Analiza
“Drug Consumption Data Set”
Skupa podataka

Anđela Janošević 315/2016

Anđela Križan 136/2016

Septembar 2019

Sadržaj

Sažetak

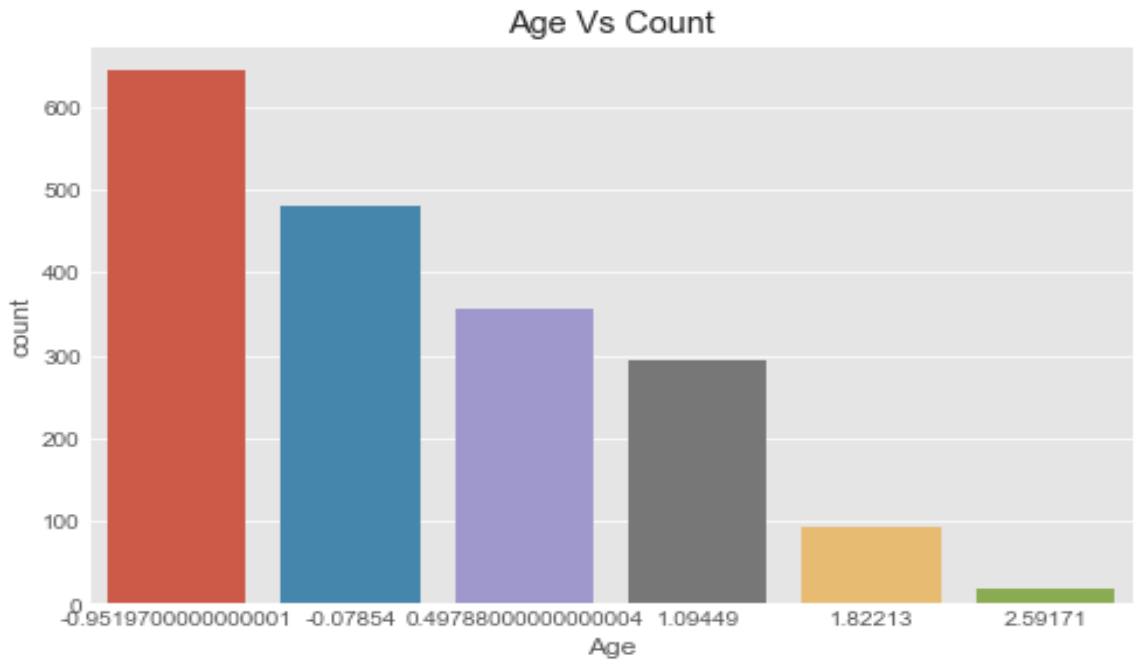
U ovom radu predstavićemo rezultate dobijene anketiranjem na temu konzumacije određenih supstanci. Prvo ćemo opisati dobijeni skup podataka. Zatim ćemo opisati postupak pretprocesiranja podataka. Uporedićemo rezultate za istu klasifikaciju, dobijene uz pomoć različitih algoritama.

Opis i vizualizacija skupa podataka

"Drug Consumption Data Set" je skup podataka koji sadrži zapise za 1885 ispitanika. Za svakog ispitanika je poznato 12 atributa opisa ličnosti:

- NEO-FFI-R (neurotičnost, ekstraverzija, otvorenost za iskustvo, prihvatljivost i savesnost)
- BIS-11 (impulsivnost)
- ImpSS (težnja za osećajem)
- Nivo obrazovanja
- Starost
- Pol
- Država prebivališta
- Nacionalnost

Među ispitanicima najviše ima ljudi starosti od 18 do 24 godina.

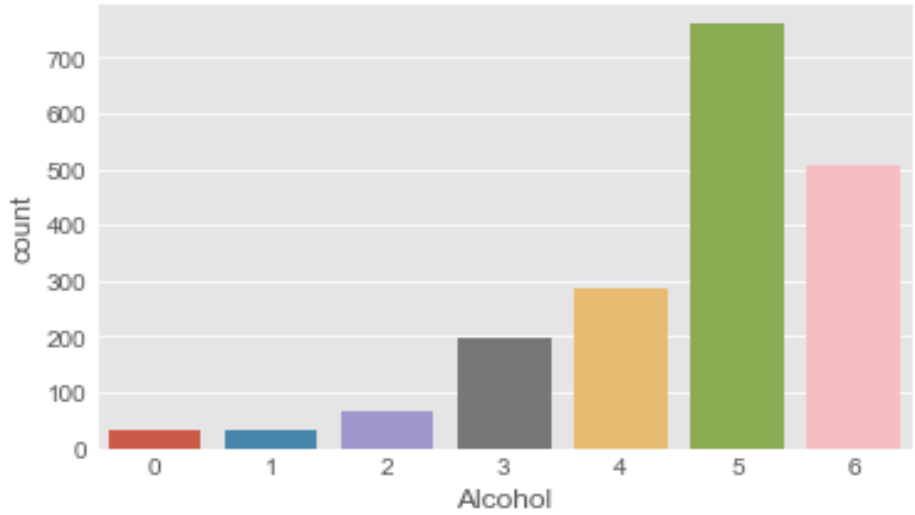


Učesnici su ispitivani u vezi njihove upotrebe 18 legalnih i ilegalnih droga:

