

# 고급 통계 및 머신러닝 기법을 활용한 학생 성적 예측 분석

우동협, 정명훈, 임재성, 김동환, 강성우

October 12, 2024

## Contents

<b>1 서론</b>	<b>3</b>
1.1 배경 및 목적 . . . . .	3
1.2 데이터셋 개요 . . . . .	3
1.3 분석 접근 방법 . . . . .	3
<b>2 통합 성적(G) 생성 및 분석</b>	<b>3</b>
2.1 크론바흐 알파를 통한 내적 일관성 검증 . . . . .	5
2.2 주성분 분석을 통한 통합 성적(G) 도출 . . . . .	6
2.3 G의 분포 및 특성 분석 . . . . .	6
<b>3 데이터 전처리 및 고급 탐색적 데이터 분석 (EDA)</b>	<b>7</b>
3.1 데이터 클리닝 및 전처리 . . . . .	7
3.1.1 범주형 및 수치형 변수 식별 . . . . .	7
3.1.2 범주형 변수 인코딩 . . . . .	8
3.1.3 수치형 변수 스케일링 . . . . .	8
3.2 이상치 처리: IQR 방법 및 윈저화(Winsorization) 적용 . . . . .	9
<b>4 탐색적 데이터 분석 (EDA)</b>	<b>10</b>
4.1 범주형 변수 빈도 분석 . . . . .	10
4.2 기본 통계 및 기술 통계 분석 . . . . .	12
4.2.1 수치형 변수 기술 통계 . . . . .	12
4.2.2 범주형 변수 기술 통계 . . . . .	13
4.3 데이터 요약 및 관계 탐색 . . . . .	13
4.3.1 성별과 학업 성취도 . . . . .	13
4.3.2 부모의 교육 수준과 성적 간의 관계 . . . . .	14
4.3.3 학습 시간과 성적 간의 관계 . . . . .	14
4.3.4 결석 횟수와 성적 간의 관계 . . . . .	14
4.3.5 인터넷 접근성과 성적 간의 관계 . . . . .	15
4.3.6 수치형 및 범주형 변수 상관관계 분석 . . . . .	15
4.3.7 다변량 분석: 앤드류 커브 (Andrews Curves) . . . . .	16
<b>5 예비 통계 분석 및 가설 검정</b>	<b>16</b>
5.1 학습 시간과 성적 간의 상관관계 분석 . . . . .	17
5.2 어머니의 교육 수준에 따른 성적 차이 분석 . . . . .	17
5.3 결석 횟수와 성적 간의 회귀 분석 . . . . .	17
5.4 성별에 따른 성적 차이 검정 . . . . .	17

5.5	다중 비교에 대한 FDR 제어 . . . . .	18
5.6	예비 분석의 요약 및 모델 구축을 위한 시사점 . . . . .	18
<b>6</b>	<b>머신러닝 모델 개발 및 평가</b>	<b>19</b>
6.1	Random Forest 모델 . . . . .	19
6.2	Gradient Boosting 모델 . . . . .	19
6.3	XGBoost 모델 . . . . .	19
6.4	LightGBM 모델 . . . . .	19
6.5	CatBoost 모델 . . . . .	20
6.6	Linear Regression 모델 . . . . .	20
6.7	Ridge 모델 . . . . .	20
6.8	Lasso 모델 . . . . .	20
6.9	모델 비교 및 최종 평가 . . . . .	20
6.10	하이퍼파라미터 튜닝 및 교차검증 . . . . .	21
6.11	특성 중요도 분석 . . . . .	21
6.11.1	주요 인사이트 도출 . . . . .	22
6.12	모델 비교 및 최종 모델 선정 . . . . .	23
<b>7</b>	<b>모델 해석 및 설명 가능성</b>	<b>23</b>
7.1	SHAP (SHapley Additive exPlanations) 해석 . . . . .	24
7.1.1	이론적 배경 . . . . .	24
7.1.2	SHAP 특성 중요도 (Bar Plot) . . . . .	24
7.1.3	SHAP 특성 영향 (Dot Plot) . . . . .	25
7.2	SHAP Dependence Plot . . . . .	25
7.3	부분 의존성 플롯 (Partial Dependence Plot, PDP) . . . . .	26
7.4	LIME (Local Interpretable Model-agnostic Explanations) . . . . .	27
7.5	모델 해석의 중요성 . . . . .	28
<b>8</b>	<b>응용: SHAP와 최적 수송 이론을 결합한 개인화된 학업 성취도 향상 전략</b>	<b>28</b>
8.1	SHAP와 최적 수송 이론의 결합 . . . . .	29
8.2	최적 수송 문제 설정 . . . . .	29
8.3	비용 함수 정의 및 해석 . . . . .	29
8.4	이론적 증명 . . . . .	29
8.5	개인화된 학습 피드백의 구성 . . . . .	30
8.6	알고리즘 . . . . .	30
8.7	실험 결과: 실제 건국대학교 학생을 대상으로 피드백 생성 . . . . .	31
8.7.1	학생 1의 개인화된 전략 . . . . .	31
8.7.2	학생 50의 개인화된 전략 . . . . .	31
8.8	한계점 및 향후 연구 방향 . . . . .	32
<b>9</b>	<b>결론</b>	<b>32</b>

# 1 서론

본 보고서는 포르투갈 중등학교 학생들의 성적 예측을 목적으로 다양한 머신러닝 기법을 활용하여 성적에 영향을 미치는 요인들을 분석하고 예측 모델을 개발하는 보고서이다. 사용된 데이터셋은 포르투갈 중등학교 두 곳에서 수집된 성적 및 학생 정보를 포함하고 있으며, 학생들의 인구통계학적 정보, 가정 환경, 학습 관련 변수 등을 바탕으로 성적을 예측하는 것이 목적이다. 더 나아가, 이 보고서는 SHAP와 최적 수송 이론을 결합하여 건국대학교 학생들의 학업 성취도 향상을 위한 개인화된 전략을 제시하는 새로운 방법론 것을 목표로 한다.

## 1.1 배경 및 목적

학생들의 성적에 영향을 미치는 요인들은 다양하며, 이를 분석하여 성적을 예측하는 것은 교육 정책 수립과 개인 맞춤형 학습 전략을 세우는데 중요한 역할을 한다. 특히 최근 머신러닝 기법의 발전으로 인해 더 정교하고 정확한 예측 모델을 개발할 수 있게 되었다.

본 보고서의 목적은 포르투갈 중등학교 학생들의 데이터를 바탕으로 성적을 예측하고, 예측된 모델을 건국대학교 학생들에게 적용하여 성적 향상에 실질적으로 기여할 수 있는 방안을 도출하는 데 있다. 이를 통해 학생들의 성적을 높일 수 있는 구체적인 피드백과 맞춤형 학습 전략을 제시하고자 한다.

## 1.2 데이터셋 개요

본 분석에 사용된 데이터셋은 포르투갈의 두 개 중등학교 학생들의 성적 및 관련 정보를 포함하고 있다. 이 데이터셋은 아래와 같은 주요 특징을 가지고 있다:

- 총 33개의 변수로 구성되어 있음
- 학생의 인구통계학적 정보, 가정 환경, 학습 관련 변수를 포함
- G1, G2, G3로 표현되는 세 번의 성적 평가 결과 포함

## 1.3 분석 접근 방법

본 보고서에서는 다음과 같은 단계로 데이터 분석을 수행한다:

- 탐색적 데이터 분석 (EDA)을 통한 데이터 특성 파악
- 고급 통계 기법을 활용한 변수 간 관계 분석
- 다양한 머신러닝 모델 개발 및 성능 비교
- 최적 모델 선정 및 해석
- 실제 적용 가능성 평가

이러한 분석을 통해, 학생들의 성적에 영향을 미치는 주요 요인들을 식별하고, 효과적인 성적 예측 모델을 제시하고자 한다.

# 2 통합 성적(G) 생성 및 분석

본 보고서에서는 G1, G2, G3 세 번의 학기별 성적을 하나의 통합 성적(G)으로 결합하여 분석을 진행하였다. 이 과정에서는 성적 간의 일관성을 평가하기 위해 크론바흐 알파 (Cronbach's Alpha)를 활용하였으며, 주성분 분석(PCA)을 통해 주요 성분을 추출하여 성적 데이터를 축소하였다. 최종적으로 통합 성적(G)의 분포와 특성을 분석하였다.

변수 이름	설명
school	학생이 다니는 학교
sex	학생의 성별
age	학생의 나이
address	거주지 유형
famsize	가족 규모
Pstatus	부모의 동거 상태
Medu	어머니의 교육 수준
Fedu	아버지의 교육 수준
Mjob	어머니의 직업
Fjob	아버지의 직업
reason	이 학교를 선택한 이유
guardian	보호자
traveltime	통학 시간
studytime	주간 학습 시간
failures	과거 학급 실패 횟수
schoolsup	추가 교육 지원 여부
famsup	가족의 교육 지원 여부
paid	과목 관련 추가 유료 수업 여부
activities	방과 후 활동 여부
nursery	유치원 출석 여부
higher	고등 교육 희망 여부
internet	인터넷 사용 가능 여부
romantic	연애 관계 여부
famrel	가족 관계의 질
freetime	방과 후 여가 시간
goout	친구와의 외출 빈도
Dalc	평일 음주량
Walc	주말 음주량
health	건강 상태
absences	결석 횟수
G1	1차 시험 성적
G2	2차 시험 성적
G3	3차 시험 성적

Table 1: 데이터셋의 변수 및 설명

## 2.1 크론바흐 알파를 통한 내적 일관성 검증

먼저, G1, G2, G3 성적 간의 일관성을 평가하기 위해 크론바흐 알파(Cronbach's Alpha)를 계산하였다. 크론바흐 알파는 다중 항목으로 구성된 척도의 신뢰도를 측정하는 지표로, 값이 0.7 이상이면 일관성이 높다고 평가한다.

크론바흐 알파를 성적 통합 과정에 사용한 이유는 다음과 같다:

- **내적 일관성 평가:** G1, G2, G3 성적이 동일한 구성개념(학생의 수학 능력)을 일관되게 측정하고 있는지 확인할 수 있다. 높은 크론바흐 알파 값은 세 성적이 유사한 특성을 측정하고 있음을 나타낸다.
- **신뢰도 검증:** 여러 학기에 걸친 평가가 신뢰할 만한 결과를 제공하는지 검증할 수 있다. 이는 통합 성적(G)의 신뢰성을 뒷받침하는 근거가 된다.
- **통합의 타당성 확보:** 높은 크론바흐 알파 값은 세 성적을 하나의 지표로 통합하는 것이 통계적으로 타당함을 보여준다.
- **측정 오차 최소화:** 여러 측정치를 통합함으로써 개별 측정의 오차를 상쇄하고, 더 안정적인 성적 지표를 얻을 수 있다.

본 보고서에서 사용한 크론바흐 알파 계산 함수는 다음과 같은 수식을 기반으로 한다.:

$$\alpha = \frac{N \cdot \bar{r}}{1 + (N - 1) \cdot \bar{r}} \quad (1)$$

여기서:

- $N$ 은 항목의 수 (본 보고서에서는 3, G1, G2, G3)
- $\bar{r}$ 은 항목 간 평균 상관계수

해당 수식은 항목 간 평균 상관계수를 사용하여 크론바흐 알파를 계산한다.

계산 결과, 크론바흐 알파 값은 0.95로 매우 높게 나타났으며, 이는 G1, G2, G3 성적이 동일한 학업 성취도를 일관되게 반영하고 있음을 시사한다.

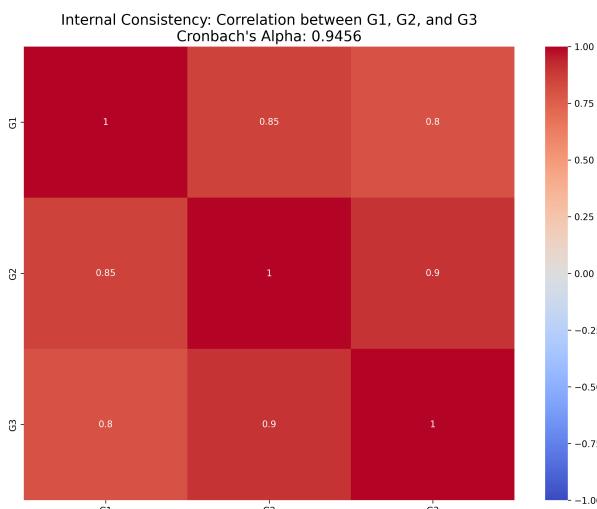


Figure 1: 크론바흐 알파 분석 결과: G1, G2, G3의 내적 일관성

그림 1에서 알 수 있듯이, 세 성적 간의 내적 일관성이 매우 높아 통합 성적(G)으로의 결합이 타당함을 확인할 수 있다.

