2024년 2학기 통계분석 및 실습

머신러닝과 딥러닝 모델을 활용한 ELL 학습자 에세이 자동 채점 모델

학번: 202111168

학과: 응용통계학과

이름: 김다빈

날짜: 2024/12/09

목차

머신러닝과 딥러닝 모델을 활용한
ELL 학습자 에세이 자동 채점 모델

1. 서론

1) 배경 및 필요성

ELL(English Language Learners)는 영어를 제2언어로 배우며, 영어 교육을 별도로 받을 기회를 가진 학생들을 의미한다. 이들은 기본적인 글쓰기 연습의 기회가 부족하여 글쓰기 능력 향상이 어려운 상황이다. 또한, 기존의 자동화 모델들은 ELL 학생들에게 맞춤형 피드백을 제공하는 데 한계가 있다. 이에 따라, 본 프로젝트는 자동화된 피드백 모델을 통해 ELL 학생들의 글쓰기 능력을 효과적으로 지원하고자 한다.

2) 목적

본 프로젝트의 목적은 8학년부터 12학년까지의 **ELL 학생들의 영어 글쓰기 능력을 평가하는 모델을 구축**하는 것이다. ESL(English as a Second Language) 학생들은 다른 학생들에 비해 글쓰기 경험이 부족한 경우가 많아, 이들에게 정확한 피드백을 제공할 수 있는 모델이 필요하다. 이를 통해 ELL 학생들이 더욱 풍부한 글쓰기 기회를 가질 수 있도록 하고, 교사들의 채점 부담을 경감시키는 것이 본 프로젝트의 궁극적인 목표이다.

2. 본론

1) 데이터 수집

-kaggle(https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data)

구분	내용
train.csv	· text_id으로 데이터가 고유하게 구분되며, 이후 7개의 평가기준(cohesion, syntax, vocabulary, phraseology, grammar, conventions)이 포함되어있다. · 총 3,911개의 에세이가 존재한다.
test.csv	· full_text와 text_id만 제공된다. 해당 데이터를 가지고 모델의 성능을 검증한다. · 3개의 새로운 에세이가 존재한다.
sample_submission.csv	· 케글에 제출해 평가받을 파일이다.

2) 데이터 설명

A. 데이터 설명 개요

본 데이터는 8학년부터 12학년까지의 ELL 학생들이 작성한 에세이로 구성되어 있으며, 3911개의 에세이 데이터로 이루어져 있다(총 3912행, 8열). 각 에세이는 6가지 평가 항목(cohesion, syntax, vocabulary, phraseology, grammar, conventions)으로 평가되었고, 점수는 1점에서 5점까지 0.5점 간격으로 채점되었다. 각 컬럼에 대한 구체적인 설명은 아래와 같다.

B. 데이터 칼럼 설명

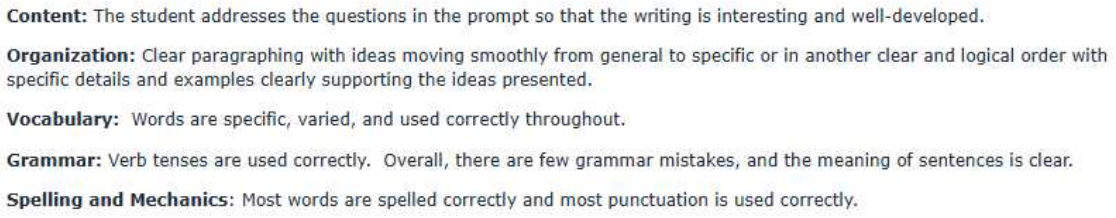
칼럼명	칼럼 설명	자료의 종류	예시
text_id	데이터를 구분 짓는 고유 식별자이다.	질적자료	없음.
full_text	각 에세이의 전체 텍스트이다.	질적자료	없음.
Cohesion	글 전체의 주제가 자연스럽게 연결되어 흐름이 잘 이루어지는 것을 의미한다.	양적자료	"첫째로"나 "그러므로"와 같은 연결어를 사용하여 문장 간의 관계를 명확하게 표현한다.
Syntax	문장 구조와 배열을 의미한다.	양적자료	"그는 뛰었다."와 같은 간단한 문장뿐 아니라, "비가 오고 있었지만, 그는 계속 뛰었다."와 같이 복잡한 구조도 사용한다.
Vocabulary	적절한 어휘 사용을 의미한다.	양적자료	"행복하다" 대신 "기쁘다", "즐겁다" 등의 다양한 표현 사용한다.
Phraseology	어구와 관용구를 얼마나 정확하고 자연스럽게 사용하는지를 평가한다.	양적자료	"새로운 장을 열다"와 같은 적절한 관용 표현 사용한다.
Grammar	문법의 정확성을 평가한다.	양적자료	“나는 학교에 갔다”와 같은 적절한 시제를 사용한다.
Conventions	표기상의 정확성을 평가한다.	양적자료	"I went to the store."에서 문장 시작은 대문자로, 문장 끝은 마침표로 마무리한다.

캐글에서 제공한 데이터는 데이터 구조와 평가 기준에 대한 명확한 설명이 부족하기에 **train 데이터에 대한 분석을 집중적으로 함으로써 본 분석에 대한 신뢰성과 타당성을 높이고자한다.**

C. 타 에세이 평가지표와 비교 분석

해당 데이터는 평가 기준과 각 항목의 가중치에 대한 정보가 명확히 제공되지 않았다. 따라서, 다른 ELL 평가 기준과 점수 분포를 바탕으로 분석을 유추할 필요가 있다.

I. 평가 기준 : 미국 내 ELL 학생들의 글쓰기 평가는 주별 교육 정책과 교재에 따라 기준이 다양하여, 통일된 평가 기준이 존재하지 않는다. 그러나 본 프로젝트에서 사용한 평가지표는 이러한 다양한 기준과 유사한 항목들로 구성되어 있어, ELL 학생들의 글쓰기 능력을 타당하게 측정할 수 있을 것으로 판단된다.



Content: The student addresses the questions in the prompt so that the writing is interesting and well-developed.

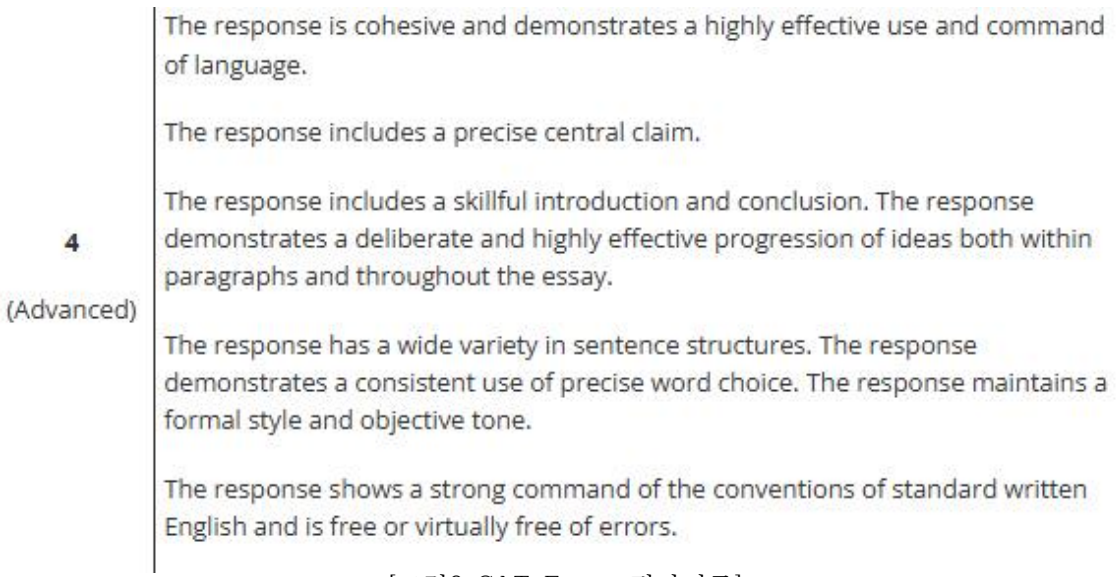
Organization: Clear paragraphing with ideas moving smoothly from general to specific or in another clear and logical order with specific details and examples clearly supporting the ideas presented.

