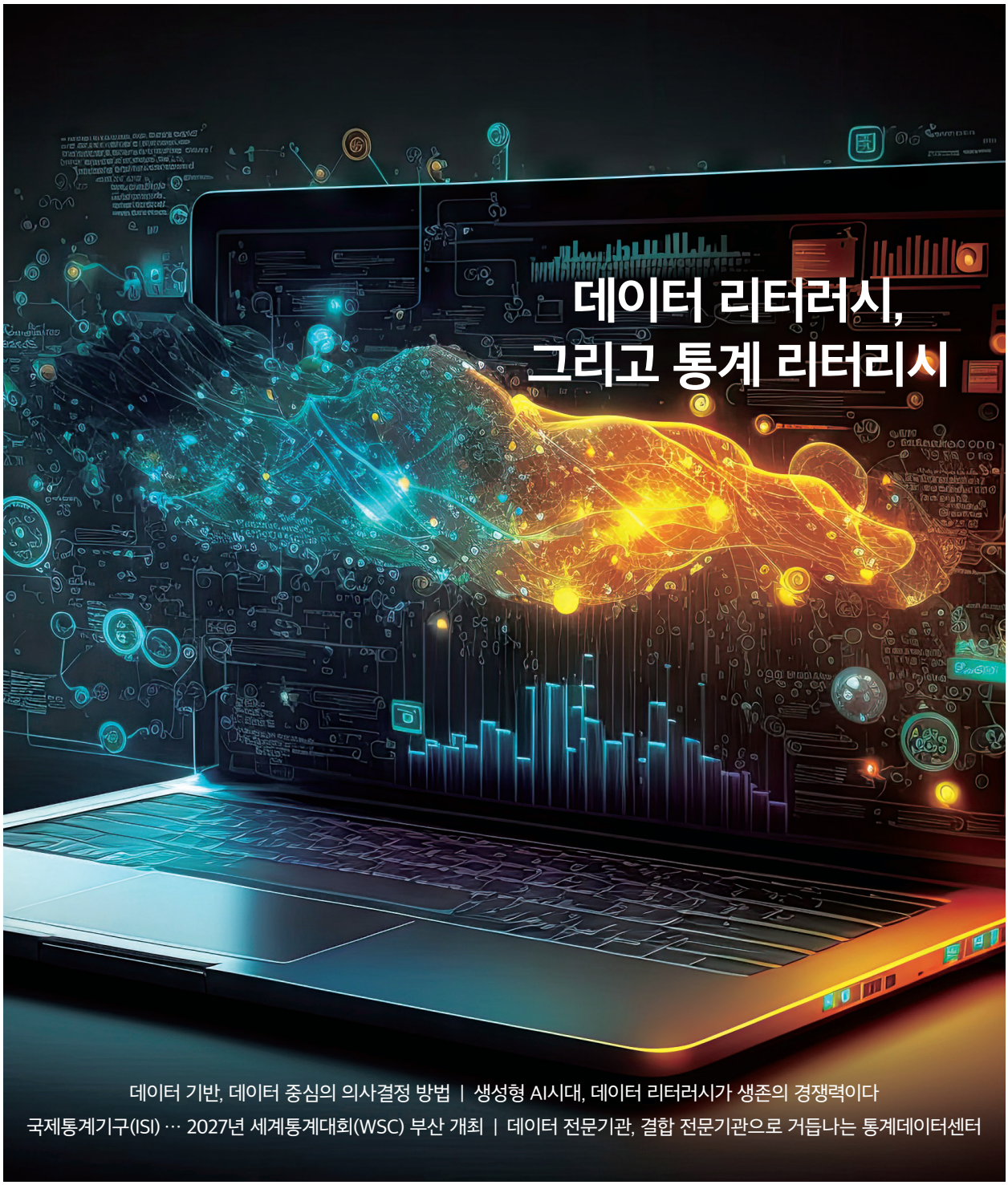
통계의 창

WINDOW OF STATISTICS

2023.
WINTER
VOL.32



데이터 기반, 데이터 중심의 의사결정 방법 | 생성형 AI시대, 데이터 리터러시가 생존의 경쟁력이다

국제통계기구(ISI) … 2027년 세계통계대회(WSC) 부산 개최 | 데이터 전문기관, 결합 전문기관으로 거듭나는 통계데이터센터

CONTENTS

통계의 창
2023. Winter
Vol.32

발행일 2023년 11월 20일

발행인 송영선

발행처 통계교육원

기획 황현식, 정명진, 김정대

주소 대전광역시 서구 한밭대로 713(월평동) 통계센터 통계교육원

전화 042-366-6151, 6152

팩스 042-366-6498

이메일 mjjung@korea.kr, haissy@korea.kr

디자인 및 인쇄 (주)피그마리온(02-516-3923)

ISSN 2005-1379
©2023. 통계교육원
※ '통계의창'에 실린 내용은 필자 개인의 의견이므로 필자의
소속기관이나 본지의 공식적인 견해를 대변하는 것은 아닙니다.

1 통계 ISSUE

2 통계 FOCUS

3 통계 EDU

01 데이터 기반, 데이터 중심의 의사결정 방법

김철수 | 디지털역량연구소장

데이터 기반, 데이터 중심의 의사결정 방법을 글로써 달라는 요청을 받자마자 필자는 ChatGPT에게 이렇게 물었다.

"데이터 기반, 데이터 중심의 의사결정 방법을 주제로 글을 쓰고 싶다. 참조할 사람 3명을 알려줘."
ChatGPT는 2021년 9월까지 인터넷에 올라온 수많은 데이터를 학습했고, 제 질문에 다음과 같이 3명을 알려주었다.

1 에릭 콜슨(Eric Colson)
Stitch Fix의 Chief Algorithms Officer로서, 데이터 사이언스를 이용하여 개인화된 의류 선택 알고리즘의 성공적인 구현에 중추적인 역할을 했습니다.

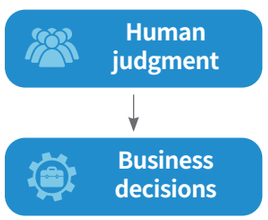
2 앤드류 응(Andrew Ng)
스탠포드 대학교 교수이며 Coursera의 공동 창립자. 그는 머신러닝과 인공 지능 분야의 전문가로서, 데이터의 중요성과 의사결정에 있어서의 머신러닝의 역할에 대한 깊은 통찰력을 갖고 있습니다.

3 힐러리 메이슨(Hilary Mason)
Fast Forward Labs의 설립자이며, 데이터 사이언스와 기계 학습의 전문가입니다. 그녀는 실제 비즈니스 문제를 해결하는데 데이터를 어떻게 활용할 수 있는지에 대한 실용적인 관점을 제공합니다.

이들은 통계, 인공지능, 빅데이터, 데이터 기반 의사결정에서 가장 유명하고 혁신적인 성과를 낸 사람들이다. 이 중에서 첫 번째 인물인 에릭 콜슨은 Stitch Fix에서 추천 알고리즘을 개발했다. 그전에는 Netflix와 Yahoo에서 데이터 과학자로 일했다. 그가 2019년 하버드 비즈니스 리뷰에 쓴 글 "What AI-Driven Decision Making Looks Like"에는 데이터 주도 의사결정에 관한 구분과 통찰이 나온다. 그는 데이터 주도 의사결정을 4단계로 나눴다. 그의 구분과 통찰을 이해해 보자.

1 단계 감·경험·연륜·노하우·속담과 같은 개인의 판단

에릭 콜슨은 사람들이 지금까지 해왔고 지금도 가장 많이 하는 의사결정 방법은 개인의 판단이라고 한다. 감이나 경험, 연륜이나 노하우 같은 것이다.



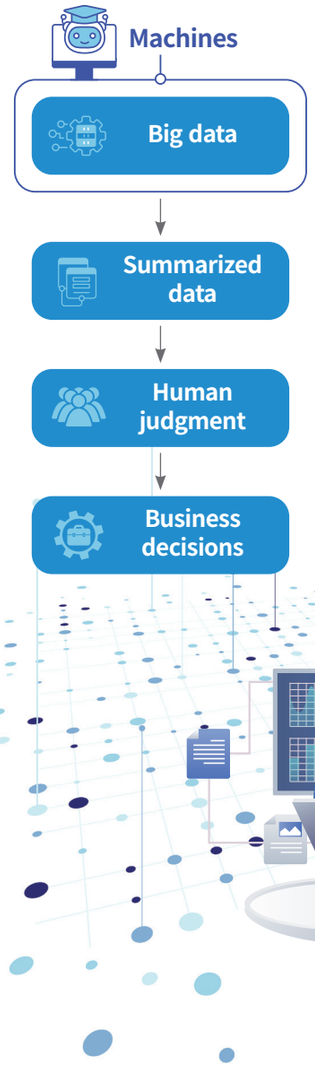
감, 경험, 연륜, 노하우 같은 것이 확고한 상식처럼 된 경우가 속담이다. 지금은 남녀노소 구분 없이 속담을 사용한다. 하지만 과거에 속담은 아무나 쓸 수 있는 것이 아니었다. 아프리카의 요류바족은 속담의 인용이나 언급은 아주 많고 다양한 경험이 바탕이 되어야 한다고 믿는다. 나이 어린 사람이 연장자 앞에서 속담을 쓰면 연장자에게 사과해야 한다. 부족에서 가장 현명한 사람은 속담을 얼마나 많이 사용하느냐에 달렸다. 속담을 사회규범과 전통적인 믿음을 강화시키는 도구로 사용하는 것이다.

티브족의 연장자는 속담은 좌절을 맛본 사람에게 용기를 북돋아주는 수단으로 생각한다. 연장자들은 젊은 사람들에게 세상을 어떻게 살 것인지 속담으로 충고한다. 어떤 속담은 유용한 지식을 얻는 법을 알려주고, 불평등한 상황에 지혜롭게 대처하는 법을 알려준다.

속담은 권위의 상징이고 권력의 표현이다. 속담으로 의사결정을 하는 것은 조직의 위계를 강화한다. 조직에서 상사가 데이터보다는 속담과 같은 감, 경험, 연륜, 노하우로 결정하는 것은 어찌보면 당연한 전통처럼 보인다. 하지만 속담은 결국 고정관념이다. 동화, 설화, 신화, 전설 같은 이야기도 모두 고정관념이고 고정관념을 강화한다. 시대가 변하고 기술이 발전하고 사람이 달라지고 시장이 바뀌는데 고정관념은 그대로다. 고정관념으로 모든 것을 결정하려는 사람을 우리는 '꼰대'라고 부른다. 꼰대의 의사결정에 대한 반감, 또는 꼰대의 고정관념 회피의 방법으로 데이터 기반, 데이터 중심의 의사결정이 나왔다고 볼 수 있다.

2 단계 데이터 분석을 통한 개인적 판단

컴퓨터(Machines)가 빅데이터를 분석해서 요약하고 사람은 요약 데이터를 가지고 여전히 감이나 경험으로 판단한다.



여기서 중요한 이슈 하나가 있다. 컴퓨터가 제시하는 요약된 데이터가 과연 정확한 것이냐 하는 문제다. 예를 들어 퀴즈를 하나 풀어보자. 지금은 많은 사람들이 배달 앱으로 음식을 주문해 먹는다. 불과 10년 전만 해도 우리는 다들 전화로 주문했다.

그때 2015년에 한 배달 앱 기업이 몇몇 치킨집의 POS 데이터를 모아 분석했다. POS는 Point Of Sales의 약자인데 쉽게 말해 카운터에 있는 계산대라고 보면 된다. 계산대를 터치하면 주문 정보가 입력되고 카드결제기와 연결되어서 결제도 가능하다. 이들 치킨집의 영업시간은 오전 11시부터 새벽 2시까지였다. 그러면 이들 치킨집의 POS 데이터를 분석했더니 우리나라 사람들은 하루 중 몇 시에 치킨을 가장 많이 주문해 먹었을까?

오후 6시? 오후 12시? 밤 9시? 월드컵 하는 시간? 아니다. 정답은 새벽 3시였다. 당시에 치킨집에는 주문 전화를 받는 전화기가 여러 대였다. 전화 한 대

만 갖고 장사할 수 없었다. 고객이 전화를 했는데 매장에서 전화를 안 받으면 고객은 이내 다른 치킨집에 전화하기 때문에 같은 전화번호로 여러 전화기를 두고 주문을 받았다.

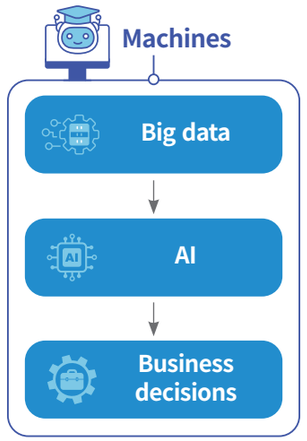
전화기는 주방 앞에 있어서 바로 주방에 주문 내용을 전달해야 했다. 그러니 카운터까지 가서 POS에 입력할 시간이 없었다. POS 입력 속도도 사람을 따라오지 못했다. 그래서 주방 앞 전화기 앞에 종이 공책을 갖다 놓고 주문 내역을 볼펜으로 적었다.

이제 영업이 2시에 끝나고 청소하고 사장이 카운터에 앉았다. 장부를 꺼내 POS에 입력하기 시작했다. 데이터는 현상, 즉 현재 상태의 기록이므로 입력하는 순간이 주문 시간이 되었다. 그래서 POS 데이터를 분석하면 새벽 3시에 다들 주문한 것으로 나온다.

컴퓨터가 아무리 좋은 모델을 사용한들, 입력된 데이터가 쓰레기면 나오는 결과도 결국 쓰레기다. Garbage in, Garbage Out이다. 실제로 몇몇 대기업이 현장에서 날고 기는 직원 수십 수백 명을 모아서 6개월간 빅데이터 분석과 파이썬 프로그래밍을 가르쳤다. 하지만 현장에 돌아간 사람들은 데이터로 의사결정할 수 없었다. 데이터가 없거나, 있어도 쓰레기거나, 고치려고 해도 못 고치는 경우가 너무 많았다. 데이터 품질이 담보되지 못한 상황이라면 컴퓨터에 빅데이터를 맡겨 의사결정하게 할 수 없다.

3 단계 인공지능을 통한 의사결정

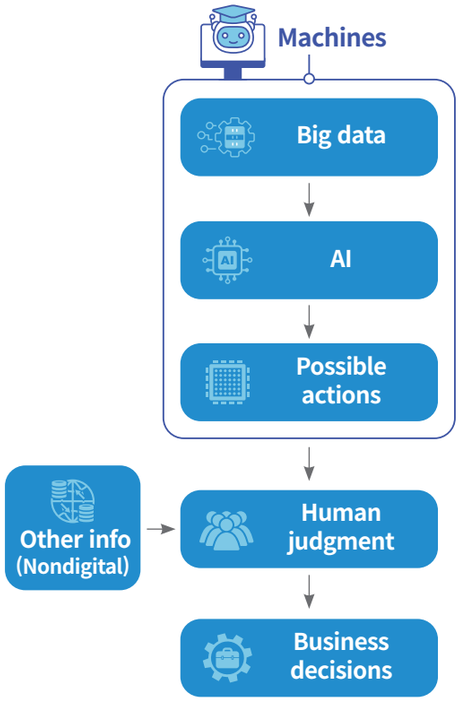
사람의 개입 없이 컴퓨터가 데이터 수집부터 의사결정까지 한다.



자율주행차가 대표적인 예다. 흔히 자율주행차는 컴퓨터가 운전한다고 생각하는데 단계가 여러 가지 있다. 1단계는 자동 브레이크나 자동 속도 조절 같은 운전 보조 기능이다. 2단계는 부분 자율 주행이어서 차선을 넘지 않거나 주차를 알아서 하는 경우다. 3단계는 조건부 자율주행이어서 자동차가 안전 기능을 직접 제어하고 탑승자의 제어가 필요하면 신호를 보낸다. 4단계는 고도 자율주행인데 비가 오나 눈이 오나 주변 환경에 관계없이 운전하고 운전자가 제어하는 것이 불필요하다. 5단계는 완전 자율주행이어서 운전자가 타지 않고도 움직이는 무인 주행차다.

문제는 자율주행이 비즈니스 의사결정이냐 하는 것이다. 운전을 하면서 언제 좌회전을 하고 언제 멈춰야 하는지 결정하는 것이 비즈니스 의사결정일까? 비즈니스에서 의사결정은 조직이나 기업의 목표와 전략을 달성하기 위해 선택해야 할 행동이나 경로를 결정하는 과정이다. 이 과정에는 권한과 책임이 따른다. 사람은 자기 권한과 책임 하에서 필요한 권한을 사용하고 필요한 책임을 진다. 이것이 자율주행차와 사람이 의사결정을 달리하는 이유다.

자율주행차는 무조건 자기를 방어하려 하겠지만 사람은 자기 팔이 잘려나가더라도 돌진하는 의사결정을 할 수 있다. 자율주행차는 최적의 타이밍을 계산하지만 사람은 이때다 싶으면 덤벼든다. 이때 자율주행차는 데이터 기반으로 의사결정하고, 사람은 감으로 의사결정한다고 볼 수 없다. 사람은 사람 나름의 권한과 책임을 사용하는 과정이 있는 것이다. 즉, 컴퓨터가 의사결정의 모든 과정을 수행한다는 것은 결국 사람이 그렇게 하도록 허용한 것이고, 그 사람이 자신의 권한과 책임 하에서 의사결정한 결과인 것이다.



4 단계 인공지능과 인문학의 교차로에서 의사결정

컴퓨터가 빅데이터를 분석하고 AI로 학습하고 테스트해서 가능한 안을 내놓고, 거기에 사람의 비디지털적인 감각을 더해서 판단한다.

컴퓨터가 빅데이터와 인공지능 기술을 활용해서 우리에게 Possible actions을 준다는 말은 곧 우리에게 '안(案)'을 낸다는 말과 같다. '안'은 문제를 해결하는 더 좋은 방법을 말한다. 보고(안), 기획(案)이라고 쓰는 이유도, 상사의 문제를 푸는 더 좋은 방법을 글로 담았기 때문이다. 일반적으로 안은 상사와 고객에게 준다. 상사에게 주는 것이 기안이고, 고객에게 주는 것이 제안이다.

이때 상사에게 보고(안), 기획(안)을 주는 것이 이른바 부서원의 역할이다. 부서원은 보고나 기획의 초안을 만들고 상사와 검토하면서 의사결정으로 나아간다. 이때 부서원은 시장 조사, 사례 조사, 문헌 연구, 실증 실험, 설문 조사 등 데이터를 기반으로 안을 만든다. 즉 데이터를 분석해서 안을 만든 것이다. 그 안을 두고 상사는 감, 경험, 연륜, 노하우로 판단한다. 비 데이터적인 분석 기법을 사용하는 것이다. 어쩌면 기술과 인문학의 교차로에서 비즈니스 의사결정을 하는 것이 최적이지 않을까? 애플의 스티브 잡스가 했던 말을 떠올려 보자. "우리가 창의적인 제품을 만든 비결은 항상 기술과 인문학의 교차점에 있고자 했기 때문입니다."

에릭 콜슨의 글이 알려주는 것은 바로 이것이다. 우리가 빅데이터 분석을 배우거나 인공지능을 연구하거나 컴퓨터 기술을 배우는 것이 핵심이 아니다. 데이터를 기반으로 데이터를 중심으로 의사결정하는 것이 핵심이 아니다. 그렇다고 사람이 감이나 경험, 연륜이나 노하우로 꼰대처럼 의사결정하는 것도 핵심이 아니다. 핵심은 조화에 있다. 둘을 합치는 것이다. 데이터로 안을 만들고 감으로 판단해서 비즈니스에 도움이 될 의사결정을 하는 것이다.

데이터 기반 의사결정이 기존의 사람의 의사결정과 대척점에 있는 것이 아니라는 말이다. 우리는 데이터와 감의 교차로에서 의사결정을 해야 한다는 말이다.

Apple's special event in March 2011
(출처:https://www.youtube.com/watch?v=Kll1MR-qNt8)

02 생성형 AI시대, 데이터 리터러시가 생존의 경쟁력이다

이재원 | 경희대학교 경영대학원 겸임교수 [1)]

1) 이재원 저서 : <마이데이터 레볼루션 - 초개인화 시대가 온다>,
<2030 데이터 리터러시 레볼루션 - 당신은 챗GPT 시대의 생존역량을 갖추었는가>

데이터 리터러시가 앞으로 10년의 무기가 된다

"세상에 데이터가 없으면 어떻게 될 것 같으냐?"고 챗GPT에게 물었다. 정리를 잘해서 딱 부러지게 대답한다. "데이터가 없으면 우리가 필요로 하는 정보를 얻을 수 없고 원하는 서비스를 이용할 수 없다"고 한다. 또한 "데이터는 경제적 가치를 가지고 있어 없으면 경제적 문제가 발생할 수 있으며, 사회적 문제를 해결할 수도 없다. "결국 데이터가 없으면 미래 예측이 어려워지며, 우리의 삶에 큰 영향을 미칠 수 있다."고 대답한다.

이렇듯 데이터가 없는 세상은 상상하기 어렵다. 더구나 디지털 시대를 살아가는 요즘, 인공지능, 블록체인, IOT, 메타버스 등 디지털 기술은 모두 데이터를 원료로 하여 움직인다. 특히 챗GPT를 포함한 생성형 AI 시대에는 데이터가 더욱 중요하다.
인공지능은 데이터로 학습하여 모델을 생성하고 데이터로 결과를 제시한다. 모든 것이 데이터로 움직인다. 따라서 일상이나 비즈니스에서 데이터에 대해 모르거나 제대로 활용하지 못한다면, 즉 데이터 리터러시로 무장하지 않으면 생존할 수 없다.

데이터 리터러시는 데이터data와 리터러시literacy의 합성어로 흔히 '데이터 문해력(文解力)'이라고 번역된다. 데이터 리터러시는 데이터를 문맥에 맞게 읽고 사용하고 소통하는 능력은 물론이고 데이터에 담겨 있는 의미를 파악하고 목적에 맞게 활용하는 능력을 의미한다. 결국 데이터를 어떻게 잘 이해하고 활용하느냐 즉, 데이터 리터러시가 미래 경쟁력의 관건으로 작용한다.

데이터 리터러시에 자신이 없는 이유는 무엇일까

글로벌 컨설팅 전문기업 액센츄어Accenture의 설문조사를 보면 데이터 리터러시에 자신감을 느끼는 사람은 21%에 불과하다고 한다. 상황이 이렇다 보니 직장인들과 학생들이 모두 코딩이나 R, 파이썬 Python 등 분석 도구 배우기에 열심이다. 기업에서도 데이터 분석 전문인력을 확충하거나 전담부서를 두어 경쟁에 뒤떨어지지 않게 노력하고 있다. 하지만 배운 분석 도구들이 실제 현업에서 생각만큼 제대로 활용되고 있지 않는 것 같다. 데이터 분석 보고서는 넘쳐나지만 굳이 분석을 안 해도 알 만한 일반적인 결론만 들어 있는 경우가 많다. 목적과 문제해결이라는 본질은 놓아둔 채 기술만 열심히 익힌 결과다.

회사에서 제공하는 데이터 교육은 대부분 프로그램을 짜보는 기술 향상 교육에 치중돼 있다. 그러나 올바른 데이터 리터러시의 함양은 도구 활용 기술뿐만 아니라 데이터에 대해 올바른 시각을 가지고 올바른 방법으로 문제를 해결하는 역량을 함께 기를 때 가능하다. 이를 위해서는 우리 회사의 데이터는 어떻게 발생되어 수집되고 관리되는지부터 알아야 한다. 내가 필요한 데이터는 어디에 어떤 형식으로 저장되는지를 먼저 알아야 필요할 때 쉽게 찾아쓸 수 있다. 분석 도구를 교육할 때도 직원들이 부서의 문제를 들고 직접 코딩해보도록 하고 성공적 활용사례를 공유하여 이를 응용할 수 있도록 해야 한다.

조직에서 데이터 활용 역량이 이미 비즈니스 성패를 가르고 있다

넷플릭스는 추천 알고리즘을 구축하기 위해 2006년부터 3년에 걸쳐 100만 달러의 우승 상금을 걸었을 정도로 알고리즘 개발에 사활을 걸었다. 그때부터 사용자들의 검색, 시청 시간 및 기록, 평점 등 행태 정보를 수집하고 취향이 비슷한 사람들의 데이터, 컨텐츠 정보, 사용 장치까지 다양한 데이터를 결합하여 활용한다. 데이터에 기반한 추천 시스템이 없었다면 지금의 넷플릭스가 있었을까?

현대자동차도 더 이상 자동차 제조회사에 머무르지 않고 있다. 차량과 관련한 다양한 이 업종 데이터를 연결하여 고객들이 편리하게 자동차를 사용하도록 생태계를 만들었다. 주유 기록, 정비 내역 등 고객이 차량 관리 내역을 한눈에 파악하고(차량관리), 실시간 차량 위치 데이터를 이용하여 차 안에서 주문과 결제(차량편의) 등도 가능하게 만들었다. 자동으로 주행 거리 포인트를 제공하며(차량정보) 보험 가입도 편리하게(차량금융) 하고 있다.

인도의 조마토(Zomato)라는 회사는 데이터 분석을 기반으로 한 배달 앱을 만들어 13억 6,000만 인도 국민의 식습관에 혁명을 일으켰다. 인도의 자가용 소유 가정이 단 2%뿐이고 남의 집에서 요리한 음식을 절대 먹으려 하지 않아 인도 국민의 90%는 식당에서 외식을 하지 않는다. 하지만 인도 국민들이 이 앱에 열광하고 있다. 클릭 몇 번의 편리함으로 조마토는 20만 개 이상의 레스토랑 파트너와 약 10만 명의 배달 파트너가 있고 지금까지 약 1억 건 이상의 배달을 주문하고 있다고 한다. 모두 음식, 배달 시간, 가격, 할인에 대한 고객의 선호와 취향을 추적해 통찰을 얻고 다양한 지역 정보와 결합하는 데이터 활용에 능한 덕분이다.

젊은층 대상 의류 브랜드인 '디스커버리' 로 유명한 국내 의류업체 에프앤에프(F&F)도 소비자들의 성향 데이터를 면밀히 분석하여 신상품 개발에 반영하였다. 2019년에는 '따뜻함'과 '커플'이란 키워드를 도출하여, '플리스'라는 겨울 커플룩을 히트시켰고 2020년에는 890g의 초경량 백팩 '라이크 에어 백팩'을 탄생시켰다. 감으로 움직였던 패션산업을 데이터를 수집·분석·활용하여 비즈니스를 혁신하였다. 이제 데이터 자산을 활용하여 비즈니스와 고객 경험을 혁신하는 것이 기업의 필수적인 경영 전략이 되어 버렸다.

데이터 중심의 경영체제와 기업문화가 선행되어야

카지노로 유명한 해러스 그룹의 게리 러브먼 회장은 데이터 지향적인 조직 문화를 구축하기 위해 혼신의 노력을 기울였다. 그는 “데이터 분석을 통해 알아낸 것인가?”라는 질문을 직원들에게 자주 했다고 하며 직원들은 누구나 이를 뒷받침하도록 데이터 분석에 입각한 증거를 제시해야했다. 심지어 러브먼은 “우리 회사에서 해고되는 사유는 3가지다. 첫째는 절도, 둘째는 성희롱, 셋째는 데이터 없이 말하는 것이다.”라고 말했을 정도로 데이터를 기업 경영의 가장 중요한 핵심으로 삼았다.

국내 숙박 플랫폼의 대표 주자인 야놀자도 “야놀자만의 일하는 방식인 와이코드Y-CODE를 만들어 ‘데이터가 모든 판단의 기준이다.’라고 명시하여 조직의 사고와 행동의 기준으로 제시하고 있다. 특히, 이 회사는 데이터 관련 업무를 IT 부서뿐만 아니라 기획·영업·마케팅 부서에서도 직접 쿼리를 짜고 대시보드를 만들어 데이터를 분석할 수 있도록 지원하고 있다.

데이터 활용 최강 기업으로 유명한 에어비앤비는 호스트(숙박 제공자)의 과거 행동을 기반으로 호스트의 선호도를 학습하는 알고리즘을 통해 머신러닝 모델을 개발했고 검색엔진에 적용했다. 그 결과 예약 전환율이 약 3.75% 상승해 더 많은 거래가 성사됐다고 한다. 또한 이 회사는 데이터 접근성, 데이터 도구 활용, 데이터 교육이 데이터 기반 의사결정에 가장 중요하다고 보고 2016년 데이터 유니버시티를 설립했다. 설립한 이후 약 400개 이상의 과정이 진행됐고 4,000명 넘는 임직원의 대다수가 하나 이상의 과정을 수강했다고 한다. 특히 현업 맞춤형 교육인 데이터 U 인텐시브는 사업부 단위별로 부서에서 사용하는 데이터를 가지고 교육한다.

이렇듯 사례를 든 기업들 모두 최고 경영자부터 솔선수범하고 데이터를 경영의 기준으로 삼도록 조직문화를 바꾸었다. 또한 구성원이 데이터에 쉽게 접근할 수 있게 하고 매일의 업무에서 데이터에 기반한 의사결정을 하도록 교육하고 지원하였다.
과거에는 데이터라 하면 데이터 전문가나 전담부서의 일로 인식됐으나 지금은 고객과 현장의 문제해결을 위해서 전 직원의 데이터 리터러시를 반드시 높여 나가야 한다. 지금부터라도 변화를 만들어 조직의 모든 문화를 데이터 중심으로 바꿔 나가야 한다. 이를 위해 데이터에 쉽게 접근할 수 있도록 시스템과 도구, 거버넌스 등 인프라를 구축하고 직원들의 평가, 보상, 교육에 힘써야 할 것이다. 데이터 리터러시를 성공적으로 향상한 조직은 이것이 반복적 프로세스의 결과라는 것을 잘 알고 있다. 작은 것부터 시작해 피드백을 통해 지속적으로 향상해 나가는 것이 좋겠다.

직원들도 퍼스널(Personal)한 데이터 경쟁력을 갖추어야

작년에 한 직원을 칭찬하고 점심을 사준 적이 있다. 작업 시간이 오래 걸리는 일들을 간단히 프로그램을 고쳐 나머지 직원들의 일을 크게 줄여줬기 때문이다. 이 직원은 며칠 고민한 끝에 새로운 함수를 찾아내서 한꺼번에 몇만 줄씩 올라가도록 간단한 프로그램으로 문제를 해결해버렸다. 그 덕분에 기존 작업 시간이 1~2시간에서 5분으로 줄었다고 하니 참 감사한 일이다. 개인이나 직원들의 데이터 활용 역량이 높아져야 특근도 없어지고 올바르게 문제해결도 할 수 있다. 그렇게 하기 위해서는 데이터를 잘 활용할 수 있도록 직원들의 역량이 높아져야 한다. 가트너는 “데이터를 언어처럼 배워라.”라고 강조한다. 우리가 모국어를 자유롭게 사용하듯이 이제 데이터도 모국어처럼 배워 능숙하게 활용해야 하는 시대다. 그러기 위해서는 다음의 사항은 반드시 머리에 넣어두자.

❶ 첫째, 현장에서 보면 수집한 데이터를 정리하여 그럴듯한 결론만 제시하는 경우가 종종 있는데 문제해결 중심의 결과가 나오도록 데이터를 분석해야 한다. 이를 위해 다양한 관점에서의 가설이 필요하다. 강력한 가설이 되기 위해서는 하나의 궁금증에서 시작해 그 데이터를 보다 보니 다른 것이 궁금해지는 꼬리에 꼬리를 무는 가설일수록 좋다.

❷ 둘째, 데이터 기초 분석 방법과 데이터 분석 도구 등에 대한 이해도 중요하다. 통계의 기초, 데이터 마이닝, 머신러닝과 딥러닝에 대해 개념만이라도 알아두자. R이나 파이썬과 같은 고급 분석도구를 모두 다 배울 필요는 없다. 때에 따라 엑셀이나 SQL로도 분석할 수 있다. 쉬운 것부터 시작해 자꾸 경험 해보고 익숙해진 다음 고급 분석 도구에 도전하면 된다.

❸ 셋째, 데이터 시각화와 스토리텔링도 활용해 소통 능력을 높여 보기 바란다. 본인에게 맞는 시각화 도구를 선택해 차트나 그래프를 자주 그려보고 위치나 색상도 강조해보고 하면서 이런 문제를 표현하려면 이런 기법이 좋더라는 것을 체득해야 한다.

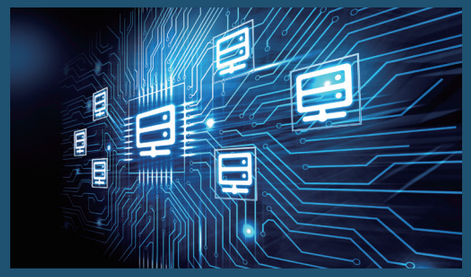
❹ 넷째, 비판적으로 사고하고 분석의 시야를 넓혀야 한다. 데이터 분석을 통한 정답은 하나가 아니라 여러 가지가 있을 수 있으므로 입체적 사고가 필요하다. 그러기 위해 현업 업무 외에도 시장과 고객을 이해할 수 있는 마케팅, 심리학, 경영학 등에 대한 다양한 분야의 책도 읽어두면 도움이 될 것이다.

앞서 얘기했듯이 데이터 리터러시는 전문가들만을 위한 영역이 아니다. 데이터를 접하는 현업에 있는 실무자, 대학생, 주부, 청소년 모두가 올바른 관점과 필요한 역량을 갖추고 다양한 데이터를 기반으로 일상생활에서 문제해결 경험을 많이 쌓아가는 것이 데이터 리터러시를 높이는 지름길이다.

03 데이터 리터러시, 그리고 통계 리터리시

윤석용 | 명지대 AI빅데이터융합연계 교수

리터러시(Literacy)라는 용어 자체는 다소 생경하지만, 단어를 이해하는 데는 그렇게 어렵지 않다. 케임브리지 영영사전에서 리터러시는 “읽고 쓰는 능력(the ability to read and write)”으로 정의되어 있다. 좀 더 풀어 보면, 문장을 읽어 이해하는 독해력(讀解力)과 의사를 표현하기 위한 문장력(文章力)이라고 할 수 있고, 최근에는 독해력을 문해력(文解力)으로 표현하기도 한다. 그런데 문해력은 읽고 이해하는 능력을 넘어 데이터를 찾고, 데이터를 분석하여 이해하고, 인사이트를 발견하고, 여러 사람과 소통하고, 도구를 이용한 계산 및 통계처리 등을 포함하는 조금은 더 포괄적인 역량을 말한다.



최근 인터넷에서 많이 회자되었던 ‘심심한 사과’, ‘사흘’ 등의 표현이 바로 문해력과 관련된 이슈의 일면이라고 할 수 있는데, 2022년 EBS가 조사한 문해력 수준 진단 결과, 조사대상자의 27.5%가 문장 이해에 어려움이 있다는 것이다. 독서와 사고보다는 직관적이고 단편적인 동영상과 앞뒤 문맥 없이 축소된 단어들에 익숙한 인터넷 시대의 단면이기도 하다.

국립국어원의 우리말샘에서 보면 새롭게 등록된 디지털 리터러시, 미디어 리터러시라는 단어가 있는데, 여기서 디지털 리터러시를 이렇게 정의하고 있다. “디지털 시대에 필수적으로 요구되는 정보의 이해 및 표현 능력”.

20세기 들어 인터넷의 출현은 우리 모두의 삶을 빠른 속도 위에 올려놓았다. 수일 또는 수주에 걸려 전달되던 편지가 키보드의 클릭만으로 수초 내에 전달되고, 데이터가 쌓여가는 속도 또한 과거와는 비교가 되지 못한다.

[그림1] AI의 발전단계

우리 시대의 주류가 되어 가고 있는 AI

아직도 인공지능(AI)이라고 하면 SF 영화 속 로봇을 떠올릴 수도 있지만, 이제 AI는 현실이 되어 누구나 쉽게 경험할 수 있는 시대가 되었다.

AI는 보통 세 단계로 나누어진다.

1단계인 ANI(Artificial Narrow Intelligence)는 데이터분석, 기계학습 알고리즘 등으로 무장하여 AI의 맛을 보여준 빅데이터 시대를 열었다.

2단계인 AGI(Artificial General Intelligence)는 특정 도메인에 국한하지 않는 인간과 유사한 일반적 지능을 갖춘 인공지능으로 아직은 멀리 있는 공상 과학 속의 단계로 생각했지만, ChatGPT의 출현은 AGI로 들어가는 초입일 수 있다고 많은 학자가 이야기하고 있다.

3단계는 레이 커즈와일(Ray Kurzweil)이 이야기한 특이점의 시대, 즉 **ASI**(Artificial Super Intelligence)로서, 강력한 지능체계를 갖추고 스스로 목표를 설정하여 지식을 강화하는, 말 그대로 영화 속 인공지능의 세상이다.

이제 AI는 강 건너의 이야기가 아닌 우리 시대의 주류가 되어가고 있고, 여기에 맞는 정보의 이해 및 표현 능력이 필요하다. 이것이 지금의 리터러시라고 할 수 있다.

지금의 AI는 수십 년 전의 인공지능과 접근방식이 전혀 다르다. 즉 프로세스의 논리적 구현이 아닌, 데이터 기반의 학습으로 ChatGPT를 통하여 경험한 것처럼 빅데이터의 활용과 컴퓨팅 파워를 이용하여 AI를 구현함으로써 AI의 무한한 능력의 확대를 보여주고 있다. 따라서 지금은 AI의 기반이라 할 수 있는 데이터와 데이터 분석으로 들어가 데이터 리터러시를 이해하는 것이 무엇보다 중요한 시점이 되었다.

시티즌 데이터 사이언티스트 (Citizen Data Scientist, CDS)

데이터의 존재만으로 데이터의 가치를 이야기하는 것은, 금이 매장된 산의 존재만으로는 그 가치를 알 수 없고 채굴을 통해서만이 금의 절대적 가치 평가가 가능하듯이, 데이터도 존재만으로는 제한적인 의미를 갖지만, 데이터 분석을 통해 인사이트를 찾는다면 데이터의 절대적 가치를 평가받을 수 있다.

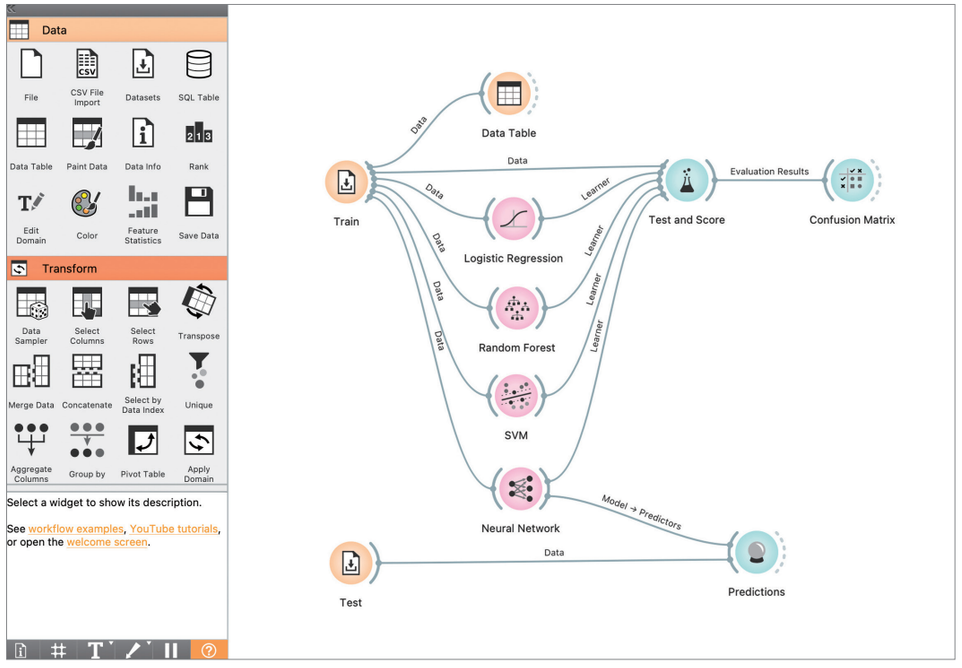
데이터를 분석하기 위해서는 많은 전문적 지식이 요구되는데, 이러한 전문가를 데이터과학자(Data Scientist), 데이터분석가(Data Analyst)라고 한다. 그러나 이러한 전문가에게 데이터 리터러시를 이야기하는 것은 대장장이에게 쇠메의 사용 방법을 이야기하는 것과 같다. 데이터 리터러시는 자기의 본연의 업무가 있고, 이를 수행하면서 발생한 데이터를 스스로 분석하고 인사이트를 찾아 활용할 수 있는 역량으로, 데이터 분석 전문가를 제외한 우리 대다수라고 할 수 있다. 이를 시티즌 데이터 사이언티스트(Citizen Data Scientist, CDS)라고 부른다.

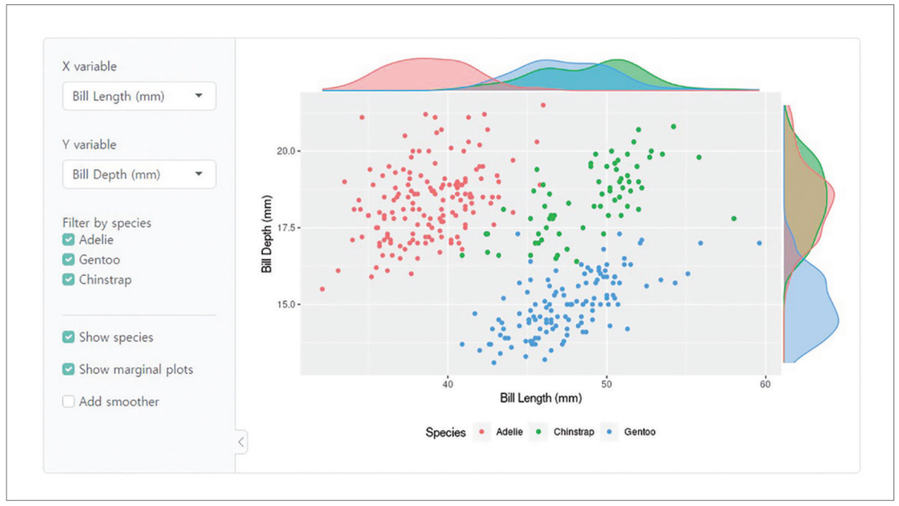
CDS가 갖춰야 할 데이터 리터러시를 정리하면,

❶ **첫째,** 자신의 데이터를 수집하고 분석할 수 있는 역량이다.
수행하는 업무에서 발생하는 데이터의 가치를 인식하고, 이를 의미 있게 수집·저장할 수 있고, 수집된 데이터를 이용하여 통계분석, 데이터시각화 등의 탐색적데이터분석(EDA)을 수행하고, 필요시 Orange3와 같은 CDS 도구를 이용하여 예측모형을 쉽게 만들어 업무에 적용해 보는 것이다.

❷ **둘째,** 데이터 분석 결과로 보고서를 만들어 의사

[그림2] Orange3를 이용한 데이터 분석





[그림3] Shiny를 이용한 분석결과 시스템화

소통에 활용하거나 Streamlit, Shiny 등의 간단한 도구를 이용하여 데이터 분석 결과를 시스템화해 보는 것이다. CDS 입장에서 기계학습·딥러닝·시스템화가 가능할 수 있을지 의문을 품을 수도 있지만, CDS 도구는 바로 이러한 상황을 쉽게 풀어주는 툴이라고 할 수 있다.

통계리터러시의 필요성

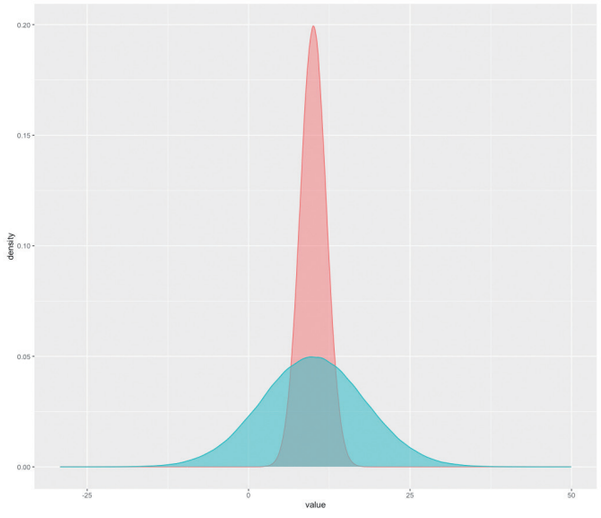
이쯤에서 통계 리터러시에 관한 이야기를 할 필요가 있다. 데이터분석은 통계에 기반하고 있고 통계의 지원 없이는 데이터 분석 결과에 대하여 확신을 가질 수 없다. 통계학은 수학이나 다른 고전 학문에 비하여 그렇게 역사를 길게 보지 않는다. 아직도 많은 발전이 기대되는 학문으로 해석될 수 있다.

가끔 AI 강의 중 “평균이 무엇인가요?”라는 질문을 던져보면, 대부분 “모든 요소의 값을 더하고 요소의 개수로 나눈 값입니다.”라고 대답한다. 나의 질문은 “What is the mean?”이었지만 대답은 “How to calculate the mean?”에 대한 대답이었다. 그렇다면 정확하게 평균이 무엇이고 어떻게 사용해야 하는지 알지 못하고 있다는 결론에 이르게 된다. 평균과 분산을 이해하는 것은 데이터 분석에서 중요한 요소인데, 정확히 이해를 못 하고 있다는 방증이라고도 할 수 있다.

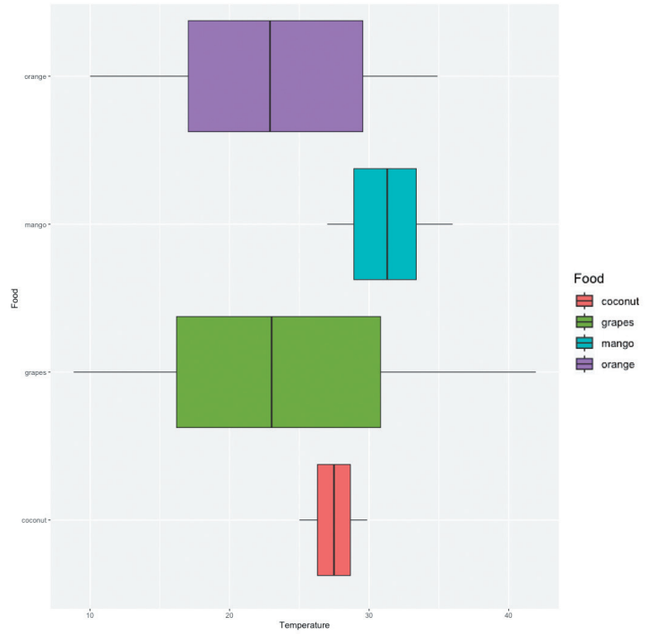
데이터가 갖고 있는 정보의 양과 분산, 데이터의 대푯값, 데이터의 분포, 데이터 간의 상관관계, 데이터의 차원, 가설 및 검정 등을 이해해야만 데이터 분석 결과를 확신할 수 있기 때문이다. 기술분석은 평균값, 중앙값, 최빈값, 분산, 범위, 분위값, 왜도, 첨도 등을 나열하는 것으로 끝나는 것이 아니기 때문이다. 가끔 공공기관에서 데이터 리터러시 강의를 하면서, 공공에서 발표되는 모든 데이터 요약 자료에 Box Plot 등의 시각화 요소나 기초통계량을 같이 병기하면 리터러시 차원에서 좋을 것 같다고 이야기하곤 한다. 바로 통계 리터러시가 필요하기 때문이다.

데이터도 없고 통계에 대한 리터러시도 낮은 상태에서 AI를 이야기할 수 있을까? 단지 AI 응용시스템의 사용자로 국한한다면 그럴 수도 있다. 그러나 CDS 차원에서 데이터 리터러시, 통계 리터러시는 필수 문해력이라고 할 수 있다.

데이터 리터러시나 통계 리터러시는 그렇게 어려운 주제가 아니다. 데이터의 가치를 이해하고 데이터와 통계에 기반한 과학적인 의사결정이 조직문화로 자리 잡을 수 있다면 어렵지 않게 리터러시를 확보할 수 있다. 모두가 데이터분석전문가가 될 필요는 없다. 그러나 모두가 데이터 리터러시, 통계 리터러시의 역량은 확보해야 하는 시대임은 분명하다.



[그림4] 평균과 분산



[그림5] Box Plot을 이용한 데이터 시각화

국 제 통 계 기 구 (I S I) …
2027년 세계통계대회(WSC) 부산 개최

최봉수 | 통계청 국제협력담당관 사무관

한국 통계청은 2027년 제66차 국제통계기구(ISI)[1] 세계통계대회(WSC)[2] 유치에 성공하여 2027년 7월 중 부산 BEXCO에서 대회를 개최한다. 이는 우리나라가 2001년 제53차 서울대회에 이어 26년만에 동 대회를 다시 유치하는 쾌거를 이룬 것이다.

1887년부터 2년마다 개최되는 국제통계기구(ISI) 세계통계대회(WSC)는 전세계의 저명한 통계학자, 각 국 정부·국제기구 및 민간기업의 통계전문가들이 모여, 통계에 관한 이론 및 실무적 발전을 논의하고, 통계 관련 일자리 및 지식공유의 기회를 제공하는 '통계인의 올림픽'이다.

치열했던 유치 과정을 거쳐 2023년 7월 유치국으로 결정

이번 유치 결정은 지난 7월 7일 국제통계기구(ISI) 집행위원회에서 의결되었으며, 우리나라(부산)는

집행위원 만장일치로 유치국으로 결정되었다.

우리나라는 2027년 세계통계대회(WSC) 유치를 신청한 40여개 국 중 최종 후보국인 일본, 싱가포르, 태국과 치열하게 경합하여 선정되었으며, 2023년 7월에 열린 제64차 캐나다 대회(오타와, '23.7.16.~20.)에서 우리나라가 제66차 대회 유치국으로 공식적으로 발표되었다.

그간 유치를 위한 과정을 살펴보면, 2022년말 유치 희망 도시(서울, 부산)를 신청하고 ISI 집행위 심사를 거쳐 최종 유치 신청서 제출 국가 중 하나로 선정되었다. 2023년 3월 한국통계학회, 컨벤션센터 등과 협업하여 공식 입찰신청서를 제출 후 ISI 실사단 현지 방문(5월), 집행위 최종 회의 등을 거쳐 2023년 7월 한국 선정 결과 발표 및 2023 ISI WSC(캐나다 오타와)대회에서 공식 선포를 하였다.

이번 성과는 그간 한국 통계청이 국제연합(UN), 경제협력개발기구(OECD) 등 의장국으로서의 활동을 통해 지속가능발전목표(SDGs), 데이터혁신 등 국제적 통계논의를 주도하고, 국제개발협력(ODA) 확대를 통한 개도국의 통계역량강화 지원 등 국제사회에서의 기여를 인정받은 결과로 평가된다.

한국 통계청은 그간 제53차 ISI WSC(2001), 제3차 및 제6차 OECD 세계포럼(2009, 2018), UN공조 국제세미나 등 다수의 대규모 국제회의를 성공적으로 개최하고 진행한 경험이 있으며, 이러한 대규모 국제행사를 개최한 경험과 노하우 등이 국제사회에서 인정을 받아 이번 2027년 ISI 세계통계대회를 유치하는 계기가 되었다.

1) 국제통계기구(International Statistical Institute, ISI)는 통계의 국제교류 증진을 위해 각 국가 및 국제기구의 통계작성기관, 저명한 통계학자로 구성된 국제통계 조직

2) 세계통계대회(World Statistics Congress, WSC)는 '23년 캐나다(제64차), '25년 네덜란드(제65차)에서 개최될 예정

ISI(국제통계기구)는 무엇을 하는 곳인가?

동 대회를 주최하는 ISI에 대해 간략히 살펴보면, 1885년 런던에서 각 국가의 통계작성 시 통일된 기준이 필요하다는 인식하에 창설되었으며, 상설사무국은 네덜란드 통계청(Voorburg 소재)에 입주해 있으며, 통계이론, 작성방법, 활용에 대한 전문지식의 교환 등 통계의 국제교류 증진을 위해 각 국가 및 국제기구의 통계작성기관, 저명한 통계학자 등으로 구성되어 있다.

주요 활동으로는 매 2년마다 세계통계대회 개최, 각종 산하분과연구회 운영 및 활동을 통해 통계인의 국제적 교류 촉진, 통계인 상호간 전문지식의 교환 및 지식 향상 등을 추진하고 있으며, 주요 분과연구회로는 베르누이학회(1975), 조사통계연구회(1973), 계산통계연구회(1977), 공식통계연구회(1985), 통계교육연구회(1991), 국제산업통계학회(1992), 국제환경통계학회(1989) 등 7개의 분과가 있다.

세계통계대회(WSC)에서는 무엇을 하는가?

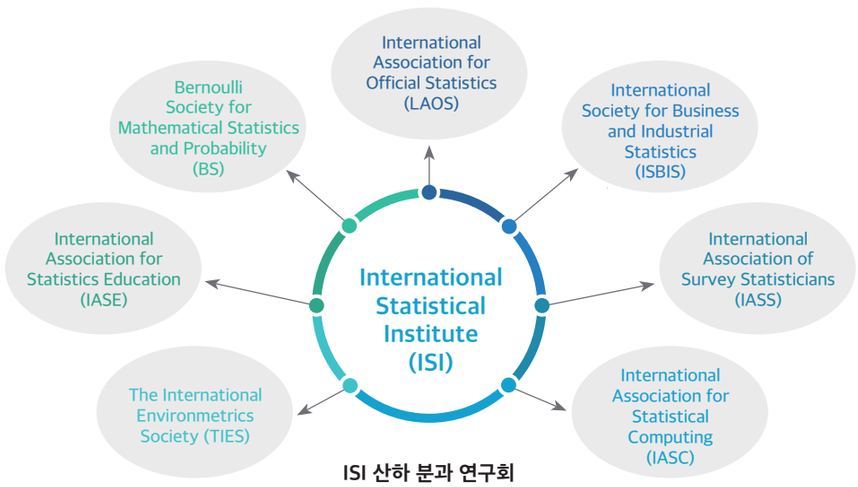
ISI 세계통계대회는 학계, 민간단체, 정부 및 국제기구 등의 통계종사자가 약 6일간의 회의기간 동안 900여 편의 방대한 논문이 발표되며, 학계, 정부 및 국제기구가 공동으로 추진하고 있으며 공식 언어는 영어와 불어이다.

대회의 주요 일정을 살펴보면 주최국 고위층과 ISI 회장단의 개회선언, 기조연설, ISI 총회, 학술회의, 행정회의, 각 국-국제기구 통계기관장 간담회, 문화행사 등 다양한 행사를 개최하며, 학술회의로는 초청논문세션(Invited Papers Session), 특별주제세션(Specal Topic Session), 기고논문세션(Contributed Papers Session) 등이 있다.

과거 국가별 개최 현황 및 2001년 ISI 대회 개최 결과

연도별 개최국 현황을 살펴보면, 아시아 지역에서는 일본이 3회(1930, 1960, 1987), 인도 2회(1951, 1977), 중국 2회(1995, 2013), 필리핀, 말레이시아가 개최했었다. 우리나라는 1969년 제37차 런던대회부터 정부대표단이 참석하고 있으며, 2001년에는 한국의 서울 COEX에서 제 53차 세계통계대회가 개최된 바 있다.

지난 2001년 제53차 ISI WSC*를 성공적으로 개최하여, 참석한 전세계 약 2,700명의 통계전문가들에



ISI 산하 분과 연구회

ISI WSC 세계통계대회

개최장소	참가국	참가자수	발표논문(서울대회)		
			전체	초청(IP)	기고(CP)
53차대회대한민국(서울, 2001)	116개국	2,628명	938	250건	688
52차대회핀란드(헬싱키, 1999)	95개국	2,091명			
51차대회터키(이스탄불, 1997)	95개국	1,608명			

게 활발한 정보공유와 만남의 場을 제공한 바 있다.

▸제53차 ISI WSC(World Statistics Congress) : 2001. 8.22.(수)~8.29(수), 서울 코엑스

또한, 2001년 53차 ISI 대회에서는 노벨경제학 수상자(2명) 초청 특별강연, 189개 주제에 대한 논문발표대회, 통계조사방법론 특별강좌(5개) 및 9개의 위성회의 및 세미나를 개최한 바 있으며, 행사 규모면에서도 직전의 다른 대회와 비교해서 참가 인력, 발표 논문수 등에서 양적, 질적으로 많은 확대가 있었다.

2027년 세계통계대회 개최의 의의

2027년 한국의 부산 BEXCO에서 개최될 세계통계대회에서는 지역통계협력 강화를 위해 대륙별 통계협의체 구축 및 각 국과의 양자간, 다자간 통계기관장 회의 개최를 추진하고 UN 등과 공조하여 인공지능(AI) 및 빅데이터 활용 등 혁신적 통계작성 기법, 기후변화통계 등 다양한 현안에 대한 특별회의 개최는 물론, 노벨경제학 수상자 등 세계적인 석학들을 초빙하여 특별 강연도 실시할 예정이다.

이번 세계통계대회는 포용적이고 지속적인 글로벌 통계발전을 위한 개도국의 통계작성 역량 강화와 혁신적인 통계기법 개발 기반 마련 및 민·관·학의 통계생산자, 이용자와 관련 기관간의 교류를 촉진시

키고, 이론적 통계 기법 및 실무적 통계의 발전에도 적극 기여할 것으로 기대된다.

또한, 세계적인 통계대회를 유치함으로써, 국내에 인공지능(AI), 빅데이터 활용 등 혁신적 통계기술 연구 및 도입을 확대·촉진하고, 국내 통계전문가들에게 국제무대에서의 연구성과 발표 및 국내외 기업에의 취업 기회를 제공하는데 크게 기여할 것으로 기대된다.

요컨대 2001년 제53차 세계통계대회 개최 이후 2번째로 개최하는 2027년 ISI 세계통계대회는 지난 26년간 한국의 눈부신 통계분야 발전 사례를 전 세계에 공유하여 국제통계발전에 다시 한번 더 기여하고 OECD, UN 등 국제기구는 물론, 국제통계사회에서 한국의 영향력을 확대할 수 있는 좋은 기회가 될 것이다.

데이터전문기관, 결합전문기관으로 거듭나는
통 계 데 이 터 센 터

임정주 | 통계청 마이크로데이터과 사무관

데이터 활용의 시대 도래

데이터 3법이 시행('20.8.5.)되면서 4차 산업혁명 시대 신성장 동력인 데이터를 활용할 수 있는 기반이 마련됨에 따라 데이터 이용 수요가 급증하는 가운데 데이터 활용의 핵심인 가명정보[1] 활용에 대한 관심이 높아지고 있는 상황이다.
특히 통계작성, 과학적 연구, 공익적 기록보존 등을 목적으로 하는 경우에는 정보 주체의 동의 없이 가명정보를 연계·분석하는 길이 열렸는데, 그 역할을 '데이터전문기관'과 '결합전문기관'이 수행하게 되었다.

모든 데이터가 연결되는 '디지털 플랫폼 정부'를 추진하는 현 정부에서 가명정보의 활용은 국민, 기업, 정부가 함께 사회문제를 해결하고 데이터 연계·융합을 통한 새로운 가치 창출을 위해 반드시 필요한 것으로 여겨지고 있다.

통계청 통계데이터센터는 개인정보보호 및 데이터 결합 역량을 인정받아 최근 양 기관[2]의 지위를 모두 획득하여 가명정보 결합을 통한 데이터 산업 발전을 견인할 토대를 마련하였다.
인구·가구·사업체 등 국가의 핵심 통계데이터 보유기관인 통계청은 데이터 활용의 시대적 요구에 부응하고 디지털플랫폼 정부 구현에 적극 기여할 수 있는 새로운 역할을 부여받게 된 것이다.

1) 이름 등 개인정보를 암호화하여 특정 개인을 알아볼 수 없게 처리한 정보(예: 홍길동, 25세, 공무원 → AG3EF8, 20대, 공무원)
2) 개인정보보호법에 근거한 '결합전문기관', 신용정보법에 근거한 '데이터전문기관'

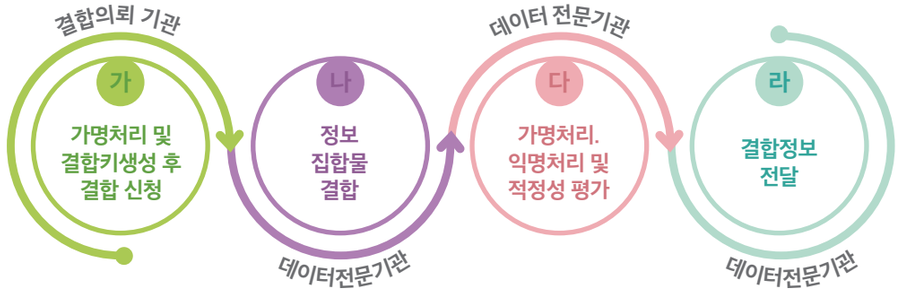
데이터전문기관과 결합전문기관이란?

「데이터전문기관」은 신용정보법 제26조의 4에 따라 정보집합물[3] 결합 및 가명·익명 처리 적정성 평가를 전문적으로 수행하는 기관을 말한다.
금융분야의 가명정보 결합이 가능해짐에 따라 은행·카드·보험·금융투자 등 다양한 금융업권에서 체계적으로 관리되는 정형데이터와 인구가구정보·통신정보·보건의료정보 등 다른 산업 분야의 다양한 형태의 데이터를 서로 융합하여 금융분야의 혁신성장을 이끌 수 있게 되었다.
금융위원회는 2023년 7월 제14차 정례 회의를 통해 통계청을 포함하여 BC카드, 삼성 SDS, 삼성카드, 신한은행, 신한카드, LG CNS, 쿠콘 등 8개 기관을 데이터전문기관으로 추가 지정하여 기존 4개 전문기관(금융보안원, 금융결제원, 신용정보원, 국세청)을 12개로 확대하였다.

최근 D-테스트베드[4] 운영 관련 한국핀테크지원센터에서 통계청이 보유한 인구·가구 정보의 제공 가능 여부 검토를 요청하는 등, 통계청은 타 기관에 비하여 인구·가구 분야에서 우월한 데이터를 보유하고 있어 관련 데이터 수요는 증가할 것으로 보인다.
지난 3년간 데이터전문기관은 4개 기관이 총 287건의 정보집합물 결합을 완료하였고, 금융분야와 비금융분야 이종의 데이터 결합 수요는 꾸준할 것으로 예상되어 통계청 데이터전문기관(통계데이터센터)의 역할에 기대가 큰 시점이다.

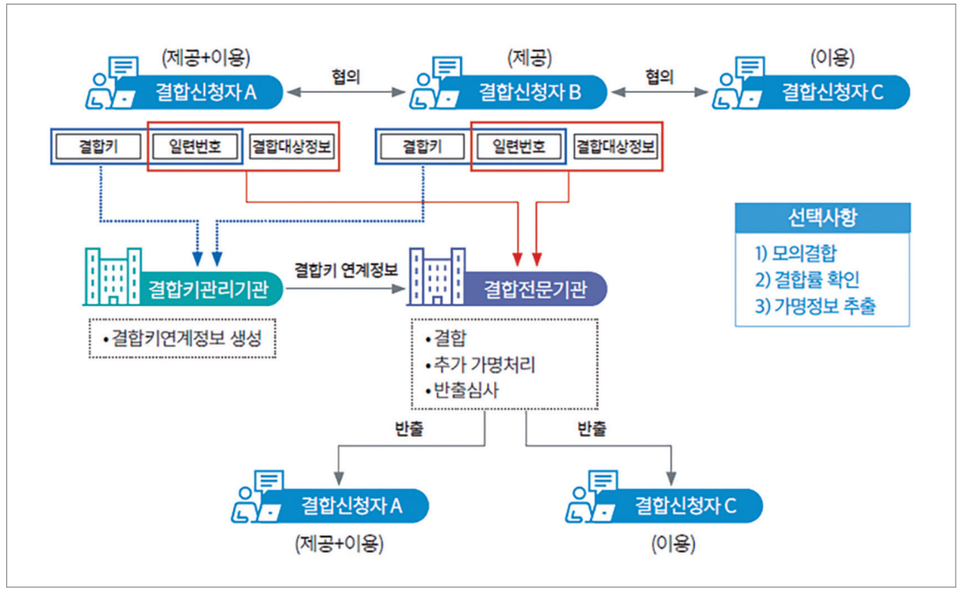
「결합전문기관」은 개인정보보호법 제28조의3 제1항에 따라 서로 다른 개인정보처리자 간의 가명정보 결합을 수행하기 위해 개인정보위 또는 관계 중앙행정기관의 장이 지정하는 기관을 말한다.
통계청은 2020년 11월 결합전문기관에 지정되어 암 질병(폐암) 치료효과 분석(사망원인자료 활용)을 시작으로, 서울시 거주 1인 가구 및 가구원수별 통계(인구·가구 정보 활용) 및 제주 한달살이 분석(인구가구통계등록부 등 활용) 등 다양한 분야에서 가명정보 결합을 추진하여, 정책을 뒷받침하는 맞춤형 분석을 가능하게 하였다.

현재 건강보험 급여에 따른 소득분배영향 분석(가계금융복지조사 활용)을 위한 데이터 결합을 진행하고 있고, 알코올 중독자 치료기기 효과 분석을 위해 건강보험공단에 사망원인자료를 제공하여 관련 연구를 지원하고 있다.
지난 3년 동안 결합전문기관은 총 51건의 결합을 진행하였으며, 통계청은 그 중 9건을 수행하여 전체



데이터 전문기관 업무체계(출처 : 금융분야 가명·익명처리 안내서)

3) 정보를 체계적으로 관리하거나 처리할 목적으로 일정한 규칙에 따라 구성되거나 배열된 둘 이상의 정보
4) 스타트업, 예비창업자 등이 핀테크 아이디어의 사업성, 실현 가능성 등을 검증할 수 있도록 별도의 테스트 기간 동안 금융·비금융 결합 데이터 및 테스트 환경을 지원하는 사업



결합전문기관 업무체계(출처 : 가명정보처리 가이드라인)

결합전문기관 및 데이터전문기관 비교(※ 통계청, 국세청, 삼성SDS, LG CNS, BC카드 등 5개 기관은 양 기관 지위 모두 획득)

구분	결합전문기관	데이터전문기관
근거법령	개인정보 보호법	신용정보의 이용 및 보호에 관한 법률
소관기관	개인정보보호위원회	금융위원회
목적	통계작성, 과학적 연구, 공익적 기록보존	통계작성, 연구, 공익적 기록보존
업무	가명정보 결합(개인정보처리자 간)	정보집합물 결합 (신용정보회사등과 제3자)
대상	비금융데이터+비금융데이터	금융데이터+(비)금융데이터
결합키 관리	별도 결합키관리기관(한국인터넷진흥원)	데이터전문기관에서 처리
자기보유 정보결합 (자체결합)	제3자 제공 목적으로만 자신이 보유한 정보를 자신이 직접 결합 가능. 단 외부전문가를 통해 결합대상정보의가명처리 수준 등 검토해야함	이해상충 가능성이 없고 적정성 평가를 타 전문기관이 수행하는 경우 자가결합 가능
지정	'20년 11월 결합전문기관 지정(개보위)	'23년 7월 데이터전문기관 지정(금융위)
지정기관 현황 ('23.8월말 기준)	통계청, 국세청, KCA, KICA, 롯데정보통신, 한국지역정보개발원, 한국사회보장정보원, 삼성SDS, LG CNS, CJ, 국민건강보험, 건강보험심사평가원, 국립암센터, 한전KDN, 한국도로공사, 한국데이터산업진흥원, NIA, DOUZONE, SK C&C, BC카드, KERIS국가정보자원관리원 등 **총 22개 기관**	통계청, 국세청, 신용정보원, 금융보안원, 금융결제원, BC카드, 삼성 SDS, LG CNS, 삼성카드, 신한은행, 신한카드, 쿠콘 등 **총 12개 기관**

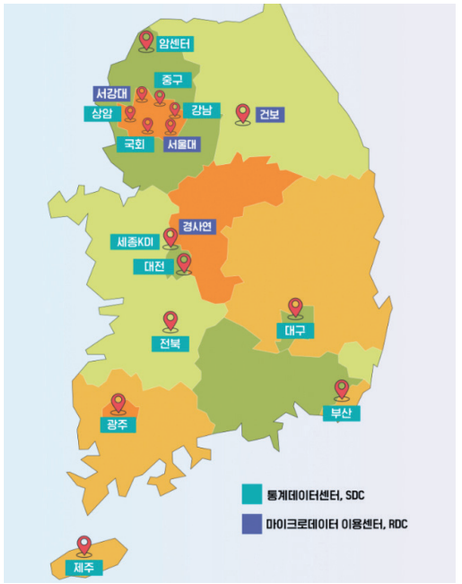
결합사례에서 17.6%를 차지하고 있다. 결합전문기관은 데이터 전문기관에 비해서 활용도는 조금 떨어지나 정부가 가명정보 활용 확대방안('23년 7월)을 통해 가명정보관련 세부 활용기준 마련과 전문인력을 양성하는 등 적극 지원하기로 함에 따라 향후 그 활용은 확대될 것으로 기대된다.

통계청은 결합전문기관과 데이터전문기관의 지위를 겸하고 있는 중앙행정기관으로서, 가명정보를 통하여 데이터 활용 활성화에 기여하고자 조직과 예산 등 전문기관 운영을 위해 필요한 인프라를 구축해 나갈 것이다. 특히 실제 전문기관을 운영하고 있는 '통계데이터센터'는 국민 모두가 안전하게 데이터와 친숙해질 수 있는 공간이 되도록 그 역할을 다할 것으로 생각된다.

국민 편의를 위한 센터 일원화 추진

통계청의 데이터센터는 통계데이터센터(SDC)와 마이크로데이터 이용센터(RDC)로 구분되어 있어, 일반 국민들이 이용하는데 약간의 혼란을 주고 있다는 지적을 받아 왔다.

이에 따라 최근 RDC를 SDC로 통합하는 센터 일원화 작업을 추진하고 있으며, 지역 RDC를 SDC로 통합하는 작업을 지속적으로 추진할 예정이다. 2023년 상반기에는 제주센터, 전북센터, 국립암센터를 통합 개소하여 전국적으로는 12개의 센터를 운영함으로써 지역별 연구자들의 접근성을 제고하고 있다. 12개의 지역센터는 데이터 안심존으로서 전문기관에서 결합된 데이터 결합물을 결합신청·의뢰자가 안전하게 심층분석을 할 수 있는 분석공간을 제공하고 아울러 지역의 데이터허브로서 역할을 강화하도록 더욱 노력 할 것이다.



SDC 및 RDC 현황('23년 8월말 기준)

통계데이터센터 현황('23년 8월말 기준)

경인권	• **서울중구센터**(한국데이터산업진흥원 7층) • **서울강남센터**(서울세관 별관 4층) • **국립암센터**(국립암센터 연구동 7층)	• **서울상암센터**(S-Plex 센터 15층) • **서울국회센터**(국회도서관 1층)
충청권	• **본부/대전센터**(통계센터 13층)	• **세종KDI센터**(KDI 1층)
동남권	• **부산센터**(부산통합청사 1층)	
동북권	• **대구센터**(SW융합테크비즈센터 2층)	
호남권	• **광주센터**(광주통합청사 4층) • **전북센터**(전북테크비즈센터 6층)	• **제주센터**(제주연구원 2층)

데이터 시대의 통계학, 데이터 기반 통계교육

탁병주 | 전주교육대학교 수학교육과 교수

2016년 3월, 구글 딥마인드 챌린지 매치를 통해 세계적인 프로 바둑기사 이세돌 9단을 4:1로 꺾은 바둑 프로그램 알파고(AlphaGo)는 챗GPT(ChatGPT)가 등장하기 전까지 한동안 인공지능의 대명사로 불렸다. 수가 한정되어 있는 장기나 체스에 비해 바둑은 상식적으로 계산이 도저히 불가능한 수준의 경우의 수가 존재하기에, 그동안 바둑의 수읽기는 사실상 '직관'이나 '통찰'과 같은 모호한 용어로 설명할 수밖에 없었고 이는 인간 고유의 사유(思惟)로 인식되어 왔다.

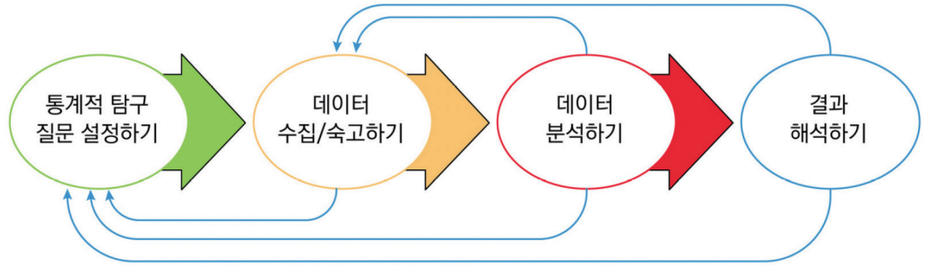
그러나 컴퓨터에 수많은 기보들을 말 그대로 '때려 넣었을' 뿐인데 인공지능은 머신러닝 기술을 이용하여 거기서 일정한 패턴을 찾아내고, 대국을 통해 기보가 추가될 때마다 그 패턴을 갱신하는 방식으로 인간의 수읽기를 능가해버렸다. 컴퓨터의 메모리와 처리 가능 용량이 천문학적으로 늘어남에 따라 우리 사회가 데이터(data)의 위력을 실감하고 충격에 빠지게 된 것이다. 이것이 벌써 7년 전 일이다.

알파고와 이세돌의 대국

데이터가 인간 사회에 미치는 영향은?

데이터의 가장 큰 가치는 합리적인 예측에 있다. 그리고 현대 사회에서 무언가를 예측한다는 것은 인간에게 더욱 중요해지고 있는데, 이는 2019년 말부터 약 3년여 간 지속되었던 코로나 팬데믹을 통해서도 확인할 수 있다. 국내 코로나19 확진자가 매우 적었던 2020년에만 해도 확진자 발생에 따라 국민 모두가 숨죽이며 일상생활의 대부분을 자제해야 했고, 역학조사를 통해 확진자의 동선을 공개하는 과정에서 수많은 개인정보 유출 논란이 있어 왔다. 그러나 2022년 이후에는 확진자가 하루에도 수만 명씩 발생함에도 개인 위생에 주의하라는 안내가 고작이다. 백신 접종의 유무도 중요한 변수이기는 하지만, 이는 데이터의 축적으로 코로나19에 대한 많은 정보를 알게 되었고 그만큼 현상에 대한 예측 또한 가능해졌기 때문이기도 하다.

세상은 온통 불확실한 것들로 가득 차있고, 사람들은 예측할 수 없는 불확실성에 불안함을 느낀다. 심지어, 예고된 위험보다도 위험에 대한 불확실성에 더욱 심리적인 불안감을 크게 가지기도 한다. 그러나 데이터는 그럴 듯한 예측을 가능케 함으로써 100%는 아니지만 어느 정도 예측이 가능하고 적어도 그 예측이 참인지 거짓인지 확인하기 전까지는 지각된 통제감을 가질 수 있게 하여 불안 증세를 완



화시켜 준다.
데이터가 있으면 적어도 마음은 편안해진다는 것이다. 현대 사회가 개인의 행복에 많은 가치를 부여하고 있다는 점을 고려하면, 데이터의 진정한 가치는 인간의 정서를 안정시켜준다는 심리적인 측면에 기대는 부분이 크다고도 볼 수 있다.

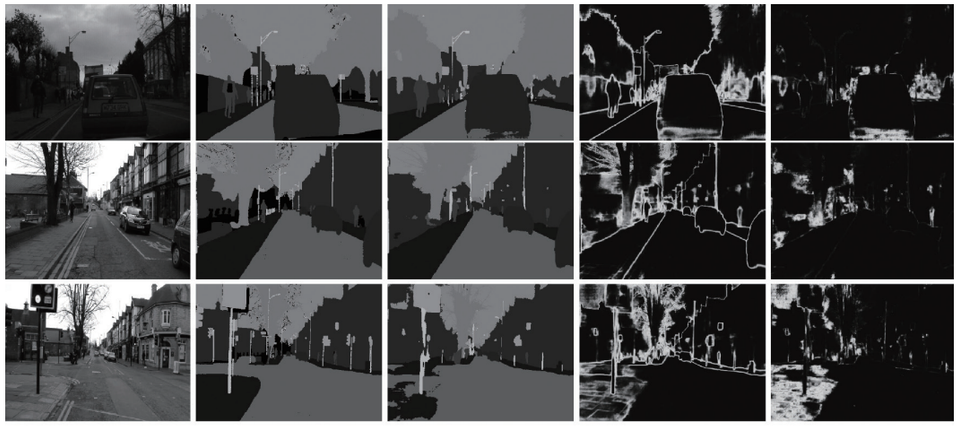
데이터 시대, 빈도주의에서 베이즈주의로

그러나 데이터에 기반을 두어 의사결정을 한다는 것은, 실은 확실성과 정확성을 포기한다는 뜻이기도 하다. 1, 3, 5, 7, 9, □에서 빈 칸에 들어갈 알맞은 수는 앞서 제시된 1, 3, 5, 7, 9라는 데이터를 바탕으로 패턴을 파악하여 자연스럽게 11이라고 추측할 수 있다. 그러나 반드시 11이라는 법은 없다. 1, 3, 5, 7, 9, 101, 103, 105, 107, 109, 201, 203, 205, 207, 209, …와 같이 크게 보면 국소적으로 보았을 때와는 다른 새로운 패턴이 존재할 수도 있다. 가장 고전적인 데이터과학이라 할 수 있는 통계학에서도 다양한 추리와 예측을 시도하지만, 언제나 틀릴 가능성을 내포할 수밖에 없다. 그럼에도 통계학을 '과학'으로 인정하는 이유는 데이터에서 패턴

토머스 베이즈 (사진출처:wikipedia)

을 찾아 모델링하거나 데이터를 이용하여 불확실성을 수량화하는 과정에서 수학의 연역적인 '형식'을 빌렸기 때문이다.

수리과학으로서 갖추어야 하는 연역적인 연단에 많은 가치를 부여함에 따라, 오랫동안 칼 피어슨(Karl Pearson, 1857~1936), 로널드 피셔(Ronald Fisher, 1890~1962), 예지 네이만(Jerzy Neyman, 1894~1981)과 같은 빈도주의자들에 의해 수리통계학이 20세기 초까지 통계학의 헤게모니를 장악해왔다. 그러나 제2차 세계 대전 이후 실용주의(pragmatism)가 대두하면서, 비록 논리적 기



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

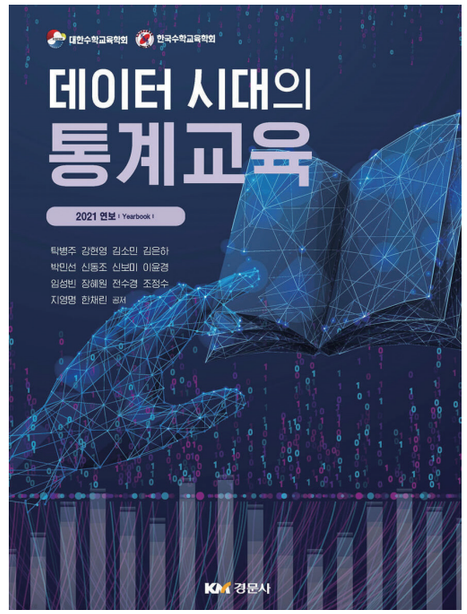
반은 굳건하지 못하나 데이터를 이용하여 무언가를 예측하는 데 탁발한 위력을 보여주는 베이지안 통계학이 대두한다.

특히, 오늘날 빅데이터 처리 기술과 머신러닝의 발달로 인해 베이지안 통계학을 활용한 데이터 기반 예측은 불확실성을 제어하고 인간의 불안을 잠재워주는 강력한 도구로서의 기능을 하고 있다. 새로운 정보를 받아들이면 기존의 믿음을 갱신하는 자연스러운 인간의 사고방식을 수학적으로 표현하고자 했던 토머스 베이즈(Thomas Bayes, 1701~1761)의 아이디어가 현대에 와서야 빛을 발하게 된 것이다. 데이터를 기반으로 인간의 사고방식을 구현하는 인공지능의 등장, 데이터 시대가 도래한 것이다.

데이터 시대, 데이터 기반 통계교육의 철학

통계학은 태생적으로 데이터에 기반을 둔다. 그러나 그럼에도 필자는 통계교육의 나아갈 방향을 일컫기 위해 굳이 '데이터 기반'이라는 수식어를 붙이고자 한다. 이는 통계학이 데이터에 기반을 둠에 따라 수학과는 상반된 비결정론적 세계관, 맥락 의존성과 같은 특성을 지니고 있음에도, 이러한 통계학 고유의 논리와 성격이 학교 교육에서는 거의 고려되지 않았기 때문이다. 게다가 최근에는 머신러닝의 등장으로 데이터의 유형과 본성이 변화하였고 분석 방법 또한 다변화하고 있기에, 빅데이터를 포함한 데이터 소양을 아우를 수 있는 통계교육의 새로운 지향점을 모색할 필요가 있다. 이를 데이터 기반 통계교육(data-driven statistics education)이라 명명하고자 한다.

그동안 통계교육 연구자들은 데이터 기반 통계교육을 실천하기 위해서 학생들에게 '통계적 추리'와 '통계적 문제해결'의 경험을 충분히 제공해야



한다고 주장해왔다. 수학의 논리와 형식을 강조하는 관점에서 통계적 추리는 곧 형식적인 통계적 추정을 의미하는 것이었고, 따라서 초등학교와 중학교에서 이루어지는 통계교육은 추리보다 기술(description)에 초점이 맞춰져 있다.
그러나 실용주의적 기조가 강조됨에 따라 베이지안 통계학이 대두하였듯이, 통계교육 역시 학문적으로 엄밀히 인정될 만큼의 형식성은 갖추지 못하더라도 그 나름의 조리와 정연함을 바탕으로 추리하고 그 근거에 대해 추론해보는 비형식적인 통계적 추리가 초등학교 수준에서부터 강조될 필요가 있다. 또한, 통계학은 본래 문제해결을 위한 학문이라는 점을 고려할 때 탐구 문제를 설정하고 자료를 수집, 분석하여 결과를 해석하는 일련의 과정을 순환적으로 경험하는 것이 무엇보다도 중요하다. 구체적으로, 데이터 기반 통계교육을 실천하기 위해서는 다음과 같은 시사점을 고려해야 한다(이경화 외, 2021)[1)]

1)이경화, 유연주, 탁병주(2021). 데이터 기반 통계교육을 위한 수학과 교육과정 재구조화 방향 탐색. 학교수학, 23(3), 361-386.

● 데이터를 직접 수집, 정리하여 집단의 특성을 기술하는 것뿐만 아니라, 이미 수집, 정리된 데이터를 이해하고 평가하며 분석하는 경험을 제공할 필요가 있다.

● 이미 수집된 데이터를 보고 데이터가 최초에 어떤 목적과 방법으로 수집된 것인지를 알아내는 추론 기회를 제공할 필요가 있다.

● 데이터의 편향(bias)을 인지하고 통계적 추리 과정에서 이로 인해 오류가 발생할 수 있음을 유의하도록 지도해야 한다.

● 통계적 문제해결을 위하여 정형화된 데이터뿐만 아니라 비정형화된 데이터를 활용하는 사례들을 다룰 필요가 있다.

● 통계적 탐구 질문에 따라 데이터를 직접 수집하는 경험뿐만 아니라, 이미 다른 목적에 의해 형성된 데이터를 바탕으로 통계적 탐구 질문을 도출하는 경험을 제공해야 한다.

● 학생들이 데이터의 절차적인 처리보다는 문제해결의 목적과 데이터의 맥락에 주목할 수 있도록 테크놀로지를 적극 활용할 수 있어야 한다.

● 통계적 문제해결 과정에서 발생할 수 있는 데이터의 왜곡, 개인정보 보호 등 윤리 문제의 중요성을 함께 지도해야 한다.

수학과 교육과정의 개정과 학교 통계교육의 변화

2022년 12월, 교육부에서 발표한 새로운 수학과 교육과정에서는 미래 지향적 수학 학습 내용 반영을 주요 개정 방향으로 설정하면서 '실생활 자료 중심의 통계교육 내용 재구조화'를 추구하였다. 그 결과, 다소 부족하기는 하지만 성취기준이나 교수-학습 방법 및 유의사항에 통계적 추리, 문제해결과 관련된 요소들이 일부 반영되었다(통계의 창 2023년 여름호 中 「2022 개정 수학과 교육과정에 나타난 실용통계교육의 방향」 참고).

예를 들어, 초등학교 5~6학년군 성취기준으로 "탐구 문제를 설정하고, 그에 맞는 자료를 수집, 정리하여 적절한 그래프로 나타내고 해석할 수 있다", "자료를 이용하여 가능성을 예상하고, 가능성에 근거하여 적절한 판단을 내릴 수 있다"와 같은 문장을 추가하여 초등학교 수준에서부터 데이터를 이용하여 문제를 해결하고 이를 바탕으로 추측해보는 경험을 충분히 제공하기 위해 노력하였다. 유의사항으로도 "자료를 수집할 때, 간단한 설문조사, 실험이나 관찰, 공공 자료 활용과 같은 방법 중 탐구 목적에 적합한 것을 결정하게 한다", "자료 수집의 목적과 수집한 자료의 특성에 맞는 그래프로 적절히 표현되었는지, 또는 정보를 왜곡하는 오류가 포함되어 있지는 않은지 등을 비판적으로 판단하게 할 수 있다"와 같은 문장을 통해 통계교육의 목적으로서 데이터 소양 함양에 조금 더 초점을 맞추고자 노력하였다.

그러나 이와 같이 데이터의 유형과 본성이 변화하고 그 가치가 높아지는 데이터 시대에 맞게 데이터 기반 통계교육이 학교 현장에 제대로 안착하기 위해서는 무엇보다도 데이터에 대한 교사와 학생들의 접근성이 높아져야 한다. 특히, 문제로 확연히 느껴질 만한 맥락이 주어질 때 학생의 입장에서 데이터의 가치를 더욱 크게 인식할 수 있음에도 불구하고, 교육부의 검정을 통과해야 하는 수학 교과서에서는 학생들이 문제라고 느낄 만한 상황을 부정적이라는 이유로 회피하는 경향이 있다. 어른의 시각으로 학생들을 멸균실에 가두려 드는 현재의 교육 문화에서는 학생들이 다양한 데이터를 접하고 자유롭게 다룰 수 있는 기회가 매우 제한적인 셈이다. 이러한 점에서 통계청에서 제공하고 있는 국가통계포털(KOSIS)이나 통계지리정보서비스(SGIS), KOSIS 통계놀이터 등의 서비스는 현재의 경직된 통계교육 문화를 탈피하고 데이터 기반 통계교육의 실천을 위한 좋은 초석이 될 수 있다.

실제로 학교 현장의 많은 선생님들이 통계청에서 제공하고 있는 다양한 공공 자료를 통계 수업에서 활용하는 사례가 점점 많아지고 있는 만큼, 모쪼록 데이터 기반 통계교육의 철학이 학교 교육에 뿌리내림으로써 모든 이들이 데이터의 가치를 인식하고 이를 합리적으로, 그리고 비판적으로 활용할 수 있는 소양이 갖추어지기를 희망한다.

KOSIS 통계놀이터

AI 스마트팜 기술이 농업의 패러다임을 바꾼다

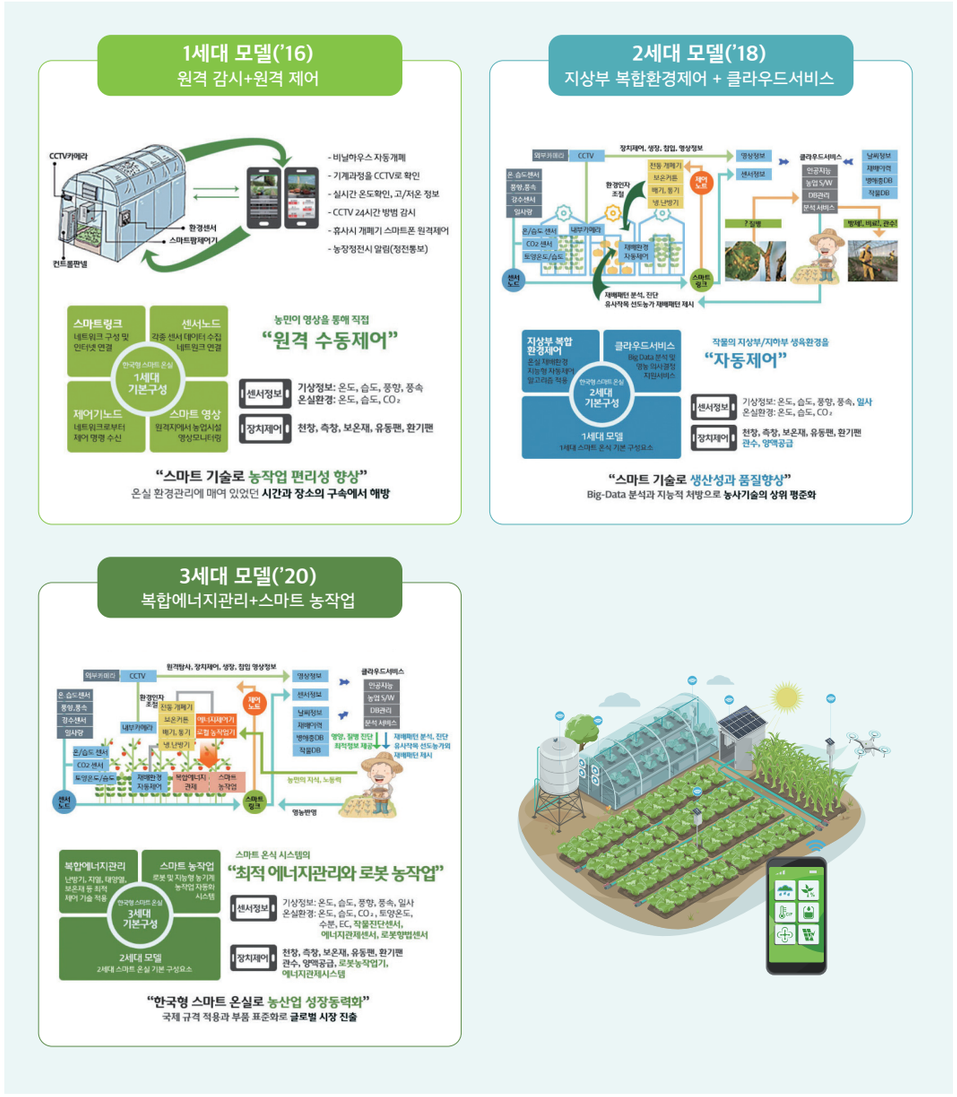
김준수 | 주식회사 어밸브 수석연구원

21세기에 들어 산업계에 가장 큰 영향을 미친 과학기술 하나를 손꼽아야 한다면 단언컨대 인공지능(AI, Artificial Intelligence)일 것이다. 인공지능이란 데이터를 통하여 인간이 분석하는 방식과 유사하게 추론, 학습 및 행동할 수 있는 컴퓨팅 기술을 의미한다. 이는 단지 엔지니어링 분야뿐만 아니라 바이오테크, 언어학 그리고 철학과 심리학을 포함한 광범위한 학문에 적용되고 있다. 인공지능 기술은 주어진 알고리즘과 데이터를 기반으로 학습하여 해당 데이터에 근거한 결정을 내리는 머신 러닝(machine learning)부터 발전하여 현재는 알고리즘 계층을 통해 자체적으로 학습을 하고 지능적인 결정을 내릴 수 있는 딥 러닝(deep learning)까지 발전하였다. 그렇다면 인공지능이라는 강력한 과학기술은 농업이라는 분야에 얼마나 많은 변화를 가져왔고 앞으로 농업혁신에 어떠한 변화를 가져올 수 있는지 살펴보자.

인공지능을 활용한 농업혁신의 필요성

인공지능 기술이 산업의 다방면에 적용되어 혁신을 이끌고 있는 반면 1차 산업인 농업에 대해서는 기술적인 혁신이 더디다. 예로부터 인간의 노동력이 주를 이루었던 농업은 종자를 선별하는 과정에서부터 씨앗을 심고, 작물을 재배하고 최종적으로 선별 및 수확하는 일련의 과정을 농업 전문가들이 자신만의 노하우에 근거하여 행해왔다. 공산품과는 달리 각각의 살아있는 생물체인 작물을 기르는 과정이기 때문에 표준화된 데이터가 아직까지 충분하지 않으며 이 때문에 농업의 완전한 자동화는 어려운 실정이다.

그럼에도 다양한 스마트팜 기술이 농업에 접목되며 생산성을 극대화하고 농업에 필요한 인력을 절감하여 기술을 통한 농업혁신의 가능성을 보여주고 있다. 대표적으로 IoT 기술이 농가에 보급



되어 온도 조절, 광량 조절, 관수 개폐 등을 핸드폰 터치 스크린만으로 조절할 수 있게 되었으며 센서들은 미세한 환경변화도 측정하여 사전에 작물 로스(loss)가 발생하는 것을 예방한다.

다만, 현재로서는 농업혁신이 농장의 부분 자동화에 그쳐있고 상당한 부분은 사람의 노동력 없이는 이행되기 어렵다. 이는 작물을 '관찰'하고 그 상태를 '파악'하여 최종적인 '판단'을 내리는 것이 여전히 사람에 의해 이루어지기 때문이다. 이러한 일련의 과정을 인공지능이 대체할 수 있게 되어야 비로소 농장 자동화의 길이 열릴 것이며, 그렇기 때문에 현재 농업혁신의 가장 큰 관문이 인공지능인 것이다.

국내 스마트팜 현황

우리가 소위 말하는 스마트팜이란 기존 농장에 IoT, ICT 정보통신 기술을 결합하여 원격·자동으로 작물의 생육환경을 관리하는 기술 복합체를 의미한다. 기술적으로 다변화하고 있는 스마트팜도 하나의 정형화된 기술이 아니며 세대에 따라 기술적인 발전을 하고 있다.

1세대 스마트팜은 IoT 기술의 등장과 함께 '원격 수동제어'를 현실화하였지만, 여전히 농부들의 직접적인 관찰과 판단을 요한다. 현재 실현되고 있는 2세대 스마트팜은 복합환경제어와 클라우드 서비스를 통하여 재배 과정의 '자동 제어'를 가능케 했다.

인공지능 및 작업 로봇이 탑재되어 농장의 완전 자동화를 가능하게 하는 3세대 스마트팜은 아직까지 실현되지 않았지만 우리가 나아가야 할 방향성을 제시해주고 있다. 스마트팜의 완전 자동화를 위해서는 인공지능이 사람의 눈을 대신하여 자체적인 판단을 내리고, 로봇이 사람의 팔을 대신하여 인공지능의 판단에 따른 조치를 취해야 한다.

진정한 의미의 스마트팜이 되기 위해서는 인공지능의 개발뿐만 아니라 인공지능이 내린 판단을 이행해줄 수 있는 하드웨어 시스템도 함께 구축되어야 한다는 뜻이다. 그러한 '스마트'한 농장이 현실화가 되어야 비로소 우리는 침체하고 있는 농산업에 대한 완전한 기술적인 해결책을 고안해냈다고 말할 수 있을 것이다.

국내 스마트 딸기 유리온실(출처:어밸브)

그렇다면 우리나라 스마트팜의 현 위치는 어떠할까. 현재로서는 1세대 스마트팜 시설에 머물러있지만, 2세대 스마트팜 기술 개발을 통해 인공지능이 농사짓는 시대로 나아가기 위해 다방면으로 노력하고 있다. 1세대 스마트팜을 도입한 농가에서는 모든

국내 수직농장(출처:어밸브)

국내 스마트팜(출처:어밸브)

농사 환경을 농가에서 수동으로 관리해야 하지만 2세대 스마트팜이 도입되면 의사결정을 모두 인공지능이 내려주기 때문에 농장주의 편리성이 극대화된다. 만약 빠른 시일 내에 우리나라에서 2세대 스마트팜 기술이 정착된다면 해외 각국에 한국형 스마트팜의 글로벌 수출까지 노려볼 수 있을 것이다.

국외 스마트팜 살리나스 밸리

국외 스마트팜 Hortiplan

해외 스마트팜 현황

유리온실, 수직농장(실내형 식물공장)에 한정된 국내 스마트팜과는 달리 해외에서는 더 다양한 형태로 시스템 농업이 개발되고 있다. 일례로, 미국의 살리나스 밸리에서는 노지에 무인드론을 활용하여 농약을 자동으로 살포하고 벨기에 유리온실에서는 로봇을 사용하여 농장 관리에 필요한 투입 인력을 최소화하고 있다. 국내 스마트팜에 비해 해외에서 더 다양한 기술들이 상용화되고 있는 가장 큰 이유는 외국에서는 넓은 농업 부지가 확보되기 때문이다. 넓은 부지로부터 발생하는 수확량의 차이가 투자자본 규모의 차이를 만드는 것이다.

해외에서는 이처럼 스마트팜과 관련하여 더 큰 규모에서 더 다양한 R&D 연구가 이루어지고 있으며 농업 선진국의 위상을 보여주고 있다. 아직 우리나라에서 대부분의 대규모 농업혁신 사업은 정부 주도 개발 사업으로 진행되고 있다. 경상도, 전라도 지역에 집중적으로 산재되어 있는 좁은 농업 부지만으로는 대기업이 자본을 투입할 조건이 안되며 중소기업들이 성장할 동력이 확보되기 어렵기 때문이다. 이러한 환경적인 제약은 앞으로 우리나라만의 우수한 기술력과 적극적인 민관·산학 협력을 통해 극복해 나가야 할 것으로 보인다.

세대별 스마트팜 도표 통합(출처 스마트팜연구개발자집단)

구분	1세대	2세대	3세대
목표 실현시기	현재	2025년 → 2030년	2030년 → 2040년
목표효과	편의성 향상 '좀 더 편하게'	생산성 향상 '덜 투입, 더 많이'	지속가능성 향상 '누구나 고생산·고품질'
주요기능	원격 시설제어	정밀 생육관리	전주기 지능·자동관리
핵심정보	환경정보	환경정보, 생육정보	환경정보, 생육정보, 생산정보
핵심기술	통신기술	통신기술, 빅데이터/AI	통신기술, 빅데이터/AI, 로봇
의사결정/제어	사람	사람/컴퓨터	컴퓨터
대표 예시	 스마트폰 온실제어 시스템	 빅데이터 기반 지능형 생육관리소프트웨어	 무인자율형 로봇농장

2세대 스마트팜 실현을 위한 과제

인공지능 기술 개발에 따라 국내외 농업벤처들의 2세대 스마트팜을 위한 경주가 진행되고 있다고 해도 과언이 아니다. 2세대 스마트팜 실현의 가장 큰 걸림돌은 데이터의 표준화다. 인공지능이 자체적인 판단을 내릴 수 있게 하기 위해서는 인공지능이 학습할 수 있는 표준화·정형화된 데이터가 필요하다. 농업의 특성상 작물별, 시설별, 환경별로 수집되는 데이터의 종류가 제각각이기 때문에 아직은 표준화된 데이터풀(data pool)이 형성되었다고 보기 어렵다. 이에 따라 정부 지자체에서 표준화된 기준을 세우기 위한 노력이 계속되고 있고 국내 여러 기업에서는 자체적으로 작물 데이터풀을 수집하여 완성도 높은 인공지능 모델을 구축하려는 노력이 빛나고 있다.

이러한 노력이 축적되어 머지않은 미래에 인공지능 모델을 활용한 3세대 스마트팜이 실현될 것이며 이는 우리나라가 농업 선진국으로 자리매김하는 것에 디딤돌 역할을 할 것이다. 현재로서는 스마트팜 기업들간의 협력 뿐만 아니라 엔지니어링, 로보틱스, 정보통신 기업 및 국가 연구 기관과의 스마트팜 컨소시움(consortium)을 형성하는 것을 우선순위로 두어야 할 때이다. 나아가 농업 선진국들에 뒤처지지 않고 자국 내 농업혁신을 이루기 위해 국가와 국민 차원의 관심과 지원이 절실하다.

빅데이터는 우리가 모르는 것에 대해 얼 마 나 말 해 줄 수 있 나

오세욱 | 한국언론진흥재단 책임연구위원

개인적으로 야구를 참 좋아한다. 한 팀을 30년 넘게 좋아하고 있는데 요새 성적이 별로라 최근에는 자주 보지는 않고 있다. 야구를 워낙 좋아했지만, 좋아하는 팀 성적에 따라 일희일비하는게 싫어 떠나려 한 적도 있었다. 그 시기 사회과학방법론 수업을 들었고 통계가 무엇인지 어렴풋이 알게 되었다.

숫자를 갖고 참 많은 것을 해석해 볼 수 있다. '머니볼'이라는 영화를 본 다음에는 내가 배웠던 통계를 야구에 적용해 보는 재미에 빠지게 됐다. 야구 경기에서 측정된 모든 기록은 숫자로 남아 있기에 다양한 방식으로 계산하고 이를 바탕으로 향후 성적을 예측해 볼 수 있었다. 그런데 이 예측은 항상 빗나갔다. 분명히 통계적 예측 모델에 따르면 내가 좋아하는 팀의 성적은 최상위권이어야 했지만 결과는 항상 실망이었다. 야구의 기록은 정말 많은 것을 얘기해주지만 기록 외 변수가 너무 많았다.

예를 들어, 그 팀의 운영진, 감독과 선수들이 서로 얼마나 신뢰하는지는 숫자로 표현되지 않는다. 게다가 난 특정 한 팀을 너무 좋아해서 숫자를 우리 팀에 유리한 방향으로 해석하는 오류도 자주 범했다.

데이터 중심 사회에서 중요해지는 것은 결국 해석력

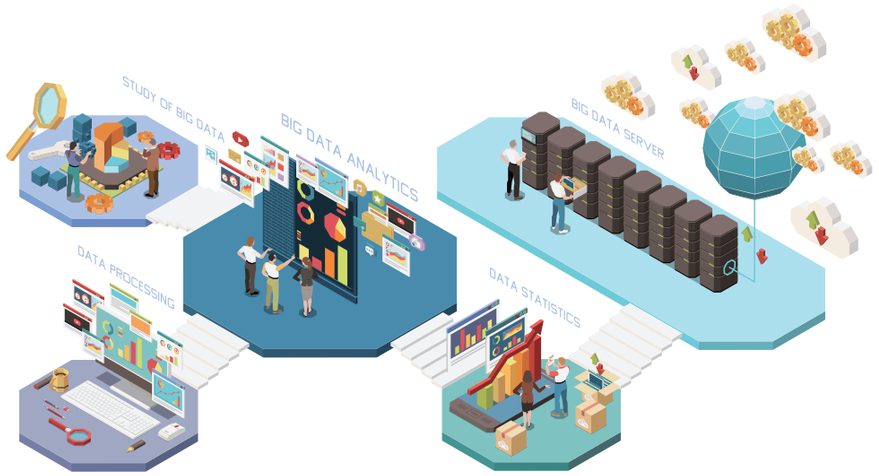
우리 대부분은 경험하고 있지만, 말로 하는 것보다 구체적인 숫자로 얘기할 때 상대방을 설득하기가 쉽다. 숫자는 말보다 정확하고 객관적이고 공정하다는 일반적 인식이 있기 때문이다. 데이터 중심 사회가 되면서 이러한 경향은 더욱 강해지고 있다.

우리가 알고 있는 대부분의 기업은 가능한 많은 데이터를 모으려고 한다. 소비자를 이해하고 설득하기 위한 가장 강력한 수단은 숫자이기 때문이다. 네이버 등 테크 기업들의 서비스를 이용하기 위해 회원 가입을 할 때 반드시 동의해야 하는 약관에는 수집하는 개인 정보 목록이 제시돼 있다. 뭐 이런 것까지 수집하나 할 정도로 상세하게 제시돼 있지만, 우리 대부분은 무심코 동의하고 있다. 이렇듯 상세하게 정보를 수집하는 이유는 하나다. 가능한 숫자를 많이 확보해야 그 사람의 행동을 예측하기 쉽고 이 예측에 따른 추천을 통해 자신들이 원하는 목적 달성이 쉬워지기 때문이다. 우리는 숫자로 이루어진 테크 기업들의 데이터베이스 속에서 무언가로 범주화되고 있다.

문제는 이 숫자가 그 사람을 정확히 보여주지 않는다는 점이다. 내가 야구를 얼마나 좋아하는지를 숫자로 표현하기는 어렵다. 나도 숫자로 표현하기 어려운 것이 제대로 수집될 수는 없다. 다만, 내가 얼마나 좋아하는지를 다양한 측정 방법을 통해 유추할 뿐이다. 야구 기사를 얼마나 봤고, 커뮤니티에는 얼마나 방문했고, 댓글을 달았는지, 내가 좋아하는 팀에 '좋아요' 등을 클릭했는지 등을 수집한 뒤 이를 계량화해 숫자로 표현한다.

당연히 야구를 잘 아는 사람일수록 어디서 어떤 정보를 수집해야 하는지를 잘 안다. 야구를 좋아하는 사람만이 아는 경로들이 있기 때문이다. 제대로 된 숫자는 그 분야를 아는 사람이 가장 잘 알아본다. 같은 숫자를 보더라도 여러 갈래로 해석이 달라지는 이유이기도 하다. 매주 발표되는 지지율이 각자의 정치적 성향에 따라 달리 해석되는 것이 대표적이다.

과학으로서 통계는 이러한 문제점을 해결하기 위해 신뢰도를 검증하여 얼마나 믿을 수 있는지를 수치화하고 해석할 때 유의할 점 등을 서술한다. 아무리 잘 설계한 조사를 실시하더라도 통계 결과는 놓치



는 부분이 있을 수밖에 없다. 숫자가 모든 것을 말해주지 못하기 때문이다. 대한민국 인구주택 총조사 결과는 정말 많은 것을 제시해주지만 우리 국민 전체의 삶을 그대로 보여주고 있다고 말하기는 어렵다. 결국 해석이 필요하다.

해석에 있어 중요한 것은 지식과 경험

이 해석에 있어서 중요한 것이 숫자가 놓치는 부분들을 간파하는 능력이다. 이는 사실 해석하는 사람이 그 분야에 대해 얼마나 지식과 경험을 갖고 있는지에 달려 있다. 지식은 쌓으면 어느 정도 해결할 수 있지만 경험은 다른 문제다. 남자와 여자의 삶은 아무리 감정이입을 하더라고 서로 다른 경험이 될 수밖에 없다. 내가 좋아하는 야구 팀에 유리하게 통계 모델을 적용하는 것처럼 각자 경험에 따라 같은 숫자도 다르게 해석하는 경우가 많다. 서울의 물가는 언론에서 지역의 물가보다 더 중요한 통계로 다뤄진다. 정말 많은 사람들이 아파트에 살고 있기에 거의 대부분의 부동산 가격 관련 기사는 아파트 매매가를 주로 다룬다. 2020년 '주거실태조사' 통계에 따르면 아파트 거주 비율은 51.1%다. 언론에 보도되는 각종 통계들은 그 통계를 다루는 언론 종사자들의 평균적 경험을 반영하는 경우가 많다. 엄마가 그렇게 부러워하는 '엄마 친구 아들 혹은 딸'은 통계적으로 봤을 때 극히 작은 수이지만, 우리 대부분은 그 존재를 알고 있다.

이러한 해석의 문제를 해결할 수 있다는 주장 중 하나가 '빅데이터'다. 통계가 일반적으로 표본 추출을 통해 만들어지는데 반해 '빅데이터'는 표본이 아닌 전수에 가까운 수집을 하기 때문에 더 정확하고 더 많은 것을 말해준다고 한다.

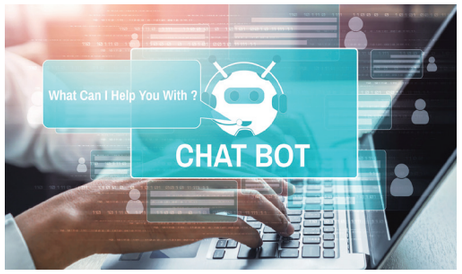
'빅데이터'는 정말 '빅'인가

우리는 지금 데이터 시대를 살고 있다. 더 많은 데이터가 더 많은 사실들을 얘기해 줄 것이라는 일반적인 믿음이 우리 사회를 지배하고 있다. '빅데이터'라는 단어가 일상적으로 사용되고 있으며, 이 '빅데이터'가 기존에 인간이 갖고 있던 통찰력과 직관, 사고 능력을 뛰어넘어 더 많은 것을 얘기해주는 것처럼 간주되고 있다. 빅데이터 분석을 했다고 하면 더 나은 분석일 것이라는 일반적 믿음 아래에서 단순한 데이터 분석도 빅데이터 분석을 한 것처럼 포장되고 있기도 하다. '빅'이란 크다는 의미의 영어 형용사다. '빅데이터'는 우리 말로 풀자면 '큰 데이터'이다. 그런데 크다는 것이 얼마나 크다는 것인지 명확히 정의하기 어렵다. 전 세계 데이터 양은 매년 두 배 이상 증가하는 것으로 알려져 있지만, 그 증가 속도는 더욱 빨라지고 있다.[1)]

지난 2001년 미국의 시장조사 및 컨설팅 기업인 가트너(Gartner)가 내린 정의[2)]에 따르면, '빅데이터'는 전례 없이 빠른 속도로 쏟아져 나오는 다양한 종류의 데이터를 의미한다. '빅데이터'는 기술의 발전으로 새로운 데이터 출처로부터 수집된 큰 규모의 데이터 세트라고 할 수 있다. 하지만, 절대적 의미에서 '빅'은 정의할 수 없으며 이전보다 클 뿐이다.

빅데이터는 전체 세상 일부분에 대한 근사치

제이콥스(Jacobs, 2009)[3)]는 이에 따라 "어떠한 지점에서든 그 당시 일반적이었으며 유효성이 증명된(tried-and-true) 방식을 넘어서 바라볼 것을 강요하는 데이터"라고 '빅데이터'를 정의한다. '빅데이터'는 절대적이 아니라 상대적이며 시대에 따라 다르다. 1980년 초반에는 수천 개의 테이프를 자동으로 입출력할 수 있는 '테이프 몽키'가 필요했던 데이터였으며 1990년대에는 엑셀과 데스크톱 PC의 범위를 초월해 유닉스 워크스테이션 기반의 소프트웨어를 필요로 하는 데이터가 '빅데이터'라는 것이다. 지금은 DBMS에 넣어 데스크톱 환경의 통계 및 시각화 패키지의 도움으로 분석하기에는 너무나 큰 데이터를 의미한다. '빅데이터'는 현재 시점에서 큰 데이터일 뿐이지 모든 것을 말해주는 데이터가 아니다.



또한, 데이터가 아무리 크다 해도 한계가 있다. 기본적으로 데이터는 '연산가능성(computability)'(Boyd & Crawford, 2012)[4)]을 기반으로 한다. 숫자화될 수 있어야 한다는 것으로 계산이 가능해야만 분석이 가능하다. '엄마를 얼마나 좋아해'라는 질문에 대한 답은 숫자로 정의될 수 없다. 모든 것을 아는 것처럼 보이는 '챗GPT'의 한계를 우리는 이미 체험하고 있다.

"챗GPT는 현재 웹의 흐릿한 그림(ChatGPT Is a Blurry JPEG of the Web)"[5)] 이라는 공상과학 작가 테드 창(Ted Chiang)의 지적처럼 이른바 '빅데이터'로 우리가 얻을 수 있는 것은 전체 세상 일부분에 대한 근사치에 불과하다. 숫자로 된 야구의 기록이 모든 것을 말해준다면 야구를 좋아할 필요가 없을 것이다. 야구는 '모르기' 때문에 재미있다.

1) https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
2) https://www.gartner.com/en/information-technology/glossary/big-data
3) Jacobs, A. (2009). The Pathologies of Big Data. Communications of the Acm, 52(8), pp. 36-44.
4) Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information Communication & Society, 15(5), 662-679.
5)Ted Chiang (2023. 2. 9.). ChatGPT Is a Blurry JPEG of the Web. The New Yorker. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

학교 현장에 빅데이터 활용 교육 도입이 필요하다

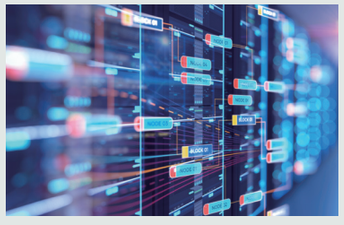
최우성 | 다산고등학교 교장

빅데이터의 역기능을 순기능으로 바꾸기 위한 과제

2016년 10월 촛불 집회 당시, 언론사의 가장 큰 골칫거리는 참가 인원 수였다. 경찰과 집회 주최 측 추산 인원 차이가 컸기 때문이다. 이 문제는 집회장 근처 편의점의 '카드 결제 내역'을 확인하거나 '통신사의 데이터', 즉 빅데이터를 활용하여 집회에 참가한 인원수를 집계하면 쉽게 해결된다. 당시 이 방법으로 집회에 참가한 인원을 비교적 정확하게 집계할 수 있었다.

무엇보다 빅데이터의 가치를 드높인 사건은 구글의 독감 유행 예측 이벤트였다. 구글은 사용자들의 '독감' 검색량 추이를 분석하여 독감 증상을 보이는 사람들이 북미 지역에 많다는 사실을 알아냈고, 곧 독감 유행이 닥칠 것이라고 예측하였다.

이처럼 빅데이터는 데이터의 크기, 다양성, 속도, 정확성, 가치 등의 속성을 가지고 설명할 수 있다. 요즘은 온라인 쇼핑몰에서도 빅데이터를 활용하여 시장 흐름을 예견하고 구매자의 욕구를 데이터로 만들어 실제 구매로 이어질 수 있게 제품을 추천한다.



그러나 빅데이터는 역기능도 상존하고 있다.

CCTV는 인간의 행동을 디지털 장비에 저장하고 있으며, 통신사는 지피에스(GPS), 위치 추적 등으로 휴대전화 사용자의 모든 동선을 알고 있다. 페이스북에 올린 사진, 신용카드 결제 내역, 검색한 내용 등은 빅데이터로 누적됨과 동시에 '빅브라더'라는 특정한 조직에 노출이 된다. 한마디로 인간의 디지털 족적이 곳곳에 남게 되는 것이다. 그리고 이는 악용될 소지가 있다.

이에 따라 빅데이터의 소유권과 저작권 분쟁이 문제로 등장하고 있다. 일상적인 이야기와 사진 또는 동영상 등의 개인 저작물이 공유 기능에 의해 배포되는 경우 수익을 볼 수 있는 구조로 변모할 수 있는데, 빅데이터의 소유권과 저작권 분쟁이 생기는 것이다.

이렇듯 성큼 다가온 4차 산업 혁명 시대에 빅데이터 분석을 위해 사용되는 통계가 매우 중요하게 여겨지고 있다. 따라서 빅데이터 분석은 오류(평균치의 함정)에 빠지지 않도록 상당한 전문성이 요구되며, 섬세한 주의가 필요하다. 사실 빅데이터의 역기능을 제거하고 순기능을 보장한다면, 인간 삶은 더욱더 윤택해진다. 그러므로 역기능을 순기능으로 바꿀 수 있는 빅데이터 활용 교육이 필요하다.

삶과 연계하여 창의적인 인재로 성장할 수 있는 발판이 필요

미래 사회를 준비하는 우리 초중고 교육은 교과서에 활자화된 데이터만을 학습하는 단계에 머물러서는 안 된다. 다양한 디지털 도구를 가지고 빅데이터를 활용하는 교육이 필요하다.

아직도 일선 학교에서는 수학의 '미분과 적분', '방정식과 부등식' 등을 교육 과정에 맞게 알맞은 공식을 사용하여 문제를 풀어보고 있다.

그러나 교실 수업에서 디지털 도구를 활용하게 된다면, 학생들은 어려운 문제를 손쉽게 풀 수 있는 경험을 획득

할 수 있게 된다. 이는 단순한 문제 풀이에 그치는 것이 아니라 삶과 연계하여 창의적인 인재로 성장할 수 있는 발판이 되어 줄 것이다.

빅데이터를 활용한 교육은 빅데이터를 수집, 분석하여 학생들의 학습 경험을 향상하는 것을 목표로 한다. 학생들의 학습 수준, 관심사, 학습 패턴 등을 파악하고, 이를 바탕으로 개인화된 학습 콘텐츠와 학습 방법을 제공할 수 있다.
이러한 이유 등으로 빅데이터를 활용한 교육은 현실과 동떨어진 교육이 아니라 현재 이루어지고 있는 교육 과정 속에서 녹여내야 한다.

빅데이터 활용 교육, 미래 사회를 대비하는 필수 과제

빅데이터는 4차 산업 혁명의 핵심 기술로, 다양한 분야에서 활용되고 있다. 교육 분야에서도 빅데이터를 활용한 교육이 활발히 이루어지고 있다. 빅데이터 활용 교육은 학생들의 학습 경험을 향상하고, 미래 사회에 필요한 인재를 양성하는 데 중요한 역할을 한다.

빅데이터 활용 교육의 대표적인 사례로는 학습 분석(Learning Analytics)이 있다. 학습 분석은 학생들의 학습 관련 데이터를 수집하고 분석하여, 학생들의 학습 수준, 관심사, 학습 패턴 등을 파악하는 것이다. 이를 바탕으로 개인화된 학습 콘텐츠와 학습 방법을 제공할 수 있다.

예를 들어, 학생들의 과제 제출 기록, 시험 성적, 온라인 활동 등을 분석하여 학생들의 학습 수준을 파악할 수 있다. 또한, 학생들의 SNS 활동, 웹 검색 기록 등을 분석하여 학생들의 관심사를 파악할 수 있다. 이러한 데이터를 분석하여 학생들에게 적합한 학습 콘텐츠와 학습 방법을 제공할 수 있다면,
학생들의 학습 효율을 높일 수 있다.

빅데이터 활용 교육은 학생들의 창의력과 문제 해결 능력을 향상시키는 데에도 도움이 되며, 학생들에게 다양한 관점과 사고방식을 제공할 수 있다. 예를 들어, 역사적 사건의 원인을 분석하거나, 사회 현상의 문제를 해결하기 위해 빅데이터를 활용할 수 있다. 이러한 과정을 통해 학생들은 창의적인 사고력을 키우고, 복잡한 문제를 해결하는 능력을 배울 수 있다.

빅데이터 활용 교육 도입을 위해 해결해야 할 과제

그러나 빅데이터 활용 교육은 아직 초기 단계에 있으며, 해결해야 할 과제들도 존재한다. 빅데이터를 수집하고 분석하기 위해서는 전문적인 기술과 지식이 필요하다. 또한, 빅데이터 활용 교육이 학생들의 사생활 침해로 이어질 수 있다는 우려도 있다.

구체적인 교육 방안으로는 다음과 같은 것들이 고려될 수 있다.

- 빅데이터 활용 교육을 위한 전담 부서를 설치하고, 교육 정책을 수립 및 시행한다.
- 빅데이터 활용 교육을 위한 교사 연수를 확대하고, 교사들이 전문성을 갖출 수 있도록 지원한다.
- 학교에 필요한 빅데이터 분석 도구와 기자재를 지원한다.
- 빅데이터 활용 교육을 위한 교육 플랫폼을 구축하고, 교사와 학생들이 이를 활용할 수 있도록 한다.

빅데이터 활용 교육이 활성화되기 위해서는 정부와 교육계의 노력이 필요하다. 정부는 빅데이터 활용 교육을 위한 교육 플랫폼을 구축하고, 교육청은 디지털 도구를 활용할 수 있도록 기자재를 보급하고 교사 연수를 실시해야 한다. 또한, 빅데이터 활용 교육의 윤리적 기준을 마련하고, 학생들의 사생활 침해를 예방하기 위한 대책을 마련해야 한다.
빅데이터 활용 교육은 미래 사회를 대비하는 필수 과제이다. 학생들이 빅데이터를 활용한 교육을 통해 미래 사회에 필요한 역량을 키울 수 있도록 정부와 교육계의 지속적인 노력이 필요하다.

증거기반 의사결정을 위해 통계안목을 갖추자

최성호 | 경기대학교 진성애교양대학 학장

토론과 의사결정에서 증거를 기초로 하는 문화 필요

과거 정책결정자나 행정 관료들이 지시와 회의에 의해 보고서를 쓰면 정책이 되는 시절이 있었다. 언제부터인가 정부정책은 물론 사회와 개인의 모든 의사결정이 반드시 증거를 기반으로 해야 한다는 주장이 힘을 얻고 있다. 2000년 UN의 새천년 개발목표(Millennium Development Goals: MDGs 2015)는 증거에 기반을 둔 의사결정을 강조하였고 2017년 OECD 개발협력보고서 '발전을 위한 데이터' 2030 지속가능개발목표(SDGs 2030)가 데이터와 증거에 기반을 둔 우선순위와 전략의 선택에 의해서만 달성될 수 있다고 하였다.

미국은 2018년 데이터 접근의 개선과 평가역량 강화에 의한 증거 기반 확충을 위하여 '증거기반 정책결정법'(Evidence-based Policy-making Act)을 제정했다. 우리나라도 '데이터기반행정 활성화에 관한 법률'("데이터기반 행정법")을 2020년 말부터 시행하였다. 기업과 개인의 경우에도 목적 달성에 효과적인 의사결정은 데이터에 기반을 두어야 한다. "우리 편은 다 맞고 상대측은 모두 틀리다"는 맹목적 진영대립이 난무하는 우리 사회에서 토론과 결정이 적정한 데이터와 통계를 기초로 해야 함은 더할 나위 없이 긴요하다.

상황판단과 의사결정에 통계를 활용하는 사례

의사결정에 대한 통계 개념의 이해와 활용의 간단한 사례를 들어보자.

첫째. 한 자동차회사가 동북아와 북미에 공장을 신설한다고 하자.

2021년의 구매력평가 환율 기준 1인당 국내총생산(GDP)이 동북아에서 한국 4만 4천 달러, 일본 4만 1천 달러, 중국 1만 8천 달러이다. 또한 북미의 경우 미국 6만 4천 달러, 캐나다 4만 8천 달러, 멕시코 1만 9천 달러이다. 두 지역의 평균 소득이 각각 4만 4천 달러, 3만 4천 달러로 북미가 동북아의 1.3배 수준이니 유사한 공장을 지었다면 옳은 결정일까.

이 경우에서는 '평균의 평균'이 빠질 법한 함정을 피해 가중평균을 계산해야 평균소득이 동북아의 2.5배에 이르는 북미에 고가 세단, 동북아에 중저가 세단 공장을 건설하는 결정이 가능할 것이다. 대푯값으로 산술평균이 가장 많이 쓰이지만, 증가율에 관해서는 기하평균이, 속도의 경우 조화평균의 개념이 적절하다.

둘째. A펀드와 B펀드 중에 투자처를 고민한다고 하자.

A펀드는 연말에 코스피 지수가 3천 이상이면 100% 수익, 미달이면 90% 손실을 주고, B펀드는 3천 이상이면 10% 수익, 미달이면 0% 수익을 준다. 코스피 지수가 3천 이상일 확률은 50%이니 기대수익률은 둘 다 5%이다.

이런 경우 수익률의 평균 이외에 분산 또는 표준편차를 비교하여야 위험을 고려한 합리적 결정에 이를 것이다. 물론 이 사례와 같은 경우는 별로 없으며 고수익·고위험, 저수익·저위험의 투자 대안 간 선택이 일반적이다.

셋째. 1936년 미국 대선을 앞두고 두 기관이 여론조사를 했다.

리터러시 다이제스트는 유권자 1천만 명에게 우편으로 설문지를 보내 238만 명의 회신을 받은 결과 공화당 후보인 알프레드 랜던 후보의 당선을 예측하였다. 그러나 민주당 후보 프랭클린 루스벨트의 당선을 맞춘 것은 단지

수천 명의 표본을 대상으로 조사한 갤럽이었다.
이 사례는 잘 설계된 표본조사가 전수조사보다 나을 수 있다거나, 여론조사에 표본의 크기보다 표본 추출의 방법이 더 중요하다는 사실을 말해준다. 이 조사를 계기로 파산하여 최고 여론조사기관의 위상을 갤럽에게 내어준 리터러시 다이제스트사가 자동차등록부와 전화번호부에서 추출한 표본이 당시 상류층에 편중되었던 문제가 실패를 초래했다.

이같이 간단한 사례 이외에도 데이터기반 행정법 등의 입법으로 부처, 기관, 부문 간 데이터 연계·통합이나 데이터 활용 혁신, 정책효과 평가가 가능하게 되었다. 경제, 보건, 교육, 교통 등 빅데이터 풀의 구축과 공유·활용을 통하여 데이터 분석, 최적대안 탐색, 선제적·맞춤형 행정서비스 제공, 정보집약 민간서비스 개발이 확대될 것으로 전망된다.

현대판 프로크루스테스:경계하고 조심해야[1)]

통계 활용의 확대에 대한 긍정적 기대에도 불구하고 의사결정의 근거로 제시되는 통계를 왜곡하거나 조작하는 프로크루스테스를 조심해야 하는 상황이 빈번하다. 2022년 제20대 대통령선거를 앞두고 시행된 여러 여론조사의 결과가 서로 엇갈려 유권자들에게 혼란을 주었다. 선거가 두 달도 남지 않은 1월 14일 두 건의 여론조사는 같은 기간에 거의 동일한 조사기관들이 동일한 방법으로 조사하였음에도 윤석열 후보와 이재명 후보의 예측 지지율이 39%:33%와 29%:37%인 정반대의 조사결과를 발표하였다.

1) 여기서 프로크루스테스는 필자의 공저(송인창·최성호, 통계안목: 세상을 바로 보는 힘, 바틀비, 2023)에서 통계의 왜곡과 조작을 고대 그리스 신화에 나오는 '프로크루스테스의 침대'에 비유한 표현이다.

초미의 관심을 받으면서 유수의 여론조사 업체들이 조사한 결과가 이럴진대 특정 정당이나 후보가 비용을 부담하는 조사들은 왜곡 가능성이 높다. 따라서 조사기관은 통계전문가를 보유하여 믿을만한지, 표본의 수나 추출방법, 응답률은 적절한지, 설문의 구성과 조사방식에는 문제가 없는지 확인해야 한다. '특정 조사기관 효과(House effect)'는 불가피하므로 서로 다른 조사기관, 여러 후보들 간의 지지율 차이보다 같은 기관이 조사한 특정 후보자의 지지율 변화에 주목하는 것이 유용하다.

최근 감사원이 발표한 지난 정부의 부동산가격과 고용, 그리고 소득분배 통계의 조작 의혹은 사실 여부를 떠나 국가통계기구가 작성하거나 승인한 통계마저 의심받고 있다는 참담한 상황을 보여주고 있다. 2019년 말 발생한 코로나-19의 대응과정에서 각종 통계가 정책 혼란을 부추겼다는 평가도 외면하기 어렵다.

증거기반 의사결정을 위한 우리 사회의 과제

한국 사회에 증거에 기반을 둔 의사결정이 보편적으로 정착하기 위하여 몇 가지 과제가 있다. 우선 정확성과 신뢰성, 시의성을 가진 데이터가 생산되어야 한다. 전문성과 정치적 중립성을 가진 국가통계기구의 역할이 기본이다.

미국은 여러 기관이 국가통계를 작성하는 분산형 통계조직을 채택하고 있다. 의회 관리예산처(OMB) 내 정보규제청(OIRA)의 통계과학정책실(SSP)이 총괄기능을 수행한다. 미국통계학회 회장을 지낸 캐서린 월맨(Katherine K. Wallman)은 통계 예산·인력의 배분, 중복 조정과 통계기준 일치 등을 총지휘하는 SSP실장인 통계수석(Chief Statistician)을 1992년에서 2017년까지 25년 동안 맡았다. 부시와 클린턴, 아들 부시, 오바마를 포함한 4명 대통령의 행정부에서 근무했음을 의미하며, 국가통계기구의 정치적 중립과 독립성을 상징한다.

우리나라의 경우 1990년 이후 통계청장의 평균 임기가 2년도 채 안되더니 급기야 2018년에는 소득분배 등 통계

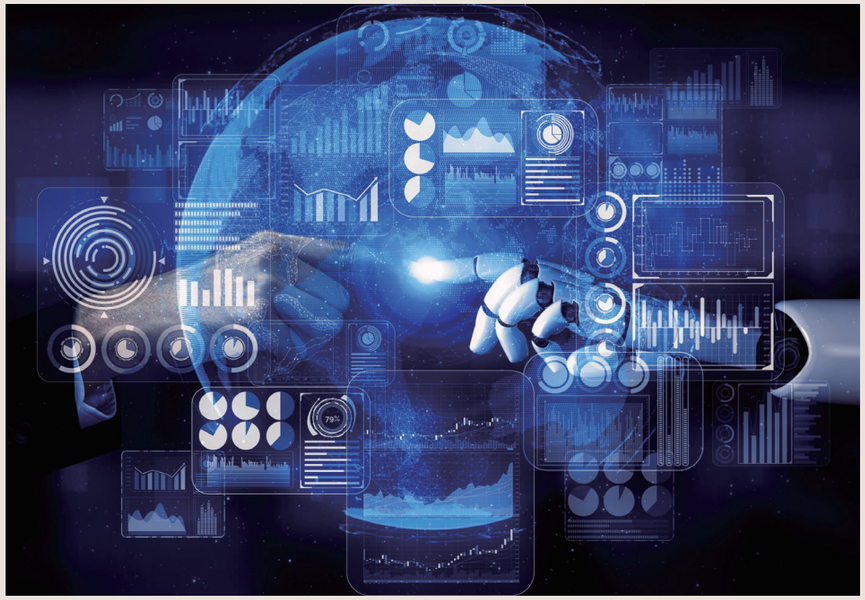
가 정책성과를 나타내지 못했다는 이유로 갑자기 통계청장이 경질되었다는 시비가 제기되었다.
또한 생산된 데이터가 체계적으로 축적되고 공유될 수 있는 인프라가 구축되어야 한다. 공공기관이 생산한 데이터를 적극적으로 공개하고 민간 데이터 또한 공유할 수 있는 인센티브를 마련해야 한다. 통계청의 통계정보 원포털, 서울시의 빅데이터 캠퍼스, 행정안전부의 행정·공공기관 정보 클라우드 전환, 미국 A사나 한국 S사의 빅데이터 클라우드 구축 논의 등이 그것이다. 한편 제도적 인프라도 중요하다. 예컨대 여론조사의 경우 선거여론조사에 관해서는 중앙선관위 규제나 한국기자협회의 선거여론조사준칙이 있다. 그러나 정책 결정에 중대한 영향을 미칠 수 있는 일반 여론조사에 대해서도 미국 여론조사협회, 영국 임프레스(Impress)·시장조사학회 등 사례처럼 민간의 실효성 있는 자율규제 장치가 정착되어야 한다.

한편 공공정책을 비판하고, 정치과정에 참여하며, 스스로 의사결정을 내리는 시민의 통계안목 함양이 무엇보다 중요하다. 앞에 언급한 미국의 Wallman은 통계 문해력을 소통과 비판은 물론, 개인이나 공공의, 직업적이거나 개인적인, 모든 의사결정에 통계를 활용하는 능력으로 정의하였다. 국가통계가 의혹의 대상이 된 현실에서 각종 언론과 정당, 사회단체, 정치인들이 쏟아내는 여론조사 결과를 그대로 믿기는 어렵다. 민주사회에서 다양한 의사결정이 증거에 기반을 두고 합리적으로 이루어지기 위해서는 시민의 세련된 통계안목이 긴요하다.

시민의 통계안목은 통계문해능력(statistical literacy)의 기초교육을 통해 개선할 수 있다.

첫째, 기초 통계 개념을 정확하게 이해하여야 한다.
평균 등 대표 값(representative value)과 함께 분산도(variation), 그리고 비율을 나타내는 백분율(percentage)과 백분율 점수(percentage point) 등의 개념을 정확하게 이해해야 한다. 다양한 주장을 전개하는데 표나 그래프를 통한 데이터 시각화(data visualization)가 중요하지만 통계의 해석을 왜곡하거나 과장할 수 있어 주의하여야 한다.

둘째, 기초 통계 개념을 도구로 작성된 경제사회 지표에 대한 기초 지식을 갖추어야 한다.
1인당 국내총생산(GDP), 물가지수, 실업률, 국제수지 같은 경제지표의 정의와 집계과정도 중요하다. 지역, 산업, 행정 등의 부문별 빅데이터 플랫폼 구축을 위해서는 부문별 이론도 반영되어야 한다. 지형, 위치, 이동 등의 빅데



이터 분석에 의해 친환경차 충전, 재난예방 시설, 행정서비스센터 등의 최적입지를 도출할 수도 있다. 기업들이 환경·사회·지배구조 친화경영(ESG)을 실천하기 위해 데이터를 개발하고 지표체계를 구축하기도 한다.

셋째, 모집단과 표본의 이론으로 설문조사의 유용성과 한계를 알아야 한다.
그에 따라 올바른 조사방법을 설계하여 실행하고 조사결과를 적절하게 해석해야 한다.

빅데이터·인공지능(AI) 시대에 통계문해 능력의 함양이 개인과 사회의 생존과 번영을 위한 필수 조건이다. 통계를 대할 때 예리한 비판의 눈으로 통계가 말하는 바를 한 번, 또 한 번 찬찬히 따져봐야 한다. 특히 프로크루스테스가 입맛에 맞게 왜곡한 통계가 아닌지 의심해 보는 과정이 필수적이다. 또한 의사결정자는 기존 통계에 안주하지 말고 현실 문제를 파악하여 해결책을 제시하는데 더 적합한 통계가 있는지 항상 고민해야 한다. 시민들도 스스로의 통계안목을 유지하는 방법을 모색해야 한다. 자신의 관심, 업무, 자산관리 등에 관한 통계플랫폼을 구축하여 관리함을 생활화할 수도 있다. 무엇보다 바람직한 정책과 성공적인 인생은 좋은 통계와 정확한 분석을 기반으로 설계될 수 있다는 인식이 확산되어야 한다.

Machine Learning의 효과적 운영을 위한 조건

김주환 | SAS Korea 이사

모델의 개발과 서비스 운영 통합의 필요성

Machine Learning은 복잡한 문제들을 효율적으로 풀어낼 수 있는 도구로 다양한 분야에 활용되고 있으며 그 활용 영역도 점차 확대되어 가고 있다. 하지만, Machine Learning 모델을 시스템에 적용하여 운영하는 것에는 많은 어려움이 있고 이에 대부분 Machine Learning 모델 개발까지는 진행되나 시스템에 적용하여 운영 서비스 제공까지는 적용되지 못하는 경우가 많이 있다. Machine Learning 모델을 시스템에 적용 시 발생하는 어려움에는 모델 운영 적용, 모델 성능 모니터링, 모델 관련 데이터 관리, 모델 적용 관련 협업 등의 어려움을 들 수 있다.

모델 운영 적용 어려움

Data Scientist는 필요 데이터 정의, 데이터 수집, 데이터 정제, 모델링 단계로 모델 개발을 진행하며 이를 시스템에 적용하기 위해서는 일/주/월 단위로 이전 데이터가 적재되어야 운영에 모델을 적용할 수 있다. 이를 위해서는 API, Batch Process 개발, 운영 Infrastructure 설계 및 적용 등 다양한 시스템 요소를 고려하여야 하지만 Data Scientist는 운영 서비스 개발에 능숙함이 부족한 경우가 많고 이에 개발한 모델을 운영 시스템까지 적용하는데는 어려움이 있다.

모델 성능 모니터링 어려움

개발한 모델을 서비스 개발자의 지원을 받아 운영 시스템에 적용한 후에는 운영하며 모델 성능의 변화를 지속적으로 모니터링하여야 하고 모델의 성능이 떨어지는 경우 개선하는 단계가 필요하다. 모델 성능이 모니터링되지 않으면 모델 성능이 떨어진 경우 과거 개발된 모델이 계속 적용되어 모델의 활용에 문제가 발생하며, 이에 모델 성능을 모니터링 할 수 있는 환경 구축이 필요하다.

모델 관련 데이터 관리 어려움

운영 중인 모델에 성능 이상이 모니터링되면 모델 학습 시 사용된 데이터와 운영에 적용되고 있는 데이터 간에 어떠한 차이가 발생한 것인지 비교 분석해야 한다. 하지만 데이터는 운영 중에 계속 변하고 있고 모델 학습도 반복해서 업데이트가 이루어져 과거 모델 학습 시 활용된 데이터를 확보하는데 어려움이 있다.

모델 적용 관련 협업의 어려움

모델을 개발하고 운영 서비스에 적용하기 위해서는 서비스 개발자와의 협업이 필요하다. 하지만 서비스 개발자는 Machine Learning 모델을 이해하기 어렵고 Data Scientist는 시스템 인프라 설계부터 서비스 운영까지를 이해하기 어렵다. 이에 서비스 개발자, Data Scientist는 서로 모르는 영역을 논의하여야 하고 이런 과정을 모델 개발과 업데이트 할 때마다 반복해야 하는 어려움이 발생한다.

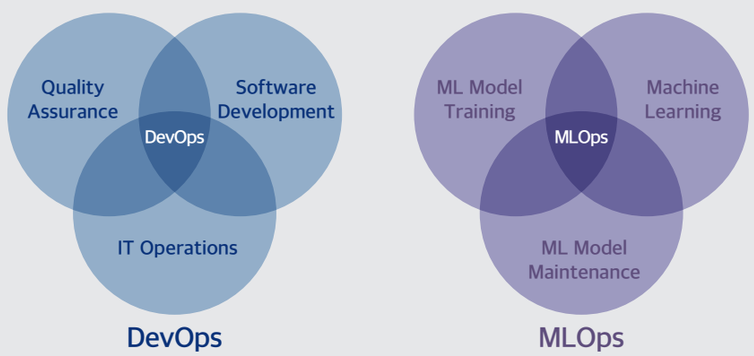
Machine Learning은 업무나 기업 입장에서는 다양한 비즈니스적인 문제나 요구사항을 해결하기에 유용한 도구이다. 하지만, 모델 개발 외에도 모델 배포 및 운영 적용, 향후 모델의 지속적인 유지 관리, 모델 적용 후 업무에서 나타나는 효과성 입증, 정확성, 수익성 등 모델이 개발되어 배포 후 안정기까지 가기에는 많은 어려운 과정이 존재한다.

이와 같은 어려움을 고려하여 Machine Learning 모델의 개발과 서비스 운영(Operations)을 통합한 방법론이 MLOps이며 모델 개발과 안정적인 모델 운영을 위한 협업 방식이다. 즉, MLOps는 모델 설계부터 운영 적용, 상용화까지 가장 빠른 시간 안에 가장 적은 위험 부담으로 Machine Learning 모델이 적용될 수 있도록 기술적인 문제를 최소화하는 방법론이다.

MLOps 개념 : 모델에 대한 통합·배포·학습을 구현하고 자동화

MLOps는 모델 개발에 IT운영 적용을 고려한 개념으로 Machine Learning 모델의 운영 적용을 위해 DevOps 원칙을 적용한 것이다. DevOps는 소프트웨어 개발(Development)과 IT운영(Operations)의 합성어로 소프트웨어 개발자와 정보기술 전문가 간의 협업과 소통, 통합을 강조한 것으로 기획과 개발, 테스트, 운영, 모니터링이 유기적으로 연결된 방법론이다.
DevOps의 목표는 변하는 요구사항을 반영하여 더 나은 서비스를 제공하는 것으로 개발과 운영이 분리된 환경에서는 개발이 지연되거나 배포 후 문제가 발생하지만 DevOps는 개발자와 운영자가 책임을 공유함으로 오류를 줄여 비용을 절감할 수 있고 문제 발생 시 즉시 처리할 수 있다.

MLOps는 모델 개발(Machine Learning)과 IT운영(Operations)의 합성어로 Machine Learning 모델을 안정적이고 효율적으로 배포 및 유지 관리하는 것을 목표로 하는 방법론이며, Machine Learning 모델에 대해 지속적인 통합, 지속적인 배포, 지속적인 학습을 구현하고 자동화한다.
DevOps가 운영을 위해 보다 조직 간 협조적으로 구축해 가기 위한 체계를 의미한다면 MLOps는 Machine Learning 모델이 기계 스스로 자료의 수집부터 서비스 수행까지 모두 통제할 수 있도록 자동화하는 것을 의미한다.



[표1.] DevOps, MLOps 비교

MLOps 단계

MLOps는 Machine Learning 모델을 운영에 적용하여 시스템화 하는 것을 목표로 Machine Learning 분야와 소프트웨어 및 데이터 엔지니어링 분야 간 협업을 하며 데이터 수집, 모델 학습 및 개발, 모델 배포, 모델 버전 관리, 지속적인 데이터 평가 및 추적, 모델 성능 모니터링, 피드백의 과정으로 진행된다. 즉, MLOps는 Data Scientist와 서비스 운영 전문가와의 협업 및 커뮤니케이션을 통해 모델 배포 및 자동화, 거버넌스 규정, 확장성, 모니터링 및 관리, 재현성을 목표로 진행된다.

MLOps의 단계는 ML(모델학습) 단계와 Ops(운영) 단계로 나누어진다.
ML인 모델학습 단계에서는 데이터 수집, 전처리, 모델학습 및 개발, 평가의 과정으로 진행되고 Ops인 운영 단계에서는 개발된 모델의 배포, 모델 모니터링, 테스트, 피드백의 과정으로 진행된다.
위의 단계에서와 같이 MLOps는 Machine Learning, Data, Infrastructure 등을 모두 포함한 개념으로 데이터, 모델, Infrastructure 등 모든 시스템이 유기적으로 돌아가도록 하는 것이 중요 요소이다.
위의 MLOps 단계를 표로 표현하면 다음과 같다.

[표2.] MLOps 단계

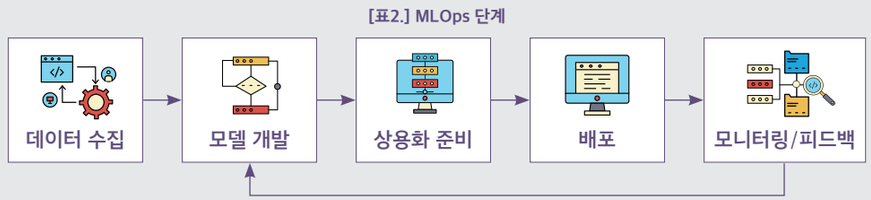


표2의 MLOps 단계를 세부적으로 설명하면 다음과 같다.

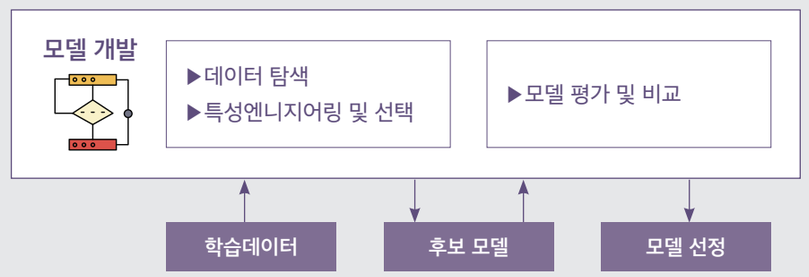
데이터 수집

모델 개발에 필요한 데이터를 정의하고 원천데이터에서 데이터를 수집하고 추출하는 단계로 모델 개발에 필요한 테이블 및 필드, 기간 정의 및 수집하는 과정이다.

모델 개발

수집된 데이터에 대해 탐색적 데이터 분석(EDA)을 수행하여 모델에 적용될 데이터를 이해하며 이를 기반으로 데이터를 정제하며 학습할 데이터를 준비한다. 학습데이터는 학습, 검증, 테스트 데이터로 분할하여 구성하고, 다양한 Machine Learning 알고리즘 적용 및 하이퍼파라미터를 조정하여 후보 모델을 개발한다. 개발된 후보 모델은 준비된 평가 데이터에 적용하여 모델을 평가한 후 최종 모델을 선정한다.

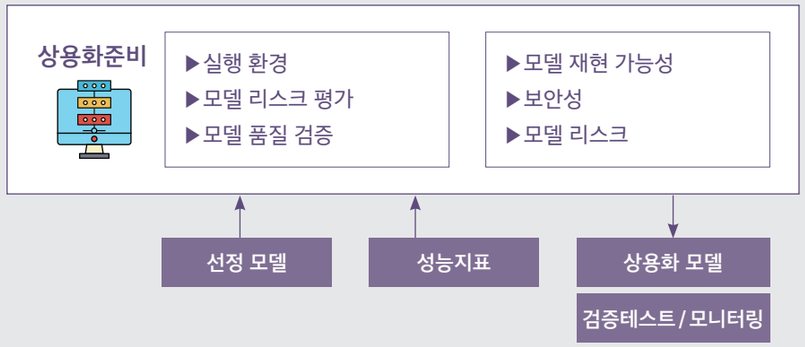
[표3.] 모델 개발 단계



상용화 준비

Machine Learning 개발 환경과 서비스에 적용(운영)하는 상용 환경은 크게 다를 뿐만 아니라 상용 환경에서 모델 작동 시 리스크가 있어서 상용 전환 과정에서는 테스트 및 잠재적 리스크가 적절히 경감되도록 고려가 필요하다. 이에 상용화 준비 단계에서는 모델 성능을 검증하고 배포에 적합한 수준인지 검토한다.

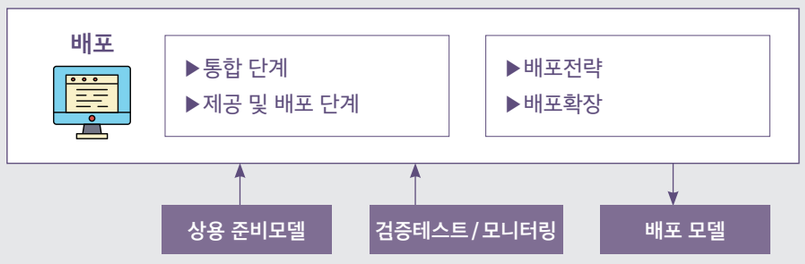
[표4.] 상용화 준비 단계



배포

상용(운영) 준비된 모델에 대해 모델을 검증하고 상용 배포 시 리스크가 낮은지를 판단하며 상용(운영) 전략을 수립하여 준비된 모델을 배포한다.

[표5.] 배포 단계

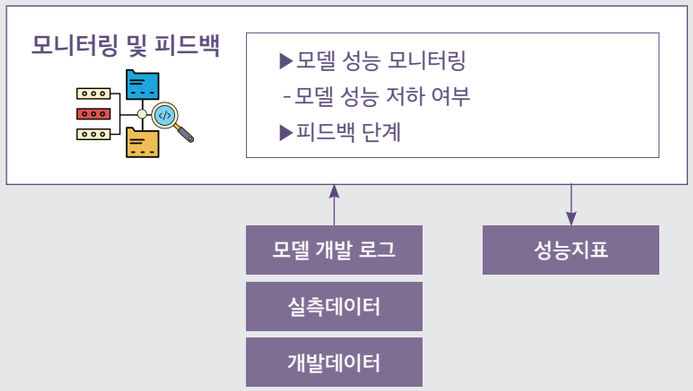


모니터링 및 피드백

Machine Learning 모델을 상용(운영) 배포한 후 모델 성능 저하가 빠르게 일어나거나 아무 경고 없이 업무에 부정적인 영향을 미치고 나서 늦은 대응을 하는 경우가 있을 수 있다.
이에 모델을 지속적으로 성능 모니터링하는 단계는 중요한 단계이며 모니터링 결과에 따라 개선 필요시 피드백을 통해 모델을 개선

적용한다. 중요한 점은 MLOps의 단계는 일회성이 아니라 배포 이후 모델 성능을 모니터링하고 필요시 모델 개발 단계로 신속하게 전환하여 모델 학습 및 개발, 배포, 모니터링 단계를 수행하도록 하는 반복 과정이라는 점이다.

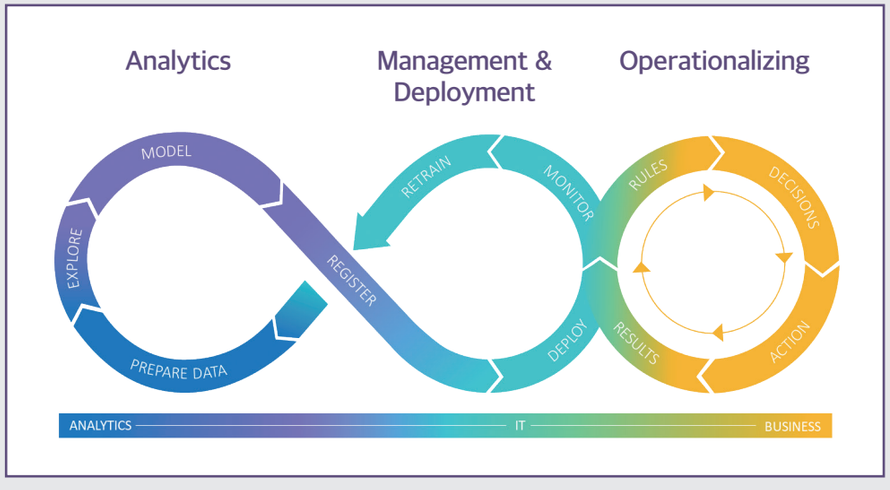
[표6.] 모니터링 및 피드백 단계



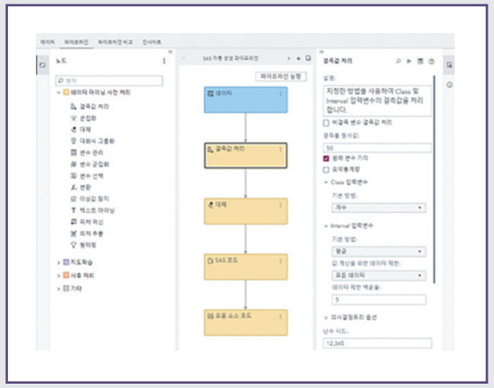
MLOps 적용 사례

MLOps의 흐름은 분석, 배포 및 모니터링, 운영의 단계로 MLOps 솔루션은 MLOps 흐름을 고려하여 구성되어 있으며 SAS를 중심으로 사례를 소개하고자 한다.

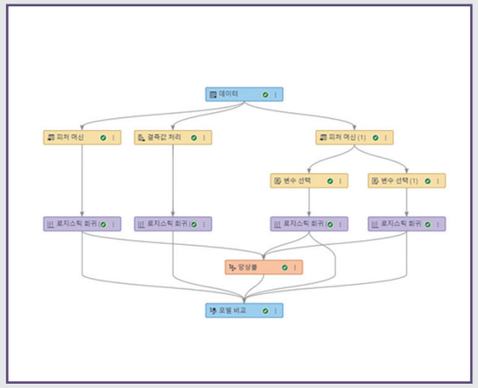
[표7.] MLOps 흐름도



① 수집된 데이터를 가공하여 분석데이터 구성 및 모델링 단계 (Analytics)



분석데이터 구성 (데이터 정제 및 가공)

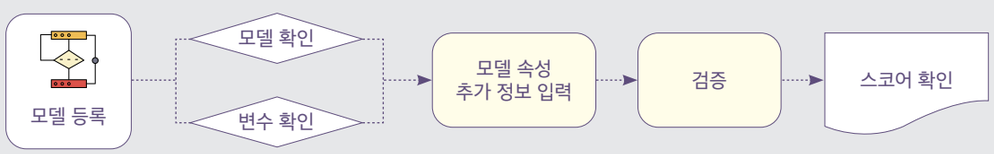


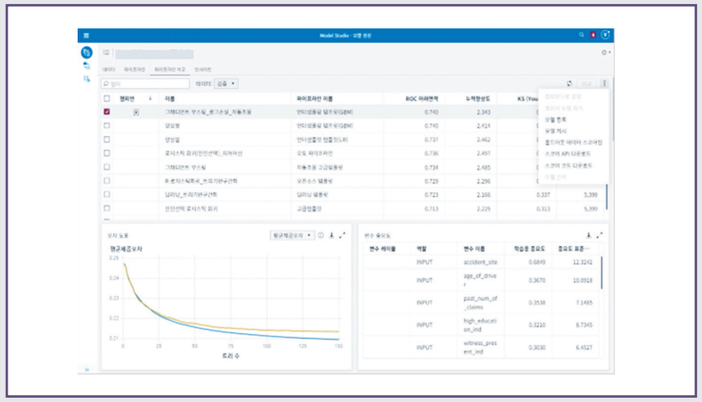
모델링 과정 (Pipeline 형태)

② 성능 비교를 통한 최종 모델 선정(Analytics)

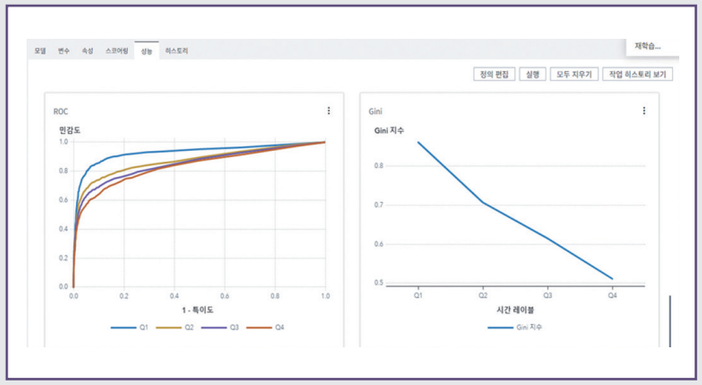
- 다양한 알고리즘 중 성능이 좋은 모델을 최종 챔피언 모델로 선정

③ 모델 배포 및 등록 (Management & Deployment)



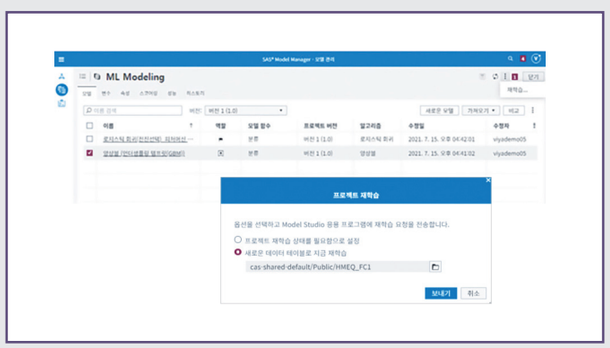


최종 챔피언 모델 배포 및 등록

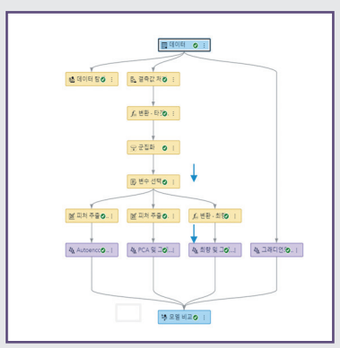


모델 성능 모니터링 (모니터링 단위별)

④ 모델 성능 모니터링하고 피드백하여 재학습하는 단계 (Operation)

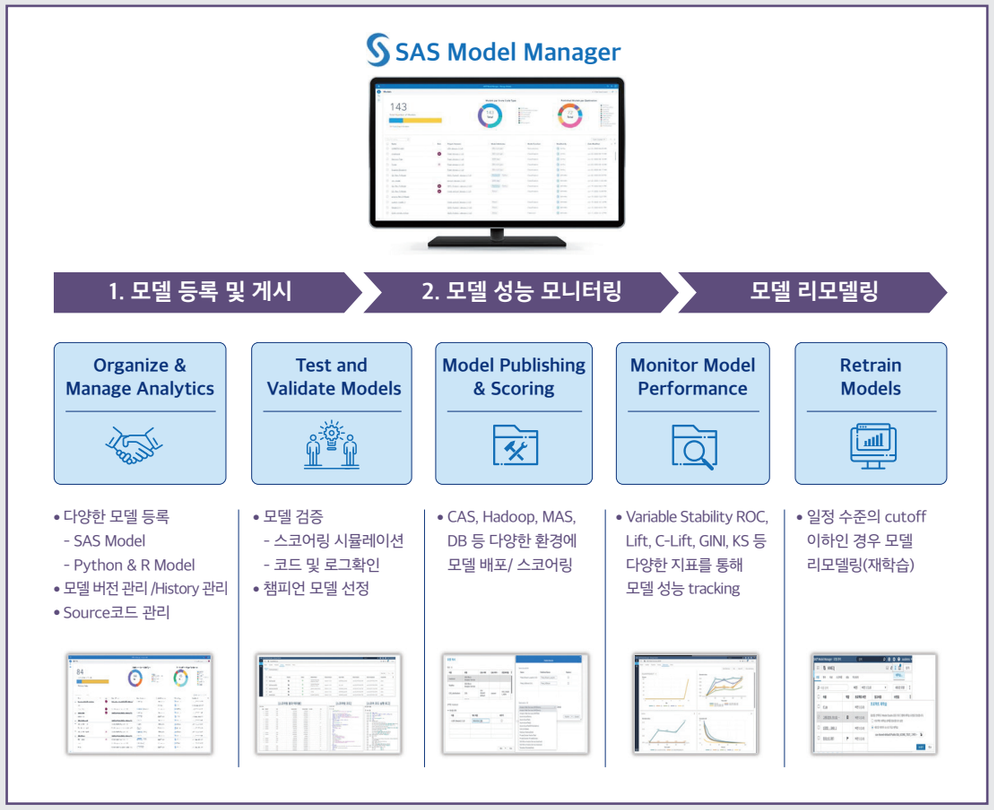


성능 모니터링 후 재학습 적용

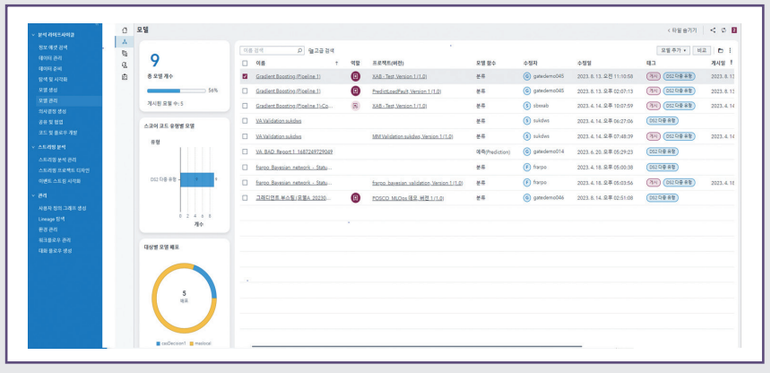


모델 재학습

모델 배포(게시), 모델등록, 재학습하는 단계에 대해 정리하면 다음과 같다.



⑤ 모델에 대한 정보 및 상태를 파악할 수 있도록 대시보드 제공



업무 및 기업에서는 데이터 정제 및 가공, 모델 학습, 평가, 배포 및 서비스, 모니터링, 재학습까지의 전체 라이프 사이클을 통합 지원하는 MLOps 환경을 요구하고 있으며 SAS MLOps 솔루션은 하나의 시스템 내에서 통합 지원하고 있다.

MLOps 활용에 대한 기대 효과

Machine Learning을 업무에 도입하고 투자한다면 기존에 해결하기 어려웠던 비즈니스 문제들을 해결해 갈 수 있다. 하지만 Machine Learning 방법과 다르게 MLOps는 모든 비즈니스 문제를 직접 해결할 수 있는 것은 아니다. MLOps 가치는 Machine Learning으로 개발된 모델이 안정적이고 빠르게 운영에 적용해 갈 수 있다는 점 즉, Machine Learning에 투자한 것이 안정적인 운영을 통해 실질적인 가치 창출로 이어질 수 있도록 하는 점에 있다.

MLOps는 단기간에 가시적 결과를 내기 위한 용도는 아니고 다양한 분야의 관계자 참여를 필요로 하지만 Machine Learning 업무가 확장될수록 그 가치와 효과 증가를 기대할 수 있다. 또한 모델 배포와 운영에 들어가는 인력과 소요시간을 줄여 비용 절감의 효과를 기대할 수 있고 양질의 서비스를 빠른 시간 안에 업무에 적용할 수 있는 효과를 기대할 수 있다.

MLOps는 Machine Learning 모델의 개발과 적용에 가속화를 제공하므로 모델 개발의 생산성 및 관리에 수요가 많은 업무와 산업(금융, 제조, 유통 등) 분야에서 관심과 활용이 높아지고 있으며 지속적으로 그 활용도는 높아질 것으로 예상된다.

(참고자료) 1. MLOps 도입 가이드, 한빛미디어, 2022 2. SAS MLOps 영역 업무 자료

통계로 바라보는 세상이야기

신동헌 | 도서출판 지일북스 대표

아름답고 소중한 우리 한글, 가장 많이 사용하는 말은?

국립국어원 '2020 국민의 언어 의식 조사'에 따르면, 2020년 기준 국어에 대한 관심도는 55.4%로 2010년 이후 증가하고 있으며, 말하기(78.5%), 언어 예절(73.9%) 분야에 관심이 높은 것으로 나타났고, 우리 국민의 50.9%는 평소 국어를 바르게 사용한다고 생각하고 있으며, 욕설(46.9%)보다 비속어(48.1%)를 더 자주 사용하는 것으로 나타났습니다. 또한, 가장 많이 사용하는 말로는 표준어(56.7%)가 방언(43.3%)보다 높았는데요, 방언을 사용하는 사람과 대화할 때 친근하다고(79.9%) 느끼는 것으로 조사되었습니다. 이번 조사는 전국에 거주하는 만 20세 이상 만 69세 이하의 성인 남녀 5,000명을 대상으로 가구 방문 조사하였으며, 2005년에 처음 실시한 이후 5년 주기로 진행되었습니다.

가을은 독서의 계절! 가장 많이 빌려 본 책은?

문화체육관광부 '2021 국민독서실태조사'에 따르면, 성인이 독서를 하는 목적으로 새로운 지식·정보 습득이 1위를 차지했고, 교양을 쌓고 인격을 형성하기 위해서, 마음의 위로와 평안을 얻기 위해서 순으로 나타났습니다. 국립중앙도서관의 '인기대출도서'에 따르면, 올해 9월 11일까지 전국 공공도서관에서 대출한 도서는 39,363건으로 '불편한 편의점'이 1위를 차지하였고, 30,235건으로 '아버지의 해방일지'가 그 뒤를 이었습니다. 통계청의 '사회조사' 결과, 독서인구 비율은 2013년 62.4%에서 2015년 56.2%, 2017년 54.9%, 2019년 50.6%, 2021년 45.6%로 감소 추세이며, 1인당 평균 독서권수도 같은 기간 11.2권에서 9.3권, 9.5권, 7.3권, 7.0권으로 줄어들고 있습니다.

가을에 즐기는 지역 축제, 관심 유형 1위는?

문화체육관광부와 한국관광공사가 발표한 2022년 문화관광축제 빅데이터 분석 보고서를 보면, 2022년 개최된 21개 문화관광축제의 총 방문객 수는 코로나19 팬데믹 시기인 2019년 대비 무려 19.7% 증가하였고, 21개 축제 개최 전후 4주 대비 파급효과를 분석한 결과, 일 평균 방문객 수는 현지인 42.6%, 외지인 139.6%, 외국인은 45.1% 증가한 것으로 나타났으며, 총 내비게이션 검색 건수 또한 급격히 증가하였는데요. 지난 2018년은 56,153건이었던 것에 반해 2022년은 109,093건으로 무려 94.28%나 늘었으며, 2022년 축제 기간 외지인들의 목적지 유형별 내비게이션 검색 건수 중 가장 큰 비중을 차지한 것은 바로 '음식(27.0%)'으로 나타났습니다.

월별 국내여행 횟수 가장 높은 달은 '9월'

웨더아이(기상전문 IT기업)에 따르면 올해는 10월 1일 설악산을 시작으로 중부지방에서는 10월 19~20일, 지리산과 남부지방에서는 10월 20~26일 사이에 첫 단풍을 볼 수 있을 것으로 예상하였는데요. 올해 9월의 일평균 기온이 평년보다 높았고, 10월의 일 평균 기온 또한 평년과 비슷하거나 조금 높을 것으로 예상되면서 첫 단풍 또한 평년보다 늦게 시작할 것으로 예측되었습니다. 문화체육관광부의 2022년 국민여행조사에 따르면 2022년 국내여행의 연간 경험률은 전년대비 0.3%p 증가한 94.2%였으며, 국내여행 경험률을 월별로 살펴보면 9월이 54.3%로 1년 중 가장 높았고, 다음으로는 8월(53.1%), 10월(50.8%), 7월(50.3%) 등의 순으로 나타났습니다.

다문화 가구 40만 시대, 가장 어려운 건 '언어'

통계청의 '2022 인구주택총조사'에 따르면, 다문화 가구는 39만 9천 가구로 전년 대비 3.6% (1만 4천 가구) 증가한 것으로 나타났습니다. 또한, 33만 5천 가구를 기록했던 2018년부터 다문화 가구 수가 꾸준히 증가하였는데요, 다문화 대상자들의 국적은 한국계 중국인이 32.3%로 가장 많았고, 뒤를 이어 베트남 21.8%, 중국 19.0%, 필리핀 5.4%, 일본 3.6% 순으로 나타났습니다. 여성가족부의 '2021 전국 다문화가족 실태조사'에 따르면, 결혼 이민자 및 귀화자를 대상으로 지난 1년간 한국생활에서의 어려운 점을 물은 결과, '언어문제'가 22.9%로 1위를 차지했고, '경제적 어려움'이 21.0%, '외로움'이 19.6% 순으로 뒤를 이었습니다.

대한민국을 위한 새로운 혁신, 디지털플랫폼정부

디지털플랫폼정부는 이번 정부의 핵심 정책 과제인데요. 통계청은 지난 8월 30일 '디지털플랫폼정부와 국가통계의 역할'을 주제로 국가통계발전 포럼을 개최하였는데요, 고진 디지털플랫폼정부 위원회 위원장은 '새로운 혁신, 디지털플랫폼정부'라는 기조연설에서 인공지능, 데이터로 만드는 세계 최고의 디지털플랫폼정부 추진을 위해 부처 간 데이터 칸막이 해소, 인공지능·데이터 기반 과학적 행정 등을 강조하였습니다. 이어진 전문 세션에서는 국가통계 생산, 활용, 서비스, 미래 대응 등 4개 분야에서 다양한 발표와 토론을 진행하였는데요. 이번 포럼에는 중앙행정기관·공공기관·지자체·연구기관·대학·민간기업 등 100여 개 기관의 500여 명이 참석하였습니다.

문화 강국 '대한민국', 해외에서 이미지는?

한국국제문화교류진흥원에서 전 세계 26개국 만 15~59세 현지인 25,000명을 대상으로 조사한 '2023 해외한류실태조사' 결과에 따르면, '한국은 문화 강국이다'라는 항목에 '그렇다', '매우 그렇다'라고 응답한 비율이 53.8%로 전체의 절반이 넘는 비중으로 나타났습니다. 또한, 한국에 대해 가장 먼저 떠오르는 이미지로 K-POP(14.3%)을 꼽았는데요. 한국 문화 콘텐츠의 이용 경험자를 대상으로 조사한 결과, 한국 문화콘텐츠가 '매우 마음에 든다', '마음에 든다'라고 응답한 비율은 예능 76.5%, 드라마 76.3%, 영화 75.7%, 음식 74.2% 등 각 분야별로 70%를 상회할 정도로 한국 문화콘텐츠는 세계적으로 높은 호감도를 보이고 있음을 확인할 수 있었습니다.

대한민국, 저출산·고령화 심각하다! 대책은?

2022년 합계출산율 0.78명, 출생통계 작성 이래 최저. 지난 8월 통계청에서 발표한 2022년 출생통계를 보면, 작년 우리나라에서 태어난 출생아 수는 24만 9천 명으로 전년 대비 4.4% 감소하였습니다. 출생아 수의 감소와 함께 여성 1명이 평생 낳을 것으로 예상되는 평균 출생아 수인 '합계출산율'도 0.78명으로 줄었으며, 인구 1천 명당 출생아 수를 뜻하는 '조출생률' 또한 4.9명으로 전년 대비 0.2명 감소하였고, 2022년 전체 출생아 중 혼인 중의 출생아 비중은 96.1%, 혼인 외의 출생아 비중은 3.9%로 나타났습니다. 정부는 심각해지는 저출산, 고령화 상황에 대해 첫만남 이용권, 부모급여, 가정양육수당지원 등 출산 육아정책을 확대 지원하고 있습니다.

9월 셋째주 토요일 청년의 날! 청년들의 삶은?

청년기본법 제7조에 따르면, 청년의 날은 청년발전 및 청년지원을 도모하고 청년문제에 대한 관심을 높이기 위하여 지정된 법정 기념일입니다. 국무조정실에서 발표한 '청년삶실태조사'에 따르면, 청년들은 자신의 미래실현 가능성에 대해 '어느 정도 미래를 실현할 수 있다'라는 답변이 19~24세, 25~29세, 30~34세 각각 87.9%, 88.2%, 86.5%로 모든 연령층에서 가장 높은 비중을 차지했습니다. '완벽하게 실현할 수 있다'는 답변은 19~24세 청년들이 8.0%로 가장 높았고, '전혀 실현할 수 없다'는 답변은 19~24세 청년들이 4.1%로 가장 낮았습니다. 삶의 만족도는 19세~24세가 6.82점으로 가장 높았고, 30~34세가 6.76점, 25~29세가 6.60점을 기록했습니다.

청년 3명 중 1명만 '결혼' 긍정적으로 생각

통계청에서는 '사회조사'를 통해 청년층의 결혼, 출산, 노동 등에 대한 가치관 변화를 분석하여 발표하였습니다. 2022년 결혼에 대해 긍정적으로 생각하는 청년의 비중은 10년 전 56.5%보다 20.1%p나 감소한 36.4%로, 청년 3명 중 1명만이 결혼에 대해 긍정적으로 생각하는 것으로 나타났습니다. 2022년 청년들이 생각하는 '결혼하지 않는 사람들이 결혼을 하지 않는 주된 이유'에 대한 응답을 살펴보면 '결혼자금 부족(33.7%)'을 가장 큰 원인으로 꼽았습니다. 다음으로는 '결혼의 필요성을 못 느낌(17.3%)', '출산·양육 부담(11.0%)', '고용상태 불안정(10.2%)', '결혼 상대 못 만남(9.7%)' 순이었습니다. (청년 연령은 <청년기본법>에 따라 19~34세에 해당함)

9월 1일 통계의 날, 근대 통계의 시작은 언제?

가을의 시작을 알리는 9월의 첫날 '9월 1일'은 '통계의 날'입니다. 1896년 9월 1일, 고종황제는 서양의 근대적인 제도를 받아들이자는 갑오개혁의 일환으로 '호구조사규칙'을 공표하였고, 1925년에 간이 국세조사를 실시했으며, 1948년 '제1회 총인구조사 시행령'을 공포하였습니다. 이를 기념하기 위해 1995년부터 매년 9월 1일을 '통계의 날'로 지정하였습니다. 올해로 제29회 통계의 날을 맞아 통계청에서는 기념식, 전시회, 통계인의 밤 등 다양하고 다채로운 행사를 개최하였는데요. '제29회 통계의날 기념식'에서 이형일 통계청장은 "통계를 시의성 있게 이용자 중심의 통계서비스를 제공하고, 국가통계제도를 정비하고 개선하겠다"면서 방향을 제시하였습니다.

2027년, 부산에서 열리는 통계 올림픽 '세계통계대회'

세계통계대회(ISI World Statistics Congresses)는 전 세계의 저명한 통계학자, 각국의 정부와 국제기구의 통계 업무 종사자들이 한 자리에 모여 통계에 관한 이론과 실무적인 문제, 추후 발전 방향을 논의하는 자리입니다. 세계통계대회는 1887년 이탈리아 로마에서 국제통계협회에 의해 처음 개최된 이후 2년마다 열리고 있습니다. 우리나라는 2001년 제53차 서울 대회에 이어 26년 만에 세계통계대회를 다시 유치하게 되었는데요. 지난 7월에 열린 제64차 캐나다 대회에 이어 65차 네덜란드를 거쳐, 66차는 2027년 부산에서 대회가 개최될 예정입니다. 40여 개국과 치열하게 경합한 끝에 국제통계기구(ISI) 집행위원회에서 만장일치로 선정되어 더욱 의미있는 쾌거라고 할 수 있습니다.

예산에서 금산, 지역소멸의 해법을 찾는 '로코노미'

최근 방송에서 큰 관심을 끌었던 예산시장 프로젝트가 이번엔 금산 세계인삼축제로 이어질 전망인데요. 해당 프로젝트는 지역축제와 연계하는가 하면 지역마케팅연구소를 설립하는 등 지속가능성을 실험대에 올려놓은 상태입니다. 엠브레인의 '로코노미 활용 식품 관련 U&A 조사'결과에 따르면, 로코노미 식품 구매 경험은 81.6%로 이미 많은 사람들이 구매 경험이 있는 것으로 밝혀졌습니다. 구매는 주로 오프라인에서 이뤄졌는데요. 지역 상품 판매 매장에서 구매했다는 응답이 49.8%로 절반 가까이 차지하며, 로코노미 식품이 지역 경제에 크게 이바지하고 있음을 보여주고 있습니다. 로코노미란 지역을 뜻하는 로컬과 경제를 의미하는 이코노미를 합성한 신조어를 말합니다.

환경도 자원도 지키는 신재생 에너지

한국에너지공단의 '신·재생에너지 산업통계 결과안내'에 따르면, 2017년 884만 toe였던 에너지 생산량이 2021년 1천 400만 toe을 기록했는데요. 우리 정부는 2021년 기준, 총 발전량 중 신재생에너지의 비중은 8.3%지만 2036년까지 24.7%로 늘리겠다고 선언하여, 앞으로 신재생에너지 사업에 대한 정책적인 지원을 전망하고 있습니다. 또한, 보고서에 따르면 정부뿐만 아니라 기업에서도 신재생에너지를 주목하고 있습니다. 2021년 신재생에너지 사업체 수는 10만 7,833개, 종사자 수는 14만 953명, 매출액은 28조 8,807억으로 나타났는데요. 각각 전년 대비 31.7%, 19.0%, 13.4% 증가한 수치로, 전반적인 신재생에너지 사업의 규모가 커지고 있음을 확인할 수 있습니다.

사람 사는 곳 절반 이상이 아파트! 인기 이유는?

국토교통부의 '주거실태조사'에 따르면, 우리 국민의 주거유형에서 가장 많은 비중을 차지한 건 '아파트' 였습니다. 2006년 41.8%였던 아파트 비중이 2021년에는 51.5%로 늘어났습니다. 또한, 통계청의 '2020 인구주택총조사'를 보면, 1인가구는 50.3%인 334만 가구로 절반을 넘었으며, 혼자사는 주된 이유는 '본인 직장'(34.3%)로 가장 높았고, 가구별 평균 거주기간은 8.7년이었으며, 전국에 빈집은 총 151만 호인데요, 사유로는 매매·임대·이사(42.9%), 가끔 이용(27.1%), 미분양·미입주(13.9%) 순으로 나타났습니다. 우리 국민이 아파트를 좋아하는 이유로는 놀이터, 어린이집, 도서관, 독서실, 헬스장 등의 부대시설과 원활한 주차장, 쓰레기 처리 시설 등의 편리함을 꼽았습니다.

시 지역 상반기 고용률 당진시(71.0%) 2위, 1위는?

통계청에서 발표한 '2023년 상반기 지역별고용조사 시군구 주요고용지표' 결과에 따르면, 9개 도 시지역의 취업자는 1,385만 1천명으로 전년동기대비 24만명 증가, 고용률은 61.9%로 전년동기대비 0.8%p 상승한 것으로 나타났습니다. 시 지역 중 제주특별자치도 서귀포시가 72.0%로 가장 높은 고용률을 보였으며, 충청남도 당진시(71.0%), 경상북도 영천시(67.6%) 등에서 높게 나타났습니다. 아울러 9개 도 군지역의 취업자는 210만 5천명으로 전년동기대비 1만 1천명 증가, 고용률은 68.7%로 전년동기대비 0.3%p 상승한 것으로 나타났고, 군 지역의 고용률 1위는 경상북도 청송군(82.1%)이었으며, 전라남도 신안군(78.6%), 전라북도 장수군(77.8%) 등에서 높게 나타났습니다.

경영자의 데이터 리더십을 위한 키 차트(Key Chart) 접근법

강양석 | Deep Skill 대표

나는 경영자로서 데이터가 흐르는 조직 만들기를 시도 한 적이 있다. 코스닥 상장사 4개를 대상으로 대략 3년에 걸친 대형 프로젝트였다. 경영철학부터 사업계획수립, 목표관리, 회의, 평가, 보상, (역)채용 모든 체계를 데이터가 잘 흐를 수 있도록 꿰는 것이 핵심이었다.
2016년에 코스닥 상장사로는 비교적 이르게 전략기획팀을 폐지하고 데이터과학 팀도 만들었었다. 결과는 실패였다. 이유야 많겠지만, 조직이 데이터를 쓴다는 것이 개인과 팀과 어떻게 달라야 하는지 잘 알지 못했던 것 같다. 양질의 데이터, 내공 있는 분석가, 빛나는 성공사례가 전부가 아닌 것 같다. 4차 산업혁명과 디지털 대전환이 요구하는 조직의 변화는 우리가 생각하는 것보다 훨씬 깊다. 최고 의사결정자의 철학부터 데이터 관리 인프라까지 모조리 바뀌어야 가능한 일이다.

이제 경영자들이 데이터를 봐야 하는 시대가 오고 있다

그런 의미에서 보면, 그간 우리 사회의 데이터 관련 담론은 지나치게 분석가 중심으로 흐른 경향이 있다. 이제는 경영자들이 데이터를 어떻게 봐야 하는지 풍부한 담론이 나와야 할 때가 된 것이다. 분석가의 데이터에서 경영자의 데이터로. 그도 그럴 것이 데이터의 힘은 디지털 기술과 기업 경영의 근간을 통째로 바꿔 놓고 있다. 심지어는 기업의 형태 그 자체까지도 말이다. 이런 변화의 깊이를 제대로 이해하지 못한 경영자는 전투(분석)에서는 승리하지만, 전쟁(혁신)에서는 패배할 수도 있다. 단순히 멋진 분석 사례 하나 둘 나온다고 기뻐할 일이 아니라는 것이다.

분석가들보다는 경영자들이 먼저 데이터에 대한 접근이 필요하다. 더 정확히 얘기하면 전문경영인(CEO)보다는 오너(Owner)이다. 왜냐하면 변화가 깊을수록 기업 경영 체질 자체를 바꿔야 하는데 전문경영인들은 여간해선 무한책임을 동반하는 일을 도맡기가 구조적으로 힘들기 때문이다. 경영철학부터 모든 경영체계를 모두 손볼 수 있는 전문 경영인은 많지 않다. 데이터의 담론은 그만큼 심대하게 대우받을 자격이 충분하다.
그렇다고, 분석가 중심의 데이터 담론은 그 자체로 잘못된 것은 아니다. 모든 일에는 시작이 있듯 기업은 데이터 기반 혁신 사례를 원했고 그들이 가진 기법과 전문성 중심으로 데이터에 대한 관점

도 자연스럽게 형성되었기 때문이다. 그래서 대표이사들도 R과 파이썬을 공부하면서 데이터의 세상에 발을 들여놓기 시작했다.

이러한 분위기가 무르익어 2020년경부터 국내에서도 최고데이터책임자(CDO: Chief Data Officer)들이 생겨나기 시작했고, 시쳇말로 가장 '섹시한' 직업이 되었다. 하지만, 내가 많은 기업 경영자들에게 듣는 말과는 괴리가 있다. 일단, 최고 의사결정자와 최고데이터책임자가 대화가 어렵다는 토로가 많다. 전자는 데이터 리터러시가 부족하고 후자는 비즈니스 리터러시가 부족해서이겠지만 더 근본적인 문제는 경영자들이 데이터가 만들어 내는 인문적이고 경영체계 관점의 기초 담론에 인색했던 게 더 크다.

예를 들면, 최근 국내 7조원 규모의 코스피 상장사 대표이사는 다음과 같은 요구를 조직장들에게 한 적이 있다. '우리 회사 ERP 등을 보면 이미 데이터가 많잖아요? 그러니, 너무 거창한 거 말고, 소소해도 좋으니 기존의 데이터로 작은 분석부터 시작해 봅시다.' 얼마나 자상한 지시인가? 작은 성공부터 만들자는 멋진 철학도 좋고 말이다. 다만, 이 친절한 지시도 현장에서는 굉장한 혼선을 유발한다. 왜냐하면 '있는' 데이터로만 분석을 한다는 것은 실제 실천하기 정말 어려운 개념이기 때문이다. 쉽게 말해, 지금 당장 냉장고 문을 열어 무턱대고 쌓여 있는 식재료를 모아 요리를 만들어 보라는 것과 같기 때문이다. 실제로 분석에 참여한 많은 직원들이 '난 꼭 이 문제를 풀고 싶은데 기존 데이터가 엉망이어서 안되겠는데? 그럼 있는 데이터에 문제를 끼워 맞추어야 하나?'라는 푸념이 여기저기 쏟아졌다. 결국, 이 지시는 철회되었고 난 그 혼란을 외부 자문 위원으로서 생생히 목도했다. 뭘 먹고 싶은지를 정해야 장을 보는 것이다. 장 먼저 보고 뭘 먹을지 정하는 게 아니고 말이다.

너무 많은 분석은 조직을 마비시킨다

하나 더 예를 들어보자. 경영자의 기초 담론이 얼마나 중요한지 말이다. 모든 것이 빠르게 변화는 요새 경영자들은 주야장천 외친다. '혁신적인 사고를 해야 합니다.' 또는 '데이터 기반 과학적 의사결정을 해야 합니다.'라고 말이다. 그런데 이 두 메시지는 오묘하게 상충관계에 있다. 혁신적 사고는 미래의 산물이고, 데이터는 과거의 산물이기 때문에 혁신적 사고일수록 이를 지지할 데이터란 건 없을 공산이 크니 말이다. 좀 더 체계적으로 설명하면, 데이터 기반 의사결정은 직관과 분석의 배합 비율에 따라 크게 3가지로 나뉜다. 데이터 중심(Data-Driven), 데이터 참조(Data-informed) 그리고 데이터 영감(Data-Inspired) 의사결정이 그것이다. 뒤로 갈수록 분석보다 직관의 힘이 더 크다. 경영자가 이 셋을 구분하지 못하면 모든 문제를 데이터로 풀어야 한다는 강박이 생기고 이는 조직 전체를 지적 유희에 빠지게 한다. 이를 딱 표현한 말이 '너무 많은 분석은 조직을 마비시킨다. (Too much Analysis creates Paralysis.)'이다. 경영자는 조직이 어떤 성격의 문제를 주로 고민하는지를 이해하고 '직원들이 회귀분석보다 게스티메이션(guesstimation)을 먼저 훈련할 수 있었으면 좋겠어요.'라고 말할 줄 알아야 한다.

경영자가 R과 파이썬을 배울 때가 아니다. 데이터 기반 혁신의 변화관리 로드맵을 짤 때다. 인류 역사상 전대미문의 코로나라는 비극이 끝난 2023년, 본격적으로 디지털 혁명이 시작되고 있다. 모든 업종과 업태에서 거대한 변화가 불어 닥치는 이때 크건 작건 한 조직을 운영하고 있다는 건 얼마나 가슴 벅차고, 힘에 벅찬 일인가? 이 변혁의 시기에 경영자는 기존 그 어떤 선배 경영자보다 고민이 많을 것이다. 4차 산업 혁명의 한복판에서 변화관리 전략을 짜야하니 말이다. 태풍 속에서 종이비행기를 접는 심정이 이런 걸까?

경영자들의 문제의식을 구체적으로 끌어내기 위한 방식

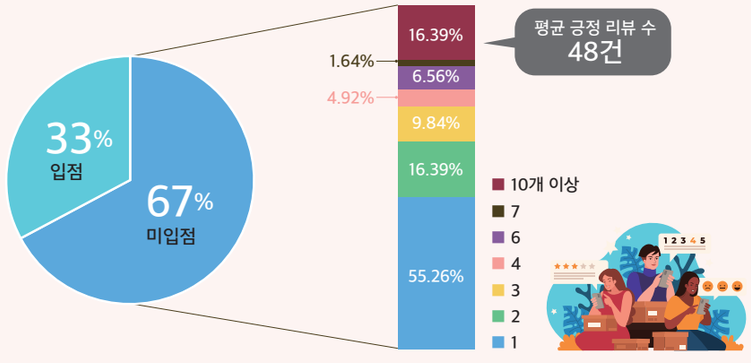
빅데이터 시대에 (당신에게 모든 데이터가 확보되었고, 최고의 데이터 과학자가 함께한다면) 어떤 문제를 풀고 싶냐고 물어보면 뚜렷한 주관이 담긴 답을 하는 경영자가 많지 않다. 자신이 풀고 싶은 문제를 명확히 얘기하지 못하는 경영자가 어떻게 조직을 이끌겠는가? 풀고 싶은 문제를 중심으로 모든 경영자원이 기획되고 투자되기 때문이다. 실제로 디지털 대전환이 본격적으로 논의되기 시작하던 2018년 글로벌 경영 컨설팅사 맥킨지의 조사에 따르면 디지털 대전환 성공 요인 중 압도적으로 중요하게 꼽힌 것이 '경영진이 혁신 스토리를 명확하게 제시할 수 있는가 여부(Management team established clear change story for transformation)'였다.

큰 격차로 그 다음으로 언급된 것이 적절한 기술과 협업 문화 등이었다. 즉, 모든 이야기의 시작은 뭐니 뭐니 해도 리더가 상상을 얼마나 구체화하여 전달하느냐에 있다는 것이다.

상상을 구체화한다고 하면 매우 막연하게 느껴질 것이다. 그래서 내가 경영자들의 문제의식을 구체적으로 끌어내기 위해 간단하지만 독특한 방식을 사용한다. 키 차트(Key Chart) 그리기가 그것이다. 키 차트는 데이터 분석 기획 초기 단계에서 그리는 가설적 차트로서 '만약 데이터 기반 문제 해결이 성공한다면 그 결과를 하나의 차트로 표현해 보면 어떻게 될까요?'에 대한 답이다. 어떤 기술, 데이터, 기법이 사용될지도 모르지만 그 결과물을 미리 하나의 차트로 그려보는 것이다. 이게 가능할까 싶지만 가능하고 효과도 크다. 리더가 가진 현장적인 고민의 지향성, 혁신성, 실효성, 구체성 그리고 실현 가능성이 한 번에 판가름 나기 때문이다. 나아가 기술, 데이터, 기법을 몰라도 혁신을 기획하는데 전혀 지장이 없다는 걸 스스로 깨닫게 된다는 점도 적지 않은 소득이고 말이다.

예를 들어 다음의 국내 선진 유통업체 임원이 그린 키 차트를 보자. 분석의 제목은 '고객 니즈가 빠르게 변하는 화장품 브랜드의 고객 선호도를 월별 평가 분석하여 신규 취급 필요 브랜드를 도출하고 상품 구색 경쟁력을 강화함'이다.

2023년 8월 현재 고객 선호도가 높은 화장품 브랜드의 약 70%는
당사에서 취급하지 않고 있어 취급을 적극적으로 고려할 필요가 있음

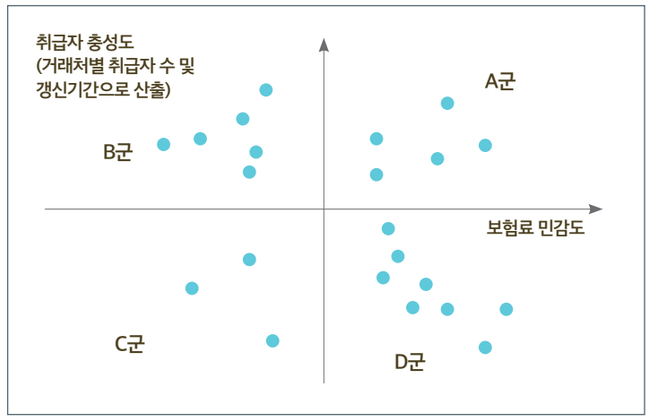


키 차트(Key Chart)는 분석 기획단계에서 분석 결과를 단 하나의 가설적 차트로 표현한 것으로
경영자의 데이터 리더십의 핵심 역량이라고 볼 수 있다. (출처 딥스킬)

어떤 데이터를 어떻게 수집해서 어떤 분석을 진행할지는 전혀 안 나와 있지만 그 모든 활동이 마무리되었을 때 결국 어떤 말을 하고 싶어하는지는 명확히 알 수 있다. '최근 잘 나가는 브랜드를 우리가 취급하지 못하고 있다니까!' 라는 메세지를 주고 싶어하는 의도를 정확히 알 수 있기 때문이다. 주목할 점은 이 차트가 본격적인 분석을 하기 전에 기획 초기 단계에서 나왔다는 것이다. 적어도 데이터 과학자는 이 차트를 보며, '결국 선호도를 어떻게 구체화하는지가 핵심이겠구나!'라는 논의의 방향을 빨리 잡을 수 있게 된다. 의도에 따른 쟁점을 신속히 정하게 된 것이다. 이렇게 분석의 초기 단계에서 작고 핵심적인 부분부터 정해, 점점 더 명확히 해야 할 개념들을 확장해나가다 보면 논의의 집중도를 유지한 채 효과적인 기획을 마무리 할 수 있게 된다.

초기 키 차트를 데이터 과학자와 논의하다 보면 좀 더 구체화된 키 차트를 얻을 수 있게 된다.

손해율 불량한 국내근재보험 계약의 보험료 민감도
취급자충성도 분석을 통해, 각 군별 손해율 개선방안 도출하여 수익성 개선



누적 손해율 100% 이상 계약자 mapping

구분	대응 방안
A군	취급자와 계약별 손해율 개선 방안 사전 협의 하에 보상한도액 인하 등 보험조건 개선 유도
B군	취급자와 협의 하에 보험료 인상 및 수수료 소폭 인상
C군	보험료 인상
D군	보상한도액 인하 등 보험조건 개선 유도하되 이탈 대응

키 차트는 논의를 통해 계속 구체화 될 수 있으며 그 구체화되는 과정에서
데이터 과학자와 경영자는 효율적인 기획을 하게 된다. (출처 딥스킬)

모 금융그룹의 임원이 작성한 이 키 차트에서 흥미로운 점은 분석의 기준(보험료 민감도와 취급자 충성도)을 명확히 제시하고 있다는 점이다. 물론 '취급자 충성도'는 또 어떻게 정의하나요 라는 꼬리에 꼬리를 무는 질문에 맞닥뜨릴수 있지만 그 과정 자체가 논의가 진전되고 있다는 증거이다.

왜냐하면 데이터 과학자들이 정작 힘들어하는 부분은 대용량의 데이터를 가공하는 법을 몰라서가 아니라, 이런 현장의 통찰이 그대로 녹아 들어 있는 분석의 '관점(View Point)'을 정하지 못해서 훨씬 힘들어하기 때문이다.

키 차트 그리기가 주는 시사점

키 차트 실습은 여러모로 흥미로운 시사점을 남긴다.

첫째는 생각보다 고급 디지털 기술과 데이터 과학적 요소를 반드시 필요로 하는 문제는 많지 않다는 것이다.

우리 팀의 경험에 따르면 고급 분석 기법과 빅데이터를 필요로 하지 않는 주제가 약 70%에 달하는 것 같다. 이 점은 디지털 환경과 투자에 대한 매우 중요한 시사점을 제공하고 있다. 우리가 디지털 및 데이터 기반 트렌드 및 성공사례를 접하는 경로를 잘 살펴보면 대부분 그러한 서비스를 수행하는 전문가들의 입을 통할 때가 많다. 그들의 설명은 늘 잘 짜여 있고 체계적이라는 장점이 있지만 다소 거창하고 두렵고 엄두가 나질 않는다. 오죽하면 종교, 학원, 컨설팅 서비스의 공통점은 '공포를 파는 서비스'라는 말까지 나오겠는가? 한마디로 말하면 조직은 그런 두려움에 직면하기 이전에 자신의 문제를 쭉 펼쳐 놓고 본 적이 없는 상태에서 쫓기듯 혁신을 강요당하게 된다는 것이다. 서툴고 거칠지만 당장 우리가 풀고 싶은 문제가 무엇인지 살펴보는 기회가 필요하다. 그러면 현업 리더들이 당장 풀고 싶어하는 문제들의 모음집을 얻을 수 있다. 그 모음집에서 어떤 기술, 어떤 데이터가 필요할지에 대한 청사진이 나온다. 기술에서 시작하면 딱 과투자하기 십상이다.

둘째는 리더들이 문제정의 하는 것을 매우 어려워한다는 것이다.

물론 키 차트라는 형식이 생소할 수도 있겠지만, 핵심은 혁신을 글과 그림으로 구체화해보는 습관이 부족하다는 데 있다. 혁신이 표현되는 것은 중요하다. 실제로 나는 디지털 기술과 주요 데이터 과학 성공 사례를 설명하기보다 '엄격한 글쓰기(technical writing)'에 더 공을 들인다. 글을 쓰면 생각이 도망갈 구석이 없다고 하듯 리더가 혁신을 글로 표현하면 생각이 구체화되고 논의가 가능해지며 자신의 생각의 사각지대를 스스로 탐지할 수 있는 기회를 얻게 된다. 그래서, 주위 전문가들의 산출물을 검토하는 능력 못지 않게 모든 혁신의 시원성을 가진 가장 작은 생각의 단위인 키 차트 그리기만큼은 그릴 줄 알아야 한다. 이런 리더의 글쓰기를 가장 강조한 사람을 꼽으라면 아마존의 창업자 제프 베조스를 들 수 있다. 전 세계에서 데이터를 가장 잘 쓰는 기업으로 구글과 꼭 함께 언급되는 아마존 말이다. 그는 리더의 글쓰기 능력에 광적으로 집착했다고 전해지는데 실제 키 차트로 혁신을 표현해 보는 연습을 해보면 실감할 수 있다. 읽지 않

아도 읽히게 혁신을 표현하는 리더와 십여 분을 얘기해도 무슨 말을 하는지 이해할 수 없게 하는 리더가 분명 존재한다는 것을 말이다. 혁신적인 기획이냐 아니냐는 고사하고 일단 데이터 과학자가 알아듣게 자신의 의도를 표현할 줄 아는 경영자가 적다는 의미다.

셋째는 현장 경영진의 문제를 구체화하는 능력이 데이터 기반 혁신의 한계를 결정한다는 것이다.

데이터 기반 혁신 기업들의 혁신 과정을 살펴보면 크게 3가지 단계로 구분 지어 볼 수 있는데 첫번째 단계는 데이터 전문가들에 의해 혁신 사례가 나오는 단계이고 둘째는 현업 부서 곳곳에서 고유의 문제의식을 바탕으로 데이터 기반 문제해결 사례가 나오는 단계이다. 그리고 마지막 단계는 첫 번째와 두 번째 단계가 융합되어 거대한 기업 기억(Corporate Memory)을 갖게 되는 단계가 된다. 기업 기억은 한 조직이 크고 작은 자신의 문제해결 경험을 통합하여 거대한 지식을 자산화한 결과물이다. 당연히 우리 조직에 가장 특화된 인공지능의 트레이닝 재료가 된다는 점에서 매우 중요하다.

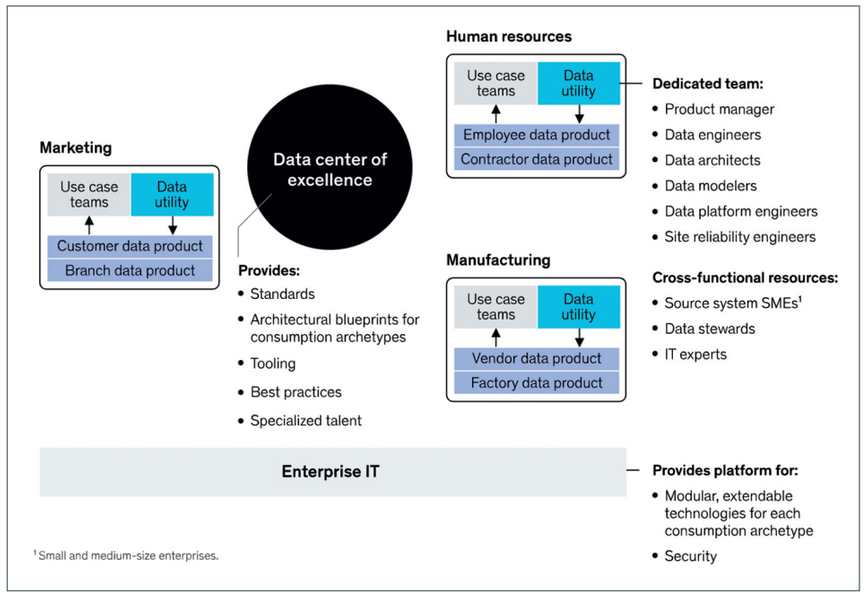
그렇다면, 이 3개 단계 중 가장 넘기 힘든 구간은 어디일까? 단연 2번째 단계이다. 왜냐하면 어떤 팀이 키 차트 산출물을 가장 원할지를 보면 알 수 있는데 그게 바로 데이터 과학자 팀이기 때문이다. 왜 데이터 과학자들은 현장 경영진의 키 차트 실습 결과물을 그렇게 원할까? 2번째 단계에서 반드시 필요한 문제집을 자신들만의 상상만으로는 채울 수 없기 때문이다. 실제로 임원진 데이터 리더십 워크샵을 직접 주관하는 데이터 과학자팀도 많았으며 심지어 그 결과물로 차년도 사업계획을 세우려는 디지털 본부장도 여럿 보았다. 이런 현상을 살펴보면 현장 리더의 과제를 상상하는 능력만큼 혁신이 일어난다 해도 지나침이 없다.

문제로 정의하는 과정이 모든 이야기의 시작

현장 경영진의 상상력이 데이터가 흐르는 조직의 핵심이라는 입장은 다양한 사례에 의해서 뒷받침 되고 있다. 데이터 기반 조직 변화관리 전문가이자 현재 호주 국영 통신기업 NBN CO의 CDO(Chief Data Officer)인 소니아 보이예(Sonia Boije)는 2021년 맥킨지와의 인터뷰에서 다음과 같이 말했다. “데이터 유스케이스(Data Use case)를 지속적으로 수집하고 관찰하다 보면 정말 중요한 데이터가 무엇인지 보입니다.” 여기서 그가 말하는 유스케이스가 바로 앞서 키 차트로 표현된 ‘문제’에 해당한다. 쉽게 말해 “누울 자리를 보고 다리를 뻗어야 한다”는 얘기다.

최근 그는 필자가 속한 딥스킬 팀과의 인터뷰에서 자신이 이끌었던 팀은 크게 데이터 유스케이스 파트[1]와 데이터 유틸리티 파트[2]로 구성하고, 그 중 유스케이스 팀은 “어떤 문제를 풀 것인가?”, 유틸리티 팀은 “어떤 데이터를 쌓고 관리할 것인가?”에 관한 업무를 전담한다고 했다. 팀의 구성에서부터 문제와 데이터의 균형을 강조한다는 의미다.

Managing data like a product requires the right operating model.
(출처:맥킨지앤컴퍼니)



1) https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/unlocking-success-in-digital-transformations
2) https://www.mckinsey.com/capabilities/quantumblack/our-insights/how-to-unlock-the-full-value-of-data-manage-it-like-a-product

맥킨지는 2021년 글 “어떻게 데이터의 잠재 가치를 깨울 것인가?(how to unlock the full value of data)”에서 데이터 전문가 조직(Data center of excellence)을 운영할 때 문제를 담당하는 유스케이스 팀(Use case teams)과 데이터를 담당하는 데이터 유틸리티 팀(Data Utility)을 구성하는 게 기본이라고 설명했다.

여기서 더욱 흥미로운 점은 데이터 전문가 조직이 굳이 데이터 분석가(Analyst)들로 구성될 필요가 없다는 점이다. 소니아 보이예 역시 데이터 과학자가 아닌 마케터 출신이었다. 오히려 현업의 문제를 더 잘 이해하는 전문가와 데이터 분석을 지원하는 데이터 엔지니어들로 구성하는 게 나을 수 있다. 2022년 현재 국내, 아주 극소수 기업에서도 이런 데이터 분석가 없는 데이터 전문 조직이 태동하고 있다.

LG전자가 대표적이다. LG전자 내 데이터 전문조직은 일부 데이터 분석가가 있기는 하지만 주로 데이터 엔지니어와 데이터 거버넌스 전문가 등으로 구성돼 있다. 당연히, 이 팀의 모토는 ‘데이터로 (데이터 과학자뿐만 아닌) 전 구성원이 성과를 내게 하는 것’이다. 이런 모토에서 문제와 데이터의 균형감을 놓치지 않기 위해 노력하고 있음을 알 수 있다.
또한 이처럼 데이터 분석 성숙도가 높은 조직에서는 분석가는 현업 팀의 구성원이라고 보는 시각이 강하며 인재개발(Human Resource Development)부서와의 유대감을 중요하게 생각한다. 실제 실무 문제에 기반한 데이터 교육이 공허하지 않다고 믿기 때문이다. 물론, 데이터 기반 혁신을 이제 막 시작하는 모든 기업이 이런 고도화된 역할 구조를 처음부터 갖는 것은 어렵겠다.
하지만 균형감을 놓치지 않기 위한 방안을 치열하게 고민해야 한다. 경영진을 중심으로 무수한 혁신을 상상하고 문제로 정의하는 과정이 모든 이야기의 시작이다. 어렵고 복잡하고 무겁게 시작하려 하지 말고 그냥 종이, 연필, 지우개를 가지고 혁신의 제목과 키 차트를 그려보자.

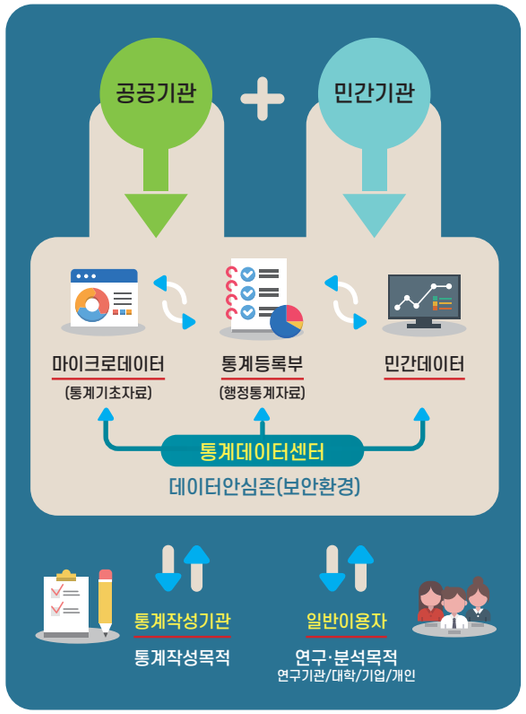
SDC
Statistics
Data Center

행정통계자료와 민간자료를 한곳에!

통계데이터센터 서비스

통계데이터센터가 새로운 서비스로
정보화 사회를 선도합니다.

행정자료를 수집하여 가공한 행정통계자료(통계등록부),
통계청이 제공하는 승인된 통계기초자료(마이크로데이터) 등
통계자료뿐만 아니라 민간자료까지 한 곳에서 분석이 가능한 통계데이터센터(SDC)



1 분석 플랫폼 제공 서비스
- 분석시스템 · 통계패키지 제공
- 통계자료(통계등록부 · 통계기초자료) 및 민간자료, 이용자 반입자료 연계 · 분석

2 전문가 분석지원 서비스
- 분석 경험이 없는 이용자를 위한 데이터 분석 지원
- 센터 이용 상담 및 데이터 분석 자문

3 주문형 분석서비스
- 시간 및 거리상 센터 방문이 어렵거나 직접 자료분석을 하기 힘든 이용자를 위한 서비스
- 센터 이용자료를 활용하여 연계 · 분석 후 이용자가 원하는 형태로 결과를 제공

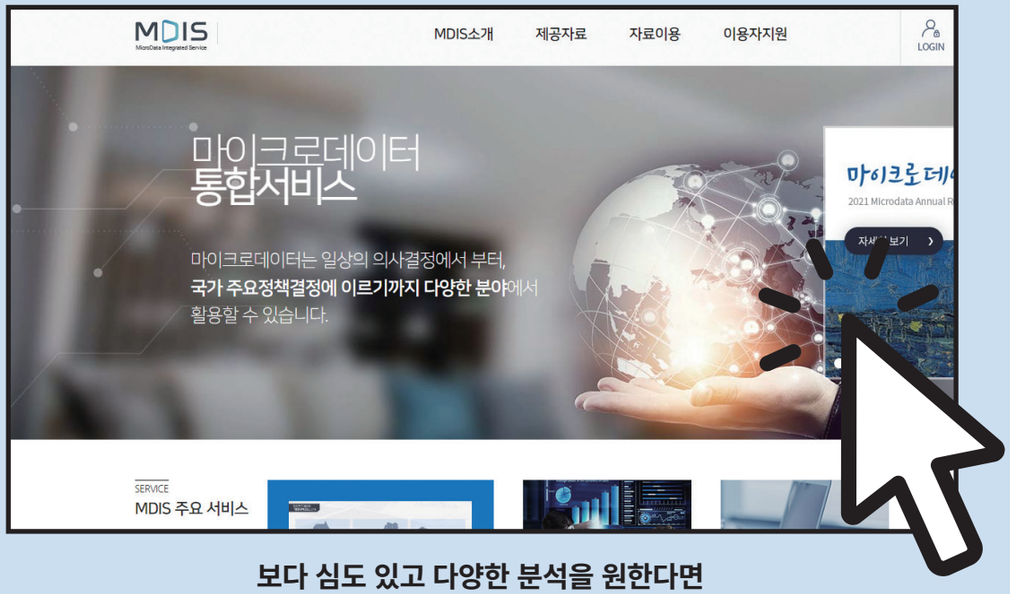
4 명부 서비스
- 분석센터로 방문하여 자료분석 및 표본설계를 통해 데이터 반출

5 이용자 교육 서비스
- 이용자 교육 홈페이지 운영
- 통계분석 프로그램 및 분석사례 교육
- 매년 통계데이터 활용대회 개최

※ RDC 제공자료도 이용 가능합니다.

통계청, 정부부처, 지방자치단체, 연구기관 등

모든 기관의 마이크로데이터를 한 곳으로



보다 심도 있고 다양한 분석을 원한다면
지금 바로 MDIS를 클릭해 보세요.

■ 서비스 소개 (2023년 5월 기준)

가. 서비스명 : 마이크로데이터통합서비스(MDIS, mdis.kostat.go.kr)
나. 제공 통계 수 : 21개 주제별 총 357종 통계 제공
(통계청 50종 및 통계작성기관 307종)
다. 제공 형태 : 마이크로데이터(통계에 따라 사람, 사업체, 가구 기반 자료)

기준		주요 통계
통계청	인구 · 가구	경제활동인구조사, 가계동향조사, 국내인구이동통계, 사망원인통계, 가계금융복지조사, 지역별고용조사, 인구주택총조사, 인구동향조사, 생활시간조사, 사회조사 외 8종
	사업체·농어가	전국사업체조사, 광업제조업조사, 농가경제조사, 기업활동조사, 농림어업총조사, 농산물생산비조사, 경제총조사, 어가경제조사, 운수업조사 외 14종
	행정통계	귀농어귀촌인통계, 영리법인기업체행정통계, 신혼부부통계, 주택소유통계, 중장년층행정통계, 퇴직연금통계, 일자리행정통계, 기업생멸행정통계, 육아휴직통계
통계작성기관		전국다문화가족실태조사, 가족실태조사, 자동차주행거리통계, 직종별사업체노동력조사, 보육실태조사, 기상관측통계, 국민여가활동조사, 외래관광객조사, 한부모가족실태조사, 청소년종합실태조사 외 297종

■ 서비스 내용

가. 구분 : 자료의 민감성 정도에 따라
공공용, 인가용으로 구분 운영

나. 수수료
- 무료 : 공공용 자료
- 인가용 : 선택제 수수료 부과

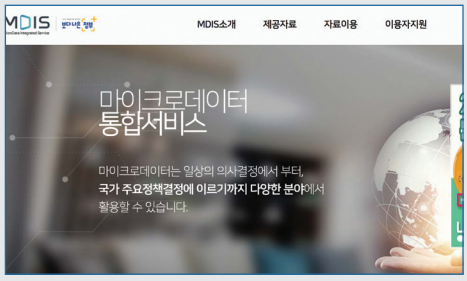
다. 서비스 방법
- 추출 · 다운로드 : MDIS 포털에서 직접 무료 다운로드
- 원격접근서비스 : 승인 후 이용자가 집·사무실 등에서 통계청 서버 접속 후 활용
- 이용센터 : 승인 후 지정된 장소를 방문·활용

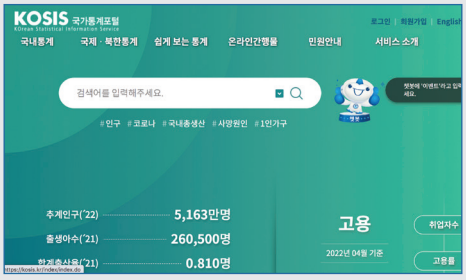
■ 문의

- 연락처 : 재단법인 한국통계진흥원
- 전화 : (02) 512-0167 FAX :(02) 515-0240
- 주소 : (우) 06097
서울특별시 강남구 선릉로 612, 6층
- E-mail : MDIS@stat.or.kr

통계청에서 국가통계를 활용하세요!

통계청은 통계개발·활용·교육에 필요한 모든 정보와 도움을 제공합니다.
다양한 국가통계정보 제공 사이트를 활용하세요.





통계교육원

sti.kostat.go.kr

국내 유일의 국가통계교육 전문기관

통계 작성 및 활용 전문통계과정,
기관맞춤형과정, e-러닝 과정

통계데이터센터

data.kostat.go.kr

행정통계자료와 민간자료를 한곳에

행정통계자료(통계등록부), 민간자료의
연계·융합이 가능한 데이터 플랫폼

MDIS

mdis.kostat.go.kr

원하는 자료를 직접 분석 및 요청

온라인으로 추출/다운로드 선택 시
공공용 마이크로데이터를 무료로 분석 활용 가능

KOSIS

kosis.kr

국가통계 쉽게 찾기

국내, 국제, 북한의 주요 통계를
한 곳에 모아 알기 쉽게 분류해 제공

SGIS

sgis.kostat.go.kr

지도 위 통계정보 살펴보기

인구, 가구, 주택, 사업체 통계 등 각종 통계를
지도(GIS) 위에서 한눈에 파악

통계청
통계교육원