|  |
| --- |
| Математички факултет у београду |
| Истраживање података из базе пулсара |
| Семинарски рад из предмета истраживање података 1 |
|  |
| **Стеван Поповић** |
| **2.6.2023.** |

Увод

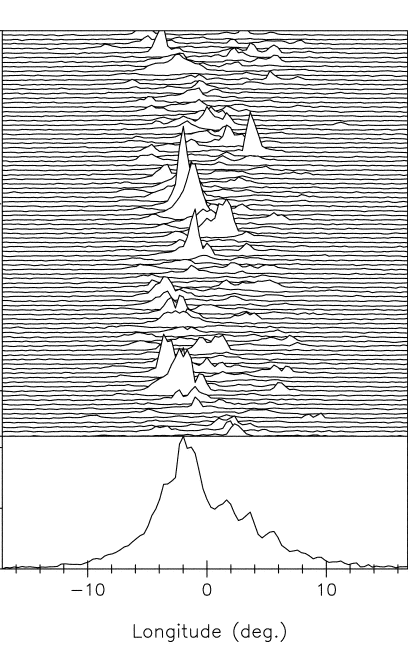
У овом раду обрађиваћемо базу података преузету са [kaggle.com](https://www.kaggle.com/datasets/charitarth/pulsar-dataset-htru2) сајта која садржи информације о мерењима радио таласа из свемира у потрази за пулсарима. Пулсар је тип неутронске звезде кој производи јаке електромагнетне таласе на својим половима и ротира се великом брзином. Ми са Земље те радио таласе детектујемо периодично у моментима када је један од полова пулсара окренут ка нама, стога нам делује као да он пулсира. Одатле назив (eng. **puls**e **r**adio signal).

На том скупу података приказаћемо неколико метода за класификацију и кластеровање, и на крају ћемо направити и правила придруживања.

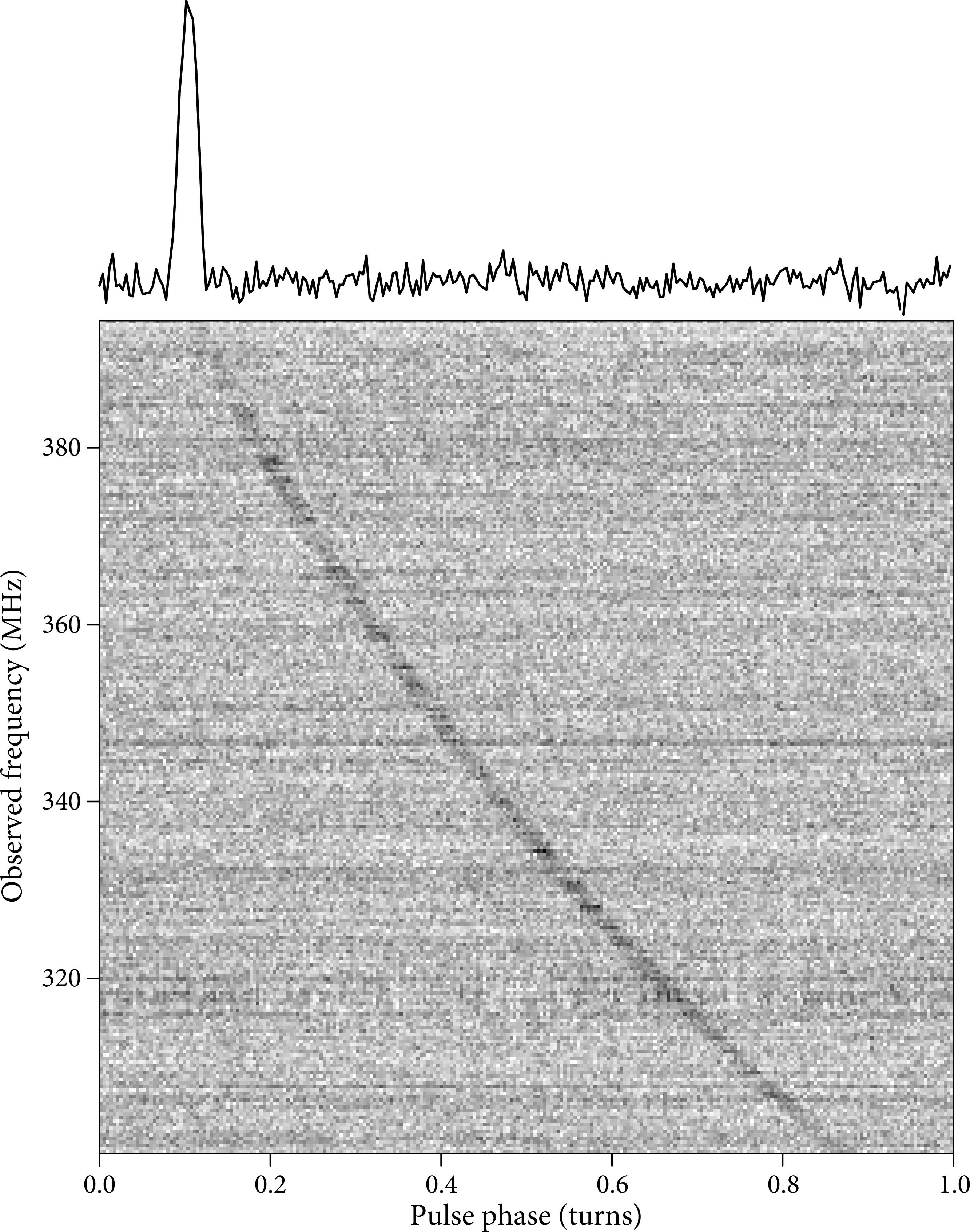
|  |
| --- |
|  |

Експлоративна анализа

База података налази се у csv (eng. comma separated values) датотеци са седамнаест хиљада врсти и девет колона. Свака врста садржи једну инстанцу мерења, осам од девет колона садрже непрекидне нумеричке податке добијене мерењем и девета колона садржи информацију о класи, тојест да ли су вредности добијене мерењем сигнала пулсара или не.

