Machine Learning with Python

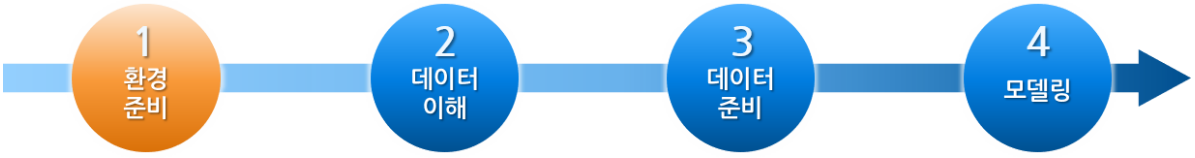
Life is too short, You need Python

실습 내용

- 머신러닝 모델링을 위한 코딩은 무조건 할 수 있어야 합니다.
- 코딩 내용을 자세히 알지 못해도 무작정 코딩을 진행해봅니다.
- Admission 데이터를 대상으로 모델링 해서 대학원 입학 여부를 예측해 봅니다.
- kNN 알고리즘을 사용합니다.

1.환경 준비

- 기본 **라이브러리**와 대상 **데이터**를 가져와 이후 과정을 준비합니다.



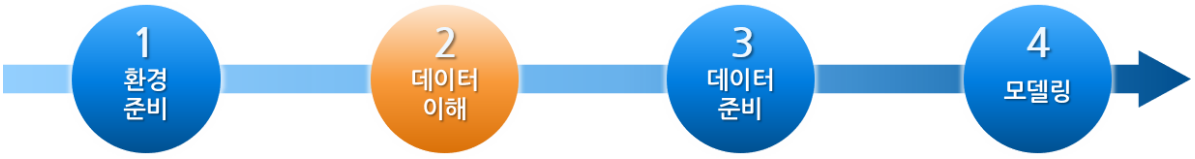
In [1]:
```
# 라이브러리 불러오기
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
```

```
warnings.filterwarnings(action='ignore')
%config InlineBackend.figure_format = 'retina'
```

In [2]:
```
# 데이터 읽어오기
path = 'https://raw.githubusercontent.com/Jangrae/csv/master/admission_simple.csv'
data = pd.read_csv(path)
```

2.데이터 이해

- 분석할 데이터를 **충분히 이해**할 수 있도록 다양한 **탐색** 과정을 수행합니다.



In [3]:
```
# 상/하위 몇 개 행 확인
data.head()
```

Out[3]:

	GRE	TOEFL	RANK	SOP	LOR	GPA	RESEARCH	ADMIT
0	337	118	4	4.5	4.5	9.65	1	1
1	324	107	4	4.0	4.5	8.87	1	1
2	316	104	3	3.0	3.5	8.00	1	0
3	322	110	3	3.5	2.5	8.67	1	1
4	314	103	2	2.0	3.0	8.21	0	0

데이터 설명

- GRE: GRE Scores (out of 340)
- TOEFL: TOEFL Scores (out of 120)
- RANK: University Rating (out of 5)
- SOP: Statement of Purpose Strength (out of 5)
- LOR: Letter of Recommendation Strength (out of 5)
- GPA: Undergraduate GPA (out of 10)
- RESEARCH: Research Experience (either 0 or 1)
- ADMIT: Chance of Admit (either 0 or 1)

In [4]:
```
# 하위 몇 개 행 확인
data.tail()
```

Out[4]:

	GRE	TOEFL	RANK	SOP	LOR	GPA	RESEARCH	ADMIT
495	332	108	5	4.5	4.0	9.02	1	1
496	337	117	5	5.0	5.0	9.87	1	1
497	330	120	5	4.5	5.0	9.56	1	1
498	312	103	4	4.0	5.0	8.43	0	0
499	327	113	4	4.5	4.5	9.04	0	1

