Analiza indeksa ljudskog razvoja i njegovih komponenti korišćenjem metoda mašinskog učenja

Dragana Radojičić
Ekonomski fakultet, Univerzitet u Beogradu
Beograd, Republika Srbija
dragana.radojicic@ekof.bg.ac.rs
0000-0001-7850-2623

Mladen Stamenković
Ekonomski fakultet, Univerzitet u Beogradu
Beograd, Republika Srbija
mladen.stamenkovic@ekof.bg.ac.rs
0000-0002-3838-878X

***Apstrakt* - Indeks ljudskog razvoja je pokazatelj koji služi za rangiranje zemalja prema nivou njihovog ljudskog razvoja i predstavlja meru napretka zemlje u pogledu kvaliteta životnog standarda njenih stanovnika. U okviru ovog istrazivanja koristimo bazu podataka koja pruža informacije o indeksu ljudskog razvoja za 195 zemalja, kao i podatke o očekivanom životnom veku, predviđenim godinama školovanja i bruto nacionalnom dohotku, koji će biti ključni za dalja istraživanja. Ideja ovog rada je koristeći različite tehnike mašinskog učenja analiziramo komponente indeksa ljudskog razvoja, kao i socio-ekonomskih faktora koji utiču na razvoj zemalja. Rezultati grupisanja metodom K-srednjih vrednosti ukazuju da zemlje sa višim indeksom ljudskog razvoja i bruto domaćim proizvodu po glavi stanovnika pripadaju klasterima koji se razlikuju u poređenju sa onima sa nižim vrednostima, naglašavajući značajne socio-ekonomske razlike između dobijenih klastera. Dalje, posmatrane podatke analiziramo koristeći algoritam slučajnih šuma kako bi ispitali uticaj posmatranih komponenti na indeks ljudskog razvoja.**

Ključne reči – Mašinsko učenje, Indeks ljudskog razvoja, Analiza glavnih komponenti, Klasterizacija metodom K-srednjih vrednosti, Algoritam slučajnih šuma.

I. UVOD

Indeks ljudskog razvoja, skraćeno IHR, (engl. Human Development Index (HDI)) predstavlja meru koja se koristi za procenu razvoja država, uzimajući u obzir tri ključna aspekta, naime: zdravlje stanovništva, obrazovanje, i poslednja komponenta je bruto nacionalni dohodak po glavi stanovnika (izražen u američkim dolarima). IHR služi kao sredstvo za poređenje socioekonomskog stanja različitih zemalja, i može pružiti korisne informacije za analizu politika posmatranih zemalja i unapređenje kvaliteta života. IHR se određuje kao geometrijska sredina normalizovanih indeksa tri ključne komponente. Zemlje se na osnovu vrednosti IHR-a obično svrstavaju u četiri kategorije:

- Vrlo visok ljudski razvoj (IHR viši ili jednak od 0,800)
- Visok ljudski razvoj (IHR od 0,700 do 0,799)
- Srednji ljudski razvoj (IHR između 0,550 i 0,699)
- Nizak ljudski razvoj (IHR ispod 0,550).

IHR je važna mera za praćenje napretka zemalja u pogledu ljudskog razvoja, ali se često koristi u kombinaciji s drugim indikatorima kako bi se dobila sveobuhvatna slika o stanju u određenoj zemlji. Program Ujedinjenih naroda za razvoj (UNDP) je 1990. godine u godišnjem izveštaju o ljudskom razvoju (Human Development Report) prvi put predstavio IHR. Kasnije, 2010.godine UNDP uvodi IHR prilagođen nejednakostima (engl. Inequality-adjusted Human Development Index (IHDI)), koji uzima u obzir nejednakosti u području zdravlja, obrazovanja i dohotka unutar pojedinih zemalja. IHR je prihvaćen kao značajnija mera razvoja, a njegova važnost i relevantnost su prepoznatljive. Kako bi se IHR proširio i obuhvatio više aspekata ljudskog razvoja, UNDP je uveo dodatne indekse, kao što su Indeks humanog razvoja prilagođen za nejednakost polova, Indeks humanog razvoja prilagođen za nejednakost raspodele, itd. UNDP i dalje radi na modernizaciji IHR-a kako bi odražavao savremene izazove, poput: klimatskih promena, digitalne revolucije, ekoloških razmatranja i itd.