Пошто су пулсари, као и све друго у свемиру, на великој удаљености од нас, позадинско зрачење уноси јак шум с којим се астрономи морају некако изборити. То раде снимајући радио таласе неко време, тражећи било какав периодичан пулс. Затим поделе снимак на интервале које одговарају периоду и саберу сигнал. У добијеној суми позадинско зрачење се углавном поништи јер је насумично, док се сабрани пулсеви појачају.

Слика 1: сирови сигнали из свемира (горе ) и њихова сума (доле)



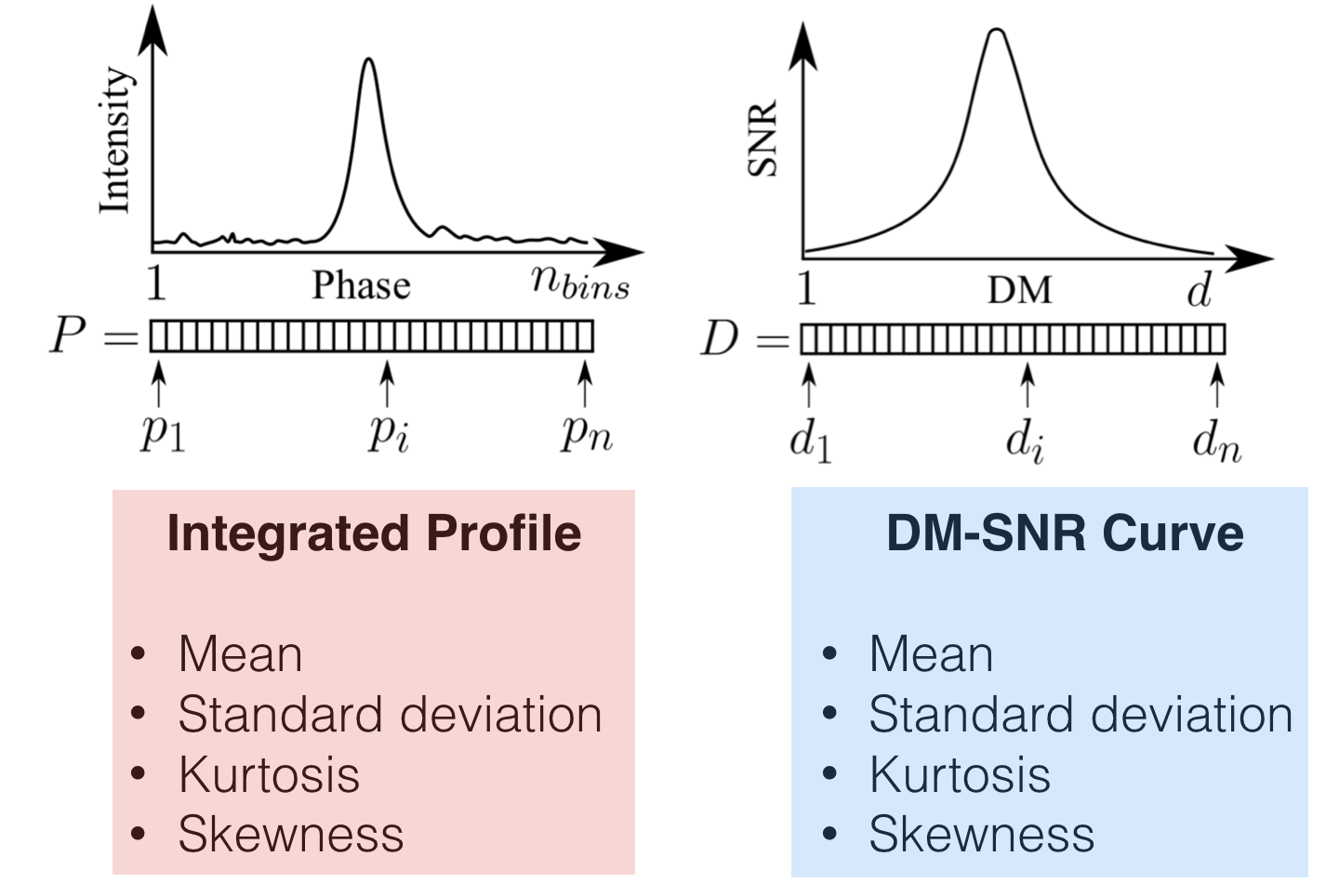
Слика 2: дисперзија (доле) сигнала (горе)

Други проблем који се јавља при мерењу је дисперзија: различите фреквенције путују различитом брзином кроз свемир, те се и то мора узети у обзир. Због тога поред самог измереног таласа астрономи направе још једну криву која приказује размеру шума и сигнала у односу на меру дисперзије.

