

목차

Topic

- 주제 선정 배경
 - 데이터 선정
 - 주제 선정

Analysis

- Modeling Plan
 - 모델 생성
- 모형 성능 비교





O1 Topic

- 주제 선정 배경
 - 데이터 선정
 - 주제 선정

주제 선정 배경

"한국 청소년, 진로 고민 1위..경험은 꼴찌"



E IV-10 한국청소년의 현재의 고민·걱정거리

	-	-	-
-	위:	D/ /	
-	Z.,	70.1	100
	-	701	-

내용	전혀 걱정하지 않는다	약간 걱정한다	다소 심각하게 걱정한다	매우 심각하게 걱정한다	전체
공부·학교 성적 문제	11,9	44,8	31,5	11,9	100(1502)
진로 ·진학 문제	8,2	32,9	35.0	24.0	100(1502)
가정의 경제적 형편	16,6	40,9	27.6	15,0	100(1502)
가족과의 불화	49,9	30,6	14,3	5,3	100(1502)
사귀는 사람(남자친구, 여자친구) 문제	42.7	36,3	14,5	6,5	100(1502)
친구와의 관계	34,3	44,5	16,8	4,4	100(1502)
나의 성격문제	23,9	48,2	21,4	6,5	100(1502)
나의 외모·신체 관련 문제	17.3	48,3	23,6	10,7	100(1502)



청소년들의 가장 큰 고민은 '진로' 진로를 탐색하고 스스로 결정을 내릴 수 있는 능력이 부족하기 때문 -> 진로결정능력에 영향을 주는 요인 분석

데이터 선정

통계청 > 사회 > 청소년종합실태조사(2017)

- ◢ 🚞 사회
 - ▷ 📶 지자체사회조사
 - ▷ 국민다문화수용성조사
 - ▷ 국민생활안전실태조사
 - ▷ 국제결혼중개실태조사
 - ▷ 남해군노인실태조사
 - ▶ 다문화청소년패널조사
 - ▷ 양성평등실태조사
 - ▶ 📹 여성가족패널조사
 - ▷ 용인시경제지표조사
 - ▷ 전국다문화가족실태조사
 - ▷ 청소년매체이용및유해환경실
 - ◢ 🚞 청소년종합실태조사
 - **4** [=] 2017
 - □ 🗐 양육자(제공)
 - ☑ 🗊 청소년(제공)
 - L 항목조회



만 13-19세의 현재 재학중인 중·고등학생을 대상으로 선정

- 변수 '연령(만)'의 값이 13-19인 관측치만 추출
- 최종 학력이 고등학교 졸업 이상인 관측치 삭제
- 학업 중단 경험 '있다'고 답한 관측치 삭제
- 대학 관련 변수, 학업 중단 관련 변수 삭제
- 취업 후를 전제로 한 변수 삭제

주제 선정



반응변수 : 청소년기는 진로를 탐색할 시기이기 때문에 <u>'장래 희망 직업의 유무'와</u> <u>'미래에 스스로 진로를 결정할 수 있는가'를 함께 반응변수로 선정해야 한다고 판단</u>

진로 및 직업선택

아래의 각 문항을 잘 읽고 자신의 생각과 일치하는 카의 번호에 표시해 주십시오.

구 분	전혀 그렇지 않다	그렇지 않은 편이다	그런 편이다	매우 그렇다
1) 나는 분명한 인생목표가 있다	0	2	3	4
2) 진로나 직업을 선택할 시기가 되면 내가 정한 기준에 의하여 선택할 것이다	0	2	3	4
3) 내가 희망하는 직업을 미래에 가질 수 있다고 생각한다	0	2	3	4



이 문항들이 청소년기의 진로결정능력을 잘 반영한다고 판단 세 문항의 평균을 '진로결정능력'으로 명명하고 반응변수로 선정

반응변수 1 : 진로결정능력

반응변수 2 : 장래 희망 직업 여부

주제 선정



주제 : 청소년의 진로 고민을 효과적으로 도울 방법을 제안한다.



