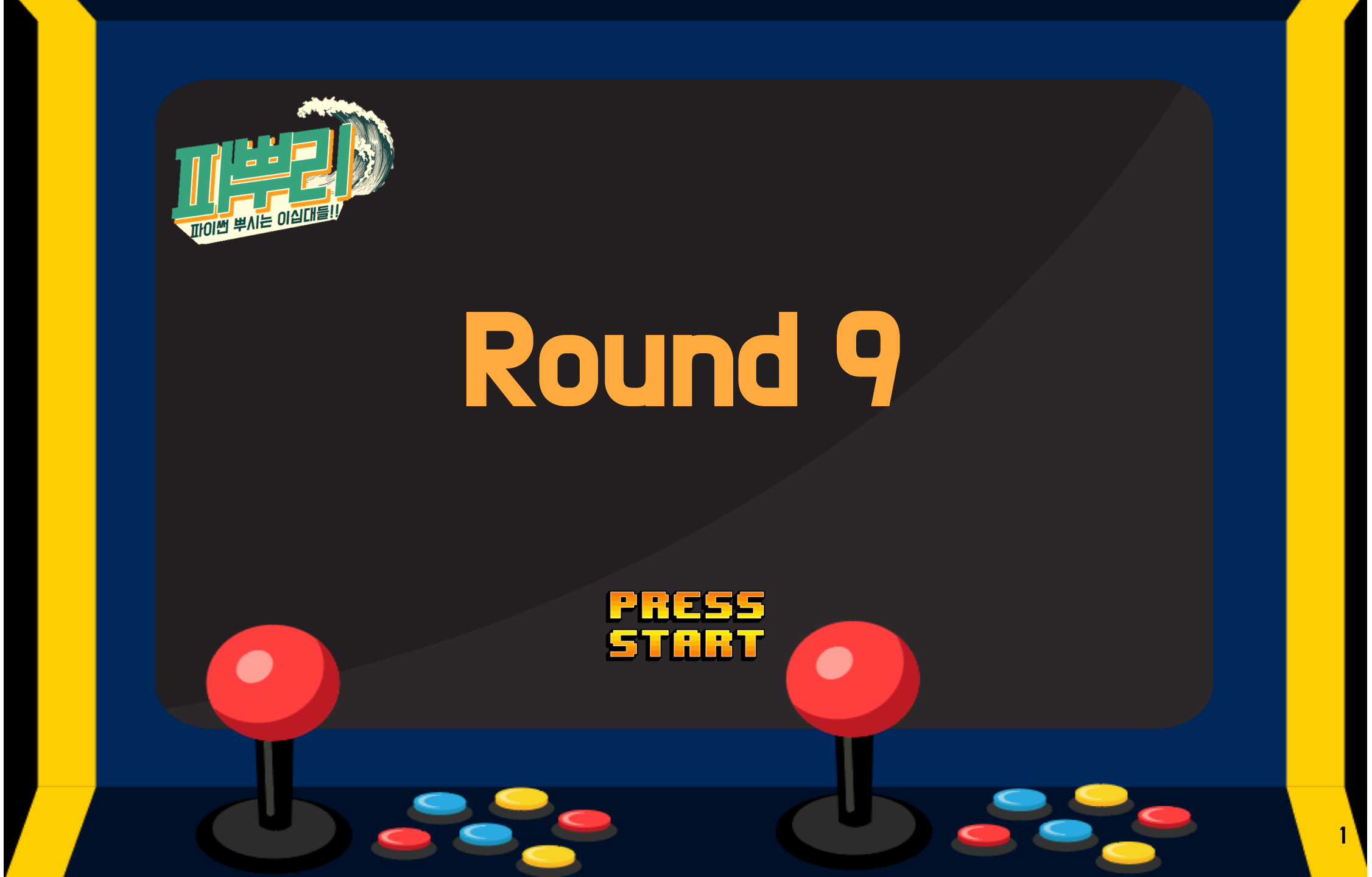

파뿌리
파이썬 뿌시는 이십대들!!
Round 9
PRESS
START

파뿌리
파이썬 뿌시는 이십대들!!

