Размити множества,
зимен семестър 2023/2024

Курсов проект:
Размита клъстеризация на факторите за щастието в различните държави

Име: Надежда Францева
Фн: 8MI3400357
Специалност: Изкуствен Интелект

1. Въведение в решавания проблем и цел на проекта

Целта на проекта е да се изследва щастието в различните държави по света и да се разбере как различните фактори, като брутен вътрешен продукт (БВП) на глава от населението, социална подкрепа, продължителност на здравословния живот и други, влияят върху него. За тази цел се използва размит клъстерен анализ - Fuzzy C-means, който позволява категоризиране на държавите в групи в зависимост от техните общи характеристики. Програмата има допълнителна задача като разширение на алгоритъма да бъде възможен избор на броя клъстери за групиране.

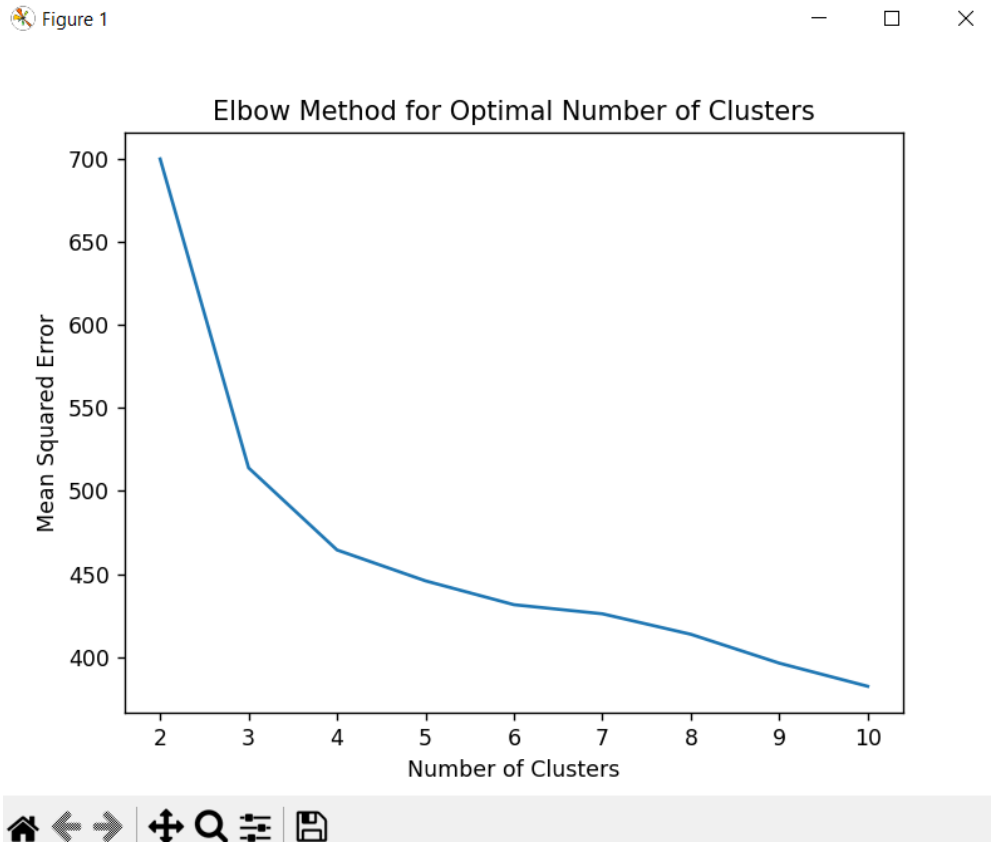
2. Теоретична постановка и описание на използвания алгоритъм

В проекта се използва размит клъстерен алгоритъм, който е метод за категоризиране на данни в групи (клъстери), като в отличие от стандартния клъстерен алгоритъм, позволява елементите да принадлежат към повече от един клъстер с определена вероятност. За целта се използва - Fuzzy C-means, за да изчислява степента на принадлежност на всеки запис към всеки отделен клъстерен център със стойност между 0 и 1. Той е особено полезен, когато точките не са ясно дефинирани към един клъстер и когато искаме да имаме идея за степента на принадлежност на всяка точка към всеки клъстер. Основните стъпки на Fuzzy C-means алгоритъма:

1. Избор на брой клъстери (брой на центровете на клъстерите), брой на итерациите и параметърът, който контролира степента на размиване (m).
2. Инициализация на центровете на клъстерите. Обикновено се прави чрез случайно избрани точки от данните.
3. За всяка точка се изчислява степента на принадлежност към всеки клъстер, използвайки формулата на Евклидово разстояние и параметъра на размиване (m). Тези степени на принадлежност се изчисляват итеративно.
4. Изчисляват се новите центрове на клъстерите, като се използват степените на принадлежност.
5. Продължава се итеративния процес от стъпки 3 и 4, докато критериите за спиране са изпълнени (например, когато се достигне максимален брой итерации).

Алгоритъмът се опитва да минимизира общата грешка, която е сумата на квадратите на разстоянията от всеки обект до всеки клъстер с теглата, дадени от степените на принадлежност. Той се основава на идеята, че всяка точка има степен на принадлежност към всеки клъстер вместо да бъде стриктно асоциирана с един клъстер, както при стандартния k-means алгоритъм.

Тези изчисления се извършват в функцията fuzzy_kmeans (X, n_clusters, m, max_iter=100), която приема като данни факторите за клъстериране – имената на колоните, броя клъстери, броя тестове, максималния брой итерации (или обновления) при изпълнение на алгоритъма за кластеризация. Итерацията привършва работа, когато се достигне условие за терминиране (при нашия случай имаме фиксиран брой итерации - 100). За избора на подходяща бройка клъстери използваме метода Elbow, които начертава сумата от квадрати спрямо броя на клъстерите и след това търсим точката, където скоростта на намаляване рязко се променя. Тази точка се нарича точка "лакът", което показва оптималния брой клъстери – в нашия случай 4:



Фигура 1. Анализ за избор на брой клъстери с метод на Лакътя

При изобразяването на клъстерите в двумерното пространство се взимат две по две от характеристиките на нашите данни.

3. Описание на данните и предпроцесна обработка

Използваните данни са извлечени от платформата Kaggle:

https://www.kaggle.com/datasets/unsdsn/world-happiness

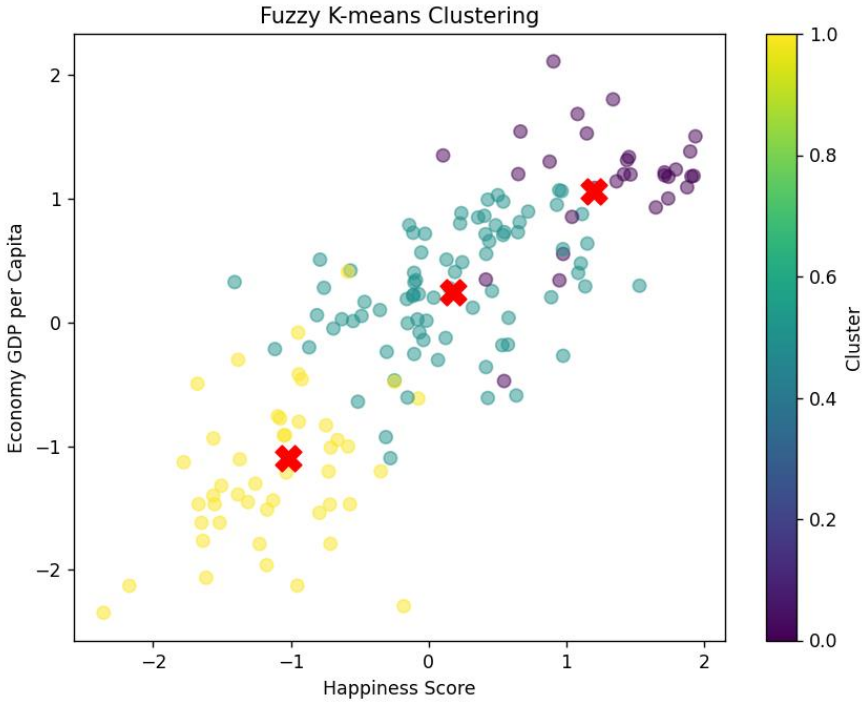
Данните, които се използват в проекта, представляват данни за щастието в различни държави през 5 години: 2015, 2016, 2017, 2018, 2019 година. Те са записани във файл със .csv формат. Има около 160 записа (instances) с по 12 характеристики (features) за всяка от петте години. В тях се съдържат следните показатели:

