

Data Science: 데이터 분석 프로세스

소개

데이터 분석을 한마디로 정의하면, “데이터의 패턴을 파악해 미래를 예측하고, 이로써 유리한 의사결정을 하기 위한 일련의 과정”이라 할 수 있습니다. 기업에 실질적인 성과를 가져다줄 의사결정은 몇몇 전문가의 거창한 분석만으로는 안 됩니다. 비즈니스의 전 부문에서 모든 업무를 보다 합리적으로, 더 좋은 성과가 기대되는 방향으로 결정하고 추진할 때 비로소 기업은 성장할 수 있습니다.

또한, 모든 구성원이 ‘분석에 적합한 데이터’가 무엇인지 이해해야 양질의 데이터가 쌓입니다. 제아무리 뛰어난 데이터 사이언티스트도 엉망진창인 데이터에서 빛나는 보석을 찾아낼 수 없습니다. 비전문가인 직원들이 데이터를 이해함으로써 기업이 확보하는 데이터의 품질이 높아집니다. 그제야 비로소 전문가의 고차원적인 분석도 가능해집니다.

데이터 분석의 프로세스와 데이터 취합을 다루는 이번 콘텐츠를 통해 데이터 사이언스 역량을 키워보시기 바랍니다.



목차

주요 내용	03
데이터 분석 프로세스	04
데이터 취합	07
데이터 전처리	08
데이터 전처리 필수 용어	10
요약(Summary)	11



주요 내용



데이터 분석은 데이터를 효과적으로 요약하는 기술입니다.



데이터 분석을 위해 가장 먼저 (1)분석의 목적을 정하고, (2)목적에 맞게 원하는 데이터를 수집합니다.



데이터를 확보한 후에는 (3)데이터 전처리 과정을 진행하고, 탐색적 데이터 분석 작업의 과정을 거칩니다.



(4)데이터 모델링을 통해 데이터를 요약하고, (5)데이터를 시각화하여 다른 사람들과 공유합니다.



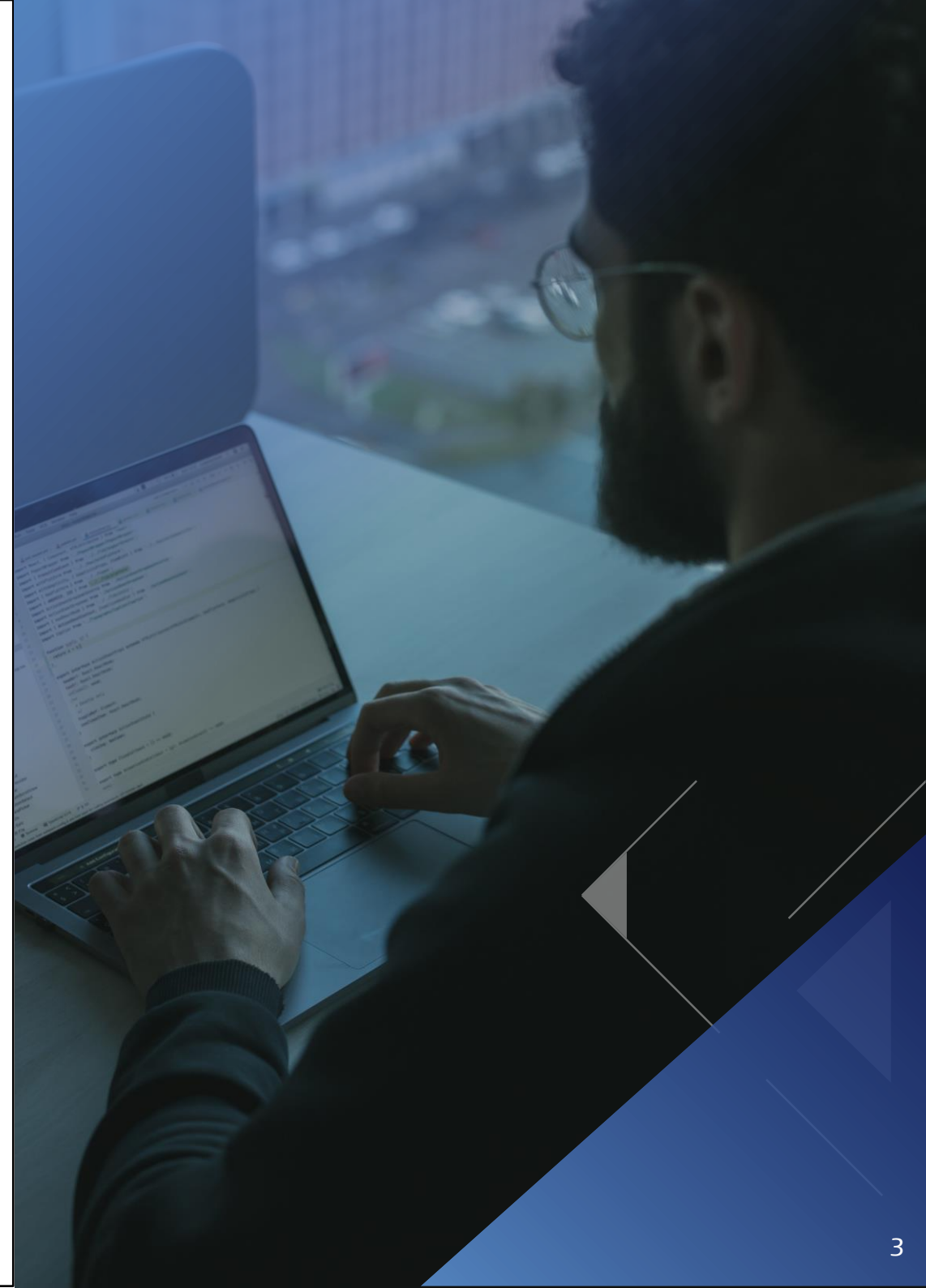
데이터 취합하는 방법으로는 (1)고객에게 직접 묻기 (2)시스템 활용하여 결과 데이터 확보 (3)외부 환경에서의 데이터 수집 (웹 크롤링)이 있습니다.