《 Round 9 》

- 분류분석 개요
- 분류분석 실습
- 텀 프로젝트 개요

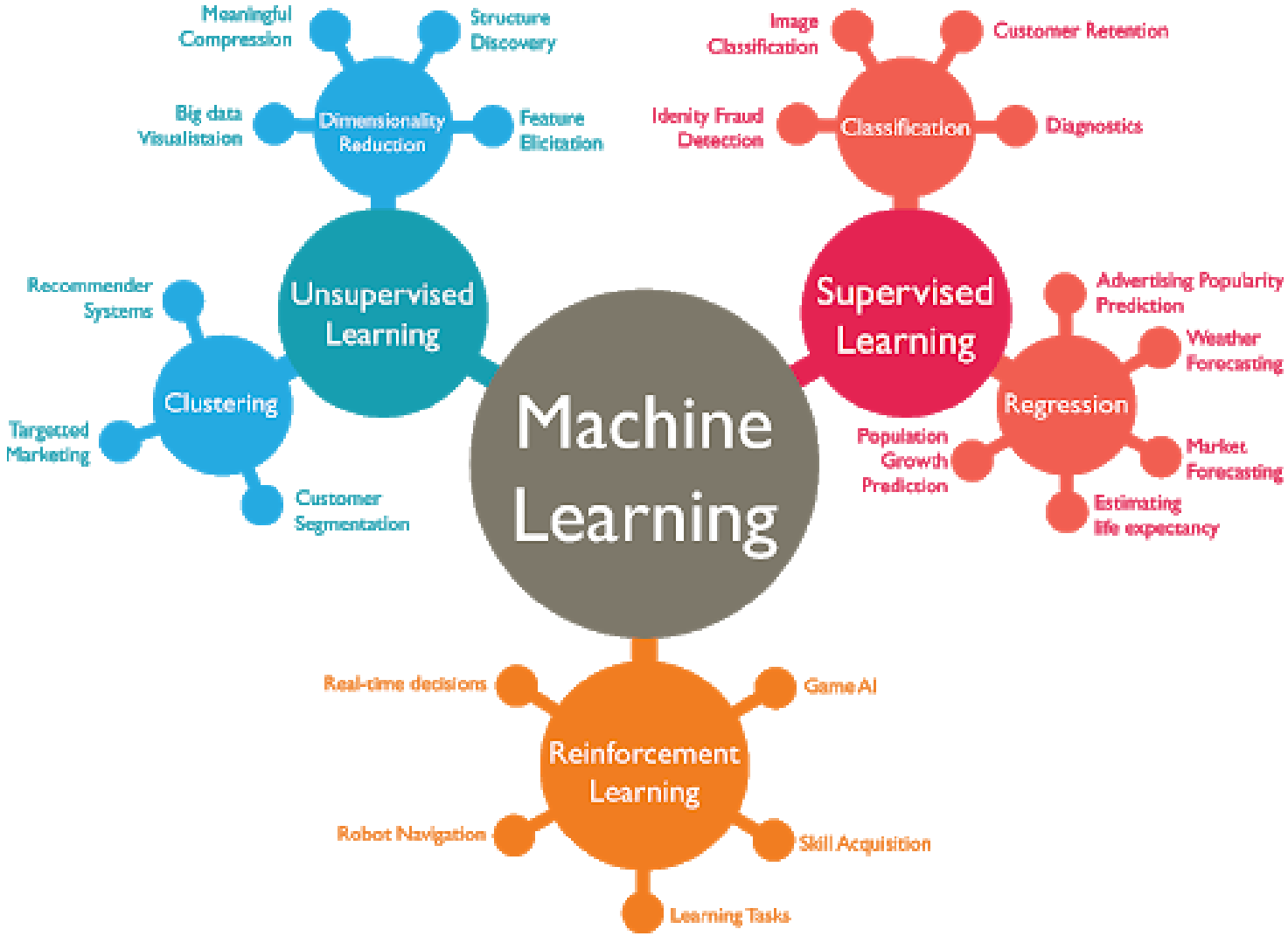
New Assignment

파뿌리
파이썬 뿌시는 이십대들!!

