기계학습기초 미니 프로젝트 보고서

주제: 학생들의 적응도를 통한 온라인 교육 분석

(Analysis of Online Education Through Students Adaptability Level)

자연과학대학 수학과

201621144

허 욱

**목차**

I. 아이디어 제안 ................................................................................................................03

1. 아이디어 선정 배경

2. 프로젝트 초기 목적

3. 프로젝트 기대 효과

II. 프로젝트 소개 ...............................................................................................................04

1. 주제

2. 방글라데시 교육 현황

III. 데이터 .............................................................................................................................05

1. 데이터셋

2. 데이터 인코딩

3. 데이터 적용

IV. 기계학습 알고리즘 모델 .........................................................................................07

1. 지도학습 - Decision Tree .

2. 지도학습 - Knn

3. 비지도학습 - k-means Clustering

4. 지도학습 - Decision Tree2

V. 프로젝트 결과 ...............................................................................................................12

1. 프로젝트 요약

2. 프로젝트 결론

VI. 참고자료 .........................................................................................................................14

**I. 아이디어 제안**

1. 아이디어 선정 배경 : 코로나19로 인한 교육 불평등 및 수업 운영의 어려움

- 코로나19 펜데믹으로 2020년부터 초,중,고,대학교뿐만 아니라 사교육기관에서도 비대면 온라인 수업을 실시함.

- 전국교직원노동조합이 실시한 설문조사에서 비대면 온라인 수업의 가장 큰 문제로 학습격차 심화(61.8%)를 지적했고, 한국교육학술정보원(KERIS)이 전국 초,중,고 교사들을 대상으로 실시한 설문조사에서도 응답자 80%가 코로나 19로 원격 수업이 진행된 이후 학생들 간의 학습 격차가 심해졌다고 응답함.

- 코로나 19 초창기 급작스러운 온라인 교육 시행으로 서울시교육청이 지정한 10곳의 '온라인 수업 시범학교'의 수업운영에 인터넷 연결 불안정, 기기 부족 등의 문제가 많이 발생하여 온라인 수업에 대해 혼란이 클 것이라는 우려가 많았음.

- 급작스러운 비대면 온라인 교육에 학생들이 잘 적응했을지에 대한 궁금증과 더불어 적응을 못했을 것이라는 예상과 함께 온라인 교육 적응도에 영향을 미치는 요인을 확인하고 분석해보고자 아이디어를 선정하게 됨.

2. 프로젝트 초기 목적

(1) 학생들이 온라인 교육(수업)에 얼마나 적응했는지를 확인하기 위함.

(2) 온라인 교육의 문제를 제기할 수 있고 개선되어야 함을 주장하기 위한 근거로 제시하기 위함.

3. 프로젝트 기대 효과

(1) 온라인 수업이 오프라인 수업만큼의 '보편적 교육' 역할을 할 수 있도록 그 방법들을 마련하고, 이를 통해 학생들이 온라인 수업을 통해 직면했던 어려움이나 여러 문제들에 대한 해결책을 고민할 수 있는 자료가 될 수 있음.

(2) 코로나 엔데믹으로 접어드는 시기에 오프라인 교육과 온라인 교육을 병행해도 괜찮은지에 대한 해답을 제시할 수 있고, 온라인 교육에 대한 학생들의 적응도를 향상시킬 수 있는 방법을 찾는 데 도움이 되는 자료가 될 수 있음.