Mnogi istraživači su pokazali interesovanje za proučavanje Indeksa ljudskog razvoja zbog njegovog značaja kao sveobuhvatne mere ljudskog blagostanja i razvoja jedne zemlje. Njihove studije često se usmeravaju na analizu faktora koji utiču na IHR, njegove promene kroz vremena i uticaj na oblikovanje politike i održivi razvoj. U radu [1] autori analiziraju IHR, razmatrajući njegove komponente, strukturu i kritike, i predlažu alternativne indekse za merenje ljudskog razvoja bazirane na poboljšanju komponenti indeksa. Studija u radu [2] koristi kvantitativni pristup sa deskriptivnom analizom i prostornom regresijom kako bi se ispitali faktori koji utiču na indeks ljudskog razvoja u Indonezii pre i tokom pandemije COVID-19. U literaturi [3] istražuje se razvoj Indeksa ljudskog razvoja, kao mera društveno-ekonomskog napretka, i prvi put se uvodi Indeks političke slobode. Rad [4] ističe da rast BDP-a značajno poboljšava IHR doprinoseći poboljšanju blagostanja ljudi, na osnovu analiza na podacima iz Indonezije.

Ograničenja i nedostaci izveštaja o humanom razvoju analizirani su u radu [5], s posebnim osvrtom na odstupanja izveštaja od njegove prvobitno predviđene uloge, i na to da IHR ne uspeva verno da odrazi stvarnost koju bi trebalo da meri. Brojni naučni radovi ispituju potencijalna poboljšanja indeksa, pri čemu je ova tema prvi put pokrenuta neposredno nakon izveštaja iz 1990. godine u nekoliko akademskih članaka ([6], [7], [8], [9], [10], itd.). Kako je mašinsko učenje počelo da se primenjuje u različitim domenima, istraživači su takođe iskoristili njegov potencijal da istraže indeks ljudskog razvoja. Klasifikacija IHR-a u literaturi često uključuje subjektivno prosuđivanje i podložna je kritici, autori rada [11] koriste klasterizaciju metodom K-srednjih vrednosti i algoritme K-medoida vođene podacima da grupišu IHR u tri klastera, minimizirajući subjektivnost u procesu klasifikacije. U radu [12] autori koristeći klasterizaciju metodom K-srednjih vrednosti dolaze do tri različite grupe regencija u Centralnoj Javi: oblasti visog IHR, oblasti srednjeg IHR i oblasti niskog IHR. Koristeći dva modela slučajnih šuma,

jedan za regresiju i jedan za klasifikaciju, studija [13] analizira indeks ljudskog razvoja u cilju razumevanja dinamike globalnog razvoja. Autori rada [14] su implementirali različite algoritme mašinskog učenja za predviđanje indeksa ljudskog razvoja i testirali su klasterizaciju metodom K-srednjih vrednosti i hijerahijska klaster analizu da grupiše vrednosti IHR indikatora iz 186 zemalja u četiri oznake. Studija u radu [15] grupiše odabrane HDI indikatore u 6 klastera u svim okruzima u Istočnoj Nusa Tengari, kako bi bile identifikovane oblasti za sprovođenje odgovarajućih politika.

U ovom radu ideja je da korišćenjem algoritma klasterizacije metodom K-srednjih vrednosti identifikujemo klastere zemalja prema njihovim sličnostima u razvojnim pokazateljima i izdvojimo grupe zemalja koje imaju sličan nivo razvoja.

