**Data Mining Report**

**경희대학교 산업경영공학과**

의사결정나무 분석을 이용한

청소년 흡연 예측 대한 모델 제안



과목명

담당교수

팀원

동영상URL

데이터마이닝

진창호

산업경영공학과 2016100937 김성수  
산업경영공학과 2016100946 김효준  
산업경영공학과 2016100966 윤태웅  
산업경영공학과 2016100989 정일도  
산업경영공학과 2016101000 홍성희

* Business Understanding

[ ✓ Business Problem → Data Science Problem ]

BP: 흡연 가능성이 높은 청소년에 대한 효율적 흡연 예방 교육의 필요성   
DSP: Classification Analysis 기법 중 Decision Tree를 이용해 청소년 흡연자의 Pattern을 파악하여, 향후 흡연을 할 것으로 예측되는 청소년을 선별해 내어 선별적인 흡연 예방 교육을 통해 보다 효율적인 청소년 흡연 예방 효과를 얻는다.

흡연은 단일 요소로서 국민건강에 가장 많은 피해를 주는 위험요인이며, 또한 예방할 수 있는 가장 큰 사망원인으로, 전체 사망원인의 16%를 기여한다고 하며 (맹광호,1988) 폐암, 심근경색증, 만성 폐쇄성 폐 질병, 위궤양 및 기타 여러 악성 종양의 직접적인 원인이 된다고 알려져 있다(심홍석외,2000). [[1]](#endnote-1) 직접흡연 뿐만 아니라 간접흡연에 의한 질병 발생 확률도 무시할 수 없는 수준이라는 것도 극명하다.[[2]](#endnote-2) 이처럼 흡연이 건강에 미치는 부정적인 영향은 매우 크다. 이에 따라서 우리는 금연 클리닉 서비스를 제공하는 보건복지부의 국가금연지원센터를 대상으로 삼아 향후 흡연을 할 것으로 예측되는 청소년을 선별해 내는 모델을 제공하고자 한다.

청소년, 그 중에서 특히 15세에 흡연을 시작한 사람의 경우, 25세 이후 흡연을 시작한 사람에 비해 60세 폐암 발병률이 약 3배에 달한다고 보고된 바 있다.[[3]](#endnote-3) 그리고 18세에 흡연을 시작한 청소년을 대상으로 코호트 분석을 한 결과, 65세까지는 전체 누적 사망자의 21.2%, 75세까지는 27.0%, 85세까지는 28.2%에 이르는 등 흡연에 기인한 질병으로 사망하는 질병부담이 상당한 비중을 차지하고 있다.[[4]](#endnote-4) 다시 말해 담배로 인한 암 발생의 위험은 얼마나 오래, 많이 피웠는가 보다는 얼마나 일찍 피우기 시작했는가가 중요하다는 것이다.  
 또한 19세 이하인 경우는 20세 이상인 경우에 비하여 금연이 1.49배 더 어렵다는 연구 결과를 통해, 흡연습관이 형성되기 이전에 청소년의 흡연을 예방하기 위한 노력은 매우 중요하다는 것을 알 수 있다.[[5]](#endnote-5) 이미 흡연을 시작한 청소년에 대해 금연을 돕는 것도 중요하지만, 언급된 연구 결과에 따르면 처음부터 아예 흡연을 시작하지 않게 하는 것이 우선이다.  
 이에 따라 우리는 국민건강영양조사의 데이터를 중심으로 청소년들의 흡연에 유의미한 영향을 끼치는 요인들을 찾아 의사결정나무 모델을 수립해 흡연 가능성이 높은 청소년을 찾아 선제적인 흡연 예방 교육을 제공해 청소년기의 흡연 시작을 사전에 예방하고자 한다.

