Praktikum Data Preprocessing

Praktikum Decision Tree

Ulima Inas Shabrina(2110181048)

Latihan 6 - #Assignment
(Decision Tree dengan nilai numerik)

1. dataset ← titanic.csv
2. test_dataset ← titanic_test.csv
3. train_data ← ambil dataset kolom fitur (Sex, Age, Pclass, Fare). Lakukan pengisian missing value pada fitur Age dengan nilai mean dari masing-masing class
4. test_data ← ambil test_dataset kolom fitur (Sex, Age, Pclass, Fare).
5. train_label ← ambil dataset kolom kelas (Survived)
6. test_label ← titanic_testlabel.csv
7. Lakukan klasifikasi test_data terhadap train_data dengan Decision Tree, dan berapakah error rationya?
8. Tampilkan hirarki dari Decision Tree

Soal 1. Menampilkan data dari titanic.csv

In [1]:
```
import pandas as pd

dataset = pd.read_csv('titanic.csv')
dataset
```

Out[1]:

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Soal 2. Menampilkan data dari titanic_test.csv

In [2]:
```
test_dataset = pd.read_csv('titanic_test.csv')
test_dataset
```

Out[2]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

418 rows × 11 columns

Soal 3. Mengambil dataset kolom fitur (Sex, Age, Pclass, Fare) dan Melakukan pengisian missing value pada fitur Age dengan nilai mean dari masing-masing class

In [11]:
```
train_data = pd.DataFrame(dataset, columns = ['Sex', 'Age', 'Pclass', 'Fare'])
male_data = train_data['Sex'] == 'male'
female_data = train_data['Sex'] == 'female'
average_male = train_data[male_data]['Age'].mean()
average_female = train_data[female_data]['Age'].mean()
train_data.loc[male_data, 'Age'] = train_data.loc[male_data, 'Age'].fillna(average_male)
train_data.loc[female_data, 'Age'] = train_data.loc[female_data, 'Age'].fillna(average_female)

print('Rata - rata umur Male -->', average_male)
print('Rata - rata umur Female -->',average_female)
print('\n\nTabel')
print(train_data)
```

```
Rata - rata umur Male --> 30.72664459161148
Rata - rata umur Female --> 27.915708812260537


Tabel
        Sex        Age  Pclass     Fare
0      male  22.000000       3   7.2500
1    female  38.000000       1  71.2833
2    female  26.000000       3   7.9250
3    female  35.000000       1  53.1000
4      male  35.000000       3   8.0500
..      ...        ...     ...      ...
886    male  27.000000       2  13.0000
887  female  19.000000       1  30.0000
888  female  27.915709       3  23.4500
889    male  26.000000       1  30.0000
890    male  32.000000       3   7.7500

[891 rows x 4 columns]
```

Soal 4. Mengambil test_dataset kolom fitur (Sex, Age, Pclass, Fare) dan menghilangkan baris data yang terdapat missing values

In [4]:
```
test_data = test_dataset[['Sex', 'Age', 'Pclass', 'Fare']].dropna()
test_data
```

Out[4]:

	Sex	Age	Pclass	Fare
0	male	34.5	3	7.8292
1	female	47.0	3	7.0000
2	male	62.0	2	9.6875
3	male	27.0	3	8.6625
4	female	22.0	3	12.2875
...	...	...	...	...
409	female	3.0	3	13.7750
411	female	37.0	1	90.0000
412	female	28.0	3	7.7750
414	female	39.0	1	108.9000
415	male	38.5	3	7.2500

331 rows × 4 columns

Soal 5. Mengambil dataset kolom kelas (Survived)

In [5]:
```
train_label = pd.DataFrame(dataset, columns = ['Survived'])
train_label
```

Out[5]:

	Survived
0	0
1	1
2	1
3	1
4	0
...	...
886	0
887	1
888	0
889	1
890	0

891 rows × 1 columns

Soal 6. Menampilkan data dari titanic_testlabel.csv

In [6]:
```
test_label = pd.read_csv('titanic_testlabel.csv')
test_label
```

Out[6]:

	PassengerId	Survived
0	892	0
1	893	1
2	894	0
3	895	0
4	896	1
...	...	...
413	1305	0
414	1306	1
415	1307	0
416	1308	0
417	1309	0

