1. 빅데이터 분석 기획

1.1 빅데이터의 이해

[문1] '큰 용량과 복잡성으로 기존 애플리케이션이나 툴로는 다루기 어려운 데이터셋의 집합'을 무엇이라고 하는가?

해설 빅데이터에 대한 설명입니다. 위키백과에서는 빅데이터를 '기존 데이터베이스 관리도구의 능력을 넘어서는 대량(수십 테라바이트)의 정형 또는 심지어 데이터베이스 형태가 아닌 비정형의 데이터 집합조차 포함한 데이터로부터 가치를 추출하고 결과를 분석하는 기술'이라고 정의하고 있습니다.
정답 빅데이터

[문2] '추론과 추정의 근거를 이루는 객관적 사실' 또는 '현실에서 관찰하거나 측정하여 수집한 사실'을 무엇이라고 하는가?

해설 데이터(data)는 '현실 세계에서 측정하고 수집한 사실이나 값'을 의미합니다. 반면 정보(information)는 '어떠한 목적이나 의도에 맞게 데이터를 가공 처리한 것'입니다. 객관적인 것인가, 주관적 목적과 의도가 개입했는가에 따라 두 개념의 차이를 구분해야 합니다.
정답 데이터

[문3] '지식을 도출할 때 사용하는 데이터'이며 '데이터의 가공 및 데이터 간 관계를 통해 패턴을 인식하는 것'을 무엇이라고 하는가?

해설 정보에 대한 설명입니다. 정보는 '가공처리된 것'이라는 핵심 키워드에 유념해야 합니다.
정답 정보

[문4] '데이터를 통해 도출된 다양한 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험을 결합해 내재화한 것'을 무엇이라고 하는가?

해설 지식은 정보를 구조화 또는 분류한 결과물이라는 것이 핵심 키워드입니다.
정답 지식

[문5] '지식의 축적과 아이디어가 결합된 창의적 산물'을 무엇이라고 하는가?

해설 지식과 창의적인 아이디어가 결합된 것을 지혜라고 합니다.
정답 지혜

[문6] 빅데이터의 특징으로서 말하는 3V는 무엇인가?

해설 3V는 양, 다양성, 속도를 각각 의미합니다. 최근에는 5V 또는 8V도 말하고 있으나 3V가 공통적인 빅데이터의 특징이라고 할 수 있습니다.
정답 Volume(크기), Variety(다양성), Velocity(속도)

[문7] '자신에 관한 정보가 언제, 어떻게 그리고 어느 범위까지 타인에게 전달되고 이용될 수 있는지를 정보 주체가 스스로 결정할 수 있는 권리'를 무엇이라고 하는가?

해설 개인정보의 주인(소유자)인 내가 주체적으로 정보가 어디까지 전달되어 활용될 수 있는지를 자유롭게 결정할 수 있는 권리입니다. 개인정보를 내가 선택적으로 제공할 수 있어야 합니다.
정답 개인정보 자기결정권

[문8] '정보 주체가 개인 데이터에 대한 열람, 제공 범위, 접근 승인 등을 직접 결정함으로써 개인의 정보 활용 권한을 보장하고 데이터 주권을 확립하는 패러다임 또는 정책'을 무엇이라고 하는가?

해설 마이 데이터는 개인정보 자기결정권이라는 권리를 기반으로 나의 정보를 스스로 통제할 수 있는 일련의 과정이나 패러다임, 정책을 의미합니다.
정답 마이 데이터

[문9] '데이터 과학자에게 요구되는 역량으로서 비판 능력, 호기심, 커뮤니케이션, 스토리텔링, 시각화에 해당하는 기술'을 무엇이라고 하는가?

해설 하드 스킬과 소프트 스킬 중 소프트 스킬은 눈에 보이지 않는 정성적인 측면이 강하고, 중에서 인문학에 가까운 특성을 가지고 있습니다. 반면, 하드 스킬은 수행 절차와 IT 기술이 포함된 정량적인 측면이 강하고, 엔지니어링에 가까운 특성을 가지고 있습니다.
정답 소프트 스킬

[문10] 지식의 한 종류로서 '겉으로 드러나지 않는 지식'은 무엇인가?

해설 지식은 크게 암묵지와 형식지로 나뉩니다. 암묵지는 말로 표현하기 어려운 지식으로, 말이나 글로 표현이 애매한 지식이며 반면, 형식지는 말이나 글로 표현할 수 있고 전달과 공유가 가능한 형상화된 지식입니다. 즉, 암묵지는 정리되지 않은 내면의 깊은 지식이고, 형식지는 명확하게 요약 정리된 지식이라고 할 수 있습니다.
정답 암묵지

[문11] '데이터 분석 조직의 유형으로, 분석 전담 조직이 우선순위에 따라 진행하며 일부 분석 업무가 중복되거나 이원화될 수 있는 조직 구조'는 어떤 구조인가?

해설 데이터 분석 조직은 집중 구조, 기능 구조, 분산 구조로 구분될 수 있습니다. 집중 구조는 데이터 분석을 수행하는 DSCoE(데이터 분석 전담 조직)이 있고, 마케팅 조직이나 재무 조직에서 분석하는 업무와 DSCoE 조직의 업무가 겹칠 가능성이 있습니다. 기능 구조는 데이터 분석 조직이 없고, 각 마케팅 조직과 재무 조직 등에서 알아서 분석 역할을 수행합니다. 분산 구조는 DSCoE 조직에서 업무 조직으로 데이터 분석을 수행하는 인력을 직접 배치하기 때문에 신속한 결과가 도출됩니다. 또한, 조직별로 업무의 우선순위에 따라 분석을 수행하고 그에 따른 모범사례를 공유하거나 참조할 수 있습니다.
정답 집중 구조

[문12] '합리적인 의사결정을 방해하는 요소들 중에서, 문제의 표현 방식에 따라서 동일한 사건이나 상황임에도 개인의 선택과 판단에 의해서 다르게 받아들이게 되는 현상'을 무엇이라고 하는가?

해설 프레이밍 현상은 같은 상황을 다르게 생각하는 것을 말합니다.
정답 프레이밍 현상

[문13] '사업, 부서, 혹은 개인 차원의 목표가 달성되었는지 그 실적을 추적하기 위한 측정 가능한 정량적 성과지표'는 무엇이라고 하는가?

해설 KPI에 대한 설명이며, 이는 수치로 표현 가능한 목표입니다. 다음 해당 부서는 작년대비 상반기 매출을 30% 이상 향상 등이 KPI의 예입니다.
정답 핵심 성과 지표(KPI: Key Performance Indicator)

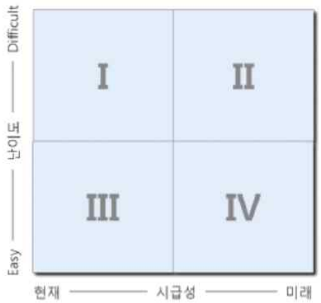
[문14] ‘데이터로부터 의미 있는 정보를 추출해 내는 학문‘을 무엇이라고 하는가?

해설 데이터 사이언스에 대한 설명이며, 사전적 정의는 ‘데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야’라고 할 수 있습니다.

정답 데이터 사이언스

1.2 데이터 분석 계획

[문15] 데이터 분석 과제 추진시 고려해야 하는 우선순위 평가 기준으로서 아래와 같은 사분면으로 시급성과 난이도를 구분할 수 있다. 난이도가 우선인 상황에서 의사결정 순서와 시급성이 우선인 상황에서 의사결정 순서를 각각 순서대로 언급하시오.



해설 데이터 분석을 수행할 때는 난이도와 지금 당장 분석해야 하는지에 대한 시급성으로 순서를 정할 수 있습니다. 난이도가 우선이라면, 당장 수행해야 할 과제를 기준으로 쉬운 것(Ⅲ)부터 어려운 것(Ⅰ)을 수행한 후 장기적으로 어려운 난이도의 과제(Ⅱ)를 나중에 수행합니다. 단, 난이도가 낮은 Ⅳ는 미래에 분석을 수행하는 시급하지 않은 과제이므로 선택에서 제외됩니다. 반면 당장 수행해야 하는 시급성이 우선이라면, 쉬운 과제를 기준으로 시급한 순(Ⅲ → Ⅳ)으로 수행한 후, 어려운 과제(Ⅱ)를 수행합니다. 단, 시급성이 높은 Ⅰ은 고난이도의 과제이기 때문에 당장 분석 수행이 어려울 가능성이 높으므로 선택하지 않습니다. 즉, 경영진의 의사결정을 통해서 난이도를 조율하고 적용 우선순위를 조정해야할 과제인 것입니다.

정답 난이도가 우선인 상황에서 의사결정 순서: Ⅲ → Ⅰ → Ⅱ
시급성이 우선인 상황에서 의사결정 순서는 Ⅲ → Ⅳ→ Ⅱ

[문16] '문제가 주어져 있어 해결 방안을 바로 탐색해야 하고 해법을 찾기 위해 각 과정이 체계적으로 단계화되어 수행하는 분석 과제 발굴 방식'은 무엇이라고 하는가?

해설 하향식 접근방법(top down approach)은 문제가 주어지고 이에 대한 해법을 찾기 위해 각 과정이 체계적으로 단계화되어 수행하는 방식입니다. 하향식 접근방법은 문제탐색 단계→문제정의 단계→해결방안 탐색단계→타당성 검토단계 등 4단계로 진행됩니다.

정답 하향식 접근 방식

[문17] '문제 정의 자체가 어려운 경우 디자인 사고 또는 비지도학습 등을 사용하여 데이터 분석을 수행하면서 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식'은 무엇이라고 하는가?

해설 어떤 분석을 수행할지에 대해서 아무것도 결정된 바가 없을 때, 분석 대상을 먼저 정하는 작업을 수행합니다. 이러한 방식에서는 디자인적 사고, 비지도학습 등 탐색적 방법을 활용한 상향식 접근 방법이 적합합니다.
정답 상향식 접근방식

[문18] 상향식 접근 방식의 발산단계와 하향식 접근 방식의 수렴단계를 반복하여 과제를 발굴하는 방법을 무엇이라고 하는가?

해설 디자인 사고란, 인간을 관찰하고 공감하여 이해한 뒤, 다양한 대안을 찾는 확산적 사고와, 주어진 상황에 최선의 방법을 찾는 수렴적 사고의 반복을 통하여 혁신적 결과를 내는 창의적 문제 해결 방법입니다.
정답 디자인 사고(Design Thinking)

[문19] 빅데이터 분석을 기획하고자 할 때, 분석하려는 대상은 정해졌으나 어떻게 분석해야 할지 모르는 경우에 사용하는 유형은?

해설 무엇을 분석할지 정해졌지만, 어떻게 분석해야 할지 모르는 경우는 솔루션을 찾아내는 방향으로 분석을 기획합니다. 또한, 분석 대상과 방법의 인지 여부에 따라 사용되는 솔루션, 최적화, 통찰, 발견도 반드시 알고 있어야 합니다.
정답 솔루션

[문20] 데이터 분석 시에 분석 과제가 정해지고 분석 역량은 확보한 상태지만, 분석기법이나 시스템을 신규 도입해야 하는 환경에서 수행하는 방법은 무엇인가?

해설 데이터를 분석할 수 있는 기술 수준이 있지만 새로운 시스템을 도입해야 하는 상황에 필요한 수행 방법은 시스템 고도화입니다.
정답 시스템 고도화

[문21] 조직의 경영목표와 전략의 효과적 지원을 위해 중장기적으로 마스터 플랜을 수립하고, IT 사업 도출과 로드맵을 수립하는 활동은 무엇인가?

해설 시스템 구축을 위한 장기적인 계획을 세우는 전반적인 활동을 ISP라고 합니다.
정답 정보전략계획 또는 ISP

[문22] 전산시스템을 필요로 하는 곳으로부터 하청을 받아 시스템의 기획 개발, 유지보수, 운영 등을 대신해주는 업종을 무엇이라고 하는가?

해설 보통 SI업체라고 하며 외주로 개발 전과정 또는 일부를 진행하는 업종입니다.
정답 SI(System Integration, 시스템 구축)

[문23] 다음은 무엇을 설명하는 것인가?
"데이터를 분석하기 위한 데이터 마이닝 방법론으로 [단계/일반화 태스크/세분화 태스크/프로세스 실행]의 4가지 구성요소와 [업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 → 전개]로 이루어진 6개의 절차로 이루어진 방법론"

해설 대표적인 데이터 분석 방법론은 CRISP-DM과 KDD, SEMMA라는 3가지가 있습니다. 이중 CRISP-DM은 4가지 구성요소와 6개의 절차로 이루어져 있습니다. KDD와 SEMMA에 대해서도 숙지하고 있어야 합니다.
정답 CRISP-DM

[문24] 다음은 무엇을 설명하는 것인가?
"Fayyad가 프로파일링 기술을 기반으로 통계적인 패턴/규칙이나 지식을 찾기 위해 체계적으로 정리한 방법으로, [데이터 세트 선택 → 데이터 전처리 → 데이터 변환 → 데이터 마이닝 → 데이터 마이닝 결과 평가] 단계로 수행하는 방법론"

해설 KDD 데이터분석 방법론에 대한 설명입니다.
정답 KDD

[문25] 다음은 무엇을 설명하는 것인가?
"SAS사에서 자사의 기술로 데이터 마이닝 기능을 구성하여 [샘플링 → 탐색 → 수정 → 모델링 → 검증]이라는 5단계로 정리한 통계 중심의 마이닝 방법론"

해설 SEMMA 데이터분석 방법론에 대한 설명입니다. SEMMA는 KDD와 동일한 5단계 절차로 구성된 통계 기반의 방법론입니다.
정답 SEMMA

[문26] 기업에서 소비자들을 자신의 고객으로 만들고, 지속적인 고객으로 유지하고자 고객 관련 정보를 분석하고 저장하는 정보시스템을 무엇이라고 하는가?

