Klasterovanje nad skupom podataka o saobraćajnim nesrećama u Velikoj Britaniji (2005-2017)

Seminarski rad u okviru kursa
Istraživanje podataka 1
Matematički fakultet, Univerzitet u Beogradu

Jovan Mirkov
148/2016
mi16148@matf.bg.ac.rs

15.08.2019

Sažetak

Cilj ovog seminarskog rada je klasterovanje nad datim skupom podataka. Korišćene su različite metode klasterovanja obrađene u programskom jeziku Python i IBM-ovom alatu SPSS Modeler. Najbolje rezultate analize je dao algoritam *BIRCH* i *hijerarhisko klasterovanje*.

Sadržaj

1 Opis skupa podataka

Vlada Velike Britanije prikuplja i objavljuje na godišnjem nivou detaljne informacije o saobraćajnim nesrećama širom zemlje. Podaci dolaze sa internet stranice (dato u literaturi) otvorenih podataka Vlade Velike Britanije gde ih je objavilo ministarstvo saobraćaja.

Skup podataka sastoji se od dve csv datoteke:

- Accident_Information.csv: svaki red u datoteci predstavlja jedinstvenu saobracajnu nesreću koja sadrži različita svojstva vezana za nesreću kao kolone. Podaci su skupljani u rasponu godina 2005-2017.

- Vehicle_Information.csv: svaki red u datoteci predstavlja sudelovanje jedinstvenog vozila u jedinstvenoj prometnoj nesreći, s različitim kolonama svojstava vozila i putnika kao kolona. Podaci su skupljani u rasponu godina 2004-2016.

Dva gore spomenuta skupa podataka mogu se povezati putem jedinstvenog identifikatora saobracajne nesreće (kolona Accident_Index).

2 Predprocesiranje podataka

Tabela Accident_Information.csv ima 2047256 redova i 34 kolone a tabela Vehicle_Information.csv ima 2177205 redova i 24 kolone. Veliki broj atributa je kategoričkog tipa što nam govori da će biti dosta transformacija u numeričke tipove.

Koristivši python biblioteku "missingno" uvideli smo da neke kolone u naša dva csv fajla imaju previše nedostajućih vrednosti da bi imalo smisla vršiti neke inputacije i zato ih uklanjamo. Uklanjamo atribute za koje imamo veliki procenat vrednosti koje govore o tome da dati atribut nije zabeležen, veliki procenat istih vrednosti (uglavnom više od 98%) ili jednostavno predstavljaju informaciju korisnu isključivo za domen arhiviranja događaja a ne kao u našem slučaju za klasterovanje. Pošto je skup svakako obiman, i nakon što smo gore navedenom analizom ukolnili veći deo nedostajućih vrednosti, slobodno možemo ukloniti i cele redove gde se javljaju nedostajuce vrednosti jer ih ima zanemarljivo malo.

Nakon ovog dela ciscenja stanje je sledece: Tabela Accident_Information.csv ima 1289781 redova i 23 kolone, dok tabela Vehicle_Information.csv ima 1707518 redova i 9 kolone.

Može se reći da se broj atributa drastično smanjio ali sam odabir atributa prilikom izrade skupa je zasluzan za to, mada još uvek imamo više od milion redova.

Koristeci atribut 'Accident_index' vršimo unutrašnje spajanje po toj koloni. Ovim gubimo veliki broj redova jer se godine prikupljanja podataka prvobitna dva skupa ne podudaraju, 2005-2017 za Accident_Information.csv, 2004-2016 za Vehicle_Information.csv.

Atribut	Tip	Opis
Accident_Index	string	identifikacioni broj saobraćajne nesreće
1st_Road_Class	string	vrsta puta
Accident_Severity	string	ozbiljnost nesreće
Did_Police_Officer_Attend_Scene_of_Accident	float	reagovanje policije
Junction_Control	string	čime je raskrsnica kontrolisana
Junction_Detail	string	detalji o raskrsnici
Latitude	float	geografska širina lokacije
Light_Conditions	string	uslovi osvetljenosti
Local_Authority_(District)	string	lokalne vlasti prema distriktima
Local_Authority_(Highway)	string	lokalne vlasti prema autoputevima
Longitude	float	geografska dužina lokacije
Number_of_Casualties	int	broj žrtvi
Number_of_Vehicles	int	broj vozila
Police_Force	string	policijska snaga prema lokaciji sudara
Road_Surface_Conditions	string	uslovi površine puta
Road_Type	string	tip puta
Special_Conditions_at_Site	string	specijalne okolnosti na putu
Speed_limit	float	ograničenje brzine
Weather_Conditions	string	vremenske prilike
InScotland	string	da li se desilo u Škotskoj
Daytime	string	doba dana
Date	string	datum
Hour	int	sat u danu
Age_Band_of_Driver	string	godine vozača
Driver_Home_Area_Type	string	tip mesta
Junction_Location	string	raskrsnica
Sex_of_Driver	string	pol vozača
Vehicle_Leaving_Carriageway	string	pozicija vozila u odnosu na put nakon sudra
Vehicle_Manoeuvre	string	akcija koja je dovela do sudara
Vehicle_Type	string	tip vozila
X1st_Point_of_Impact	string	mesto udara na automobilu

Slika 1: Korišćeni atributi

Nakon ovog dela čišćenja imamo samo jednu tabelu i to sa 1088249 redova i 30 kolona.

Sledeća faza čišćenja se sastoji od transformacije kategorickih atributa u numeričke, ali samo onih za koje ima smisla napraviti takvo preslikavanje u numerički tip tako da se nakon transformacije mogu smatrati rednim(ordinalnim) atributom.

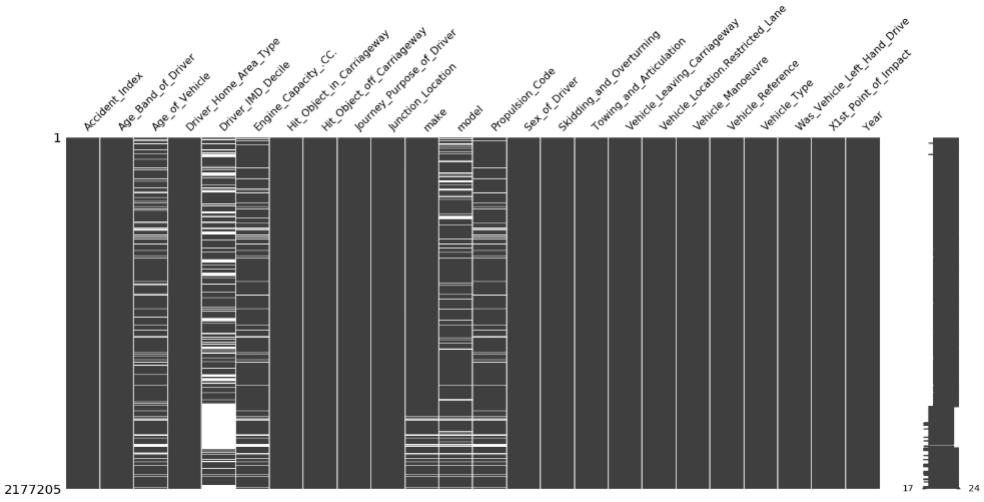
Sve ostale kategoričke atribute transofrmišemo tako što brišemo tu kolonu i za svaku jedinstvenu vrednost te kolone pravimo novu. Ove nove kolone imaju vrednost '1' ako za dati red važi svojstvo koje odgovara novoj flag koloni, a inače '0'.

