Pokok Bahasan X
Analisis Group Means Titanic Dataset

Kode Pokok Bahasan: TIK.RPL03.001.006.01

Deskripsi Pokok Bahasan:
Membahas tentang Studi Kasus Data Mining pada Python dengan Titanic Dataset

No	Elemen Kompetensi	Indikator Kinerja	Jml Jam	Hal
1	Memahami proses analisis group means	Mampu Lakukan analisis group means berdasarkan nilai survival sebagai grouping variable pada Python	1	12
2	Menerapkan decision tree untuk membangun model	Mampu melakukan pemodelan prediksi dengan decision tree	2	15

TUGAS PENDAHULUAN
Hal yang harus dilakukan dan acuan yang harus dibaca sebelum praktikum :
1. Menginstal Python pada PC masing-masing praktikan.
2. Menginstal Python Studio pada PC masing-masing praktikan.

DAFTAR PERTANYAAN
1. Apa tujuan melakukan analisis group means?
Analisis Diskriminan adalah salah satu tehnik analisa Statistika dependensi yang memiliki kegunaan untuk mengklasifikasikan objek beberapa kelompok. Pengelompokan dengan analisis diskriminan ini terjadi karena ada pengaruh satu atau lebih variabel lain yang merupakan variabel independen.
2. apa keunggulan decision tree?
Memahami kasus dan seluruh aspek yang terkait. Menggambarkan kerangka berfikir yang sistematis. Menggambarkan struktur pengambilan keputusan yang dilakukan desicion maker sepanjang tahapan atau urutan waktu termasuk seluruh kemungkinan keputusan dan outcome. 1) Mudah dibaca dan ditafsirkan tanpa perlu pengetahuan statistik; 2) Mudah disiapkan tanpa harus menghitung dengan perhitungan yang rumit; 3) Proses Data Cleaning cenderung lebih rapih

Penjelasan Dataset

Training set harus digunakan untuk membuat model machine learning Anda. Untuk the training set, kami memberikan hasil (juga dikenal sebagai "ground truth) untuk setiap penumpang. Model Anda akan didasarkan pada "fitur" seperti jenis kelamin dan kelas penumpang. Anda juga dapat menggunakan rekayasa fitur untuk membuat fitur baru.

Test set harus digunakan untuk melihat seberapa baik performa model Anda pada data yang tidak terlihat. Untuk test set, kami tidak memberikan ground truth untuk setiap penumpang. itu tugas Anda untuk memprediksi hasil ini. Untuk setiap penumpang dalam test set, gunakan model yang Anda latih untuk memprediksi apakah mereka selamat atau tidak tenggelamnya Titanic.

Kami juga menyertakan gender_submission.csv, sekumpulan prediksi yang mengasumsikan semua dan hanya penumpang wanita yang selamat, sebagai contoh bagaimana seharusnya tampilan file pengiriman.

Data Dictionary

VariableDefinitionKey
Survival 0 = No, 1 = Yes,
pclass Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd,
Sex,
Age in years,
sibsp # of siblings / spouses aboard,
the Titanic parch # of parents / children aboard
the Titanic Ticket number fare Passenger

Source : https://www.kaggle.com/c/titanic/data?select=train.csv

LAB SETUP

Hal yang harus disiapkan dan dilakukan oleh praktikan untuk menjalankan praktikum modul ini.

1. Menginstall library yang dibutuhkan untuk mengerjakan modul.
2. Menjalankan Python.

ELEMEN KOMPETENSI I

Deskripsi:
Memahami proses analisis group means

Kompetensi Dasar:
Mampu Lakukan analisis group means berdasarkan nilai survival sebagai grouping variable pada Python

Latihan 1.1.1

Penjelasan Singkat :
Pada latihan ini anda akan diminta untuk Lakukan analisis group means berdasarkan nilai survival sebagai grouping variable.