최종 목표 : 1. '진로결정능력'에 영향을 주는 주요 요인 분석

2. '장래 희망 직업 여부'에 영향을 주는 주요 요인 분석



변수변환과 생성 주관적 응답을 카테고리 별로 모음



새로운 변수	변수 설명	변환
q1	주관적 안녕감	I.건강 1.주관적 안녕감에 대한 응답의 평균, 0~1 척도
feeling	정서	I.건강 7.정서에 대한 응답의 평균, 0~1 척도
q8a	참여도	청소년 정책 참여도에 대한 응답의 평균, 0~1 척도
q8b	필요도	청소년 정책 필요도에 대한 응답의 평균, 0~1 척도
experience	청소년 활동 경험 횟수	문화, 과학, 봉사 등 청소년 활동 경험 횟수의 총합
facility	시설 경험 횟수	청소년수련관, 도서관, 전시관 등 시설 경험 횟수의 총합
program_know	청소년 프로그램 인지도	청소년 프로그램 인지도에 대한 응답의 평균, 0~1 척도
q13	양육자와의 활동	양육자와의 활동 횟수에 대한 응답의 평균, 0~1 척도
q14	양육자와의 관계	양육자와 함께 보내는 시간에 대한 응답의 평균, 0~1 척도
q15	양육자의 관심	양육자의 관심 정도에 대한 응답의 평균, 0~1 척도
parents_help	부모님의 부담 비용에 대한 의견	부모님이 언제까지 비용을 부담해야 하는지에 대한 응답의 평균, 0~1 척도
marry	결혼관	결혼관에 대한 응답의 평균, 0~1 척도, 1에 가까울수록 개방적
agree	동의 정도	동성애, 이혼, 미혼모, 입양에 대한 응답의 평균, 0~1 척도, 1에 가까울수록 개방적
society1	사회관	우리 사회에 대한 응답의 평균, 0~1 척도
multiculture	다문화관	다른 문화에 대한 개방정도, 0~1 척도
reunify	통일관	통일의 필요성에 대한 응답의 평균, 0~1 척도
school_satis	학교생활 만족도	학교생활 만족도에 대한 응답의 평균, 0~1 척도

변수변환과 생성



오른쪽 그림과 같이 정도를 나타내는 응답은 0~1의 범위를 갖도록 척도 변환하였다. 0으로 갈수록 부정적, 1로 갈수록 긍정적임을 의미한다.

- 해당 변수: meal1~3, stress1~2, q6a, q6b, feeling, q8a, q8b, program_know, q13, q14, q15, family_satis, parents_help, marry, agree, society1~3, vote_age, multiculture, reunify, school_satis, school_score, alone_time, q31~q32, edu_want, y

오른쪽 그림과 같이 경험의 유무를 나타내는 응답은 없을 경우 0, 있을 경우 1로 표기하였다. 횟수를 묻는 경우, 유무에 대한 변수는 지우고, '없다'의 응답을 0으로 표기하였다.

- 해당 변수: experience, club_offline, club_online, facility, y4, parttime, job_training

전혀 그렇지		그런	매우
그렇지 않다	않은 편이다	편이다	그렇다
•	0 2		4

없다	있다	횟수
0	0	ಶ

변수변환과 생성



- sleep_time : 잠에 드는 시각을 연속적으로 나타내기 위해 0시~4시를 24시~ 28시로 변환
- city_size : 값이 커질수록 대도시임을 나타내기 위해 척도를 반전시킴

> tab	le(yo	uth\$\	7)					
0 938	1 262	2 91	3 26	4 2	20 13	21 36	22 458 1	23 158
				1				
> tab	le(yo	uth\$	sleep_	time)				
20 13	21 36	22 458	23 1158	24 938	25 262	26 91	27 26	28 2



- experience : 청소년 활동 경험 횟수 중 유일하게 200건을 넘어서는 관측치인 576건을 이상값으로 처리
- career_plan : 청소년이 선택할 수 없는 답안인 '이직'응답을 결측치로 처리