Vocabulary: Words are specific, varied, and used correctly throughout.

Grammar: Verb tenses are used correctly. Overall, there are few grammar mistakes, and the meaning of sentences is clear.

Spelling and Mechanics: Most words are spelled correctly and most punctuation is used correctly.

[그림1 Glendale Community College의 에세이 평가기준]



4 (Advanced)	The response is cohesive and demonstrates a highly effective use and command of language. The response includes a precise central claim. The response includes a skillful introduction and conclusion. The response demonstrates a deliberate and highly effective progression of ideas both within paragraphs and throughout the essay. The response has a wide variety in sentence structures. The response demonstrates a consistent use of precise word choice. The response maintains a formal style and objective tone. The response shows a strong command of the conventions of standard written English and is free or virtually free of errors.

[그림2 SAT Essay 평가기준]

II. 가중치에 대한 분석 : 미국의 SAT나 TOEFL과 같은 평가에서는 논증력과

문법의 중요도가 다르게 적용되는 경우가 있다. 이를 참고하여, 본 데이터에서도 평가 지표별로 가중치를 두지 않고 동일한 비중으로 분석을 진행한다.

III. ELL 학생들의 언어적 배경 : 미국에서 ELL 학생들의 모국어는 주로 스페인어로, 전체 ELL 학생의 76.6%가 스페인어를 사용한다. 그 외에 아랍어, 중국어, 베트남어 등도 주요 모국어로 나타나므로,본 데이터를 분석할 때 스페인어 사용자가 많을 것이라는 전제를 바탕으로 한다.

IV. 다른 평가의 점수와의 비교 : WIDA ACCESS 시험의 데이터는 보호되어 있어 공개된 점수 분포가 없다. 그러나 SAT와 TOEFL 등의 시험에서 글쓰기 평가의 [그림3, 그림4] 평균 점수와 분포를 참고하여, 본 데이터의 중앙값과 평균값이 실제 ELL 학생들의 글쓰기 점수 분포와 유사할 것이라는 가정을 세울 수 있다.

점수	백분위
1600-1520	99+
1510-1290	99-90
1280-1190	89-90
1180-1120	78-70
1110-1060	69-60
1050-1010	58-50
1000-960	48-40
950-910	38-31
900-840	29-20
830-780	18-11
770-630	10-1
620-400	1-

〈자료: 칼리지보드〉

[그림3 전체 응시자 SAT 백분위 점수대]

분류	읽기	듣기	말하기	쓰기	총점
전체 응시자	22.3	22.3	20.8	21.2	87
고등학생	20.3	20.9	20.8	20.4	82
대학생	21.5	21.8	21.0	21.0	85
대학원생	23.4	23.2	21.0	21.7	89

[그림4 응시자의 현 교육 수준으로 분류한 TOEFL iBT 시험 평균 점수]

V. 평가원 수 : 케글 데이터에서는 에세이를 채점한 채점자의 수가 명시되어

있지 않다. 일반적으로 SAT나 ACT와 같은 표준화 시험은 두 명의 채점자가 각각 독립적으로 에세이를 채점하며, 두 채점자의 점수 차이가 일정 수준 이상일 경우 세 번째 채점자가 추가로 검토하는 방식을 따른다. 반면, 학교 과제는 대개 한 명의 교사가 채점하고, 전국 대회나 연구 평가와 같은 중요한 평가는 두 명 이상의 채점자가 참여하는 경우가 많다. 이를 바탕으로 볼 때, 대규모 시험이나 대회일수록 채점자가 2명 이상일 가능성이 높다. 현재 데이터는 8학년에서 12학년 사이의 ELL 학생들이 학교에서 수행한 에세이를 바탕으로 언어 능력을 평가하기 위해 수집된 것으로, 일반적으로 학교에서 수행되는 과제는 교사 한 명이 채점하는 경우가 많다. 따라서 이 데이터의 채점자는 1명일 가능성이 높다고 판단할 수 있다.

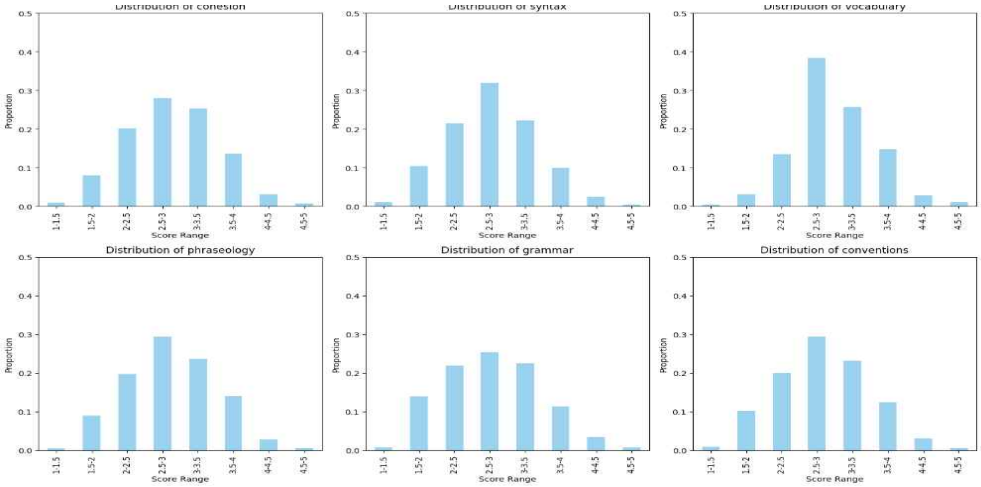
D. **변수별 분포 및 구조 파악**

I. 데이터 결측치와 이상치 유무

해당 데이터에는 결측치와 이상치가 없는 것으로 파악된다. 여기서 이상치란 1점부터 5점까지 0.5 간격을 벗어나는 값을 의미한다.

II. 변수별 분포

[그림 5]를 통해 모든 평가지표의 점수 분포가 정규분포를 따르며, 각 평가지표의 분포가 유사함을 확인할 수 있다. 또한, [그림 6]에서는 모든 평가지표의 분포 구조가 유사하다는 점을 다시 한 번 확인할 수 있다.



