Diabetes 130-US hospitals for years 1999-2008

Projekat iz Istraživanja podataka 1 [R274]

Milica Tošić 105/2019

Jun 2023

Sadržaj

1 Uvod

U ovom radu bavimo se analizom baze podataka "Diabetes 130-US hospitals for years 1999-2008" . Ova baza podataka sadrži informacije o dijabetesu i bolničkom lečenju pacijenata sa dijabetesom u 130 bolnica u Sjedinjenim Američkim Državama od 1999. do 2008. godine. Nad ovim podacima izvršićemo razne algoritme klasifikacije i klasterovanja radi grupisanja pacijenata sa istim karakteristikama i radi nekih predikcija, a i takođe ćemo koristiti softer SPSS radi dubljeg razumevanja podataka i nekih statističkih analiza.

2 Podaci sa kojim radimo

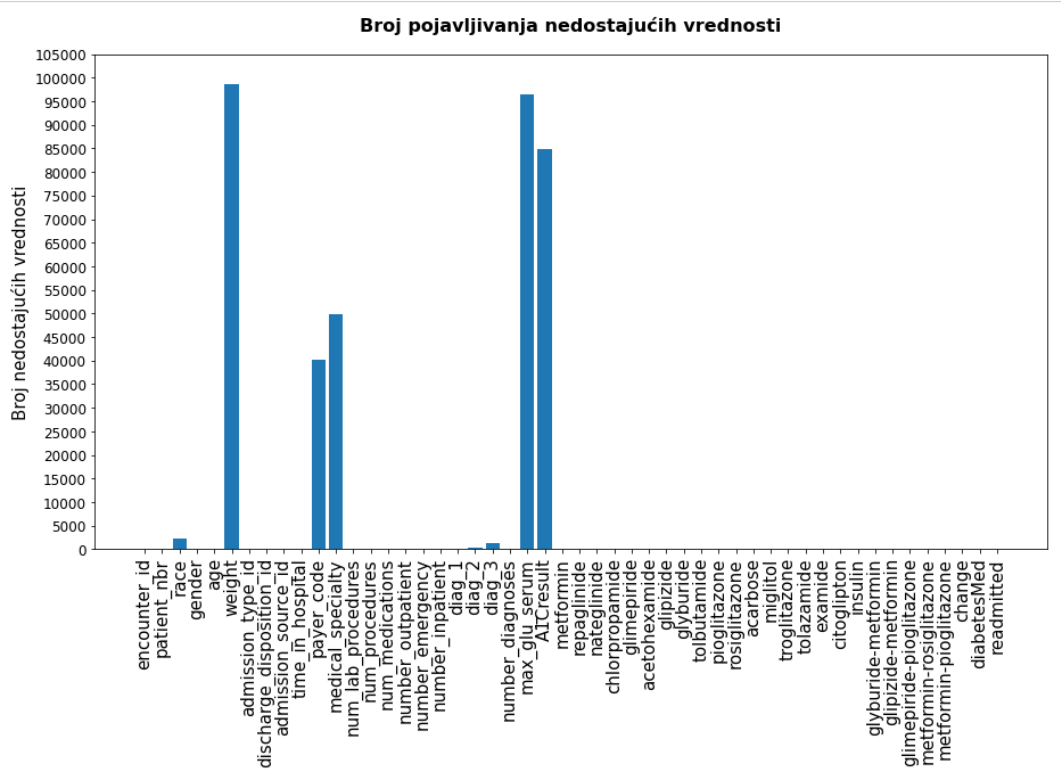
Podataka sa kojma radimo ima puno. Ova baza podataka ima 101,766 instanci i 50 atributa.