II. DESKRIPTIVNA ANALIZA

A. Baza podataka

Za potrebe ovog istraživanja korišćena je baza podataka „Human Development Index and Components 2021“, preuzeta sa sajta https://www.kaggle.com/ (datum pristupa: 01.12.2024.godine). Baza sadrži podatke o indeksu ljudskog razvoja, kao i podatke o njegovim komponenetama: očekivanom životnom veku, prosečnom broju godina školovanja i bruto nacionalnom dohotku, za 195 zemalja u 2021. godini. Posmatrana baza sadrži podatke o indikatorima za procenu i poređenje nivoa razvoja i kvaliteta života u različitim zemljama sveta.

B. Analize baze i deskriptivne statistike

Pre početka same analize podataka, proveravamo da li baza sadrži nedefinisane ili nedostajuće podatke. Da bismo sprečili negativan uticaj na rezultate, uklanjamo sve redove u kojima se pojavljuje barem jedna nedostajuća numerička vrednost. Kako bismo dobili jasniju sliku o posmatranim podacima, pregledajmo deskriptivne statistike numeričkih karakteristika prisutnih u posmatranoj bazi podataka koje su tabelarno prikazane na Slici 1.

	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita
count	191.0000	191.0000	191.0000	191.0000	191.0000
mean	0.7206	71.3147	13.5304	8.9838	20249.0942
std	0.1507	7.6465	2.9260	3.1732	21825.2641
min	0.3850	52.5000	5.5000	2.1000	732.0000
25%	0.5995	65.7500	11.6000	6.2500	4593.0000
50%	0.7390	71.7000	13.4000	9.3000	12306.0000
75%	0.8350	76.7000	15.6000	11.5000	30079.5000
max	0.9620	85.5000	21.1000	14.1000	146830.0000

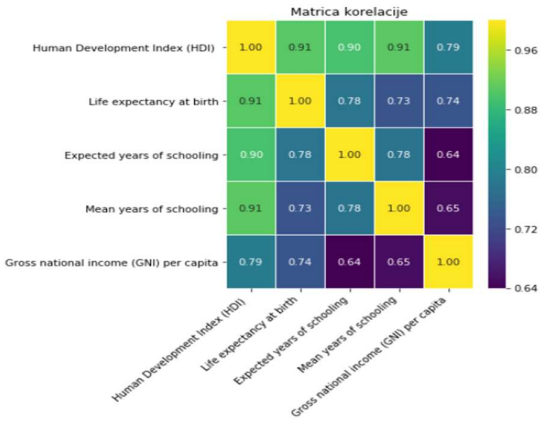
Slika 1. Prikaz deskriptivnih statistika numeričkih atributa

Konkretno, izdvajamo i podatke za Srbiju, što je prikazano na Slici 2.

Country	HUMAN DEVELOPMENT	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita
Serbia	VERY HIGH	0.802	74.2	14.4	11.4	19123.0

Slika 2. Podaci o indikatorima o IHR za Srbiju

Kako bi se prikazala međusobna veza između posmatranih karakteristika i eventualna redundancija, na Slici 3 prikazane su međusobne individualne korelacije varijabli.



Slika 3. Međusobne individualne korelacije varijabli

Možemo primetiti jaku i veoma jaku korelaciju između varijabli, što je zapravo sasvim očekivan rezultat, s obzirom da posmatrani indikatora mere slične aspekte.

III. KLASTERIZACIJA METODOM K-SREDNJIH VREDNOSTI

Algoritam klasterizacije metodom K-srednjih vrednosti (eng. K-means) omogućava analizu razvojnih pokazatelja i prepoznavanje zemalja koje se svrstavaju u slične razvojne grupe. Klasterizacija metodom K-srednjih vrednosti je algoritam nenadgledanog učenja koji omogućava grupisanje opservacija u K klastera prema njihovim sličnostima.