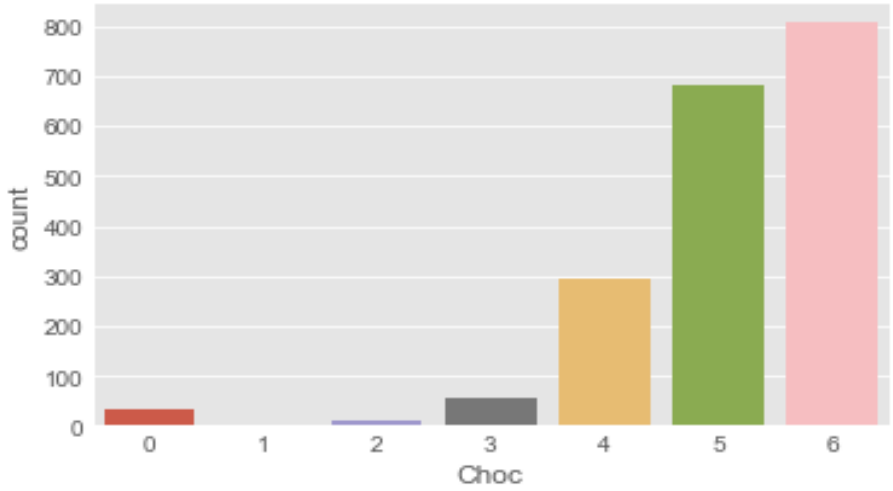
- Alkohol
- Amfetamin
- Amil nitrat
- Benzodiazepin
- Kanabis
- Čokolada
- Kokain
- Kofein
- Krek
- Ekstazi
- Heroin
- Ketamin
- Legalne droge
- LSD
- Metadon
- Gljive
- Nikotin
- Semeron

Za svaku supstancu ispitanik odgovora o periodu konzumiranja: nikada(CL0), pre više od decenije (CL1), u poslednjoj deceniji (CL2), u poslednjih godinu dana (CL3), u poslednjih mesec dana(CL4),u poslednjih nedelju dana (CL5) i prethodni dan (CL6). U nastavku su histogrami koji prikazuju broj ispitanika za svaki period konzumiranja. Izabrali smo dve lakse i dve teže supstance za prikaz.