Atributi naše baze:
1. encounter_id: ID susreta
2. patient_nbr: broj pacijenta
3. admission_type _id: ID tipa prijema
4. discharge_disposition_id: ID stanja pri otpustu
5. admission_source_id: ID izvora prijema
6. time_in_hospital: vreme provedeno u bolnici (u danima)
7. num_lab_procedures: broj laboratorijskih procedura
8. num_procedures: broj medicinskih procedura
9. num_medications: broj propisanih lekova
10. number_outpatient: broj ambulantnih poseta
11. number_emergency: broj hitnih slučajeva
12. number_inpatient: broj hospitalizacija
13. number_diagnoses: broj dijagnoza
14. race: etnička pripadnost
15. gender: pol
16. age: starost
17. weight: težina
18. payer_code: kod plaćanja
19. medical_specialty: medicinska specijalnost
20. diag_1: prva dijagnoza
21. diag_2: druga dijagnoza
22. diag_3: treća dijagnoza
23. max_glu_serum: maksimalna vrednost glukoze u krvi
24. A1Cresult: rezultat A1C testa
25. metformin: upotreba metformina
26. repaglinide: upotreba repaglinida
27. nateglinide: upotreba nateglinida
28. chlorpropamide: upotreba hlorpropamida
29. glimepiride: upotreba glimepirida
30. acetohexamide: upotreba acetohexamida
31. glipizide: upotreba glipizida
32. glyburide: upotreba gliburida
33. tolbutamide: upotreba tolbutamida
34. pioglitazone: upotreba pioglitazona
35. rosiglitazone: upotreba rosiglitazona
36. acarbose: upotreba akarboze
37. miglitol: upotreba miglitola
38. troglitazone: upotreba troglitazona
39. tolazamide: upotreba tolazamida
40. examide: upotreba exameda
41. citoglipton: upotreba citogliptona
42. insulin: upotreba insulina
43. glyburide-metformin: kombinacija gliburida i metformina
44. glipizide-metformin: kombinacija glipizida i metformina
45. glimepiride-pioglitazone: kombinacija glimepirida i pioglitazona
46. metformin-rosiglitazone: kombinacija metformina i rosiglitazona
47. metformin-pioglitazone: kombinacija metformina i pioglitazona
48. change - kromena lekarskog tretmana
49. diabetesMed - da li je pacijent koristio lekove za dijabetes
50. readmitted - donovni prijem (da li je pacijent ponovo primljen u bolnicu)

Prvih 13 atributa su numerički, a svi ostali su kategorički.

3 Priprema podataka za klasifikaciju

3.1 Nedostajuće vrednosti



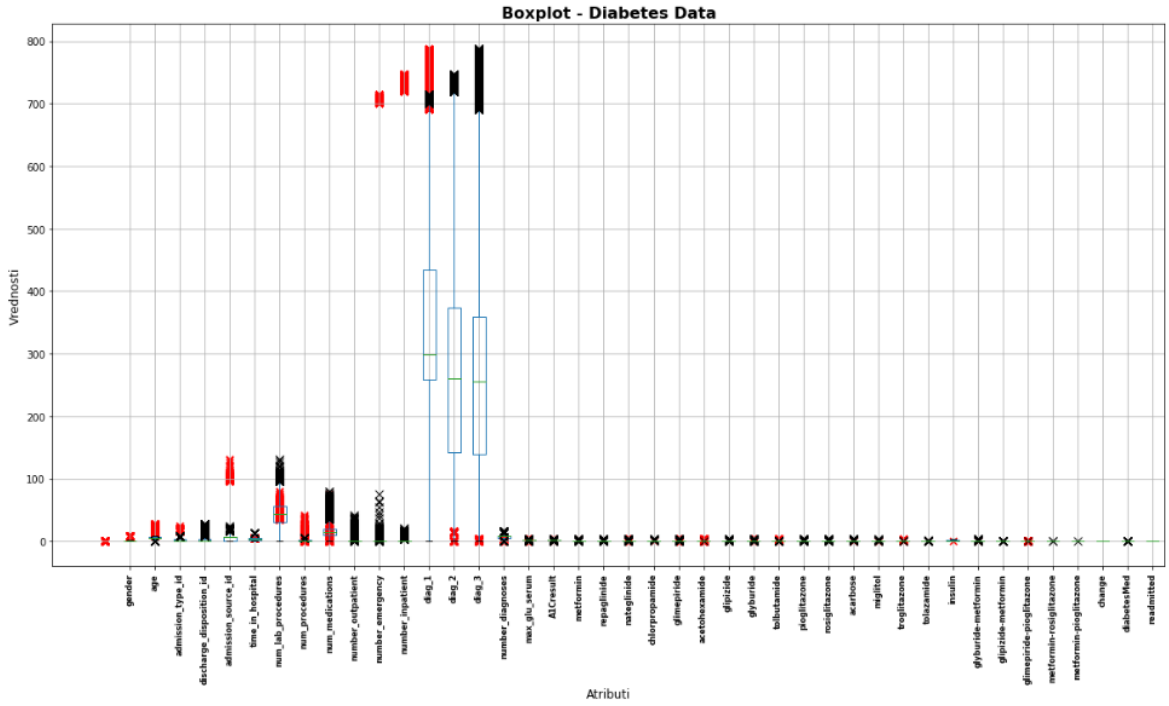
Grafikon koji prikazuje nedostajuće vrednosti.

