모집단을 어떤 특성에 따라 서로 겹치지 않는 여러 개의 층으로 분할한 후 각 층에서 표본 을 단순 무작위 추출법에 따라 추출하는 방법(계층 간 이질성, 계층 내 동질성)

-층화표본추출

전산시스템을 필요로 하는 곳으로부터 하청을 받아 시스템의 기획 개발, 유지보수, 운영 등 을 대신해주는 업종

-SI

인간의 간섭을 가능한 최소한으로 하여 금융 서비스나 투자 관리를 온라인으로 제공하는 투 자 자문역할의 일종이다. 수리적 규 칙이나 알고리즘에 기반한 디지털 금융 서비스를 제공한 다.

-로보 어드바이저

통계적인 가설검정에서 사용되는 기준값으로 α로 표시

-유의수준

구조화되어 있지 않은 문서를 단어로 나누는 과정

-토큰화

관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한뒤 적합도를 측정해 내는 분석 방 법이다. / 독립변수를 조 작하면서 종속변수가 어떻게 변하는지를 보며 두 변인의 관계를 파 악하는 것

-회귀분석

데이터에 개인을 식별할 수 있는 정보가 있는 경우 일부 또는 전체를 삭제하거나 일부를 대 체 처리함으로써 개인을 식별할 수 없게 하는 것

-데이터 비식별화

데이터 분석 시에 분석과제가 정해지고 분석 역량은 확보한 상태지만, 분석 기법이나 시스템을 신규 도입해야 하는 환경에서 수행하는 방법

-시스템 고도화

관심을 갖고 있는 모집단의 특성을 나타내는 대푯값

-모수

정밀도와 재현율의 조화평균으로, 정밀도와 재현율 중 한쪽만 클 때보다 두 값이 골고르 클 때 큰 값이 된다.

-F1스코어

가설 검정에서 사용된 샘플 데이터로부터 계산된 표본 통계량. 이것으로 P-value를 계산하 며 귀무가설을 기각할 것인지 판별

-검정통계량

사생활 침해를 방지하기 위해 데이터에 포함된 개인정보를 삭제하거나 알아볼 수 없는 형태 로 변환하는 방법

-익명화

관측 또는 실험으로 얻은 자료의 평균값으로 결측값을 대치해서, 불완전한 자료를 완전한 자료로 만드는 방법

-평균대치법

데이터를 통해 도출된 다양한 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험 을 결합해 내재화한 것

-지식

대용량의 정형 및 비정형 데이터를 저장하고 손쉽게 접근할 수 있게 하는 대규모 저장소

-데이터 호수

의사 결정나무 중 연속형 타깃변수(또는 목표변수)를 예측하는 의사결정나무를 무엇이라고 하 는가?

-회귀나무

분석 대상 데이터 집합에서 준식 별자 속성이 동일한 레코드가 적어도 K개 이상 존재하도록 제한하는 개인정보 보호 기법

-K-익명성

구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임 워크. 간단하게 설명하자면, 한명이 4주 작업할 일을 4명 이 나누어 1주일에 끝내는 것

-맵리듀스

전체 데이터 중 분석에 필요한 데이터만 선택적으로 이용하는 것

-샘플링

로지스틱 회귀분석에서 어떠한 일이 일어날 확률을 일어나지 않을 확률로 나누어 log를 취 하고 이를 0~1의 값이 아닌 (-무한대, 무한대) 범위에서 선형함수를 시그모이드 함수로 변환 하는 방법은 무엇인가?

-로짓변환

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

-정보통신망법

고정된 훈련 데이터 세트와 테스트 검증데이터 세트로 평가하여 반복적으로 튜닝할 시 테스트 데이터에 과적합되는 결과가 생기는 것을 방지하는 방 법

-교차검증

가설검정은 귀무가설과 대립가설 중에서 하나의 가설을 양자택일한다. 그래서 1-α는 귀무가 설을 채택시키므로, 1-α의 영역을 “(1)”이라고 부르고, 반대로 α는 귀무가설을 기각(탈락)시 키므로, α 의 영역을 “(2)”이라고 부른다.

