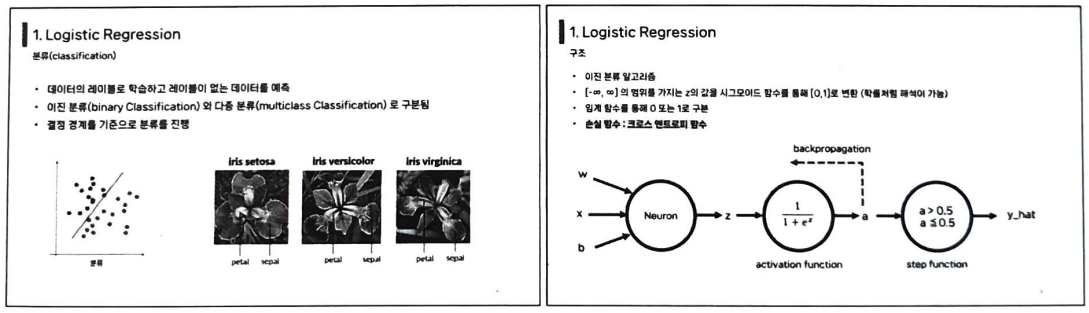
산학교육센터 T-SUM 주차별 활동 보고서

팀명	TSUANMI	장소	과학기술 2관 225호	일시 (주차)	4주차, 2024. 04. 30.(화), 19시 ~ 21시 30분 (150분)
활동 주제	이진 분류				

활동 내용

[로지스틱 회귀 학습]

4회차에서는 이진 분류 수업을 진행했다. 이전 회차까지 선형 회귀에 대한 수업을 진행했는데, 선형 회귀는 한 종속 변수와 여러 독립 변수 간의 관계를 파악하는 것으로, 집값 예측, 당뇨 수치 예측 등이 있다. 이러한 회귀 모델을 통해 특정 변수의 값을 분류할 수 있다. 예를 들어, 당뇨 수치를 통해 당뇨병이 있는지/없는지 집값을 예측한 값을 통해 집을 구매할 것인지/아닌지 등으로 분류할 수 있다. 이러한 분류 문제는 단순히 두 개로 분류하는 이진 분류 모델이 아니라, 여러 카테고리로 분류하는 다중 분류로도 이어질 수 있다. 당뇨병 수치를 더 잘게 나누어 매우 높음, 높음, 정상, 낮음, 매우 낮음과 같이 5개의 카테고리로 나눌 수 있다. 이러한 분류 모델은 이전에 배운 선형 회귀 모델에서 조금의 확장으로 분류 모델을 설계할 수 있다. 이전에 만든 선형 회귀 모델의 선형인 값을 비선형으로 바꿔주는 활성화 함수 하나만을 추가하면 분류 모델이 만들어진다. backpropagation 과정도 선형 회귀와 동일하고, 미분된 값도 동일하기 때문이다. 즉, 로지스틱 회귀는 새로운 개념이 아닌, 이전에 배운 개념의 추가한 것임을 멘티들에게 강조했다. 이론 수업이 끝난 후 실제 코드를 작성해봄으로써 이론 내용을 한 번 더 정리하고 코드로 분류 모델을 어떻게 구축하는지를 학습했다.



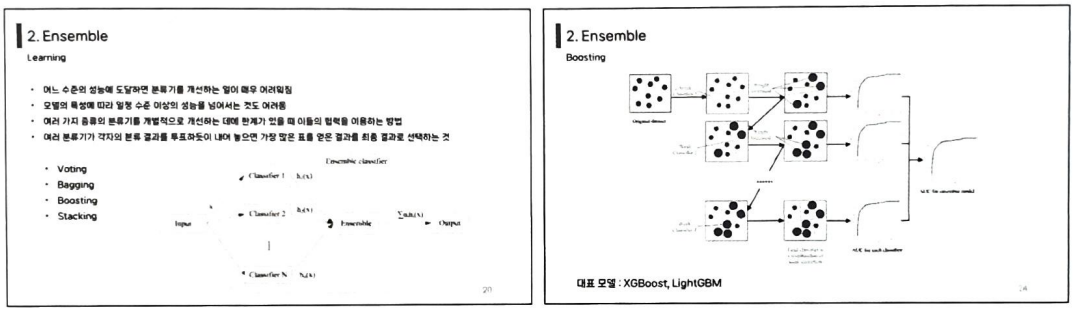
[그림 1] (좌) 분류 정의 (우) 로지스틱 회귀 구조

```
1 from sklearn.linear_model import LogisticRegression
2 from sklearn.metrics import accuracy_score
3
4 model = LogisticRegression()
5 model.fit(X_train, y_train)
6
7 pred = model.predict(X_test)
8 print(f'Acc : {accuracy_score(y_test, pred)}')
executed in 58ms, finished 20:28:48 2024-04-30
Acc : 0.956140350877193
```