Те две криве описују се на исти начин, користећи четири њихове особине:

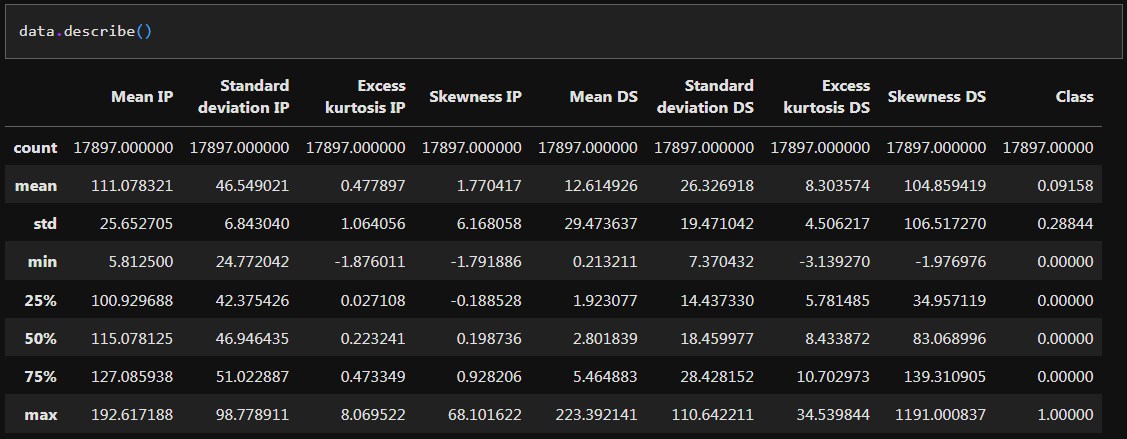
* Средња вредност (аритметичка средина)
* Стандардна девијација (просечно одступање од средине)
* Куртоза (мера која описује колико је талас зашиљен)
* Искривљеност (мера која описује колико је талас накривљен у једну страну)

Слика 3: добијене две криве и њихове особине

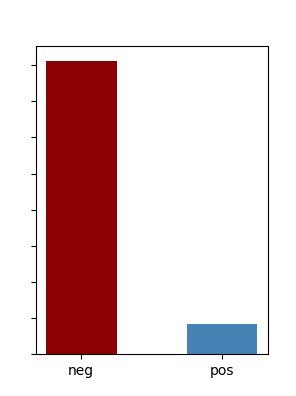


Приликом рада називи колона записани су скраћено на енглеском језику.

Недостајућих вредности нема тако да се не морају допуњавати одговарајућим вредностима нити има инстанци које треба избацити.

