Универзитет у Београду

Електротехнички факултет

Катедра за Сигнале и системе

Машинско учење

-пројекат-

Професор:

Доц. др Предраг Тадић

Студент:

Милан Симовић, 3169/2022

Јануар 2023. год.

Садржај:

Задатак: Потребно је препознати да ли неко коришћење кредитне картице представља превару или не. Скуп података којим располажем преузет је са линка https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud. На том линку се налазе и публикације у којима је рађено над истим подацима.

Решење:

Увод

На располатању имам 1 000 000 примера, сваки са по седам обележја. Прва три обележја су континуална, а остала четири бинарна. Како циљна променљива може имати две могуће вредности (превара или није превара) реч је о проблему бинарне класификације.

Опис обележја дат је у табели 1. Надаље ћу се на обележја позивати помоћу њиховог редног броја (табела 1).

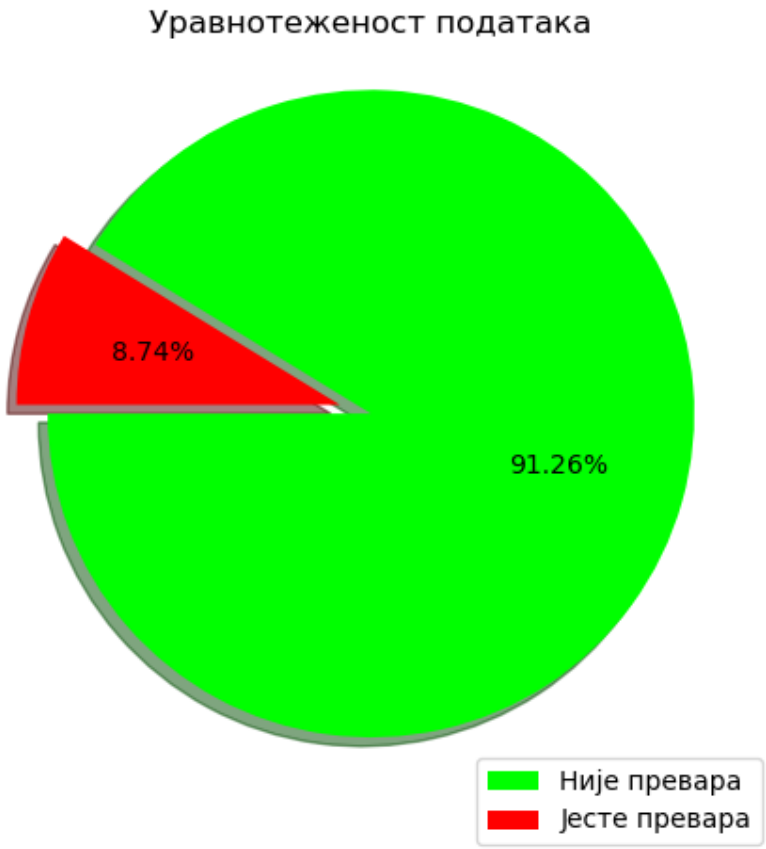
Обележја

Ред. бр.	Опис
1.	Растојање места трансакције од куће власника картице
2.	Растојање места трансакције од места претходне трансакције
3.	Однос тренутне цене и медијане плаћања
4.	Куповина у истој продавници
5.	Коришћен чип
6.	Унет пин
7.	Онлајн наруџбина