**II. 프로젝트 소개**

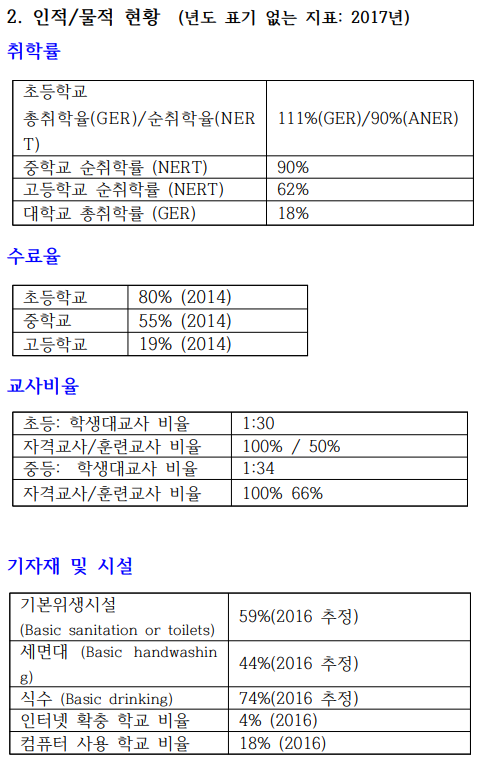
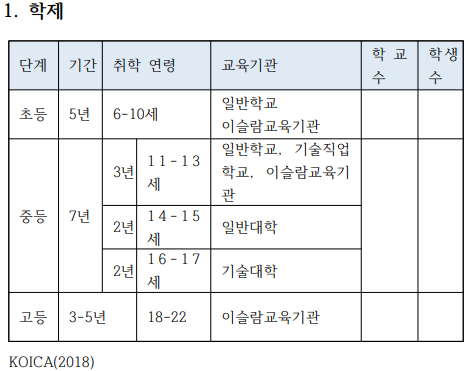
1. 주제 : 학생들의 적응도를 통한 온라인 교육 분석

- 2020년 12월부터 2021년 2월까지 방글라데시 학생들의 교육 적응도에 대한 온라인 설문결과 데이터셋을 바탕으로 온라인 교육을 분석함.

2. 방글라데시 교육 현황

(1) 정부 정책 방향 : 초,중,고등교육 의무화, 문맹퇴치, 과학기술 중점 교육을 통한 인간개발지수 향상 및 지식기반사회 구축, 해외 취업 확대 및 여성 취업률 향상, 일반 교육의 질적 개선 및 IT 분야 투자 확대 등을 통한 노동생산성 향상 등을 경제성장 추진 전략으로 채택함(관계부처 협동, 2016).

(2) 성인 및 중등교육 참여율 향상 필요 : 보편적 교육기회 확대 및 교육 참여율 향상을 위해 중,고등학교 IT 교육 강화 등을 추진함(관계부처 협동, 2016).



[ UNESCO GEM report (2019) ]

**III. 데이터**

1. 데이터셋

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | ⇦⇧13개 feature pieplots    3개 target value pieplot ⇨ |  |

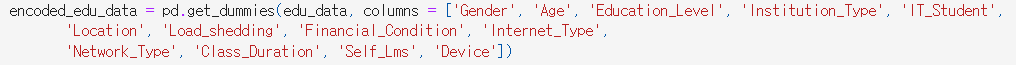
- 2020년 12월부터 2021년 2월까지 방글라데시 학생들의 교육 적응도 데이터셋으로 13개의 feature variable과 1205개의 인스턴스, target value 는 3개로 데이터에 라벨링되어 있음.



2. 데이터 인코딩

- 데이터셋이 모두 범주형 값을 갖는 Categorical Data 이므로 문자열 데이터의 범주형 값들을 숫자형으로 표현하기 위해 인코딩을 진행

(1) label 인코딩 : 파이썬 내장 map 함수를 활용 ⇩example

(2) One-Hot 인코딩 : Pandas 의 get\_dummies 함수를 활용 ⇩example

3. 데이터 적용

(1) label 인코딩

- label 인코딩은 범주형 Feature 값들을 0부터 1씩 증가하는 값으로 변환하는 방식이기 때문에 범주에 속한 값 간에 서열(순위)이 있는 변수(예를 들면, 성적, 직급, 나이 등)의 경우 모델 적용에 문제가 없지만 범주에 속한 값 간에 서열이 없는 변수(예를 들어 혈액형, 성별 등)의 경우 label 인코딩 시 숫자의 차이가 발생하게 됨.

- 따라서 숫자의 차이가 모델에 영향을 주지 않는 의사결정트리에 label 인코딩 데이터셋을 적용함.