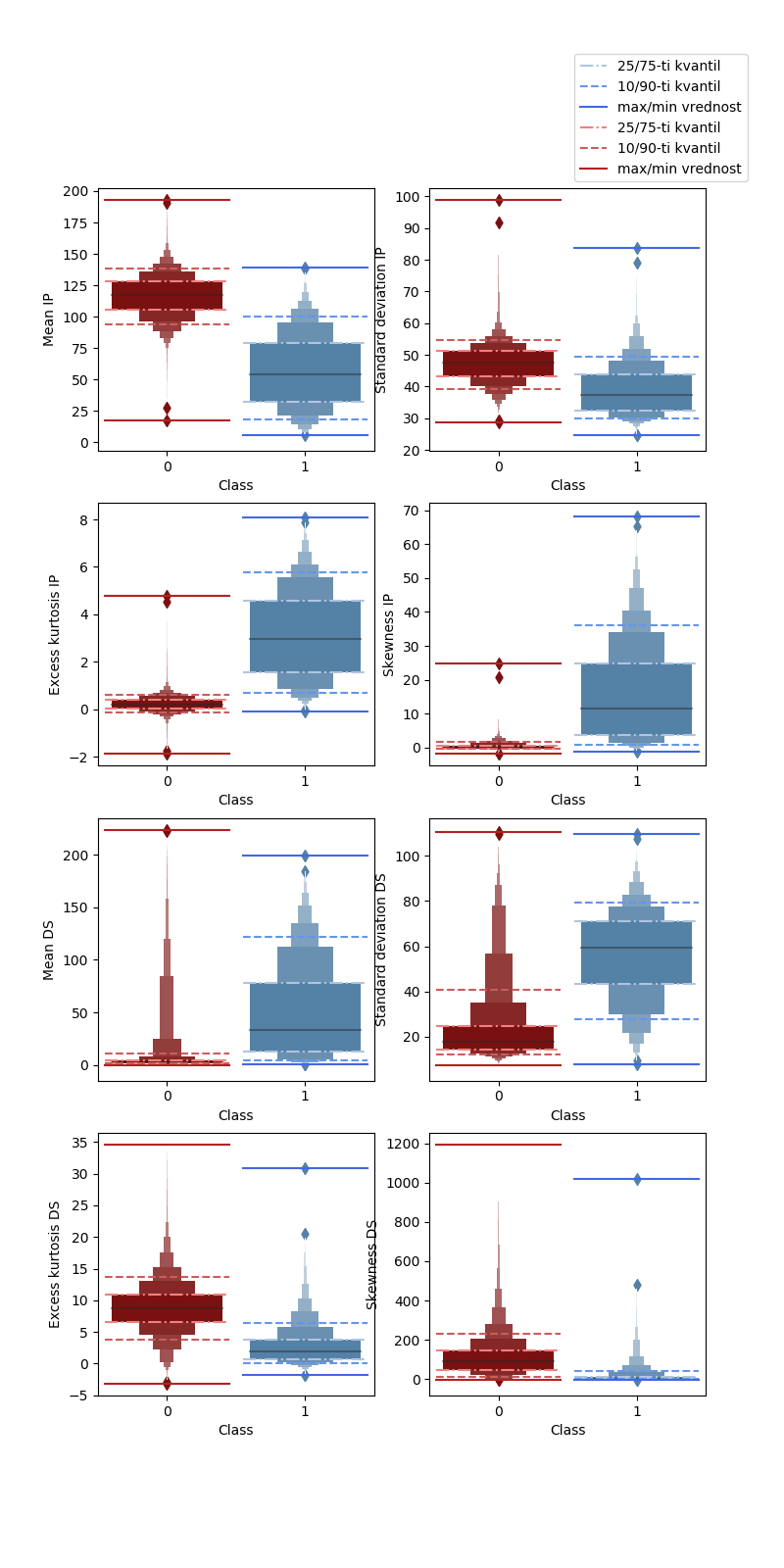


Слика 4: табела неких вредности које описују податке



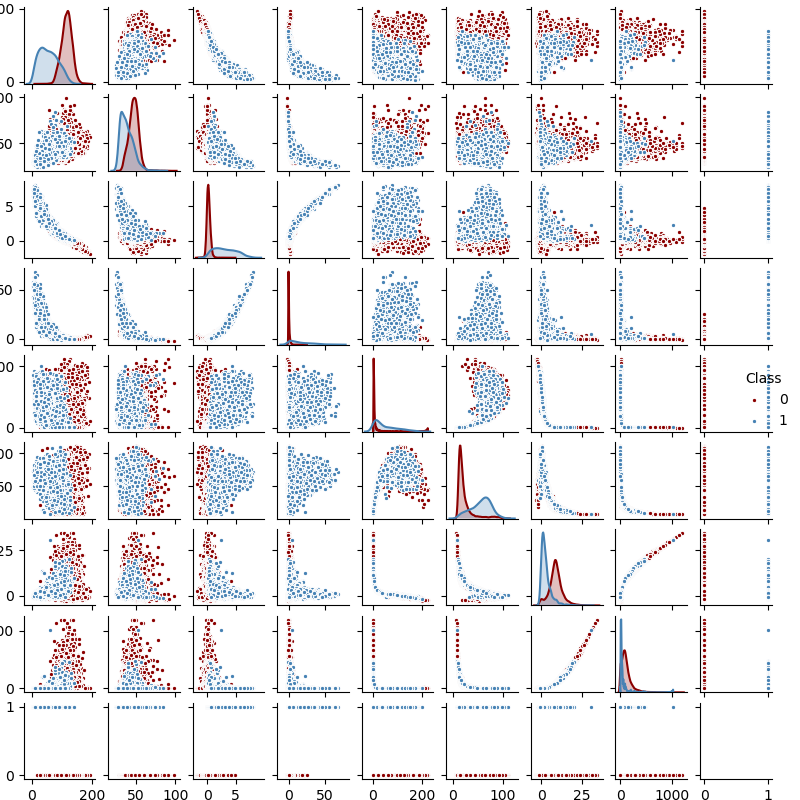
Слика 5: приказ класне небалансираности

На слици 4 можемо видети да је просек класног атрибута око 0.09 што значи да је 9% инстанци снимљених сигнала пулсар. На то треба обратити пажњу, и на адекватан начин користити моделе који су осетљиви на небалансираност података на пример навођењем тежина или балансирањем података неким од алгоритама (undersampling, oversampling, SMOTE).



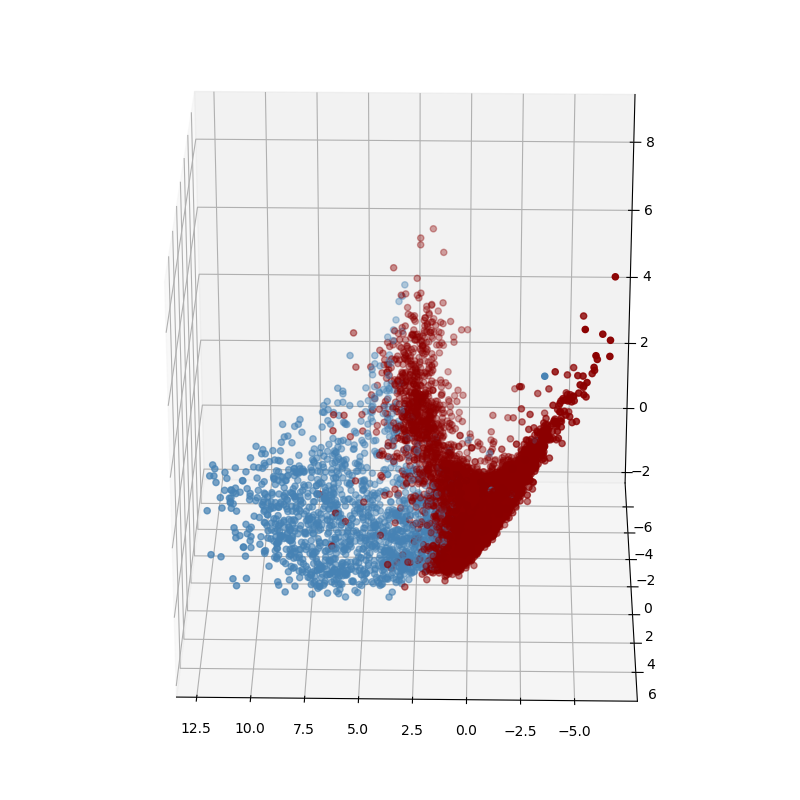
Подаци се могу приказати на више начина. У раду коришћен је boxenplot који податке приказује, на слици 6, слично свећицама, али су “кутије” подељене на више сегмената које одговарају одређеним квантилима. На слици 7 нацртани су распореди свих инстанци по два атрибута, за сваки пар атрибута. На главној дијагонали и последњој врсти де факто се налазе еквивалентни прикази са слике 6.

Слика 6: boxenplot



Слика 7: распоред инстанци по паровима свих атрибута

Последњи приказ података, слика 8, којим се бавимо у овој етапи рада је приказ главних компоненти, користећи PCA алгоритам (енг. Principal Component Analysis). PCA алгоритам ради тако што у простору података са n димензија (у нашем случају 8) нађе праву по којој подаци имају највећу варијансу, и њу постави за прву главну компоненту. Наставља да ради итеративно на преосталом простору са n-1 димензијом. Резултат је линеарна трансформација система која може бити корисна због добијене могућности пројектовања података са пуно димензија тако да остане очувано што више варијансе на нешто видљиво и схватљиво људском оку и уму.



