

Titanic

NO.

Library

```
{ import numpy as np
  import pandas as pd
  import seaborn as sns
  sns.set('darkgrid')
  style =
```

```
{ from sklearn.ensemble import RandomForestClassifier
  from sklearn.preprocessing import OneHotEncoder, LabelEncoder, StandardScaler
  from sklearn.metrics import roc_curve, auc
  from sklearn.model_selection import StratifiedKFold
```

```
import string
```

```
import warnings
```

```
warnings.filterwarnings('ignore')
```

```
SEED = 42
```

```
def concat_df(train_data, test_data):
```

```
    return pd.concat([train_data, test_data], sort=True).reset_index(
        drop=True)
```

```
def divide_df(all_data)
```

```
    return all_data.iloc[:890], all_data.iloc[891:].drop(['Survived'],
        axis=1)
```





## 1.2.1 Age.

```
df_all_corr = df_all_corr.sort_values(ascending=False, kind='quicksort', inplace=True)
```

```
df_all_corr.rename(columns={'Age_0': 'Feature 1', ... 3, inplace=True)
```

```
df_all_corr[df_all_corr['Feature 1'] == 'Age']
```

index 별

median값

찾아 올림.

```
age_by_pclass_sex = df_all.groupby(['Sex', 'pclass']).median()['Age']
```

```
df_all['Age'] = df_all.groupby(['Sex', 'pclass'])['Age'].apply(lambda x: x.fillna(x.median()))
```

## 1.2.2 Embarked.

```
df_all[df_all['Embarked'].isnull()]
```

```
df_all['Embarked'] = df_all['Embarked'].fillna('S')
```

## 1.2.3 Fare

['Fare']

```
med_fare = df_all.groupby(['pclass', 'parch', 'sibsp'])['Fare'].median()[3][0]
```

['Fare']

## 1.2.4 Cabin.

```
df_all['Deck'] = df_all['Cabin'].apply(lambda s: s[0] if pd.notnull(s) else 'N')
```

```
df_all_decks = df_all.groupby(['Deck', 'Pclass']).count().drop(columns=['...'])
```

```
def get_pass_dist(df):
```

```
    # Creating a dictionary for every passenger class count in every deck.
```

```
    deck_counts = {'A': {}, 'B': {}, ..., 'T': {}}
```

```
    decks = df.columns.levels[0]
```

```
    for deck in decks:
```

```
        for pclass in range(1, 4):
```

```
            try:
```

```
                count = df[deck][pclass][0]
```

```
                deck_counts[deck][pclass] = count
```

```
            except KeyError:
```

```
                deck_counts[deck][pclass] = 0
```

```
    # Percentage.
```

```
    df_decks = pd.DataFrame(deck_counts)
```

```
    deck_percentages = {}
```

```
    for col in df_decks.columns:
```

```
        deck_percentages[col] = [(count / df_decks[col].sum()) * 100
```

```
            for count in df_decks[col]]
```

```
    return deck_counts, deck_percentages
```



```
def display_pclass_dist(percentages):
```

```
    Runtime  
    df_percentages = pd.percentages, transpose)
```

```
    deck_names = ('A', ..., 'T')
```

```
    bar_count = np.arange(len(deck_names))
```

```
    bar_width = 0.85
```

```
    pclass1 = df_percentages[0]
```

```
    pclass2 = df_percentages[1]
```

```
    pclass3 = df_percentages[2]
```

```
    plt.figure(figsize=(20,10))
```

```
    plt.bar(bar_count, pclass1, color='#659494', edgecolor='white',  
            width=bar_width, label='passenger class 1')
```

```
    plt.bar(" " , bottom=pclass1 " ")
```

```
    plt.bar(" " , bottom=pclass1 + pclass2
```

각 줄의 label 사이의 간격

```
    plt.xlabel('Deck', size=15, labelpad=20)
```

```
    plt.ylabel('Passenger Class Percentage', size=15, labelpad=20)
```

```
    plt.xticks(bar_count, deck_names) → X의 label 설정
```

```
    plt.tick_params(axis='x', labelsize=15)
```

```
    " (" " "y", " ")
```

```
    plt.legend(loc='upper left', bbox_to_anchor=(1,1), prop={'size':15})
```

```
    plt.title(' ', size=18, y=1.05)
```

```
    plt.show()
```

```
all_deck_count, all_deck_per = get_pchss_dist(df_all_decks)
display_pchss_dist(all_deck_per)
```

