# Welcome to 30 Days ML | Day 21

# Import Library

In [2]:
```
import numpy as np
import pandas as pd
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
import seaborn as sns
```

# Import Dataset

In [3]:
```
df = pd.read_csv('train.csv')
```

In [4]:
```
df.head()
```

Out[4]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Emb |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | |

In [5]:
```
df = pd.read_csv('train.csv')[['Age','Pclass','SibSp','Parch','Survived']]
```

In [6]:
```
df.head()
```

Out[6]:

| | Age | Pclass | SibSp | Parch | Survived |
|---|---|---|---|---|---|
| **0** | 22.0 | 3 | 1 | 0 | 0 |
| **1** | 38.0 | 1 | 1 | 0 | 1 |
| **2** | 26.0 | 3 | 0 | 0 | 1 |
| **3** | 35.0 | 1 | 1 | 0 | 1 |
| **4** | 35.0 | 3 | 0 | 0 | 0 |

# Drop NA Value

In [7]:
```
df.dropna(inplace=True)
```

In [8]:
```
df.sample(5)
```

Out[8]:

| | Age | Pclass | SibSp | Parch | Survived |
|---|---|---|---|---|---|
| **663** | 36.0 | 3 | 0 | 0 | 0 |
| **498** | 25.0 | 1 | 1 | 2 | 0 |
| **342** | 28.0 | 2 | 0 | 0 | 0 |
| **136** | 19.0 | 1 | 0 | 2 | 1 |
| **884** | 25.0 | 3 | 0 | 0 | 0 |

# Separate X and Y

In [9]:
```
X = df.iloc[:,0:4]
y = df.iloc[:,-1]
```

In [10]:
```
X.head()
```

Out[10]:

| | Age | Pclass | SibSp | Parch |
|---|---|---|---|---|
| **0** | 22.0 | 3 | 1 | 0 |
| **1** | 38.0 | 1 | 1 | 0 |
| **2** | 26.0 | 3 | 0 | 0 |
| **3** | 35.0 | 1 | 1 | 0 |
| **4** | 35.0 | 3 | 0 | 0 |

# Check Accuracy for Logistic Regression

In [11]:
```
np.mean(cross_val_score(LogisticRegression(),X,y,scoring='accuracy',cv=20))
```

Out[11]: 0.6933333333333332

# Applying Feature Construction

# Create New Column

In [12]:
```
X['Family_size'] = X['SibSp'] + X['Parch'] + 1
```

In [13]:
```
X.head()
```

Out[13]:

| | Age | Pclass | SibSp | Parch | Family_size |
|---|---|---|---|---|---|
| **0** | 22.0 | 3 | 1 | 0 | 2 |
| **1** | 38.0 | 1 | 1 | 0 | 2 |
| **2** | 26.0 | 3 | 0 | 0 | 1 |
| **3** | 35.0 | 1 | 1 | 0 | 2 |
| **4** | 35.0 | 3 | 0 | 0 | 1 |

# Apply New Function

In [14]:
```
def myfunc(num):
    if num == 1:
        #alone
        return 0
    elif num >1 and num <=4:
        # small family
        return 1
    else:
        # large family
        return 2
```

In [15]:
```
myfunc(4)
```

Out[15]: 1

# Apply M Function

In [16]:
```
X['Family_type'] = X['Family_size'].apply(myfunc)
```

In [17]:
```
X.head()
```

Out[17]:

| | Age | Pclass | SibSp | Parch | Family_size | Family_type |
|---|---|---|---|---|---|---|
| **0** | 22.0 | 3 | 1 | 0 | 2 | 1 |
| **1** | 38.0 | 1 | 1 | 0 | 2 | 1 |
| **2** | 26.0 | 3 | 0 | 0 | 1 | 0 |
| **3** | 35.0 | 1 | 1 | 0 | 2 | 1 |
| **4** | 35.0 | 3 | 0 | 0 | 1 | 0 |

# Drop unwanted columns

In [18]:
```
X.drop(columns=['SibSp','Parch','Family_size'],inplace=True)
```

In [19]:
```
X.head()
```

Out[19]:

| | Age | Pclass | Family_type |
|---|---|---|---|
| **0** | 22.0 | 3 | 1 |
| **1** | 38.0 | 1 | 1 |
| **2** | 26.0 | 3 | 0 |
| **3** | 35.0 | 1 | 1 |
| **4** | 35.0 | 3 | 0 |

# Review Accuracy after Feature Construction

In [20]:
```
np.mean(cross_val_score(LogisticRegression(),X,y,scoring='accuracy',cv=20))
```

Out[20]: 0.7003174603174602

# Feature Splitting (New Topic)

# Review Import Data

In [21]:
```
df = pd.read_csv('train.csv')
```

In [22]:
```
df.head()
```

Out[22]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Emb |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | |

In [23]:
```
#Use Name Column
```

In [24]:
```
df['Name']
```

Out[24]:
```
0                                Braund, Mr. Owen Harris
1      Cumings, Mrs. John Bradley (Florence Briggs Th...
2                                 Heikkinen, Miss. Laina
3           Futrelle, Mrs. Jacques Heath (Lily May Peel)
4                               Allen, Mr. William Henry
                                ...
886                                Montvila, Rev. Juozas
887                         Graham, Miss. Margaret Edith
888             Johnston, Miss. Catherine Helen "Carrie"
889                                Behr, Mr. Karl Howell
890                                  Dooley, Mr. Patrick
Name: Name, Length: 891, dtype: object
```

# Separate Salutation

In [25]:
```
df['Title'] = df['Name'].str.split(', ', expand=True)[1].str.split('.', expand=True)[
```

In [26]:
```
df['Name'].str.split(', ', expand=True)[1].str.split('.', expand=True)[0]
```

Out[26]:
```
0        Mr
1       Mrs
2      Miss
3       Mrs
4        Mr
       ...
886     Rev
887    Miss
888    Miss
889      Mr
890      Mr
Name: 0, Length: 891, dtype: object
```

In [27]:
```
df[['Title','Name']]
```

Out[27]:

| | Title | Name |
|---|---|---|
| **0** | Mr | Braund, Mr. Owen Harris |
| **1** | Mrs | Cumings, Mrs. John Bradley (Florence Briggs Th... |
| **2** | Miss | Heikkinen, Miss. Laina |
| **3** | Mrs | Futrelle, Mrs. Jacques Heath (Lily May Peel) |
| **4** | Mr | Allen, Mr. William Henry |
| **...** | ... | ... |
| **886** | Rev | Montvila, Rev. Juozas |
| **887** | Miss | Graham, Miss. Margaret Edith |
| **888** | Miss | Johnston, Miss. Catherine Helen "Carrie" |
| **889** | Mr | Behr, Mr. Karl Howell |
| **890** | Mr | Dooley, Mr. Patrick |

891 rows × 2 columns

# Review Analysis after Feature Splitting

In [28]:
```
np.mean(cross_val_score(LogisticRegression(),X,y,scoring='accuracy',cv=20))
```

Out[28]:
```
0.7003174603174602
```