1. Opis dataset-a

Dataset je kreiran spajanjem 2 različita dataseta, jednog koji sadrži podatke o državama, i drugog koji sadrži procenat osoba koje puše u određenim državama. Datasetovi su spojeni po državama koje su se preklapale u oba dataseta.

Sastoji se od 25 kolona (atributa) I sadrži 114 obzervacija. Prva kolona sadrži države, a ostale kolone geografske, demografkse, socijalne, ekonomske i slične parametre tih država. U poslednjoj koloni se nalazi procenat ukupnog stanovništva datih država koji konzumira duvan, odnosno preciznije, puši. Taj podatak ćemo koristiti kao verovatnoću da nasumično odabrana osoba iz određene države puši.

1.1 Opis atributa

1. country - država

2. Region – region

3. Surface area km2 – površina države u izražena kvadratnim kilometrima

4. Population in thousands – broj stanovnika izražen u hiljadama

5. Population density (per km2) – gustina naseljenosti po kvadratnom kilometru

6. Sex ratio (m per 100f) – odnos muškog i ženskog stanovništva, broj muškaraca na 100 žena

7. GDP: Gross domestic product (million current US$) – bruto domaći proizvod izražen u milionima dolara

8. GDP growth rate (annual %, const. 2005 prices) – procentualni rast BDP-a na godišnjem nivou

9. GDP per capita (current US$) – BDP po stanovniku izražen u dolarima

10. Economy: Industry (% of GVA) – procenat bruto dodate vrednosti iz industrije

11. Employment: Industry (% of employed) – procenat zaposlenog stanovništva u industriji

12. Employment: Services (% of employed) – procenat zaposlenog stanovništva u uslužnim delatnostima

13. Unemployment (% of labour force) – procenat nezaposlene radne snage

14. Labour force participation (female/male pop. %) – odnos žena naspram muškaraca od ukupne radne snage izražen procentualno

15. Population growth rate (average annual %) – procentualni rast populacije na godišnjem nivou

16. Urban population (% of total population) – procenat urbanog stanovništva

17. Urban population growth rate (average annual %) – procentualni rast urbane populacije na godišnjem nivou

18. Fertility rate, total (live births per woman) – broj rođene dece po ženi

19. Population age distribution (0-14 / 60+ years, %) – odnos mladog i starog stanovništva izražen u procentima

20. International migrant stock (000/% of total pop.) – broj migranata izražen u promilima

21. Education: Government expenditure (% of GDP) – procenat BDP koji država izdvaja za obrazovanje

22. Education: Primary gross enrol. ratio (f/m per 100 pop.) – odnos ženskog i muškog stanovništva koji upišu osnovni stepen obrazovanja

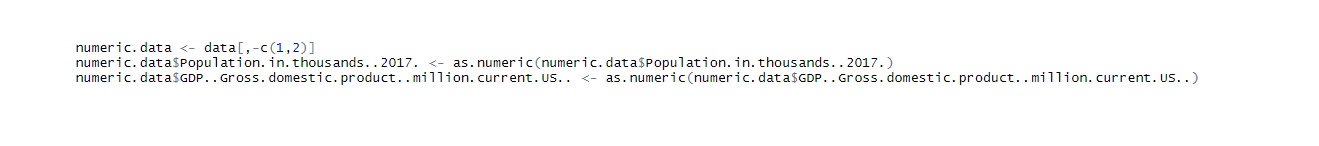
23. Education: Secondary gross enrol. ratio (f/m per 100 pop.) - odnos ženskog i muškog stanovništva koji upišu srednji stepen obrazovanja

24. Education: Tertiary gross enrol. ratio (f/m per 100 pop.) - odnos ženskog i muškog stanovništva koji upišu viši stepen obrazovanja

25. Total Smoking Rate – procenat stanovnika koji puši

2. Sređivanje podataka

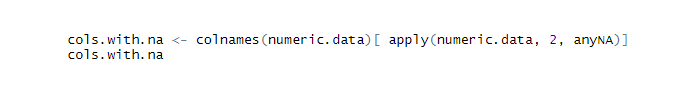
Za rad sa neuronskom mrežom koristićemo samo numerićke podatke, tako da ćemo kreirati novi dataset, koji će biti podskup originalnog, bez prve dve kolone. Takođe, dve varijable su tipa int, a ne numeric, tako da ćemo i njih pretvoriti u odgovarajući tip.