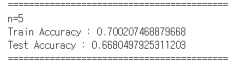
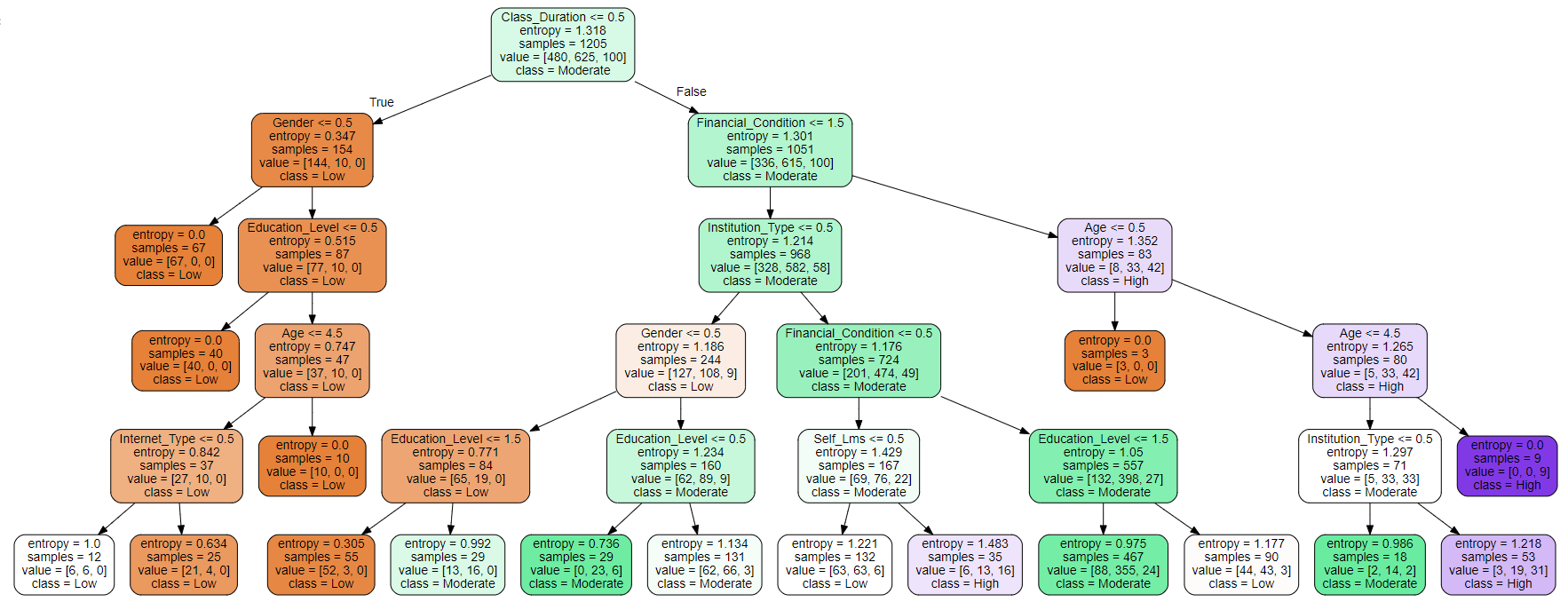
(2) One-Hot 인코딩

- One-Hot 인코딩은 범주형 변수를 0 또는 1 값을 가진 하나 이상의 새로운 feature로 변환하는 방식이기 때문에 전체적인 기계학습 모델에서 적용하기 좋은 알고리즘임.

- distance 기반의 Knn 알고리즘은 변수들의 숫자 차이가 결국에 distance에 영향을 주고 이는 분류에 영향을 주기 때문에 One-Hot 인코딩 데이터셋을 적용함.