해설 CRM은 고객이 중심인 IT 시스템으로서, 고객 정보를 활용해서 마케팅 정보를 결정하는 고객 관리시스템입니다.
정답 CRM(Customer Relationship Management)

[문27] 기업이 시간과 비용을 최적화 시키기 위해 외부 공급업체와 연계하여 통합한 정보시스템을 무엇이라고 하는가?

해설 SCM(Supply Chain Management)에 대한 설명으로서 제품의 생산과 유통 과정을 하나의 통합망으로 관리하는 공급망 관리 시스템을 말합니다.
정답 SCM(Supply Chain Management)

[문28] 재무, 제조, 소매유통, 공급망, 인사 관리, 운영 전반의 비즈니스 프로세스를 자동화하고 관리하는 시스템을 무엇이라고 하는가?

해설 ERP(Enterprise Resource Planning)에 대한 설명으로서 기업에 있는 모든 인적 자원, 물적 자원을 효율적으로 관리하는 전사적 자원 관리 시스템을 의미합니다.
정답 ERP(Enterprise Resource Planning, 전사적자원 관리)

[문29] 여러 사람이 공유하여 사용할 목적으로 체계화해 통합, 관리하는 데이터의 집합 / 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합

해설 데이터베이스에 대한 설명입니다. 데이터베이스는 데이터의 집합이며 이를 통합·관리하기 위한 개념 및 도구입니다.
정답 데이터베이스

[문30] 다음은 무엇에 대한 설명인가?
“데이터베이스의 테이블이 어떻게 구성되는지, 어떤 정보를 담고 있는지에 대한 기본적인 구조를 정의하는 것”

해설 스키마에 대한 설명입니다. 정확하게는 데이터베이스 스키마(database schema)라고 하며, 데이터베이스에서 자료의 구조, 자료의 표현 방법, 자료 간의 관계를 형식 언어로 정의한 구조입니다. 데이터베이스 관리 시스템(DBMS)이 주어진 설정에 따라 데이터베이스 스키마를 생성하며, 데이터베이스 사용자가 자료를 저장, 조회, 삭제, 변경할 때 DBMS는 자신이 생성한 데이터베이스 스키마를 참조하여 명령을 수행하게 됩니다.
정답 스키마(scheme)

[문31] 다음은 무엇에 대한 설명인가?
“데이터베이스를 관리하며 응용 프로그램들이 데이터베이스를 공유하며 사용할 수 있는 환경을 제공하는 소프트웨어”

해설 데이터베이스 관리 시스템(database management system, DBMS)은 다수의 사용자들이 데이터베이스 내의 데이터를 접근할 수 있도록 해주는 소프트웨어 도구의 집합입니다.
정답 DBMS(DataBase Management System)

[문32] 사용자의 의사결정에 도움을 주기 위해 수집된 대량의 비즈니스 데이터베이스로서 주제 중심적으로 통합적이며 시간성을 가지는 비휘발성 데이터의 집합을 무엇이라고 하는가?

해설 데이터 웨어하우스에 대한 설명입니다. 데이터 웨어하우스란 사용자의 의사 결정에 도움을 주기 위하여 기간시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스입니다. 줄여서 DW로도합니다.
정답 데이터 웨어하우스(Data Warehouse, DW)

[문33] 다음은 무엇에 대한 설명인가?
“데이터 웨어하우스에서 추출한 데이터를 특정 주제영역으로 분석 후 그 결과를 조직이나 팀에서 활용하도록 제공한 데이터로서 작은 규모의 데이터 웨어하우스라고도 할 수 있다.”

해설 데이터 마트(Data Mart, DM)는 데이터 웨어하우스(Data Warehouse, DW) 환경에서 정의된 접근계층으로, 데이터 웨어하우스에서 데이터를 꺼내 사용자에게 제공하는 역할을 하는 데이터집합니다.
정답 데이터 마트(Data Mart)

[문34] 다음은 무엇에 대한 설명인가?
“전사 차원의 모든 데이터에 대해 정책 및 지침, 표준화, 운영 조직 및 책임 등의 표준화된 관리 체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축하는 것”

해설 데이터 거버넌스(data governance)는 기업에서 사용하는 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책과 프로세스를 다루며 프라이버시, 보안성, 데이터품질, 관리규정 준수를 강조하는 개념 및 프로세스입니다.
정답 데이터 거버넌스

[문35] 당사자의 동의 없는 개인정보 수집 및 활용하거나 제3자에게 제공하는 것을 금지하는 등 개인정보보호를 강화한 내용을 담아 제정한 법률

해설 개인정보보호법에 대한 설명입니다.
정답 개인정보보호법

[문36] 정보통신망의 개발과 보급 등 이용 촉진과 함께 통신망을 통해 활용되고 있는 정보보호에 관해 규정한 법률

해설 정보통신망법에 대한 설명입니다.
정답 정보통신망법

[문37] 2020년 1월에 국회에서 통과된 이른바 데이터 3법이라 불리는 3가지는 무엇인지 적으시오.

해설 데이터 3법은 개인정보보호법, 정보통신망법, 신용정보법입니다. 데이터의 이용을 활성화하기 위해서 데이터 3법 개정안이 발의되었습니다. (발의 일시: 2018.11.15)
정답 개인정보보호법, 정보통신망법, 신용정보법

1.3 데이터 수집 및 저장 계획

[문38] 다음은 무엇에 대한 설명인가?
“데이터에 개인을 식별할 수 있는 정보가 있는 경우 일부 또는 전체를 삭제하거나 일부를 대체 처리함으로써 개인을 식별할 수 없게 하는 것”

해설 개인정보를 숨기거나 변환하는 것을 '비식별화'라고 합니다. 비식별화 방법 외에, 가명처리, 총계, 데이터 범주화, 데이터 삭제, 데이터 마스킹이 있습니다. 각 특징을 이해하고 있어야 합니다
정답 데이터 비식별화

[문39] 개인을 식별할 수 있는 데이터를 다른 값으로 대체하여 식별할 수 없게 하는 비식별화 방법을 무엇이라고 하는가?

해설 개인의 식별 가능한 데이터를 다른 값으로 대체하여 비식별화하는 방법은 가명처리입니다
정답 가명처리

[문40] 사생활 침해를 방지하기 위해 데이터에 포함된 개인정보를 삭제하거나 알아볼 수 없는 형태로 변환하는 방법

해설 익명화에 대한 설명입니다.
정답 익명화

[문41] 분석 대상 데이터 집합에서 준식별자 속성이 동일한 레코드가 적어도 K개 이상 존재하도록 제한하는 개인정보 보호 기법

해설 k-익명성에 대한 설명입니다.
정답 k-익명성

[문42] K-익명성의 동질설 문제나 배경지식을 이용하는 문제를 해결하기 위하여 익명성을 향상시키는 방법

해설 L-다양성에 대한 설명입니다.
정답 L-다양성

[문43] 동질 집합에서 민감정보의 분포와 전체 데이터 집합에서의 민감정보 분포가 유사한 차이를 보이게 만드는 기법

해설 T-접근성에 대한 설명입니다.
정답 T-접근성

[문44] 데이터의 결측값을 처리하는 방법 중 이것은 보통 m번 대체를 수행하고 그에 따른 m개의 자료가 생성되면 이를 각각 분석하는 방법이다.

해설 다중 대체법에 대한 설명입니다.
정답 다중 대체법

[문45] 통계 값을 적용하여 특정 개인을 식별할 수 없도록 하는 비식별화 방법을 무엇이라고 하는가?

해설 총계처리에 대한 설명입니다.
정답 총계처리

[문46] 특정 정보를 해당 그룹의 대푯값 또는 구간값으로 변환하는 비식별화 방법을 무엇이라고 하는가?

해설 데이터 범주화에 대한 설명입니다.
정답 데이터 범주화

[문47] 데이터의 전부 또는 일부분을 대체값(공백, 노이즈 등)으로 변환하는 비식별화 방법 / 개인의 사생활 침해를 방지하고 통계 응답자의 비밀사항은 보호하면서 통계자료의 유용성을 최대한 확보할 수 있는 데이터변환 방법은?

해설 데이터 마스킹에 대한 설명입니다.
정답 데이터 마스킹

[문48] 다음은 무엇에 대한 설명인가?
“고정된 구조로 정해진 필드에 저장된 데이터. 엑셀 스프레드시트, RDBMS(관계형 데이터베이스), CSV 파일 형태가 대표적이다”

해설 고정된 구조(fixed structure)로 된 데이터를 정형 데이터라고 합니다.
정답 정형 데이터

[문49] 다음은 무엇에 대한 설명인가?
“고정된 필드에 저장되어 있지는 않지만, 데이터와 메타데이터, 스키마 등을 포함하는 데이터(XML, HTML, JSON 등). 규칙을 가지고 있어 필요 시 정형 데이터로 변형 가능한 데이터”

해설 XML, HTML, JSON 등의 형태로 데이터와 메타데이터 등이 같이 포함된 데이터는 반정형 데이터입니다.
정답 반정형 데이터

[문50] 정해진 구조가 없고 고정된 필드에 저장되어 있지 않은 데이터는 무엇이라고 하는가?

해설 문자, 이미지, 영상 등 아직 정형화되지 않은 구조의 데이터를 비정형 데이터라고 합니다.
정답 비정형 데이터

[문51] 다음은 무엇에 대한 설명인가?
“데이터에 관한 구조화된 데이터로서 어떤 목적을 가지고 만들어진 데이터를 말한다. 데이터 그 자체가 아니라, 자료의 속성, 구조 등을 설명하는 데이터이다”

해설 메타데이터에 대한 설명입니다.
정답 메타데이터

[문52] 절대적인 영점이 존재하고 순서와 의미가 포함된 데이터 측정 척도는 어떤 척도인가?

해설 데이터의 척도는 크게 명목척도, 서열척도, 등간척도, 비율척도로 구분됩니다. 이 중 비율척도는 가장 많은 양의 정보를 얻을 수 있으며, 절대영점, 즉 0이 ‘정말로 없다’는 의미를 가진 수치형 자료가 될 수 있습니다. 그 외에 명목척도, 서열척도, 등간척도의 정의와 특징도 알아두어야 합니다.
정답 비율척도

[문53] 인터넷상에 제공되는 다양한 웹사이트로부터 소셜 네트워크 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠를 수집하는 기술을 무엇이라고 하는가?

해설 인터넷에 있는 웹사이트에서 정보들을 모두 긁어오는 기술로, 텍스트 마이닝을 수행하기 전의 학습데이터로 사용하기 위한 기반 자료로 활용할 수도 있습니다.
정답 크롤링

[문54] 데이터에서 잡음을 제거하기 위해 추세에서 벗어나는 값들을 변환하고, 거칠게 분포된 데이터를 부드럽게 만들기 위한 데이터 변환 기술은 무엇이라고 하는가?

해설 평활화는 기존의 극단적인 데이터 분포를 부드럽게 바꿔주는 기술입니다. 다음 그림과 같이 극단적인 값들 간의 차이를 보다 작게 차이가 나도록 바꿔주는 것입니다. 명암 값의 범위가 0~10이라고 가정하면, 기존 명암 값은 0~10까지 극단적인 데이터 분포를 보이고 있으나 이를 평활화한 오른쪽은 2~8까지의 유사한 명암 값으로 부드럽게 나타남을 알 수 있습니다.
정답 평활화

[문55] 다음은 무엇에 대한 설명인가?
“이것은 비즈니스 측면에서 일반적으로 ‘공동 활용의 목적으로 구축된 유무형의 구조물’을 의미한다. 수집된 데이터를 가공, 처리, 저장해두고 이 데이터에 접근할 수 있도록 API를 공개한다. 그러면 다양한 서드파티 사업자들이 비즈니스에 필요한 정보를 추출해 활용하게 되고 빅데이터는 그 자체로 이 역할을 수행하게 된다.”

해설 플랫폼은 컴퓨터의 아키텍처, 운영 체제, 프로그램 언어, 그리고 관련 런타임 라이브러리 또는 GUI를 포함하며, 소프트웨어가 구동 가능한 하드웨어 아키텍처나 소프트웨어 프레임워크의 종류입니다
정답 플랫폼

[문56] 다음은 무엇에 대한 설명인가?
“대용량의 정형 및 비정형 데이터를 저장하고 손쉽게 접근할 수 있게 하는 대규모 저장소”

해설 데이터 호수(Data Lake, 데이터 레이크)는 데이터가 분석이나 마이닝에 필요할 때까지 모든 유형의 데이터를 보관할 수 있는 대규모 데이터 저장소
정답 데이터 호수(Data Lake)를 의미합니다.

[문57] 대용량 파일을 저장하고 처리하기 위해서 네임 노드(master)와 데이터 노드(slave)로 구성된 파일 시스템을 무엇이라고 하는가?

