연구보고 20-R08-1

청소년 빅데이터 체계 구축 및 활용방안 연구 데이터 분석 보고서 1

: 청소년 비만에 대한 비정형 빅데이터 연구

송태민

한국청소년정책연구원
National Youth Policy Institute

연구보고 20-R08-1

청소년 빅데이터 체계 구축 및 활용방안 연구 데이터 분석 보고서 1_ 청소년 비만에 대한 비정형 빅데이터 연구

저 자 송태민

연 구 진 **연구책임자_** 송태민(삼육대학교 교수)

한국청소년정책연구원
National Youth Policy Institute

Publisher

발간사

급속한 사회환경 변화와 청소년 현안의 복잡성 증가로 인해 다양한 정형, 비정형 빅데이터에 기반한 청소년 정책수립과 미래예견적 의사결정의 중요성이 강조되고 있습니다.

본 연구는 이러한 중요성에 기반하여 문헌고찰, 분석방법에 관한 콜로키움, 외국의 빅데이터 체계 구축 및 활용 사례 검토, 청소년 핵심 영역 온톨로지 개발, 전문가 심층면접, 빅데이터 체계 구축 및 활용 방안 등에 관한 전문가 조사, 비정형 빅데이터 분석, 정형 빅데이터 분석 등을 통해 청소년 대상 빅데이터 체계 구축의 가능성을 파악하고 방안을 제시합니다.

청소년 연령 규정의 다양성과 모호성, 청소년 관련 이슈에 대한 위기 중심·잔여적 접근의 한계, 비정형 빅데이터의 광범위성 등을 고려하여 연구의 범위는 광의의 보편적 청소년복지 관점에서 청소년기본법 상의 청소년 연령인 9세~24세 청소년을 대상으로 하는 공공 정형 빅데이터, 소셜 비정형 빅데이터에 초점을 맞춥니다.

본 연구가 상대적으로 빅데이터에 대한 논의가 활발하게 진행되지 못한 청소년 분야에서 빅데이터에 대한 본격적인 논의가 이루어지는 중요한 계기가 될 수 있기를 희망합니다.

2020년 12월

한국청소년정책연구원

원장 **김 현 철**

Abstract

국문초록

본 연구의 목적은 첫째, 소셜 빅데이터를 활용하여 청소년의 비만 검색과 관련한 문서를 수집하고 주제분석과 감성분석을 통하여 청소년의 비만에 대한 미래신호를 탐지하는 것, 둘째, 머신러닝 기술을 활용하여 청소년의 다이어트 성공, 실패 예측모형을 개발하는 것이다.

비만 관련 토픽은 온라인 채널에서 매일 매시간 수집하여 총 120만 7,531건의 온라인 문서(Text)를 분석대상으로 하였다. 수집된 문서 내에 '청소년, 학생, 소아, 아동, 어린이, 학령전기, 중학생, 초등학생, 학령기, 10대, 고등학생'의 키워드가 포함된 문서를 청소년 문서(11,963건)로 분류하여 연구대상으로 하였다. 소셜 빅데이터를 수집하기 위해서는 크롤러(Crowler)를 사용하였고, 비만 토픽은 모든 관련 문서를 수집하기 위해 'Obesity', 'Diet'를 사용하였다.

주요 연구결과는 다음과 같다.

첫째, 청소년 비만 관련 운동 요인에 대한 성공 감정은 식이요인 55.6%, 운동요인 52.4%였다.

둘째, 다이어트 주요 이슈에 대한 감성과 DoD 증가율을 분석한 결과 저지방식사, 저열량식사, 해조류섭취, 근력운동, 유제품 이슈의 경우 리스크의 관리가 주요한 대응 방안이 될 수 있다.

셋째, 다이어트 실패와 독립변수 간의 연관성 예측에서 '유산소, 저지방 식사, 균형 잡힌 식사, 과일섭취'의 경우 다이어트 성공 확률이 2.17배 높은 것으로 나타났다.

연구결과를 바탕으로 비정형 빅데이터 분석을 위한 개선방안을 제시하였다.

연구요약

1. 연구목적

- 본 연구는 첫째, 소셜 빅데이터를 활용하여 청소년의 비만(다이어트) 검색과 관련된 문서를 수집하고, 주제분석과 감성분석을 통하여 청소년의 비만(다이어트)에 대한 미래신호를 탐지한다. 둘째, 머신러닝 기술을 활용하여 청소년의 다이어트 성공/실패 예측모형(인공지능)을 개발한다.

2. 연구방법

- 비만(다이어트) 관련 토픽의 수집은 2011. 1. 1~2013. 12. 31까지 온라인 채널에서 매일 매시간 수집하여 총 120만 7,531건의 온라인 문서(Text)를 분석대상으로 하였다. 수집된 문서 내에 '청소년, 학생, 소아, 아동, 어린이, 학령전기, 중학생, 초등학생, 학령기, 10대, 고등학생'의 키워드가 포함된 문서를 청소년 문서(11,963건)로 분류하여 연구대상으로 하였다. 본 연구의 소셜 빅데이터를 수집하기 위해서는 크롤러(Crawler)를 사용하였고, 비만(다이어트) 토픽은 'Obesity', 'Diet'를 사용하였다. 온라인 문서의 노이즈를 없애기 위한 불용어로 '배송비만, 관리비만' 등을 사용하였다.

3. 주요결과

- 첫째, 청소년 비만(다이어트) 관련 운동 요인에 대한 성공 감정은 식이요인(55.6%), 운동요인(52.4%)으로 나타났다.
- 둘째, 운동 요인 키워드의 중요성을 나타내는 단어 빈도에서 자전거 운동은 중요하게 평가되지 않으나 해당 주제에 대한 확산 정도인 문서빈도는 높게 나타나, 자전거 운동에 대한 청소년의 관심이 높아지는 것으로 보인다. 반면 스포츠는 중요한 키워드임에도 확산은 낮아 스포츠에 대한 청소년의 관심이 낮아지는 것으로 보인다.
- 셋째, 채소섭취, 걷기, 유제품, 유연성운동, 유산소운동은 단어빈도는 높으나 DoV의 증가율의 중앙값이 낮게 나타나 채소섭취, 걷기, 유제품, 유연성운동, 유산소운동에 대한 청소년의 관심을 높일 수 있는 방안이 마련되어야 할 것이다. 걷기, 유연성운동, 현미식사는 문서빈도는 높으나 DoD의 증가율의 중앙값과 낮게 나타나 걷기, 유연성운동, 현미식사에 대한 청소년의 관심을 높일 수 있는 방안이 마련되어야 할 것이다. 특히, 걷기운동은 단어빈도와 문서빈도에서 높게 나타났지만 증가율은 낮게 나타나 걷기운동에 대한 청소년의 관심을 높일 수 있는 방안이 마련되어야 할 것이다.
- 넷째, 약신호(2사분면)에는 댄스, 근력운동, 저지방식사, 해조류섭취, 자전거가 포함된 것으로 나타났으며, 특히 해조류섭취는 높은 증가율을 보이고 있어 해조류섭취 키워드는 시간이 지날 수록 강한 신호로 발전해 갈 수 있다. 따라서 이에 대한 과학적 근거와 다양한 식단 마련이 필요할 것으로 보인다.
- 다섯째, 다이어트 주요 이슈에 대한 감성과 DoD 증가율을 분석한 결과 저지방식사, 저열량식사, 해조류섭취, 근력운동, 유제품 이슈의 경우 리스크의 관리가 주요한 대응 방안이 될 수 있다. 다이어트식단, 현미식사, 유연성, 원푸드식사, 육류섭취 키워드는 이슈의 홍보를 강화하는 것이 청소년 등의 이슈에 대한 이해도를

높이는 방안이 될 수 있다. 과일섭취, 유산소, 걷기, 스포츠, 운동치료 키워드는 이슈에 대한 설계에 있어 잘못된 부분이 없었는지 다시 점검할 필요가 있다.

- 여섯째, training data 와 test data를 5:5로 학습하여 모델을 평가한 결과 정확도와 민감도에서 서포트벡터머신 모형이 가장 우수한 것으로 나타났다. 서포트벡터머신 모형을 이용하여 실제 데이터의 독립변수만으로 종속변수를 예측하고, 실제 데이터의 종속변수와 예측데이터의 종속변수가 동일한 학습데이터를 생성한 결과, 예측 데이터의 다이어트 성공은 53.89%로 나타났다.
- 일곱째, 랜덤포레스트 예측 모형이 다이어트 성공유무를 예측(성공, 실패)하는데 가장 큰 영향을 미치는 입력변수는 채소섭취로 나타났으며 성공 예측확률은 46.12%로 나타났다.
- 마지막으로 다이어트 실패와 독립변수 간의 연관성 예측에서 {유산소, 저지방식사, 균형 잡힌 식사, 과일섭취} 경우 다이어트 성공 확률은 2.17배 높은 것으로 나타났으며, {걷기, 저지방식사, 균형잡힌 식사}의 경우 다이어트 실패 확률이 2.11배 높은 것으로 나타났다.

4. 정책제언

- 첫째, 온라인 채널에서 청소년들이 언급하는 비만(다이어트)과 관련한 용어는 이론적 배경하에 분류된 온톨로지의 전문용어도 사용하지만 온라인 채널 이용시점에서 자주 사용하는 구어체나 속어를 사용할 수 있기 때문에 비만(다이어트) 온톨로지는 용어의 추가 등 수정·보완이 지속적으로 이루어져야 할 것이다.
- 둘째, 인공지능 개발을 위해 머신러닝 모형에 사용된 데이터에 대한 지속적인 개선(update)이 필요하다. training data를 학습하여 분석된 머신러닝 모형은 test data로 실행했을 때, 실제의 분류와 예측의 분류는 다르게 나타

날 수 있다. 따라서 모형의 예측률을 높이기 위해서는 본고에서 제시한 실제 분류와 예측분류가 동일한 케이스만 선택(selection)하여 양질의 학습데이터로 생성한 후, 지속적으로 추가하게 되면 이들 양질의 학습데이터를 다시 학습하게 될 경우, 우수한 다이어트 예측모형(인공지능)이 개발 될 것으로 본다.

차 례

청소년 빅데이터 체계 구축 및 활용방안 연구 데이터 분석 보고서 1_ 청소년 비만에 대한 비정형 빅데이터 연구

연구보고 20-R08-1

표 목차

그림 목차

제1장 서 론

1 서 론

세계보건기구(WHO)에서는 비만을 '지방조직에 비정상적으로 많은 지방이 쌓여 건강이 나빠지는 상태'로 정의하고 있다(Freedland & Aronson, 2005). 2013년 미국 의학 협회에서는 비만을 질병으로 공식 선언하였고(Cowley, Brown & Considine, 2016), 비만을 새로운 공중보건 문제이자 심각한 건강 문제로 정의하고 있다(Doll, Petersen & Stewart-Brown, 2000; Perichart, Balas, Schiffman, Barbato & Vadillo, 2007). WHO에서는 전 세계 14억 성인들이 비만 및 과체중으로 인하여 영향을 받고 있다고 보고한다(Blümel et al., 2015). 한국의 경우 비만으로 인한 사회경제적 비용은 11조 4,679억 원이며 이 가운데 의료비 손실은 51.3%(5조 8,858억원)이다. 비만과 관련한 질병으로는 당뇨병이 22.6%로 가장 높았고, 이어 고혈압, 허혈성심장질환, 관절염 순이다(National Health Insurance Service[NHIS], 2019). 비만은 영양이 과대하고 신체적 활동량이 부족한 원인 외에도 생물학적, 환경적 요인이 건강행태에 영향을 준다고 보고된다(Egger & Swinburn, 1997). 비만을 해결하기 위한 촉진 요인으로 지역사회 역할이 강조되고 있다(Schulz et al., 2005). 비만은 당뇨병, 고혈압, 고지혈증, 관절염 등의 만성질환의 위험을 증가시키며, 협심증, 심근경색증, 뇌졸중 등의 심혈관계 질환의 증가를 초래하여 조기사망의 원인이 된다(Doll et al., 2000; Nationals Heart Lung and Blood Institute[NHLBI], 1998). 비만은 만성질환(당뇨병, 고지혈증, 관절염, 고혈압), 수면 무호흡증, 신생물(암) 등의 질병과 정신

적인 문제 등에 대한 위험요인을 증가시키며(Perichart et al., 2007), 비알콜성 지방간염, 근골격근 등의 질병과도 관계되어 있다(Flegal & Troiano, 2000).

한국에서도 식생활의 서구화로 인한 영양과다, 운동부족 등의 생활방식 변화로 비만 인구가 지속적으로 증가하고 있다(정영호, 조숙자, 임희진, 2010). 비만의 요인으로는 유전, 내분비 장애요인 외에도 잘못된 식습관, 운동부족, 스트레스 등의 환경적 요인을 들 수 있고, 특히 생활수준의 향상과 식생활이 변화함으로써 최근 환경 요인에 의한 비만 환자가 증가 추세에 있다(Kim et al., 2017). 비만과 관련된 다양한 요인 중 생활습관 요인은 비만의 주요 영향 중의 하나로 식습관, 식생활 태도 및 운동습관 등의 다른 영향 요인들과 상호작용하여 비만을 유발하는 것으로 보고된다. 비만 대상자들은 영양이나 운동 등의 생활습관을 통하여 자신의 행동을 수정하거나, 약물이나, 한약을 복용하는 등의 다양한 방법을 시도하고 있으나, 대부분 효과가 적거나 일시적으로 비만이 개선되는 것으로 나타나 체중을 관리하는데 어려움을 겪고 있는 것으로 나타난다. 체중을 관리하는데 있어 실패하는 경우가 잦을수록 자아존중감이 낮아지거나 보다 소극적인 상태가 되거나 다른 사람 간의 관계가 어렵거나, 나아가 식사를 거부하거나 폭식을 하는 증상을 나타내기도 한다(안혜영·임숙빈·홍경자·허명행, 2007).

비만이 우울증의 위험을 증가시키며, 우울감이 비만을 발생시키는 예측 인자로 작용하는 것으로 나타났다(Luppino et al., 2010). 우울 상태는 폭식의 효과로 인해 체중 증가의 위험율을 높이며, 기분장애 및 불안장애를 위한 약물처치 역시 체중증가를 초래할 수 있다. 비만 클리닉을 내원한 여성을 대상으로 비만정도에 따른 우울과의 관계에 매개요인으로 작용하는 사회적 체형 불안과 스트레스 효과를 검증한 결과, 비만 정도보다는 사회적 체형 불안이 스트레스를 매개로 우울을 야기하는 것으로 나타났다. 과체중이나 비만인 경우 정상체중을 가진 사람보다 제2형 당뇨병이 발생하기 쉽고 체질량지수의 증가에 따라 인슐린저항성, 고인슐린혈증과 대사성 증후군 발생이 많은 것으로 나타났다(안근희·임미자·이혜진·김

권범·한경아·민경완, 2004). 어린이와 청소년들 사이에서 텔레비전 시청은 비만과 직접적인 관련이 있으며, 신체 활동 감소 및 건강하지 못한 식생활 습관 및 행동은 아동기 비만을 야기하는 원인이 될 수 있고(Lowry, Wechsler, Galuska, Fulton & Kann, 2002), 뿐만 아니라, 좌식생활을 많이 할수록 대사증후군의 위험도가 증가하며, 성인의 좌식습관이나 신체활동량과 대사증후군은 유의하게 관련성이 있는 것으로 연구 되었다(Edwardson et al., 2012).

비만을 치료하는 방법으로는 운동요법, 약물요법, 식이요법 등이 이용되고 있다. 약물요법은 보통 식욕을 억제하거나 갑상선을 치료하는 것으로 약물요법의 효능이 일시적으로 좋아지기는 하나, 부작용에 대한 위험이 있고 약 10% 이상의 체중이 줄어드는 효과를 얻기가 쉽지 않기 때문에 대부분 식이와 운동요법과 함께 약물요법을 같이 사용하는 것을 효과적인 방법으로 말하고 있다(이영숙, 2011). 기본적으로 비만을 치료하는 방법은 식사요법, 운동요법 및 행동을 수정하는 요법이며 약물로 치료하는 방법은 식사요법과 운동요법을 보조적으로 치료하는 방법으로 행동수정이나 생활습관을 교정하는 것만으로 체중을 줄이는 것이 효과적인 방법이 아니기 때문에 많은 수의 환자에게는 약물로 치료하는 방법을 병행하여 사용하고 있다(Kim, Park & Song, 2017). 비만과 관련된 요인을 분석하는 것은 매우 중요하다. 비만에 영향을 주는 원인은 개인의 생활습관과 주변의 환경, 그리고 개인에 대한 행동이 다양하고 복잡하게 연관되어 있어 비만을 몇 가지 단순한 원인으로만 나타내는 것은 올바르지 않다. 그리고 생활습관과 행동이론을 바탕으로 비만에 관련된 원인을 사회·심리학적인 면에서 다양하게 분석하고 있다(Kim, Kim & Kim, 1997). 뿐만 아니라 비만 문제를 해결하기 위해서는 식이영양, 운동습관, 마음가짐, 올바른 비만 정보 인식, 생활습관 개선 등의 여러 요인 등을 관리할 수 있는 서비스가 필요하다(Na, Park & Kim, 2014).