데이터 전처리는 데이터를 정제, 통합, 축소, 변환하는 과정을 거칩니다.



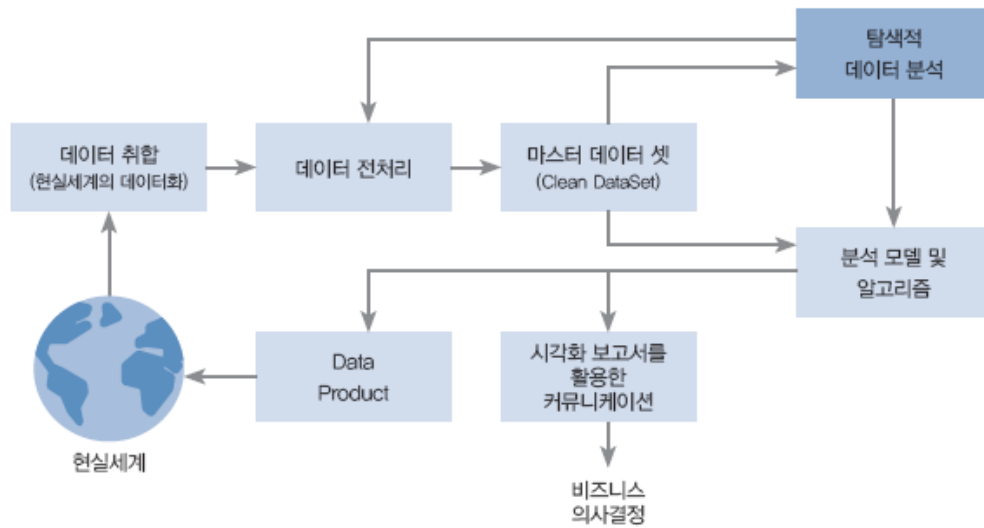
관련성, 연결성, 정확성, 충분한 규모를 기준으로 데이터를 정제, 통합, 축소, 변환해야 합니다.



데이터 분석 프로세스

데이터 분석이란, 데이터를 효과적으로 요약하는 기술이라고 정의할 수 있습니다.
이러한 데이터 분석은 데이터의 유형이나 도구, 분석 알고리즘과 무관하게 개념적으로는 언제나 유사한 과정을 밟습니다.

