**4.Domaći zadatak – Klasterizacija**

Cilj klasterizacije je za zadati skup zapisa, od kojih svaki sadrži skup atributa, i za zadatu meru sličnosti među njima, naći klastere tako da su :

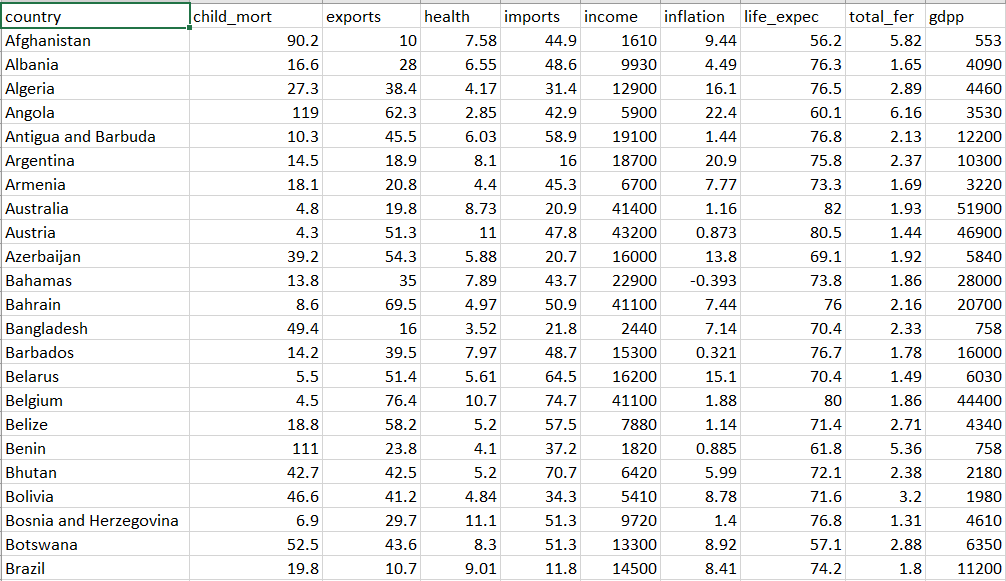
• Zapisi iz istog klastera što sličniji jedan drugom

• Zapisi iz različitih klastera što manje slični jedni drugom

**Dataset**:

Za potrebe domaćeg zadatka korišćen je „ Country-data.csv“ skup podataka (u .csv formatu), preuzet sa [https://www.kaggle.com/](https://www.kaggle.com/arjunbhasin2013/ccdata) („Humanitarian aid to underdeveloped countires“ dataset).

Na sledećoj slici prikazan je deo dataseta:



Radi se o podacima koji prikazuju socio-ekonomske i zdravstvene faktore koji određuju i utiču na razvijenost države. HEP International predstavlja humanitarnu nevladinu organizaciju koja se zalaže za borbu protiv siromaštva i pružanja pomoći i pogodnosti ljudima iz siromašnih i zaostalih država, kao i pružanja pomoći državama koje su zadesile prirodne katastrofe i nesreće. Ova organizacija kroz prethodno prikazane parameter u dataset-u zaključuje o stepenu razvijenosti država i odlučuje da izdvoji onoliko sredstava pomoći u zavinosti od stepena nerazvijenosti korespodentne države. Cilj ovog projekta će biti da se korišćenjem K-Means algoritma države grupišu u klastere po sličnosti u pogledu njihove razvijenosti, kako bi HELP International organizacija mogla lako da ustanovi koje države pripadau klasteru kome je pomoć najneophodnija.

Sadrži 10 atributa i 168 obzervacija.

**Atributi** :

*country* - naziv zemlje

*child*\_*mort* – stopa smrtonsti dece za određenu zemlju

*exports* – stopa izvoza određene zemlje

*health* – stopa zdravlja u određenoj zemlji

*imports* – stopa uvoza određene zemlje

*income* – stopa prihoda za određenu zemlju

*inflation* – inflacija

*life\_expec* – srednji ljudski vek određene zemlje

*total*\_*fer* – prosečan broj dece u određenoj zemlji

*gdpp* – bruto domaći proizvod zemlje

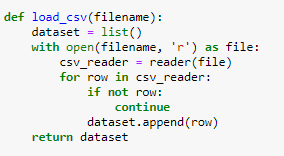
**Opis projekta:**

**Implementacija K-means algoritma**

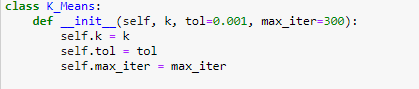
*Algoritam*:

1. Izabere se vrednost za K.
2. Nasumično se odabere K skupova karakteristika, da budu pocetni centroidi
3. Na osnovu sličnosti objekta i centroida , dodeljuju se objekti klasteru
4. Preračunavanje novih centroida klastera
5. Ovo se ponavlja sve dok se ne dodje do toga da se centroidi ne menjaju , ili je dostignut određeni broj iteracija

Funkcija **load\_csv**(filename) koristi se za čitanje podataka iz fajla csv formata.



Kreiramo **Kmeans** klasu. Prosleđuje se inicijalno K, kao sto je rečeno u opisu algoritma. Vrednost tol je tolerancija, što će govoriti da je kreirnje klastera optimizovano, ako se centroid ne pomera više od vrednosti tolerancije. Vrednost max\_iter je da se ograniči broj ciklusa izvršavanja.