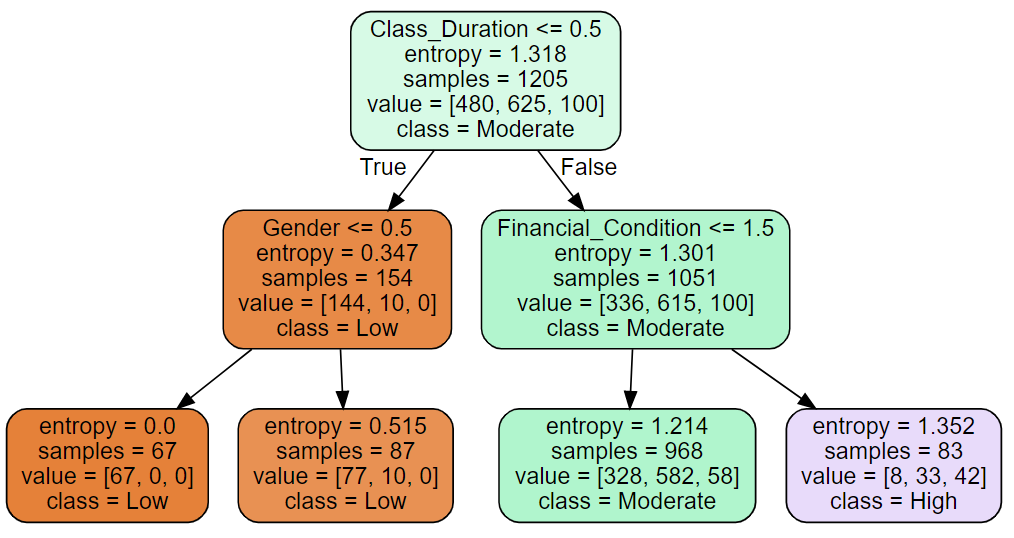
**IV. 기계학습 알고리즘 모델**

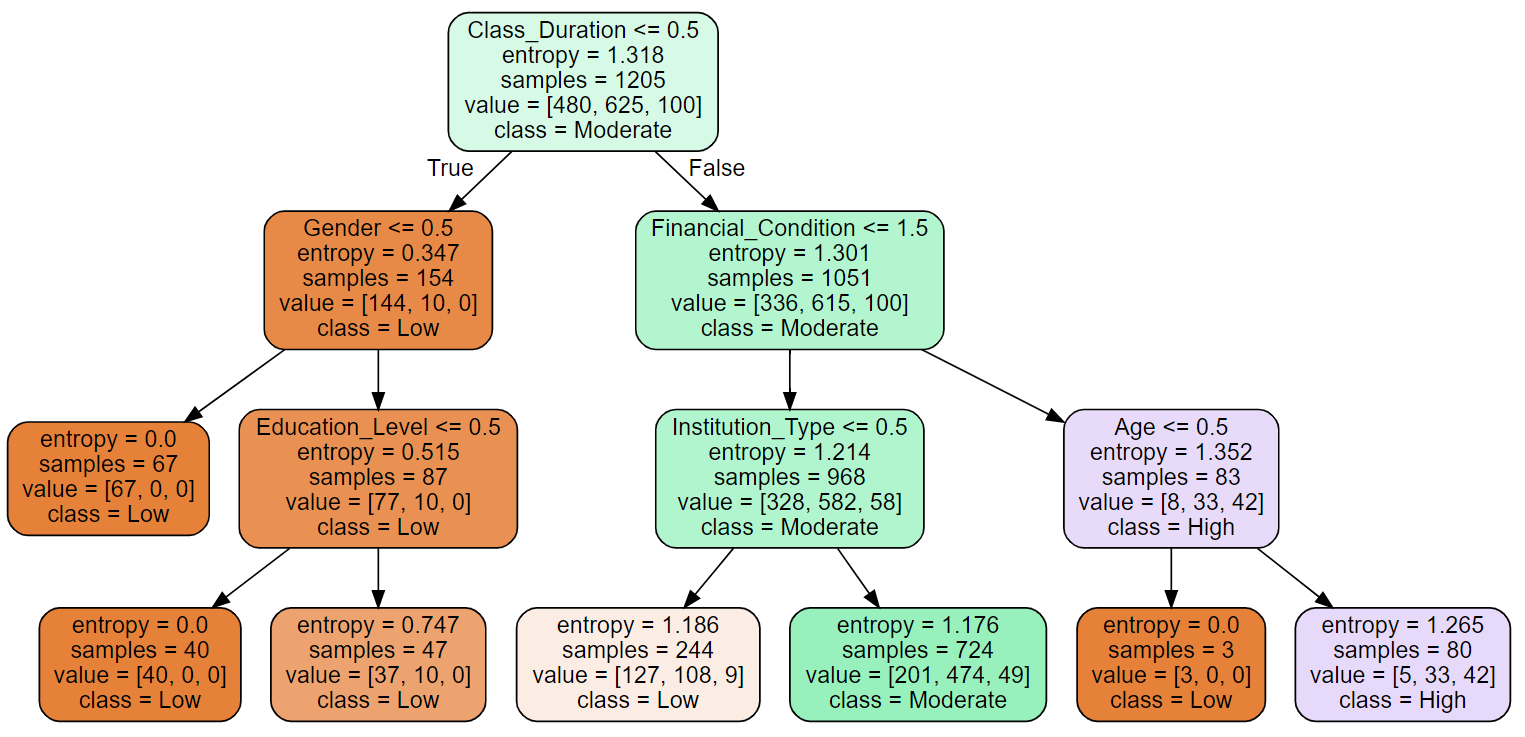
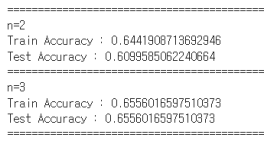
1. 지도학습 - Decision Tree(의사결정트리)

