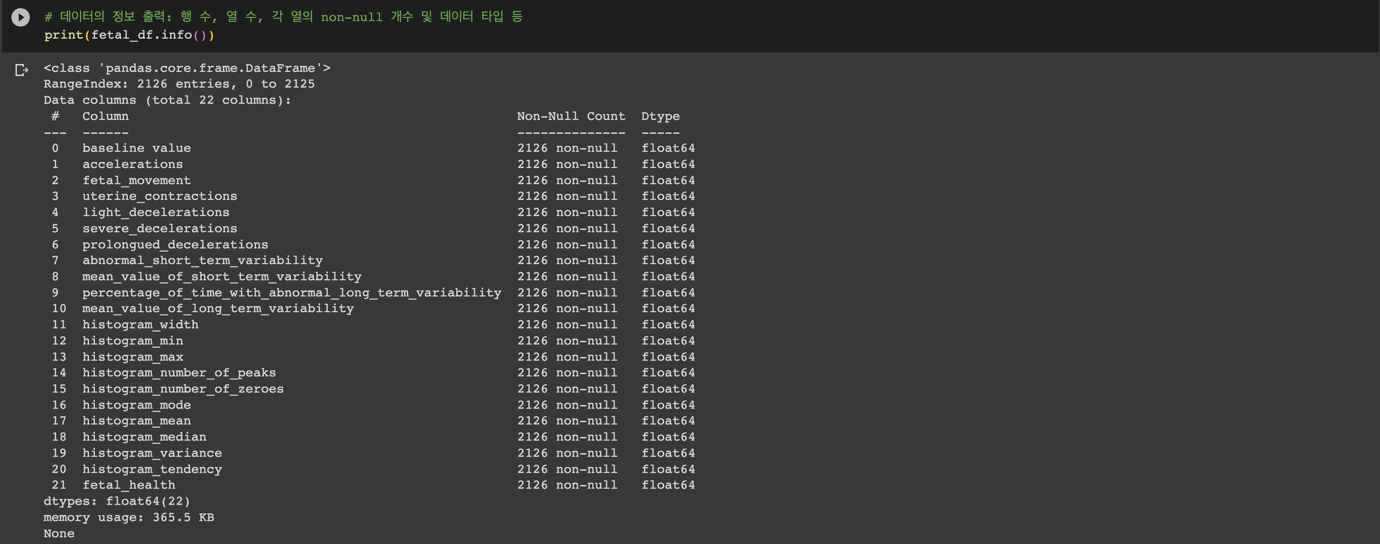


그림 1 데이터 정보 확인하기

각 데이터 변수의 기술 통계를 출력했다.

