



6주차: 데이터 분석 방법/데이터 마이닝의 이해

ChulSoo Park

School of Computer Engineering & Information Technology

Korea National University of Transportation



학습목표 (6주차)

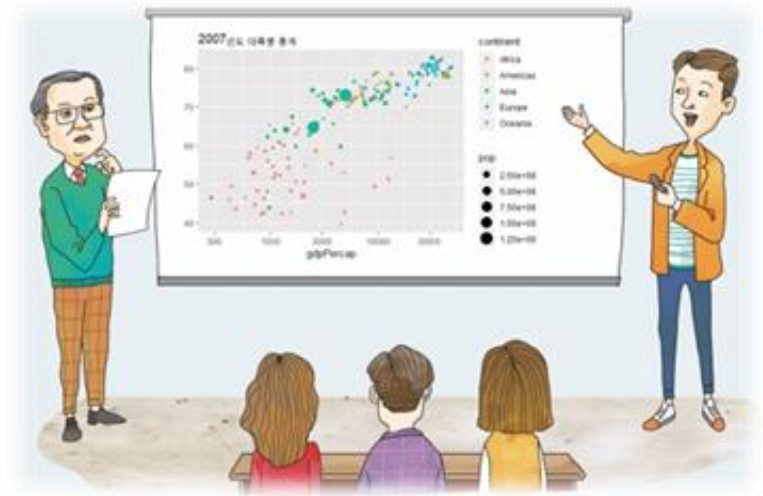
- ❖ 데이터 과학, 데이터 마이닝, 기계 학습 개념 이해
- ❖ 데이터 분석 유형의 이해
- ❖ 데이터 분석 모델 학습
- ❖ 데이터 분석 기법의 이해
- ❖ 데이터 분석 도구 파악
- ❖ 데이터 분석 사례 고찰



07

CHAPTER

데이터 분석 방법 과 데이터 마이닝



데이터 사이언스 개론(김화중,홍릉과학출판사), 데이터 과학 입문(최대우외2명,한국방송통신대학교 출판부)

CONTENTS

개론 7.1 데이터 분석 유형
개론 7.2 데이터 분석 모델
개론 7.3 기계 학습
개론 7.4 분석 모델
개론 7.5 데이터 분석 도구

입문 7.1 데이터 과학에서 마이닝의 역할
입문 7.2 데이터 마이닝의 개념
입문 7.3 데이터 마이닝 관련 분야
입문 7.4 데이터 마이닝 기법 및 도구
입문 7.5 데이터 마이닝 적용 사례