(1) 엔트로피 기준, 트리의 max\_depth=5 로 설정한 의사결정트리

- IG(Information Gain) 값 기반의 feature\_importance 기준으로 결정노드가 위에서부터 정해지는데 Class Duration(일간 수업 시간), Financial Condition(가정 재정 상태) 등의 변수로 데이터가 크게 분류됨을 확인할 수 있음.

- max\_depth=5 로 제한한 의사결정트리의 test 정확도는 약 0.67

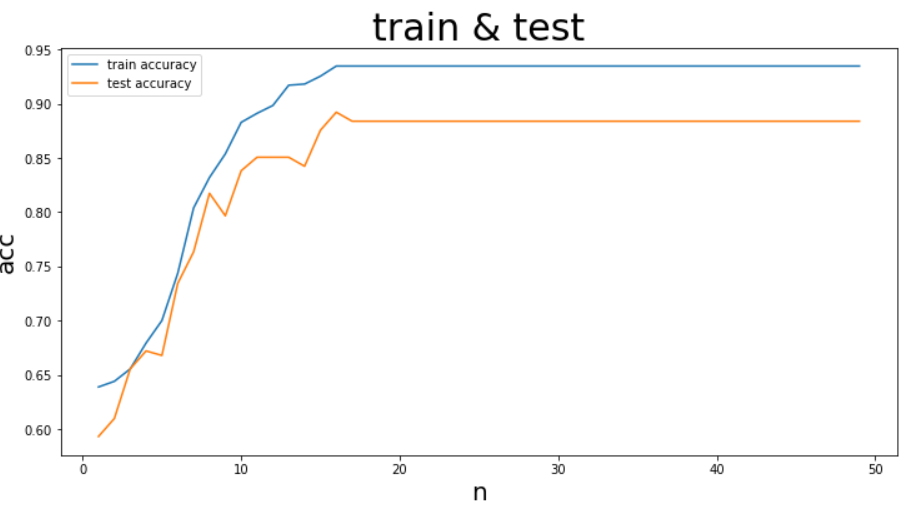
(2) 엔트로피 기준, 트리의 max\_depth=2, 3 으로 설정한 의사결정트리



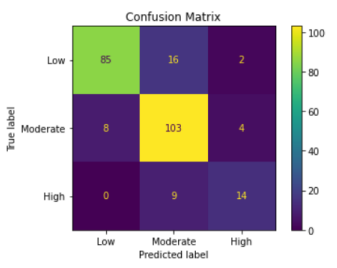
- max\_depth를 각 2와 3 으로 제한한 의사결정트리의 test 정확도는 트리의 깊이가 더 깊은 5일때보다 낮은 약 0.61 과 0.66 을 각각 보임.

(3) max\_depth 값 변화에 따른 Test Accuracy

- 의사결정트리가 오버피팅이 되기 쉬운 알고리즘이기 때문에 최적의 max\_depth 값 설정을 위해 max\_depth 값의 변화에 따른 Test 정확도 그래프를 시각화함.

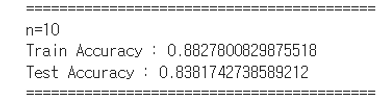
 - max\_depth =16 일 때 약 0.89로 최대의 test 정확도를 보임.

- 트리의 깊이가 깊어질수록 오버피팅됨을 train 정확도나 test 정확도 값이 유지되거나 떨어지는 모습을 통해 확인할 수 있으므로 트리의 깊이를 고려하여 max\_depth =10 일 때의 모델을 최적의 모델이라고 판단함.