데이터 분석 프로세스는 다음과 같습니다.



데이터 분석 프로세스

(1) 분석 목적(가설) 수립

가장 먼저 데이터 분석의 목적을 정하고 문제를 정의해야 합니다. 목적이 없다면 데이터는 어떤 것도 말해줄 수 없습니다. 데이터 자체는 중립적이며, 우리의 의사결정에 영향을 주는 인사이트가 도출될 때 비로소 데이터 분석 작업이 가치를 지니는 것이기 때문입니다. 따라서 무엇을 알고 싶은지, 무엇을 예상하고자 하는지, 무엇을 하고 싶은지 등 분석의 목적이 가장 먼저 정해져야 합니다.

데이터 분석을 통해 좋은 답을 얻으려면 “다음 주 월요일 내 주식 가격은 얼마일까?” 처럼 회귀(Regression) 알고리즘을 염두에 둔 구체적인 질문을 던져야 합니다.

“와인을 구매하는 내 고객은 프랑스 와인을 선호할까, 이태리 와인을 선호할까?”와 같이 고객이 적절하게 분류되는 것이 의미 있는 상황이라면 클러스터링(Clustering) 알고리즘을 적용해서 데이터를 분류하는 시도가 필요합니다.

(2) 측정 도구의 개발 및 데이터 수집

데이터 분석의 목적이 정해지면 원하는 데이터를 수집합니다. 가령, 내일 날씨를 알고 싶으면 일주일 전후의 날씨나 일 년 전 해당 일의 날씨 데이터를 수집하는 식입니다.

사람들의 의견을 묻는 서베이로 데이터를 모으려면 설문지를 만들고, 객관적인 실험을 통해 데이터를 확보하려면 실험 시나리오와 환경을 설계하는 등 측정 도구를 개발합니다.

(3) 데이터 전처리 및 탐색적 데이터 분석

데이터 확보 후에는 수집된 데이터를 전반적으로 훑어 보면서 눈에 띄는 특징이나 관련성이 있는지, 혹시 누락되거나 잘못된 데이터가 있는지 등을 살펴봅니다. 다양한 경로로 수집된 데이터는 단위가 다르거나, 같은 내용을 가리키되 표현이 다르거나, 전체를 왜곡시킬 수 있는 매우 특이한 데이터가 포함되는 등 어수선하기 마련입니다. 이를 분석하기 좋게 다듬는 작업을 데이터 전처리라고 합니다. 이 과정에서 처음에는 미처 생각하지 못했던 아이디어를 얻을 수도 있고, 분석의 목적이 바뀌거나 구체화되기도 하며, 가설이 세워지기도 합니다.

데이터 전처리를 거친 정제된 데이터를 선입견 없이 다각도로 살펴보는 작업을 탐색적 데이터 분석이라 합니다. 간단한 데이터를 다룬다면, 탐색적 데이터 분석만으로도 쉽게 가설을 뒷받침할 패턴을 찾아낼 수 있습니다. 그러나 데이터 세트의 변수가 너무 많고 복잡하다면 특징만 추려 모델링하고, 분석 모델의 정확도를 높이는 훈련이 필요합니다.

취합한 데이터에서 가설을 뒷받침할 패턴을 발견하면, 단순히 수집된 표본 데이터만의 특징인지, 전체에 적용될 수 있는 경향성인지 확인하는 과정이 필요합니다. 추론 통계를 활용한 이 단계를 가설 검정 또는 유의성 검정이라고 합니다.