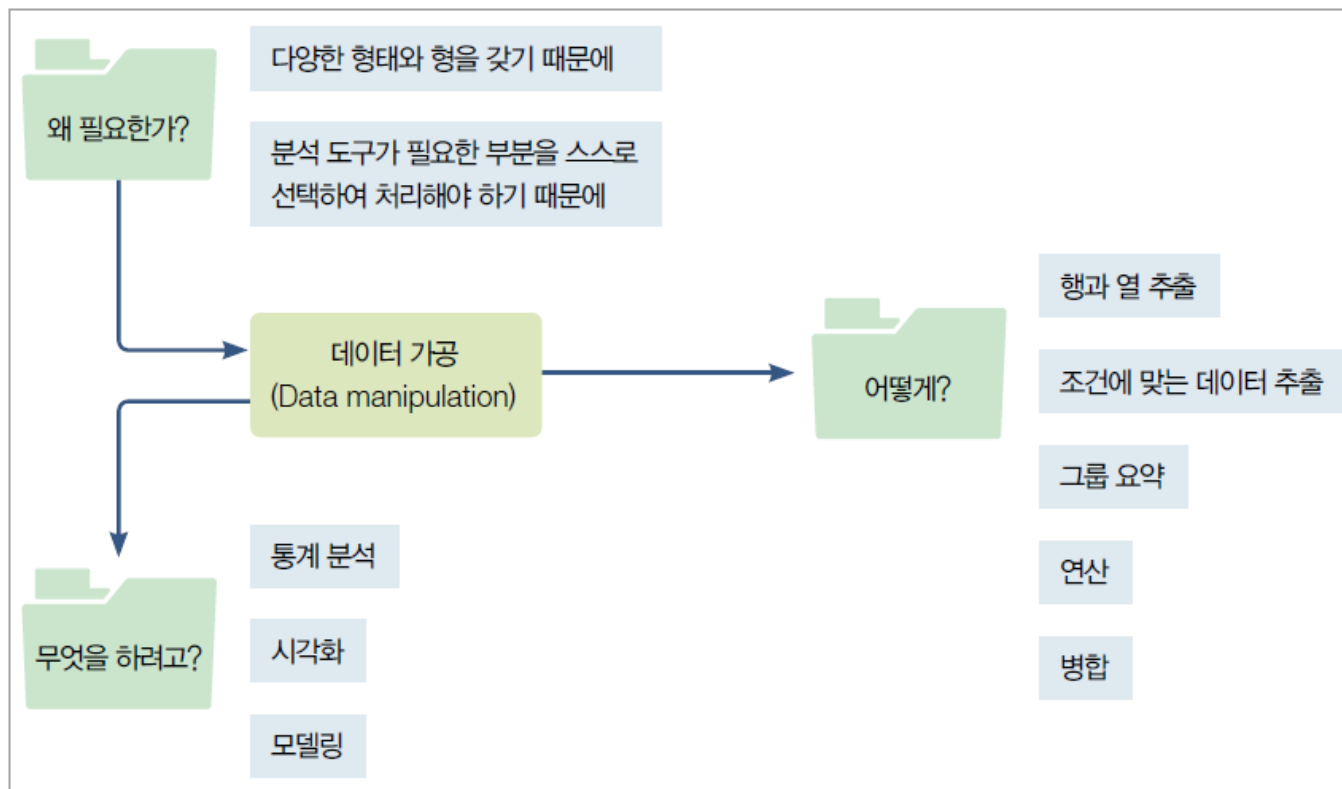
Review

- 잘 정제되고 다듬어진 데이터는 큰 가치가 있지만, 정리되지 않은 데이터는 의미 추출이 어려울 뿐 아니라 잘못된 결론에 이르게 할 수도 있음



Review

- 데이터를 보다 효과적으로 분석하기 위해 데이터를 만지고 변형하는 작업이 바로 데이터 가공(data wrangling)
- 디지털화된 데이터와 R 같은 분석 도구를 활용해 쉽고 빠르게 데이터 가공 작업을 수행할 수 있음



Review

■ 행/열 단위의 연산

- group_by 함수를 이용하면 데이터 프레임에 포함된 factor형 속성을 활용해 전체 데이터를 그룹으로 분류 가능
- 보통 summarise 함수를 연이어 사용해 그룹별 통계 지표를 한번에 산출

```

Console C:/RSources/
> summarize(gapminder, pop_avg=mean(pop))
# A tibble: 1 x 1
  pop_avg
  <dbl>
1 29601212
> summarize(group_by(gapminder, continent), pop_avg=mean(pop))
# A tibble: 5 x 2
  continent pop_avg
* <fct>      <dbl>
1 Africa    9916003.
2 Americas  24504795.
3 Asia      77038722.
4 Europe    17169765.
5 Oceania   8874672.
> summarize(group_by(gapminder, continent, country), pop_avg=mean(pop))
`summarise()` has grouped output by 'continent'. You can override using the `.groups` argument.
# A tibble: 142 x 3
# Groups:   continent [5]
  continent country      pop_avg
  <fct>      <fct>      <dbl>
1 Africa    Algeria    19875406.
2 Africa    Angola      7309390.
3 Africa    Benin       4017497.
4 Africa    Botswana     971186.
5 Africa    Burkina Faso  7548677.
6 Africa    Burundi     4651608.
7 Africa    Cameroon    9816648.
8 Africa    Central African Republic 2560963
9 Africa    Chad        5329256.
10 Africa   Comoros     361684.
# ... with 132 more rows

```

■ 데이터 정렬과 검색(2)

- 정렬과 검색을 통해 데이터를 자세히 관찰 가능
- arrange 함수 : arrange() orders the rows of a data frame by the values of selected columns.
- arrange 함수를 사용해 데이터를 총 판매량의 평균 가격을 기준으로 정렬하면, 판매량 순위는 물론 최대치를 기록한 연도와 지역을 알아낼 수 있음

Console C:/RSources/ ↗

```
> arrange(avg_data, desc(v_avg))
# A tibble: 432 x 5
# Groups:   region, year [216]
  region      year type      v_avg p_avg
  <chr>    <int> <chr>    <dbl> <dbl>
1 TotalUS    2018 conventional 42125533. 1.06
2 TotalUS    2016 conventional 34043450. 1.05
3 TotalUS    2017 conventional 33995658. 1.22
4 TotalUS    2015 conventional 31224729. 1.01
5 SouthCentral 2018 conventional 7465557. 0.806
6 west       2018 conventional 7451445. 0.981
7 California 2018 conventional 6786962. 1.08
8 west       2016 conventional 6404892. 0.916
9 west       2017 conventional 6279482. 1.10
10 California 2016 conventional 6105539. 1.05
# ... with 422 more rows
> |
```