(4) Confusion Matrix(혼동행렬) 시각화

- 최적이라고 판단한 max\_depth =10 일 때의 Confusion matirx 시각화 결과

- train\_test\_split을 통한 의사결정트리 학습과정에서 test\_size를 0.2 로 설정하였기 때문에 test 데이터의 인스턴스 수는 241개임.

 - 모델이 Low를 85개, Moderate를 103개, High를 14개 예측 성공하였고 정확도를 구해보면 acc=(85+103+14)/241 0.8382 로 코딩을 통해 구한 정확도와 일치함.

(5) k-fold Cross Validation(k겹 교차검증)

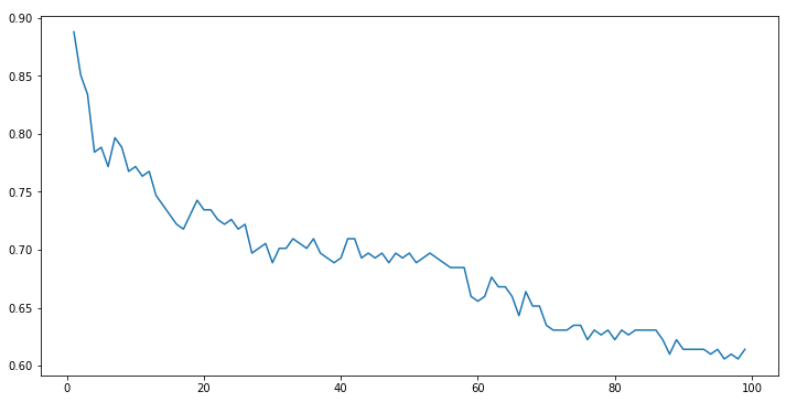
- k-fold 교차검증을 통해 모델의 성능을 확인함.

- k=2부터 49까지의 k-fold 교차검증 결과 최대 정확도 약 0.86 을 보임.

- k=10 일 때 약 0.85를 보임.

2. 지도학습 - Knn(K-Nearest-Neighbors)

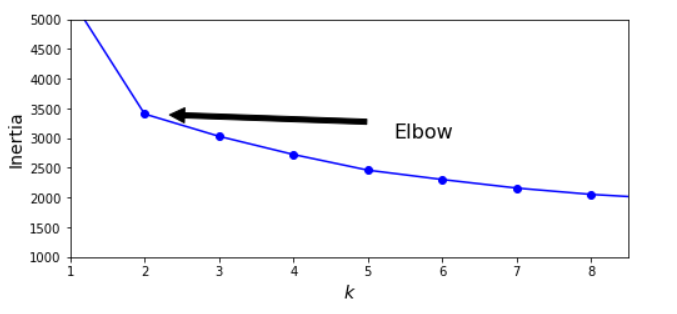
- Knn 알고리즘의 경우 고차원일수록 거리측도의 유효성이 떨어지고, 고차원 데이터 산점도 시각화가 불가능한 점 등을 미루어 보았을 때 프로젝트 문제 분석에 좋은 알고리즘은 아니라고 판단하여 정확도만 확인함.

 - k 값에 따른 정확도 시각화 결과, k 값이 커짐에 따라 언더피팅이 발생함을 확인할 수 있음.

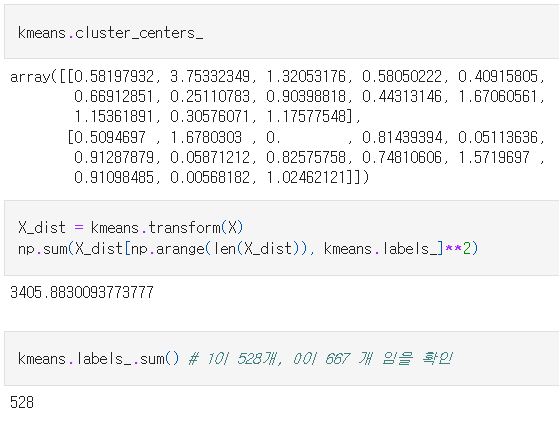
- 범주형 변수 데이터이므로 hamming distance 적용하였고, k=5 로 설정한 knn 모델의 정확도는 약 0.79 를 보임.