- Country - държава
- Happiness Rank - ранг на страната въз основа на резултата за щастие
- Happiness Score - показател, измерен през съответната година, като на хората от извадката беше зададен въпросът: „Как бихте оценили щастието си по скала от 0 до 10, където 10 е най-много щастие ?“.
- Lower Confidence Interval - нисък интервал на доверие
- Upper Confidence Interval - горен интервал на доверие
- Economy (GDP per Capita) - икономика (БВП на глава от населението) - степента, в която БВП допринася за изчисляването на оценката за щастие
- Family - степента, в която семейството допринася за изчисляването на резултата за щастие
- Health (Life Expectating) - здраве (очаквана продължителност на живота) - степента, в която очакваната продължителност на живота е допринесла за изчисляването на резултата за щастие
- Freedom – свобода - степента, в която свободата е допринесла за изчисляването на резултата за щастие
- Generosity – щедрост - степента, до която щедростта е допринесла за изчисляването на резултата за щастие
- Trust (Government Corruption) - доверие (правителствена корупция) - степента, в която възприятието за корупция допринася за резултата за щастие
- Dystopia Residual - остатък от дистопия - степента, в която остатъкът от дистопията е допринесла за изчисляването на резултата за щастие

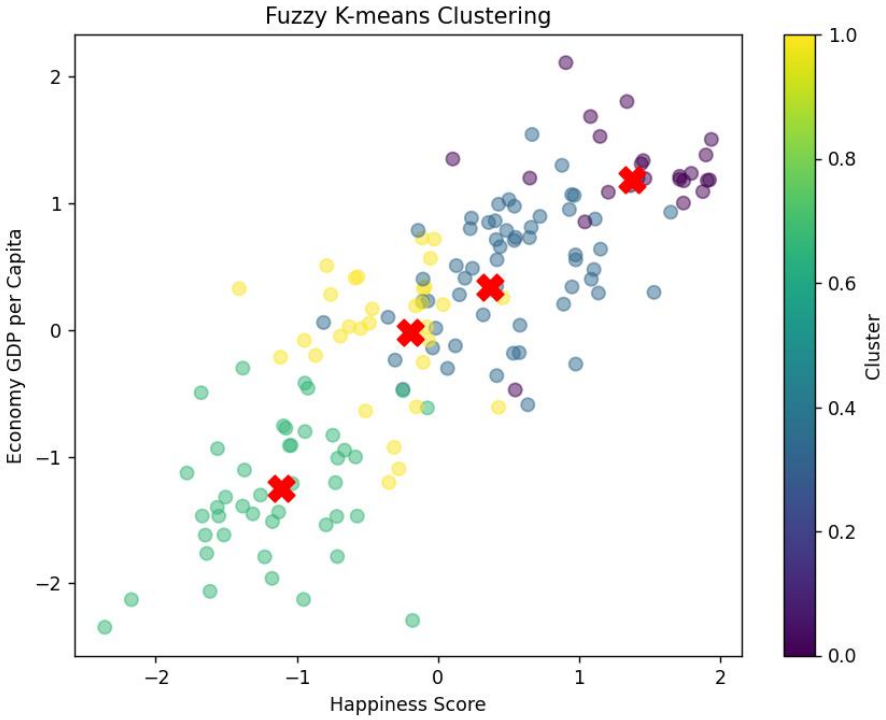
Стойностите на характеристиките за конкретен запис са числа (integer). Предпроцесната обработка на данните включва изчистване на липсващите стойности и скалиране на данните за подготовка за анализ.