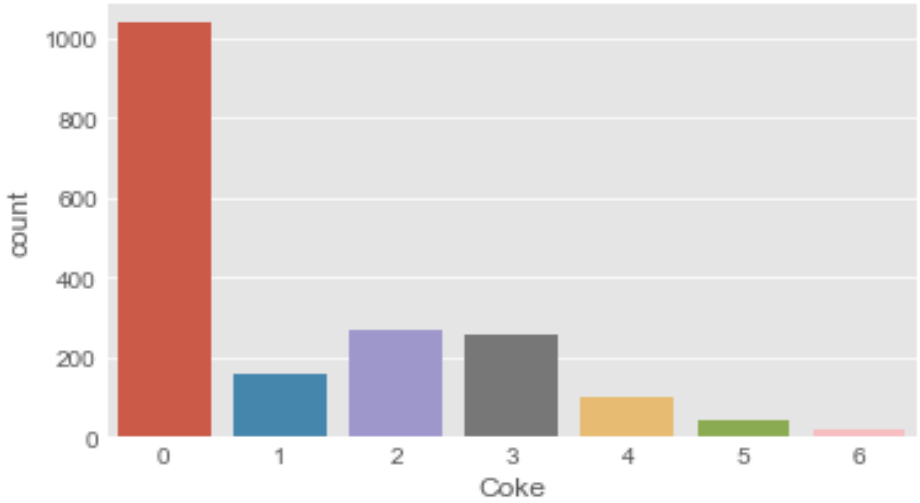
- Alkohol



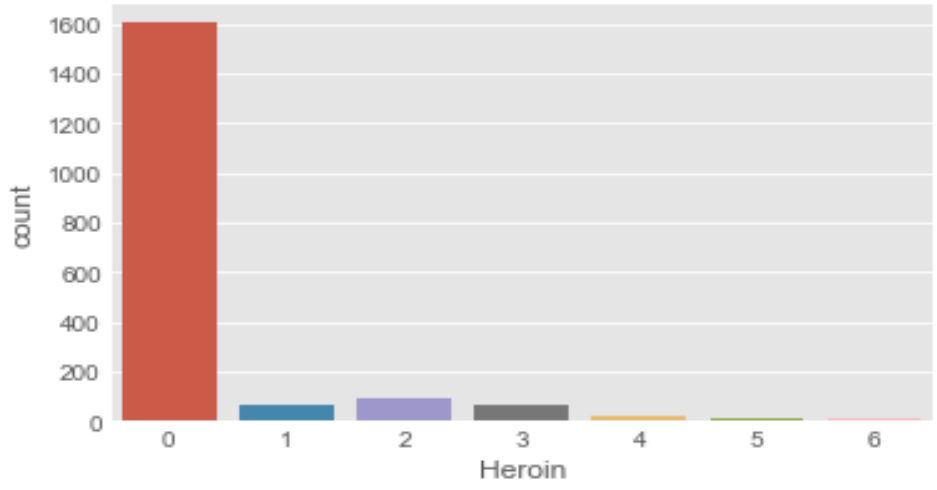
- Čokolada



- Kokain



- Heroin



Korišćeni alati

Za pretprocesiranje koristile smo jezik Python, biblioteke NumPy i Pandas. Za klasifikaciju koristile smo IBM SPSS i jezik Python, biblioteka Scikit-learn. Biblioteka Matplotlib jezika Python kao i IBM SPSS korišćeni su za vizualizaciju atributa.

Pretprocesiranje

Originalni podaci su bili sa ekstenzijom .data te je bilo potrebno prebaciti ih u CSV format. Skup podataka nije sadržao nedostajuce vrednosti. Obzirom da su nazivi atributa nedostajali ubacile smo ih u prvu vrstu. Radi lakšeg rukovanja podacima, klase CL0-CL6 smo zamenile brojevima 0-6. Mogle smo da rešimo problem korišćenjem svih 7 klasa, ali zbog prirode problema smo uradile binarizaciju i to:

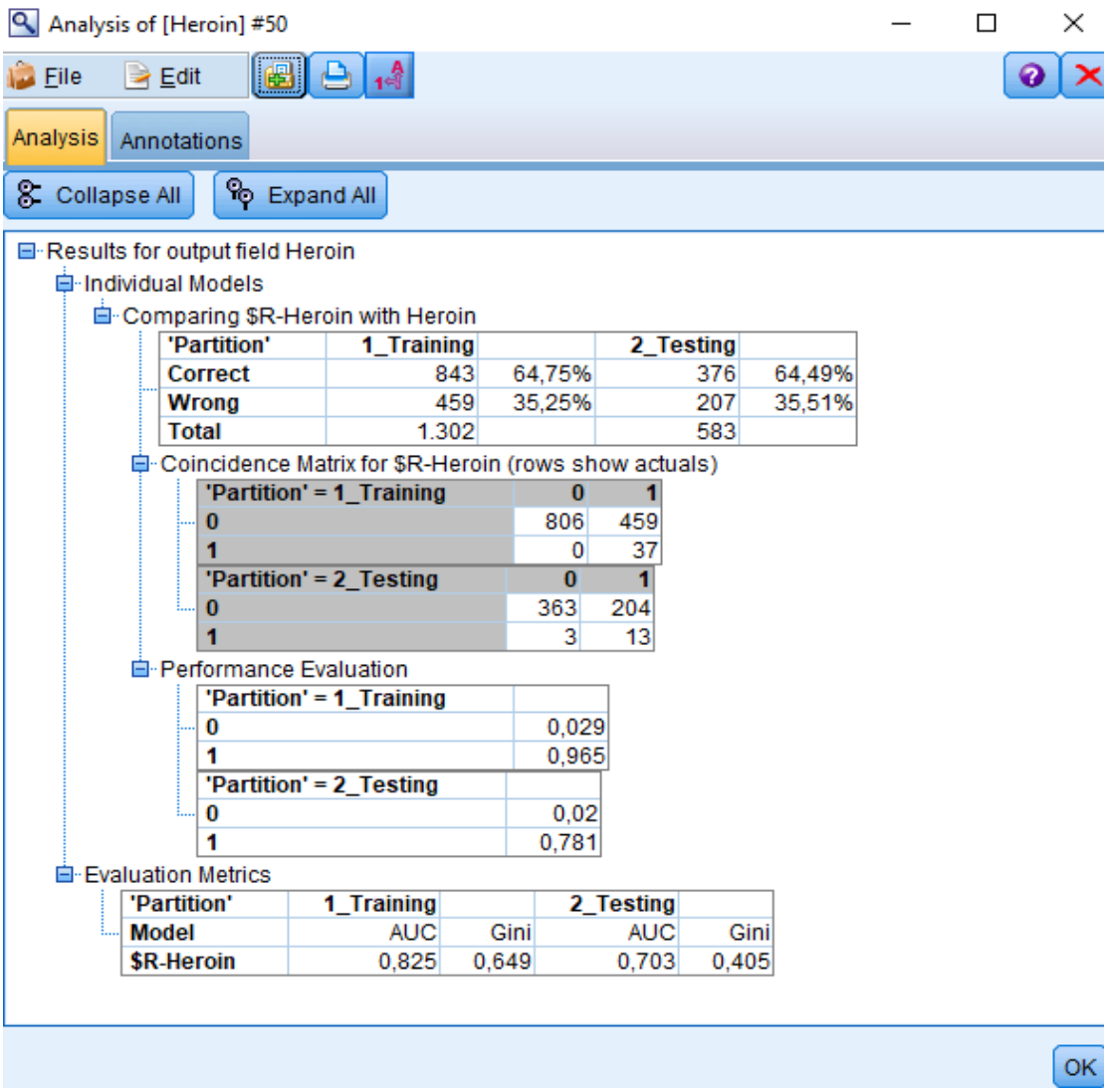
- Za lakše supstance kao što su alkohol,nikotin,čokolada i kofein dodelile smo klasu 1 osobama koje su ih koristile u poslednjih nedelju dana, ostalima je dodeljena klasa 0.
- Za ostale, teže supstance, dodelile smo klasu 1 osobama koje su konzumirale u poslednjih mesec dana, dok je ostalima dodeljena klasa 0.

