МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут інформаційних технологій

Кафедра інженерії програмного забезпечення

**ЛАБОРАТОРНИЙ ПРАКТИКУМ 1.**

Підготував: студент групи ПДМ-51

Гапей Максим Юрійович

Перевірив: викладач

Садовенко Володимир Сергійович

Київ 2024

**Мета роботи:** створити модель з кластеризацією та без за допомогою Birch. Провести нормалізацію датасету.

**Частина 1**

Import the required libraries/packages. Scikit Learn provides the module for direct implementation of BIRCH under the cluster class packages:

import pandas as pd

pd.set\_option('display.max\_rows', None)

import numpy as np

from matplotlib import pyplot as plt

import seaborn as sns

sns.set()

from sklearn.datasets import make\_blobs

from sklearn.cluster import Birch

Read the data thru the pd.read\_csv into wine\_data. Determine the dimensions using shape and the result contains 63 rows (one for each customer) and three columns (the name of the customer, their purchases of Pinot Noir, and their purchases of Champagne):

wine\_data = pd.read\_csv("wine\_data.csv")

wine\_data.head()

print(wine\_data.shape)

wine\_data.describe()

X = wine\_data.iloc[:,[1,2]].values

print(X.shape)

wine\_data.info()

(63, 3)

(63, 2)

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 63 entries, 0 to 62

Data columns (total 3 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 name 63 non-null object

1 Pinot Noir 63 non-null int64

2 Champagne 63 non-null int64

dtypes: int64(2), object(1)

memory usage: 1.6+ KB

Create a scatter plot with data. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables:

plt.figure(figsize=(10,6))

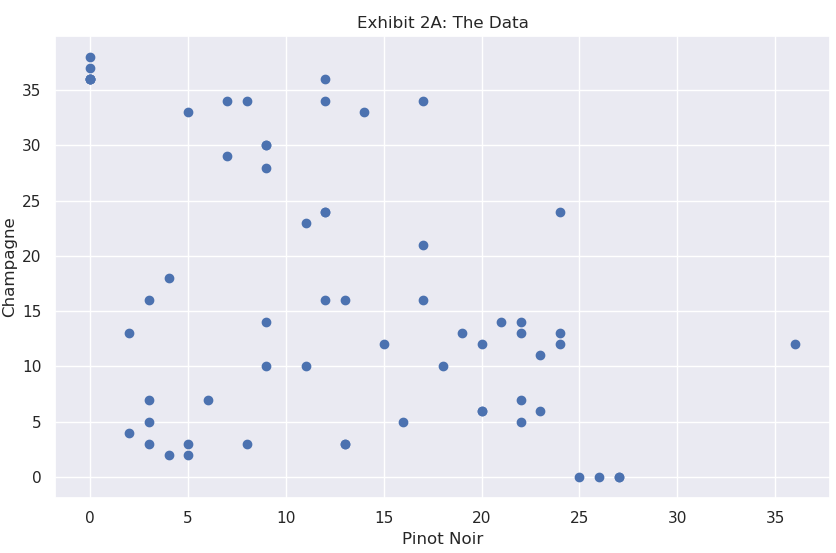
plt.scatter(wine\_data['Pinot Noir'], wine\_data['Champagne'])

plt.title("Exhibit 2A: The Data")

plt.xlabel("Pinot Noir")

plt.ylabel("Champagne")

plt.show()



Determine the clusters. Dendrogram is used to determine the clusters and thru it we are able to see that there are 3 possible clusters. The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold.

import scipy.cluster.hierarchy as sch

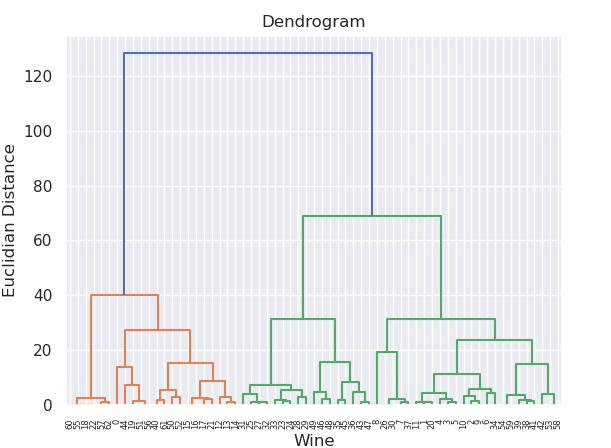
dendrogram = sch.dendrogram(sch.linkage(X,method='ward'))

plt.title("Dendrogram")

plt.xlabel("Wine")

plt.ylabel("Euclidian Distance")

plt.show()