Табела 1 – Опис обележја

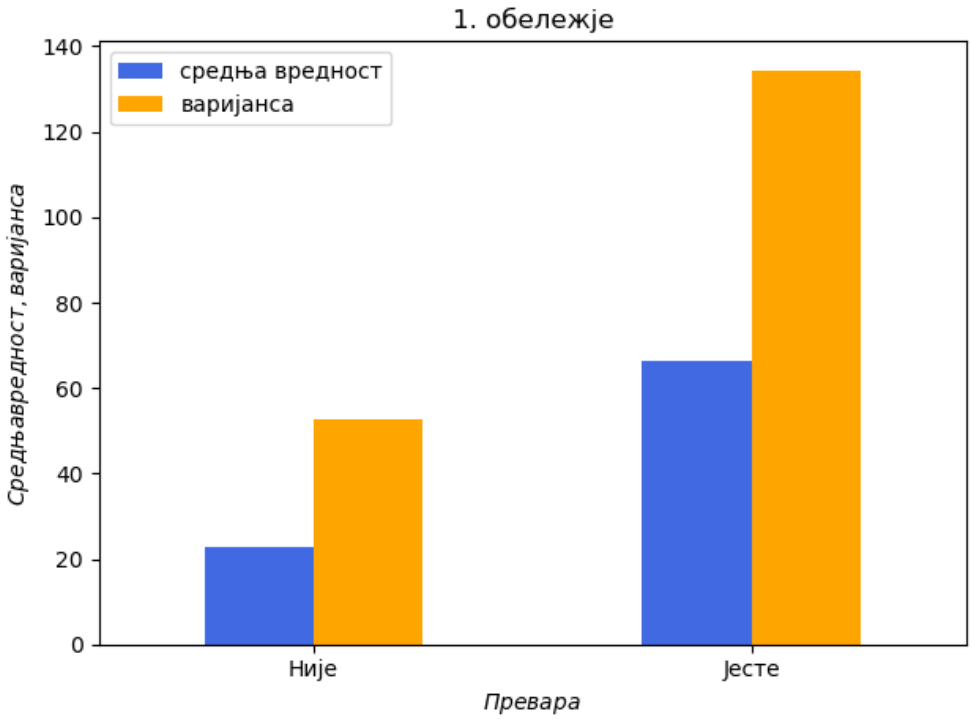
Експлоративна анализа података

Скуп података је небалансиран, што се може видети на дијаграму приказаном на слици 1.