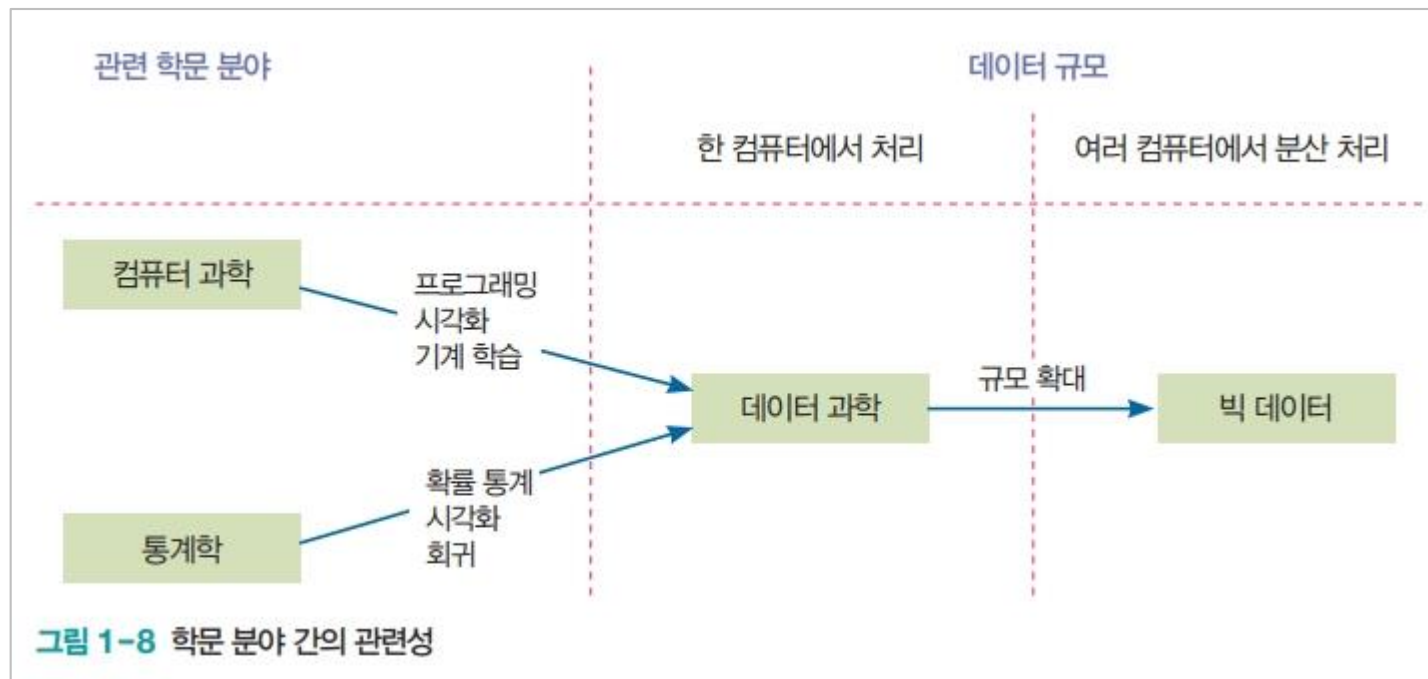
GROUP BY & SORT



Preview

■ 데이터 과학은 다학제 분야

- 컴퓨터 과학: 프로그래밍 언어, 현대적 시각화 기법, 기계 학습 등
- 통계학: 요약 통계, 확률 통계 기법, 시각화 도구, 회귀 기법 등
- 빅데이터: 분산 처리, 하둡 등
- 데이터 마이닝 : 데이터로부터 유용한 지식을 찾아내는 과정
- 이 책에서는 굳이 데이터 과학과 데이터 마이닝을 구분하지 않는다. (p.31)



Preview

■ 데이터 과학은 세상과 활발히 상호작용

- 탐색적 데이터 분석 (EDA: exploratory data analysis) 단계에서 데이터가 부족하다 판단되면 데이터 수집 단계로 돌아가 추가 수집
- 변수를 추가하여 완전 새로 수집하는 상황도 발생
 - 예) 푸드트럭 예에서 골목상권의 영향을 반영하려면 음식점 수와 초밥집이 있는지 여부를 나타내는 변수를 추가하고 새로 데이터를 수집해야 함

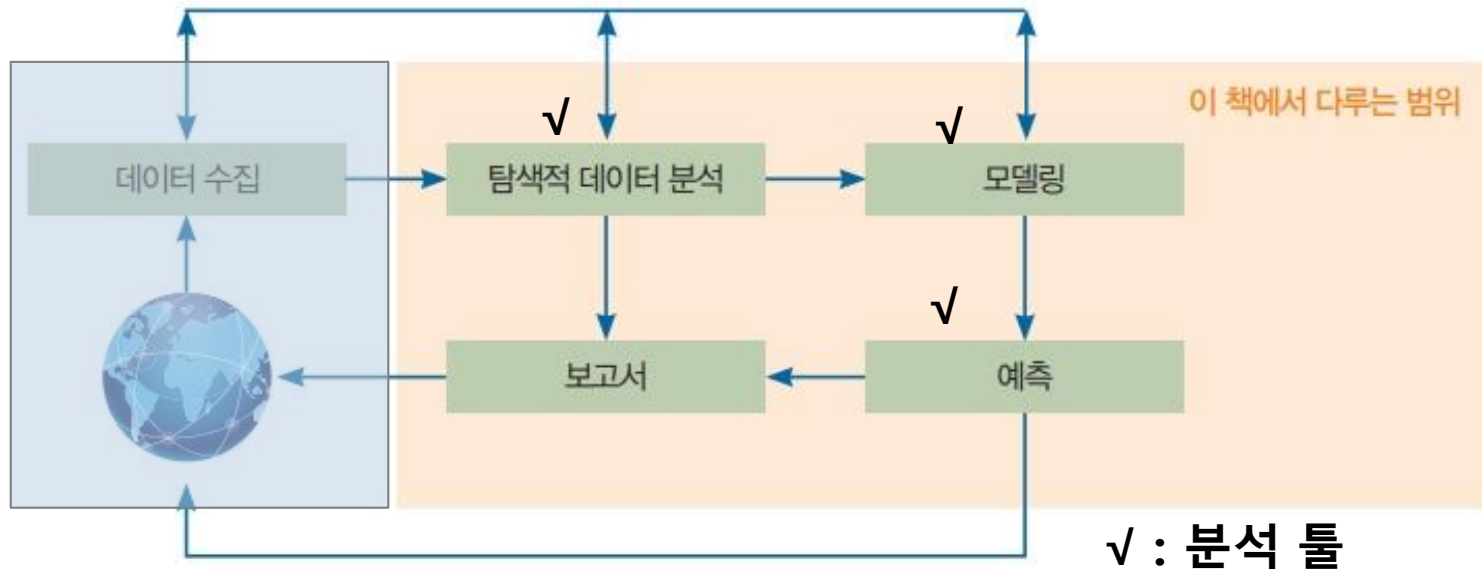


그림 1-7 세상과 상호작용하는 데이터 과학





■ 데이터 과학(data science)이란?

데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 학문.

■ 데이터 마이닝(data mining)이란?

대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 분석하여 가치 있는 정보를 추출하는 과정이다.

■ 기계 학습(機械學習) 또는 머신 러닝(machine learning)이란?

경험을 통해 자동으로 개선하는 컴퓨터 알고리즘의 연구이다. 인공지능의 한 분야로 간주된다. 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야.

