보 고 서

제목 : 통계기반 데이터 분석 중간고사

|  |  |
| --- | --- |
| **과 목 명** | **통계기반 데이터 분석** |
| **학 과** | **데이터 사이언스** |
| **학 번** | **A60042** |
| **이 름** | **안성민** |
| **제 출 일** | **2020년 5월 3일 (일)** |
| **담당교수** | **정화민 교수님** |

# 

# 

# 과제내용 1) NCS 통계기반 데이터 분석 코딩 1번

**1. 데이터 분석 예시에 대한 p-값 산출**

**(1) 예시 데이터 불러오기**

예시 데이터가 “diet.csv” 파일형식으로 **[그림 1-1]**과 같이 저장되어 있다 고 가정

|  |  |
| --- | --- |
|  | **[그림 1-1]**  **diet.csv 파일 저장내용** |

예시 데이터 확인

|  |
| --- |
|  |
| **[소스파일 Ex00.R]** |

**(2) 검정 통계량 분포를 확인하여 p-값을 계산**- 두 집단이 차의 평균이 0인지 아닌지를 검정, T 검정 통계량을 바탕으로 쌍체 t검정을 수행하여 p-값을 계산한다.

|  |  |
| --- | --- |
|  | |
| **[소스파일 Ex00.R]** | |
|  | **[그림 1-2]**  **다이어트 전후 차이에 대한 box plot 함수 수행 결과** |

**T 통계량 계산**

|  |
| --- |
|  |
| **[소스파일 Ex00.R]** |

**양측 검증 및 단측 검증을 수행하여 가설 채택을 한다.**

|  |
| --- |
|  |
| **[소스파일 Ex00.R]** |

**귀무가설을 채택 혹은 기각한다.**

**유의수준과 p-값을 활용하여 귀무가설 채택 여부를 결정하게 된다.**

1. **데이터 분석 예시에 대한 귀무가설 채택 여부 결정  
   양측검정, 단측검정 모두 p-값이 유의수준보다 작으므로, 유의수준 5% 기준으로 귀무가설을 기각하고 대립가설을 채택하게된다.**

# 과제내용 2) NCS 통계기반 데이터 분석 코딩 2번

1. **기술통계 분석**

**(1)데이터 내용 및 구조파악**- 데이터와 관련된 기본적인 내용들 확인

|  |
| --- |
|  |
| **[소스파일 Ex01.R]** |

**(2)요약데이터 확인**

|  |
| --- |
|  |
| **[소스파일 Ex01.R]** |

**(3)그래프를 통한 데이터의 이해  
분석하고자 하는 데이터를 그래프 등을 통한 시각화 기법을 활용하여 데이터에 대한 이해도를 높임**

|  |  |
| --- | --- |
|  | |
| **[소스파일 Ex01.R]** | |
|  | **좌🡪우,위🡪아래 순서**  **히스토그램 그래프**  **Boxplot 그래프**  **두 변수간 산점도 그래프** |
|  | **시계열 데이터** |

# 과제내용 3) NCS 통계기반 데이터 분석 코딩 3번

1. **상관분석**

R의 corr 함수를 통해 상관계수를 파악하고, 그래프를 통해 독립변수 간의 상관관계 분석을 수행

|  |  |
| --- | --- |
|  | |
| **[소스파일 Ex02.R]** | |
|  | EuStockMarkets 데이터의 변수들 간의 상관계수들에 대한 상관계수 행렬 플롯 (corrplot() 함수 수행 결과) |

# 과제내용 4 NCS 통계기반 데이터 분석 코딩 4번

1. **주성분 분석(PCA)**

대표적인 훈련 데이터인 iris를 이용하여 주성분 분석을 수행

|  |  |
| --- | --- |
|  | |
| **[소스파일 Ex03.R]** | |
|  | iris 데이터 변수들을 통해 도출된 주성분 1과 2 (PC1 및PC2)를 이용한 산점도 (biplot() 함수 수행 결과) |

# 과제내용 5) NCS 통계기반 데이터 분석 코딩 5번

1. **분류 : 로지스틱 회귀분석**

독립 변수들의 선형 결합을 통해 사건의 발생 여부등을 분류하기 위한 목적으로 사용되는 통계기법,  
선형 회귀분석과는 다르게 종속변수가 범주형인 데이터인 경우에 사용된다.

|  |  |
| --- | --- |
|  | |
| **[소스파일 Ex04.R]** | |
|  | summary 함수를 통해 모델의 결과를 확인  결과에서 p-value가 0.05보다 작은  변수는 통계적으로 유의미하다고 해석할 수 있는 변수 |