해설 데이터 노드를 지배하는 네임 노드는 클라이언트의 요청을 받습니다. 받은 요청은 데이터 노드로 전달하며, 데이터 노드는 네임 노드로부터 받은 명령을 처리하는 역할을 합니다. 또한, 네임 노드는 동일한 데이터 값을 3번 중복 저장함으로써, 발생할 수 있는 장애 및 데이터 손실을 막을 수도 있습니다.
정답 하둡 분산 파일 시스템 또는 HDFS

[문58] 솔루션 제조사 또는 제삼자의 소프트웨어로 제공되는 도구로서, 시스템 간 연동을 통해서 실시간으로 데이터를 수신할 수 있는 기능을 제공하는 인터페이스 기술

해설 API는 상호 간의 정보를 적극적으로 소통하기 위한 것으로, API 호출을 통해서 요청하고 응답을 받게 되는 데이터 교환 기술입니다. 네이버의 지도 데이터도 외부에서 API를 활용하여 사용할 수 있습니다. 예를 들어 지도 정보가 필요한 사용자는 지도 기관의API를 호출하면, 필요한 지도 이미지와 관련 정보를 응답(리턴)하게 됩니다. 즉, 각 기관에서 만든 API들로 원하는 정보를 입맛에 맞게 활용할 수 있는 것입니다.
정답 API

[문59] 다양한 원천으로부터 데이터를 수집하고, 공통 형식으로 변환한 후, 데이터 저장소 또는 데이터 웨어하우스에 적재하는 기술을 무엇이라고 하는가?

해설 ETL이란 여러 곳에 분산된 데이터들 중에서 필요한 데이터를 가져오고(추출), 이후 필요한 형식과 값들로 변환해서, 최종 저장소에 저장(적재)하는 일련의 기술입니다. 보통 OLTP환경이 아닌 OLAP 환경에서 ETL 기술을 적극적으로 사용하고 있습니다.
정답 ETL

[문60] TCP/IP 통신을 수행하고, 서버와 클라이언트 간의 파일을 전송하기 위한 프로토콜은 무엇이라고 하는가?

해설 서로 떨어진 위치에서 파일을 전달하거나 수신하기 위한 파일 전송 규약입니다.
정답 FTP

[문61] 관계형 데이터베이스를 SQL을 사용해 CRUD(Create, Read, Update, Delete)를 수행하고 관리할 수 있는 소프트웨어를 무엇이라고 하는가?

해설 RDBMS에 대한 설명입니다.
정답 RDBMS

[문62] 다음은 무엇에 대한 설명인가?
“Not Only SQL의 약자로 SQL을 사용하는 전통적인 관계형 데이터베이스 시스템보다 상대적으로 제한이 덜한 데이터 모델을 기반에 둔 분산 데이터베이스 기술 / 데이터 저장을 위한 스키마가 필요 없으며 조인 연산을 지원하지 않는다.”

해설 NoSQL에 대한 설명입니다.
정답 NoSQL

[문63] 데이터 원천으로부터 데이터를 추출 및 변환하여 데이터 웨어하우스 등에 데이터를 적재하는 작업은 무엇이라고 하는가?

해설 ETL에 대한 설명입니다.
정답 ETL(Extraction, Transformation and Load)

[문64] 다음은 무엇에 대한 설명인가?
“웹을 운영하는 주체가 누구나 사용할 수 있게 공개한 데이터를 개발자나 사용자가 수집해 사용하는 기술을 의미”

해설 Open API에 대한 설명입니다.
정답 Open API(Application Programming Interface)

[문65] 다음은 무엇에 대한 설명인가?
“대규모 분산 시스템 모니터링을 위해 에이전트와 컬렉터 구성을 통해 데이터를 수집하고 수집된 데이터를 하둡 파일 시스템(HDFS)에 저장하는 기능을 제공하는 데이터 수집 기술”

해설 Chukwa(척와)에 대한 설명입니다.
정답 Chukwa(척와)

[문66] 다음은 무엇에 대한 설명인가?
“RDBMS와 하둡 사이의 데이터를 이동시켜주는 애플리케이션”

해설 Apache Sqoop(스쿱)에 대한 설명입니다.
정답 Apache Sqoop(스쿱)

[문67] 다음은 무엇에 대한 설명인가?
“분산 환경에서 대량의 로그 데이터를 효과적으로 수집하여 합친 후 다른 곳으로 전송할 수 있는 신뢰할 수 lT는 서비스”

해설 Apache Flume(플럼)에 대한 설명입니다.
정답 Apache Flume(플럼)

[문68] 빅데이터 저장 기술로 컴퓨터 네트워크를 통해 공유하는 여러 호스트 컴퓨터의 파일에 접근할 수 있게 하는 파일 시스템은?

해설 분산 파일 시스템에 대한 설명입니다.
정답 분산 파일 시스템

[문69] 다음은 무엇에 대한 설명인가?
“실시간으로 기록 스트림을 게시, 구독, 저장 및 처리할 수 있는 분산 데이터 스트리밍 플랫폼”

해설 Apache Kafka(카프카)에 대한 설명입니다.
정답 Apache Kafka(카프카)

[문70] 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크. 간단하게 설명하자면, 한명이 4주 작업할 일을 4명이 나누어 1주일에 끝내는 것

해설 MapReduce(맵리듀스)에 대한 설명입니다.
정답 MapReduce(맵리듀스)

[문71] 다음은 무엇에 대한 설명인가?
“각종 사물에 센서와 통신 기능을 내장하여 인터넷에 연결하는 기술. 즉, 무선 통신을 통해 각종 사물을 연결하는 기술을 의미한다”

해설 사물인터넷(Internet of Things, IoT)에 대한 설명입니다.
정답 사물인터넷(Internet of Things, IoT)

[문72] 빅데이터 분석에 경제성을 제공해준 기술은? / 인터넷상의 서버에서 데이터 저장, 처리, 네트워크, 콘텐츠 사용 등 서로 다른 물리적인 위치에 존재하는 컴퓨팅 자원을 가상화 기술을 통해 IT 관련 서비스를 한번에 제공하는 혁신적인 컴퓨팅 기술은 무엇이라고 하는가?

해설 클라우드 컴퓨팅에 대한 설명입니다.
정답 클라우드 컴퓨팅

[문73] 빅데이터 저장 기술로 관계형 데이터베이스 관리 시스템으로 하나의 데이터베이스를 여러 개의 서버상에 구축하는 시스템은 무엇이라고 하는가?

해설 데이터베이스 클러스터에 대한 설명입니다.
정답 데이터베이스 클러스터

2. 빅데이터 탐색

2.1 데이터 전처리

[문74] 데이터의 결측값(치)을 채우거나 이상값을 제거하여 데이터 품질을 높이는 과정을 무엇이라고 하는가?

해설 결측값, 이상치 등을 처리하는 과정을 데이터 정제과정이라고 합니다.
정답 데이터 정제

[문75] 대부분의 데이터가 주로 분포된 범위에서 많이 벗어난 값은 무엇이라고 하는가?

해설 이상값은 말 그대로 정상이 아닌 이상이 있는 값입니다. 정상적인 위치에 분포하지 않고 범위를 넘어서는 값이 이상값입니다. 다음 그림과 값이 나이가 -1, 199인 경우는 상식 수준에서도 이상값임을 알 수 있습니다.
정답 이상값 또는 이상치

[문76] 어떤 값도 대상 변수에 입력되지 않아, 비어 있는 값을 무엇이라고 하는가?

해설 결측값은 공백은 아니며, 입력이 누락되어 비어 있는 값입니다. 파이썬에서는 NaN으로 표시됩니다.
정답 결측값 또는 결측치

[문77] 평균으로부터 표준편차의 K배만큼 떨어져 있는 값들을 이상값으로 판단하는 방법은 무엇이라고 하는가?

해설 ESD는 이상값을 찾는 방법입니다. 분석 데이터에서 평균과 표준편차를 계산한 후, 평균으로부터 3×표준편차 이상이거나 -3×표준편차 이하인 경우를 이상값으로 처리합니다. 이상값을 인식하는 방법에는 ESD, 기하평균을 활용한 방법, 사분위수를 활용한 방법이 있습니다. 기하평균을 활용한 방법은 기하평균에서 2.5×표준편차 이상 떨어진 값을 이상치로 판단합니다. 사분위수를 활용한 방법은 1사분위수로부터 -1.5×OR 이하이거나, 3사분위수로부터 1.5×IQR 이상인 값을 이상치로 판단합니다.
정답 ESD(Extreme Studentized Deviation)

[문78] 데이터 표본을 4개의 동일한 부분으로 나눈 사분위수에서 3사분위수에서 1사분위수를 뺀 것은?

해설 먼저 데이터를 동일한 4개 영역으로 구분합니다. 다음 그림과 같이 1~99까지 데이터가 있다고 가정하면, 1~24, 26~49, 51~74, 76~99까지 4개 영역으로 나눕니다. 전체 중에 25%에 해당하는 1사분위수는 25, 50%에 해당하는 중위값은 50, 75%에 해당하는 3사분위수는 75입니다. 여기서 IQR은 3사분위수 1사분위수를 계산한 것으로 50입니다.
정답 IQR 또는 사분위수 범위

[문79] 다수의 클래스 데이터를 일부만 선택해 데이터 비율을 맞추는 불균형 데이터 처리 기법은 무엇이라고 하는가?

해설 데이터의 비율에 차이가 나면, 데이터 분석 결과가 데이터양이 많은 쪽으로 왜곡될 가능성이 커집니다. 따라서 다음 그림과 같이 남성 데이터양이 여성 데이터양보다 월등히 많으면, 데이터양이 많은 쪽을 적은 쪽으로 맞춰서 데이터를 분석하는 기법입니다.
정답 언더 샘플링

[문80] 소수의 클래스 데이터를 늘려서 데이터 비율을 맞추는 불균형 데이터 처리 기법은 무엇이라고 하는가?

해설 다음 그림과 같이 여성 데이터양이 남성 데이터양보다 훨씬 적은 경우, 데이터양이 적은 쪽을 많은 쪽에 맞춰서 데이터를 분석하는 기법입니다.
정답 오버 샘플링

[문81] 데이터를 전처리하는 과정에서 단위를 변환하거나 변수를 결합하거나 표현형식을 변환해서 기준 변수를 새롭게 정의하는 변수는 무엇이라고 하는가?

해설 파생변수는 우리가 직접 특정 함수를 통해서 새로운 값을 만들거나, 특정 조건을 만족하는값을 변형하는 방식으로 변수를 생성하는 것입니다. 다음 그림은 나이 변수를 기준으로 20세 이상인 경우는 성인이라고 간주하고 Y로 설정, 그렇지 않으면 N으로 설정하는 '성인 여부'라는 파생변수를 만드는 예시입니다. 반면, 요약변수는 수집 데이터를 분석해서 종합하거나 통합한 변수를 의미합니다. 총 구매 금액, 구매 횟수 같은 변수 등이 합계나 카운트를 수행한 요약변수의 사례입니다.
정답 파생변수

[문82] 관측 또는 실험으로 얻은 자료의 평균값으로 결측값을 대치해서, 불완전한 자료를 완전한 자료로 만드는 방법은 무엇이라고 하는가?

해설 단순 대치법에는 완전 대치법, 평균대치법, 단순확률 대치법이 있습니다. 완전 대치법은 결측치가 존재하는 데이터는 삭제하는 것이며, 단순확률 대치법은 추정된 통계량에 확률값을 부여하는 대치 방법입니다.
정답 평균 대치법

[문83] 표본조사에서 흔히 사용하는 방법으로, 무응답을 현재 진행 중인 연구에서 비슷한 성향을 가진 응답자의 자료로 대치하는 방법은 무엇이라고 하는가?

해설 핫덱은 데이터의 누락된 값(결측치)을 처리하는 방법입니다. 결측치가 발생한 데이터와 유사한 성향을 가진 응답자의 값으로 결측치를 보정하는 전처리 방법입니다.
정답 핫덱 대체법

[문84] 대체할 자료를 현재 진행 중인 연구에서 얻는 것이 아니라, 외부 출처 또는 이전의 비슷한 연구에서 가져오는 방법은 무엇이라고 하는가?

해설 콜드덱은 데이터에 누락된 값(결측치)을 처리하는 방법입니다. 핫덱과 비슷하지만 결측치를 대체할 값을 과거의 연구 자료에서 가져오는 등 외부에서 가져오는 방식을 말합니다.
정답 콜드덱 대체법

[문85] 정규화 방법 중 원 데이터의 분포를 유지하면서 정규화하는 방법으로서 모든 데이터를 0과 1 사이의 값으로 변환하는 기법은?

해설 min-max 정규화 방법에 대한 설명입니다.
정답 최소-최대(Min-Max) 정규화

[문86] 비지도학습 과정 중에 변수들의 정보를 유지하면서 변수의 개수를 줄이는 방법은 무엇이라고 하는가?

해설 차원 축소는 데이터가 가지고 있는 정보를 유지하면서 데이터를 줄이는 데이터 전처리 과정의 방법으로, PCA(주성분 분석), SVD(특이값 분해) 등의 알고리즘이 있습니다.
정답 차원축소

[문87] 데이터의 분포를 잘 설명하는 변수 선택 방법 중, 비어 있는 상태에서 시작하여 점진적으로 하나씩 추가하는 방법은 무엇이라고 하는가?