[그림 2] 로지스틱 회귀 코드

[앙상블 기법 학습]

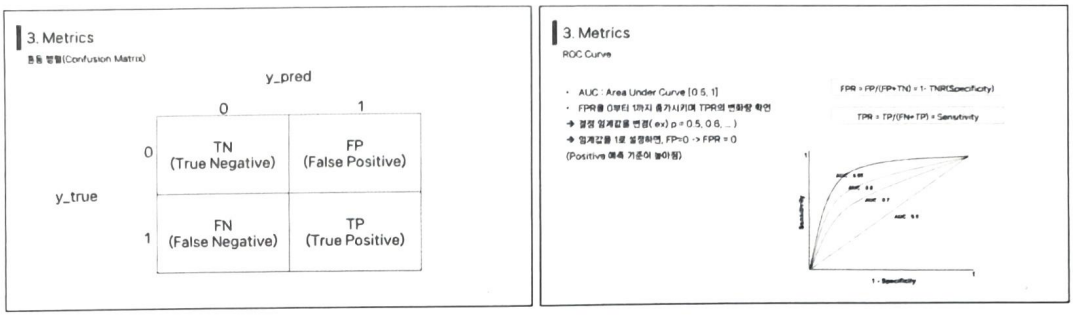
다음으로 앙상블 모델에 대해 학습했다. 우리가 의사 결정을 할 때 한 명의 전문가의 의견보다 여러 명의 전문가들의 의견을 모아서 결정하는 것이 더 나은 경우가 많다. 마찬가지로, 앙상블은 여러 모델의 의견을 모아서 더 나은 예측을 할 수 있도록 한다. 즉, 앙상블은 여러 개의 다양한 기본 모델을 조합하여 더 강력한 모델을 만드는 기법이다. 주로 사용되는 앙상블 기법으로는 Voting, Bagging, Boosting, Stacking 등이 있다. 이러한 다양한 방식으로 다수의 모델을 조합하여 예측을 수행하며, 그 특징에 따라 다양한 상황에 적용될 수 있다. 멘티들에게 앙상블 기법을 강조하여 수업을 진행했는데, 앞서 말했듯이, 단순히 로지스틱 회귀 모델 혹은 SVM 모델 등의 단일 모델만 사용해서 데이터 분석을 진행하지 않는다. 이는 좋은 결과를 보이지 못할 뿐만 아니라, 모델의 특성 중요도를 파악할 수 없는 등의 문제가 있다. 그래서, 이번 경진대회에서도 앙상블 모델을 사용하여 예측을 수행할 것이기 때문에 멘티들에게 앙상블 기법에 대해 잘 이해할 수 있도록 최대한 자세하고 반복적으로 설명해주었다.



[그림 3] (좌) 앙상블 기법 (우) Bagging 과정

[평가 지표 학습]

