합격의 선을 넘는

2026

선넘는

ADsP

데이터분석 준전문가

'시험장 필수품'

시험에 꼭 나오는 시크릿 요약집

아답터의
Youtube 무료 강의

문풀 올인원
APP 제공

공석민 편저

쏠터북스

1 과목

데이터의 이해 – 데이터와 정보/데이터베이스

1 데이터와 정보

- **데이터의 정의:** 가공되지 않은 상태의 객관적 사실(Fact)
- **정보의 정의:** 데이터로부터 가공 및 의미 부여가 된 자료
- **데이터의 특성:** 현실을 반영하는 '존재적 특성'과 의미 부여가 가능한 '당위적 특성'
- **DIKW 피라미드:**
  - 데이터(Data): 가공 전의 객관적 사실
  - 정보(Information): 데이터를 통한 패턴 인식
  - 지식(Knowledge): 패턴을 통한 의사결정 활용
  - 지혜(Wisdom): 창의적인 전략 도출
- **데이터의 유형:**
  - 형태: 정형(RDB, 엑셀), 반정형(HTML, JSON, XML), 비정형(SNS, 유튜브, 음원)
  - 표현: 정량적(수치, 기호), 정성적(언어, 문자)
  - 분석 목적: 수치형(연속형, 이산형), 범주형(명목형, 순서형)
- **지식의 상호작용:** 암묵지(개인 습득)와 형식지(문서화)의 공통화, 표출화, 연결화, 내면화 과정

2 데이터베이스의 정의와 특징

- **데이터베이스 특징:** 공용 데이터, 통합된 데이터, 저장된 데이터, 변화되는 데이터(무결성 유지)
- **데이터베이스 설계 절차:** 요구조건 분석 → 개념적 설계(ERD) → 논리적 설계(모델링) → 물리적 설계(물리공간 저장 구조)
- **관계형 DBMS vs NoSQL:** 테이블 구조(MySQL, Oracle) vs 비정형 데이터 처리(MongoDB, Redis 등)
- **주요 용어:** 스키마(구조 명세), 메타데이터(데이터를 설명하는 데이터), 데이터 사전(구조 정보 저장소), 인스턴스(데이터), 인덱스(탐색 도구)

1 과목

데이터베이스 활용 / 빅데이터의 가치와 미래

1 데이터베이스 활용

- **기업 내부 활용:** CRM(고객 분석), SCM(공급망 최적화), ERP(경영 효율화), RTE(신속 의사결정), BI(리포트 도구), BA(통계적 통찰), KMS(지식 통합), 블록체인(분산 저장)
- **데이터웨어하우스(DW):** 중앙 저장소. 주제지향성, 데이터 통합, 시계열성, 비휘발성

2 빅데이터의 이해 및 가치

- **3V (Gartner):** Volume(규모), Variety(다양성), Velocity(속도)
- **빅데이터 비유:** 석탄/철, 원유, 렌즈, 플랫폼
- **빅데이터의 변화:** 전수조사, 사후처리, 양 중심 분석, 상관관계 분석 지향
- **분석 기법:** 회귀분석, 분류분석, 연관규칙, 유전자 알고리즘, 기계학습, 감정분석, 소셜 네트워크 분석, 텍스트 마이닝

3 위기 요인과 통제 방안

- **위기 요인:** 사생활 침해, 책임 원칙 훼손(범죄 예측 등), 데이터 오용(분석 결과 맹신)
- **통제 방안:** 사용자 책임 전환, 결과 기반 책임 고수, 알고리즈미스트(피해자 구제 전문인력) 활용
- **데이터 3법:** 개인정보보호법, 정보통신망법, 신용정보법. 가명정보 개념 도입
- **비식별화:** 가명처리, 총계처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹

2 과목

데이터분석 기획의 이해 – 방향성 / 방법론

1 분석 기획 방향성 도출

- **분석 대상(What)과 방법(How) 유형:**
  - 최적화(Optimization): 대상과 방법 모두 인지
  - 통찰(Insight): 대상 미인지, 방법 인지
  - 솔루션(Solution): 대상 인지, 방법 미인지
  - 발견(Discovery): 대상과 방법 모두 미인지
- **기획 시 고려사항:** 가용 데이터 파악, 유스케이스 탐색, 장애요소 사전 계획 수립