Radi eksperimentisanja sa različitim strukturama tipova u tabeli, nakon dalje segmentacije tabela dobijamo sledeće četiri tabele:

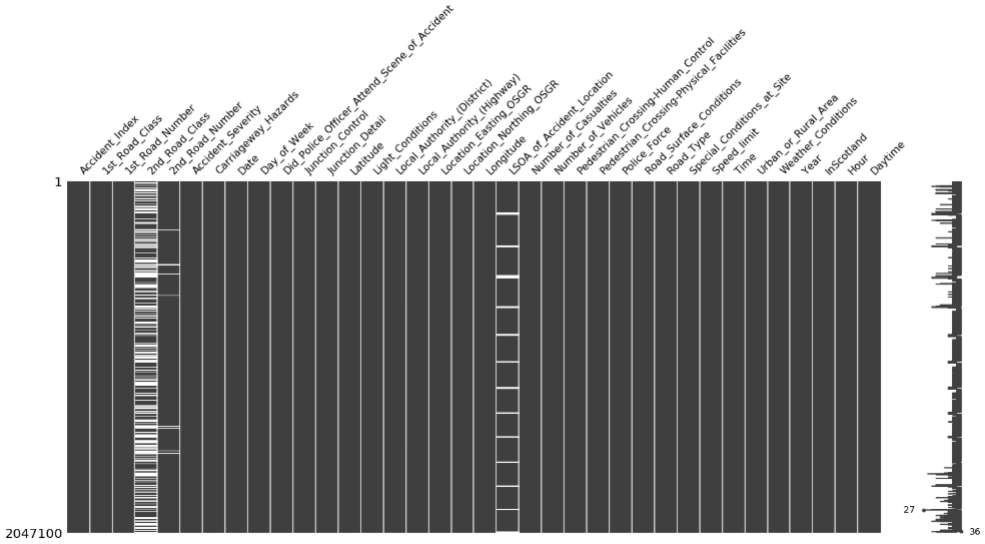
- A - tabela sa atributima koji su eksperimentisanjem pri odabira atributa pokazali dobre rezultate
- B - tabela sa atributima koji su prvobitno bili numečki i atributi koje smo prethodno pretovrili u redne
- C - tabela sa isključivo flag atributima dobijenih iz kategoričkih atributa koji nisu mogli biti prebačeni u redne
- D - prethodne dve tabele zajedno

3 Vizuelizacija podataka

Već smo spominjali Python biblioteku *missingno* koja nam je pomogla da uočimo kolone sa velikim brojem nedostajućih vrednosti. Iz priložene slike možemo videti i koje su to kolone.

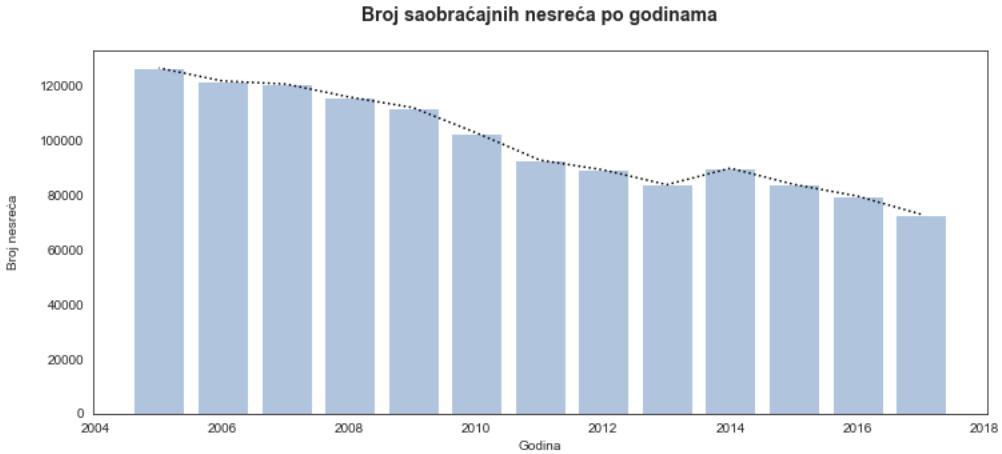


Slika 2: Nedostajuće vrednosti u tabeli Vehicle_Information.csv

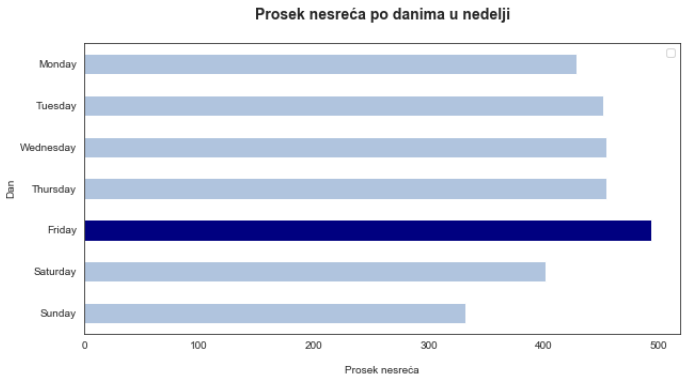


Slika 3: Nedostajuće vrednosti u tabeli Accident_Information.csv

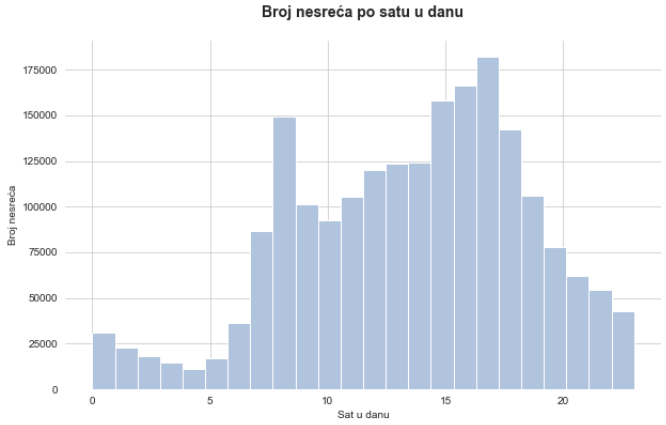
Kroz sledeći niz vizuelizacija smo odgovorili na pitanja koja se prva javljaju kad uočimo atribute koji predstavljaju vremenske odrednice, tačnije, atribute 'Date' i 'Hour'. Nakon završene viuelizacija možemo ukloniti ove atribute jer je atribut 'Hour' već predstavljen preko atributa 'Daytime' a atribut 'Date' nam neće biti potreban. Primećujemo da broj nesreća opada tokom godina.



