Titanic Dataset Project

About Titanic Dataset

Tenggelamnya Titanic adalah salah satu bangkai kapal paling terkenal dalam sejarah.

Pada tanggal 15 April 1912, selama pelayaran perdananya, RMS Titanic yang secara luas dianggap “tidak dapat tenggelam” tenggelam setelah bertabrakan dengan gunung es. Sayangnya, tidak ada sekoci yang cukup untuk semua orang di dalamnya, mengakibatkan kematian 1502 dari 2224 penumpang dan awak.

Meskipun ada unsur keberuntungan yang terlibat dalam bertahan hidup, tampaknya beberapa kelompok orang lebih mungkin bertahan hidup daripada yang lain.

Project ini membangun model prediktif untuk menjawab pertanyaan: "orang seperti apa yang lebih mungkin bertahan?" menggunakan data penumpang (yaitu nama, umur, jenis kelamin, kelas sosial ekonomi, dll).

Train.csv akan berisi detail subset penumpang di pesawat (tepatnya 891) dan yang terpenting, akan mengungkapkan apakah mereka selamat atau tidak, kumpulan data `test.csv` berisi informasi serupa dan tujuan dari project ini untuk memprediksi hasil ini.

Titanic Project

1. Exploratory Data Analysis
2. Data Cleaning
3. Feature Scaling
4. Data Modelling
5. Confusion Matrix
6. Prediction
7. Conclusion

1. Exploratory Data Analysis

```
[4] # printing first five rows of the data set
    train.head()
```

	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] # number of rows and columns
    train.shape
```

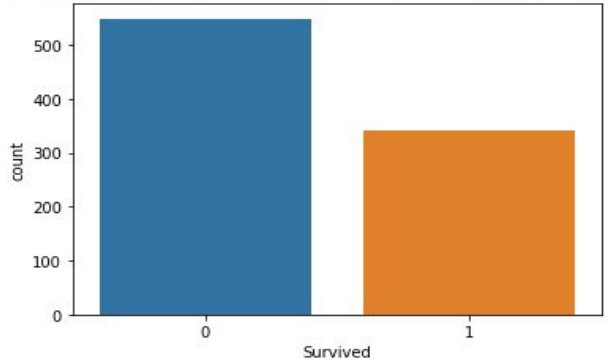
```
(891, 12)
```

Dataset titanic berisi 891 baris dan 12 kolom data training, yang berisi informasi tentang orang yang berada didalam kapal titanic. Referensi dari dataset ini bersumber dari kaggle. Selanjutnya dataset ini akan dilakukan analisa machine learning dengan menggunakan metode regresi logistik. Tujuan dari analisa ini adalah membuat prediksi apakah penumpang kapal titanic akan survived atau tidak.

```
[12] count of Survived
0    549
1    342
Name: Survived, dtype: int64
% of Survived
0    61.616162
1    38.383838
Name: Survived, dtype: float64
```

```
# countplot
sns.countplot(x='Survived', data=train)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1e56ce0700>
```



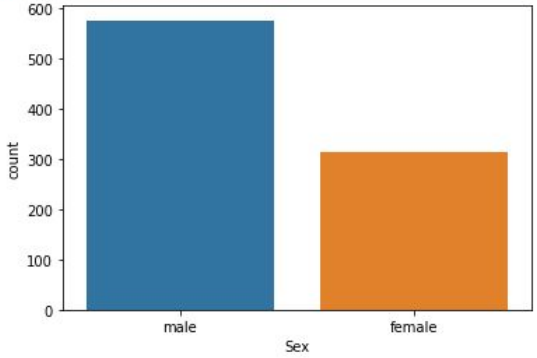
Count of Survived

Jumlah penumpang yang survived adalah 342 (38,38 %), sementara jumlah penumpang yang tidak survived adalah 549 (61,62%).