4. Експериментални резултати

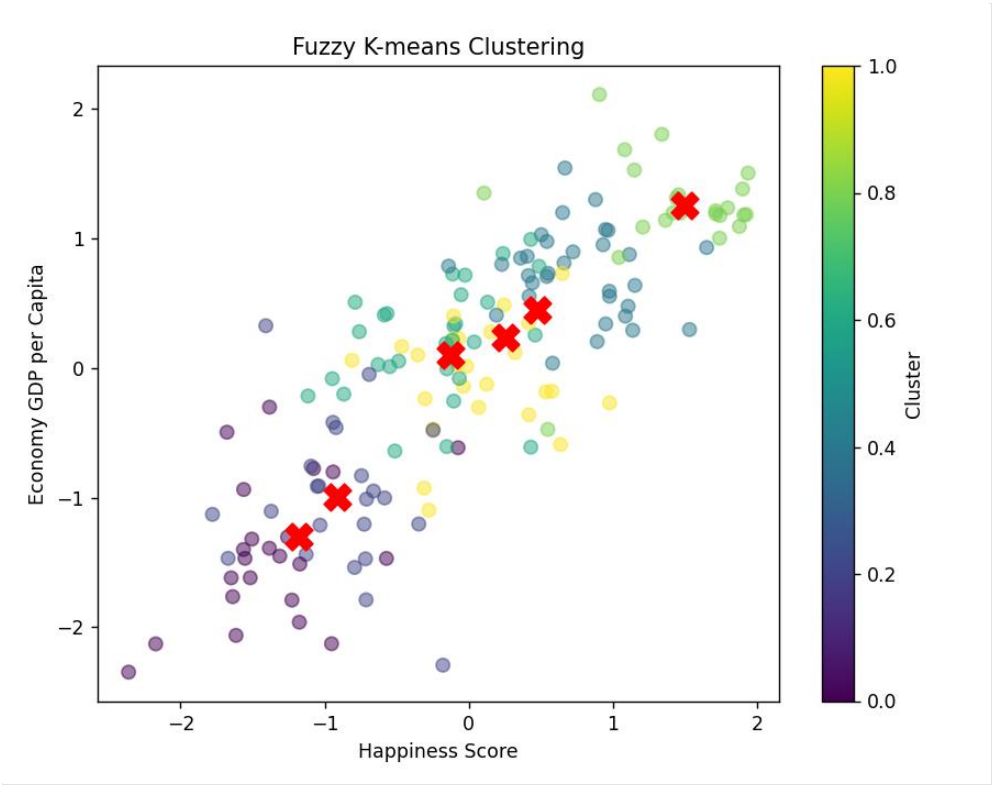
След извършване на размит клъстерен анализ се получават клъстери от държави с подобни характеристики по отношение на щастието на гражданите им. Визуализира се разпределението на държавите в тези клъстери и се анализират функциите на членство за всяка точка спрямо центровете на клъстерите.



Фигура 2. Резултати след групиране на декомпозираните данни по Economy (GDP per Capita) и Happiness Score с Fuzzy C-means на 3 отделни клъстера