해설 전진 선택법은 데이터의 특징을 가장 잘 나타낼 수 있는 변수부터 차근차근 선택하면서 분석하는 기법입니다. 다음 그림과 같이 성적과 관련된 다섯 가지 데이터 변수(키, 통학 거리, 혈액형, 휴대폰 사용시간, 부모 수입)가 있다고 가정해봅니다. 이때 성적과 가장 밀접한 '휴대폰 사용시간' 변수를 선택하고, 다음으로 '부모 수입' 변수를 선택하고, 마지막으로 '통학 거리' 변수를 선택하면서 데이터 분석을 수행하는 변수 선택 기법입니다.
정답 전진 선택법

[문88] 전체 모형에서 가장 적은 영향을 주는 변수부터 하나씩 제거하는 방법 / 최적방정식을 선택하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수를 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하는 방법은 무엇인가?

해설 후진 제거법은 데이터의 특징을 가장 잘 나타내지 못하는 변수부터 하나씩 제거하면서 분석하는 기법입니다. 다음 그림과 같이 성적과 관련된 다섯 가지 데이터 변수(키, 통학 거리, 혈액형, 휴대폰 사용시간, 부모 수입)가 있다고 가정해봅시다. 처음에는 5개 변수부터 시작하지만, 이후에는 성적과 가장 거리가 먼 '혈액형' 변수를 제거하고, 다음으로 '키' 변수를 제거하는 변수 제거 기법입니다.
정답 후진 제거법

[문89] 변수를 연속적으로 추가 혹은 제거하면서 AIC가 낮아지는 모델을 찾는 방법은?

해설 전진 선택법과 후진 제거법을 연속적으로 반복하면서 AIC 기준으로 모델을 최종 모델을 선택하는 방법을 단계적 방법(stepwise method)이라고 합니다.
정답 단계적 방법

2.2 데이터 탐색

[문90] 데이터를 이해하고 의미 있는 관계를 찾아내기 위해 데이터의 통곗값과 분포 등을 시각화하고 분석하는 등 본격적으로 데이터 분석을 시작하기 전에 자료를 직관적으로 통찰하는 활동을 무엇이라고 하는가?

해설 EDA, 탐색적 데이터 분석에 대한 설명입니다.
정답 EDA(Exploratory Data Analysis, 탐색적 데이터 분석)

[문91] 데이터의 분포를 나타내는 통계량으로 치우진 정도를 의미하는 지표를 무엇이라고 하는가?

해설 왜도는 왼쪽이나 오른쪽으로 쏠렸는지는 의미하는 지표입니다. 왼쪽으로 쏠리면 왜도는 양의 값을 가지고, 오른쪽으로 쏠리면 음의 값을 가집니다. 왜도와 첨도는 분석하는 데이터의 분포가 정규분포에서 어느 정도 벗어나 있는지를 확인하는 지표입니다.
정답 왜도

[문92] 데이터가 분포의 중심에 어느 정도 몰려 있는가를 측정할 때 사용하는 척도를 무엇이라고 하는가?

해설 첨도는 평균을 중심으로 얼마만큼 가까이 몰려 있는지에 대한 뾰족한 정도를 나타내는 지표입니다. 평균값에 많이 몰릴수록 양의 첨도를 가지고 넓게 퍼져 있을수록 음의 첨도를 가집니다.
정답 첨도

[문93] 데이터의 최댓값에서 최솟값을 뺀 것으로 순서 통계량의 산포를 의미

해설 최대값-최소값으로 계산한 통계량은 범위(range)입니다
정답 범위(range)

[문94] 2개의 변수 간 상관 정도를 나타내는 지표로 상관관계의 경향성은 알 수 있으나, 선형성의 강도는 확인하지 못하는 해당 지표를 무엇이라고 하는가?

해설 공분산에 대한 설명입니다. 두 변수 간의 상관정도를 파악하는 대표적인 통계량은 상관계수(correlation coefficient)로서 -1 ~ +1 사이로 표현되어 상관의 강도를 알 수 있습니다. 공분산은 상관계수로 변환되기 전에 원데이터의 단위대로 상관정도를 파악하기 때문에 단위가 표준화되지 않아 강도는 알 수 없습니다. 정리하면 두 변수의 관계를 파악한 값은 공분산(covariance)이지만 값이 표준화되지 않은 상관정도이며, 상관계수는 -1 ~ +1 사이로 표준화한 상관정도입니다.

정답 공분산(covariance)

2.3 통계기법 이해

[문95] 수집한 데이터를 요약, 묘사, 설명하는 통계 기법을 무엇이라 하는가?

해설 기술통계(descriptive statistics)에 대한 설명입니다. 통계적 방법은 크게 기술통계와 추론통계로 구분됩니다. 기술통계는 데이터의 경향을 전반적으로 요약하는 데에 목적이 있습니다.
정답 기술통계

[문96] 수집한 데이터를 바탕으로 모수에 대하여 추론 또는 예측하는 통계 기법을 무엇이라 하는가?

해설 추론통계에 대한 설명입니다. 모집단으로부터 얻은 통계량으로 모집단의 특징, 즉 모수를 가늠하는 방법입니다.
정답 추론(추측) 통계

[문97] 우리는 모집단을 조사하기 위해 추출한 모집단의 일부 원소를 이용한다. 통계자료의 획득 방법 중 모집단을 조사하기 위해 추출한 집단을 무엇이라 하는가?

해설 모집단에서 일부를 뽑아 조사 또는 분석을 하는데, 이때 뽑혀진 집단을 표본(sample)이라고 합니다.
정답 표본 집단, 샘플

[문98] 전체 데이터 중 분석에 필요한 데이터만 선택적으로 이용하는 것을 무엇이라 하는가?

해설 모집단에서 일부를 뽑는 과정, 또는 추출 방법을 표본추출(sampling)이라고 합니다.
정답 표본추출, 샘플링(sampling)

[문99] 관심을 갖고 있는 모집단의 특성을 나타내는 대푯값을 무엇이라 하는가?

해설 모집단의 특성을 모수(parameter)라고 하며, 전체를 조사 또는 분석하지 않는 한 알 수가 없는 값입니다.
정답 모수

[문100] 다음은 무엇에 대한 설명인가?
“표본을 조사하여 얻은 데이터로 표본의 특징을 수치화한 값으로서 모수를 추정하기 위해 구하는 표본 값들에 대한 용어”

해설 표본으로부터 얻어진 여러 가지 값들을 총칭하여 통계량(statistic)이라고 합니다.
정답 통계량

[문101] 다음은 무엇에 대한 설명인가?
“현재까지 주장되어 온 가설로서, 기존과 비교하여 변화나 차이가 없음을 나타내는 가설, 또는 실험, 연구를 통해 기각하고자 하는 어떤 가설”

해설 귀무가설은 대부분의 사람들이 믿는 것들, 상식적으로 주장되는 것들이라고 이해하면 쉽습니다. 예를 들어 '매년 장마는 6월 말~7월 초에 시작된다.'가 이에 해당합니다.
정답 귀무가설

[문102] 다음은 무엇에 대한 설명인가?
“표본을 통해 확실한 근거를 가지고 입증하고자 하는 가설 또는 실험, 연구를 통해 증명하고자 하는 새로운 아이디어 혹은 가설”

해설 대립가설은 귀무가설이 틀렸다고 판단할 때, 대안으로 선택하는 가설입니다. 즉, 우리가 데이터로 증명하고 싶은 새로운 가설을 말합니다. 귀무가설이 매년 장마는 6월 말~7월 초에 시작된다.'라면, 대립가설은 매년 장마는 6월 말~7월 초에 시작되지 않는다.'입니다.
정답 대립가설

[문103] 다음은 무엇에 대한 설명인가?
“가설 검정에서 사용된 샘플 데이터로부터 계산된 표본 통계량으로서 이것으로 p-value를 계산하며 귀무가설을 기각할 것인지 판별한다.”

해설 귀무가설과 통계량이 얼마나 다른가를 계산한 값들을 검정통계량이라고 합니다. 분석마다 계산방법이 다르지만 의미는 동일합니다.
정답 검정통계량

[문104] 통계적인 가설검정에서 사용되는 기준값으로 α로 표시하는 것을 무엇이라고 하는가?

해설 가설을 기각 혹은 채택하겠다는 연구자의 기준을 유의수준(significance level)이라고 하며 α로 표시합니다.
정답 유의수준

[문105] 다음은 무엇에 대한 설명인가?
"귀무가설 분포에서 검정통계량보다 극단적인 값이 관측될 확률값으로서 이 값이 작을수록 검정통계량이 귀무가설의 내용에 적합하지 않음을 나타낸다."

해설 p-value(유의확률)은 귀무가설이 맞는다는 전제하에 표본에서 귀무가설이 맞는다고 증명해줄 값들이 나올 확률을 의미합니다. 따라서 p-value 값이 작을수록 귀무가설이 맞는다는 의심이 높아지게 됩니다. 일반적으로 p-value는 0.05. 즉, 5% 이하인 경우에 이론적으로 귀무가설을 기각하고 대립가설을 채택하게 됩니다. p-value는 확률값이므로 0~1 사이의 범위를 가집니다.
정답 유의 확률(P-value)

[문106] 귀무가설이 참일 때 귀무가설을 기각하는 오류를 무엇이라고 하는가?

해설 1종 오류에 대한 설명입니다.
정답 1종 오류(알파 오류)

[문107] 귀무가설이 거짓일 때 귀무가설을 채택하는 오류는 무엇이라고 하는가?

해설 2종 오류에 대한 설명입니다.
정답 2종 오류(베타 오류)

[문108] 대립가설이 참일 때 귀무가설을 기각하고 대립가설을 채택할 확률은 무엇이라고 하는가?

해설 검정력(檢定力, statistical power)에 대한 설명이며, 이는 대립가설이 사실일 때, 이를 사실로서 결정할 확률입니다.
정답 검정력

[문109] 괄호 안에 들어갈 말을 순서대로 작성하시오.
"가설검정은 귀무가설과 대립가설 중에서 하나의 가설을 양자택일한다. 그래서 1-α는 귀무가설을 채택시키므로, 1-α의 영역을 "((1))"이라고 부르고, 반대로 α는 귀무가설을 기각(탈락)시키므로, α의 영역을 "((2))"이라고 부른다."

해설 귀무가설을 채택하기 위한 신뢰구간은 1-α이며, 반대로 α(유의수준)보다 작으면 기각하게 됩니다. 따라서 1-α를 채택역, α를 기각역이라고 각각 부릅니다.
정답 (1)채택역, (2)기각역

[문110] 변숫값들의 합을 변수의 총 개수로 나눈 값은 무엇이라고 하는가?

해설 평균에 대한 설명입니다.
정답 평균

[문111] 분산을 제곱근한 값은 무엇인가?

해설 $\sqrt{분산}$ 으로 계산하며, 평균으로부터 얼마만큼 떨어져 있는지를 나타내는 지표입니다.
정답 표준편차

[문112] 표본평균들의 표준편차를 무엇이라고 하는가?

해설 여러 번 반복해서 조사 혹은 분석을 할 때마다의 평균들이 이루는 분포에서, 표준편차는 표준오차를 의미합니다.
정답 표준오차

[문113] 통계적 추정을 할 때 표본자료 중 모집단에 대한 정보를 주는 독립적인 자료의 수는 무엇인가?

해설 자유도(degree of freedom)에 대한 설명입니다.
정답 자유도

[문114] 평균으로부터 흩어진 편차의 제곱합을 무엇이라고 하는가?

해설 분산은 평균에서 얼마나 떨어져 있는지를 수치화하는 지표입니다. 사람들이 모여 있으면 분산되지 않았다.'라고 하고, 사람들이 널리 퍼져 있으면 많이 분산되었다.' 라고 말하듯이 평균에 몰려 있으면 분산값이 작은 것이고 퍼져 있으면 분산값이 큰 것입니다.
정답 분산

[문115] 모든 데이터 값을 크기 순서로 오름차순 정렬할 때, 중앙에 위치한 데이터 값을 무엇이라고 하는가?

해설 중위값은 1등부터 99등까지 성적순으로 세웠다고 가정하면, 성적이 50등인 경우가 성적의 중위값입니다.
정답 중위값 또는 중앙값

[문116] 다음 데이터들의 중앙값(중위값)운 무엇인가?
'1, 10, 90, 200'

해설 전체 데이터가 홀수인 경우는 $\frac{n+1}{2}$번째가 중앙값이고, 짝수인 경우는 $\frac{n}{2}$, $\frac{n}{2}+1$번째의 평균이 중앙값이 됩니다. 따라서 이 문제는 다음과 같이 2번째와 3번째의 평균을 구하여 중앙값을 계산합니다. 따라서 (10+90)/2 = 45가 됩니다.
정답 45

[문117] 표본의 수가 무한히 커지면 표본의 분포와 관련 없이 표본 평균은 정규분포를 따른다는 원칙은 무엇인가?

해설 중심극한정리는 데이터가 많으면 많을수록 결국은 정규분포 형태와 가까워진다는 이론입니다. 이는 전체 데이터가 아닌 일부만을 추출한 표본 데이터로 전체 데이터를 추정할 수 있다는 수학적인 근거입니다.
정답 중심극한정리

[문118] 표본을 추출하는 방법 중에서 각 계층을 고루 대표할 수 있도록 표본을 임의로 추출하는 방법은 어떤 추출법인가?

