

17

## 21회 기출 분석 (1)

### [수업 목표]

이번 시간에는 21회 기출 문제를 풀어보겠습니다

### [수업 개요]

# 21회 기출 중 전처리에 대해서 풀어보겠습니다.

0:00

**17.1 21회 기출문제 풀이 (1)**

**17.1.1 머신 러닝 (50점)**

학생 성적에 관한 데이터셋은 제공 394행짜리 소규모 데이터.

1-1. 시각화 포함 탐색적 자료분석(EDA)  
 1-2. 결측치 식별하고 결측치를 예측하는 두 가지 방법 정도를 쓰고, 선택한 이유를 설명.  
 1-3. 범주형 변수 인코딩이 필요한 경우를 식별하고, 변환을 적용하시오. 선택한 이유를 설명.

```
In [1]: import pandas as pd
import numpy as np
df = pd.read_csv("./student_data.csv")
df
```

```
Out[1]:
```

	school	sex	paid	activities	famrel	freetime	goout	Dalc	Walc	health	absences	grade
0	GP	F	no	no	4.0	3.0	4.0	1.0	1.0	3.0	6.0	5
1	GP	F	no	no	5.0	3.0	3.0	1.0	1.0	3.0	4.0	5
2	GP	F	yes	no	4.0	3.0	2.0	2.0	3.0	3.0	10.0	7
3	GP	F	yes	yes	3.0	2.0	2.0	1.0	1.0	5.0	2.0	15
4	GP	F	yes	no	4.0	3.0	2.0	1.0	2.0	5.0	4.0	6
...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	yes	no	5.0	5.0	4.0	4.0	5.0	4.0	11.0	9
391	MS	M	no	no	2.0	4.0	5.0	3.0	4.0	2.0	3.0	14
392	MS	M	no	no	5.0	5.0	3.0	3.0	3.0	3.0	3.0	10
393	MS	M	no	no	4.0	4.0	1.0	3.0	4.0	5.0	0.0	11
394	MS	M	no	no	3.0	2.0	3.0	3.0	3.0	5.0	5.0	8

### [EDA]

# profiling 에서 탐색적 자료분석 인사이트를 얻고 시각화를 해주시면 됩니다.

02:25

jupyter 17.1.21회기출분석(1) (autosaved)

File Edit View Insert Cell Kernel Navigate Widgets Help

Trusted Python 3 (ipykernel)

1-1. 시각화 포함 탐색적 자료분석(EDA)  
 1-2. 결측치 식별하고 결측치를 예측하는 두 가지 방법 정도를 쓰고, 선택한 이유를 설명.  
 1-3. 범주형 변수 인코딩이 필요한 경우를 식별하고, 변환을 적용하시오. 선택한 이유를 설명.

```

In [1]: import pandas as pd
import numpy as np
df = pd.read_csv("../student_data.csv")
df

```

Out [1]:

	school	sex	paid	activities	famrel	freetime	good	Delc	Walc	health	absences	grade
0	GP	F	no	no	4.0	3.0	4.0	1.0	1.0	3.0	6.0	5
1	GP	F	no	no	5.0	3.0	3.0	1.0	1.0	3.0	4.0	5
2	GP	F	yes	no	4.0	3.0	2.0	2.0	3.0	3.0	10.0	7
3	GP	F	yes	yes	3.0	2.0	2.0	1.0	1.0	5.0	2.0	15
4	GP	F	yes	no	4.0	3.0	2.0	1.0	2.0	5.0	4.0	6
...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	yes	no	5.0	5.0	4.0	4.0	5.0	4.0	11.0	9
391	MS	M	no	no	2.0	4.0	5.0	3.0	4.0	2.0	3.0	14
392	MS	M	no	no	5.0	5.0	3.0	3.0	3.0	3.0	3.0	10
393	MS	M	no	no	4.0	4.0	1.0	3.0	4.0	5.0	0.0	11
394	MS	M	no	no	3.0	2.0	3.0	3.0	3.0	5.0	5.0	8

17.1.2 1-1. 시각화 포함 탐색적 자료분석(EDA)

```

In [2]: df.info()

```

<class 'pandas.core.frame.DataFrame'>  
 RangeIndex: 395 entries, 0 to 394  
 Data columns (total 12 columns):  
 # Column Non-Null Count Dtype

[결측치 대체 및 범주형 변수 변환]

# KNN을 이용한 결측치 대체법은 꼭 기억해주시기 바랍니다.

08:46

jupyter 17.1.21회기출분석(1) (autosaved)

File Edit View Insert Cell Kernel Navigate Widgets Help

Trusted Python 3 (ipykernel)

1. 결측치 존재 여부  
 2. 데이터 타입 설명  
 3. 종속변수 분포 설명  
 4. 독립변수 상관관계 설명  
 5. 종속변수와 독립변수의 상관관계 설명  
 6. 유의할 점 (pandas\_profiling의 warning 2 설명)

17.1.3 1-2. 결측치 식별하고 결측치를 예측하는 두 가지 방법 정도를 쓰고, 선택한 이유를 설명. **1**

17.1.4 1-3. 범주형 변수 인코딩이 필요한 경우를 식별하고, 변환을 적용하시오. 선택한 이유를 설명.

1-2 답

1) 단순 대체법 : 수치형 변수라면, 각 행의 평균이나 중앙값을 사용하여 결측치를 보간할 수 있으며, 명목형, 범주형 변수라면 최빈값을 사용하여 대체할 수 있습니다.

2) KNN을 이용한 결측치 대체 : 보간법 중 결측치가 없는 행들의 최근접 이웃 데이터를 통해 결측치가 있는 변수 대체를 할 수 있습니다.

[다음 수업 예고]

다음 시간에는 모델링 및 모델 평가 기출 문제를 풀어보겠습니다