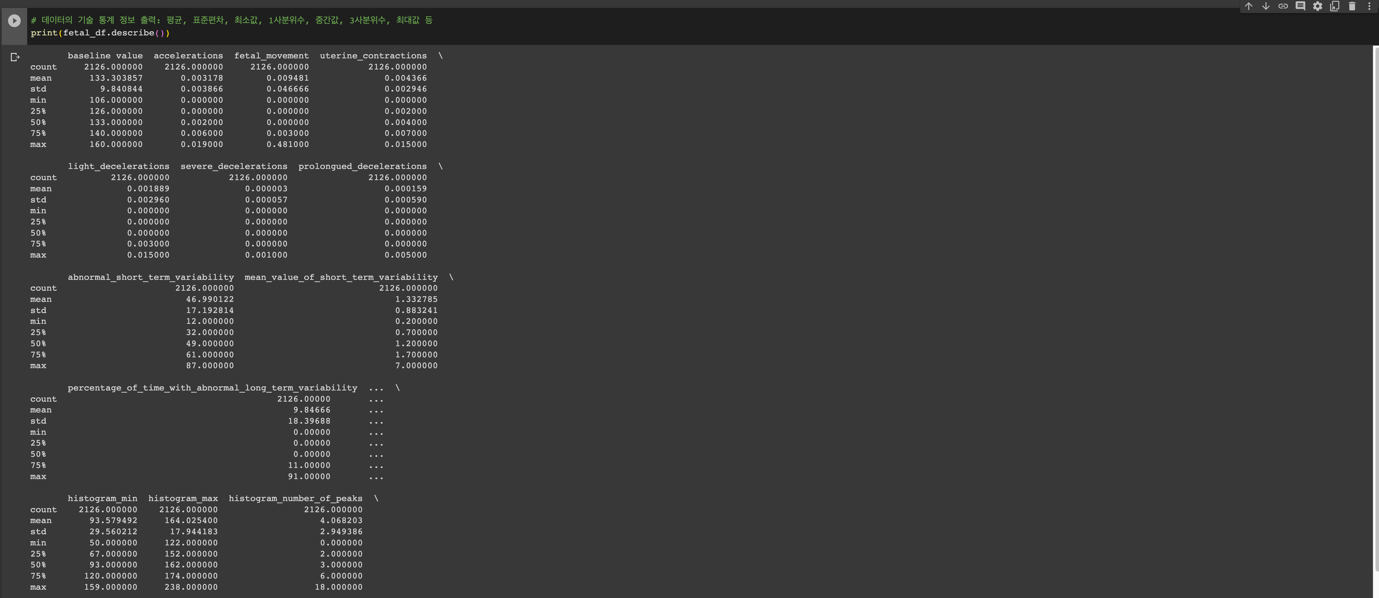


그림 2 기술 통계 정보 출력

타겟 변수로 사용할 태아 건강 상태인 fetal\_health 부분의 분포를 표로 확인한다.  
각 피쳐의 분포를 표로 나타냈다.



그림 3 타켓 변수 분포 확인

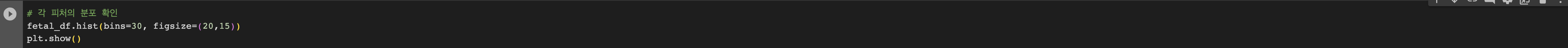


그림 4 피쳐 확인 코드

도표, 평면도이(가) 표시된 사진

자동 생성된 설명

그림 5 피쳐 그래프 결과

각 피쳐의 다중 공선성 확인을 위해 히트맵을 작성해서 시각화로 확인했다.

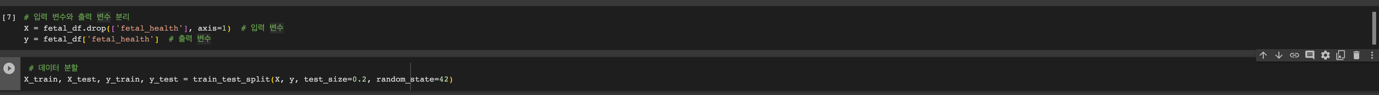
그래프, 텍스트, 소프트웨어, 도표이(가) 표시된 사진

자동 생성된 설명

그림 6 히트맵

1. 예측 분석

입력 변수와 출력 변수를 분리하고, 데이터를 훈련 데이터와 테스트 데이터로 분할했다.

**그림 7 변수 분리, 데이터 분리**

* 1. Decision Tree

DecisionTreeClassifier를 이용해 의사 결정 트리 모델을 생성한다.

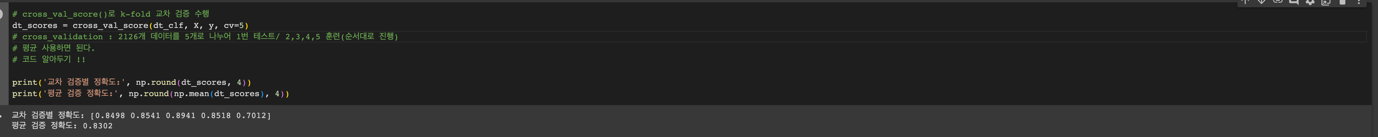
그림 8 DecisionTreeClassifier

의사 결정 트리의 예측 수행, 평가 수행을 진행했다. accuracy, precision, recall, f1 score을 출력했다. roc\_auc 같은 경우 코드에 에러가 발생해 주석 처리했다. 정확도는 92%가 나왔다.

텍스트, 멀티미디어 소프트웨어, 소프트웨어, 그래픽 소프트웨어이(가) 표시된 사진

자동 생성된 설명그림 9 의사결정트리의 예측, 평가 수행

cross\_val\_score()로 k-fold 교차 검증을 수행하고, 2126개의 데이터를 5개로 나누어 각각 순서대로 테스트를 진행했다. 그 후 교차 검증별 정확도와 평균 검증 정확도의 결과를 각각 출력했다.

그림 10 cross\_val\_score() 진행, 교차 검증별 정확도, 평균 검증 정확도 출력

타겟 데이터인 ‘fetal\_health’를 제외한 나머지 columns의 이름을 가져왔다. 실행 결과 : ['baseline value', 'accelerations', 'fetal\_movement', 'uterine\_contractions', 'light\_decelerations', 'severe\_decelerations', 'prolongued\_decelerations', 'abnormal\_short\_term\_variability', 'mean\_value\_of\_short\_term\_variability', 'percentage\_of\_time\_with\_abnormal\_long\_term\_variability', 'mean\_value\_of\_long\_term\_variability', 'histogram\_width', 'histogram\_min', 'histogram\_max', 'histogram\_number\_of\_peaks', 'histogram\_number\_of\_zeroes', 'histogram\_mode', 'histogram\_mean', 'histogram\_median', 'histogram\_variance', 'histogram\_tendency']

그림 11 열 이름 가져오는 코드

그 후, 의사 결정 트리를 그래프로 구현했다.