Слика 8: Анализа главних компоненти

Класификација

Пре надгледаног учења као што је класификација, неопходно је поделити податке на тренинг и тест скуп. Одабрана размера за ту поделу у овом раду је 25-75%, у корист тренинг података. Важно је стратификовати поделу по атрибуту класе како би остала једнако балансирана.

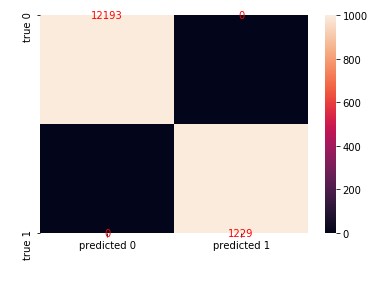
У раду користе се два алгоритма за класификацију. Први је CATR (енг. classification and regression tree) кој прави бинарно дрво одлучивања. Битно је сетити се да су нам класе небалансиране, те се размера класа мора унети као један од параметара. Ту размеру алгоритам узима у обзир када рачуна нечистоћу чворова. Коришћена мера нечистоће је Гинијева. Она се рачуна тако што прво израчуна Симпсонов индекс **λ**. То је вероватноћа да се два насумично одабрана елемента у чвору налазе у истој класи (узевши у обзир да су класе небалансиране). Одузимањем Симпсоновог индекса од броја један добија се вероватноћа да су две насумично одабране инстанце у различитим класама.

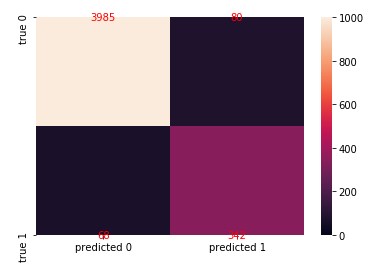
Остали параметри су накнадно подешени, коришћењем претраге оптималног модела из скупа неколико задатих могућих вредности.

На слици 9 може се видети наивно стабло пре додавања параметара.

Слика 9: наивно стабло



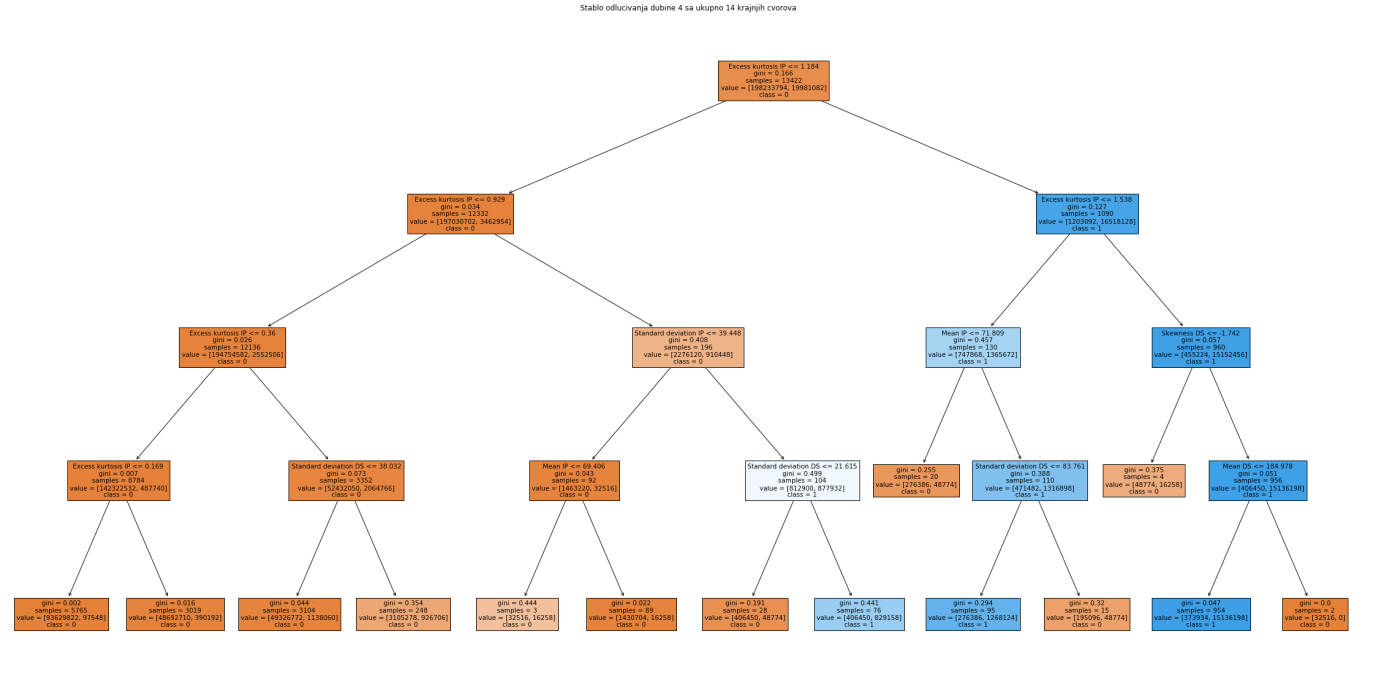
Овај модел је очекивано претрениран, што се може видети на слици 10