```
[ ] count of sex
male      577
female    314
Name: Sex, dtype: int64
% of Sex
male      64.758698
female    35.241302
Name: Sex, dtype: float64
```

```
[ ] # countplot
sns.countplot(x='Sex', data=train)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe24251f100>
```



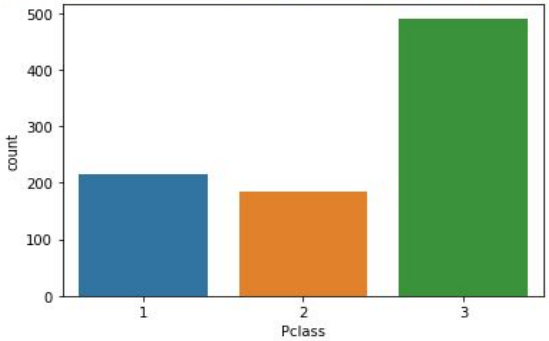
Count of Sex

Jumlah penumpang dengan jenis kelamin pria adalah 577 (64,75%) dan jumlah penumpang dengan jenis kelamin wanita adalah 314 (35,25%).

```
count of Pclass
3    491
1    216
2    184
Name: Pclass, dtype: int64
% of Pclass
3    55.106622
1    24.242424
2    20.650954
Name: Pclass, dtype: float64
```

```
sns.countplot(x='Pclass', data=train)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe2425184c0>
```



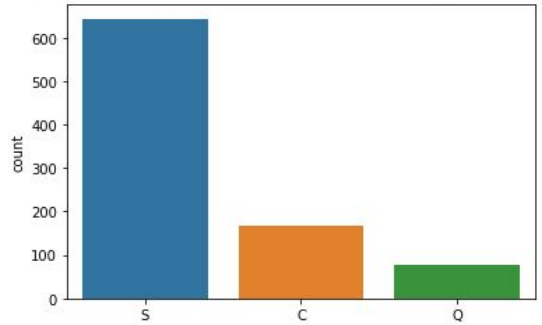
Count of Pclass

Terdapat 3 kategori kelas penumpang. Penumpang dengan kategori kelas 1 berjumlah 216 (24,24%), penumpang dengan kategori kelas 2 berjumlah 184 (20,65%), dan penumpang dengan kategori kelas 3 berjumlah 491 (55,11%)