# Passenger in the T deck is changed to A

```
idx = df_all[df_all['Deck'] == 'T'].index,
df_all.loc[idx, 'Deck'] = 'A'
```

```
df_all_decks_survived = df_all.groupby(['Deck', 'Survived']).count(),
                        drop(columns=[.....]),
                        rename(columns={ 'None': 'Count' }),
                        transpose()  ↳ 기존 colname,  ↳ 바뀐 colname.
```

```
def get_survived_dist(df):
```

# Creating a dictionary for every survival count in every deck.

```
surv_counts = {'A': { }, ... }
```

```
decks = df.columns.levels[0]
```

```
for deck in decks:
```

```
    for survive in range(0, 2):
```

```
        surv_counts[deck][survive] = df[deck][survive].count()
```

```
df_surv = pd.DataFrame(surv_counts)
```

```
surv_percentages = { }
```

```
for col in df_surv.columns:
```

```
    surv_percentages[col] = (df_surv[col].sum() / df_surv[col].count()) * 100
```

```
    for count in df_surv[col].values:
```



```
return surv_counts, surv_percentages
```

```
def display_surv_dist(percentages):
```

```
    {
```

와 거의 동일.

```
all_surv_count, all_surv_per = get_survived_dist(df_all_decks_surv)
```

```
display_surv_dist(all_surv_per)
```

value 바뀜. 

```
df_all['Deck'] = df_all['Deck'].replace(['A', 'B', 'C'], 'ABC')
```

```
}
```

```
df_all['Deck'], value_counts()
```

# Dropping the Cabin feature

```
df_all.drop(['Cabin'], axis=1, inplace=True)
```

```
df_train, df_test = divide_df(df_all)
```

```
dfs = [df_train, df_test]
```

```
for df in dfs:
```

```
    display_missing(df)
```

### 1.3 Survival Distribution

```
survived = df_train['Survived'].value_counts()[1]
```

```
not_survived = " " [0]
```

```
survived_per = survived / df_train.shape[0] * 100
```

```
not " not " " "
```

```
plt.figure(figsize=(10,8))
```

```
sns.countplot(df_train['Survived'])
```

### 1.4 Correlations

```
fig, ax = plt.subplots(nrow=2, figsize=(20,20))
```

```
sns.heatmap(df_train.drop(['PassengerID'], axis=1).corr(),
```

```
ax=ax[0], annot=True, annot_kws={'size': 14},
```

```
square=True, cmap='coolwarm')
```

```
" df_test "
```



## 1.5 Survival Distribution in Features.

```
cont_features = ['Age', 'Fare']
```

```
surv = df_train['Survived'] == 1
```

```
fig, axs = plt.subplots(ncols=2, nrows=2, figsize=(20, 20))
```

```
plt.subplots_adjust(right=1.5)
```

```
for i, feature in enumerate(cont_features):
```

```
    # Distribution of survival in feature
```

```
    sns.distplot(df_train[~surv][feature], label='Not Survived', hist=True,  
                 color=' ', ax=axs[0][i])
```

```
    sns.distplot(df_train[surv][feature], label='Survived',  
                 " ")
```

```
axs[0][i].set_xlabel(" ") } X2 label None. 1/4  
[1] " ( )
```

```
for j in range(2):
```

```
    axs[i][j].tick_params(axis='x', labelsize=20)
```

```
    " y'
```

## 1.5.2 Categorical Features

```
cat_features = [ ~ ]
```

```
fig, axs = plt.subplots(ncols=2, nrows=3, figsize=(20,20))
```

```
plt.subplots_adjust(right=1.5, top=1.25)
```

```
for i, feature in enumerate(cat_features, start=1):
```

```
    plt.subplot(2, 3, i)  # Categorical label
```

```
    sns.countplot(x=feature, hue='Survived', data=df_train)
```

## 1.6 Conclusion

```
df_all = concat_df(df_train, df_test)
```

```
df_all.head()
```



## 2. Feature Engineering

### 2.1 Binning the continuous features

#### 2.1.1 Fare

13개의 구간으로 나눈다.

```
df_all['Fare'] = pd.qcut(df_all['Fare'], 13)
```

```
fig, axes = plt.subplots(figsize=(22, 9))
```

```
sns.catplot(x='Fare', hue='Survived', data=df_all)
```

} Option

#### 2.1.2 Age

```
df_all['Age'] = pd.qcut(df_all['Age'], 10)
```

```
fig, axes = plt.subplots(figsize=(22, 9))
```

```
sns.catplot(x='Age', hue='Survived', data=df_all)
```

}