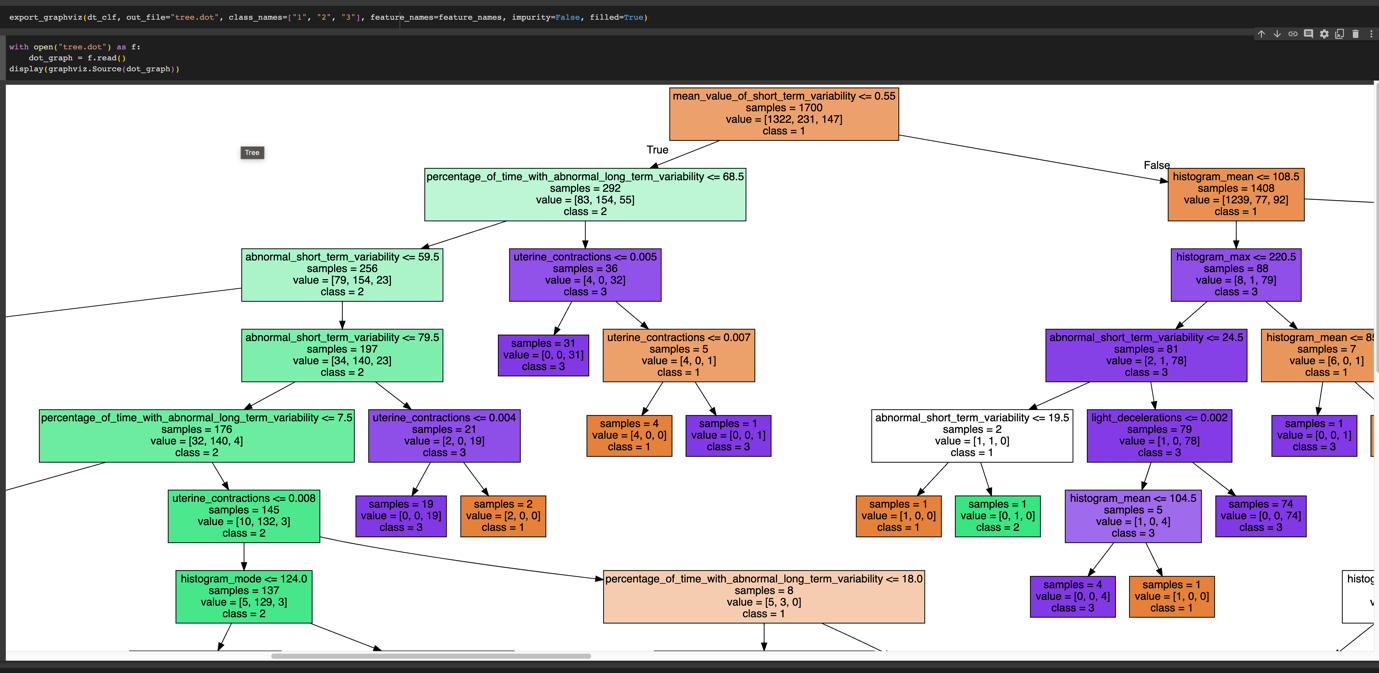


그림 12 Decision Tree 시각화

각 피쳐의 중요도를 출력하고 가로 막대 그래프로 시각화했다.