[그림5 각 평가지표에 대한 점수별 비율]

	cohesion	syntax	vocabulary	phraseology	grammar	conventions
count	3911.000000	3911.000000	3911.000000	3911.000000	3911.000000	3911.000000
mean	3.127077	3.028254	3.235745	3.116850	3.032856	3.081053
std	0.662542	0.644399	0.583148	0.655997	0.699841	0.671450
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	2.500000	2.500000	3.000000	2.500000	2.500000	2.500000
50%	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000
75%	3.500000	3.500000	3.500000	3.500000	3.500000	3.500000
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

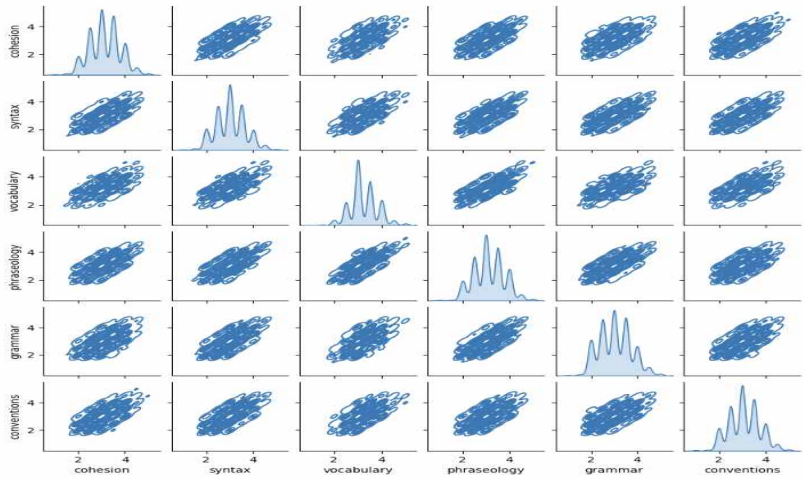
[그림6 각 평가지표에 대한 점수 기초 통계량]

III. 변수 간 상관관계, 연관성 파악

일부 변수 간에는 상관관계가 강하게 나타나는 반면, 연관성이 없어 보이는 변수도 있어 실제로 어떤 관련성을 갖고 있는지 확인하기 위해 분석을 진행하였다.

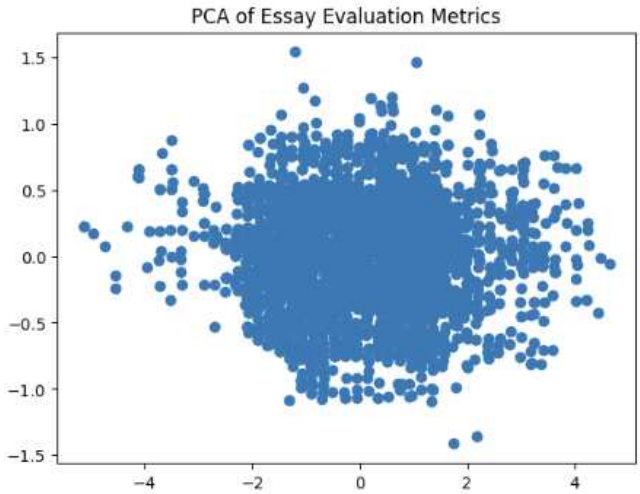
- 산점도와 커널 밀도 함수를 통해 변수 간 상관관계를 파악한 결과, 모든 변수의 상관계수가 0.67 이상의 값을 보였다. 또한, 비선형 관계나 다른 특이한 추이는 나타나지 않았다.

[그림7 변수별 상관관계에 대한 히트맵]



[그림8 변수별 상관관계에 대한 커널밀도함수]

- PCA 분석을 통해 변수 간 관계성을 추가로 분석해보자.
에세이 데이터의 변수들이 많고, 변수 간 상관관계가 존재하는 것으로 보여 차원 축소를 통해 데이터를 보다 간결하게 시각화하고 주요 패턴을 파악하기 위해 PCA를 진행하였다.



[그림9 데이터에 대한 주성분 분석 PCA]

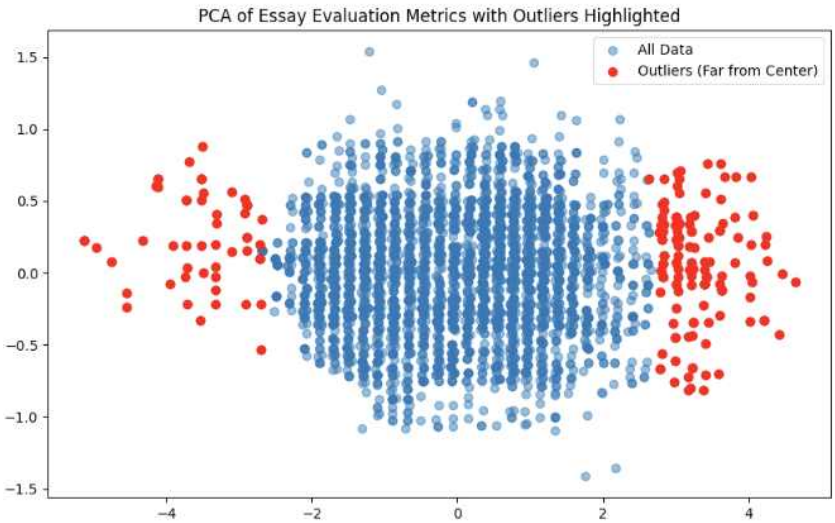
아래 [그림10]을 통해, 두 개의 주성분만으로 데이터의 80% 이상의 설명력을 확보할 수 있으며, 데이터가 대부분 중앙에 몰려 있음을 확인할 수 있다. 이는 에세이들에 대한 점수가 전반적으로 크게 차이나지 않고 대부분 유사한 평가 점수를 받았음을 의미한다. 중앙에 모인 것은 모든 점수가 대체로 고르게 분포한다는 의미로 해석될 수 있다.

```
[0.73864783 0.06596901]
```

[그림10 PCA에대한 설명 분산 비율]

- PCA 결과 추가 분석

이제 PCA 결과에서 중앙에 몰려 있지 않은 데이터들의 특징을 알아보자. 상위 5% 거리의 데이터들만 추출하여 분석한 결과[그림 11], 해당 데이터들은 점수 간 차이가 커서 중앙에서 벗어난 것이 아니라, 특정 평가 기준에서 점수가 모두 극단적으로 낮거나 높은 경우 PCA 중앙에 몰려 있지 않음을 파악할 수 있었다.거나 높은 경우 pca 중앙에 몰려있지 않음을 파악할 수 있었다.