Skup za treniranje čini 70% podataka, a skup za testiranje 30%.

Klasifikacija

Za klasifikaciju smo koristile više algoritama. U Python-u smo koristile Random Forest, Decision Tree, Cross Validation i KNN algoritam. Preciznost svakog algoritma smo prikazale kao prosek preciznosti svake supstance posebno. Primetile smo da za supstance koje malo ljudi koristi, kao što je heroin, algoritam radi sa velikom preciznošću, uz jedan problem. Algoritam ne klasifikuje dobro zavisnike, ali to ne utiče na preciznost jer ih ima malo. Nismo uspele da pronađemo biblioteku koja bi podržala menjanje cene misklasifikacije, zbog toga smo nastavile klasifikaciju u IBM SPSS-u. Pored navedenih algoritama, u SPSS-u smo koristile i SVM,CRT,Neural Net.

Kao primer klasifikacije težih supstanci uzele smo heroin i kokain i dobile sledeće rezultate:



'Partition'	1_Training		2_Testing	
Correct	843	64,75%	376	64,49%
Wrong	459	35,25%	207	35,51%
Total	1.302		583	

'Partition' = 1_Training	0	1
0	806	459
1	0	37

'Partition' = 2_Testing	0	1
0	363	204
1	3	13

'Partition' = 1_Training	
0	0,029
1	0,965

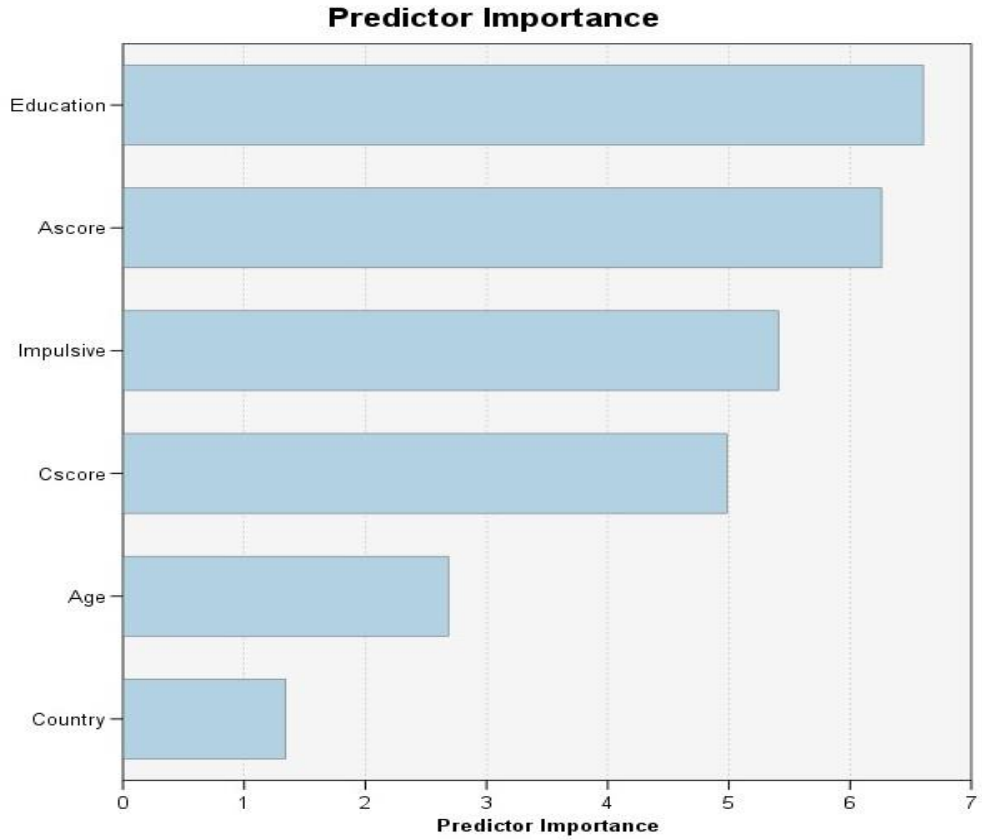
'Partition' = 2_Testing	
0	0,02
1	0,781

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$R-Heroin	0,825	0,649	0,703	0,405