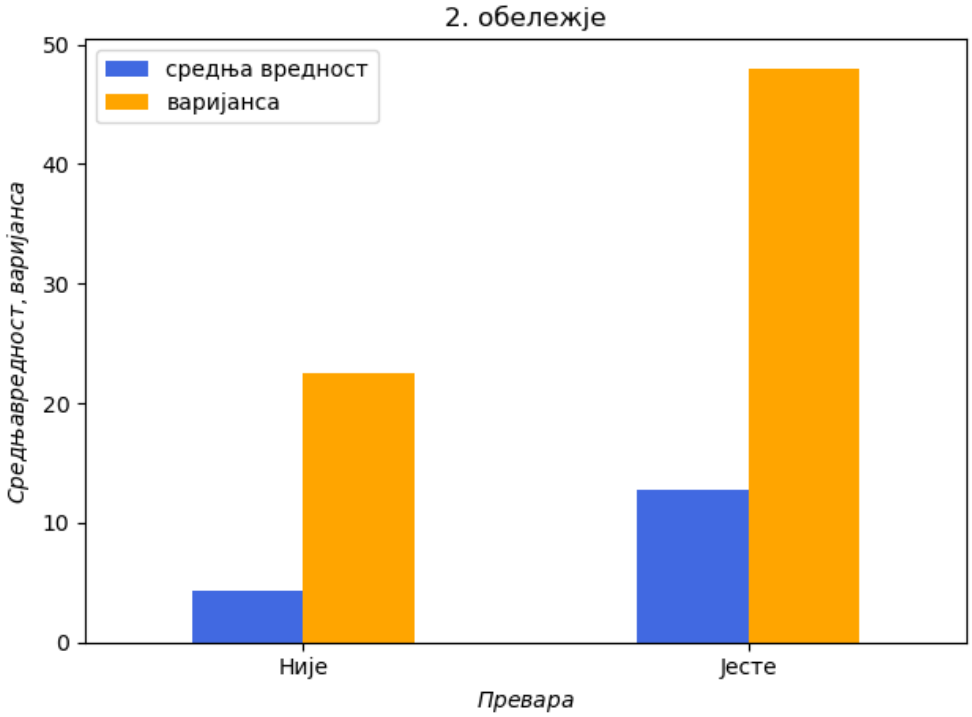


Слика 1 – Уравнотеженост података

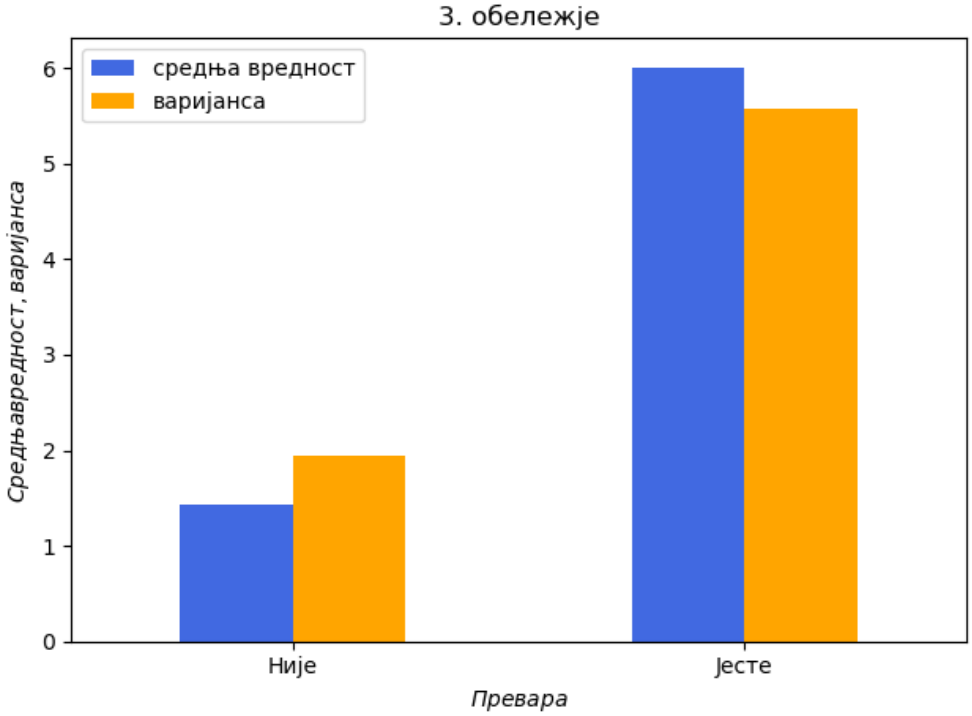
Најпре је извршена процена средње вредости и варијансе континуалних предиктора по класама. Графици тих вредности дати су на сликама 2-4.