[그림11 중앙에서 떨어진 상위 5%를 빨간색으로 표시한 PCA]

또한, PCA 결과에서 왼쪽으로 떨어진 데이터와 오른쪽으로 떨어진 데이터에 대한 추가 분석을 진행하였다. 왼쪽으로 떨어진 데이터는 점수가 극단적으로 낮은 데이터[그림 12]들이고, 오른쪽으로 떨어진 데이터는 점수가 극단적으로 높은 데이터[그림 13]임을 확인할 수 있었다.

```
왼쪽 이상치 데이터들의 각 평가 기준별 평균값:
cohesion        1.567308
syntax          1.567308
vocabulary      1.875000
phraseology     1.730769
grammar         1.653846
conventions     1.567308
dtype: float64
```

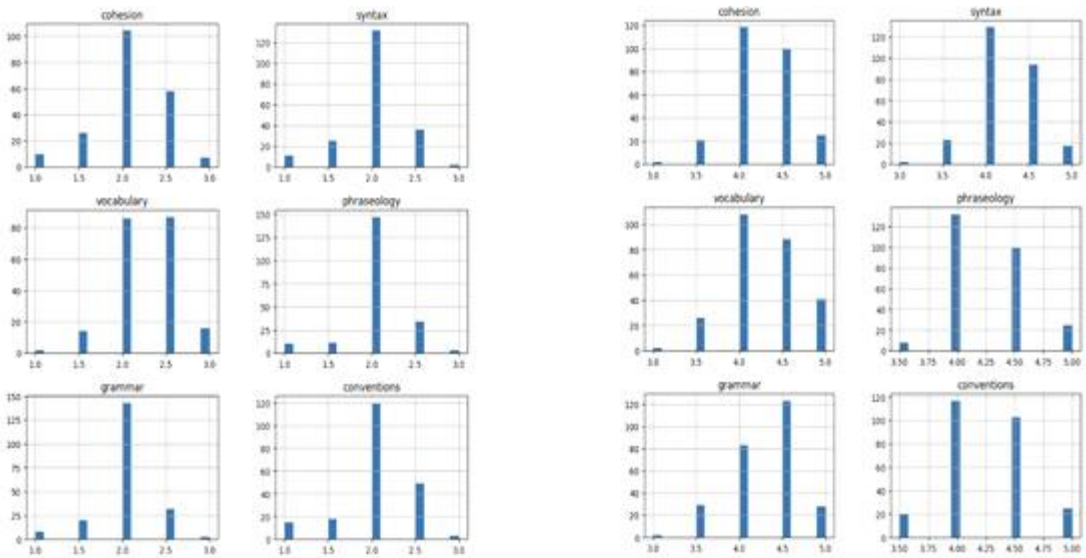
[그림12 왼쪽 이상치 데이터들의 각 평가기준별 평균값]

```
오른쪽 이상치 데이터들의 각 평가 기준별 평균값:
cohesion        4.367647
syntax          4.415441
vocabulary      4.507353
phraseology     4.466912
grammar         4.477941
conventions     4.422794
dtype: float64
```

[그림13 오른쪽 이상치 데이터들의 각 평가기준별 평균값]

특히, 오른쪽 데이터가 왼쪽 데이터에 비해 아래로 더 넓게 분포된 현상에 대해 추가적인 분석을 진행한 결과, 이러한 분포의 차이는 다음 두 가지 요인에서 비롯된 것으로 나타났다.

첫째, 데이터 분포의 다양성 차이가 주요 원인으로 분석되었다. 왼쪽 데이터는 평균적으로 2점을 중심으로 좁게 분포된 반면, 오른쪽 데이터는 4점을 중심으로 상대적으로 더 넓게 퍼져 있었다. 오른쪽 데이터는 각 평가 기준에서 점수가 4점 이상으로 다양하게 분포되어 있어, PCA 변환 시 더 넓은 변동성을 나타낸 것으로 보인다.



[그림14 왼쪽 데이터의 변수별 분포와 오른쪽 데이터의 별수별 분포]

둘째, PCA2에 대한 기여도가 오른쪽 데이터가 왼쪽 데이터에 비해 2배 이상 높게 나타난 점도 주요 원인으로 파악되었다. 오른쪽 데이터는 PCA2 기여도 합계가 7.55로 계산되었으며, 이는 왼쪽 데이터의 PCA2 기여도 합계인 3.62와 비교했을 때 2배 이상의 차이를 보인다.

Left Outliers PCA2 Contributions (Mean):

cohesion	-1.300655
syntax	0.004465
vocabulary	-0.228081
phraseology	0.226601
grammar	1.474466
conventions	-0.384216

Right Outliers PCA2 Contributions (Mean)

cohesion	-2.668840
syntax	0.009437
vocabulary	-0.432959
phraseology	0.478308
grammar	3.152020
conventions	-0.809727

[그림15 PCA2에 대한 왼쪽 데이터의 기여도와 오른쪽 데이터의 기여도]

결론적으로, 오른쪽 데이터가 왼쪽 데이터에 비해 아래로 더 넓게 분포된 이유는 높은 점수 구간에서 평가 기준들의 분포가 다양하게 퍼져 있고, PCA2에 대한 기여도가 상대적으로 높기 때문이다.

- 클러스터링 분석을 수행해보았다.
에세이 데이터가 특정 그룹으로 나뉠 가능성이 있다고 판단하여, 클러스터링을 통해 비슷한 패턴을 가진 데이터를 그룹화하였다. 이를 통해 각 에세이가 어떤 기준으로 묶이는지, 학생들의 평가 점수가 유사한 특성을 가진 에세이들을 분류할 수 있는지 확인하고자 하였다.
분석 결과, 다음과 같은 군집 특성을 발견할 수 있었다.
> 0번 군집: 모든 평가지표에서 낮은 점수를 받은 에세이들로 구성됨.
> 1번 군집: 모든 평가지표에서 높은 점수를 받은 에세이들로 구성됨.
> 2번 군집: 모든 평가지표에서 중간 점수를 받은 에세이들로 구성됨.

하나의 평가지표에서 낮은 점수를 받은 에세이는 다른 평가지표에서도 낮은 점수를 받는 경향이 있음을 확인할 수 있었다. 이는 데이터가 매우 잘 정제되어 있으며, 평가 점수 간 일관성이 높다는 것을 의미한다.

E. 점수를 잘 받은 사람과 못 받은 사람들의 에세이에 대한 분석

높은 점수를 받은 상위 10개의 에세이와 낮은 점수를 받은 하위 10개의 에세이에 대한 분석을 진행하였다. 에세이 평가에 대한 전문가가 아니기 때문에, 정량적 평가가 가능한 syntax와 grammar는 분석할 수 있지만, 정성적 평가가 필요한 다른 요소에 대한 평가는 미흡할 수 있다는 점을 고려해야 한다.
우선, 상위 10개의 에세이를 분석해본 결과, 논리적으로 잘 연결되어 있으며 문장 구조가 명확하고 안정적이었다. 어휘는 비교적 풍부하고 적절하게 사용되

었으며, 어구 표현도 대체로 적절하게 이루어졌고, 문법적 오류가 많지 않음을 확인할 수 있었다. 물론 일부 문법적 오류와 철자 오류는 존재하지만, 반복적인 철자 오류나 눈에 띄는 문법적 오류가 없다는 것이 상위 10개 에세이의 특징이다.
하위 10개의 에세이 분석 결과, 여러 감점 요인이 존재했다. 예를 들어, 동일한 문장이 여러 번 반복되어 응집력이 떨어지거나, 스페인어로 작성된 에세이도 있었다. 또한, 대부분이 비문이 많고 어휘 선택이 매우 제한적인 경우가 많았다.
에세이를 직접 분석한 결과, 문법적 오류 1개당 0.5점 감점과 같이 명확하게 정해진 감점 기준은 없지만, 눈에 띄는 오류가 많을수록 점차 점수가 깎이는 일관된 채점 방식이 있음을 파악할 수 있었다.

F. 평가 기준 별 점수 차이
데이터 분석 결과, 평가 기준 별 점수 차이의 최대값은 2점이며, 이러한 차이를 보이는 데이터는 총 12개에 불과하다. 또한, 점수 차이가 1.5점 이상인 데이터는 전체의 9%에 불과하여, 대부분의 에세이 점수 분포가 매우 일관되고 선형적으로 나타난다는 사실을 확인할 수 있다.