해설 층화 추출법은 모집단을 비슷한 성질을 갖는 2개 이상의 층으로 구분한 후, 각 층에서 무작위로 데이터를 추출하여 표본을 만드는 방법입니다. 반면 집락 추출법은 모집단을 군집으로 나눈 후 에 각 군집에서 표본을 추출하는 방법입니다.
정답 층화 추출법

[문119] 다음은 표본추출법 중 무엇에 대한 설명인가?
“지역별 매출액, 영업이익률, 판매량과 같이 수치로 명확하게 표현되는 데이터로, 그 양이 크게 증가하더라도 이를 DBMS에 저장, 검색, 분석하여 활용하기가 용이하다. 정량적 데이터 번호를 부여한 샘플을 나열하여 k개씩 n개의 구간을 나누고 첫 구간에서 하나를 임의로 선택한 후에 k개씩 띄어서 표본을 선택하고 매번 k번째 항목을 추출하는 표본 추출 방법”

해설 계통추출방법에 대한 설명입니다.
정답 계통추출방법

[문120] 두 개 이상의 집단 간 비교를 수행하고자 할 때, 집단 내의 분산, 총 평균과 각 집단의 평균 차이에 의해 생긴 집단 간의 분산 비교로 얻은 분포를 이용하여 가설 검증하는 통계기법은 무엇이라고 하는가?

해설 다수의 집단 데이터가 있는 경우에 집단 간의 퍼져 있는 정도와 집단 안에서 데이터가 퍼져 있는 정도를 F 분포로 계산하여 가설 검증하는 것이 분산분석입니다. 분산분석은 일원 분산분석, 이원 분산분석, 다변량 분산분석으로 나뉩니다.
정답 분산 분석

[문121] 정규 분포의 평균을 측정할 때 주로 사용되는 분포로 모집단의 분산(혹은 표준편차)이 알려져 있지 않은 경우에 정규분포 대신 이용하는 확률분포를 무엇이라고 하는가?

해설 t-분포에 대한 설명입니다.
정답 t-분포

[문122] 총 평균과 각 집단의 평균 차이에 의해 생긴 집단 간 분산 비율을 나타내는 분포를 무엇이라고 하는가?

해설 F-분포에 대한 설명입니다.
정답 F-분포

[문123] 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률 분포는 무엇이라고 하는가?

해설 포아송분포에 대한 설명입니다.
정답 포아송분포

3. 빅데이터 모델링

3.1 분석 모형 설계

[문124] 분석용 데이터를 이용한 가설 설정을 통해 통계모델을 만들거나 기계학습을 이용한 모델을 만드는 과정을 무엇이라고 하는가?

해설 모델링(modeling)에 대한 설명입니다. 본격적인 분석에 앞서 데이터→분석알고리즘→결과활용 등의 프로세스를 구체적으로 정립하는 과정입니다.
정답 모델링

[문125] 다음은 무엇에 대한 설명인가?
“이것은 인공지능의 한 분야로 간주된다. 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이다. 학습 데이터로 학습한 알려진 특성을 활용해 ‘예측’하는 방법이다.”

해설 인공지능을 구현하기 위한 방법론으로 머신러닝에 대한 설명입니다.
정답 기계 학습 =머신러닝

[문126] 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계 학습의 한 분야는 무엇인가?

해설 딥러닝에 대한 설명 중 하나입니다. 높은 수준의 추상화는 고차원으로서 깊고(deep) 넓은(wide) 신경망을 구현한다는 의미입니다.
정답 딥러닝

[문127] 대규모로 저장된 데이터 안에서 통계적 규칙이나 패턴을 분석하여 가치 있는 정보를 추출하는 과정을 무엇이라고 하는가?

정답 데이터 마이닝
해설 데이터에서 숨겨진 의미들을 찾아나가는 과정을 데이터 마이닝이라고 합니다. 즉, 데이터를 캐다라는 영문 의미와도 직결됨을 알 수 있습니다.

[문128] 입력 데이터에 대한 정답인 레이블이 없는 상태에서 데이터가 어떻게 구성되었는지를 알아내는 기계학습 방법을 무엇이라고 하는가?

정답 비지도학습
해설 비지도학습은 정답이 없는 데이터에서 유의미한 패턴을 추출하는 학습방법입니다. 왼쪽의 그림처럼 유사한 특징을 지닌 데이터가 분포된 상태에서 오른쪽의 그림으로 데이터를 세 개의 유형으로 군집화한 것이 비지도학습의 예입니다.

[문129] 입력 데이터에 대한 정답, 즉 레이블이 포함되어 있는 상태에서 데이터가 어떻게 구성되었는지 학습하는 기계학습 방법을 무엇이라고 하는가?

정답 지도학습
해설 지도학습은 정답이 있는 데이터로 학습을 시키는 방법입니다. 쉽게 말하면, 학교에서 선생님이 문제를 출제하고 나서 문제에 대한 정답을 알려주는 방식이라고 이해할 수 있습니다. 이를 통해 새로 주어진 데이터가 있으면 정해진 레이블(정답)에 따라 분류할 수 있거나, 연속된 그래프의 값을 예측(회귀)할 수 있습니다.

[문130] 다음은 무엇에 대한 설명인가?
“대규모로 저장된 데이터 속에서 분석을 통해 유의미한 패턴과 규칙을 찾아내는 과정 데이터 마이닝 모델이 학습 데이터를 과하게 학습하여 훈련용 데이터에 대한 성능은 높게 나오지만, 테스트 데이터에 대한 성능은 낮게 나오는 것”

정답 과(대)적합
해설 과적합(overfitting)에 대한 설명입니다. 학습데이터에는 좋지만 테스트나 다른 데이터에 잘 맞지 않는 문제를 과적합되었다고 합니다.

[문131] 모델이 너무 단순해서 학습 데이터조차 제대로 예측하지 못하는 경우를 무엇이라고 하는가?

정답 과소적합
해설 과소적합(underfitting)에 대한 설명으로 학습데이터에 조차 잘 맞지 않는 모델을 말합니다.

[문132] 모델에서 외적인 요소로서, 데이터 분석을 통해서 얻어지는 값이 아닌 사용자가 직접 설정해주는 값이며, 경험에 의해 결정 가능한 값은 무엇인가?

정답 하이퍼 파라미터
해설 머신러닝(기계학습)과 딥러닝을 수행하는 과정에서 우리가 직접 수동으로 설정해주어야 할 부분이 있습니다. 하이퍼 파라미터들은 설정을 통해서 데이터의 모델 성능을 향상시켜 최적의 훈련 모델을 구현하는 데 필수적인 값입니다. 실무에서는 여러 번 하이퍼 파라미터들을 변경하면서 최적의 값 을 찾는 노력이 필요합니다.

[문133] 데이터 학습을 위해 차원이 증가하면서 학습데이터 수가 차원의 수보다 적어져 성능이 저하되는 현상으로서, 차원이 증가함에 따라(=변수의 수 증가) 데이터 사이의 거리가 멀어져 데이터의 밀도가 낮아져서 빈 공간에 0으로 채워져서 모델의 성능이 하락하는 현상을 무엇이라고 하는가?

정답 차원의 저주
해설 차원의 저주에 대한 설명입니다.

3.2 분석 기법 적용

[문134] 구간이나 비율에 의미가 담긴 수치적 두 변수 간의 선형적 연관성을 계량적으로 파악하기 위한 통계적 기법이며, 일반적으로 선형적인 관계 정도를 측정하는 척도는 무엇인가?

정답 피어슨 상관계수
해설 피어슨 상관계수는 두 연속형 변수 간에 선형 관계가 얼마나 강한지를 보여주는 지표입니다. 대표적인 사례로는 키와 몸무게 변수 간에 피어슨 상관계수를 구하는 것으로, 키가 크면 몸무게도 증가하는지 등에 대한 선형성 수준을 확인할 수 있습니다. 계수의 값이 +1에 가까울수록 양의 상관관계를, -1에 가까울수록 음의 상관관계를 의미하며, 0은 두 변수 간에 관계가 없음을 나타냅니다.

[문135] 이것은 데이터 안의 두 변수 간의 관계를 알아보기 위해 사용하는 값이다. 두 변수간의 공분산으로는 음과 양의 관계를 파악할 수 있으나 관계 정도를 확인하기는 힘들다. 그래서 각 변수의 표준편차를 곱하여 공분산을 나누어 -1에서 1사이의 값으로 표준화하여 두 변수 간의 관계 정도를 확인 할 수 있도록 수치화 한 이것을 활용한다. 이것은 무엇인가?

정답 상관 계수
해설 상관 계수(correlation coefficient)에 대한 설명입니다. 공분산을 -1에서 1사이의 값으로 표준화하여 변수 간 강도를 비교할 수 있습니다.

[문136] 비선형적인 관계도 파악할 수 있는 상관계수는 무엇인가?

정답 스피어만 상관계수
해설 스피어만 상관계수에 대한 설명입니다. 비선형적 또는 서열적 자료 간의 관계에 적합합니다.

[문137] 하나 이상의 독립변수들이 종속변수에 미치는 영향을 추정할 수 있는 통계기법으로, 변수들 사이의 인과관계를 밝히고 독립변수를 조작하면서 종속변수가 어떻게 변하는지를 보며 두 변인의 관계를 파악함으로써 모형을 적합 시켜 관심 있는 변수를 예측하거나 추론하는 분석 방법을 무엇이라고 하는가?

정답 회귀 분석
해설 회귀 분석은 변수들 사이에 수학적인 관계성을 찾아 이를 수학적인 모형으로 도출하고, 이렇게 도출된 수학적 모형으로 원하는 변수를 예측하는 데이터 분석 방법입니다. 기초적인 사례로 키와 몸무게 변수 간의 수학적 회귀 모형을 들 수 있습니다. 이를 통해 키가 180cm인 경우와 키가 2m인 경우 등에 대해 몸무게를 예측할 수 있게 됩니다.

[문138] 회귀분석에서 사용된 모형의 일부 설명 변수가 다른 설명 변수와 상관 정도가 높아 데이터 분석 시 부정적 영향을 미치는 것(회귀분석의 기본 가정인 독립성에 위배하는 문제를 무엇이라고 하는가?

정답 다중공선성
해설 다중공선성은 분석하려는 독립변수들 간의 상관성이 높아서 통계적인 결과의 정확도를 떨어뜨리는 속성을 말합니다. 만약 태어난 월(月), 태어난 계절, 태어난 분기 등이 독립변수로 이루어졌다고 가정하면, 태어난 월 2월, 태어난 계절 겨울, 태어난 분기 1분기는 상관관계가 매우 높은 변수들이므로 다중공선성이 있다고 판단할 수 있습니다. 따라서 세 개의 변수 중에 대표적인 변수를 하나 선택하여 다중공선성을 제거하는 방식으로 해결할 수 있습니다.

[문139] 회귀분석의 5가지 기본 가정에 대해 쓰시오

정답 선형성, 독립성, 등분산성, 정규성, 비상관성
해설 회귀분석은 독립변수와 종속변수가 직선, 즉 선형적인 관계에 있어야 적용이 가능하며(선형성), 측정값은 어떤 경향이나 시간적 추세가 없이 독립적으로 측적된 것이하여 합니다(독립성). 또한 X와 Y의 관계가 구간에 따라 분산이 유사해야 하고(등분산성), 실제값과 예측된 값의 차이는 정규분포의 형태를 띄어야 합니다(정규성), 독립변수들 간에는 지나치에 상관이 적어야 합니다(비상관성)

[문140] 다음 괄호 안에 들어갈 단어를 순서대로 기입하시오
“우리는 모집단의 실제값과 회귀선과의 차이인 ((1)을 알아낼 수 없기에 표본에서 나온 관측값과 회귀선의 차이인 ((2))를 이용해 분석을 수행한다”

정답 (1)오차, (2)잔차
해설 실제값과 회귀선(예측값)과의 차이는 오차라고 하며 이를 파악하기 위해 잔차 분석을 수행합니다.

[문141] 전체 변동 중 회귀모형에 의해 설명되는 변동의 비율로, 표본에 의해 추정된 회귀식이 주어진 자료를 얼마나 잘 설명하는지를 보여주는 값으로서 주어진 데이터에 회귀선이 얼마나 잘 맞는지, 적합 정도를 평가하는 척도이자 독립변수들이 종속변수를 얼마나 잘 설명하는지 보여주는 지표는 무엇이라고 하는가?

정답 결정계수
해설 결정계수 혹은 설명력에 대한 설명입니다. 실제 측정값이 회귀선에 가까울수록 결정계수는 1, 즉 100%에 가까운 예측력을 보입니다.

[문142] 회귀모형의 계수를 추정하는 방법으로써 잔차제곱합을 최소화하는 계수를 찾는 방법은?

정답 최소제곱법
해설 실제값과 예측값의 차이를 최소화하면서 회귀선을 찾는 방법을 최소제곱법이라고 하며, 선형회귀분석 방법의 핵심입니다.

[문143] 시계열 분석에서 평균이 일정하지 않은 비정상 시계열을 정상 시계열로 바꾸기 위해서, 현재 시점에서 이전의 시점을 빼는 방법은 무엇이라고 하는가?

정답 차분
해설 시계열 자료에서 추세를 없애기 위해 이전 시점의 값과 현재 시점의 값을 빼서 새로운 변수를 만드는 것을 차분이라고 합니다. 시계열을 추세가 없는 안정적인(stationary) 자료로 만들기 위한 방법입니다.