In [5]:
```
# 변수 확인
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   GRE       500 non-null    int64  
 1   TOEFL     500 non-null    int64  
 2   RANK      500 non-null    int64  
 3   SOP       500 non-null    float64
 4   LOR       500 non-null    float64
 5   GPA       500 non-null    float64
 6   RESEARCH  500 non-null    int64  
 7   ADMIT     500 non-null    int64  
dtypes: float64(3), int64(5)
memory usage: 31.4 KB
```

In [6]:
```
# 기술통계 확인
data.describe().T
```

Out[6]:

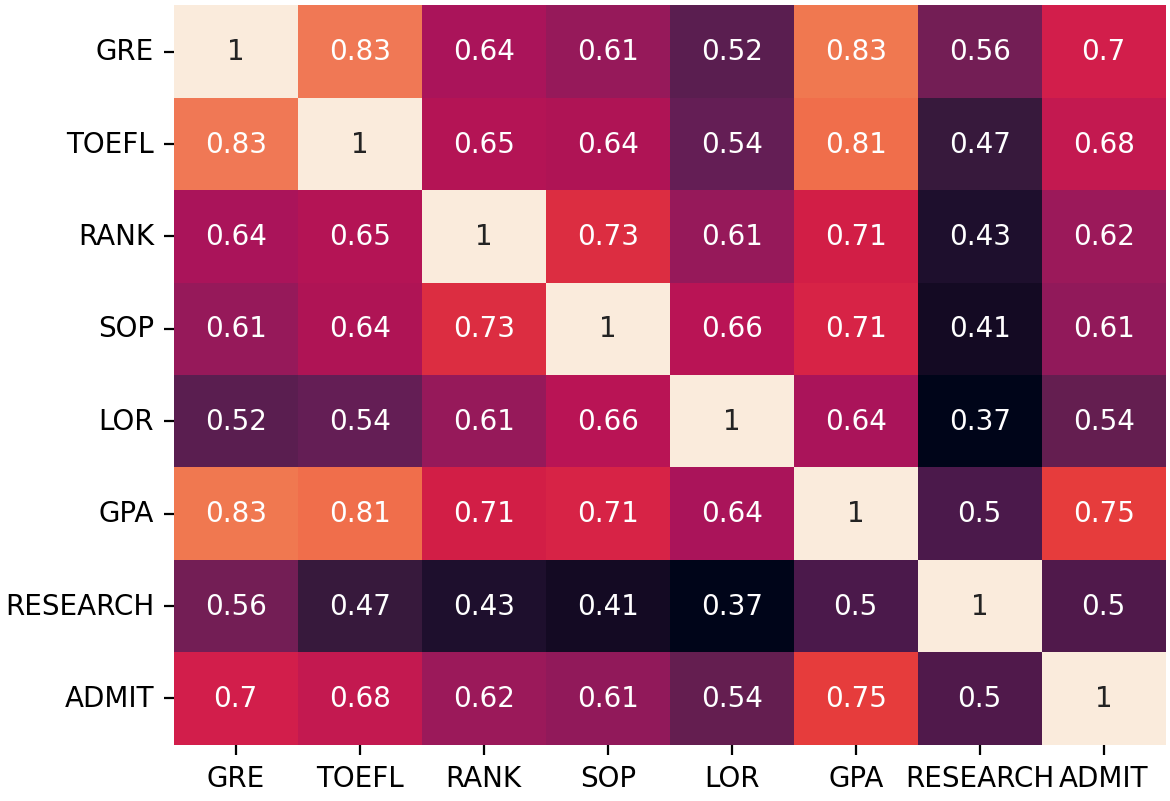
	count	mean	std	min	25%	50%	75%	max
GRE	500.0	316.47200	11.295148	290.0	308.0000	317.00	325.00	340.00
TOEFL	500.0	107.19200	6.081868	92.0	103.0000	107.00	112.00	120.00
RANK	500.0	3.11400	1.143512	1.0	2.0000	3.00	4.00	5.00
SOP	500.0	3.37400	0.991004	1.0	2.5000	3.50	4.00	5.00
LOR	500.0	3.48400	0.925450	1.0	3.0000	3.50	4.00	5.00
GPA	500.0	8.57644	0.604813	6.8	8.1275	8.56	9.04	9.92
RESEARCH	500.0	0.56000	0.496884	0.0	0.0000	1.00	1.00	1.00
ADMIT	500.0	0.43600	0.496384	0.0	0.0000	0.00	1.00	1.00