2 분석 방법론

- **구성요소:** 절차, 방법, 도구와 기법, 템플릿과 산출물
- **모델 유형:**
  - 폭포수: 이전 단계 완료 후 진행
  - 나선형: 반복을 통한 위험 요소 제거 및 점진적 완성
  - 애자일: 짧은 주기의 반복적 개발과 고객 피드백 반영
- **KDD 절차:** 데이터 선택 → 전처리 → 변환 → 마이닝 → 평가
- **CRISP-DM 절차:** 업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 → 전개

3 빅데이터 분석 방법론 5단계

- **기획(Planning):** 비즈니스 이해, 범위 설정, 프로젝트 정의, 위험 계획
- **준비(Preparing):** 필요 데이터 정의, 데이터 스토어 설계, 수집 및 정합성 점검
- **분석(Analyzing):** 데이터 준비, 텍스트 분석, 탐색적 분석(EDA), 모델링, 평가 및 검증
- **시스템 구현(Developing):** 설계 및 구현, 시스템 테스트
- **평가 및 전개(Deploying):** 모델 발전 계획 수립, 프로젝트 평가 및 보고

2 과목

분석 과제 발굴 / 마스터 플랜 / 거버넌스

1 분석 과제 발굴

- **하향식 접근법(Top-Down):** 문제 탐색(STEEP, 비즈니스 모델 기반) → 문제 정의 → 해결방안 탐색 → 타당성 검토(경제적, 데이터, 기술적)
- **상향식 접근법(Bottom-Up):** 문제 정의가 어려울 때 데이터를 통한 사물 인식. 비지도학습 및 프로토타이핑 접근법 활용
- **디자인 싱킹:** 공감 → 문제 정의 → 아이디어 도출 → 프로토타입 → 테스트

2 분석 프로젝트 관리 및 마스터 플랜

- **관리 요소:** 데이터 양, 속도, 데이터/분석 복잡성, 정확도와 정밀도(Trade-Off)
- **마스터 플랜 우선순위:**
  - 시급성: 비즈니스 효과(Return)
  - 난이도: 투자비용 요소(Investment: 데이터 규모, 다양성, 속도)
- **전략 로드맵:** 체계 도입 → 유효성 검증(파일럿) → 확산 및 고도화

3 분석 거버넌스 체계

- **구성요소:** 조직, 프로세스, 시스템, 데이터, 교육 및 마인드 육성
- **분석 준비도:** 분석 기법, 인력 및 조직, IT 인프라 등 6개 영역 진단
- **분석 성숙도:** CMMI 기반 5단계 수준 진단
- **분석 수준 진단 결과:** 도입형, 준비형, 정착형, 확산형
- **데이터 거버넌스 체계:** 데이터 표준화, 관리 체계, 저장소 관리, 표준화 활동
- **분석 조직 구조:** 집중 구조(전담), 기능 구조(현업 직접), 분산 구조(인력 배치)

3 과목

데이터분석 - 데이터 마트 / 전처리

1 데이터 마트(DM)

- **데이터 마트:** 특정 주제 중심의 소규모 데이터 웨어하우스
- **요약변수:** 수집 정보의 단순 종합 (ex 월간 수입)
- **파생변수:** 특정 의미를 부여한 사용자 정의 변수 (ex 고객구매등급)

2 탐색적 자료 분석(EDA)

- **4가지 주제:** 저항성 강조, 잔차 계산, 자료변수의 재표현, 그래프를 통한 현시성

3 결측값 처리

- **단순대치법:**
  - 완전분석법(삭제), 평균 대치법(조건부/비조건부)
  - 단순 확률 대치법: Hot-Deck(현재 데이터셋 유사치), Cold-Deck(외부 유사치), Nearest Neighbor
- **다중대치법:** 대치 → 분석 → 결합 과정을 여러 번 반복하여 분산 과소 추정 문제 해결

4 이상값(Outlier) 검색

- **ESD (Extreme Studentized Deviation):** 평균 ±3 × 표준편차를 벗어나는 값
- **사분위수(IQR):** Q1 - 1.5 × IQR보다 작거나 Q3 + 1.5 × IQR보다 큰 경우
  - 상한, 하한, 최솟값, 최댓값, 1~3사분위값을 표현하며 평균은 표현하지 않음
- **Z-Score:** 데이터 표준화 후 임계값 초과 여부 판단
- **DBScan:** 밀도 기반으로 밀도가 낮은 데이터를 이상값으로 판단

3 과목

데이터분석 – 통계학 개론 / 추론 통계

1 자료의 척도와 확률

- **질적 척도:** 명목(성별), 순서(학년)
- **양적 척도:** 등간(온도), 비율(무게 - 절대 0 존재, 사칙연산 가능)
- **확률의 기초:** 덧셈 정리, 조건부 확률 P(A|B), 독립 사건, 배반 사건
- **확률 변수:** 이산확률변수(셀 수 있음), 연속확률변수(셀 수 없음)