머신러닝 모델을 만든 뒤, 그 모델의 성능을 평가하는 것도 중요하다. 모델의 정확도와 같은 평가 지표를 확인하고 그 모델을 그대로 사용할 것인지 아니면 조금 더 학습시키고 사용할 것인지 정해야 한다. 특히, 분류 모델에서는 이 과정이 매우 중요하다. 단순히 정확도만을 보고 모델을 평가하면 안 된다. 예를 들어, 정확도가 90%라고 해서 반드시 좋은 모델은 아니다. 예를 들어, 100명의 환자 중 98명이 양성, 2명이 음성이라고 하자. 만약 이 모델을 정확도만을 보고 환자의 양성/음성 여부를 예측한다면, 단순히 100명 모두 양성이라고 해도 정확도는 98%가 되기 때문에 좋은 모델이라고 착각할 수 있다. 그렇기에 단순히 정확도만 보는 것이 아닌, 다른 지표 또는 혼동 행렬(confusion matrix)을 확인함으로써 모델을 평가하는 것이 좋다. 그렇다고 해서, 계속해서 모델을 복잡하게 만들면 이 모델은 오로지 훈련 데이터에서만 좋은 성능을 보이고 실제 예측 데이터에서는 좋은 성능을 보이지 못하는 과적합이 발생할 수 있다. 그렇기에 모든 일반적인 성능을 내는 모델을 만드는 것이 중요하다. 이 내용은 인공지능 분야에서 매우 중요한 내용이고, 새로운 모델을 만들 때, 자신이 만든 모델의 성능을 설명해야 하므로 이 내용을 정확하게 알고 있는 것이 중요하다. 멘티들에게 이 내용을 강조하여 단순히 수치만을 보고 모델의 성능이 좋다고 판단하지 않도록 했다. 이론 수업 후 간단한 퀴즈를 직접 손으로 계산해 보면서 평가 지표를 이해하도록 했고, 그 후 코드를 통해 그 계산한 값이 맞는지 확인해보았다.



[그림 4] (좌) 혼동 행렬 (우) ROC Curve

```
y_test = [0, 0, 1, 1, 0, 1, 1, 0, 0, 1]
pred =   [0, 1, 1, 1, 1, 0, 0, 1, 1, 1]

print(f'accuracy_score : {accuracy_score(y_test, pred)}')
print(f'precision_score : {precision_score(y_test, pred)}')
print(f'recall_score : {recall_score(y_test, pred)}')
print(f'f1_score : {f1_score(y_test, pred)}')
```

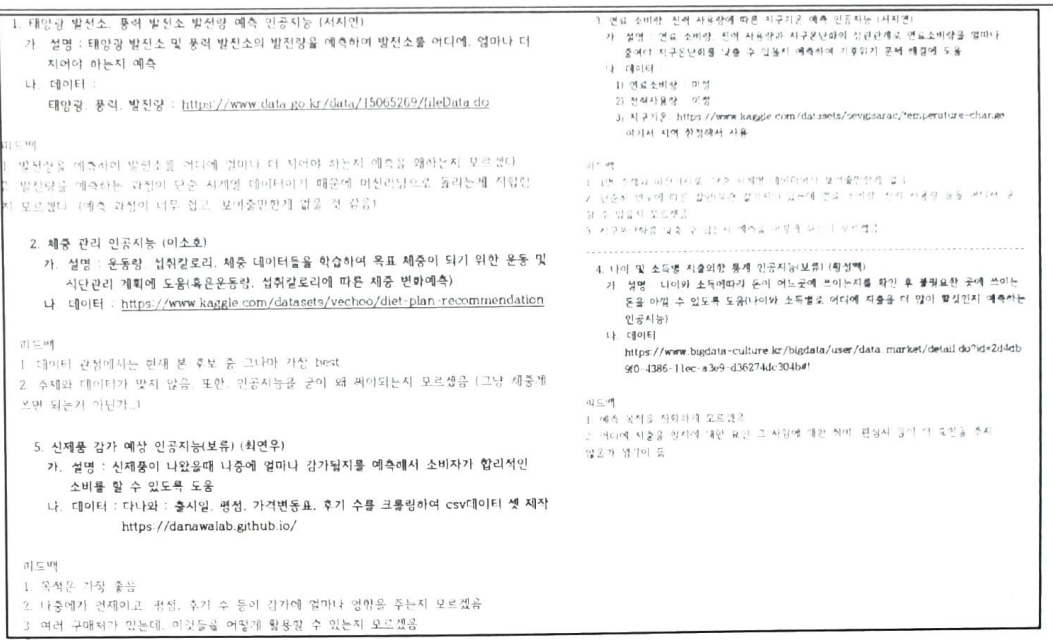
executed in 25ms, finished 20:16:15 2024-04-30

```
accuracy_score : 0.4
precision_score : 0.42857142857142855
recall_score : 0.6
f1_score : 0.5
```