《 Round 9 》

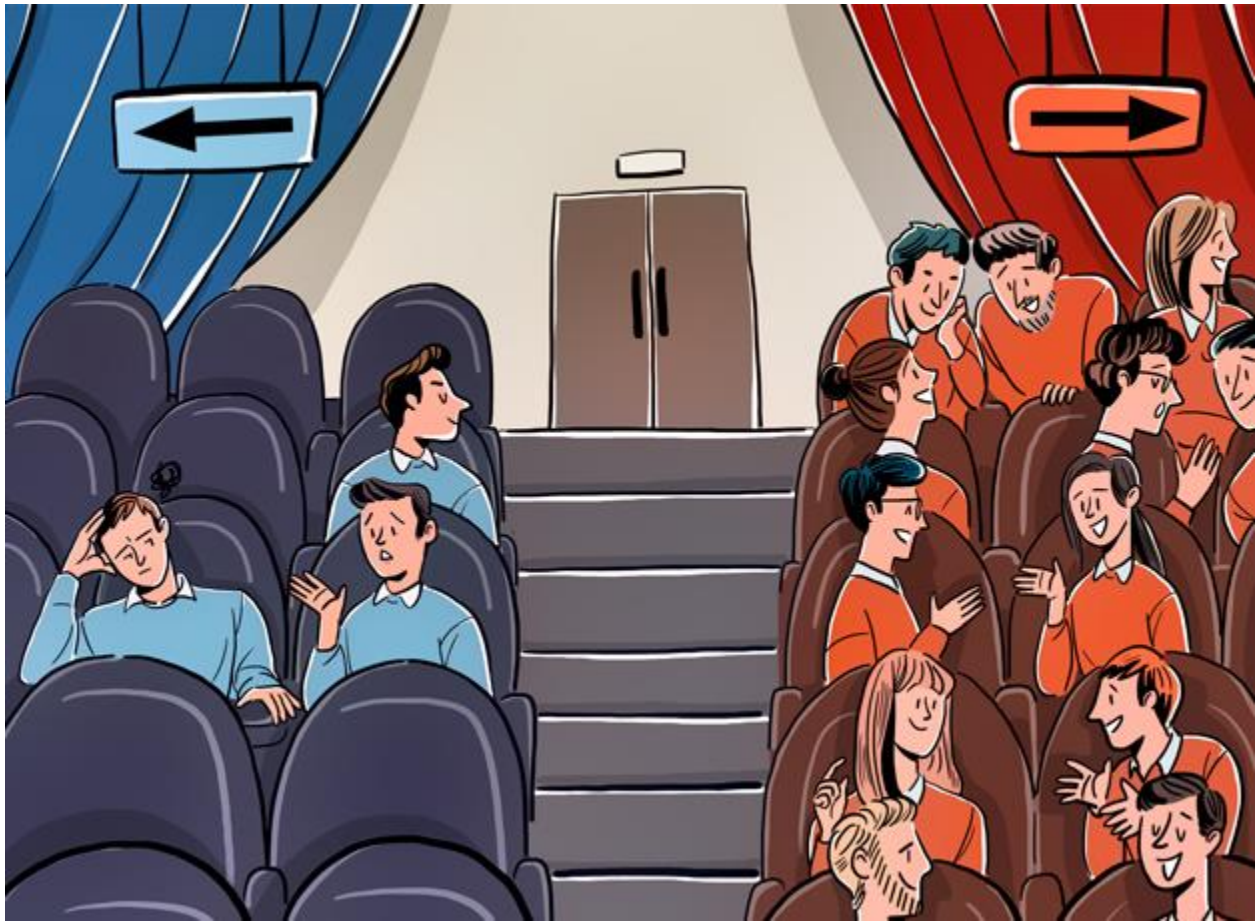
- 분류분석 개요 《
- 분류분석 실습
- 텀 프로젝트 개요

Machine Learning?