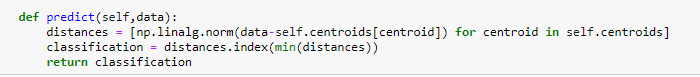


Metoda **fit**(self,data) koristi se za određivanje klastera. Prosleđuju joj se svi podaci. U centroids se čuvaju centroidi. Početni centroidi jesu prva dva uzorka iz prosleđenih podataka, koji se dodjeljuju kroz for petlju.Zatim se sa for petljom prolazi kroz max\_iter. Na početku su to prazne klasifikacija. Zatim se prolazi kroz atribute (karakteristike) iz dataseta i računa rastojanje (promenljia distances) svakog atributa od postojećih centroida i na taj način vrši njihova klasifikacija.

Sledeći korak, kao što je na početku rečeno, je preračunavanje novih centroida. Treba kreirati nove centroide i odrediti pomeranje trenutnih centroida. Da bi se proverilo koliko je optimizovano pomeranje, korste se trenutni i prethodni centroid. Na početku je pretpostavka da je optimizovano. Zatim se centroidi upoređuju sa prethodnim. Ukoliko su u pragu tolerancije, optimizovano je. Ukoliko nisu, promenljiva optimized se postavlja na False i nastavlja se početna for petlja u opsegu od 0 do self.max\_iter.

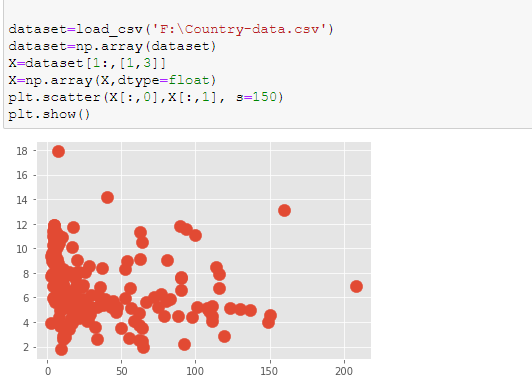


Metod **predict** određuje kom klasteru pripada koji podatak. Ova metoda na identičan način klasifikuje, kao kod dodeljivanja atributa klasterima na početku.

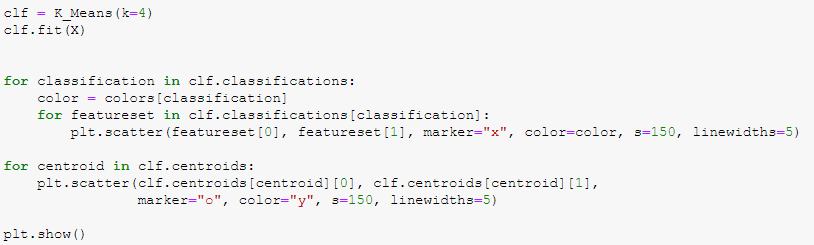


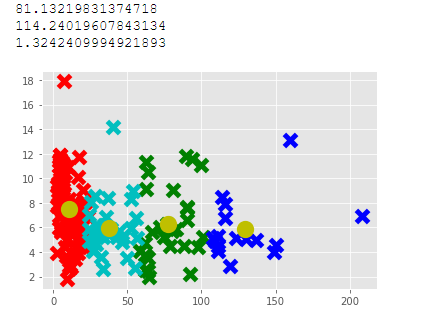
Učitavamo dataset. Na početku uzimamo kolone 1 i 3, što znaci pravimo klastere, tj grupišemo države prema tome kolika je stopa smrtnosi dece, a kolika zdravstvena stopa za korespodentnu državu (***child\_mort, health)***.

Crtamo grafik kako bismo videli kako su naši podaci raspoređeni. Vidi se da su podaci većine država međusobno na malim rastojanjima, ali da takođe postoje države gde je stopa smrtnosti dece daleko veća u poređenju sa ostalim državama.



Kreiramo K\_Means objekat.Kreiraćemo **4 klastera**. Koristimo fit metodu za klasterovanje. Koristimo scatter metodu za crtanje grafika, kako bismo prikazali klastere. Takođe žutom bojom obeležavamo svaki centroid.





**Zaključak**:

Klasteri označeni plavom i zelenom bojom predstavljaju države gde je stopa smrtnosti dece iznadprosečna(zeleni klaster) ili ekstremno velika (plavi klaster- Avganistan, Pakistan, Angola, Centralno Afrička republika, Haiti, Burkina Faso, Kongo, Mali, Mozambik, Gvineja…), za razliku od država gde je stopa smrtnosti dece ispod prosečna(aqua klaster) ili ekstremno mala (crveni klaster). Takođe, plavi i zeleni klaster predstavljaju države koje imaju ispodprosečnu zdravstvenu stopu (u većini su – Pakistan, Centralno Afrička Republika, Kongo, Sri Lanka, Turkemnistan…), za razliku od crvenog klastera gde većina država ima veliku zdravstvenu stopu razvoja. Po navedenim kriterijumima, lako se mogu izdvojiti države iz plavog klastera, kao i pojedine države zelenog klastera (sa niskom zdravstvenom stopom, a ekstremnijom stopom smrtnosti dece) koje HELP International vladina organizacija može smatrati najugroženijim i kojima je najneopohodnija njihova pomoć.