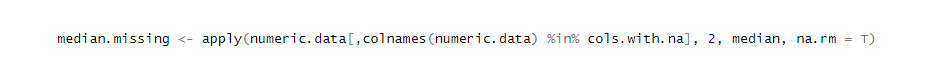
Sada imamo dataset koji se sastoji isključivo od numeričkih podataka. Sledeći korak je da proverimo da li postoje nedostajuće vrednosti (NA).



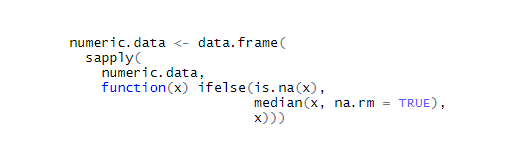
Proverom dolazimo do podatka da postoje 32 NA vrednosti. Neuronska mreža ne može da radi sa nedostajućim vrednostima, tako da je to problem koji moramo da rešimo pre početka rada. Jedan od načina je da se potpuno uklone redovi u kojima se nalaze nedostajuće vrednosti, ali to bi značajno oštetilo naš dataset. Zbog toga ćemo se opredeliti za zamenu, a ne uklanjanje NA vrednosti. S obzirom da naši podaci nemaju normalnu raspodelu, najbolje rešenje je zameniti NA vrednosti sa medijanama odgovarajućih kolona. Za početak ćemo napraviti listu imena svih kolona koje sadrže NA vrednosti.



Zatim ćemo iskoristiti tu listu, kako bi izračunali medijanu za te kolone.



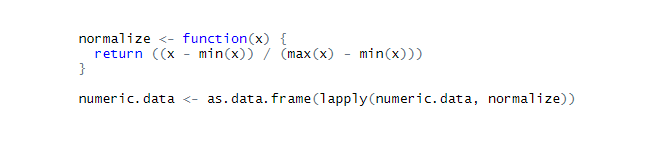
I na kraju konačno možemo da zamenimo NA vrednosti sa medijanom, primenom funkcije.



Sada kada imamo potpun numerički dataset, vreme je za normalizaciju podataka. Normalizacija predstavlja transliranje vrednosti podataka na opseg od 0 do 1. To je dobra praksa kod primene neuronskih mreža, a i u mašinskom učenju generalno. Odnos i distribucija podataka ostaće nepromenjeni, ali ćemo eliminisati prisustvo ogromnih brojeva, koji mogu da dovedu do grešaka i nepreciznosti prilikom primene neuronske mreže. Normalizacija se vrši primenom sledeće formule:



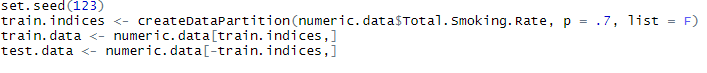
Ispisaćemo funkciju koja vrši normalizaciju podataka, i primeniti je na ceo dataset.



Dataset je sada u potpunosti spreman za rad.

# 3. Kreiranje neuronske mreže

Pre svega, podelićemo dataset na train i test podatke.