변수변환과 생성



불필요한 변수 제거

제거한 변수	변수 설명	제거 이유
V8~10, V12~14	취침/기상시각의 분, 주말	분석의 용이함을 위해 취침시각과 기상시각은 주중데이터의 '시'만 사용함
V55, V56	청소년 활동 경험 참여 이유, 만족도	청소년 활동에 참여하지 않은 경우에 의한 결측치가 많음
V57, V58	참여 희망 활동 (1,2순위)	너무 많은 범주
V127	자율 학습 시간(분)	분석의 용이함을 위해 자율 학습 시간(시)만 이용
V129~132	평일 방과후 주요 활동(1,2순위)	너무 많은 범주
V146~151	학업 중단 관련 질문	학업 중단의 경우 분석에서 배제
V154	장래 희망 직업	너무 많은 범주
V155~156	직업 선택 기준(1,2순위)	너무 많은 범주
V162~164	직업 훈련 관련 질문	직업 훈련 경험이 없는 경우에 의한 결측치가 많음
V166	조사 참여 양육자와의 관계	양육자에 대한 다른 정보가 없으므로 제거
V167~168	교육정도	이미 관측치를 중학생과 고등학생으로 선정하여 제거
V169~170	연령별, 성별	다른 변수와 중복

결측값 처리



- KNN 방법으로 결측값이 포함된 관측치와 가장 가까운 10개의 관측치 사용.
- 'Cluster' 패키지의 daisy함수를 이용하여 변수 type에 관계없이 적용 가능한 표준화된 거리 측도 사용
- 수치형 변수의 경우 중앙값, 범주형 변수의 경우 최빈값으로 대체
- 처리한 변수 : q8a(청소년 정책 참여도), q8b(청소년 정책 필요도), experience(청소년 활동 경험 횟수), club_online(온라인 상 가입 단체 개수), career_plan(향후 진로 계획)

최종 분석 데이터 생성



최종 분석 데이터

관측치	변수
2778개	607H

> dim(teen) [1] 2778 60



범주형 변수 범주 설명

변수명	범주 설명 변수명		범주 설명
	1-TV, 비디오 시청 2-집에서 인터넷(게임 포함) 하기 3-라디오나 음악 듣기 4-잡지나 책(만화 포함) 읽기	희망 교육 수준	1-고졸 이하 2-전문대졸(2~3년제) 3-대졸(4년제) 4-대학원 이상(석/박사) 5-모름/무응답
방과후 주 활동 (저녁식사 전 /후)	방과후 5-집에서 가족/친구와 대화 주 활동 6-집안일(청소 등) 돕기 7-그냥 쉬거나 잠자기 8-집에서 숙제 등 공부	향후 진로 계획	1-상급 학교 진학 2-취업 3-창업 4-가업 7- 아직 정하지 못하였다
		주 양육자 학력	1- 초졸 이하 2-중졸 3-고졸 4-대졸 이상



