혁신성장 청년인재 집중양성
이미지 분석 인공지능 서비스 개발 실무 과정

Semi-Project for Part 4
Training & testing traditional ML algorithms (Titanic survival analysis)

Daeyeon Jo
repositivator@gmail.com

1. Blank notebook for this semi-project

ML for Titanic survival prediction (Blank)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [1]:

```
import warnings
warnings.filterwarnings("ignore")

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# from sklearn import ?
# from sklearn.metrics import ?
```

Data info

- **PassengerId** : Unique ID of passenger
- **Survived** : 0 = No, 1 = Yes
- **pclass** : Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **sibsp** : # of siblings & spouses aboard the Titanic
- **parch** : # of parents / children aboard the Titanic
- **ticket** : Ticket number
- **cabin** : Cabin number
- **embarked** : Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

1. Preparing dataset (2번부터 실습

In [12]:

```
data_df = pd.read_csv('titanic.csv')
data_df.head(3)
```

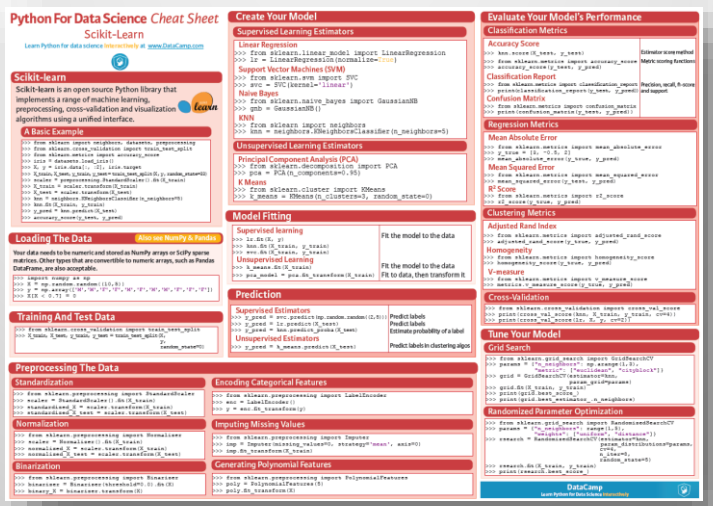
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S

2. Possible pathways for data preprocessing

+ Check & adjust data for handling Missing data & Outlier
+ Select important columns (or just use all columns & improve your model later)
+ Change characters to numbers (Binary num, Class num, One-hot vector, etc.)
+ (If applicable & useful) Select features with Tree-based models
+ (If applicable & useful) Modify the scale of features with StandardScaler / MinMaxScaler
+ (If applicable & useful) Reduce dimension with PCA