## 2.2 주성분 분석을 통한 통합 성적(G) 도출

세 번의 성적 평가( $G_1, G_2, G_3$ )를 하나의 통합 성적( $G$ )으로 축소하기 위해 주성분 분석(PCA)을 실시하였다. PCA는 데이터를 보다 단순한 구조로 변환하여 주요한 변동 요인을 추출하는 기법이다. PCA 분석 결과, 첫 번째 주성분이 전체 변동성의 90%를 설명하는 것으로 나타났다. 이를 통해  $G_1, G_2, G_3$  세 성적을 단일 성적 지표인  $G$ 로 축소하는 것이 타당하다는 결론을 도출하였다.

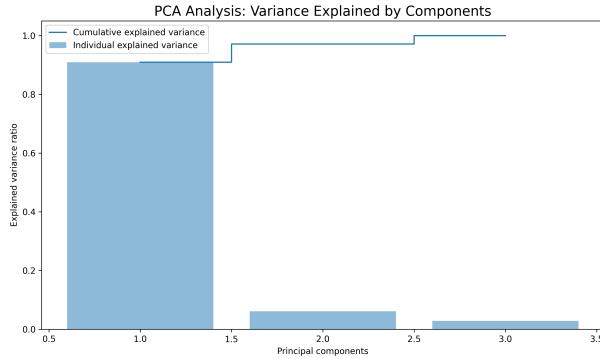


Figure 2: 주성분 분석(PCA) 결과: 첫 번째 주성분이 90%의 분산을 설명

그림 2는 PCA 분석 결과를 시각화한 것으로, 첫 번째 주성분이 대부분의 변동성을 설명하는 것을 보여준다.

## 2.3 G의 분포 및 특성 분석

통합 성적( $G$ )을 도출한 후, 해당 성적의 분포 및 특성을 분석하였다. 통합 성적( $G$ )은 0에서 30점 사이의 값으로 변환되었으며, 이를 시각화한 히스토그램을 통해 학생들의 성적 분포를 확인할 수 있다.

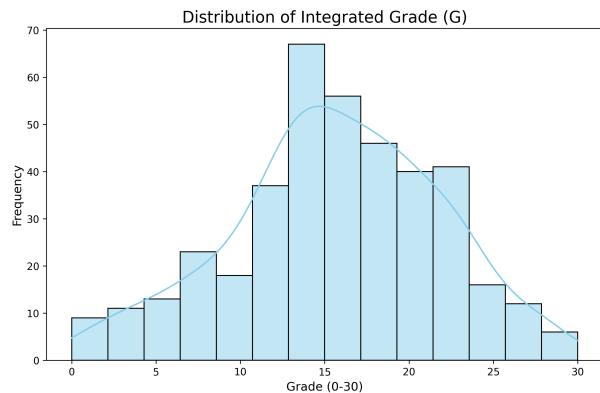


Figure 3: 통합 성적( $G$ )의 분포: 0 30점 사이의 성적 분포

그림 3은 통합 성적( $G$ )의 분포를 나타낸 것으로, 대부분의 학생들이 15점에서 25점 사이에 분포하고 있음을 확인할 수 있다.

또한, 통합 성적( $G$ )과 원래 성적( $G_1, G_2, G_3$ ) 간의 상관관계도 분석하였다. 결과적으로,  $G$ 는  $G_1, G_2, G_3$  각각과 높은 상관관계를 보였으며, 이는 통합 성적이 원래 성적을 충분히 잘 설명하는 지표임을 보여준다.

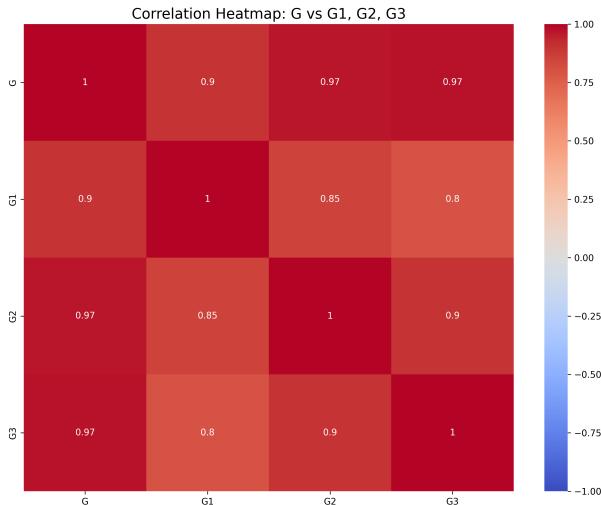


Figure 4: 통합 성적(G)과 G1, G2, G3 간의 상관관계 히트맵

그림 4은 G와 G1, G2, G3 간의 상관관계를 나타내며, 높은 상관계수(0.9 이상)를 확인할 수 있다. 이를 통해 통합 성적(G)이 원래 성적을 대표할 수 있는 유효한 지표임을 확인하였다.

이와 같은 분석 결과를 바탕으로, 통합 성적(G)을 활용한 예측 모델 개발 및 분석을 이어나갈 예정이다.

### 3 데이터 전처리 및 고급 탐색적 데이터 분석 (EDA)

#### 3.1 데이터 클리닝 및 전처리

데이터 분석에 앞서, 데이터 클리닝 및 전처리 작업을 수행하였다. 결측치는 데이터셋에 존재하지 않으므로 따로 처리하지 않았다. 그러나, 이상치 탐지를 위해 IQR(Interquartile Range) 방법을 사용하여 일부 변수에서 이상치를 탐지하였다. 이때, 이상치가 많은 absences와 같은 변수를 제거하지 않고, 원저화(Winsorization) 기법을 사용하여 값을 제한하는 방법을 채택하였다. 그 후, 데이터의 범위를 표준화하기 위해 Robust Scaling을 적용하였다.

##### 3.1.1 범주형 및 수치형 변수 식별

먼저, 범주형 변수와 수치형 변수를 구분하였다. 이는 각 변수에 적합한 인코딩 및 스케일링을 적용하기 위함이다. 분석을 통해 아래와 같은 범주형 변수와 수치형 변수를 식별하였다.

school	sex
address	famsize
Pstatus	Mjob
Fjob	reason
범주형 변수:	guardian
	schoolsup
	famsup
	activities
	higher
	romantic

age	Medu
Fedu	traveltime
studytime	failures
수치형 변수:	famrel
	freetime
	goout
	Dalc
Walc	health
absences	G

### 3.1.2 범주형 변수 인코딩

범주형 변수 중 순서형 데이터는 없었으므로, 모든 범주형 변수에 대해 원-핫 인코딩(One-Hot Encoding)을 수행하였다. 이 인코딩 기법은 각 범주를 이진 변수를 사용하여 표현하며, 순서성이 필요하지 않은 데이터에 적합하다.

### 3.1.3 수치형 변수 스케일링

수치형 변수의 값 범위가 서로 다르기 때문에 이를 표준화하였다. 특히, age, absences, G는 다른 수치형 변수에 비해 값의 범위가 크게 달랐다. 따라서 변수 간의 편향을 줄이기 위해 Z-스코어 표준화를 적용하였다.

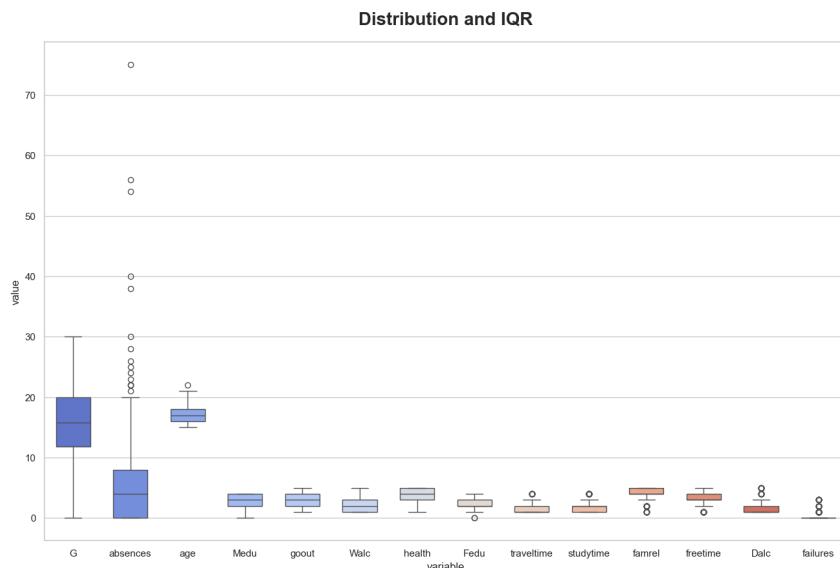


Figure 5: 표준화 전 수치형 변수들의 분포 (Boxplot)

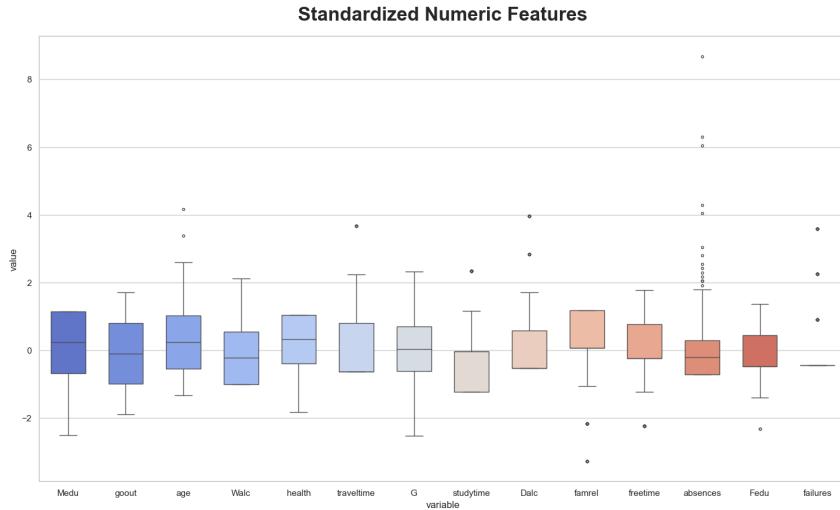


Figure 6: 표준화 후 수치형 변수들의 분포 (Boxplot)

### 3.2 이상치 처리: IQR 방법 및 원저화(Winsorization) 적용

이상치는 IQR(Interquartile Range) 방법을 사용하여 탐지하였으나, 이상치를 제거하는 방식은 전체 데이터의 약 40%가 손실되어 적절하지 않다고 판단하였다.

Original data shape: (395)

Data shape after outlier removal: (226)

따라서, 이상치를 제거하는 대신 원저화(Winsorization) 기법을 적용하여 이상치 값을 상한 및 하한으로 제한하고, 이상치가 학습에 미치는 영향을 줄이도록 하였다. 그 후, Robust Scaling을 적용하여 이상치의 영향을 최소화하였다.

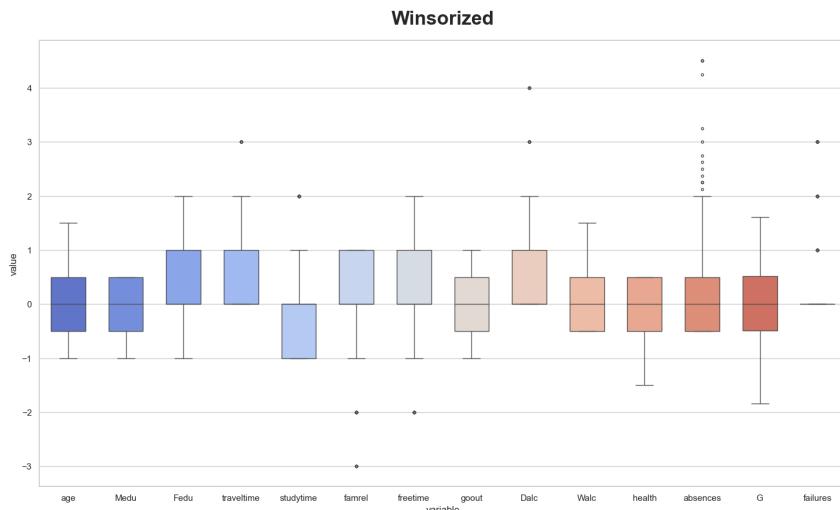


Figure 7: Winsorization + Robust Scaling 적용 후

특히 **absences** 변수는 이상치가 많았으며, 이로 인해 모델 성능에 부정적인 영향을 미칠 수 있다고 판단하였다. 따라서 먼저 이상치를 포함한 상태에서 모델을 학습시키고, 이후 **feature importance** 분석을 통해 이 변수의 중요도를 확인한 후, 필요 시 제거할 계획이다.