캐글 데이터는 구조와 평가 기준에 대한 설명이 부족하기에, train 데이터에 대한 철저한 분석을 통해 데이터의 신뢰성과 구조적 특징을 파악하고자 하였다. **이를 통해 데이터의 일관성과 안정성을 확인하고, 모델 구축에 적합한 데이터임을 검증했다.**

3) 데이터 전처리

텍스트 데이터 전처리는 대소문자 통일, 줄임말 처리 등을 통해 데이터 품질을 개선하여 모델이 일관된 패턴을 학습하도록 돕기 때문에, 전처리를 시작해보려고 한다.

A. train 데이터와 test 데이터의 full_text를 합친다.

에세이 데이터를 한 번에 전처리하기 위해 train과 test 데이터의 full_text를 결합한다.

B. 소문자 변환
모든 텍스트를 소문자로 변환하여 대소문자 차이로 인한 오류를 제거한다.

C. 특수 문자 및 숫자 제거
영문자가 아닌 모든 문자와 숫자를 제거하여 분석에 필요한 단어만 남긴다.

D. 해시태그 제거
텍스트에서 해시태그 기호(#)를 제거하여 불필요한 기호를 정리한다.

E. 단어 길이 제한
너무 짧거나 긴 단어는 분석에 유의미하지 않을 가능성이 높으므로, 3자 미만 및 7자 초과 단어를 제거하여 텍스트의 일관성을 높인다.

F. 빈번 단어 제거
각 텍스트에서 상위 25개의 빈번 단어를 제거하여, 분석에 유의미한 단어들이 남도록 한다. 이는 빈번 단어들이 의미 있는 패턴을 흐리게 하거나 중요도가 떨어지는 단어일 가능성이 높기 때문에 수행하는 단계이다.

G. 줄임말 탐색 및 변환
모든 줄임말이 풀어쓰기로 변환된 텍스트 데이터가 생성된다.

H. 최종적으로 빈도수 1인 단어 제거 및 텍스트 재구성
각 텍스트 항목에서 빈도수 1인 단어를 제거하고 나머지 단어만 이어붙여서 최종 텍스트 데이터를 완성한다.

4) 데이터 TF-IDF 벡터화

TF-IDF는 특정 문서에서 자주 등장하면서 전체 문서에서는 드물게 나타나는 단어에 높은 가중치를 부여하여 중요한 단어를 부각시키는 기법이다. BoW보다 단어의 중요도를 잘 반영한다는 점에서 해당 벡터화 기법을 사용하였다. 문맥 정보를 반영할 수 있는 다른 벡터화 기법도 있지만, 텍스트 데이터를 처음 다루기에 TF-IDF를 선택하였다.

- stop_words='english': 일반적으로 의미가 적거나 분석에 큰 도움이 되지

않는 불용어(stop words)를 자동으로 제거하여, 분석의 효율을 높인다.
- max_df=0.5: 전체 문서의 50% 이상에서 등장하는 단어를 제거하여, 분석에 불필요한 빈번 단어를 배제한다.
- min_df=0.01: 전체 문서의 1% 미만에서 등장하는 단어를 제거하여, 지나치게 드문 단어를 배제한다.

5) 다중출력회귀 기법

A. 다중출력회귀 모델의 개요

다중출력회귀(Multi-output Regression)는 여러 종속 변수를 동시에 예측하는 회귀 기법이다. 단일 회귀 모델이 하나의 종속 변수에 대한 값을 예측하는 것과 달리, 다중출력회귀는 하나의 모델로 다수의 출력 변수(예측값)를 생성할 수 있다는 점에서 차이가 있다. 이 모델은 출력 간 상관관계가 존재하거나 다차원적인 예측이 필요한 경우에 특히 효과적이다.
다중출력회귀의 핵심 아이디어는 각 종속 변수를 별도로 예측하지 않고, 종속 변수들 간의 관계와 특징을 동시에 고려하여 모델이 학습하는 데 있다. 이 방법을 통해, 모델은 각 변수를 단독으로 예측할 때보다 더 높은 예측 정확성을 기대할 수 있으며, 결과 또한 일관성을 갖출 수 있다.
다중출력회귀 모델의 주요 장점은 다음과 같다. 첫째, 여러 종속 변수를 별도의 모델로 학습하는 대신 하나의 모델에서 예측이 가능하므로, 계산 효율이 높다. 둘째, 각 예측값 간의 관계를 자연스럽게 반영할 수 있어, 단일 회귀 모델을 여러 개 사용하는 것보다 결과가 일관적이다. 셋째, 모든 변수를 종합적으로 고려하여 예측하기 때문에 상호 관계에 대한 통찰을 얻을 수 있다.
본 자료 분석에서는 에세이에 대한 다양한 평가 지표들(cohesion, syntax, vocabulary 등)이 연속형 변수로 구성되어 있고, 각 평가지표 간 상관관계가 존재한다고 판단하였다. 특히, 평가 항목 중 종속 변수로 삼을 만한 단일 변수가 없기 때문에, 에세이를 독립 변수로 두고 다중출력회귀를 통해 상관관계가 높은 여러 평가지표를 동시에 예측하는 것이 가장 적합하다고 보았다. 이를 통해, 각 평가 지표가 에세이의 어떤 특성을 반영하는지 명확히 파악하고, 에세이 평가에 대한 일관된 예측을 도출하고자 한다.

B. 다중출력회귀 모델 선정

본 프로젝트에서는 에세이 평가를 위한 다중출력회귀 모델로 **SVR(Support Vector Regressor), 랜덤 포레스트 회귀(Random Forest Regressor), XGBoost**를 선택하였다. 이 모델들은 각기 다른 방식으로 데이터를 분석하여 예측 성능을 높일 수 있는 잠재력을 가지고 있으며, 중간 규모의 텍스트 데이터에 적합하다. 각 모델의 선택 이유와 배제된 다른 모델들에 대한 이유는 다음과 같다.

- SVR (Support Vector Regressor)
선택 이유: SVR은 중간 규모 데이터에서 우수한 성능을 보이며, 텍스트 기반 회귀 분석에 자주 사용되는 모델이다. 현재 데이터는 적당한 크기를 가지며, SVR은 에세이 평가지표 간의 미세한 차이를 잘 반영할 것으로 기대된다.

- 랜덤 포레스트 회귀 (Random Forest Regressor)
선택 이유: 랜덤 포레스트 회귀는 복잡한 상호작용을 잘 반영할 수 있는 모델로, 본 프로젝트의 평가지표들(cohesion, syntax, vocabulary 등) 간의 상호작용을 효과적으로 반영할 수 있다. 또한, 중간 규모 데이터에 적합하고, 이산형 데이터에서도 우수한 성능을 보여줄 것으로 기대된다.

- XGBoost
선택 이유: XGBoost는 일반적으로 대규모 데이터에 적합한 모델이지만, 다양한 하이퍼파라미터 튜닝 옵션을 제공하여 예측 성능을 최적화할 수 있다. 중간 규모 데이터에서도 우수한 성능을 보일 가능성이 있어 이번 프로젝트에서 시도해보고자 한다.

C. 다중출력회귀 기법의 적용

세 가지 모델 모두 MultiOutputRegressor 메소드를 사용하여 다중출력 회귀로 확장하여 분석을 진행하였다.