In [7]:
```
# 상관관계 확인
data.corr()
```

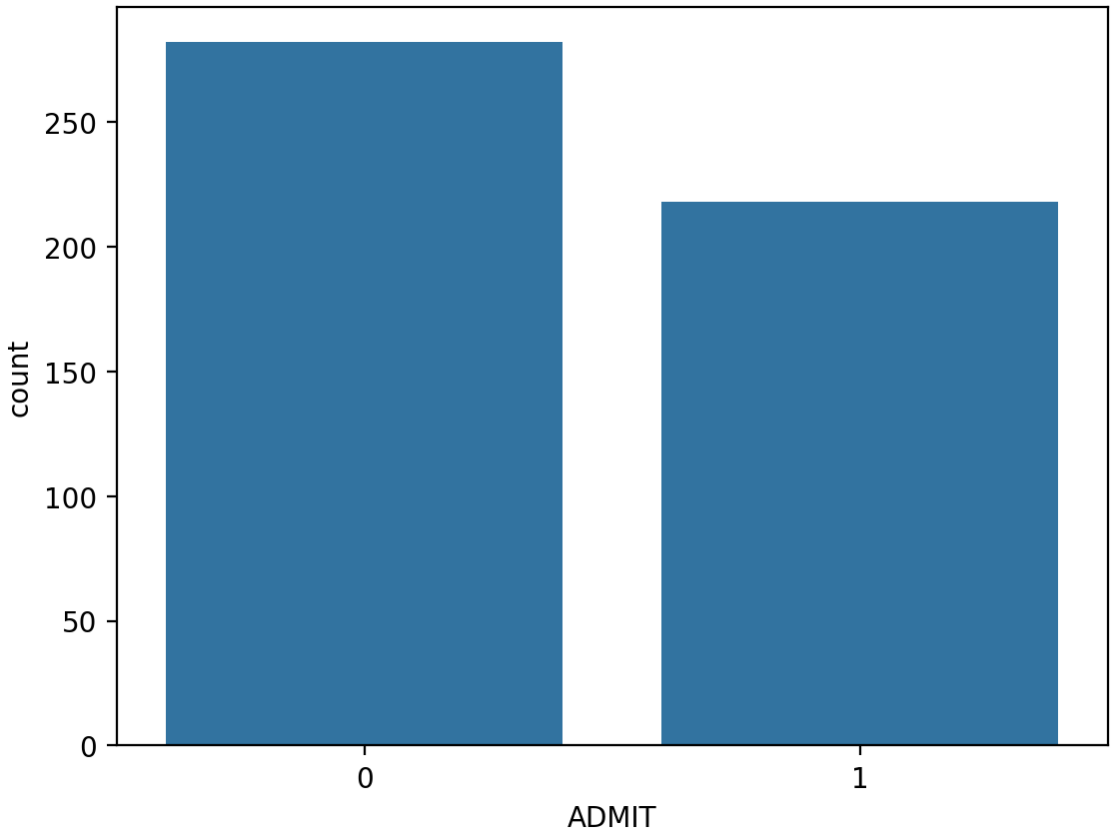
Out[7]:

	GRE	TOEFL	RANK	SOP	LOR	GPA	RESEARCH	ADMIT
GRE	1.000000	0.827200	0.635376	0.613498	0.524679	0.825878	0.563398	0.701671
TOEFL	0.827200	1.000000	0.649799	0.644410	0.541563	0.810574	0.467012	0.680503
RANK	0.635376	0.649799	1.000000	0.728024	0.608651	0.705254	0.427047	0.618367
SOP	0.613498	0.644410	0.728024	1.000000	0.663707	0.712154	0.408116	0.606876
LOR	0.524679	0.541563	0.608651	0.663707	1.000000	0.637469	0.372526	0.536527
GPA	0.825878	0.810574	0.705254	0.712154	0.637469	1.000000	0.501311	0.752196
RESEARCH	0.563398	0.467012	0.427047	0.408116	0.372526	0.501311	1.000000	0.503104
ADMIT	0.701671	0.680503	0.618367	0.606876	0.536527	0.752196	0.503104	1.000000

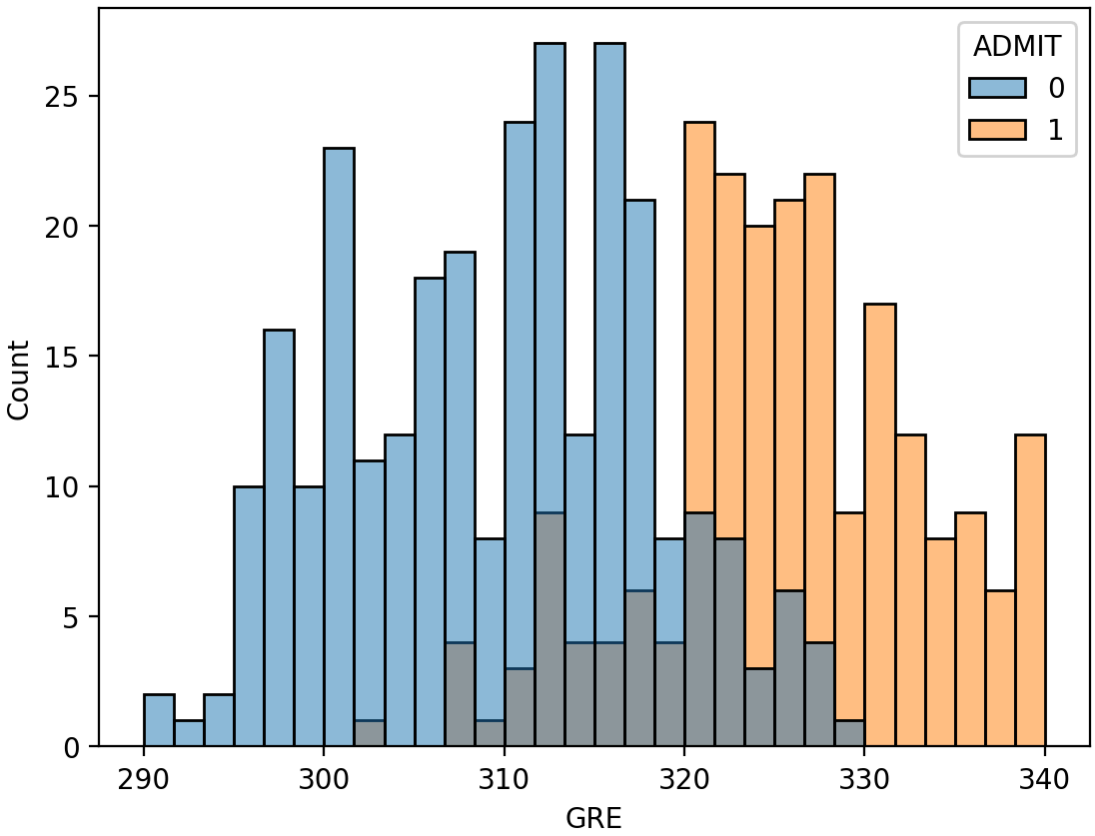
In [11]:
```
# 상관관계 시각화
sns.heatmap(data.corr(), annot=True, cbar=False)
plt.show()
```