Dodatnom proverom zvanično zaključujemo da su atributi: race, weight, payer_code, medical_specialty, diag_1, diag_2 , diag_3, max_glu_serum, A1Cresult oni atributi koji imaju nedostajuće vrednosti. Neki od ovih atributa mogu biti manje relevantni za ciljnu varijablu readmitted : race (etnička pripadnost), weight (težina) i payer_code (kod osiguravajuće kompanije), pa bi se mogli obrisati iz skupa atributa. S druge strane, atributi diag_1 (primarna dijagnoza), diag_2 (sekundarna dijagnoza), diag_3 (tercijarna dijagnoza), max_glu_serum (najviša izmerena glikemija) i A1Cresult (prosečna glikemija) mogu biti važni jer se tiču dijagnostike i tretmana dijabetesa, što može biti povezano sa ponovnim prijemom u bolnicu, pa te nedostajuće vrednosti menjamo sa najčešćom vrednošću u koloni.
Pored navedenih kolona možemo da obršemo i kolone encounter_id i patient_nbr kojima je (po informacijama sa sajta baze) 'uloga' ID, pa ih iz tog razloga možemo obrisati. Takođe, još dve kolone koje možemo obrisati su examide, citoglipton jer imaju samo vrednost 'No' i time ne pružaju nikakvu korisnu informaciju za dalju analizu.

3.2 Enkodiranje

Radi lakšeg rada nad podacima, kategoričke atribute ćemo konvertovati u numeričke. Koristićemo LabelHotEncoder. LabelEncoder dodeljuje numeričke vrednosti jedinstvenim kategoričkim vrednostima, bez potrebe za kreiranjem dodatnih kolona, kao na primer OneHotEncoder. Ovo može rezultirati bržom obradom, što je jedan od razloga zašto sam se ja i odlučila za njega(OHE radio jako dugo - skoro sat vremena).

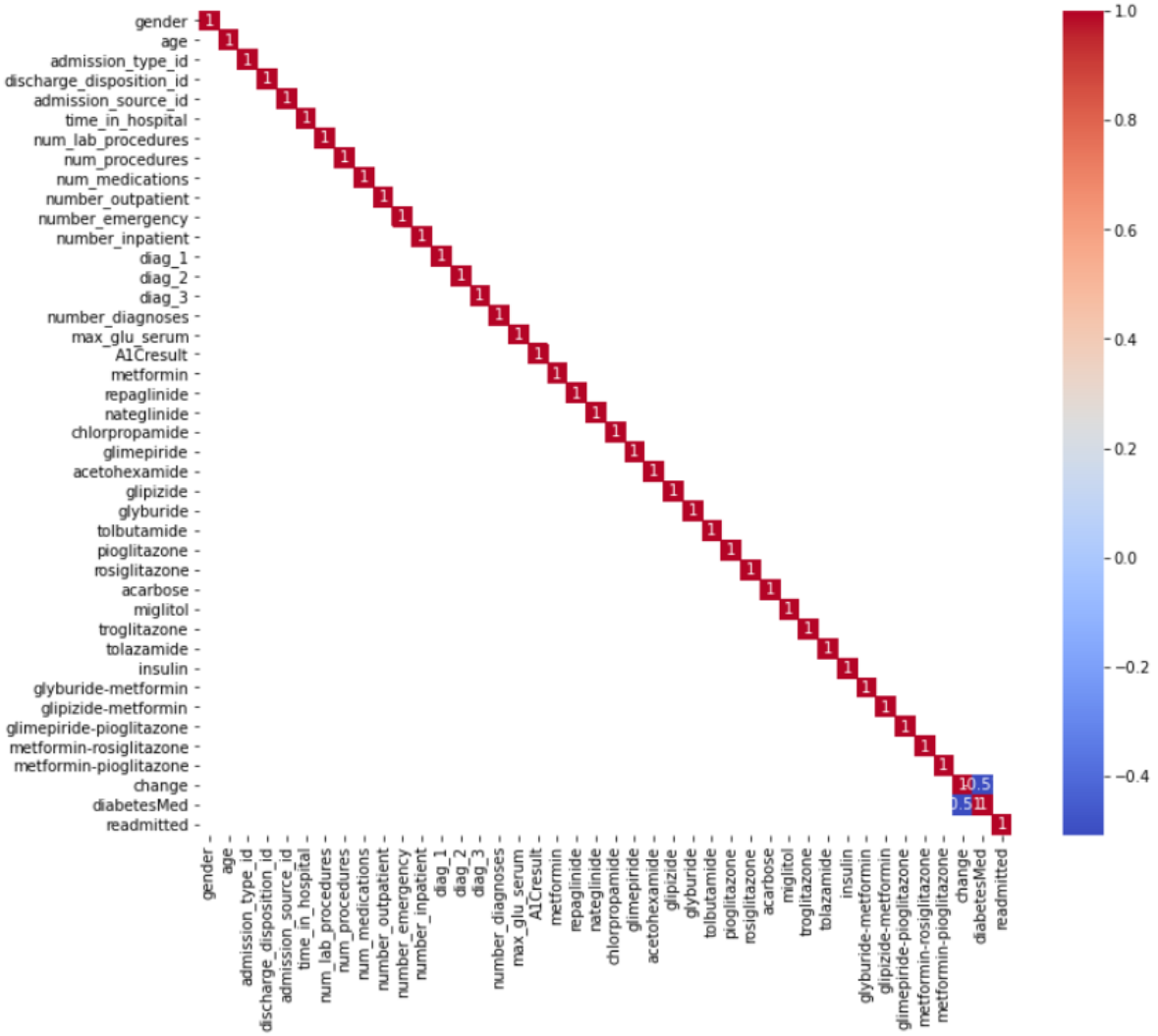
3.3 Rad sa outlier-ima



Boxplot koji prikazuje outlire-e.