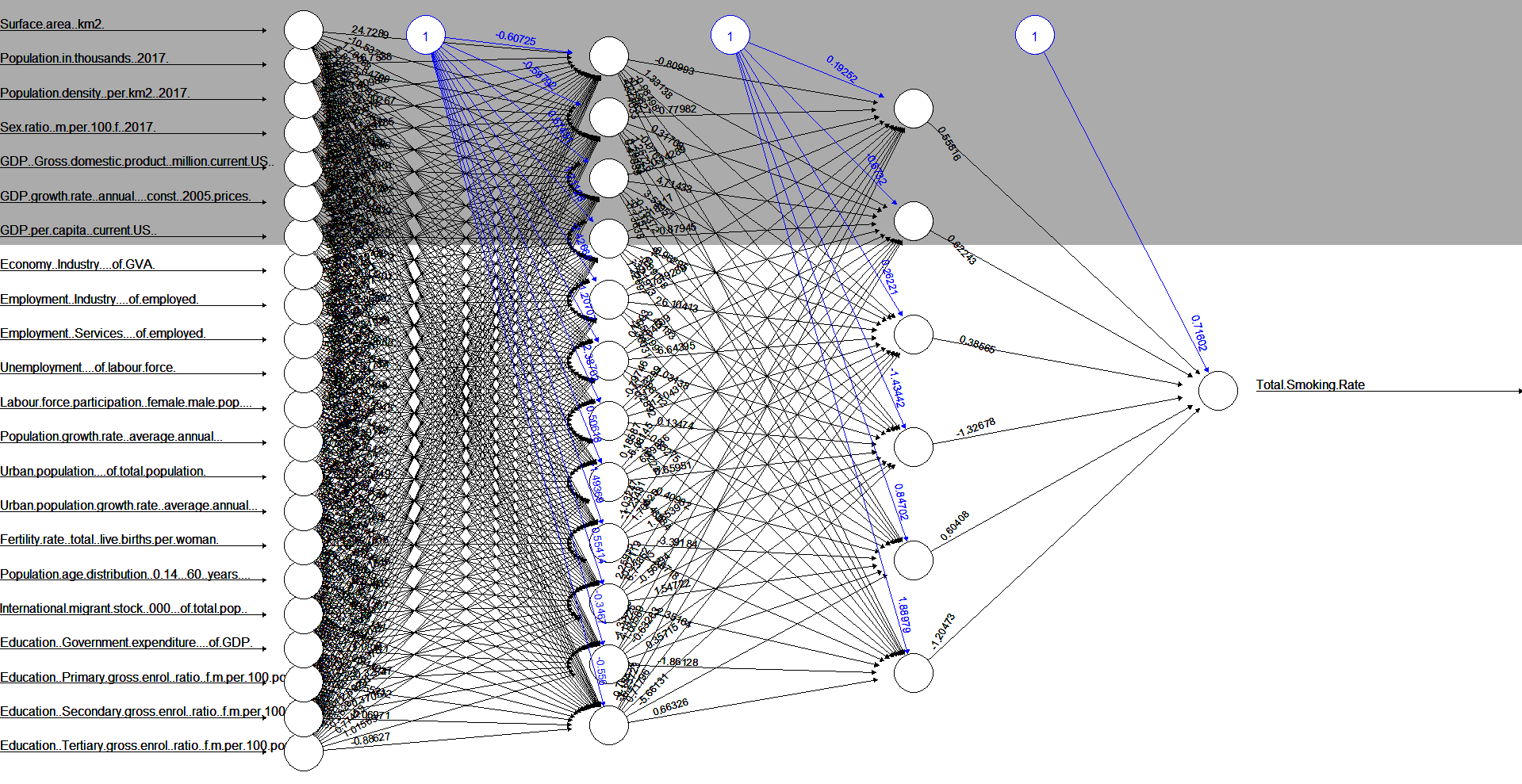


70% dataseta je uzeto za train, a 30% za test podatke.

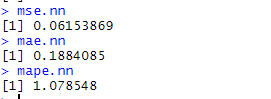
Zatim pravimo neuronsku mrežu. U ovom slučaju ona ima 2 skrivena sloja, jedan sa 12 I drugi sa 6 skrivenih neurona. Ne postoji precizno definisano pravilo za određivanje broja skrivenih slojeva I neurona u njima, pa su oni nasumično odabrani.



Nakon kreiranja mreže, možemo da je grafički prikažemo:



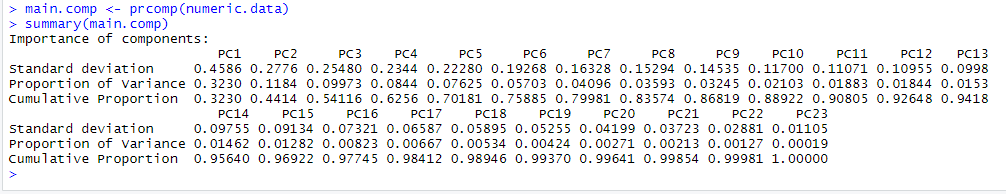
Nakon vršenja predikcija, izračunali smo mean squared error, mean average error I mean average percentage error, kako bismo videli Koliko je mreža dobra.



Vidimo da greške nisu male, što znači da naša neuronska mreža ne vrši veoma preciznu predikciju. Kako bismo eliminisali promenljive koje su međusobno korelisane, kao i one koje jako malo doprinose rezultatu predikcije, izvršićemo analizu glavnih komponenti, i zatim napraviti novu neuronsku mrežu koristeći samo novodobijene varijable.

# 4. Analiza glavnih komponenti

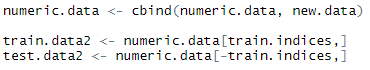
Analizom glavnih komponenti dobićemo linearnu i nekorelisanu kombinaciju originalnih varijabli. Prva varjabla obuhvata najveći deo varijabiliteta, sledeća najveći deo preostalog varijabiliteta, i tako do kraja. Radi smanjenja broja varijabli, kao i zanemarivanja onih koje imaju jako mali značaj za mrežu, uzećemo varijable koje obuhvataju približno 90% varijabiliteta.