Слика 2 – Средња вредност и варијанса за прво обележје

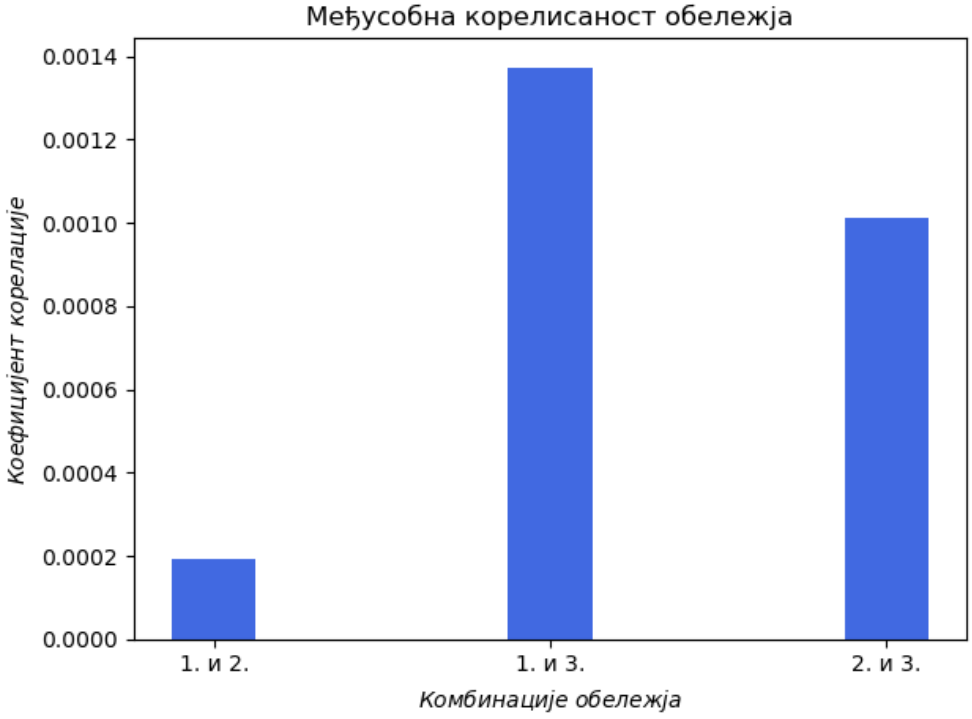


Слика 3 – Средња вредност и варијанса за друго обележје



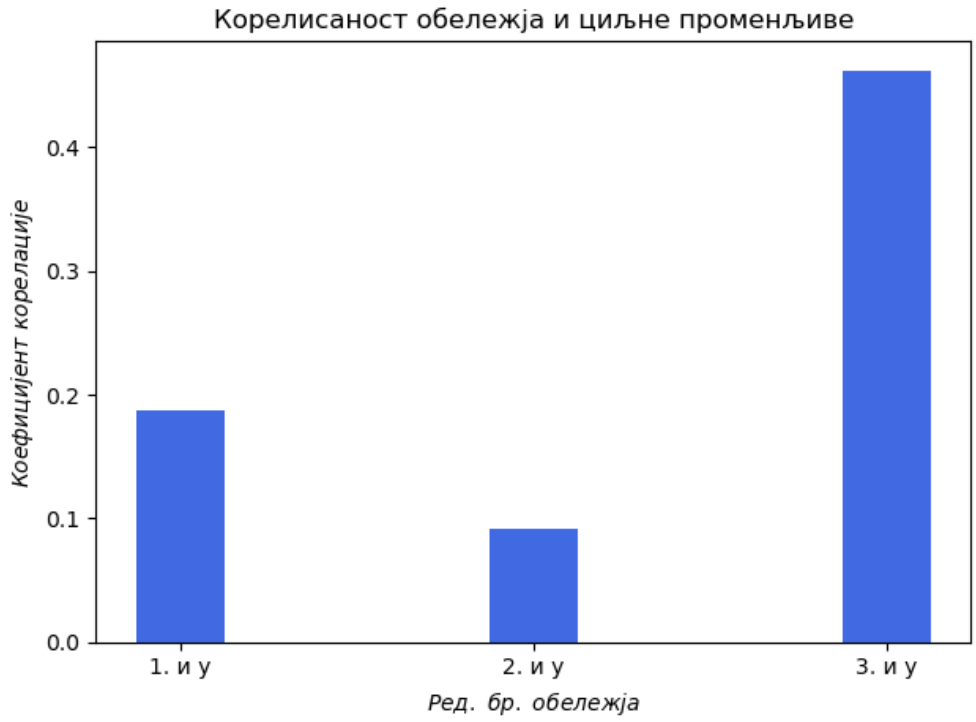
Слика 4 – Средња вредност и варијанса за треће обележје

График међусобне корелисаности континуалних обележја дата је на слици 5. Са тог графика јасно видимо да ниједна два обележја нису значајно међусобно корелисана, па нијдно не морамо искључити из разматрања на самом почетку.