Klasterizacija metodom K-srednjih vrednosti se može predstaviti u sledećim fazama:

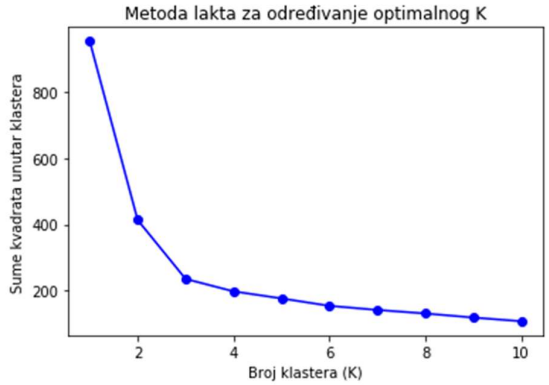
- *Odabir broja klastera* (može se odrediti na osnovu prethodnog znanja, metodom lakta, analizom podataka, itd.),
- *Inicijalizacija K početnih centroida* (središnjih tačaka) za klastere (obično se nasumično odredi K tačaka),
- *Dodeljivanje opservacija klasterima* (svaka opservacija iz posmatrane baze dodeljuje se najbližem klasteru najčešće korišćenjem Euklidske distance),
- *Ažuriranje centara klastera* (centroida svakog klastera se ponovo izračunava kao prosek svih opservacija koje su u tom klasteru),
- *Ponovljavanje koraka 3 i 4* (Ovi koraci se ponavljaju dok se pozicije centroida više ne menjaju ili dok se ne dostigne maksimalni broj iteracija).

Ovaj algoritam spada među najstarije algoritme za klasterovanje – njegova osnovna ideja datira iz 1956. godine (videti [16]), iako je formalno dobio naziv u radu [17].

Prednosti algoritma klasterizacije metodom K-srednjih vrednosti su jednostavnost za implementaciju i razumevanje, efikasnost, brzina, itd. Međutim, algoritam ima i nedostatke, kao što su: potrebno je unapred odrediti broj klastere, osetljivost na outliere, zavisnostost od odabira inicijalnih tačaka, itd. U radu [18] autori predlažu poboljšani algoritam

klasterizacije metodom K-srednjih vrednosti, koji kombinuje algoritam najveće minimalne udaljenosti i tradicionalni algoritam klasterizacije metodom K-srednjih vrednosti, koji prevazilazi nedostatke tradicionalnog algoritma klasterizacije metodom K-srednjih vrednosti za inicijalizaciju početnih centroidnih tačaka. Rad [19] prikazuje sveobuhvatnu sliku i analizu algoritma klasterizacije metodom K-srednjih vrednosti, ispitujući njegove prednosti i nedostatke.

Postupak klasterovanja metodom K-srednjih vrednosti počinjemo selekcijom numeričkih kolona, i standardizacijom njihovih vrednosti, zatim biramo broj klastera metodom lakta. Posmatranjem grafikona prikazanog na Slici 3 optimalan broj klastera K determinišemo koristeći metod lakta. Na grafikonu su prikazane promene suma kvadrata odstupanja tačaka od centroida u klasterima, i dalje tražeći prevojnu tačku određujemo broj klastera. Primećujemo da je uočljivo prelamanje linije oko vrednosti 3 na *x* osi, i nakon toga algoritam se pokreće za K=3, pa se podaci raspoređuju u tri klastera.