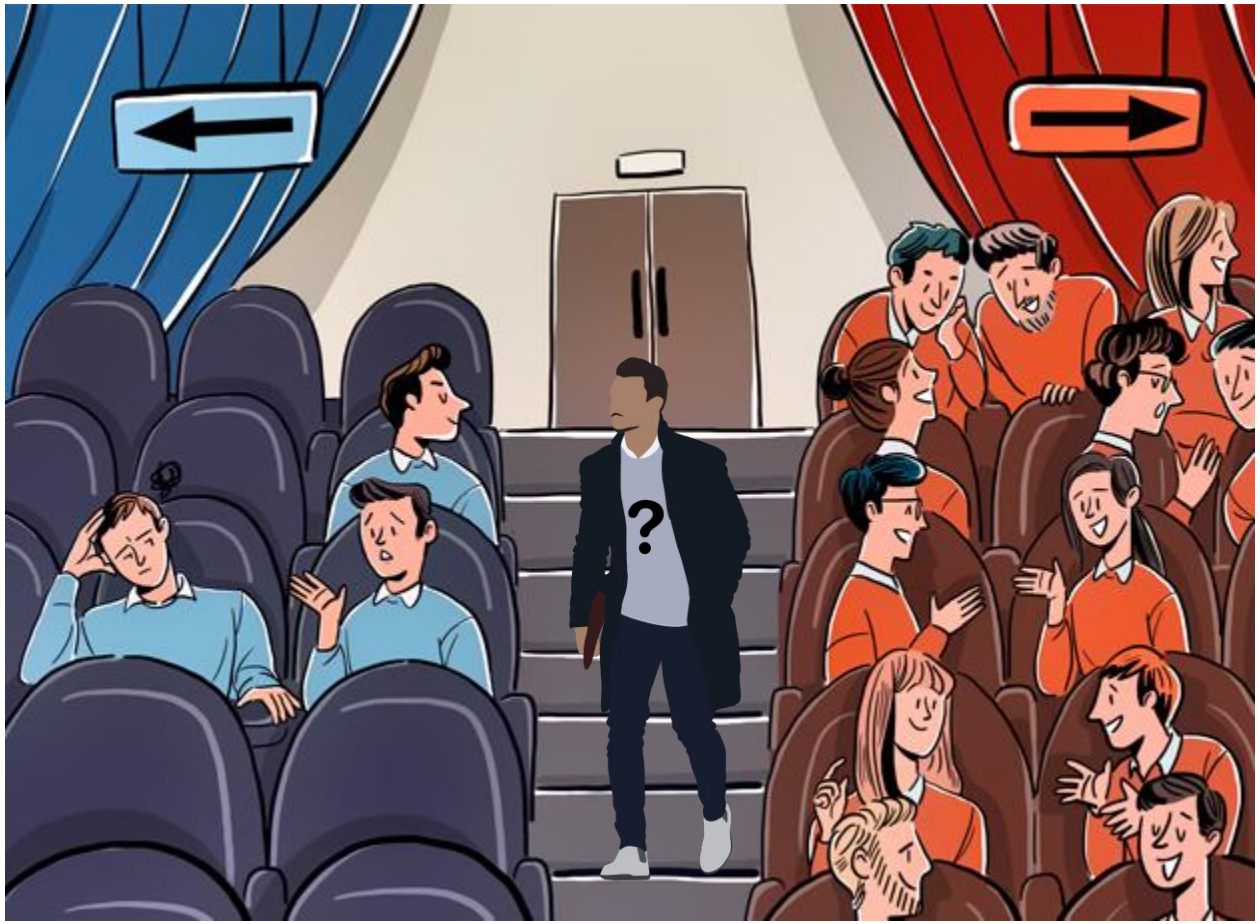
파뿌리
파이썬 뿌시는 이십대들!!

Classification?

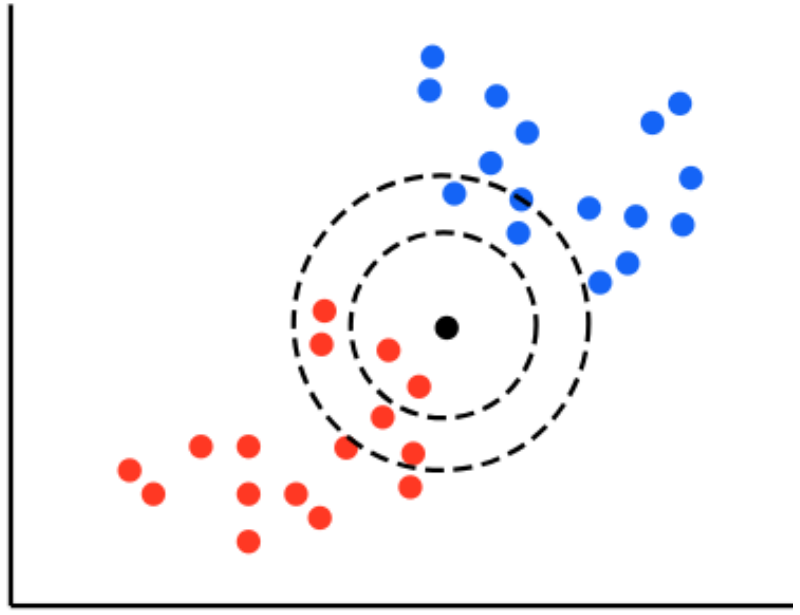


파뿌리
파이썬 뿌시는 이십대들!!