[문144] 시계열 분석의 특성으로써, 시점에 상관없이 시계열의 특성이 일정함을 의미하는 속성은 무엇이라고 하는가?

정답 정상성(stationarity)
해설 정상성은 모든 시점의 평균이 일정하고, 뚜렷한 추세가 없어서 미래에도 과거와 동일하다는 속성입니다. 정상성을 만족하는 세 가지 조건은 다음과 같습니다.
(1) 평균이 일정하다.
(2) 분산이 시점에 의존하지 않는다.
(3) 공분산은 시차에만 의존하고 시점 자체에는 의존하지 않는다.

[문145] 시계열 분석의 기본이 되는 중요한 개념으로 시계열의 평균과 분산이 일정하고 일정한 추세가 없는 것을 무엇이라 하는가?

정답 정상 시계열
해설 정상성(stationarity)이 확보된 시계열 데이터를 정상 시계열 자료라고 합니다.

[문146] 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법으로 분석목적에 따라 특정 요인만 분리 분석하거나 제거하는 방법은 무엇이라고 하는가?

정답 분해 시계열
해설 시계열 자료 속에 있는 추세, 계열요인, 변동 등을 나눠서, 즉 분해해여 세부적인 정보를 얻기 위한 것을 시계열 자료를 분해(decompose)한다고 합니다.

[문147] 다음은 무엇에 대한 설명인가?
“변수들의 자기상관성을 기반으로 한 시계열 모형으로 현시점의 자료를 p시점 전의 과거 자료를 통해 설명할 수 있는 모델이다. 자기 자신의 과거 값이 이후 자신의 값에 영향을 준다. 현시점의 자료가 k 시점 이전의 유한개의 과거 자료로 설명할 수 있는 모형으로 자기 자신의 과거 값이 이후 자신의 값에 영향을 주기 때문에 이런 이름이 지어졌다.”

정답 자기회귀모형(AutoRegressive, AR모형)
해설 시계열분석은 기본적으로 회귀분석을 응용한 것입니다. 다만 타변수(독립변수)보다 자기 변수의 이전 시점 값을 사용하기 때문에 자기회귀라고 합니다. 여기서 자기는 자동이라는 의미가 아니라 자기 자신을 의미합니다.

[문148] 다음은 무엇에 대한 설명인가?
“현재 데이터가 과거 백색잡음의 선형 가중합으로 구성된다는 모형으로 시간이 갈수록 관측치의 평균값이 지속해서 증가하거나 감소하는 시계열모형이다. 백색잡음 과정은 서로 독립이고 평균이 0인 확률변수이므로 항상 정상성을 만족한다.”

정답 이동평균모형(Moving Average, MA모형)
해설 선형 가중합이라는 것이 결국 평균을 낸다는 것입니다. 다만 단순한 이전 시점들의 평균이 아니라 가까운 시점에 더 많은 가중치를 두기 때문에 가중합 혹은 가중평균을 냅니다.

[문149] 다음은 무엇에 대한 설명인가?
“데이터가 비정상성이 아닌 증거를 나타내는 경우에 적용되며, 초기 차분 단계(모델의 "통합된" 부분에 해당)를 한 번 이상 적용하여 비정상성을 제거할 수 있다. 분기 , 반기, 연간 단위로 다음 지표를 예측하거나 주간, 월간 단위로 지표를 리뷰하여 경향을 분석하는 기법이다.”

정답 자기회귀누적 이동평균모형(AutoRegressive Integrated Moving Average, ARIMA모형)
해설 AR모형과 MA모형을 통합(Integrated)한 모형으로서 ARIMA모형에 대한 설명입니다.

[문150] 반응변수가 범주형인 경우에 적용하는 회귀 모형으로, 설명변수(독립변수)의 값이 주어질 때 각 범주에 속할 추정확률을 기준치에 따라 분류하는 목적으로 사용하는 분석 기법을 무엇이라고 하는가?

정답 로지스틱 회귀 분석
해설 종속변수가 범주형 변수인 경우에 로지스틱 회귀 분석으로 종속변수에 대한 예측 모델을 만들 수 있습니다. 종속변수(Y)는 남/여, 성공/실패, 암/정상, 지연/정상, 합격/불합격 등과 같이 0 또는 1로 표현할 수 있는 값입니다.

[문151] 로지스틱 회귀분석에서 어떠한 일이 일어날 확률을 일어나지 않을 확률로 나누어 log를 취하고 이를 0~1의 값이 아닌 (-무한대, 무한대) 범위에서 선형함수를 시그모이드 함수로 변환하는 방법은 무엇인가?

정답 로짓 변환
해설 범주형 종속변수이므로 이를 확률과 로그 변환하여 선형회귀에 적용하는 방법이 로지스틱 모델의 핵심 과정입니다. 여기서 선형함수를 시그모이드 함수로 변환하는 방법을 로짓 변환이라고 합니다. 시그모이드는 로지스틱 함수와 동일한 말로서 S자 모양이라는 뜻입니다.

[문152] 베이즈 정리와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전 정보와 데이터로부터 추출된 정보를 결합하고 베이즈 정리를 이용하여 어떤 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘은 무엇인가?

정답 나이브 베이즈 분류
해설 나이브 베이즈 분류에 대한 설명입니다. 베이즈 정리와 특징에 대한 조건부 독립이라는 핵심적 방법론을 활용합니다.

[문153] 각 데이터들이 가진 속성들로부터 분할 기준 속성을 판별하고, 분할 기준 속성에 따라 트리 형태로 모델링하는 분류 예측 모델은 무엇인가?

정답 의사결정나무
해설 의사결정나무는 데이터 특성을 나무 모양의 가지로 활용해서 예측값을 찾아가는 모형입니다. 다음 그림과 같이 상식적인 수준에서 이해하기 쉽고, 결과의 해석이 매우 쉬운 특징을 가지고 있습니다.

[문154] 의사결정나무 형성 과정 중에서 오차를 크게 할 위험이 높거나 부적절한 추론 규칙을 가지고 있는 가지 또는 불필요한 가지를 제거하는 방법은?

정답 가지치기
해설 가지치기는 의사결정나무의 분기들이 많아서 데이터의 과적합이 발생할 위험을 방지하는 방법입니다. 전체 의사결정나무를 생성한 후에 적절한 가지를 잘라내며, 잘라진 가지는 버리는 것이 아닌 정상 가지들과 합치는 개념으로 이해하면 좋습니다.

[문155] 의사결정나무 중 연속형 타깃변수(또는 목표변수)를 예측하는 의사결정나무를 무엇이라고 하는가?

정답 회귀나무
해설 분류의 문제가 아닌 회귀의 문제를 의사결정나무에 적용할때는 회귀나무라고 별도로 칭합니다.

[문156] 의사결정 나무에서 더 이상 분기되지 않도록 하는 규칙은?

정답 정지규칙
해설 정지규칙에 대한 설명입니다.

[문157] 기계학습의 한 분야로 공간상에서 최적의 분리 초평면을 찾아서 분류 및 회귀를 수행할 수 있고, 데이터를 분리하는 초평면 중에서 데이터들과 거리가 가장 먼 초평면을 선택하여 분리하는 지도 학습 기반의 이진 선형 분류 모델로서 훈련 시간이 상대적으로 느리지만 다른 기계학습방법 대비 과대적합의 가능성이 낮은 모델은?

정답 서포트 벡터 머신 또는 SVM
해설 서로 다른 데이터를 분리하기 위해 가장 적당한 선이나, 면, 초평면(결정경계)을 찾아서 분류하는 방법이 SVM입니다. 데이터를 분리하는 초평면과 가장 가까이 있는 데이터는 서포트 벡터라고 하고, 초평면과 서포트 벡터 간의 거리는 마진이라고 합니다.

[문158] 여러 개의 결정 트리 분류기(같은 알고리즘이 여러 개)가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 수한한 뒤 최종적으로 모든 분류기가 보팅을 통해 예측을 결정하는 앙상블 알고리즘은?

정답 랜덤 포레스트
해설 의사결정나무를 앙상블로 여러개(수십~수백개) 활용하는 방법은 랜덤 포레스트입니다. 랜덤 포레스트는 의사결정나무를 뻥튀기한 후, 각 의사결정나무에서 도출된 결과들을 취합하고 다수결을 통해 결론을 얻는 방법입니다. 다음 그림에는 7개의 의사결정나무가 있고, 5개는 A 그룹이라는 결과를 2개는 B 그룹이라는 결과를 도출했습니다. 결과적으로 A 그룹이 5개로 50% 이상의 의사결정나무가 손을 들어주었으므로, A 그룹의 값을 최종 결론으로 삼습니다.

[문159] 동일하거나 다른 학습 알고리즘을 사용해서 여러 모델을 학습하는 개념으로서 주어진 자료로부터 여러 개의 예측모형들을 만든 후 예측모형들을 조합하여 하나의 최종 예측 모형을 만들어 분류 정확성을 향상시키는 기법은?

정답 앙상블 기법
해설 앙상블은 여러 개의 모형을 만들어서 나중에 결합하여 최종 결과를 생성합니다. 대표적인 앙상블 기법으로 배깅, 부스팅이 있습니다.

[문160] 크기가 같은 표본을 여러 번 단순임의 복원 추출하여 분류기를 생선한 후 앙상블하는 기법이며 모델의 안정성을 높이기 위하여 분석 데이터로부터 여러 개의 단순 복원 임의 추출하여 다수결을 통해 최종의 예측 모델을 도출하는 알고리즘이다. 각각의 분류기가 모두 같은 유형의 알고리즘 기반이지만, 데이터 샘플링을 서로 다르게 가져가면서 학습을 수행해 보팅을 수행하는 방법을 무엇이라고 하는가?

정답 배깅
해설 앙상블 방법 중 배깅(bagging)에 대한 설명입니다. 배깅은 부트스트랩과 결합의 합성어입니다. 즉, 단순 복원 임의추출을 통한 여러 개의 부트스트랩을 만들고, 각 부트스트랩에서 모델링을 수행합니다. 이후 부트스트랩의 모델링 결과를 결합해서 최종적인 예측 모형을 산출합니다. 여기서 부트스트랩은 원본 데이터에서 중 복을 허용한 복원 임의추출을 수행한 표본 데이터를 의미합니다.

[문161] 성능이 약한 학습기를 여러 개 연결하여 순차적으로 학습하여, 정답을 맞히지 못한 부분에 가중치를 부여함으로써 강한 학습기를 생성하는 앙상블 기법으로 모델의 정확성을 높이기 위해 오분류된 개체들에 가중치를 부여함으로써 새로운 분류규칙을 생성 및 반복하여 약한 분류 모델을 강한 분류모델로 변형하는 알고리즘은 무엇인가?

정답 부스팅
해설 앙상블 방법 중 부스팅(boosting)에 대한 설명입니다. 부스팅은 최초에는 같은 가중치로 모델링을 수행하지만, 이후에 순차적으로 모델링을 반복하면서 잘못된 결과에는 가중치를 높게 부여합니다. 가중치가 높게 부여된 잘못된 결과는 이후에 더 잘 분류하여 모형의 정확도를 향상시키도록 합니다. 배경에 비하여 순차적으로 모델링을 수행하기 때문에 속도는 느리지만, 보다 정확한 결과를 도출할 수 있습니다.

[문162] 주어진 자료에서 단순 복원 임의추출 방법을 활용하여 동일한 크기의 표본을 여러 개 생성하는 샘플링 방법으로 재표본추출 방법의 일종으로 중복추출을 허용하여 주어진 자료에서 단순 랜덤 복원추출 방법을 활용하여 동일한 크기의 표본을 여러개 생성하는 샘플링 방법은?

정답 부트스트랩
해설 부트스트랩은 중복을 허용해서 임의로 추출하는 표본 데이터 생성 방법입니다.

[문163] 고객의 대규모 거래데이터로부터 함께 구매가 발생하는 규칙을 도출하여, 고객이 특정 상품 구매 시 이와 연관성 높은 상품을 추천하는 분석으로 어떤 변인 간에 주목할 만한 상관관계가 있는지를 찾아내는 방법은?

정답 연관규칙 분석(장바구니분석)
해설 연관규칙 분석에 대한 설명입니다.

[문164] 비지도학습의 연관규칙 분석 기법 중에서 발생 항목 집합에서 연관 관계를 찾아내는 것으로, 일명 장바구니 분석이라고도 불리는 알고리즘은?

정답 Apriori 알고리즘
해설 Apriori 알고리즘은 지지도, 신뢰도 향상도의 과정을 거쳐 두 개의 아이템 간 연관관계를 수치로 확인하는 알고리즘입니다.

[문165] 연관규칙에서 두 아이템의 연관규칙이 유용한 규칙일 가능성의 척도를 무엇이라고 하는가?

정답 신뢰도
해설 A를 구매한 경우에 B를 구매한다는 가정 하에서 신뢰도는 A를 구매한 건수로 A와 B를 모두 구매한 건수를 나누어 구합니다. 신뢰도의 결과가 높을수록 의미 있는 규칙일 가능성이 큽니다. 이는 조건부 확률이라고 부를 수 있습니다.

[문166] 연관규칙 분석에서 품목간 상관관계를 기준으로 규칙의 예측력을 평가하는 지표로서 A→B의 연관 규칙에서 임의로 B가 구매되는 경우에 비해 A와의 관계가 고려되어 구매되는 경우의 비율로서 연관규칙에서 두 아이템의 연관 규칙이 우연인지 아닌지를 나타내는 척도는?