소셜미디어의 인기는 지난 몇 년 동안 엄청나게 증가하였으며, 최근에는 병원, 의료 및 보건전문가에 의하여 소셜미디어의 사용이 크게 성장하고 있다(Marjolijn L. Antheunis, Kiek Tates, Theodoor E, & Nieboer, 2013).

Kent 등(2016)의 연구에서 2012년 중 2개월이라는 기간 동안 소셜미디어(페이스북, 트위터) 상에서의 비만(뚱뚱한, 비만, 과체중, 지방)과 암 관련 단어를 포함하는 3,702건의 게시 글을 추출하여 분석한 결과, 비만과 암 사이의 인과관계 및 연관성에 중점을 둔 게시 글이 59%였으며, 긍정 및 부정적인 감정이 포함된 게시 글은 전체 게시 글의 31% 비중을 차지하였다. 페이스북에 비하여 트위터에 비례적으로 부정적인 감정의 게시 글이 많은 것으로 나타났다. Ashrafian 등(2014)의 논문에서는 소셜 네트워크 서비스는 비만 및 과체중 환자, 의료서비스 제공자 사이에서 유효한 정보 매체를 교환 할 수 있으며, 잠재적으로 체중 감량의 결과를 제공 할 수 있다고 강조했다. 2013년 미국 심장 협회(Jennifer, Tracie, Elizabeth, Richard & Alex, 2013)에서도 어린이와 청소년들의 소셜 미디어 사용량이 증가함에 따라 문자메세지, 트윗 등을 포함한 온라인 네트워킹이 비만 예방을 할 수 있는 강력한 도구가 될 수 있다고 밝혔다.

이처럼, 비만에 대해 일반 대중들의 관심이 증가하고 많은 사람들이 온라인 채널과 소셜미디어를 사용하여 비만에 대한 유익한 정보를 얻고 상호 간에 의견을 공유한다. 온라인 채널 상에서 생성되고 있는 비만과 관련된 주제에 대해 빅데이터를 수집하고 분석하고 처리하고 비만과 관련된 많은 사람의 인식과 대응방법, 사회적 현상을 보다 실질적이고 구체적으로 파악한다면 비만과 관련된 의미를 찾아낼 수 있을 것이다. 본 연구는 첫째, 소셜 빅데이터를 활용하여 청소년의 비만(다이어트) 검색과 관련된 문서를 수집하고, 주제분석과 감성분석을 통하여 청소년의 비만(다이어트)에 대한 미래신호를 탐지한다. 둘째, 머신러닝 기술을 활용하여 청소년의 다이어트 성공/실패 예측모형(인공지능)을 개발한다.

제2장 연구방법

1. 연구자료
2. 자료분석

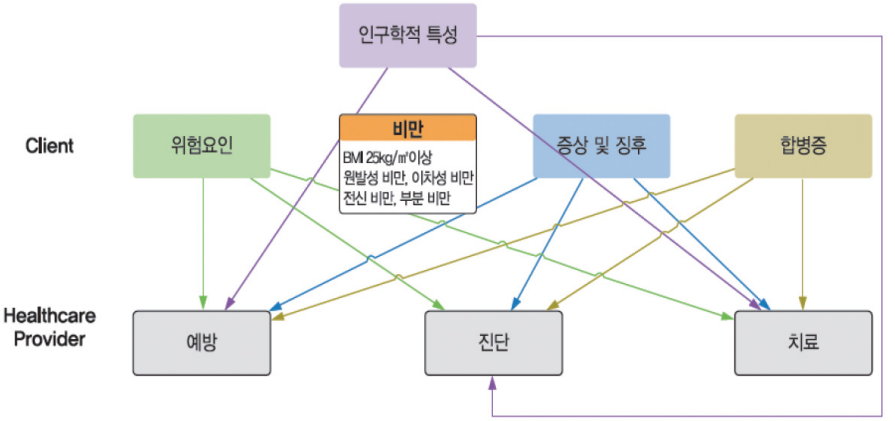
2 연구방법

1. 연구자료

본 연구는 한국에서 서비스되고 있는 온라인 뉴스채널, 블로그, 카페, 트위터, 게시판에서 총 207개의 온라인 채널을 통하여 수집된 문자(텍스트) 기반의 버즈(웹문서)를 소셜빅데이터(Social Bigdata)로 정의하였다. 비만(다이어트) 관련 토픽의 수집은 2011. 1. 1~2013. 12. 31까지 온라인 채널에서 매일 매시간 수집하여 총 120만 7,531건의 온라인 문서(Text)를 분석대상으로 하였다. 그리고 수집된 문서 내에 '청소년, 학생, 소아, 아동, 어린이, 학령전기, 중학생, 초등학생, 학령기, 10대, 고등학생'의 키워드가 포함된 문서를 청소년 문서(11,963건)로 분류하여 연구대상으로 하였다. 본 연구의 소셜 빅데이터를 수집하기 위해서는 크롤러(Crawler)를 사용하였고, 비만(다이어트) 토픽은 모든 관련 문서를 수집하기 위해 'Obesity', 'Diet'를 사용하였다. 온라인 문서의 노이즈를 없애기 위한 불용어로는 '배송비만, 관리비만' 등을 사용하였다.

그림 II-1 청소년 비만과 관련된 소셜 빅데이터의 분석 절차와 방법

소셜 미디어에서 표현되는 문서는 텍스트 형태의 비정형 데이터이기 때문에 이를 보다 효과적으로 수집·분석하기 위해서는 분석틀이 필요하며, 분석틀의 내용으로는 관련된 Topic이 어떠한 개념과 항목들로 구성되어 있는지와 각각의 개념 간 관계와 관련한 정의가 필요하다. 따라서 개념과 항목들의 관계를 반영한 온톨로지를 개발하여야 한다. 온톨로지는 관심이 있는 Topic에 대해 shared concepts (공유하는 개념)을 formalizing(형식적)하고 representing(나타내기)하기 위한, computer-interpretable knowledge model(컴퓨터가 해석이 가능한 지식 모델)(Kim, Park, Min & Jeon, 2013)이다. 따라서 온라인상의 비만(다이어트) 주제에 대한 소셜 빅데이터를 수집하기 위해서는 수집된 빅데이터 자료를 구분하고 사용하기 위한 분석의 도구로서, 비만과 관련한 Topic을 분류하고 비만을 관리할 수 있는 온톨로지와 용어의 체계를 개발해야 한다. 본 연구의 종속변수와 독립변수의 분류를 위한 주제분석은 비만(다이어트) 온톨로지를 개발하여(Kim, Park & Song, 2017) 사용하였다.



* 출처: Kim AR, Park HA, Song TM.(2017). Development and Evaluation of an Obesity Ontology for Social Big Data Analysis. Health Inform Res, 23(3), 159-168.

그림 II-2 비만 온톨로지(비만의 진단, 예방과 치료 프로세스)

1) 종속변수(Labels)

본 연구의 종속변수로는 비만(다이어트) 온라인 문서에 대한 감성분석을 통하여 다이어트 '성공유무'를 사용하였다.

비만(다이어트) 감성분석은 감성어 사전을 사용하여 긍정(Positive)감정[비만예방하다, 하체비만(탈출하다·다이어트하다), 복부비만(해결하다·관리하다·빼다·벗어나다), 다이어트(성공하다·효과적이다·올바르다·빠르다·추천하다 등), 즉 비만 탈출의 긍정적 의미]은 성공(Success)으로, 부정(Negative)[다이어트(무리하다·실패하다·잘못되다·포기하다·좋지 않다 등), 즉 비만 탈출의 부정의 의미]은 실패(Failure)로 분석하였다. 비만 감정은 통계분석을 위하여 긍정의 감정을 가진 문서는 '1', 부정의 감정을 가진 문서는 '0'으로 코드화하였다.

2) 독립변수(Feature Vectors)

본 연구의 독립변수로는 개발된 온톨로지와 이론적 배경에서 비만(다이어트)에 영향을 주는 요인을 중심으로 비만(다이어트)와 관련한 운동요인, 식이요인을 사용하였다.

비만(다이어트)와 관련한 운동요인은 걷기(달리기, 걷기, 계단, 계단오르기), 유산소(유산소운동), 유연성(스트레칭, 유연성운동, 요가), 스포츠(농구, 배구, 축구, 핸드볼, 스쿼시, 테니스, 라켓스포츠, 배드민턴), 댄스(에어로빅댄스, 에어로빅, 수중에어로빅, 댄싱), 근력운동(근력운동, 바벨, 덤벨), 자전거(하이킹, 자전거타기, 자전거, 좌식싸이클, 실내자전거), 운동치료(운동처방, 운동밴드, 짐볼, 저항성운동, 로잉머신)의 8개 요인으로 해당 운동 요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

비만(다이어트)과 관련한 식이요인은 저열량식사(저칼로리식사, 저칼로리, 저열량, 지중해식다이어트, 칼로리제한다이어트, 저탄수화물다이어트, 채식다이어트), 원푸드식사(마시는다이어트, 액체다이어트, 원푸드다이어트, 단일식품다이어트), 저지방식사(저지방, 저지방식사, 저지방식다이어트, 저지방다이어트), 다이어트식단(덴마크다이어트, 황제다이어트, 단백질파우더, Zone다이어트), 균형잡힌식사(무기질, 비타민, 칼슘, 단백질), 현미식사(잡곡밥, 잡곡류, 현미), 해조류섭취(미역, 다시마, 파래, 김), 채소섭취(야채, 녹황색채소, 채소, 채식, 양배추, 깻잎, 시금치, 상추, 쑥갓, 배추), 과일섭취(수박, 포도, 복숭아, 자두, 단감, 사과, 딸기, 과일), 육류섭취(쇠고기, 살코기, 기름기없는고기, 사태찜, 장조림, 고단백식사, 고단백다이어트, 고단백식이), 유제품(요구르트, 요플레, 저지방우유, 보통우유, 우유, 치즈, 전지분유), 소금섭취(나트륨, 소금)의 12개 요인으로 해당 식이요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

2. 자료분석

본 연구의 소셜 빅데이터에서 수집된 문자 형태의 온라인 문서에 대해 주제분석(text mining)을 위해서는 문서빈도(Document Frequency)와 단어빈도(Term Frequency)를 분석해야 한다. 단어빈도의 분석은 해당 문서에서 출현된 단어의 빈도를 구하고, 각각의 문서에 대해 나타나는 빈도를 합하여 분석할 수 있다. 문서빈도의 분석은 특정한 단어가 나타나는 문서의 수를 말하며, 주제분석(text mining)에서 중요정보를 추출하기 위해서 TF-IDF(Term Frequency-Inverse Document Frequency) 기법을 이용하고 있다. Spärck(1972)는 많이 나타나지 않은 단어에 대해 가중치를 높게 부여하는 역문서빈도(Inverse Document Frequency, $IDF_j = \log_{10}(\frac{N}{DF_j})$) 방법을 제시하였다. 단어빈도 처리에서 많이 나타나지 않은 단어에 대해 가중치를 높게 부여한다면, 단어의 빈도와 역문서의 빈도를 결합하는 'TF-IDF=$TF_{ij} \times IDF_j$'를 분석하여 단어의 중요도 지수인 가중치를 적용할 수 있다. Yoon(2012)은 온라인 뉴스채널의 문서를 수집하고 분석하여 주제분석을 통하여 생성되어진 단어빈도(Term Frequency, TF)와 문서빈도(Document Frequency, DF)를 Hiltunen(2008)이 제안한 신호, 이해, 이슈로 연계하여 분석하였다.

본 연구의 비만(다이어트)에 대한 주요 신호를 탐색하기 위하여 Yoon(2012)이 제시한 온라인 문서 내에서 출현한 비만(다이어트) 관련 주요 요인의 Keyword Emergence Map(KEM)과 Keyword Issue Map(KIM)을 분석하고 분석된 요인의 포트폴리오를 활용하여 약한 신호를 구별하였다. KEM은 Visibility(가시성)를 나타내는 것으로 DoV[Degree of Visibility: $DoV_{ij} = (\frac{TF_{ij}}{NN_j}) \times \{1 - tw \times (n - j)\}$]를 분석하고, KIM은 Diffusion(확산)의 정도를 나타내는 것으로 DoD[Degree of Diffusion: $DoD_{ij} = (\frac{DF_{ij}}{NN_j}) \times \{1 - tw \times (n - j)\}$]를 분석

가능하다. 여기서 'NN: 전체 문서수, TF: 단어빈도, DF: 문서빈도, tw: 시간가중치[1], n: 전체 시간구간, j: 시점'을 의미하고 있다.

1) 머신러닝 연구방법

본 연구의 다이어트 성공/실패에 대한 예측모형을 개발하기 위하여 머신러닝의 학습방법은 지도학습 방법과 비지도학습 방법을 사용하였다. 본 연구에 사용된 지도학습 알고리즘은 Naïve Bayes Classification 모형, logistic regression 모형, random forests 모형, neural networks 모형, support vector machines 모형, 그리고 decision trees 모형을 사용하였고, 비지도학습 알고리즘은 연관규칙을 사용하였다.

머신러닝은 이미 존재하는 데이터를 학습한 다음 학습을 통하여 기존에 알고 있는 속성을 바탕으로 신규 데이터에 대하여 예측할 수 있는 Value(값)를 발견하는 것이다. 따라서 머신러닝은 결과값을 추정하기 위하여 확률추론과 Data(데이터)를 기반으로 자동적으로 학습을 하는 알고리즘이다. 머신러닝의 학습방법은 대개 supervised learning(지도학습)과 unsupervised learning(비지도학습)으로 나뉜다.

지도학습은 학습데이터 속에 Lables(종속변수)가 존재하는 상태에서 Feature vectors(독립변수)와 Labels를 참고하여 Learning(학습)하여 모델을 만든 후, Labels이 존재하지 않는 새로운 데이터의 Feature vectors만으로 추정된 Labels를 나타낸다[그림 V-3]. 지도학습에 포함되는 머신러닝 알고리즘으로는 Naïve Bayes Classification 모형, logistic regression 모형, random forests 모형, neural networks 모형, support vector machines 모형, 그리고 decision trees 모형이 있다.

---

1) 본 연구에서 시간 가중치는 0.05를 적용함

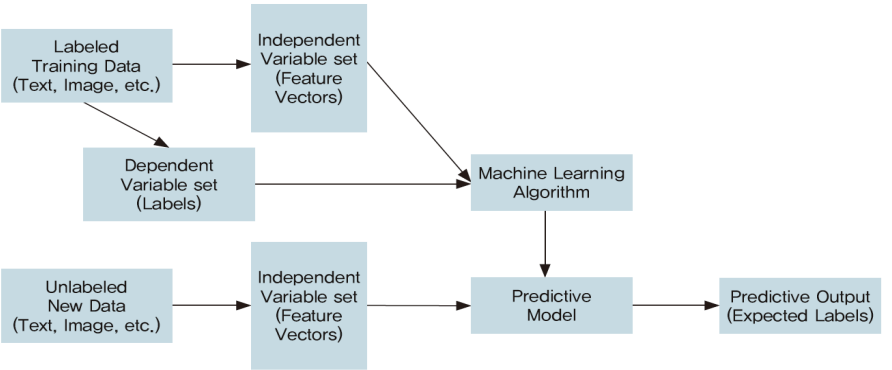


그림 II-3 지도학습 모델링

비지도학습은 학습데이터 속에 Labels가 없는 상태에서 Feature vectors만으로 학습하고 모형을 만든 후, Labels가 속하지 않는 새로운 데이터의 Feature vectors만으로 추정된 Labels를 나타낸다. 비지도학습에 포함되는 머신러닝 알고리즘으로는 association analysis(연관분석)와 cluster analysis(군집분석) 등이 있다.