[그림 5] 다양한 평가 지표 코드

[3주차 경진대회 활동 피드백]

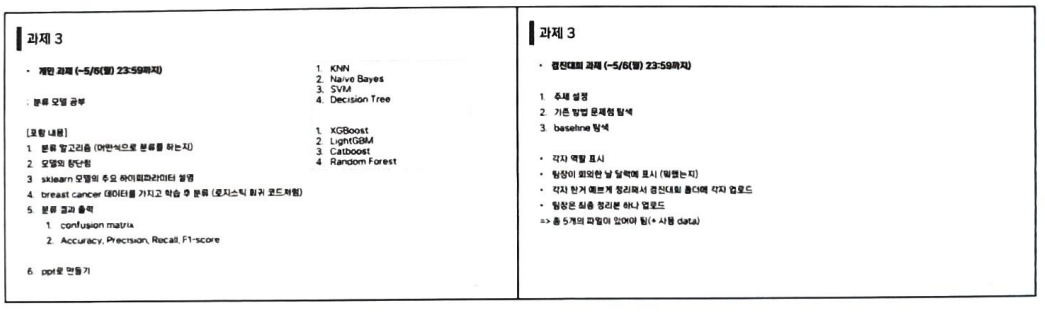
머신러닝 수업이 끝나고 멘티들이 약 3주 동안 준비한 경진대회 내용에 대해 피드백을 해주었다. 멘티들에게 이번 4주차 활동 전까지 경진대회 주제를 생각해오라고 공지했으며, 이번 회차에 그에 대한 피드백을 해주었다. 멘티들은 멘토링 활동 후 남아서 회의를 진행했으며, 활동이 없는 날에도 지속적으로 회의를 진행했다. 그 결과, 각자 5개의 주제를 정했으며, 그중 가장 좋은 주제 5개를 골라 그 주제에 대한 데이터를 탐색하여 멘토에게 보고서 형식으로 제출했다. 각 주제에 대한 피드백은 [그림 6]과 같다. 요약하자면, 아직 멘티들이 '머신러닝을 활용한 경진대회' 에 대해 정확하게 파악하지 못했다. 인공지능으로 어떠한 target 값을 예측을 할 수 있지만, 단순히 예측하는 것에 초점을 맞추다 보니 데이터 분석의 의미를 잃는 경우가 많다. 왜 인공지능 방법을 사용해야 되는지, 기존의 방법 말고 새로운 방법을 왜 사용해야되는지, 이것을 예측함으로써 얻을 수 있는 의의는 무엇인지 등을 생각하지 않다 보니, 주제와 데이터가 부실해지는 경우가 많다. 현재 멘티들도 이러한 상황에 있어, 이러한 피드백을 해주며, 단순히 인공지능을 사용하는 것이 아니라, 현실 세계의 문제점을 해결하는 것에 초점을 맞추어 경진대회 주제를 선정하도록 조언해주었다. 또한, 멘티들이 데이터를 선정할 때, 단순히 데이터의 '크기' 만을 보고 선정하는 것이 아니라, 데이터의 '특성(feature)' 을 잘 보고 결정하도록 피드백 해주었다. 결국 데이터의 품질이 좋아야 좋은 결과가 나오는데, 데이터가 어떤 정보를 담고 있는지 명확하게 알아야 유의미한 분석을 진행할 수 있기 때문에, 데이터를 유심히 살펴보도록 강조했다.