Langkah-Langkah Praktikum:

Target (class) : *Survival (1=survived; 0= not survived)*
Data :*titanic.csv,*

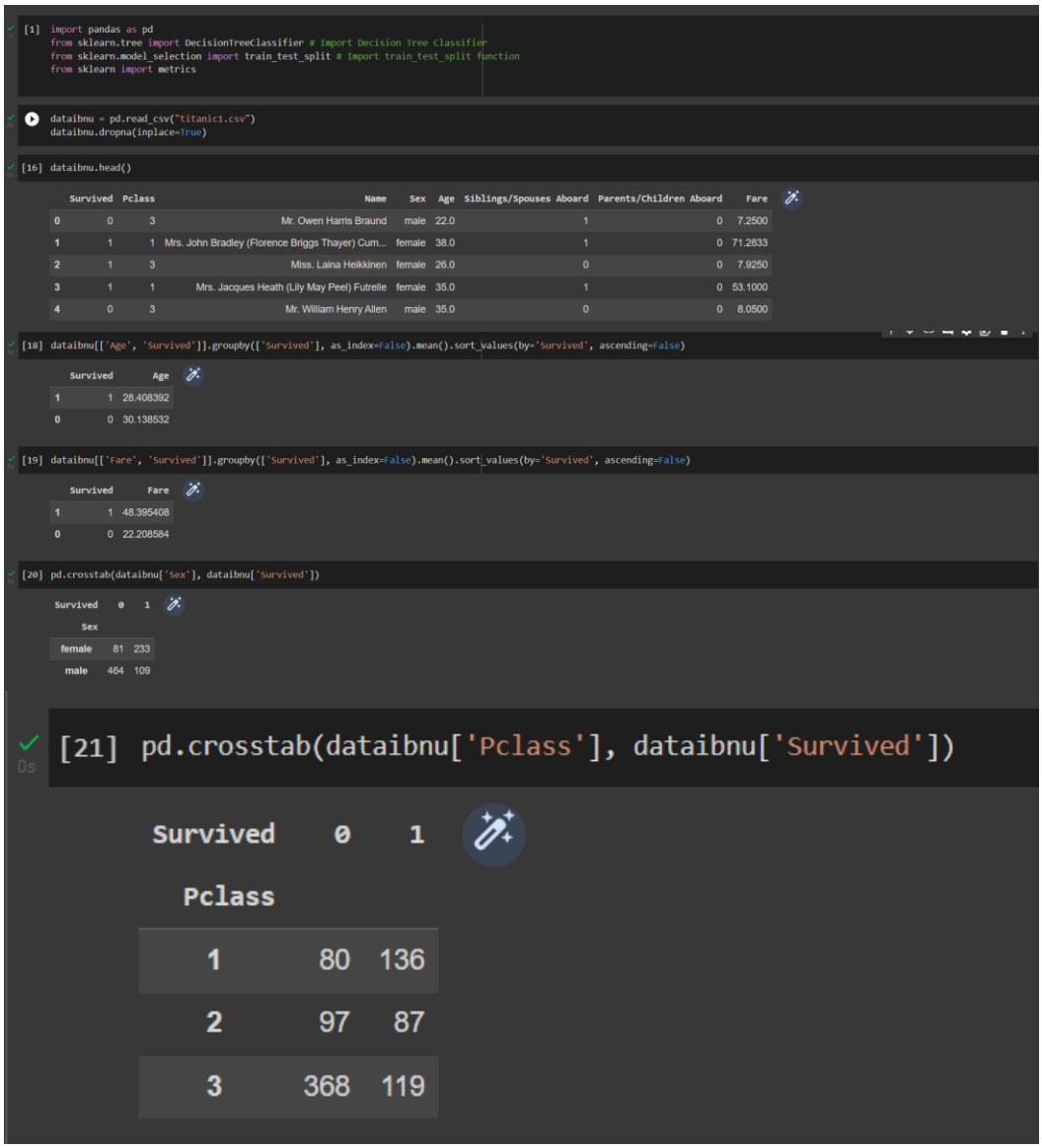
```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier # Import Decisio
n Tree Classifier
from sklearn.model_selection import train_test_split # Import tr
ain_test_split function
from sklearn import metrics
```

```
data_namapraktikan = pd.read_csv("titanic.csv")
data_namapraktikan.head()

data_namapraktikan.dropna(inplace=True)
```

```
data_namapraktikan
[['Age', 'Survived']].groupby(['Survived'], as_index=False).mean
().sort_values(by='Survived', ascending=False)
```

```
data_namapraktikan
[['Fare', 'Survived']].groupby(['Survived'], as_index=False).mea
n().sort_values(by='Survived', ascending=False)
```

```
pd.crosstab(data['Sex'], data['Survived'])


pd.crosstab(data['Pclass'], data['Survived'])
```



```
[1] import pandas as pd
    from sklearn.tree import DecisionTreeClassifier # Import Decision Tree Classifier
    from sklearn.model_selection import train_test_split # Import train_test_split function
    from sklearn import metrics

    dataibnu = pd.read_csv("titanic1.csv")
    dataibnu.dropna(inplace=True)

[16] dataibnu.head()
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250
3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.0	1	0	53.1000
4	0	3	Mr. William Henry Allen	male	35.0	0	0	8.0500