Train data / test data 생성

- 모델 적합을 위한 Train data(70%)와 모델 평가를 위한 test data(30%)로 자료 분리

최종 분석 데이터

변수명	변수 설명	변수 유형	변수명	변수 설명	변수 유형
age	연령(만)	numeric	vote_age	투표 연령 동의 정도	numeric(0~1척도)
q1	주관적 안녕감	numeric(0~1척도)	multiculture	다문화관	numeric(0~1척도)
sleep_time	취침 시각(시)	numeric	reunify	통일관	numeric(0~1척도)
wakeup_time	기상 시각(시)	Numeric	school_satis	학교생활 만족도	Numeric(0~1척도)
meal1	아침 식습관	numeric(0~1척도)	school_score	학업 성취도	Numeric(0~1척도)
meal2	점심 식습관	Numeric(0~1척도)	academy_time	사교육 시간(일주일 총 시간)	Numeric
meal3	저녁 식습관	Numeric(0~1척도)	self_study	하루 혼자 공부한 시간(시간)	Numeric
q4	지난 1주간 신체활동 시간(시간)	Numeric	alone_time	평일 방과후 혼자 있는 시간	Numeric(0~1척도)
stress1	스트레스_일상 중	Numeric(0~1척도)	afterschool1	방과후 주 활동_저녁식사 이전	categorical(1~13)
stress2	스트레스_연속 2주 이상	Numeric(0~1척도)	afterschool2	방과후 주 활동_저녁식사 이후	categorical(1~13)
q6a	건강_신체	Numeric(0~1척도)	q31_1A	진로교육경험_'진로와 직업'과목	categorical(유:1/무:0)
q6b	건강_정신	Numeric(0~1척도)	q31_2A	진로교육경험 _현장학습 또는 견학	categorical(유:1/무:0)
feeling	정서	Numeric(0~1척도)	q31_3A	진로교육경험_상담	categorical(유:1/무:0)
q8a	청소년 정책 참여도	Numeric(0~1척도)	q31_4A	진로교육경험_진로 검사	categorical(유:1/무:0)
q8b	청소년 정책 필요도	numeric(0~1척도)	q31_5A	진로교육경험_인터넷 또는 동영상	categorical(유:1/무:0)
experience	청소년 활동 경험 횟수	numeric	q31_6A	진로교육경험_동아리 활동	categorical(유:1/무:0)
club_offline	오프라인 상 가입 단체(동아리) 개수	numeric	q31_7A	진로교육경험_초청 강연	categorical(유:1/무:0)
club_online	온라인 상 가입 단체 개수	numeric	q31_8A	진로교육경험_재량 활동	categorical(유:1/무:0)
facility	시설 경험 횟수	numeric	q32_1A	진로체험경험_진로 박람회	categorical(유:1/무:0)
program_know	청소년 프로그램 인지도	numeric(0~1척도)	q32_2A	진로체험경험_청소년 수련관, 사회복지관 프로그램	categorical(유:1/무:0)
q13	양육자와의 활동	numeric(0~1척도)	q32_3A	진로체험경험_산업체 프로그램	categorical(유:1/무:0)
q14	양육자와의 관계	numeric(0~1척도)	q32_4A	진로체험경험_대학 프로그램	categorical(유:1/무:0)
q15	양육자의 관심	numeric(0~1척도)	edu_want	희망 교육 수준	categorical(1~4)
family_satis	가정생활 만족도	numeric(0~1척도)	career_plan	향후 진로 계획	categorical(1~8)
parents_help	부모의 자녀에 대한 비용부담 범위	numeric	parttime	아르바이트 경험	categorical(유:1/무:0)
marry	결혼관	numeric(0~1척도)	job_training	직업 훈련 경험	categorical(유:1/무:0)
agree	동의 정도	numeric(0~1척도)	sex	성별	categorical(남:1/여:2)
society1	사회관	numeric(0~1척도)	parents_edu	주 양육자 학력	categorical
society2	사회 신뢰도	numeric(0~1척도)	y1	진로결정능력	numeric(0~1척도)
society3	사회 문제 관심도	numeric(0~1척도)	y2	장래 희망 직업 여부	범주형(유:1/무:0)



Analysis

- Modeling Plan
- 모델 생성
- 모형 성능 비교

03

Modeling Plan



Modeling 목표

- 1. 수치형 반응변수 '청소년의 진로 결정 능력' 향상 회귀분석
- 2. 범주형 (binary) 반응변수 '장래희망 결정 여부'를 결정하는 분류 규칙 찾기 분류분석



후보 모델

다음과 같은 후보 모델을 train data를 이용하여 fitting

Regression	Classifacation
Linear Regression	LDA
Regression Tree	Logistic Regression
Random Forest	

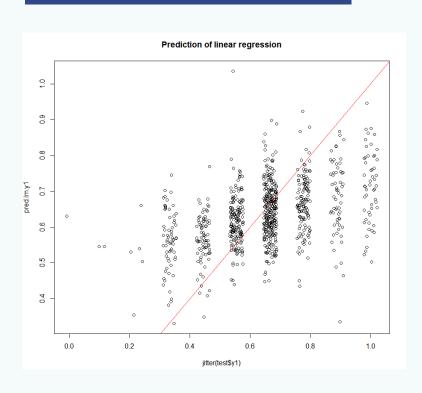


모델 평가

각 후보 모델에 대해 test data에서의 성능을 비교하고 모델 분석 및 평가 시행

- 모델 평가 기준: Regression - test data에서의 MSE / Classification - test data에서의 오분류율

1. Linear Regression Model



- '진로 결정 능력'을 반응변수로 하는 선형 회귀 적합
- Stepwise regression 방법론을 이용하여 모형에 유용한 변수 선택
- MSE: test error = 0.0246