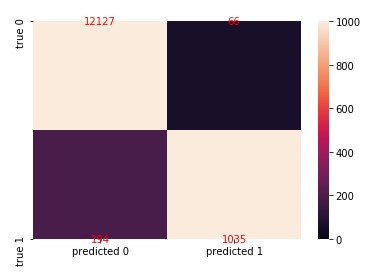
Слика 10: матрице конфузије тренинг скупа (лево) и тест скупа (десно)

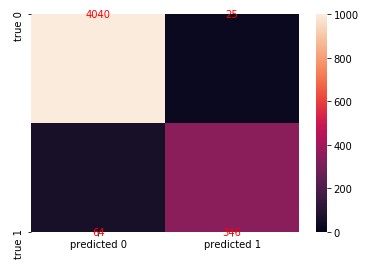
Скупови унутар којих се врши претрага параметра је уређена тројка:

* У првом скупу ради се претрага оптималног параметра *max\_depth*, кој одређује максималну будину стабла
* У другом се претражује оптимална вредност за параметар *min\_samples\_split*, кој одређује минималну количину инстанци у чвору потребну да би се он разгранао
* У трећем се претражује оптимална вредност за параметар *min\_impurity\_decrease*, кој одређује колико минимално мора да се смањи нечистоћа при потенцијалном гранању да би се оно остварило

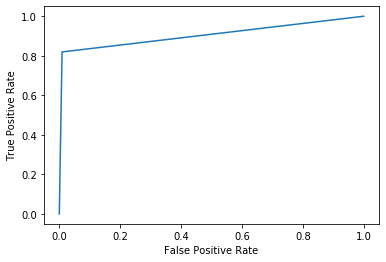
Претрага пролази кроз 440 кандидата и издваја онај са најбољом оценом. Стабло најбољег кандидата и неке његове оцене могу се видети на сликама 11, 12, 13

Слика 11: добијено стабло одлулчивања претрагом оптималних параметара





Слика 12: матрице конфузије тренинг скупа (лево) и тест скупа (десно)

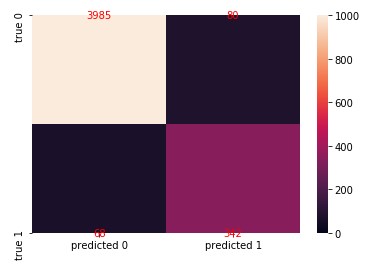
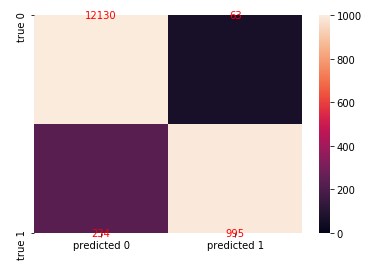


Слика 13: РОК крива

Опширније оцене могу се наћи у свесци са кодом.

Други метод за класификацију који је коришћен у раду је класификатор насумичним стаблима. Он припада групи ансамбла, тојест метода који раде тако што уместо да оптимизују параметре једног модела провуку податке кроз више “довољно добрих” модела. На пример, ако се ради о бинарној класификацији, довољно је да сваки од модела унутар ансамбла ради с прецизношћу од бар 51%. Приликом процене класе којој инстанца припада, за одговор се узима оно што највећи број појединачних модела предвиди.

Процес прављења модела аналоган је претходном. На слици 14 могу се видети матрице конфузије тог модела.



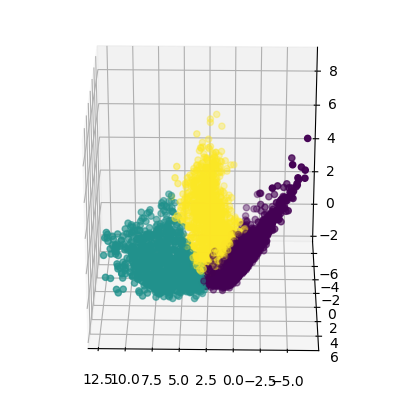
Слика 14: матрице конфузије тест (лево) и тренинг (десно) података

Кластеровање

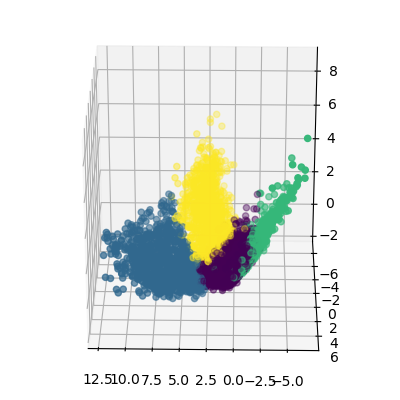
С обзиром да је скуп података предвиђен за класификацију, питање је колико смисла ће имати кластери које добијемо. Занимљиво би било видети да ли ће ненадгледано учење успети да издвоји пулсаре од осталих инстанци што је урађено, и уочити неке смислене групе унутар скупа пулсара, што је остављено за наредни рад. Резултати се зато пореде са сликом 8.

Треба узети у обзир да коришћени алгоритми за кластеровање користе раздаљине међу инстанцама те је неопходно нормализовати податке.