Slika 4: Broj saobraćajnih nesreća po godinama

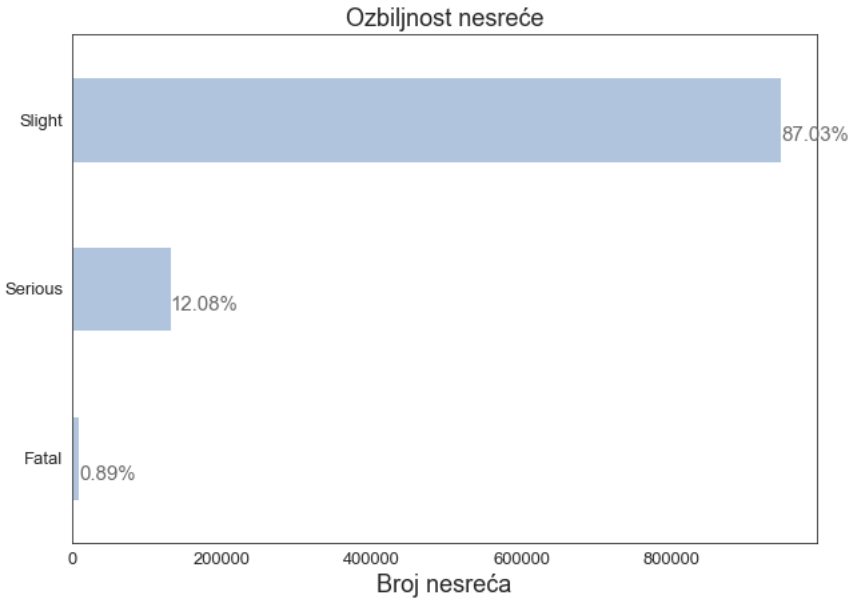


Slika 5: Prosek nesreća po danima u nedelji



Slika 6: Broj nesreća po satu u danu

Korisno je i videti procente težina slučaja nesreće.



Slika 7: Ozbiljnost nesreće

Za sledeći niz vizuelizacija koristimo atribute 'Police_Force ', 'Local_Authority_(District)' i 'Local_Authority_(Highway)' preko kojih ćemo njihovim iscrtavanjem dobiti mapu segmentisanu po atributu koji budemo prosledili kao skup kategorija prema kome se menja boja na slici. Korišćenje funkcije *scatter* nemamo mogućnost da korigujemo boje jer je argument za tu vrstu već iskorišćen tako da boje možda neće biti odabrane tako da se vidi razlika izmedju dobijenih površina. Na slici nema navođenja koja boja predstavlja koju kategoriju zbog preglednosti jer ih ima previše (ivideti slike u Dodatku A).

Na mapama se mogu videti različite oblasti delovanja policijskih snaga, lokalnih vlasti gledano po distriktima i lokalnih vlasti gledano na osnovu odgovornosti prema putevima. Nakon ovih vizuelizacija uklanjamo te atribute jer nose prevelik broj kategorija i samim tim bi doprineo velikom broju kolona u slučaju prebacivanja u flag atribute.

Matrica korelacije (videti u Dodatku A) nam ukazuje na neke korelacije:

- uslovi osvetljenosti - deo dana
- uslovi - uslovi na putevima
- ogranicenje brzine - tip puta
- ogranicenje brzine - tip okoline
- broj povredjenih - broj automobila koji ucestvuju

4 Klasterovanje

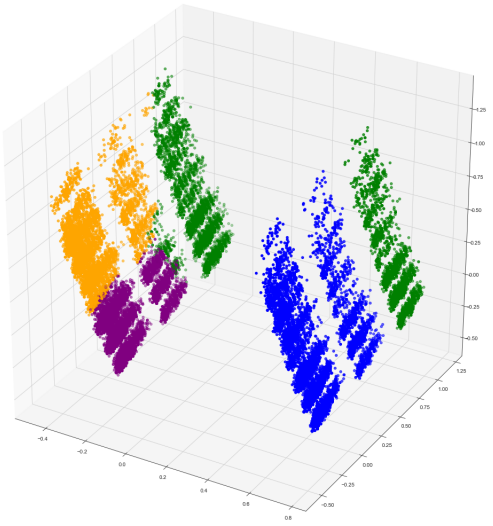
U svim narednim algoritmima su se koristili uzorci odgovarajuće veličine jer je skup obiman, ali izbor veličine uzoraka se birao do gornje granice koja ne uzrokuje 'MemoryError'. Za sve algoritme je korišćena analiza glavnih komponenti (PCA) za bolju vizuelizaciju.

4.1 K - sredina

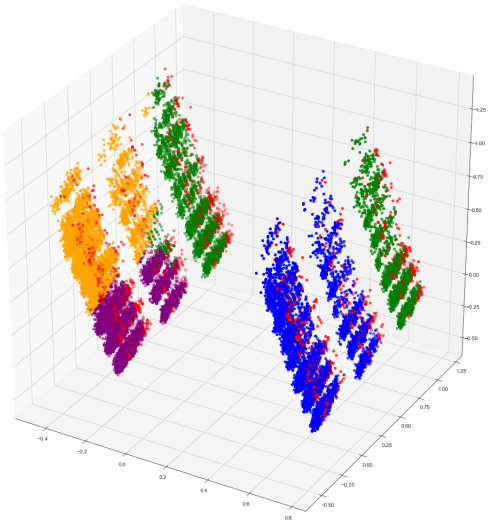
Korišćeno je *pravilo lakta* za utvrđivanje broja klastera kao parametar ovog algoritma. Rezultati senka koeficijenta pokazuju solidne rezultate za neke skupove. Za skup B je prikazano dodatno menjanje klastera prilikom menjanja broja klastera kao parametar algoritma.