I. SVR (Support Vector Regressor)

SVR은 데이터를 가장 잘 설명할 수 있는 최적의 선을 찾고, 설정된 오차 범

위(마진) 안에 최대한 많은 데이터를 포함하도록 학습한다. 이 과정에서 마진 밖의 데이터에 대해 패널티를 부여하여 모델이 과적합하지 않도록 조정한다.

SVR(Support Vector Regressor)을 기반으로 한 다중출력 회귀 모델(MultiOutputRegressor)을 사용하여 에세이 평가 점수를 예측한 결과, 모델의 R^2 점수는 약 0.865로 나타났고 MSE는 0.058값을 나타냈다. 이는 모델이 전체 데이터의 약 86.5%를 설명할 수 있다는 의미로, 높은 설명력을 지닌 모델임을 보여준다. R^2 값이 1에 가까울수록 모델의 예측이 실제 값과 잘 일치한다는 것을 의미하기 때문에, 0.865라는 값은 다중출력 예측 모델이 에세이의 여러 평가지표를 동시에 예측하는 데 있어 우수한 성능을 보이고 있음을 나타낸다.

II. 랜덤 포레스트 회귀 (Random Forest Regressor)

랜덤 포레스트는 여러 개의 의사결정 나무를 생성한 후, 각 트리의 예측값을 평균 내어 최종 예측값을 산출한다. 각 트리는 데이터를 무작위로 샘플링하여 학습하며, 이로 인해 과적합을 방지하고 모델의 안정성을 높인다.

랜덤 포레스트 회귀 모델(Random Forest Regressor)을 기반으로 다중출력 회귀 분석을 수행한 결과, 훈련 데이터에 대한 R^2 점수는 0.887로 나타났고, MSE는 0.048값을 나타냈다. 이는 모델이 훈련 데이터의 약 88.7%를 설명할 수 있다는 의미로, 에세이 평가 점수 예측에 있어 높은 설명력을 지니고 있음을 보여준다.

III. XGBoost

XGBoost는 부스팅(Boosting) 방식을 사용하여 작동하며, 여러 개의 약한 모델(의사결정 나무)을 순차적으로 학습시킨다. 이전 모델이 만든 오차를 다음 모델이 학습하면서 점점 더 정확한 예측값을 만들어낸다.

XGBoost 회귀 모델(XGBoost Regressor)을 기반으로 다중출력 회귀 분석을 수행한 결과, 모델의 R^2 점수는 약 0.867로 나타났고 MSE는 0.057값을 나타냈다. 이는 모델이 전체 데이터의 약 86.7%를 설명할 수 있다는 의미로, 에세이 평가 점수 예측에 있어 우수한 설명력을 지니고 있음을 보여준다.

XGBoost 모델은 일반적으로 대규모 데이터에서 높은 성능을 보이는 모델로, 이번 프로젝트에서도 평가지표 간 상관관계를 잘 반영하며 높은 예측력을 발휘하였다. 이를 통해 XGBoost 모델이 에세이 데이터의 평가 점수 예측에 효과적임을 확인할 수 있었다.

D. 다중출력회귀 기법의 적용 분석 결과

I. SVR (Support Vector Regressor)

테스트 데이터에 대한 예측 점수는 [그림14]와 같이 도출되었다.

```
Predicted scores for the test data:
 [[2.80528341 2.7086882  3.1246479  2.95630579 2.60169406 2.55961524]
 [3.27353371 3.00215939 3.0209024  2.77393673 2.87368202 3.17669341]
 [3.50274661 3.37189758 3.58778033 3.28601734 3.25614067 3.27430488]]
```

[그림 16 SVR 모델을 사용한 테스트 데이터에대한 예측 점수]

II. 랜덤 포레스트 회귀 (Random Forest Regressor)

테스트 데이터에 대한 예측 점수는 [그림15]와 같이 도출되었다.

```
Predicted scores for the test data:
 [[3.    2.7   3.175 3.17  2.925 2.755]
 [3.045 2.965 3.05  2.8   2.76  3.165]
 [3.595 3.42  3.5   3.415 3.085 3.31 ]]
```

[그림 17 랜덤 포레스트 회귀 모델을 사용한 테스트 데이터에대한 예측 점수]

III. XGBoost

테스트 데이터에 대한 예측 점수는 [그림16]와 같이 도출되었다.

```
Predicted scores for the test data:
 [[3.1085002 2.666383  3.5787616 2.9735332 2.6200178 2.8696952]
 [3.0813165 3.1203916 3.4049063 3.101281  3.4389532 3.1833398]
 [3.2406096 3.843515  3.4682503 4.069343  3.8067489 3.2121122]]
```

[그림 18 XGBoost 모델을 사용한 테스트 데이터에대한 예측 점수]

이와 같은 다중출력회귀 모델의 적용 결과를 통해, 각각의 모델이 에세이 평가 점수 예측에서 높은 설명력을 보였으며, 테스트 데이터에 대해서도 신뢰할 수 있는 예측값을 제공하였다.

6) Train 데이터의 적합성 검증을 위한 실험

train 데이터의 구조적 안정성과 신뢰성은 이전 분석을 통해 확인되었으나, 데이터가 실제 모델 학습 및 예측에 적합한지를 검증하기 위해 추가적인 실험을 수행하였다. 이를 위해 train 데이터에 변형을 가하고 test 데이터에서 예측 성능을 분석하는 실험 결과를 다루겠다. 이 실험은 다음 두 가지로 구성되며, 각각의 결과를 통해 train 데이터의 적합성을 확인하였다.

- 상위권 점수 데이터를 제거하고 모델을 구축한 뒤, 제거된 데이터를 test 데이터로 사용하여 예측 성능을 평가.
- 상위권 점수를 test 데이터로 사용하며, 해당 데이터에 오류를 추가하여 모델의 민감도와 데이터 적합성을 평가.

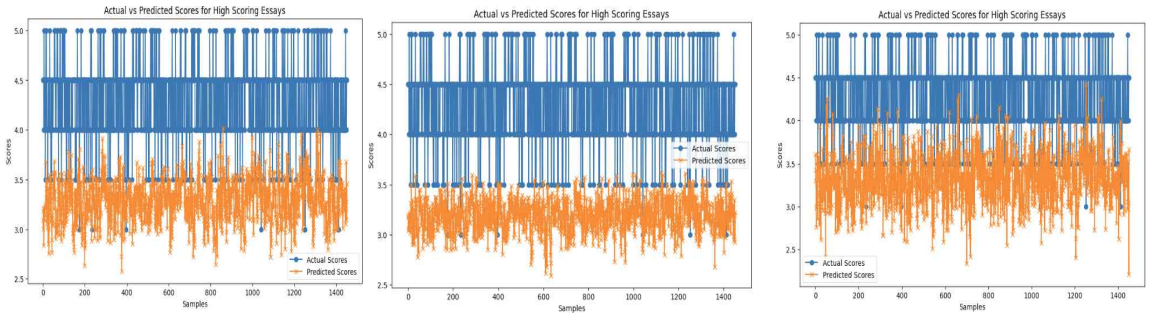
A. 상위권 점수 데이터 제거 실험

I. 실험 설계

Train 데이터에서 평균 점수가 4점 이상에 해당하는 고득점 데이터를 제거하고, 나머지 데이터로 모델을 학습하였다. 제거된 상위권 데이터는 test 데이터로 설정하여, 모델이 상위권 점수를 얼마나 잘 예측하는지 확인하였다. 본 실험은 train 데이터가 올바르게 구축되었다면, 제거된 고득점 데이터를 모델이 제대로 예측하지 못할 것이라는 가정에 기반한다.