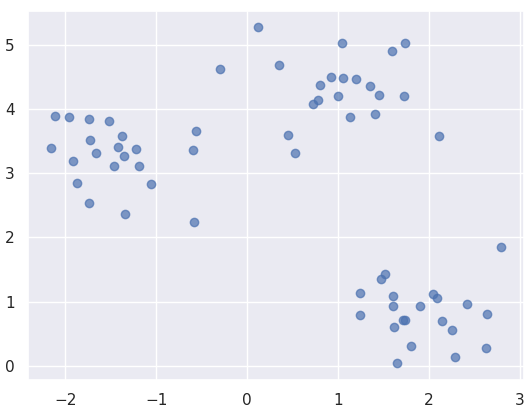
**

Generating a dataset using make blobs. We use *scikit-learn* to generate data with nicely defined clusters.

X, clusters = make\_blobs(n\_samples=63, centers=3, cluster\_std=0.50, random\_state=0)

plt.scatter(X[:,0], X[:,1], alpha=0.7, edgecolor='b')

plt.show()



Next is to initialize and train the model.

brc = Birch(branching\_factor=50, n\_clusters=3, threshold=1.5)

brc.fit(X)

Birch(threshold=1.5)

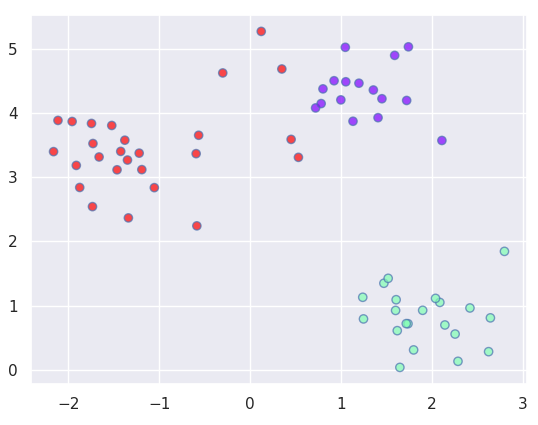
Predict method to obtain a list of points and their respective cluster.

labels = brc.predict(X)

Plot the data points using a different color for each cluster.

plt.scatter(X[:,0], X[:,1], c=labels, cmap='rainbow', alpha=0.7, edgecolors='b')

plt.show()

****

BIRCH Model without clusters determined. Create a BIRCH model then fit the data set in the model.

model = Birch(branching\_factor=50, n\_clusters=None, threshold=1.5)

model.fit(X)

Birch(n\_clusters=None, threshold=1.5)

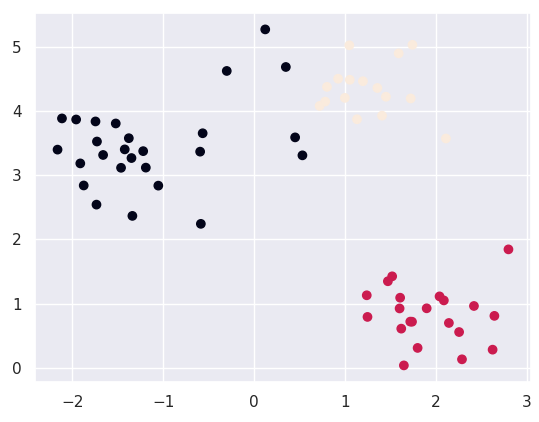
Create a prediction of the dataset using the generated model.

pred = model.predict(X)

Making the scatterplot to check the results.

plt.scatter(X[:,0], X[:,1], c=pred)

plt.show()



**Частина 2**

from sklearn.preprocessing import StandardScaler

import pandas as pd

data = pd.read\_csv('dataset.csv')

num\_features = ['InstallmentCredit', 'Yearly Period', 'Class']

for feature in num\_features:

data[feature] = data[feature].apply(lambda label: 0 if label=='?' else label)

X = data[num\_features]

scaler = StandardScaler()

X\_normalized = scaler.fit\_transform(X)

print(X\_normalized)

[[-0.33623128 -1.11677851 1.31508711]

[-0.88904815 -0.97504435 -1.15953917]

[ 0.00549133 -0.15119038 -1.15953917]

...

[-0.89200006 -1.68157698 nan]

[-0.75124352 -0.80235046 nan]

[-0.1505662 -0.26312192 nan]]

**Висновки:** Під час виконання практичної роботи, було створено модель з кластеризацією та без за допомогою Birch. Проведено нормалізацію датасету.