```
[18] dataibnu[['Age', 'Survived']].groupby(['Survived'], as_index=False).mean().sort_values(by='Survived', ascending=False)
```

	Survived	Age
1	1	28.408392
0	0	30.138532

```
[19] dataibnu[['Fare', 'Survived']].groupby(['Survived'], as_index=False).mean().sort_values(by='Survived', ascending=False)
```

	Survived	Fare
1	1	48.395408
0	0	22.208584

```
[20] pd.crosstab(dataibnu['Sex'], dataibnu['Survived'])
```

Survived	0	1
Sex		
female	81	233
male	464	109

```
[21] pd.crosstab(dataibnu['Pclass'], dataibnu['Survived'])
```

Survived	0	1
Pclass		
1	80	136
2	97	87
3	368	119

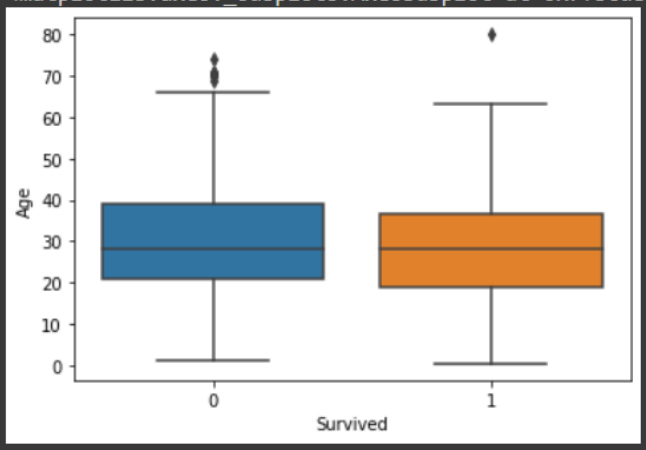
Lakukan analisis boxplot untuk setiap variable predictor yang bersifat numerik. Jelaskan maknanya

Boxplot bertujuan untuk melakukan analisis dari dua data yang disatukan, bagaimana dari kedua data itu bisa kita temukan hasil yang ingin kita lihat, sepertii dalam kasus titanic ini kita lihat umur dari seluruh penumpang dan apakah mereka selamat atau tidak, begitu juga dengan fare(tarif) apakah mempengaruhi keselamatan dari penumpang.

```
import seaborn as sns
sns.boxplot(x ='Survived', y ='Age', data = data_namapraktikan)


sns.boxplot(x ='Survived', y ='Fare', data = data_namapraktikan)
```

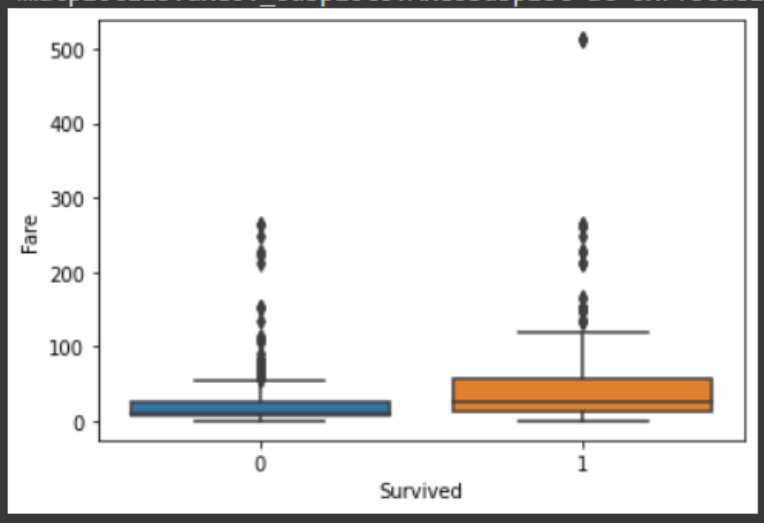
```
import seaborn as sns
sns.boxplot(x ='Survived', y ='Age', data = dataibnu)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcd8265a90>
```