Vrednosti senka koeficijenta prema skupovima opsanim u predprocesiranju:

- Tabela A: 0.39
- Tabela B: 0.36
- Tabela C: 0.25
- Tabela D: 0.24



Slika 8: 4-means



Slika 9: 5-means

4.2 DBSCAN

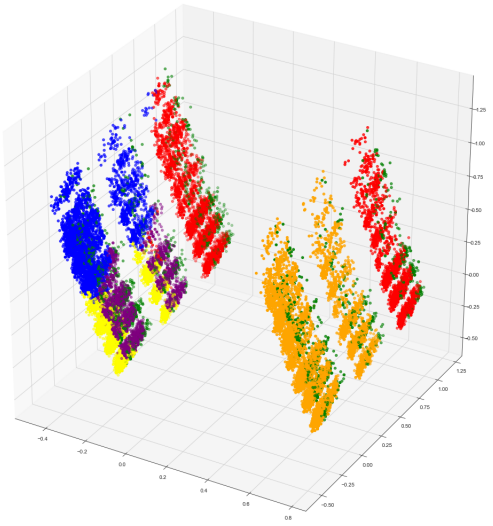
Radi jednostavnosti predstavljanja i prema isprobavanju raznih *epsilon* vrednosti za parametar algoritma biće prikazani rezultati samo za vrednot 0.3. Rezultati su veoma loši, što će pokazati senka koeficijent.

Vrednosti senka koeficijenta prema skupovima opsanim u predprocesiranju:

- Tabela A: 0.32
- Tabela B: 0.16
- Tabela C: 0.19
- Tabela D: -0.32

4.3 Hijerarhisko klasterovanje

Prema senka koeficijentu ovaj algoritam daje ubedljivo najbolje rezultate, u tabeli se nalaze te vrednosti prema kombinacijama mera bliskosti i kriterijuma za određivanje blizine. Kao vizuelizacija neće biti iscrtavano deljenje skupa na klastere redom od prve podele pa na dalje jer rezultati nisu vizuelno zadovoljavajući a takođe bi to podrazumevalo mnogo slika. Umesto toga biće prikazani dendagrami.



Slika 10: 6-means

4.4 BIRCH

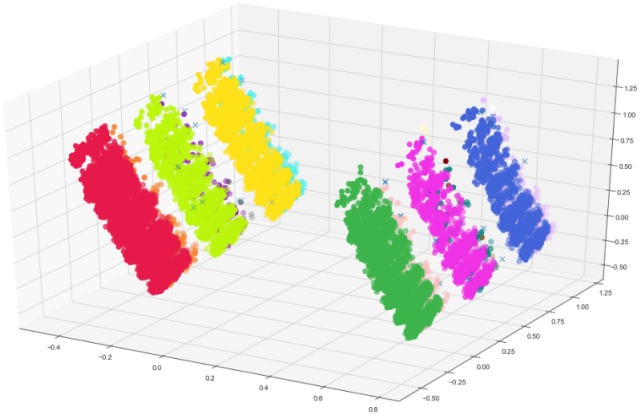
BIRCH (*balanced iterative reducing and clustering using hierarchies*) algoritam je još poznat i kao Two-Step klasterovanje jer se sastoji iz dve faze: faza učitavanja podataka u memoriju uz pravljenje *cluster-feature* drveta i faza izvršavanja nekog drugog algoritma klasterovanja (uglavnom *k-means* ili *hijerarhisko*)nad listovima *cluster-feature* drveta uz dodatno profinjavanje klastera (pomera pojedinačne redove skupa ako su u listovima dodeljenje pogrešnom klasteru).[**d**] Dobijamo dobre rezultate za skupove sa bez flag atributa (skup A i B).

Vrednosti senka koeficijenta prema skupovima opsanim u predprocesiranju:

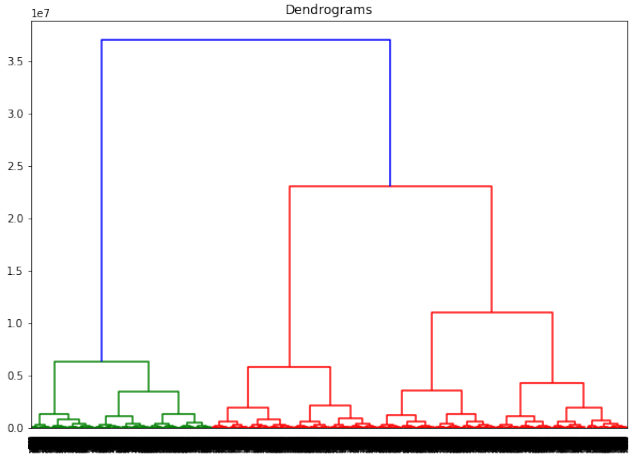
- Tabela A: 0.72
- Tabela B: 0.55
- Tabela C: 0.31
- Tabela D: 0.37

4.5 Kohonen

Algoritam Kohonen, još poznat kao *self-organizing map (SOM)* koristi veštačke neuronske mreže za klasterovanje. Senka koeficijent kod



Slika 11: DBSCAN nad tabelom B



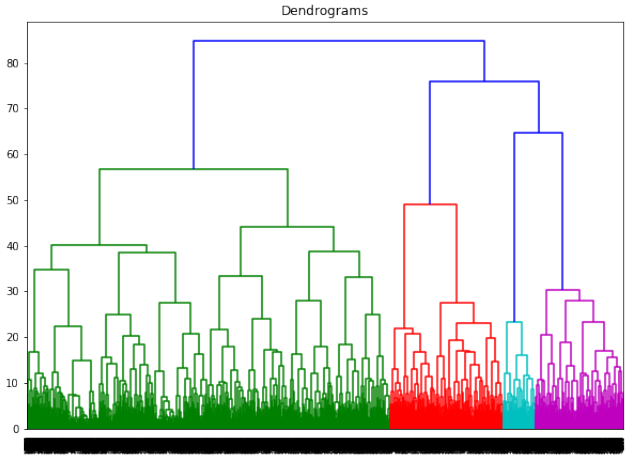
Slika 12: Skup A

naših ulaznik tabela je tokom izvršavanja sa razlicitim parametrima pokretanja bio između -0.5 i 0.25, ali je ipak skup A dao dovoljno dobre rezultate tako da ćemo ih prikazati.

5 Zaključak

Najbolje rezultate su ostvarili *hijerarhisko klasterovanje* i *Birch klasterovanje*. Najgore rezultate je ostvarilo DBSCAN. Izbor različitih tabela prema tipu podataka je uticao na rezultate.

