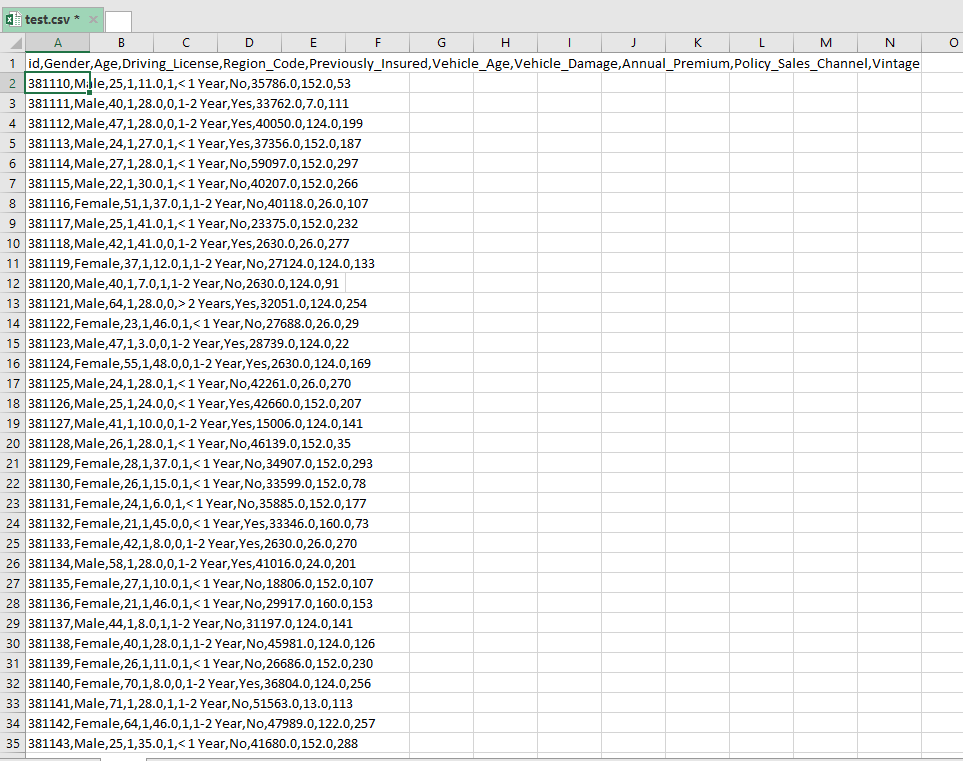
Застосування нечіткої кластеризації C-середніх для аналізу даних на мові Python.

Файл із вхідними даними:



**Хід роботи**

Для використання алгоритму нечіткої кластеризації потрібна множина точок, що утворюють між собою скупчення в умовному просторі. Для того, щоб була можливість зобразити результати графічно та інтерпретувати їх, обмежимось двовимірним простором, на якому утворені кластери можна виділити неозброєним поглядом.

Дана таблиця даних із полями: Id, Стать, Вік, Наявність ліцензії, Код регіону і т. д. Для показників, що можуть бути логічно пов’язані та представлені чисельно, застосуємо наступний алгоритм:

1. Імпорт бібліотек

import numpy as np  
from fcmeans import FCM  
from matplotlib import pyplot as plt  
import csv

2. Створюємо порожній масив для попарних ознак

X = []

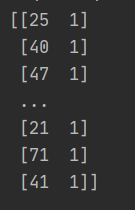
3. Зчитуємо csv-файл і виділяємо дві ознаки, що нас цікавлять. В даному випадку ми з’ясовуємо, як пов’язані вік людини та наявність у неї ліцензії водія. Обрані ознаки приєднуються до списку X у форматі [a, b]

with open('test.csv') as File:  
 reader = csv.DictReader(File)  
 for row in reader:  
 X.append([int(row['Age']), int(row['Driving\_License'])])

4. Приводимо отриманий масив до формату numpy.array

X=np.array(X)  
print(X)