Outlier-e brišemo komandom data=data[outliers].

3.4 Provera međusobne korelisanosti između atributa



Heatmap-a koja pokazuje međusobnu korelisanost atributa sa pragom 0.5.

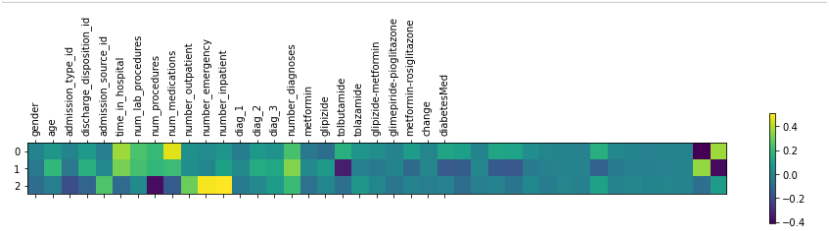
Kao što vidimo na slici ni jedan atribut nije jako koreliran sa nekim drugim (da je prag na primer veći od 0.7 ili 0.8), stoga na osnovu ovoga ne možemo zaključiti da li još neke dodatne kolone možemo obrisati.

3.5 Feature selection

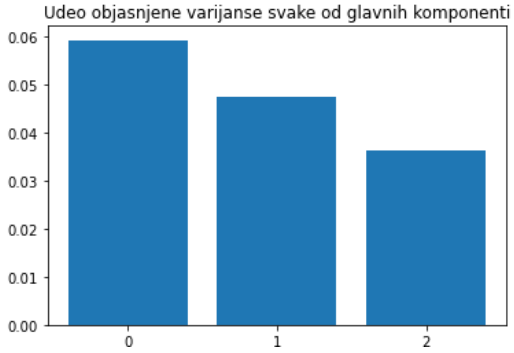
Koristeći funkcije SelectKBest, chi2 uradili smo feature selection da bismo odredili koji atributi su bitni, a koji ne. Za k sam stavila 25 jer je to malo više od 50% atributa, da ne bih imala gubitak velikog broja inforamcija, ali i da opet pokušamo da optimizujemo donekle. Podatke koji su preostali(koje feature selection nije izvukao) sam izbrisala.

3.6 PCA algoritam

Iskorišćen PCA algoritam. Glavne komponente su linearna kombinacija polaznih atributa . I njihove koeficijente možemo videti na prvoj slici ispod.



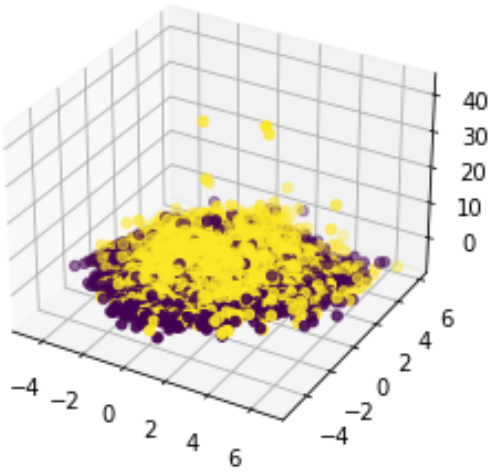
Matrica komponenata.



Udeo objašnjene varijanse.

Podaci sami po sebi očigledno imaju nisku varijabilnost. Prva komponenta može objasniti samo $\sim 6\%$, druga $\sim 5\%$, a treća $\sim 4\%$ varijabilnosti(raznolikost ili rasprostranjenost podataka) u podacima.

Kumulativna suma udela objašnjene varijanse.



Vizuelizacija podataka.

Glavne komponente objašnjavaju 14% ukupne varijanse podataka.

4 Klasifikacija

Podatke smo podelili na trening i test preko funkcije train_test_split importovane iz sklearn.model_selection. Takođe, podatke smo normalizovali koristeći funkciju Normalizer importovanu iz sklearn.preprocessing.

4.1 Stablo odlučivanja (DecisionTreeClassifier)

Koristeći GridSearchCV iz sklearn.model_selection dobijamo sledeći skup parametara i accuracy tj. tačnost:
DecisionTreeClassifier(criterion='entropy', max_depth=5)
{'criterion': 'entropy', 'max_depth': 5, 'min_samples_split': 10}

8811	2162
5382	3999

Matrica konfuzije za stablo odlučivanja.