II. 결과 및 분석

제거된 상위권 데이터를 test 데이터로 사용하여 예측한 결과, 모델은 상위권 점수를 중간 점수대로 예측하는 경향을 보였다. 아래 그림과 표는 각 모델의 예측 결과를 정리한 것이다.



[그림19 각 모델(좌측부터 순서대로, SVR, 랜덤포레스트,XGBoost)의 예측 결과 비교]

모델	Training R^2	Training MSE	Test High Scores R^2	Test High Scores MSE
SVR(Support Vector Regressor)	0.877	0.043	-6.707	1.164
Random Forest Regressor	0.884	0.040	-7.968	1.352
XGBoost	0.877	0.043	-6.476	1.128

[표1 각 모델의 예측 결과 비교]

III. 결과 해석

Train 데이터에서 상위권 데이터를 제거한 상태로 학습된 모델은 상위권 점수를 정확히 예측하지 못하고, 중간 점수대로 예측하였다. 또한 Test 데이터에 대한 R^2 값이 극단적으로 낮고, MSE 값이 매우 큰 결과를 나타냈다.
이는 train 데이터가 고득점 데이터를 포함한 상태에서만 모델이 올바르게 구축되었다는 것을 의미한다. 따라서, train 데이터가 고득점 데이터를 포함한 상태에서는 적절하게 구축되었다고 판단할 수 있다.

추가적으로 같은 방법으로 하위권 점수를 제거하여 모델을 구축한다음 제거된 하위권 점수를 test 데이터에 넣어 예측해본결과 이 또한 상위권 점수와 마찬가지로 대체적으로 중간점수대로 예측함을 확인해볼 수 있다.

B. 상위권 데이터에 오류를 추가한 민감도 실험

I. 실험 설계

상위권 점수(모든 평가지표에서 만점 데이터)를 test 데이터로 설정하고, 해당 데이터에 다음과 같은 오류를 단계적으로 추가하였다.

- 철자 오류: 단어의 철자를 고의적으로 틀리게 수정.
- 문법 오류: 시제 불일치, 주어-동사 불일치 등의 문법적 오류 추가.

모델이 점수에 변화를 반영하는지 확인하여, train 데이터가 모델 학습에 적합

했는지 평가하였다.

II. 결과 및 분석

(SVR)

항목	cohesion	syntax	vocabulary	phraseology	grammar	conventions
Essay 1	4.229	4.276	4.211	4.103	4.088	4.249
Essay 2	4.027	3.912	4.199	3.833	3.906	3.932

(Random Forest Regressor)

항목	cohesion	syntax	vocabulary	phraseology	grammar	conventions
Essay 1	4.37	4.48	4.265	4.32	4.27	4.245
Essay 2	4.4	4.215	4.4	4.275	4.07	4.395

(XGBoost)

항목	cohesion	syntax	vocabulary	phraseology	grammar	conventions
Essay 1	4.837	5.005	4.776	4.646	4.717	4.673
Essay 2	4.764	4.367	4.462	4.747	4.266	4.842

[표2 각 모델의 만점 에세이에 대한 예측 점수 비교]

III. 결과 해석

본 실험을 통해, train 데이터는 모델 학습에 적합하게 구성되었으며, 평가 항목별 데이터 품질과 학습 구조가 적절히 반영되었음을 확인할 수 있었다. 평가 기준 간의 강한 상관관계로 인해 문법적, 철자 오류가 다른 항목에도 영향을 미친다는 것을 확인하였다. 이는 학습된 모델이 단순한 독립적 요소 평가가 아니라, 전체적인 언어적 품질을 기반으로 평가한다는 점에서 높은 신뢰성을 보여준다.

본 2개의 실험을 통해, **train 데이터는 모델 학습에 적합하게 구축되었으며, 평가 항목별로 데이터 품질과 학습 구조가 적절히 반영되었음을 확인할 수 있다.** 이러한 결과는 train 데이터가 특정 점수 구간(고득점 데이터)을 포함하여

학습될 때 가장 효과적이며, 모델의 신뢰성과 일반화 성능을 보장할 수 있음을 시사한다.

7) 딥러닝을 활용한 점수 예측 모델

딥러닝 기반 **RoBERTa 모델**을 활용하여 ELLs의 에세이에 대한 점수를 예측하자.
RoBERTa 모델이란 Google에서 개발한 BERT 모델을 기반으로 Facebook AI Research에서 더욱 개선한 자연어 처리(NLP) 모델이다. RoBERTa는 Transformer 아키텍처를 기반으로 설계된 모델로, 텍스트의 문맥을 이해하기 위해 셀프 어텐션(Self-Attention) 메커니즘을 사용하며 문장의 앞뒤 문맥을 동시에 학습하여 단어가 문장에서 가지는 정확한 의미를 파악할 수 있다. RoBERTa는 BERT의 학습 방식과 데이터 활용 방식을 개선한 모델로, 더 많은 데이터를 학습하고 최적화된 학습 방식을 통해 더 강력한 언어 이해 능력을 제공한다. 특히, 문맥을 이해하는 능력이 뛰어나 텍스트 기반 에세이 점수 예측과 같은 작업에 적합하다.

A. 데이터 전처리

영문자가 아닌 모든 문자와 숫자를 제거하여 텍스트를 정리하였다.

B. 벡터화

RoBERTa의 사전 학습된 토크나이저(RobertaTokenizer)를 사용하여 텍스트 데이터를 토큰화하였다. 텍스트는 최대 길이(max_len=128)로 패딩되며, 초과된 텍스트는 잘렸다.

C. 딥러닝 모델 적용

사전 학습된 RoBERTa 모델(roberta-base)을 기반으로 하며, RoBERTa의 마지막 출력층 위에 선형 회귀 레이어를 추가하였다. RoBERTa의 셀프 어텐션 메커니즘을 통해 문맥 정보를 학습하고, 선형 회귀 레이어에서 최종 점수를 예측하였다. 학습은 3개의 에포크 동안 수행되었으며, 배치 크기는 16, 학습률은 2e-5로 설정되었다. 옵티마이저는 Adam을 사용하였으며, 손실 함수는 평균

제곱 오차를 사용하였다. 마지막으로 학습 과정에서 GPU를 활용하여 계산 속도를 최적화하였다. 최종적으로 테스트 데이터의 예측은 아래 그림과 같다.

	text_id	cohesion	syntax	vocabulary	phraseology	grammar	conventions
0	0000C359D63E	3.073160	2.831918	3.256041	3.310897	2.876477	2.627572
1	000BAD50D026	3.040500	2.457518	2.810586	2.632758	2.222659	2.604402
2	00367BB2546B	3.940084	3.571915	3.743716	4.050148	3.584963	3.095850

[그림 20 RoBERTa 모델을 통한 테스트 데이터 예측 결과]

3. 결론

1) 프로젝트 요약

본 프로젝트는 ELL 학생들의 글쓰기 능력을 평가하기 위해 머신러닝과 딥러닝 기법을 활용한 자동 채점 모델을 구축하고자 진행되었다. 데이터 분석을 통해 train 데이터의 신뢰성과 구조적 안정성을 검증하였으며, 다중출력회귀 모델(SVR, 랜덤 포레스트, XGBoost)을 적용해 각 평가지표(cohesion, syntax, vocabulary 등)를 예측하였다. 랜덤 포레스트 모델이 가장 높은 설명력을 보였으며, train 데이터를 변형하거나 오류를 추가하여 데이터 적합성을 실험적으로 검증한 결과, 데이터가 모델 학습에 적합하게 구축되었음을 확인하였다. 또한, RoBERTa 기반 딥러닝 모델을 적용하여 문맥 이해를 바탕으로 한 점수 예측에서 높은 성능을 보였다. 본 프로젝트는 ELL 학생들에게 효과적인 피드백을 제공할 수 있는 자동 채점 모델의 가능성을 입증하였으며, 교사의 채점 부담을 경감하고 학생들의 글쓰기 학습을 지원할 수 있는 기반을 마련하였다.