 [ ✓ Why is it important? ]

흡연과 이로 인해 야기되는 질병들은 과거부터 현재까지 심각한 사회적 문제로 이어지고 있다. 청소년기에 흡연을 시작한 흡연자들의 암 발병률은 그렇지 않은 흡연자들에 비해 현저히 높고, 이는 자연스럽게 흡연에 의한 암으로 인한 사망률을 높이는 결과를 가져온다. 따라서 흡연 예방 교육을 통해서 청소년기의 흡연율을 낮출 수 있다면 미래의 암에 의한 사망률 또한 낮출 수 있을 것이고, 암 등 흡연에 기인한 여러 질병으로 인한 사회 문제를 완화하는데 큰 기여를 할 수 있을 것이다.  
 뿐만 아니라, 유사한 연구[[6]](#endnote-6)가 2014년 발행되었지만 이는 담뱃값이 인상되기 전에 진행된 연구이다. 2015년 01월 01일 담뱃값 인상 이후 청소년 흡연 예측 모델 관련 연구는 발행된 바 없다. 그러므로 본 연구를 통해 담뱃값 인상 이후의 데이터를 기반으로 현재 상황을 잘 대변하는 의사결정나무 모델을 수립하고자 한다.

본 연구를 통해 우리는 청소년 흡연자의 패턴을 파악하여, 이를 기반으로 의사결정나무 모델을 수립할 것이다. 해당 모델을 통해 청소년들의 환경요인과 정신 건강 상태 등을 입력 자료로 이용하여 해당 청소년이 흡연을 할 것으로 예측되는지 여부를 파악해 국가금연지원센터에 제공하여 보다 더 효율적인 흡연 예방 교육을 통한 청소년 흡연율 감소를 꾀할 수 있다. 더 나아가 이 모델을 학교, 교육청 등의 교육 기관에 제공하여 청소년 흡연 예방에 활용할 수 있다.

* Data Understanding

[국민건강영양조사 원시자료의 Strength, Limitation, Benefit, Cost]

Strength: 원시자료 구성이 건강설문조사 ⋅ 검진조사 ⋅ 영양조사 별로 구분이 잘 되어 있어 필요한 데이터에 대한 접근성과 활용성이 뛰어나다. 또한 연령층(성인, 청소년, 소아)의 분류가 잘 이루어져 있어 우리가 필요한 청소년에 대한 데이터를 추출해내기 용이하다. 또한 질병관리본부 건강영양조사과에서 공개한 공공 데이터이므로 해당 데이터 전반에 대한 신뢰도가 높다.

Limitation: 국민건강영양조사 설문조사 응답 시 청소년의 경우 부모를 동반해 진행한다. 이러한 외부적 요인의 작용으로 인해 설문조사의 특정 변수에 대한 신뢰도가 떨어질 가능성이 존재한다. 제7기로 지정된 16년, 17년, 18년도의 설문 항목들 중 다른 항목이 존재해 일부 데이터를 사용할 수 없다.

Benefit: 질병관리본부에서 조사해, 대대적으로 공개된 공공 데이터이므로 사용시 별도의 비용이 발생하지 않는다. 18년도의 데이터뿐만 아니라 같은 제 7기로 지정한 16년, 17년도의 데이터도 무료로 가져와 동일한 가중치를 적용해 분석할 수 있다.

Cost: 736개의 column(2018년 기준) 수와 다수의 결측치를 처리하는 데에 많은 시간과 노력이 소요된다. 예를 들어 성인인 경우와 청소년인 경우 응답해야 하는 설문 항목이 다르기 때문에 각 Data Object마다 특정 Attribute value가 Missing Value가 존재한다.

DSP를 해결하기 위해 청소년 흡연 시작 요인에 해당하는 요소들을 이전에 발행된 논문들의 연구 결과에 입각하여 선정한다. 이 요인들 중 국민건강영양조사 Data set에 존재하는 Attribute들을 1차적으로 선정한다. 해당 Attribute들 중 청소년 흡연 시작 요인이 될 수 있다고 판단한 Attribute들은 다음과 같다. (추후 상관 분석, 주성분 분석 등의 기법들을 통해 청소년 흡연과 유의미한 관계를 가진 Attribute들이 추가 또는 제외될 수 있다.)