Classification?

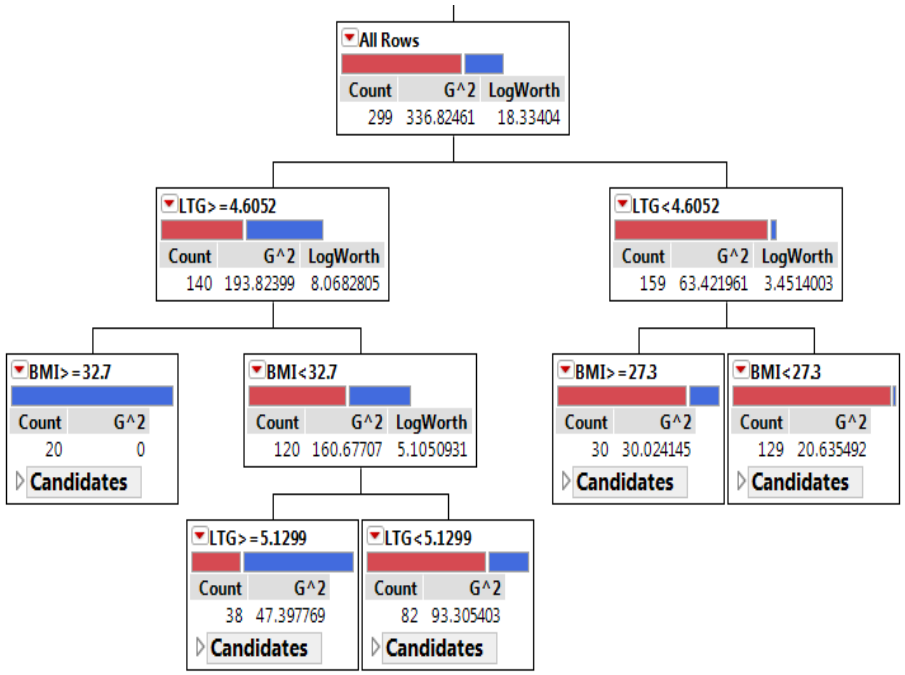


K Nearest Neighbours(KNN)



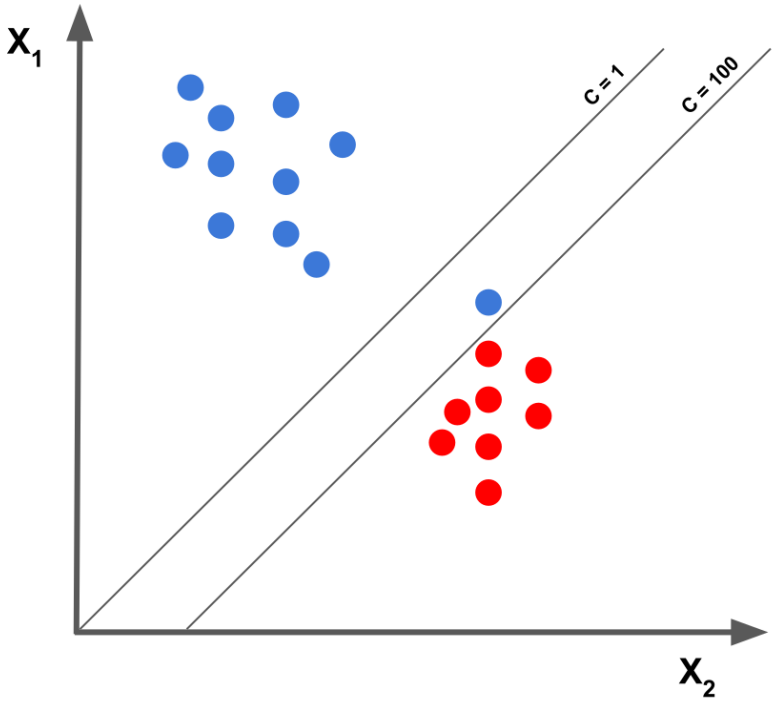
- 거리 기반의 분류 방법
- 어떠한 관측치와 가장 가까운 K개의 이웃으로 분류를 결정
- 과반수의 이웃을 따르기 때문에 K는 홀수 일 때 유리함.
- 이상치에 강하다는 특징.