## 4 탐색적 데이터 분석 (EDA)

탐색적 데이터 분석(EDA)은 데이터의 특성과 구조를 파악하고, 변수 간의 관계를 이해하여 향후 모델링 과정에서 중요한 요소를 발견하는 과정이다. 본 분석에서는 주요 변수들의 분포, 변수 간 상관관계, 그리고 성적 변화 패턴을 파악하고자 한다. 이를 통해 데이터에 숨어있는 인사이트를 도출하고, 분석 및 모델링 전략을 수립할 수 있다.

### 4.1 범주형 변수 빈도 분석

범주형 변수의 분포는 데이터의 불균형 여부를 확인하고, 학습 과정에서 해당 변수가 모델 성능에 미치는 영향을 평가하는 중요한 요소 중 하나이다. 아래는 주요 범주형 변수들의 빈도 분포를 시각화한 결과이다.

**Distribution of Categorical Features**

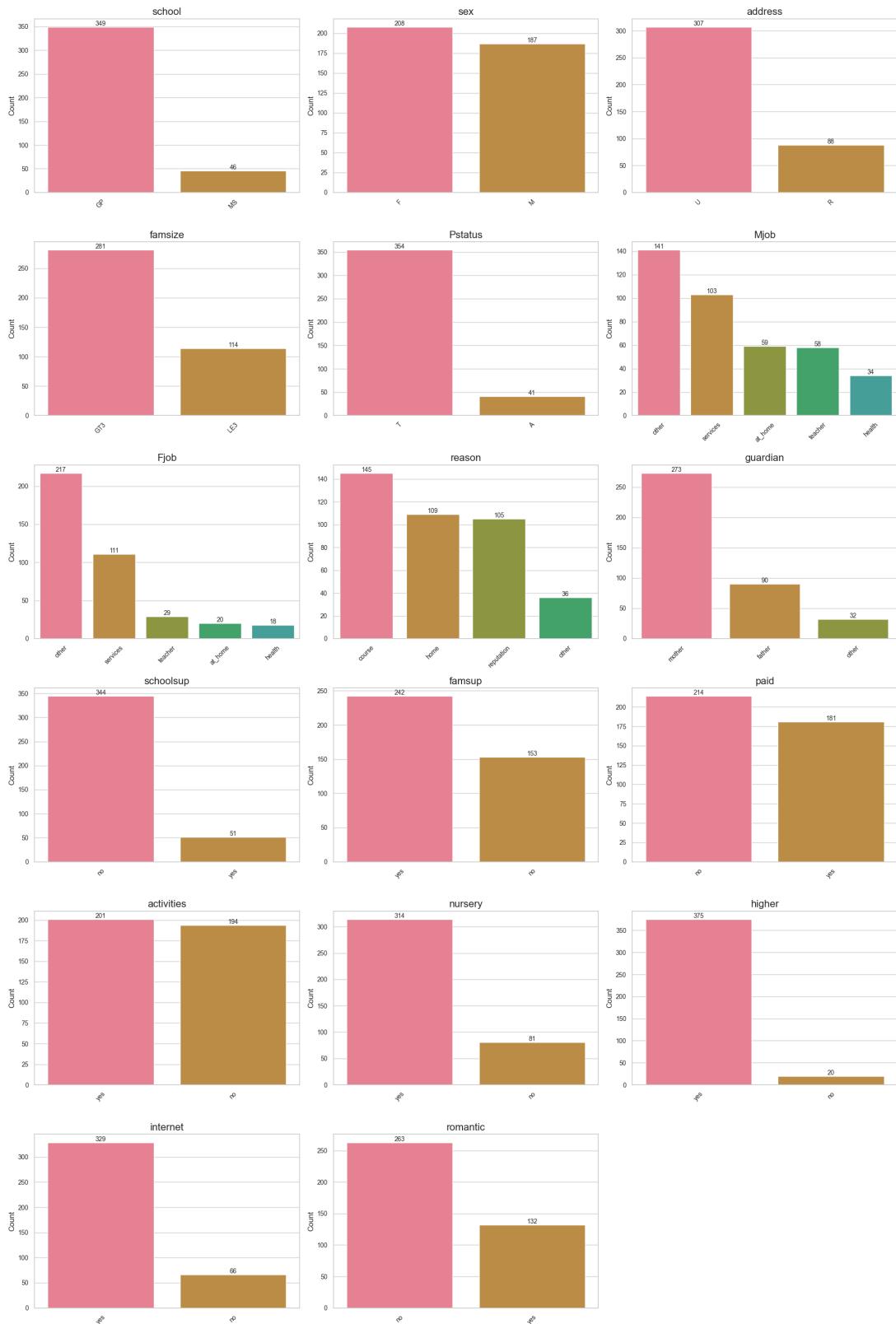


Figure 8: 범주형 변수들의 빈도 분포

**1. school** 변수 두 학교(Gabriel Pereira, Mousinho da Silveira)에 대한 데이터가 불균형하게 분포되어 있으며, Gabriel Pereira 학교의 데이터가 대부분을 차지하고 있다. 이를 고려하여 성능 분석 시

평가할 필요가 있다.

2. **sex** 변수 성별은 비교적 균등하게 분포되어 있어 성별에 따른 성적 차이를 분석하기 적합하다.

3. **address** 변수 거주지 유형(도시, 농촌) 간 데이터가 불균형하게 분포되어 있다. 도시 거주 학생이 압도적으로 많은 비율을 차지하며, 이는 거주지 유형에 따른 학업 성취도 차이가 모델에 미치는 영향을 평가하는 데 중요한 요인으로 작용할 수 있다.

4. **Mjob, Fjob (부모의 직업)** 부모의 직업 분포는 특정 직업군에 집중되어 있으며, 특히 어머니의 직업에서 'other' 카테고리가 가장 큰 비율을 차지하고 있다. 이는 모델이 특정 직업군에 과도하게 영향을 받을 수 있음을 시사하므로, 학습 시 이 변수의 기여도를 평가할 필요가 있다.

5. **higher (대학 진학 희망)** 대학교 진학 희망 여부는 'Yes'가 대부분을 차지하는 불균형한 분포를 보인다. 이는 대부분의 학생이 대학 진학을 희망하고 있음을 나타내며, 해당 변수의 중요성을 평가하는 것이 필요하다.

6. **internet, romantic** 변수 인터넷 접근 여부와 연애 관계 여부는 불균형하게 분포되어 있으며, 인터넷 접근이 가능한 학생들이 대부분을 차지한다. 이로 인해 인터넷 접근이 성적에 미치는 영향이 크게 나타날 수 있으므로, 모델 학습 시 해당 변수의 중요도를 평가해야 한다.

## 4.2 기본 통계 및 기술 통계 분석

### 4.2.1 수치형 변수 기술 통계

아래 표는 데이터셋에서 수치형 변수들의 기술 통계를 보여준다. 변수마다 평균(mean), 표준편차(std), 최솟값(min), 사분위수(25%, 50%, 75%), 최댓값(max), 왜도(skew), 첨도(kurtosis) 값을 계산하였다.

변수	count	mean	std	min	25%	50%	75%	max	skew	kurtosis
age	395.0	16.70	1.28	15.0	16.0	17.0	18.0	22.0	0.47	-0.00
Medu	395.0	2.75	1.09	0.0	2.0	3.0	4.0	4.0	-0.32	-1.09
Fedu	395.0	2.52	1.09	0.0	2.0	2.0	3.0	4.0	-0.03	-1.20
traveltime	395.0	1.45	0.70	1.0	1.0	1.0	2.0	4.0	1.61	2.34
studytime	395.0	2.04	0.84	1.0	1.0	2.0	2.0	4.0	0.63	-0.01
failures	395.0	0.33	0.74	0.0	0.0	0.0	0.0	3.0	2.39	5.00
famrel	395.0	3.94	0.90	1.0	4.0	4.0	5.0	5.0	-0.95	1.14
freetime	395.0	3.24	0.99	1.0	3.0	3.0	4.0	5.0	-0.16	-0.30
goout	395.0	3.11	1.11	1.0	2.0	3.0	4.0	5.0	0.12	-0.77
Dalc	395.0	1.48	0.89	1.0	1.0	1.0	2.0	5.0	2.19	4.76
Walc	395.0	2.29	1.29	1.0	1.0	2.0	3.0	5.0	0.61	-0.79
health	395.0	3.55	1.39	1.0	3.0	4.0	5.0	5.0	-0.49	-1.01
absences	395.0	5.71	8.00	0.0	0.0	4.0	8.0	75.0	3.67	21.72
G	395.0	10.91	3.32	3.0	8.0	11.0	13.0	19.0	0.24	-0.69

Table 2: 수치형 변수의 기본 통계와 왜도, 첨도

#### 4.2.2 범주형 변수 기술 통계

다음 표는 범주형 변수의 고유값 개수, 최빈값(mode), 최빈값 빈도 및 최빈값 비율을 나타낸다.

변수	고유값 개수	최빈값	최빈값 빈도
school	2	GP	349
sex	2	F	208
address	2	U	307
famsize	2	GT3	281
Pstatus	2	T	354
Mjob	5	other	141
Fjob	5	other	217
reason	4	course	145
guardian	3	mother	273
schoolsup	2	no	344
famsup	2	yes	242
paid	2	no	214
activities	2	yes	201
nursery	2	yes	314
higher	2	yes	375
internet	2	yes	329
romantic	2	no	263

Table 3: 범주형 변수들의 기술 통계량

### 4.3 데이터 요약 및 관계 탐색

데이터의 기본 통계 및 분포를 파악한 후, 변수 간의 관계를 탐구하였다. 성별, 부모의 교육 수준, 학습 시간, 결석 횟수, 인터넷 접근성 등이 성적(G1, G2, G3)에 미치는 영향을 중점적으로 분석하였다.

#### 4.3.1 성별과 학업 성취도

성별에 따른 학업 성취도(G)의 차이를 분석하였다. 남학생과 여학생의 성적 평균을 비교하여 성적 차이가 유의미한지 살펴보았다.

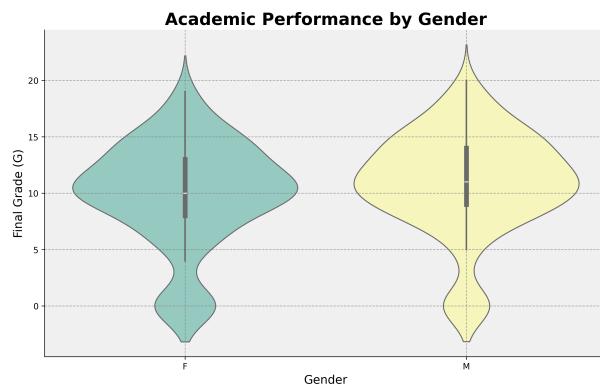


Figure 9: 성별 분포에 따른 성적 차이 분석

여성 학생의 성적 분포는 남학생에 비해 다양성이 크고, 일부 여학생의 성적이 더 낮은

경향이 있지만, 성적의 중앙값에서는 두 성별 간 큰 차이가 없다. 전체적으로 볼 때, 성별에 따른 성적 차이는 크지 않다고 판단된다.

#### 4.3.2 부모의 교육 수준과 성적 간의 관계

부모의 교육 수준이 높은 학생이 성적이 더 우수할 가능성이 높다는 가설을 세우고 이를 검토하였다. 어머니(Medu)와 아버지(Fedu)의 교육 수준이 성적에 미치는 영향을 분석하고, 부모의 교육 수준이 높은 학생들의 성적 평균을 비교하였다.