-채택역,기각역

여러 개의 동일한 종류 또는 서로 다른 모형들의 예측/분류 결과를 종합하여 최종적인 의사결정에 활용하는 기법

-앙상블 기법

데이터상의 주석 작업으로 딥러닝과 같은 학습 알고 리즘이 무엇을 학습하여야 하는지 알려 주는 표식 작업

-어노테이션

전체 데이터에서 1개 샘플만을 테스트 데이터 집합에 사용하고 나머지 (N-1)개는 학습 데이터 집합에 사용하는 교차 검증 방법

-LOOCV

말뭉치에서 자주 등장하지만, 분석에 있어 기여하는 바가 없는 단어

-불용어

특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석

-감정분석

정규 분포의 평균을 측정할 때 주로 사용되는 분포로 모 집단의 분산(혹은 표준편차)이 알려 져 있지 않은 경우에 정규분포 대신 이용하는 확률분포는?

-T-분포

대규모 분산 시스템 모니터링 을 위해 에이전트와 컬렉터 구성을 통해 데이터를 수집하고 수 집된 데이터를 하둡 파일 시스템(HDFS)에 저장하는 기능을 제공하는 데이터 수집 기술

-척와

데이터 분석 조직의 유형으로, 분석 전담 조직이 우선순위에 따라 진행하며 일부 분석 업무가 중복되거나 이원화될 수 있는 조직 구조

-집중구조

비선형적인 관계도 파악할 수 있는 상관계수는?

-스피어만 상관계수

상향식 접근 방식의 발산단계와 하향식 접근 방식의 수렴단계를 반복하는 과제 발굴 방법

-디자인사고

비지도 학습의 연관규칙 분석 기법 중에서 발생 항목 집합에서 연관 관계를 찾아내는 것으로, 일명 장바구니 분석이라고도 불리는 알고리즘

-apriori알고리즘

MSE(오차의 제곱에 대한 평균)에 루트를 씌운 것

-RMSE

통계적 추정을 할 때 표본자료 중 모집단에 대한 정보를 주는 독립적인 자료의 수

-자유도

여러 사 람이 공유하여 사용할 목적으로 체계화해 통합, 관리하는 데이터의 집합 / 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합

-데이터베이스

전체 변동 중 회귀모형에 의해 설명되는 변동의 비율로, 표 본에 의해 추정된 회귀식이 주어 진 자료를 얼마나 잘 설명하는지를 보여주는 값 / 주어진 데이터에 회귀선이 얼마나 잘 맞 는지, 적합 정도를 평가하는 척도이자 독립변수들이 종속변수를 얼마나 잘 설명하는지 보여 주는 지표

-결정계수

다양한 데이터를 통합적으로 분석하여 기업 의사결정권자가 합리적인 의사결정이 가능하도 록 지원하는 일련의 활동

-BI

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

-요인분석

분산 환경에서 대량의 로그 데이터를 효과적으로 수집하여 합친 후 다른 곳으로 전송할 수 있는 신뢰할 수 IT 서비스는

-아파치플럼

시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법 / 분석목적에 따 라 특정 요인만 분리 분석하거나 제거하는 작업을 함

-분해시계열

자연어 분석 작업의 대상이 되는 대량의 텍스트 문서들을 모아놓은 집합

-말뭉치

연관 규칙에서 두 아이템의 연관 규칙이 우연인지 아닌지를 나타내는 척도

-향상도

기업의 목표를 달성하기 위 한 성과지표

-KPI

다층 신경망에서 은닉층이 많아 학습이 이루어지지 않는 문제는 무엇인가?

-기울기소실문제

빅데이터 저장 기술로 관계형 데이터베이스 관리 시스템으로 하나의 데이터베이스를 여러 개의 서버상에 구축하는 시스템은?

-데이터베이스클러스터

시계열 분석의 기본이 되는 중요한 개념으로 시계열의 평균과 분산이 일정하고 일정한 추세 가 없는 것을 무엇이라 하는가?