Decision Tree



- 불순도를 기반으로 한 분류 방법
- 중요성이 높은 변수부터 불순도가 최소가 되는 지점을 분기
- DT 자체가 설명력을 확실히 가져간다는 것이 장점
- LGBM, XGBoost 등 DT를 기반으로 한 모델의 성능적인 강점도 많이 두드러지는 추세.

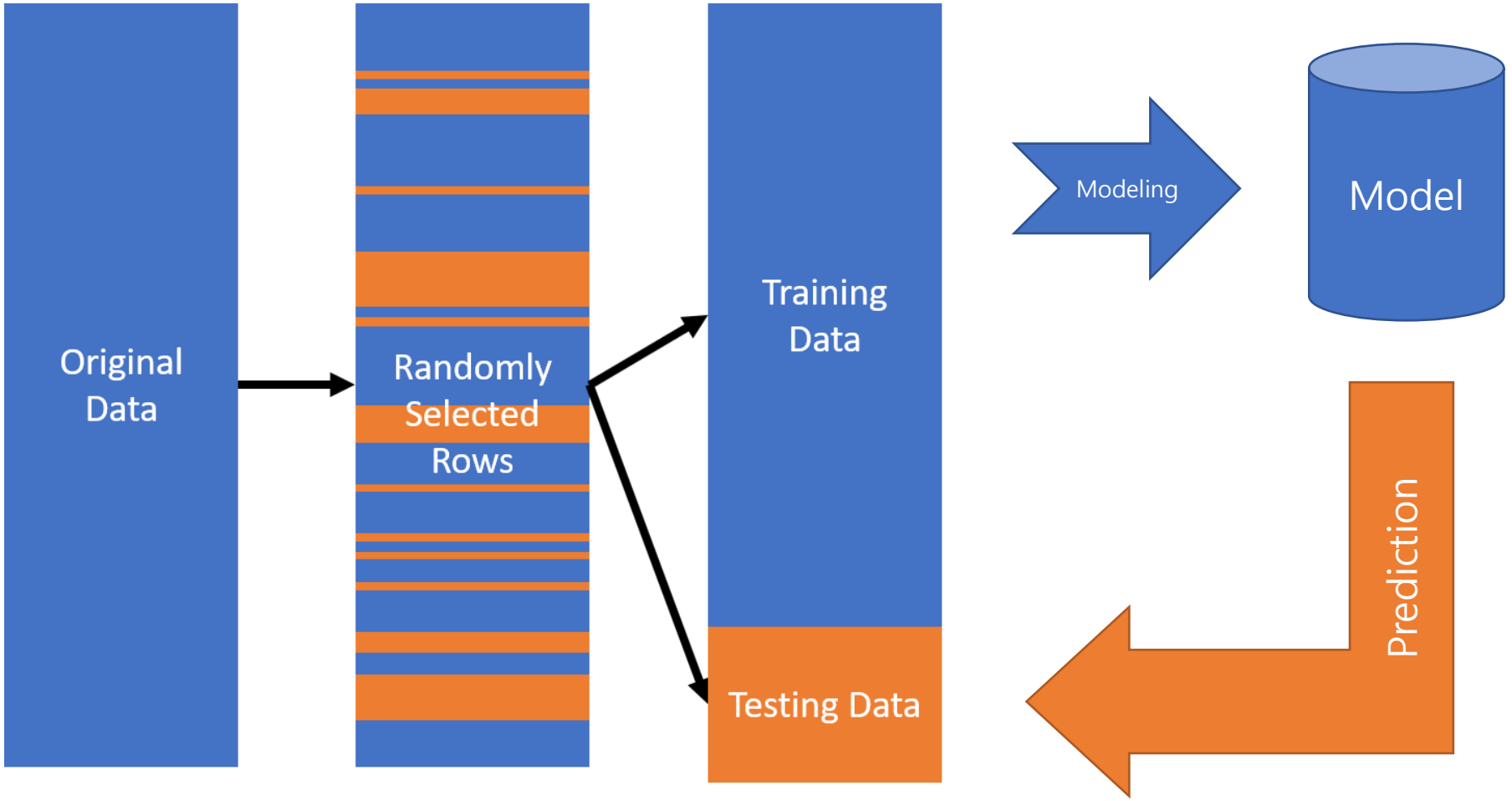
Support Vector Machine



- 두 집단 사이의 마진을 기준으로 한 분류법
- 서포트 벡터 간의 마진이 최대가 되는 선을 기준으로 분류
- 파라미터 C를 통해 오차허용을 유연하게 정의해줄 수 있음
- 대체적으로 빠르고 높은 성능

