Практична робота №6

Тема: КЛАСИФІКАЦІЯ ТА КЛАСТЕРИЗАЦІЯ.

Мета: ознайомитись з

- методами класифікації та кластеризації і навчитися, що використовують дерева прийняття рішень і інструментами фактичного аналізу методами головних компонент та методами найбільшої подібності.

Виконав: ІП-13 Ах Хадам М.Р.

Перевірила: Лінуссова Т.А.

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
pd.options.display.float_format = '{:.2f}'.format
```

Основне завдання

1. Скачати потрібні дані.

Завантажимо та обробимо дані

In [2]:
```
df = pd.read_csv("data/titanic.csv")
df.isnull().sum()
```

Out[2]:
```
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
```

Побудуємо теплову карту для пропущених значень кожного з стовпців

In [3]:
```
sns.heatmap(df.isnull(), cbar=False)
```

Out[3]:
```
<Axes: >
```

Побудуємо гістограми усіх важливих пасажирів для кожного з стовпців

In [4]:
```
df[df['Survived'] == 1].hist(figsize=(8, 8))
plt.show()
df[df['Survived'] == 1]['Embarked'].hist()
```

Заповнимо пропущені дані відповідними характеристиками

In [5]:
```
df = df.fillna(df.mean(numeric_only=True))
df['Cabin'] = df['Cabin'].fillna('')
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode[0])
df = df.set_index('PassengerId')

# for corr matrix (basic & fancy) 1)
df['Sex'] = df['Sex'].replace({'male': 0, 'female': 1})

df.isnull().sum()
```

Out[5]:
```
Survived    0
Pclass      0
Name        0
Sex         0
Age         0
SibSp       0
Parch       0
Ticket      0
Fare        0
Cabin       0
Embarked    0
dtype: int64
```

In [6]:
```
df
```

Out[6]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
PassengerId											
1	0	3	Braund, Mr. Owen Harris	0	22.00	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.00	1	0	PC 17599	71.28	C85	C
3	1	3	Heikkinen, Miss. Laina	1	26.00	0	0	STON/O2. 3101282	7.92		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.00	1	0	113803	53.10	C123	S
5	0	3	Allen, Mr. William Henry	0	35.00	0	0	373450	8.05		S
...	...	...	...	...	...	...	...	...	...	...	...
887	0	2	Montvila, Rev. Juozas	0	27.00	0	0	211536	13.00		S
888	1	1	Graham, Miss. Margaret Edith	1	19.00	0	0	112053	30.00	B42	S
889	0	3	Johnston, Miss. Catherine Helen "Carie"	1	29.70	1	2	W./C. 6607	23.45		S
890	1	1	Behr, Mr. Karl Howell	0	26.00	0	0	111369	30.00	C148	C
891	0	3	Dooley, Mr. Patrick	0	32.00	0	0	370376	7.75		Q

891 rows × 11 columns

Отже, дослідні при первинній обробленній, пусті значення були заповнені

2. Для даних по титаніку titanic.csv побудувати модель, в якій можна визначити, чи вижиле пасажир, заповнивши решту параметрів.

Побудуємо кореляційну матрицю та діагрему розсіювання

Видалимо стовп стаття

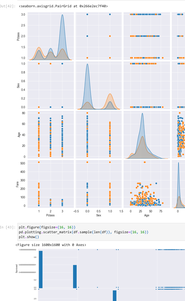
In [7]:
```
df = df.drop(['Name', 'SibSp', 'Parch', 'Ticket', 'Cabin', 'Embarked'], axis=1)
df
```

Out[7]:

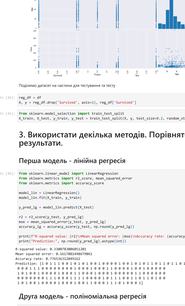
	Survived	Pclass	Sex	Age	Fare
PassengerId					
1	0	3	0	22.00	7.25
2	1	1	1	38.00	71.28
3	1	3	1	26.00	7.92
4	1	1	1	35.00	53.10
5	0	3	0	35.00	8.05
...	...	...	...	...	...
887	0	2	0	27.00	13.00
888	1	1	1	19.00	30.00
889	0	3	1	29.70	23.45
890	1	1	0	26.00	30.00
891	0	3	0	32.00	7.75

891 rows × 5 columns

In [8]:
```
fig, ax = plt.subplots()
pearson_corr = df.corr(method='pearson')

ax.set_title('Correlation matrix for Pearson method')
sns.heatmap(pearson_corr, annot=True)

plt.show()
```



In [9]:
```
fig, ax = plt.subplots()
spearman_corr = df.corr(method='spearman')

ax.set_title('Correlation matrix for Spearman method')
sns.heatmap(spearman_corr, annot=True)

plt.show()
```



In [10]:
```
sns.pairplot(df, hue='Survived')
```

Out[10]:
```
<seaborn.axisgrid.PairGrid at 0x26a1ec7f40>
```

In [11]:
```
plt.figure(figsize=(8, 16))
sns.boxplot(x='Sex', y='Survived', data=df)
plt.show()
```


In [12]:
```
reg_df = df
X, y = reg_df.drop('Survived', axis=1), reg_df['Survived']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

3. Використати декілька методів. Порівняти результати.

Перша модель - лінійна регресія

In [13]:
```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error
from sklearn.metrics import mean_absolute_error

model_lin = LinearRegression()
model_lin.fit(X_train, y_train)

y_pred_lg = model_lin.predict(X_test)

rl = r2_score(y_test, y_pred_lg)
mse = mean_squared_error(y_test, y_pred_lg)
accuracy_lg = accuracy_score(y_test, np.round(y_pred_lg))

print("Accuracy rate: {accuracy_lg}".format(accuracy_lg=accuracy_lg))
print("Predictions", np.round(y_pred_lg), np.asfarray(y))
```

```
Accuracy rate: 0.7988763066202091
Predictions [1.0 1.0 0.0 1.0 0.0 0.0 0.0 0.0 1.0 1.0 1.0 1.0 0.0 0.0 0.0 1.0 0.0 1.0
 0.0 1.0 1.0 0.0 1.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 0.0 0.0 1.0 0.0 1.0 0.0
 1.0 0.0 0.0 0.0 0.0 0.0 1.0 1.0 1.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
 0.0 0.0 0.0 1.0 1.0 0.0 1.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0
```

```
16-1461
```