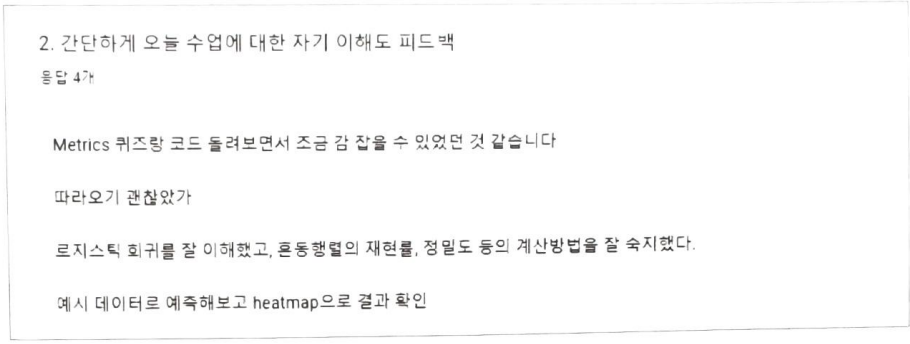
[그림 6] 3주차 경진대회 활동 피드백

[4주차 과제]

4주차 과제로 개인 과제, 경진대회 과제를 부여했다. 개인 과제로는 다양한 분류 모델을 각자 2개씩 조사해서 ppt로 만드는 것이다. 이번 주차에서 로지스틱 회귀 모델을 배웠는데 분류 모델은 매우 많아서 이를 활동 시간에 모두 다루는 것은 무리가 있다. 그래서 멘티들이 각자 공부하고 자료를 만듦으로써 자신의 학습 능력을 향상시키고, 다양한 분류 모델에 대한 깊은 이해도를 갖출 수 있다. 또한, 이러한 과정을 통해 멘티들은 자신이 조사한 모델의 장단점, 주요 특성, 성능 등을 교류하며, 여러 분류 모델 간의 비교 분석을 통해 문제 해결에 적합한 모델을 선택하는 능력을 개발할 수 있을 것으로 기대된다. 경진대회 과제로는 이번 피드백을 통해 주제 구체화, 데이터 선정, 주제 목적 등을 다시 생각해보는 것이다. 이번 과제를 통해 경진대회 주제 및 목적이 매우 구체화될 것으로 기대되고, 이는 전반적인 데이터 분석의 과정을 설계하고 분석하는데 큰 도움이 될 것으로 기대된다.



[그림 7] 4주차 과제

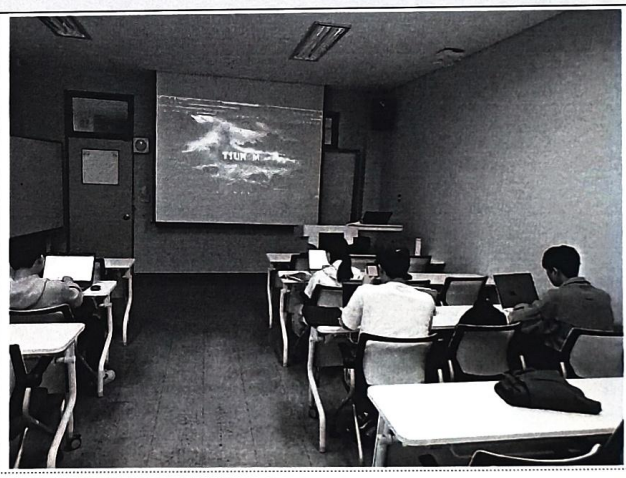


[그림 8] 멘티 셀프 피드백

멘티별 활동내용	멘티 이름	활동 내용 및 역량 증가 정도
	황성백	수학적 이해력이 매우 높고, 수업 내용을 잘 이해하고 있는 것으로 보인다.
	이소호	질문을 자주 하고, 대답도 열심히 하면서 수업에 매우 적극적으로 임한다.
	서지연	큰 어려움 없이 수업 내용을 잘 따라오는 것으로 보인다.
	최연우	프로젝트에 적극적으로 임하며, 자신의 아이디어를 적극적으로 잘 표현한다.

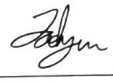
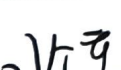
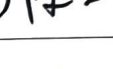
증빙 사진

[시작]



[끝]

참여자 서명부

구분	소속	학번	성명	서명
멘토	컴퓨터융합소프트웨어학과	2019270641	박재현	
멘티	컴퓨터융합소프트웨어학과	2022270608	황성백	
멘티	컴퓨터융합소프트웨어학과	2022270611	이소호	
멘티	컴퓨터융합소프트웨어학과	2022271323	서지연	
멘티	컴퓨터융합소프트웨어학과	2023270639	최연우	
지도교수		노경숙 교수님