2 기초 통계량

- **중심경향성:** 평균, 중앙값, 최빈값
- **분산 정도:** 범위, 분산, 표준편차, 사분위수(IQR), 변동계수(CV)
- **관계 측정:** 공분산(상관 정도), 상관계수(-1~1 사이의 표준화된 상관성)
- **첨도와 왜도:** 첨도(뾰족한 정도), 왜도(비대칭 정도, 0이면 대칭)
  - 왜도 > 0: 최빈값 < 중앙값 < 평균값 (오른쪽 꼬리)
  - 왜도 < 0: 최빈값 > 중앙값 > 평균값 (왼쪽 꼬리)

3 확률 분포

- **이산확률분포:** 베르누이, 이항, 기하, 음이항, 초기하, 다항, 포아송분포
- **연속확률분포:** 균일, 정규, t(작은 표본 평균 검정), 카이제곱(모분산 검정), F(분산 비교), 지수, 감마, 베타분포
- **중심극한정리:** 표본이 30개 이상이면 표본평균분포가 정규분포를 이룸

4 추론 및 가설 검정

- **점추정 조건:** 불편성, 효율성, 일치성, 충족성
- **가설 검정:**
  - 귀무가설(H0: 차이가 없다), 대립가설(H1: 차이가 있다)
  - 유의수준(α): 귀무가설이 참일 때 기각하는 1종 오류 허용 한계
  - p-value < α 이면 귀무가설 기각
- **t – 검정:** 단일표본(평균 검정), 대응표본(동일 모집단 전후 비교), 독립표본(서로 다른 모집단 비교)

3 과목

데이터분석 - 회귀 분석 / 다변량 분석

1 회귀 분석 (Regression)

- **개념:** 독립변수(x)와 종속변수(y) 간의 상관관계 분석
- **최소제곱법:** 잔차 제곱합(SSE)이 최소가 되는 회귀계수 추정
- **결정계수(R^2):** 총 변동(SST) 중 모형에 의해 설명되는 변동(SSR)의 비율
- **회귀 분석 종류:** 단순, 다중, 다항 회귀분석
- **규제 회귀:** 릿지(L2 - 제곱합), 라쏘(L1 - 절댓값), 엘라스틱넷(둘 다 사용)
- **회귀 분석 가정:** 선형성, 등분산성, 정규성, 독립성

2 다중공선성 및 모형 선택

- **진단:** 상관계수 0.8 이상, VIF 10 이상, 조건수 30 이상
- **해결:** 변수 제거, 차원 축소, 규제 회귀 활용
- **최적 방정식 탐색:** 전진선택, 후진제거, 단계별 선택법(AIC, BIC 고려)
  - AIC, BIC는 작을수록 좋음
- **분산분석(ANOVA)표:** F통계량으로 모형의 유의성 검정. 데이터 수 = 자유도 + 1

3 상관분석 및 주성분 분석(PCA)

- **피어슨 상관분석:** 양적 척도, 연속형 변수, 선형 관계 측정
- **스피어만 상관분석:** 서열 척도, 순서형 변수, 비선형 관계 가능
- **주성분 분석:** 차원 축소 기법. 분산이 가장 큰 축이 첫 번째 주성분
  - 각 주성분은 서로 직교(독립)
  - 스크리플롯에서 기울기가 완만해지기 바로 전까지를 개수로 선택
- **다차원 척도법(MDS):** 거리 정보를 보존하며 시각화. Stress 값이 0에 가까울수록 좋음

3 과목

데이터분석 – 시계열 예측 / 데이터 마이닝 개요

1 시계열 분석

- **변동 요인:** 추세 요인, 계절 요인, 순환 요인, 불규칙 요인
- **정상성 조건:** 시점에 관계없이 일정한 평균과 분산 유지
  - 차분: 평균 안정화 (현 시점 - 이전 시점)
  - 변환: 로그/제곱근 변환으로 분산 안정화
- **백색 잡음:** 평균 및 분산 일정, 자기상관 없음, 정상성 만족

2 시계열 모형

- **자기회귀(AR):** 과거 값이 미래를 결정. 부분자기상관함수(PACF) 활용
- **이동평균(MA):** 백색잡음의 선형결합. 자기상관함수(ACF) 활용
- **ARIMA(p, d, q):** AR(p), MA(q), 차분 횟수(d)의 결합
  - d = 0 이면 ARMA, p = 0 이면 IMA, q = 0 이면 ARI 모델