418 rows × 2 columns

Soal 7. Melakukan klasifikasi test_data terhadap train_data dengan Decision Tree

In [21]:
```
import pandas as pd
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
import graphviz

tr = DecisionTreeClassifier()
train_data['Sex'] = train_data['Sex'].replace(['male', 'female'], [1.0, 0.0])
test_data['Sex'] = test_data['Sex'].replace(['male', 'female'], [1.0, 0.0])
tr.fit(train_data, train_label)
class_result = tr.predict(test_data)

accuration = tr.score(train_data, train_label)
error_ratio = round((1-accuration)*100,2)
accuration = round((accuration)*100,2)

print('Score --> ', accuration, '%')
print('\nError ratio --> ', error_ratio, '%')
```

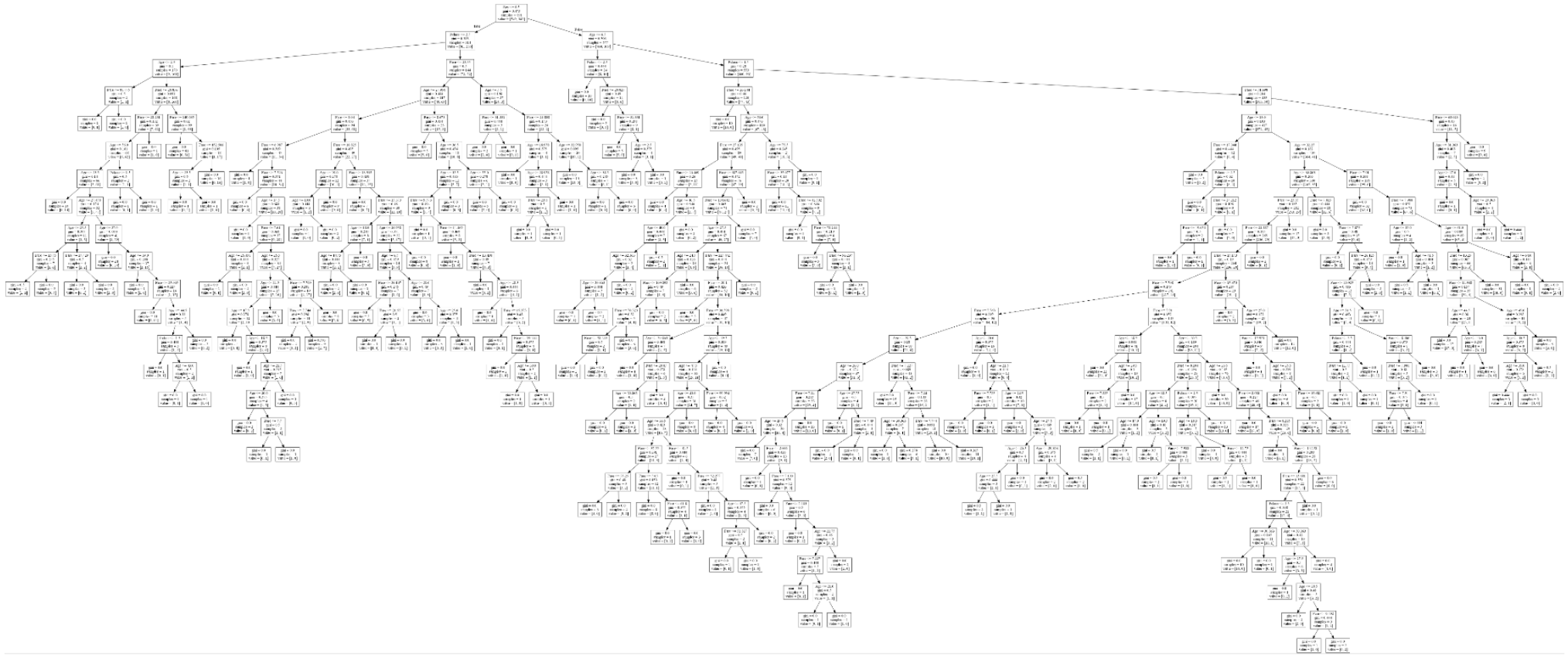
```
Score -->  97.98 %

Error ratio -->  2.02 %
```

Soal 8. Menampilkan hirarki dari Decision Tree

In [19]:
```
import os
os.environ["PATH"] += os.pathsep + 'C:\Program Files\Graphviz\bin'

dot_data = tree.export_graphviz(tr, out_file=None, feature_names=train_data.columns.values)
graph = graphviz.Source(dot_data, format="png")
graph.render(view=True)
```



Terimakasih