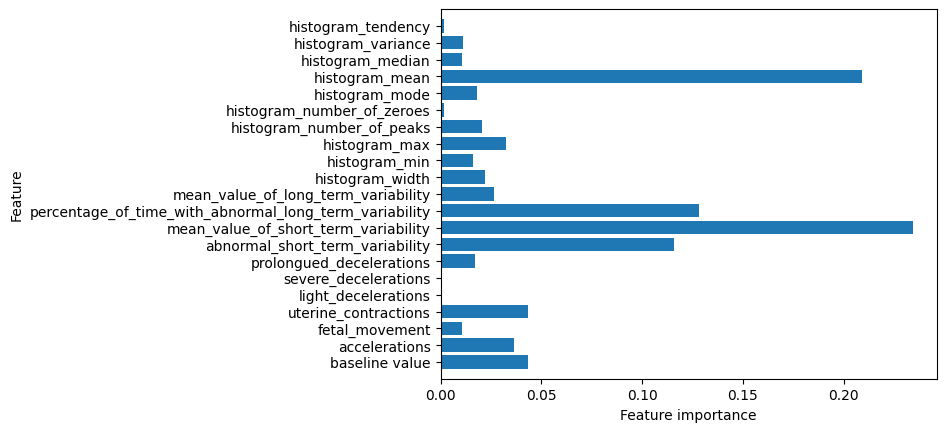


그림 13 가로막대 그래프 시각화

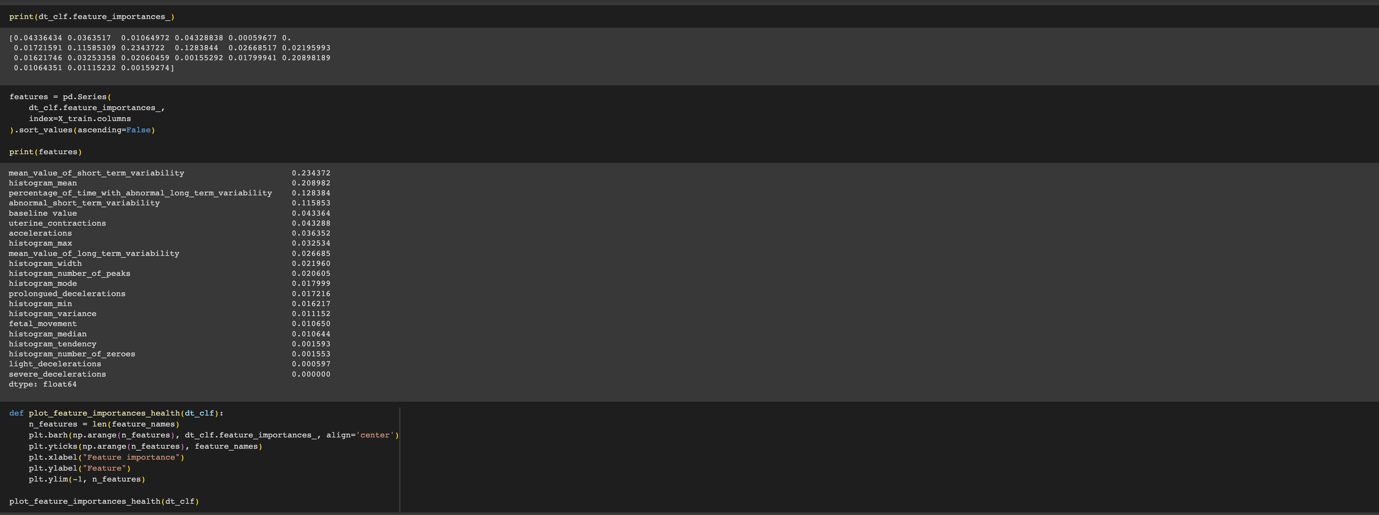


그림 14 importance 구현 코드

* 1. Random Forest

의사 결정 트리를 Ensemble 학습으로 진행한 RandomForest는 RandomForestClassifier로 정의하고 위의 의사 결정 트리와 유사하게 accuracy, precision, recall, f1 score을 출력한다. 정확도 같은 경우는 의사 결정 트리의 정확도인 92%보다 높은 94%가 나왔다.

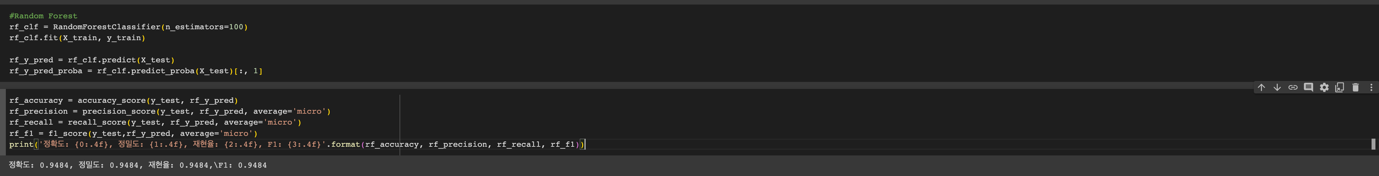


그림 15 Random Forest 적용, 정확도 측정

* 1. Support Vector Machine

Support Vector Machine 같은 경우, 위의 Decision Tree나 Random Forest와 달리 변수의 정규화가 필요해 StandardScaler()를 이용해 정규화를 해준 뒤 SVM을 적용해줬다. 그 후 모델을 학습하고 정확도 측정과 성능 평가를 진행했다. 정확도는 90%가 나왔다. 사실 데이터 탐색을 했을 때, 타겟 데이터가 불균형을 이루기 때문에 SVM의 정확도는 의미가 없다고 봐야 한다.

