**문항1. 데이터 분석과 빅데이터 분석의 차이점에 4가지에 대해서 설명하시오.**

데이터분석유형

1. 기술적분석

 -점수 카드, 대시보드에 기초하여 과거에 무슨 일이 발생했는지를 밝혀줌

-그 결과를 어떻게 해석할 것인가? 해석 결과에 따라 어떻게 행동할지 결정하게 됨

 -이때 행동의 결정을 판단하는 것은 인간이 수행함

1. 진단적분석
* 과거에 축적된 행동 데이터를 토대로 인과관계를 찾아냄
* 그 결과를 어떻게 해석할 것인가? 해석 결과에 따라 어떻게 행동할지 결정하게 됨
* 이때 행동의 결정을 판단하는 것은 인간이 수행함
1. 예측분석
* 통계학적 모델을 사용해 미래에 무슨 일이 어느 정도의 확률로 일어날지 예측함
* 기존의 데이터베이스에서 과거의 데이터를 참조하는 데이터 마이닝의 확장 기능에 해당함
1. 처방적분석
* 차세대분석법 이라고도 불림
* 예측되는 사태를 위해서 무엇을 하면 좋을지 처방함
* 앞선 3가지 분석의 조합이 이루어짐 (1. 과거에 일어난 것을 해석하는 기술적 분석 2. 그것이 왜 일어났는지를 밝히는 진단적 분석 3. 그리고 앞으로 어떤 일이 일어날지를 추측하는 예측분석)
* 그 다음에 취해야 할 최선의 행동은 인간 대신 컴퓨터가 고찰함
* 이러한 처방적 분석에서는 주로 인공지능이 이용되는 경우가 많음 ( 분석-가설-실시-학습을 반복함으로써 제안하는 최적 행동에 대한 정밀도가 올라가게 됨)
* 빅데이터 분석유형

1.데이터분석가

-비즈니스에 대한 이해가 필요하며, 문제를 올바르게 정의하고 분석 과정의 오류를 수정할 수 있어야 함

2. 분석 기반 기술

- 오픈소스 소프트웨어의 구조를 이해하고 오픈소스를 사용하여 프로그래밍할 수 있는 능력이 필요함

3. 알고리즘 분석과 분석 모델 이해

- 다양한 알고리즘을 구성하고, 모델링 기법을 선택 및 개발하여 분석과정에서 나타날 수 있는 파라미터 및 분석 모델을 최적화할 수 있어야 함

4. 데이터 통찰력

- 수집된 데이터를 정제하고 분석을 위한 데이터로 변환, 삭제,업데이트 할 수 있어야 함

- 어떤 질문을 던져서 어떠한 과정으로, 어떻게 데이터를 이용하고, 어떤 결과물을 만들어 내는지는 데이터를 넓고,깊고,다양하게 보는 통찰력에 달려 있음

데이터 유형은 정형화의 정도에 따라 3가지로 구분한다.
1. 정형 데이터는 고정된 필드에 저장된 데이터이다. 즉 데이터베이스의 정해진 규칙에 맞게 데이터 중에서 수치만으로 의미 파악이 쉬운 데이터를 의미한다. 예를 들면 표를 그려 놓고 채워 넣는 형식의 데이터로 이름, 나이, 주민등록번호, 카드번호 등 주로 숫자와 짧은 단어로 구성된 것이다.
2. 반정형 데이터는 고정된 필드에 저장되어 있지는 않지만 메타데이터나 스키마를 포함하는 데이터를 의미한다. 즉, 고정된 양식은 없으나 어느 정도 구조가 정해져 있는 데이터라고 할 수 있다.
3. 비정형 데이터는 고정된 필드에 저장되어 있지 않은 데이터를 의미한다. 정해진 규칙이 없어서 값의 의미를 쉽게 파악하기 어렵다. 빅데이터의 약 85% 가량은 형태가 비정형 데이터에 해당하며 향후에 더욱 증가할 것이다.