3. 비지도학습 - k-means Clustering(k-평균 군집화)

(1) k 설정

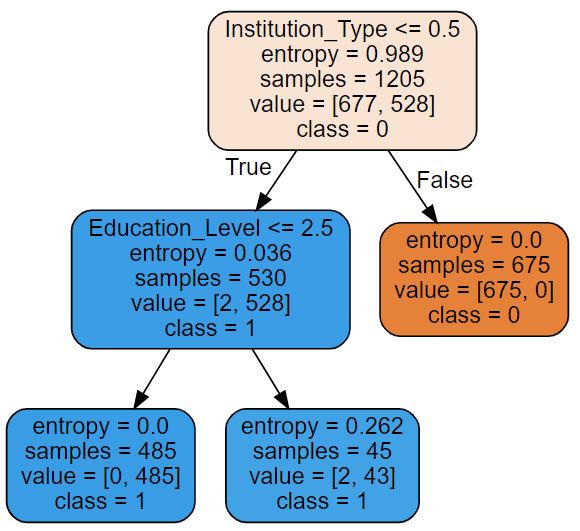
 - 데이터의 기존 라벨을 고려하지 않고 군집화를 진행함.

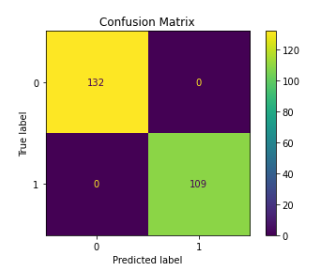
- 군집의 개수를 정하기 위해 elbow method 를 활용하였고, elbow point k=2 이기 때문에 2개의 군집으로 데이터셋을 나누는 군집화를 진행함.

(2) k=2 Clustering

- 군집화 결과 '0'군집으로 으로 677개의 데이터, '1'군집으로 528개 데이터가 클러스터링되었음을 확인할 수 있음.

4. 지도학습 - Decision Tree2

 - 군집의 center 값들로는 군집을 나누는 기준(변수의 유의미성)을 파악하기 어렵기 때문에 '0'과 '1'을 데이터를 나누는 새로운 라벨로 정하고 이 새로운 라벨을 활용하여 의사결정트리 모델을 생성해봄.

 - clustering label 자체로 데이터가 pure하게 나뉘어져 있을 것이므로, max\_depth=2 로 작게 설정하여 의사결정트리 모델을 시각화한 결과, Institutuion type(교육기관 성격), Education Level(학력)에 따라 데이터가 크게 나뉨을 확인할 수 있고, 분류가 잘 되었음을 확인할 수 있음.

- 혼동행렬을 시각화한 결과, test 데이터셋에

대해서는 완벽하게 분류되었음을 확인할 수 있음.

**V. 프로젝트 결과**

1. 프로젝트 요약

(1) 의사결정트리

- label 인코딩 데이터셋을 적용했고 기존 데이터의 라벨을 기준으로 의사결정트리 모델을 생성했을 때, 데이터를 나누는 유의미한 변수들은 Class Duration(일간 수업 시간), Financial Condition(가정 재정 상태) 등이었음.

- k=2 Clustering 기반의 군집 라벨을 데이터의 새 라벨로 하여 생성한 의사결정트리 모델에서 데이터를 나누는 유의미한 변수들은 Institutuion type(교육기관 성격), Education Level(학력)이었음.

- k-fold 교차검증을 통해 모델 성능의 향상을 확인함.

- 의사결정트리 모델은 오버피팅이 쉽게 발생한다는 점, 최적의 max\_depth 값 설정에 어려움이 있는 점 등을 한계점이었음.

(2) knn 알고리즘

- One-Hot 인코딩 데이터셋을 적용했기 때문에 기존 데이터셋보다 더 고차원 데이터 였으므로 거리측도의 유효성은 더 떨어졌을 뿐만 아니라 시각화가 불가능하다는 점에서 한계점이었음.

- PCA(주성분분석)을 통한 차원 축소로 고차원 데이터의 문제를 해결할 수 있다고 생각하지만 기존 feature의 의미를 잃어버린다는 점에서 고려해봐야함.