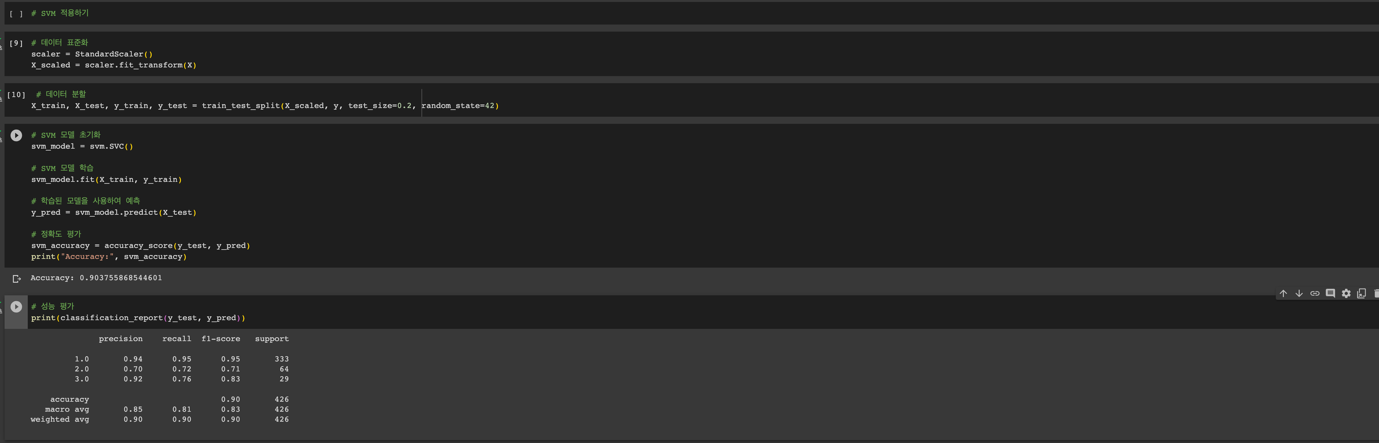


그림 16 Support Vector Machine 측정 결과

* 1. Logistic Regression

Logistic Regression 같은 경우도 로지스틱 회귀 모델을 생성하고, 훈련해준 뒤 정확도 평가와 성능 평가를 진행했다. 정확도는 88%가 나왔다.

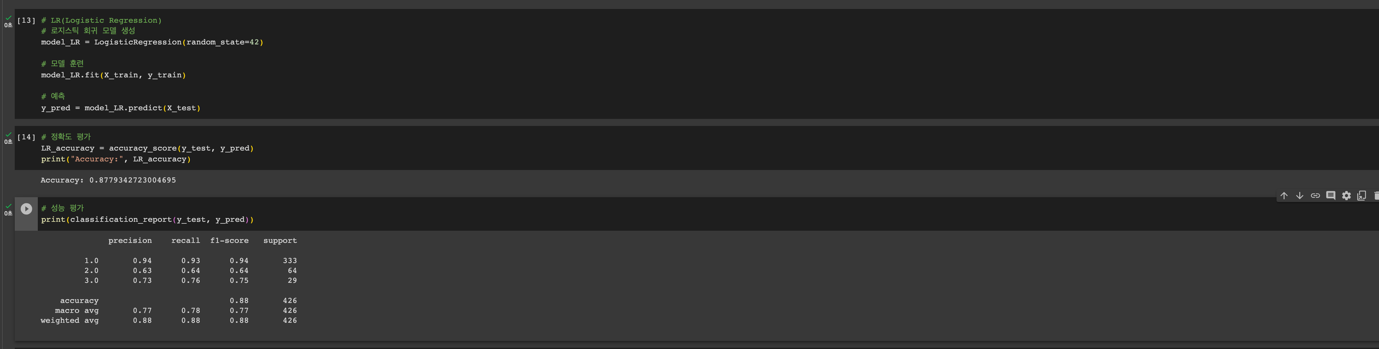


그림 17 Logistic Regression 적용

* 1. Deep Neural Network

DNN 같은 경우, 입력층과 출력층 사이에 2개의 은닉층(각각 32개 노드, 16개 노드)를 두고 DNN을 설계했다. 그리고 EarlyStopping을 사용해 조기 멈춤도 설정했다. 정확도는 78%가 나왔다.

그림 18 DNN 성능평가

텍스트, 소프트웨어, 멀티미디어 소프트웨어, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그림 19 DNN 설계 과정

1. 테스트 데이터로 예측한 결과 군집화

군집을 위해 실루엣 계수를 측정했다.



그림 20 실루엣 계수 측정

그 후, fetal\_health의 각 실루엣 계수의 평균을 출력했다.

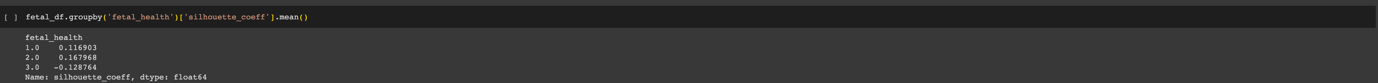


그림 21 실루엣 계수 평균 출력

데이터 시각화 강의 시간에 배운 실루엣 계수 면적 시각화 함수를 이용해서 시각화를 진행했다. cluster 개수를 2개~5개로 설정해 면적을 시각화했을 때, 4개의 cluster가 실루엣 계수 라인에 전체적으로 면적이 들어가서 cluster가 4개일 때가 가장 적합한 것으로 보인다.

