REPORT

인천대학교 INCHEON NATIONAL UNIVERSITY

우울증 위험 예측 모델 개발 및 주요 요인 분석

과목명	프로그래밍언어
담당교수	정지혁교수님
학과	산업경영공학과
학년	3학년
학번	202202884
이름	이준형
제출일	2024-12-14

1. 서론

1.1 연구 배경

우울증은 현대 사회에서 가장 널리 퍼져 있는 정신 건강 문제 중 하나로, 특히 청소년과 대학생들에게 심각한 영향을 미칩니다. 세계보건기구(WHO)에 따르면, 우울증은 전 세계적으로 주요 사망 및 장애 원인 중 하나로 꼽히며, 특히 15~29세 인구 사이에서 자살의 주요 원인으로 작용하고 있습니다.

학생들은 학업, 사회적 압박, 미래에 대한 불안 등 다양한 스트레스를 경험하며, 이러한 스트레스는 우울증 발병의 위험 요인으로 작용할 수 있습니다. 특히, 학생일때 정신 건강 문제는 이후 삶 전반에 걸쳐 부정적인 영향을 미칠 가능성이 높아 조기 예측 및 예방이 중요합니다.

그래서 저는 학생을 대상으로 한 Depression Student Dataset을 분석하여, 우울증과 관련된 주요 요인을 파악하고 이를 통해 우울증 위험을 효과적으로 예측하는 모델을 개발하고자 합니다.

1.2 연구 목적

연구의 주요 목적:
1. 데이터 전처리를 통해 각각 분석기법에 알맞은 데이터셋 생성 후 대학생 우울증의 주요 영향을 미치는 요인을 데이터 분석을 통해 도출.
2. 1번에서 말한 분석기법(상관 분석 및 카이제곱검증)을 활용하여 주요 변수와 우울증 간의 관계를 평가.
4. 분석 결과를 바탕으로 우울증 위험을 예측하는 모델의 성능을 평가하고, 실제로 예측 실행,이를 개선하기 위한 전략 제안.

2. 데이터 개요

2.1 데이터 출처

-데이터셋:[Depression StudentDataset]
(https://www.kaggle.com/datasets/ikynahidwin/depression-student-dataset/data)
- 출처: Kaggle
- 데이터 설명: 데이터셋은 대학생의 우울증 여부와 관련된 다양한 사회적 및 개인적 요인을 포함하며, 총 502개의 레코드와 11개의 변수로 구성되어 있습니다.

2.2 데이터 항목

- Gender: 성별 (남성, 여성)
- Age: 나이
- Academic Pressure: 학업 스트레스 수준 (1~5: 낮음~높음)
- Study Satisfaction: 학업 만족도 (1~5: 낮음~높음)
- Sleep Duration: 수면 시간 (Less than 5 hours, 5-6 hours 등 범주형)
- Dietary Habits: 식습관 (Healthy, Moderate, Unhealthy)
- Have you ever had suicidal thoughts?: 자살 충동 경험 여부 (Yes/No)
- Study Hours: 하루 평균 공부 시간
- Financial Stress: 재정적 스트레스 수준 (1~5: 낮음~높음)
- Family History of Mental Illness: 가족 정신 질환 병력 여부 (Yes/No)
- Depression: 우울증 여부 (Yes/No)

2.3 데이터의 특성

- 총 데이터 개수: 502개
- 결측치 여부: 없음
- 데이터 타입: 수치형, 범주형, 서열형 혼합
- 데이터는 다양한 요인들 간의 상호작용을 탐구하기 적합하며, 특히 우울증 여부를 예측하는데 필요한 중요한 변수를 포함하고 있습니다.
- 데이터셋의 주요 특성은 우울증과 연관된 다차원적인 요인을 제공하여, 개인적 요인(성별,나이 등)과 환경적 요인(학업 및 재정적 스트레스)을 분석 가능하게 합니다.