4.2 Slučajne šume (RandomForestClassifier)

Kombinuje više stabala odlučivanja. Svako drvo se trenira na slučajno izabranom podskupu podataka, a klasifikacija se obavlja na osnovu glasanja svakog drveta. Random Forest ima sposobnost smanjenja preprilagođavanja i poboljšanja sposobnosti rada nad neviđenim podacima, u odnosu na pojedinačno stablo odlučivanja.

Koristeći GridSearchCV dobijamo:
RandomForestClassifier(min_samples_split=10, n_estimators=300)
max_depth: None, min_samples_split: 15, n_estimators: 300

8170	2803
4446	4935

Matrica konfuzije za slučajne šume.

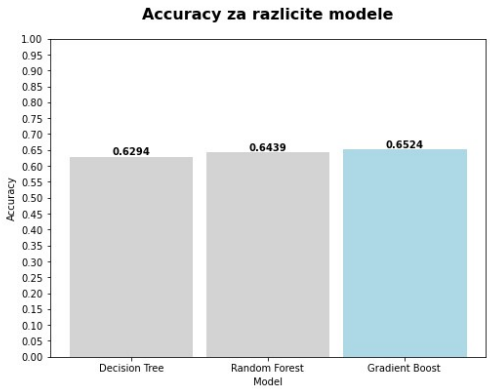
4.3 Klasifikator pojačavanjem gradijenta (GradientBoostingClassifier)

U svakoj iteraciji, algoritam računa grešku između trenutnih predikcija modela i stvarnih vrednosti ciljne promenljive. Zatim se računa gradijent ove greške kako bi se odredio pravac u kojem treba poboljšati model.

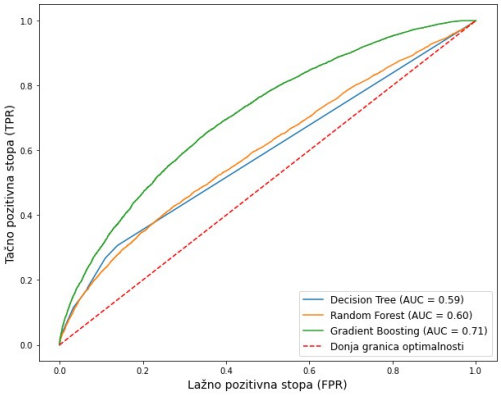
Koristeći GridSearchCV dobijamo:
GradientBoostingClassifier(learning_rate=0.1,min_depth = 5,min_samples_split=10,
n_estimators=300)
'max_depth': None, 'min_depth: 5, 'n_estimators': 300

8096	2877
4198	5183

Matrica konfuzije za klasifikator pojačavanjem gradijenta.



Tačnost za sva 3 modela.



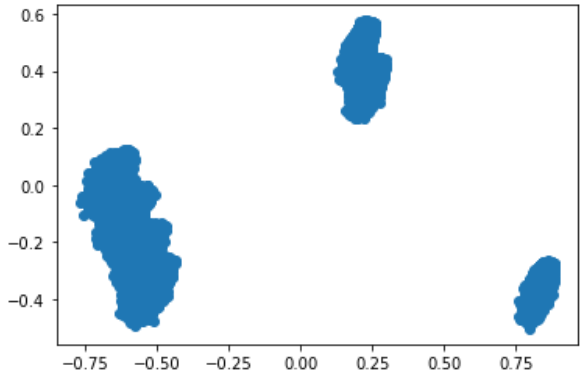
Poređenje modela preko ROC krive.

Donja granica optimalnosti na grafiku ROC krive je zapravo prava y=x. Ne bismo voleli da kriva ide ispod te prave, odnosno voleli bismo da kriva bude sto bliza gornjem levom ćošku, tj. tački (0,1).

5 Klasterovanje

Kao i kod klasifikacije, odradićemo preprocesiranje: brisanje nedostajućih vrednosti, enkodiranje, feature selection.

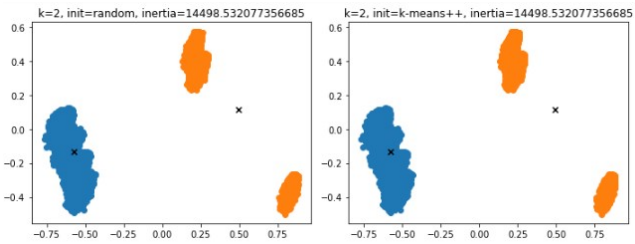
Takođe, iskorišćen je i PCA algoritam. Na sledećoj slici vidimo da su podaci, raspoređeni tako, da ćemo najverovatnije koristiti tri klastera u algoritmima klasterovanja. MinMaxScaler omogućava normalizaciju svih atributa na isti opseg, čime se obezbeđuje da svi atributi imaju sličan uticaj prilikom PCA analize.