Слика 5 – Међусобна корелисаност обележкја

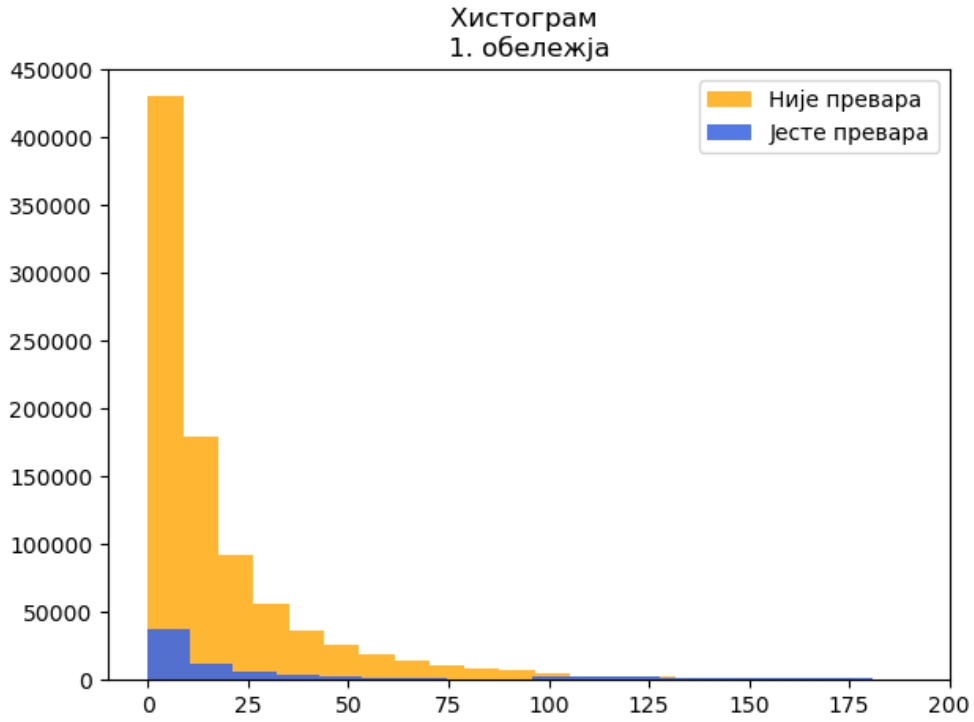
На слици 6 приказан је график корелисаности сваког од континуалних обележја са циљном променљивом.



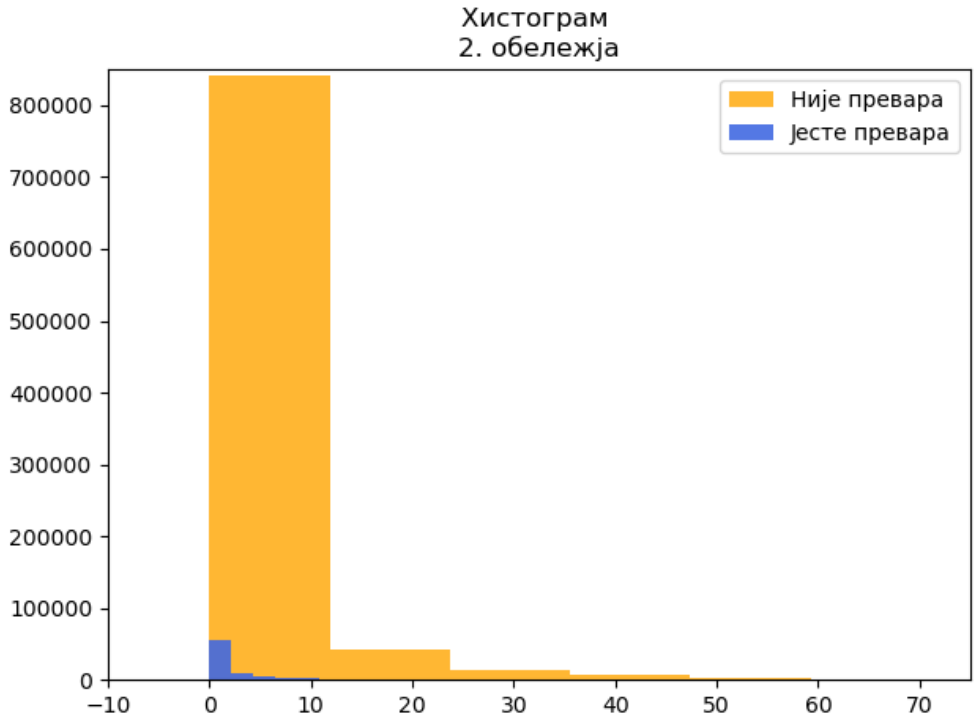
Слика 6 – Корелисаност обележја и циљне променљиве

Са графика изнад видимо да је од континуалних обележја треће најбоље корелисано са циљном променљивом. То ће у једном делу овог рада бити искоришћено.