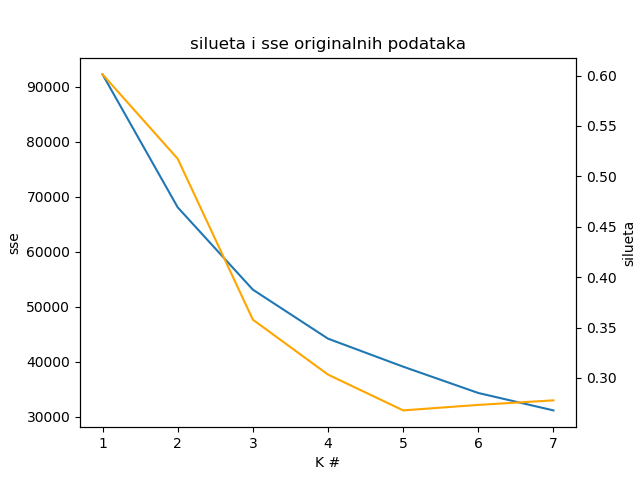
Први алгоритам који је коришћен је ка средина. Он ради тако што одреди ка насумичних кандидата за центроиде, и све инстанце подели у класе чији су центроиди представници на основу тога ком центроиду су најближе. Затим итеративно ажурира положаје центроида на тежишта добијених кластера и итеративно понавља процес до услова заустављања који је углавном у виду броја инстанци које су промениле кластер између итерација или раздаљина између кандидата за центроиде између итерација.

Пошто се свакако користи PCA за приказивање добијених кластера, урађена су два приступа: једном се подаци кластерују па се приказују коришћењем главних компоненти, други пут се кластеровање врши на самим компонентама и такво приказује. Поступак је поновљен за број кластера од 2 до 8. Разлике између два наведена приступа нема, на овом скупу података резултати су идентични. Приказ овог процеса може се видети у свесци, а на слици 15 је издвојена подела на три кластера, где уочавамо кластер приказан зеленом бојом који подсећа на класу пулсара.

Слика 15: добијени кластери користећи ка средина са три кластера

На слици 16 налази се график силуете и суме квадрата грешака који одговарају кластеровањима која су наведена горе. Примећујемо “лакат” у околини поделе података на четири кластера, која се може видети на слици 17.

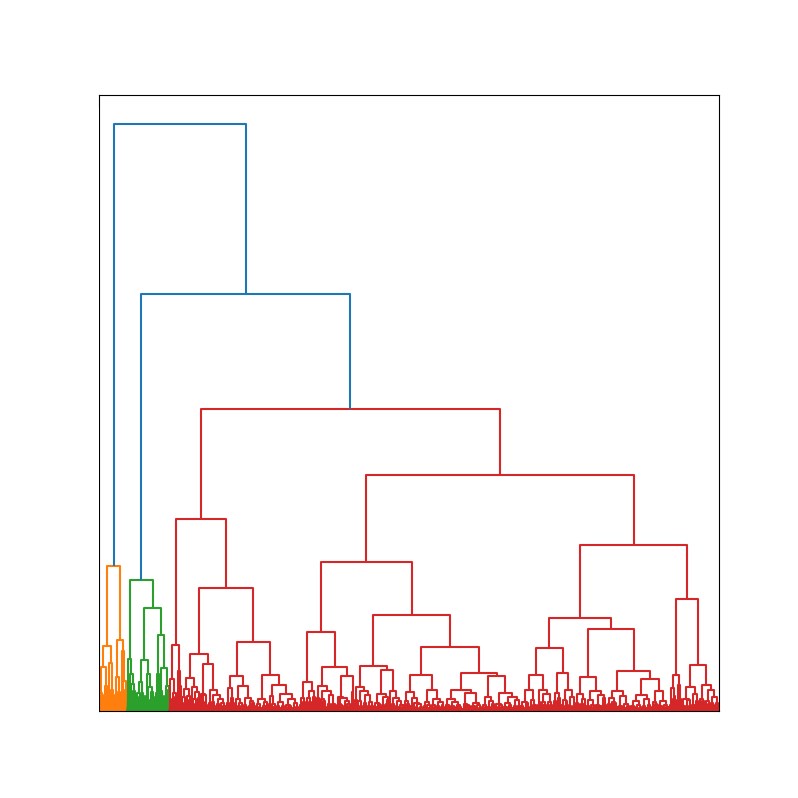
Слика 17: добијени кластери користећи ка средина са четири кластера



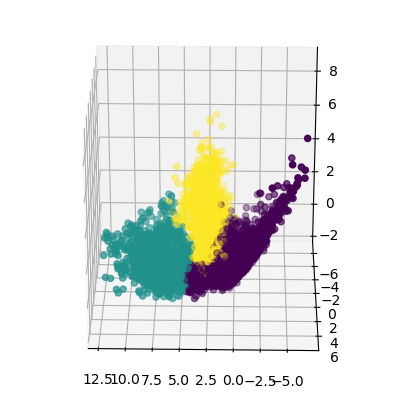
Слика 16: график силуете и суме квадрата грешака

Други приступ кластеровању који је коришћен је хијерархијско кластеровање. У скупу података проналазе се два најближа елемента, групишу се у један кластер и затим се процес итеративно понавља с тим што се у свакој наредној итерацији претходно груписане инстанце посматрају као један објекат. При томе раздаљина између два објекта се може рачунати на више начина као што су раздаљина између два најближа члана, два најдаља члана, раздаљина између просечних позиција чланова унутар кластера. У овом раду коришћен је Вордов метод који не тражи два најближа кластера већ проба да споји свака два и бира онај пар при чијем ће спајању сума варијанси нових кластера бити најмања.

На слици 18 може се видети тако добијена хијерархија. Обојена су три подстабла јер при спајању два од та три кластера варијанса расте преко подразумеваног критеријума заустављања који је у коришћеној имплементацији једнак седам десетина максималне раздаљине. На слици 19 налазе се та три кластера у простору. У свесци могу се видети прикази података при заустављању кластеровања на нивоима за 2, 4 и 5 кластера.