-정상시계열

장소나 지역에 따른 데이터의 분포를 표현하는 것

-공간시각화

평균을 0으로 표준편차를 1로 변환하는 스케일링법

-표준화

연관규칙 분석에서 품목간 상관관계를 기준으로 규칙의 예측력을 평가하는 지표는? / A→B 의 연관 규칙에서 임의로 B가 구매되는 경우에 비해 A와의 관계가 고려되어 구매되는 경우 의 비율이다.

-향상도

고객의 대규모 거래데이터로부터 함께 구매가 발생하는 규칙을 도출하여, 고객이 특정 상품 구매 시 이와 연관성 높은 상품을 추천하는 분석 / 어떤 변인 간에 주목할 만한 상관관계가 있는지를 찾아내는 방법

-연관분석

거리의 최대/최소값을 측정하는 군집 간 거리 측정 방법

-최장/최단연결법

실제 Positive인 값 중 Positive로 분류한 비율

-재현율

오차의 제곱에 대한 평균

-MSE

2개의 변수 간 상관 정도를 나타내는 지표로 상관관계의 경향성은 알 수 있으나, 선형성의 강도는 확인하지 못하는 해당 지표

-공분산

수집한 데이터를 요약, 묘사, 설명하는 통계 기법

-기술통계

평균으로부터 K\*표준편차만큼 떨어져 있는 값들을 이상값으로 판단하는 방법

-ESD

정규화 방법 중 원 데이터의 분포를 유지하면서 정규화하는 방법은? / 모든 데이터를 0과 1 사이의 값으로 변환하는 기법은?

-최소최대정규화

새로운 사건이 속할 범주를 찾아내는 일

-유형분석

코호넨 맵이라고도 불리며, 인공신경망을 기반으로 차원축소와 군집화를 동시에 수행할 수 있는 알고리즘은? / 코호넨에 의해 제시되었으며, 비지도 신경망으로 고차원의 데이터를 이 해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화하는 클러스터링 방법은 무 엇인가?

-자기조직화지도

데이터베이스의 테이블이 어떻게 구성되는지, 어떤 정보를 담고 있는지에 대한 기본적인 구 조를 정의하는 것

-스키마

주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률 분포는?

-포아송분포

전체 항목 중 A와 B가 동시에 포함되는 항목의 비율은?

-지지도

개체들의 유사성과 비유사성을 측정하여 N차원 공간에 점으로 표현하여 개체들 간의 집단화를 시각적으로 표현하는 분석 방법

-다차원척도법

일반 상용 서버로 구성된 클러스터에서 사용할 수 있는 분산 파일 시스템과, 대량의 자료를 처리하기 위한 분산 처리 시스템을 제공하는 오픈소스 프레임워크

-하둡

데이터베이스를 관리하며 응용 프로그램들이 데이터베이스를 공유하며 사용 할 수 있는 환경 을 제공하는 소프트웨어

-DBMS

시계열 모델 중에서 자신의 과거 값을 사용하여 설명하는 모형으로, 백색잡음의 현재 값과 자기 자신의 과거 값을 선형 가중합으로 수행하는 정상 확률 모형

-이동평균모형

반응변수가 범주형인 경우에 적용하는 회귀 모형으로, 설명변수(독립변수)의 값이 주어질 때 각 범주에 속할 추정확률을 기준치에 따라 분류하는 목적으로 사용하는 분석 기법

-로지스틱회귀분석

포본평균들의 표준편차

-표준오차

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

-데이터거버넌스

가장 단순한 종류의 교차검증 방법으로 데이터를 랜덤으로 추출해 학습 데이터와 테스트 데 이터로 나누는 것 / 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 검증 용 자료로 사용하는 방법은 무엇인가?

-홀드아웃

현 시점의 자료값에서 전 시점의 자료를 빼는 방법은?