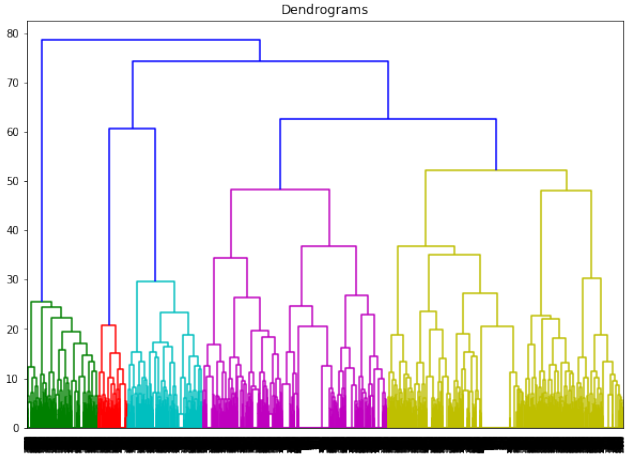
Tabela C sa isključivo flag atributima je dobila najbolje rezultate u odnosu na ostale tabele pri *hijerarhiskom klasterovanju*, dok je kod drugih algoritama dobila loše ili jedva osrednje. Tabela D je kod svakog algoritma davala najlošije rezultate. Tabele A i B su uglavnom davale osrednje i dobre rezultate tako da se može zaključiti da je u opštem



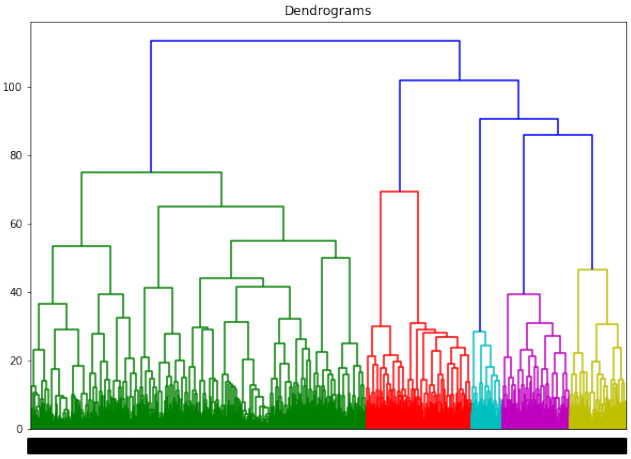
Slika 13: Skup B

slučaju najbolje koristiti tabele sa (originalno) numeričkim vrednostima i dodavanje nekih kategoričkih koji se mogu transformisati u redni atribut može poboljšati rezultate.