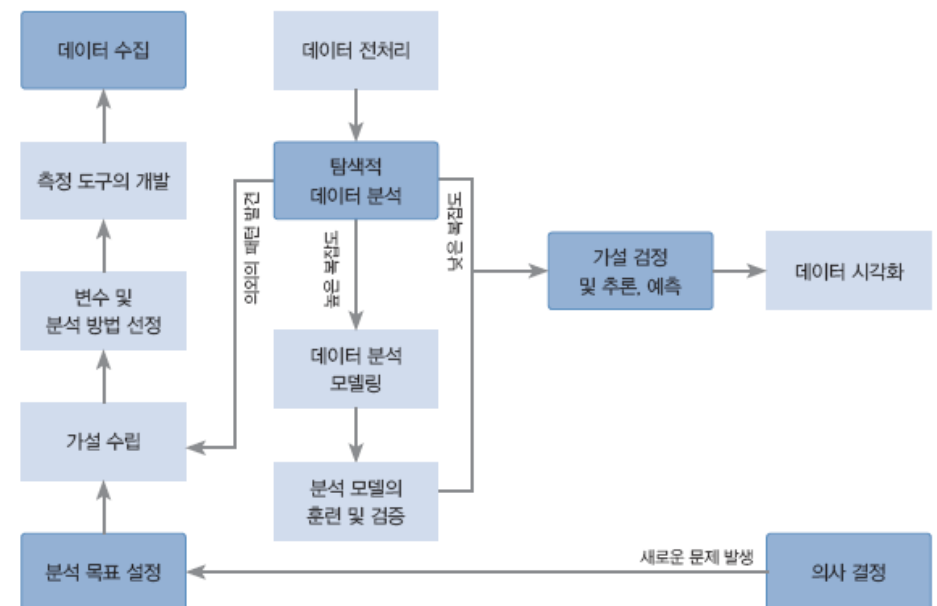
(4) 데이터 모델링

데이터 분석의 목적과 데이터의 특징이 명확히 파악되었다면, 데이터 요약과 정인 모델링이 필요합니다. 데이터 모델링이란, 문제의 특징을 잡아내 간결하게 만드는 과정입니다. 마치 복잡한 지도를 간단한 약도로 표현하는 것과 같습니다.

(5) 분석 결과 시각화 및 공유

데이터 분석의 목적은 데이터를 기반으로 어떤 현상에 대한 더 나은 이해를 얻어내거나 데이터 기반 추론을 통한 미래 예측을 하는 것입니다. 궁극적으로는 데이터 분석을 통해 도출해낸 정보를 타인과 교환함으로써 공통의 이해와 의사결정을 공유하는 것이라고 할 수 있습니다.

이를 위해서 데이터 시각화, 즉 누구나 알아보기 쉽고, 이해하고 동의할 수 있도록 설득력 있게 데이터 분석의 결과를 다양한 값과 그림으로 표현하는 단계가 필요합니다. 시각화된 분석 결과 보고서는 데이터를 직접 분석한 데이터 사이언티스트 뿐만 아니라 그 밖의 사람들도 의사결정을 진행할 때 유용하게 활용할 수 있습니다.



[데이터 분석 실무 흐름도]

데이터 취합

분야를 막론하고 데이터를 수집하는 방법은 크게 2가지로 나눌 수 있습니다.

- ① **Primary Data** : 당면한 구체적 목적을 해결하기 위해 새로 수집되는 자료
- ② **Secondary Data** : 다른 목적이나 용도로 과거에 수집되었던 조직화된 정보

<데이터를 얻는 효과적인 방법>

1. 고객에게 직접 묻기



사회과학에서는 관련 데이터를 설문지, 인터뷰, 심층 면접 등 여러 방법으로 취합해서 확보한 데이터를 기반으로 한 결론을 도출하는 방식을 많이 사용합니다. 최근에는 구글 설문지와 같은 온라인 서베이 도구를 활용하는 경우가 많아지고 있습니다.

2. 시스템 활용하여 결과 데이터 확보



고객의 방문 시각, 평균대기 시간, 일일 방문 고객 수, 평균 수신율, 방문 목적, 체류 시간, 적립액 등 다양한 객관적 지표를 얻어내는 것을 말합니다. 시스템을 이용해서 이러한 데이터를 문자와 숫자로 기록합니다.

개인 차원에서는 엑셀 표로, 기업 차원의 경우에는 CRM(Customer Relationship Management), ERP(Enterprise Resource Planning) 시스템을 운영해서 데이터베이스에 기록할 수 있습니다. 다양한 데이터를 구성원들이 쉽게 볼 수 있도록 BI(Business Intelligence) 도구들이 활용되기도 합니다.