2) 적합한 모델 선정

본 프로젝트에서는 다중출력 회귀 모델(SVR, 랜덤 포레스트, XGBoost)과 딥러닝 모델(RoBERTa)을 활용하여 ELL 학생들의 에세이 평가 점수를 예측하였다.
분석 결과, 딥러닝 모델 **RoBERTa**가 가장 적합한 모델로 판단되었다. RoBERTa는 트랜스포머(Transformer) 기반의 사전 학습된 자연어 처리 모델

로, 방대한 학습 데이터를 바탕으로 텍스트의 문맥적 관계를 가장 잘 이해할 수 있다. 이러한 특성은 cohesion, phraseology와 같은 정성적 평가 기준에서 높은 신뢰성을 제공하며, **다양한 평가 항목 간의 상호작용을 정확히 반영할 수 있다.**

추가적으로 머신러닝 모델 중에서는 [그림20]와 같이 **랜덤 포레스트**(Random Forest)가 가장 안정적인 성능을 보였다. 랜덤 포레스트는 Train 데이터에서 가장 높은 설명력(R^2: 0.887)과 낮은 오차(MSE: 0.048)를 기록하며, 학습 데이터에 대한 높은 적합성을 보여주었다.

따라서, 문맥 이해와 정성적 평가가 중요한 상황에서는 RoBERTa가 가장 적합하며, 효율성과 계산 비용을 중시하는 환경에서는 랜덤 포레스트를 활용하는 것이 적절하다고 판단된다. 이 두 모델은 각각의 강점을 살려, 프로젝트 목표에 따라 유연하게 활용될 수 있다.

[그림 21 각 모델을 통한 테스트 데이터 예측 결과]

따라서 본 분석자는 딥러닝 모델인 **RoBERTa가 테스트 데이터의 문맥적 이해**

와 평가 항목 간의 관계를 가장 잘 반영한다고 판단하였기 때문에 테스트 데이터에 대한 최종 점수 예측은 RoBERTa 모델을 사용하여 아래 표와 같이 구성된다.

ID	cohesion	syntax	vocabulary	phraseology	grammar	conventions
0000C359D63E	3.07316	2.831918	3.256041	3.310897	2.876477	2.627572
000BAD50D026	3.0405	2.457518	2.810586	2.632758	2.222659	2.604402
00367BB2546B	3.940084	3.571915	3.743716	4.050148	3.584963	3.09585

[표3 RoBERTa 모델을 통해 예측한 test 데이터 점수]

만약 분석자가 모델의 효율성을 더 중요시한다면 랜덤포레스트 머신러닝을 사용해 아래 표와 같이 점수가 구성될 수 있다.

ID	cohesion	syntax	vocabulary	phraseology	grammar	conventions
0000C359D63E	3	2.7	3.175	3.17	2.925	2.755
000BAD50D026	3.045	2.965	3.05	2.8	2.75	3.165
00367BB2546B	3.595	3.42	3.5	3.412	3.085	3.3

[표4 랜덤포레스트 모델을 통해 예측한 test 데이터 점수]

3) 기대효과 및 한계

본 프로젝트를 통해 구축된 자동 채점 모델은 ELL 학생들에게 즉각적이고 맞춤형 피드백을 제공함으로써 글쓰기 능력 향상에 실질적인 도움을 줄 수 있을 것으로 기대된다. 또한, 교사의 채점 부담을 줄여 더 많은 학생에게 개별적인 지도를 가능하게 하고, 공정하고 일관된 평가 기준을 제공할 수 있다.
다만, 본 모델은 비교적 제한된 데이터셋과 특정 학습 환경에 기반하고 있어 다양한 언어적 배경과 글쓰기 스타일을 포괄하는 데 한계가 있을 수 있다. 추가적으로, 평가 기준 간의 세부적인 가중치나 채점자의 주관적 판단을 반영하지 못한다는 점은 실제 사용 시 개선이 필요한 부분으로 판단된다.

4) 수업에서 느낌점

본 프로젝트를 통해 처음으로 학습 데이터 구축과 평가라는 작업을 직접 수행해 보았다. 처음에는 이러한 작업이 낯설었고, 왜 이 과정이 필요한지에 대한 명확한 이해가 부족해 혼란스럽게 느껴지기도 했다. 특히, 학습 데이터가 학습에 적합하게 구성되었는지 판단하고 이를 검증하기 위한 다양한 실험을 설계하는 과정은 익숙하지 않은 도전이었다.
그러나 교수님과의 주 1회 면담을 통해 이러한 작업의 필요성과 중요성을 점차 이해할 수 있었다. 교수님은 데이터 품질과 모델 성능 평가의 중요성을 강조하며, 구체적인 피드백을 통해 어떻게 학습 데이터의 적합성을 검증할 수 있는지 여러 방향을 제시해 주셨다. 이를 통해, 단순히 주어진 데이터를 가지고 모델을 돌리는 단계를 넘어, 데이터의 구조적 완전성과 학습 모델의 신뢰성을 검증하는 작업이 왜 중요한지를 깨달았다.
이번 경험은 데이터 분석 및 모델 학습 과정에서 체계적인 평가와 검증의 중요성을 배우는 계기가 되었으며, 앞으로의 프로젝트에서 이를 실무에 적용할 자신감을 갖게 되었다.

(추가적으로 교수님께 드리는 피드백)

주1회 수업시간을 활용해 수업이 아닌 1:1면담으로 더 많은 개인별 피드백을 받을 수 있어서 정말 의미있는 수업이었습니다. 한 학기 동안 열심히 제 프로젝트에 대해 피드백해주시고 고민해주셔서 감사했습니다.

4. 참고문헌

1. 그림1 Glendale Community College의 에세이 평가기준 : https://www.glendale.edu/academics/academic-divisions/esl-credit/esl-123-final-essay-scoring-rubric
2. 그림2 SAT Essay 평가기준 : https://blog.prepscholar.com/sat-essay-rubric-full-analysis-writing-strategies
3. 그림3 전체 응시자 SAT 백분위 점수대 : https://blog.naver.com/PostView.naver?blogId=upgradecampus&logNo=223483337201
4. 그림4 응시자의 현 교육 수준으로 분류한 TOEFL iBT 시험 평균 점수 : https://www.prnewswire.com/kr/news-releases/u0065u0074u0073u002Cu002Du0032u0030u0032u0033u002DuC138uACC4u002DuD1A0uD50Cu002DuC2DCuD5D8u002DuC131uC801u002DuBCF4uACE0uC11Cu002DuBC1CuAC04-302217657.html