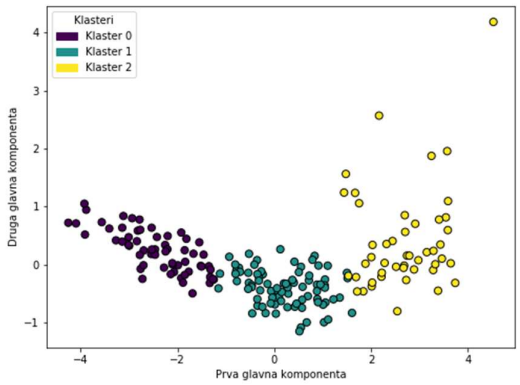
Slika 3. Determinisanje optimalnog broja klastera

Kako bismo utvrdili specifičnosti i razlike između klastera izračunavamo prosečne vrednosti obeležja po klasterima, a rezultati su prikazani na Slici 4. Treba imati na umu da indeksiranje serija i nizova u programskom jeziku Python počinje od broja 0, pa je na Slici 4 prvi klaster je indeksiran '0', itd. Možemo zaključiti da treći klaster obuhvata zemlje sa najvišim vrednostima svih posmatranih karakteristika, dok se u prvom klasteru nalaze zemlje sa najnižim vrednostima posmatranih parametara.

Cluster	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita
0	0.535250	62.866667	10.363333	5.220000	3576.800000
1	0.748798	72.116667	13.992857	9.915476	14480.404762
2	0.906723	80.665957	16.746809	12.123404	51842.872340

Slika 4. Prosečne vrednosti karakteristika u dobijenim klasterima

S obzirom da su naši podaci visokodimenzionalni, kako bismo redukovali dimenzionalnost i vizuelno prikazali klastere, a da sačuvamo što više varijanse iz originalnog skupa podataka, primenjujemo analizu glavnih kompomnenti. Analiza glavnih komponenti omogućava projekciju višedimenzionalnih podataka uz pomoć najvažnijih glavnih komponenti. Prikaz raspodela zemalja po klasterima u odnosu na prve dve glavne komponente je prezentovan na Slici 5.



Slika 5. Vizualizacija klasterovanja metodom K-srednjih vrednosti pomoću glavnih komponenti

IV. SLUČAJNE ŠUME

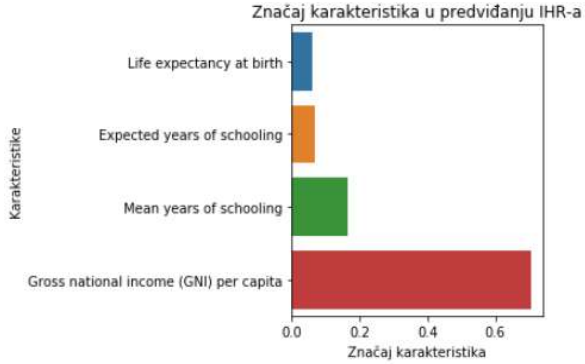
Slučajne šume (engl. random forest) je algoritam mašinskog učenja zasnovan na ansambl metodi, što zapravo znači da kombinuje više modela kako bi poboljšao tačnost i pouzdanost pri rešavanju kompleksnih zadataka. Ovo je algoritam proste agregacije i baziran je na treniranju više stabala odlučivanja (tzv. ansambli), koristeći nasumično izabrane podskupove stabala, a ponekad i nasumične podskupove posmatranih karakteristika.

Algoritam slučajne šume se može predstaviti u sledećim fazama:

- *Generisanje više stabala odlučivanja:* kreira veliki broj stabala odlučivanja, zatim se svako stablo obučava na različitim podskupovima podataka.
- *Slučajni izbor karakteristika:* slučajnim izborom selektuje se podskup karakteristika iz skupa svih prisutnih karakteristika.
- *Treniranje stabala odlučivanja:* svako stablo se trenira na drugačijem podskupu podataka i karakteristika.
- *Agregacija rezultata:* u slučaju regresionih zadataka rezultati pojedinačnih stabala se uprosečavaju, a u slučaju klasifikacionih zadataka odlučuje na osnovu većinskog glasanja.