In [54]:
```
sns.countplot(x=data['ADMIT'])
plt.show()
```

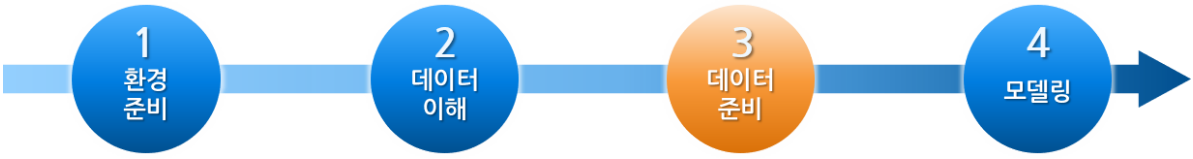


In [58]:
```
sns.histplot(x=data['GRE'], hue=data['ADMIT'],bins=30)
plt.show()
```



3.데이터 준비

- **전처리** 과정을 통해 머신러닝 알고리즘에 사용할 수 있는 형태의 데이터를 준비합니다.



1) x, y 분리

- 우선 target 변수를 명확히 지정합니다.
- target을 제외한 나머지 변수들 데이터는 x로 선언합니다.
- target 변수 데이터는 y로 선언합니다.
- 이 결과로 만들어진 x는 데이터프레임, y는 시리즈가 됩니다.
- 이후 모든 작업은 x, y를 대상으로 진행합니다.

In [12]:
```
# target 확인
target = 'ADMIT'

# 데이터 분리
x = data.drop(target, axis=1)
y = data.loc[:, target]
```

2) 학습용, 평가용 데이터 분리

- 학습용, 평가용 데이터를 적절한 비율로 분리합니다.
- 반복 실행 시 동일한 결과를 얻기 위해 random_state 옵션을 지정합니다.

In [74]:
```
# 모듈 불러오기
from sklearn.model_selection import train_test_split

# 7:3으로 분리
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=1, strat
```

In [75]:
```
# 데이터 분포 확인
print(y_train.value_counts(normalize=True))
print(y_test.value_counts(normalize=True))
```

```
ADMIT
0    0.562857
1    0.437143
Name: proportion, dtype: float64
ADMIT
0    0.566667
1    0.433333
Name: proportion, dtype: float64
```

4.모델링

- 본격적으로 모델을 **선언**하고 **학습**하고 **평가**하는 과정을 진행합니다.
- 우선 **회귀** 문제인지 **분류** 문제인지 명확히 구분합니다.

- 회귀 문제 인가요? 분류 문제인가요?
- 회귀인지 분류인지에 따라 사용할 알고리즘과 평가 방법이 달라집니다.
- 우선 다음 알고리즘과 평가 방법을 사용합니다.
  - 알고리즘: KNeighborsClassifier
  - 평가방법: accuracy_score

In [76]:
```
# 1단계: 불러오기
from sklearn.neighbors import KNeighborsClassifier #최근접
from sklearn.metrics import accuracy_score
```

In [77]:
```
# 2단계: 선언하기
model = KNeighborsClassifier()
```

In [78]:
```
# 3단계: 학습하기
model.fit(x_train, y_train)
```

Out[78]:
```
KNeighborsClassifier()
```

In [79]:
```
# 4단계: 예측하기
y_pred = model.predict(x_test)
```

In [80]:
```
print(y_test.values[:20])
print(y_pred[:20])
```

```
[1 0 0 1 0 1 1 1 1 1 1 0 1 0 1 0 0 1 0 0]
[1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 0 0 1 0 0]
```

In [95]:
```
# 5단계: 평가하기
print('accuracy_score', accuracy_score(y_test, y_pred)) # 86%
```

```
accuracy_score 0.8466666666666667
```

In [96]:
```
y_base = np.array([y_train.mode()[0]] * len(y_test))
```

In [97]:
```
print('accuracy_score', accuracy_score(y_test, y_base)) # base 56%
```

```
accuracy_score 0.5666666666666667
```

In [ ]:
```

```