2024년 제3회 청소년 IT경시대회 기출문제

# 데이터분석 부문

시행 | 2024년 9월 28일 (토)

주최 | 한국정보기술진흥원　　후원 | 서울특별시의회

< 수험자 유의사항 >

1. 본 진흥원이 주최하는 경시대회의 시험문제는 저작권법에 따라 보호되는 저작물이며, 시험문제의 일부 또는 전부를 무단 복제, 배포, (전자)출판 하는 등 저작권을 침해하는 경우 저작권법에 의하여 민·형사상 불이익을 받을 수 있습니다.
2. 신분증을 지참하지 않은 자는 시험에 응시할 수 없습니다.
 - 인정 신분증의 범위: 주민등록증, 여권, 청소년증
3. 시험 중에는 어떠한 통신기기 및 전자기기(휴대전화, 스마트폰, 태블릿PC, 스마트워치, 이어폰, 전자사전 등)도 소지 및 사용할 수 없으며, 적발 시에는 부정행위로 처리됩니다.
4. 시험 중에는 퇴실 할 수 없으며, 응시자 이외에는 시험장에 출입하실 수 없습니다.
5. 부정한 방법으로 시험에 응시하거나 시험에서 부정행위를 한 자에 대해서는 해당 회차의 시험을 정지시키거나 합격을 무효로 하며, 이후 3년간 본 진흥원에서 주최하는 시험에 응시하실 수 없습니다.

1. 다음 코드에서 발생하는 오류의 원인은 무엇인가?

```
import numpy as np
a = np.array([1, 2, '3'])
b = a + 1
```

① 문자열과 정수를 더할 수 있다.
② 배열에 정수를 더할 수 있다.
③ numpy는 정수를 지원하지 않는다.
④ numpy는 문자열을 지원하지 않는다.
⑤ 배열 크기가 맞지 않는다.

2. 다음 프로그램의 실행 결과는 무엇인가?

```
import numpy as np

arr = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
result = arr.sum(axis=0)
print(result)
```

① [1 2 3]　　② [5 7 9]
③ [6 15 24]　　④ [7 8 9]
⑤ [12 15 18]

3. 다음 중 numpy에서 브로드캐스팅이 발생하지 않는 코드는 무엇인가?

```
a = np.array([[1, 2], [3, 4]])
b = np.array([1, 2])
c = np.array([1, 2, 3])
```

① a + b　　② a * b
③ a + 3　　④ a + c
⑤ b * 2

4. 다음 중 데이터를 분석하는 방법 중 하향식 방법에 대한 설명으로 옳은 것을 고르시오.

① 데이터에 기반하여 문제를 탐색한다.
② 미시적 관점의 과제가 주어진다.
③ 빠르게 변화하는 환경에서 다양한 시도를 한다.
④ 전통적 과제 발굴 방식으로 빠르게 적용한다.
⑤ 데이터 기반의 의사결정체제가 있으면 다양한 시도가 가능하다.

5. 다음 코드에서 apply 함수의 목적은 무엇인가?

```
import pandas as pd
df = pd.DataFrame({'A': [1, 2, 3], 'B': [4, 5, 6]})
df['C'] = df['A'].apply(lambda x: x ** 2)
print(df)
```

① 모든 행을 제거한다.
② A열의 각 값을 제곱하여 C열에 저장한다.
③ B열의 값을 제거한다.
④ 새로운 열 D를 추가한다.
⑤ 데이터프레임을 정렬한다.