Ključna ideja je da se za treniranje svakog stabla koriste različiti uzorci podataka i atributa, što omogućava smanjenje varijanse i poboljšanje tačnosti. Slučajne šume su korisne i imaju široku primenu zbog prednosti koje ima ovaj algoritam kao sto su: velika otpornost na nedovoljno prilagođavanje (engl. underfitting) i preterano prilagođavanje (engl. overfitting), otpornost na promene u podacima, pruža dobre rezultate i kada neki podaci nedostaju, može da oceni važnost svake posmatrane karakteristike u bazi podataka, itd. S druge strane ovaj algoritam ima i nedostatke kao što su: proces treniranja je nekada računarski zahtevan, gubitak interpretabilnosti, zahteva puno memorije, itd. S obzirom da se algoritam slučajnih šuma može koristiti za evaluaciju važnosti karakteristika, koristimo algoritam da analiziramo uticaj posmatranih komponenata iz naše baze na indeks humanog razvoja. Najpre definišemo novi kategorički atribut „HDI Rank“ i dodajemo ga u posmatranu bazu za svaku prisutnu opservaciju, tako što u zavisnosti od intervala kome vrednost IHR-a pripada (pomenutih u prvom delu) preslikavamo u vrednost ’4’ ukoliko je vrednost iz kategorije ’Vrlo visok ljudski razvoj’, u ’3’ ukoliko je vrednost iz

kategorije ’Visok ljudski razvoj’, u ’2’ ukoliko je vrednost iz kategorije ’ Srednji ljudski razvoj’, u ’1’ ukoliko je vrednost iz kategorije ’ Nizak ljudski razvoj’. Možemo uočiti da bruto nacionalni dohodak po glavi stanovnika (engl. Gross national income per capita) ima najznačajniji uticaj na "IHR Rank", pa samim tim i na vrednost IHR-a, što je i logično, imajući u vidu samu prirodu njegove definicije. Prikaz uticaja posmatranih atributa na varijablu 'HDI Rank' je prikazan na Slici 6. Posmatrani model zasnovan na algoritmu slučajnih šuma pokazuje da bruto nacionalni dohodak po glavi stanovnika ima izrazito značajan uticaj na IHR, što se poklapa i sa metodologijom izračunavanja IHR-a. Važnost karakteristika u modelu slučajne šume pokazuje koliki uticaj svaka pojedinačna karakteristika ima na donošenje odluka unutar modela.



Slika 6. Uticaj karakteristika na IHR

V. ZAKLJUČAK

Budući da indeks ljudskog razvoja odražava socioekonomski status zemlje, analiza njegovih komponenti i uticaja na sam indeks je ključna za identifikaciju oblasti koje zahtevaju poboljšanje kvaliteta života. Primenom metode K-srednjih vrednosti identifikove su grupe zemalja sličnim profilima razvoja koje se odlikuju sličnim razvojnim karakteristikama. Analiza je pokazala da zemlje sa višim vrednostima IHR-a i bruto nacionalnog dohodka po glavi stanovnika formiraju posebne klastere u odnosu na zemlje sa nižim vrednostima, upravo to ukazuje na socio-ekonomske razlike između posmatranih zemalja. Dalje, primenom algoritma slučajne šume ocenjujemo značaj karakteristika i zaključujemo da bruto nacionalni dohodak po glavi stanovnika ima najsnažniji uticaj u ocenjivanju kategorije IHR-a. Dalje, primenom algoritma slučajne šume ocenjujemo značaj karakteristika i zaključujemo da bruto nacionalni dohodak po glavi stanovnika ima najsnažniji uticaj u ocenjivanju kategorije IHR-a. S obzirom da tehnike mašinskog učenja mogu ukazati i izdvojiti obrasce između IHR-a i socio-ekonomskih indikatora, daljim razvojom i primenom tih modela na podacima o IHR-a i njegovim komponentama mogu se izvesti relacije što može dati smernice za unapredjenja.

LITERATURA

[1] F. Noorbakhsh. “The human development index: some technical issues and alternative indices,” Journal of International Development: The Journal of the Development Studies Association, vol. 10.5, str. 589-605, 1998.