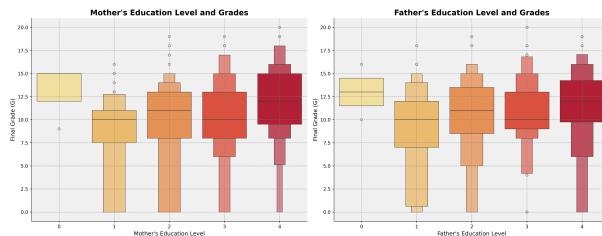


Figure 10: 부모의 교육 수준과 성적 간의 관계

어머니와 아버지의 교육 수준이 높을수록 학생의 성적이 전반적으로 상승하는 경향을 보인다. 그러나 그래프를 통해 확인할 수 있듯이, 어머니와 아버지의 교육 수준에 따른 성적 상승 효과는 큰 차이를 보이지 않으며, 두 부모 모두 비슷한 정도로 성적 향상에 기여하고 있음을 알 수 있다.

#### 4.3.3 학습 시간과 성적 간의 관계

학습 시간(studytime)이 성적에 미치는 영향을 분석하였다. 주간 학습 시간이 긴 학생들이 성적이 더 높은지 여부를 분석하고, 특히 학습 시간이 10시간 이상인 학생들이 성적 상위 그룹에 속하는 비율을 비교하였다.

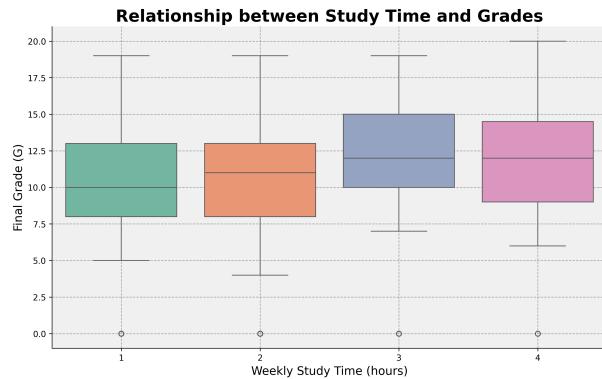


Figure 11: 학습 시간과 성적 간의 관계

학습 시간 길어질수록 증가하는 경향이 보이긴하나 3시간 이상인 경우에도 일정 수준의 변동성이 있으며, 모든 학생이 더 많이 공부할수록 성적이 높아지는 것은 아니라는 점을 알 수 있다.

#### 4.3.4 결석 횟수와 성적 간의 관계

결석 횟수(absences)가 성적에 미치는 부정적인 영향을 분석하였다. 결석 횟수가 많을수록 성적이 낮아지는 경향이 있는지 여부를 확인하였다.

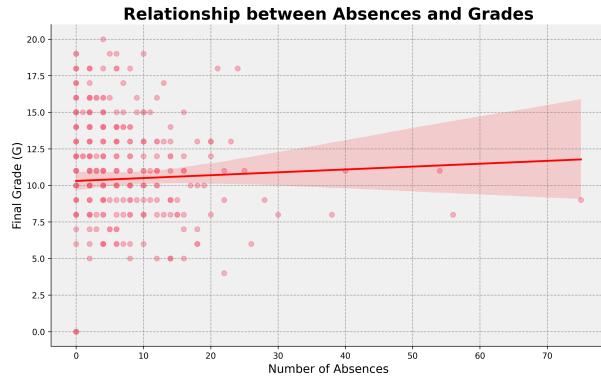


Figure 12: 결석 횟수와 성적 간의 관계

그 영향은 크지 않다. 일부 결석이 많음에도 성적이 우수한 학생들도 있다.

#### 4.3.5 인터넷 접근성과 성적 간의 관계

인터넷 접근 여부가 성적에 미치는 영향을 분석하였다. 인터넷에 접근할 수 있는 학생들이 성적 상위 그룹에 더 많이 속하는지 여부를 검토하였다.

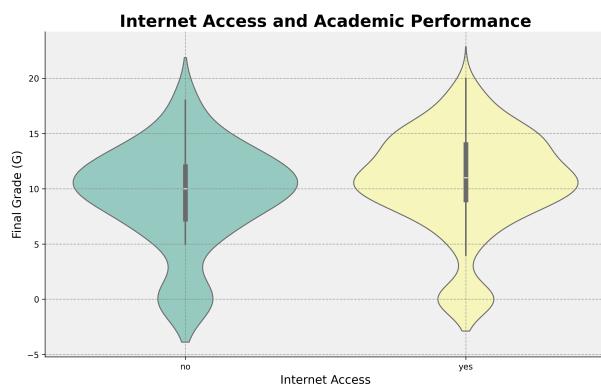


Figure 13: 인터넷 접근성과 성적 간의 관계

인터넷 접근이 가능한 학생들이 더 높은 성적을 얻는 경향이 있다. 인터넷이 없는 학생들은 성적 변동이 크고, 일부는 낮은 성적을 기록한다.

#### 4.3.6 수치형 및 범주형 변수 상관관계 분석

수치형 변수와 범주형 변수가 성적에 미치는 영향을 분석하였다.

Correlation Matrix of Numeric Variables

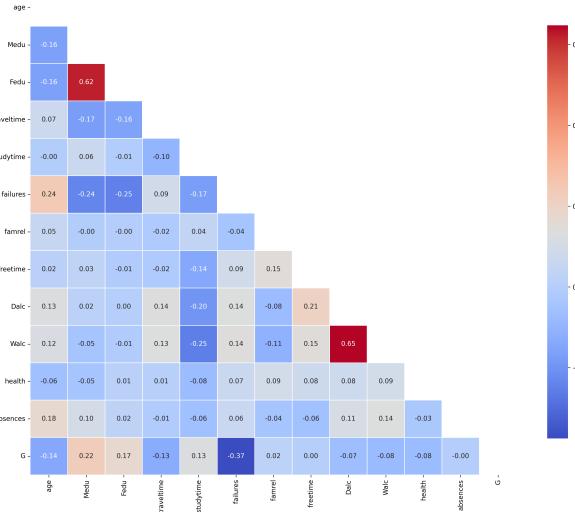


Figure 14: 주요 변수와 성적 간의 상관관계 (수치형 및 범주형)

상관 행렬에서 G1, G2, G3 성적 간의 높은 상관관계가 확인되며, ‘failures’ 변수는 성적과 음의 상관관계를 가진다.

#### 4.3.7 다변량 분석: 앤드류 커브 (Andrews Curves)

앤드류 커브(Andrews Curves)를 사용하여 각 학생의 특성을 다변량 분포로 시각화하고, 성적에 따른 군집의 특성을 파악하였다. 이를 통해 변수 간의 복잡한 상호작용을 탐구하였다.

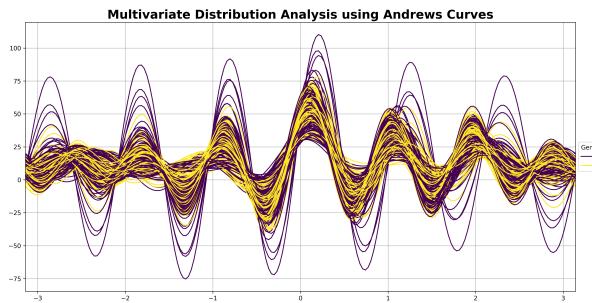


Figure 15: 앤드류 커브를 사용한 다변량 분포 분석

여성 학생의 변동성이 더 크며, 성별 간 데이터의 변동 패턴에서 일부 차이가 나타난다. 이상으로 데이터의 기본 분포 분석 및 주요 변수 간의 상관관계를 파악하는 EDA 작업을 마쳤다. 이를 기반으로 고급 통계 기법 및 머신러닝 모델을 적용하여 성적 예측 모델을 개발할 예정이다.

## 5 예비 통계 분석 및 가설 검정

본 보고서에서는 학생들의 성적에 영향을 미치는 주요 요인들을 파악하고, 이후 모델 구축에 활용하기 위해 사전적으로 다양한 통계 분석과 가설 검정을 수행하였다. 이러한 예비 분석은 모델링 과정에서 고려할 변수 선정과 변수 간의 관계를 이해하는 데 중요한 역할을 한다.

## 5.1 학습 시간과 성적 간의 상관관계 분석

학습 시간(studytime)과 성적(G) 간의 상관관계를 파악하기 위해 피어슨 상관분석을 실시하였다.

**귀무가설 (H0):** 학습 시간과 성적 간에는 상관관계가 없다.

**대립가설 (H1):** 학습 시간과 성적 간에는 상관관계가 있다.

분석 결과, 상관계수는 0.135로 나타났으며, p-값은 0.017로 유의수준 0.05보다 작게 나타났다. 이는 귀무가설을 기각하고 대립가설을 채택할 수 있음을 의미한다. 따라서, 학습 시간과 성적 간에는 유의미한 양의 상관관계가 있다고 할 수 있다.

변수	상관계수 (r)	p-값
학습 시간	0.135	0.017

Table 4: 학습 시간과 성적 간의 피어슨 상관분석 결과

## 5.2 어머니의 교육 수준에 따른 성적 차이 분석

어머니의 교육 수준(Medu)에 따른 성적의 평균 차이를 검정하기 위해 일원분산분석(ANOVA)을 실시하였다.

**귀무가설 (H0):** 어머니의 교육 수준에 따른 성적의 평균 차이는 없다.

**대립가설 (H1):** 어머니의 교육 수준에 따른 성적의 평균 차이가 있다.

ANOVA 결과, F-통계량은 6.236, p-값은 0.000으로 나타났다. 이는 유의수준 0.05에서 귀무가설을 기각할 수 있음을 의미한다. 따라서, 어머니의 교육 수준에 따라 학생들의 성적에 유의미한 차이가 있다고 할 수 있다.

요인	F-통계량	p-값
어머니의 교육 수준	6.236	0.000

Table 5: 어머니의 교육 수준에 따른 성적의 ANOVA 결과

## 5.3 결석 횟수와 성적 간의 회귀 분석

결석 횟수(absences)가 성적에 미치는 영향을 파악하기 위해 단순 선형 회귀분석을 실시하였다.

**귀무가설 (H0):** 결석 횟수는 성적에 영향을 미치지 않는다.

**대립가설 (H1):** 결석 횟수는 성적에 유의미한 영향을 미친다.

분석 결과, 결석 횟수의 회귀 계수에 대한 p-값은 0.896으로 나타나 유의수준 0.05보다 크게 나타났다. 따라서 귀무가설을 기각하지 못하며, 결석 횟수는 성적에 유의미한 영향을 미치지 않는 것으로 판단된다.

변수	회귀계수	표준오차	p-값
상수항	0.523	0.014	0.000
결석 횟수	-0.014	0.105	0.896

Table 6: 결석 횟수와 성적 간의 회귀 분석 결과

## 5.4 성별에 따른 성적 차이 검정

성별(sex)에 따른 성적의 평균 차이를 검정하기 위해 독립표본 t-검정을 실시하였다.

**귀무가설 (H0):** 성별에 따른 성적의 평균 차이는 없다.

**대립가설 (H1):** 성별에 따른 성적의 평균 차이가 있다.

분석 결과, t-통계량과 p-값이 NaN으로 나타났다. 이는 남학생 또는 여학생 그룹의 표본 크기가 너무 작거나 분산이 없어 검정이 수행되지 않았을 가능성이 있다. 따라서 성별에 따른 성적의 평균 차이에 대해 유의미한 결론을 도출할 수 없었다.

## 5.5 다중 비교에 대한 FDR 제어

본 보고서에서는 여러 변수에 대한 가설 검정을 동시에 수행함에 따라 다중 비교로 인한 1종 오류(Type I error)의 증가 가능성이 있다. 이는 여러 번의 통계적 검정을 수행할수록 우연에 의해 유의미한 결과로 나타날 확률이 높아지기 때문이다. 이러한 문제를 보정하기 위해 Benjamini/Hochberg 방법을 사용하여 FDR(False Discovery Rate)을 제어하였다.

FDR은 다중 검정 시 거짓 발견의 비율을 의미하며, Benjamini/Hochberg 방법은 주어진 유의수준 내에서 이 비율을 제어하는 절차이다. 이를 통해 검정의 통계적 파워를 유지하면서도 다중 비교로 인한 오류를 효과적으로 통제할 수 있다.