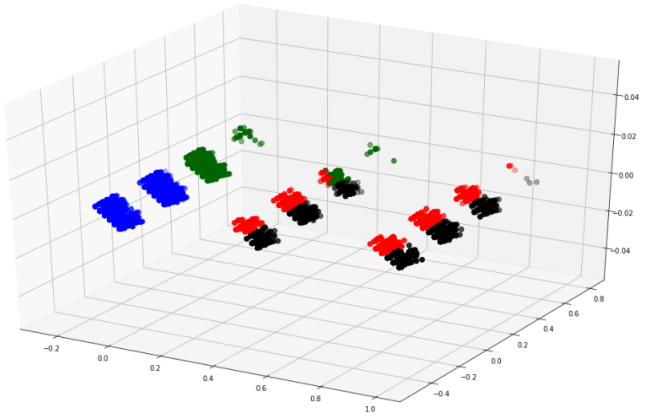
A Dodatak A



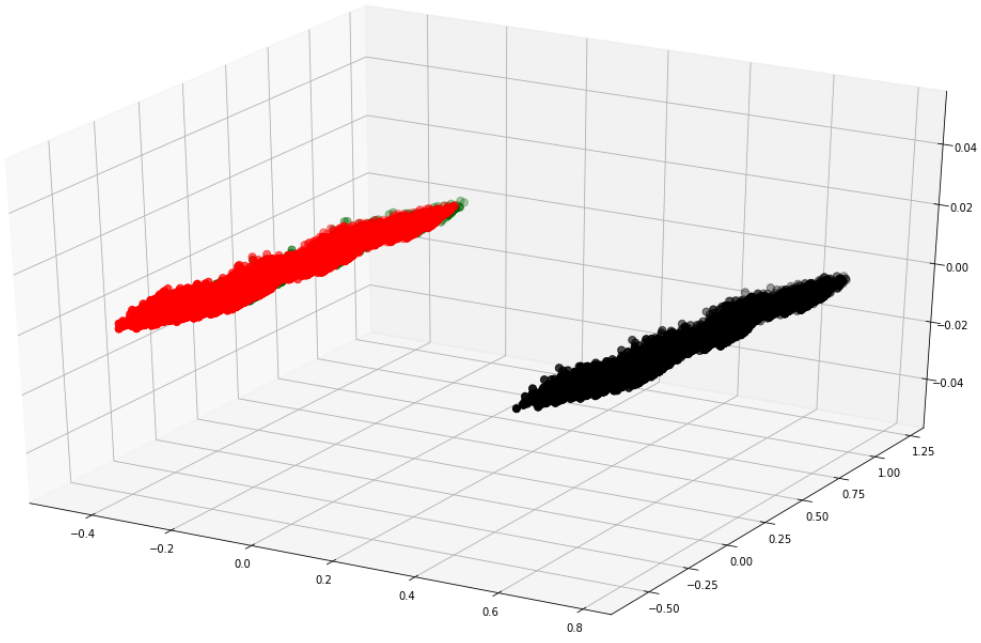
Slika 14: Skup C



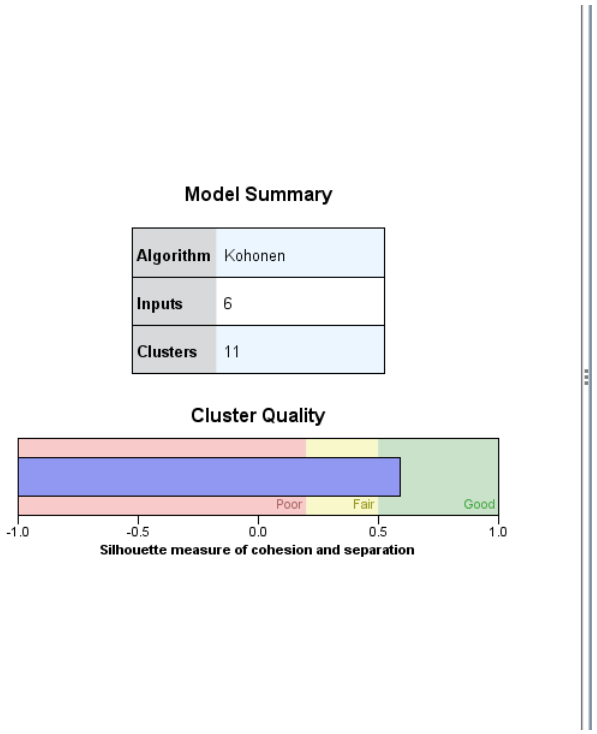
Slika 15: Skup D

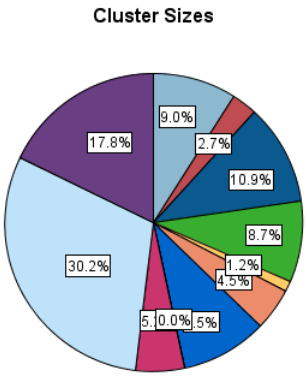


Slika 16: BIRCH nad skupom A



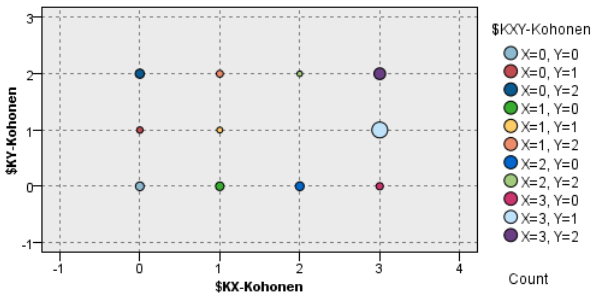
Slika 17: BIRCH nad skupom B



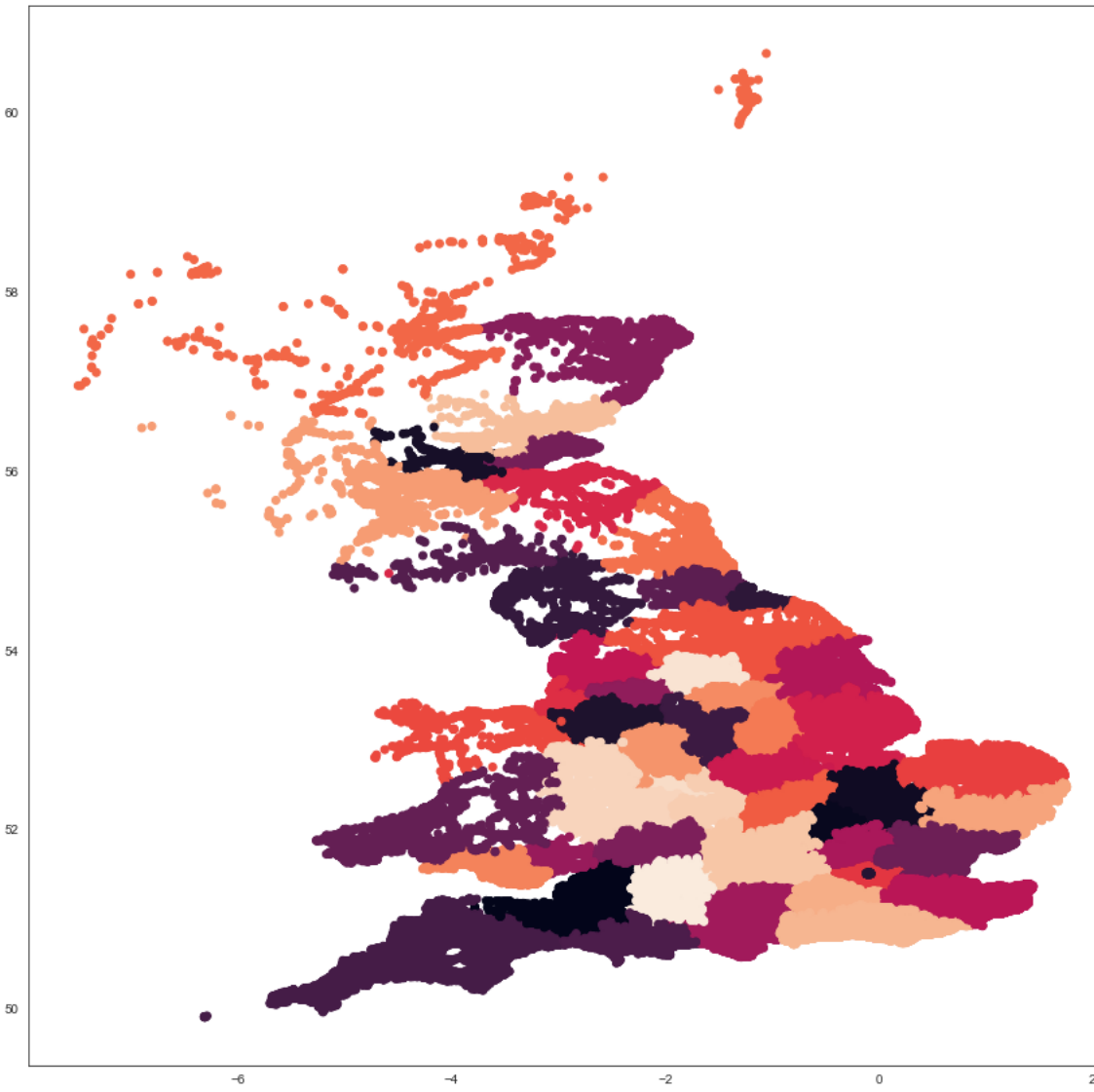


Size of Smallest Cluster	15 (0%)
Size of Largest Cluster	32828 (30.
Ratio of Sizes: Largest Cluster to Smallest Cluster	2,188.53