정답 향상도
해설 향상도에 대한 설명입니다.

[문167] 연관규칙에서 특정 아이템이 전체 데이터에서 발생하는 척도이자 전체 거래 중에서 A, B 아이템이 동시에 포함된 거래 비율은?

정답 지지도
해설 지지도는 장바구니 분석인 Apriori 알고리즘에서 세 가지 척도 중 하나입니다. 2개의 아이템(A,B)이 모두 포함된 건수를 전체 건수로 나눠서 지지도를 계산합니다. 지지도 계산값이 높을수록 2개의 아이템이 포함된 건수가 많음을 알 수 있습니다.

[문168] 주어진 각 개체들의 유사성을 분석해서 높은 대상끼리 일반화된 그룹으로 분류하는 기법을 무엇이라고 하는가?

정답 군집분석
해설 군집분석에 대한 설명입니다. 유사성이 높은 개체들을 그룹화하는 대표적인 비지도학습 방법입니다.

[문169] 다음은 무엇에 대한 설명인가?
“주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. 이 알고리즘은 자율 학습의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행한다.”

정답 k-평균 군집화 알고리즘
해설 k-평균 군집화 알고리즘, k-mean 방법이라고 합니다. 군집분석은 계층적 방법과 k-means의 비계층적 방법으로 나뉩니다.

[문170] 다음은 무엇에 대한 설명인가?
“군집 간의 거리를 측정하는 방법 중에서 군집의 방향으로 인한 오차 제곱합의 증가량이 최소가 되도록 군집을 형성하는 군집 간 거리 측정 방법”

정답 와드 연결법
해설 와드(ward) 연결법은 군집 간의 거리가 아닌 군집 내 편차의 제곱합을 기반으로 군집을 만들어가는 방법입니다. 와드 연결법은 크게 계층적 군집 내에 속한 방법으로 개별 데이터를 하나의 군집이라고 간주한 후, 두 군집을 순차적으로 묶어 큰 군집으로 생성하는 방식입니다. 군집으로 만드는 기준에 따라서 다음 그림과 같은 연결법들이 있습니다.

[문171] 군집 분석에서 군집 내의 응집도와 군집 간의 분리도를 계산하여 1에 가까울수록 군집 결과가 좋은 것이고, -1에 가까울수록 군집 결과가 좋지 않은 것으로 해석하는 지표를 무엇이라고 하는가?

정답 실루엣
해설 실루엣은 각각의 군집이 잘 분리되었는지를 나타냅니다. 1에 가까울수록 다른 군집끼리는 멀리 떨어져 있고, 같은 군집에서 데이터들이 똘똘 뭉쳐 있다는 의미입니다.

[문172] 동일한 상대적 거리를 가진 실수 공간의 점들로 대상들을 배치시키는 방법으로서 개체들의 유사성과 비유사성을 측정하여 N차원 공간에 점으로 표현하여 개체들 간의 집단화를 시각적으로 표현하는 분석 방법은 무엇이라고 하는가?

정답 다차원 척도법
해설 다차원 척도법은 분석 대상의 데이터들 간의 유사한 정도를 시각화한 통계기법입니다. 만약 시중에 판매되는 네 가지 라면이 있다고 가정하면, 각 라면 브랜드가 고객 관점에서 어떻게 자리 잡고 있는지를 5점 척도(1점 ~ 5점)로 분석합니다. A 라면과 B 라면이 얼마나 비슷한지에 대한 개별 쌍으로 유사성을 2차원 그래프에 거리의 간격으로 표현하는 것입니다. 즉, 시각적으로 그려진 그림으로 브랜드별로 차별화 여부를 확인할 수 있습니다.

[문173] 데이터 안에 관찰할 수 없는 잠재적인 변수가 존재한다고 가정하는 차원축소기법. 모형을 세운 뒤 관찰 가능한 데이터를 이용하여 해당 잠재 요인을 도출하고 데이터 안의 구조를 해석하는 기법은?

정답 요인분석
해설 요인분석에 대한 설명입니다. 요인분석의 핵심 키워드는 잠재적 공통 요인을 도출하는 것입니다.

[문174] 상관성이 높은 변수들을 선형 결합하여 기존의 상관성이 높은 변수들을 요약하고자 축소하는 기법으로서, 차원을 단순화하여 상관성 있는 변수들 간의 복잡한 구조를 분석하는 차원 축소에 해당하는 통계기법은 무엇이라고 하는가?

정답 주성분 분석
해설 분석하려는 데이터는 수학적으로 다차원 또는 다양한 방향에 분포해 있습니다. 따라서 데이터의 특성을 보다 잘 드러내는 주인공 차원을 찾아내는 주성분 분석을 수행해야 합니다. 다음 그림에서는 x축과 y축의 2차원에서 나뭇잎 모양으로 분포된 데이터를 확인할 수 있습니다. 해당 데이터에서 주성분 분석을 통해 주성분 1과 주성분 2로 데이터의 독특한 특징을 추출할 수 있습니다.

[문175] 여러 변수의 변량을 서로 상관성이 높은 변수들의 선형 조합으로 만든 새로운 변수로 요약 및 축소하는 기법은 무엇이라고 하는가?

정답 PCA(Principal Components Analysis)
해설 주성분분석, PCA에 대한 설명입니다. 변수의 선형 조합이라는 방법으로 변수를 축소합니다.

[문176] 고차원에 존재하는 데이터 간의 거리를 최대한 보존하면서 데이터 간의 관계를 저차원으로 축소해 시각화하는 방법은 무엇이라고 하는가?

정답 t-SNE(t-분포 확률적 임베딩, Stochastic Neighbor Embedding)
해설 t-SNE에 대한 설명입니다. 기존의 PCA 등보다 확률분포 상에서 발생확률로 골고루 분포하여 더 자연스러운 시각화를 가능하게 합니다.

[문177] 행렬의 크기가 다른 M*N 행렬에 대해 세 행렬의 곱으로 분해하는 것으로 데이터 압축 등의 많은 분야에서 활용되며 M*N 차원의 행렬 데이터에서 특잇값을 추출하고 이를 통해 주어진 데이터 세트를 효과적으로 축약할 수 있는 차원 축소 기법은?

정답 특잇값 분해(Singular Value Decomposition, SVD)
해설 특잇값 분해에 대한 설명입니다. 협업필터링 추천 알고리즘에서도 활용됩니다.

[문178] 두 벡터 사이의 각도를 이용하여 벡터간의 유사 정도를 측정하는 방식은?

정답 코사인 유사도
해설 코사인 유사도에 대한 설명입니다. 코사인 유사도는 두 값의 거리가 아닌 각도로 유사도를 측정합니다.

[문179] 다양한 문서 자료 내 비정형 텍스트 데이터에 자연어 처리(NLP) 기술 및 문서처리 기술을 활용해 인사이트를 도출하는 기술을 무엇이라고 하는가?

정답 텍스트 마이닝
해설 텍스트 마이닝(text mining)에 대한 설명입니다.

[문180] 자연어 분석 작업의 대상이 되는 대량의 텍스트 문서들을 모아놓은 집합은?

정답 말뭉치(corpus)
해설 코퍼스 혹은 말뭉치라고 합니다.

[문181] 구조화되어 있지 않은 문서를 단어로 나누는 과정을 무엇이라고 하는가?

정답 토큰화
해설 토큰화 혹은 토크나이징(tokenizing)이라고 합니다.

[문182] 말뭉치에서 자주 등장하지만, 분석에 있어 기여하는 바가 없는 단어를 무엇이라고 하는가?

정답 불용어(stopword)
해설 이를 불용어라고 합니다.

[문183] 텍스트 마이닝의 전처리 과정으로 어형이 변형된 접사를 제거하고, 단어의 원형 또는 어간을 분리하는 과정을 무엇이라고 하는가?

정답 스테밍
해설 비정형 데이터에서 유의미한 정보를 추출하기 전, 비정형 데이터의 전처리 과정으로 스테밍 과정을 수행합니다. 즉, 단어의 원형으로 변경하는 것입니다.

[문184] 주관적인 의견이 포함된 데이터에서 사용자가 게재한 의견과 감정을 나타내는 패턴을 분석하는 기법을 무엇이라고 하는가?

정답 오피니언 마이닝
해설 오피니언 마이닝은 웹사이트나 소셜미디어에서 사람들의 의견이나 댓글, 게시글 등을 분석하여 여론이나 감정, 평판 등을 도출하는 데이터 분석 기법입니다.

[문185] 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법을 무엇이라고 하는가?

정답 워드 클라우드(Word Cloud)
해설 워드 클라우드에 대한 설명입니다.

[문186] 특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석을 무엇이라고 하는가?

정답 감성분석
해설 감성분석에 대한 설명입니다.

[문187] 오피니언 리더, 즉 영향력 있는 사람을 찾아낼 수 있으며, 고객 간 소셜 관계를 파악하는 방법을 무엇이라고 하는가?

정답 소셜 네트워크 분석
해설 소셜 네트워크 분석에 대한 설명입니다.

[문188] 최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 매커니즘을 통해 점진적으로 진화시켜 나가는 방법을 무엇이라고 하는가?

정답 유전자 알고리즘
해설 유전자 알고리즘은 다윈의 적자생존 이론을 기반으로 진화를 통해 최적화를 거치는 방법입니다. 진화를 통해 보다 좋은 해들을 만들어가는 과정을 거치게 됩니다.

[문189] XOR과 같이 선형 분리할 수 없는 문제에 약점을 지닌, 입력층과 출력층으로만 구성된 최초의 인공신경망을 무엇이라고 하는가?

정답 퍼셉트론
해설 퍼셉트론은 중간층(뉴런)과 출력층의 구분 없이 단순하게 구성된 인공신경망으로, Exclusive OR (XOR) 계산이 불가능한 문제가 있습니다. 이에 XOR 계산이 가능한 다층 퍼셉트론이 탄생한 계기가 됩니다.

[문190] 신경망 학습에서 오차를 출력층에서 입력층으로 전달하여 가중치와 편향을 갱신하는 알고리즘을 무엇이라고 하는가?

정답 역전파 알고리즘
해설 일반적으로 입력층에서부터 가중치를 업데이트하면서 출력층으로 결팟값을 도출하는 방식으로 학습합니다. 이와 같이 임의로 설정한 가중치로 계산하는 순전파 방식은 많은 오차가 발생하여 낮은 성능을 보입니다. 이에 오차를 줄이기 위하여 출력층에서 입력층 방향으로 가중치를 업데이트하는 방식을 사용합니다. 이를 역전파라고 합니다.

[문191] 다층 신경망에서 은닉층이 많아 학습이 이루어지지 않아 활성화 함수인 시그모이드 함수에서 편미분을 진행할수록 0으로 근접해지는 현상을 무엇이라고 하는가?

정답 기울기 소실
해설 인공신경망에서 학습을 수행하는 경우, 출력층에서 입력층 방향으로 거꾸로 되돌아가면서 가중치를 조정하는 역전파 과정을 거칩니다. 이때 가중치를 조정하는 방법으로 미분값(기울기)을 계산하는데, 입력층에 가까워질수록 미분값이 0에 가까워집니다. 따라서 가중치 조정에 영향을 주지 않는 문제가 발생하며, 이를 기울기 소실이라고 합니다.
이는 활성화 함수로 사용하는 시그모이드 함수의 특성으로 기울기 소실이 발생한 것이며, ReLU 함수 등의 다른 활성화 함수를 사용하여 해당 문제를 해결할 수 있습니다.

[문192] 인공신경망에서 입력층과 출력층 사이에 위치하여 내부적으로 동작하는 계층을 무엇이라고 하는가?

정답 은닉층
해설 인공신경망은 입력층과 출력층 사이에 외부에서 접근할 수 없는 은닉층을 구성하고 있습니다. 은닉층 수가 늘어날수록 신경망의 복잡도는 향상되며, 이를 다층 신경망이라고 부릅니다. 다음 그림은 입력층에 3개 노드, 1개의 은닉층에 2개 노드, 출력층에 3개의 노드가 있습니다.

[문193] 인공신경망 모델에서 입력 신호의 총합을 출력 신호로 변환하는 함수로, 입력받은 신호를 얼마나 출력할지 결정하고 다음 단계에서 출력된 신호의 사용 여부를 결정하는 함수를 무엇이라고 하는가?

정답 활성화 함수
해설 활성화 함수는 입력값을 출력값으로 내보낼 때, 선형이 아닌 비선형 방식으로 값을 변환할 수 있습니다. 이를 통해서 다층 퍼셉트론의 은닉층들을 구성하고, XOR 문제를 해결할 수 있습니다.

[문194] 신경망 학습에서 평균 제곱 오차 또는 교차 엔트로피 오차를 사용하여 현재의 상태를 나타내는 지표를 무엇이라고 하는가?

정답 손실함수
해설 손실함수는 비용함수라고도 불리며, 실제값과 예측값이 얼마나 차이가 나는지를 나타내는 수치입니다. 결국은 실제값과 예측값의 차이를 줄이는 방향으로 데이터 학습이 진행되어야 합니다. 그럼 실제값과 예측값을 어떻게 비교할까요? 다양한 방법이 있지만, 대표적으로 오차의 제곱에 대한 평균인 MSE, MSE에 루트를 씌운 RMSE 등이 있습니다.