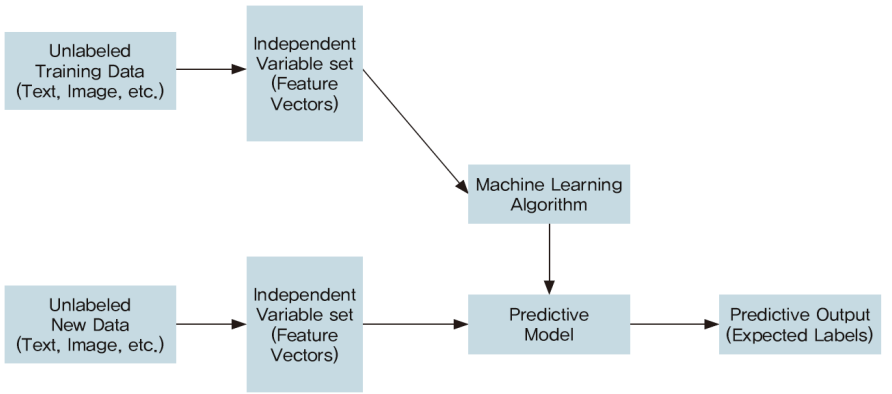


그림 II-4 비지도학습 모델링

본 연구에 사용된 지도학습 알고리즘은 Naïve Bayes Classification 모형, logistic regression 모형, random forests 모형, neural networks 모형, support vector machines 모형, 그리고 decision trees 모형을 사용하였다.

Naïve Bayes Classification 모형은 조건의 확률값에 대한 규칙인 Bayes 정리를 바탕으로 하여 분류하여 학습하는 방법을 말한다. Bayes 정리는 이전확률값에 특별한 event(사건)가 있을 경우 그 사건의 확률값이 변경될 수 있다는 것이다. 즉 'posterior probability(사후확률)는 prior probability(사전확률)을 통하여 추정가능 하다'라는 뜻에 기반하여 모형의 분류를 예측하고 있다. 즉, P(A|B)는 B가 생성되었을 때 A가 생성될 확률, P(B|A)는 A가 생성되었을 때 B가 생성될 확률, P(A,B)는 A와 B가 같이 생성될 확률, P(A)는 A가 생성될 확률, P(B)는 B가 생성될 확률을 나타내고 있다. Naïve라는 뜻은 단순하고 어리석다는 의미를 나타낸다. 따라서 Naïve Bayes는 분류를 빠르고 쉽게 하기 위하여 분류하는데 사용되는 속성들이 상호간에 확률론적으로 서로 독립이라는 것을 가정하고 있기 때문에 확률론적으로 서로 독립이라는 기본 가정이 위배될 때에는 Error를 발생시킬 수 있다. 따라서 Naïve Bayes는 속성을 많이 포함하고 있는 데이터에 대하여 속성 상호 간의 관련성을 살펴보게 되면 매우 복잡해지는 경우가 발생 할 수 있기 때문에 단순화하여 real time(실시간)으로 예측을 빠르게 판단할 수 있게 사용된다. 따라서 스팸 mail의 구분이나 질병예측의 분야에서 대부분 이용되고 있다.

logistic regression 모형은 Feature vectors는 양적인 데이터를 가지며 Labels는 다변량인 비선형 regression 모형이다. 통상적으로 regression 모형의 적합도의 검정방법은 잔차에 대한 제곱의 합을 최소화시키는 최소자승법을 이용하지만 logistic regression 모형은 Event의 생성 가능성을 많게 하는 확률값인 likelihood(우도비)를 최대한 크게하는 최대우도추정방법을 이용한다. logistic regression 모형은 Feature vectors(예측변수)가 Labels에 영향을 주는 값인 승산에 대한 확률값인 odds ratio(오즈비)로 검정한다. 따라서 Labels의 범주가

(0과 1)인 binary(이분형) logistic regression 모형을 추정하기 위한 확률값의 비율에 대한 승산의 비율에 대한 변화량을 예측하는 것이다. multinomial(다항) logistic regression 모형은 Feature vectors는 양적인 데이터를 가지며, Labels의 범주가 3개 이상인 여러 개의 범주를 가지고 있다.

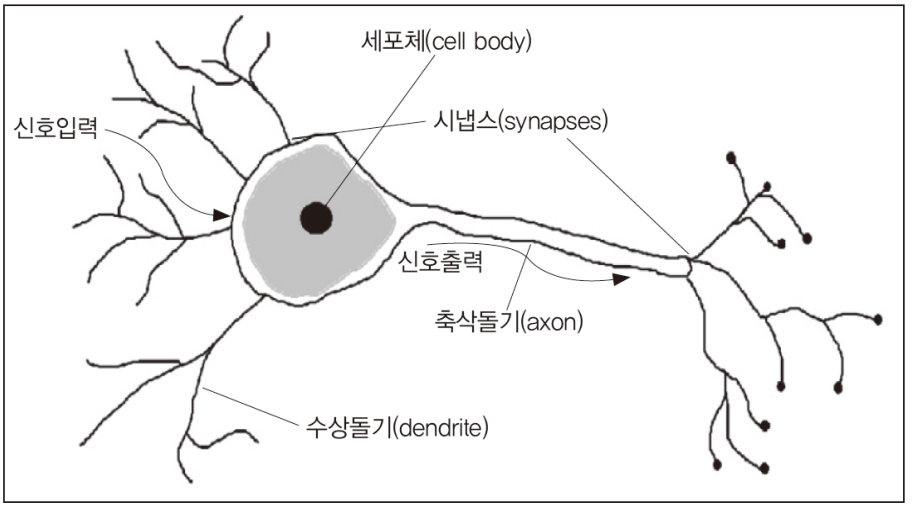
decision trees 모형은 결정된 규칙에 따라서 Tree 구조로 도식화하여 예측과 분류를 실행하는 방법이다. decision trees 모형은 segmentation(세분화), classification(분류), clustering(군집화), forecasting(예측) 등의 목적으로 사용하고 있다. decision trees 모형의 장점으로는 Tree구조에서 예측된 변수가 목표된 변수를 설명 하는 부분에 있어 얼마나 중요한 것인지를 쉽게 알아볼 수 있고 두 개 이상의 변수가 결합되어서 목표가 되는 변수에 영향을 미치는지를 쉽게 판단할 수 있다.

기타 지도학습 알고리즘은 random forests 모형, neural networks 모형, support vector machines 모형 등이 있다.

random forest(랜덤포레스트) 모형은 Breiman(2001)이 제안 한 것으로, 학습데이터에서 여러 개의 예측된 모델들을 생성한 후, 예측된 모델들을 결합하여 한 개의 최종 예측된 모델을 생성하는 기계학습을 위한 ensemble(앙상블) 방법 중 한 개다. 따라서 분류의 정확도가 높고 이상값에 둔감하며 계산이 신속하다는 장점이 있다(Jin & Oh, 2013). 처음 개발된(Breiman, 1996) 앙상블 알고리즘은 Bagging(배깅)이다. Bagging은 decision trees 의 단점인 '최초의 분리되는 변수가 변경되면 마지막 decision trees가 완전하게 변경되어 예측에 대한 저하를 가져올 수 있고, 동시에 예측된 모형에 대한 해석이 어려운' 안정되지 않은 Learning 방법을 삭제함으로써 예측에 대한 영향을 향상하기 위한 분석 방법이다. 즉, 이미 있는 데이터에 대하여 많은 bootstrap(붓스트랩) 데이터를 만들어 예측된 모형을 생성한 후, 예측된 모형을 결합하여 마지막 모형을 만든다. 랜덤포레스트는 학습 데이터에서 n개의 데이터를 사용하여 붓스트랩 표본을 만들어

input variable들 중 일부에 대해서 random(무작위) 하게 추출하여 decision trees를 만들고, decision trees를 선형적으로 결합하여 마지막 학습기를 생성한다. random forest 모형은 Variable(변수)에 대한 중요도 지수를 제공한다. 즉, 특정 변수에 대한 중요도 지수는 특정 변수를 내포하지 않은 경우에 보다, 특정 변수를 내포할 경우에 예측된 오차값이 적어지는 정도를 나타내는 것이다. random forest는 terminal node(단노드)가 있을 경우에 terminal node의 majority(과반수)로 Labels의 분류를 결정한다. random forest에서 MSE (Mean Decrease Accuracy)는 가장 Robust(강건한) information(정보)에 대한 것으로 정확도로 표현한다. IncNodePurity(Mean Decrease Gini)는 최선의 분류에 대한 손실과 관련한 함수로서 중요도를 나타낸다.

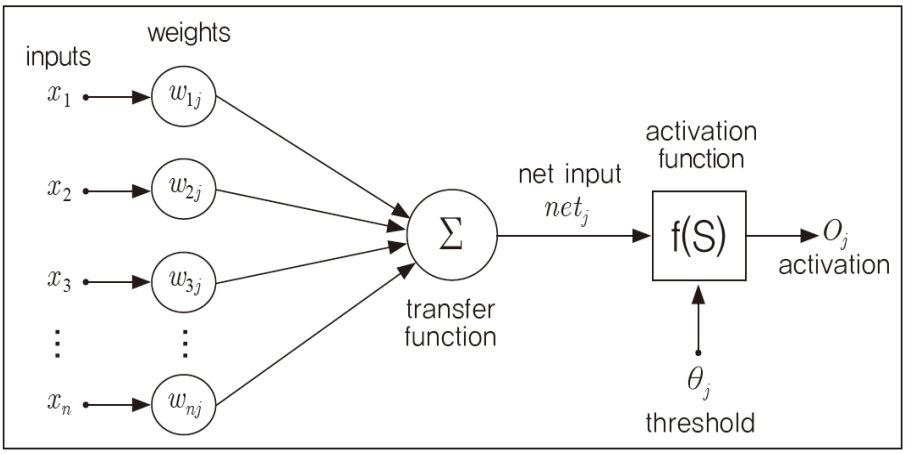
Artificial neural network(인공신경망모형)는 인간의 신경체계와 같은 생물학적인 신경체계망의 동작 방식을 기본으로 하는 머신러닝 모델의 한 종류로 인간의 두뇌에서 의사를 결정하는 형태를 모방(copy)하여 나누는 모델이다. 인간의 Biological Neural Network(신경망)는 약 250억 개의 뉴런(신경세포)으로 구성되어 있다. 특히 뉴런은 한개의 cell body(세포체)와 cell body의 돌기인 한개의 axon(축삭돌기)와 많은 dendrite(수상돌기)로 형성되어 있으며, 뉴런 간의 정보교류는 synapses(시냅스)라고 부르는 연결부위를 통하여 이루어진다. synapses는 뉴런의 signal을 바로 전달하는 것은 아니다. signal의 세기가 Threshold(임계치) 이상으로 나타나야 signal을 전달하게 된다. 따라서 cell body는 dendrite에서 input(입력)된 signal을 축적함으로써 Threshold에 전달되고, output signal을 axon에 전달하고 axon 끝단의 synapses를 통해 이웃한 신경세포인 뉴런으로 전달된다.



* 출처: Biological Neural Network,
https://cogsci.stackexchange.com/questions/7880/what-is-the-difference-between-biological-andartificial-neural-networks. 2020.4.29. 인출

그림 II-5 Biological Neural Network

Neural network은 사람의 뇌의 구조를 모방하여 만든 지도학습의 한가지 방법으로 많은 신경세포들을 서로 간에 연결시켜 입력값(input value)에 대하여 최적의 출력값(output value)을 예측할 수 있다. 따라서 Neural network은 두뇌의 가장 기본적인 단위인 신경세포와 같이 training data(학습데이터)에서 signal을 입력 받아서 그 값이 특정한 임계점에 도달하게 되면 output을 발생시킨다.



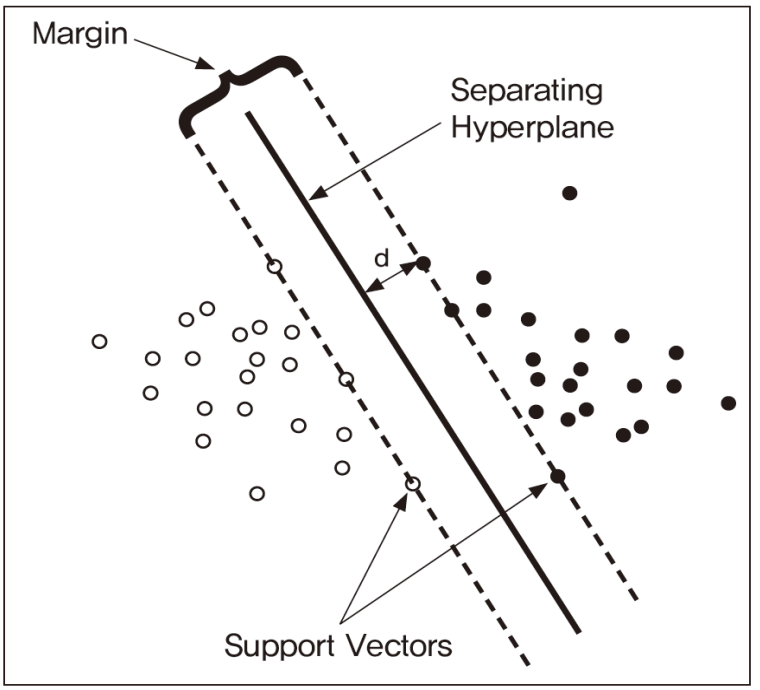
* 출처: Artificial neural network, https://commons.wikimedia.org/wiki/File:Artificial_neural_network.png, 2020.4.29. 인출

그림 II-6 Artificial Neural Network

Minsky와 Papert(1969)는 선형적인 문제에서만 해결 할 있는 퍼셉트론(perceptron)을 제안하였다. 퍼셉트론은 단층신경망에 은닉층을 추가하여 일반화로 형성된 비선형함수로 분류하였다. 그리고, Rumelhart, GHinton & Williams (1986)는 출력층에서 나타날 수 있는 error를 역으로 전파(back propagation) 함으로써 은닉층에 대해 학습 가능한 역전파 알고리즘을 제시하였다. 그리고 딥러닝(deep learning)을 통하여 심층 신경망을 생성할 수도 있다. 딥러닝은 input과 output 사이에 여러 개 숨어 있는 레이어가 존재하는 다층신경망이다. 다층신경망 모델을 구성할 경우의 고려사항은 다음과 같이 여러 가지가 있다. 첫째, input variable의 value의 범위를 결정하여야 한다. Neural network 모형에 적합한 학습데이터를 생성하기 위해서는 범주형 변수로 변환하여야 하며, 모든 범주에서는 일정한 수(빈도) 이상의 값이 존재해야 한다. 그리고 연속형 변수는 범주형 변수로 변경하고, 변수의 범주의 값이 0~1에 존재하도록 변경한다. 둘째, 은닉층

의 수와 은닉노드의 수를 적당한 수로 결정하여야 한다. 은닉층의 수와 은닉노드의 수가 많게 되면 가중계수의 수가 상대적으로 많아져 overfit(과적합)이 될 가능성이 있다. 따라서 Neural network 모델을 구성할 때 대부분 은닉층의 수는 한개로 하고 은닉노드의 수는 충분하게 하여 은닉노드의 수를 한개씩 감소시키면서 분류기의 정확하면서 은닉노드의 수가 가급적 작은 모델을 선택한다.

Support Vector Machine(SVM) 모형은 Cortes와 Vapnic(1995)이 제안한 것으로 지도학습 머신러닝의 일종이다. SVM 모형은 회귀와 분류에 모두 사용된다. Logistic regression 모형은 input value가 있을 때 output value에 대한 조건부적인 확률값을 추정하는 데 비하여, SVM 모형은 확률을 추정하지 않고 직접적으로 분류하여 결과만 예측한다. 따라서 SVM 모형은 분류 효율이 높아 빅데이터(모집단)를 학습할 경우 확률추정을 하는 방법들 보다 전반적으로 예측력이 높다. SVM은 두개의 집단(y:1 또는 y:-1)의 경계선을 가로지르는 두 개의 support vector(초평면)에서 두개의 집단 경계선에 존재하는 데이터 사이의 거리의 차이(margin)가 최대가 되는(잘못된 분류를 최소시키는) 모델을 최종 결정하게 된다.



* 출처: Support Vector Machines, https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf, 2020.4.29. 인출

그림 II-7 Support Vector Machine Classification

본 연구에 사용된 비지도학습 알고리즘은 연관분석을 사용하였다.