그리고 클러스터를 할당해 군집 시각화 결과를 확인했다. 일전 강의자료의 코드를 참고했는데 ‘c=labels’를 넣으면 에러가 해결되지 않아 지우고 실행해서 원하지 않는 K-means가 나왔을 것으로 예상한다.

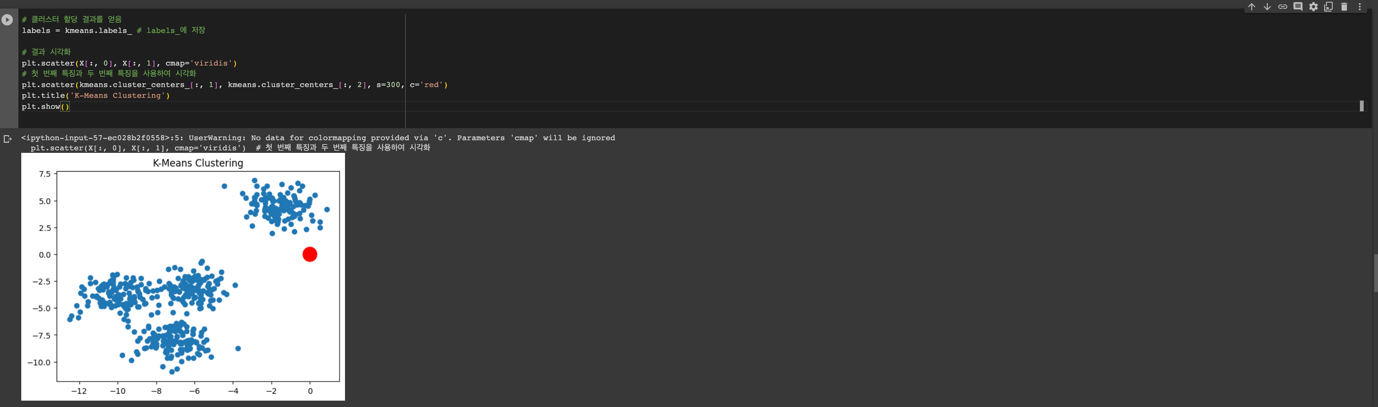


그림 22 K-means 적용

KMeans(n\_clusters=3, init='k-means++', max\_iter=300, random\_state=0)를 이용해 Kmeans를 정의하고, labels를 출력했다.



그림 23 Kmeans++ 정의

PCA(주성분 분석)을 이용해 pca\_x, pca\_y를 도출했다.

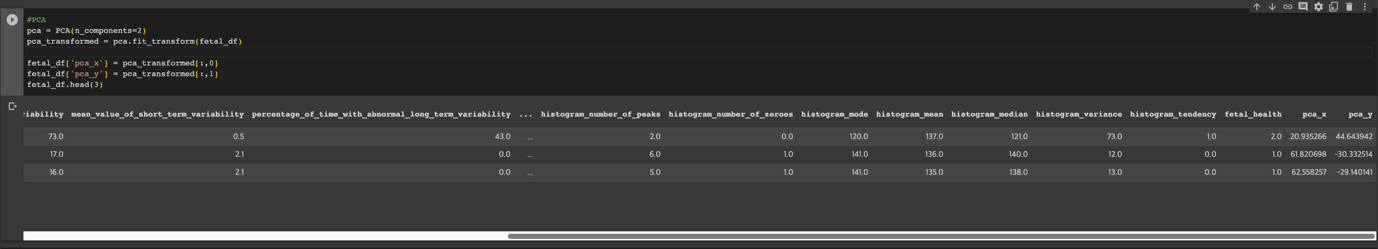


그림 24 PCA\_X, PCA\_Y 도출

군집 값이 1, 2, 3인 경우마다 별도의 인덱스로 추출했다. 그리고 각 해당하는 인덱스에 각 군집 레벨의 pca\_x, pca\_y 값을 추출하였다. 시각화 시 각 마커를 o, s, ^로 표시했다. 시각화 결과를 확인했을 때, 군집이 정확히 구분되지 않았고, 어떤 클러스터는 값이 모이지 않고 이상치로 흩어진 부분이 많았다.



그림 25 군집 시각화 결과

이번에는 K-Medoids를 이용해 군집 시각화를 진행했다. 이 또한, 군집 결과가 이상하게 된 것을 확인할 수 있다.