변수	원래 p-값	FDR 보정된 p-값	유의미한 결과
학습 시간 (studytime)	0.017	0.025	예
어머니의 교육 수준 (Medu)	0.000	0.000	예
결석 횟수 (absences)	0.896	0.896	아니오

Table 7: 다중 비교에 대한 FDR 보정 결과

FDR 보정 결과, 학습 시간과 어머니의 교육 수준은 성적과 유의미한 관계가 있는 것으로 나타났으며, 결석 횟수는 유의미한 관계가 없는 것으로 확인되었다. 이는 다중 비교 문제로 인한 1종 오류를 효과적으로 통제하였으며, 결과의 신뢰성을 높였음을 의미한다.

## 5.6 예비 분석의 요약 및 모델 구축을 위한 시사점

이상의 예비 분석 결과를 종합하면 다음과 같다.

- **학습 시간**은 성적과 유의미한 양의 상관관계를 보였다. 이는 모델 구축 시 중요한 예측 변수로 고려될 수 있음을 시사한다.
- **어머니의 교육 수준**은 성적에 유의미한 영향을 미쳤다. 따라서 부모의 교육 수준도 모델에 포함하는 것이 적절하다.
- **결석 횟수**는 성적에 유의미한 영향을 미치지 않는 것으로 나타났다. 이는 모델 구축 시 해당 변수를 제외하거나 추가적인 분석을 통해 재검토할 필요가 있다.
- **성별**에 따른 성적 차이는 통계적으로 유의미한 결론을 도출할 수 없었다. 데이터의 한계로 인해 모델에서의 성별 변수 활용은 신중히 고려해야 한다.
- **다중 비교 문제**에 대해 FDR 보정 방법을 적용하여 1종 오류를 통제하였으며, 이를 통해 변수 선택의 신뢰성을 향상시켰다.

본 보고서의 예비 통계 분석을 통해 학습 시간과 어머니의 교육 수준이 성적 예측에 있어 중요한 변수임을 확인하였다. 이러한 결과는 이후 진행될 모델 구축 과정에서 변수 선택과 모델의 구조를 결정하는 데 있어 근거로 활용될 것이다. 이를 바탕으로 보다 정확하고 신뢰성 있는 예측 모델을 개발하여 교육 현장에서의 효과적인 활용 방안을 제시하고자 한다.

## 6 머신러닝 모델 개발 및 평가

본 보고서에서는 다양한 머신러닝 모델을 개발하여 학생들의 성적을 예측하고, 이들의 성능을 비교 평가하였다. 각 모델의 성능은 RMSE(Root Mean Square Error) 지표를 통해 평가되었으며, 라이브러리에서 제공하는 기본 파라미터로 각 모델을 실행한 결과를 바탕으로 분석을 진행하였다.

### 6.1 Random Forest 모델

랜덤 포레스트(Random Forest)는 여러 개의 결정 트리(decision trees)를 결합하여 예측의 안정성을 높이는 앙상블 모델이다. 과적합을 방지하고, 여러 트리를 통해 다양한 샘플에 대한 예측을 할 수 있는 강점이 있다.

본 보고서에서 사용된 Random Forest 모델은 기본 파라미터로 학습되었으며, 그 결과는 다음과 같다:

- RMSE: 0.183682
- MAE: 0.146346

### 6.2 Gradient Boosting 모델

Gradient Boosting Trees는 순차적으로 약한 학습기를 결합하여 성능을 향상시키는 앙상블 기법이다. 각 단계에서 잘못 예측된 샘플에 더 많은 가중치를 부여하여 모델을 개선한다.

Gradient Boosting 모델은 기본 파라미터로 학습되었으며, 그 결과는 다음과 같다:

- RMSE: 0.187382
- MAE: 0.145621

### 6.3 XGBoost 모델

XGBoost는 Gradient Boosting 알고리즘을 개선한 모델로, 더 빠르고 정확하게 학습할 수 있다. 특히 효율적인 메모리 사용과 과적합 방지 기능을 제공한다.

XGBoost 모델은 기본 파라미터로 학습되었으며, 그 결과는 다음과 같다:

- RMSE: 0.208828
- MAE: 0.163830

### 6.4 LightGBM 모델

LightGBM은 대규모 데이터셋에서 학습 속도가 빠르고 효율적인 앙상블 모델로, 리프 중심의 트리 성장 방식을 사용하여 성능을 극대화한다.

LightGBM 모델은 기본 파라미터로 학습되었으며, 그 결과는 다음과 같다:

- RMSE: 0.192700
- MAE: 0.158672

## 6.5 CatBoost 모델

CatBoost는 범주형 변수를 자동으로 처리하는 강점을 가진 모델로, Gradient Boosting 기반의 앙상블 모델이다.

CatBoost 모델은 기본 파라미터로 학습되었으며, 그 결과는 다음과 같다:

- RMSE: 0.186662
- MAE: 0.146696

## 6.6 Linear Regression 모델

선형 회귀(Linear Regression)는 변수 간의 선형 관계를 기반으로 예측하는 간단한 모델이다.

Linear Regression 모델은 기본 파라미터로 학습되었으며, 그 결과는 다음과 같다:

- RMSE: 0.184108
- MAE: 0.150354

## 6.7 Ridge 모델

Ridge 회귀는 선형 회귀의 변형으로, L2 정규화를 통해 과적합을 방지한다.

Ridge 모델은 기본 파라미터로 학습되었으며, 그 결과는 다음과 같다:

- RMSE: 0.182919
- MAE: 0.149427

## 6.8 Lasso 모델

Lasso 회귀는 L1 정규화를 통해 변수를 선택하는 기능을 추가한 선형 회귀 모델이다.

Lasso 모델은 기본 파라미터로 학습되었으며, 그 결과는 다음과 같다:

- RMSE: 0.206378
- MAE: 0.165036

## 6.9 모델 비교 및 최종 평가

각 모델의 성능을 RMSE, MAE기준으로 비교한 결과는 다음 표와 같다:

모델	RMSE	MAE
Random Forest	0.183682	0.146346
Gradient Boosting	0.187382	<b>0.145621</b>
XGBoost	0.208828	0.163830
LightGBM	0.192700	0.158672
CatBoost	0.186662	0.146696
Linear Regression	0.184108	0.150354
Ridge	<b>0.182919</b>	0.149427
Lasso	0.206378	0.165036

Table 8: 모델별 성능 비교 (RMSE, MAE)

## 6.10 하이퍼파라미터 튜닝 및 교차검증

기본 설정으로 수행된 모델들에서 앙상블 모델의 성능이 기대보다 낮았던 이유는 하이퍼파라미터 최적화가 부족했기 때문으로 판단되었다. 이에 따라, 각 모델의 성능을 향상시키기 위해 **하이퍼파라미터 튜닝**을 진행하고, 이를 통해 모델의 일반화 성능을 평가하였다.

모델의 일반화 성능을 검증하기 위해 **5-fold 교차 검증(cross-validation)**을 사용하였다. 교차 검증은 데이터를 여러 개의 폴드(fold)로 나누어 각 폴드에 대해 모델을 학습하고 평가하는 방식으로, 모델이 새로운 데이터에 대해 얼마나 잘 일반화되는지를 확인하는 데 유용하다.

하이퍼파라미터 튜닝 후 각 모델의 성능을 RMSE, MAE로 평가하였으며, 그 결과는 다음과 같다:

모델	RMSE	MAE
RandomForest	0.179681	<b>0.141293</b>
GradientBoosting	0.183857	0.143913
XGBoost	0.184051	0.145959
<b>LightGBM</b>	<b>0.178568</b>	0.145329
CatBoost	0.185512	0.147295
Ridge	0.180556	0.145878
Lasso	0.206378	0.165036

Table 9: 하이퍼파라미터 튜닝 후 5-fold 교차 검증을 통한 모델 성능 평가 (성능이 가장 좋은 값은 볼드처리)

RandomForest의 우수한 성능은 이 모델이 교육 데이터의 이상치와 복잡한 패턴을 효과적으로 처리함을 시사한다. 반면, Lasso의 낮은 성능은 변수를 단순히 지워버리는 접근이 이 데이터셋에 적합하지 않음을 나타낸다. 이는 각 변수의 복잡한 영향을 고려하는 것이 중요함을 의미하며, 향후 특성 중요도 평가 시 이러한 결과를 바탕으로 더 세밀한 분석이 필요할 것이라고 판단하였다.

## 6.11 특성 중요도 분석

특성 중요도 분석을 통해 모델이 예측 성능을 향상시키기 위해 각 변수를 얼마나 많이 사용하는지를 확인할 수 있다. 아래 그림은 랜덤 포레스트 모델의 특성 중요도를 나타낸다. 분석 결과, **실패 횟수(failures)**와 **결석 일수(absences)**가 성적 예측에서 가장 중요한 변수로 확인되었다.

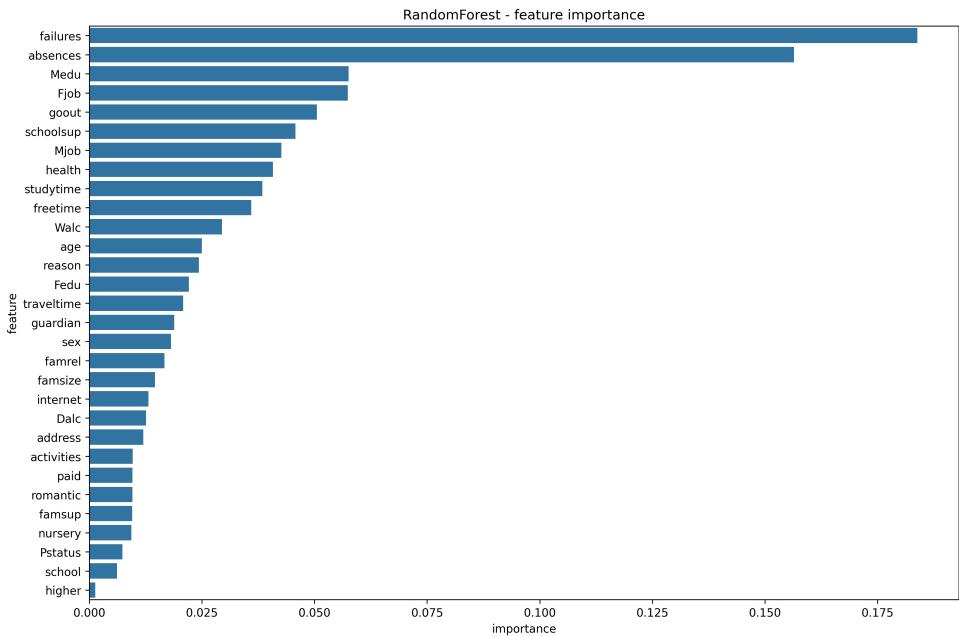


Figure 16: 랜덤 포레스트 모델의 특성 중요도 분석

### 6.11.1 주요 인사이트 도출

위 그림에서 확인할 수 있듯이, 실패 횟수(failures)와 결석 일수(absences)가 성적 예측에서 가장 중요한 특성으로 나타났다. 이는 학생들이 과거 학업에서 실패를 많이 경험했거나, 결석 일수가 많으면 성적에 부정적인 영향을 미친다는 점을 반영하고 있다.

**실패 횟수(failures)와 성적의 관계** 실패 횟수(failures)는 성적 예측에서 가장 중요한 변수로, 실패 경험이 많을수록 성적에 부정적인 영향을 미치는 경향이 있다. 이는 학업에 대한 심리적 부담감이나 학습 동기 저하와 관련이 있을 수 있다. 교육적인 지원이나 추가 학습 지원이 필요한 학생들을 조기에 식별하는 데 도움이 될 수 있다.

**결석 일수(absences)와 성적의 관계** 결석 일수(absences)와 성적 간의 관계를 분석한 결과, 결석 일수는 두 번째로 중요한 변수로 나타났다. 결석 일수는 앞서 언급한 것처럼 이상치 처리 과정에서 문제가 되었으나, 윈저화(Winsorization)를 통해 이상치를 조정함으로써 여전히 중요한 변수로 남아 있다. 결석 일수가 많을수록 학습 기회가 줄어들고 성적에 부정적인 영향을 미치는 경향이 뚜렷하다. 이는 결석이 학생의 학업 성취에 중요한 요소임을 시사하며, 교육 현장에서 결석률을 줄이는 노력이 필요함을 강조한다.