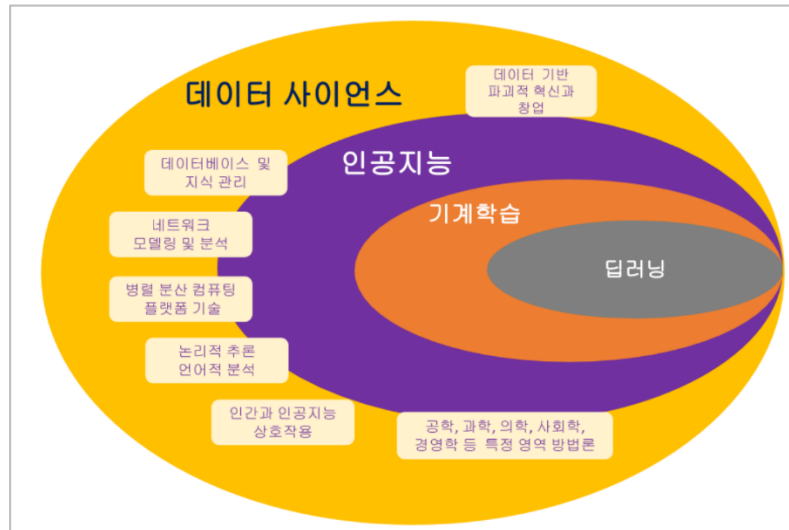
기계 학습과 데이터 마이닝은 종종 같은 방법을 사용하며 상당히 중첩된다. 기계 학습은 훈련 데이터(Training Data)를 통해 학습된 알려진 속성을 기반으로 예측에 초점을 두고 있다.

데이터 마이닝은 데이터의 미처 몰랐던 속성을 발견하는 것에 집중한다.



Preview

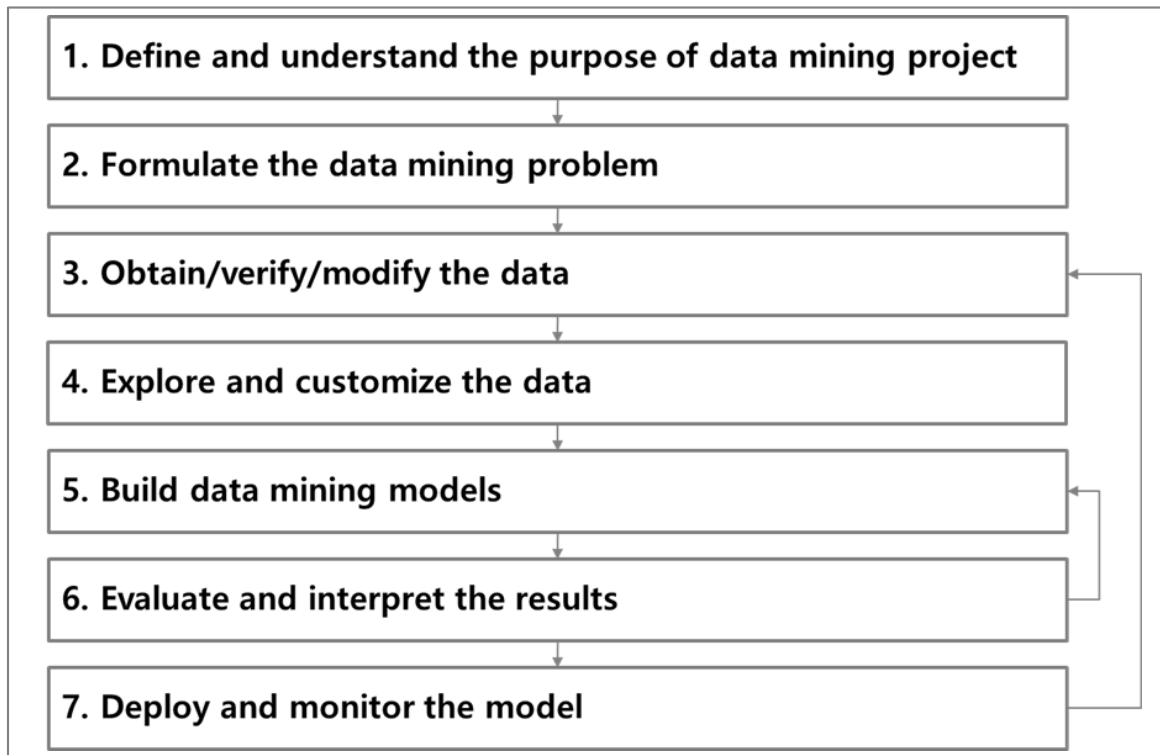
- ① **과학적 방법론** : 데이터사이언스는 딥러닝, 기계 학습, 관계 및 논리적 분석, 인공지능, 통계적 분석 등 대용량 데이터로부터 통찰력과 지식을 얻고 추리하기 위한
- ② **풀려고 하는 문제 영역의 지식을 바탕으로** 다양한 형식의 방대한 원천 데이터의 획득, 정제, 모델링, 통합 관리, 복합 분석, 기계 학습, 시각화 등 일련의 과정을 통해 인간과 사회에 유용한 디지털 솔루션을 만들어 적용하고
- ③ **지속적으로 개선하는 공학적 측면**을 포괄하는 새로운 학문이다.



