Basic ML Pipeline using sklearn

```
# linear regression
# sklearn basics
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import pandas as pd
```

```
#read data
mtcars = pd.read_csv("https://gist.githubusercontent.com/seankross/a412dfbd88b3db70
mtcars.head()
```

	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

Split data

```
# prepare data
X = mtcars.drop(["model", "mpg"], axis=1)
y = mtcars["mpg"]

# split data
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size = 0.25, random_state = 42 #set.seed() #
)

#result ออกมา 4 ส่วน X_train =0.75
```

```
X_train.shape # ( row, column )
#y_train.shape # ( row, column )
```

```
(24, 10)
```

Predict New Data (Scoring)

```
# train model
model = LinearRegression()      #ประกาศ class linear
model.fit(X_train, y_train)

# test model
p = model.predict(X_test)      #predic
print(p)
```

```
[19.816545   10.98232893 16.31616932 27.16613904 28.59706508 18.29855129
 14.85758111 27.41057736]
```

Model Evaluation

```
model.score(X_test, y_test) # R2
```

```
0.7856209608689562
```

Build model

Desicion Tree Regression

```
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor

tree_model = RandomForestRegressor()

tree_model.fit(X_train, y_train)

p = tree_model.predict(X_test)
print(p)

test_r2 = tree_model.score(X_test, y_test)
print(test_r2)
```

```
[19.784 11.894 16.224 28.921 21.946 18.372 15.051 30.412]
0.8957158153958411
```

** Wrap up "Sklearn" **

- 1.Load data set
- 2.Split data
- 3.initiate model
- 4.fitting
- 5.scoring