하지만 가설 검정 결과에서는 결석 일수와 성적 간에 유의미한 상관관계가 나타나지 않았다. 이는 가설 검정이 단순한 상관관계를 평가하는 데 그치며, 다른 변수와의 상호작용을 고려하지 않기 때문일 수 있다. 반면, 머신러닝 모델에서는 여러 변수들이 동시에 고려되며, 결석 일수와 성적 사이의 비선형적 관계나 상호작용이 반영될 수 있다. 따라서 결석 일수가 다른 요인들과 결합하여 성적에 중요한 영향을 미치는 것으로 보인다. 이러한 결과는 결석 일수가 단순한 상관관계로는 충분히 설명되지 않더라도, 다변량 모델에서는 성적 예측에 중요한 역할을 할 수 있음을 시사한다.

**어머니의 교육 수준(Medu)** 어머니의 교육 수준(Medu) 또한 성적 예측에 중요한 영향을 미치는 변수로 나타났다. 이는 부모의 교육 수준이 자녀의 학습 환경에 중요한 영향을 미친다는 기준 보고서 결과와 일치한다. 특히 어머니의 교육 수준이 높을수록 학습 지원이 더 잘 이루어져 성적이 우수할 가능성성이 높음을 보여준다.

기타 중요한 변수 이 외에도 아버지의 직업(Fjob), 외출 빈도(goout), 추가 교육 지원(schoolsup), 주간 학습 시간(studytime) 등도 중요한 특성으로 나타났다. 이는 학생들의 생활 환경, 부모의 지원, 그리고 학습 시간이 성적에 중요한 영향을 미친다는 것을 의미한다.

**결론** 특성 중요도 분석을 통해 확인된 변수들은 학업 성취에 큰 영향을 미치는 요인들로, 특히 실패 횟수(failures)와 결석 일수(absences)가 성적 예측에 가장 큰 영향을 미치는 것으로 확인되었다. 결석 일수와 관련된 이상치 문제는 원저화를 통해 효과적으로 처리되었으며, 그 결과 중요한 예측 변수로 남아 있음을 확인할 수 있었다. 이러한 결과는 학업 성취에 영향을 미치는 다양한 요인들을 고려하여 교육 정책을 설계하는 데 중요한 시사점을 제공한다.

## 6.12 모델 비교 및 최종 모델 선정

최종적으로, 다양한 머신러닝 모델들 간의 성능을 비교한 결과, 앞서 Random Forest 모델이 가장 우수한 성능을 기록하였다. 비선형 데이터의 복잡한 패턴을 잘 포착하여, 학생 성적 예측에 매우 적합한 모델로 평가되었다. Gradient Boosting Trees와 다른 앙상블 모델들도 준수한 성능을 보였으나, RandomForest에 비해 약간의 성능 차이가 있었다.

모델	최종 RMSE
<b>RandomForest</b>	<b>0.179681</b>
Gradient Boosting Trees	0.183857
XGBoost	0.184051
<b>LightGBM</b>	<b>0.178568</b>
CatBoost	0.185512

Table 10: 최종 모델 성능 비교

이후, 최종적으로 소프트보팅(Soft Voting) 기법을 적용하여 여러 모델의 예측을 결합하였으며, 가중치는 각각 RandomForest에 0.7, Gradient Boosting Trees에 0.2, LightGBM에 0.1로 설정하였다.

이전에 학습한 랜덤 포레스트 모델의 특성 중요도를 기반으로 중앙값을 기준으로 주요 특성을 선택하여 학습을 진행하였다. 선택된 주요 특성은 ['reason', 'goout', 'Medu', 'traveltime', 'schoolsup', 'failures', 'Fjob', 'studytime']이다.

다음은 특성 선택 전후의 성능 비교 결과이다:

- **전체 변수를 사용한 경우** Final Ensemble Model - RMSE: 0.177365, MAE: 0.142754, MAPE: 0.639951
- **특성 선택 후 사용한 경우** Final Ensemble Model - RMSE: 0.195342, MAE: 0.154398, MAPE: 0.764300

위의 결과를 통해, 특성 선택 후 성능이 오히려 저하된 것을 확인할 수 있었다. 이는 각 특성들이 데이터 예측에 중요한 역할을 하고 있음을 의미하며, Lasso 모델에서 일부 변수의 중요도가 과소평가되었을 가능성을 시사한다. 따라서, 모든 특성을 포함한 학습이 이 데이터셋에서는 더 나은 성능을 보이는 것으로 분석되었다.

## 7 모델 해석 및 설명 가능성

모델의 예측 성능뿐만 아니라 예측 결과를 해석할 수 있는 능력도 매우 중요하다. 특히 교육 분야에서 예측 결과는 정책 결정과 관련된 중요한 정보를 제공할 수 있기 때문에, 그 과정에서 왜 특정한 결과가 나왔는지를 설명할 수 있어야 한다. 이 섹션에서는 SHAP(SHapley

Additive exPlanations), 부분 의존성 플롯(Partial Dependence Plot, PDP), LIME(Local Interpretable Model-agnostic Explanations) 등의 해석 기법을 사용하여 트리 기반 모델의 예측을 어떻게 해석할 수 있는지 설명한다.

## 7.1 SHAP (SHapley Additive exPlanations) 값 분석

SHAP는 협력 게임 이론의 Shapley 값으로 모델의 예측에 각 특성이 어떻게 기여하는지를 정량적으로 나타내는 방법이다.

### 7.1.1 이론적 배경

Shapley 값은 협력 게임 이론에서 각 플레이어가 게임의 결과에 기여한 정도를 공정하게 분배하는 방법이다. 이를 머신러닝에 적용하면, 각 특성이 모델의 예측 결과에 어떻게 기여했는지를 계산할 수 있다.

### 7.1.2 SHAP 특성 중요도 (Bar Plot)

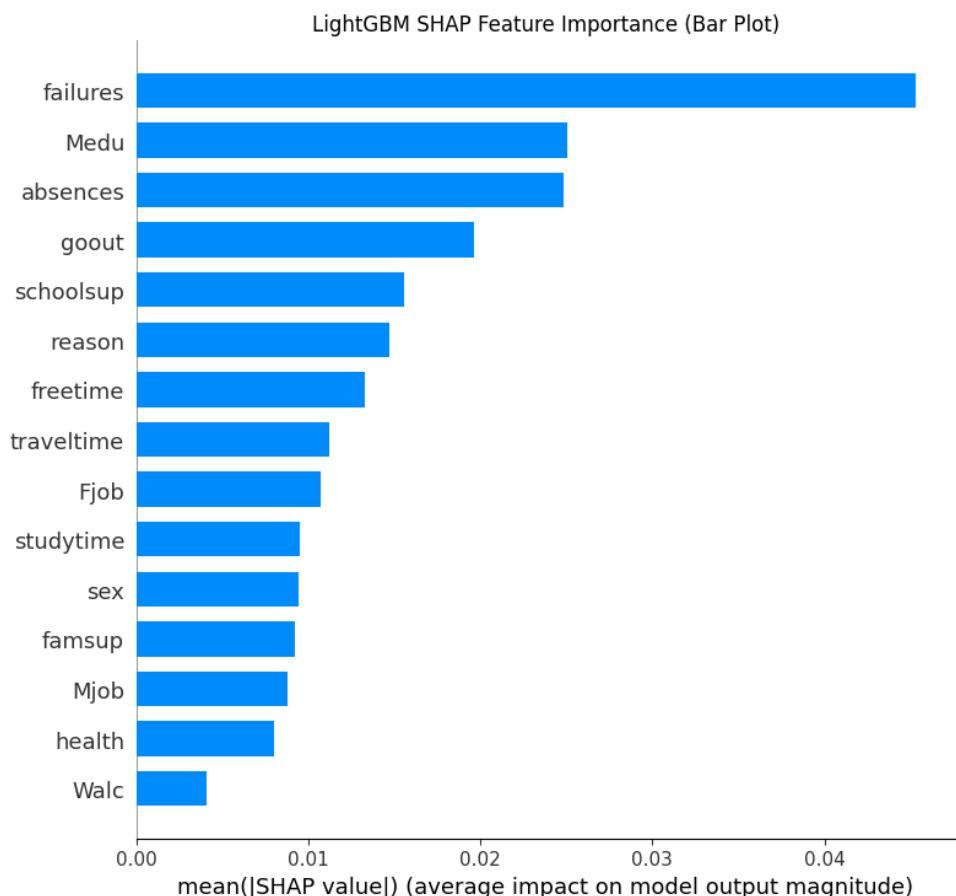


Figure 17: SHAP 특성 중요도 (Bar Plot)

그림 17는 각 특성의 중요도를 SHAP 값으로 나타낸 결과이다. 결석 일수(absences), 어머니의 교육 수준(Medu), 실패 횟수(failures)가 성적 예측에서 가장 큰 영향을 미치는 변수임을 알 수 있다.

### 7.1.3 SHAP 특성 영향 (Dot Plot)

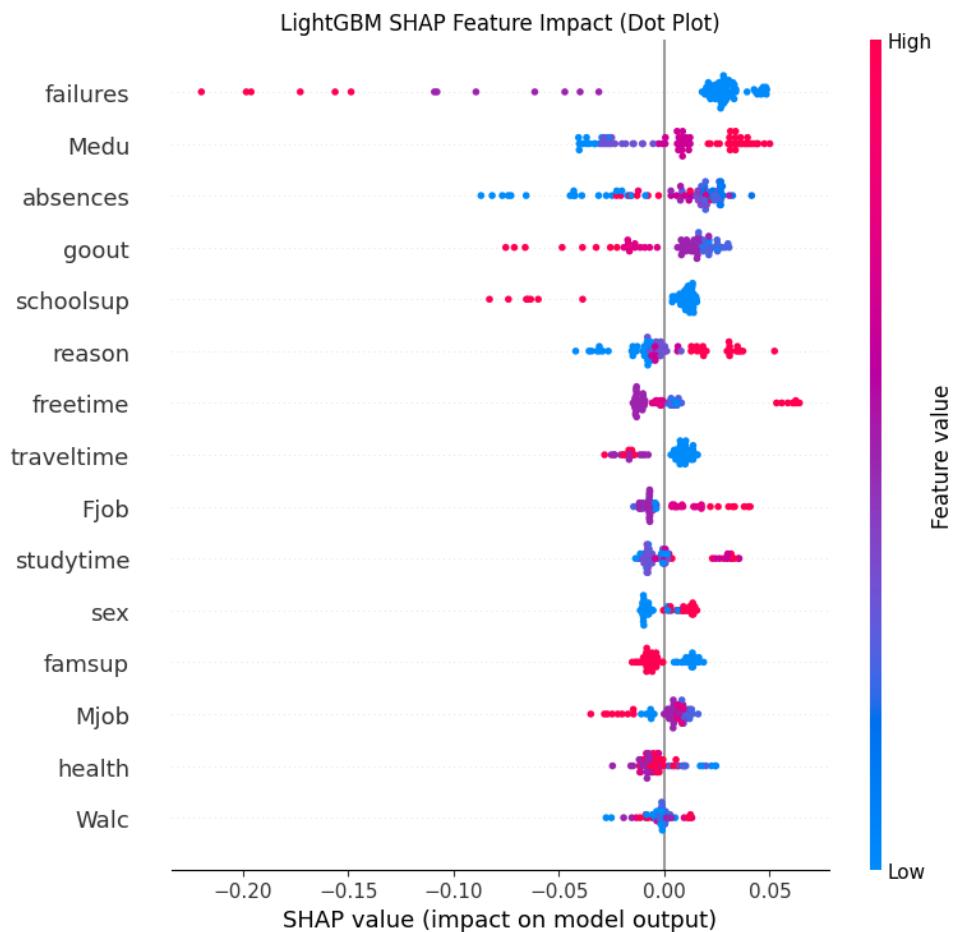


Figure 18: SHAP 특성 영향 (Dot Plot)

그림 18은 각 특성 값이 모델의 예측에 미치는 영향을 보여준다. SHAP 값의 크기와 방향에 따라 성적 예측이 어떻게 변화하는지를 시각적으로 나타낸다. 실패 횟수(failures)는 높을 수록 부정적인 영향을 주며, 어머니의 교육 수준(Medu)은 높을 수록 긍정적인 영향을 주는 것을 알 수 있다.