개론 7 데이터 분석 방법

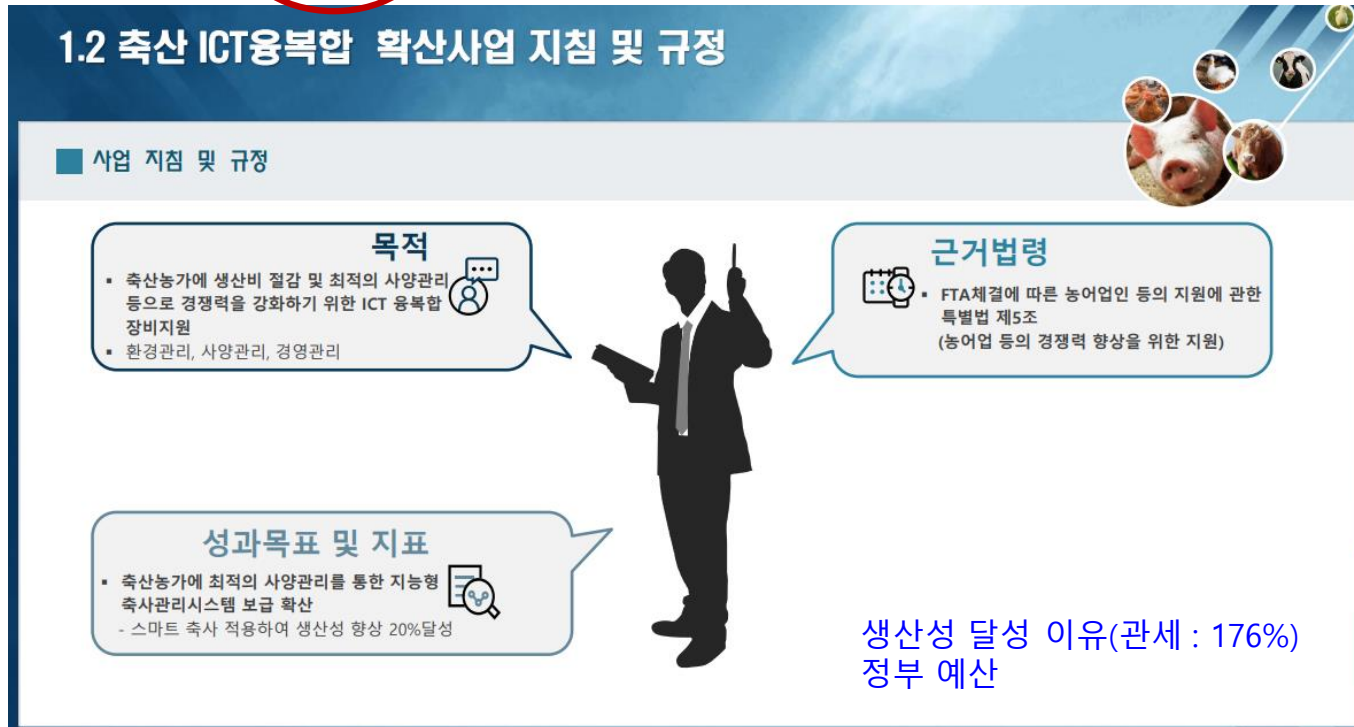
- 데이터 사이언스 프로젝트에서 **문제를 명확히 정의**하고 **데이터를 준비**했으면 답(가치)을 얻기 위해서 **데이터 분석**을 해야 한다.(p.166)
- 데이터 사이언스의 전체 과정에서 70~80%의 시간은 데이터를 모으고 준비하는데 소요된다.(p.32)

데이터 마이닝 Process



개론 7 데이터 분석 방법

■ 스마트팜의 목적



자유무역협정(FTA)란?

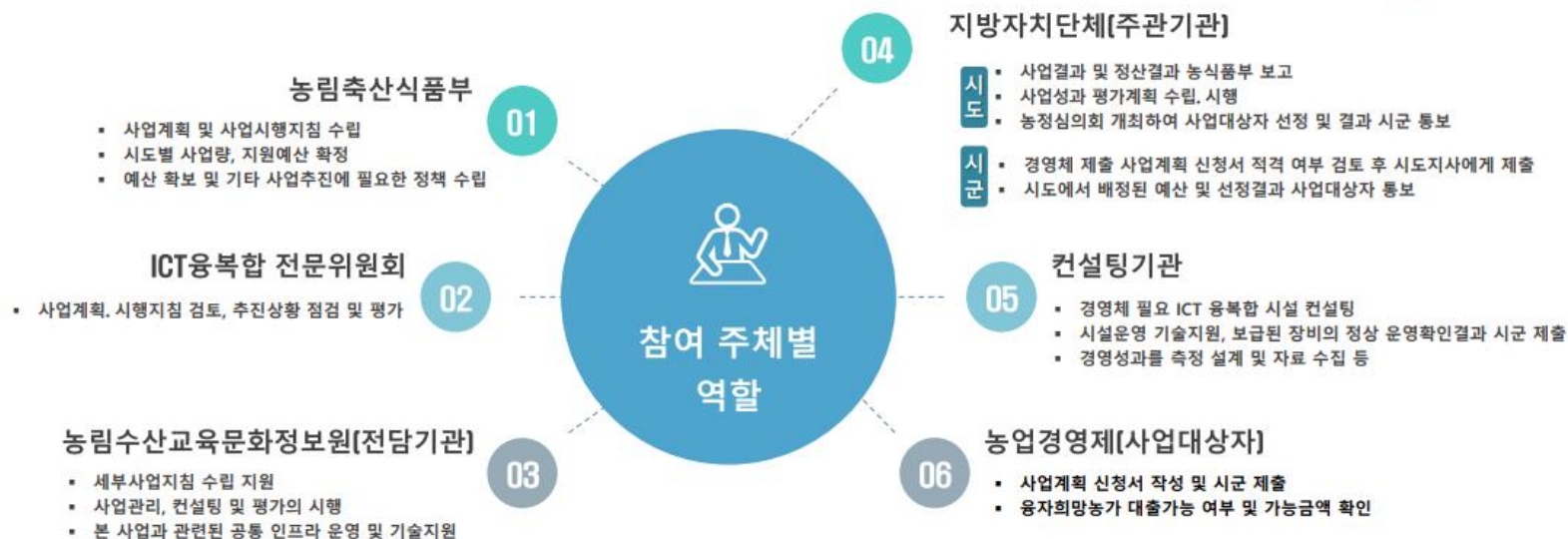
자유무역협정(FTA)은 협정을 체결한 국가 간에 상품/서비스 교역에 대한 관세 및 무역장벽을 철폐함으로써 배타적인 무역특혜를 서로 부여하는 협정입니다.