6. 다음 코드를 실행시켰을 때, 그래프의 모양에 대한 설명으로 옳은 것을 고르시오.

```
import matplotlib.pyplot as plt
import numpy as np

x = np.linspace(0, 10, 100)
y = np.sin(x)
plt.plot(x,    y,    linestyle='--',    color='r',
marker='o')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.title('Sine Wave')
plt.grid(True)
plt.show()
```

① 선 그래프이며, 빨간색 점선으로 그려진다
② 선 그래프이며, 파란색 실선으로 그려진다
③ 산점도 그래프이며, 빨간색 실선으로 그려진다
④ 산점도 그래프이며, 파란색 점선으로 그려진다
⑤ 그래프가 그려지지 않는다

7. 다음 코드의 실행 결과를 고르시오.

```
import pandas as pd

data  =  {'Name':  ['Alice',  'Bob',  'Charlie',
'David'],
       'Age': [25, 30, 35, 40],
       'Score': [85, 90, 88, 75]}

df = pd.DataFrame(data)

df['Score'] += 5
df['Age'] *= 2

result = df[df['Age'] > 60]['Score'].mean()

print('%.1f' % result)
```

① 85.5　② 86.0
③ 86.5　④ 87.0
⑤ 87.5

8. 데이터 분석에서 귀무 가설에 대한 설명으로 가장 올바른 것을 고르시오.

① 두 변수 간의 관련성이 없다는 가설
② 모집단에 대한 완전한 정보를 가지고 있다는 가설
③ 데이터 집합의 평균값이 0이라는 가설
④ 표본의 크기가 작아서 유의미한 결과를 도출할 수 없다는 가설
⑤ 특정 사건이 발생할 확률이 0이라는 가설

9. 다음 코드의 실행 결과를 작성하시오. (힌트: 실행 결과의 형태는 "High 정수" 이며, 이 중 정수 부분만 정답으로 작성하세요.)

```
import numpy as np
import pandas as pd

data = {'A': [1, 2, 3, 4, 5],
        'B': [6, 7, 8, 9, 10]}

df = pd.DataFrame(data)

df['C'] = np.where(df['A'] > 3, 'High', 'Low')
df['D'] = df['A'] * df['B']

result = df[df['D'] > 30]['C'].value_counts()

print(result)
```

10. 아래 지문을 읽고 빈 칸에 들어갈 단어를 순서대로 <보기>에서 골라 작성하세요. (부분 점수 있음)

<지문>
모델링이란 분석용 데이터를 이용한 가설 설정을 통하여 통계모형을 만들거나 (1)을 이용한 데이터의 분류, 예측, 군집 등의 기능을 수행하는 모형을 만드는 과정입니다. (1)은 지도학습(Supervised Learning)과 비지도학습(Un-supervised Learning) 등으로 나뉘어 다양한 알고리즘을 적용할 수 있습니다.

데이터 정규화에서 데이터를 분할 할 때에는 아래와 같이 분할할 수 있습니다.

· (2) : 모델을 생성하기 위한 데이터
· (3) : 최소 예측 오차를 갖는 모델을 결정하기 위한 데이터
· (4) : 생성한 모델을 평가하기 위한 데이터 (최근 데이터)

모든 데이터를 사용하여 모델을 생성하는 경우, 사용된 데이터에 과적합 되는 문제가 발생하기 때문에 향후 모델을 사용하는 경우, 예측력이 매우 떨어지는 현상이 발생할 수 있습니다. 그렇기 때문에 모델의 안정성과 예측력의 비교를 위해서 데이터를 (2) / (3)로 구분하여 사용하며, (4)로 평가합니다.

데이터 세트에서 (5)를 이용하여 다양한 알고리즘을 거쳐 모델링을 진행하게 됩니다. 모델링의 결과 중 가장 우수한 알고리즘을 선정하고, 일부 변수를 제외한 최적의 모델 선정 과정을 거치게 됩니다. 데이터에 대해 모델을 학습한다는 것은 데이터에 기반해 최적화된 모델 파라미터를 알아내는 것을 말합니다.

<보기>
검증데이터 과적합 기계학습 데이터마이닝 설명변수 수치형파라미터 예측정확도 적합도 테스트데이터 테스트베드 품질관리 학습데이터

2024년 제3회 청소년 IT경시대회 기출문제

데이터분석 부문 정답

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| **1** | **5** | **4** | **4** | **2** |
| 6 | 7 | 8 | 9 | |
| **1** | **3** | **1** | **2** | |
| 10 | | | | |
| **기계학습, 학습데이터, 검증데이터<br>테스트데이터, 설명변수** | | | | |