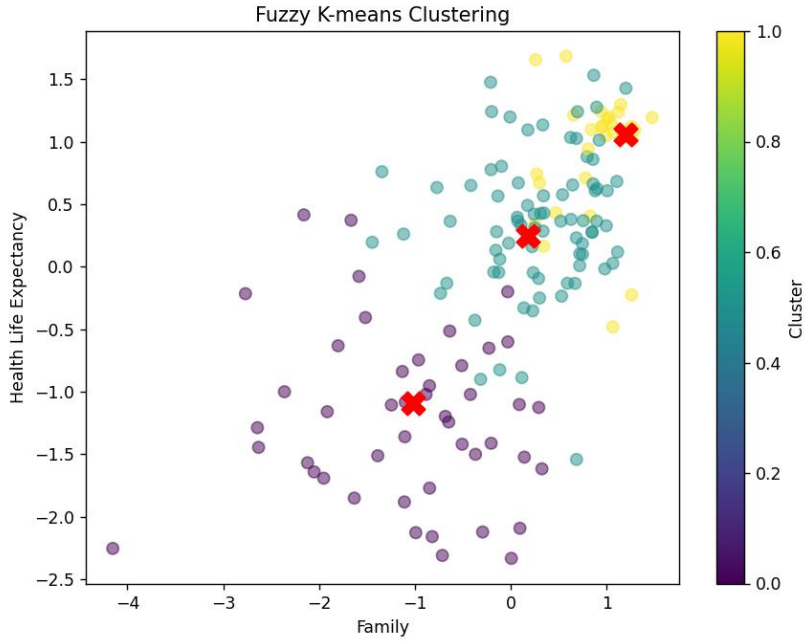


Фигура 3. Резултати след групиране на декомпозираните данни по Economy (GDP per Capita) и Happiness Score с Fuzzy C-means на 4 отделни клъстера

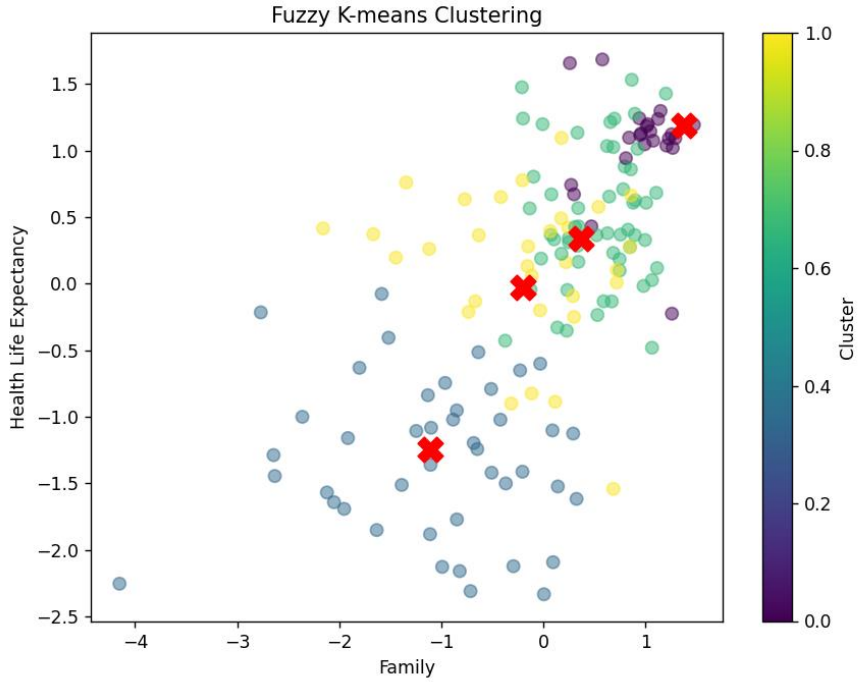


Фигура 4. Резултати след групиране на декомпозираните данни по Economy (GDP per Capita) и Happiness Score с Fuzzy C-means на 6 отделни клъстера