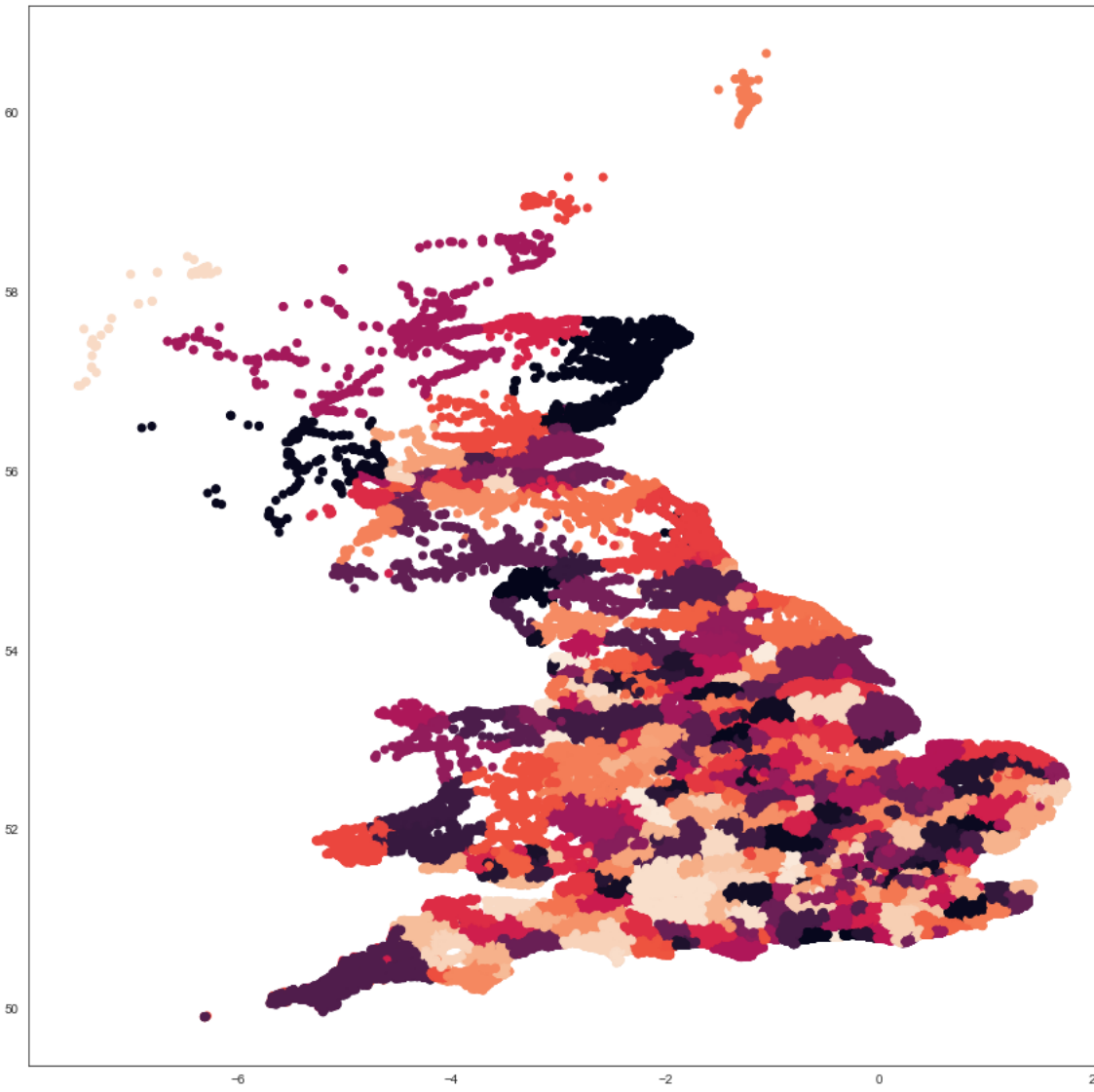
Slika 18: Kohonen nad skupom A



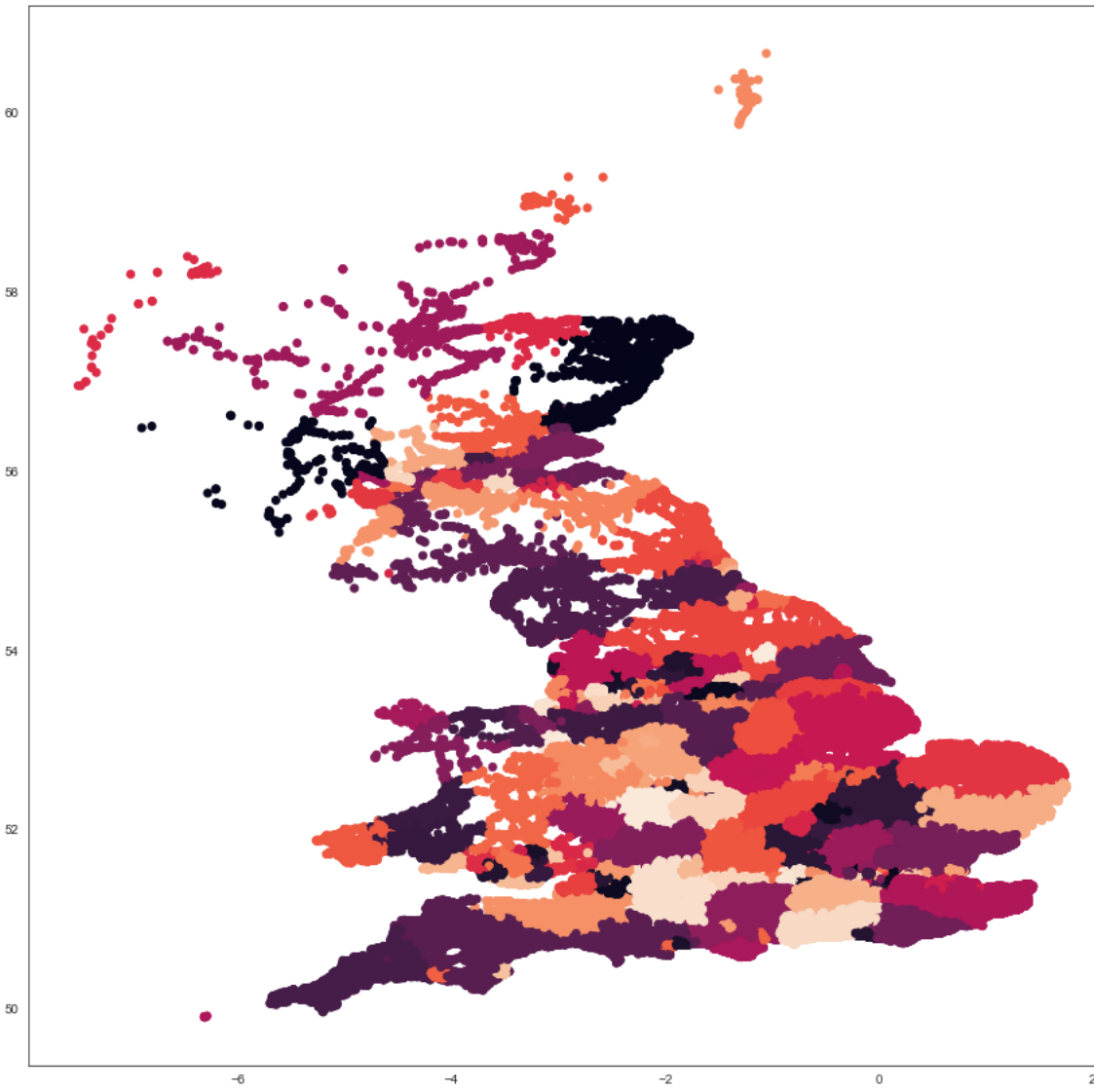
Slika 19: Kohonen nad skupom A



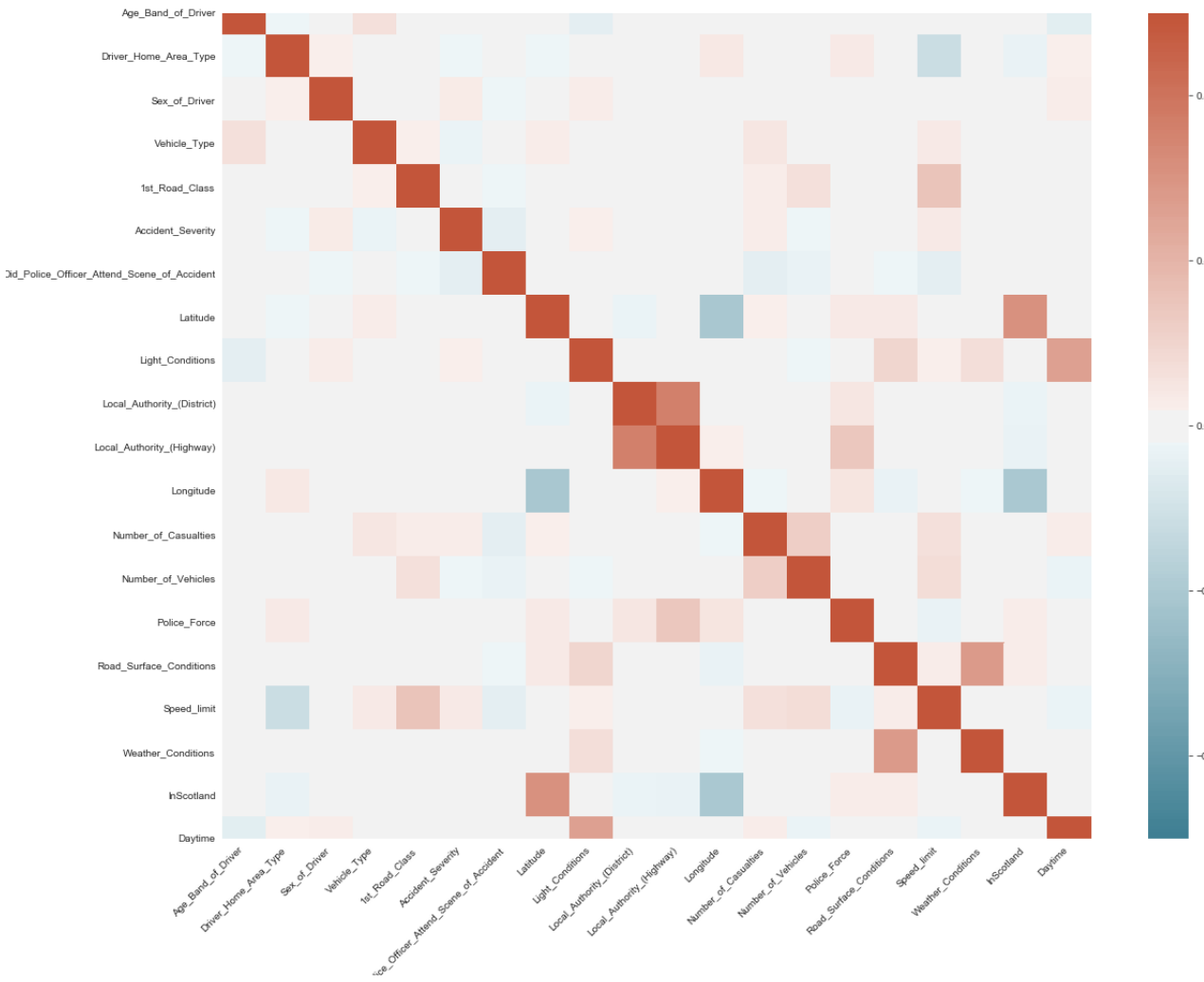
Slika 20: Broj nesreća po satu u danu



Slika 21: Broj nesreća po satu u danu



Slika 22: Broj nesreća po satu u danu



Slika 23: Matrica korelacije

Mera Bliskosti	Kriterijum blizine	Senka koeficijent
complete	manhattan	0.5097087763755905
complete	cosine	0.16750151022711413
complete	euclidean	0.4856298172240169
average	manhattan	0.5131697581898554
average	cosine	0.17375056896553856
average	euclidean	0.5131697581898554
single	manhattan	0.3422291925712226
single	cosine	0.17424488432428845
single	euclidean	0.34222919319883977
ward	euclidean	0.5201150420814326

Slika 24: Skup A

Mera Bliskosti	Kriterijum blizine	Senka koeficijent
complete	manhattan	0.568777956787152
complete	cosine	0.4599095757295255
complete	euclidean	0.5568353964183733
average	manhattan	0.4690401620371774
average	cosine	0.40953785490593625
average	euclidean	0.4563790186062456
single	manhattan	0.4733724709492776
single	cosine	0.42051193207808946
single	euclidean	0.5057733277849132
ward	euclidean	0.6748279826802646

Slika 25: Skup B

Mera Bliskosti	Kriterijum blizine	Senka koeficijent
complete	manhattan	0.6430676121495346
complete	cosine	0.39106511801624044
complete	euclidean	0.5837036614261488
average	manhattan	0.6029121812013397
average	cosine	0.4914907727372172
average	euclidean	0.5652597304303723
single	manhattan	0.6131801289777862
single	cosine	0.5577143555021572
single	euclidean	0.5597472427792276
ward	euclidean	0.6681278469117078

Slika 26: Skup C

Mera Bliskosti	Kriterijum blizine	Senka koeficijent
complete	manhattan	0.3718745999845559
complete	cosine	0.3316180511061783
complete	euclidean	0.3874155384367804
average	manhattan	0.3838092875388943
average	cosine	0.19726662136198822
average	euclidean	0.3122276796766169
single	manhattan	0.29813640690047716
single	cosine	0.18645570067420733
single	euclidean	0.18204031436291043
ward	euclidean	0.4313820280353939

Slika 27: Skup D