FTA는 그 동안 유럽연합(EU)이나, 북미자유무역(NAFTA)등과 같이 인접 국가나 일정한 지역을 중심으로 이루어졌기 때문에 흔히 지역무역협정(RTA:Regional Trade Agreement)이라고도 부릅니다 (출처:관세청 홈페이지)

개론 7 데이터 분석 방법

1.2 축산 ICT융복합 확산사업 지침 및 규정



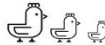
■ 참여 주체별 역할



개론 7 데이터 분석 방법

1.3 축산 ICT융복합 확산사업 지원 장비 안내

양돈, 양계, 낙농분야 주요 내용

분야	양돈	양계(산란계/육계)	양계(종계)	낙농
지원대상	축산업등록 경영체	축산업등록 경영체	축산업등록 경영체	축산업등록 경영체
지원내용	<ul style="list-style-type: none">▪ 군사급이기▪ 자동급이기▪ 사료믹스급이기▪ 컴퓨터엑상급이기▪ 돈선별기▪ 사료빈관리기▪ 음수관리기▪ 모돈발정체크기▪ 환경관리기 (온도, 습도, 정전, 화재 등) (팬, 쿨링패드, 냉방기, 난방기) 농장 기상대 (온도, 습도, 풍향, 풍속 등)▪ CCTV, 차량출입장치▪ 약취저감장치▪ 분뇨처리장치▪ 양돈생산경영관리 프로그램▪ PC	<ul style="list-style-type: none">▪ 부화기▪ 자동급이기(사료빈관리기 포함)▪ 자동급수기(음수관리기 포함)▪ 난선별기▪ 체중기▪ 사료빈관리기▪ 음수관리기▪ 계사환경관리기 (온도, 습도, 정전, 화재 등) (팬, 쿨링패드, 냉방기, 난방기 등)▪ 농장 기상대 (온도, 습도, 풍향, 풍속)▪ CCTV, 차량출입장치▪ 약취저감장치▪ 분뇨처리장치▪ 양계생산경영관리 프로그램▪ PC	<ul style="list-style-type: none">▪ 자동포유기▪ 체중측정기▪ 발정탐지기/분만알리미▪ 착유기▪ 로봇착유기▪ 자동급이기(개체급이정보 포함)▪ 음수관리기▪ 사료빈관리기▪ 조사료분석기/유성분 분석기▪ TMR배합기▪ 환경관리기, 농장기상대 (온도, 습도, 강우, 풍속, 화재 등) (팬, 안개분무기, 천장개폐기, 원치커텐)▪ CCTV▪ 차량출입장치▪ 분뇨처리장치▪ 낙농생산경영관리 프로그램▪ PC	  