‘평소 스트레스 인지 정도[BP1]’[[7]](#endnote-7) ⋅ ‘최근 1년동안 1주 동안 아침식사 빈도[L\_BR\_FQ]’[[8]](#endnote-8) ⋅ ‘주중 하루 평균 수면시간[Total\_slp\_wk]’[[9]](#endnote-9) ⋅ ‘가정실내 간접흡연 노출 여부[BS9\_2]’ ⋅ ‘공공기관실내 간접흡연 노출여부 [BS13]’[[10]](#endnote-10) ⋅ ‘1년간 자살 생각 여부[BP6\_10]’ ⋅ ‘2주이상 연속 우울감 여부[BP5]’ ⋅ ‘평소 하루 앉아서 보내는 시간(분) [BE8\_2]’ ’⋅‘담배 한두 모금 피운 경험[BS10\_1]’

이 중 우리는 청소년의 흡연 가능성 여부에 대해 관심이 있으므로 Target Attribute를 ‘담배 한두 모금 피운 경험[BS10\_1]’으로 삼고자 한다.

* (Expected) Data Preparation

[ 새로운 변수 생성 및 할당 ]

간접흡연 여부: ‘공공기관실내 간접흡연 노출여부 [BS13]’와 ‘가정실내 간접흡연 노출 여부[BS9\_2]’ 두 Attribute의 Value 중 하나라도 ‘1(예)’이 있으면 1, 이외의 경우는 0으로 값을 할당한다. 이 새로운 변수명을 “passive\_smoking”으로 지정한다.  
 스트레스 여부: ‘평소 스트레스 인지 정도[BP1]’의 Value가 1,2,3(조금이라도 받는 것)이면 ‘스트레스 받음(1)’, 4(거의 받지 않는다)는 ‘스트레스를 받지 않음(0)’으로 할당하였다. 8(비해당) & 9(모름)라고 응답한 event는 분석대상에서 제외한다. 이 새로운 변수명을 “stress\_status”라고 한다.

[ Model 수립에 사용될 Attribute 선정 ]

청소년을 대상으로 이루어진 한 연구에서는 대상자의 개인적 특성에서 성별, 학교급, 주관적 학업성적, 가족적 특성에서 거주형태, 아버지의 학력, 어머니의 학력, 주관적 경제적 상태 그리고 정신건강 특성에서는 우울감, 주관적 행복감, 스트레스 인지, 주관적 수면 충족이 흡연 경험에 영향을 미쳤다. [[11]](#endnote-11)또한 다른 연구에서는 아침식사는 음주, 흡연, 운동 등과 같은 건강 행위와의 상관성이 있다는 것을 밝혀냈다.[[12]](#endnote-12)

위의 논문에 의해 우리는 흡연에 영향을 미치는 사항으로서 학업성적, 경제상태, 아버지 학력, 어머니 학력, 음주 정도, 자살 생각률, 우울감, 스트레스 정도, 주관적 수면 충족, 주관적 건강 인지 정도, 가족구성원 흡연상태등을 선택한다.

부모의 최종학력과 청소년기의 음주의 경우 국민건강영양조사 데이터에 따르면 16년, 17년에는 성인을 대상으로 조사됐고, 18년에만 청소년을 포함하여 조사됐기 때문에 사용이 어려울 것으로 판단되어 분석대상에서 제외했다.

결론적으로, 국민건강영양조사에서 사용할 Attribute로는 평소 스트레스 인지 정도[BP1], 2주이상 연속 우울감 여부[BP5], 1년간 자살 생각 여부[BP6\_10], 최근 1년동안 1주 동안 아침식사 빈도[L\_BR\_FQ], 주중 하루 평균 수면시간[Total\_slp\_wk], 가정실내 간접흡연 노출 여부[BS9\_2]를 선택할 것이다.