# 과제내용 6) NCS 통계기반 데이터 분석 코딩 6번

1. **예측 : ARIMA 모델**
2. 데이터의 확인 및 이해

|  |  |
| --- | --- |
|  | |
| **[소스파일 Ex05.R]** | |
|  | AirPassengers 시계열 데이터에 대해 시간에 따른 항공기 탑승승객 수에 대한 플롯 (plot() 함수 수행 결과) |
|  | 계절성(seasonality), 추세 (trend), 불확실성 (random) 요소로 분해해서 그래프를 확인 |
|  | |
| **[소스파일 Ex05.R]** | |
|  | 시계열 데이터를 diff()함수와 log()함수 등을 활용하여 안정적인 시계열 데이터로 변환 |
|  | |

1. 예측 모델 생성

ARIMA 모델은 AR 모델과 MA 모델을 결합한 것으로 세 가지 파라미터 p, d, q가 필요하다. 파라미터를 구하기 위해서는 KPASS test 4, ACF, PACF를 실제 그려서 적절한파라미터를 찾아야 하지만, 이와 같은 작업을 자동으로 해주는 auto.arima 함수를 사용하면 p, d, q를 자동으로 구해낼 수 있다

|  |  |
| --- | --- |
|  | |
| **[소스파일 Ex05.R]** | |
|  | AirPassengers 데이터를 바탕으로 생성한 ARIMA모델 기반 10년 예측 데이터 플롯 |

# 과제내용 7) NCS 통계기반 데이터 분석 코딩 7번

1. **군집화 : K 평균 군집화**

(1) 군집 수 K 결정

- iris 데이터를 이용하여 K 평균 군집화 모델을 구축  
군집화를 수행하기 전에 먼저 군집 수 K를 결정해야 한다. K-평균 군집화를 수행하기위해서는 먼저 K값을 결정하고 각 군집에 대한 초기 군집 중심을 결정해야 한다. 초기군집중심에 가까운 개체들이 해당 군집으로 소속되고 이후 소속된 개체들을 반영하여군집 중심 값이 갱신된다. 이와 같은 과정을 반복하여 최종적으로 K 평균 군집화가 이루어진다. 군집 수 결정은 실루엣값을 이용하는데 실루엣 폭(Silhouette width) 값이 가장 큰 값을 찾아 군집 수로 결정하는 방법이 보편적이다.

|  |  |
| --- | --- |
|  | |
| **[소스파일 Ex06.R]** | |
|  | iris 데이터를 K=2인 군집화 수행 후 실루엣 계  산 결과에 대한 플롯  Average silhouette width 값을 확인할 수 있  다. iris 데이터에서 **군집을 2**로 했을 때는 실루엣값이 **0.68**이다. |
|  | |
| **[소스파일 Ex06.R]** | |
|  | iris 데이터를 K=3인 군집화 수행 후 실루엣 계  산 결과에 대한 플롯  Average silhouette width 값은 0.57로, k=2 일때보다 낮은 값을 가짐을 확인할 수 있다. |

(2)K 평균 군집화 모델 생성

|  |
| --- |
|  |
| **[소스파일 Ex06.R]** |
| **iris 데이터는 setosa, versicolor, virginica 세 종류의iris 꽃을 분류한 데이터로 실제 군집 수를 3으로 가정하고 K 평균 군집화를 수행한 경우가 더욱 정확도가 높다.** |
| **[수행결과]** |

# 과제내용 8) NCS 통계기반 데이터 분석 코딩 8번

1. **파생변수를 활용한 분석모델 확장**

|  |
| --- |
|  |
| **[소스파일 Ex07.R]** |
| 기존 데이터 세트에 age20s, ages30s, ages40s, age50s과 같은 파생변수가 생성된 것을 확인할 수 있다. |
| **[수행결과]** |

# 과제내용 9) NCS 통계기반 데이터 분석 코딩 9번

1. **앙상블 기법들을 활용하여 분석모델을 확장**

**(1)배깅 (Bagging)**

- 배깅은 예측 모델의 변동성이 큰 경우 예측모델의 변동성감소를 위해 사용

Bootstraping aggregating : 주어진 데이터에 대해서 여러 개의 부트스트랩 자료를 생성하고 각 부트스트립 자료를 모델링 한 후 결합하여 최종 예측 모델을 산출하는 방법.