개론 7 데이터 분석 방법

01

축산 빅데이터 플랫폼 개요

I 개요 및 구축 목적

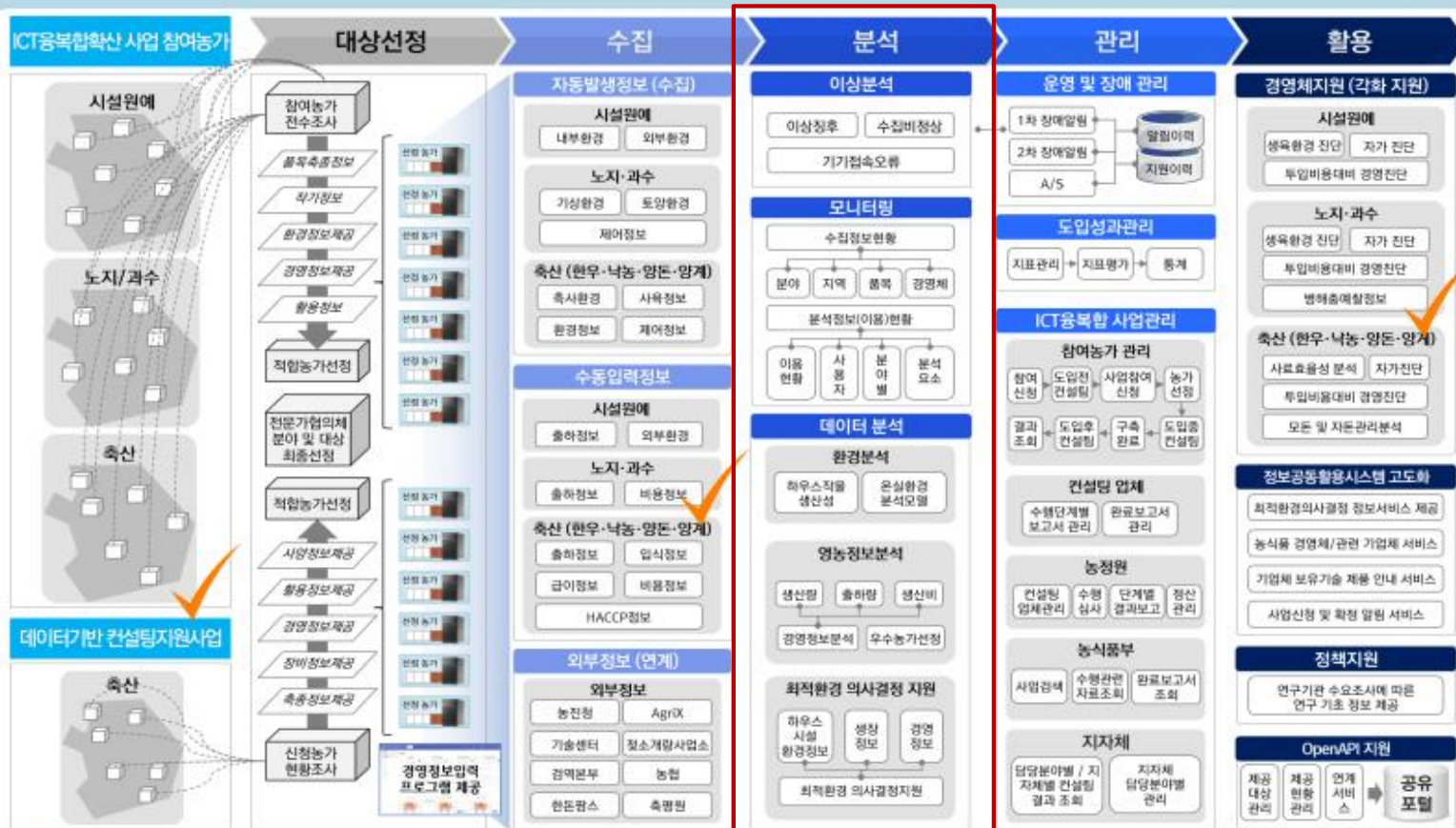
II

III

IV

축산 빅데이터 플랫폼 공통교육

농식품 ICT 적용 **축산 농가의 빅데이터를 수집하여**
농가 소득 향상을 위한 다양한 분석·활용 데이터 서비스 제공



개론 7 데이터 분석 방법

국가 기술 자격증 : 빅데이터분석기사(필기)

필기과목명	주요항목	세부항목	세세항목
빅데이터 분석 기획 (1)	빅데이터의 이해	빅데이터 개요 및 활용	빅데이터의 특징
			빅데이터의 가치
			데이터 산업의 이해
			빅데이터 조직 및 인력
		빅데이터 기술 및 제도	빅데이터 플랫폼
			빅데이터와 인공지능
	데이터분석 계획	분석방안수립	개인정보 법·제도
			개인정보 활용
		분석 작업 계획	분석 로드맵 설정
			분석 문제 정의
	데이터 수집 및 저장 계획	데이터 수집 및 전환	데이터 분석 방안
			데이터 확보 계획
			분석 절차 및 작업 계획
			데이터 수집
		데이터 적재 및 저장	데이터 유형 및 속성 파악
			데이터 변환
빅데이터 탐색 (2)	데이터 전처리	데이터 정제	데이터 비식별화
			데이터 품질 검증
		분석 변수 처리	데이터 적재
			데이터 저장
			데이터 정제
			데이터 결측값 처리
	데이터 탐색	데이터 탐색 기초	데이터 이상값 처리
			변수 선택
			차원 축소
			파생변수 생성
		고급 데이터 탐색	변수 변환
			불균형 데이터 처리
빅데이터 탐색 (2)	통계기법 이해	기술통계	데이터 탐색 개요
			상관관계 분석
			기초통계량 추출 및 이해
			시각적 데이터 탐색
		추론통계	시공간 데이터 탐색
			다변량 데이터 탐색
	빅데이터 결과분석 (4)	빅데이터 결과분석	비정형 데이터 탐색
			데이터 요약
			표본추출
			확률분포
	빅데이터 결과분석 (4)	빅데이터 결과분석	표본분포
			점추정
			구간추정

필기과목명	주요항목	세부항목	세세항목
빅데이터 분석 기획 (1)	분석모형 설계	분석 절차 수립	분석모형 선정
			분석모형 정의
		분석 환경 구축	분석모형 구축 절차
			분석 도구 선정
	분석기법 적용	분석기법	데이터 분할
			회귀분석
			로지스틱 회귀분석
			의사결정나무
			인공신경망
		고급 분석기법	서포트벡터머신
			연관성분석
			군집분석
			범주형 자료 분석
빅데이터 분석 기획 (1)	분석모형 설계	분석기법	다변량 분석
			시계열 분석
			베이지안 기법
			딥러닝 분석
	분석기법 적용	고급 분석기법	비정형 데이터 분석
			양상블 분석
			비모수 통계
			평가 지표
		분석모형 평가 및 개선	분석모형 진단
			교차 검증
			모수 유의성 검정
			적합도 검정
			과대적합 방지
빅데이터 분석 기획 (1)	분석모형 설계	분석기법	매개변수 최적화
			분석모형 융합
			최종모형 선정
			분석모형 해석
	분석기법 적용	고급 분석기법	비즈니스 기여도 평가
			시공간 시각화
			관계 시각화
			비교 시각화
		분석결과 해석 및 활용	인포그래픽
			분석모형 전개
			분석결과 활용 시나리오 개발
			분석모형 모니터링
			분석모형 리모델링