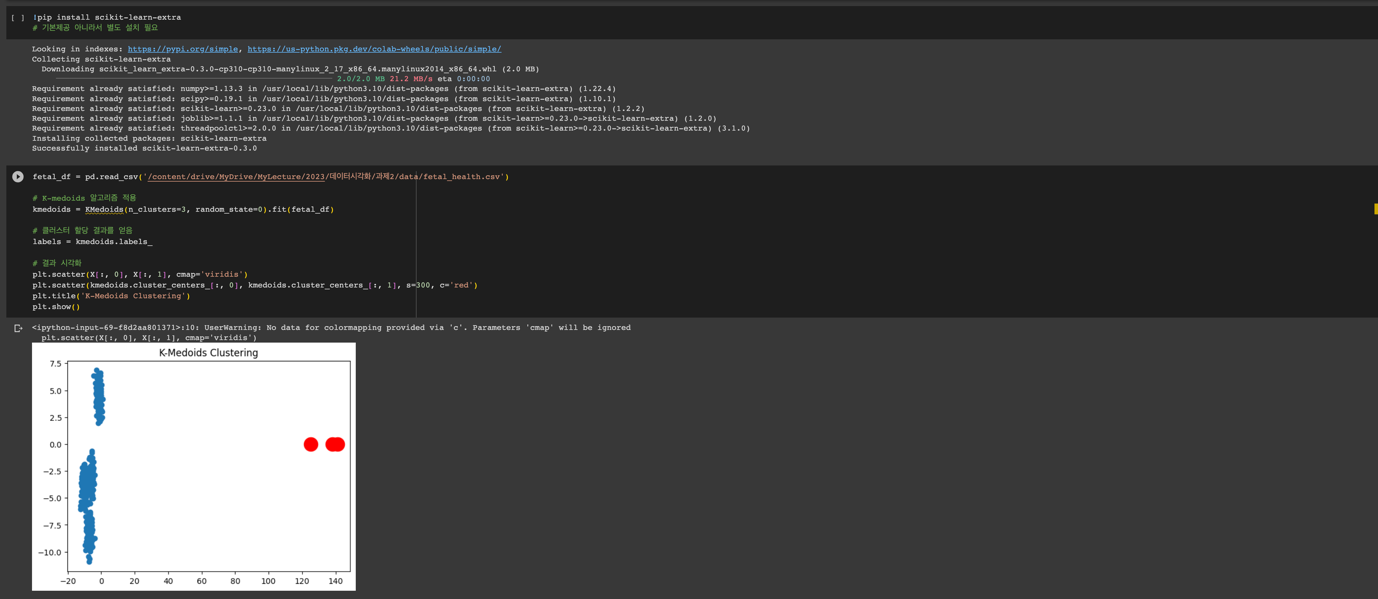


그림 26 K-Medoids 실행 결과

K-means, K-medoids의 결과가 다 제대로 군집이 되지 않았다. 그래서 fetal\_df에 있는 변수를 일부만 가져온 뒤 2개의 변수를 선택해 scatter 시각화를 통해 군집화를 다시 진행했다.

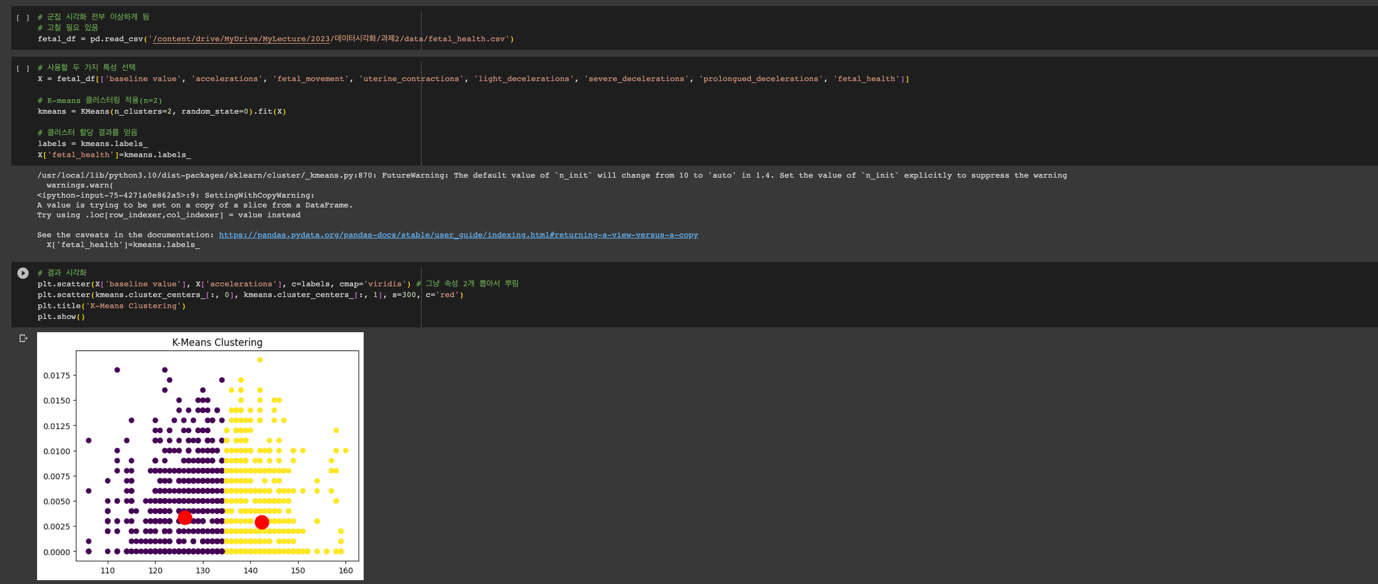


그림 27 두 개의 변수로 K-Means

PCA를 이용해 주성분 분석을 한 후 각 시각화를 진행했다. 시각화가 2개의 결과로 나와서 3개일 때의 방법을 알지 못했다.