연관분석은 용량 큰 Database에서 변수들 간의 관계가 의미가 있는지를 살펴보기 위한 방법이다. 연관분석은 기본적인 통계적 가설검정의 과정이 요구되지 않으며 빅데이터 내에 내포되어 연관규칙을 발견하는 것이다. 연관규칙은 '기저귀를 사는 남자가 맥주를 같이 구입한다.'라는 장바구니 분석에서 사용되는 기법으로, 데이터에서 발생하는 키워드도 이러한 장바구니 분석에 적용하여 분석할 수 있다. 빅데이터 분석에서 사용되는 연관분석은 한 개의 recode에서 속하는 두 개 이상의 변수들 간의 상호연관성을 찾는 것이다. 연관분석은 동시에 같이 발생한

변수들의 집합에 대한 연관규칙과 관련 조건을 발견하는 분석기법이다. 연관규칙의 평가는 전체 문서를 분석하여 산출되는 confidence(신뢰도), support(지지도), 그리고 lift(향상도)로 평가할 수 있다. support는 전체 문서에서 발생되는 연관규칙(X→Y)과 관련되는 데이터에 대한 비율을 말한다. confidence는 변수 X를 내포하는 레코드에서 변수 Y가 포함되는 레코드의 비율을 뜻한다. lift는 변수 X가 존재하지 않을 경우 변수 Y에 대한 확률이 변수 X가 존재할 경우, 변수 Y에 대한 확률이 증가하는 비율을 말한다. lift는 커질수록 변수 X에서 발생하는 여부가 변수 Y가 발생되는 여부에 영향을 주게 된다. 따라서 support는 빈번하게 발생하지 않는 규칙에 대해 제한하는데 이용된다. confidence는 변수들의 관련 정도를 분석하는 데 사용될 수 있다. lift는 발생되는 연관규칙(X→Y)에서 변수 X가 있을 때 변수 Y가 발생되는 비율을 나타낸다. 연관분석은 분석하는 사람이 정한 최소의 support를 만족시키는 frequent item set(빈발항목집합)을 생성한다. 그리고, 생성된 규칙에 대해 최저 confidence의 기준을 정하고 lift 가 1 이상인 것을 규칙으로 선택한다(Park, 2013).

머신러닝 모형의 평가하는 방법으로는 training data와 test data로 나누어(예 7:3)하여 training data에서 생성된 모형함수를 test data에 실행했을 경우 나타나는 오분류표와 Receiver Operation Characteristic(ROC) Curve로 평가하였다.

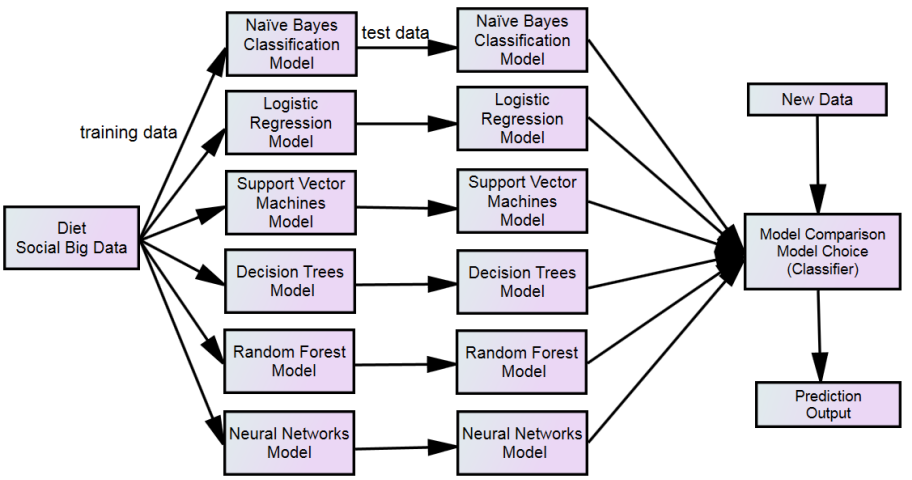


그림 II-8 청소년 비만 소셜 빅데이터 예측모델 평가

본 연구의 머신러닝 모형의 평가는 training data에서 생성된 모형 함수를 test data에서 실행했을 경우에 생성되는 분류 정확도를 사용하였다.

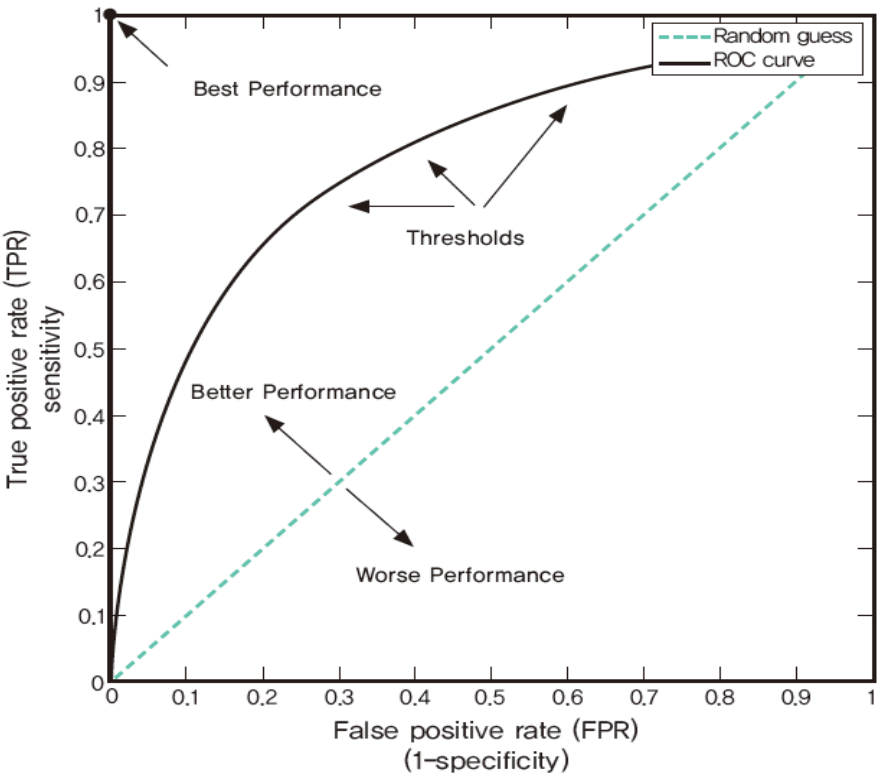
표 II-1 오분류표

분류집단 / 실제집단	O(Negative)	1(Positive)
O(Negative)	N_{00}	N_{01}
1(Positive)	N_{10}	N_{11}

* N: 전체 데이터 수

Accuracy(정확도)는 전체 데이터 중에서 정확하게 분류된 비율이다. error rate(오류율)는 잘못 분류된 비율이다. Sensitivity(민감도)는 부정적으로 분류된 자료 중에서 올바르게 분류된 자료의 비율이다. Specificity(특이도)는 긍정적 자료 중에서 올바르게 분류된 자료의 비율이다. Precision(정밀도)는 잘못 분류된 자료 중에서 실제 잘못 분류된 자료의 비율이다. 그리고 본 연구에서의 머신러닝

모형의 평가는 여러 개의 절단값에서 특이도와 민감도 간의 관계를 나타내는 ROC Curve를 사용하였다.



* 출처: Hassouna, M., Tarhini, A., Elyas, T. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. International Business Research, Vol 8(6), pp. 224-237., Vol 8(6), pp. 224-237.

그림 II-9 ROC curve

본 연구의 기술통계와 decision tree 분석은 SPSS 24.0을 사용하였고 머신러닝 모델링과 연관분석은 R 3.6.1 version을 사용하였다.

제3장 연구결과

3 연구결과

1. 소셜 빅데이터 기반 다이어트[2] 미래신호 탐색

다이어트 관련 성공감정은 근력운동, 유연성운동, 댄스, 유산소운동, 걷기 등의 순으로 높게 나타났다. 다이어트 관련 식이요인에 대한 성공 감정은 다이어트식단, 저지방식사, 저열량식사, 해조류섭취, 현미식사, 유제품 등의 순으로 높게 나타났다. 다이어트 관련 채널별 성공감정은 Board, Blog, Café, News, Twitter의 순으로 높게 나타났다.

표 Ⅲ-1 비만(다이어트) 관련 운동/식이요법/질병/치료의 감정 교차분석

단위: 빈도(%)

Factor	Item	Emotion		Total
		Success	Failure	
운동요인	걷기	418(50.5)	409(49.5)	827
	유산소	318(52.0)	294(48.0)	612
	유연성	454(58.9)	317(41.1)	771
	스포츠	99(44.0)	126(56.0)	225
	댄스	76(53.1)	67(46.9)	143
	근력운동	199(62.4)	120(37.6)	319
	자전거	143(48.5)	152(51.5)	295
	운동치료	38(27.7)	99(72.3)	137
	계	1,745(52.4)	1,584(47.6)	3,329

2) 비만과 다이어트에 관한 온라인 문서를 수집하여 다이어트 성공유무를 예측하는 모델을 개발하는 것이 연구의 목적이기 때문에 연구결과에서는 비만(다이어트)을 다이어트로 통일하여 사용한다.

Factor	Item	Emotion		Total
		Success	Failure	
채널	Twitter	253(31.5)	551(68.5)	804
	Blog	1,376(52.2)	1,260(47.8)	2,636
	Café	485(47.0)	547(53.0)	1,032
	Board	5(50.0)	5(50.0)	10
	News	981(37.5)	1,634(62.5)	2,615
	계	3,100(43.7)	3,997(56.3)	7,097
성공유무		3,100(43.7)	3,997(56.3)	7,097
식이 요인	**저열량식사**	247(64.2)	138(35.8)	385
	원푸드식사	166(58.2)	119(41.8)	285
	저지방식사	135(71.4)	54(28.6)	189
	다이어트식단	117(75.5)	38(24.5)	155
	균형잡힌식사	1,369(54.9)	1,124(45.1)	2,493
	현미식사	301(60.9)	193(39.1)	494
	해조류섭취	143(63.3)	83(36.7)	226
	채소섭취	1,359(51.5)	1,278(48.5)	2,637
	과일섭취	643(54.4)	540(45.6)	1,183
	육류섭취	165(56.9)	125(43.1)	290
	유제품	504(58.3)	361(41.7)	865
	소금섭취	269(49.3)	277(50.7)	546
	계	5,418(55.6)	4,330(44.4)	9,748

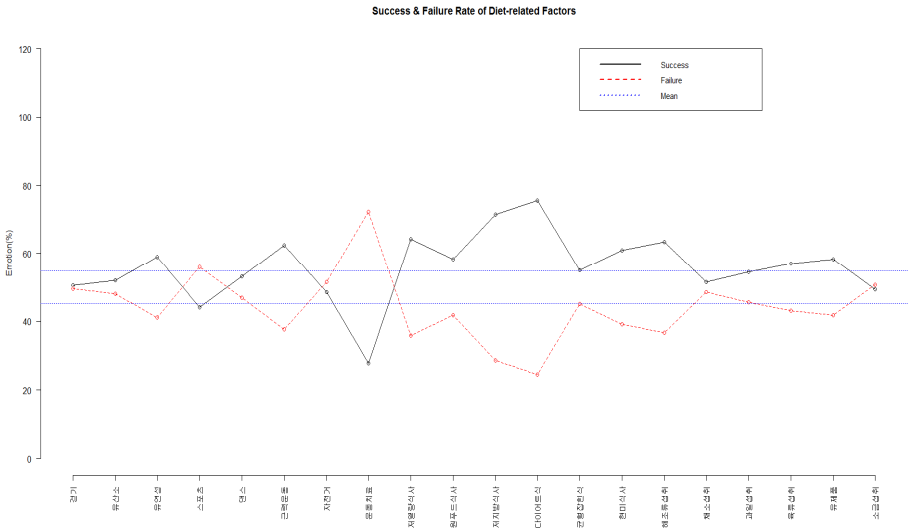


그림 Ⅲ-1 비만(다이어트) 관련 운동/식이요법의 성공유무 변화

2. 미래신호 탐색 다이어트 관련 키워드의 단어 및 문서 빈도 분석

다이어트 관련 단어빈도와 문서빈도 그리고 단어의 중요도 지수를 고려한 문서빈도를 분석한 결과는 다음과 같다. 다이어트에 대한 운동, 식이요법, 치료, 질병의 신호(키워드)의 변화를 살펴보았다. 단어빈도에서는 채소섭취, 균형잡힌식사, 과일섭취, 걷기, 유제품, 유연성, 소금섭취, 현미식사, 유산소, 저열량식사 등의 순위로 나타나고 있다. 운동 요인은 걷기, 식이요법 요인은 채소섭취가 중요한 키워드로 나타났다. 특히, 운동요인 중 자전거 운동[TF(15위)→DF(12위)]은 많이 확산되고 있는 것으로 나타났다. 중요도 지수를 고려한 단어 빈도에서는 운동 요인은 걷기와 유연성 운동, 식이요법 요인은 채소섭취, 균형잡힌식사가 중요한 키워드로 나타났다.

표 Ⅲ-2 온라인 채널의 비만(다이어트)의 키워드 분석

순위	TF		DF		TF-IDF	
	키워드	빈도	키워드	빈도	키워드	빈도
1	채소섭취	5,335	채소섭취	3,007	채소섭취	3,620
2	균형잡힌식사	5,024	균형잡힌식사	2,772	균형잡힌식사	3,586
3	과일섭취	1,896	과일섭취	1,319	과일섭취	1,965
4	걷기	1,275	유제품	971	걷기	1,544
5	유제품	1,254	걷기	883	유제품	1,466
6	유연성	931	유연성	808	유연성	1,163
7	소금섭취	772	소금섭취	664	소금섭취	1,030
8	현미식사	654	유산소	622	현미식사	945
9	유산소	622	현미식사	514	유산소	848
10	저열량식사	439	저열량식사	407	스포츠	688
11	스포츠	393	근력운동	328	저열량식사	679
12	육류섭취	390	자전거	310	육류섭취	653
13	근력운동	387	육류섭취	303	근력운동	635
14	해조류섭취	350	원푸드식사	288	해조류섭취	625
15	자전거	342	스포츠	255	자전거	570
16	원푸드식사	289	해조류섭취	235	원푸드식사	491
17	저지방식사	201	저지방식사	200	저지방식사	373
18	다이어트식단	179	다이어트식단	157	다이어트식단	351
19	댄스	151	운동치료	151	댄스	300
20	운동치료	151	댄스	149	운동치료	299
합계		21,035	합계	14,343	합계	21,830

키워드의 년도별 순위의 변화는 2011년에는 유산소 운동이 5위에서 2012년과 2013년에는 7위로 나타나 유산소운동의 중요성이 감소되는 것으로 나타났다. 반면 2011년에 저열량식사는 15위에서 2012년과 2013년에 10위로 나타나 시간이 갈수록 다이어트와 관련된 저열량식사에 대한 이슈가 증가하고 있는 것으로 나타났다.

표 Ⅲ-3 온라인 채널의 비만(다이어트)의 월별 키워드 순위변화(TF기준)

순위	2011년	2012년	2013년
1	채소섭취	채소섭취	균형잡힌식사
2	균형잡힌식사	균형잡힌식사	채소섭취
3	과일섭취	과일섭취	과일섭취
4	걷기	걷기	걷기
5	유제품	유제품	유제품
6	유연성	소금섭취	유연성
7	유산소	유연성	현미식사
8	소금섭취	현미식사	소금섭취
9	현미식사	유산소	유산소
10	스포츠	저열량식사	저열량식사
11	육류섭취	스포츠	근력운동
12	근력운동	근력운동	육류섭취
13	해조류섭취	육류섭취	해조류섭취
14	자전거	자전거	자전거
15	저열량식사	원푸드식사	스포츠
16	원푸드식사	해조류섭취	원푸드식사
17	다이어트식단	저지방식사	저지방식사
18	운동치료	다이어트식단	댄스
19	저지방식사	운동치료	운동치료
20	댄스	댄스	다이어트식단

3. 다이어트 관련 키워드의 미래신호 탐색

비만(다이어트) 키워드에 대한 평균단어 빈도와 DoV의 증가율을 다음과 같이 분석하였다. 채소섭취, 걷기, 유제품, 유연성운동, 유산소운동은 높은 빈도를 보이고 있으나 DoV 증가율(0.063)에 대한 중앙값보다 적게 산출되어 시간이 지날수록 signal은 약해지는 것으로 산출되었다. 균형잡힌식사, 과일섭취, 소금섭취, 현미식사는 빈도는 높게 나타났으며, DoV 증가율은 중앙값보다 높게 나타나 시간이 갈수록 빠르게 신호가 강해지는 것으로 나타났다. DoD의 증가율과 평균문서 빈도를 산출한 결과 걷기, 유연성운동, 현미식사는 문서빈도는 높으나 DoD의 증가율의 중앙값(0.018) 보다 낮게 나타났다.