개론 7 데이터 분석 방법

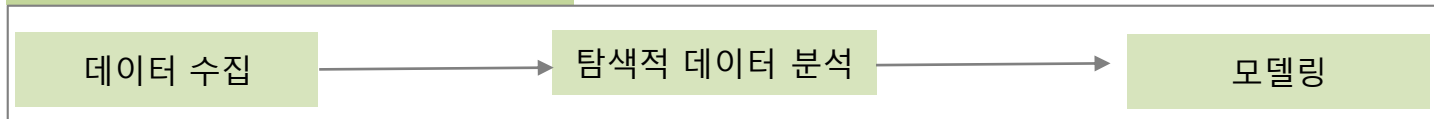
국가 기술 자격증 : 빅데이터분석기사(실기)

실기과목명	주요항목	세부항목	세세항목
빅데이터 분석 실무	데이터 수집 작업(1)	데이터 수집하기	정형, 반정형, 비정형 등 다양한 형태의 데이터를 읽을 수 있다.
			필요시 공개 데이터를 수집할 수 있다.
	데이터 전처리 작업(2)	데이터 정제하기	정제가 필요한 결측값, 이상값 등이 무엇인지 파악할 수 있다.
			결측값과 이상값에 대한 처리 기준을 정하고 제거 또는 임의의 값으로 대체할 수 있다.
		데이터 변환하기	데이터의 유형을 원하는 형태로 변환할 수 있다.
			데이터의 범위를 표준화 또는 정규화를 통해 일치시킬 수 있다.
			기존 변수를 이용하여 의미 있는 새로운 변수를 생성하거나 변수를 선택할 수 있다.
	데이터 모형 구축 작업(3)	분석모형 선택하기	다양한 분석모형을 이해할 수 있다.
			주어진 데이터와 분석 목적에 맞는 분석모형을 선택할 수 있다.
			선정모형에 필요한 가정 등을 이해할 수 있다.
		분석모형 구축하기	모형 구축에 부합하는 변수를 지정할 수 있다.
			모형 구축에 적합한 형태로 데이터를 조작할 수 있다.
			모형 구축에 적절한 매개변수를 지정할 수 있다.
	데이터 모형 평가 작업(4)	구축된 모형 평가하기	최종 모형을 선정하기 위해 필요한 모형 평가 지표들을 잘 사용할 수 있다.
			선택한 평가지표를 이용하여 구축된 여러 모형을 비교하고 선택할 수 있다.
			성능 향상을 위해 구축된 여러 모형을 적절하게 결합할 수 있다.
		분석결과 활용하기	최종모형 또는 분석결과를 해석할 수 있다.
			최종모형 또는 분석결과를 저장할 수 있다.

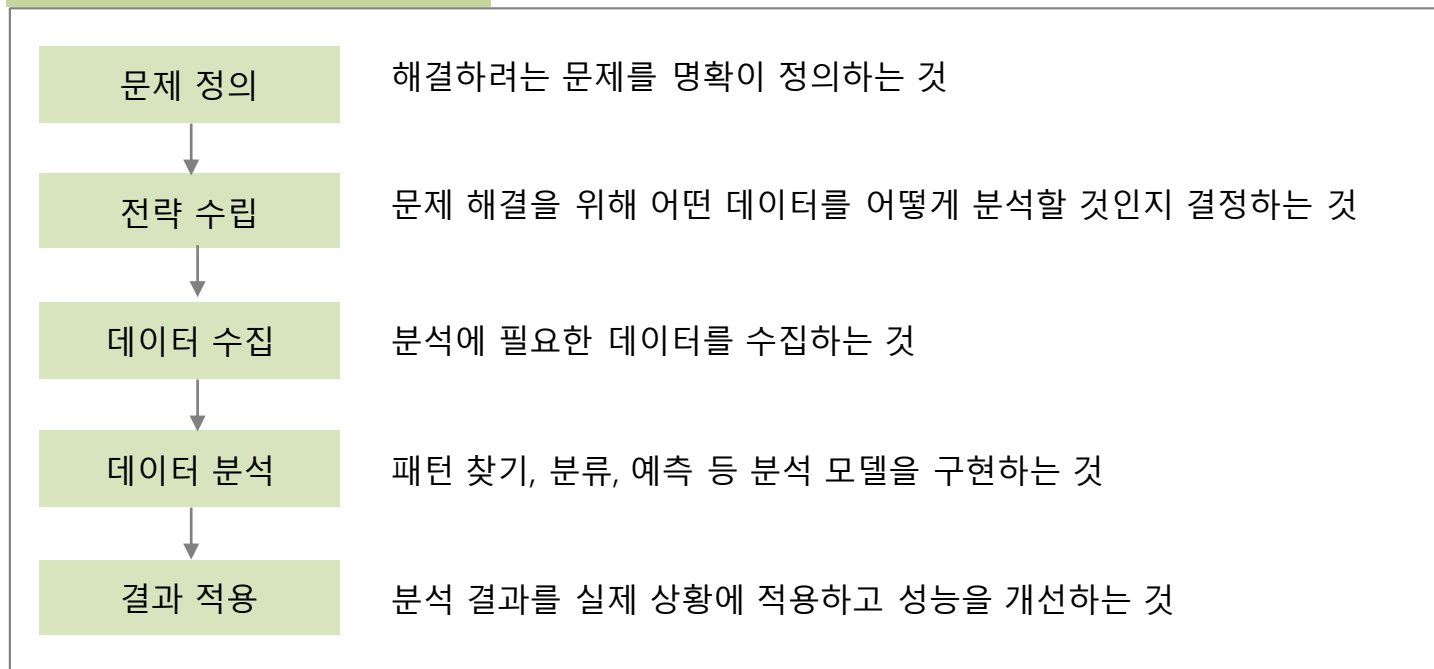


개론 7 데이터 분석 방법

데이터 과학의 절차(그림 1- 5)



데이터 사이언스 프로세스



Thank you