## 7.2 SHAP Dependence Plot

특정 변수의 값 변화가 모델의 예측에 미치는 영향을 다른 변수와의 상호작용과 함께 시각화하는 방법이다.

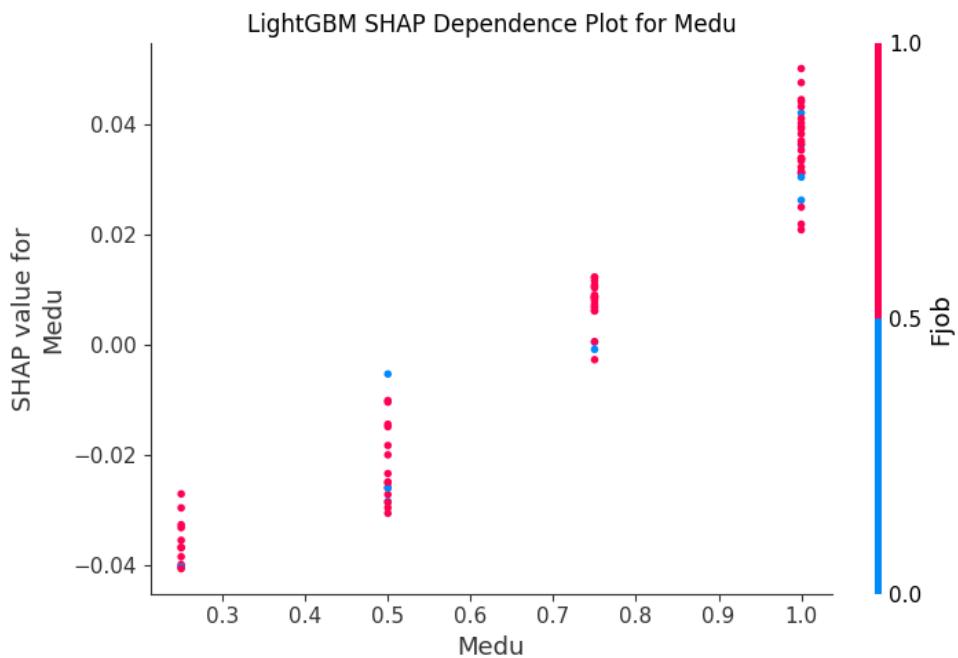


Figure 19: SHAP Dependence Plot: 어머니의 교육 수준(Medu)과 아버지의 직업(Fjob)

그림 19는 아버지의 직업이 있고 어머니의 교육 수준(Medu)이 높을수록 성적 예측에 긍정적인 영향을 미친다는 사실을 보여준다. 즉, 아버지가 직업이 있고 교육 수준이 높은 어머니를 둔 학생일수록 성적이 더 높게 예측된다는 것을 알 수 있다.

### 7.3 부분 의존성 플롯 (Partial Dependence Plot, PDP)

PDP는 특정 특성과 목표 변수 간의 관계를 시각적으로 보여주는 도구이다. 다른 특성들이 고정된 상태에서 특정 특성의 값이 변화할 때 모델의 예측이 어떻게 변하는지를 보여준다.

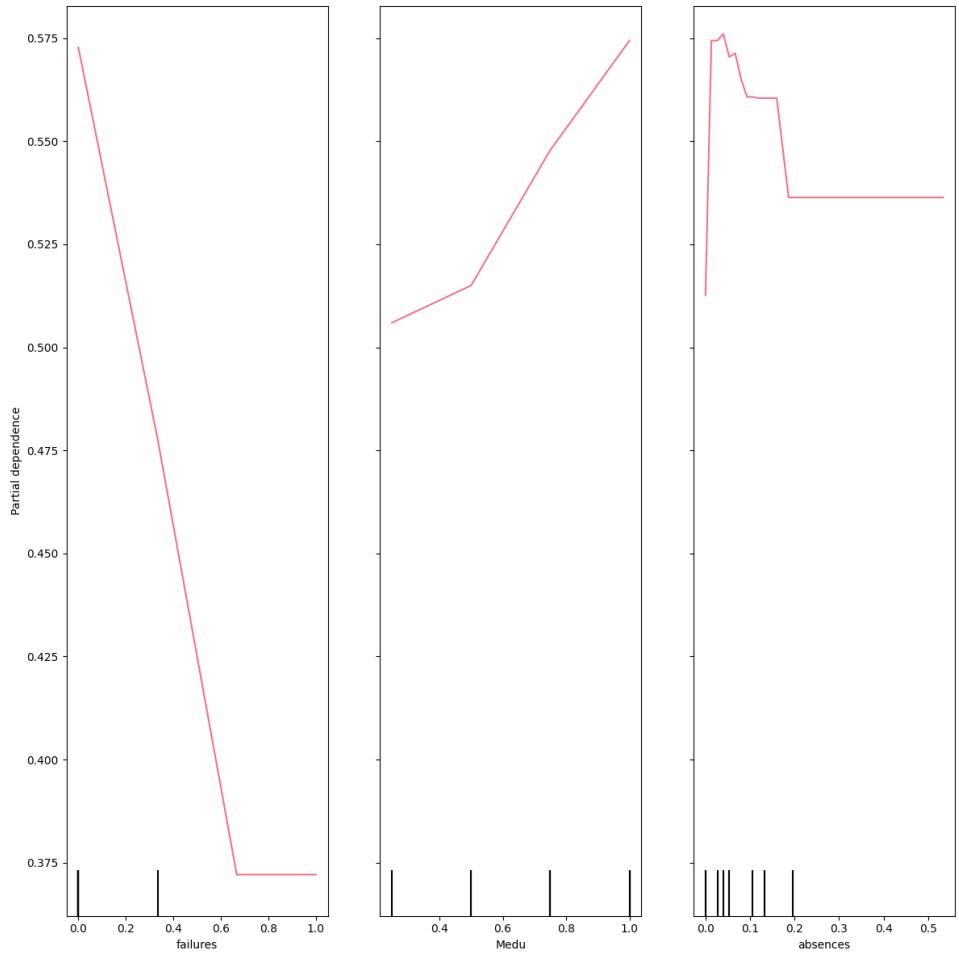


Figure 20: 부분 의존성 플롯 (PDP)

그림 20에서, 결석 일수가 증가할수록 성적 예측이 감소하는 경향을 확인할 수 있다. 반면, 어머니의 교육 수준(Medu)이 높아질수록 성적 예측이 증가하는 경향을 보인다.

#### 7.4 LIME (Local Interpretable Model-agnostic Explanations)

LIME은 복잡한 모델의 예측을 지역적으로 해석하기 위한 방법이다. 모델의 예측을 로컬한 관점에서 설명 가능한 선형 모델로 근사하여, 개별 예측에 어떤 변수가 가장 크게 영향을 미쳤는지 파악할 수 있다.

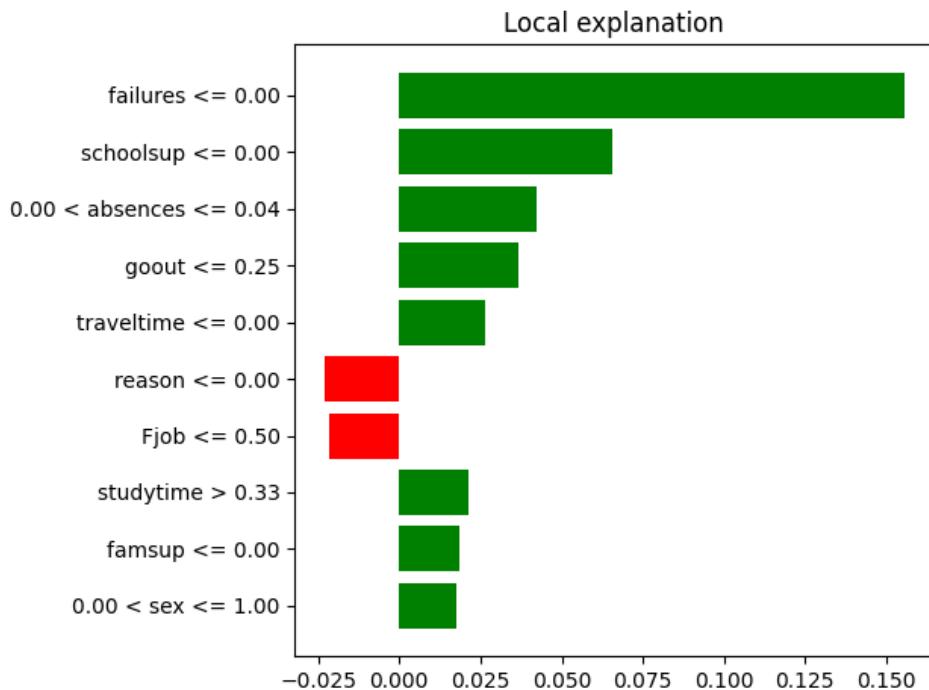


Figure 21: LIME 분석 결과

그림 21는 특정 학생에 대한 예측 결과를 해석한 것으로, 결석 일수와 학교 지원(schoolsup)이 이 학생의 성적 예측에 크게 기여한 것을 확인할 수 있다. 결석 일수는 성적 예측을 낮추는 주요 요인이고, 학교 지원은 성적 예측을 긍정적으로 영향을 미쳤다.

## 7.5 모델 해석의 중요성

이와 같은 다양한 해석 기법을 사용하면 모델의 예측 과정을 더욱 명확히 설명할 수 있으며, 교육 정책과 같은 실제 문제에 이를 적용할 수 있다. 결석 일수, 어머니의 교육 수준, 실패 횟수와 같은 변수들이 성적에 중요한 영향을 미친다는 점을 고려할 때, 출석률을 높이기 위한 정책이나 학업 실패 경험을 극복할 수 있는 프로그램을 제공하는 것이 학생들의 성적 향상에 도움이 될 수 있다.

모델 해석은 단순히 상관관계 이상의 인사이트를 제공하며, 이를 바탕으로 데이터 기반의 의사결정을 내릴 수 있는 강력한 도구이다. SHAP, PDP, LIME와 같은 기법들은 이러한 해석을 가능하게 만들어 머신러닝 모델을 “블랙박스”에서 “투명한 상자”로 바꾸는 역할을 한다.

이와 같은 해석 기법을 통해 우리는 학생 성적에 영향을 미치는 요인들을 정확히 파악하고, 더 나은 교육 전략을 세울 수 있을 것이다.

## 8 응용: SHAP와 최적 수송 이론을 결합한 개인화된 학업 성취도 향상 전략

최종적으로 우리는 SHAP와 최적 수송 이론을 결합하여 학생들의 학업 성취도를 향상시키는 개인화된 전략을 제시한다. 이 절에서는 제안된 이론적 개념과 방법론에 대한 상세한 설명과 증명을 다룬다.

## 8.1 SHAP와 최적 수송 이론의 결합

Shapley Additive exPlanations (SHAP)는 예측 모델이 각 특성에 대해 예측에 기여한 정도를 공정하게 분배하는 방법론이다. 각 특성에 할당된 SHAP 값은 해당 특성의 기여도를 나타내며, 특성의 중요도를 측정하는 데 효과적으로 사용된다. 본 보고서에서는 SHAP 값을 이용하여 각 학생의 성적을 예측하는 데 중요한 특성을 도출하고, 이를 통해 개선 경로를 제시하기 위해 최적 수송 이론을 적용하였다.

최적 수송 문제는 현재 상태에서 목표 상태로 이동하는 데 필요한 최소 비용을 계산하는 문제이다. 이를 성적 향상 문제에 적용하면, 현재 학생의 특성 벡터  $x$ 에서 목표 성적을 가지는 특성 분포  $\nu$ 로 이동하기 위한 최적 경로를 찾을 수 있다.

## 8.2 최적 수송 문제 설정

학생의 현재 특성 벡터  $x \in \mathcal{X}$ 가 주어졌을 때, 목표 성적  $y_{\text{target}}$ 을 달성하기 위한 목표 분포  $\nu$ 로의 이동을 최소 비용으로 해결하는 최적 수송 문제를 정의한다. 여기서  $\delta_x$ 는 학생의 현재 상태를 나타내는 디락 델타 측도이며,  $\nu$ 는 목표 성적을 만족하는 특성 분포를 의미한다.