-차분

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 6개의 절차로 이루어진 방법론

-CRISP DM

번호를 부여한 샘플을 나열하여 k개씩 n개의 구간을 나누고 첫 구간에서 하나 를 임의로 선 택한 후에 k개씩 띄어서 표본을 선택하고 매번 k번째 항목을 추출하는 표본 추출 방법

-계층추출법

실제값과 평균값 차이 제곱

-SST

혼동행렬을 활용한 평가지표 중에서 민감도와 동일한 계산식이며, 모형의 완전성을 평가하는 지표

-재현율

고객과의 관계를 지속적으로 강화하기 위한 정보시스템

-CRM

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

-요인분석

데이터를 이해하고 의미 있는 관계를 찾아내기 위해 데이터의 통곗값과 분포 등을 시각화하 고 분석하는 것

-EDA

기업이 시간과 비용을 최적화 시키기 위해 외부 공급업체와 연계하여 통합한 정보시스템

-SCM

중요 정보를 하나의 그래픽으로 표현하여 정보를 쉽게 이해할 수 있도록 만드는 시각화 기 법

-인포그래픽

임의의 개수로 소집단을 나누고 소집단의 중심으로부터 각 개체까지의 거리를 산출하고 나서, 각 개체를 가장 근접한 소집단에 배정한 이후, 해당 소집단의 중심좌표를 업데이트하는 군집화 방식

-K-평균 군집화

오피니언 리더, 즉 영향력 있는 사람을 찾아낼 수 있으며, 고객 간 소셜 관계를 파악

-소셜 네트워크 분석

지식을 도출할 때 사용하는 데이터 / 데이터의 가공 및 데이터 간 관계를 통해 패턴을 인식 하는 것

-정보

계층적 군집분석에서 군집 내의 오차제곱합에 기초하여 거리를 측정하는 방법은? / 오차제 곱합의 합에 비해 증가한 정도가 작아지는 방향으로 군집하는 방법은?

-와드연결법

가설 검정에서 사용된 샘플 데이터로부터 계산된 표본 통계량. 이것으로 P-value를 계산하 며 귀무가설을 기각할 것인지 판별

-검정통계량

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 알고리즘

-나이브베이즈분류

군집의 중심점을 기준으로 거리를 측정하는 군집 간 거리 측정 방법

-중심연결법

귀무가설이 옳다는 가정하에서 얻은 통계량이 귀무가설을 얼마나 지지하는지를 나타내는 확률

-유의확률

데이터에 관한 구조화된 데이터 / 어떤 목적을 가지고 만들어진 데이터 / 데이터 그 자체가 아니라, 자료의 속성, 구 조 등을 설명하는 데이터

-메타데이터

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

-클라우드컴퓨팅

분석용 데이터를 이용한 가설 설정을 통해 통계모델을 만들거나 기계학습을 이용한 모델을 만드는 과정

-모델링

우리는 모집단의 실제값과 회귀선과의 차이인 (1)을 알아낼 수 없기에 표본에서 나온 관측값 과 회귀선의 차이인 (2)를 이용해 분석을 수행한다.

-오차,잔차

작은 규모의 데이터 웨어하우스는? / 데이 터 웨어하우스에서 추출한 데이터를 특정 주제영 역으로 분석 후 그 결과를 조직이나 팀에서 활용하도록 제공한 데이터

-데이터 마트

데이터의 전부 또는 일부분을 대체값(공백, 노이즈 등)으로 변환하는 비식별화 방법 / 개인 의 사생활 침해를 방지하고 통계 응답자의 비밀사항은 보호하면서 통계자료의 유용성을 최 대한 확보할 수 있는 데이터변환 방법은?

-데이터마스킹

회귀모형의 계수를 추정하는 방 법으로써 잔차제곱합을 최소화하는 계수를 찾는 방법은?

-최소제곱법

큰 용량과 복잡성으로 기존 애플리케이션이나 툴로는 다루기 어려운 데이터셋의 집합

-빅데이터

주어진 자료에서 단순복원 임의추출 방법을 활용하여 동일한 크기의 표본을 여러개 생성하는 샘플링 방법

-부트스트랩

Positive로 예측한 것 중에서 실제 값이 Positive인 비율

-정밀도

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

-다중대체법

데이터가 비정상성이 아닌 증거를 나타내는 경우에 적용되며, 초기 차분 단계(모델의 "통합 된" 부분에 해당)를 한 번 이상 적용하여 비정상성을 제거할 수 있다. / 분기 , 반기, 연간 단위로 다 음 지표를 예측하거나 주간, 월간 단위로 지표를 리뷰하여 경향을 분석하는 기법