Iz tabele možemo da uoćimo da je već kod 11. varijable obuhvaćeno 90.8% ukupnog varijabiliteta, pa ćemo za novu mrežu iskoristiti samo prvih jedanaest varijabli.

# 5. Kreiranje neuronske mreže sa varijablama iz analize glavnih komponenti

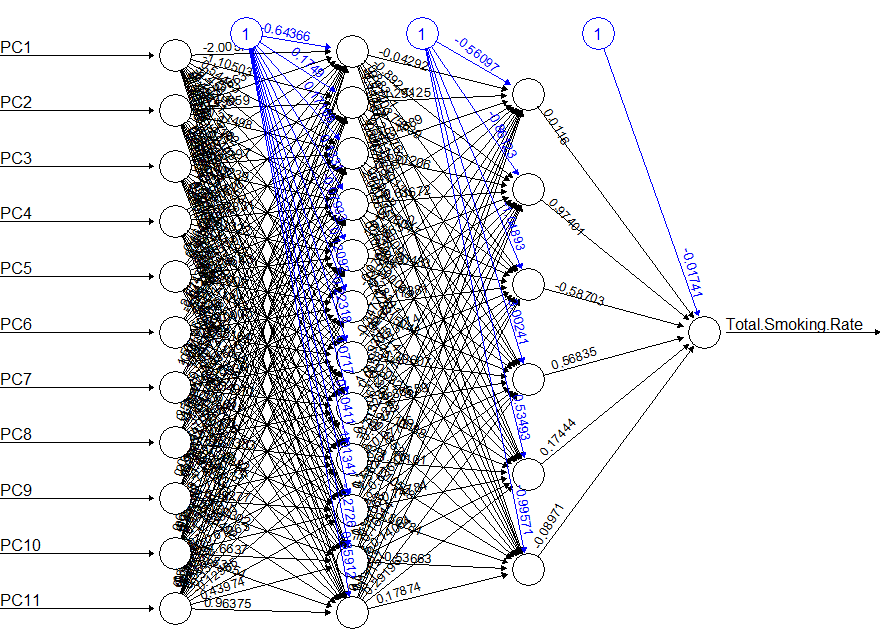
Za početak ćemo dodati nove varijable na ranije korišćen dataset, i izvršiti novu podelu na train i test podatke.



Zatim kreiramo novu neuronsku mrežu, koja će da koristi samo nove varijable za predikciju. Broj skirvenih slojeva i neurona ostavićemo nepromenjenim.

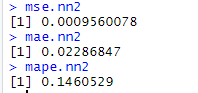


Grafički ćemo prikazati i novu mrežu:



Možemo uočiti da sada imamo samo 11 ulaznih neurona, odnosno ulaza, koji predstavljaju varijable dobijene iz analize glavnih komponenti.

Nakon vršenja predikcije, ponovo računamo 3 vrste greške, radi provere koliko mreža dobro vrši predikciju, kao i radi poređenja sa prethodnom mrežom.



Greške su izuzetno male, i neuporedivo manje od prethodne neuronske mreže.

# 6. Zaključak

Neuronska mreža kreirana korišćenjem svih podataka, pokazala se kao veoma neprecizna. Na osnovu toga može se zaključiti da u originalnom datasetu postoji veliki broj korelisanih varijabli, kao i veliki broj varijabli koje nemaju veliki uticaj na predikciju. Zbog toga smo se odlučili da odradimo analizu glavnih komponenti, kako bismo dobili nekorelisane varijable koje opisuju preko 90% varijabiliteta originalnih podataka, i time značajno smanjili broj podataka i isfiltirali sve one podatke koji smetaju pri predikciji. Nakon analize dobili smo 11 varijabli pomoću kojih smo napravili novu mrežu, koja se pokazala kao mnogo bolja od prve, sa jako malim greškama.

|  |  |  |
| --- | --- | --- |
|  | **Neuronska mreža**  **(sve varijable)** | **Neuronska mreža**  **(samo varijable iz pca)** |
| **Mean Squared Error**  **(MSE)** | 10.50330 | 5.50694 |
| **Mean Absolute Error**  **(MAE)** | 2.51758 | 1.72301 |
| **Mean Absolute Percentage Error**  **(MAPE)** | 0.13532 | 0.09533 |