1. Linear Regression Model

- 추정된 회귀 계수의 의미 해석
- 유의수준 5%에서 변수 선택

(-) 효과

변수명	변수 설명	계수 크기
sleep_time	취침 시각(시)	0.01
wakeup_time	기상 시각(시)	0.01
stress2	스트레스_연속 2주 이상	0.04
parents_help	부모의 자녀에 대한 비용부담 범위	0.04
q31_4A1	진로교육경험_진로 검사_유	0.03
q32_1A1	진로체험경험_진로 박람회_유	0.02
career_plan3	향후 진로 계획_창업	0.18
career_plan7	향후 진로 계획_미결정	0.05

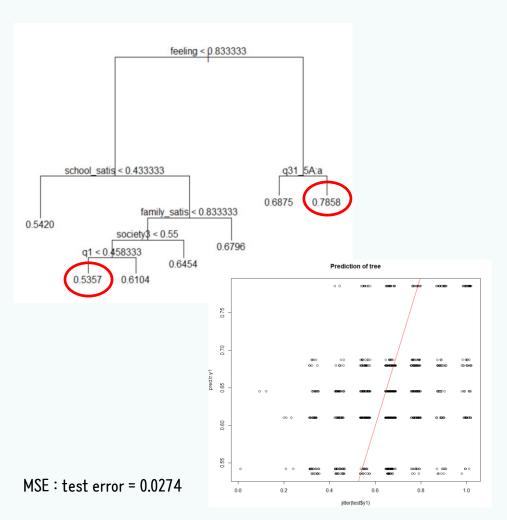
(+) 효과

변수명	변수 설명	계수 크기
age	연령	0.01
feeling	정서	0.22
q8b	청소년 정책 필요도	0.03
program_know	청소년 프로그램 인지도	0.05
q13	양육자와의 활동	0.04
family_satis	가정생활 만족도	0.07
society3	사회 문제 관심도	0.06
multiculture	다문화관	0.07
reunify	통일관	0.06
school_satis	학교생활 만족도	0.13
school_score	학업 성취도	0.06
self_study	하루 혼자 공부한 시간(시간)	0.01
q31_2A1	진로교육경험 _현장학습 또는 견학_유	0.02
q31_5A1	진로교육경험_인터넷 또는 동영상_유	0.03
edu_want	희망 교육 수준	0.05
job_training1	직업 훈련 경험_유	0.07



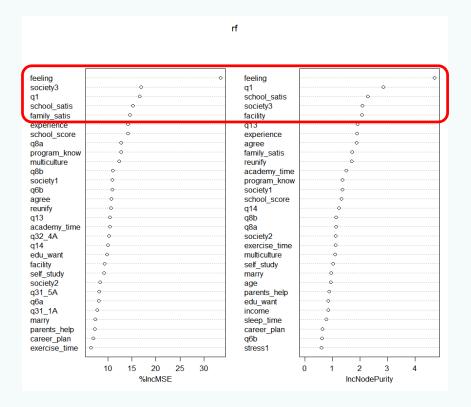
2. Regression Tree Model

- '진로 결정 능력'을 반응변수로 하는 트리 모형 적합
- tuning parameter들은 기본값 사용
- 중요 변수: splitting variables
- ⇒ feeling(정서), school_satis(학교생활 만족도), q31_5A(진로교육경험_인터넷 또는 동영상), family_satis(가정생활 만족도), society3(사회 문제 관심도), q1(주관적 안녕감)
- feeling(정서)이 0.83이상, q31_5A(진로교육경험_인터넷 또는 동영상)가 1(경험 유)일 때 진로 결정 능력이 가장 높음
- Feeling(정서)이 0.83 이하, school_satis(학교생활 만족도)가 0.43 이상, family_satis(가정생활 만족도)가 0.83 이하, society3(사회문제 관심도)이 0.55 이하, q1(주관적 안녕감)이 0.46 이하일 때 진로결정능력이 가장 낮음