```
[ ]
count of Embarked
S    644
C    168
Q     77
Name: Embarked, dtype: int64
% of Embarked
S    72.440945
C    18.897638
Q     8.661417
Name: Embarked, dtype: float64
```

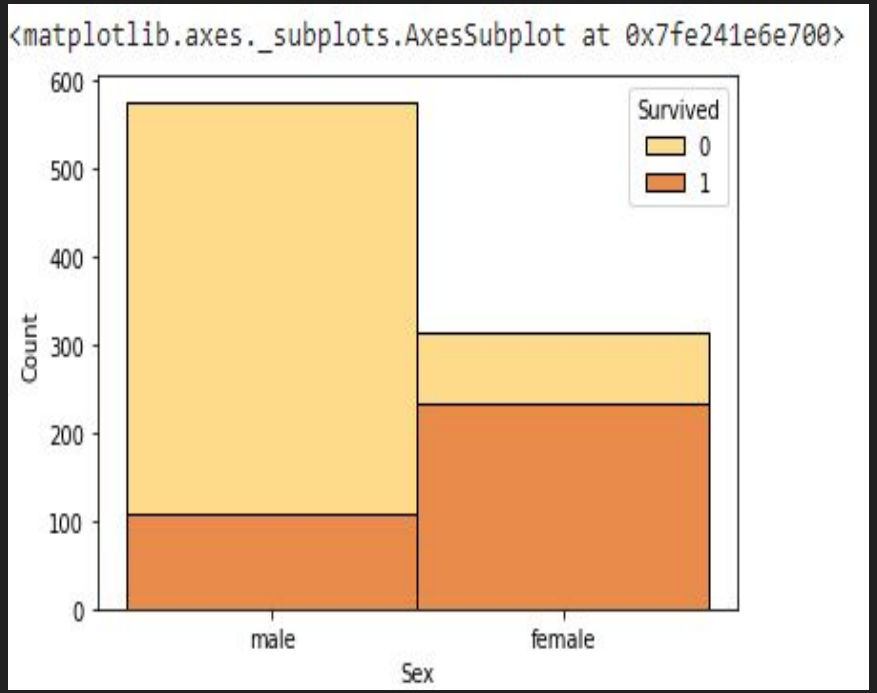
```
sns.countplot(x='Embarked', data=train)
```

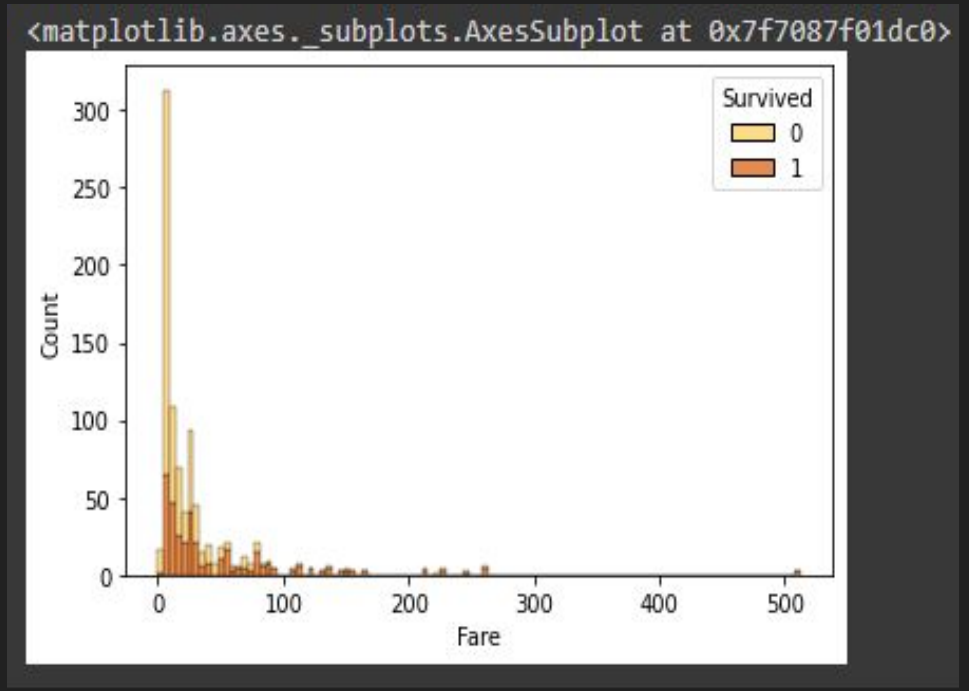
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe241f7e9a0>
```



Count of Embarked

Terdapat 3 kategori embarked penumpang. Penumpang dengan kategori embarked ‘S’ berjumlah 644 (72,44%), penumpang dengan kategori embarked ‘C’ berjumlah 168 (18,90%), dan penumpang dengan kategori embarked ‘Q’ berjumlah 77 (8,66%)





Visualizing Survival Based on The Gender

Berdasarkan visualisasi data diatas dapat diambil kesimpulan bahwa penumpang dengan gender wanita lebih banyak survived dibandingkan penumpang dengan gender pria

Visualizing Survival Based on The Fare

Penumpang dengan tarif tiket yang lebih murah lebih mungkin meninggal. Dengan kata lain, penumpang dengan tiket lebih mahal, dan status sosial yang lebih penting, tampaknya akan diselamatkan terlebih dahulu.

2. Cleaning The Train Dataset

```
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
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

Number of Null Values in Dataset

Dataset train memiliki nilai null pada beberapa kolomnya, kolom age memiliki 177 nilai null, kolom cabin memiliki 687 nilai null, dan data embarked memiliki 2 nilai null

Dropping Unwanted Columns

Beberapa kolom didalam dataset train dihapuskan karena tidak dilibatkan didalam pengamatan, kolom yang dihapus adalah Name, Ticket, Cabin, PassengerId.