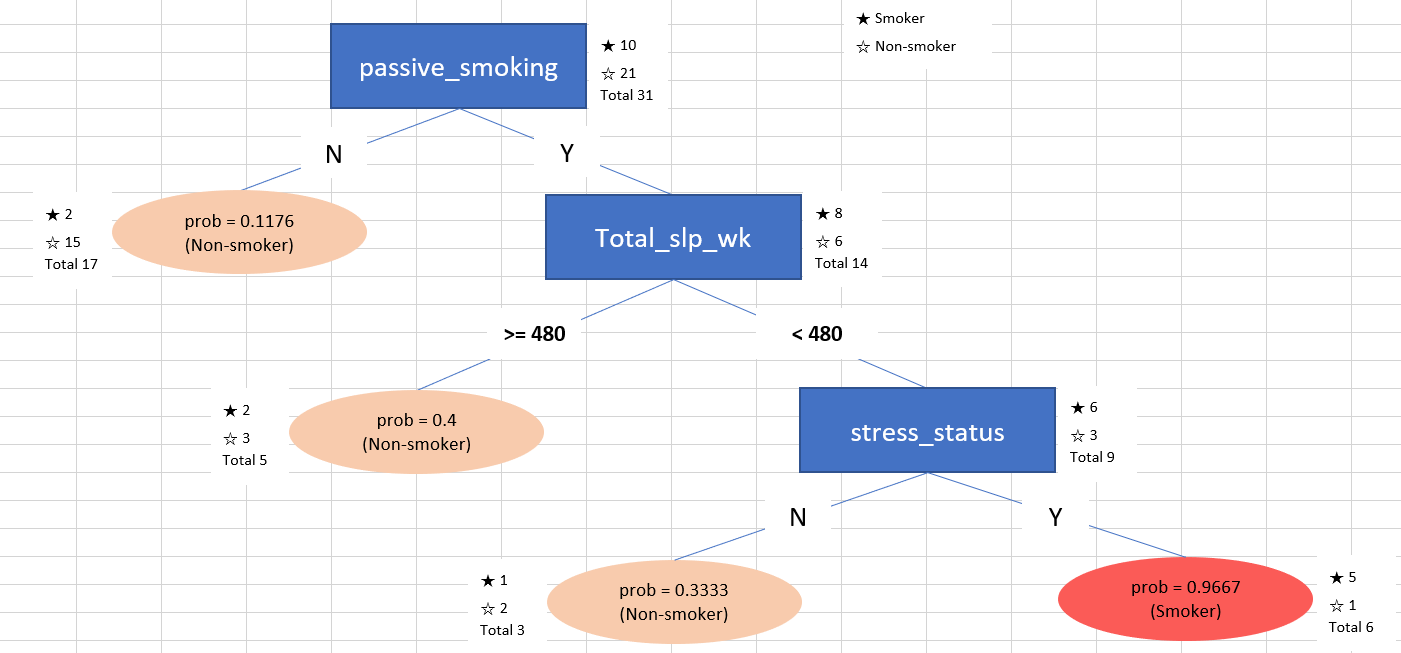
또한 우리는 청소년의 흡연 가능성 여부에 대해 관심이 있는 것이므로 담배를 한 두 모금이라도 피운 것으로 간주한다. 따라서 해당 모델에서의 Target Attribute는 청소년의 담배 한두모금 피운 경험[BS10\_1]의 여부이다.

[ 결측치 처리 ]

결측치에 대한 처리 방법 중 현재까지 사용할 수 있는 방법은 결측치가 존재하는 행 자체를 제거하는 방법과 평균값 대체를 통한 추론이 있다. 이에 따라 두 경우 모두를 수행해보고 더 타당한, 즉 결측치 처리 중 발생한 데이터의 변경에 의한 정보 손실이 더 적은 방법을 이용할 것이다.

* (Expected) Modeling

우리는 [국민건강영양조사] 데이터를 Train Data와 Test Data로 각각 임의로 31개, 9개로 선정하여 모델 과적합 등의 문제를 발견하고 대응 방안을 마련한다. 각 attribute에 따라 흡연 대상자를 분류 및 분석할 수 있어야 함으로 modeling 방법으로 Classification Analysis를 선택한다.

Attribute의 수가 증가할수록 Model Complexity가 증가하여 적정 수준을 넘게 되면 Overfitting이 될 가능성이 커진다. 따라서 우리는 Decision Tree의 Depth를 3까지만 하기로 설정했다. Interior Nodes는 사각형, Leaf nodes는 타원으로 [그림1] 처럼 표시하였다.   
  


[그림1]

모델에서 Informative Attribute를 구하기 위해 각 attribute(간접흡연, 수면시간, 스트레스)의 Information Gain을 계산하였다.  그 결과는 다음과 같다.

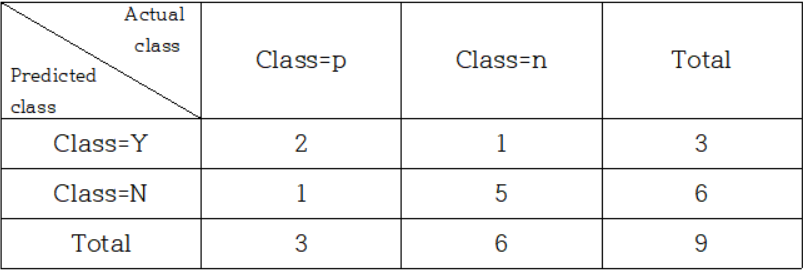
* Parent entropy = - - = 0.90717
* stress\_statusAttr(entropy) = {-{- = 0.79645
* passive\_smokingAttr(entropy) = {-{-0.73151
* total\_slp\_wkAttr(entropy) = {-{- = 0.85962
* [간접흡연] IG = 0.90717-0.73151 = 0.17566
* [수면시간] IG = 0.90717-0.85962 = 0.04755
* [스트레스] IG = 0.90717-0.79645 = 0.11072

  가장 높은 IG가 나온 [간접흡연] attribute를 Informative Attribute로 선정하였다. [간접흡연] attribute로 분류한 후, 두 번째 Attribute를 선정할 때도 같은 방법을 사용하였다. 또한, Class label을 Smoker 과 Non-Smoker로 설정하였다.

우리가 관심 있는 Class label이 흡연 여부이므로 각 노드의 probability 값을 Smoker의 수로 계산하였다.

* (Expected) Evaluation

[국민건강영양조사] 데이터 중 9개를 임의로 추출하여 Test Data를 구성하였다. 이를 Decision Tree 모델에 적용하여 Actual Class와 Predict Class 표로 만들고 Confusion Matrix를 [그림2]와 같이 만들었다.



[그림 2]

또한, Confusion Matrix를 토대로 다음과 같은 척도를 직접 계산함으로써 F-measure과 Accuracy를 도출할 수 있었다. F-measure를 통해 Precision과 Recall을 잘 통합하여 classifier가 얼마나 정확하게 분류하는지 판단할 수 있었고 Accuracy를 통해 정확성을 한 번에 나타낼 수 있었다.

1. TP rate(Recall) = TP / (TP+FN) = 0.6667
2. FN rate = FN / (TP+FN) = 0.3333
3. FP rate = FP / (FP+TN) =  0.1667
4. TN rate(Specificity) = TN / (FP+TN) = 0.8333
5. Precision = TP / (TP+FP) = 0.6667
6. F-measure = 2 \* {Precision\*Recall / (Precision + Recall)} = 0.6667
7. Accuracy = Numbers of correct decisions made / Total number of decisions made = 0.7778

-Valid: 수면시간, 간접흡연, 스트레스 각각 Information Gain을 구하여 모델을 수립하였고, 적절한 기준에 따라 Stopping Condition을 설정하여, Overfitting 과 Underfitting을 예방하였으므로 이 모델을 Valid하다.

-Reliable: [국민건강영양조사] 데이터의 일부를 직접 추출하여 분석을 하였으므로 타당하며 분석에 필요한 attribute를 선정하는데 있어서 논문을 바탕으로 선별하였기에 신뢰할 만하다. 또한 모델을 평가하는 가장 평이한 측도인 Accuracy가 0.7778 이므로 적절하다.

-limitation : 청소년의 가정 소득에 대한 attribute를 사용하고 싶었으나 [국민건강영양조사] 데이터 셋에는 성인을 대상으로만 조사하여 사용 할 수 없었다.  또한 청소년의 음주여부에 대한 attribute를 사용하고 싶었으나 16, 17년도에만 청소년을 대상으로 실시하였고 18년에는 성인에 대해서만 실시하여 사용할 수 없었다.

* 참고문헌

1. 이재정. (2007). 직간〮접 흡연과 흡연관련 질병 발생 위험도 연구. 석사학위논문, 연세대학교, 서울 [↑](#endnote-ref-1)
2. 박민아, 김미예, 하영선. (2014). 청소년 간접흡연 예방 프로그램의 개발 및 효과. 지역사회간호지, 1(25), 45 [↑](#endnote-ref-2)
3. 보건복지부. (2017). 청소년흡연. <http://www.nosmokeguide.go.kr/lay2/bbs/S1T33C109/H/22/view.do?article_seq=223&tag_name=&cpage=1&rows=10&condition=TITLE_TAG_CONTS&keyword=%EC%B2%AD%EC%86%8C%EB%85%84&cat=&rn=4> [↑](#endnote-ref-3)
4. 고숙자, 정영호. (2013). 청소년 흡연으로 인한 질병부담: 사망자 수를 중심으로. 보건복지포럼. 25 [↑](#endnote-ref-4)
5. 최지선. (2020). 청소년의 사회경제적 수준과 흡연환경이 흡연시작, 금연시도에 미치는 영향. 석사학위논문, 충남대학교, 대전 [↑](#endnote-ref-5)
6. 김원경. (2014). 청소년 성별에 따른 흡연경험 관련변인 연구. 미래청소년학회지, 11(2), 63-85 [↑](#endnote-ref-6)
7. 이현주, 김덕진. (2017). 청소년의 정신건강 특성이 흡연에 미치는 영향. 한국웰니스학회지, 12(3), 423-431 [↑](#endnote-ref-7)
8. 박아름, 이유현, 김윤진, 이상엽, 이정규, 정동욱, 탁영진, 이승훈, 황혜림, 이은경, 김규리. (2018). 40-64세 성인에서 아침식사 횟수와 건강 행위와의 관련성: 제 6기 국민건강영양조사, 2013-2014년. Korean J Fam Pract, 8(2), 286-291 [↑](#endnote-ref-8)
9. 이현주, 김덕진. (2017). 청소년의 정신건강 특성이 흡연에 미치는 영향. 한국웰니스학회지, 12(3), 423-431 에서 재인용 [↑](#endnote-ref-9)
10. 임명실, 이무식, 나백주, 홍지영, 유세종 (2010). 일부 초중학생의 간접흡연이 흡연에 미치는 영향. 한국산학기술학회 학술대회논 문집, 1209-1212 [↑](#endnote-ref-10)
11. 이현주, 김덕진. (2017). 청소년의 정신건강 특성이 흡연에 미치는 영향. 한국웰니스학회지, 12(3), 423-431 에서 재인용 [↑](#endnote-ref-11)
12. 4 박아름, 이유현, 김윤진, 이상엽, 이정규, 정동욱, 탁영진, 이승훈, 황혜림, 이은경, 김규리. (2018). 40-64세 성인에서 아침식사 횟수와 건강 행위와의 관련성: 제 6기 국민건강영양조사, 2013-2014년. Korean J Fam Pract, 8(2), 286-291 에서 재인용  
     [↑](#endnote-ref-12)