3. 외부 환경에서의 데이터 수집 (웹 크롤링)



외부 데이터는 대개 온라인에 공개된 데이터를 참고하게 되는데, 이러한 데이터를 가져오는 작업을 크롤링(Crawling)이라고 합니다. 보통 봇, 크롤러를 이용해 자동 처리합니다. 필요한 크롤링 봇을 만드는 것 또한 데이터 분석 기술의 한 분야를 차지하기도 합니다.

크롤링은 어려운 작업처럼 보이지만, 흔히 쓰는 엑셀 (Excel) 프로그램을 이용해 누구나 할 수 있습니다. 엑셀 2010 PRO 이후 버전을 쓰는 사람이라면 [데이터] 탭에서 "외부 데이터 가져오기"에서 데이터 가져올 곳을 "웹"으로 선택하면 웹 크롤링을 할 수 있습니다. 물론 복잡한 데이터 수집은 파이썬이나 R 등을 이용해 별도로 만들어야 하지만, 간단한 데이터 수집은 엑셀 크롤러를 이용해 누구나 간단하게 할 수 있습니다.

데이터 전처리

여러 출처로부터 다양한 데이터를 수집하게 되면, 데이터 형식이 제각각이거나, 결측치가 존재하는 등 분석에 바로 활용하기 어려운 경우일 가능성이 큼니다. 이를 분석에 알맞은 형태로 가공하는 작업을 **데이터 전처리**라고 합니다. 데이터를 정제, 통합, 축소, 변환하는 과정을 거칩니다.

- (1) **데이터 정제**: 데이터 내의 잡음을 제거하고, 일관성 결여를 교정
- (2) **데이터 통합**: 여러 데이터의 원천에서 발생하는 데이터를 일관성 있는 데이터 저장소나 통합 파일에 저장
- (3) **데이터 축소**: 집계, 중복제거, 군집화 등의 과정을 통한 데이터 크기 축소
- (4) **데이터 변환**: 데이터를 하나의 표현 형식에서 다른 형식으로 변환



데이터 전처리

그렇다면, 데이터를 어떤 기준으로 정제, 통합, 축소, 변환하는 것일까요?

- (1) **관련성**: 분석하기 전, 원하는 바를 구체적으로 정하고 분석 목적에 맞는 데이터를 정리해야 합니다.
- (2) **연결성**: 각 자료는 주어진 변수에 적합한 데이터를 갖는 것이 좋습니다. 데이터가 활용 가치를 가지려면 누락된 값이 없거나 최소한이어야 합니다.
- (3) **정확성**: 데이터 자체에 오류가 있다면 아무리 좋은 알고리즘이라도 원하는 결과를 얻기 어렵습니다.
- (4) **충분한 규모**: 기초적인 통계든 머신러닝이든 데이터의 양이 충분하지 않으면 원하는 결과를 도출하기 어렵습니다. 특히 머신러닝에서는 훈련 데이터와 검증 데이터를 구분하는 경우가 많으므로 데이터의 규모를 어느 정도 확보하는 것이 필수적입니다.



데이터 전처리 시, 반드시 알아야 할 용어

(1) 변수 (Variable): 측정할 수 있는 수량, 품질, 속성 (ex. 근무 직원의 연령)

(2) 값 (Value): 변수 측정 시 변수의 상태. 변수의 값은 첫 측정치에서 다른 측정치로 변경될 수 있음. (ex. 근무 직원의 연령 = 27세, 30세)

(3) 관찰 (Observation): 유사한 조건에서 시행되는 일련의 측정 세트. 관측치에서는 각각 다른 변수 및 여러 값이 포함됨.

(4) 표 (Tabular): 열과 행으로 구성된 데이터 요약 형식. "열"에는 각각의 변수가 배치되고, "행"에는 각 관찰 항목에 대한 값이 배치됨. (ex. 집값과 관련성이 있는 요인 표)



요약(Summary)

오늘날 사람들은 온라인에서 활동하고, 수많은 데이터가 온라인에 쌓이고 있습니다. 더불어 추적기술과 분석 기술이 발달하면서 기업은 온라인에 남겨진 고객의 발자취를 따라 그들의 마음을 짐작하고 읽어낼 수 있게 되었습니다. 이러한 추세로 인해 데이터 분석에 대한 수요는 나날이 높아지고 있고, 데이터 분석을 배우고자 하는 사람도 계속 증가하고 있습니다. 개인의 라이프스타일을 토대로 개개인의 행동 맥락을 알고리즘이 파악하여 맞춤형 마케팅을 펼치는 데이터 분석 역량은 이미 대기업에서는 매우 중요해진 지 오래입니다.