Learning library

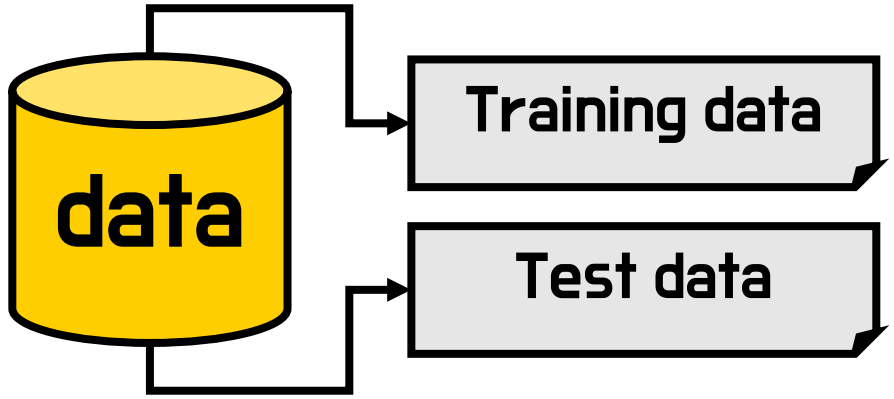
Data splitting



Data splitting

```
## dataset을 training data와 test data로 분할
# train_test_split(독립변수, 종속변수, test data 사이즈(%), 랜덤 추출 시드값)
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=10)

print('train data 개수: ', len(X_train))
print('test data 개수: ', len(X_test),"\n")
```



파뿌리
파이썬 뿌시는 이십대들!!

《 Round 9 》

- 분류분석 개요
- 분류분석 실습 《
- 텀 프로젝트 개요

지표 해석 - Confusion Matrix

		Predicted		
		Yes	No	
Actual	Yes	2 (True +ve)	1 (False -ve)	2/(2+1)=⅔ **Recall (Sensitivity)**
	No	2 (False +ve)	3 (True -ve)	3/(3+2)=⅗ **(Specificity)**
		2/(2+2)=50% **(Precision)**		**Accuracy=** (2+3)/(2+1+2+3)= 5/8

source code :

https://github.com/koptimizer/Python_Breakers/blob/master/season2/code/claSample.ipynb

KNN

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, Y_train)
Y_hat = knn.predict(X_test)

from sklearn import metrics
knn_confusion_matrix = metrics.confusion_matrix(Y_test, Y_hat)
print(knn_confusion_matrix)
```

```
[[13  0  0]
 [ 0 13  1]
 [ 0  4 14]]
```

```
knn_report = metrics.classification_report(Y_test, Y_hat)
print(knn_report)
```

```
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        13
  versicolor       0.76      0.93      0.84        14
   virginica       0.93      0.78      0.85        18

    accuracy                           0.89        45
   macro avg       0.90      0.90      0.90        45
weighted avg       0.90      0.89      0.89        45
```

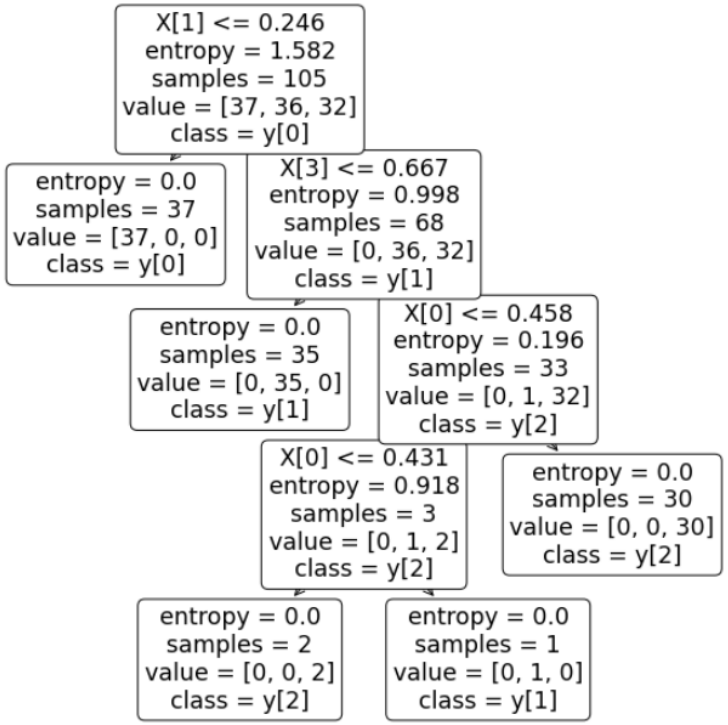
DT