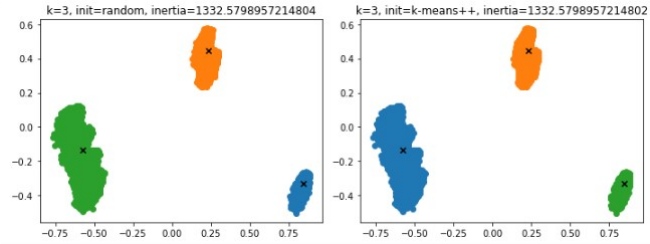
PCA algoritam.

5.1 KMeans - algoritam k sredina

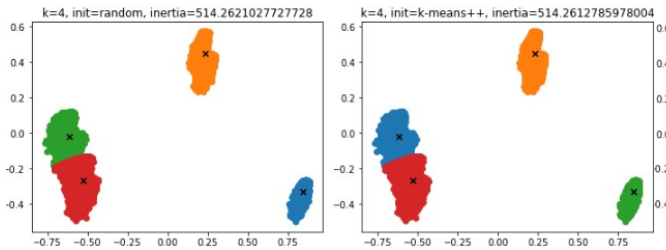
Iskorišćen algoritam K sredina. Takođe, preko plota, u for petlji, prikazana je promena sa povećanjem broja klastera. Već preko PCA smo mogli da pretpostavimo da će optimalni broj klastera biti 3, a to možemo videtu i ovde.



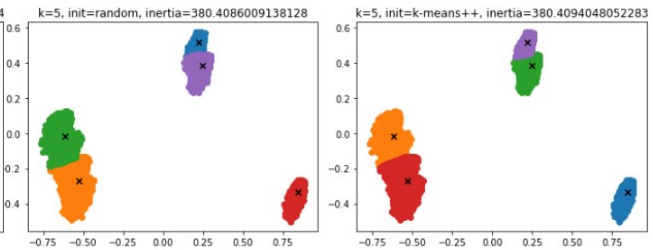
KMeans za k=2.



KMeans za k=3.

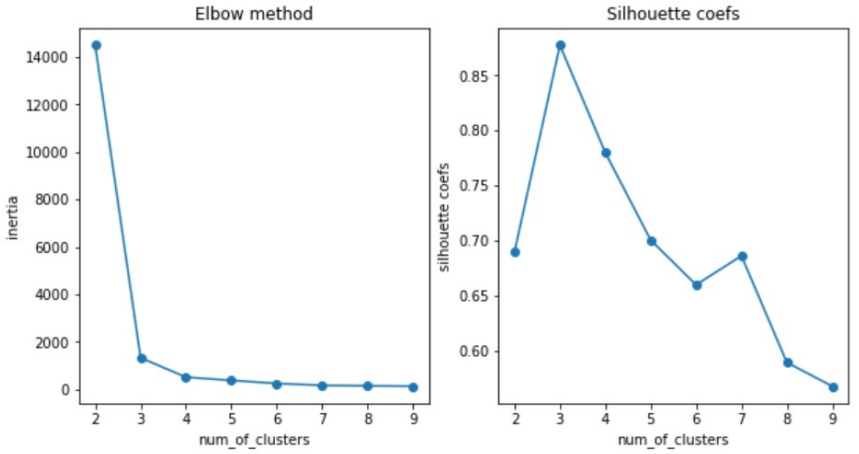


KMeans za k=4.



KMeans za k=5.

Najpre ćemo naš zaključak potvrditi metodom lakta. Posmatramo prvi broj posle kog pad više nije jako strm. To je broj 3.
Na slici ispod, takođe je prikazan i grafik koeficijenata siluete na kom se takođe najbolje vidi da je vrednost siluete najveća kada se koriste 3 klstera.



Metod lakta i koeficijenti siluete

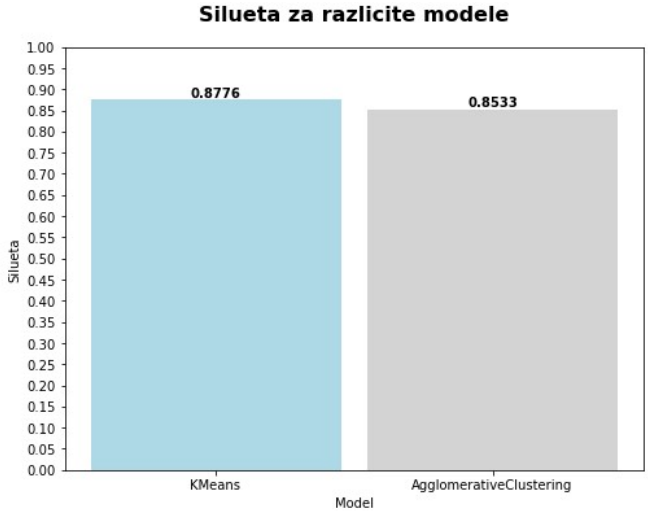
```
[0.6901420642124987,
 0.877571248132134,
 0.7800686438336438,
 0.7007530691313438,
 0.6601227571613532,
 0.6864072917084068,
 0.5896503743910007,
 0.5680922011162881]
```

Vrednosti siluete prikazane numerički.