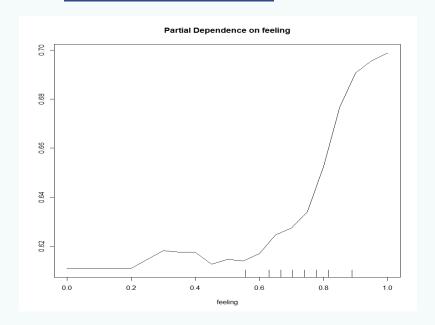
3. Random Forest



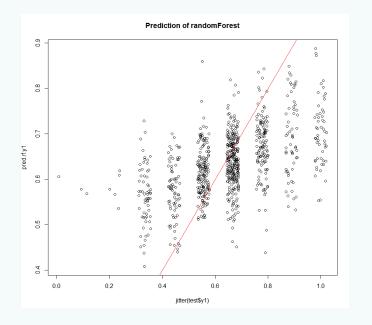
- '진로 결정 능력'을 반응변수로 하는 Random Forest 모형 적합
- tuning parameter들은 기본값 사용
- 가장 중요한 변수 **feeling(정서)**
- society3(사회문제에 대한 관심도), q1(주관적 안녕감), school_satis(학교생활 만족도) 변수도 꽤 높게 나타났으나 정서의 영향력이 월등해 보임



3. Random Forest



- feeling(정서)에 대한 partial plot
- 본인의 정서에 대해 긍정적인 답변을 할 때 진로 결정 능력도 높아짐



- MSE: test error = 0.0223

모델 생성: Classification



1. LDA

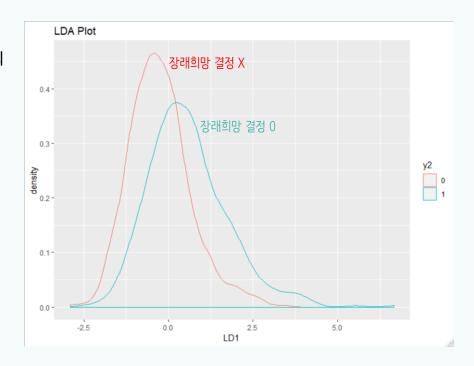
- '장래희망 결정 여부'에 대한 LDA 모형 적합
- experience(청소년활동경험횟수), facility(시설경험횟수) 변수의 평균의 그룹간 차이가 크다

	Group mea	
	청소년활동경험횟수	시설경험횟수
장래희망 결정 X	5.92	12.06
장래희망 결정 0	10.16	17.34

- Confusion matrix

	predicted 0	predicted 1
test data 0	539 (0.91)	51 (0.09)
test data 1	195 (0.87)	28 (0.13)

- 오분류율: 0.295



모델 생성: Classification



2. Logistic regression

- 범주형 변수 '장래희망 결정 여부'를 반응변수로 하는 logistic 모형 적합
- Stepwise 과정을 통해 AIC를 최소화하는 모형 선택

q31_2A1

q32_3A1

career_plan7

(+) 요끠	(+	-)	효	고	
--------	----	----	---	---	--

변수명변수 설명odds ratiosociety2사회 신뢰도 (0~1 척도)0.39multiculture다문화관 (0~1 척도)0.52

진로교육경험_현장학습이나 견학_유

진로체험경험_산업체 프로그램_유

향후 진로 계획_아직 없다

(-) 효과

0.66

0.76

0.21

변수명	변수 설명	odds ratio
club_offline	오프라인상 가입 단체(동아리) 수	1.34
q13	양육자와의 활동 (0~1 척도)	2.57
society3	사회 문제 관심도 (0~1 척도)	1.81
reunify	통일관 (0~1 척도)	3.10
school_satis	학교생활 만족도 (0~1 척도)	6.19
self_study	혼자 공부하는 시간(시간)	1.12
q31_1A1	진로교육경험_'진로와 직업' 과목_유	1.54
q31_8A1	진로교육경험_재량 활동_유	1.37
q32_4A1	진로체험경험_대학 프로그램_유	1.45
career_plan2	향후 진로 계획_취업	1.86
parttime1	아르바이트 경험_유	1.46
sex2	성별_여성	1.34