데이터 분석은 데이터를 효과적으로 요약하고, 패턴을 파악하여 미래를 예측하는 과정입니다. 이러한 데이터 분석의 목적은 더 나은 의사결정입니다. 분석을 위한 분석은 의미가 없기에 실질적인 이익을 가져오는 분석의 목적이 명확해야 합니다. 데이터를 수집했다면 데이터 전처리를 통해 정제된 데이터를 선입견 없이 다각도로 살펴보는 작업을 거쳐 데이터의 패턴을 찾아냅니다. 데이터 분석의 목적과 데이터의 특징이 명확히 파악되었다면, 데이터 모델링을 통해 데이터를 요약하고, 마지막으로 분석의 결과를 시각화하여 보기 좋게 표현해야 합니다.

데이터 분석의 목적은 데이터를 기반으로 어떤 현상에 대한 더 나은 이해를 얻어 내거나 데이터 기반 추론을 통한 미래 예측을 하는 것입니다. 궁극적으로는 데이터 분석을 통해 도출해낸 정보를 타인과 교환함으로써 공통의 이해와 의사결정을 공유하는 것입니다.



마소캠퍼스는 양질의 교육을 합리적인 가격으로 제공하는
온오프믹스형 IT 실무교육센터입니다.

마소캠퍼스 교육과정의 특징



실무 사례를 토대로 이뤄지는 지식 전달



PC와 Mobile로 언제 어디서나 편리하게 보는 온라인 강의



IT 세상의 트렌드를 따라 잡는 웹북 서비스

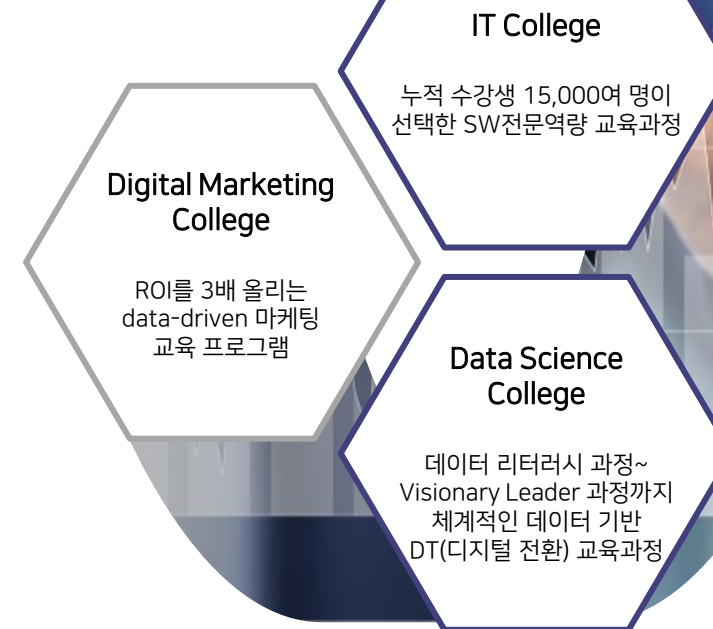
더 많은 콘텐츠 보기

디지털마케팅&데이터분석
국내 1위 교육기관

마소캠퍼스

- 네이버,삼성,LG...173개 기업과 81개 대학이 선택한 전문성
- 월평균 수강생 750명, 속이 꽉 찬 83개의 실무 향상 프로그램
- ROI 증대 효과가 검증된 Best Practice 위주의 커리큘럼
- 비슷한 주제의 타기관 대비 80% 수준인 합리적인 가격
- PC, 모바일로 오프라인과 동일한 경험을 하는 O2O 강의
- 최대 1,000명이 동시 접속 가능한 마소웨бина 시스템

Actionable Content



<https://www.masocampus.com>