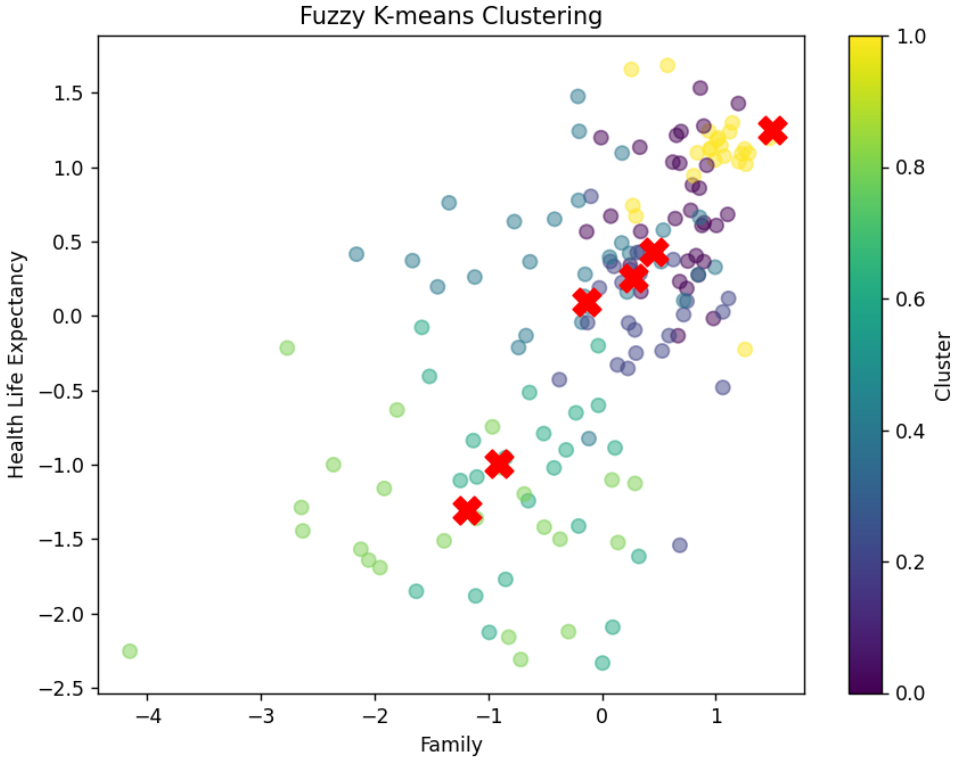
Предвид получените резултати достигаме до заключението, че при 4 броя клъстери точността на групирането е най-висока. Това може да се забележи на средната фигура. Има ясно изразени бройки, при които междуклъстерното пространство достига локален минимум и би бил подходящ избор за оптимален брой.



Фигура 5. Резултати след групиране на декомпозираните данни по Health (Life Expectating) и Family с Fuzzy C-means на 3 отделни клъстера



Фигура 6. Резултати след групиране на декомпозираните данни по Health (Life Expectating) и Family с Fuzzy C-means на 4 отделни клъстера



Фигура 7. Резултати след групиране на декомпозираните данни по Health (Life Expectating) и Family с Fuzzy C-means на 6 отделни клъстера

5. Основни изводи

Извършеният анализ позволява да се разбере как различните фактори влияят върху щастието на гражданите в различните държави. Въпреки че има различни подходи за анализ на този проблем, размитата клъстеризация предоставя възможност за по-гъвкаво и детайлно категоризиране на данните, което може да води до по-добро разбиране на взаимосвързаността между различните фактори и щастието на гражданите. Възможни подобрения и развития на проекта биха били:

- по-добра предпроцесна обработка на данните
- избиране на най-значими характеристики. Може да се постигне например чрез след прилагането на PCA
- оптимизация на размития алгоритъм - например смяна на критерия за спиране на итерациите, за да не се попада в локален екстремум
- добавяне на размит извод - при постъпване на нови данни, те да могат да се причислят към някой клъстер. Може да се постигне например чрез импликация на Мамдани
- добавянето на тегла към конкретни характеристики, върху които искаме да акцентираме

6. Списък на използваните технологии

- Кодът е реализиран на езика python
- Основни използвани библиотеки за:
    - обработка на данни - pandas, numpy
    - за част от предпроцесната обработка, тренирането и оценката – sclearn
    - за генериране на триъгълни функции – scfuzzy
    - за визуализация – matplotlib

7. Списък на използваната литература

- Fuzzy Clustering articles:

https://www.geeksforgeeks.org/ml-fuzzy-clustering/

- Fuzzy C-means clustering algorithm documentation:

https://fuzzy-c-means.readthedocs.io/en/latest/

- Elbow Method for Finding the Optimal Number of Clusters in K-Means:

https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

- Dataset:

https://www.kaggle.com/datasets/unsdsn/world-happiness

8. Приложение: Код на програмната реализация

Кодът е може да бъде намерен в архива на проекта.