→ 유의수준 5%에서 유의한 변수들의 계수 해석 odds ratio란 x 1 증가시 odds의 변화 비율

모델 생성: Classification



2. Logistic regression

Confusion matrix

	predicted 0	predicted 1
test data 0	533 (0.90)	57 (0.10)
test data 1	176 (0.72)	67 (0.28)

- 오분류율: 0.280

모델 성능 비교



반응변수 : 진로결정능력



반응변수 : 장래희망 결정 여부

- MSE

	Test error
Linear regression	0.0246
Tree	0.0274
Random Forest	0.0223

- 오분류율

	Test data
LDA	0.295
Logistic regression	0.280

-> 대부분 비슷한 test error를 보임

Test data에 대한 mse가 가장 작은 Random forest와 Logistic regression 모델을 최적모형으로 선택! 나머지 모델도 의미 있으므로 이어서 설명



- 결론 및 해결방안

Conclusion: 결론 및 해결방안

반응변수1(진로결정능력)의 결론

- regression tree와 random forest모형 적합 결과 정서와 주관적 안녕감, 사회문제에 대한 관심도, 학교생활 만족도 등이 중요한 변수로 선택
- linear regression 결과, 긍정적인 정서 함양이 중요
- 또한, 스트레스 값이 높을 수록 낮은 진로결정능력을 보임
- 이러한 분석 결과들을 종합해보면 청소년의 진로결정 능력은 단순히 진로교육 경험의 유무보다는 전반적인 심리 상태와 삶에 대한 만족도에 큰 영향을 받는다는 것을 알 수 있음
- 공통적으로 인터넷을 통한 진로교육이 효과적



반응변수2(장래희망 직업 유무)의 결론

- LDA모형과 logistic regression 분석 결과, 대체적으로 청소년 활동 경험, 시설 경험, 오프라인 가입 단체(동아리) 수 등 체험 경험이 많을수록 장래희망을 결정한 경우가 많았음
- 그러나 logistic regression 모형 결과, 현장학습이나 견학, 산업체 프로그램과 같은 직업에 대한 직접적인 체험은 장래희망을 결정하는 데 중요한 경험임에도 불구하고 장래희망 결정으로 이어지지 않았음
- ⇒ 장래희망 결정에 방해 요소라고 볼 수는 없음 오히려 경험을 통해 진로를 탐색하고 수정하는 과정 중에 있기 때문이라고 봐야 함
- ⇒ 장래희망 직업을 갖는 것에 목표를 두기보다 스스로 진로를 탐색하고 결정할 수 있는 진로결정능력을 키우는 것이 더 중요함

Conclusion: 결론 및 해결방안

해결방안

- 현재 학교 내 진로 교육은 경험과 정보 제공을 통해 장래희망을 결정하는 것에만 초점이 맞춰져 있음 진로를 탐색하고 결정할 수 있는 힘(진로결정능력)을 길러줘야 함
- 획일적이고 진부한 진로교육은 축소하고 현장체험학습이나 인터넷 강의 등 효과적인 진로교육을 추천
- 청소년의 자아에 대한 만족도와 사회에 대한 호기심, 적극성을 길러주는 프로그램 활성화
- 입시에 매몰된 학습 패턴을 벗어나 사회 다방면에 관심을 가질 수 있도록 교육
- 청소년의 학부모, 양육자에게 가정 내 분위기와 양육자와의 활동시간의 중요성을 피력하고 올바른 가정생활을 이끌어줄 수 있도록 교육

한계점

- 한국 사회에서 진로결정능력에 대한 관심은 크지 않기 때문에 그것을 측정할 객관적인 변수가 없었다.
- 반응변수의 값이 이산적이라 풍부한 해석을 하는 데에 어려움이 있었다. 특히, 장래희망 결정 여부의 경우, 특정한 장래희망이 없더라도 관심사나 방향성을 가지고 있다 하더라도 표현되지 않았다는 한계가 있다.