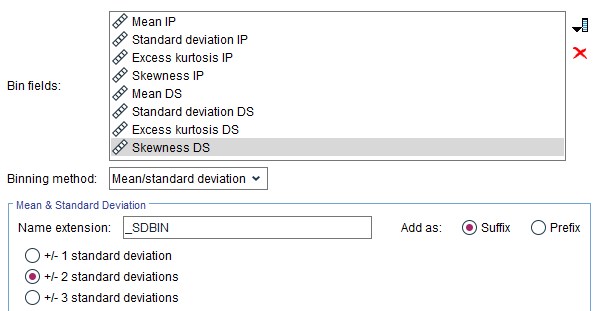
Слика 18: дендрограм хијерархијског кластеровања са три кластера издвојена својим бојама



Слика 19: три кластера добијена хијерархијским кластеровањем

Правила придруживања

Користећи априори принцип генерисана су нека правила придруживања. Пре тога, пошто априори принцип прима само дискретне податке потребно је дискретизовати све непрекидне, што су у овом случају сви подаци сем класног. На слици 20 приказан је тај процес: подаци су подељени по броју стандардних девијација колико одступају од средине своје колоне.

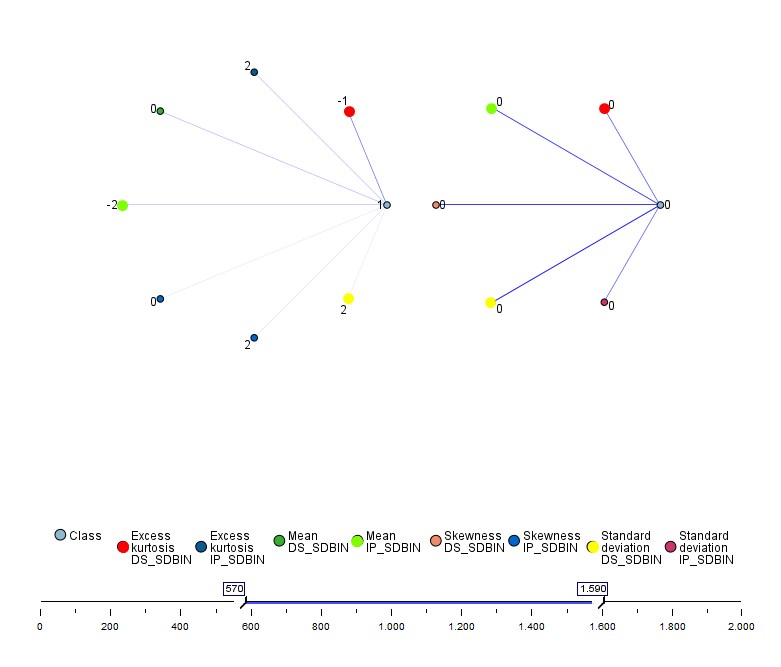


Слика 20: дискретизација података

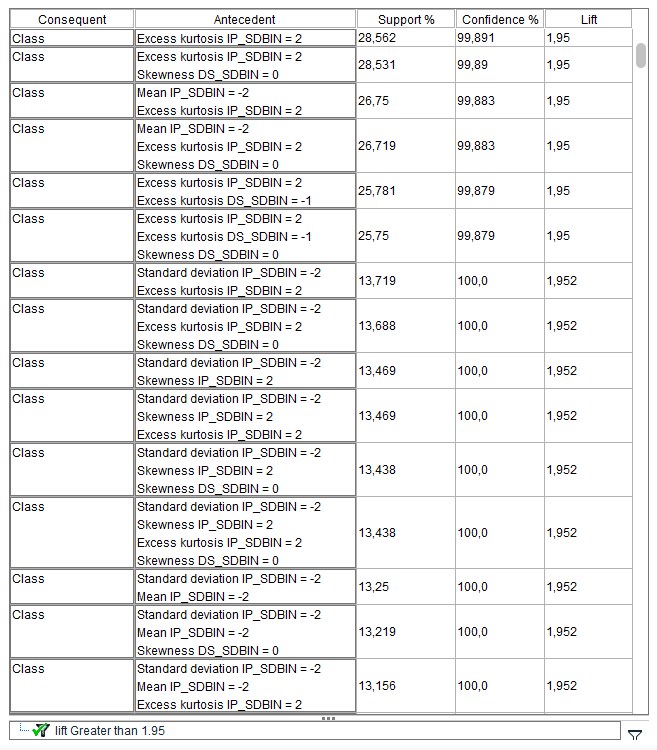
Затим се случајним непотпуним избором (енг. undersampling) балансира број негативне класе са позитивном. На слици 21 приказан је усмерени мрежни график на коме се види јачина везе припадања одређеним корпама инстанце ка њеном припадању класи. На основу њега могуће је извести нека очекивања која ће се правила појавити, на пример нешто попут:

*excess kurtosis DS = -1, excess kurtosis IP = 2 => class = 1*

На слици 22 приказана су нека од генерисаних правила придруживања. Могуће је приметити да се заиста у некој мери поклапају на очекивањима насталих на основу графика са слике 21.



Слика 21: график јачине веза између припадности инстанци корпама и класама



Слика 22: генерисана правила придруживања

Закључак

Дати скуп података није гломазан, нема сувише инстанци те није рачунарски захтевно применити разне моделе на њега. Такође нема сувише атрибута па је и приказивање података у човеку јасном облику било лако. Инстанце су довољно јасно раздвојене својим особинама до те мере да не само да је класификацијом добијен задовољавајући резултат већ су и све пробане методе кластеровања нашле нешто што подсећа на класну поделу података. Априори принципом нађено је неколико правила придруживања са добрим оценама подршке поузданости и lift-a, што је још један показатељ да су подаци лепо класно раздвојени.

За даљи рад остављено је испитати добија ли се нешто занимљиво и смислено даљом поделом пулсара на подкластере тојест подкласе, мада се ту поставља питање постоји ли основ за тако нешто у контексту астрофизике.

Референце: