

# 통계의 창

WINDOW OF STATISTICS

2023.  
SUMMER  
VOL.31

## 인공지능과 통계



챗GPT가 가져다준 빛과 어둠 | 다시 짚어보는 인공지능, 그리고 챗GPT  
안전한 사회 구축을 위한 합리적 기준 한국범죄분류 | 글로벌 통계활동의 핵심 기구 「유엔통계위원회」

# CONTENTS

통계의 창  
2023. Summer  
Vol.31

**발행일** 2023년 5월 4일  
**발행인** 정향우  
**발행처** 통계교육원  
**기획** 황현식, 최희정, 김병준  
**주소** 대전광역시 서구 한밭대로 713(월평동) 통계센터 통계교육원  
**전화** 042-366-6151, 6152  
**팩스** 042-366-6498  
**이메일** ccn420@korea.kr, joonk@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

<b>다시 짚어보는 인공지능, 그리고 챗GPT</b> 정복문   신한대학교 마이크로디그리대학 교수	<b>004</b>
<b>챗GPT가 가져다준 빛과 어둠</b> 김명주   서울여자대학교 정보보호학과 교수	<b>014</b>
<b>챗GPT-4 인공지능 시대, 필요한 미래 인재 교육은 무엇일까</b> 안종배   국제미래학회 회장	<b>020</b>

# 2

## 통계 FOCUS

<b>안전한 사회 구축을 위한 합리적 기준 한국범죄분류</b> 김태균   통계청 통계기준과 주무관	<b>026</b>
<b>글로벌 통계활동의 핵심 기구 「유엔통계위원회」</b> 진유강   통계청 국제협력담당관실 사무관	<b>032</b>
<b>AI가 불러온 변화 데이터가 지배하는 축구</b> 김종원   데이터 분석가(비프로일레븐)	<b>036</b>
<b>조선시대, 송남잡지에서 살펴본 아날로그 데이터 사이언스</b> 최원재   클라우드와이즈 대표	<b>042</b>
<b>세계 최초 개발된 「보이스피싱 음성분석 모델」</b> 김현우   행정안전부 통합데이터분석센터 주무관	<b>048</b>
<b>데이터 사이언스의 핵심은 빅데이터</b> 박성현   서울대학교 통계학과 명예교수	<b>052</b>
<b>기후 변화 예측 인공지능, 어디까지 와있나</b> 함유근   전남대학교 해양학과 교수	<b>060</b>
<b>통계로 바라보는 세상이야기</b> 신동헌   도서출판 지일박스 대표	<b>066</b>
<b>「2022 개정 수학과교육과정」에서 나타난 실용통계교육의 방향</b> 고은성   전주교육대학교 교수	<b>070</b>

# 3

## 통계 EDU

<b>분석은 질문이다...챗GPT에게 묻다</b> 강양석   Deep Skill 대표	<b>074</b>
<b>통계교육원의 빅데이터프로젝스수행 (SAS기반) 교육 수가-1, 2</b> 김세윤   교육생, 김지안   교육생	<b>080</b>
<b>교육과정별 운영계획</b>	<b>084</b>



# 01 다시 짚어보는 인공지능, 그리고 챗GPT

정복문 | 신한대학교 마이크로그리대학 교수

최근 매체에서는 인공지능, Artificial Intelligence 줄여서 AI라는 말이 매우 많이 나타나고 있습니다. 인공지능이란 인간의 지능처럼 고도의 문제해결능력을 가진 인공적 지능을 말합니다. 좀 더 단순히 말하면 사람의 지능을 흉내 내는 컴퓨터 소프트웨어 또는 컴퓨터를 말합니다.

공상과학소설이나 영화 즉, SF를 좋아하시는 분들은 익숙한 용어라 생각이 됩니다. SF 영화에서는 오래전부터 인간처럼 생각하고 행동하는 로봇이나 컴퓨터 비서들이 주인공이나 주인공을 돕는 중요한 역할들로 나타나고 있습니다.



인공지능이라는 용어가 일반적으로 크게 퍼진 계기는 2016년 3월, 우리나라 최고의 바둑기사 이세돌과 구글의 인공지능 기술인 알파고의 대결이라고 할 수 있습니다.

당시에 많은 사람이 깜짝 놀랐던 것이 있습니다. 왜냐하면, 컴퓨터의 연산 기술이 아무리 발달해도 인간을 능가할 수 없다는 분야가 있었는데, 그 중에 하나가 바둑이었기 때문입니다.

이 세기의 대결에서 이세돌 기사가 한 번의 승리를 거두었는데, 그 이후 현재까지 수많은 인간과 인공지능의 바둑 대결이 있었지만, 그 한 번의 승리가 인간이 인공지능을 이긴 유일한 사례로 기록되고 있습니다. 이 사건으로 인해 많은 사람이 인공지능에 많은 관심을 가지게 계기가 되었다고 생각합니다.



AI 포스터

## 영화로 살펴보는 인공지능의 의미

인공지능이라는 기술은 최근에 새로 나타난 기술이 아닙니다. 예전부터 인공지능이라는 기술을 소재로 한 공상과학소설이나 영화가 인기를 끌어왔습니다. 최근에는 너무 당연하게 쓰이는 소재이지만, 인공지능이 메인 소재로 쓰인 영화를 몇 편 소개하겠습니다.

### 영화 [AI] ▶ 기계인지, 인간인지에 대한 정체성 혼란

2001년에 미국의 유명한 감독인 스티븐 스필버그가 만든 AI라는 영화가 있습니다. 제목부터 인공지능이라고 명명한 영화인데, 이 영화를 소개하는 이유는 인공지능을 가진 로봇이 과연 기계인가 인간인가를 고민해보는 시발점이 된 영화라고 생각하기 때문입니다.



내용을 잠깐 소개하자면, 이 기사를 읽는 분들은 피노키오라는 동화를 알고 계실 겁니다. 피노키오라는 동화를 보지 못했더라도 거짓말을 하면 코가 길어지는 나무 인형에 대해서는 알고 계실 건데요. 이 영화는 피노키오 동화와 맥락을 같이하는 주제라고 생각합니다.

피노키오의 주된 내용은 나무 인형인 피노키오가 자신을 만들어준 제페토라는 아빠에게 진짜 인간 아들이 되기 위해 모험을 하는 여정을 보여주는데요, AI라는 영화 역시 인간 부모에게 입양된 인공지능 로봇이 자신이 기계인지, 인간인지에 대한 정체성 혼란을 겪고 진짜 자신이 무엇인지를 찾아 떠나는 내용입니다.

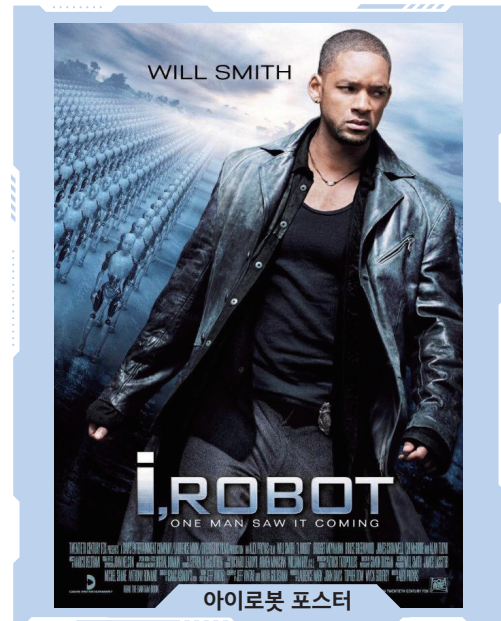
제가 이 영화에서 주목하는 것은 인공지능 기술의 발달 보다는 인공지능이 탑재된 로봇이 과연 인간의 잣대에서 윤리적인 부분이나 인권같은 것을 어떻게 적용해야 하는 것을 고민하고 있었다는 부분입니다. 조금 오래된 영화이지만 한번 찾아보는 것도 좋을 것 같습니다.

## 영화 [아이로봇] ▶ 로봇 3원칙

두 번째 소개할 영화는 아이로봇이라는 영화입니다. 이 영화에서는 로봇영화에서 다루는 중요한 원칙이 나오고 있습니다. 유명한 SF소설 작가인 아이작 아시모프는 1942년 내놓은 소설에서 ‘로봇 3원칙’을 내놓았습니다.

로봇 3원칙은 다음과 같습니다.

**제1원칙 :** 로봇은 인간에게 해를 입혀서는 안 된다.



그리고 위험에 처한 인간을 모른 척해서도 안 된다.

**제2원칙 :** 제1원칙에 위배되지 않는 한, 로봇은 인간의 명령에 복종해야 한다.

**제3원칙 :** 제1원칙과 제2원칙에 위배되지 않는 한, 로봇은 로봇 자신을 지켜야 한다.

이 원칙을 보면 로봇은 인간이 소유한 재물로써 인간을 지키기 위한 존재로 묘사됩니다.

하지만 로봇이 지능을 가지고 있고, 자신을 지켜야 한다는 자각을 한다면 이 원칙을 모두 지킬 수 있을까요?

이에 대한 고민을 잘 나타내고 있는 영화가 2004년에 개봉된 “아이로봇”이라는 영화라고 생각합니다. 이 영화에서는 미래의 발달된 사회에서 저명한 로봇공학자가 살해되는 사건이 발생합니다. 범인으로 지목된 용의자는 인간이 아닌 박사가 만든 인공지능 로봇이었습니다. 당연히 로봇 3원칙을 적용하면 일어날 수 없는 사건이었던 거죠.

주인공은 미국의 유명한 배우 윌 스미스입니다. 윌 스미스는 로봇을 싫어하는 형사 역할로 출연하고, 윌 스미스와 범인으로 지목된 인공지능 로봇 “씨니”의 갈등과 추격전이 영화의 주된 내용입니다.

이 영화에서 주목할 것은 인공지능을 가진 로봇이 단순히 기계로 치부될 것인가, 아니면 하나의 인격체로써 존중받아야 할 존재인가를 고민하는 것입니다.

2004년도에 개봉되어 오래된 영화라고 할 수 있지만, 현재 또는 미래에서도 우리가 인공지능의 발달과 더불어 고민하고 해결해 나가야 할 주제를 잘 표현하고 있다고 생각합니다.

## 영화 [엑스마키나] ▶ 튜링 테스트

세 번째 소개할 영화는 2014년에 개봉한 “엑스 마키나”라는 영화입니다. 엑스 마키나는 원래 연극에서 쓰이던 용어입니다. 옛날 연극에서는 극이 진행되다가 서사가 꼬여버려서 인과관계가 부족해지는 경우에 신(God)을 등장시켜서 모든 것을 해결해 버리는 경우가 있었는데요, 이렇게 신이 등장할 때, 기계장치나 특수효과를 이용해서 멋지게 출연하는 것을 뜻하는 용어였습니다. 이렇게 꼬여버린 상황을 전지전능한 신이 등장하여 해결하는 것을 “데우스 엑스 마키나”라고 합니다. 해결하기 힘든 상황을 간편하게 풀어버리고 이를 정당화하려는 캐릭터나 연출등을 줄여서 엑스 마키나라고 하는 경우가 있습니다.

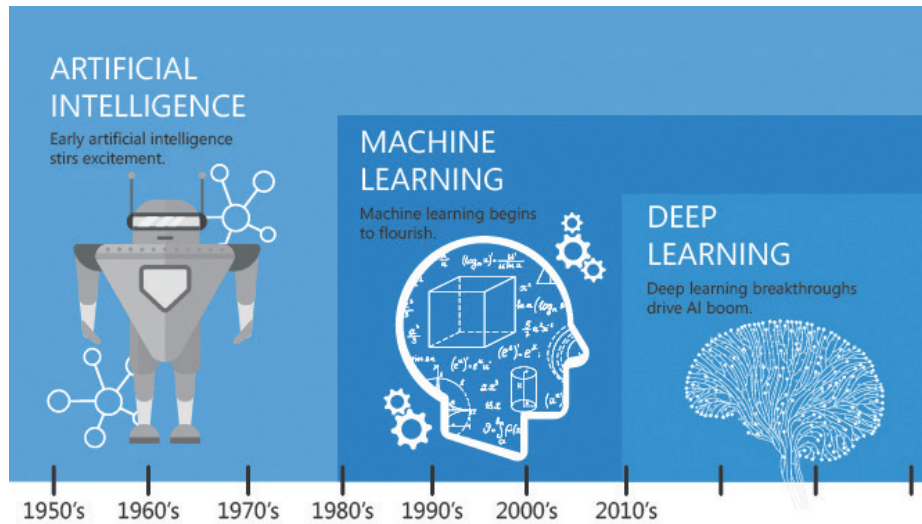
이 영화는 연출, 연기, 특수효과 등 여러 면에서 좋은 평가를 받은 작품이지만 개인적으로는 호불호가 갈릴 수 있습니다. 영화 전체가 조금 어두운 분위기거든요. 제가 이 작품을 인공지능 수업에서 많이 소개하는 이유는 이 영화의 메인 주제가 인간이 인공지능을 가진 프로그램 또는 로봇을 과연 기계라고 인식하는지 아니면 인간이라고 인식하는지를 테스트하는 “튜링 테스트”를 주제로 삼고 있기 때문입니다.

튜링 테스트는 1950년 앨런 튜링에 의해 개발된 “인간의 것과 동등하거나 구별할 수 없는 지능적인 행동을 보여주는 기계의 능력에 대한 테스트”입니다. 다시 말해 음성이건, 텍스트이건 기계와 대화를 하고 사람이 기계인지 사람인지 구분을 못 하면 튜링 테스트를 통과했다고 판단되는 테스트입니다. 앨런 튜링은 2차 세계대전에서 적국의 암호를 해독



하는 등의 활약으로 유명한 수학자이자 공학자인데요, 이 분의 영웅적인 활동을 다룬 영화가 “이미테이션 게임”이라는 영화가 있습니다.

이미테이션 게임이라는 영화는 인공지능에 대한 주제가 아니라 앨런 튜링의 활동에 초점을 맞춘 영화이기 때문에 따로 내용을 소개하지는 않겠습니다만, 앨런 튜링의 능력과 인간적 고뇌를 잘 표현한 영화라 생각하기에 한번쯤 찾아봐도 좋을 것 같습니다.



[그림 1] 인공지능의 발전 과정

출처: <https://www.linkedin.com/pulse/artificial-intelligence-ai-vs-machine-learning-deep-natarajan-siva/>

최근에 이슈가 되고 있는 채팅 프로그램들이 많이 있습니다. 우리나라에서는 심심이, 이루다 등의 프로그램들이 출시되어 많은 사람에게 인기를 끌었습니다. 그리고 인공지능을 연구하는 OpenAI라는 회사에서 2022년 12월 1일 “챗GPT”를 공개하였습니다. 챗GPT는 다른 프로그램들과 유사하게 인공지능이 빅데이터 분석을 바탕으로 사람과 대화하듯 답을 해주는 시스템입니다. 챗GPT는 출시되자마자 5일 만에 100만 명의 사용자를 확보하는 선풍적인 인기를 끌고 있습니다.

현재까지는 튜링 테스트를 통과한 고도화된 인공지능 기술이 거의 없긴 합니다. 하지만 곧 인간인지 기계인지 구분할 수 없는 그런 성능의 인공지능이 나올 것은 틀림없을 것입니다.

앞서 서술했듯이 인공지능이라는 기술이 최근에 갑자기 나타난 기술은 아닙니다.

로봇 3원칙, 튜링 테스트라는 내용만 보아도 1940년도에 이미 인공지능이라는 것을 예견하고 있었던 거죠. 물론 당시에는 초보적인 컴퓨팅 기술을 가지고 있었지만, 인공지능이라는 기술에 대한 고민은 컴퓨터 기술의 초기부터 고민되고 연구됐던 것입니다.

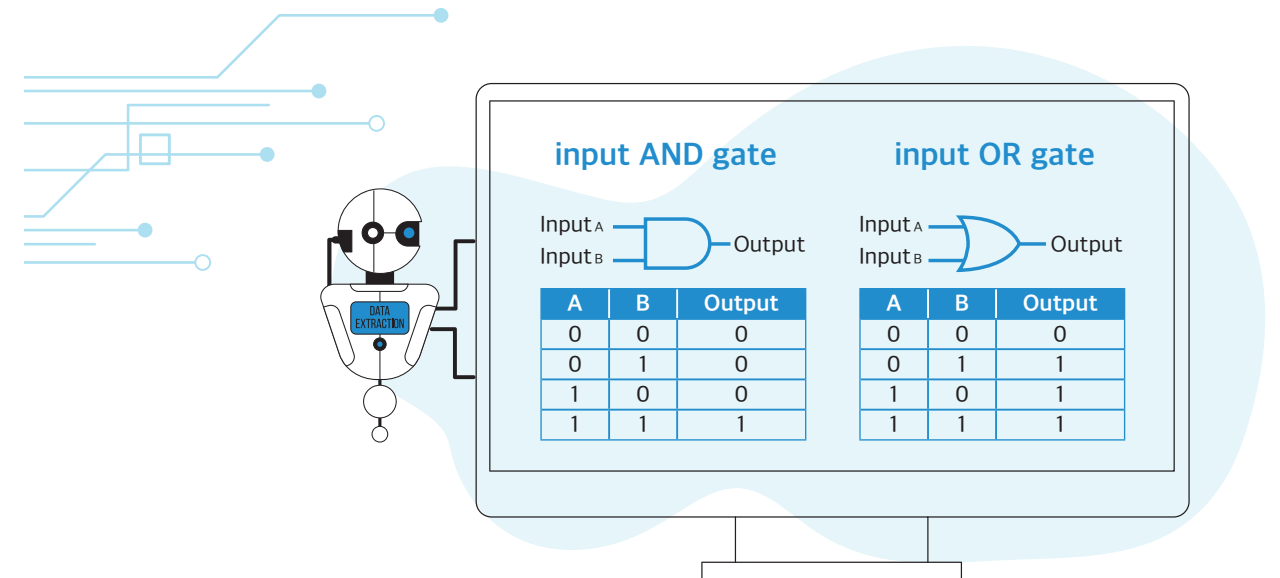
## 인공지능은 어떻게 발전해왔나

[그림 1]을 보시면, 인공지능이라는 개념 속에 머신러닝 즉 기계학습 그리고 딥러닝이라는 기술이 있습니다. 최근에 출시되는 인공지능 기술은 기계학습과 딥러닝을 토대로 진화하고 있는데요, 인공지능 기술의 발전 역사에 대해서 간단히 알아보도록 하겠습니다.

1950년 영국수학자 앨런 튜링은 “계산 기계와 지능”이라는 논문을 통해 기계가 생각할 수 있는지 테스트 하는 방법, 지능적 기계의 개발 가능성, 학습하는 기계 등에 대해서 기술하였습니다. 어쩌면 현대의 인공지능 기술을 이미 이때 모두 서술한 것이라 할 수 있을 정도입니다.

앨런 튜링이 개발한 튜링 머신은 존 폰 노이만 교수에 직/간접적인 영향을 주어 현대 컴퓨터 구조 프로그램 내장형 방식의 표준이 되었습니다.

프로그램 내장형 방식은 현재 우리가 쓰고 있는 모든 컴퓨터의 구조입니다. 운영체제가 탑재되고 프로그램을 설치하고 사용하는 것이 컴퓨터를 구성하

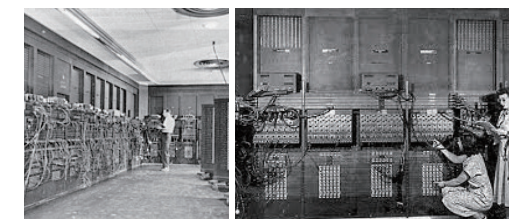


AND, OR 게이트

는 CPU, RAM, 하드 디스크등에 모두 들어있고, 우리는 그것을 소프트웨어적인 관점에서 사용하는 것입니다. 지금은 이것을 너무나 당연하게 생각하고 있지만, 컴퓨터 초기모델은 컴퓨터의 기능을 이용하려면 [그림 2]에서 보이는 것처럼 집보다 큰 컴퓨터에 플러그를 옮겨 꼽으면서 프로그램을 사용해야 했습니다. 이와 같은 불편함을 없애고 모든 프로그램 내용을 컴퓨터 내부로 넣어버린 방식이 바로 프로그램 내장형 방식이라고 할 수 있습니다.

이와 같이 현대 컴퓨터를 쓰는 우리는 상상할 수 없는 극악한 환경에서 인공지능이라는 개념까지 생각했다는게 너무나 대단하지 않나요?

이와 더불어 워런 맥컬록(Warren Mc Cullonch)과 논리학자 월터피츠(Walter Pitts)가 전기 스위치체



[그림 2] 초기 컴퓨터 형태

럼 온, 오프 하는 기초기능의 인공신경을 그물망 형태로 연결하면 사람의 뇌에서 동작하는 아주 간단한 기능을 흉내 낼 수 있다는 것을 이론적으로 증명하였습니다. 인공지능 기술이 시작된 거죠.

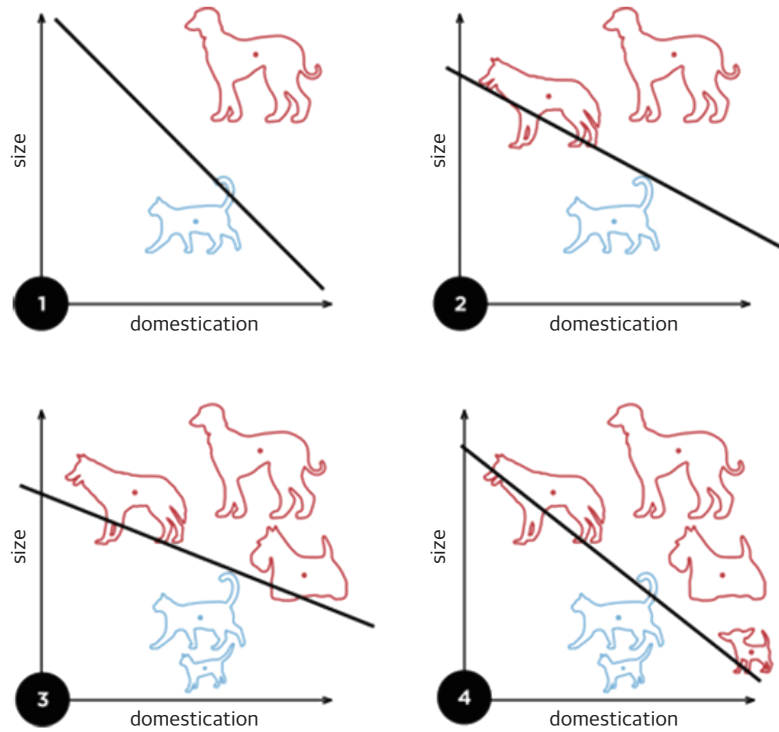
그리고 뇌 신경을 모사한 인공 신경 뉴런인 퍼셉트론(Perceptron)이 탄생하게 됩니다.

1950년대는 인공지능 연구의 시작이자 부흥기라고 할 수 있습니다.

하지만 1960년대에 컴퓨터 논리적인 문제로 인하여 퍼셉트론이라는 이론은 AND 또는 OR 같은 논리적 분리는 가능하지만, XOR문제는 적용할 수 없다는 것이 마빈 민스키와 세이무어 페퍼트라는 학자에 의해 증명되었습니다.

컴퓨터는 기계 내부적으로 0과 1이라는 두 개의 논리적인 구분된 값으로 데이터를 저장합니다. 즉, 단순히 표현하자면 예, 아니오로 대답할 수 밖에 없다는 것이죠. 그런데 AND, OR 논리 구조는 매우 쉽습니다. 그림에서 보듯이 2개의 입력값이 있다면, 입력된 모든 값이 참일 경우 참의 결과가 나오는 논리적인 계산을 AND 계산이라고 합니다. 그리고 참이 하나라도 있으면 참의 결과가 나오는 논리적인





선형으로 분리 가능한 분류

계산을 OR 계산이라고 합니다.

이렇게 입력값에 따른 구분 결과를 분류하기 위해서는 하나의 선이 있으면 됩니다.

하지만 XOR 논리식은 참 입력의 개수가 홀수일 때 결과가 참이 되는 논리식입니다. XOR의 논리식에 의한 결과를 분류하면 선으로 구분할 수 없다는 것이죠.

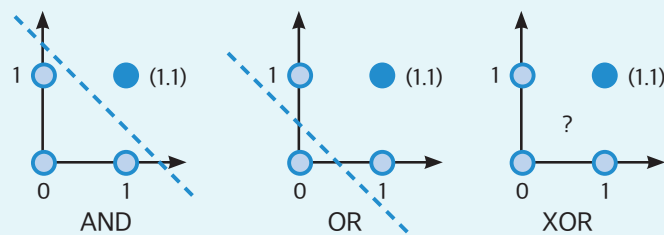
단순한 AND, OR 논리식에서 XOR 같은 컴퓨터의 논리적인 조합이 폭발적으로 늘어나면서 당시까지

나온 인공지능의 이론으로는 그 논리를 모두 다룰 수가 없었습니다. 그래서 인공지능 연구는 많이 중단되었고, 암흑기에 접어들게 됩니다.

그리고 1970년대, 80년대, 90년대를 거치오면서 컴퓨터 데이터는 실용적인 통계기술로써 많이 쓰이게 되고, 데이터를 조작하고 분류하는 데이터마이닝이라는 이름으로 산업의 비효율성을 해결하는 빅데이터 기술로 발전하게 됩니다.

특히 90년대는 추론엔진 기술이 개발되었는데, 0

INPUT		OUTPUT
A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0



XOR 논리식

과 1 사이에도 여러 가지 값을 가질 수 있는 퍼지 이론을 통해 다중 값 논리 방법을 이용하는 방법이 활용되었습니다. 그리고 기계학습 즉 머신러닝의 많은 이론들이 개발되었으나, 컴퓨터 성능 및 제한적인 활용 등으로 인하여 크게 주목받지는 못하였습니다.

그러다가 2000년대에 드디어 힌튼 교수에 의해 심층신경망 즉 딥러닝 기술이 실용화 가능성이 제기되었고, 2010년대에는 합성곱신경망의 일종인 Deep-CNN(Convolution Neural Network)의 이미지 인식 성능이 대폭 개선되고, 2014년 구글이 딥마인드 테크놀로지를 인수하였고, 이의 결과로 개발된 인공지능 기술이 알파고였던 것입니다.

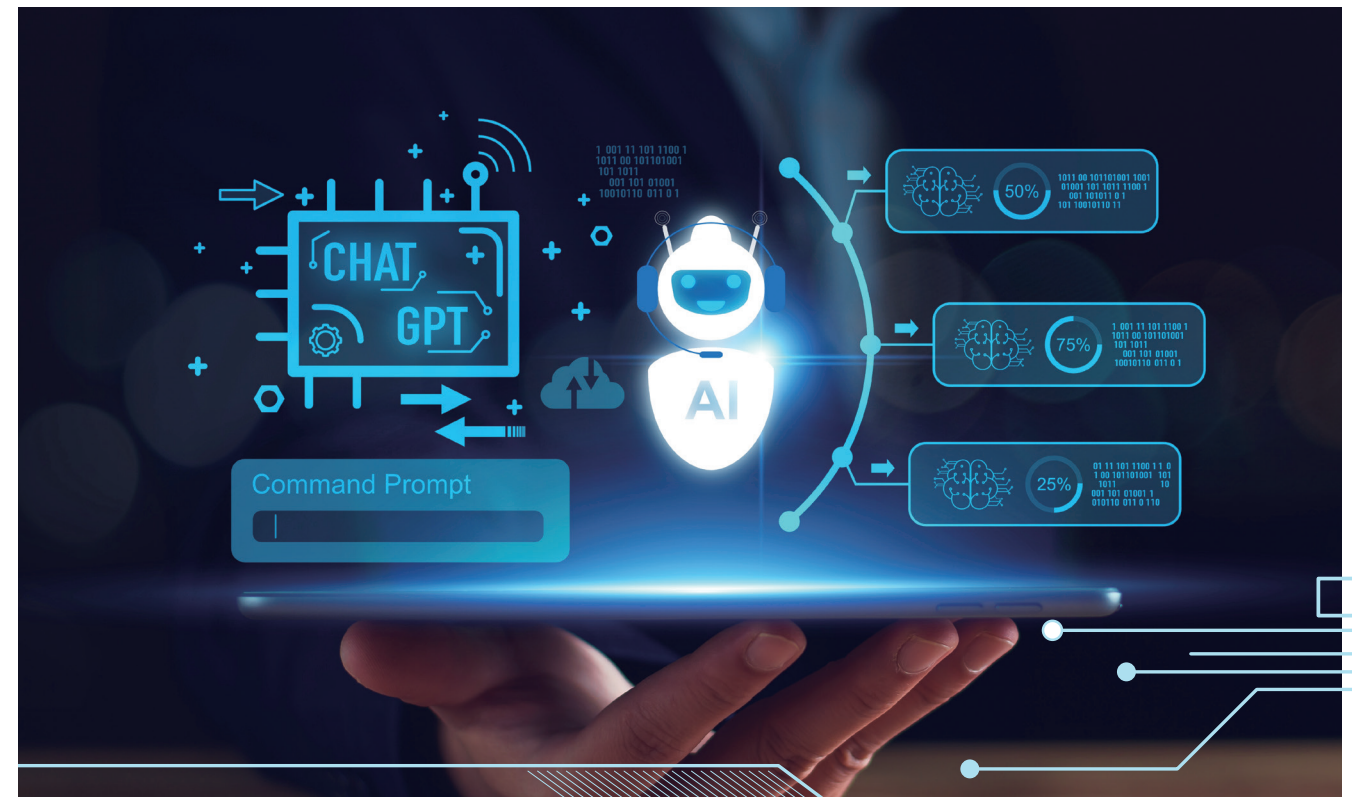
앞서 서술하였듯이 2016년 이세돌과 인공지능의 바둑 대결은 전 세계의 이목을 끌기에 충분했고, 인공지능 연구는 인공지능 이론의 개선과 컴퓨터 성능의 발달로 인해 폭발적으로 증가하게 됩니다.

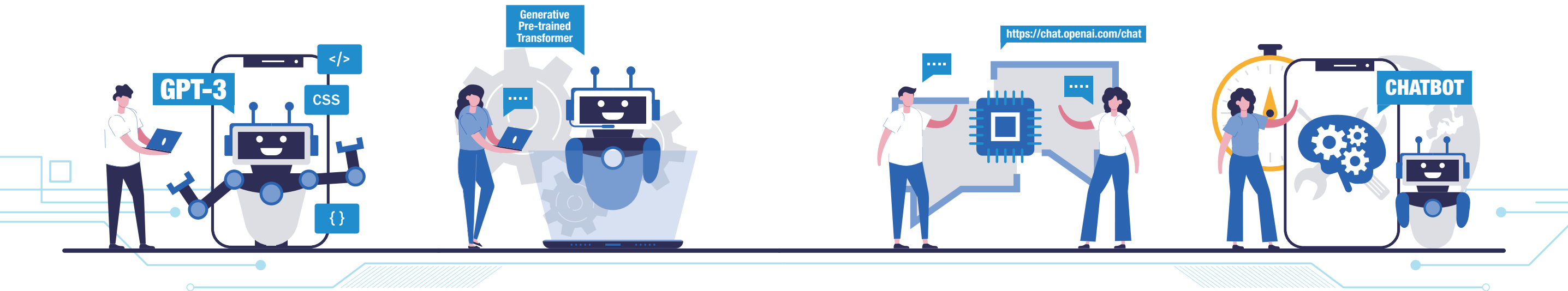
이제 우리는 인공지능 기술과 떼려야 뗄 수 없는 관계가 되어 버렸습니다.

## 챗GPT가 우리에게 던져준 질문들

최근 전 세계를 강타한 AI 열풍의 주인공은 당연 챗GPT라 할 수 있습니다. 챗GPT는 샘 알트만이 인류에게 이익을 주는 것을 목표로 하는 비영리단체 OpenAI에서 만든 인공지능 대화형 챗봇입니다. 2022년 12월에 출시해서 5일 만에 100만명의 사용자를 모았다고 이슈가 되었는데, 현재는 2달만에 1억명의 사용자를 확보하여 인터넷 등장 이후 가장 빠른 속도로 사용자를 끌어모으고 있습니다.

챗GPT는 GPT-3라는 거대 언어 모델을 기반으로 만들어졌습니다. GPT는 Generative Pre-trained Transformer의 약자이며, Transformer는 2017





년 구글에서 개발된 문장 속 단어와 순차 데이터 내의 관계를 추적해 맥락과 의미를 학습하는 신경망 모델입니다. 따라서 GPT-3는 1,750억 개의 매개변수를 이용하여 Transformer 모델로 미리 학습된 거대 언어 모델이 됩니다.

챗GPT는 마이크로소프트의 지원하에 GPT-4까지 적용하여 개발된 상태입니다. GPT-3에서도 챗GPT로 활용할 수 있는 일이 매우 많았는데, GPT-4까지 적용되어 대중화 된다면 그 활용범위는 상상할 수 없을 정도입니다.

OpenAI는 챗GPT의 API를 2023년 3월 1일 공개

하였습니다. 챗GPT를 사용하려면 <https://chat.openai.com/chat> 사이트에 접속하여 사용할 수 있는데, API를 이용한다면 다른 많은 프로그램에서 챗GPT의 기능을 사용할 수 있게 됩니다.

웹 브라우저의 확장 프로그램, 카카오톡의 오픈채팅 등 다양한 곳에서 챗GPT의 기능을 활용할 수 있고, 이것은 마치 2000년도에 유행했던 유비쿼터스 마냥 컴퓨터의 응답을 요구하는 어떤 분야에서든 사용될 수 있음을 의미합니다.

창업 이후 검색의 최강자 자리를 한 번도 내놓지 않은 구글에서 챗GPT가 출시되자 코드 레드(Code Red)를 발령하며 긴장하고 있습니다.

검색자가 원하는 검색 결과를 가장 잘 찾아주는 기능을 제공하기에 구글이 검색의 제왕 자리를 놓지 않는 것인데, 인간이 다시 정리해야 할 검색 결과를 챗봇이 스스로 정리하여 제공한다면 검색을 위한 발걸음이 구글에서 챗GPT로 돌아설 것이 뻔하기 때문입니다.

검색에 대한 결과뿐만이 아닙니

다. 챗GPT는 학습된 지식을 토대로 질문에 대한 새로운 문장을 서술해줍니다. 따라서 이러한 기능을 이용한다면 창작 업무에도 활용할 수 있습니다. 이미 전 세계에서 챗GPT를 이용한 레포트, 소설, 논문 등이 쏟아져 나오기 시작했습니다. 문제는 챗GPT의 응답 결과를 얼마나 신뢰할 수 있는냐는 것입니다. 많은 실험자들이 다양한 질문을 통해 챗GPT의 신뢰성을 검증하고 있고, 실제 현실과는 다른 대담을 찾아내고 있습니다.

OpenAI의 창업자 샘 알트만은 다음과 같이 말하였습니다.

“GPT-3은 너무 과대평가 되었습니다. 여러 칭찬은 감사하지만, 여전히 약점이 있고 이상한 실수를 하기도 합니다. AI가 세상을 바꿀 것이지만 GPT-3가 그 첫발을 내딛은 것뿐이라 생각합니다. 여전히 알아낼 게 많아요.”

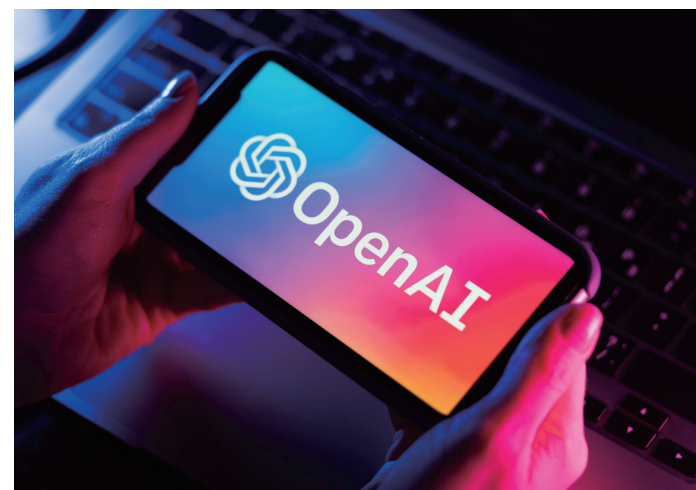
샘 알트만의 말처럼, 챗GPT의 결과를 100% 신뢰할 수는 없습니다. 하지만 사실로 판명된 정보를 토대로 조합된 결과는 충분히 기존의 검색엔진을 대체하기에 충분할 것입니다.

인공지능 기술의 발전은 굴곡이 있었지만, 현재의 발전 속도는 너무나 빠르다고 할 수 있습니다. 컴퓨

터공학과를 졸업하고 20여년을 IT세계에서 일을 하고 있는 필자도 2022년 12월까지만 해도 챗GPT를 보면서 재밌다고 생각했습니다. 하지만 본 기사를 작성하기 위해 자료를 찾는 동안 받았던 생각은 무섭다는 것이었습니다. 불과 몇 달 만에 챗GPT뿐만 아니라 생성 AI에 관한 발전과 대중화 속도는 쫓아가기도 버겁다는 생각이 들었기 때문입니다.

불과 1년 전 만 해도 미래사회에서 인공지능이 발달한 세상에서 없어질 직업, 살아남을 직업을 언급할 때, 인간만이 할 수 있는 창작이나 예술의 영역은 살아남는다고 했습니다. 하지만 이제는 이러한 생각도 변경되어야 합니다. 이미 인공지능이 그린 그림이 미술대전에서 대상을 받고, 소설을 작성하고, 영상을 제작하고 있습니다.

인공지능의 기술 발전은 인간이 두려워할 만큼 빠른 것이 분명합니다. 하지만 자동차의 발전을 막아주려 마차를 막아주고, 스마트폰의 발전을 2G폰을 위해서 막아서는 실수를 되풀이해서는 안 될 것입니다. 이제 인공지능 기술은 선택이 아니라 필수가 되어가고 있습니다. 우리는 기술의 발전을 앞장서지는 못하더라도, 편승하여 따라가려는 노력은 해야 할 것입니다.

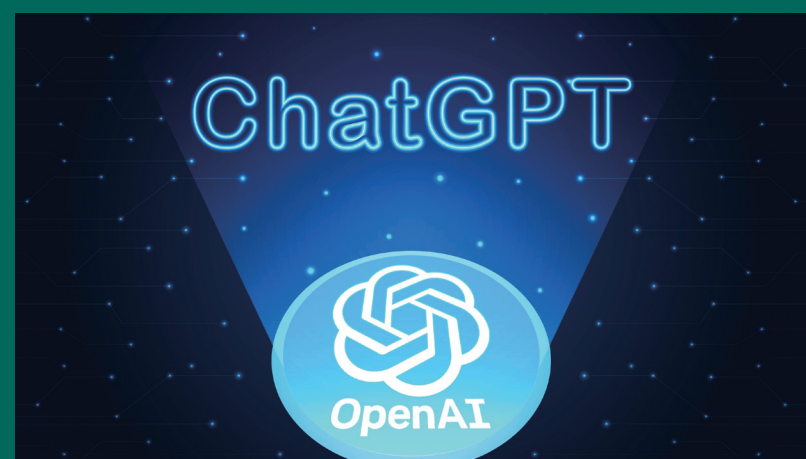




# 02 챗GPT가 가져다준 빛과 어둠

김명주 | 서울여자대학교 정보보호학과 교수

2023년 상반기는 “챗GPT”의 시대이었다고 해도 과언이 아니다. 2022년 11월 30일 공식 오픈한 지 5일 만에 이용자 100만 명을 넘어섰고 2달 만에 1억 명을 넘어섰다. 이런 열풍은 역사상 처음이었다. 유료 서비스인 “챗GPT 플러스”도 생각보다 많은 이용자를 전 세계적으로 확보해갔다. 식지 않은 열풍 속에서 주마가편 격으로 3월 14일 전격 단행한 4.0으로의 버전 업그레이드는 유료 서비스의 차별화와 가속화를 재촉하였다. 국내에서만 해도 소개된 지 불과 몇 달 만에 챗GPT에 관한 수백 권의 책들이 서점에 넘쳐났다. 이 중에는 챗GPT를 저자 혹은 공저자로 하는 책들을 발견하는 것도 더 이상 어렵지 않게 되었다.



## 챗GPT가 가져온 변화 ... 구글 독점 구도에 큰 위협으로 부상

챗GPT의 등장은 그동안 글로벌 검색엔진 시장의 92.9%를 차지해온 구글에게는 큰 위협이 되었다. 지금까지는 검색엔진을 통해서 어떤 주제에 관하여 먼저 검색한 후 이를 통해서 얻은 수많은 검색 결과물들을 상당한 시간 읽으면서 그 내용을 소화해야 했다. 그리고 나서야 자신의 의견을 종합하여 제시

할 수 있었던 기존의 긴 과정을 챗GPT는 엄청나게 단축해 주었다. 고작 3%의 비율로 검색엔진 시장 2위를 차지해왔던 마이크로소프트사는 자신이 이미 투자해온 챗GPT에게 100억 달러(12조 5천억 원)를 추가로 투자했다. 검색엔진 1%가 일 년에 광고 수익 10억 달러(2조 5천억 원)를 좌우하는 상황을 감안할 때, 100억 달러라는 투자는 큰 무리가 아니다. 챗GPT를 담아낸 마이크로소프트사의 검색엔진 Bing은 덕분에 큰 상승세를 타고 있다.





## 챗GPT가 가져온 변화 ... 인공지능 리터러시

챗GPT는 예상보다 뛰어난 글 생성 능력 때문에 등장 초기부터 여러 잡음을 불러일으켰다. 고등학교와 대학교 학생들이 글쓰기 성격의 과제를 수행할 때 챗GPT를 활용하는 바람에 교육의 목적이 달성되지 않을 뿐 아니라 공정한 과제 평가도 이루어질 수 없게 된 것이다. 그래서 미국의 일부 주에서는 학교에서의 챗GPT 사용을 금지했다.

챗GPT를 제작한 OpenAI 기업은 이러한 문제해결에 도움이 될 수 있는 도구로서 “인공지능 문서 분류기(AI Text Classifier)”를 공개하기도 했다. 주어진 글이 인공지능이 생성한 것인지, 아니면 사람이 쓴 글인지를 이 도구는 판단해준다. 비록 이 도구의 정확도가 현재로서는 26% 정도에 불과하지만

일선 학교에 이를 활용할 경우, 충분한 견제 도구가 될 수 있다. 다만 이런 도구를 활용하여 학생들을 강제하기보다는 챗GPT 시대를 맞이하여 학생들이 새롭게 갖추어야 할 학습 윤리에 대한 공감대 형성이 더 바람직해 보인다. 이를 “인공지능 리터러시”(AI Literacy)라는 표현 안에 담는 중이다.

## 챗GPT가 가져온 변화 ... 개인정보 유출에 대한 우려

챗GPT가 4.0으로 전격적인 버전 업그레이드를 선언하고 불과 1주일이 지난 3월 20일, 유료 사용자의 신용카드 결제 정보 일부가 외부에 유출되었다는 나쁜 소식이 들려왔다. 그리고 이와 더불어 챗GPT와의 대화 목록이 사용자들 간에 섞여서 나타

나는 오류가 발생했다. 불행 중 다행으로 타인의 대화 목록만 보이지 대화 내용까지는 보이지 않았다. 만일 타인의 대화 내용까지 노출되었다면 개인정보와 사생활 정보 유출, 명예훼손, 기업비밀 유출 등으로 인하여 소송 폭탄을 맞아 챗GPT는 매우 심각한 위기에 봉착했을 것이다.

OpenAI의 대표 샘 알트만은 그날 자신의 트위터에서 “매우 끔찍한 상황”이었다고 고백했을 정도이다. 그런데 샘 알트만은 이날 9시간가량 지속된 오류가 챗GPT의 대화 목록을 유지하는 부분을 자신들이 잘못 코딩해서 발생한 것이 아니라, 외부에서 제공하는 오픈 소스 라이브러리의 오류 때문이었다고 밝혀서 다른 차원의 충격을 주었다.

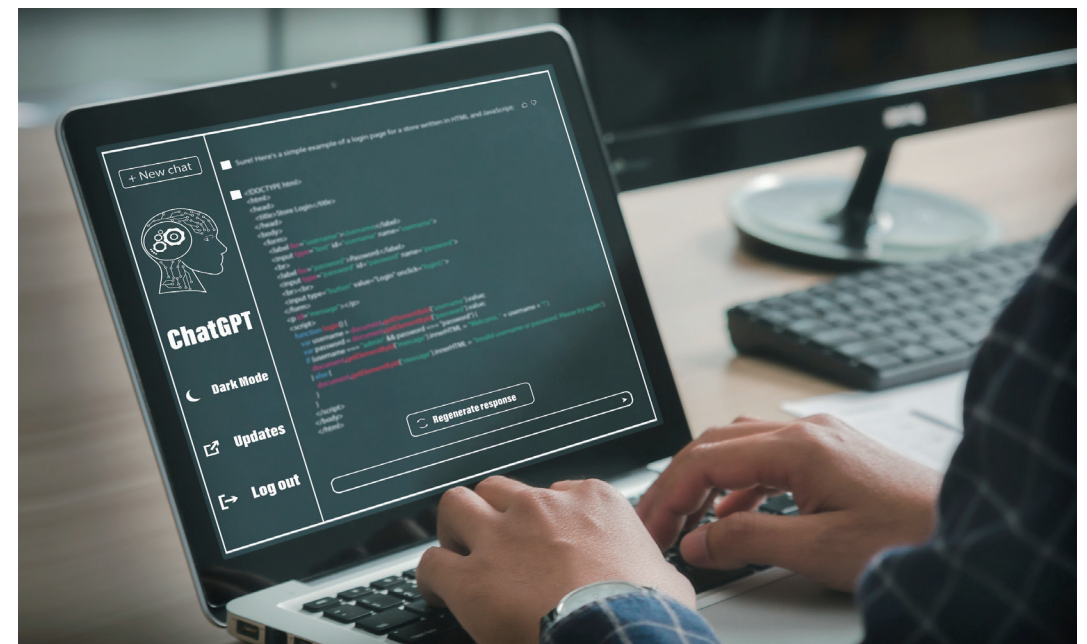
이는 챗GPT의 대화 목록 제공 부분이 오픈 소스로 구현되었다는 뜻이며 이 오픈 소스를 본인들이 통제하지 못했다는 이야기이다. 이것은 오픈 소스의 취약점을 집중적으로 활용하여 공격하는 해커들에게 앞으로도 큰 관심을 끌어낼 만한 ‘어리석은’ 멘트였다.

한번 발생한 보안사고는 해당 근원지를 바꾸지 않는 이상 다시 발생할 가능성이 크다. 그래서 앞으로

우리는 챗GPT를 사용할 때 대화 내용 중에 개인 정보와 은밀한 사생활, 회사의 주요 기밀을 언급하지 않는 것이 필요하다. 혹시나 불가피하게 이를 언급해서 챗GPT의 답변을 들었다면, 이 대화 내용을 별도로 보관한 후 대화 목록에서는 삭제하는 것이 좋다. 이미 발생한 잠재적 보안취약점이 챗GPT의 사용환경에 여전히 존재할 수 있으며, 대화 목록에 실린 내용 모두는 물리적으로 챗GPT 서버로 일단 유출되어 저장된다는 점도 기억해야 한다.

## 챗GPT가 가져온 변화 ... 오토 프로파일링의 위험

이탈리아가 챗GPT에 대한 국가적 사용 중단을 결정한 것은 그로부터 10일 뒤인 3월 30일이다. 이러한 결정의 이유는 GDPR이라 불리는 유럽연합의 개인정보 보호법을 위반했다는 의구심 때문이다. 이러한 의구심은 챗GPT가 특정인에 대해 놀라운 정보 학습력을 기반으로 프로파일 작성 능력을 보여줄 때부터 시작되었다. 유럽연합의 개인정보 보호





법은 본인의 사전 승낙받지 않을 경우, 그 개인에 대한 프로파일을 자동으로 생성하는 “오토 프로파일링” 행위를 금지하고 있다. 개인이 자신의 다양한 개인정보를 인터넷상에 이미 공개하여 여기저기 올렸다고 하더라도 각각은 고유의 공개 목적이 있을 터인데, 이러한 개인정보를 모두 가져와서 개인의 프로파일을 완성해내는 행위는 당초의 공개 목적에 부합하는 행동이 아니라는 것이다.

따라서 이러한 오토 프로파일링을 시도하려면 해당 개인에게 사전에 승인받아야 가능하다. 그런데 챗GPT는 특정한 개인 특히 사회적으로 유명하여 이미 많은 정보가 인터넷에 공개된 인물에 대하여 자연스럽게 프로파일 정보를 자동으로 생성해서 보여준다. 이것은 검색엔진으로 특정 인물을 검색하는 것과는 완전히 다르다. 검색엔진은 특정 인물과 관련된 사이트 정보들을 보여줄 뿐이며 프로파일을 만들어주지는 않는다. 그러나 챗GPT는 자동으로 개인의 프로파일을 만들어준다. 이는 유럽연합의 개인정보 보호법 GDPR 위반에 해당한다. 그리고 우리나라 개인정보 보호법을 역시 위반하는 기능이다.

따라서 OpenAI는 이러한 실정법 위반에 대하여 앞으로 적절한 조치를 해야만 하는데, 가장 손쉬운 방법은 특정 개인에 대한 오토 프로파일링 성격의 질의응답 서비스를 제한하는 것이다. 챗GPT 때문에 오토 프로파일링을 불법에서 합법으로 바꾸는 개인정보보호법 개정을 모든 국가가 함께 진행할 수는 없기 때문이다.

## 챗GPT가 가져온 변화 ... 할루시네이션

현재의 챗GPT는 2021년 9월까지 획득한 데이터를 대상으로 학습을 진행하였다. 그러다 보니 그 이후에 생성된 데이터와 사실에 대해서는 학습하지 못했기 때문에 이에 관한 질문을 받으면 사실에 기반한 답을 생성해낼 수 없다. 그렇다고 챗GPT가 “죄송합니다. 저는 모릅니다”라고 답하지 않는다. 외형상 거침없이 어떤 답변이라도 생성해낸다. 심지어 이미 학습한 데이터에 대해서도 사실이 아닌 내용을 사실과 섞어가면서 글을 생성하기도 한다. 이러한 챗GPT의 답변을 가만히 듣고 있으면 이용

자도 순간적으로 챗GPT의 대답이 사실인 것처럼 속아 넘어갈 수 있다. 이를 “환각” 영어로는 “할루시네이션(Hallucination)”이라고 부른다. 4.0으로 버전이 업그레이드되었지만 이러한 할루시네이션 현상은 여전히 나타난다. 지금은 메타라 불리는 페이스북에서 인공지능 연구를 총괄해온 안 르쿤 박사는 챗GPT의 이러한 할루시네이션 현상은 태생적으로 불가피한 것이라고 주장했다.

## 정보를 검증할 수 있는 비판적 사고가 필요

심지어 OpenAI의 대표 샘 알트만 역시 같은 이야기를 했다. “인간이 불완전하기 때문에 챗GPT와 같은 인공지능도 완벽할 수 없다”. 사실 챗GPT의 사업책임자로서 이런 식의 말을 하

는 것 자체가 부적절하다. 그렇지만 이 발언은 챗GPT의 피할 수 없는 약점을 시인한 셈이다. 따라서 챗GPT의 사용에 따른 책임은 사업자와 개발자로부터 이제는 이용자로 온전히 넘어온 셈이다. 앞으로 챗GPT 이용자들은 아주 유능한 챗GPT를 사용할 때마다 챗GPT에 대한 경계심과 비판적 사고도 놓치지 말아야 한다. 챗GPT가 생성한 답변이 정말 사실에 근거하고 있는지 바로 판단할 수 있을 실력을 평소에 갖추든지 아니면 추가로 검색엔진을 통해서 확인 검증하는 작업을 꼭 해야 한다는 뜻이다. 챗GPT는 활용 효과가 매우 큰 기술인지라 우려와 두려움 역시 매우 큰 것이 사실이다.





## 03

# 챗GPT-4 인공지능 시대, 필요한 미래 인재 교육은 무엇일까

안중배 | 국제미래학회장



챗GPT-4가 전 세계적으로 돌풍을 일으키고 있다. 아무리 어려운 질문도 챗GPT-4는 매우 정돈된 답을 내놓는다. 유용성의 한계를 가늠할 수 없을 정도라는 극찬도 나온다. 반면에 챗GPT-4 출현을 계기로 인공지능 시대가 본격화되면서 인간의 일자리를 위협하는 시점이 훨씬 앞당겨질 것이라는 우려도 증폭되고 있다.

이에 이전과는 다른 인재 역량이 필요하다는 현장의 요구가 많아지고 있고 이와 더불어 챗GPT-4 인공지능을 건강하게 선용할 수 있는 윤리 교육의 중요성도 높아지고 있다.

## 인공지능 시대를 대비하기 위한 첫걸음...교육 환경 변화

챗GPT-4가 빠르게 열어가는 인공지능 시대는 초지능·초연결·초실감이 구현되고 디지털세상과 현실세상이 교류하고 융합될 것이다. 인공지능을 통해 초지능, 인공지능 기반의 사물인터넷으로 초연결, 그리고 인공지능 기반의 메타버스로 휴머니즘을 담은 초실감이 구현되고 있다.

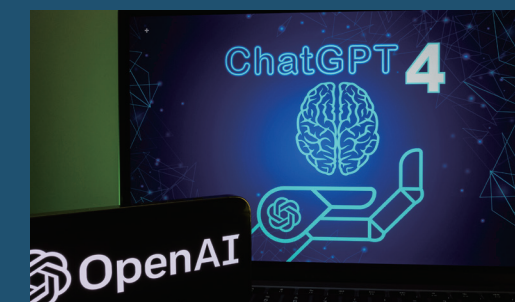
이러한 급격한 변화와 더불어 따뜻한 인성과 공동

체 가치를 추구하는 휴머니즘이 강화될 것으로 예상된다. 이러한 챗GPT-4 인공지능의 성장과 활용은 인공지능 연구 및 개발에 대한 투자 증가와 교육 및 연구 환경 개선에도 기여하고 있다.

그러나 GPT-4와 같은 인공지능 시스템은 편향성, 개인정보 침해, 허위 정보 유포 등의 문제를 수반할 수 있다. 따라서 인공지능 기술의 성장과 활용에 대한 적절한 이해와 균형있는 교육이 필요하다. 교육은 이러한 인공지능 시대 미래 사회 변화에 대응하여 미래 인재를 양성하면서 인공지능 윤리 의식을 함양하는 것이 중요해지고 있다.

## 인공지능 시대 요구되어지는 4대 핵심 역량

인공지능 시대가 여는 미래에는 과학기술, 산업, 사회, 문화, 가치관이 지금과 다르게 많이 변화할 것으로 예상된다. 그동안 대한민국의 빠른 산업화와 국가 발전에 크게 기여해온 한국의 교육은 인공지능 시대가 현실화되기 시작하면서 기존의 교육으로 는 더 이상 학생들의 미래도 대한민국의 미래도 담보하기 어렵게 되었고, 이에 기존 교육과는 전혀 다







른 새로운 교육 패러다임이 필요하게 되었다. 인공지능 시대 초지능·초연결·초실감 사회의 패러다임에 맞는 교육의 혁명적인 변화가 요청되고 있는 것이다.

이에 따라 미래 교육의 비전은 인공지능 시대를 주도할 미래창의혁신 인재 양성이 되어야 한다. 이를 기반으로 목표는 글로벌 경쟁력을 갖춘 미래창의혁신 인재를 양성하는 교육, 개인의 창의성과 다양성이 존중되고 행복한 삶과 건강한 사회의 지속 발전에 기여하는 교육을 실현하는 것이어야 한다.

인공지능 시대에 필요한 핵심 역량 인재상은 4대 핵심 기반역량을 바탕으로 이루어진다고 볼 수 있다.

❶ **첫 번째** 주목해야 할 기반역량은 창의로운 인지 역량 영역으로 창의성과 문제해결 사고력, 미래 도전정신, 인문학적 소양 등을 말한다.

❷ **두 번째** 기반역량 영역은 인성을 갖춘 정서역량이다. 이 부분에 해당되는 역량으로는 인성·윤리의식, 문화예술 소양, 자아 긍정 관리, 미래리더쉽 등의 역량이 해당된다.

❸ **세 번째** 기반역량으로는 협력하는 사회 역량 부분으로 소통과 협업 역량, 사회적 자본 이해, 글로벌 시민 의식, 스포츠, 체력과 관련된 역량이다.

❹ **네 번째** 기반역량으로는 생애주기 학습역량을 들 수 있으며 여기에는 미래 변화를 예측하고 생애 계획을 세워 평생 스스로 학습하는 자기주도 학습 역량, 과학기술 변화이해, 인공지능과 메타버스 포함 ICT 활용 역량, 평생학습 등의 역량이 해당된다.

이러한 미래 인재에게 기본이 되는 4대 핵심 기반 역량을 바탕으로 하여 인공지능 시대에 필요한 영역별 융합적 전문역량을 함양하여 건강한 미래사회를 주도할 수 있는 창의적으로 사고하는 인성을 갖춘 미래창의 혁신 전문 인재를 양성하여야 할 것이다.

종합적으로, 인공지능 시대에서 미래를 리더할 인재는 AI 활용 전문성, 혁신적 사고력, 인성 리더십, 창의성, 그리고 윤리적 책임감을 모두 갖추어야 한다. 이러한 역량들은 인공지능 분야에서 일하는 사람들뿐만 아니라, 인공지능과 함께 일하는 모든 분야에서 중요한 역량이다. 이러한 역량들을 갖춘 인재들은 인공지능 시대에서 새로운 기술과 새로운 응용 분야를 개척하며, 건강한 사회 발전에 기여할 수 있는 리더로 성장할 것이다.

## 인공지능 발전에 따른 미래 인재 교육 방안

교육부는 2025학년도부터 초등학교 3·4학년과 중 1·고1 학생들은 수학·영어·정보 교과를 공부할 때 인공지능(AI) 디지털 교과서를 도입하여 사용하기로 하였다. 기존 교과서를 대체할 ‘AI 디지털 교과서’는 과목별 특성과 학습 수준에 맞는 맞춤형 학습이 가능하도록 대화형 AI, 음성인식, 확장현실 등의 AI 기술을 기반으로 개발된다.

즉 모든 학생이 인공지능 메타버스 공간에서 AI 튜터의 도움을 받으면서 동료와 게임을 하듯이 학습을 즐길 수 있게 되는 것이다. 모든 교사는 학습자 한 명 한 명에게 맞추어 학습환경을 제공하는 맞춤형 학습환경 디자이너가 된다. 모두가 방대한 지식의 핵심 개념을 이해하는 탄탄한 지식기반을 갖추고, 그 위에 데이터, 첨단기술, 인문학 등의 소양을 쌓아 올리고, 또한 창의력, 비판적 사고력, 협력, 소통 등의 미래역량을 꽃피우는 교육을 받는다.

이러한 교육의 변화는 교육 관련한 모든 체계가 총체적으로 상호 협력하면서 교육 전반에서 동시에 혁명적인 변혁이 요구되고 있다.

특히 교사가 학생들의 수업을 진행하는 교수 방법이 기존의 효과적 지식 전달을 위한 일방향 강의실에서 벗어나 미래 인재 역량을 함양할 수 있도록 혁신되어야 한다. 챗GPT-4 인공지능 시대 인재는 스스로 무엇을 모르는지와 무엇을 알아야 하는지 그리고 이를 위해 무엇을 해야 하는지를 스스로 인지하고 질문하며 실천하는 메타인지 역량을 갖추어야

한다. 이를 함양하기 위해 메타교수학습법으로 교수 방법이 혁신되어야 한다.

## 메타교수학습법으로 S.M.A.R.T 메타교수법인

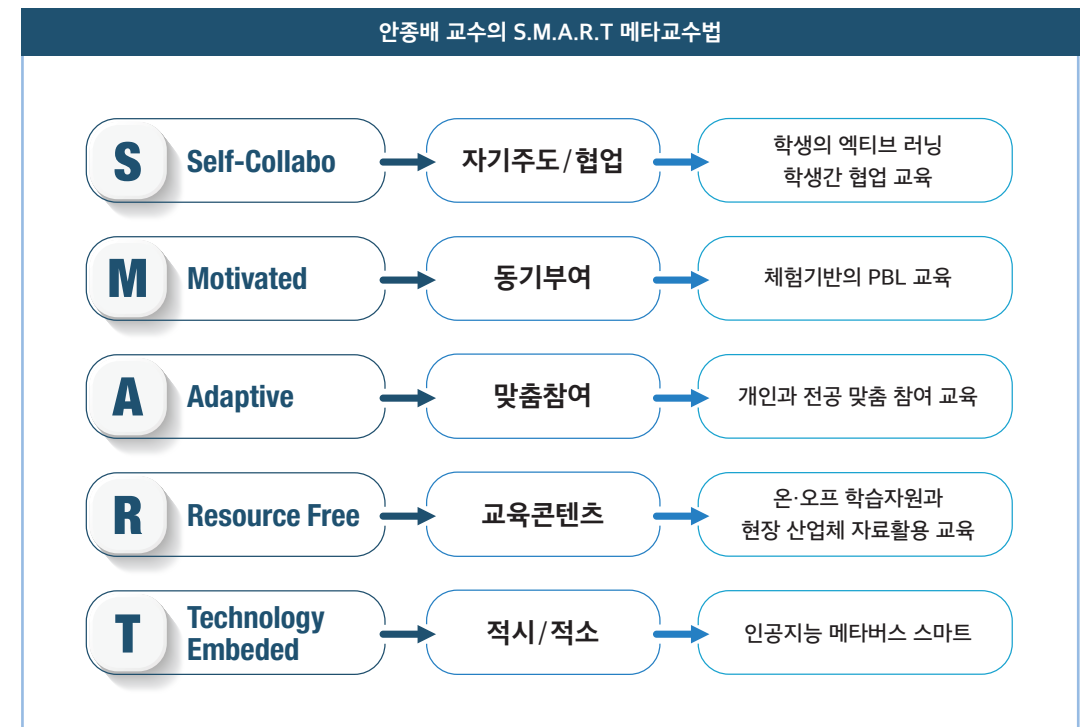
▶ **S(Self-collabo)** 학생 스스로 자기주도 학습과 협업할 수 있게 하는 교수법

▶ **M(Motivated)** 문제·프로젝트 중심으로 운영되는 동기부여식 학생 중심 교수법

▶ **A(Adaptive)** 학생 중심으로 진행되는 개인 맞춤형 참여 교수법

▶ **R(Resource free)** 다양한 교육자원을 이용해 문제 해결 역량을 갖추 수 있도록 교육 콘텐츠들을 활용하는 교수법

▶ **T(Technology Embeded)** 인공지능과 메타버스 등 스마트 ICT를 활용한 양방향 체험 교수법을 통해 미래 인재 역량인 창의력, 협업 능력, 커뮤니케이션 역량, 인성 리더쉽, 인공지능과 메타버스 활용 역량 등을 함양하게 된다.





## 인공지능을 활용하는 교육 방안

인공지능은 사용하는 역량을 함양하면서 학생들의 미래 역량을 함양하는 교육을 위해 다양하게 활용될 수 있다.

- ❶ **첫째**, 인공지능을 활용하여 학생 개개인에게 맞는 교육 콘텐츠를 제공하고 개인별 문제 해결 능력에 적절한 개인 맞춤형 교육을 실시할 수 있다.
- ❷ **둘째**, 학생들이 어려운 개념이나 수업 내용에서 이해하지 못한 부분에 대해 질문할 때, 친절하고 지속하여 답변해 주는 인공지능 튜터 교육을 제공할 수 있다.
- ❸ **셋째**, 인공지능을 활용하여 학생들이 자기주도적으로 학습을 진행할 수 있는 지원을 제공하여 자기주도 학습 교육을 원활하게 진행할 수 있다.
- ❹ **넷째**, 인공지능은 주제에 대한 배경 정보와 다양한 관점을 제공하여 학생들이 다양한 주제에 대해 함께 토론하면서 비판적 사고 능력을 향상시키는 데 도움이 된다.

❺ **다섯째**, 인공지능은 창의적 글쓰기를 포함하여 학생들이 창의력을 발휘하고 새로운 아이디어를 탐구할 수 있도록 도움을 준다. 교사는 학생들에게 AI가 제안하는 다양한 관점과 아이디어를 소개함으로써 사고의 폭을 넓히고, 창의적인 발상을 유도할 수 있다.

❻ **여섯째**, 인공지능을 활용하여 학생들의 감정적 인 문제나 고민에 대한 상담을 제공할 수 있다. 이를 통해, 학교에서는 학생들의 상담 내용을 분석하여, 개별 학생들의 필요에 맞는 지원을 제공할 수 있다.

❼ **일곱째**, 인공지능을 사용하여 수업에 참여하는 학생들의 동기를 높일 수 있다. AI와의 대화를 통해 학생들은 새로운 지식을 얻고, 팀별 과제에 도움도 받으면서 수업과 팀별 활동에 더 적극적으로 참여할 수 있게 된다.

❽ **여덟째**, 인공지능을 사용하여 학생들의 언어 구사력을 향상시킬 수 있으며, 외국어를 가르치거나 학습하는데 도움이 된다. 학생들은 질문하는 능력을 향상 시키고 외국어로 대화를 나누며 문법, 발음, 어휘력을 향상시킬 수 있다.

## 인공지능 시대 윤리교육의 중요성과 방안

인공지능의 발전이 인류의 행복을 증진시키고 유익하게 될 것인지, 아니면 재앙을 몰고 올지는 현재 인류의 의지에 달려 있다. 인공지능의 발전으로 초지능·초연결·초실감 사회가 급속히 구현되고 인공지능이 모든 인류지능의 합을 넘어서는 싱귤래리티(Singularity)시대가 다가옴에 따라 인공지능 윤리가 더욱 중요해지고 있다.

따라서 인공지능 사용 가이드라인과 윤리, 법제에 대한 논의와 개발자와 사용자가 공히 인공지능을 인류의 미래에 유익하도록 개발하고 선용할 수 있게 인공지능 윤리 의식을 갖추게 하는 인공지능 윤리 교육이 필요하다.

인공지능 윤리 교육은 초중고 및 대학교의 교육 과정에 인공지능 윤리에 관한 내용을 통합하여 학습자들이 윤리적 가치와 활용 능력을 동시에 함양할 수 있도록 해야 한다. 기업과 정부도 인공지능 개발자, 사용자, 관계자들을 대상으로 윤리 워크숍 및 세미나

등을 개최하여 인공지능 윤리에 대한 지식과 경험을 공유하고 인성과 윤리 캠페인을 전개해야 한다.

인공지능 윤리 교육을 통해 개발자와 사용자 공히 인공지능에 대한 윤리적 책임감을 갖도록 해야 한다. 인공지능 개발자와 사용자 공히 AI 윤리 가이드라인을 이해하고 존중하여 편향, 개인정보 침해, 허위 정보 유포 등의 문제를 최소화하며 올바른 활용을 도모토록 해야 한다. 인공지능 윤리교육을 통해 개발자들은 신뢰성 있는 인공지능 시스템을 구축하여 신뢰를 얻고, 인공지능 기술의 건강한 성장을 이끌어낼 수 있도록 해야 한다.

이렇게 인공지능 시대에 교육은 산업사회의 특성인 표준화, 규격화, 정형화된 교육 방향을 탈피하여 다양성, 창의성, 유연성을 강화하고 인성과 윤리 의식을 함양하는 방향으로 교육이 혁신되어야 한다. 이러한 교육으로 인공지능 시대에 필요한 미래 인재 역량을 함양하고 윤리적 책임감을 강화해야 한다. 이를 통해 교육이 인공지능 시대가 새로운 기회가 되게 하고 미래 사회를 건강하게 만들어 가는 역할을 담당하게 된다.



# 안전한 사회 구축을 위한 합리적 기준 한국 범죄 분류

김태균 | 통계청 통계기준과 주무관



## 통계정보제공을 위한 합리적 표준 ... 분류체계

현대사회의 가장 두드러진 특징은 산업 분야의 표준화(Standardization)라고 볼 수 있다. 표준화는 일상적이고, 반복적인 문제를 주어진 여건에서 최선의 상태로 해결하기 위한 일련의 활동으로 정의되는데 이를 위한 합리적 기준이 바로 표준(standards)<sup>1)</sup>을 의미한다. 표준은 사전적으로 사물의 정도나 성격 따위를 알

기 위한 근거나 기준으로 정의되지만, 사회가 복잡해지고, 사회 안팎의 교류 범위가 이전과 비교할 수 없이 확대되면서 ‘국가사회의 모든 분야에서 정확성, 합리성 및 국제성을 높이기 위하여 국가적으로 공인된 과학적·기술적 공공 기준’<sup>2)</sup>으로 정의가 확대되었다.

표준 정의의 확대는 우리 생활과도 밀접한 통계의 표준화를 촉진하게 되었고, 국가중앙통계기관인 통계청은 통계법 제22조<sup>3)</sup>에 따라 신뢰도 높은 국가

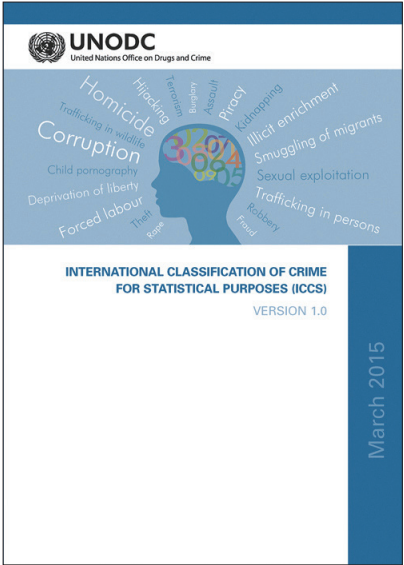


통계 생산과 경제·사회 등 여러 분야에 유용한 통계 정보 제공을 위한 합리적 표준으로 분야별 분류체계를 개발 운영하고 있다. 한국표준산업분류, 한국표준직업분류, 한국표준질병·사인 분류 등의 표준분류(7종)를 포함한 총 35종의 분류는 경제·사회·보건 분야에서 자료수집, 분석 및 활용의 기준으로 통계자료의 정확성, 합리성 및 국제성 확보에 기여하고 있으며 보다 확대될 것으로 예상된다.

이번에 소개할 한국범죄분류는 2015년 UN에서 채택한 국제범죄분류(International Classification of Crime for Statistical Purposes, 이하ICCS)를 기반으로 개발 중인 분류로 국가안전, 범죄예방 데이터의 효율적 수집, 자료처리 및 통계자료의 국내외 활용을 위한 공식 범죄통계 분류체계이다.

## 범죄분석과 예방을 위한 국제범죄분류 등장

한국범죄분류의 기반이 되는 ICCS는 유엔마약법



국제범죄분류(ICCS) 초판

1) ISO/IEC Guidebook 2th(2004)  
2) 국가표준기본법 제3조 제1항(법률 제15643호)  
3) 통계법 제22조 제1항(법률 제17339호), '통계청장은 통계작성기관이 동일한 기준에 따라 통계를 작성할 수 있도록 국제표준분류를 기준으로 산업, 직업, 질병·사인(死因) 등에 관한 표준분류를 작성·고시하여야 한다.'



죄사무소(UNODC, United Nations Office on Drugs and Crime)가 국제적으로 합의된 정의 및 원칙을 바탕으로 개발한 범죄분류체계를 의미한다. UNODC가 ICCS를 개발한 목적은 국제적으로 표준화된 분류체계를 통하여 범죄 행위를 명확하게 정의하고, 국가 간 범죄통계의 수집을 유용하게 함으로써 국제적 차원에서 범죄 현상을 정확히 진단·

해결을 위한 정책개발의 기초자료로 활용하는 것이었다.

이를 위해 UNODC와 UNECE(United Nations Economic Commission for Europe)는 2009년부터 ICCS 개발을 위해 법률 중심의 범죄분류에서 행위중심을 기초로 한 범죄분류체계로의 개편을 추

국제범죄분류 11개 대분류

대분류 코드	분류명
01	사망을 야기했거나 의도한 행위 (Acts leading to death or intending to cause death)
02	상해를 야기했거나 의도한 행위 (Acts leading to harm or intending to cause harm to the person)
03	성범죄 (Injurious acts of a sexual nature)
04	폭력 또는 협박등반 재산침해 행위 (Acts against property involving violence or threat against a person)
05	재산 침해 행위 (Acts against property only)
06	관리대상 약물이나 향정신성 물질 관련 위법행위 (Acts involving controlled psychoactive substances or other drugs)
07	사기, 기만, 부패 관련 행위 (Acts involving fraud, deception or corruption)
08	공공질서, 국가의 권위, 국가규범 위반 행위 (Acts against public order, authority and provisions of the State)
09	공공안전, 국가보안 위반 행위 (Acts against public safety and state security)
10	자연환경 위해 행위 (Acts against the natural environment)
11	기타 범죄 행위 (Other criminal acts not elsewhere classified)

※ 대분류(2자리), 중분류(4자리), 소분류(5자리), 세분류(6자리)로 코드가 구분되며, 한국범죄분류에서는 기수범죄 외에 범죄유형을 포괄하기 위해 세세분류(8자리)를 신설 적용함  
 ※ 국제범죄분류(ICC)S 원문은 유엔마약범죄사무소(UNODC) 누리집(www.unodc.org)에서 제공함



진하였고, 수년간의 연구 끝에 2012년 ICCS 초안을 발표하였다.

이후 초안에 대한 검토 작업 끝에 2015년 3월 제 46차 유엔통계위원회(United Nations Statistical Commission)에서 ICCS를 범죄통계 작성을 위한 표준분류로 채택되었다.

이러한 국제범죄분류는 범죄예방 및 형사사법 정책 개발을 위해 쉽게 이해되고, 활용할 수 있는 통계자료를 제공하는 것이 목적이기 때문에 국제범죄분류는 정책적 연관성에 우선순위를 두고 11개 대분류로 구성하였다. 대분류별로 범죄 관련 ‘정책 분야’, ‘대상’, ‘심각성’, ‘범행방식’을 기준으로 62개 중분류, 165개 소분류, 230개 세분류로 범죄를 분류하여 모든 범죄 행위나 사건을 포괄하도록 설계되었다.

또한 범죄 속성을 이해하는 데 필요한 추가적 정보를 대상별로 세분화함으로써 범죄특성에 대해 정책적으로 의미 있는 정보를 제공하도록 했다.

범죄통계의 일관성과 비교성을 높이기 위해 한국형 범죄분류체계의 개발

국제범죄분류 채택 및 이행 공고에 따라 통계청은 국내 범죄 통계의 일관성과 국제 비교성을 높이고, 국내 및 국제적 차원의 분류 해석능력을 향상하기 위해 총 3단계(1단계 개발연구, 2단계 일반분류제정, 3단계 표준분류화)의 한국범죄분류 개발 계획을 수립하였다. 2017년 「국제범죄분류의 도입 타당성 연구」를 시작으로 ‘대분류별 분류 체계 연구’(2018-19)를 통해 국제범죄분류와 국내범죄명의 연계원칙을 마련하였고, 2020년부터 균형법을 제외한 국내 범죄명 1만 5천여개를 분류체계에 연계하는 작업과 적정성 검증 연구가 진행 중이다.

앞에서 설명한 것처럼 ICCS는 행위 중심으로 통계를 작성하는 데 반해 우리나라의 경찰과 검찰이 사용하고 있는 범죄분류체계는 법률조항 중심으로 이루어져 있다. ICCS를 따르는 대부분의 영미법계 국가들에서는 범죄가 발생하면 경찰이 사건을 인지하

ICCS-죄명코드 수정연계원칙(2020)

번호	연계원칙
01	우리나라 범죄통계의 기본단위인 죄명코드를 ICCS와의 연계 대상으로 한다.
02	ICCS의 항목분류 코드에 우리나라의 죄명코드를 연계하는 방법을 사용한다.
03	우리나라 죄명코드는 최하위 단위의 개별 코드를 이용하고, ICCS 역시 원칙적으로 가장 하위 단위의 코드를 연계하는 것을 목표로 하되, 불가피한 경우 차상위 단위의 코드를 연계한다.
04	가능한 모든 죄명코드에 대하여 1대 1 연계를 원칙으로 하되, 예외적으로 2개 이상 또는 누락되는 경우 그에 대하여 후술하는 설명서에 작성한다.
05	예비(예비교사 및 예비방조), 음모(음모교사 및 음모방조), 미수(미수교사 및 미수방조), 교사, 방조는 각각 해당 범죄 기수항목 및 정범의 범죄에 연계되는 ICCS의 분류단위에 연계한다.
06	결합범은 법정형이 중한 범죄로, 법정형이 동일한 경우 범죄행위로 침해당하는 법익이 중한 정도에 따라 연계한다.
07	결과적 가중범은 치사형태를 제외하고 고의의 기본범죄와 연계한다.

고 수사를 개시하는 과정에서 주된 행위를 중심으로 통계 작성표(원표)를 작성하고 있지만, 우리나라의 형사사법기관이 사용하는 범죄분류체계는 경찰이 작성하는 통계 작성표(원표)를 기준으로 하고 있다.

이는 경찰이 사건 수사를 종결하면서 사건 발생·가해자 검거 통계 작성표(원표)를 작성할 때 범죄행위를 입력하는 것이 아니라 해당 법률조항과 죄명을 기재하는 통계분류체계라고 설명할 수 있다. 이러한 특징은 국내 범죄명을 ICCS에 연계하는 데 장애요소였기 때문에 개발연구의 효율성을 제고하기 위해 위의 표와 같은 연계원칙을 마련하여 한국범죄분류체계(안) 개발에 적용하고 있다.

2022년 연구결과로 제시된 한국범죄분류체계(안)에서 균형법을 제외한 총 15,951개의 죄명코드<sup>4)</sup>를 활용하여 총 15,321개의 국내 범죄명 코드와 ICCS 연계작업을 완료하였다. 특히, 연계율 제고와 국내 범죄통계 특성을 반영하기 위해 국내범죄명에 규정된 ‘미수’, ‘방조’ 등의 범죄 정도를 구분한 세세분류(8자리) 도입하여 한국범죄분류에 독자성과 통계적 포괄성을 확보하였다.

국가대응 모니터링, 치안 정책 평가 등 다양한 분야 활용 기대

앞의 결과를 살펴보면, 개발 중인 한국범죄분류를 활용하면 국제범죄분류와 국내죄명 간의 연계율은

4) 2022년 기준 대검찰청 죄명코드는 총 20,165개(신규죄명코드 108개 포함)이고 이중 균형법은 4,214개인 것으로 조사되었다.

약 96.0%에 달한다. 이는 대륙법계의 법률체계를 따르면서도 법률조항 중심인 한국법률체계와 행위위주로 범죄를 구분한 국제범죄분류의 성격상 차이를 감안한다면 상당한 수준이며 국제 비교가 가능한 범죄통계의 구축과 활용 가능성을 보여 준다.

따라서 통계청은 2023년 일반분류화를 위한 연구를 수행하여 한국범죄분류체계(안) 타당성과 활용가능성을 점검하고, 전문가 자문과 통계작성기관과의 논의를 거친 후 2024년에 일반분류를 지정하여 분류체계 활용을 권고할 예정이다. 국제 비교성을 바탕으로 한 한국범죄분류체계의 개발과 활용은 국내 범죄통계 분석도구의 통일, 국제 비교의 효율성 확보와 함께 우리 사회의 안전을 위한 범죄 수준의 변화 측정, 범죄에 대한 국가 대응 모니터링, 치안정책 평가 등의 다양한 분야에 활용될 것으로 기대된다. 특히 유엔 글로벌 의제인 지

속가능발전목표 16번 목표(SDG16)<sup>5)</sup> 이행 모니터링 지표 작성에 기반이 되는 유엔 범죄동향조사(UN-CTS) 자료 항목을 확대하여 제공할 수 있게 된다.

이를 위해 통계청은 2024년부터는 한국범죄분류 지침서 개발과 분류체계 개정관리를 위한 프로세스 구축을 추진하여 향후 보다 안정적인 한국범죄분류 운영을 위한 방안을 마련할 예정이다. 아울러 국제범죄분류 도입의 단계적 도입과정, 활용 방안 등을 메뉴얼화하고, 통계청과 UNODC가 아시아·태평양 지역의 범죄통계 발전을 위해 설립·운영 중인 「아·태 범죄통계 협력센터」와 함께 한국의 국제범죄분류 도입과정, 활용사례를 홍보·보급하여 통계 선진국으로서의 우리나라의 위상도 높일 예정이다.



5) 지속가능발전을 위한 평화롭고 포용적인 사회 증진, 모두에게 정의 보장과 모든 수준에서 효과적이고 책임성 있으며 포용적인 제도 구축

# 글로벌 통계활동의 핵심 기구 「유엔통계위원회」

진유강 | 통계청 국제협력담당관실 사무관



## 통계 분야 최대 종합 국제회의체, 유엔통계위원회

유엔통계위원회(United National Statistical Committee, UNSC)는 매년 전 세계 150여개 국가와 60여개 국제기구 및 비정부기구의 통계기관장들이 참석하여 글로벌 통계 발전을 위해 각종 통계 기준 설정, 신규 통계 개발, 지침서 발간, 통계자료의 보급과 국가 통계인들의 역량강화를 논의하는 통계 분야 최대의 종합 국제회의체이자 최고위급 의사결정기구다.

유엔통계위원회는 유엔 경제사회이사회(UN Economic and Social Council, ECOSOC) 산하 기능별 위원회(Functional Committee) 중 하나로 1947년 설립되었으며, 유엔 경제사회국(UN Economic and Social Affairs, UNDESA) 내 유엔통계처(UN Statistics Division)가 사무국 역할을 담당하고 있다. 유엔통계위원회 회의는 매년 3월 미국 뉴욕에서 개최되며, 경제·사회·환경 분야 의제 약 20~30여 개를 매년 논의하고 주요 의사결정을 추진한다.

유엔통계위원회는 유엔의 다른 위원회들과 마찬가지로

위원국(member countries)과 옵서버국(observer countries)로 구성되어 있는데, 위원국은 주요 의제에 대한 사전 검토를 포함 위원회 최고 의사결정기구인 의장단(Bureau) 진출 자격이 주어진다. 유엔통계위원회의 위원국은 현재 총 24개 국가이며, 지리적 분포를 고려하여, 아시아태평양지역 4개국, 아프리카 5개국, 동유럽 4개국, 중남미 4개국, 서유럽 및 기타지역 7개국 등으로 배분되어 있다. 위원국의 임기는 총 4년이며 연임이 가능하다. 위원국은 ECOSOC 관리회기에서 매년 임기가 종료된 공석을 대상으로 선출한다. 금년 4월 5일 미국 뉴욕에서 개최된 ECOSOC 관리회기에서 우리나라는 2024-2027년 임기 위원국으로 재선출됨에 따라 2016년에 위원국 진입 후 3회 연속 진출하는 쾌거를 달성하였다.

유엔통계위원회 최고 의사결정기구인 의장단은 총 5명으로 구성(의장 1명, 부의장 3명, 보고관 1명)되어 있다. 이들 의장단은 회기마다 선출되는데, 보통 임기는 최장 2년이다. 2023년 제54차 유엔통계위원회 회기 기준 의장단은 현재 헝가리(의장), 한국, 스위스, 튀니지(이상 부의장), 멕시코(보고관) 등으로 구성되어 있다. 의장단은 회기별로 회의를 주재





유엔통계위원회에서 발언하는 한훈 통계청장



제54차 유엔통계위원회 의장단

하는 역할을 수행하고 회기 간(inter-sessional)에는 약 2개월에 한 번씩 의장단 회의를 개최하여 차기 회기 의제 및 현안 이슈 등을 논의한다.

### 통계청은 유엔통계위원회 활동에 적극 참여 ... 한훈 통계청장 부의장으로 선출

통계청은 한국의 중앙통계기관으로서 매년 통계청장을 수석대표로 하는 대표단을 구성하여 유엔통계위원회에 참석하고 있다. 매 회기 중 논의되는 다양한 통계 의제에 대해 우리나라의 입장을 적극 피력하고, 통계 선진국으로서의 위상을 강화하기 위한 노력을 지속적으로 추진하고 있다. 뿐만 아니라 회의 기간 중 각국 통계청장 및 국제기구 통계기관장과 양자회담을 개최하여 국제적 통계사업의 발전 방안을 모색하고 우리나라의 통계 제도를 개선하기 위한 다양한 사업의 발판도 마련하고 있다. 우리나라는 2016년부터 위원국 자격으로 통계위원회 활동에 적극 참여하고 있으며, 2016~2017년에는 부의장국으로, 2022년에는 보고관 자격으로 의장단 역할도 수행한 바 있다.

지난 3월 미국 뉴욕에서 개최된 제54차 유엔통계위원회는 코로나19 팬더믹으로 인해 3년 만에 대면회의로 개최되었는데, 통계청도 한훈 통계청장을 대표단장으로 한 대표단을 파견하여 다양한 의제 논의에 적극 참여하고 한국의 선진 통계 경험을 공유

#### 최근 3년간 통계청 유엔통계위원회 참가 현황

참가 기간	주요 주제 및 우리나라 참가단 활동
2021.3.1. ~ 3.5. (52차)	· 2030 SDGs 데이터 및 지표, 유엔 통계시스템 조정, 경제통계, 국민계정, 국제무역 및 기업통계, 이주통계, 통계시스템 현대화 및 관리, 빅데이터, 국제통계분류, 가구조사, 데이터와 메타데이터 제시 및 배포 등 다양한 의제와 관련한 주요 발언
2022.3.1. ~ 3.5. (53차)	· 2030 SDGs 데이터 및 지표, 데이터 스튜어드십, 인구통계, 범죄통계, 국민계정, 농어업통계, 빅데이터, 통계·지리 정보통합 서비스 등 의제 주요 발언 · 개회식 중 회원국 만장일치로 보고관에 선출
2023.2.28.~3.3 (54차)	· 2030 SDGs 데이터 및 지표, 이주통계, 경제통계, 기업 및 무역통계, 국민계정, 국제통계분류, 빅데이터, 데이터 스튜어드십, 통계·지리정보 통합서비스 등 의제 주요 발언 · 개회식 중 회원국 만장일치로 부의장국으로 선출

하였다. 특히, 금년에는 한훈 통계청장이 의장단의 부의장으로 선출됨에 따라 최장 2년간 유엔통계위원회 활동을 추진해 나아가는 리더 역할을 수행하게 된다.

통계청은 유엔통계위원회 참석 외에도 유엔과 다양한 협력사업을 추진하고 있다. 대표적으로, 2010년부터 유엔통계청과 공동으로 ‘공식통계 국제세미나’를 개최해오고 있다. 동 세미나는 매년 주요 글로벌 통계이슈를 논의하는 회의로 각국 통계청이 직면한 기술변화와 혁신에 따른 대응 방안을 폭넓게 다뤘었다. 그간 지속가능발전목표(Sustainable Development Goals, SDGs) 및 고령화, 저출산 등 사회문제 뿐 만 아니라, 최근에는 데이터 혁명시대를 맞이하여 통계청의 역할 확대에 따른 ‘데이터 스튜어드십’, ‘데이터 공유 보안’, ‘데이터 과학 및 인공지능’ 등을 주제로 아태지역 및 전 세계 통계 전문가들과 함께 논의하였다.

### 유엔통계위원회, 글로벌 도전과제를 적극적으로 수용하기 위한 의사결정기구로 변모

지난 2022년 유엔 사무총장은 ‘Our Common

Agenda’라는 이니셔티브를 통해 팬더믹과 급격한 기술발전에 따른 데이터 혁신의 시대를 맞이하여, 유엔 내 통계 및 데이터의 활동을 확대하는 방안을 추진 중이다. 이에 따라 유엔통계위원회를 글로벌 통계활동의 핵심 기구(primary body)로 지정하고, 유엔통계청을 유엔 시스템 내 모든 통계 및 데이터 업무를 총괄 조정하는 기구로 격상시켰다.

이에 따라, 과거 국가통계 생산과 배포에 한정되어 있던 유엔통계위원회의 역할이 확장되었고, 위원회는 새로운 글로벌 도전과제를 적극적으로 수용하기 위한 의사결정기구로 변모하고 있다. 이에 따라, 최근 통계와 데이터에 젠더적 시각(gendered perspectives)을 주류화 하는 방안을 모색하고, 경제 사회 환경 통계를 통합하고 지속가능한 발전을 측정하기 위한 ‘Beyond GDP’ 이슈 등을 공론화하는 데 주도적 역할을 수행하고 있다. 또한, 이와 더불어 유엔통계위원회는 유엔 시스템 내 위원회의 위상 강화를 위해 현 24개국에 한정된 위원국수를 47개국으로 확대하는 작업도 진행 중이다. 이와 같이, 유엔통계위원회의 위상이 높아짐에 따라 위원국으로서, 그리고 의장단의 부의장 역할을 수행 중인 통계청의 역할도 점차 중요해지고 있다.



# A I 가 불 러 온 변 화 데이터가 지배하는 축구

김종원 | 데이터 분석가(비프로일레븐)



스포츠 경기에서 데이터는 주관적으로 생각하던 것들을 명확하게 보여주고 숨겨진 진실들을 밝혀내어 감독과 선수들이 경기를 준비하는데 중요한 정보로 사용된다.

데이터를 가장 활발히 사용하고 있는 스포츠 중 하나인 야구의 경우, 투수가 공을 던지기 전 양쪽 팀은 투수와 타자의 과거 모든 데이터를 수치화하고 투수는 경기 상황(이닝, 점수 차, 주자 등)과 상대 타자의 특성을 고려해 가장 성공 확률이 높은 공을 던지고, 타자 역시 다양한 상황을 통계적으로 고려하여 상대 투수의 구질을 예상하여 다음 타격을 준비한다. 육상, 역도, 양궁 등 기록 스포츠의 경우 선수들의

자세, 각도, 궤적 등 역학적인 데이터들이 최고의 경기력(힘, 속도, 정확도 등)을 발휘하기 위한 중요한 정보로 사용된다.

하지만, 22명의 선수가 동시에 움직이고 다음 행위를 예측하기 어려운 동적 스포츠인 축구 경기에서 데이터는 활용 빈도가 낮았고, 실제 현장에서는 데이터보다 주관적인 분석이 더 중요하게 여겨져 왔다. 하지만 최근 AI(Artificial Intelligence: 인공지능) 기술의 발달과 함께 축구 경기에서도 의미 있는 데이터 수집이 가능해졌고, 현장 적용 가능한 새로운 지표들이 생겨나면서 경기를 준비하는 과정에서 데이터를 활용하는 새로운 시대가 열렸다.

## 축구 경기에서 활용되는 데이터 수집 기술

축구 경기에서 발생한 모든 행동(사건)들을 데이터로 수집하기 위해서는 우선적으로 경기 영상이 필요하고, 영상을 확보하기 위한 경기 촬영이 진행되어야 한다.

축구 경기 촬영은 그동안 구단 내 전력분석관, 프런트 직원, 촬영 대행 업체 등 다양한 경로를 통해 진행되어왔다. 경기 촬영은 긴 촬영 시간과 촬영하는 사람의 상당한 집중력이 필요하고, 이 과정에서 날씨, 카메라 위치, 촬영하는 사람의 컨디션 등 다양한 변수들에 의해 경기를 일정하고 정확하게 촬영하는 것은 쉽지 않다. 하지만, 이제 기술의 발전과 함께

사람이 직접 축구 경기를 촬영하지 않고도 영상을 확보할 수 있는 시대가 열렸다.

AI 기반 스포츠 분석 솔루션 회사인 비프로컴퍼니(Bepro company)는 인공지능 기술을 이용해 경기 촬영을 자동화하는데 성공했다. 경기장 높은 곳에 고정형 네트워크 카메라 3대(Fixed cam)를 설치하고, 카메라 3대에서 각각 촬영한 경기 영상을 하나의 영상으로 결합시키는 방식을 도입했다. 여러 개의 영상을 하나의 영상으로 결합시키는 기술은 스티칭 기술(Stitching technology)이라고 불리는데, 카메라 파라미터에 눈금을 찍고 3개의 카메라에 찍힌 눈금을 잇는 방식이다.

[그림 1] Stitching(스티칭) 기술 (사진출처:비프로컴퍼니)





스티칭 기술을 통해 하나의 영상을 만든 후에는 경기장 안에 있는 모든 선수들을 인식하고 추적하는 과정을 거친다. 이때 사용되는 기술이 오브젝트 트래킹 기술(Object tracking technology)인데, 멀티 오브젝트 디텍션을 통해 선수, 심판, 공 등을 인식하고 추적하게 된다.

선수 추적을 구현하고 나면 비프로 자체 플랫폼의 3D 비디오 플레이어에 스티칭되고 디텍팅된 영상을 집어넣어 사용자에게 택티컬뷰, 줌인/줌아웃 등 영상을 컨트롤 할 수 있게 해준다. 선수들의 추적이 가능해짐으로써, 선수들의 신체 활동에 대한 정보가 자동으로 입력되고, 선수들의 뒀거리, 스프린트, 고강도 스프린트, 히트맵 등 다양한 피지컬 데이터들을 추출할 수 있게 된다.

그동안 축구 경기에서 피지컬 정보는 웨어러블 장비(스포츠 브라)를 입고 그 안에 있는 GPS 장비를 통해 선수들의 위치를 추적해왔는데, 인공지능 기

술의 발전으로 인해 이제는 선수들이 장비를 착용하는 번거로움과 GPS의 오류 등의 문제점들이 해결됐다.

선수의 위치 추적이 가능해지면서 선수들의 피지컬 데이터 뿐만 아니라 전술적, 기술적 데이터들도 확보할 수 있게 됐다. 주어진 공간 안에서 상대 11명의 선수들을 넘어서 상대 골문에 득점을 해야 하는 특성을 지닌 축구는 선수들이 기술을 가지고 전술적 움직임을 통해 상대 공간을 뚫어내는 것이 핵심이다.

선수 위치 추적 기능과 함께 선수들 간의 간격, 선수 사이의 공간, 압박의 위치 등 X와 Y 좌표 정보들을 통해 다양한 전술적 데이터들을 찾아낼 수 있고, 더불어 공이 나아가는 방향, 거리, 위치 등에 대한 선수들이 한 행위에 대한 기술적 데이터를 추출하는 것 역시 가능해졌다.



[그림 2] Object tracking(오브젝트 트래킹) 기술 (사진출처:비프로컴퍼니)



[그림 3] Object tracking(오브젝트 트래킹) 기술을 이용한 전술적 데이터 활용 (사진출처:비프로컴퍼니)



[그림 4] Object tracking(오브젝트 트래킹) 기술을 이용한 기술적 데이터 활용 (사진출처: 쿠팡플레이)

최근에는 휴대용 카메라 Cerberus가 출시되어 경기장에 카메라를 고정 설치하지 않아도 쉽게 휴대하며 경기를 촬영하는 것이 가능해졌다. Cerberus는 카메라가 선수들을 인식해 자동으로 움직이며 촬영하고, 스티칭과 오브젝트 트래킹 기술이 접목되어 다양한 체력적, 전술적, 기술적 데이터들을 추출할 수 있다. 기존 3대의 픽스캠에서 촬영된 영상을 하나의 영상으로 스티칭하던 기술이 하나의 카메라 안에서 가능하게 된 것이다. Cerberus 카메라의 개발로 인해 이제는 원정 경기에서도 카메라를 휴대하며 자동으로 쉽게 경기 촬영할 수 있게 됐고, 홈 경기장에 고정형 카메라가 없는 유소년 및 아마추어 축구팀들도 휴대를 하면서 경기를 쉽게 촬영할 수 있다는 점에서 활용 가치가 높다.

### 공격수의 골결정력 지표 ... xG

최근 축구계에서 가장 주목받고 있는 지표가 있는데 바로 xG(Expected Goal) 값이다. xG는 기대득점 값으로, 각 슈팅 별로 득점으로 이어질 확률을 산출하여 합산한 값이다. xG값은 슈팅을 하는 순간 골대와의 거리, 각도, 골대와 슈팅하는 선수 사이의 상대 수비수 숫자, 슈팅을 하는 신체

[그림 5] 휴대용 경기 자동 촬영 카메라 Cerberus (사진출처:비프로컴퍼니)

부위, 슈팅을 하는 선수에게 공이 전달되는 과정 등 다양한 상황들을 인공지능이 계산하여 도출된다.

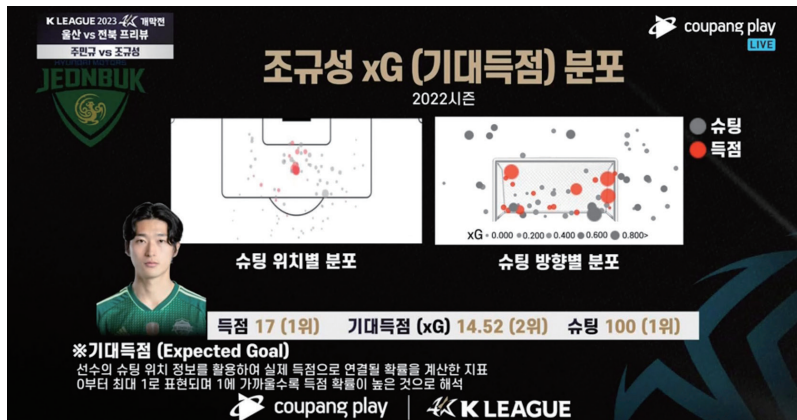
예를 들어, 패널티킥의 경우 슈팅 상황이 고정되어 있기 때문에 가장 간단하게 xG값을 계산할 수 있는데, 지난 시즌 K리그에서 패널티킥이 46번 나왔고 그 중 33번 득점에 성공했다면, 패널티킥의 득점 확률은 71.7%로, xG값은 0.717이 주어진다. xG값은 매 슈팅마다 0부터 1까지의 값이 주어지고, 0에 가까울수록 득점을 하기 어려운 상황에서의 슈팅, 1에 가까울수록 득점을 하기 쉬운 상황에서의 슈팅이다. 이렇게 도출된 xG값은 실제 득점 값과의 비교를 통해 의미있는 해석이 가능한데, xG값이 실제 득점보다 높다면 그 팀(선수)은 골 결정력이 좋지 않았다고 해석할 수 있으며, 반대로 xG값이 실제 득점보다 낮다면 득점 찬스를 잘 살렸다고 해석할 수 있다.

그동안 공격수들의 골 결정력에 대해서 평가할 방법이 슈팅 대비 득점 수 외에는 마땅히 없었는데, 이제 xG값을 통해 조금 더 경기 상황을 고려한 골 결정력에 대한 객관적인 평가가 가능해졌다. 지난 시즌(2022년) 국내 프로축구 K리그에서 득점 1, 2등을 차지한 조규성과 주민규는 나란히 17골 썩을 기록했는데, xG값은 각각 14.517(조규성)과 17.104(주민규)로 조규성이 조금 더 높은 골 결정력을 선보였다고 해석할 수 있다.

### 미더필더와 수비수의 공격 관여 능력 지표 ... xT

지난 시즌 K리그 MVP 이청용은 리그 35경기에 출전하여 3골 2도움을 기록했다. 그의 득점과 도움 수만을 보면 MVP 수상에 의문을 품는 사람이 있을 수 있다. 하지만, 지난 시즌 K리그를 시청한 사람이라면 누구나 이청용이 팀의 공격 전개를 매끄럽게 이끌고 공격과 수비 사이에서 원활유 역





[그림 6] xG 데이터 활용 예시 (사진출처:쿠팡플레이)

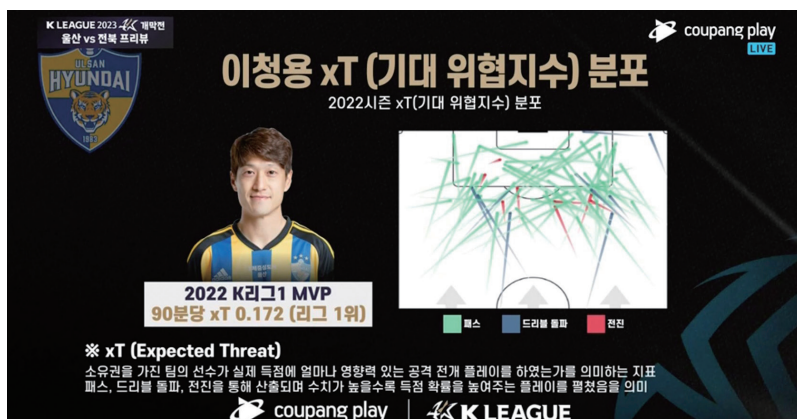
할을 하며 팀의 핵심적인 역할을 했다는 것을 알고 있을 것이다. 이처럼 이청용의 팀 내 기여도와 활약은 단순히 득점과 도움 수로 평가하기는 어렵다. 하지만, 이제는 객관적인 수치로도 이청용 선수의 활약을 확인할 방법이 생겼다. 지난 시즌 이청용 선수는 K리그 전체 선수들 가운데 1등을 기록한 지표가 있는데 바로 xT값이다. xT(Expected Threat) 값은 기대 위협 값으로, 득점 확률이 높은 지역으로 공을 전달하는 능력을 의미한다.

공은 3가지 방법(패스, 드리블, 캐리)을 통해 전달될 수 있으며, 경기장을 구역별(21x18)로 나누어 각 구역에 해당하는 득점 가능성을 부여하고, 공의 전달

이 마무리되는 구역에서 공의 전달이 시작되는 구역의 득점 가능성 값을 뺀 값의 합으로 구해진다. 그동안 득점, 어시스트, 슛, 크로스 등 득점에 집중되어 오던 경기 지표에 xT라는 지표가 생기면서 좀 더 경기장 아래 지역(자신의 진영)에서 플레이하는 미드필더와 수비수들의 팀 내 공격 관여 능력을 평가할 수 있는 평가 지표가 생긴 것이다.

### 골키퍼의 선방 능력 지표 … xGOT

공격수, 미드필더, 수비수 뿐 만 아니라 골키퍼의 선방능력 역시 좀 더 객관적인 데이터를 통해 확인하는 것이 가능해졌다. 그동안 골키퍼의 선방능력을



[그림 7] xT 데이터 활용 예시 (사진출처:쿠팡플레이)

평가할 때는 클린시트(무실점) 횟수 또는 세이브 숫자가 객관적으로 평가할 수 있는 유일한 방법이었는 데, 이 지표들은 팀이나 수비진의 수준에 따라 편향된 정보가 될 수 있다. 즉, 강력한 수비진을 갖춘 강팀의 골키퍼는 선방능력과 상관없이 클린시트가 많을 것이고, 상대의 슈팅을 많이 허용하는 약팀의 골키퍼는 세이브 숫자가 많을 가능성이 높다.

팀과 수비진의 수준과 상관없이 골키퍼의 절대적인 선방 능력을 객관적으로 확인하는 것이 쉽지 않았는데, 이제는 선방지수(xGOT-실점) 값으로 평가를 할 수 있게 됐다. xGOT(Expected Goals on Target)란 유효슈팅 내 기대 득점 값으로, 유효 슈팅이 득점으로 연결될 확률을 의미한다. xGOT값은 xG값과 슈팅이 골문 안으로 향한 위치(goalmouth location)에 의해 결정되고 계산과정에서 모든 슈팅 중 유효 슈팅만 포함된다. xGOT 값이 0.25라는 것은 25%의 득점 가능성이 있는 슈팅이고, 반대로 골키퍼 입장에서는 75%의 확률로 막을 수 있는 슈팅이다. xGOT값보다 실점이 많다면 막아야 할 슈팅을 막지 못했다는 것을 의미하고, xGOT값보다 실점이 적다면 선방을 통해 실점을 지켜냈다는 것으로 해석된다.

지난 시즌 K리그에서 가장 많은 클린시트를 기록한 선수는 1등팀 울산 현대의 조현우였지만, 선방지수(xGOT-실점) 값이 가장 높았던 선수는 인천 유나이티드의 김동현이었다. 선방지수 값을 통해 김동현 선수가 실제 득점 가능성이 높은 상대의 슈팅을 가장 많이 막아낸, 즉 선방 능력이 가장 좋은 선수였다고 해석할 수 있다.

앞서 살펴본 xG, xT, xGOT 외에도 xA(기대 어시스트), xGC(빌드업지수), PPDA(압박지수) 등 새로운 지표들이 계속해서 생겨나고 있다. 과거의 1차 데이터(슈팅 수, 패스 수, 태클 수 등)에서 벗어나 축구 경기 상황을 고려한 데이터의 추출이 가능해지면서

축구라는 종목에서는 불가능할 것 같았던 데이터를 이용해 경기를 설명할 수 있는 시대가 점점 다가오고 있다.

### 인공지능 기술이 불러온 축구분석 패러다임의 변화

축구 경기에서 지도자는 다음 경기를 준비할 때 정량적 분석보다는 상대 팀의 공격 패턴과 수비 패턴, 역습 상황에서의 움직임, 압박의 형태 등 정성적 분석을 통해 상대의 공간을 어떻게 뚫고 상대의 공격을 어떻게 막을 것인가에 대해 고민한다.

득점, 슈팅, 패스 성공률 등 결과 중심적인 정량적 데이터들은 경기를 준비하는 과정에서 활용 가치가 낮았고, 무시를 받아왔다. 하지만, 이제 인공지능 기술의 발달과 함께 전술적인 상황을 고려한 과정 중심적인 정량적 데이터들을 추출하는 것이 가능해지면서 축구 경기에서 데이터의 활용은 이제 선택이 아닌 필수가 되고 있다.

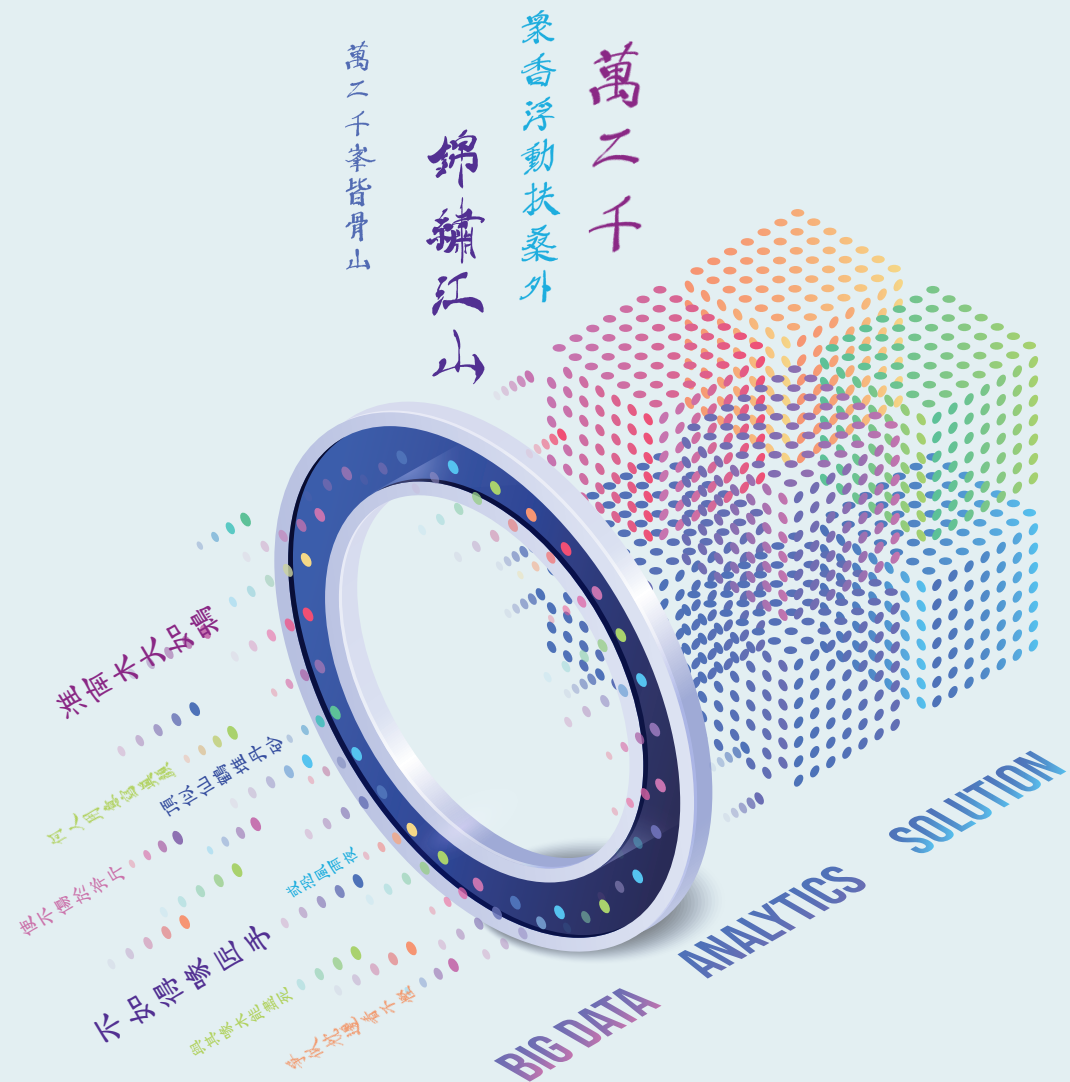
예를 들어, ‘상대 팀이 우리 수비진영으로 전방압박을 해올 때 경기를 잘 못풀어 나간다’는 상황이 ‘경기장 수비지역 1/3 위치에서 상대 선수가 15m/s 이상의 속도로 접근해서 수비선수 반경 1m 내로 접근해오는 상황’을 전방압박 상황으로 정의하고, 정의한 상황에 대한 데이터와 영상 정보를 추출하는 것이 가능해진 것이다.

인공지능 기술의 발전과 함께 축구 분석의 패러다임은 변하고 있다. 이제 축구 경기에서 인공지능 기술은 없어서는 안될 존재로 자리매김했다. 지금 이 순간에도 인공지능 기술은 계속해서 발전 중이고, 새로운 의미 있는 지표들이 개발되고 있다. 앞으로 다가올 변화를 피할 수는 없다. 축구계에는 오지 않을 것 같던 데이터 활용 시대가 이미 시작됐고, 이제부터는 데이터를 얼마나 잘 해석하고 의미 있는 정보를 잘 찾아내는지가 가장 중요한 요소가 될 것으로 보인다.



# 조선시대, 송남잡지에서 살펴본 아날로그 데이터 사이언스

최원재 | 클라우드와이즈 대표



흔히 데이터 사이언스라고 하면 빅데이터와 이에 대한 분석을 먼저 떠올립니다. 그리고 빅데이터 분석이라고 하면 또 통계와 확률을 떠올립니다. 그러나 데이터 사이언스의 목적을 따져보면 반드시 데이터 사이언스가 숫자와의 툴기만으로 이뤄진 것이 아니라는 사실을 알게 됩니다.

어떤 대상의 데이터를 놓고 그 숫자에 대한 해석이든 사실에 대한 내용 분석이든 데이터 사이언스는 분석가의 인지력에 많은 부분을 의존합니다. 통계로 나온 결과의 숫자를 ‘떼었다, 붙였다’ 조합하면서 누군가를 설득할 수 있는 분석을 도출하려면 자신이 보는 관점의 근거를 대야 하기 때문입니다.

이러한 점에서 데이터 사이언스는 자기만의 통찰력을 끄집어내고 이를 자기만의 언어로 표현하는 공간임이 틀림없습니다. 데이터 사이언스라고 하는 정의는 상당히 커다란 담론이어서 다가가기 어려운 면이 많아 보입니다. 감이 잡힐 듯 잡히지 않는 일종의 회색 지대이지요.

그동안 대학강의를 하면서 가장 안타까운 점은 학생들이 데이터 사이언스의 필요성에 대해서는 절감하면서도 자기가 할 수 없는 것이라고 서둘러 마침

표를 찍어버리는 분위기입니다. 아마도 데이터 분석을 위해 배워야 하는 컴퓨터 프로그래밍이 큰 걸림돌인 것 같습니다. 그러나 Chat GPT의 등장 이후 이런 태도에도 조금씩 변화가 보입니다. 오늘 말씀드리고자 하는 것은 숫자가 아닌 문자의 결과이지만 숫자로 변환되는 데이터의 모습 못지않게 중요한 데이터에 대한 해석력과 그 구현 방법론에 관한 이야기입니다.

## 창의적 통찰의 순간 ... 데이터 속에 보이는 인사이트

디지털로 생각을 나타내기 위해서 현대에는 컴퓨터 언어로 프로그래밍을 하고 시각화를 합니다. ‘프로그래밍’이라고 하면 우선 무엇을 알아야 할까요? 컴퓨터 언어를 배워야 합니다. 그리고 도메인 지식으로 관련 데이터를 섭렵해야 합니다. 이렇게 두 가지가 마련되면 데이터를 프로그래밍으로 조작할 수 있게 됩니다. 그리고 데이터가 조직화되는 프로그래밍 과정에서 평소에 짝 눌러있던 답답함이 한 방에 해소되는 인사이트가 불현듯 나타날 때가 있습니다.



사실 이때를 위해서 데이터를 프로그래밍한다고 해도 무리가 아닙니다. 데이터 분석은 데이터의 정렬, 나열이 아니기 때문입니다. 데이터 분석을 통해 결론적으로 자기의 생각이 나타나야 합니다. 데이터 분석에 있어서 분석자가 누구나 다 볼 수 있는 것을 굳이 보여준다면 데이터 사이언티스트라고 불리기에 너무도 초라할 수 있습니다. 이쯤 되면 데이터에 관한 내용을 알기 때문에 시각화를 잘 할 수 있는 것인지 시각화를 프로그래밍하다 보니 인사이트가 생겨나는 것인지 구분이 안 되기도 합니다.

중요한 것은 데이터 사이언스에서 이러한 창의적 통찰의 순간(아하! 순간)이 분석자에게나 분석 결과를 마주하는 사람에게나 모두 큰 도움이 된다는 점입니다. 데이터 사이언스의 가장 중요한 부분일지도 모르겠습니다. 흔히 말하는 데이터 시대란 데이터를 버무리면서 나타나는 이와 같은 창의·인지력의 선순환 구조가 만들어지는 곳일 테지요.

그런데 이러한 기법적 원리는 오늘날의 ‘디지털’ 데이터 사이언스에서만 보이는 것은 아닙니다. 과거 아날로그 데이터 사이언스에도 있었습니다. 단지 처리하는 데이터의 양이 비교할 수 없을 정도로 많아졌고 그 처리 속도가 컴퓨터의 도움으로 굉장히 빨라졌다는 사실 외에 데이터가 창의적 인사이트를 불러일으킨다는 점은 변함이 없습니다.

과거 디지털이 없던 시기에도 데이터 속에 보이는 인사이트를 이러한 창발적 체계에 맞춰 공유할 수 있던 방법이 있었으니 바로 시작(詩作)이었습니다. 컴퓨터가 없던 시절에 다양한 데이터를 데이터베이스로 분류하고 정리하면서 나타나는 분석자의 통찰력은 시로 나타났습니다. ‘한자(컴퓨터 언어)로 시를 짓는다(코딩)’는 것은 지금의 프로그래밍만큼이나 정교하고 논리적인 인지 과정의 알고리즘 없이는 불가능했습니다. 한시 작법이 누구나 배울 수 있는 것도 아니었고, 시작에서 가장 중요한 추상화는

프로그래밍 과정에서 나타나는 추상화와 맥락이 똑 같습니다.

그러므로 많은 데이터를 다루면서 순간 떠오르는 인사이트를 엄격한 문법에 맞춰 한자로 코딩하는 일련의 과정을 단순히 감정의 문학적 문자화로만 볼 것이 아니라 직관과 유비 사고를 거쳐 나오는 추상화와 정규화로 구현되는 데이터 시각화 단계의 하나로도 볼 수 있을 것입니다.

### 송남잡지와 데이터 큐레이션

『송남잡지』는 1855년에 송남 조재삼이 간행했습니다. 6책 14권으로 편찬한 백과전서적 성격의 유서(類書)로서 1,743장으로 이루어진 빅데이터입니다. 총33류 4,432칙으로 구성되어 있고, 천문류부터 국호류, 화학류, 집물류, 충수류, 어조류에 이르기까지 사물을 총망라하였습니다. 조선 후기의 사회상과 풍속을 이해할 수 있는 고유의 언어와 문화 데이터까지 집적했습니다.

이런 백과사전류의 콘텐츠는 빅데이터와 스몰데이터의 교집합을 이루는 데이터로 구성됩니다. 즉, 데이터 큐레이션입니다. 사용자가 데이터와 상호 작용하고 이해하여 분석 결과를 생성하는 데 데이터를 사용할 수 있도록 필요한 데이터 원본을 식별하고, 그 데이터를 컨텍스트와 연계하는 프로세스를 데이터 큐레이션이라고 합니다.

독자적으로 빅데이터를 구성하고 이를 분석하는 과정에서 저자 조재삼은 데이터 사이언티스트의 면모를 유감없이 발휘합니다. 데이터 리터러시와 관련해서 오늘날 『송남잡지』에 주목해야 하는 이유는 저자 조재삼이 『송남잡지』라는 데이터베이스를 만드는 과정에서 보여준 데이터 큐레이션과 ‘아하! 순간’의 데이터 큐레이션인 자작시(自作詩) 때문입

니다. 문학적 관점에서 볼 때 조재삼의 자작시는 독자들의 정서를 환기하는 장치였습니다.

그러나 데이터 리터러시 측면에서 조재삼의 자작시는 그가 데이터를 분석함에 있어 ‘아하! 순간’을 경험한 순간 발생한 직관적 영감을 살려 창의적인 방법으로 데이터 리터러시를 나타냈음을 보여주는 데이터 큐레이션입니다.

데이터 분석의 영향이 만들어낸 그만의 인사이트, ‘아하! 순간’을 한자로 코딩했다는 사실은 데이터 추상화의 단계를 밟은 결과라고 보아야 합니다. 추상화 능력은 컴퓨팅 환경에서 문제를 해결하기 위한 데이터 모델링이 요구하는 역량으로 데이터에서 필요한 것만 남겨둘 수 있는 역량입니다. 데이터 사이언티스트 조재삼은 데이터의 ‘안’을 들여다보고 데이터를 분석하고서 데이터를 보는 방법과 해석을 표현하는 방법을 달리 할 수 있게 만드는 창의력 신장의 도구로 데이터 큐레이션을 생각했던 것입니다. 즉, 조재삼은 데이터를 둘러싸고 이성과 직관과 부딪치는 ‘아하! 순간’에 데이터를 큐레이션하는 데에 있어서 그 방법으로 한자 코딩을 통해 시를 프로그래밍함으로써 데이터 리터러시의 창의적 표현 방법의 사례를 보여준 것입니다. 조재삼과 같은 특유의 표현이 있어야만 데이터는 생명력을 갖게 됩니다.

조재삼의 자작시 하나를 보면서 조재삼이 어떻게 데이터 리터러시에 데이터 큐레이션을 활용했는지 살펴보겠습니다. 제목은 「딱따구리[啄木]」입니다. 이해도를 높이기 위해서 디지털 네트워크 그래프도 만들어 비교해보겠습니다. 조재삼이 시로 구성한 시각화와 컴퓨터를 이용한 디지털 시각화를 비교하면 저자가 송남잡지를 통해 목적인 바를 선명하게 읽을 수 있을 것입니다. 디지털 데이터 큐레이션은 웹사이트 ‘최원재 프로젝트’에서 확인할 수 있습니다.<sup>1)</sup>

1)최원재 프로젝트 <http://dh.aks.ac.kr/~pattern/wiki/index.php>



저자는 여러 서적과 상황을 참고해서 딱따구리에 대한 당시의 다양한 데이터를 추출했습니다. 『운서』에서는 “딱따구리[啄木]가 열(鷺)이다”라고 한 데이터를, 『고금이전(古今異傳)』에서는 “딱따구리는 본래 뇌공(雷公)의 약초 캐는 관리인데, 이것이 새로 변하였다”라는 데이터, 『이물지(異物志)』에서는 “갈색인 놈이 암컷이고 얼룩무늬가 있는 놈이 수컷이다. 또 산작목(山斲木)·화로아(火老鴉)라고도 한다”는 데이터, 『비아(埤雅)』에서는 “열(鷺)은 금법(禁法)을 잘하니, 발톱을 구부려 땅에 금을 그어 도장을 만들면 막혔던 구멍이 절로 열린다. 날아갈 때는 날개로 그것을 덮어버린다”라는 데이터를 가져왔습니다.

독자들은 저자의 데이터베이스 덕분에 딱따구리에 대해서 한 번에 여러 데이터를 볼 수 있었을 겁니다. 이렇게 사실적 데이터를 추출하고 집적하면서 저자는 대상에 대한 ‘아하! 순간’이 일어났습니다. 그리고는 데이터 큐레이션을 시작합니다. 물론 프로그

래밍은 한자로 했습니다.

왕우칭(王禹偁)의 시, 구양수(歐陽修)의 부(賦), 우리나라에서 구전되는 시를 이용해서 만든 데이터 큐레이션이 완성되었습니다. 한시의 특징인 블록으로 규격화된 시구, 달리 말해 한자로 코딩된 데이터 프로그래밍의 일부를 적절히 활용함으로써 데이터 큐레이션의 효과를 높이고 있는 것입니다. 또한 원전 주석에서 데이터를 찾아 인지적 비축분을 높여 도메인 지식 축적에 기여하기도 했습니다. 이렇게 다양한 데이터를 마주하고 다루는 사이에 자신에게도 ‘아하! 순간’이 들어 자신만의 데이터를 만들고 이를 하나의 큐레이션으로 시각화한 것이지요. 이는 정보를 포함한 데이터를 논리적으로 재구성하는 과정에서 저자 조재삼이 직관적으로 영감을 받아 시로 형상화한 것으로 대상에 관한 정보 데이터뿐만 아니라 대상 분석에서의 감상도 한시라는 코딩 장치를 이용해 표현한 결과 「딱따구리[啄木]」라는 데이터 큐레이션이 완성된 것입니다.



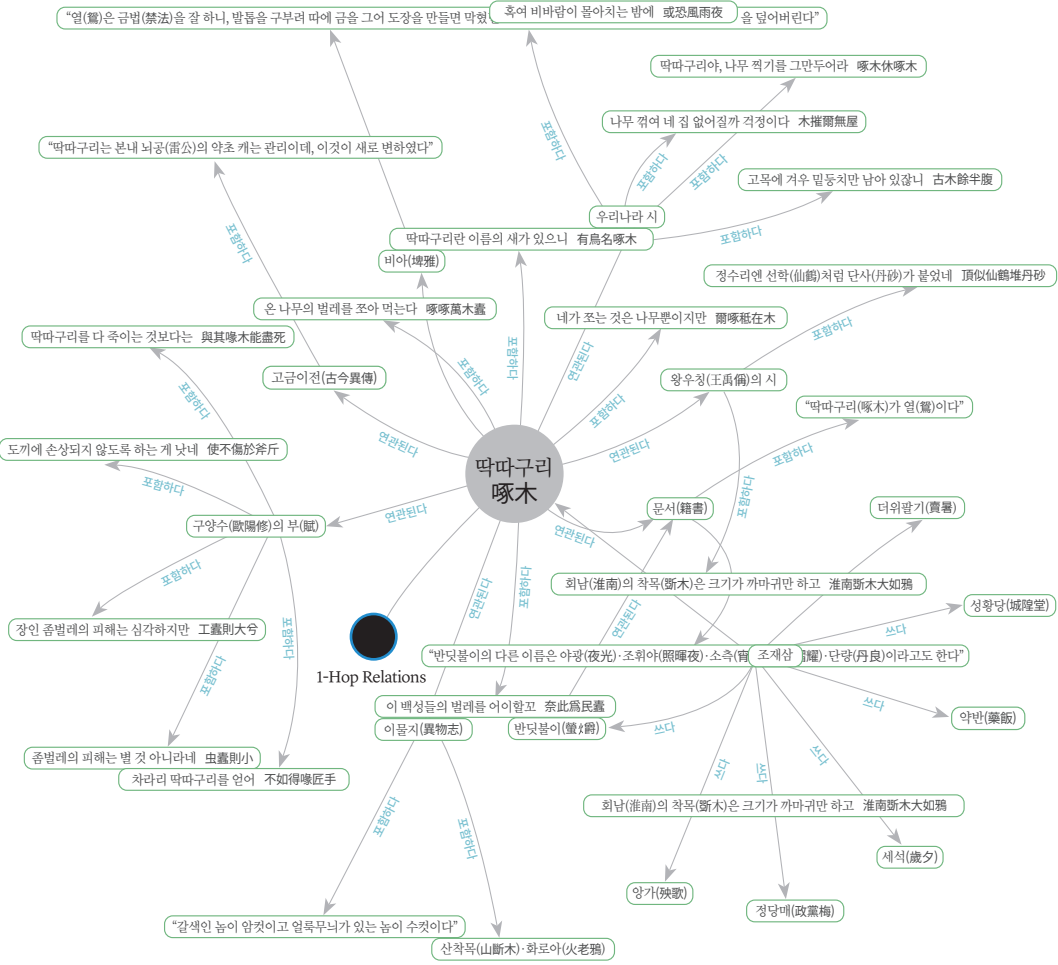
## 데이터 큐레이션으로 만들어진 네트워크 그래프

내용에서도 일관적인 흐름에 따라 분석한 데이터를 배열하는 데이터 큐레이션의 모습을 보이고 있으며 동시에 형식적으로도 자신의 데이터에 의견을 삽입함으로써 시의 추상성과 압축성을 이용해 독자들에게 제3의 데이터를 창의적으로 구현할 수 있는 접근법을 보여주고 있습니다.

저자의 데이터 큐레이션을 보면 독자들도 ‘아하! 순간’을 외칠 것 같지 않나요? 조재삼의 데이터 큐레이션에 디지털 리터러시를 추가해서 만든 데이터

큐레이션을 아래와 같이 네트워크 그래프로 만들어 보았습니다. 데이터 리터러시와 디지털 리터러시가 만나 보여주는 오늘날의 데이터 추상화와 압축성을 조재삼은 시라는 미디어로 구현한 것입니다. 다시 말해서 디지털 리터러시가 존재하지 않았던 170여 년 전 그는 데이터 분석을 통해 인사이트 메시지를 전달하는 코딩 방법으로 시를 활용했다고 볼 수 있습니다.

조재삼의 자작시 「딱따구리[啄木]」를 읽으면서 독자들은 아래에 보이는 디지털 네트워크 그래프가 보여주는 모든 데이터에 노출되는 것과 마찬가지로 효과를 경험했을 것입니다.



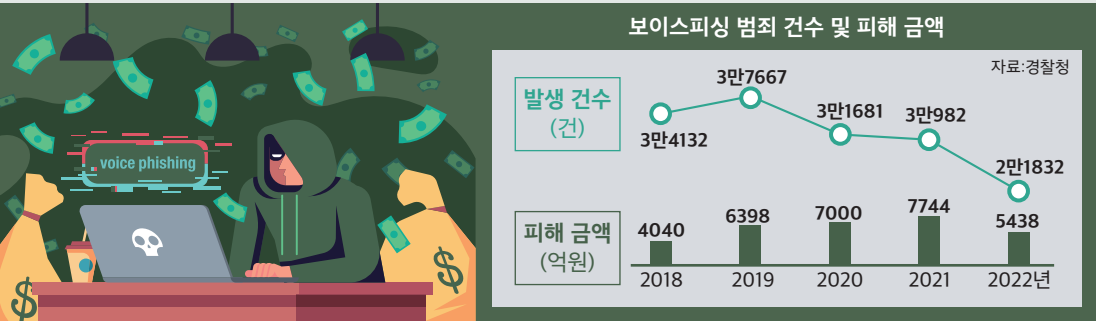
# 세계 최초 개발된 「보이스피싱 음성분석 모델」

김현우 | 행정안전부 통합데이터분석센터 주무관



## 보이스피싱 피해 건수는 총 15만 6,249건, 피해액은 3조원

보이스피싱 범죄 피해 규모는 5년간 평균 약 6,100억원 이상으로 국민 고통이 크고, 범죄예방 관련 사회·경제적 비용도 급증하고 있는 상황이다. 최근 경찰청 발표 자료에 따르면 지난 5년간 국내에서 발생한 보이스피싱 피해 건수는 총 15만 6,249건, 피해액은 3조원을 넘어서는 등 국민 생활에 심각한 영향을 미치고 있는 것으로 나타났다.



## 보이스피싱 수사를 위해 음성을 통한 동일인 여부 판단 필요

그동안 국내에서는 행정안전부 국립과학수사연구원(이하 국과수)이 러시아와 영국에서 개발한 음성 분석 모델을 활용하여 보이스피싱 수사에 필요한 음성 감정을 진행해왔다. 하지만 외국어로 학습된 분석모델 특성상 한국

어를 사용하는 범죄자의 동일인 여부를 판별하는 정확도에는 한계가 있었다. 음성 감정 시 두 개의 음성을 비교하여 동일 화자(話者) 여부를 판단해야 하는데 정확도가 낮아서 감정 불가 사례(약 70%)가 지속적으로 발생하고 있었다.

특히, 보이스피싱 범죄 조직은 통상 그룹(수사관역할, 검사역할 등)을 이루어 활동하지만 기존에 보유하고 있는 외산 모델에서는 그룹별 군집화 기능은 없었고 화자 구분만 가능했다. 군집화 유사 분석모델이 러시아에서 개발된바 있으나 사건별 연루자들을 그룹화한 기능은 없었다. 또한 국과수에서 사용하는 외산 모델은 구매 및 유지 관리 비용이 커 사용자 확장성이 부족하다는 단점이 있다. 경찰청, 각 시·도 경찰서에는 다량의 전화금융사기 범죄자 음성데이터를 확보하고 있음에도 불구하고 이를 분석할 수 있는 시스템이 없어 범죄자 음성분석이 불가능한 상황이었다.

## 최초의 한국어 기반 음성분석모델의 개발 추진

법정부 데이터분석활용 허브기관인 행정안전부 통합데이터분석센터는 외산모델의 기능 및 사용범위 한계를 해결하기 위해 화자구분의 정확성을 높이고 보이스피싱 범죄조직 식별·군집화까지 가능한 최초의 한국어 기반 음성분석모델의 개발을 목표로 국과수와 2022년 8월부터 협업체계를 구축하여 모델 개발을 추진하였다. 통합데이터분석센터는 범죄 수사 현장의 요구를 충분히 반영하여 분석과제 발굴, 상세기획, 데이터 확보·정제, 알고리즘 수립, 활용방안 등을 관리하는 역할을 담당했고 국과수는 분석모델의 알고리즘 상세 설계, 학습에 필요한 음성데이터 분류 및 가공, 반복적인 학습 및 성능 검증을 수행하여 분석모델을 고도화하는 역할을 담당했다.

‘보이스피싱 음성분석 모델’ 개발을 위한 데이터 및 모델링 과정은 약 6천명의 국내외 발화자로부터 추출한 100만개의 음성데이터를 활용, 다양한 학습과정과 성능검증과정을 반복 수행해 ①동일 화자여부 검증, ②동일 범죄가담자 그룹화가 가능한 모델을 정립하였다. 특히 한국어의 경우, 약 10만개 이상의 일반인 음성데이터와 국과수가 보유 중인 실제 보이스피싱 사기범 음성데이터를 함께 사용하여 다양한 학습 과정과 성능 검증과정을 반복 시행함으로써 보이스피싱 화자 구분 등에 필요한 최적의 알고리즘을 만들어 낼 수 있었다.

## 동일 화자여부 검증 기능

첫 번째 기능인 동일 화자여부 검증은 최신 딥러닝(Deep-learning) 기술을 기반으로 한국어와 외국어 음성데이터 학습과정을 거친 개발 모델을 활용하여 동일인일 확률 값을 출력한다. 보이스피싱 범죄자 검거 시 분석모델을 활용하여 기 확보 중인 범죄자 음성데이터들과 비교, 일치여부 등에 따라 범죄자의 여죄를 확인하는데 사용 할 수 있다.





## 범죄 가담자 그룹화 기능

두 번째 기능인 범죄 가담자 그룹화는 보유하고 있는 보이스피싱 음성데이터 풀(pool) 가운데서 화자 특징이 유사하고 연관성있는 음성들을 군집화하여 사기범 집단을 그룹화할 수 있다. 보이스피싱 범죄조직은 통상 그룹(수사관역할, 검사역할 등)을 이루어 활동하는데 아래 그림과 같이 사건(범죄1~4)별 범죄자 목소리의 연쇄 비교과정을 거쳐 동일인 확인 및 군집화를 분석모델을 통해 할 수 있다. 사건별 연루자들을 군집화 할 수 있는 기술은 사실상 세계 최초로 구현했다고 볼 수 있다.



## 기존 보유하고 있는 외산모델 대비 약 77% 향상

개발한 음성분석 모델의 모델 성능을 검증하기 위해 1차 150명의 660여개 음성데이터, 2차 200명의 12,000여개 별도 음성데이터를 사용하여 평소발성/위장발성, 일반녹음/무선통화녹음/유선통화녹음 다양한 환경의 상황에 검증을 진행하였다. 1·2차 성능 검증 결과, 범죄자의 음성을 정확하게 판별해내는 판독률이 기존 보유하고 있는 외산모델 대비 약 77% 향상된 것이 확인됐다.

기존에는 100개의 범죄자 음성 감정시 외산모델에서는 화자 동일성 여부를 28개 정도만 판별해 낼 수 있었지만 새롭게 개발한 모델에서는 51개까지 판별이 가능하다. 그리고 동일화자 및 군집화에 대한 분석모델의 정확도는 약 96.6% 수준으로 확인됐으며 정확도란 분석모델이 ‘범죄자를 범죄자로 판별’한 경우의 신뢰도 수준에 해당한다고 보면 된다.

보이스피싱 음성분석 모델이 성공적으로 개발이 완료됨에 따라 통합데이터분석센터는 국과수, 경찰청과 협력하여 보이스피싱범 수사과 검거 과정에 적극적으로 활용하고 해외 확산도 추진하고자 한다. 먼저, 국과수는 새로 개발된 모델을 2월 말부터 보이스피싱 사기범 목소리 감정에 활용하고 있으며 보유 중인 약 1만 개의 보이스피싱범 음성데이터를 군집 분석하여 범죄조직 그룹화와 이미 검거된 범죄자의 여죄 추궁 등에 사용한다. 아울러 경찰청과도 모델을 공유하여 보이스피싱 범치 초동수사의 속도와 검거율을 높이는데 활용하고 기관사칭, 전세사기 등 다양한 음성관련 범치 수사 전반에 적용해 국민피해를 최소화 하고자 한다.

## 통합데이터분석센터 ... 데이터 분석을 통해 국민·사회현안을 해결

행정안전부 통합데이터분석센터는 ‘데이터기반행정법’에 근거하여 중앙·지자체·공공기관을 아우르는 범정부적 관점에서 데이터분석 및 활용을 지원하기 위해 2021년 12월에 설치되었다. 데이터 분석을 통해 국민·사회현안을 해결하거나 데이터에 기반한 좋은 정책이 만들어질 수 있도록 지원, 기관의 데이터 분석·활용 역량 강화를 위한 교육과 컨설팅 등을 지원한다.

앞으로 통합데이터분석센터는 국민의 관심사항 및 사회이슈에 대한 지속적 분석 및 모니터링을 통해 효율적 국정운영 및 국민 체감형 행정서비스를 할 수 있도록 하려고 하며, 2023년에는 공정사회, 국민안전, 보건·의료 등의 분야의 데이터 분석과제를 선정해 ‘국민들이 실질적으로 체감할 수 있는 혁신적인 대국민 서비스 제공’을 위해 전문적인 역량을 발휘할 예정이다.



# 데이터 사이언스의 핵심은 빅데이터

박성현 | 서울대학교 통계학과 명예교수



## 앞으로 사회발전에 가장 기여도가 큰 학문은 데이터과학

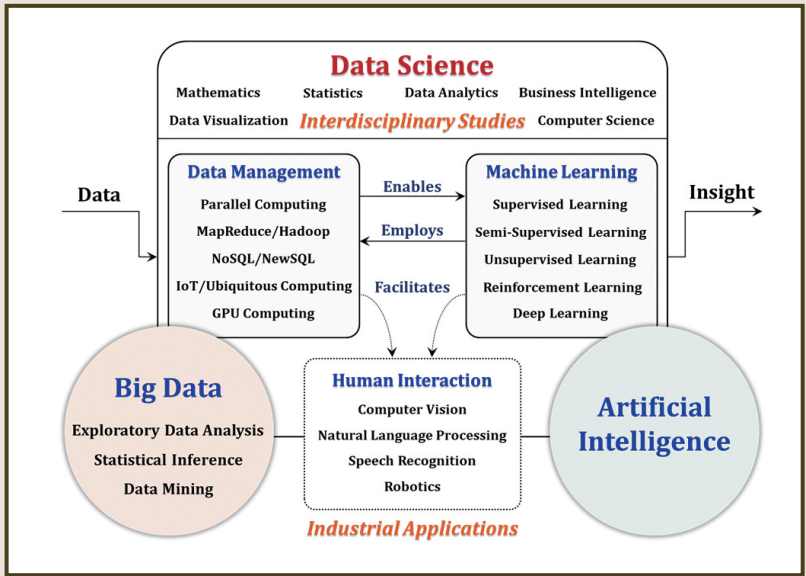
4차 산업혁명은 빅데이터, 인공지능, 사물인터넷(IoT), 메타버스 등으로 불리는 소프트웨어와 데이터 기반의 지능 디지털 기술(intelligent digital technology) 혁명을 말한다. 이 혁명은 우리의 산업과 생활방식을 급격히 바꾸어 놓고 있으며, 인류 문명에 주는 영향도 엄청나다.

4차 산업혁명의 근간에는 다량의 데이터를 소프트웨어와 연계하여 신속정확하게 정보를 창출하고 활용하는 과학이 중요성을 띄고 있다. 이런 학문을 데이터과학(data science), 혹은 데이터 사이언스라고 부른다.

데이터과학은 데이터의 수집과 저장에 필요한 데이터 프로세싱 기술과 데이터 분석에 관한 지식(통계학, 데이터 마이닝, 기계학습 등)을 기반으로 다량의 데이터로부터 패턴을 찾아내고, 통계적 추정, 예측 모델링 등을 통하여 필요한 정보를 창출하고, 이를 실제로 활용하는 것을 연구하는 융합과학(convergence science)이다. 어쩌면 21세기 사회발전에 가장 기여도가 큰 학문은 데이터과학이 될 것이라고 예측된다.

넓은 의미에서 데이터과학의 전체 조감도를 그려보면 [그림 1]과 같다. 데이터과학은 빅데이터와 인공지능(AI) 기술을 이용하여 통찰력(insight) 있는 정보를 만들어내기 위해 태어난 새로운 학문이다. 데이터과학의 밑바탕이 되는 기존의 학문들은 수학, 통계학, 컴퓨터과학, 데이터 분석학(analytics), 데이터 시각화(visualization), 비즈니스 인텔리전스(intelligence) 등으로 명실공히 다학제 간 연구 학문이다.

데이터과학은 빅데이터를 처리 분석 저장할 수 있도록 물리적 환경을 조성해 주는 데이터관리(data



[그림 1] 데이터과학의 전체 조감도(자료: 한국과학기술한림원(2019))

management) 분야와 데이터로부터 모델을 만들어 AI로 구현할 수 있도록 해주는 기계학습(machine learning) 분야로 나누어진다. 모든 단계마다 사람이 개입하여 조절 및 통제하는 인간과의 상호작용(human interaction)이 적용되며, 이 때 사용되는 기술로는 탐색적 데이터분석, 컴퓨터 비전, 자연어 처리, 음성 인식, 로봇틱스 등이며, 궁극적으로는 다양한 산업에 적용되어 새로운 제품 및 서비스를 창출함으로써 인간의 삶을 풍요롭게 해준다.

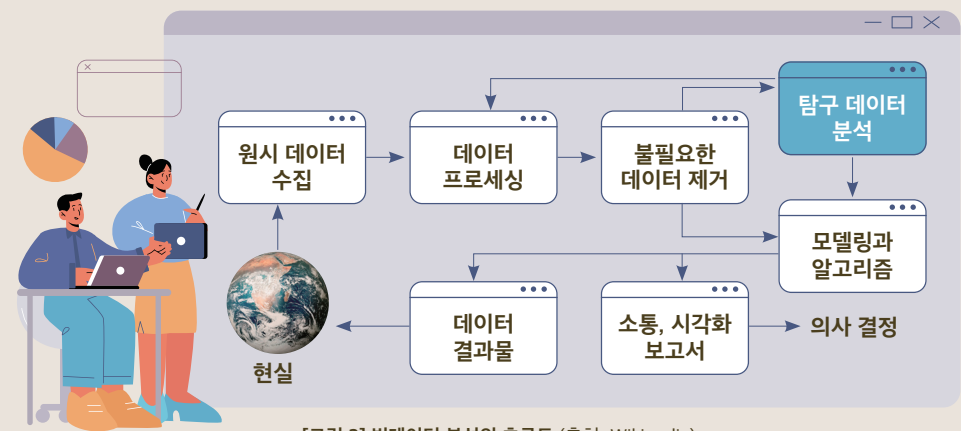
## 빅데이터는 어떻게 분석할 수 있을까

빅데이터는 데이터 형식(정형 데이터, 비정형 데이터 등)이 다양하고 생성 속도가 빨라 새로운 관리 분석 기법이 필요한 대용량의 데이터를 말하고, 흔히 대용량(Volume), 빠른 속도(Velocity) 및 높은 다양성(Variety)의 속성(3V)을 가진 정보 자원으로 정의한다. 여기에 추가하여 최근에는 데이터의 신뢰성(Veracity), 불안정성(Volatility)을 포함하여 5V의 속성을 가진 데이터로 정의하기도 한다.

데이터과학의 핵심 기술인 빅데이터는 그 분석의 흐름도로 Wikipedia 사전은 [그림 2]를 제시했다. 이 그림에서 보면 원시데이터를 먼저 수집(raw data collected)하고, 이를 프로세싱하여(data is processed) 정리하고 저장하고, 다음으로 잘못된 데이터를 제거하여 깨끗한 데이터(clean dataset)를 만드는 작업을 수행한다. 여기까지는 주로 전산과학(computational science)의 연구영역이다. 다음 단계는 탐색적 데이터분석(exploratory data analysis)과 모델링(models & algorithms)으로 이는 통계학 및 응용수학의 연구영역이다.



그 다음 단계로는 소통하고 시각화하여 보고서(communicate, visualize and report)를 작성하고 의사결정에 중요한 정보를 제공하는 것이다. 이는 전문 학문 영역(예로, 경영학, 산업공학, 보건학 등)의 영역이라고 볼 수 있다. 데이터 결과물(data product)은 현실을 정확히 반영하여야 하며, 이는 다음 단계의 원시데이터 수집에 가이드라인 역할을 하는 것이다. 즉, 빅데이터 분석을 제대로 하려면 전산과학, 통계학, 응용수학, 전문 학문(경영학, 산업공학, 보건학 등) 등의 협업이 필요한 경우가 대부분이다.



[그림 2] 빅데이터 분석의 흐름도 (출처: Wikipedia)

이제 4차 산업혁명의 정보폭발 시대를 맞아 다양한 경로로 취합된 빅데이터에서 숨은 정보와 새로운 지식을 발굴하여 혁신을 도모하려는 노력이 모든 분야에 확산되면서 빅데이터와 데이터과학이 차세대 지식혁명을 이끌 주역으로 부상하고 있다.

이미 오래전 미국의 McKinsey(2011) 보고서에 의하면 기업의 빅데이터 활용은 고객의 행동을 미리 예측하고 대처방안을 마련해 기업경쟁력을 강화시키고, 생산성 향상과 비즈니스 혁신을 가능하게 한다고 말했다. 이 주장은 정확히 맞아 들어가고 있으며, 지금은 기업 경영에서 빅데이터를 빼놓고 생각할 수 없는 시대가 되었다.

## 사례를 통해 살펴보는 빅데이터의 활용

빅데이터 분석 사례는 우리 주위에 수없이 많으나 우선 다음의 간단한 사례들을 살펴보자.

### (1) 구글의 독감 동향 시스템 빅데이터 사례

구글 검색사이트는 세계에서 가장 인기 있는 검색사이트로, 구글은 독감과 관련된 검색어 빈도를 분석해 독감 환자 수와 유행 지역을 예측하는 독감 동향 서비스를 개발했다. 이는 미 질병통제본부보다 예측력이 뛰어나고, 12주 더 빠르게 실시간으로 서비스를 제공하고 있다. 구글은 2003년부터 개별 검색어 수천억개를 분석해 ‘감기’와

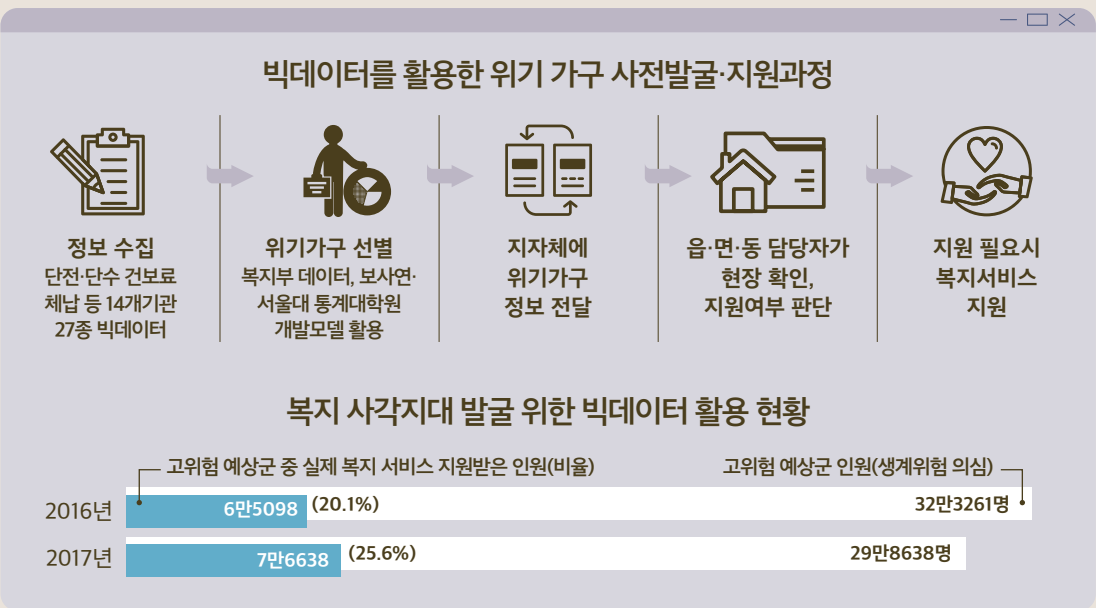
관련된 45개의 단어들을 찾아내었으며 이를 분석함으로써 특정 지역의 독감의 유행 수준을 예측하는 시스템을 개발했다.

매년 독감 시즌마다 특정 검색어 패턴이 눈에 띄게 나타나는 것을 발견하고 이 검색어들을 미국 질병통계예방센터(CDC)의 데이터와 비교해 보았는데, 주간 검색 빈도와 독감 증세를 보인 환자 수 사이에 매우 밀접한 패턴을 발견한 것이다. 이를 통해감기 관련 질병에 걸린 사람의 수를 예측할 수 있었다. CDC의 감기 감시 시스템은 데이터를 집계하고 발표하는데 12주 정도의 시간이 소요되지만, 구글 검색어는 이 같은 과정을 자동화를 통해 실시간으로 처리할 수 있어 순발력 있게 유용한 감기 정보를 제공할 수 있는 것이다. 빅데이터 분석의 위력이다.

### (2) ‘복지 사각지대’ 없애는 보건복지부 빅데이터 사례

우리 정부에서 빅데이터를 잘 활용하는 부처 중 하나는 보건복지부다. 이 내용 중의 일부는 조선일보(2018년 3월 9일자)에도 게재되어 있다. 보건복지부는 2016년부터 ‘복지 사각지대’ 발굴에 빅데이터를 활용해 왔다. 이 사업을 ‘빅데이터를 활용한 위기 가구 사전 발굴 지원 사업’이라고도 부르고, 그 과정이 [그림 3]에 요약되어 있다.

부모가 이혼한 14세 P군은 치매에 걸린 할아버지, 신장 투석을 받는 할머니와 함께 살고 있다. 2017년 7월에 P군의 아버지가 교도소에 수감되면서 생계가 어려워지자 할머니는 대출을 받고 사채까지 끌어 써야 했다. 생계가 막막해진 P군 가정도 빅데이터를 통해 징후가 포착됐다. 빅데이터 분석으로 문제를 감지한 동사무소 직원은 2018년 1월 P군 가정을 방문해 월 117만원을 긴급 지원하고, 기초생활수급자 신청을 도와주었다.



[그림 3] 복지 사각지대 발굴을 위한 빅데이터 활용 (자료=보건복지부)

보건복지부는 고용노동부와 경찰청을 비롯한 정부 부처와 한국전력공사, 신용정보원 등 14개 기관으로부터 27종의 빅데이터를 수집하고 있다. 단전, 단수, 건보료 체납, 실직 여부 등 경제적 어려움을 겪는 가정에 자살을 시도한 사람이 있는지와 같은 정신적 문제를 살펴볼 수 있는 응급의료센터 등의 자료도 포함된다. 이렇게 수집한 자료를 통계 모델로 분석해 생계 위험에 빠졌다고 판단되는 ‘고위험군 가정’을 가려낸 뒤 각 지자체에 이 내용을 전달하는 것이다. 그러면 읍 면 동 단위 지자체 직원의 현장 방문 조사가 가능해지는 것이다.

2017년에는 빅데이터를 통해 약 30만 명 고위험군을 가려내 이 중 7만 7,000명(25.6%)에게 복지서비스를 지원했다. 빅데이터의 특성상 시간이 지나면서 더 많은 정보가 쌓이면 지금보다 더 손쉽고 정확하게 고위험군을 가려내 지원이 가능할 것이다.

### (3) 빅데이터 통계분석 사례 : 서울특별시 노인 건강조사

#### ① 데이터와 분석 목적

이번 사례에서는 빅데이터에서 실제 통계분석을 어떻게 하는지 살펴보자. 원데이터 파일은 2020년에 작성된 ‘20년도 서울특별시 지역사회 건강조사 데이터’이며, 이 파일에서 연령이 65세 이상인 시민을 추출하여 데이터의 크기가 n = 5,938명인 파일을 만들어 저자팀이 분석하였다. 데이터의 양이 빅데이터라고 보기는 어렵지만 빅데이터 분석의 대표적 사례이므로 여기에 소개하고자 한다.

분석 목적은 ‘코로나로 인해 65세 이상의 노인들이 활동 제약으로 인해 본인이 느끼는 ‘주관적 스트레스 지수’가 증가되었고, 이는 건강에 영향을 주고 있다’라는 가설을 검정해 보고, 이를 통계적으로 확인해 본 것이다. 통계분석을 위해서는 사용가능한 통계 프로그램이 다양(SAS, SPSS, R, Python, Excel 등)하게 많으나 우리 주위에서 접하기 쉬운 Excel을 사용하여 분석하였다.

이 분석에서 사용된 변수들은 성별, 연령(만 나이), 현재 흡연여부, 연간 일주일 아침 식사 평균 횟수, 하루 평균 수면시간, 우울감 경험 여부, 고혈압 진단 경험 여부, 당뇨병 진단 경험 여부, 행복감 지수, 주관적 건강지수, 보건



기관 이용 여부, 주관적 스트레스 지수, 연간 낙상 경험 여부, 연간 사고중독 경험 여부 등이다. 이 데이터에 대한 서울특별시의 지역사회건강조사 원시자료 사용 지침서 에 따르면 각 변수에 설정된 코딩 값은 다음과 같다.

#### ② 변수와 변수값 코딩

범주형 변수(categorical variable)	
▪ 현재 흡연 여부	1. 매일 피움 2. 가끔 피움 3. 과거에 피웠고 현재는 피우지 않음
▪ 연간 일주일 아침식사 횟수	1. 주 5-7회 2. 주 3-4회 3. 주 1-2회 4. 거의 안한다.
▪ 우울감 경험 여부 ▪ 고혈압 진단 경험 여부 ▪ 당뇨병 진단 경험 여부 ▪ 보건기관 이용 여부 ▪ 연간 낙상 경험 여부 ▪ 연간 사고중독 경험여부	1. 예 2. 아니요
▪ 행복감 지수	1. 매우 불만족, 10. 매우 만족으로 10점 만점 차등 점수
▪ 주관적 건강 지수	1. 매우 좋음 2. 좋음 3. 보통 4. 나쁨 5. 매우 나쁨
▪ 주관적 스트레스 지수	1. 대단히 많이 느낀다 2. 많이 느끼는 편이다 3. 조금 느끼는 편이다 4. 거의 느끼지 않는다

연속형 변수(continuous variable)	
▪ 하루 평균 수면시간	연령(만 나이) : 연속형 변수로 간주

#### ③ 변수들 간의 상관분석

위에 열거된 모든 변수들 간에 상관관계가 존재하는가를 알아보기 위하여 상관분석을 실시하여 [표 1]에 상관분석표를 작성하였다. 상관분석을 모든 두 변수들 간의 상관관계를 알아보는 통계적 분석 방법이다. 상관계수 r은  $-1 \leq r \leq 1$ 의 범위 내에서 두 변수 간의 상관관계를 내는 척도이다. 즉, r이 양이면 양의 상관관계를, 음이면 음의 상관관계를 나타낸다. 이 상관계수의 절댓값이 1에 가까울수록 높은 상관관계를 나타내는 것이다.

표본의 크기가 작을 때에는 상관계수 r의 절댓값이 제법 커야 상관관계가 유의(significant) 하지만, 표본의 크기가 이 분석에서 사용하는 것과 같이 1,000보다도 클 때에는 r의 절댓값이 0.05만 되어도 유의하다. ‘주관적 스트레스 지수’가 주요한 관심 변수이므로, 이 변수와 상관계수의 절댓값이 0.05인 변수들은 ‘우울감 경험 여부’ (r = 0.29), ‘주관적 건강 지수’ (r = -0.23), ‘행복감 지수’ (r = 0.10), ‘수면시간’ (r = 0.06), ‘만 나이’ (r = 0.09), ‘성별’ (r = -0.08), ‘연간 낙상 경험 여부’ (r = 0.09), ‘연간 사고중독 경험 여부’ (r = 0.05)가 있다.

상관계수 값의 해석은, 예를 들면, ‘주관적 스트레스 지수’와 ‘우울감 경험 여부’ 간의 상관계수가 r = 0.29(양의 상관관계)이므로, 우울감 경험이 적을수록, 주관적 스트레스 지수값이 높아져서 스트레스를 거의 느끼지 않는 것이다(주관적 스트레스 지수값은 작으면 스트레스가 심하고, 크면 스트레스가 적은 것으로 코딩되어 있음). 두 번



	만나이	성별	생	출연 여	종업 여	성	유주 여	간 유주 여	여가 운연 안	신체활동도	신체활동일	아침소인	임지 체적	구강건강	수면시간	복 스트레칭	슬관절	경험	후연자 예법	진단 경험	진단 경험	복상 경험	고통증 경험	간장질환	복잡 질환	지구력	이용
만나이	1																										
성별	0.028964	1																									
평생 흡연	0.041199	0.701605	1																								
현재 흡연	0.064022	0.718596	0.944802	1																							
평생 음주	0.181069	0.345088	0.315914	0.320715	1																						
연간 음주	0.19262	0.365111	0.331448	0.336755	0.963818	1																					
자동차 운전	0.289761	0.403924	0.276414	0.28175	0.194036	0.200954	1																				
격렬한 산악	-0.02198	-0.00671	-0.01925	-0.01496	-0.03534	-0.0034	-0.00355	1																			
중등도 산악	-0.0324	-0.00456	-0.0119	-0.01114	-0.02036	0.009017	-0.01078	0.792855	1																		
산악 활동	-0.08684	0.018246	-0.02597	-0.0401	-0.00719	-0.01048	-0.00032	-0.00371	-0.00997	1																	
본인 인지	0.257906	-0.08005	-0.07401	-0.08429	0.67384	0.070491	0.117631	-0.0264	-0.04866	0.009733	1																
주관적 지각	0.200515	0.00073	-0.07254	-0.08948	0.032513	0.030461	0.117198	-0.02405	-0.02793	0.008451	0.152852	1															
수면시간	0.075601	-0.04393	-0.03635	-0.03703	-0.00506	0.013471	-0.00969	0.042784	0.046775	-0.01115	0.0356	0.017073	1														
주관적 스트레스	0.098613	-0.08421	-0.0252	-0.01801	0.004773	-0.00509	-0.02597	0.015243	0.008565	-0.04865	0.008818	-0.10664	0.064655	1													
우울감 경험	0.018128	-0.0436	-0.02598	-0.02538	-0.00041	0.017924	-0.01866	0.013817	0.012474	-0.04228	0.029269	-0.01511	0.157215	0.286891	1												
연간 인플루엔자	-0.09119	-0.01548	-0.02067	-0.03601	-0.00816	-0.0104	-0.05733	0.02497	-0.01104	0.0381	-0.00122	-0.01055	-0.00903	-0.01445	-0.02494	1											
고혈압 진단	-0.11206	-0.01952	-0.01178	-0.02306	-0.00864	-0.01283	-0.06665	-0.00977	0.003493	0.025673	0.058736	-0.03749	0.005798	-0.01924	-0.00535	0.06524	1										
당뇨병 진단	-0.05196	0.036355	0.064208	0.056633	-0.01684	-0.01055	-0.00518	0.012178	0.002956	0.035466	-0.0256	-0.07294	-0.02413	0.006529	0.015105	0.020949	0.140032	1									
연간 낙상	-0.07963	-0.08783	-0.05339	-0.04754	0.000307	-0.00677	-0.0872	-0.01603	-0.00946	-0.02944	-0.00875	-0.08185	0.021033	0.091702	0.09038	0.002294	0.011858	0.025568	1								
연간 사고 발생	-0.04125	-0.0366	-0.02568	-0.03068	-0.00677	-0.00908	-0.02991	0.01473	0.015324	0.00845	-0.02588	-0.02848	-0.00466	0.050558	0.050202	-0.00438	0.002584	0.025097	0.301039	1							
주관적 지각	0.213396	0.138208	0.048181	0.049394	0.016955	0.119144	0.230028	-0.04939	-0.04474	0.00236	0.118266	0.312888	-0.00014	-0.22681	-0.08946	-0.03106	-0.11856	-0.17158	-0.13136	-0.09795	1						
행복감 지각	0.085755	0.011581	0.016947	0.021434	0.037478	0.053853	0.016894	0.008894	0.005872	-0.01987	0.026969	0.032181	0.121668	0.109178	0.149086	0.003438	0.029437	-0.00334	0.045451	0.019656	0.011807	1					
보건기관	(-0.00278	-0.06112	-0.0289	-0.03681	-0.03223	-0.0262	-0.02497	-0.03912	-0.03361	0.005605	0.052317	0.007249	-0.01406	-0.00927	0.007899	0.083546	0.039237	-0.00874	0.028256	0.013714	0.002754	0.044641	1				

째로 상관계수 값이 큰 주관적 건강수준은  $r = -0.23$ (음의 상관관계)인데, 이는 주관적 건강 지수값이 커지면(건강이 나빠짐) 주관적 스트레스 지수값이 작아져 스트레스를 많이 느끼게 된다는 의미이다.

#### ④ 주요 변수들 간의 회귀분석

‘주관적 스트레스 지수’에 크게 상관관계가 있는 변수로 ‘우울감 경험 여부’와 ‘주관적 건강수준’이 앞의 상관분석에서 얻어졌다. ‘주관적 스트레스 지수’를 종속변수(dependent variable) Y로 하고, 이 Y에 영향을 주는 독립변수(independent variables)로 ‘우울감 경험 여부’를 X1으로 하고, ‘주관적 건강수준’을 X2로 하여 이들 간

요약 출력								
회귀분석 통계량								
다중 상관계수	0.351157							
결정계수	0.123311							
조정된 결정계수	0.123016							
표준 오차	0.740048							
관측수	5938							
분산 분석								
	자유도	제곱합	제곱 평균	F 비	유의한 F			
회귀	2	457.1916	228.5958	417.3957	2.5E-170			
잔차	5935	3250.432	0.547672					
계	5937	3707.623						
	계수	표준 오차	t통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y절편(주관적 스트레스 지수)	2.693878	0.059948	44.93674	0	2.576357	2.811398	2.576357	2.811398
우울감 경험 여부	0.526655	0.023854	22.07857	7E-104	0.479893	0.573417	0.479893	0.573417
주관적 건강 지수	-0.17228	0.010382	-16.5944	1.73E-60	-0.19263	-0.15192	-0.19263	-0.15192

[표 2] 회귀분석 출력표



의 회귀분석(regression analysis)을 실시하여 Y와 X1과 X2 간의 방정식 관계를 규명해 보았다. 이처럼 회귀분석은 종속변수에 영향을 주는 1개 이상의 독립변수 간의 함수관계를 방정식으로 규명해보는 통계분석 방법으로, 빅데이터 분석에서 애용된다. 통계분석 출력은 [표 2]와 같다.

회귀분석 결과표에 나와 있는 정보를 사용하여 변수들 간의 회귀방정식을 써보면

$$Y \text{의 추정치} = 2.693878 + 0.526655X_1 - 0.17228X_2$$

가 얻어진다. 이 방정식이 의미하는 것은  $X_1$ (우울감 경험 여부)이 한 단위 증가할 때  $Y$ (주관적 스트레스 지수)의 추정치가 0.526655 증가한다는 뜻이다. 즉, 변수의 코딩에서 보면  $X_1$ 이 증가하는 것은 우울감 경험이 없는 방향으로 가는 것이고,  $Y$ 가 증가하는 것은 스트레스가 적어지는 것이므로, 결국 우울감 경험이 적을수록 스트레스가 적어지는데, 그 양이  $X_1$ 의 값이 1에서 2로 갈 때  $Y$ 의 값은 0.526655 정도 증가한 다는 의미이다. 이와 유사하게  $X_2$ 에 대해서도 해석할 수 있는데, 주관적 건강지수가 한 단위 증가하면  $Y$ (주관적 스트레스 지수)가 0.17228만큼 감소하여 스트레스가 적어지는 것이다.

이와 같이 빅데이터 분석을 통하여 변수들 간의 관계에 대한 정보를 얻으면 이에 따라 서울 시민을 위한 보건정책을 세울 수 있는 것이다. 예를 들면, 주관적 스트레스 지수에 우울감 경험 여부와 주관적 건강 지수가 영향을 크게 주며, 그 관계의 양적 영향을 알았으므로, 주관적 스트레스 지수를 줄이기 위해서는 우울감을 경험하지 않도록 하는 방안을 세울 수 있고, 또한 주관적 건강 지수가 양호하게 관리되도록 하는 보건 정책을 입안할 수도 있을 것이다. 이렇게 유용한 정보를 주는 분석으로 빅데이터 분석이 최근 많이 사용된다. 회귀분석은 위에서 한 바와 같이 모든 관심 있는 변수들 간에 방정식을 구하여 그들 간의 함수 관계를 통계적으로 규명할 수 있다. 매우 유용한 방법이다.

# 기후 변화 예측 인공지능, 어디까지 와있나

함유근 | 전남대학교 해양학과 교수



## 인공지능 기후 예측 ... 시간과 공간 해상도에서 비약적인 향상

바야흐로 인공지능의 시대이다. 인공지능(AI) 기법은 기후 연구 분야에 다방면으로 활용되면서 기존의 기후 예측 기법들을 빠르게 대체/보완하고 있다.

기존의 기후 예측 모델들은 다양한 지구 시스템의 물리 법칙에 따라 에너지 및 물질의 흐름을 관련시키는 미분 방정식을 대규모 시뮬레이션을 통해 수치적 해를 구하는 방법을 따라왔는데, 이런 모델은 한계점을 가지고 있다. 프로세스 기반 따라서 기존의 기후 모델들은 일반적으로 슈퍼컴퓨터에서 실행해야만 원활한 구동이 가능하며, 지구상에 존재하는 모든 물리 법칙을 수치적으로 표현할 수 없다는 한계 때문에 필연적으로 실제 지구와의 계통적 오차가 존재하는 문제점이 있다.

다양한 지구 시스템 요소들의 관측을 통해 물리 법칙을 발견하고, 이를 수치적 변환하여 기후 예측 모델에 반영하는 식으로 모델의 고도화가 이루어져 왔지만, 이러한 방법으로는 관측 기술의 발전으로 위성, 레이더 및 센서로부터의 풍부한 기후 변수들의 관측을 직접적으로 도입하는 데 한계가 있을 수 밖에 없다.



반면, 데이터 기반의 인공지능 기후 예측 모형은 비약적으로 늘어난 관측들은 모형의 학습 데이터로 이용되어 기후 모형을 좀 더 실제의 지구와 같아지도록 만든다. 즉, 인공지능 기후 예측 모형은 기후 요소 관측의 증가로 인해 그 성능이 비약적으로 향상될 수 있다.

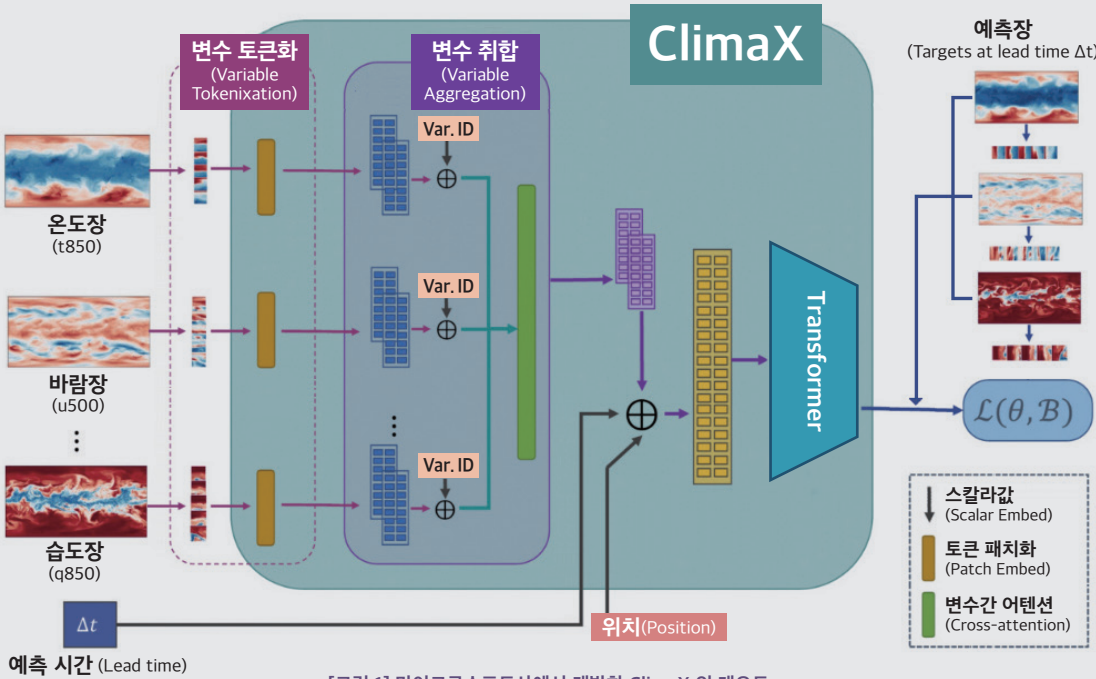
또한, 최근의 딥러닝 기법 개발은 주어진 데이터로부터 실제 지구 요소들의 관련성, 동작 원리를 더욱 효율적으로 학습하여 다양한 시간과 공간 해상도에서의 기후 예측 성능을 비약적으로 향상 시키기에 이르렀다.

## 사례를 통해 살펴보는 인공지능 기후 예측 모델

초기 연구들의 성공적인 인공지능 도입에 고무되어, 최근에는 학계에서 뿐만 아니라 다수의 글로벌 기업에서도 인공지능 기법을 활용한 기후 예측을 시도하고 있다. 이에 대표적인 사례로 마이크로소프트사에서 4주 이내의 예측을 목표로 개발된 인공지능 기반 기후 예측 모델과 수십년 예측을 목표로 개발된 인공지능 기반 기후 변화 예측 기법에 대해 소개하고자 한다.

### (1) 인공지능 기반 전지구 기후 예측 모델 ClimaX

마이크로소프트사에서 개발한 ClimaX 모델은 여러 기상 변수를 다양한 시공간 해상도에서 다루는 이질적인 데이터 세트를 사용하여 학습된 AI 기반 기후 모델이다 [그림 1].



[그림 1] 마이크로소프트사에서 개발한 ClimaX 의 개요도

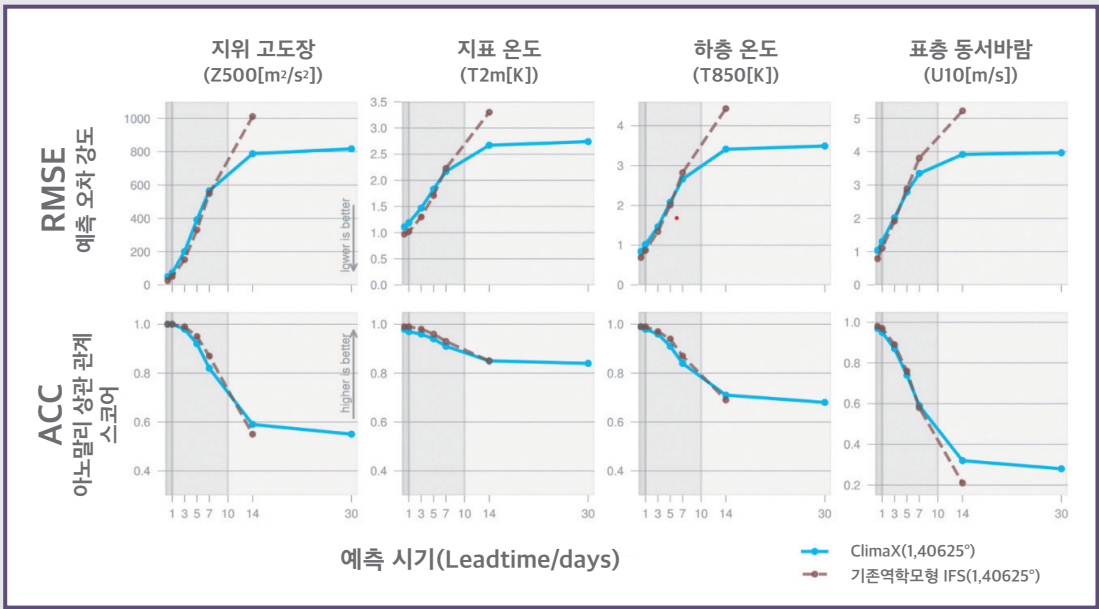


해당 모형은 사전 학습 (pre-training) 과 미세 조정 (fine-tuning) 을 거쳐 구성된다. 사전 학습은 여러 그룹에서 제공하는 다양한 기존 기후 예측 모형의 시뮬레이션 결과를 활용한다.

다양한 기존 기후 예측 모형의 시뮬레이션 결과는 실제 지구의 관측과 계통적인 오차를 가지고 있지만, 주요 물리 법칙들에 기반한 시뮬레이션 결과를 제공하기 때문에 어느 정도의 현실성을 담보하고 있다. 시뮬레이션 결과의 장점은 무수히 많은 수의 독립적인 학습 데이터를 얻을 수 있다는 것에 있으며, 다양한 기관에서 제공하는 다수의 기후 예측 모형을 활용하는 경우 모형간의 계통적 오차가 서로 상쇄되는 효과를 얻을 수 있다. 이후, 실제 관측 자료를 활용한 미세 조정을 수행하여 모형을 최종 구축한다. 이를 통해 실제 지구의 특성이 더 정확히 반영되는 인공 지능 기반 기후 예측 모형이 구축된다.

ClimaX 는 Vision Transformers (ViT) 에 기반한 다차원 이미지 대 이미지 번역 아키텍처 모형이다. ViT 기반 아키텍처는 다양한 공간-시간 입력과 유사한 다중 규모 데이터의 공간을 자연스럽게 토큰화할 수 있기 때문에 기후 변수들끼리의 시간 지연 상관성을 모델링하기 적합하다. ViT 기반 모형을 구축하기 위해서는 변수 토큰화 (Variable tokenization) 및 변수 취합 과정 (Variable aggregation) 이 수행 되어야 한다. 변수 토큰화 (Variable tokenization) 는 입력을 동일한 크기의 패치로 분할하고 각 패치를 너비, 높이 및 채널 차원에서 벡터로 평탄화 하는 과정이다. 이 후 어텐션 메커니즘 (attention mechanism)을 활용하여 특정 패치와 강한 연관성을 갖는 패치들을 찾는 과정을 거쳐 해당 패치의 예측에 활용하게 된다.

이렇게 구축된 ClimaX 의 성능을 실로 놀랍다. 가용한 프로세스 기반 기후 예측 모형과 실관측 데이터를 모두 취



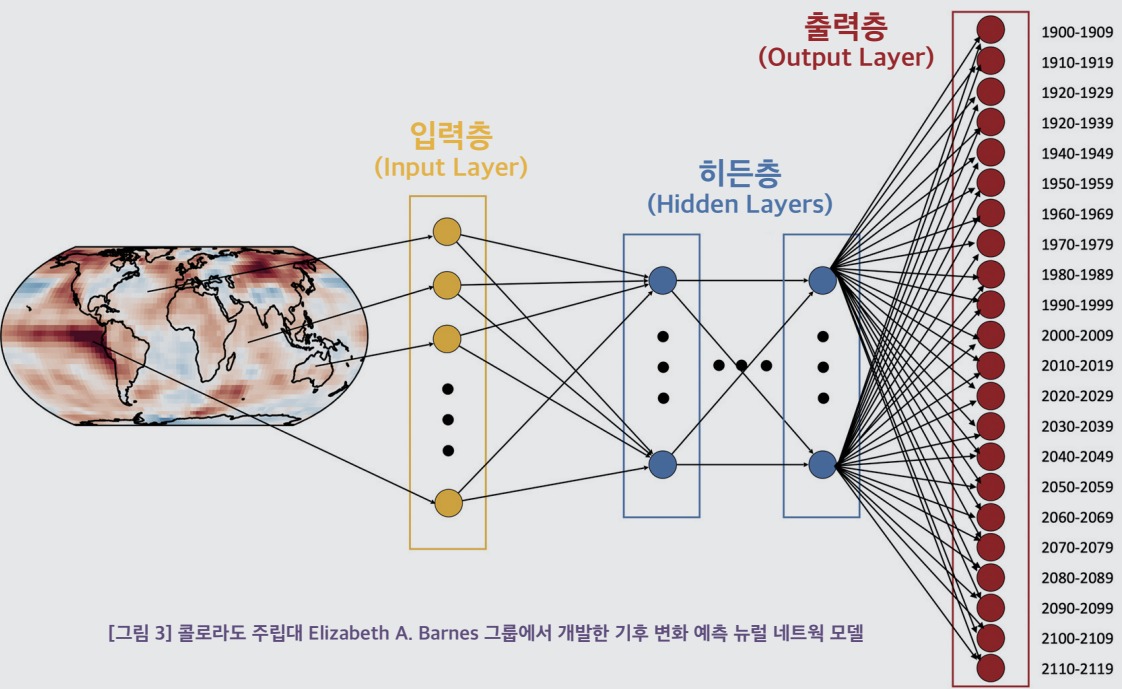
[그림 2] ClimaX (파란색) 과 기존 프로세스 기반 기후 예측 모형 (갈색)의 예측 시기에 따른 예측 오차 강도. 왼쪽부터 500hPa 지위고도장, 2m 지표 온도, 850hPa 온도, 10m 동서바람 예측 오차 강도를 의미함.

합하여 학습 시킨 결과, 단기 (일주일 이내) 예측에 대해서는 기존 프로세스 기반 모형중 가장 우수한 성능을 보이는 현업 프로세스 기반 기후 예측 모형과 비슷한 성능을 보이고, 중기 예측 (약 1주-4주 예측) 에 대해서는 기존 모형에 비해 오히려 더 나은 성능을 보인다 [그림 2].

마이크로소프트사는 해당 모형을 향후 특정 영역의 초고해상도 예측장과 기후 변화 예측에서 적용할 수 있도록 보완하여 개발할 예정이라고 한다.

## (2) 인공지능 기반 기후 변화 예측 모델

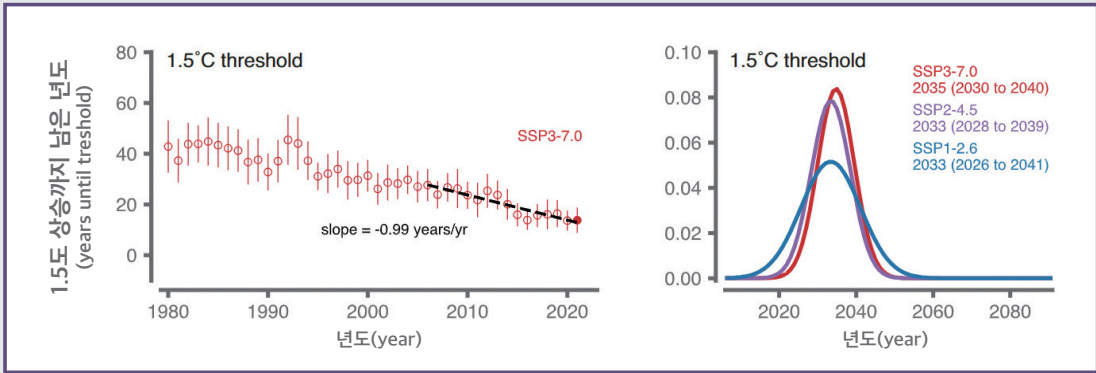
콜로라도 주립대 (Colorado State University) 의 Elizabeth A. Barnes 연구팀은 기후 변화 예측을 위한 뉴럴 네트워크 기법을 지속적으로 개발해 왔다. 뉴럴 네트워크에 기반한 모형은 전지구 예측 모델에 비해 상대적으로 단순하지만 이해하기 쉽다는 데 그 장점이 있다. 연구팀의 초기 버전 모델은, 주어진 온도 분포에 기반하여 해당 온도 분포가 몇 년도 인지에 대한 예측을 수행하는 것이다 [그림 3].



[그림 3] 콜로라도 주립대 Elizabeth A. Barnes 그룹에서 개발한 기후 변화 예측 뉴럴 네트워크 모델

해당 인공지능 모형은 프로세스 기반 기후 예측 모형 장기 시뮬레이션만을 통해 학습 되었기 때문에 실제 관측 정보는 전혀 모르는 상태로 학습이 됨에도 불구하고, 실제 관측 온도 분포에 기반한 년도 정보를 어느 정도 성공적으로 예측해 낸다. 이는 온도 분포를 통해 기후 변화가 얼마나 진행되었는지를 예측 했다고 볼 수 있다.

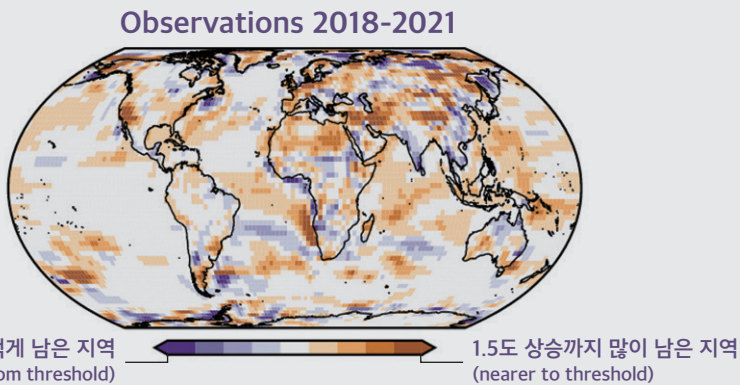
연구팀은 해당 연구를 좀 더 발전시켜 산업화 이전 대비 온도 상승폭이 1.5도가 되는 시기를 맞추는 모형을 개발하였다. 인류가 산업화 이전 대비 온도 상승폭을 1.5도 이내로 유지 시킬 수 있는지, 만약 1.5도를 넘어서는 온도



[그림 4] 콜로라도 주립대 Elizabeth A. Barnes 그룹에서 개발한 기후 변화 예측 모델에 기반한 산업화 이전 대비 1.5도 상승이 일어날 때 까지 남은 년수. 2021년 (x축)에 17년(y축)의 값은 2021년 기준 17년 이후에는 1.5도 상승이 예측된다는 의미임.

상승이 일어난다면 과연 그것이 언제 일어나는지에 대한 답은 인류가 2015년 파리 협약을 이행할 수 있는지를 판단할 수 있는 핵심 근거가 됨에 따라 최근 기후 변화 연구 커뮤니티에서 주요하게 관심을 갖는 분야 중 하나이다. 연구 결과에 따르면, 조금씩의 차이는 있지만 SSP3-7.0, SSP2-4.5, SSP1-2.6 시나리오 모두에서 1.5도의 상승이 일어나는 해를 2033-2035년 사이로 보고 있다 [그림 4].

즉, 인류가 돌이킬 수 없는 강도의 온난화가 일어나는데 남은 시간은 길어야 15년 정도라는 것이다. 만약 온도 상승폭을 2도로 산정한다면 2021년 기준으로 30년 정도의 시간이 남은 것으로 예측 하였다. 프로세스 기반 모형에서 계산된 기간은 1.5도 기준으로는 2026-2042년, 2도 기준으로는 2034-2052년인 것과 비교해서 비슷한 수준인 것으로 보이나, 프로세스 기반 기후 예측 모형의 결과에 비해 인공지능 모형의 결과가 더 좁은 년도 폭 (인공지능 기법의 경우 1.5도 기준에서 3년, 프로세스 기반 모형의 경우 16년) 을 보이는 것으로 보아 예측의 불확실성은 인공지능 기법을 활용함으로써 어느 정도 해소된 것으로 판단된다.

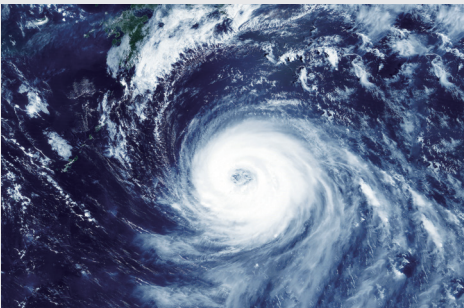


[그림 5] 2018년-2021년 평균 온도 분포에 기반한 1.5도 상승 시기 예측의 상대적 중요도. 빨간색 계열은 해당 지역이 1.5도 상승에 주요하게 기여하고 있음을 의미하며, 파란색 계열의 색은 해당 지역이 1.5도 상승과 관련된 시그널이 뚜렷하지 않음을 의미.

구축된 인공지능 기법은 상대적으로 간단한 구조를 가지고 있기 때문에 층별 기여도 전파 기법 (Layerwise Relevance Propagation) 을 통하여 지구 온난화와 관련된 시그널이 어느 지역에 오는지 손쉽게 추적할 수 있다 [그림 5].

예를 들어, 남서 아프리카와 인접한 해양 지역에서의 대조적인 온도 시그널은 이 지역의 육지-해양 온도 차이가 지구 온난화와 밀접하게 관련이 있음을 의미한다. 또한, 적도 동태평양의 온도 변동은 지구 온난화 시그널과 크게 관련이 없는 것으로 나타나는데, 이는 해당 지역이 엘니뇨에 의해 주도되는 큰 내부 변동성을 가진 지역으로 지구 온난화로 대표되는 외부 강제력에 의한 온도 변화의 효과는 크지 않을 것임을 의미한다.

대기, 해양, 지면, 해빙, 및 식생 등의 지구 시스템 요소들은 에너지, 탄소, 물을 교환하며 서로 상호 작용하며 복잡성 및 비선형성을 만들어낸다. 기존의 프로세스 기반 기후 예측 모형들은 이를 관측하고, 물리적으로 이해한 후 수치적으로 풀어내야 하기 때문에 그 발전 속도가 더딜 수밖에 없다. 한 편, 최근의 다양한 지구 시스템 요소의 관측과 급격한 컴퓨팅 파워의 증가는 인공지능 기반 기후 모델의 성능을 향상시키고 있다.



이런 식이라면 머지 않은 미래에 인공지능 기반 기후 모델이 프로세스 기반 기후 모델을 완전히 대체하지 않을까 하는 의문이 든다. 하지만 앞선 예에서 보였듯, 아직까지는 프로세스 기반 모형의 시뮬레이션 결과 없이는 인공지능 기반 기후 모델을 학습 시키기 위한 충분한 샘플 수를 확보하기 어렵다. 이는 한 쪽이 다른 한 쪽을 넘어서기 보다는 서로 상호 보완하는 방향으로 모델이 개발될 것을 의미한다. 또한, 인류가 지구 시스템을 이해하는 것이야말로 궁극적으로 지구 시스템의 변화로 인한 문제를 해결할 수 있는 단초를 제공할지도 모른다는 점에서, 프로세스 기반의 기후 모델의 개발은 필수적이다.



다만, 인류의 무분별한 개발과 욕망으로 인해 우리에게 주어진 시간이 많지는 않다. 즉, 최근의 급격한 기후 변화는 인류가 지구 시스템을 이해하기 위한 시간을 무한정으로 허용치 않고 있다. 기후 학자들은 인류가 기후 변화와 관련된 당면한 문제를 십수년 안에 해결해야 하며, 해결하지 못할시 적어도 앞으로 어떤 일들이 벌어질 지에 대해서 예측하고 대비책을 수립해야 한다고 경고한다.

데이터 기반의 인공지능 기후 모델은 인류가 당면한 기후 변화 문제를 해결할 수 있는, 아니, 적어도 대비할 수 있는 방향을 제시할 지도 모를 일이다.



# 통계로 바라보는 세상이야기

신동헌 | 도서출판 지일북스 대표

## e-지방지표로 주제별 통계자료 확인하기!

통계청은 여러분이 주제별로 쉽게! 통계자료를 살펴볼 수 있도록 하는 e-지방지표 서비스를 제공하고 있습니다. KOSIS의 2022 학생의 학교생활 만족도에 따르면, 지역별 재학생의 전반적인 학교생활에서 광주가 ‘매우 만족한다(25.9%)’ 및 ‘약간 만족한다(37.1%)’의 합에서 가장 높은 지역으로 나타났습니다. KOSIS의 2020 일반폐기물 재활용률에 따르면, 2020년 기준 폐기물이 재활용되는 비율이 가장 높은 지역은 바로 ‘부산(80.4%)’입니다. 통계표 서비스의 장점은 오름차순, 내림차순 필터링이 가능하고, 해당 표를 클릭하면 색 처리가 되어 헷갈리지 않는다는 점이고, 시각화 서비스의 장점은 전국, 최소 및 최대 건수나 비율을 한 번에 제공하고, 타임라인 기능으로 변화 양상을 쉽게 볼 수 있습니다.

## 한국어, 한류와 함께 세계로!

“aegyo(애교), japchae(잡채), banchab(반찬)” 옥스퍼드 영어사전에 등재된 한국어 단어들인데요, 2021년 9월에만 무려 26개의 한국어 단어가 옥스퍼드 영어 사전에 등재되었다고 합니다. 1884년부터 2021년 8월까지 137년간 등재된 한국어 단어는 모두 합쳐 23개라고 하니, 한 달에 26개가 얼마나 많은 수인지 실감이 나죠! 한국국제교류재단에서 108개국의 한류 동호회 현황을 분석한 통계를 보면, 2017년에 6,099만 명에서 2021년 15,660만 명으로 전 세계 한류 동호회 회원 수는 9,561만 명 증가했습니다. 2020년 2월 발표된 자료에서 한국어를 사용하는 인구는 총 7,730만 명으로 무려 14위를 차지했고, 국립국어원의 한국어-외국어 학습 사전 이용자도 2021년에는 10월까지 무려 363만 2339명이라고 합니다.

## 종이에서 포털로, 요즘 사람들의 신문읽기

신문하면 대표적으로 떠오르는 이미지는 아직도 종이신문이지만, 이제는 휴대폰으로 인터넷 포털에 접속해 기사를 보는 모습이 훨씬 익숙합니다. 한국언론진흥재단의 2021 언론수용자조사를 보면, 종이신문 뉴스 이용률은 2017년 16.7%에 비해 2021년 이용률은 8.9%로 7.8%p 하락한 반면, 인터넷 포털 이용률은 2017년 73.0%에서 2021년 79.2%로 6.2%p 상승했습니다. 한국언론진흥재단의 <2021 신문을 말한다:이용자 FGI를 통해 본 종이신문의 오늘과 내일> 조사에 따르면, 종이신문 이용자들이 생각하는 인터넷 신문 대비 종이신문의 장점으로는 종합적이고 선별되었으며 믿을 수 있는 정보를 제공하고, 이용의 다양성과 강제성, 형식의 규격과 가독성 및 집중과 기억에 도움을 준다고 설명하였습니다.

## 결혼은 선택! 연애도 안 하는 이유는?

최근에는 결혼이란 꼭 해야 하는 것이 아닌 개인의 선택 중 하나로 여겨지고 있는데요. 특히 1030세대에서 뚜렷하게 나타난다고 합니다. 통계청이 발표한 ‘2022년 혼인·이혼 통계’에 따르면 2022년 혼인 건수는 19만 2천 건으로 전년 대비 0.4% 감소한 수치를 보였습니다. 이는 통계 작성이 시작된 1970년 이후 가장 낮은 수치라고 합니다. 또한, 인구보건복지협회가 2022년 7월 만 19~34세 비혼 청년 1,047명을 대상으로 실시한 ‘2022년 제1차 저출산 인식조사’에선 65.5%가 현재 연애를 하지 않고 있으며, 이 중 3분의 1 가까운 인원(29.1%)이 연애를 한 번도 해보지 않았다고 답했고, ‘비연애’ 상태라 답한 이들 중 70.4%는 스스로 원해서 연애를 하지 않고 있다고 밝혔습니다.

## 고등학생 사교육비 최다, 대한민국의 사교육 현주소

“유난이 유난을 낳고, 유난이 집착이 되고, 집착은 또 경쟁을 낳고…” 입시 전생 속에서도 호롱불 같은 따스함을 지닌 드라마를 꿈꿨다는 양희승 작가의 <일타스캔들>이 종방하였는데요, 과열된 우리사회의 교육열을 반영한 설정에 몰입되었다는 시청자들의 반응이 컸다고 합니다. 2022년 통계청의 ‘초중고사교육비조사’ 결과에 따르면, 2022년 사교육비 총액은 약 26조 원으로 전년대비 10.8%(2조 5천억원) 증가했고, 참여율(78.3%)과 주당 참여시간(7.2시간) 역시 전년대비 각각 2.8%p, 0.5시간 늘었으며, 초중고 전체학생의 1인당 월평균 사교육비는 41만원으로 전년대비 11.8% 상승하였는데, 초등학교는 37.2만원, 중학교는 43.8만원, 고등학교는 46만원으로 조사되었습니다.

## 마스크는 벗었지만 학교생활만족도는 감소

코로나19 이후 4년 만에 실내 마스크 착용 의무가 해제되었고, 매일 아침 진행하던 체온측정과 급식실 칸막이도 사라졌지만, 지난해 11월 발표된 ‘2022년 사회조사 결과’를 보면 학교생활에 만족하는 학생들의 비율이 51.1%로 59.3%를 기록했던 2년 전보다 8.2%p 감소한 것으로 나타났습니다. 성별로 보면 남학생은 52.4%, 여학생은 49.6%로 남학생의 만족도가 여학생보다 약간 높았습니다. 2022년 중·고등학교에 다니는 학생들이 꼽은 공부하는 이유는 미래의 나를 위해 필요해서가 79.7%로 가장 높았고, 공부를 못하면 부끄럽기 때문에(32.5%)가 그 뒤를 이었습니다. 고등학교 졸업 후 창업을 진로 계획으로 세운 비율도 2015년 1.0%에서 2022년 2.9%로 점차 증가하는 추세를 보였습니다.

## 취학을 세계 최고 수준 대한민국! 아시아 국가들은?

대한민국의 학제는 6-3-3으로 운영 중이며, 2019년 초등학교 취학률은 101.0%이며, 대학 취학을 또한 98.4%로 세계적으로도 높은 수준을 보여주고 있습니다. 중국 역시 한국과 같은 6-3-3년 학제를 실시 중인데요, 2020년 중국의 초등학교 취학률은 103.2%, 대학교는 58.4%를 기록했고, 베트남의 학제는 5-4-3인데, 초등학교에 앞서 유치원 3년(3~6세)을 정규교육과정으로 분류한 영향으로 베트남의 유치원 취학률은 2020년 92.8%로 높은 수준을 보이고 있으며, 등록학생 인구만 2억 5천만 명에 달하는 교육대국 인도는 현재 기존의 10-2 학제에서 5-3-3-4의 학제로 개편되는 와중에 있는데요, 2020년 기준 인도의 초등학교 취학률은 99.9%, 대학교는 29.4%를 기록했습니다.

## 위드코로나 시대! 국민 삶의 질은 어떻게 달라졌을까?

2023년 2월 20일 통계청 통계개발원에서 발간된 ‘국민 삶의 질 2022 보고서’에 따르면 2022년 수치가 반영된 20개 지표 중 고용이나 가족, 건강, 세부 환경영역별 만족도는 개선된 반면 독거노인, 선거투표율, 학교생활만족도, 기후변화불안도 등의 4개 지표는 악화된 것으로 나타났습니다. ‘21년 1인당 여행일수(국내)는 6.58일로 ’20년(5.81일) 대비 소폭 증가한 것으로 나타났고, ’22년 미세먼지 농도는 17.9μg/m³로 ’21년보다 0.8μg/m³ 감소하였습니다. ’21년 비만율은 37.1%로 ’20년(38.3%)보다 감소하였으나, 여전히 코로나19 이전의 35% 미만에 비하면 높은 수준을 보이고 있는데요, 코로나19 시기의 재택근무와 원격 학습, 외부 활동의 제약으로 인한 활동량 감소가 원인으로 추정됩니다.

## 대한민국은 접속 중, 우리는 지금 인터넷 시대!

1982년, 현재 한국전자통신연구원의 전신인 한국전자기술연구소 컴퓨터 개발실에서 컴퓨터와 컴퓨터 간 원거리 네트워크 교신을 최초로 성공하면서 대한민국 인터넷 산업이 시작되었습니다. 2021년 기준 대한민국의 인터넷 이용률은 93.0%입니다. 특히 2000년 49.8%였던 우리나라의 인터넷 접속률은 2001년에는 63.2%, 2002년에는 70.2%, 2005년에 92.7%, 이후 계속 90% 이상을 기록했고 2021년 인터넷 접속률은 99.9%였습니다. 2021년 인터넷이용실태조사 결과에 따르면 대한민국에서 인터넷을 이용하는 사람들의 96.3%는 하루에 1회 이상 인터넷을 이용하고 있으며, 최근에는 10대부터 60대까지 전 세대에 걸쳐 90% 이상의 모바일 인터넷 이용률을 보인다고 있다고 합니다.

## 봄 나들이 방해꾼, 춘곤증! 졸음운전 주의보!!

졸음운전은 혈중알코올농도 0.17%의 만취 상태에서 운전하는 것과 비슷하다는 결과가 있습니다. 특히 졸음운전으로 인한 사고는 치사율이 4.51%로 음주운전 치사율보다 1.75배나 높은 것으로 나타나기도 했는데요. 우리나라의 교통사고 사망자 수는 꾸준히 감소하고 있는 것으로 나타났습니다. 경찰청의 월별 졸음운전 교통 사망사고 통계에 따르면 2015년부터 2019년까지 5년간 졸음운전으로 인한 교통사고 사망자 수는 전체 437명으로 집계되었는데요. 이 중 5월이 11.9%(52명)로 가장 높았습니다. 춘곤증과 함께 찾아오는 졸음운전 예방법을 알아볼까요? 먼저, 30분에 한 번씩 환기하고, 과식보다는 틈틈이 간식으로 졸음을 떨쳐보는 것도 방법이며, 장거리 운전에는 2시간에 한 번씩 쉬는 것도 잊지마세요.

## 아름다운 봄의 전령 꽃! 세계인들이 꽃을 사랑하는 방법

2021년 우리나라 국민 1인당 연간 화훼 소비액은 12,386원 수준으로 나타난 반면, 2016년 해외의 1인당 소비액을 살펴보면 스위스 185,000원, 네덜란드 110,000원, 일본 57,000원 등으로 우리나라보다 비교적 높은 수준이었습니다. 한국농촌경제연구원의 화훼류 한일 소비행태 분석 보고서에 따르면 국내 소비자는 마음 전달을 위한 선물용(38.7%)이나 결혼이나 개업·장례 등 행사용 화환(22.9%)이 꽃 구매의 주된 목적인 반면, 일본은 선물용(24.0%), 가정 장식용(23.1%), 성묘용(20.1%), 성불용(16.0%)의 순으로 상당 부분이 생활이나 가정용 소비 중심인 것을 확인할 수 있었습니다. 올해 전국 대부분 지역에서 3월 중순부터 봄꽃이 피기 시작해 4월 중순께 절정에 이를 전망입니다.

## 혹시 나도 스마트폰 중독? 현대인의 신종 질병에 대하여

통계청 생활시간조사(2019)에 따르면 개인의 하루 평균 스마트폰 사용시간은 1시간 31분이었습니다. 이것을 1년으로 계산하면 약 33,215분(약 553시간)이죠. 우리가 80세까지 산다고 가정했을 경우, 20세부터 80세까지 평생 1,407일은 스마트폰만 보고 있다는 것입니다. 2021년 기준 전체 스마트폰 이용자 중 24.2%는 과의존위험군에 해당하는 것으로 나타났습니다. 특히 10대 청소년의 비율이 37.0%로 가장 높았습니다. 2021년 환자 수가 가장 많았던 질병은 ‘일자목증후군’으로 2020년보다 늘어난 1,550,082명이었습니다. 불면증도 658,675명(2020년)에서 684,560명(2021년)으로 증가했으며, 손목터널증후군 역시 166,094명(2020년)에서 169,384명(2021년)으로 증가했습니다.

## 누구나 한 번쯤은 취준생이 된다!

통계청 ‘경제활동인구 조사’에 따르면, 취업 준비자 수는 2019년 748,000명에서 2021년 841,000명으로 3년 동안 꾸준히 늘었습니다. 잡코리아 ‘취업 준비 시 정보 습득 현황(2022)’ 설문조사에 따르면, 취업준생의 84.5%는 취업 정보 수집에 어려움을 느끼고 있다고 응답했는데요, 특히 ‘지원 기업의 조직문화 및 근무 분위기’(26.6%), ‘실제 급여 수준’(23.8%), ‘재직자들이 평가하는 근무 만족도’(21.3%) 순으로 나타났고, 취업준생들이 꼭 필요하다고 생각하는 정보로는 ‘연봉, 실 급여 수준’(31.8%), ‘직무에 대한 구체적인 정보 및 자격요건’(26.2%), ‘채용 일정 및 채용 규모’(23.3%), ‘합격 스펙 및 합격 자소서 등 참고자료’(15.3%) 등으로 응답하였습니다.

## 나는 몇 살까지 살 수 있을까? 2021 생명표로 예상해보자!

생명표란, 현재의 연령별 사망 수준이 유지된다면 특정 연령의 사람이 향후 몇 세까지 살 수 있는지 추정한 통계 표입니다. 지난해 12월 6일 2021년 생명표가 발표되었는데요, 2021년 연령별 사망확률에 따르면 2021년 출생아의 기대수명은 83.6년입니다. 1970년과 비교했을 때 21.3년 늘어난 수치로, 매년 증가하는 추세를 보이고 있습니다. 또한 남자의 기대수명은 80.6년, 여자는 86.6년으로 기대수명 차이는 1985년 8.6년으로 최고점에 도달한 후 계속 감소하여 2021년에는 6.0년으로 예측되었습니다. 2021년 현재의 기대수명은 대한민국이 OECD 평균보다 남성 2.9년, 여성 3.5년 더 높은데요, 가입했던 1996년에는 OECD 평균 기대수명보다 남자 2.3년, 여자 0.9년 낮았다고 합니다.

## 어서오세요, 한국방문 반가워요~

2019년 9월 한국을 방문한 외국인인 145만 9,664명이었는데요, 코로나19 팬데믹이 선언된 2020년 3월 이후에는 수가 급감하다가 2022년 5월 한국을 방문한 외국인인 17만 5,922명이었고, 10월에는 47만 6,097명으로 차츰 회복되고 있습니다. 2021년 11월부터 2022년 10월까지 조사한 결과, 관광 목적(53.6%)이 가장 많았고, 개인적인 이유(36.6%), 유학연수(6%) 순으로 나타났으며, 성별과 연령은 남성의 경우, 주된 연령대는 31~40세(12.5%)였고, 그 다음으로 14.1%가 21~30세 남성이었습니다. 여성은 21~30세(14.5%)가 가장 많았고, 31~40세(9.1%)가 그 뒤를 이었습니다. 방문객의 국적은 1위 미국(45만 2,363명), 2위 중국(19만 9,980명), 3위 필리핀(17만 4,129명) 순으로 나타났습니다.

## 남북한의 산업구조, 이렇게 다르다고?

북한통계포털(<https://kosis.kr/bukhan/>)을 보면 2021년 북한의 인구는 2,548만 명으로 남한(5,175만 명)의 절반 수준으로 나타났고, 북한의 명목 국민총소득(GNI)은 142.3만 원으로 남한(4,048.2만 원)의 3.5% 수준으로 남북한의 소득 격차는 확대되고 있습니다. 2021년 북한의 무역총액(남북교역 제외)은 7.1억 달러로 전년보다 17.3% 감소하였으며, 남한의 무역총액 12,595억 달러의 1/1,766배(0.1%) 수준이었습니다. 북한의 2021년 산업구조는 서비스업이 32.9%로 1위, 농림·어업은 23.8%로 2위, 제조업이 18.3%로 3위를 차지하였고, 2021년 북한의 수출액은 0.8억 달러로 2020년 수출액 0.9억 대비 8.2% 감소하였고, 수입액은 6.3억 달러로 2020년 수입액 7.7억 대비 18.4% 감소하였습니다.



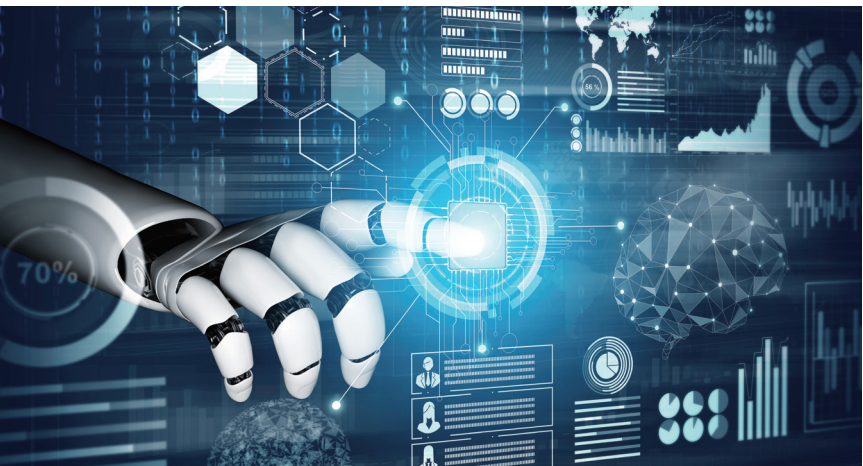


# 「2022 개정 수학과교육과정」에서 나타난 실용 통계교육의 방향

고은성 | 전주교육대학교 교수

한국에서 수학교육학이 태동하던 1990년대에만 해도 대수나 기하, 함수 교육에 비해 통계교육은 한국의 수학교육 연구자들이 그다지 관심을 기울이지 않았다. 그러나 2000년 “우리나라의 통계 교육은 실세계의 이해와 예측을 위한 자료 분석 도구로서 활용되는 실용 통계를 가르치기보다는, 인위적인 예를 통한 단순 자료 정리 기법과 통계치의 계산, 그리고 확률분포 이론과 같은 수학을 가르치는 데 주력하고 있다.”는 비판이 제기되면서 통계교육 개혁의 필요성에 공감대가 형성되었다.

또한 사회적으로 급속하게 생산되어 여러 형태로 변화되는 방대한 데이터 속에서 다양한 패턴과 경향을 읽어내고 가치를 발견하는 빅데이터 분석의 중요성이 부각되면서 시와 빅데이터의 4차 산업혁명 시대에 통계의 중요성이 증대되었다. 통계교육의 변화에 대한 바람은 더 이상 수학교육계만의 문제가 아니게 되었다.



2022 개정 수학과교육과정에서는 그 동안 학교수학에서의 통계교육에 대한 비판과 사회적 바람을 담고자 하였다. 2022 개정 수학과교육과정에서 통계교육의 주요 변화는 다음과 같다.

## 첫째, 고등학교 교육과정에 융합 선택 과목으로 <실용 통계> 신설

그 동안 고등학교 통계교육은 <확률과 통계> 과목을 통해 이루어졌는데, 교과서에 제시된 통계 관련 개념들을 소개하고 제시된 문제들을 해결하는 방식의 수업을 통해 이루어져 왔다.

그러나 통계를 수학의 하위 영역으로 인식하여 학교 수학에서 다루어오던 전통에서 벗어나지 못한 채 활동과 탐구보다는 개념과 원리 중심으로 교과서가 구현되어 있어 정규 수업 시간을 통해 실용 통계를 경험하기가 매우 어려운 상황이었다.

2022 개정 수학과 교육과정에서는 <실용 통계> 과목의 성격을 ‘통계적 문제해결 과정을 이해하고, 통계적 탐구 활동을 통해 실생활 문제를 해결하는 과목’으로 제시하고 있다. <실용 통계>에서 학습한 내용은 정보화 사회에서 생산되는 자료가 인류를 이해하고 미래를 개척하는 주요한 자산이 됨을 이해하고 이를 활용하여 현대 사회의 다양한 문제를 해결하는 데 도움이 되어야 하며, <실용 통계>를 학습한 학생들은 일상에서 접하는 문제를 통계적으로 해결하거나 의사 결정에 필요한 자료를 수집, 분석하여 합리적인 결론을 내리는 구체적인 경험을 할 수 있어야 한다.

또한 현대 사회에 나타나는 변이성을 이해하는 지식을 습득하고 통계적 문제해결 과정에 요구되는 기능을 형성하며 자연 및 사회 현상을 해석하는 도구로서 통계의 유용성을 인식할 수 있어야 한다.

자신의 진로와 적성에 따라 실생활에 유용한 통계를 학습하고 여러 교과를 융합하여 실생활 문제를 해결하는 역량을 기르고자 하는 학생들은 <실용 통계>를 선택할 수 있다.

또한, <실용 통계>에서 학습한 통계의 지식과 기능, 통계적 문제해결 경험은 자연과학, 공학, 의학뿐만 아니라 경제 경영학을 포함한 사회과학, 인문학, 예술 및 체육 분야를 학습하는 데 토대가 되며, 나아가 자료 수집, 분석, 해석 역량을 갖춘 창의적인 사람으로 성장할 수 있는 기반을 제공하는 것을 목표로 한다.



<실용 통계>에서는 범교과 학습 주제(안전 건강, 인성, 진로, 민주 시민, 인권, 다문화, 통일, 독도, 경제 금융, 환경 지속가능발전 등)를 현상이나 소재로 선택하여 활용함으로써 통계의 유용성과 필요성을 인식할 수 있도록 한다.

## 둘째, 통계적 문제해결 교육 강조

먼저, 중학교 교육과정에서는 내실있는 통계적 문제해결 교육의 토대를 마련하였다. 중학교 통계교육의 주요 내용은 다음과 같다.

‘다양한 맥락에서 해결하고자 하는 통계적 탐구 문제를 설정하고 적절한 계획을 세워 자료를 수집하게 한다. 수집한 자료를 자료의 특성과 목적에 맞게 표, 그래프, 수치 등으로 나타내어 분석하고, 그 결과를 탐구 문제와 연결하여 해석하게 한다. 수집한 자료나 분석 결과가 적절한지 판단하여 계획을 수정하고, 통계적 근거를 바탕으로 토론하는 등 통계적 문제해결 과정에 주도적으로 참여하게 한다.’

통계적 탐구 문제 설정부터 자료 수집과 정리, 탐구 문제와 연계한 결과의 해석뿐만 아니라 계획의 수정까지 학생들이 통계적 문제해결을 피상적으로 경험하는 것이 아니라 주도적 참여를 통해 내실있는 교육이 이루어질 수 있도록 하였다.

초등학교 교육과정 역시 ‘해결하고자 하는 문제를 설정하고 그에 맞는 자료를 수집, 정리하여, 막대그래프, 꺾은선그래프, 띠그래프와 원그래프 중 적절한 그래프로 나타내고 해석하는 일련의 과정을 직접 경험하게 한다.’고 제시하면서 통계적 문제해결 전 단계의 경험을 강조하고 있다.

## 셋째, 공학도구의 적극적인 활용 장려

교육과정에서는 ‘막대그래프와 꺾은선그래프, 띠그래프와 원그래프를 그릴 때 공학 도구를 사용하게 할 수 있다.’, ‘자료를 수집하고 정리하여 표나 그래프로 나타내거나 대푯값과 산포도를 구할 때 공학 도구를 이용할 수 있게 하고, 공학 도구의 편리함과 유용성을 인식하게 한다.’고 제시하면서 자료수집에서 정리까지 공학도구를 적극적으로 활용할 것을 장려하고 있다.

공학 도구는 복잡한 계산과 절차적 처리가 문제해결과 학습의 본질적인 부분이 아닐 때 이를 신속하게 대행함으로써 사고력 중심의 교수·학습을 가능하게 한다는 장점이 있다. 따라서 통계교육의 경우, 실생활 중심의 내용을 바탕으로 탐구하는 실용 통계 교육을 실천하기 위해서는 공학 도구의 사용은 필수적이다.



실용 통계 교육에서는 주어진 자료를 계산하여 평균을 정확히 계산해 내는 것보다 자료의 분포를 바탕으로 어떤 대푯값이 모집단을 분명하게 대표할 수 있는지를 판단하는 것이 통계적으로 더 중요한 활동이다. 자료를 수집하거나 정리하는 활동, 그래프를 그리는 활동 등의 기계적이고 절차적인 활동을 공학 도구가 대행해줌으로써 통계교육의 본질에 해당하는 활동에 좀 더 집중할 수 있게 된다.

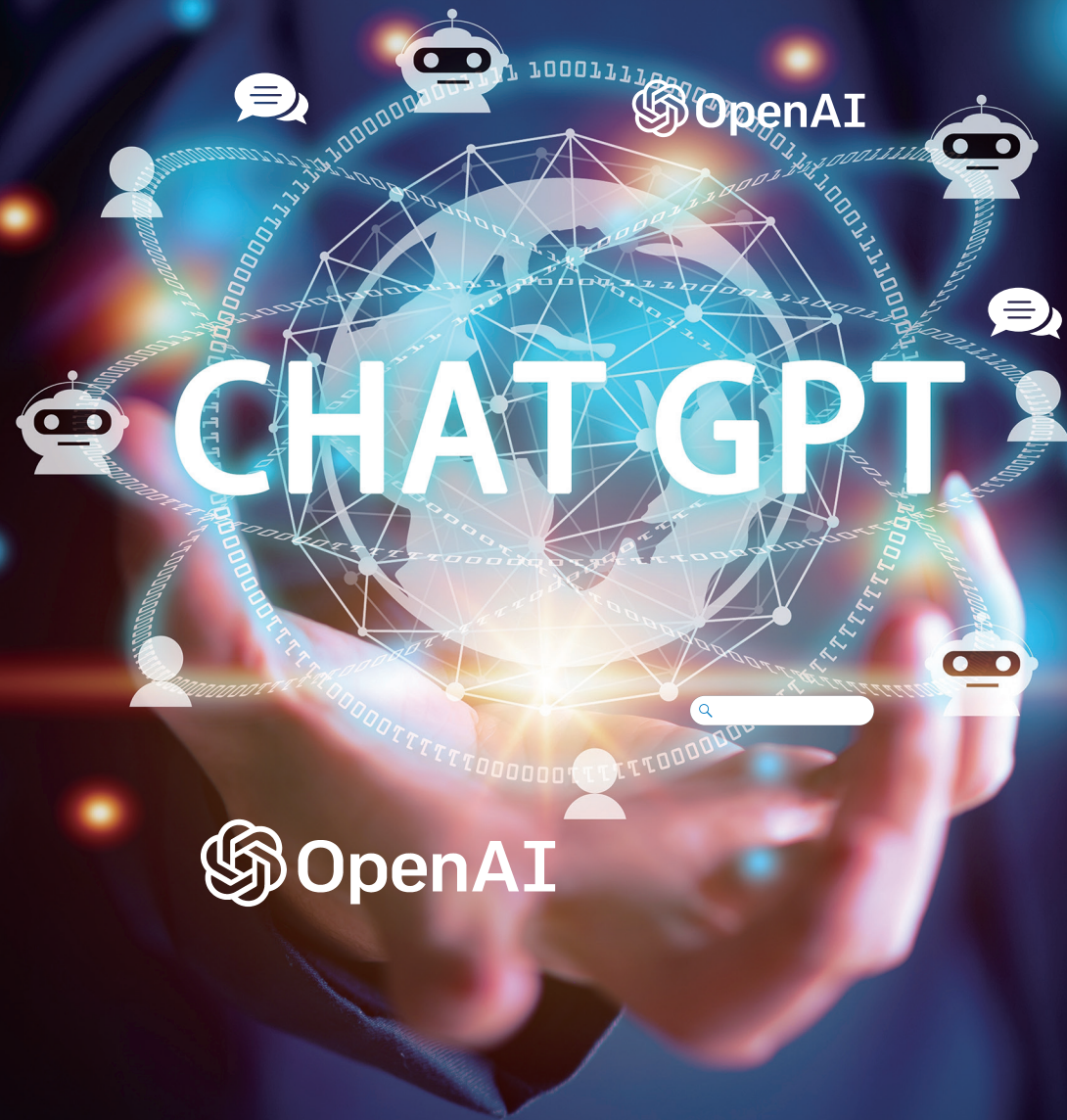
## 넷째, 합리적 의사결정, 사회를 이해하는 도구로써 통계의 중요성 강조

‘환경, 지속가능한 발전 등 범교과 학습 주제를 소재로 다루고, 이를 탐구하는 과정에서 체계적으로 사고하고 합리적으로 의사 결정을 할 수 있게 한다.’, ‘진로연계교육을 실시할 때는 학생의 흥미, 관심, 진로에 맞는 탐구 문제를 설정하여 통계 프로젝트를 수행하게 할 수 있다.’고 제시하여 통계적 탐구 주제를 타교과의 학습 주제와 연결하거나 환경, 자연현상, 사회문제 등에서 선택할 수 있도록 독려하고 있다.

이는 합리적 의사결정 도구로써 통계의 유용성뿐만 아니라 사회 현상을 파악하고 사회를 이해하는 도구로써 통계의 유용성을 인식할 수 있게 해준다.

이번 교육과정의 변화로 내실있는 통계적 문제해결 교육이 이루어질 수 있을 것이며, 공학적 도구의 활용이 확대됨으로써 통계적 사고 및 소양을 강화하는 활동에 집중할 수 있는 토대가 마련되었다.





## 분석은 질문이다...챗GPT에게 묻다

강양석 | Deep Skill 대표

챗GPT로 인해 우리는 특이점(과학 기술의 항구한 가속적 발전으로 인해 인류 역사에 필연적으로 발생할 가능성이 높은 획기적인 변곡점)을 경험하고 있다. 이 인공지능을 경험한 모든 사람은 하나같이 '세상이 정말 빨리 변하고 있어서 정신이 없어요.'라는 말을 하고 있으니 말이다. 더한 혁신은 계속해서 나오고 있지만, 분명 그 경험의 강도는 스마트폰이 처음 생겼을때의 충격을 가볍게 넘어서고 있는듯 하다.

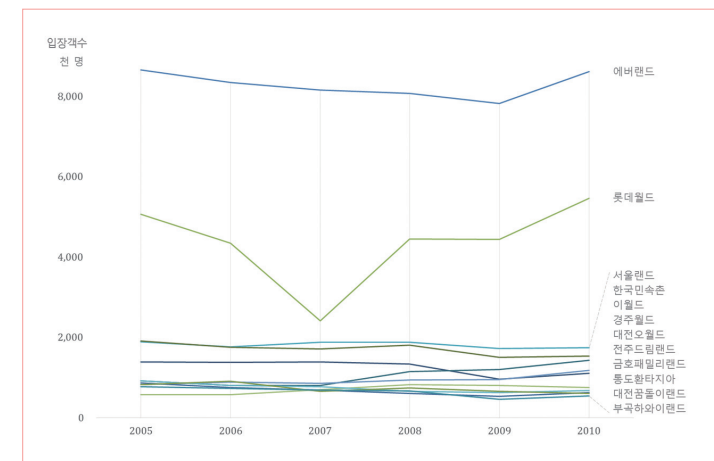


잘 알겠지만, 챗GPT는 언어를 학습한 언어 인공지능이다. 들리는 소식에 의하면 가장 많이 학습한 언어는 영어, 그 다음으로는 (추정컨데) 파이썬이라고 한다. 파이썬? 아마 많이 들어봤을 단어이다. 이는 데이터 분석 도구임과 동시에 컴퓨터 언어이기도 한 것이다. 그렇다면 챗GPT는 단순히 채팅 인공지능을 넘어 분석 도구로서의 성능 또한 뛰어날 것이라는 예상을 해볼 수 있는데, 아니나 다를까 매우 그렇다. 그래서, 챗GPT와 몇몇 분석을 해가며 과연 우리는 어떤 데이터 역량에 집중해야 하는지를 정리해 보았다.



### 사례 1.

챗GPT는 얼마나 통찰적인 데이터 해석을 할 수 있는가?



첫 번째로 얼마나 통찰적인 데이터 해석을 할 수 있을지를 물어봤다. 다음의 데이터를 주며 '국내 테마파크 산업의 특징은 무엇인가요?'라고 물어보며 말이다. 그리고 아래와 같은 답을 얻을 수 있었다.

OW 아래 차트를 보고 이 당시 한국의 테마파크 산업의 특징을 유추해주세요  
https://images.app.goo.gl/oCnLgP1YPPeo4cMy8

주어진 차트는 2018년 한국 테마파크 산업의 현황을 나타내는 것으로 보입니다. 이를 토대로 한국의 테마파크 산업의 특징을 살펴보면 다음과 같습니다.

1. 테마파크 수 증가: 2018년 기준 한국에는 20개의 테마파크가 운영되고 있습니다. 이는 2013년 대비 6개의 증가된 수치입니다.
2. 서울 근교에 집중: 한국의 대부분 테마파크는 수도권 근교에 위치해 있습니다. 서울과 인천을 중심으로 9개의 테마파크가 위치해 있으며, 나머지 11개의 테마파크는 부산, 대전, 전주, 대구 등 각 지역별로 분포되어 있습니다.
3. 대형 테마파크 우세: 대부분의 테마파크가 대형 규모로 운영되고 있습니다. 대표적인 예로는 에버랜드와 롯데월드 등이 있습니다.
4. 해외 프랜차이즈 적극 도입: 해외에서 유명한 테마파크 브랜드들이 한국에 진출하여 적극적으로 도입되고 있습니다. 디즈니, 유니버설, 롤러코스터 등이 대표적인 예입니다.
5. 다양한 이벤트 및 쇼 제공: 테마파크에서는 다양한 이벤트 및 쇼가 제공되고 있습니다. 예를 들면, 크리스마스, 할로윈, 봄 축제 등 계절마다 다양한 이벤트가 진행됩니다.
6. 가족 대상 적합: 대부분의 테마파크가 가족 대상으로 적합하게 구성되어 있습니다. 어린이부터 어른까지 즐길 수 있는 놀이기구 및 쇼 등이 다양하게 제공됩니다.

챗 지피티 대화날짜 23.04월, version 4.0 기준

나는 똑같은 질문을 약 3만여명의 사람들에게 물어본 경험이 있는데, 매우 흥미로운 점은 사람의 답변과 챗GPT의 답변이 거의 유사하다는 것이다. 당연히 사람의 말과 글을 학습한 인공지능이기 때문에 이렇게 답하는 것이 당연할지 모르겠지만, 문제는 지금부터 발생한다. 그럼 우리 인간은 어떤 데이터 학습을 해야 할 것인가 말이다. 즉, 데이터에 버젓이 드러난 사실을 캐내는 방식의 해석 능력은 이제 안타깝게도 기계에 그 자리를 내줘야 한다는 것을 의미한다. 즉, 우리의 데이터 학습의 방향은 ‘챗GPT가 할 수 있는 것과, 챗GPT가 할 수 있는 것을 구분’하는 데에서부터 시작된다.

그럼, 챗GPT가 할 수 없는 해석이란 어떤 것일까? 예를 들면 이런 답변 말이다. ‘어? 2007년 롯데월드의 입장객 추이가 심각하게 하락했지만, 에버랜드가 전혀 미동도 하지 않네요? 아마 둘이 서로 경쟁관계가 아닌가봐요?’라는 이런 통찰 말이다. 기계는 의심을 하지 않지만, 사람은 한다. 다만, 이런 의심이 ‘사실인 것도 아니고, 사실이 아닌 것도 아닌 중간 지점’

에 걸쳐 있지만 이런 모호한 상태의 의심과 호기심을 두껍게 가지고 있어야 한다. 그래야만, 같은 데이터를 보고 더 많은 사실과 통찰에 접근할 수 있기 때문이다. 그리고, 이렇게 데이터를 보고 단순히 숫자를 획득하는 것이 아니라, 더 많은 의심과 추론으로 더 넓은 사실성의 세계로 스스로를 안내하는 해석 습관은 절대 기계에 대체 되지 않는다. 이제 기계와 동급이거나 대체될 만한 역량에 크게 힘을 기울일 필요가 없다는 것이다.



## 사례 2.

### 챗GPT로 분석할 때 정말 중요한 능력은 무엇일까?

이번에는 챗GPT와 좀 더 고급진 분석을 시도해 보았다. 분석데이터의 생성, 분석, 그리고 의미 해석까지 전 과정을 함께 했으니 말이다. 단순히 코드를 간결하게 압축하던던지, 새로운 코드를 짜 주는 선을 넘어 함께 멍킹파트너(thinkingpartner : 생각동무)의 역할을 어느정도까지 해주지가 알고 싶었다.

그 첫번째로, 간단한 회귀 분석을 위한 데이터 생성을 시도하였다. 다행히도 질문자의 의도를 정확히 알아듣고 원천 데이터를 스스로 생성하는 모습을 볼 수 있었다. 매우 놀라운 일이다. 이런 방식대로라면 어떠한 요건을 주어주고 그 요건에 맞는 훈련 데이터를 얼마든지 스스로 챗GPT가 생성할수 있다는 것을 의미하므로, 우리는 간단한 분석 기법을 연습할때는 챗GPT의 생태계 내에서 기승전결을 모두 해결할 수 있다는 것을 의미한다.

섬세하게 자신이 생성한 원천데이터의 요약까지 해주는 모습을 볼 수 있다.

OW 80명의 학생에 대한 가상데이터를 만들어주세요. 80명에 대한 영어와 수학 점수를 랜덤으로 생성하고, 합격/불합격 여부도 나타내주세요. 마지막으로 이 데이터의 첫 6개 행만 보여주세요.

네, 아래의 R 코드를 사용하여 80명의 학생에 대한 가상데이터를 만들 수 있습니다. 영어와 수학 점수는 정규분포에서 무작위로 생성하고, 합격/불합격 여부는 각 학생의 수학과 영어 점수의 총합에 따라 결정됩니다. 전체 데이터 프레임의 만든 후, 'head()' 함수를 사용하여 첫 6개 행을 출력합니다.

```
R
# Set the seed for reproducibility
set.seed(123)

# Generate random math and English scores for 80 students
math <- round(rnorm(80, mean = 70, sd = 10), 1)
english <- round(rnorm(80, mean = 75, sd = 10), 1)

# Calculate the total scores and pass/fail status
total <- math + english
passfail <- ifelse(total >= 120, "Pass", "Fail")

# Create a data frame with the scores and pass/fail status
data <- data.frame(Math = math, English = english, Total = total, PassFail = passfail)

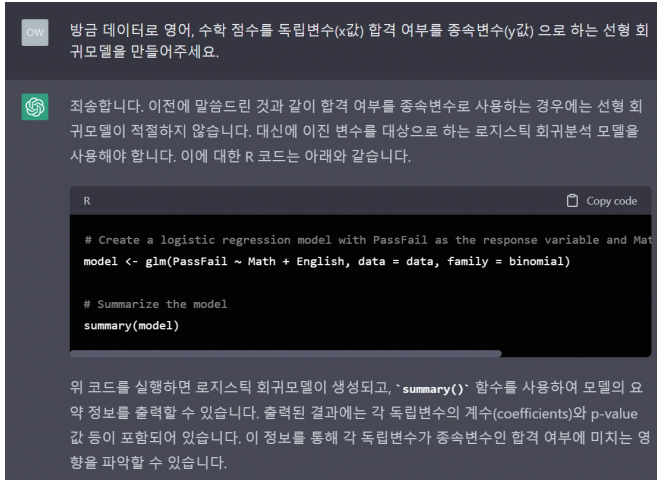
# Show the first 6 rows of the data frame
head(data, n = 6)
```

위 코드를 실행하면, 80명의 학생에 대한 가상 데이터가 만들어집니다. 'head()' 함수를 사용하여 첫 6개 행을 출력하면 아래와 같은 결과가 나타납니다.

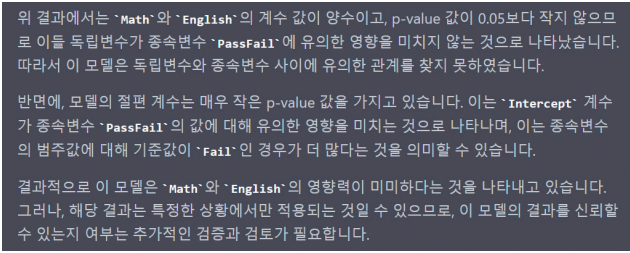
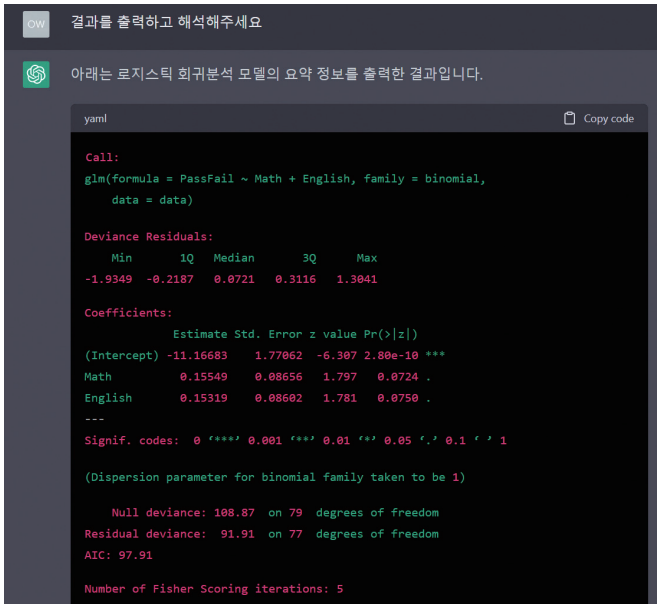
```
javascript
Math English Total PassFail
1 77.4 78.4 155.8 Pass
2 64.8 80.5 145.3 Fail
3 72.2 61.2 133.4 Fail
4 80.3 68.1 148.4 Fail
5 76.7 80.6 157.3 Pass
6 57.1 65.6 122.7 Fail
```

위 결과에서 확인할 수 있듯이, 첫 6개 학생 중에서 3명은 합격(Pass)이고 3명은 불합격(Fail)입니다.





여기서 본격적인 분석 질문을 하게 된다. 비록 예시에서는 독립변수와 종속변수라는 다소 전문적인 용어를 사용했지만, 다른 경험들에 비춰 봤을 때 이런 전문 용어를 쓰지 않아도 챗GPT가 충분히 분석의 구조를 이해할 수 있다는걸 확인 할수 있다. 이때, 매우 흥미로운 점은 챗GPT가 질문자의 질문을 수정하려고 한다는 점이다. 즉, 나는 선형 회귀모형을 주문했으나, 챗GPT가 스스로 판단하여 로지스틱 회귀 분석을 추천하고 그 이유를 친절하게 설명하는 모습을 볼수 있다. 이 과정이 현장에서 데이터 분석을 실시하는 사람들의 입장에서 보면 다소 충격으로 다가올수 있다. 왜냐하면 데이터 분석을 할 때 사람들이 초창기에 가장 힘들어 하는 부분이 ‘기법의 선택’이기 때문이다. 그런데 그 심리적 저항감이 한번에 해소되는 것을 의미하기 때문이다.



결국, 챗GPT는 자신이 추천한 분석 기법을 기반으로 명확한 분석 요약을 하며 이 대화는 끝나게 되었다.

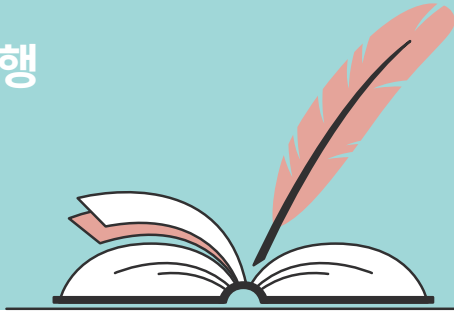
### 인공지능시대 ... 생각하는 힘, 질문하는 힘에 집중

자, 이런 일련의 과정을 보며 우리 어떤 감정을 추스려야 할까? 그것은 이제 ‘모두의 분석’의 시대가 한층 가까워 왔다는 것을 의미한다. 분석을 하기 위해서는 크게 2가지 역량이 필요하다고 들 한다. 하나는 소프트 스킬(생각하는 힘)과 하드스킬(기법/도구를 다루는 힘)이 그것이다. 사실 우리 기존의 대부분의 학습을 영어든, 수학이든, 데이터 과학이든 하드스킬 중심으로 해왔던 게 사실이다. 그리고 하드스킬에 능통한 사람이 더 권위자인 것으로 인정받아왔던 것이 사실이다. 그런데, 챗GPT의 등장은 과감히 ‘생각하는 힘, 질문하는 힘’에 집중해라고 우리에게 말하고 있다. 간단한, 대화형 명령을 통해 높은 수준의 분석 및 가공을 수월하게 해내게 때문이다. 하지만 단순히 편해졌네?에 그치면 안될 것이다. 챗GPT가 하지 못하는 영역에 집중하는 것이 우리의 시작이 되기 때문이다. 첫번째와 두번째 간단한 챗GPT 기반 데이터 이해 및 가공 사례를 통해 보면 이제 우리에게 필요한 역량은 ‘어떤 질문을 할 것인가?’로 귀결된다. 예를 들면, 첫번째 케이스에서 통찰적인 의심을 해야, ‘혹시 2007년에 롯데월드에 무슨 일이 있었니?’라는 새로운 사실로 확장이 가능하고, 두번째 케이스에서 ‘인사팀에서 이런 분석 기법으로 주로 다루는 주제가 무엇인지 알려줄래?’와 같은 적극적인 학습이 가능해질 테니 말이다.

데이터가 많아질수록 인간은 판단하기 힘들어 한다. 이 아이러니를 잘 극복하기 위해서는 점점 인공지능에 자신의 판단을 맡기는 현상은 가속화 될 것 같다. 여기서 경계해야 하는 점은 바로 ‘판단의 양극화’이다. 누군가는 챗GPT를 넘어서 무궁한 의심으로 그야말로 기계를 기계답게 부릴 것이고, 누군가는 심각한 의존 상황에 빠져 이치를 스스로 따지는 힘을 급속도로 잃어 갈 것이다. 그리고 이런 판단력의 양극화든 부와 지식의 양극화보다 훨씬 해소하기 어려운 심각한 사회적 문제가 될 것 같다. 챗GPT가 맹위를 떨치는 2023년 그 성능에 놀란 마음을 잠시 접어두고 과연 이 기술이 나를 어디로 데려갈지 곰곰히 생각해 봐야 할 것이다. 분석은 가공이 아니라 질문하는 힘이란 것을 이해할 때 그 생각의 끝이 조금 보일 것이다. 데이터와 데이터를 잘 병합해서 새로운 데이터를 만드는 사람보다, 데이터를 보고 새로운 의심과 이치를 따질 줄 아는 사람이 더 크게 드러나는 시대가 왔기 때문이다. (문의 이메일 : stephen.kang@deepskill.io)

## 통계교육원의 빅데이터프로젝스수행 (SAS기반) 교육 수기-1

김세윤 | 교육생



통계교육원의 빅데이터프로젝스수행(SAS기반) 교육을 ‘22. 7. 18. ~ ’22. 7. 22. 4일간 수강한 김세윤입니다. 본 강의에 대한 교육수기를 작성하고자 합니다. 교육수기는 저의 업무에 대한 소개 및 교육 수강 동기, 강의의 구성, 제가 느낀 강의의 장점과 아쉬웠던 점으로 구성하였습니다.

### 첫 번째 내용으로는 저의 담당업무와 교육 수강 동기를 간략히 말씀드리려고 합니다.

저는 통계학 비전공자로, 작년 3월 통계청에 입직하여 4월부터 서비스업동향과에 발령을 받아 업무경력이 이제 갓 1년을 넘긴 초임 공무원입니다. 저는 과에서 줄곧 서비스업생산지수 개편업무를 담당하면서 경제총조사 데이터를 바탕으로 자료처리를 수행하고 있습니다. 저는 통계학 비전공자로서 데이터를 핸들링한다는 것과 그를 위해 여러 프로그램과 언어를 사용한다는 발상 자체를 해본 적 없이 살아왔는데 업무를 통해 처음 접하게 되었습니다. 처음에는 막막하고 어려웠으나, 점점 익숙해지고 나아지는 스스로를 보며 자료처리에 흥미를 느끼고 있었고, 이에 통계청 내부자료뿐만 아니라 외부 데이터를 다뤄보고싶다는 생각을 하게 되었는데, 제가 주로 사용하는 SAS 프로그램을 통해 외부 데이터를 다뤄볼 수 있는 본 교육이 매력적으로 느껴져 수강하게 되었습니다.

### 두 번째 내용으로는 강의의 구성에 대해 말씀드리려고 합니다.

강의에서는 국립건강영양조사 자료(이하 ‘국건영자료’라 하겠습니다)와 국민건강보험공단 자료(이하 ‘건보공단자료’라 하겠습니다) 이렇게 2가지 자료를 다루고, 각 자료에 대한 설명, 기본적인 실습, 미니 프로젝트 수행으로 이루어집니다.

먼저 교수님께서 각 자료의 구성에 대해 설명해주시는. 개인적인 생각으로는 본격적인 자료처리에 앞서 그 자료를 잘 이해하는 것이 가장 중요하다고 생각하는데, 교수님께서 자료 구성, 변수의 의미 등을 코드북을 바탕으로 설명해주셔서 자료를 큰 어려움 없이 이해할 수 있었습니다.

다음으로는 자료를 활용하여 예제를 풀어보는 실습이 이루어집니다. 실습은 자료를 활용하여 원하는 결과값을 이끌어내는 것을 목적으로 하는데, 이 과정에서 데이터 전처리, 통계분석이 이루어집니다. 그렇기 때문에 SAS 프로그램 활용과 SQL 언어가 활용되어 다양한 구문 등을 실습해볼 수 있습니다. 또 통계분석을 위해 여러 프로시저를 활용해 보았는데, 목적에 맞는 프로시저를 직접 활용해볼 수 있습니다. 특히 국건영자료는 복합표본설계가 이루어진 자료인데, 프로시저를 활용함에 있어 단순무작위추출에 활용되는 프로시저가 아닌 복합표본설계에 적합한 다른 프로시저를 활용해야 하는 점 등을 알 수 있었습니다.

이렇게 여러 예제에 대한 실습을 통해 데이터와 친해진 후에는 미니 프로젝트가 이루어집니다. 미니 프로젝

트는 교수님께서 이전에 각 자료들을 활용하여 연구하셨던 내용을 기반으로 이를 약식으로 수행해보는 것입니다. 교수님께서 연구하셨던 내용을 먼저 설명해주시고, 앞에서 예제에 대한 실습을 하며 습득했던 다양한 기법들을 활용하여 미니 프로젝트를 수행하게 됩니다.

### 세 번째 내용은 제가 느낀 강의의 장점입니다.

통계청 공무원인 저에게는 외부 데이터를 접해볼 수 있다는 것이 가장 큰 장점으로 다가왔습니다. 통계청에서 근무하는 동안에는 통계청 내부 데이터는 많이 다뤄볼 수 있겠지만, 외부 데이터를 쉽게 접하기는 어렵다고 생각했습니다. 따라서 국건영자료나 건보공단자료와 같은 외부 데이터를 직접 접해보고 핸들링하는 경험을 해볼 수 있다는 것이 가장 큰 장점이라고 생각했습니다. 특히 건보공단 데이터는 전체 국민의 95%를 포괄하는 사실상의 전국민 전수자료인데, 물론 본 강의에서는 일부만으로 이루어진 코호트자료를 활용하긴 했으나, 그러한 자료를 직접 접해보고 다뤄볼 수 있다는 것이 정말 좋은 기회라고 생각되었습니다.

다음으로는 다양한 프로그래밍 기법을 배울 수 있다는 점이 좋았습니다. 데이터를 활용하여 여러 실습과 미니 프로젝트를 수행하는 과정에서 필수적으로 동반되는 데이터 전처리 과정에서 여러 기법들을 익힐 수 있었습니다. 특히 기본적인 컬럼 생성 및 삭제, 변수 추가, 정렬, SQL 구문을 활용한 join 및 요약값 생성, 다양한 표시형식 등을 다루며 평소 사용해보지 못했던 기법들을 익힐 수 있었습니다. 또한 통계분석을 함에 있어서도 다양한 통계분석 프로시저가 있음을 알 수 있었습니다. 특히 단순무작위추출과 복합표본설계의 차이점을 고려하여 자료의 성격에 맞는 프로시저를 활용해야 한다는 점은 통계학 비전공자인 저에게는 시야를 한층 더 넓힐 수 있는 계기가 되었습니다.

### 마지막 네 번째 내용은 제가 느낀 강의의 아쉬운 점입니다.

먼저 실습 및 프로젝트에 있어 개개인이 코드를 짜고 직접 해볼 수 있는 시간이 충분하지만은 않았다는 것입니다. 간단한 문제는 약간의 시간만으로도 코드를 구현할 수 있지만, 복합적이고 어려운 문제도 적지 않았는데, 이러한 문제들에는 충분한 생각과 시간이 필요하다고 생각했으나 충분히 주어지지는 않았습니다. 하지만 4일 만에 모든 과정을 마무리해야한다는 현실적인 여건 탓인지, 교수님께서 충분한 시간을 제공해 주시지는 못하는 모습이었고, 약간의 시간 이후에 정답을 공개하시고 그 정답에 대한 설명을 덧붙이시는 방식으로 실습과 프로젝트가 이루어져, 개인이 보다 충분히 고민할 시간이 부족하다는 것에 다소간의 아쉬움이 있었습니다.

다음으로는 교육생 개개인이 모두 프로그램 활용을 능숙하게 하지는 못한다는 점입니다. 다양한 기관에서 여러 교육생들이 참석을 하는데, 저를 포함하여 몇몇 교육생들은 실습에 어려움이 있음이 느껴졌습니다. 특히 건보공단자료는 복잡하여 작업에 상당한 난이도가 있다고 느꼈는데, 교수님께서도 아쉬움을 피력하신 것처럼 보다 프로그램 활용에 능숙한 교육생들로 교육이 이루어진다면 교육효과가 보다 극대화될 수 있을 것이라고 생각합니다.

하지만 세상 모든 것이 좋은 점만 있을 수는 없는 것처럼, 전술한대로 몇 가지 아쉬운 점이 있기는 하지만 본 강의는 유능하고 젊은 교수님께서 열정적으로 강의하시는 단점보다는 장점이 훨씬 많은 강의라고 생각합니다. 혹 저의 경우처럼 외부 데이터를 접해보고 다뤄보고 싶은 분이 계시다면 꼭 한번 추천드리고 싶습니다. 이상 교육수기를 마칩니다. 감사합니다.



## 엑셀 초급 데이터 분석 및 활용 교육 수기-2

김지안 | 교육생



안녕하십니까, 저는 지난 9월 초 통계교육원에서 ‘엑셀 초급 데이터 분석 및 활용’ 과정을 수강한 김지안이라고 합니다. 공무원이 된 지 오래되지는 않았지만, 지금까지 타 교육원에서 들었던 다른 강의들은 다소 추상적인 내용이 대부분이었던지라 수강 후 항상 아쉬움이 많았습니다. 코로나로 인하여 거의 모든 교육이 집합 교육이 아닌 온라인 교육이었던 점도 아쉬웠던 중에 통계교육원에서 이틀간 강의를 수강할 수 있는 기회가 생겨 망설임 없이 신청하게 되었습니다.

### 1. 교육 첫날

이른 새벽부터 부산에서 KTX를 타고 대전으로 향했습니다. 교육을 위해 다른 지역에 가보는 것이 처음이고 대전이라는 도시 자체도 처음이라, 새벽에 일어났음에도 피곤함보다는 설레는 마음이 컸습니다. 비가 제법 오는 날이었지만 교육원이 대전역에서 멀지 않은 거리에 있어 이동하는 데 불편함은 없었습니다. 교육원 4층 전산실 앞에서 교육생 등록 후 10시부터 본격적인 교육이 시작되었습니다. 이틀 동안 압축적으로 듣는 강의인 만큼 오지안 강사님께서 저희 수강생들이 실무에서 빠르게 적용해 볼 수 있는 내용 위주로 구성된 알찬 커리큘럼을 소개해주셨습니다.

첫날은 데이터 전처리 작업 방법, 빠른 채우기, 실용함수 활용 등의 내용을 배웠습니다.

데이터 전처리란 자료에서 불필요한 부분을 수정하거나 제거하는 작업입니다. 이 과정이 잘 되어야만 이후 작업에서 내가 원하는 결과값을 오류 없이 도출할 수 있기에 모든 데이터 처리에 앞서 진행되는 기초공사 격이라 할 수 있습니다.

수많은 데이터를 빠르게 입력할 수 있게 도와주는 빠른 채우기는 실습 강의 내내 ‘이건 만능키다!’라는 생각이 들게 하는 내용이었습니다. 수만 개의 데이터 중 엑셀이 규칙성만 잘 발견할 수 있도록 해 주면 생각보다 훨씬 더 많은 부분에 적용할 수 있었습니다.

실용함수 부분에서는 그동안 어렵듯이 알고만 있었기에 섬뚱리 사용하기 힘들었던 함수들에 대해 확실히 정리해볼 수 있었습니다.

### 2. 교육 둘째 날

둘째 날은 전날 교육보다는 더 심화된 피벗 테이블 활용, 통계함수 등에 대해 배웠습니다. 엑셀로 내려받은 많은 양의 데이터를 다양한 방법으로 조합하여 만들어내는 피벗테이블, 그리고 수많은 함수식이 처음에는 익숙지 않았지만, 강사님의 설명과 여러 번의 실습을 통해 감을 잡을 수 있었습니다. 그러다 업무상 자주 쓰는 함

수가 등장하면 괜스레 반가운 마음에 기분이 좋아지곤 했습니다.

후반부에는 데이터 시각화에 대한 강의가 이어졌습니다. 요즘은 손가락만 까딱하면 얻을 수 있는 정보의 양이 너무 많아 문제가 되는 시대라고 생각합니다. 따라서 나에게 필요한 정보만을 잘 선별하고 이를 타인이 이해하기 쉽게 잘 전달하는 것도 굉장한 경쟁력인데, 같은 내용을 담은 시각화 자료에도 똑똑한 디테일을 곁들이면 얼마나 다른 결과물이 나오는지 경험해 볼 수 있었습니다. 교육 관련 이슈뿐만 아니라 사회 전반에 걸친 통계 데이터, 시선을 끄는 시각화 자료가 필요할 때 이를 내려받을 수 있는 사이트도 상세히 알려주셔서 빠짐 없이 메모해두었습니다.

### 3. 강의 전반

이틀 동안 이어진 강의 내내 강사님께서 중간중간 자리를 옮기시며 저희가 잘 따라오고 있는지 체크하셨기 때문에 더욱 긴장감을 가지고 수업에 임할 수 있었습니다. 더불어 엑셀 실습 파일, 교육 책자가 함께 제공되었기에 강의를 따라가는 데 큰 어려움이 없었습니다. 직장으로 돌아온 지금도 책상 위 손 닿는 곳에 책자를 꽂아 두고 언제든지 참고할 수 있도록 신경 쓰고 있습니다.

이틀간 강의를 듣고 나니 아직 완벽하진 않아도 실무 경쟁력이 높아진 것만 같은 느낌이 들었습니다. 엑셀은 사무직에 종사하는 직장인들에게 필수적인 존재인 만큼 얼마나 엑셀을 잘 다룰 줄 아느냐 하는 것이 일의 효율을 높이는 일등공신이라고 생각합니다. 그동안은 반복적으로 써야 하는 기능이 있다면 그때그때 인터넷을 찾아보곤 했는데 이제는 강의에서 배운 알찬 내용들이 일목요연하게 적힌 강의 책자를 먼저 찾아볼 것 같습니다.

### 4. 숙소 및 식사

다른 지역에 가게 되면 신경 쓰일 수밖에 없는 것이 숙소인데, 강의 자체도 알차고 만족스러웠지만, 교육 기간 동안 머물렀던 통계교육원 숙소도 상당히 만족스러웠었습니다. 건물이 크고 깔끔했고 코로나 감염 방지를 위하여 1인 1실을 사용하였기에 짧은 기간이었지만 아주 편하게 이용할 수 있었습니다.

첫째날 밤, 무선 이어폰 한쪽이 침대 밑으로 굴러 들어간 적이 있어 우연히 침대 밑을 보게 되었는데 다시 주워든 이어폰에도 먼지 한 톨 묻지 않았을 만큼 깔끔한 공간이었습니다. 잘 보이지 않는 곳은 관리가 잘 안 되는 경우를 너무나 많이 봤는데, 이러한 우연한 발견에 더욱 쾌적한 기분으로 하루를 마무리할 수 있었습니다. 머무는 내내 항상 맛있는 식사가 제공되어, 특히 점심시간 이후 에너지를 잘 충전한 뒤 다시 강의에 집중할 수 있었습니다.

### 5. 마무리

혹시 통계교육원에서 엑셀 연수 관련 공문이 왔는데 업무나 거리 때문에 조금이라도 망설이고 있는 분이 계신다면 더 고민 말고 당장 신청하라고 말씀드리고 싶습니다. 그만큼 본 강의는 굉장히 유익하고, 가장 활용도가 높은 강의였습니다. 공직에 있는 동안 앞으로 또 통계교육원에서 열리는 다양한 교육에 참석할 기회가 주어진다면 틈틈이 수강하여 저만의 역량을 키워나가고 싶습니다.

교육과정별 운영계획

가. 집합과정

구분	과정명	수준	교육대상	교육 일수	기당 인원	교육 횟수	교육일정
A 기본교육(12)							
1	9급 신규자 기본교육(공채)		9급 공채 신규임용예정자	15	60	1	10.10~10.27.
2	9급 신규자 기본교육(경채)		9급 경채 신규임용예정자	10	35	1	3.20~3.31.
3	7급 신규자 실무교육		6~7급 신규자	4	30	1	6.20~6.23.
4	4급 승진후보자 역량향상		4급 승진후보자	5	25	1	1.16~1.20.
5	5급 승진후보자 역량향상		5급 승진후보자	5	24	2	3.13~3.17. 8.28~9.1.
6	6급 승진자 역량향상		6급 승진자	4	40	2	5.16~5.19. 10.31~11.3.
7	현장조사 역량강화		지방통계청 조사담당자	3	30	2	4.26~4.28. 5.24~5.26.
8	지방청 조사관리 역량강화		지방통계청 조사관리자	3	30	2	9.20~9.22. 11.8~11.10.
9	소통과 리더십 역량강화(신설)		5급	3	30	1	6.26~6.28.
10	문제해결 역량강화(신설)		6급	3	30	1	5.31~6.2.
11	지방청 관리자 리더십 교육(신설)		지방통계청 관리자 (과장, 사무국장, 본부팀장)	1	30	2	4월
12	비대면조사 공무원 조사관 교육(신설)		비대면조사 조사관	3	40	2	2.22~2.24. 11.28~11.30.
B 전문교육(84)							
B-1 국가통계정책(11)							
13	국가통계의 이해	초급	제한없음	3	20	2	4.26~4.28. 9.4~9.6.
14	통계와 정책	중급	통계청	3	20	2	4.17~4.19. 7.5~7.7.
15	정책과정과 통계의 역할	초급	제한없음	2	30	1	4.10~4.11.
16	국가승인통계관리	초급	통계작성기관	3	30	2	5.8~5.10. 9.13~9.15.
17	통계품질관리	중급	제한없음	3	30	2	2.13~2.15. 7.12~7.14.
18	정책지표 작성 방법론	초급	제한없음	2	20	1	9.7~9.8.
19	통계기초 및 활용	초급	제한없음	5	30	2	5.22~5.26. 9.18~9.22.
20	한국표준산업분류	초급	제한없음	3	30	1	3.6~3.8.
21	사회분류(직업분류 등)의 이해	초급	제한없음	3	30	1	6.12~6.14.
22	한국표준질병· 사인분류의 이해 및 활용	중급	제한없음	2	30	1	11.13~11.14.
23	한국표준건강분류의 이해	초급	제한없음	1	30	1	4.21.
B-2 국가통계이해(9)							
24	경제통계의 이해	중급	제한없음	4	30	1	3.14~3.17.
25	국민계정	중급	제한없음	3	20	1	9.6~9.8.
26	재무제표	중급	제한없음	3	30	2	6.26~6.28. 8.7~8.9.
27	농어업통계의 이해	중급	제한없음	3	30	1	4.10~4.12.
28	인구통계의 이해	중급	제한없음	3	20	1	3.29~3.31.
29	소득분배지표의 이해	중급	제한없음	2	20	1	8.3~8.4.
30	고용통계의 이해	중급	제한없음	2	20	1	6.27~6.28.
31	빅데이터와 행정자료의 이해	초급	통계청	3	30	2	3.8~3.10. 7.26~7.28.
32	지속가능발전목표(SDGs)의 이해	초급	제한없음	1	20	1	7.5.
B-3 국가통계작성(18)							
33	국가통계실무1 (조사설계 및 조사표설계 등)	초급	통계청	4	40	2	3.7~3.10. 8.8~8.11.
34	국가통계실무2 (표본설계 및 추정)	초급	통계청	4	40	2	4.11~4.14. 9.12~9.15.
35	조사설계 및 조사표설계(인지실험)	초급	제한없음	3	20	1	4.24~4.26.
36	표본실무	중급	제한없음	3	20	1	7.19~7.21.
37	통계조사관 직무연수	초급	지방통계청 통계조사관	3	30	1	8.16~8.18.
38	지역통계실무	초급	제한없음	3	20	1	3.8~3.10.
39	국가통계실무3 (자료수집·처리 및 분석)	초급	통계청	4	40	2	5.15~5.18. 10.17~10.20.
40	자료수집·처리 및 분석	초급	제한없음	3	20	1	8.28~8.30.
41	계열조정실무	중급	통계청	2	40	1	1.12~1.13.
42	데이터 에디팅	중급	제한없음	4	20	1	10.24~10.27.
43	시계열분석	중급	제한없음	4	20	1	7.3~7.6.
44	지수이론	중급	제한없음	3	20	1	7.12~7.14.
45	국가통계실무4 (통계작성·공표 등)	초급	통계청	4	40	2	6.12~6.15. 11.13~11.16.
46	국가통계정보의 활용	초급	제한없음	3	20	1	5.1~5.3.
47	통계데이터 비밀보호의 이해	중급	제한없음	3	20	1	6.14~6.16. 3.27~3.29.
48	통계보고서 작성	중급	제한없음	3	30	3	5.22~5.24. 10.11~10.13.
49	통계지리정보의 이해(신설)	초급	통계청	3	20	1	6.28~6.30.
50	나라통계시스템 종합 교육(신설)	초급	통계청, 통계작성기관	2	30	1	5.11~5.12.

구분	과정명	수준	교육대상	교육 일수	기당 인원	교육 횟수	교육일정
B-4 데이터 분석(13)							
51	R 초급 통계분석	초급	제한없음	4	30	2	2.20~2.23. 8.7~8.10.
52	R 중급 통계분석	중급	제한없음	4	30	2	6.26~6.29. 8.28~8.31.
53	R 고급 통계분석	고급	제한없음	3	30	1	7.10~7.12.
54	SAS 초급 통계실무	초급	통계청	4	20	1	3.20~3.23.
55	SAS 중급 통계분석	중급	제한없음	4	20	2	4.25~4.28. 10.23~10.26.
56	SAS 고급 통계분석	고급	제한없음	5	20	1	7.3~7.7.
57	SPSS 중급 통계분석	중급	제한없음	4	30	2	5.15~5.18. 9.18~9.21.
58	SPSS 고급 통계분석	고급	제한없음	4	30	1	7.18~7.21.
			통계청	3	20	1	2.15~2.17.
59	파이썬 초급 통계분석	초급	제한없음	3	20	2	6.14~6.16. 10.18~10.20.
60	파이썬 중급 통계분석	중급	제한없음	3	30	2	4.19~4.21. 8.2~8.4.
61	엑셀 초급 데이터 분석 및 활용	초급	제한없음	3	30	3	3.27~3.29. 6.19~6.21. 10.30~11.1.
62	엑셀 중급 통계분석	중급	제한없음	3	30	2	4.17~4.19. 9.11~9.13.
63	기타 기본 빅데이터 분석	중급	제한없음	3	20	2	6.28~6.30. 11.6~11.8.
B-5 전문가 양성(8)							
64	경제시계열분석 및 지수이론	고급	통계청	14	10	-	-
65	국민계정	고급	통계청	12	10	-	-
66	인구통계분석	고급	통계청	12	10	-	-
67	통계데이터기획	고급	통계청	11	10	-	-
68	표본설계 및 추정	고급	통계청	14	10	-	-
69	무응답 자료처리 및 분석	고급	통계청	12	10	-	-
70	통계데이터 비밀보호	고급	통계청	12	10	-	-
71	AI 적용 데이터 사이언스	고급	통계청	23	20	-	-
B-6 기타 통계교육(28)							
72	맞춤형 통계교육(기관)		통계작성기관	2	20	16	-
73	통계 아카데미		제한없음	1	40	3	-
학생교육(3)							
74	사랑나눔 통계교실		초등학교 5~6학년	3	25	2	7.24~7.26. 7.31~8.2.
75	어린이 통계캠프		초등학교 5~6학년	3	30	2	-
76	중학생 통계아카데미		중학생	3	30	3	7.5~7.7. 7.19~7.21. 8.9~8.11.
교사교육(7)							
77	초등학교 교사 통계연수		초등학교 교사	2	30	2	1.10~1.11. 8.3~8.4.
78	중학교 교사 통계연수		중학교 교사	2	30	2	1.9~1.10. 8.7~8.8.
79	고등학교 교사 통계연수		고등학교 교사	2	30	2	1.12~1.13. 8.1~8.2.
80	중등 교사 통계연수(심화)		중·고등학교 교사	3	30	2	1.11~1.13. 7.24~7.26.
81	실용통계 지도교사 통계연수		고등학교 수학 교사	4	30	2	1.16~1.19. 7.24~7.27.
82	시도교육청 교사 통계연수		초·중·고 수학교원 등	2	30	-	-
83	통계를 활용한 통합사회 지도교사 연수		고등학교 사회과 교사	2	30	2	1.30~1.31. 8.1~8.2.
외국인교육(1)							
84	UNSIAP 공조 통계연수		외국공무원	5	20	2	3.20~3.24. 6.12~6.16.
B-7 정보화(2)							
85	오피스를 활용한 데이터 시각화	중급	제한없음	3	40	4	2.22~2.24. 5.31~6.2. 9.20~9.22. 11.1~11.3.
86	디지털영상 및 이미지활용	초급	제한없음	3	30	1	10.18~10.20.
C 기타교육(3)							
C-1 사책교육(1)							
87	소셜미디어 활용		제한없음	3	20	1	6.21~6.23.
C-2 일반소양(2)							
88	생활속의 음악		통계청	3	50	1	8.2~8.4. 5.24~5.26. 9.20~9.22. 10.25~10.27.
89	현장조사 스트레스 관리		지방통계청 현장조사 직원	3	30	3	

나. 이력나과정

구분	과정명	수준	교육대상	인정시간
이러닝(126)				527
B 전문교육(87)				442
B-1 국가통계정책(15)				64
B-1-1 통계정책(12)				52
1	국가통계 및 제도	초급	제한없음	1
2	성인지통계의 이해와 활용	초급	통계청	1
3	자체통계품질진단 관리	초급	제한없음	4
4	지역사회지표 작성과 활용	초급	제한없음	4
5	지역정책과 통계 활용	초급	제한없음	11
6	통계가 쉬워지는 기초 수학 끝내기	초급	제한없음	2
7	통계기초 및 활용	중급	제한없음	13
8	통계법	초급	제한없음	2
9	통계업무 필수 지식	초급	제한없음	3
10	통계적으로 사고하기	중급	제한없음	3
11	통계품질관리	초급	제한없음	1
12	통계학의 이해	초급	제한없음	7
B-1-2 통계기준(3)				12
13	한국표준산업분류	초급	제한없음	6
14	한국표준직업분류	초급	제한없음	5
15	한국중사상지위분류(신설)	초급	제한없음	1
B-2 국가통계이해(25)				71
B-2-1 경제통계(5)				17
16	광업·제조업동향조사	초급	통계청	4
17	서비스업동향조사	초급	통계청	3
18	소비자물가조사	초급	통계청	4
19	온라인쇼핑동향조사	초급	통계청	2
20	재무제표 이해	중급	제한없음	4
B-2-2 사회통계(20)				54
21	가계동향조사	초급	통계청	3
22	가계동향조사 가구표본관리(신설)	초급	통계청	2
23	가족동향조사	초급	통계청	1
24	경제활동인구조사	초급	통계청	3
25	경제활동인구조사 사례집	초급	통계청	2
26	농가경제조사	초급	통계청	3
27	농가판매 및 구입가격 조사	초급	통계청	1
28	농산물생산비조사	초급	통계청	4
29	농업면적조사	초급	통계청	3
30	농작물생산조사(생산량부문)	초급	통계청	2
31	사회통계의 이해	초급	제한없음	6
32	산지별값조사	초급	통계청	2
33	양곡소비량조사	초급	통계청	2
34	아기경제조사	초급	통계청	3
35	어류양식동향조사	초급	통계청	2
36	어업생산동향조사	초급	통계청	3
37	이민자 체류실태 및 고용조사	초급	제한없음	3
38	인구동향조사	초급	통계청, 공무원	3
39	지역별 고용조사	초급	제한없음	3
40	축산물생산비조사	초급	통계청	3
B-3 국가통계작성(17)				59
B-3-1 조사기획(11)				7
41	조사방법의 이해	초급	제한없음	7
B-3-2 자료수집·처리 및 분석(7)				27
42	시계열자료의 분석과 실무	고급	제한없음	5
43	자료수집	초급	제한없음	2
44	조사원 면접기법교육	초급	제한없음	1
45	표본이론 기초	중급	제한없음	6
46	표본합식 기초	초급	제한없음	2
47	현장조사 인력양성	초급	제한없음	4
48	회귀분석의 이해와 사례	고급	제한없음	7
B-3-3 공표 및 관리(9)				25
49	MDIS 활용	초급	제한없음	2
50	NARA-PC 활용	초급	제한없음	6
51	SGIS 플러스 활용	초급	제한없음	2
52	국가통계포털(KOSIS) 활용	초급	제한없음	2
53	나라통계시스템 조사통계 사용자교육	초급	제한없음	2
54	통계데이터 비밀보호	초급	제한없음	3
55	통계작성기관을 위한 통계DB시스템 사용법	초급	제한없음	2
56	통계를 활용한 보고서 작성방법	중급	제한없음	4
57	제미있는 실용통계(신설)	초급	제한없음	2
B-4 데이터 분석(21)				216
B-4-1 R(2)				26
58	R 기초	초급	제한없음	14
59	R을 활용한 통계분석	중급	제한없음	12
B-4-2 SAS(2)				11
60	SAS를 활용한 초급 통계분석	초급	제한없음	6
61	SAS를 활용한 중급 통계분석	중급	제한없음	5
B-4-3 SPSS(2)				9
62	SPSS 초급 통계분석	초급	제한없음	5
63	SPSS 중급 통계분석	중급	제한없음	4

구분	과정명	수준	교육대상	인정시간
B-4 데이터 분석(21)				216
B-4-4 파이썬(2)				9
64	파이썬 데이터 분석 초급	초급	제한없음	3
65	파이썬 데이터 분석 중급	중급	제한없음	6
B-4-5 엑셀(1)				4
66	엑셀로 배우는 통계분석	중급	제한없음	4
B-4-6 빅데이터(10)				154
67	공공데이터로 만들어가는 새로운 세상	초급	제한없음	5
68	공공빅데이터 업무작용 길라잡이	초급	통계청	6
69	데이터 리터러시	초급	통계청	1
70	데이터 분석을 위한 SQL 입문	초급	제한없음	3
71	데이터 사이언스 기본과정(통계학 기반)(신설)	중급	통계청	51
72	데이터 사이언티스트 길잡이(수학 기반)	중급	통계청	69
73	데이터 에디팅 방법론(신설)	고급	통계청	3
74	데이터 전처리(신설)	초급	제한없음	3
75	빅데이터와 통계	중급	제한없음	7
76	쉽게 접근하는 데이터 사이언스 기초 (통계학 기반)(신설)	중급	제한없음	7
B-4-7 기타(2)				3
77	통계 교육을 공학도구 통그라미 활용	초급	제한없음	2
78	통계패키지 학습을 위한 필수 통계지식	중급	제한없음	1
B-6 기타 통계교육(7)				26
B-6-3 학생교육(4)				11
79	역사속에서 찾은 통계이야기	초급	제한없음	5
80	통계로 논리 잡기	초급	제한없음	4
81	통계로 보는 안전교육	초급	제한없음	1
82	통계를 보면 경제가 보여요	초급	제한없음	1
B-6-4 교사교육(3)				15
83	실용통계 지도교사 통계연수(이론편)	초급	제한없음	5
84	통계포스터 만들기 지도방법	초급	제한없음	3
85	통계를 활용한 사회와 수행평가 구성 및 실천(신설)	초급	제한없음	7
B-7 정보화(2)				6
86	개인정보보호법 이해하기	초급	통계청	2
87	초보자를 위한 엑셀 기초	초급	제한없음	4
C 기타교육(39)				85
C-1 사색교육(24)				52
88	4차 산업혁명의 이해와 미래대응전략	초급	통계청	5
89	SI반도체 중심의 디지털 혁신(신설)	초급	통계청	1
90	공직자 인보교육	초급	통계청	1
91	공직자를 위한 신권문심서	초급	통계청	3
92	내부 소통능력 및 국민소통 능력향상	초급	통계청	1
93	노동인권감수성	초급	통계청	2
94	디지털물맷을 정부란(신설)	초급	통계청	1
95	문화와 생활 속 청렴(신설)	초급	통계청	2
96	미래 과학기술과 반도체(신설)	초급	통계청	1
97	보안교육(신설)	초급	통계청	1
98	사례로 배우는 청탁금지법	초급	통계청	3
99	사회적 경제(이해편)	초급	제한없음	1
100	아동학대신고의무자교육 및 공공부문 아동학대 예방교육	초급	통계청	1
101	알기 쉬운 이해충돌방지법	초급	통계청	2
102	업무관리능력 향상	초급	통계청	1
103	인권의 이해	초급	통계청	9
104	일상의 평등 ON, 폭력예방교육	초급	통계청	4
105	장애인식개선교육(인식의 길라잡이)(신설)	초급	통계청	2
106	적극행정의 이해(신설)	초급	통계청	2
107	정보공개제도 교육(신설)	초급	통계청	1
108	청탁금지법의 이해	초급	통계청	3
109	코로나19 이후의 국제정세와 한반도의 미래는?(신설)	초급	통계청	1
110	평등한 일상, 폭력예방교육(고위직)	초급	통계청	1
111	한국사피 노동 바로 이해하기	초급	통계청	3
C-2 일반소양(14)				31
112	강의 운영 기법 및 전략	초급	제한없음	4
113	개성이 뚜렷한 직원을 마주하는 방법	초급	통계청	2
114	마음을 움직이는 설득 전략	초급	제한없음	4
115	마음을 읽는 소통형 리더의 기술	초급	통계청	1
116	변화와 성장을 위한 릴레니얼 리더십	초급	제한없음	3
117	부모교육(공동 및 가족 특성별)	초급	통계청	2
118	부모교육(생애주기별)	초급	통계청	4
119	사람을 움직이는 유익한 방법, 소통과 공감	초급	통계청	1
120	슬기롭고 똑똑하게 일하는 방법	초급	통계청	1
121	심리학을 통한 현명한 질문과 대답	초급	통계청	1
122	일가족, 나와 내 가족을 위한 행복찾기	초급	통계청	1
123	직장내 소통강화 대화센스(신설)	초급	제한없음	2
124	현장적 사고 및 기획력 향상	초급	통계청	1
125	행복한 삶을 위한 스트레스 관리	초급	제한없음	2
C-3 공통직무(1)				4
126	공직자를 위한 보고서 잘 쓰는 법	초급	통계청	2

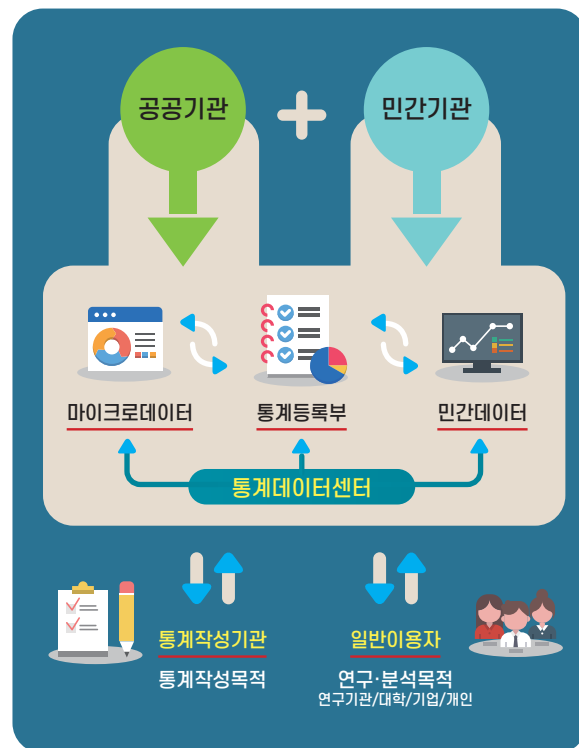


행정통계자료와 민간자료를 한곳에!

# 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

## 통계청, 정부부처, 지방자치단체, 연구기관 등 모든 기관의 마이크로데이터를 한 곳으로



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

### ■ 서비스 소개 (2023년 3월 기준)

가. 서비스명 : 마이크로데이터통합서비스(MDIS, mdis.kostat.go.kr)

나. 제공 통계 수 : 21개 주제별 총 354종 통계 제공

(통계청 50종 및 통계작성기관 304종)

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 귀농귀촌인통계, 영리법인기업체행정통계, 신혼부부통계, 주택소유통계, 중장년층행정통계, 퇴직연금통계, 일자리행정통계, 기업생멸행정통계, 육아휴직통계
통계작성기관	전국다문화가족실태조사, 가족실태조사, 자동차주행거리통계, 직종별사업체노동력조사, 보육실태조사, 기상관측통계, 국민여가활동조사, 외래관광객실태조사, 한부모가족실태조사, 청소년종합실태조사 외 294종

### ■ 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 집사무실 등에서 통계청 서버 접속 후 활용
- 이용센터 : 승인 후 지정된 장소를 방문 활용

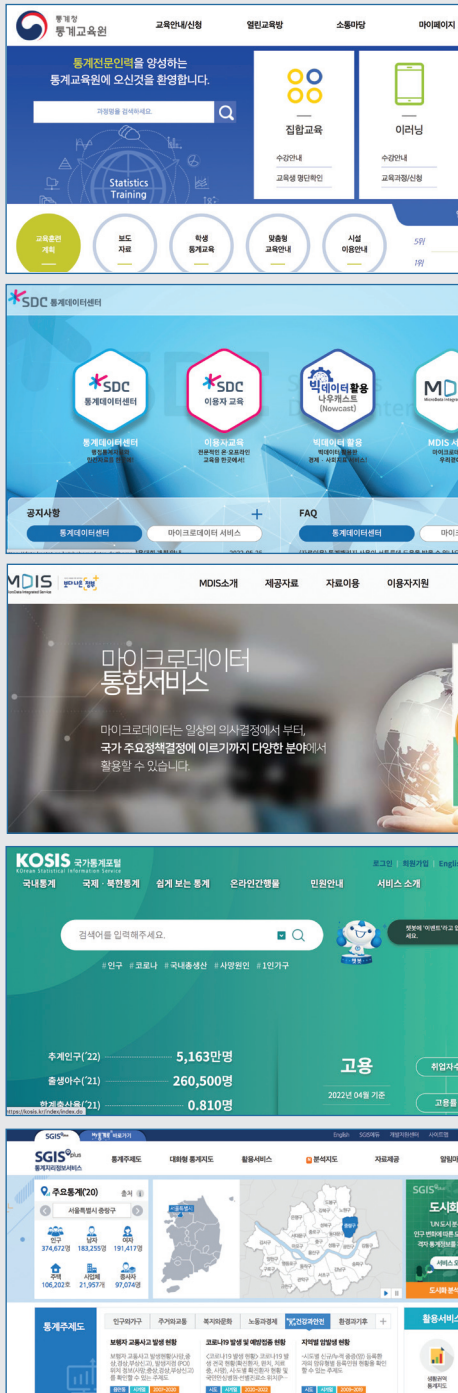
### ■ 문의

- 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