```
df1.isnull().sum()
```

```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        0
Embarked    0
dtype: int64
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	22.0	1	0	7.2500	0
1	1	1	0	38.0	1	0	71.2833	1
2	1	3	0	26.0	0	0	7.9250	0
3	1	1	0	35.0	1	0	53.1000	0
4	0	3	1	35.0	0	0	8.0500	0

Data is cleaned to have no null value

Kolom cabin sudah dihapus dalam langkah sebelumnya karena tidak dilibatkan dalam pengamatan. Untuk kolom age nlai null diisi dengan cara impute median kedalam data karena data age memiliki nilai skew. Untuk 2 nilai null didalam kolom embarked dihapus karena jumlah 2 nilai null tidak signifikan dengan jumlah total dataset 891

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
count	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000
mean	0.382452	2.311586	0.649044	0.364099	0.524184	0.382452	0.062649
std	0.486260	0.834700	0.477538	0.163160	1.103705	0.806761	0.097003
min	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	0.000000	0.271174	0.000000	0.000000	0.015412
50%	0.000000	3.000000	1.000000	0.359135	0.000000	0.000000	0.028213
75%	1.000000	3.000000	1.000000	0.434531	1.000000	0.000000	0.060508
max	1.000000	3.000000	1.000000	1.000000	8.000000	6.000000	1.000000

3. Feature Scaling

Feature scaling adalah teknik yang digunakan dalam preprocessing data untuk mengubah nilai-nilai fitur (atribut) dalam dataset ke dalam rentang yang sama. Karena age dan fare memiliki rentang yang berbeda maka dilakukan feature scaling.

```python
# Logistic regression
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression()
clf.fit(X_train, y_train)

from sklearn.metrics import accuracy_score

Y_pred = clf.predict(X_test)
accuracy_score(y_test, Y_pred)
```

```
0.8370786516853933
```

4. Data Modelling

Data dimodelkan dengan model machine learning regresi logistik, karena hasil yang diharapkan untuk memprediksi apakah penumpang akan survived atau tidak, regresi logistik adalah model yang cocok untuk memodelkan data tersebut.

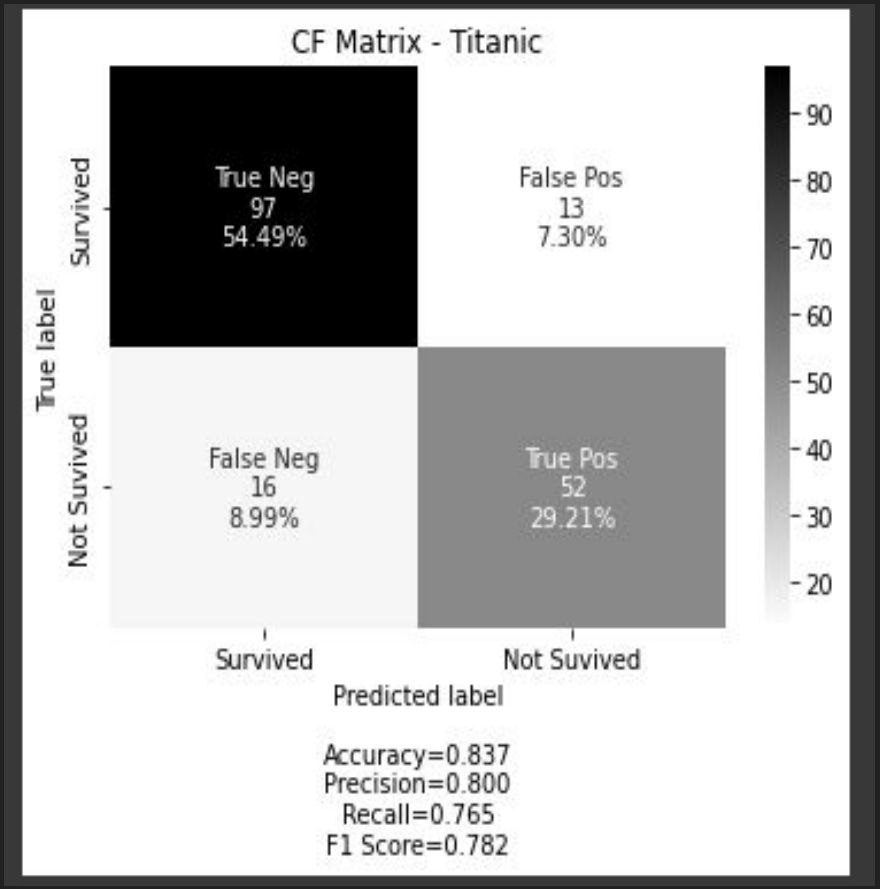
Regresi logistik pada data menghasilkan accuracy_score 83,71 % dan saat diuji dengan cross validation menghasilkan score 79,42%. Kesimpulannya model dari regresi logistik ini cukup bagus untuk memodelkan dan memprediksi data.

```python
# Cross Validation
from sklearn.model_selection import cross_val_score

scores = cross_val_score(
    clf, df1.drop(['Survived'], axis=1), df1['Survived'], cv=10, scoring='accuracy')

np.mean(scores)
```

```
0.79417888656129
```