5.2 AgglomerativeClustering - Sakupljajuce klasterovanje

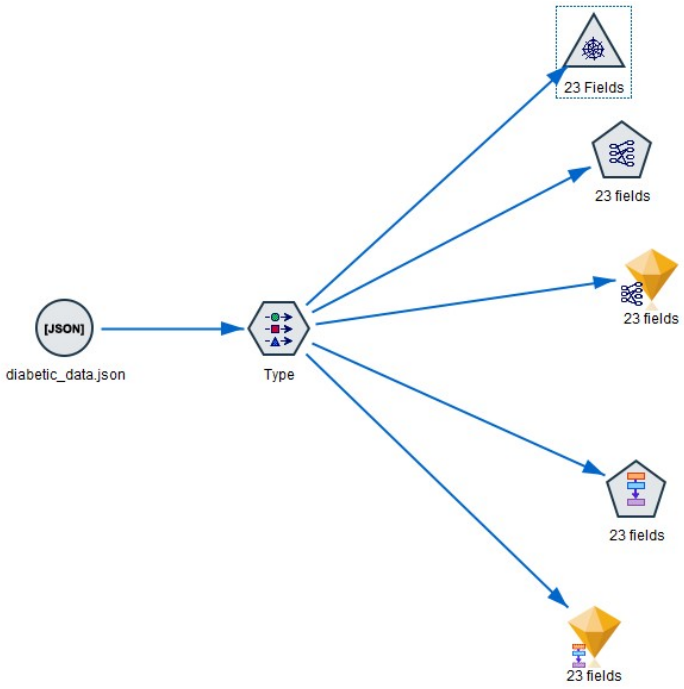
Iskorišćen algoritam Aglomerativnog klasterovanja sa atributima (n_clusters=3, linkage='single', compute _distances=True).

linkage='single': za spajanje klastera koristi se najmanja udaljenost između tačaka u dva klastera.
compute _distances=True: ovaj atribut određuje da li algoritam treba da računa udaljenosti između tačaka prilikom izvršavanja klasterovanja. Ako je postavljeno na True, udaljenosti će biti izračunate i korišćene u procesu klasterovanja. Ako je postavljeno na False, algoritam će koristiti samo na primer, povezanost tačaka u dendrogramu za formiranje klastera, a neće računati konkretne udaljenosti.
Vrednost siluete za ovaj algoritam je 0.8532505426841923, što je takodje, kao i kod K sredina, na zavidnom nivou.



Poredjenje vrednosti siluete za k=3 za oba algoritma.

6 Pravila pridruživanja



Primenjeni algoritmi u SPSS-u

6.1 Priprema podataka

Promenjen tip podataka tako da oni koji su na primer segment[a,b] postane 'Nominal' sa svim celobrojnim čvorovima tog segmenta. Na primer za atribut 'num_procedures' ima vrednosti iz segmenta [0,6], a posle ove promene uzimaće redom vrednosti 0,1,2,3,4,5, i 6. To je uradjeno da bi Apriori algoritam mogao da radi nad tim podacima. Algoritam Association rules može da radi u oba slučaja, pa smo obradili i njega. Atributi 'diag_1', 'diag_2', 'diag_3' su takvi da su im neke vrednosti kombinacija slova i brojeva pa ne mogu biti prepoznate. Njih smo izostavili iz rada. Iskorišćeno je 23 atributa, dobijenih preko Feature Selection-a u prethodne dve oblasti(bez 'diag_1', 'diag_2', 'diag_3').

6.2 Apriori algoritam

Apriori algoritam je algoritam koji u fazi generisanja čestih skupova stavki koristi osobine podrške kako bi se smanjio broj skupova stavki za koje je potrebno izračunati podršku da bi se odredilo da li je skup stavki čest.

Podrška određuje koliko se često pravilo pojavljuje u transakcijama skupa podataka:

$$\sup(X \to Y) = \frac{\sigma(X \cup Y)}{N}$$

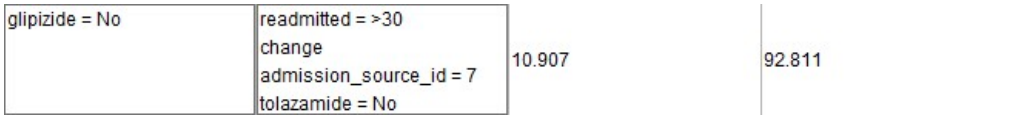
6.2.1 Neka od zanimljivih pravila dobijenih Apriori algoritmom

diabetesMed	glipizide = Steady tolazamide = No	11.157	100.0

Ako pacijent koristi lekove za dijabetes, u 100% slučajeva se koristi stabilna količina leka glipzide, a lek tolazamide se ne koristi.