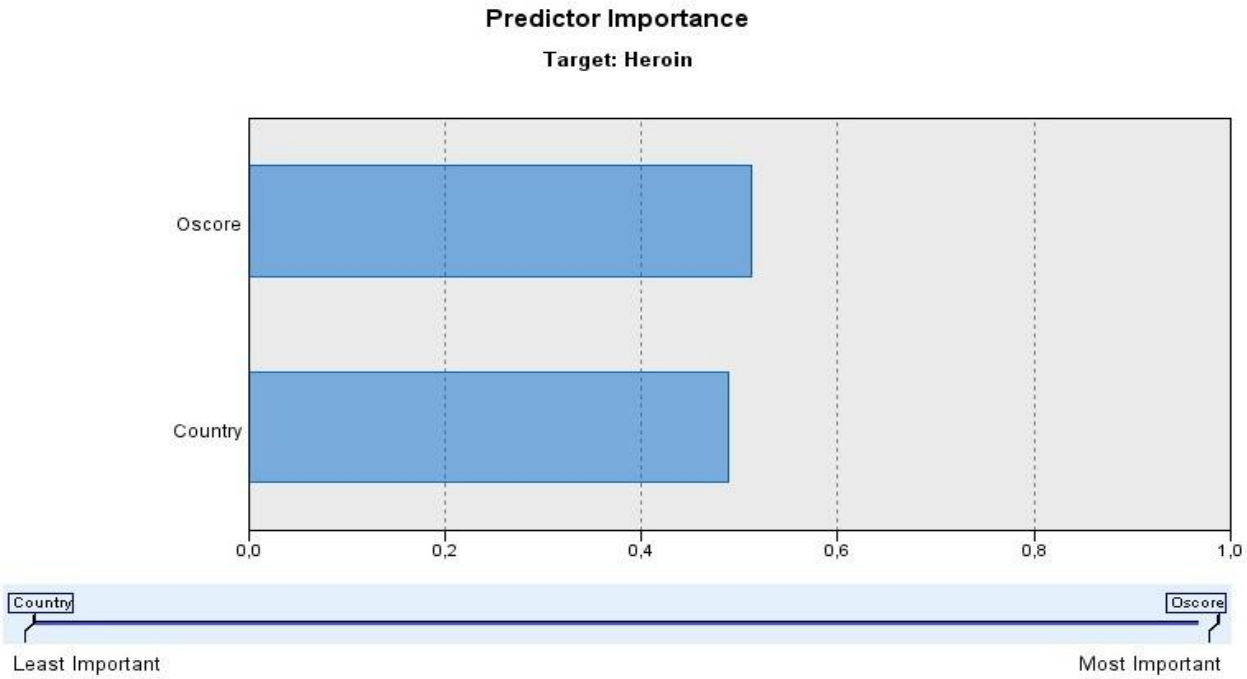
Uvođenjem cena misklasifikacija i smanjivanjem verovatnoće pronalaska zavisnika, preciznost se smanjila, ali smo propustile samo 3 od 16 zavisnika heroina.

Redosled važnosti atributa u zavisnosti od algoritma prikazani su u nastavku.