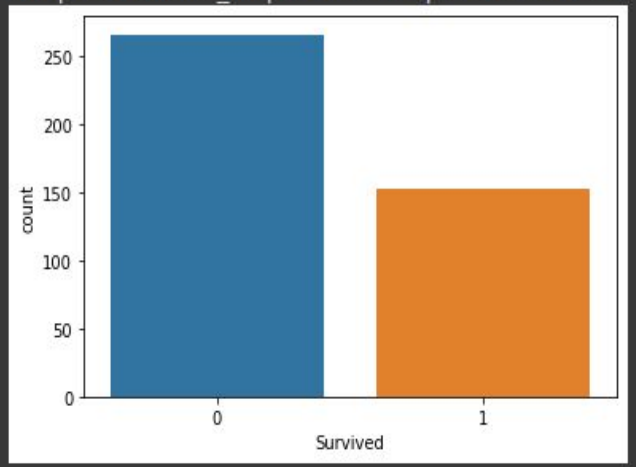


5. Confusion Matrix

Berdasarkan hasil confusion matrix ketika model regresi logistik memprediksi tidak survived dan nilai aktual tidak survived (True Positif) mendapatkan nilai 29,21%. Ketika model regresi logistik memprediksi survived dan nilai aktual survived (True Negatif) mendapatkan nilai 54.49%. Sehingga accuracy model untuk memprediksi adalah 83,70%, nilai ini cukup bagus untuk memprediksi model.

```
count of Survived
0    266
1    152
Name: Survived, dtype: int64
% of Survived
0    63.636364
1    36.363636
Name: Survived, dtype: float64
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f536382f910>
```



6. Prediction

Setelah mendapatkan model prediksi menggunakan regresi logistik selanjutnya model digunakan untuk memprediksi data, ada 418 data yang akan diprediksi. Berdasarkan hasil prediksi jumlah penumpang yang tidak survived adalah 266 penumpang (63,64%) dan jumlah penumpang yang survived adalah 152 penumpang (36,36%)

7. Conclusion

Jumlah penumpang yang tidak survived lebih banyak dibanding jumlah penumpang yang tidak survived. Penumpang dengan gender wanita lebih banyak survived dibandingkan penumpang dengan gender pria. Penumpang dengan tiket lebih mahal, dan status sosial yang lebih penting, akan diselamatkan terlebih dahulu.

Aldilah Ariwibowo

Linkedin
www.linkedin.com/in/aldilah-ariwibowo

Desainer Senior
https://github.com/aldilahariwibowo