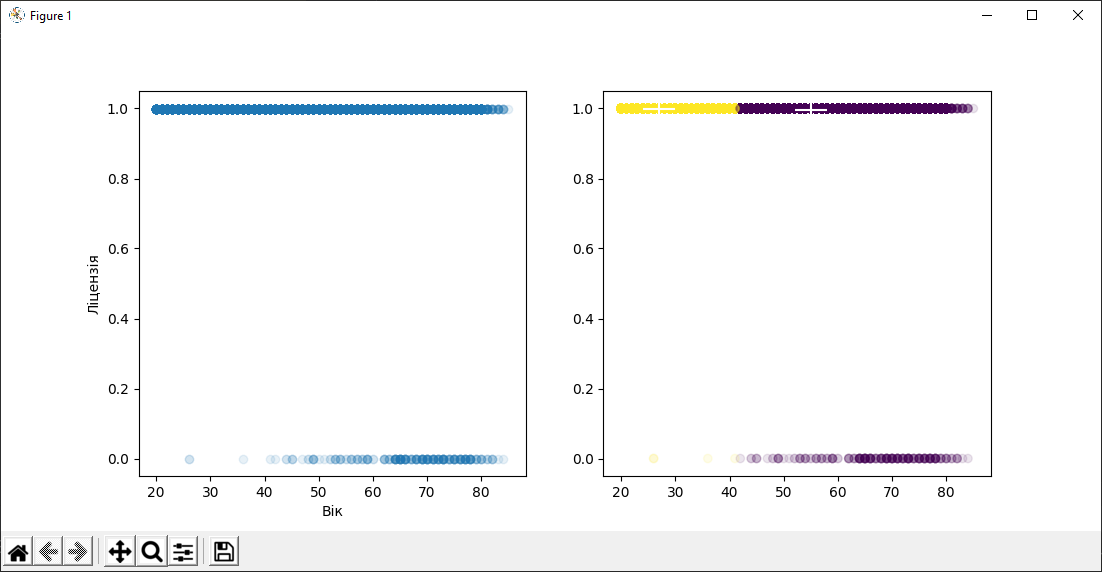
Його вигляд:



5. З бібліотекою fcmeans навчаємо модель та виводимо результат:

fcm = FCM(n\_clusters=2)  
fcm.fit(X)  
  
# outputs  
fcm\_centers = fcm.centers  
fcm\_labels = fcm.predict(X)  
  
# plot result  
f, axes = plt.subplots(1, 2, figsize=(11, 5))  
axes[0].scatter(X[:, 0], X[:, 1], alpha=.1)  
axes[1].scatter(X[:, 0], X[:, 1], c=fcm\_labels, alpha=.1)  
axes[1].scatter(fcm\_centers[:, 0], fcm\_centers[:, 1], marker="+", s=500, c='w')  
  
axes[0].set\_xlabel("Вік")  
axes[0].set\_ylabel("Ліцензія")  
  
plt.savefig('basic-clustering-output.jpg')  
plt.show()

**Результат**:

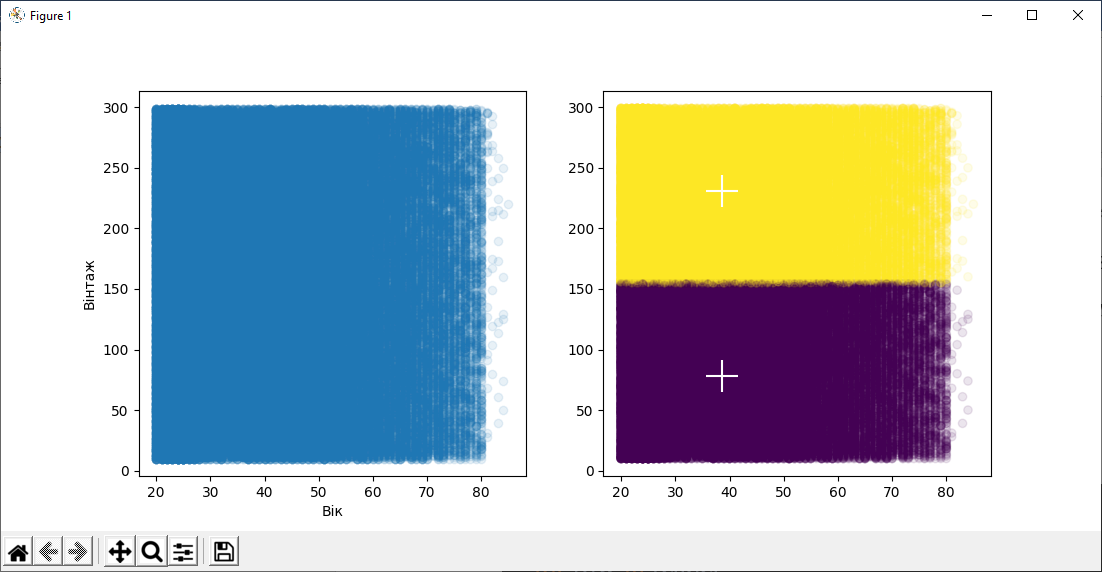


На лівому зображенні бачимо, як розподілились реалізації. Є дві групи: з наявною та відсутньою ліцензією, чим ближче входження до правої сторони, тим більший вік людини. На правому зображенні різні кольори позначають поділ на кластери, а білі плюси – їх центри. Можна побачити, що до першого кластеру віднесено людей до 40 років, майже усі з них мають ліцензію. Другий кластер починається від 40 років і розділений на дві частини: з ліцензією та без. Звідси робимо наступний висновок:

*Люди до 40 років мають посвідчення водія з більшою вірогідністю, ніж старші.*

Застосуємо алгоритм до інших ознак:

with open('test.csv') as File:  
 reader = csv.DictReader(File)  
 for row in reader:  
 X.append([int(row['Age']), float(row['Vintage'])])



Так розподіляється відношення віку людини до ступеня вінтажності її машини. За розміщенням центрів кластерів можна побачити, що густина розподілу реалізацій в них майже однакова. Звідси робимо висновок:

*Вінтажність машини та вік людини не пов’язані*

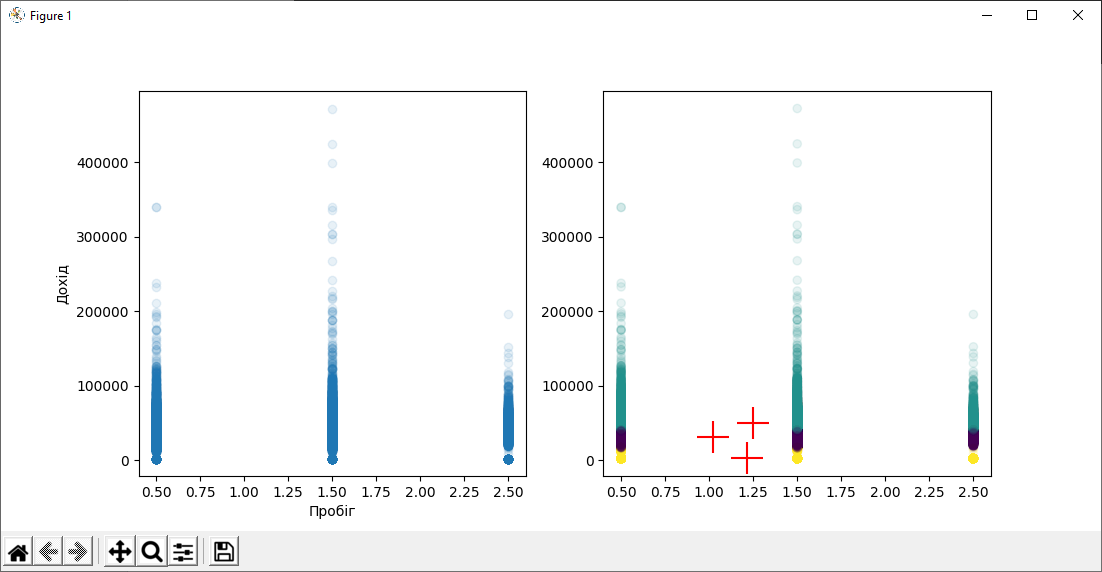
Для наступного прикладу збільшимо кількість кластерів до 3 і розглянемо відношення Щорічного доходу та Пробігу машини.

fcm = FCM(n\_clusters=3)  
fcm.fit(X)

Пробіг переведений з приблизних значень Менше 1 року або 1-2 роки до точних: 0.5, 1.5 і 2.5

with open('test.csv') as File:  
 reader = csv.DictReader(File)  
 for row in reader:  
 v\_age = 0.0  
 if row['Vehicle\_Age'] == '1-2 Year':  
 v\_age = 1.5  
 elif row['Vehicle\_Age'] == '< 1 Year':  
 v\_age = 0.5  
 else:  
 v\_age = 2.5  
 X.append([v\_age, float(row['Annual\_Premium'])])

Результат:



Бачимо, що кластери розподілились не за пробігом машини, а за щорічним доходом. З розміщення центрів кластерів можна бачити наступне:

1) Центр верхнього кластера зміщений до його нижньої границі через те, що густина реалізацій зменшується з висотою, отже:

*Чим вищий дохід, тим менша кількість людей, які його отримують*

2) Центр другого кластера зміщений вліво (до низького пробігу) більше, ніж інші два, отже:

у людей із середнім рівнем доходу пробіг машин менший, ніж у людей із високим та низьким рівнем

3) Якщо поглянути на найвищу реалізацію у кожному стовпці, що відповідає пробігу, то бачимо, що у середньому пробігу найвищий дохід – майже 500 000, у малому пробігу – близько 320 000 і у великому пробігу – менше 200 000, отже:

*Найвищий дохід мають люди із пробігом машини від 1 до 2 років.*

**Висновок**: в ході даної роботи було розглянуто реалізацію методу кластеризації C-середніх у бібліотеці fcmeans мови Python, яку було використано для розбиття бази власників автомобіля за різними ознаками та виявлено статистичні закономірності, представлені у графічному вигляді