```
[23] sns.boxplot(x ='Survived', y ='Fare', data = dataibnu)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcd81641d0>
```



ELEMEN KOMPETENSI II

Deskripsi:
Menerapkan decision tree untuk membangun model

Kompetensi Dasar:
Mampu melakukan pemodelan prediksi dengan decision tree

Latihan 1.2.1

Penjelasan Singkat :
Pada latihan ini anda akan diminta untuk Lakukan analisis group means berdasarkan nilai survival sebagai grouping variable untuk pemodelan prediksi dengan decision tree.

Langkah-Langkah Praktikum:

Buatlah decision tree menggunakan data training untuk membangun model yang dapat digunakan untuk memprediksi kelas survive.

```
!pip install six

import pandas as pd
from sklearn.tree import DecisionTreeClassifier # Import Decis
ion Tree Classifier
from sklearn.model_selection import train_test_split # Import
train_test_split function
from sklearn import metrics

data = pd.read_csv("titanic.csv")
data.head()

data.dropna(inplace=True)
data.replace({"male":0, "female":1}, inplace=True)
```

```
feature_cols = ['Sex', 'Age', 'Fare', 'Pclass']
x = data[feature_cols]
y = data.Survived
```

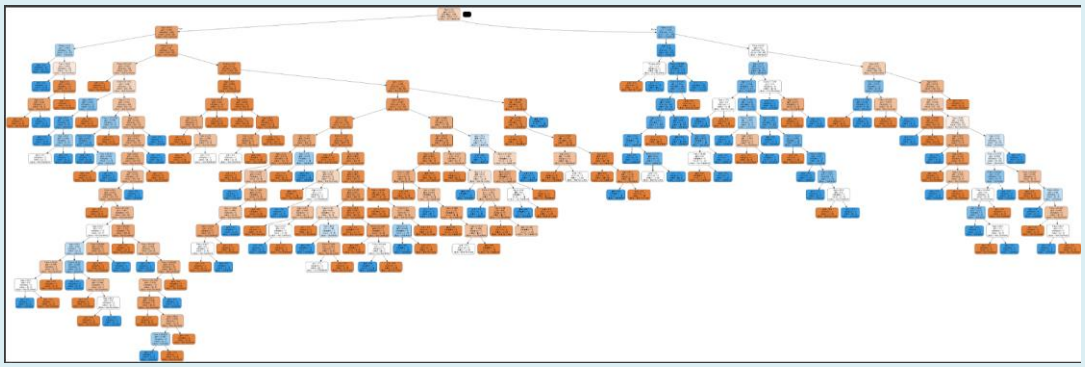
```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_s
ize=0.3, random_state=1)
```

```
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

```
feature_cols_new = X_train.columns.values.tolist()
```

```
from sklearn.tree import export_graphviz
from six import StringIO
from IPython.display import Image
import pydotplus

dot_data = StringIO()
export_graphviz(clf, out_file=dot_data,
                filled=True, rounded=True,
                special_characters=True,feature_names = feature_
cols_new,class_names=['Not Survived','Survived'])
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
Image(graph.create_png())
```

Output :



```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
pd.crosstab(y_test, y_pred)
```

Output :

```
Accuracy: 0.7640449438202247
col_0       0   1
Survived
0         126  36
1          27  78
```

CEK LIST

Elemen Kompetensi	No Latihan	Penyelesaian	
		Selesai	Tidak selesai
1	1.1.1	✓	
2	1.2.1	✓	

FORM UMPAN BALIK

Elemen Kompetensi	**Tingkat Kesulitan**	**Tingkat Ketertarikan**	**Waktu Penyelesaian dalam menit**
Memahami proses analisis group means	☐ Sangat Mudah ☐ Mudah ☐ ✓ Biasa ☐ Sulit ☐ Sangat Sulit	☐ Tidak Tertarik ☐ Cukup Tertarik ☐ Tertarik ☐ ✓ Sangat Tertarik	20
Menerapkan decision tree untuk membangun model	☐ Sangat Mudah ☐ Mudah ☐ ✓ Biasa ☐ Sulit ☐ Sangat Sulit	☐ Tidak Tertarik ☐ Cukup Tertarik ☐ Tertarik ☐ ✓ Sangat Tertarik	20