앞에서(자료분석) 제시한 미래신호 탐색절차와 같이 DoV와 DoD의 평균단어빈도와 평균문서빈도를 X축에 나타내고 DoD의 평균증가율과 DoV의 평균증가율을 Y축으로 나타낸 후, 각 값의 median(중앙값)으로 사분면으로 구분하면 2사분면에 포함되는 키워드가 약신호이다. 그리고 1사분면에 포함되는 키워드는 강신호이다. 빈도수 측면에서는 상위 10위에 DoV는 채소섭취, 균형잡힌식사, 과일섭취, 걷기, 유제품, 유연성, 소금섭취, 현미식사, 유산소, 저열량식사 순으로 포함되었고, DoV의 증가율의 중앙값(0.063) 보다 높은 증가율을 보이는 키워드는 저열량식사, 댄스, 근력운동, 저지방식사, 해조류섭취, 균형잡힌식사, 소금섭취, 현미식사 순으로 나타났으며 DoD의 증가율의 중앙값(0.018) 보다 높은 증가율을 보이는 키워드는 댄스, 저열량식사, 해조류섭취, 균형잡힌식사, 근력운동, 채소섭취, 자전거, 소금섭취 순으로 나타났다.

표 Ⅲ-4 비만(다이어트) 키워드의 DoV 평균증가율과 평균단어 빈도

키워드	DoV			평균증가율	평균단어빈도
	2011년	2012년	2013년		
채소섭취	0.225	0.262	0.234	0.028	1778
균형잡힌식사	0.193	0.231	0.265	0.173	1675
과일섭취	0.081	0.084	0.092	0.065	632
걷기	0.060	0.051	0.060	0.006	425
유제품	0.060	0.050	0.058	-0.007	418
유연성	0.049	0.036	0.037	-0.115	310
소금섭취	0.028	0.044	0.033	0.153	257
현미식사	0.028	0.026	0.035	0.148	218
유산소	0.029	0.025	0.029	0.004	207
저열량식사	0.014	0.022	0.025	0.356	146
스포츠	0.021	0.018	0.013	-0.194	131
육류섭취	0.019	0.014	0.020	0.072	130
근력운동	0.016	0.014	0.023	0.234	129
해조류섭취	0.016	0.012	0.019	0.181	117
자전거	0.015	0.014	0.017	0.061	114
원푸드식사	0.013	0.013	0.012	-0.044	96
저지방식사	0.007	0.011	0.010	0.198	67
다이어트식단	0.010	0.009	0.004	-0.350	60
댄스	0.007	0.005	0.009	0.319	50
운동치료	0.008	0.007	0.004	-0.248	50
중앙값				0.063	138.5

표 Ⅲ-5 비만(다이어트) 키워드의 DoD 평균증가율과 평균문서 빈도

키워드	DoD			평균증가율	평균문서빈도
	2011년	2012년	2013년		
채소섭취	0.175	0.208	0.220	0.123	1002
균형잡힌식사	0.154	0.195	0.209	0.170	924
과일섭취	0.084	0.091	0.086	0.018	440
유제품	0.067	0.056	0.068	0.018	324
걷기	0.062	0.053	0.058	-0.021	294
유연성	0.062	0.046	0.048	-0.111	269
소금섭취	0.038	0.055	0.039	0.071	221
유산소	0.042	0.038	0.042	0.008	207
현미식사	0.036	0.030	0.035	-0.003	171
저열량식사	0.020	0.031	0.033	0.325	136
근력운동	0.021	0.018	0.026	0.126	109
자전거	0.020	0.018	0.024	0.111	103
육류섭취	0.023	0.017	0.019	-0.058	101
원푸드식사	0.019	0.020	0.018	-0.044	96
스포츠	0.018	0.018	0.014	-0.095	85
해조류섭취	0.015	0.012	0.021	0.294	78
저지방식사	0.010	0.016	0.014	0.204	67
다이어트식단	0.013	0.011	0.006	-0.317	52
운동치료	0.012	0.011	0.006	-0.243	50
댄스	0.009	0.007	0.013	0.332	50
중앙값				0.018	122.5

KEM과 KIM에 동시에 출현되는 다이어트 관련 강신호(1사분면)에는 균형잡힌 식사, 과일섭취, 소금섭취, 저열량식사가 포함되었고, 약신호(2사분면)에는 댄스, 근력운동, 저지방식사, 해조류섭취, 자전거가 포함된 것으로 나타났다. KIM의 4사분면에 출현하는 강하지는 않지만 증가율이 약한 신호는 걷기, 유산소, 유연성으로 나타났으며, KIM의 3사분면에 출현되는 잠재신호는 원푸드식사, 스포츠, 운동치료, 다이어트식단으로 나타났다.

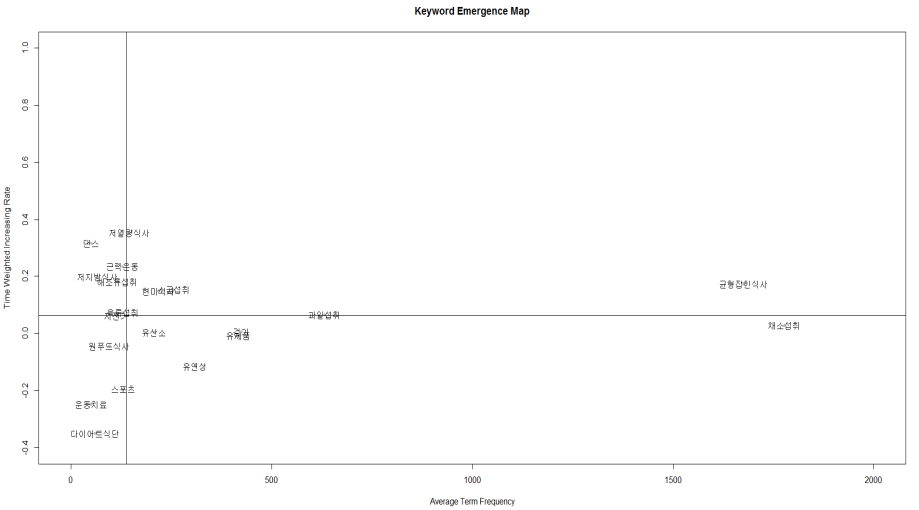


그림 Ⅲ-2 비만(다이어트) 관련 키워드 KEM

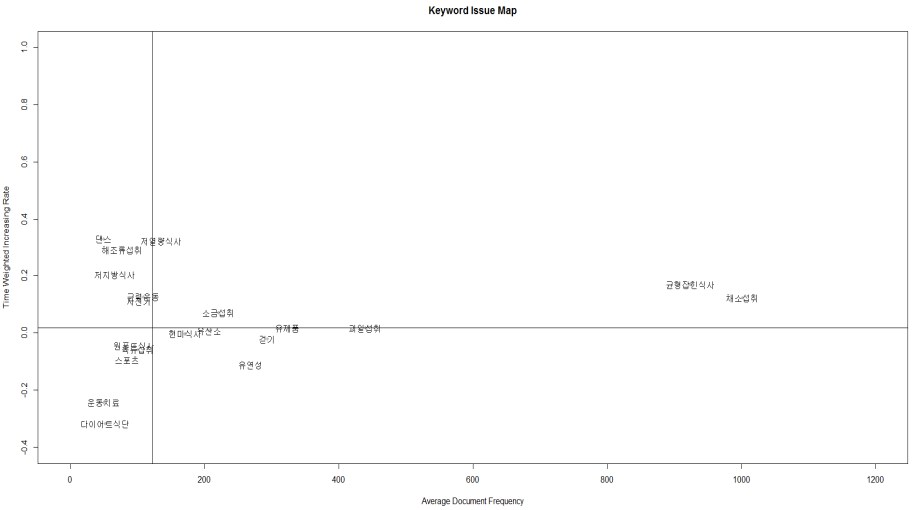


그림 Ⅲ-3 비만(다이어트) 관련 키워드 KIM

표 Ⅲ-6 비만(다이어트) 관련 키워드의 미래신호

구분	잠재신호 (Latent signal)	약신호 (Weak Signal)	강신호 (Strong signal)	강하지만 증가율이 낮은 신호 (Strong but low increasing signal)
KEM	원푸드식사, 스포츠, 운동치료, 다이어트식단	댄스, 근력운동, 저지방식사, 해조류섭취, 육류섭취, 자전거	균형잡힌식사, 과일섭취, 소금섭취, 현미식사, 저열량식사	채소섭취, 걷기, 유산소, 유연성, 유제품
KIM	원푸드식사, 육류섭취, 스포츠, 운동치료, 다이어트식단	댄스, 해조류섭취, 저지방식사, 근력운동, 자전거	채소섭취, 균형잡힌식사, 과일섭취, 유제품, 소금섭취, 저열량식사	걷기, 유산소, 유연성, 현미식사
주요 신호	원푸드식사, 스포츠, 운동치료, 다이어트식단	댄스, 근력운동, 저지방식사, 해조류섭취, 자전거	균형잡힌식사, 과일섭취, 소금섭취, 저열량식사	걷기, 유산소, 유연성

다음은 다이어트 이슈(다이어트 키워드를 이슈로 표기함)에 대한 감성분석 결과와 각 이슈의 평균 증가율을 교차하여 살펴본 것이다. 각 그림에서 1사분면은 긍정 감정(성공)과 증가율이 상대적으로 모두 높은 영역이다. DoD의 평균증가율과 감성분석 결과를 살펴보면 1사분면에 속하는 사례를 긍정 감정(다이어트 성공)이 높은 순서에 따라 나열하면, 저지방식사, 저열량식사, 해조류섭취, 근력운동, 유제품의 순이다. 2사분면은 긍정 감정(다이어트 성공)은 상대적으로 높고, 증가율이 상대적으로 낮은 영역이다. 다이어트식단, 현미식사, 유연성, 원푸드식사, 육류섭취가 2사분면에 속한다. 3사분면은 긍정 감정(다이어트 성공)과 증가율이 상대적으로 모두 낮은 영역이다. 여기에 속하는 키워드는 과일섭취, 유산소, 걷기, 스포츠, 운동치료이다. 4사분면은 긍정 감정(다이어트 성공)이 상대적으로 낮고, 증가율이 상대적으로 높은 영역이다. 여기에 속하는 키워드는 균형잡힌식사, 댄스, 채소섭취, 소금섭취, 자전거이다.

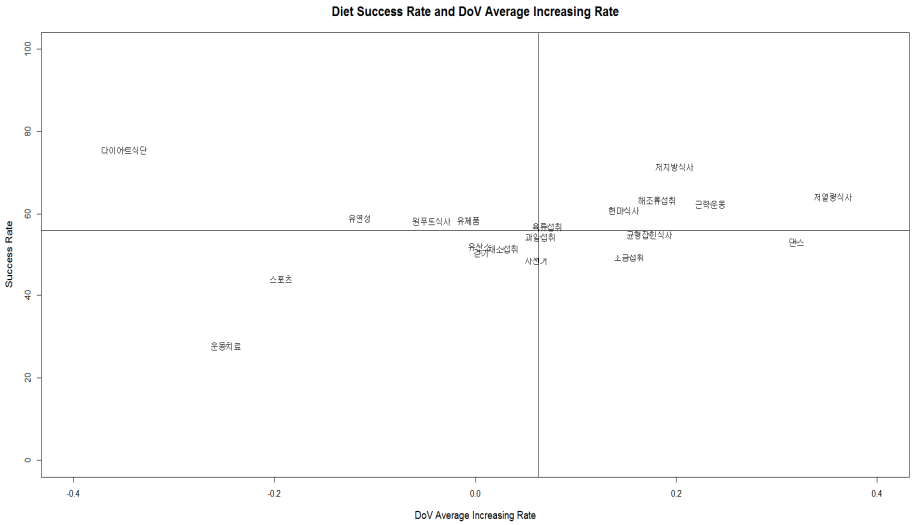


그림 Ⅲ-4 비만(다이어트) 주요 이슈에 대한 감성과 증가율(DoV 기준)

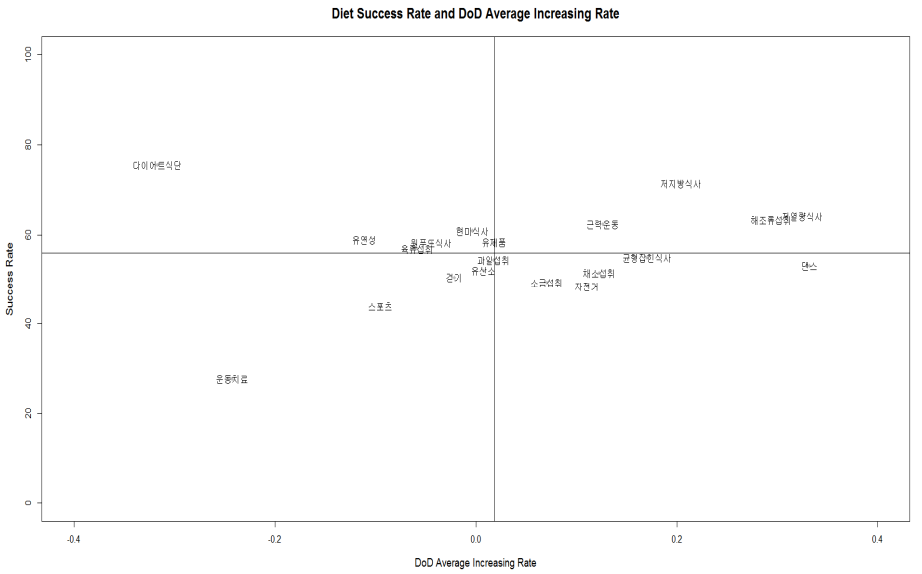


그림 Ⅲ-5 비만(다이어트) 주요 이슈에 대한 감성과 증가율(DoD 기준)

표 Ⅲ-7 비만(다이어트) 주요 이슈 관련 키워드의 감정 분석과 대응방향(DoD 기준)

구분	예상되는 리스크 관리	이슈 홍보 강화	이슈설계 점검	이슈설계 보완
	1/4 분면	2/4 분면	3/4 분면	4/4 분면
주요 키워드	저지방식사, 저열량식사, 해조류섭취, 근력운동, 유제품	다이어트식단, 현미식사, 유연성, 원푸드식사, 육류섭취	과일섭취, 유산소, 걷기, 스포츠, 운동치료	균형잡힌식사, 댄스, 채소섭취, 소금섭취, 자전거

주: 긍정 감정(성공)이 높은 순서로 배열한 것임.

4. 머신러닝을 활용한 다이어트 성공유무 예측 인공지능 개발

머신러닝을 활용한 다이어트 성공유무 예측 인공지능 개발 절차는 다음과 같다.

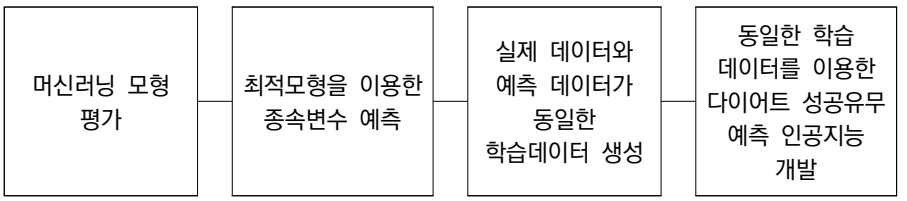


그림 Ⅲ-6 다이어트 성공유무 예측 인공지능 개발 절차

첫째, 지도학습 알고리즘을 이용하여 학습데이터(learning data)를 훈련데이터(training data)와 시험데이터(test data)로 학습하여 모형을 평가하여 최적모형을 선정한다. training data 와 test data를 5:5로 학습하여 모형을 평가한 결과 전체 데이터의 학습을 통한 모형의 평가결과 정확도와 민감도에서 서포트벡터머신 모형이 가장 우수한 것으로 나타났다.

표 Ⅲ-8 지도학습 머신러닝 알고리즘 평가(5:5)

Evaluation Index	Naïve Bayes Classification	neural networks	logistic regression	support vector machines	random forests	decision trees
accuracy	55.69	62.04	56.61	63.69	63.60	56.52
error rate	44.31	37.92	43.39	36.31	34.40	43.48
sensitivity	35.05	52.04	60.77	61.29	59.96	33.13
specificity	76.28	71.89	52.65	66.02	67.10	79.45
precision	59.58	64.60	55.06	63.65	63.72	61.25
AUC	0.59	0.66	0.60	0.66	0.69	0.57
			best accuracy		support vector machines	
			best error rate		support vector machines	
			best sensitivity		support vector machines	
			best specificity		decision trees	
			best precision		neural networks	
			best AUC (Area Under the Curve)		random forests	

둘째, 선정된 최적모델인 서포트벡터머신 모형을 이용하여 실제데이터의 독립변수만으로 종속변수를 예측한다. 셋째, 실제데이터의 종속변수와 예측데이터의 종속변수가 동일한 학습데이터를 생성한다. 실제데이터와 예측데이터의 다이어트 성공유무 교차분석 결과는 표와 같다.