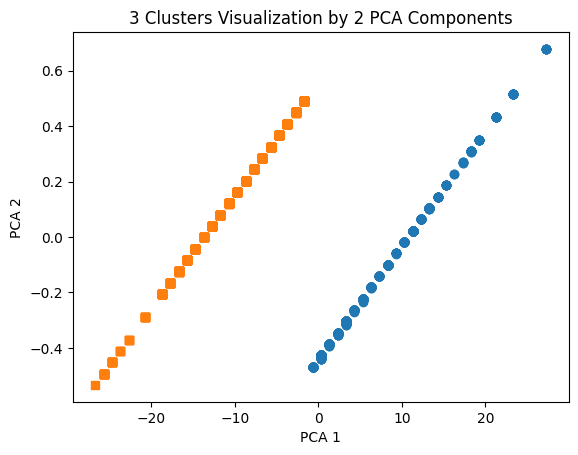


그림 28 시각화 결과

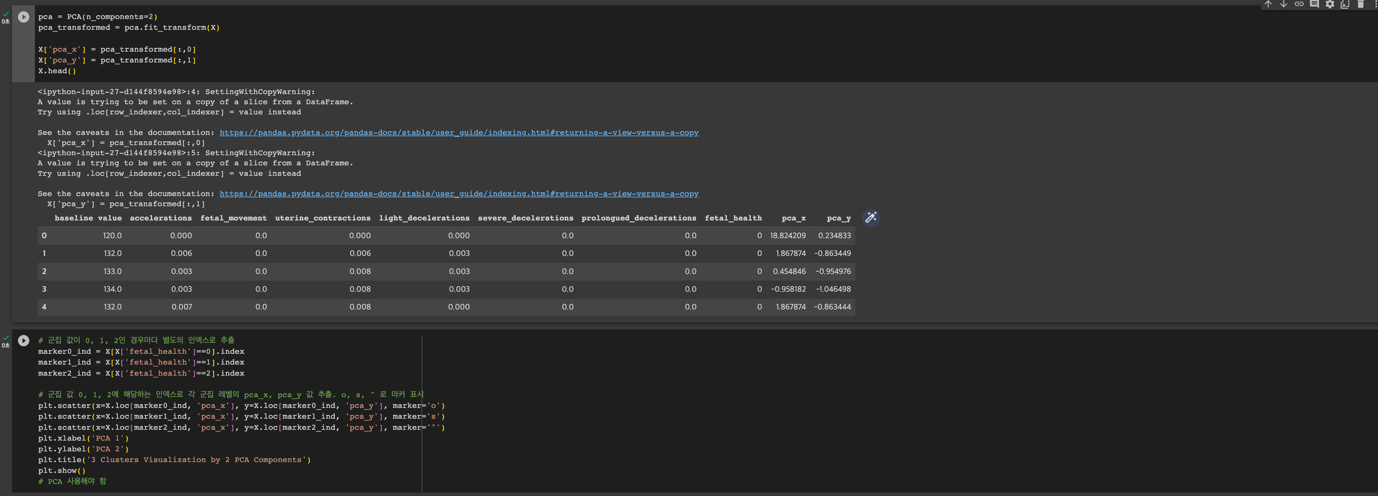


그림 29 주성분 분석 => 시각화 코드

1. 결과 해석

6개의 알고리즘 중 분류/예측의 정확도가 가장 좋은 것은 Random Forest 였고, 가장 정확도가 좋지 않은 알고리즘은 Deep Neural Network였다. 그러나, DNN은 은닉층을 더 추가한다면 좋은 알고리즘으로 발전할 가능성이 있다고 유추한다. 그러나 태아의 건강을 유추할 때 1이 가장 많은 경우였고, 2나 3은 그에 비해 희박한 숫자를 가지고 있어 예측의 정확도에 있어서는 좋은 모델은 아니라고 할 수 있다. 가중치를 줘서 데이터의 불균형을 해결 한다면 그만큼의 모델의 정확도가 달라질 것이다. 군집 같은 경우, 변수의 개수를 정해서 군집화를 해야 맞게 군집이 되었다. 군집을 하는 데 있어 PCA(주성분 분석)은 필수적인 요건임을 알 수 있다.

1. References

<https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification> (데이터 및 설명)