-자기회귀이동평균모형

문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법

-워드클라우드

이것은 비즈니스 측면에서 일반적으로 ‘공동 활용의 목적으로 구축된 유무형의 구조물’을 의 미한다. 수집된 데이터를 가공, 처리, 저장해두고 이 데이터에 접근할 수 있도록 API를 공개 한다. 그러면 다양한 서드파티 사업자들이 비즈니스에 필요한 정보를 추출해 활용하게 되고 빅데이터는 그 자체로 이 역할을 수행하게 된다.

-플랫폼

XOR 과 같이 선형 분리할 수 없는 문제에 약점을 지닌, 입력층과 출력층으로만 구성된 최초의 인공신경망

-퍼셉트론

관계형 데이터베이스를 SQL을 사용해 CRUD(Create, Read, Update, Delete)를 수행하고 관리할 수 있는 소프트웨어

-RDBMS

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

-상관계수

데이터에서 잡음을 제거하기 위해 추세에서 벗어나는 값들을 변환하고, 거칠게 분포된 데이터를 부드럽게 만들기 위한 데이터 변환 기술

-평활화

귀무가설 분포에서 검정통계량보다 극단적인 값이 관측될 확률값. 이 값이 작을수록 검정통 계량이 귀무가설의 내용에 적합하지 않음을 나타낸다.

-유의확률

배깅과 부스팅보다 더 많은 무작위성을 주어 약한 학습기를 생성한 후 이를 선형 결합하여 최종학습기를 만드는 방법

-랜덤포레스트

무응답을 현재 진행 중인 연구에서 비슷한 성향을 가진 응답자의 자료로 대치하는 방법

-핫덱대체법

분석 역량이 확보되지 않고 기존 시스템을 활용하여 분석을 수행하는 방법

-역량확보

분석 역량이 확보되지 않고 시스템을 신규도입해야 하는 분석 수행 방법

-전문업체활용

Fayyad 가 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 단계(5단계 절차)로 수행하는 방법론

-KDD

말이나 글로 표현할 수 있고 전달과 공유가 가능한 형상화된 지식

-형식지

특정 정보를 해당 그룹의 대푯값 또는 구간값으로 변환하는 비식 별화 방법

-범주화

이것은 인공지능의 한 분야로 간주된다. 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술 을 개발하는 분야 이다. / 학습 데이터로 학습한 알려진 특성을 활용해 ‘예측’하는 방법

-머신러닝

연속적인 시계열 데이터를 분석할 수 있는 신경망으로, 경사 하강법과 시간기반 오차역전파를 사용해서 가중치를 업데이트하며, 은닉층에서 재귀적인 신경망을 갖는 알고리즘

-RNN

의사결정 나무에서 더 이상 분기되지 않도록 하는 규칙은?

-정지규칙

신경망 학습에서 평균 제곱 오차 또는 교차 엔트로피 오차를 사용하여 현재의 상태를 나타내는 지표

-손실함수

예측한 y 값과 평균값의 차이 제곱

-SSR

평균에서 표준편차의 3배만큼 떨어진 값을 이상치로 판단하여 데이터의 이상값을 검출하는 방법

-ESD

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

-특이값분해

누구나 열람할 수 있는 디지털 장부에 거래 내역을 투명하게 기록하고, 여러 대의 컴퓨터에 이를 복제해 저장하는 분산형 데이터 저장기술

-블록체인

전체 변동 중 회귀모형에 의해 설명되는 변동의 비율로, 표 본에 의해 추정된 회귀식이 주어 진 자료를 얼마나 잘 설명하는지를 보여주는 값 / 주어진 데이터에 회귀선이 얼마나 잘 맞 는지, 적합 정도를 평가하는 척도이자 독립변수들이 종속변수를 얼마나 잘 설명하는지 보여 주는 지표

-결정계수

총 평균과 각 집단의 평균 차이에 의해 생긴 집단 간 분산 비율을 나타내는 분포

-F-분포