표 Ⅲ-9 실제 데이터와 예측데이터의 빈도분석

단위: n(%)

실제 데이터			예측 데이터		
Failure	Success	Total	Failure	Success	Total
2,292(50.34)	2,261(49.66)	4,553	1,693(53.89)	1,451(46.11)	3,147

넷째, 실제데이터와 예측데이터가 동일한 학습데이터(3,147건)와 랜덤포레스트 모델을 이용하여 다이어트 성공유무를 예측할 수 있는 인공지능을 개발한다. 인공지능(랜덤포레스트 예측 모델)의 다이어트 성공유무를 예측(성공, 실패)하는데 가장 큰 영향을 미치는 입력변수는 채소섭취로 나타났으며 성공 예측확률은 46.12%로 나타났다. 그리고 다이어트 성공유무 예측모델의 설명력인 결정계수(R^2, Var explained)는 86.15%로 나타났다.

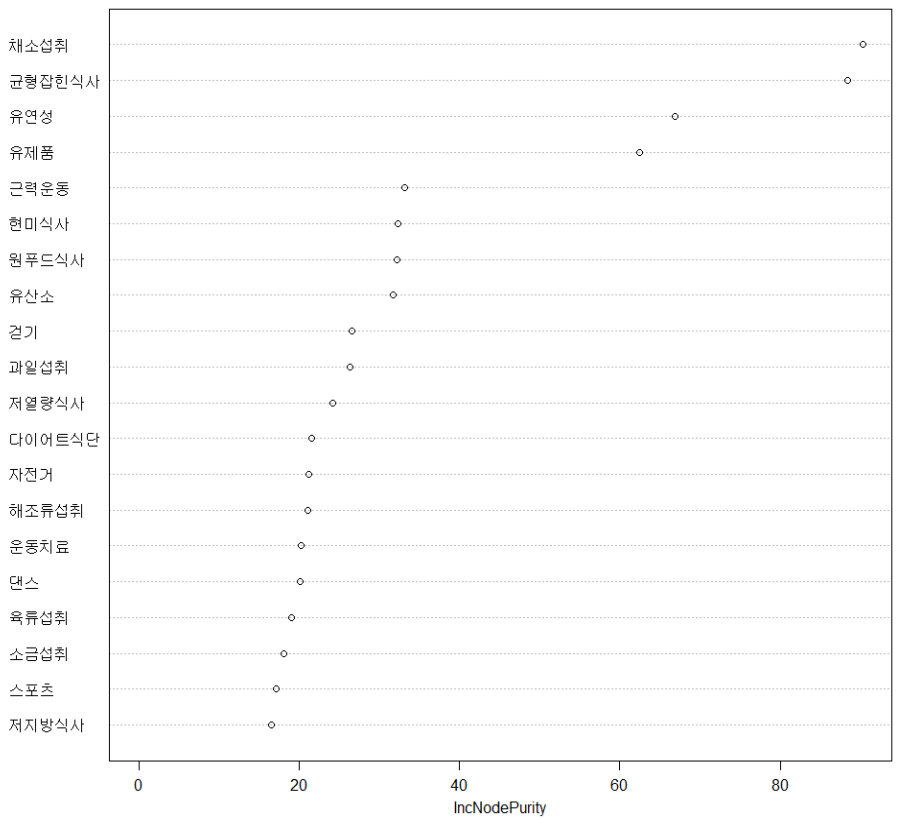


그림 Ⅲ-7 다이어트 성공유무 예측 모델(랜덤포레스트 모델)

서포트벡터머신 모형이 예측한 데이터를 이용하여 독립변수 간의 연관분석 결과를 살펴보면, 온라인 문서에서 {걷기, 유산소, 현미식식사}가 있을 경우 경우 {균형잡힌식사}를 할 경우가 1.79배 높은 것으로 나타났다. 그리고 {유연성운동, 근력운동, 과일섭취}가 있을 경우 {채소섭취}를 할 경우가 1.68배 높은 것으로 나타났다. 다이어트 성공과 독립변수 간의 연관분석에서 온라인 문서에서 {유산소, 저지방식사, 균형잡힌식사, 과일섭취}의 경우 {유산소, 저지방식사, 균형잡힌식사, 과일섭취}가 없을 때 보다 다이어트 성공확률이 2.17배 높은 것으로 나타났다. {걷기, 저지방식사, 균형잡힌식사}가 있는 경우 없을 때 보다 다이어트 성공확률이 2.11배 높은 것으로 나타났다.

표 Ⅲ-10 독립변수 간 연관규칙

rules	support	confidence	lift	count
{해조류섭취,육류섭취} => {균형잡힌식사}	0.024	1	1.793	75
{현미식사,육류섭취,소금섭취} => {균형잡힌식사}	0.025	1	1.793	80
{현미식사,유제품,소금섭취} => {균형잡힌식사}	0.029	1	1.793	92
{걷기,유산소,현미식사} => {균형잡힌식사}	0.024	1	1.793	74
{현미식사,육류섭취,유제품,소금섭취} => {균형잡힌식사}	0.023	1	1.793	71
{현미식사,채소섭취,육류섭취,소금섭취} => {균형잡힌식사}	0.025	1	1.793	80
{현미식사,과일섭취,유제품,소금섭취} => {균형잡힌식사}	0.024	1	1.793	77
{현미식사,채소섭취,유제품,소금섭취} => {균형잡힌식사}	0.029	1	1.793	91
{현미식사,채소섭취,육류섭취,유제품,소금섭취} => {균형잡힌식사}	0.023	1	1.793	71
{현미식사,채소섭취,과일섭취,유제품,소금섭취} => {균형잡힌식사}	0.024	1	1.793	77
{유연성,해조류섭취} => {채소섭취}	0.023	1	1.684	71
{유연성,근력운동,과일섭취} => {채소섭취}	0.023	1	1.684	71
{현미식사,육류섭취,소금섭취} => {채소섭취}	0.025	1	1.684	80
{육류섭취,유제품,소금섭취} => {채소섭취}	0.029	1	1.684	92
{과일섭취,육류섭취,소금섭취} => {채소섭취}	0.029	1	1.684	90
{현미식사,육류섭취,유제품} => {채소섭취}	0.033	1	1.684	105
{현미식사,과일섭취,육류섭취} => {채소섭취}	0.037	1	1.684	118
{유연성,과일섭취,유제품} => {채소섭취}	0.024	1	1.684	77

rules	support	confidence	lift	count
{균형잡힌식사,현미식사,해조류섭취,과일섭취} => {채소섭취}	0.025	1	1.684	79
{현미식사,육류섭취,유제품,소금섭취} => {채소섭취}	0.023	1	1.684	71
{균형잡힌식사,현미식사,육류섭취,소금섭취} => {채소섭취}	0.025	1	1.684	80
{과일섭취,육류섭취,유제품,소금섭취} => {채소섭취}	0.025	1	1.684	79
{균형잡힌식사,육류섭취,유제품,소금섭취} => {채소섭취}	0.029	1	1.684	91
{균형잡힌식사,과일섭취,육류섭취,소금섭취} => {채소섭취}	0.028	1	1.684	89
{현미식사,과일섭취,육류섭취,유제품} => {채소섭취}	0.030	1	1.684	94
{균형잡힌식사,현미식사,육류섭취,유제품} => {채소섭취}	0.031	1	1.684	96
{균형잡힌식사,현미식사,과일섭취,육류섭취} => {채소섭취}	0.034	1	1.684	106
{현미식사,과일섭취,유제품,소금섭취} => {채소섭취}	0.024	1	1.684	77
{유연성,균형잡힌식사,과일섭취,유제품} => {채소섭취}	0.023	1	1.684	71

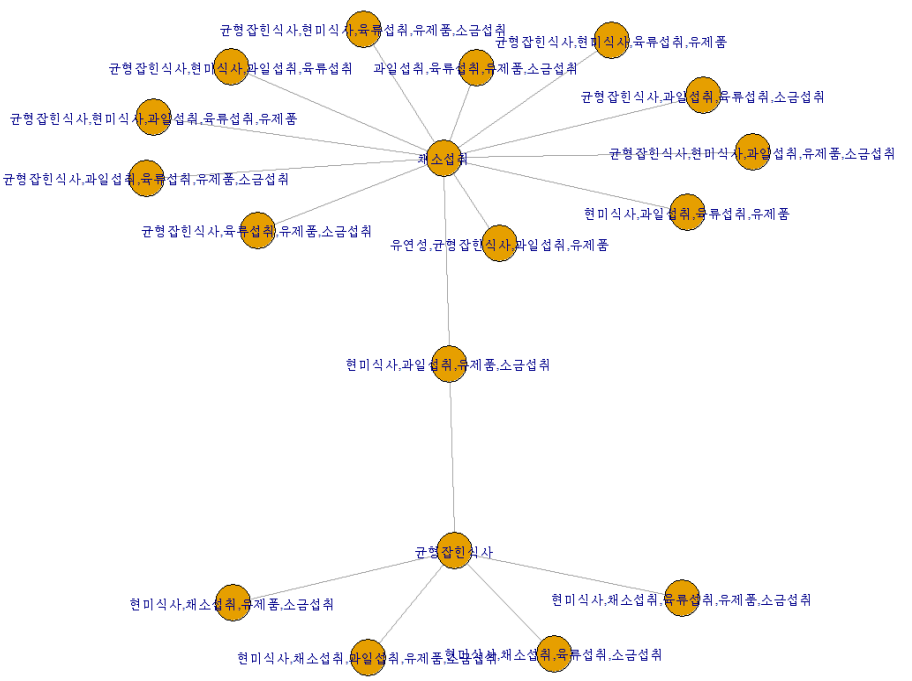


그림 Ⅲ-8 독립변수 간 연관규칙의 시각화

표 Ⅲ-11 다이어트 성공과 독립변수 간 연관규칙

rules	support	confidence	lift	count
{유산소,저지방식사,균형잡힌식사,과일섭취}=>{Success}	0.017	1.000	2.169	54
{유산소,저지방식사,균형잡힌식사,채소섭취,과일섭취}=>{Success}	0.017	1.000	2.169	54
{저지방식사,균형잡힌식사,채소섭취,유제품}=>{Success}	0.020	0.984	2.135	63
{유산소,현미식사,육류섭취}=>{Success}	0.018	0.982	2.131	56
{유산소,현미식사,채소섭취,육류섭취}=>{Success}	0.018	0.982	2.131	56
{유산소,균형잡힌식사,현미식사,육류섭취}=>{Success}	0.017	0.982	2.130	55
{유산소,균형잡힌식사,현미식사,채소섭취,육류섭취}=>{Success}	0.017	0.982	2.130	55
{유산소,저지방식사,과일섭취}=>{Success}	0.017	0.982	2.129	54
{유산소,저지방식사,채소섭취,과일섭취}=>{Success}	0.017	0.982	2.129	54
{걷기,저지방식사,균형잡힌식사}=>{Success}	0.022	0.971	2.107	68
{걷기,저지방식사,균형잡힌식사,채소섭취}=>{Success}	0.021	0.971	2.105	66
{유산소,저지방식사,균형잡힌식사}=>{Success}	0.020	0.969	2.101	62
{유산소,저지방식사,균형잡힌식사,채소섭취}=>{Success}	0.019	0.968	2.099	60
{걷기,저지방식사}=〉 {Success}	0.022	0.958	2.077	68
{걷기,저지방식사,채소섭취}=>{Success}	0.021	0.957	2.075	66
{유산소,저지방식사}=>{Success}	0.020	0.954	2.069	62
{다이어트식단,균형잡힌식사,채소섭취}=>{Success}	0.026	0.953	2.067	81
{유산소,저지방식사,채소섭취}=>{Success}	0.019	0.952	2.066	60
{저지방식사,균형잡힌식사,현미식사}=>{Success}	0.017	0.948	2.057	55
{저열량식사,균형잡힌식사,과일섭취,유제품}=>{Success}	0.021	0.942	2.043	65
{다이어트식단,균형잡힌식사}=>{Success}	0.031	0.941	2.041	96
{저열량식사,균형잡힌식사,채소섭취,과일섭취,유제품}=>{Success}	0.019	0.938	2.035	61
{저열량식사,균형잡힌식사,유제품}=>{Success}	0.029	0.938	2.033	90
{저지방식사,균형잡힌식사,유제품}=>{Success}	0.022	0.932	2.022	69
{저지방식사,균형잡힌식사,과일섭취}=>{Success}	0.025	0.929	2.016	79
{저지방식사,균형잡힌식사,채소섭취,과일섭취}=>{Success}	0.025	0.929	2.014	78
{저열량식사,균형잡힌식사,채소섭취,유제품}=>{Success}	0.023	0.923	2.002	72
{다이어트식단,채소섭취}=>{Success}	0.030	0.921	1.997	93
{저지방식사,현미식사}=>{Success}	0.018	0.919	1.994	57
{저지방식사,균형잡힌식사,채소섭취}=>{Success}	0.036	0.918	1.991	112
{저지방식사,채소섭취,유제품}=>{Success}	0.021	0.903	1.958	65
{저열량식사,균형잡힌식사,채소섭취,과일섭취}=>{Success}	0.032	0.902	1.956	101
{저지방식사,과일섭취,유제품}=>{Success}	0.017	0.900	1.952	54
{저열량식사,균형잡힌식사,과일섭취}=>{Success}	0.034	0.900	1.952	108

5. 시사점

비만은 당뇨병, 고혈압, 관절염 등 만성질환 위험 증가를 초래하여 조기 사망의 원인이 될 수 있다. 많은 사람들은 이러한 비만을 해결하기 위해 인터넷과 소셜미디어를 이용하여 비만에 관한 정보를 얻고 공유하며 의견을 교환한다. 그리고, 다이어트 성공을 위해 본인에게 적합한 다이어트 요법을 추천받고 있다. 이에, 소셜미디어 상에서 생성되고 있는 비만(다이어트)과 관련된 빅데이터를 수집하여 분석한다면, 비만과 관련된 대중의 인식과 대처방법, 사회현상을 보다 실제적으로 파악하여 본인에게 적합한 다이어트 요법을 제시할 수 있다. 기존의 연구는 이론적 모형에 근거한 바, 비만에 영향을 미치는 다양한 변인 간의 관계를 파악하는 데는 한계가 있다. 반면, 머신러닝 분석 방법은 머신러닝 알고리즘이 데이터를 학습하여 모형을 제시하기 때문에 비만에 영향을 미치는 다양한 요인을 발견할 수 있다.

이에 본 연구에서는 트위터를 비롯한 207개의 온라인 채널에서 비만(다이어트)과 관련된 온라인 문서를 수집하여 주제분석과 감성분석을 통해 독립변수를 분류하고 청소년 비만(다이어트)과 관련하여 나타나는 신호(운동, 식이)를 탐지하여 예측모형을 제시하고자 하였다. 본 연구에서 사용된 학습데이터는 2011. 1. 1~2013. 12. 31까지 비만과 관련하여 수집된 120만 7,531건의 텍스트(Text) 문서 중 '소아, 아동, 어린이, 학령전기, 학령기, 청소년, 10대, 학생, 초·중·고등학생'의 키워드가 포함된 문서를 청소년 문서(11,963건)로 분류하여 연구대상으로 하였다. 비만(다이어트)과 관련한 미래신호를 탐색하기 위해 독립변수를 대상으로 단어빈도와 문서빈도 그리고 중요도 지수를 나타내는 TF-IDF를 분석하였다. 그리고 키워드의 중요도 지수를 나타내는 KEM과 확산정보를 나타내는 KIM을 분석하여 미래신호를 탐색하였다. 머신러닝 분석기술을 이용하여 탐색된 미래신호를 중심으로 다이어트 성공유무를 예측할 수 있는 인공지능을 개발하고, 종속변수와 독립변수들 간의 연관관계를 파악하였다. 본 연구의 결과를 살펴보면 다음과 같다.

첫째, 청소년 비만(다이어트) 관련 운동 요인에 대한 성공 감정은 식이요인 55.6%, 운동요인 52.4%로 나타났다.

둘째, 운동 요인의 키워드의 중요성을 나타내는 단어빈도에서는 자전거 운동은 중요하게 평가되지 않으나 해당 주제에 대한 확산 정도인 문서빈도는 높게 나타나, 자전거 운동에 대한 청소년의 관심이 높아지는 것으로 보인다. 반면 스포츠는 중요한 키워드임에도 확산은 낮아 스포츠에 대한 청소년의 관심이 낮아지는 것으로 보인다

셋째, 채소섭취, 걷기, 유제품, 유연성운동, 유산소운동은 단어빈도는 높으나 DoV의 증가율의 중앙값이 낮게 나타나 채소섭취, 걷기, 유제품, 유연성운동, 유산소운동 대한 청소년의 관심을 높일 수 있는 방안이 마련되어야 할 것이다. 걷기, 유연성운동, 현미식사는 문서빈도는 높으나 DoD의 증가율의 중앙값이 낮게 나타나 걷기, 유연성운동, 현미식사에 대한 청소년의 관심을 높일 수 있는 방안이 마련되어야 할 것이다. 특히, 걷기운동은 단어빈도와 문서빈도에서 높게 나타났지만 증가율은 낮게 나타나 걷기운동에 대한 청소년의 관심을 높일 수 있는 방안이 마련되어야 할 것이다.

