DATA UNDERSTANDING

Disusun : Eka Dyar Wahyuni

1. Import library yang diperlukan

```
In [1]: %matplotlib inline
import numpy as np
import scipy as sp
import matplotlib as mpl
import matplotlib.cm as cm
import matplotlib.pyplot as plt
import pandas as pd
from pandas.tools.plotting import scatter_matrix
pd.set_option('display.width', 500)
pd.set_option('display.max_columns', 100)
pd.set_option('display.notebook_repr_html', True)
import seaborn as sns
sns.set(style="whitegrid")
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from sklearn.model_selection import train_test_split
```

2. Load data set, dataset yang dipergunakan ada pada ilmu, sesuaikan path pada perintah read_csv dengan posisi folder download anda

```
data = pd.read_csv('/home/eka/Downloads/train.csv')
```

Cek data set

```
In [5]: data.head()
```

Out[5]:

	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
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

3. Pemahaman data (EDA) dan preprocess data

- Statistik

```
In [6]: data.describe()
```

Out[6]:

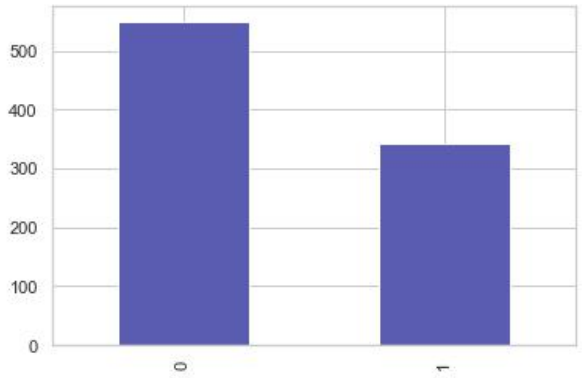
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Tugas 1 : apa yang bisa anda terjemahkan dari tabel hasil fungsi describe diatas ?

- Visual
  - Distribusi variabel

```
In [15]: data['Survived'].value_counts().plot(kind='bar')
         data['Survived'].value_counts()
```

```
Out[15]: 0    549
         1    342
         Name: Survived, dtype: int64
```



Dari grafik diatas, dapat diamati, bahwa jumlah yang meninggal adalah 549, yang selamat 342. bagaimana nilai kolom yang lain ?

- Perbandingan survival rate dengan variabel lainnya

  1) Buat fungsi untuk menghitung dan menampilkan grafiknya

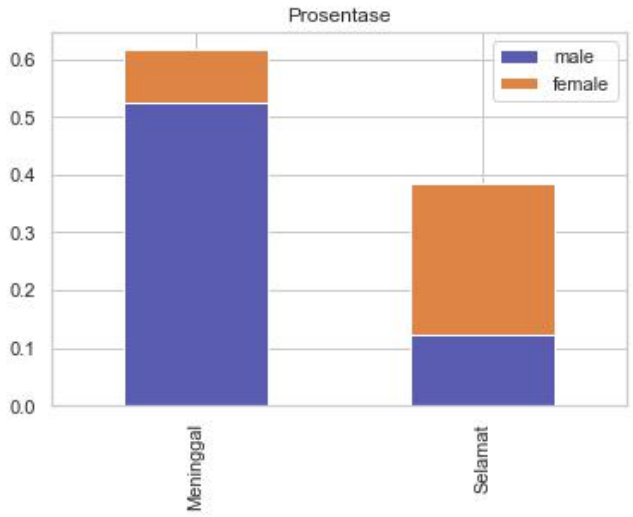
```
In [16]: def survival_stacked_bar(variable):
             died=data[data['Survived']==0][variable].value_counts()/len(data['Survived']==0)
             survived=data[data['Survived']==1][variable].value_counts()/len(data['Survived']==1)
             dataset=pd.DataFrame([died,survived])
             dataset.index=['Meninggal', 'Selamat']
             dataset.plot(kind='bar',stacked=True,title='Prosentase')
             return dataset.head()
```

  2) Panggil fungsi tersebut

```
In [17]: survival_stacked_bar('Sex')
```

Out[17]:

	male	female
Meninggal	0.525253	0.090909
Selamat	0.122334	0.261504



Tugas 2 : buat sebanyak mungkin, grafik yang menggambarkan relasi berbagai macam kolom sesuai interpretasi anda. Kumpulkan semua screenshot script python dan juga screenshot grafik yang dihasilkan

Kuailitas data

- Missing value ?

  Cek apakah ada missing value

```
In [13]: data.isnull().sum()
Out[13]: PassengerId      0
         Survived         0
         Pclass           0
         Name             0
         Sex              0
         Age            177
         SibSp            0
         Parch            0
         Ticket           0
         Fare             0
         Cabin          687
         Embarked         2
         dtype: int64
```

Untuk mengidentifikasi selamat atau tidaknya seseorang, apakah umur, kabin dan embarked diperlukan ?

Jika iya, bagaimana skenarionya ?

Skenario untuk missing value -- kolom yang mengandung missing value, didrop

Cek data

```
In [12]: data
```

Out[12]:

	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
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S

Perhatikan data diatas, apakah PassengerId mempengaruhi selamat atau tidaknya seseorang ? bagaimana dengan kolom yang lainnya ?

Tugas 3 : Bagaimana dengan dimensi kualitas data yang lain ? dan jika ada kolom yang tidak “berkualitas” bagaimana solusi mengatasinya