Bootstrap : 단순 복원 임의 추출법을 통해 원 데이터로부터 크기가 동일한 여러 개의 표본 데이터 세트

|  |
| --- |
|  |
| **[소스파일 Ex08.R]** |
| 4개의 변수중, Petal.Length 의 중요도가 가장 높게 평가되었으며, 수치는 81.10632% 임을 확인,  새로운 Session 으로 시행시, 중요도 항목은 수치상 차이가 있다. |
| **[소스파일 Ex08.R]** |
| 의사결정트리의 그래프화 |
| **[소스파일 Ex08.R]** |
|  |
| **[수행결과]** |
| 정오 그래프화 |
| **[소스파일 Ex08.R]** |
|  |
| **[수행결과]** |

**(2) 부스팅(Boosting)** - 잘못분류된 개체들에 관심을 가지고 이들을 더 잘 분류하기 위해서 잘못 분류된 개체들에 집중하여 새로운 분류 규칙을 만드는 단계를 반복하는 방법.  
 - 부스팅은 예측모델의 정확도를 향상하는 방법의 하나이다.  
 - 부스팅은  
예측변수를 차례로 생성하여, 처음 원 데이터의 객체들은 동일한 가중치에서 시작하지만, 모델링을 통한 예측변수에 의해 오 분류된 개체들에는 높은 가중치를 부여하고, 정 분류된 객체들에는 낮은 가중치를 부여하여 오 분류된 객체들이 더 잘 분류되도록 한다.

|  |
| --- |
| 의사결정트리의 그래프화 |
| **[소스파일 Ex08\_1.R]** |
|  |
| **[수행결과]** |
| 정오분류표 그래프화 |
| **[소스파일 Ex08\_1.R]** |
|  |
| **[수행결과]** |

**(3)랜덤포레스트(Random forest)**여러 개의 의사결정트리를 임의로 학습하는 앙상블 방법.  
의사결정트리를 만들 때 데이터를 복원추출로 샘플링하여 해당 데이터에 대한 모델을 만드는 방법과  
노드내에 데이터를 분기할 때 사용되는 기준을 정할 때 전체 변수를 대상으로 분기하는 것이 아니라 일부변수에 대해서 분기를 시도하여 여러 의사결정트리를 생성한다. 여러 개의 의사결정트리 모델이 만들어지면 이들의 결과들을 종합하여 다수결(majority voting) 방식 등을 이용하여 최종값을 산출한다.

|  |
| --- |
|  |
| **[소스파일 Ex09.R]** |
|  |
| **[소스파일 Ex09.R]** |
|  |
| **[수행결과]** |

# 과제내용 10) NCS 통계기반 데이터 분석 코딩 10번

1. **예측 오차를 통한 예측모델 성능 평가**

임의의 시계열 데이터가 주어졌을 때 훈련 데이터 세트와 검증 데이터 세트로 구분하고, 훈련 데이터 세트를 활용하여 ARIMA 모델을 생성, 이후 검증 데이터 세트를 이용하여 MAE, MSE, MAPE 등과 같은 예측오차를 통해 ARIMA 모델의 성능을 평가

|  |
| --- |
|  |
| **[소스파일 Ex10.R]** |
|  |
| **[수행결과]** |
|  |
| **[소스파일 Ex10.R]** |
|  |
| **[수행결과]** |
|  |
| **[소스파일 Ex10.R]** |
|  |
| **[소스파일 Ex10.R]** |
|  |
| **[수행결과]** |

# 과제내용 11) NCS 통계기반 데이터 분석 코딩 11번

1. **교차 유효성 검사를 통한 예측모델 성능 평가**

|  |
| --- |
|  |
| **[소스파일 Ex11.R]** |

# 과제내용 12) NCS 통계기반 데이터 분석 코딩 12번

1. ROC 곡선 기법을 통한 분류모델 성능 평가

|  |
| --- |
|  |
| **[소스파일 Ex12.R]** |
|  |
| **[수행결과]** |

# 과제내용 13) NCS 통계기반 데이터 분석 코딩 13번

1. **내부 평가를 이용한 군집모델 성능 평가**

**(1)데이터 준비 및 스케일링**iris 데이터를 활용하여 군집모델 생성을 위해 군집에 적합한 데이터를 준비한다.

|  |
| --- |
|  |
| **[소스파일 Ex13.R]** |

**(2) 군집모델 생성**단계(1)에서 iris 데이터를 훈련 데이터 세트와 검증데이터 세트로 나누었는데, 그 중 훈련 데이터 세트를 이용하여 3개의 군집으로 군집화 하는 군집모델을 생성

|  |
| --- |
|  |
| **[소스파일 Ex13.R]** |

**(3)검증 데이터 세트를 통한 모델 정확성 확인**

|  |
| --- |
|  |
| **[소스파일 Ex13.R]** |

**(4)내부평가 방법을 이용한 모델성능 평가**단계(3)에서 생성된 검증 데이터 세트의 군집결과를 바탕으로 dunn 지수값을 확인

|  |
| --- |
|  |
| **[소스파일 Ex13.R]** |