서울시 범죄현황 통계자료 분석 및 시각화

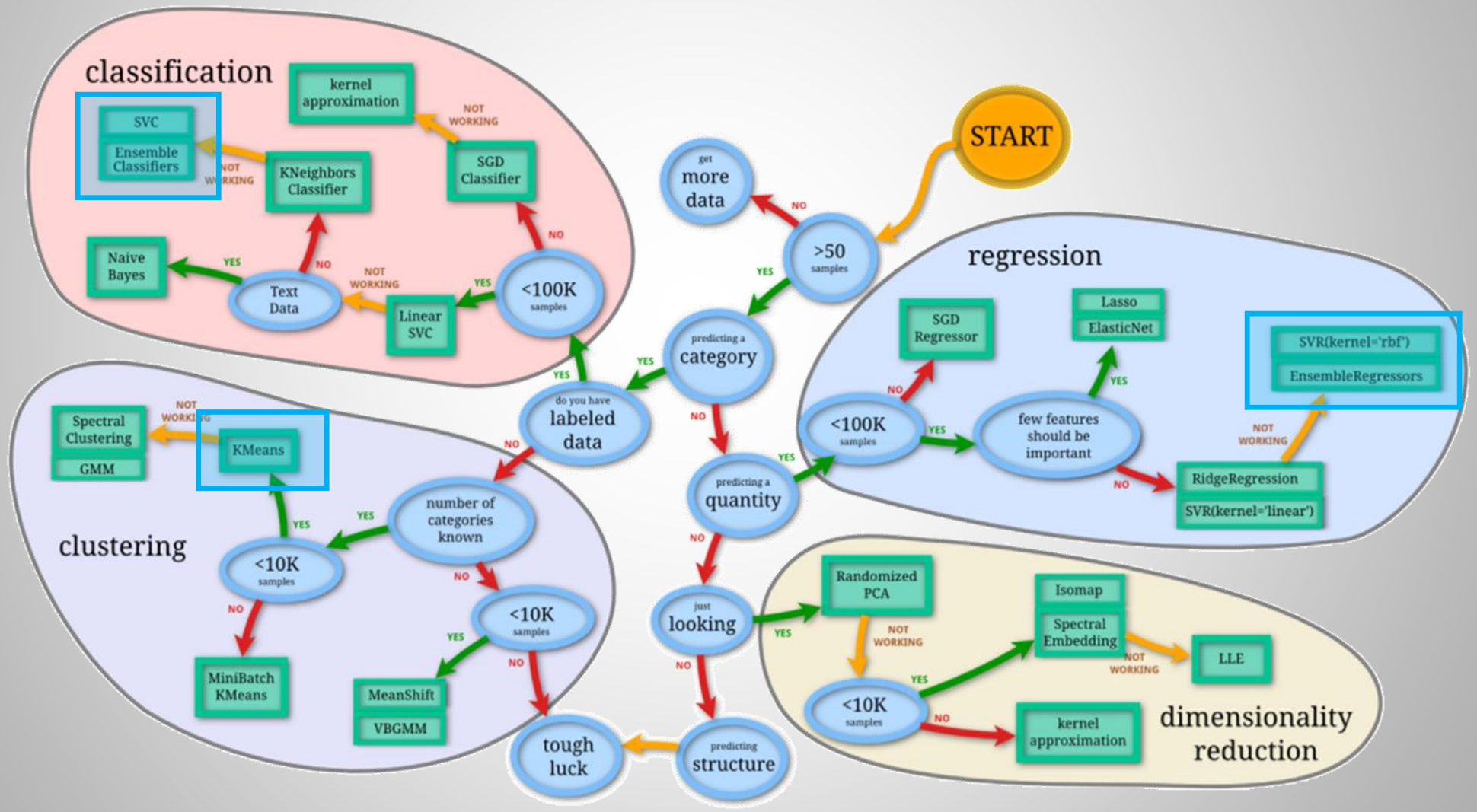
Scikit-learn practices & CheatSheet



파이썬을 활용한 기초 통계분석

풀어내려는 문제의 종류와 데이터의 타입(형태, 수)에 따른 ML 알고리즘 선택 가이드

http://scikit-learn.org/stable/tutorial/machine_learning_map/ (각 알고리즘 별 예시 코드 有)



수업 관련 공지사항

5팀 5/4팀 4팀
1팀 1/2팀 2팀 3팀

1팀 : 김은아, 문석호, 신광현, 장재원, 한상국 + 1人

2팀 : 김현하, 배소현, 윤영준, 이경희, 이재승, 하수진

3팀 : 김충희, 송윤성, 이가은, 이예랑, 이철희, 정수현

4팀 : 김채윤, 문희원, 이동규, 이주환, 이중기, 장준규

5팀 : 김지승, 이소은, 장희은, 한성준, 이대광 + 1人

* 세미프로젝트 시작 후 합류한 분들 중 파이썬이 익숙하지 않은 분들은 기존 Part 1~4의 내용들을 차례대로 실습해보고 팀 내에서 데이터에 모델을 적용하는 과정을 넓게 이해하는 것에 포커스를 맞춰주세요.

수업 관련 공지사항

* Part 1/2/4 에서 **배운 지식들을 최대한 빠짐없이 활용**하는데 초점을 맞춰주세요.

* **팀별로 자유로이 3층 내에서 자리를 옮겨서 논의하셔도 무방합니다.**

* **월 ~ 화 : 팀별 분석 작업 –〉 화요일 16:30 : 팀별 발표 및 질의응답 (15분 내외/팀)**
  : **화요일 16:25 까지** Jupyter notebook 제출 : repositivator@gmail.com

* **데이터 전처리 방법 / Model 선택 / Metric 선택 모두 자유입니다.** (배운 내용의 복습에 Focus!)

* **발표 시 포함할 사항 : 데이터 전처리 방법 & 이유 / 모델 적용 프로세스 / 모델 적용 결과**
  발표 시 제출할 사항 : 주석이 포함된 전체 코드 (.ipynb 제출, PPT 발표자료 필수 X)

* **발표 시작 시간은 변동될 수 있습니다**

혁신성장 청년인재 집중양성
이미지 분석 인공지능 서비스 개발 실무 과정

End of Document