tolazamide = No	age = [60-70) number_diagnoses = 9	10.949	100.0

Ako pacijent ne koristi lek tolazamide, onda je to u 100% slučajeva pacijent starosne dobi od 60 do 70 godina, koji je do sada bio dijagnostikovan 9 puta.



glipizide = No	readmitted = >30 change admission_source_id = 7 tolazamide = No	10.907	92.811

Ako pacijent ne koristi lek glipzide, onda je to u skoro 93% slučajeva pacijent koji je ponovo primljen u bolnicu posle više od 30 dana, koji je takođe nedavno promenio svoj lekarski tretman, koji ne koristi lek tolazamide i koji ima ID prijema 7.

Kao što smo i mogli da pretpostavimo, najviše pravila koja smo dobili su vezana za lekove i njihovo korišćenje, jer, iz nekog minimalnog medicinskog znanja možemo zaključiti da pacijent ne može koristiti neke lekove u određenom periodu života, ne može mešati više terapija i slično.

6.3 Association rules

6.3.1 Neka od zanimljivih pravila dobijenih Association rules-om

Most Interesting Rules by Confidence

					Other Evaluation Statistics			
Rank	Rule ID	Condition	Prediction	Sorted By Confidence(%)	Condition Support (%)	Rule Support (%)	Lift	Deployability (%)
1	1	admission_type_id = 1 number_outpatient = 0 metformin = Steady glipizide = No	admission_source_id = 7 diabetesMed	91.38	6.29	5.75	2.10	0.54
2	2	admission_type_id = 1 number_outpatient = 0 metformin = Steady	admission_source_id = 7 diabetesMed	91.32	7.63	6.96	2.10	0.66
3	3	admission_type_id = 1 metformin = Steady glipizide = No	admission_source_id = 7 diabetesMed	91.30	7.44	6.79	2.10	0.65
4	4	admission_type_id = 1 metformin = Steady	admission_source_id = 7 diabetesMed	91.23	9.05	8.26	2.10	0.79
5	5	admission_type_id = 1 discharge_disposition_id = 1 metformin = Steady	admission_source_id = 7 diabetesMed	91.22	5.85	5.34	2.10	0.51

Par dobijenih pravila koja se nalaze u segmentu onih sortiranih po pouzdanosti.

Most Interesting Rules by Rule Support

					Other Evaluation Statistics			
Rank	Rule ID	Condition	Prediction	Sorted By Rule Support(%)	Condition Support (%)	Confidence (%)	Lift	Deployability (%)
1	804	admission_source_id = 7 number_outpatient = 0 change	admission_type_id = 1 number_emergency = 0 metformin = No glipizide = No	17.26	25.46	67.79	2.01	8.20
2	949	admission_source_id = 7 number_emergency = 0 change	admission_type_id = 1 number_outpatient = 0 metformin = No glipizide = No	17.26	26.50	65.13	2.01	9.24
3	145	admission_type_id = 3	admission_source_id = 1	15.91	18.54	85.79	2.95	2.64
4	125	admission_type_id = 3 number_emergency = 0	admission_source_id = 1	14.74	17.12	86.13	2.96	2.38
5	346	admission_type_id = 3	admission_source_id = 1 number_emergency = 0	14.74	18.54	79.51	2.96	3.80
6	139	admission_type_id = 3 glipizide = No	admission_source_id = 1	13.91	16.19	85.90	2.96	2.28
7	516	admission_type_id = 3	admission_source_id = 1 glipizide = No	13.91	18.54	75.00	2.95	4.64

Par dobijenih pravila koja se nalaze u segmentu onih sortiranih po podršci.

7 Zaključak

Informacije su izvučene iz baze podataka za prijeme koji zadovoljavaju sledeće kriterijume:
- To su hospitalizacije (prijem u bolnicu).
- To su susreti pacijenata sa dijabetesom, odnosno oni tokom kojih je bilo uneto bilo kakvo stanje dijabetesa u sistem kao dijagnoza.
- Dužina boravka u bolnici je bila najmanje 1 dan, a najviše 14 dana.
- Laboratorijski testovi su izvršeni tokom prijema.
- Lekovi su primenjeni tokom susreta.
- Na osnovu naših podataka možemo sa najviše 65% sigurnosti da predvidimo ponovni prijem pacijenta u bolnicu.
- Lekovi koji se koriste za dijabetes tipa 2 su uzajamno povezani i medjusobno zavise jedan od drugog - ne mogu se koristiti različiti lekovi istovremeno i slično.

8 Literatura

Kaggle: Diabetes Dataset
Github repozitorijum kursa Istraživanje podataka 1: Link ka githubu
Scikit-learn Machine Learning in Python: Link ka sajtu