(3) k-means Clustering

- knn 알고리즘과 마찬가지로 distance 기반의 알고리즘이므로 프로젝트에 사용한 고차원 데이터의 Inertia(군집 응집도) 값이 크게 나왔지만 k=2 를 기준으로 군집화가 진행되었고, 2개의 군집으로 데이터를 나눌 수 있는 유의미한 변수들을 확인했기 때문에 유의미한 결과를 얻었다고 볼 수 있음.

2. 프로젝트 결론

- 변수들 간의 상관관계를 확인할 수 있는 Correlation Plot에서도 일간 수업 시간, 가정 재정상태, 교육기관 성격과 적응도 간의 약한 양의 상관관계(순서대로 0.33, 0.24, 0.27)를 보임. 교육기관 성격과 일간 수업시간 간의 상관관계(0.35)로 미루어 보았을 때, 교육기관의 차이가 학교에서 진행하는 일간 수업 시간(보충수업을 포함)에 영향이 있으며 이는 결국 온라인 교육 적응도에 영향이 있음을 시사함.

- 가정 재정상태와 적응도 사이의 상관관계를 통해 가정에서 재정 상태에 따른 온라인 교육 지원의 차이가 온라인 적응도에 영향을 어느 정도 준다는 것을 시사함.

- 프로젝트 데이터는 2020년 12월부터 2021년 2월까지의 온라인 설문결과 데이터이므로 코로나 유행을 반영한 교육적응도이고, 급작스럽게 진행되었던 불가피한 온라인 교육에 학생들이 적응을 잘 못했음을 알 수 있었음.(기존 라벨의 경우 8.3% 만이 온라인 교육에 잘 적응했다는 응답이었고, Clustering 을 통해 얻었던 2개의 라벨을 교육 적응도의 새로운 라벨로써 적응했다/적응못했다 라고 본다면 두 라벨의 차이가 크지 않았던 것으로 미루어 보아 단순히 적응했다/못했다 로 나누어도 적응하지 못한 학생 수가 많음을 알 수 있음)

- 코로나 엔데믹 시대로 접어드는 현재 온라인 교육과 오프라인 교육을 병행하려 한다면 학생들에게 오프라인만큼의 보편적 교육을 따라가기 위한 적절한 온라인 교육 지원(위에서 확인할 수 있었던 적응도에 영향을 줄 수 있었던 요인들에 대한 지원; 교수법 개선, 보충학습 확대, 재정적 지원)이 필요할 것임을 시사함.

**VI. 참고자료**

<https://www.kice.re.kr/resrchBoard/view.do?seq=632&s=kice&m=030109> [COVID-19 대응 온라인 개학에 따른 초, 중, 고등학교 원격 수업 실태 및 개선 방향 탐색]

<https://www.kaggle.com/datasets/mdmahmudulhasansuzan/students-adaptability-level-in-online-education>

<https://www.pressian.com/pages/articles/2020100511524581920>

<https://www.donga.com/news/article/all/20200430/100872622/1>

<https://todayisbetterthanyesterday.tistory.com/22?category=822147>

<https://dev-jm.tistory.com/28>

<https://hleecaster.com/ml-knn-classifier-example/>

[https://velog.io/@jhlee508/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-K-%ED%8F%89%EA%B7%A0K-Means-%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98](https://velog.io/@jhlee508/ë¨¸ì ë¬ë-K-íê· K-Means-ìê³ ë¦¬ì¦)

<https://bbdata.tistory.com/10>

[https://muzukphysics.tistory.com/entry/ML-8-%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-SVM-%EA%B8%B0%EB%B3%B8-%EA%B0%9C%EB%85%90%EA%B3%BC-%EC%9E%A5%EB%8B%A8%EC%A0%90-Support-Vector-Machine](https://muzukphysics.tistory.com/entry/ML-8-ë¨¸ì ë¬ë-SVM-ê¸°ë³¸-ê°ëê³¼-ì¥ë¨ì -Support-Vector-Machine)

<https://todayisbetterthanyesterday.tistory.com/19>

<https://inuplace.tistory.com/434>

<https://todayisbetterthanyesterday.tistory.com/38>