```
from sklearn import tree
dt = tree.DecisionTreeClassifier(criterion='entropy', max_depth=5)
dt.fit(X_train, Y_train)
y_hat = dt.predict(X_test)

from sklearn import metrics
dt_confusion_matrix = metrics.confusion_matrix(Y_test, Y_hat)
print(dt_confusion_matrix)
```

```
[[13  0  0]
 [ 0 13  1]
 [ 0  4 14]]
```

```
fig, ax = plt.subplots(figsize=(10, 10))
tree.plot_tree(dt, impurity=True, class_names = True, rounded = True)
plt.show()
```



SVM

```
from sklearn import svm
svm = svm.SVC(kernel = 'linear')
svm.fit(X_train, Y_train)
y_hat = svm.predict(X_test)

from sklearn import metrics
svm_confusion_matrix = metrics.confusion_matrix(Y_test, Y_hat)
print(svm_confusion_matrix)
```

```
[[13  0  0]
 [ 0 13  1]
 [ 0  4 14]]
```

```
svm_report = metrics.classification_report(Y_test, Y_hat)
print(svm_report)
```

```
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        13
  versicolor       0.76      0.93      0.84        14
   virginica       0.93      0.78      0.85        18

    accuracy                           0.89        45
   macro avg       0.90      0.90      0.90        45
weighted avg       0.90      0.89      0.89        45
```

파뿌리
파이썬 뿌시는 이십대들!!

《 Round 9 》

- 분류분석 개요
- 분류분석 실습
- 텀 프로젝트 개요 《

텀 프로젝트 개요

- 주제는 아래의 데이터셋 중 하나 선택해서 Regression/Classification 자유롭게
- 하나의 .ipynb 파일로 전처리 + EDA + 모델링 + 해석 및 결론을 포함해서 제출
- 평가요소
  - 데이터 분석흐름의 타당성 (20)
  - 전처리 완료된 데이터의 품질 (10)
  - 시각화의 가시성 (10)
  - 모델의 성능 및 선정이유 (20)
  - 결과 해석 및 결론의 효용성 (20)
  - 창의성 (20)
- 결과물이 자신의 Github에 포스팅 되므로 스펙과 역량을 동시에 쌓는다는 생각으로!
- 가장 훌륭한 두 분석가 님의 작품은 재능기부발표회에서 발표기회제공!

파뿌리
파이썬 뿌시는 이십대들!!

LEVEL 1

1. PENGUINS :
- 3 종류의 펭귄에 대한 데이터 344*7

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

2. TIPS :
- 어느 레스토랑에서의 손님들의 결제정보 데이터 244*7

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

파뿌리
파이썬 뿌시는 이십대들!!

LEVEL 2

1. WINE :
- 유명 와인에 대한 데이터 178*14

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065.0	0.0
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050.0	0.0
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185.0	0.0
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480.0	0.0
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735.0	0.0

2. DIANOND :
- 세공 된 다이아몬드에 대한 데이터 6000*8

	Carat Weight	Cut	Color	Clarity	Polish	Symmetry	Report	Price
0	1.10	Ideal	H	SI1	VG	EX	GIA	5169
1	0.83	Ideal	H	VS1	ID	ID	AGSL	3470
2	0.85	Ideal	H	SI1	EX	EX	GIA	3183
3	0.91	Ideal	E	SI1	VG	VG	GIA	4370
4	0.83	Ideal	G	SI1	EX	EX	GIA	3171

파뿌리
파이썬 뿌시는 이십대들!!

LEVEL 3

1. BIKE :
- 2년간 공공자전거의 대여에 관한 데이터 17379*15

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	cnt
0	1	1/1/2011	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	16
1	2	1/1/2011	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	40
2	3	1/1/2011	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	32
3	4	1/1/2011	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	13
4	5	1/1/2011	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	1

2. HOUSE :
- 특정기간동안 거래된 집에 관한 데이터 1460*81

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000

파뿌리
파이썬 뿌시는 이십대들!!

NEXT STAGE