3. 데이터 전처리

3.1 전처리 수행 작업

1. 범주형 변수의 레이블 인코딩
   - Depression 변수: Yes/No를 수치형 변수들과의 상관관계 분석을 하기 위하여 1/0으로 변환
   - Dietary Habits, Have you ever had suicidal thoughts? 등의 범주형 변수에 레이블 인코딩 적용하여 로지스틱회귀분석 사용할 수 있도록 전처리

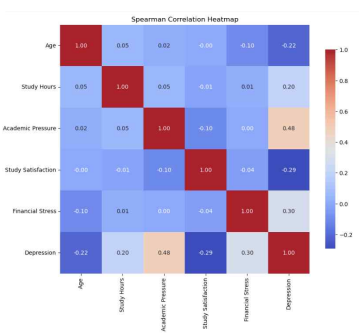
2. 서열형 변수의 Ordinal Encoding
   - Academic Pressure, Study Satisfaction, Financial Stress 등의 변수에 서열형 인코딩 적용.
   - 각 변수를 1~5의 숫자로 변환하여 분석에 활용 가능하도록 정리.

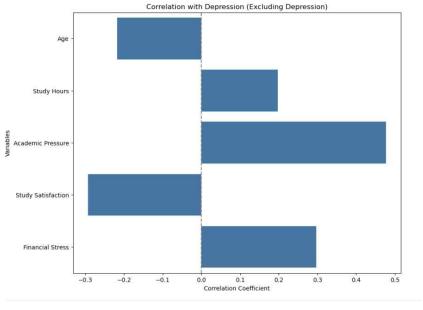
3. 결측치 및 이상치 확인
   - data.isnull().sum()을 통해 결측치 확인: 결측치 없음.

*data.copy()를 이용하여 data2와 data3를 생성
- 기존 data는 범주형 데이터끼리 카이제곱분포를 하기 위해 사용
- data2는 Depression을 인코딩하여 수치형 변수들과 상관관계 분석하기 위해 사용
- data3는 카이제곱과 상관관계 분석을 하였을 때 유의미한다고 판단되는 변수들만 범주형이면 인코딩 수치형은 그대로 유지하여 로지스틱회귀분석을 하기위하여 사용
-> data3 같은 경우 실제로 분석하였을 때 범주형변수가 유의미한 변수로 나와 범주형만 인코딩하였음

3.2 데이터 시각화

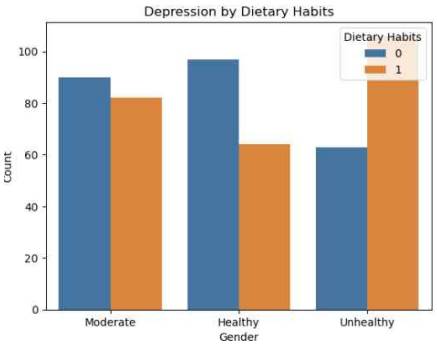
1. 상관관계 히트맵





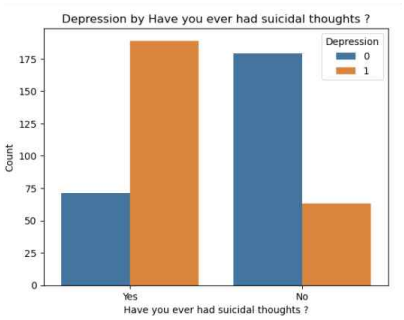
- Spearman 상관계수를 이용하여 변수 간의 상관성을 시각화.
- Depression과 가장 높은 상관성을 보이는 변수는 Academic Pressure(0.48) 및 Financial Stress(0.30).

2. 우울증과 식습관의 관계



- 우울증 여부에 따른 식습관 분포를 막대 그래프로 시각화.
- 건강하지 않은 식습관(Unhealthy)은 우울증과 높은 연관성을 보임.

3. 자살 충동 경험과 우울증



- 자살 충동 경험 여부에 따른 우울증 발생률을 시각화.
- 자살 충동 경험이 있는 학생에서 우울증 발생률이 유의미하게 높음을 확인.

4. 데이터 분석 및 성능 평가

4.1 상관 분석 결과
- Depression과 관련된 주요 변수:
- 학업 스트레스 (0.48): 가장 높은 양의 상관성을 보임.
- 재정적 스트레스 (0.30): 비교적 높은 상관성을 보임.

4.2 카이제곱 검증
- Depression과 관련된 주요 변수:
- Dietary Habits와 Depression: p-value = 0.0001 (유의미한 관계 있음)
- Have you ever had suicidal thoughts?와 Depression: p-value < 0.001 (유의미한 관계 있음)

4.3 로지스틱 회귀 분석

1. 모델 구축 및 학습
   - 독립 변수: Dietary Habits, Have you ever had suicidal thoughts?, Academic Pressure, Financial Stress (앞써 카이제곱,상관관계를 하였을 때 주요변수로 추출해낸 주요변수)
   - 종속 변수: Depression.
   - 학습 데이터와 테스트 데이터로 80:20 비율로 분리.