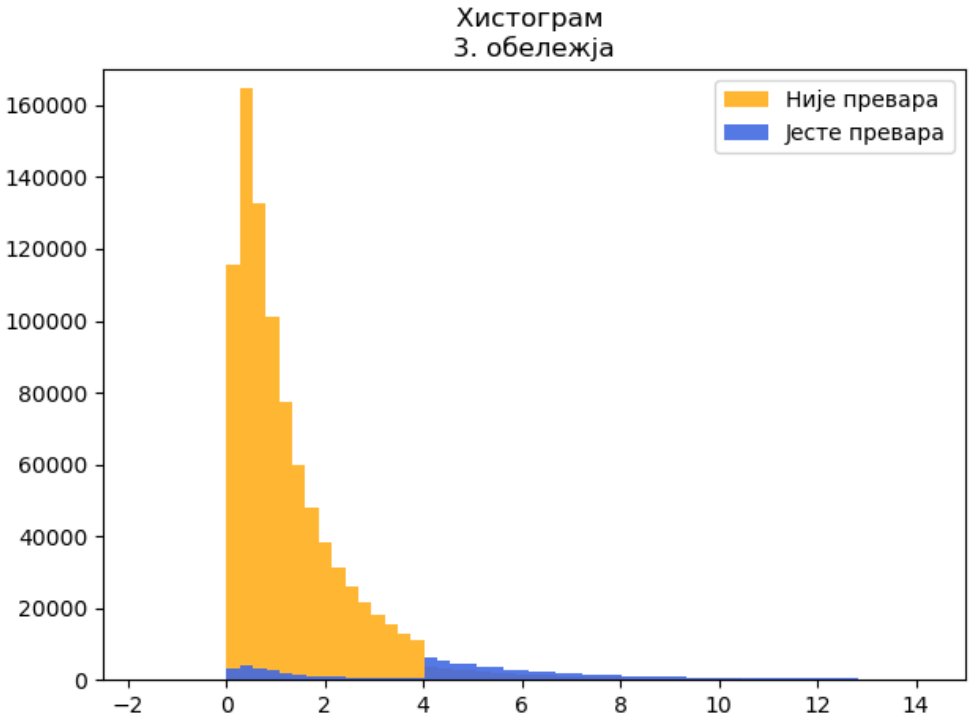
На сликама 7-9 приказани су хистограми континуалних обележја за позитивне и негативне примере.



Слика 7 – Хистограм првог обележја



Слика 8 – Хистограм другог обележја



Слика 9 – Хистограм трећег обележја

Класификација према бројности

С обзиром да је реч о небалансираном проблему, први класификатор који сам употребио примењује класификацију само на основу бројности класа. Ово је најједноставнији модел који је коришћен.

Матрица конфузије за овај проблем дата табелом 2.

	$\widehat{N}$	$\widehat{P}$
N	166605	15914
P	15934	1547

Табела 2 – Матрица конфузије (класификација према бројности)

Метрике као што су тачност, прецизност, осетљивост и *F1*-скор дати су у табели 3.