넷째, 약신호(2사분면)에는 댄스, 근력운동, 저지방식사, 해조류섭취, 자전거가 포함된 것으로 나타났으며, 특히 해조류섭취는 높은 증가율을 보이고 있어 해조류섭취 키워드는 시간이 지날수록 강한 신호로 발전해 갈 수 있다. 따라서 이에 대한 과학적 근거와 다양한 식단 마련이 필요할 것으로 본다.

다섯째, 다이어트 주요 이슈에 대한 감성과 DoD 증가율을 분석한 결과 저지방식사, 저열량식사, 해조류섭취, 근력운동, 유제품 이슈의 경우 리스크의 관리가 주요한 대응 방안이 될 수 있다. 다이어트식단, 현미식사, 유연성, 원푸드식사, 육류섭취 키워드는 이슈의 홍보를 강화하는 것이 청소년 등의 이슈에 대한 이해도를 높이는 방안이 될 수 있다. 과일섭취, 유산소, 걷기, 스포츠, 운동치료 키워드는 이슈에 대한 설계에 있어 잘못된 부분이 없었는지에 대해 다시 점검할 필요가

있다. 균형잡힌식사, 댄스, 채소섭취, 소금섭취, 자전거 키워드는 이슈가 다이어트 실패에 대한 일반국민의 반감을 가지고 올 수 있기 때문에 이슈의 설계를 보완하고 다시 점검할 필요가 있을 것이다.

여섯째, training data 와 test data를 5:5로 학습하여 모델을 평가한 결과 정확도와 민감도에서 서포트벡터머신 모형이 가장 우수한 것으로 나타났다. 서포트벡터머신 모형을 이용하여 실제데이터의 독립변수만으로 종속변수를 예측하고, 실제데이터의 종속변수와 예측데이터의 종속변수가 동일한 학습데이터를 생성한 결과, 예측 데이터의 다이어트 성공은 53.89%로 나타났다.

일곱째, 랜덤포레스트 예측 모형이 다이어트 성공유무를 예측(성공, 실패)하는데 가장 큰 영향을 미치는 입력변수는 채소섭취로 나타났으며 성공 예측확률은 46.12%로 나타났다.

마지막으로 다이어트 실패와 독립변수 간의 연관성 예측에서 {유산소, 저지방식사, 균형잡힌식사, 과일섭취}의 경우 다이어트 성공 확률은 2.17배 높은 것으로 나타났으며, {걷기, 저지방식사, 균형잡힌식사}의 경우 다이어트 실패 확률이 2.11배 높은 것으로 나타났다.

본 연구에서 제시하는 정책제언과 제한점은 다음과 같다.

첫째, 온라인 채널에서 청소년들이 언급하는 비만(다이어트)과 관련한 용어는 이론적 배경하에 분류된 온톨로지의 전문용어도 사용하지만 온라인 채널 이용시점에서 자주 사용하는 구어체나 속어를 사용할 수 있기 때문에 비만(다이어트) 온톨로지는 용어의 추가 등 수정·보완이 지속적으로 이루어져야할 것이다.

둘째, 인공지능 개발을 위해 머신러닝 모형에 사용된 데이터에 대한 지속적인 개선(update)이 필요하다. training data를 학습하여 분석된 머신러닝 모형은 test data로 실행했을 때, 실제의 분류와 예측의 분류는 다르게 나타날 수 있다. 따라서 모형의 예측률을 높이기 위해서는 본고에서 제시한 실제 분류와 예측분류가 동일한 케이스만 선택(selection)하여 양질의 학습데이터로 생성한 후, 지속적

으로 추가하게 되면 이들 양질의 학습데이터를 다시 학습하게 될 경우, 우수한 다이어트 예측모형(인공지능)이 개발 될 것으로 본다. 끝으로 본 연구의 결과가 챗봇 등으로 개발되어 실질적으로 적용된다면 개인별 맞춤형 다이어트 요법의 실시간 제공이 가능할 것으로 본다.

본 연구의 제한점은 다음과 같다.

첫째, 본 연구의 주제분석에서 사용된 비만(다이어트)관련 용어는 기존의 오프라인 조사 등을 통한 비만(다이어트)관련 용어의 조작적 정의와 다를 수 있다. 둘째, 본 연구에서 적용된 감성분석은 감성어 사전을 사용하여 감성어 사전에 포함된 문장이 있을 경우만 긍정과 부정으로 분류되었기 때문에 문서 전체의 맥락을 파악하여 긍정과 부정을 분류한 것이 아니다. 따라서 향후 문서 전체의 맥락을 파악하여 감성분석을 할 수 있는 기술 개발이 이루어 져야 할 것이다.

참고문헌

참고문헌

〈국내문헌〉

안혜영, 임숙빈, 홍경자, 허명행 (2007). 학령기 아동을 위한 멀티에이전트 비만관리 프로그램의 효과. **한국간호학회. 37**(1), 105~113

안근희, 임미자, 이혜진, 김권범, 한경아, 민경완 (2004). 비만한 제2형 당뇨병 환자의 운동요법. **대한비만학회, 13**(3), 39-47

정영호, 고숙자, 임희진 (2010). 청소년 비만의 사회경제적 비용. **보건사회연구. 30**(1), 195-219.

이영숙 (2011). 비만도 및 생활습관에 미치는 영향. **인하대학교 스포츠과학연구소, 23,** 73-88.

〈외국문헌〉

Ashrafian H. Toma T, Harling L, Kerr K, Athanasiou T, Darzi A. (2014). Social networking staregies that aim to reduce obesity have achieved significant although modest results. *Health affairs, 33*(9), 1641-1647.

Blümel JE, Chedraui P, Aedo S, Fica J, Mezones-Holguín E, et al. (2015). Obesity and its relation to depressive symptoms and sedentary lifestyle in middle-aged women, *Maturitas, 80*(1), 100-105.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 26,* 123-140.

Breiman, L. (2001). Random forest. *Machine learning, 45*(1), 5-32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *machine*

Learning, 20, 273-297.

Cowley MA, Brown WA, Considine RV. (2016). *Endocrinology: Adult & Pediatric*(Obesity: The problem and Its Management). Elsevier, 468-478.

Doll HA, Petersen SE, Stewart-Brown SL. (2000). Obesity and physical and emotional well-being: associations between body mass index, chronic illness, and the physical and mental components of the SF-36 questionnaire. *Obesity a Research Journal, 8*(2), 160-170.

Edwardson CL, Gorely T, Davies MJ, Gray LJ, Khunti K, Wilmot EG, Yates T, Biddle SJ (2012). Association of Sedentary Behaviour with Metabolic Syndrome: A Meta-Analysis. *PLoS One, 7*(4), 1-5.

Egger G. Swinburn B. (1997). An ecological approach to the obesity pandemic. *BMJ, 315*, 477-480.

Flegal KM, Troiano RP. (2000). Changes in the distribution of body mass index of adults and children in the US population. *Int J Obes Relat Metab Disord, 24*(7), 807-818.

Freedland SJ, Aronson WJ. (2005). Obesity and prostate cancer. *Urology 65*(3), 433-439.

Hassouna, M., Tarhini, A., Elyas, T. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. *International Business Research, 8*(6), 224-237.

Hiltunen, E. (2008). The future sign and its three dimensions. *Futures 40,* 247-260.

Jennifer SL, Tracie AB, Elizabeth G, Richard CW, Alex RK. (2013). Approach to the prevention and management of childhood obesity: The role of social networks and the use of social madia and related

electronic technologies. a scientific statement from the american heart association, *AHA, 127,* 260-267.

Kent, E. E., Prestin, A., Gaysynsky, A., Galica, K., Rinker, R., Graff, K., & Chou, W. Y. S. (2016). "Obesity is the new major cause of cancer": connections between obesity and cancer on Facebook and Twitter. Journal of Cancer Education, 31(3), 453-459.

Kim AR, Park HA, Song TM. (2017). Development and Evaluation of an Obesity Ontology for Social Big Data Analysis. *Health Inform Res, 23*(3), 159-168. https://doi.org/10.4258/hir.2017.23.3.159.

Kim KW, Kim YA, Kim JH. (1997). A study of the Obesity Index and Psychosocial Factors Influencing Obesity among Adolescent Girls. *Korean J Community Nutrition, 2*(4), 496-504.

Kim HY, Park HA, Min YH, Jeon E. (2013). Development of an obesity management ontology based on the nursing process for the mobile-device domain. *J Med Internet Res, 15*(6), e130. doi: 10.2196/jmir.2512.

Lowry R, Wechsler H, Galuska DA, Fulton JE, Kann L. (2002). Television viewing and its associations with overweight, sedentary lifestyle, and insufficient consumption of fruits and vegetables among US high school students: differences by race, ethnicity, and gender. *The Journal of school health, 72*(10), 413-421.

Luppino FS, Wit LM, Bouvy PF, et al. (2010). Overweight, Obesity, and Depression: A Systematic Review and Meta-analysis of Longitudinal Studies. *Arch Gen Psychiatry, 67*(3), 220-229.

Marjolijn L, Anteheunis, Kiek Tates, Theodoor E, Nieboer. (2013). Patients' and health professionals' use of social media in health care:

Motives. *barriers and expectations, 92*(3), 426-431.

Minsky, M., & Papert, S. (1969). *Perceptrons.* MIT Press, Cambridge.

Na S, Park IH, Kim HN. (2014). Development of Mobile Service and User Evaluation Method for Teenager Obesity Management, 2015 *HCI(Human Computer Interaction) Conference,* 120-124.

Nationals Heart Lung and Blood Institute. (1998). Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults, - The Evidence Report, Report No(98-4083).

Park, H. C. (2013). Proposition of causal association rule thresholds. *Journal of the Korean Data & Information Science Society, 24*(6), 1189-1197.

Perichart PO, Balas NM, Schiffman SE, Barbato DA, Vadillo OF. (2007). Obesity increases metabolic syndrome risk factors in school-aged children from an urban school in Mexico city. *Journal of the American Dietetic Association, 107*(1), 81-91.

Rumelhart DE, GHinton GE, Williams RJ. (1986). Learning representations by back-propagating errors. *Nature, 323,* 533-536.

Schulz AJ, Zenk S, Angela OY, Teretha, et al.(2005). Healthy eating and exercising to reduce diabetes: Exploring the potential of social determinants of health frameworks within the context of community-based participatory diabetes prevention. *American Journal of Public Health, 95*(4), 645-651.

Spärck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". *Journal of Documentation, 28,* 11-21.

Truman BI, Smith CK, Roy K, et al. (2011). Relational for regular reporting on health disparities and inequalities - United States.

Morbidity and Mortality Weekly Report(MMWR), 60(1), 3-10.

Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Journal Expert Systems with Applications, 39*(16), 12543-12550.

〈웹사이트 및 홈페이지 자료〉

Biological Neural Network. https://commons.wikimedia.org/wiki/File:Artificial_neural_network.png에서 2020년 4월 29일 인출.

National Health Insurance Service. Available at http://health.chosun.com/news/dailynews_view.jsp?mn_idx=281825에서 2019년 7월 15일 인출.

Support Vector Machines, https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf에서 2020년 4월 29일 인출.

ABSTRACT

The objectives of this study are as follows: 1) to collect texts related to the search of adolescents on the basis of social big data, and detect future signals on adolescent obesity through topic analysis and sentiment analysis; and 2) to develop a model for predicting the success or failure of weight loss in adolescents through the use of machine learning technology. Topics related to obesity were collected every hour of every day from online channels, and a total of 1,207,531 online texts were analyzed. Collected texts containing keywords, such as adolescents, students, infants, preschoolers, children, pre-school age, middle school students, elementary school students, school age, teens and high school students, were classified as youth texts (11,963) for research. Crawlers were used to collect social big data, and 'Obesity' and 'Diet' were used as obesity topics to collect all related texts. The main research findings are as follows. First, the feelings of success for adolescent obesity-related exercise factors were dietary factors (55.6%) and exercise factors (52.4%). Second, the analysis of sentiments on major diet issues and the increase rate of DoD found that low-fat diets, low-calorie diets, seaweed intake, muscular strength, and the risk management of dairy issues can be major countermeasures. Third, the prediction of a correlation between weight loss failure and independent variables revealed that the probability of weight loss success is 2.17 times higher in the case of aerobics, low-calorie diets, well-balanced diets and fruit intake. Based on the research results, improvement methods for analyzing unstructured big data were proposed in this study.

2020년 한국청소년정책연구원 발간자료 목록

기관고유과제

20-R01 청소년의 혐오표현 노출실태 및 대응 방안 연구 / 김영한·이유진·조아미·임성택

20-R02 청소년·청년 디지털 플랫폼노동 실태 및 대응방안 연구 / 유민상·최정원·이수정·장혜림

20-R03 Z세대 10대 청소년의 가치관 변화 연구 / 오해섭·문호영

20-R04 청소년활동의 사회적 가치 제고방안 연구 / 최용환·성유리·박윤수·김보경

20-R05 청소년지도자 양성시스템 재구축방안 연구 Ⅱ : 청소년상담사를 중심으로 / 최창욱·좌동훈·이종원·남화성·정지희

20-R06 청년 사회·경제실태 및 정책방안 연구Ⅴ / 김형주·연보라·배정희

20-R06-1 청년 사회·경제실태 및 정책방안 연구Ⅴ - 전년도 공개데이터 심층분석연구 / 김형주·연보라·배정희

20-R06-2 청년 사회·경제실태 및 정책방안 연구Ⅴ - 기초분석보고서 / 김형주·연보라·배정희

20-R07 국가 미래인적자원으로서 재외동포청소년 성장과 지원방안 연구 Ⅲ / 김경준·김정숙·윤철경

20-R08 청소년 빅데이터 체계 구축 및 활용방안 연구 / 서정아·성윤숙·송태민

20-R08-1 청소년 빅데이터 체계 구축 및 활용방안 연구 데이터 분석 보고서 1_청소년 비만에 대한 비정형 빅데이터 연구 / 송태민

20-R08-2 청소년 빅데이터 체계 구축 및 활용방안 연구 데이터 분석 보고서 2_청소년 행복 결정요인에 대한 정형 빅데이터 연구 / 홍성효

20-R09 청소년의 주거권 실태와 보장방안 연구: 사회배제 관점을 중심으로 / 김지연·김승경·임세희·최은영

20-R10 2020 아동·청소년 권리에 관한 국제협약 이행 연구 - 한국 아동·청소년 인권실태 : 총괄보고서 / 김영지·황세영·최홍일·이민희·김진호

20-R10-1 2020 아동·청소년 권리에 관한 국제협약 이행 연구 - 한국 아동·청소년 인권실태 : 심화분석보고서 / 박환보·주경필

20-R10-2 2020 아동·청소년 권리에 관한 국제협약 이행 연구 - 한국 아동·청소년 인권실태 : 기초분석보고서 / 김영지·황세영·최홍일

20-R11 지역사회 네트워크를 활용한 청소년 성장지원 정책 추진체계 구축 방안 연구 Ⅱ / 최인재·강경균·송민경·조윤정·김가희

20-R12 2020 한국아동·청소년패널조사 : 사업보고서 / 하형석·황진구·김성은·이용해

20-R12-1 2020 한국아동·청소년패널조사 : 데이터분석보고서 / 김성은·황영식

20-R13 2020 다문화청소년 종단연구 : 총괄보고서 / 양계민·장윤선·정윤미
20-R13-1 2020 다문화청소년 종단연구 : 기초분석보고서: 1기패널 / 양계민·장윤선·정윤미
20-R13-2 2020 다문화청소년 종단연구 : 기초분석보고서: 2기패널 / 양계민·장윤선·정윤미