[2] S. Astari. “Spatial Analysis of The Human Development Index in Indonesia Before and During The Covid-19 Pandemic,” IOP Conference Series: Earth and Environmental Science. IOP Publishing, str. 012002, 2024.

[3] M. Ul Haq, “Reflections on human development,” Oxford university Press, 1995.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 2003, str. 271–350.

[4] M. B. Setiawan, and A. Hakim, “Indeks pembangunan manusia Indonesia,” Jurnal Economia, vol. 9(1), str. 18-26, 2015.

[5] S. Anand, and A. Sen, “Human Development Index: Methodology and Measurement,” 1994.

[6] D. P. Doessel, and R. Gounder, “International comparisons of the standards of living and the human development index,” Discussion Papers in Economics, vol. 72, str. 1212-1217, 1991.

[7] M. Hopkins, “Human development revisited: A new UNDP report,” World Development, vol. 19(10), str. 1469-1473, 1991.

[8] N. C. Lind, “Some thoughts on the human development index,” Social Indicators Research, vol. 27, str. 89-101, 1992.

[9] G. Pyatt, “Poverty: a wasted decade,” European economic review, vol. 35(2-3), str. 358-365, 1991.

[10] H. Wang, J.H. Feil, and X. Yu, “Let the data speak about the cut-off values for multidimensional index: Classification of human development index with machine learning,” Socio-Economic Planning Sciences, vol. 87, str. 101523, 2023.

[11] R.T. Vulandari, S. Siswanti, A.K. Kusumawijaya, and K. Sandradewi, “Classification of human development index using k-means,” Indonesian Journal of Applied Statistics, vol. 2(1), str. 1-9, 2019.

[12] J. Gsim, and M.Z. Es-sadek, “Machine Learning Projections for Human Development Index Anticipation,” 2024. https://doi.org/10.21203/rs.3.rs-4376154/v1D.

[13] F.B. Khan, and A. Noor, “Prediction and Classification of Human Development Index Using Machine Learning Techniques,” In 2021 5th International Conference on Electrical Information and Communication Technology (EICT), IEEE, str. 1-6, decembar 2021.

[14] J.E. Simarmata, D. Chrisinta, and M. Purnomo, “Implementation of K-Means Clustering to Human Development Indicators in East Nusa Tenggara,” Journal of Research in Mathematics Trends and Technology, vol. 6(2), str. 46-56, 2024.

[15] Y. Li, and H. Wu, “A clustering method based on K-means algorithm,” Physics Procedia, vol. 25, str. 1104-1109, 2012.

[16] H. Steinhaus, “Sur la division des corps matériels en parties,” Bull. Acad. Polon. Sci. (in French), vol. 4 (12), str. 801–804, 1957.

[17] J. B. MacQueen, “Some Methods for classification and Analysis of Multivariate Observations,” Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press, MR 0214227, Zbl 0214.46201, str. 281–297, 1967.

[18] J. Cui, J. Liu, and Z. Liao, “Research on K-means clustering algorithm and its implementation,” In Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013), Atlantis Press, str. 1804-1806, mart 2013.

[19] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” Information Sciences, vol. 622, str. 178-210, 2023.

Analysis of human development index and its components using machine learning methods

Dragana Radojičić, Mladen Stamenković

ABSTRACT

The Human Development Index is an indicator used to rank countries according to the level of their human development and is a measure of the country's progress in terms of the quality of living standards of its inhabitants. In this research, we use a database that provides information on the human development index for 195 countries, as well as data on life expectancy, projected years of schooling and gross national income, which will be key to further research. The idea of this paper is to use different machine learning techniques to analyze the components of the human development index, as well as socio-economic factors that influence the development of countries. The results of clustering using the K-means method indicate that countries with higher human development index and gross domestic product per capita belong to clusters that differ compared to those with lower values, highlighting significant socio-economic differences between the obtained clusters. Furthermore, we analyze the observed data using the random forest algorithm in order to examine the influence of the observed components on the human development index.