3 데이터 마이닝 개요

- **학습 유형:**
  - 지도학습: 분류(KNN, SVM, 의사결정나무 등), 회귀
  - 비지도학습: 군집분석, 연관규칙, 차원축소
- **과대적합 vs 과소적합:** 복잡도와 오차 사이의 Trade-Off. 분산과 편향 관계
- **검증 방법:**
  - 홀드아웃: 훈련/검증용으로 1회 분할
  - K-fold 교차검증: k개 집단으로 나누어 교차 학습/검증
  - LOOCV: 1개 데이터로만 검증
  - 부트스트래핑: 복원추출 활용 데이터셋 생성

3 과목

데이터분석 - 분류분석 / 인공신경망(ANN)

1 분류분석 알고리즘

- **로지스틱 회귀:** 종속변수가 범주형일 때 0~1 사이 확률 도출. 오즈(Odds), 로짓(logit) 변환, 시그모이드 함수 활용
- **KNN (K-Nearest Neighbors):** 거리 기반 이웃에 따른 분류. Lazy Model
- **나이브베이즈:** 베이즈 정리를 기반 각 범주에 속할 확률. 독립성 가정 필요
- **의사결정나무(Decision Tree):** 노드 내 동질성 극대화. 지니지수(CART), 엔트로피(C4.5) 활용. 정지규칙과 가지치기로 과적합 방지
- **서포트벡터머신(SVM):** 마진 최대화 초평면 탐색. 하드마진(오류 비허용)과 소프트마진 구분

2 앙상블 (Ensemble)

- **배깅(Bagging):** 복원추출(Bootstrap) 후 투표(Voting). 랜덤포레스트(의사결정나무+배깅)가 대표적
- **부스팅(Boosting):** 오분류 데이터에 가중치 부여. AdaBoost, GBM, XGBoost, Light GBM
- **스태킹(Stacking):** 개별 모델 결과를 메타 학습기로 다시 학습

3 인공신경망 (ANN)

- **구조:** 입력층, 은닉층(하이퍼파라미터), 출력층
- **퍼셉트론:** 단층 퍼셉트론은 XOR 문제 해결 불가
- **학습 원리:**
  - 순전파: 정보가 전방 전달
  - 역전파: 가중치 수정을 통한 손실함수 최소화
  - 경사하강법: 편미분 활용 최적 해 탐색
- **활성함수:**
  - 은닉층: 시그모이드, 하이퍼볼릭 탄젠트(Tanh), ReLU(기울기 소실 극복)
  - 출력층: 시그모이드(이진분류), 소프트맥스(다중분류)
- **손실함수:** MSE(회귀), 크로스 엔트로피(분류)

3 과목

데이터분석 - 평가 / 군집분석 / 연관분석

1 분류모델 평가지표

- **오분류표(혼동행렬):** TP, FP, FN, TN
- **지표 계산:**
  - 정밀도(Precision): TP / (TP + FP)
  - 재현율(Recall): TP / (TP + FN). 민감도, TPR과 동일
  - 특이도(Specificity): TN / (FP + TN)
  - 정확도(Accuracy): (TP + TN) / (TP + FP + FN + TN)
  - F1-Score: 정밀도와 재현율의 조화평균
- **ROC 커브:** 가로축(1 - 특이도), 세로축(민감도). 면적(AUC)이 1에 가까울수록 성능 우수
- **이익도표(Lift Chart):** 등급별 반응률 및 리프트 산출

2 군집분석 (Clustering)

- **계층적 방법:** 덴드로그램 활용. 최단/최장/평균/중심/와드 연결법
- **비계층적 방법:**
  - K-평균: 군집 수 K 사전 지정. 평균으로 중심점 재설정
  - DBSCAN: 밀도 기반. 이상치에 강함. Eps, MinPts 설정
- **기타:** 퍼지군집화(확률), EM알고리즘(분포), SOM(자기조직화지도-신경망/승자독식)
- **평가지표:** 실루엣 계수(-1 ~ 1). 1에 가까울수록 군집화 우수

3 연관분석 (Association)

- **지표:**
  - 지지도: $P(A \cap B)$ (전체 중 동시 포함 비율)
  - 신뢰도: $P(A \cap B)/P(A)$ (A 구매 시 B 포함 확률)
  - 향상도: $P(A \cap B)/(P(A)P(B))$ (1보다 크면 양의 상관관계)
- **알고리즘:** Apriori(최소 지지도 활용 연산 감소), FP-Growth(트리 기반 효율 향상)

Light & Impact