협동연구과제

경제·인문사회연구회 협동연구총서 20-82-01 학교 밖 청소년 지역사회 지원방안 연구 Ⅲ : 질적패널 조사를 중심으로 / 김희진·장근영·이동훈·윤철경 (자체번호 20-R14)
경제·인문사회연구회 협동연구총서 20-82-02 학교 밖 청소년 지역사회 지원방안 연구 Ⅲ : 학교 밖 여성 청소년을 중심으로 / 오은진·장희영 (자체번호 20-R14-1)
경제·인문사회연구회 협동연구총서 20-83-01 청년 핵심정책 대상별 실태 및 지원방안 연구 Ⅲ : 청년 이직자 - 총괄보고서 / 김기헌·신동훈·변금선·고혜진·신인철 (자체번호 20-R15)
경제·인문사회연구회 협동연구총서 20-83-02 청년 핵심정책 대상별 실태 및 지원방안 연구 Ⅲ : 청년 이직자 - 심층분석보고서 / 김기헌·신동훈·고혜진·신인철 (자체번호 20-R15-1)
경제·인문사회연구회 협동연구총서 20-84-01 청소년의 건강권 보장을 위한 정책방안 연구 Ⅱ : 위기청소년 / 백혜정·임희진·송미경·김양희 (자체번호 20-R16)
경제·인문사회연구회 협동연구총서 20-84-02 청소년의 건강권 보장을 위한 정책방안 연구 Ⅱ : 위기청소년-국내·외 주요 법령 및 건강정책 분석 / 류정희·이상정·박선영·전민경 (자체번호 20-R16-1)
경제·인문사회연구회 협동연구총서 20-84-03 청소년의 건강권 보장을 위한 정책방안 연구 Ⅱ : 위기청소년 - 기초분석보고서 / 백혜정·임희진 (자체번호 20-R16-2)
경제·인문사회연구회 협동연구총서 20-85-01 청소년 미디어 이용 실태 및 대상별 정책대응방안 연구 Ⅰ : 초등학생 / 배상률·이창호·이정림 (자체번호 20-R17)
경제·인문사회연구회 협동연구총서 20-85-02 청소년 미디어 이용 실태 및 대상별 정책대응방안 연구 Ⅰ : 초등학생 - 해외사례 조사 / 정현선·심우민·윤지원·김광희·최원석 (자체번호 20-R17-1)
경제·인문사회연구회 협동연구총서 20-85-03 청소년 미디어 이용 실태 및 대상별 정책대응방안 연구 Ⅰ : 초등학생 - 기초분석보고서 / 배상률·이창호 (자체번호 20-R17-2)

연구개발적립금

20-R24 위기청소년 현황 및 실태조사 기초연구 / 황여정·이정민
20-R24-1 위기청소년 현황 및 실태조사 기초연구: 예비조사 데이터분석보고서 / 황여정·이정민·김수혜
20-R25 10대 청소년 포럼 운영 / 모상현
20-R26 코로나 19 확산 및 이후 사회변화에 따른 청소년정책의 대응방안 / 김현철

수 시 과 제

20-R18 학생 인권교육을 위한 현장 실천 강화 방안 연구 / 이정민·이종원
20-R19 청년정책 평가체계 구축을 위한 기초연구 / 배정희·김기헌
20-R20 가출청소년 지원 강화를 위한 청소년복지시설 재구조화 연구 / 황진구·김지연
20-R21 선거법 개정에 따른 청소년 정책 및 활동 지원 방안 연구 / 이창호
20-R22 청소년지도사 자격검정과목 개편 방안 연구 / 김경준·이종원·박정배
20-R23 학교밖청소년지원센터 운영모형 개발 / 김희진·백혜정

수 탁 과 제

20-R27 청소년 비즈쿨 활성화를 위한 고교교육 정책 연계방안 연구 / 강경균·안재영·황은희
20-R28 청년정책 현황 진단 및 정책추진 실효성 제고 방안 연구 / 김기헌·유민상·변금선·배정희
20-R29 (가칭)청양군 청소년재단 설립타당성 연구 / 김영한
20-R29-1 (가칭)청양군 청소년재단 설립타당성 연구(요약본) / 김영한
20-R30 2020년 학교폭력 예방교육 컨설팅 매뉴얼 / 성윤숙·양미석
20-R31 학교폭력 예방 어울림 기본 프로그램(5종)
20-R31-1 학교폭력 예방 어울림 기본 프로그램(초등학교 1~2학년용) /
성윤숙·서정아·장윤선·서고은·김성은
20-R31-2 학교폭력 예방 어울림 기본 프로그램(초등학교 3~4학년용) /
성윤숙·서정아·장윤선·서고은·김성은
20-R31-3 학교폭력 예방 어울림 기본 프로그램(초등학교 5~6학년용)/
성윤숙·서정아·장윤선·서고은·김성은
20-R31-4 학교폭력 예방 어울림 기본 프로그램(중학교용) / 성윤숙·서정아·장윤선·서고은·김성은
20-R31-5 학교폭력 예방 어울림 기본 프로그램(고등학교용) /
성윤숙·서정아·장윤선·서고은·김성은
20-R32 2019 학교폭력 예방 어울림 프로그램 적용효과 분석 / 성윤숙·양미석
20-R33 지역단위 학교폭력 예방교육 프로그램 운영현황 및 적용방안: 회복적 생활교육을
중심으로 / 성윤숙·양미석
20-R34 2018~2019 학교폭력 예방교육 연구학교 결과보고서
20-R34-1 2018~2019 학교폭력 예방교육 연구학교 결과보고서(초등학교) / 성윤숙·이윤소
20-R34-2 2018~2019 학교폭력 예방교육 연구학교 결과보고서(중학교) / 성윤숙·이윤소
20-R34-3 2018~2019 학교폭력 예방교육 연구학교 결과보고서(고등학교) / 성윤숙·이윤소
20-R35 교육과정 기반 어울림 프로그램 운영 사례 / 성윤숙·이윤소

20-R36 어울림 자유학기 프로그램(5종)
20-R36-1 어울림 자유학기 주제선택 프로그램(놀러와! 어울림 세상) / 성윤숙·이선희·정미선·선보라·이혜옥·이윤소
20-R36-2 자유학기 프로그램 동아리활동 / 성윤숙·배은정·이윤소
20-R36-3 어울림 자유학기 교과연계 주제선택 프로그램(사회) / 성윤숙·선보라·이윤소
20-R36-4 어울림 자유학기 교과연계 주제선택 프로그램(도덕) / 성윤숙·이혜옥·이윤소
20-R36-5 어울림 자유학기 교과연계 주제선택 프로그램(국어) / 성윤숙·이선희·정미선·이윤소
20-R37 외국의 학교폭력 예방교육 정책 및 프로그램 동향 / 성윤숙
20-R38 서울특별시 청소년시설 종합성과평가 평가지표 개선 용역 / 김형주·김혁진·김정주
20-R39 제1차 청년정책 기본계획 수립 연구 / 김기헌·유민상·변금선·배정희
20-R40 학교 안·밖 청소년정책 협력체계 구성 및 연계방안 연구 / 최창욱·좌동훈·남화성
20-R41 소년원학생 재범방지 통합프로그램 개발 연구 / 이유진
20-R41-1 꿈과 친구사이 : 소년원학생 재범방지 통합프로그램 매뉴얼 / 이유진
20-R42 청소년 주도적인 활동지원을 위한 법률개정 방안 연구 / 최창욱·문호영
20-R43 보호종료아동 주거지원 통합서비스 사업 사례관리사 및 자립업무종사자 직무분석 연구 / 김지연·백혜정·이상정
20-R44 2020년도 청소년 인터넷 건전이용제도 적용 게임물 평가 / 배상률·유홍식·김동일
20-R45 청소년 교육.문화공간 확충을 위한 지자체-교육청 협력 강화 방안 / 김영지·황세영·손진희·박명선·박종원·조기영
20-R46 2020년 청소년특별회의 정책과제 연구 / 최창욱·좌동훈
20-R47 농업인력 유입 확대를 위한 농촌 청소년 실태조사 / 오해섭·최홍일
20-R48 성남시 청소년시설 확충 및 개선 연구 / 황진구·남화성
20-R49 다함께돌봄센터 현장적용 프로그램 개발연구 / 황진구·좌동훈
20-R50 2020년 청소년수련시설종합평가 / 김경준·김영지·최창욱
20-R51 지역센터 관리운영을 위한 평가지표 개선방안 / 최용환·곽창규·이성규
20-R52 수원시 위기청소년 실태조사 연구 / 최용환·김보경
20-R53 2020 청소년방과후아카데미 효과·만족도 연구 / 양계민
20-R54 2020년 학교 내 대안교실 모니터링 연구 / 최인재·송원일·박지원
20-R55 2020년 대안교육 위탁교육기관 실태조사 연구 / 최인재·송원일·배수인
20-R56 복권기금 꿈사다리 장학생의 성장관리 및 성과측정 방안 연구 / 이정민·성유리·김혜원
20-R57 2020년 청소년 매체이용 및 유해환경 실태조사 / 김지연·김승경·백혜정·황여정
20-R57-1 2020년 청소년 매체이용 및 유해환경 실태조사(위기청소년 결과 분석) / 김지연·김승경·백혜정·황여정
20-R57-2 2020년 청소년 매체이용 및 유해환경 실태조사(부록: 기초통계결과표) / 김지연·김승경·백혜정·황여정

20-R58 지역사회 협력을 통한 직업계고 혁신 지원 방안 연구 – 특성화고 혁신지원 운영모형 개발 / 강경균·김영만·김용성
20-R59 지역아동센터 아동패널조사 2020 / 김희진·임희진·정윤미
20-R60 이주배경 아동·청소년 지원 지역기관 연계 시범사업 모니터링, 평가 및 모델 개발 / 연보라·최정원·김성은
20-R61 학업중단 현황 및 지원 방안 / 김성은·박하나·김현수
20-R62 환경 변화에 따른 후기청소년 정책 발전 방향 연구 / 장근영·김기헌

세미나 및 워크숍 자료집

20-S01 2020년 학교폭력 예방교육 컨설팅단 워크숍(1.30)
20-S02 2019년 학교폭력 예방교육 운영학교 성과보고회(1.31)
20-S03 2020년 제28회 청소년정책포럼 : 학교밖 청소년 취업 및 자립지원 방안(1.22)
20-S04 2019년 학교폭력 예방교육 어울림 프로그램 운영 우수사례집(1.31)
20-S05 제25차 NYPI 직원 역량강화 콜로키움: 강원국 교수의 매력적인 글쓰기 특강(4.23)
20-S06 제26차 NYPI 직원 역량강화 콜로키움: 김병완 작가의 퀀텀 독서법 및 책쓰기 특강 (4.21)
20-S07 2020년 제29회 청소년정책포럼 : 학교 밖 청소년 규모 추정 방안(6.19)
20-S08 2020년 제30회 청소년정책포럼 : 18세 선거권 이후 청소년 정치교육의 방향 및 과제(7.17)
20-S09 2020년 학교폭력 예방교육 컨설팅단 실무 역량 강화 연수(초등, 중등, 고등)(7.28)
20-S10 2020년 제31회 청소년정책포럼 : 청소년정책 관련 법제도 개선 방향(7.28)
20-S11 제30차 NYPI 직원 역량강화 콜로키움: 플랫폼 노동의 미래를 묻다 〈별점 인생〉(8.5)
20-S12 2020년 학교폭력 예방교육 컨설팅단 상반기 성과보고회(8.11)
20-S13 제31차 NYPI 직원 역량강화 콜로키움: Z세대의 SNS 마켓에서의 소비자사회화 경험에 대한 현상학적 연구: 인스타마켓을 중심으로(8.26)
20-S14 2020년 제32회 청소년정책포럼 : 포스트 코로나 시대 청소년 활동과 정책의 방향(8.18)
20-S15 청소년 이슈 관련 네트워크 분석(8.25)
20-S16 2020년 제33회 청소년정책포럼 : 학교 안밖 청소년 협력체계 구성 연계 방안(9.23)
20-S17 제32차 NYPI 직원 역량강화 콜로키움: 청소년의 혐오표현 노출실태 및 대응방안(10.6)
20-S18 2020년 대안학교 관리자 협의회 및 담당자 역량 강화 워크숍 (10.5)
20-S19 2020년 학교 내 대안교실 모니터링 위원 워크숍(10.5)
20-S20 2020년 제34회 청소년정책포럼 : 코로나 시대 청소년 성장지원의 방향(10.21)
20-S21 2020 유로필로조피 서울 대회(11.12~13)
20-S22 2020년 미인가 대안교육시설 관리자 및 담당자 역량 강화 온라인 워크숍(11.13)
20-S23 제35차 청소년정책포럼: Z세대 청소년의 가치관 변화와 정책적 대응방안(11.18)

20-S24 2020년 학교폭력 예방교육 발전방안 포럼 (해외 사이버폭력 정책 동향 및 국내 사이버폭력 예방교육 방향 모색)(12.18)
20-S25 제3회 한일진로교육포럼(11.28)
20-S26 제9회 한국아동 · 청소년패널 학술대회 자료집(11.20)
20-S27 제36차 청소년정책포럼: 우리나라 한국아동·청소년의 삶은 어떠한가?(12.1)
20-S28 제37차 청소년정책포럼: 이주배경 아동·청소년의 지원 체계의 과제와 방향(12.8)
20-S29 제1차 10대청소년목소리포럼: 코로나로 세겨진 우울의 시대 우리의 꿈과 이상을 쏘다 -동고동락, 각양각색 청소년의 희망이야기(12.5)
20-S30 제38차 청소년정책포럼: 후기청소년 지원정책의 과제와 방향(12.17)
20-S31 제39차 청소년정책포럼: 청소년 매체이용 및 유해환경 실태조사 주요 결과와 과제(12.22)
20-S32 진로위기학생 유형별 진로교육 실천과제 성과보고회 자료집(11.27)

학 술 지

「한국청소년연구」 제31권 제1호(통권 제96호)
「한국청소년연구」 제31권 제2호(통권 제97호)
「한국청소년연구」 제31권 제3호(통권 제98호)
「한국청소년연구」 제31권 제4호(통권 제99호)

기타 발간물

NYPI Bluenote 이슈 & 정책 119호 : 대안교육 위탁교육기관의 질 제고 방안
NYPI Bluenote 이슈 & 정책 120호 : 청년 핵심정책 대상별 실태 및 지원방안 연구 II : 학교 졸업예정자
NYPI Bluenote 이슈 & 정책 121호 : 민간 위탁형 공립 대안학교의 운영 현황 및 발전 과제
NYPI Bluenote 이슈 & 정책 122호 : 국가 미래인적자원으로서 재외동포청소년 성장과 지원방안 연구
NYPI Bluenote 이슈 & 정책 123호 : 다문화청소년의 심리·사회적응 지원방안 모색
NYPI Bluenote 이슈 & 정책 124호 : 정보기술을 활용한 위기청소년 사회서비스 확충 방안
NYPI Bluenote 이슈 & 정책 125호 : 미래지향적 청소년관계법 정비 방안
NYPI Bluenote 이슈 & 정책 126호 : 「청년기본법」제정 전·후 정책형성의 성과와 한계, 그리고 앞으로의 과제
NYPI Bluenote 이슈 & 정책 127호 : 청년예술인 지위 및 권리보장을 위한 '예술 활동 증명' 기준의 적정성 검토
NYPI Bluenote 이슈 & 정책 128호 : 청년종합실태조사(안) 시행을 위한 조사 설계 방향

NYPI Bluenote 이슈 & 정책 129호 : 정부 및 지자체 청년 정책참여 현황과 과제

NYPI Bluenote 통계 49호 : 청소년 '일 경험' 제도 운영 실태 및 정책방안 연구 II : 대학교 실습학기제를 중심으로

NYPI Bluenote 통계 50호 : '학교 밖 청소년'이 얼마나 있을까?

NYPI Bluenote 통계 51호 : 청소년지도사 양성제도 관련 실태조사

NYPI Bluenote 통계 52호 : 한국아동·청소년패널조사 2018 제2차년도 주요 조사 결과

연구보고 20-R08-1

청소년 빅데이터 체계 구축 및 활용방안 연구 데이터 분석 보고서 1_ 청소년 비만에 대한 비정형 빅데이터 연구

인 쇄 2020년 12월 23일
발 행 2020년 12월 30일

발행처 한국청소년정책연구원
세종특별자치시 시청대로 370
발행인 김 현 철

등 록 1993. 10. 23 제 21-500호
인쇄처 경성문화사

구독문의 : (044) 415-2125(학술정보관)
ISBN 979-11-5654-275-9 94330
979-11-5654-273-5 (세트)

연구보고 20-R08-1

청소년 빅데이터 체계 구축 및 활용방안 연구 데이터 분석 보고서 1

: 청소년 비만에 대한 비정형 빅데이터 연구

한국청소년정책연구원
National Youth Policy Institute

30147 세종특별자치시 시청대로370 세종국책연구단지
사회정책동(D동) 한국청소년정책연구원 6/7층
Social Policy Building, Sejong National Research Complex,
370, Sicheong-daero, Sejong-si, 30147, Korea
Tel. 82-44-415-2114 Fax. 82-44-415-2369 www.nypi.re.kr

94330
9 791156 542759
ISBN 979-11-5654-275-9 94330
ISBN 979-11-5654-273-5 (세트)