최적 수송 문제는 다음과 같이 정의된다:

[최적 수송 문제] 현재 학생의 특성  $x \in \mathcal{X}$ 에서 목표 분포  $\nu$ 로 이동하는 최적 수송 계획  $\pi^*$ 은 다음 식을 최소화하는  $\pi$  중 하나로 정의된다:

$$\pi^* = \min_{\pi \in \Pi(\delta_x, \nu)} \int_{\mathcal{X} \times \mathcal{X}} c(x, x') d\pi(x, x') \quad (2)$$

여기서  $c(x, x')$ 는 특성 간의 이동 비용을 나타내는 함수로, SHAP 값에 기반한 가중치로 정의된다.

## 8.3 비용 함수 정의 및 해석

비용 함수  $c(x, x')$ 는 학생의 현재 특성에서 목표 특성으로 이동하는 데 필요한 비용을 계산하는 함수이다. SHAP 값을 통해 중요한 특성에 가중치를 부여함으로써, 더 중요한 특성에 더 큰 변화가 가능하도록 한다. 비용 함수는 다음과 같이 정의된다:

$$c(x, x') = \sum_{i=1}^d w_i(|\phi_i(x)|)(x'_i - x_i)^2 \quad (3)$$

여기서  $w_i(|\phi_i(x)|)$ 는 SHAP 값  $\phi_i(x)$ 에 의해 가중치가 부여된 함수로, 특성  $i$ 의 중요도에 비례한다. 이는 각 특성의 중요도를 반영하여 최적 경로에서 중요한 특성의 변화가 더 많이 허용되도록 한다.

## 8.4 이론적 증명

### 1. 최적 수송 계획의 존재성

**정리 1** (최적 수송 계획의 존재성). 비용 함수  $c(x, x')$ 가 하한을 가지며 연속적일 때, 최적 수송 문제는 Kantorovich-Rubinstein 정리에 의해 해  $\pi^*$ 가 존재한다.

*Proof.* Kantorovich-Rubinstein 정리는 비용 함수  $c(x, x')$ 가 연속적이고 하한을 가지면, 최적 수송 문제의 해가 존재함을 보장한다. 본 보고서에서 사용된 비용 함수  $c(x, x')$ 는 SHAP 값을 기반으로 정의된 가중치를 포함하며, 각 특성의 변화에 대해 제곱 거리를 사용하므로 연속적이고 하한을 가진다. 이에 따라,  $\delta_x$ 와  $\nu$  사이의 최적 수송 계획  $\pi^*$ 는 반드시 존재하며, 이는 현실적으로 최소한의 변화를 의미한다.  $\square$

### 2. SHAP 값의 통계적 유의성

**정리 2** (SHAP 값의 통계적 의미). SHAP 값  $\phi_i(x)$ 는 특성  $i$ 가 예측에 미치는 평균 기여도를 의미하며, 이는 공정성과 일관성을 만족한다.

*Proof.* SHAP 값은 특성  $i$ 가 예측 결과에 기여한 정도를 수학적으로 정의하며, 다음 식으로 계산된다:

$$\phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (4)$$

여기서  $f(S)$ 는 특성 집합  $S$ 를 사용한 모델의 예측값을 의미한다. SHAP 값은 Lundberg 와 Lee의 보고서에 의해 공정성과 일관성을 보장하며, 특성의 중요도를 정확하게 반영하는 유의미한 통계적 도구이다.  $\square$

## 8.5 개인화된 학습 피드백의 구성

위에서 설명한 SHAP와 최적 수송 이론을 결합하여 학생에게 제공할 수 있는 피드백의 예시는 다음과 같다. 이 피드백은 SHAP 값에 따른 특성의 중요도를 바탕으로 최적 경로를 제시하며, 이를 통해 학생은 목표 성적을 달성하기 위해 어떤 특성을 변화시켜야 할지 알 수 있다.

## 8.6 알고리즘

---

**Algorithm 1** Personalized Academic Improvement Strategy using SHAP and Optimal Transport

---

**Require:** Student's feature vector  $x$ , prediction model  $f$ , target grade  $y_{target}$ , SHAP function  $\phi$

**Ensure:** Personalized improvement strategy  $S$

- 1: Calculate the optimal transport plan  $\pi^* =_{\pi \in \Pi(\delta_x, \nu)} \int c(x, x') d\pi(x, x')$
- 2: Calculate the optimal change:  $\Delta x^* = \int (x' - x) d\pi^*(x, x')$
- 3: Compute importance scores for each feature:  $I_i = |\phi_i(x)| \cdot |\Delta x_i^*|, \forall i \in \{1, \dots, d\}$
- 4: Determine the improvement direction for each feature:  $D_i = \text{sign}(\phi_i(x)), \forall i \in \{1, \dots, d\}$
- 5: Sort features in descending order based on importance scores
- 6: Generate improvement recommendations  $S$  for the top  $k$  features
- 7: **for**  $i$  in top  $k$  features **do**
- 8:     **if**  $D_i > 0$  **then**
- 9:          $S_i \leftarrow \text{"Increase feature } i\text{"}$
- 10:     **else if**  $D_i < 0$  **then**
- 11:          $S_i \leftarrow \text{"Decrease feature } i\text{"}$
- 12:     **else**
- 13:          $S_i \leftarrow \text{"Maintain feature } i\text{"}$
- 14:     **end if**
- 15: **end for**
- 16: **return**  $S$

---

개인화된 피드백 예시:

- **특성 1 (예: studytime):** 현재 값 = 2, 목표 값 = 4로 증가 필요 방향: 주간 학습 시간을 늘릴수록 성적이 향상될 가능성이 높습니다. 구체적 조언: 매일 공부하는 시간을 1시간씩 추가하여 주간 학습 시간을 늘려보세요.

- **특성 2 (예: absences):** 현재 값 = 8, 목표 값 = 2로 감소 필요 방향: 결석 일수를 줄일 수록 성적이 향상됩니다. 구체적 조언: 규칙적인 출석을 위해 학습 환경을 정비하고, 출석에 대한 동기 부여를 강화하세요.
- **특성 3 (예: goout):** 현재 값 = 5, 목표 값 = 3으로 감소 필요 방향: 친구와의 외출 빈도를 줄이면 성적이 향상될 수 있습니다. 구체적 조언: 외출 시간을 줄이고 그 시간을 학습에 투자해보세요.

예상되는 성적 향상은  $f(x + \Delta x^*) - f(x)$  점입니다.

## 8.7 실험 결과: 실제 건국대학교 학생을 대상으로 피드백 생성

본 보고서에서는 SHAP 값과 최적 수송 이론을 결합하여 각 학생의 성적을 향상시키기 위한 개인화된 학습 전략을 제안하였다. 아래는 실제 건국대학교 학생 50명의 데이터를 바탕으로 생성된 개인화된 전략 예시로, 두 명의 학생에 대한 전략을 서술하였다.

### 8.7.1 학생 1의 개인화된 전략

- **특성 1: 추가 교육 지원 여부**
  - 현재 값 = 0.89, 목표 값 = 1.00
  - 권장 조치: 증가 (중요도 점수: 0.06)
  - 조언: 추가 교육 지원 여부를 증가시키면 성적 향상에 도움이 될 수 있습니다.
  - 구체적 전략: 추가 교육 지원 여부에 대한 개인별 전략을 세워보세요.
- **특성 2: 주간 학습 시간**
  - 현재 값 = 0.00, 목표 값 = 0.01
  - 권장 조치: 증가 (중요도 점수: 0.05)
  - 조언: 주간 학습 시간을 증가시키면 성적 향상에 도움이 될 수 있습니다.
  - 구체적 전략: 주간 학습 시간에 대한 개인별 전략을 세워보세요.
- **특성 3: 건강 상태**
  - 현재 값 = 0.00, 목표 값 = 0.59
  - 권장 조치: 증가 (중요도 점수: 0.04)
  - 조언: 건강 상태를 증가시키면 성적 향상에 도움이 될 수 있습니다.
  - 구체적 전략: 건강 상태에 대한 개인별 전략을 세워보세요.

현재 예상 성적: 0.39

목표 성적: 0.78

예상 성적 향상: 0.39

### 8.7.2 학생 50의 개인화된 전략

- **특성 1: 건강 상태**
  - 현재 값 = 0.00, 목표 값 = 0.01
  - 권장 조치: 증가 (중요도 점수: 0.05)
  - 조언: 건강 상태를 증가시키면 성적 향상에 도움이 될 수 있습니다.

- 구체적 전략: 건강 상태에 대한 개인별 전략을 세워보세요.

- **특성 2: 방과 후 여가 시간**

- 현재 값 = 0.00, 목표 값 = 0.63
- 권장 조치: **증가** (중요도 점수: 0.04)
- 조언: 방과 후 여가 시간을 증가시키면 성적 향상에 도움이 될 수 있습니다.
- 구체적 전략: 방과 후 여가 시간에 대한 개인별 전략을 세워보세요.

- **특성 3: 어머니의 교육 수준**

- 현재 값 = 0.75, 목표 값 = 0.84
- 권장 조치: **증가** (중요도 점수: 0.03)
- 조언: 어머니의 교육 수준을 증가시키면 성적 향상에 도움이 될 수 있습니다.
- 구체적 전략: 어머니의 교육 수준에 대한 개인별 전략을 세워보세요.

현재 예상 성적: 0.67

목표 성적: 0.78

예상 성적 향상: 0.12

## 8.8 한계점 및 향후 연구 방향

본 보고서에서 제안한 SHAP와 최적 수송 이론을 결합한 방법론은 강력한 성적 예측 및 개선 도구로 기능할 수 있다. 그러나 몇 가지 한계점이 존재한다.

- **계산 복잡성:** SHAP 값 계산 및 최적 수송 문제 해결 과정은 대규모 데이터셋에서 계산 비용이 매우 높을 수 있다. 이를 해결하기 위해 근사 알고리즘이나 병렬 처리 기술을 도입할 필요가 있다.
- **선형성 가정:** SHAP 값은 각 특성의 기여도를 선형적으로 가정하지만, 실제 특성 간의 상호작용은 비선형적일 수 있다. 이를 보완하기 위해 비선형 상호작용을 고려한 SHAP 변형 방법을 적용할 수 있다.
- **인과관계 반영 부족:** 본 수식은 상관관계를 기반으로 한 성적 예측을 수행하며, 인과 관계를 직접 반영하지는 않는다.

## 9 결론

본 보고서에서는 SHAP와 최적 수송 이론을 결합한 개인화된 학업 성취도 향상 전략을 제안하고, 이를 학생 성적 예측에 적용한 새로운 접근 방식을 제시하였다. SHAP를 활용하여 각 특성의 중요도를 정량적으로 평가함으로써 학업 성취에 기여하는 주요 요인을 도출하였고, 최적 수송 이론을 통해 학생 개개인의 성적을 개선할 수 있는 최적 경로를 제시하였다. 실제 건국대학교 학생들의 데이터를 활용한 실험 결과, 예측 모델은 대체로 신뢰할 만한 성능을 보였으며, 주요 학습 요인이 성적에 미치는 영향을 구체적으로 분석할 수 있었다.

그 결과, 학생 개개인의 특성에 맞춘 개인화된 학습 추천 시스템의 가능성을 제시하였으며, 이를 통해 학생들의 성적 향상을 돋는 실질적인 피드백을 제공할 수 있음을 확인하였다. 향후에는 성적 예측 모델을 다양한 대학 및 학과로 확장하고, 외부 환경 요인을 반영한 예측 모델을 개발하여 더욱 정교한 개인화된 학습 전략을 제공하는 것이 중요하다.

또한, 실시간 피드백 시스템을 구축하여 학생들의 학습 성과를 지속적으로 모니터링하고, 적절한 학습 전략을 제안하는 시스템을 개발하는 것이 바람직하다. 이를 통해 학생들은 학업 성취도를 높일 수 있으며, 교육 현장에서 실제로 활용 가능한 혁신적인 학습 전략 시스템으로 자리 잡을 수 있을 것이다.

향후 과제로는 계산 효율성을 개선하고, 비선형 상호작용 및 인과관계를 반영한 모델을 더욱 정교하게 발전시켜 다양한 교육 환경에서 실질적으로 활용 가능한 개인화된 학습 전략 시스템을 구축하는 방향으로 나아가야 한다.