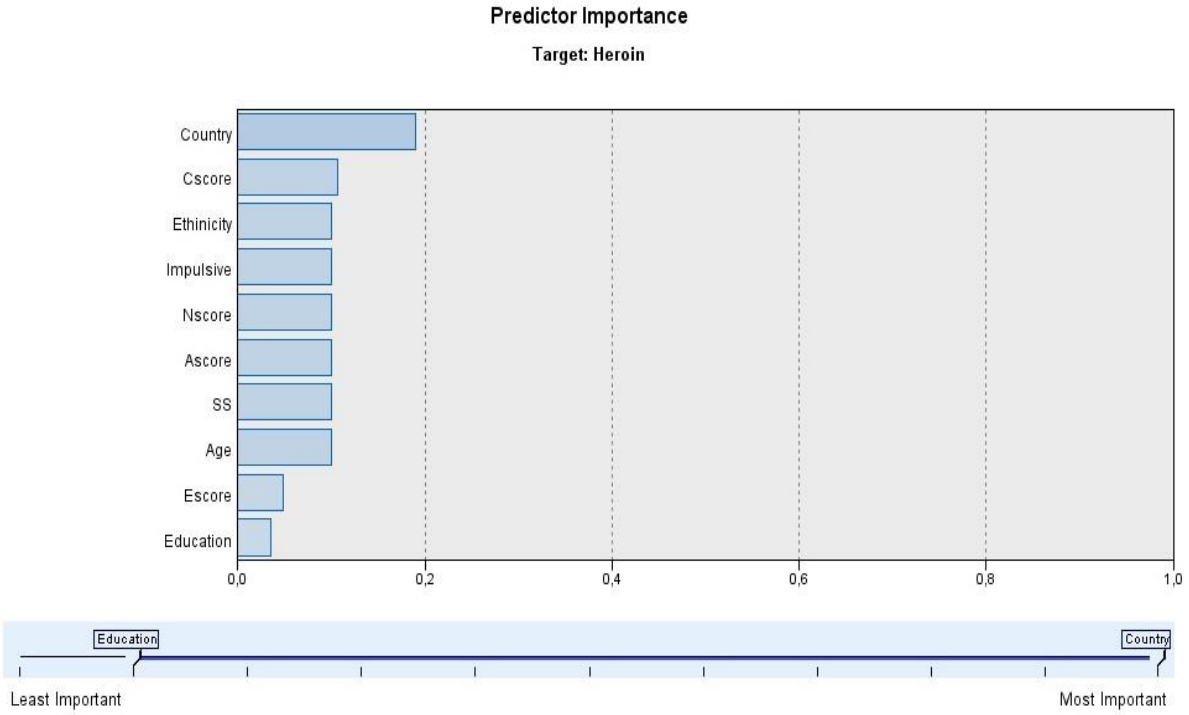
- Heroin



Random Forest

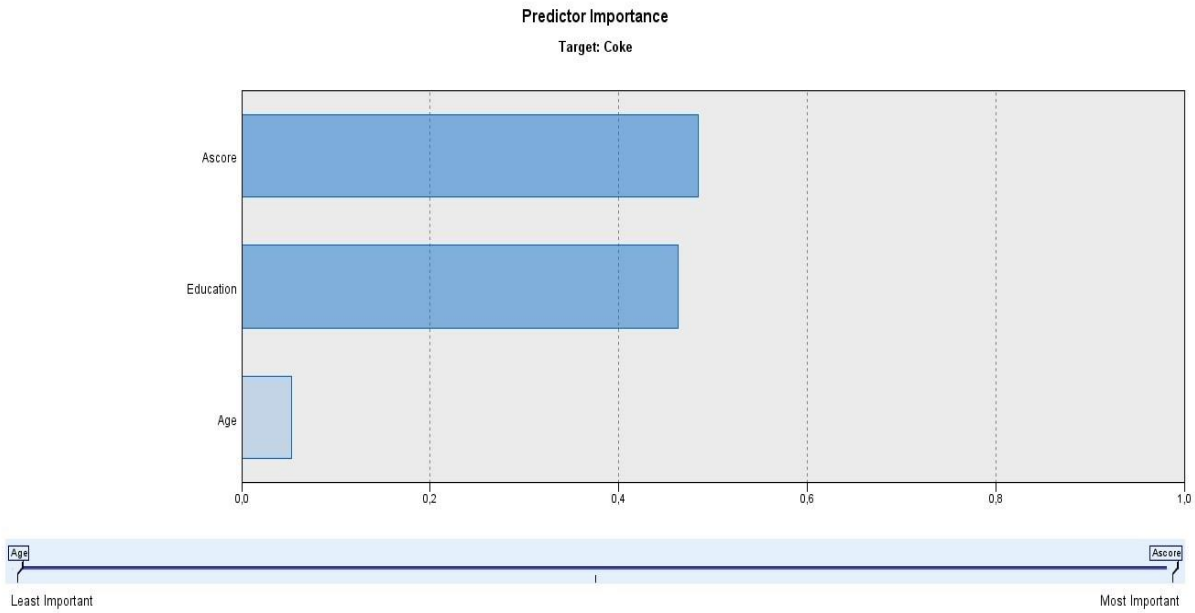


Decision Tree



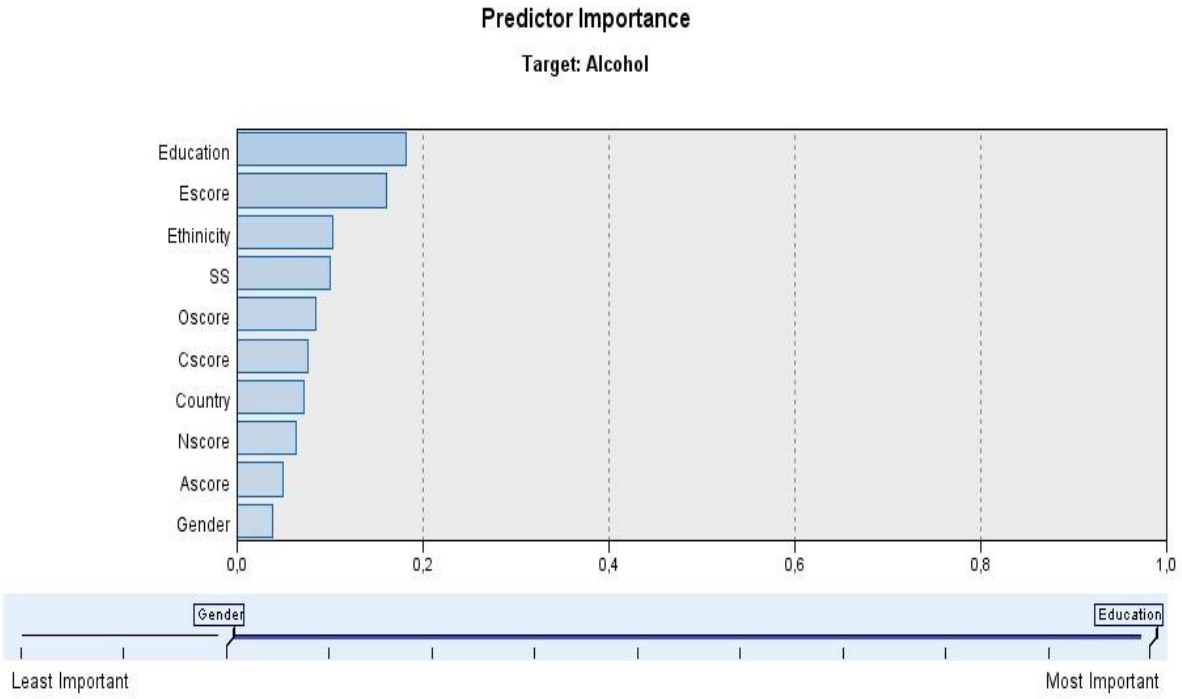
CRT

- Kokain

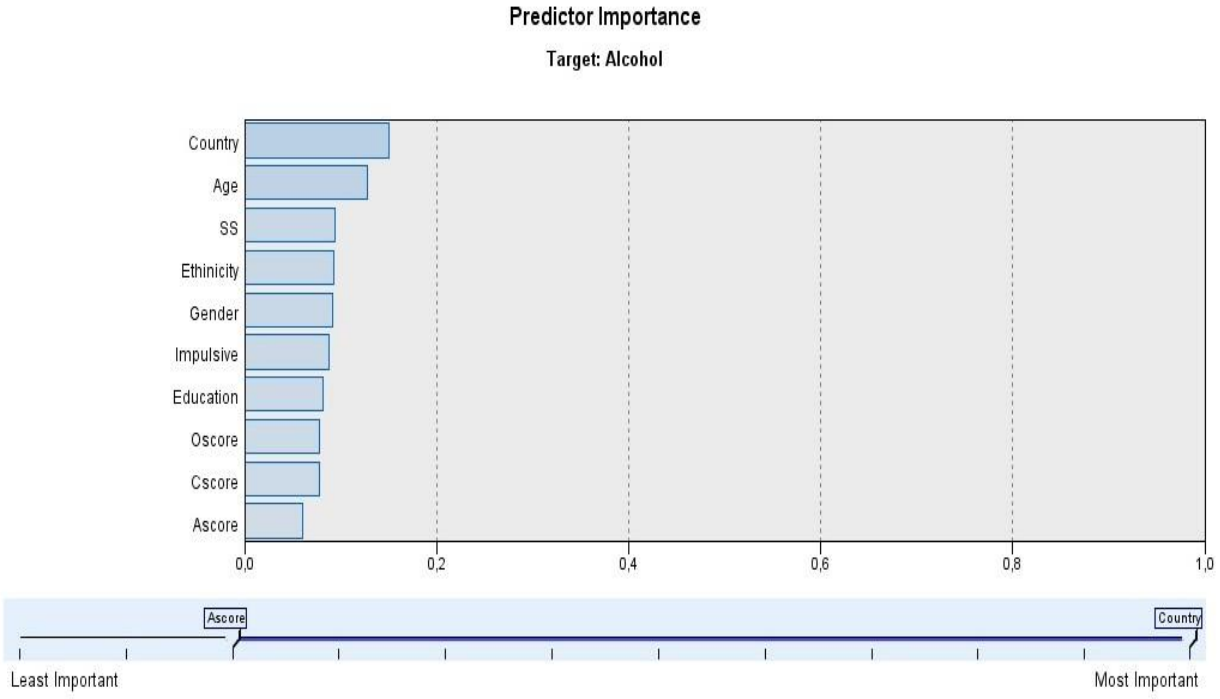


Decision Tree

- Alkohol

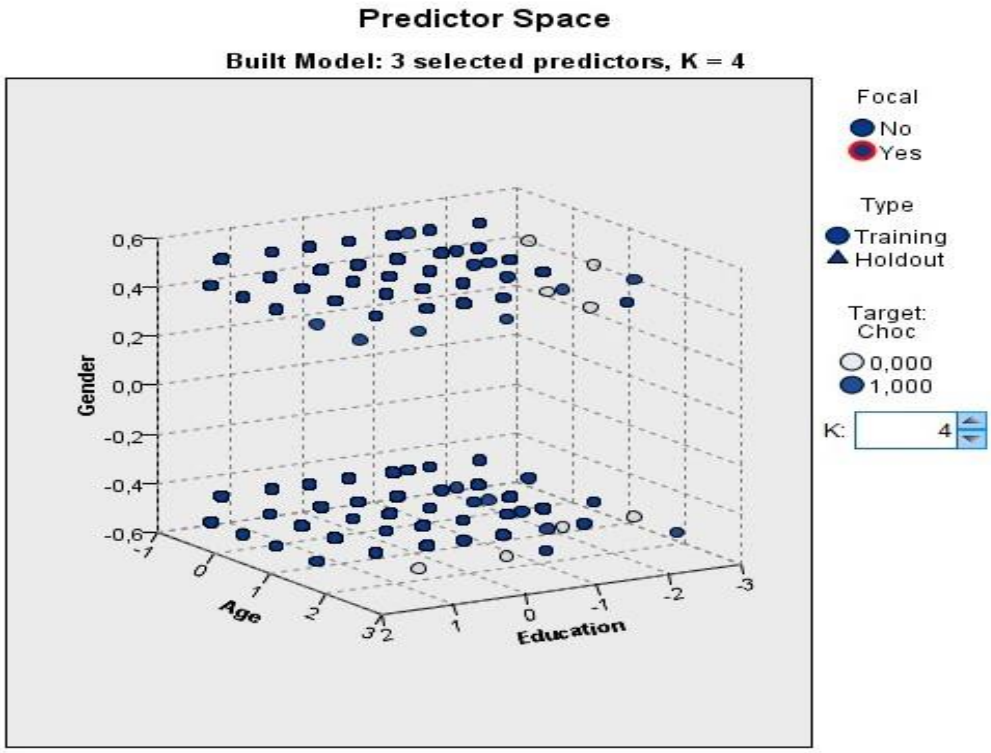


Neuron Net



Decision Tree

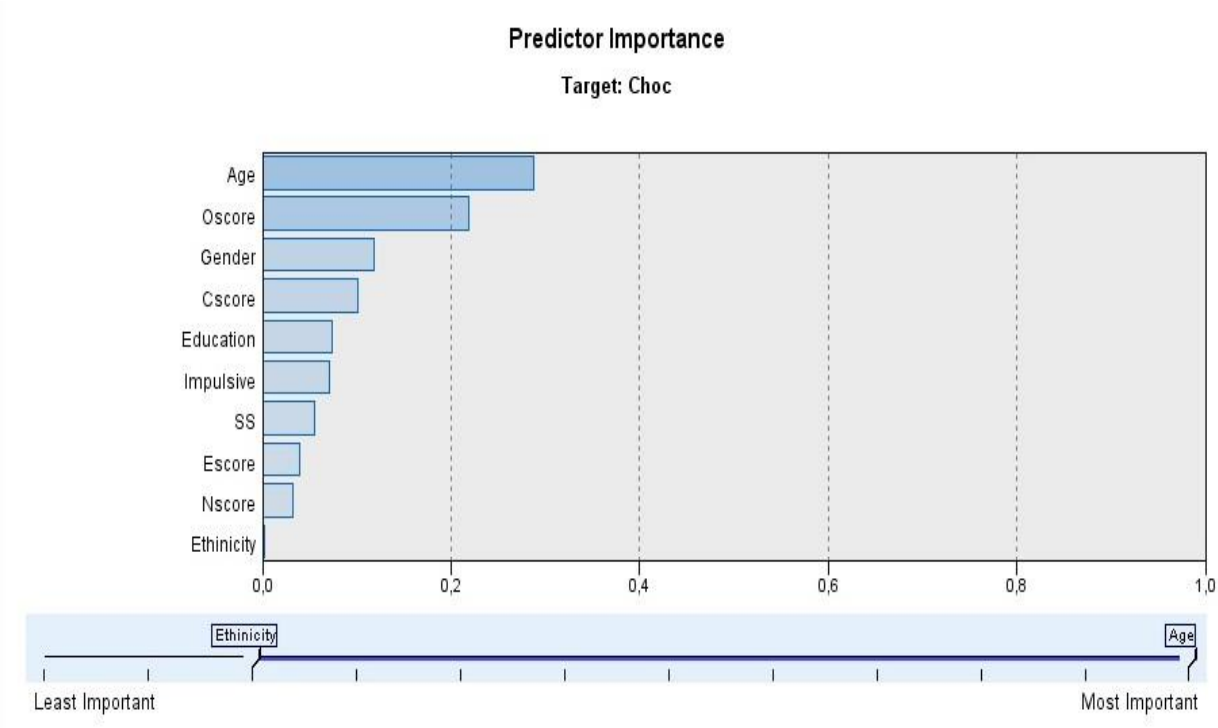
- Čokolada



Select points to use as focal records

This chart is a lower-dimensional projection of the predictor space, which contains a total of 12 predictors.

KNN



Decision Tree