데이터 분석의 유형은 사람의 개입 정도에 따라 4가지로 구분한다.
1. 기술적 분석이다. 기술적 분석이란 점수 카드나 대시보드에 기초하여 이뤄지고 과거에 무슨 일이 발생했는지를 밝혀주는 것이다. 그 후 분석 결과를 어떻게 해석하고 어떻게 행동으로 옮길지 판단하는 것은 인간이 수행한다.
2. 진단적 분석이다. 진단적 분석이란 과거에 축적된 행동 데이터를 토대로 인과관계를 찾아내는 것으로서 왜 그 일이 일어났는지를 밝히는 것이다. 이 분석 결과도 기술적 분석과 마찬가지로 어떻게 해석하고 어떻게 행동으로 옮길지 판단하는 것은 인간이 수행한다.
3. 예측 분석이다. 예측 분석은 통계학적 모델을 사용해 미래에 무슨 일이 어느 정도의 확률로 일어날지를 예측하는 것이다. 기존의 데이터베이스에서 과거의 데이터를 참조하는 데이터 마이닝의 확장 기능이라고 할 수 있다.
4. 처방적 분석이다. 처방적 분석은 차세대 분석법이라고 일컬어진다. 즉, 예측되는 사태를 위해서 무엇을 하면 좋을지 처방하는 것이다. 처방적 분석은 앞서 소개한, 과거에 일어난 것을 해석하는 기술적 분석, 그것이 왜 일어났는지를 밝히는 진단적 분석, 그리고 앞으로 어떤 일이 일어날지를 추측하는 예측 분석을 조합하여 이루어진다.

## 문항2. 데이터 전처리 기술 3가지가 무엇인지 명칭을 작성하고, 그 개념에 대해 설명하시오.(필요시, 사용되는 세부적인 기술을 설명하시오.)

1. 데이터 필터링
필터링이란 오류의 발견 후 보정과 삭제 또는 중복성 검사 등을 수행하는 것이다. 실제 사전 테스트에서 필터링 과정을 수행하게 되고, 필터링 기준을 최적화하여 활용하는 것이 일반적이다. 데이터 필터링 작업을 수행할 때에는 사전 정의된 기준에 의거하여야 하며, 오류에 대한 이력을 저장해야 한다.
2. 데이터 변환
데이터 변환은 데이터의 특정 변수를 정해진 규칙에 따라 변환하는 것이다. 데이터 변환 기술은 평활화, 집계, 정규화, 일반화, 속성 생성이다. 첫번째, 평활화란 데이터로부터 잡음제거를 위해 데이터 추세에 벗어나는 값들을 변환하는 기법이다. 두번째, 집계란 다양한 차원의 방법으로 데이터를 요약하는 기법이다. 복수의 속성을 하나로 줄이거나 유사한 데이터 객체를 줄이고 스케일을 변경하는 기법을 사용한다. 세번째, 정규화란 데이터를 정해진 구간 내에 들도록 하는 기법이다. 정규화를 할 때 사용하는 통계방식은 최소-최대 정규화, Z-스코어 정규화, 소수 스케일링이다. 네번째, 일반화란 특정 구간에 분포하는 값으로 스케일을 변화시키는 기법이다. 일반화가 잘 되어있다면 이상값이나 노이즈가 들어오더라도 크게 흔들리지 않는다.
다섯 번째, 속성 생성이란 데이터 통합을 위해 새로운 속성이나 특징을 만드는 기법이다. 이는 주어진 여러 데이터 분포를 대표할 수 있는 새로운 속성이나 특징으로 대체하여 데이터를 변경 처리하는 것이다.
3. 데이터 정제
데이터 정제는 수집된 데이터의 불일치성을 교정하기 위한 것이며, 데이터 정제 기술에는 결측치의 처리와 잡음 처리가 있다. 결측치를 처리하는 방법은 해당 레코드 무시하기, 자동채우기, 담당자 수작업하기가 있다. 잡음을 처리하는 방법은 구간화, 회기값 적용, 군집화이다.