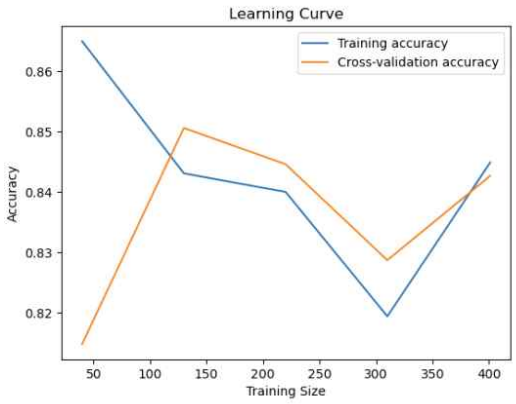
2. 모델 성능
  - 정확도 (Accuracy): 90.1%
  - 주요 변수의 유의성 (p-value):
  - 학업 스트레스: coef = 1.2947, p < 0.001
  - 자살 충동 경험: coef = 3.1597, p < 0.001
  - 식습관: coef = 0.7303, p < 0.001

3. Confusion Matrix
  - True Positive (TP): 49, True Negative (TN): 42
  - False Positive (FP): 6, False Negative (FN): 4

4. Precision, Recall, F1-Score
  - Precision: 91%
  - Recall: 90%
  - F1-Score: 90%

4.4 실제 예측 수행
- 새 데이터를 입력하여 우울증 발생 확률과 여부를 예측:
- 입력 데이터: Dietary Habits = Moderate, Have you ever had suicidal thoughts? = Yes, Academic Pressure = High, Financial Stress = High
-> 예측 결과: 우울증 발생 확률 = 88.6%, 예측 클래스 = 1 (우울증 발생)
- 입력 데이터: Dietary Habits = Healthy, Have you ever had suicidal thoughts? = No, Academic Pressure = Low, Financial Stress = Low
-> 예측 결과: 우울증 발생 확률 = 1.9%, 예측 클래스 = 0 (우울증 미발생)

4.5 모델 평가



- 우울증 여부를 예측하는 데 높은 성능을 보였으며, 주요 변수들이 통계적으로 유의미함을 확인.
- 학습 데이터에서 83.5%, 테스트 데이터에서 90.1%의 정확도를 기록하여 모델의 일반화 성 능이 양호함.
- 추가적으로 학습 곡선을 분석하여 모델이 과적합되지 않았음을 확인.

5. 결론

5.1 연구 결과
- 학업 스트레스, 재정적 스트레스, 식습관, 자살 충동 경험 등이 우울증에 큰 영향을 미치는 주요 변수로 나타남.
- 특히, 자살 충동 경험 변수는 가장 높은 회귀 계수(3.1597)를 보여 우울증 발생과 강한 연관성을 가짐.

5.2 데이터 분석의 시사점

학업 및 재정적 스트레스 감소의 중요성: 학업 스트레스와 재정적 스트레스는 우울증 발생에 주요 요인으로 작용하므로, 이를 완화할 수 있는 제도적 지원 및 정책적 개입이 필요합니다. 예를 들어, 대학 내 심리 상담 서비스 강화, 재정적 지원 프로그램 확대 등이 이에 해당할 수 있습니다.

건강한 식습관과 자살 예방 프로그램의 필요성: 건강하지 않은 식습관은 우울증 발생 가능성을 높이는 요인 중 하나로, 대학생들에게 균형 잡힌 식습관을 장려하는 캠페인 및 프로그램이 유익할 것입니다. 또한, 자살 충동 경험과 우울증의 높은 연관성을 고려할 때, 자살 예방을 위한 심리적 지원 프로그램과 상담 서비스가 필수적입니다.

5.3 한계 및 향후 연구 방향

데이터 일반화의 한계: 본 연구는 특정 데이터셋에 기반하였으며, 이를 다른 지역이나 집단에 일반화하는 데 한계가 있습니다. 향후 연구에서는 보다 다양한 배경을 가진 데이터를 수집하여 분석의 범위를 확장할 필요가 있습니다.

추가 변수의 탐색: 본 연구에서 다루지 않은 사회적 네트워크, 신체 활동 빈도 등의 변수도 우울증 발생에 영향을 미칠 수 있으므로, 이를 포함한 추가적인 데이터 분석이 요구됩니다.