[문195] 출력값 z가 여러 개로 주어지고 목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공하는 함수는?

정답 소프트맥스 함수
해설 범주가 3개 이상일 때 각각에 소갈 확률값의 합을 1로 만들기 위한 방법으로 활용됩니다.

[문196] 모델의 파라미터값을 측정하기 위해 알고리즘 구현 과정에서 사용, 주로 알고리즘 사용자에 의해 결정, 경험에 의해 결정 가능한 값이며 모델 성능 향상을 위해 조절해주는 값은?

정답 하이퍼 파라미터
해설 하이퍼 파라미터(hyper-parameter)에 대한 설명입니다.

[문197] 과대 적합을 방지하기 위해 인공지능 학습 과정에서 일부 신경망 일부만 동작하고 일부는 동작하지 않도록 하는 방법?

정답 드롭아웃
해설 드롭아웃에 대한 설명입니다.

[문198] 데이터상의 주석 작업으로 딥러닝과 같은 학습 알고리즘이 무엇을 학습하여야 하는지 알려 주는 표식 작업을 무엇이라고 하는가?

정답 어노테이션 (Annotation)
해설 어노테이션에 대한 설명입니다. 이미지나 영상에 인물, 장소 등 태그(tag)를 주고 라벨링(labeling)을 하는 과정을 어노테이션이라고 하며 학습 데이터를 만들 때 매우 중요합니다.

[문199] 기존 영상처리의 필터 기능과 신경망을 결합하여 효과적인 성능을 발휘하도록 만든 구조로서, 시각적인 이미지를 분석하는 데 사용되는 신경망 모델은?

정답 CNN 또는 합성곱 신경망
해설 CNN은 데이터의 특징을 추출해서 패턴을 파악하는 심층 심경망입니다. 구체적으로 데이터의 특징을 추출하는 합성곱 층과 도출된 여러 층의 합성곱 크기를 줄여주는 풀링 과정이 있습니다. 이를 통해 얼굴 인식 등의 이미지 인식 분야에서 널리 활용하고 있습니다.

[문200] 연속적인 시계열 데이터를 분석할 수 있는 신경망으로, 경사하강법과 시간기반 오차 역전파를 사용해서 가중치를 업데이트하며, 은닉층에서 재귀적인 신경망을 갖는 알고리즘을 무엇이라고 하는가?

정답 RNN 또는 순환신경망
해설 RNN은 시간 정보가 포함된 시계열 데이터를 대상으로 삼으며, 내부에는 반복적인 순환 구조가 존재하는 특징을 가지고 있습니다. 순환 구조를 통해 과거의 학습 결과와 가중치를 결합하여 현재 학습에 반영합니다. 즉, 현재와 과거의 학습 결과가 서로 연결되는 시간적 종속성을 지니므로, 시간정보가 포함된 음성 인식이나 필기체 인식 등에 활용됩니다.

[문201] 코호넨 맵이라고도 불리며, 인공신경망을 기반으로 차원축소와 군집화를 동시에 수행할 수 있는 알고리즘으로서 코호넨에 의해 제시되었으며, 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화하는 클러스터링 방법은 무엇인가?

정답 자기조직화지도(SOM, Self Organizing Maps)
해설 자기조직화지도에 대한 설명입니다. 타겟, 레이블이 없는 상황에서 지도학습과 유사한 결과물을 얻을 수 있어 실무에서 활용도가 높습니다.

[문202] 새로운 모델을 만들 시, 기존의 만들어진 모델을 사용하여 학습을 빠르게 하며, 모델 성능을 높이는 방법은?

정답 전이학습
해설 전이학습은 처음부터 모델을 만들지 않고 이미 검증된 모델과 그 모델에서 학습된 가중치를 이용하기 때문에 대량의 학습을 보다 빠르게 진행할 수 있고 성능 역시 보장되어 최근에 활용도가 높습니다.

4. 빅데이터 결과 해석

4.1 분석 모형 평가 및 개선

[문203] 혼동행렬을 활용한 평가지표 중에서 민감도와 동일한 계산식이며, 모형의 완전성을 평가하는 지표를 무엇이라고 하는가?

정답 재현율
해설 일명 민감도라고 불리는 재현율은 다음 식으로 계산합니다. 이후 '혼동행렬' 용어를 참고하여 TP와 TN에 대한 의미를 확인합시다.

$$재현율 = \frac{TP}{TP+FN}$$

[문204] 분석 모델에서 구한 분류의 예측 범주와 데이터의 실제 분류 범주를 교차표 형태로 정리한 행렬을 무엇이라고 하는가?

정답 혼동 행렬
해설 혼동행렬은 다음 표 기준으로 암기가 필요합니다. 그전에 각각의 의미를 파악해보도록 합시다. 먼저 가로축은 예측하는 관점이고 세로축은 실제 일어난 일에 대한 관점입니다. 이해를 돕기 위해서 여기서는 암에 걸린 환자와 정상인을 예측한다고 가정해봅시다.
TP는 암에 걸릴 것으로 예측했고 실제로도 암에 걸린 환자인 경우로 정확하게 맞춘 정답입니다. 즉, 암을 맞춘 것이므로 True 키워드가 포함됩니다. FN은 정상인이라고 예측했지만 사실은 암에 걸린 환자인 경우입니다. 오답이므로 False 키워드가 포함됩니다. FP도 암일 것으로 예측했지만 사실은 정상인인 경우입니다. 잘못 예측했으므로 오답이며, False 키워드가 포함됩니다. TN은 정상이라고 예측했고, 실제로도 정상이므로 정답입니다. 예측값과 실제값이 같으므로 True 키워드가 포함됩니다. 가로와 세로의 기준과 그 안의 긍정, 부정을 암기하여 혼동하지 않도록 합니다.

[문205] 혼동행렬을 통해서 다음 수식으로 계산하는 모델 평가지표는 무엇인가?

정답 F1 스코어
해설 F1 스코어는 정밀도와 재현율의 조화평균으로, 균형 잡히지 않은 데이터에서 모델 성능을 측정하는 데 적합합니다. $2\times\frac{precision\times recall}{precision+recall}$로 계산됩니다.

[문206] 이진 분류에서 모형이 예측한 값과 실제 값의 조합을 교차표 형태로 정리한 행렬을 무엇이라고 하는가?

정답 혼동행렬
해설 혼동행렬에 대한 설명입니다.

[문207] 체 데이터에서 올바르게 분류한 데이터의 비율은?

정답 정확도
해설 정확도

[문208] Positive로 예측한 것 중에서 실제 값이 Positive인 비율은?

정답 정밀도
해설 정밀도

[문209] 실제 Positive인 값 중 Positive로 분류한 비율은?

정답 재현율, 민감도, 참 긍정률
해설 재현율에 대한 설명입니다.

[문210] 실제 Negative인 값 중 Negative로 분류한 비율은?

정답 특이도, 참 부정률
해설 특이도에 대한 설명입니다.

[문211]: 실제 Negative인 값 중 Positive로 잘못 분류한 비율은?

정답 거짓 긍정률
해설 거짓 긍정률에 대한 설명입니다.

[문212] 정밀도와 재현율의 조화평균으로, 정밀도와 재현율 중 한쪽만 클 때보다 두 값이 골고루 클 때 큰 값이 되는 지표는 무엇인가?

정답 F1-스코어
해설 F1-스코어에 대한 설명입니다.

[문213] 고정된 훈련 데이터 세트와 테스트 검증데이터 세트로 평가하여 반복적으로 튜닝할 시 테스트 데이터에 과적합되는 결과가 생기는 것을 방지하는 방법은?

정답 교차 검증
해설 교차 검증에 대한 설명입니다.

[문214] 가장 단순한 종류의 교차검증 방법으로 데이터를 랜덤으로 추출해 학습 데이터와 테스트 데이터로 나누는 방법으로서 모형 평가방법 중 주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차 검정을 실시하는 방법으로 하나는 모형의 학습 및 구축을 위한 훈련용 자료로, 다른 하나는 성과 평가를 위한 검증용 자료로 사용하는 방법은 무엇인가?

정답 홀드아웃
해설 홀드아웃에 대한 설명입니다.

[문215] 데이터 집합을 무작위로 동일 크기를 갖는 K개의 부분 집합으로 나누고, 그 중 1개의 집합을 테스트 데이터, 나머지 (K-1)개 집합을 학습 데이터로 선정하여 분석 모형을 평가하는 기법은?

정답 K-Fold 교차 검증
해설 K-fold 교차 검증은 다음 그림과 같이 K개의 집합으로 데이터를 나눕니다. 이후 1개의 집합을 테스트용 데이터로, 나머지 K-1개의 집합은 학습용 데이터로 사용합니다. 최종적으로 K번 반복된 결과를 바탕으로 최적의 모델을 찾는 데 사용합니다.

[문216] 전체 데이터에서 1개 샘플만을 테스트 데이터 집합에 사용하고 나머지 (N-1)개는 학습 데이터 집합에 사용하는 교차 검증 방법은?

정답 LOOCV
해설 LOOCV는 전체 데이터 중에서 1개는 테스트용 데이터로, 나머지 N-1개 데이터는 학습용 데이터로 사용합니다. 즉, 데이터 수(N개)만큼의 교차 검증을 수행합니다. 이는 K-fold 교차 검증에서 부분 집합에 데이터가 샘플 1개만 들어 있는 경우와 같습니다.

[문217] 예측값과 실제값의 차이인 오차의 제곱합으로 계산하며, 회귀 모형 평가 시에 사용하는 평가지표는?

정답 SSE
해설 SST = SSE + SSR
① SST : 실제값과 평균값의 차이 제곱
② SSE : 실제값과 예측값의 차이 제곱
④ SSR : 예측한 y값과 평균값의 차이 제곱

[문218] 회귀 모형이 실제값을 얼마나 잘 반영하는지를 나타내는 비율로, 1에 가까울수록 실제값을 잘 반영하는 것으로 판단하는 평가지표를 무엇이라고 하는가?

정답 R^2 또는 결정계수
해설 결정계수는 회귀 분석한 회귀 모형식이 얼마나 정확한지 나타내는 숫자입니다. 결정계수가 0에 가까우면 정확도가 매우 떨어지고, 1에 가까우면 실제값을 많이 반영함을 알 수 있습니다.

[문219] 가로축은 혼동행렬의 거짓 긍정률로, 세로축은 혼동행렬의 사실 긍정률로 두어 시각화한 그래프를 무엇이라고 하는가?

정답 ROC 커브
해설 ROC 커브는 새로 만들어진 모델의 성능을 측정하는 그래프입니다. 다음 그림과 같이 0.50, 0.75, 0.85, 0.95처럼 수치가 올라갈수록 모델 성능이 좋음을 의미하며, 해당 그래프의 면적으로도 모델 성능을 확인할 수 있습니다. 해당 면적은 ROC 커브 아래의 영역이므로, AUC라고 부릅니다.

[문220] 적중확률(Y축,True Positive Rate, Sensitivity) 대 오경보확률(X축, False Positive Rate, 1- Specificity)의 그래프 표현이며, 민감도와 특이도를 이용하여 분류 모델의 수준을 면적으로 표현하여, 모델 평가를 가시화한 도구를 무엇이라고 하는가?

정답 ROC 커브, ROC 그래프
해설 ROC 커브에 대한 설명입니다.

4.2 분석 결과 해석 및 활용

[문221] 데이터의 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정과 기법을 무엇이라고 하는가?

정답 데이터 시각화
해설 데이터 시각화에 대한 설명입니다.

[문222] 하나 이상의 변수에 대해서 변수 사이의 차이와 유사성 등을 표현하는 방법은 무엇이라고 하는가?

정답 비교 시각화
해설 비교 시각화에 대한 설명입니다.

[문223] 장소나 지역에 따른 데이터의 분포를 표현하는 것은?

정답 공간 시각화
해설 공간 시각화에 대한 설명입니다.

[문224] 다양한 데이터를 통합적으로 분석하여 기업 의사결정권자가 합리적인 의사결정이 가능하도록 지원하는 일련의 활동을 무엇이라고 하는가?

정답 BI(Business Intelligence)
해설 BI에 대한 설명입니다.

[문225] 초기 아이디어 개발 관점 분류 중에서 생각하는 것, 기억하는 내용을 지도를 그리듯이 마음속의 생각을 확장시키면서 줄거리를 이해하며 정리하는 방식을 무엇이라고 하는가?

정답 마인드맵
해설 마인드맵에 대한 설명입니다. 전체의 조각들을 한눈에 보기 쉽게 계층적 시각화로 표현하는 방법입니다.

[문226] 중요 정보를 하나의 그래픽으로 표현해서 보는 사람들이 쉽게 정보를 이해할 수 있도록 만드는 시각화 방법은?

정답 인포그래픽스
해설 인포그래픽스는 시각적인 이미지만으로 직관적으로 정보를 바로 이해할 수 있는 방법입니다. 다음 그림은 총 20명 중 15명에게 무언가의 의미가 있음을 알 수 있습니다.

[문227] 여러 가지 변수를 비교할 수 있는 시각화 그래프 중에서 칸 별로 색상을 구분하여 데이터 값을 표현한 비교 시각화 기법은?

정답 히트맵
해설 히트맵은 열을 의미하는 히트와 지도를 의미하는 맵이 결합한 용어로서, 색상을 통해 열 분포 형식으로 나타내고 이를 통해서 직관적으로 이해하는 시각화 기법입니다.