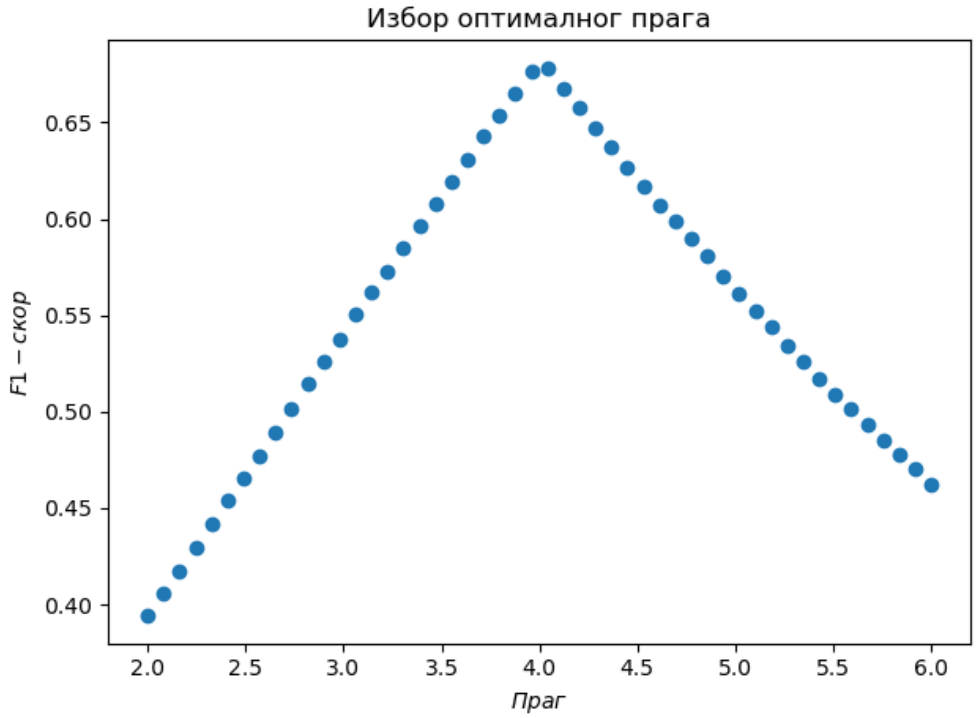
Тачност	0,84
Прецизност	0,09
Осетљивост	0,09
F1-скор	0,09

Табела 3 – Вредности метрика (класификација према бројности)

Из табеле 3 видимо да је тачност износи 0,84 што је прилично велика вредност с обзиром на једноставност класификатора. Међутим, како је реч о небалансираној класи, од интереса да нам је *F1*-скор. Та метрика представља хармонијску средину прецизности и осетљивости. У овом случају има малу вредност, што је последица погрешног класификовања класе од интереса.

Класификација на основу информативног обележја

Обележје које сам изабрао за класификацију је треће, због тога што је најбоље корелисано са циљном променљивом. Праг са којим га поредимо сам изабрао са графика приказаног на слици 10. То је вредност за коју је *F1*-скор на валидационом (тест) скупу максималан.



Слика 10 – Избор оптималног прага за треће обележје

Оптимална вредност прага је 4, а када је стандардизујем њена вредност износи 0,78.

За овако одабрану вредност прата метрике на тест скупу дате су табелама 4 и 5.

	$\widehat{N}$	$\widehat{P}$
N	175008	7511
P	4520	12961

Табела 4 – Матрица конфузије (класификација на основу информативног обележја)

Тачност	0,94
Прецизност	0,63
Осетљивост	0,74
F1-скор	0,68

Табела 5 – Вредности метрика (класификација на основу информативног обележја)

Логистичка регресија

Логистичка регресија подразумева да на основу обучавајућег скупа података одредимо параметре вектора θ помоћу којих се рачуна вредност логистичке функције неког вектора (примера). Изглед логистичке функције једног модела дат је изразом (1).

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n)}} \quad (1)$$

Потребно је обучити модела колико има и класа. Одлука којој класи припада пример доноси се на основу логистичке функије са највећом вредности.

Вредност *F1*-скор метрике која се добије када модел обучим користећи сва обележја износи 0,72. Искључујући поједина обележја из употребе резултати се или мало мењају или не мењају.

У табелема 6-9 дати су резултати на тестирајућем скупу.

	$\widehat{N}$	$\widehat{P}$
N	181219	1300
P	7370	10111

Табела 6 – Матрица конфузије (логистичка регресија – без 4. и 5. обележја)

Тачност	0,96
Прецизност	0,89
Осетљивост	0,58
F1-скор	0,70

Табела 7 – Вредности метрика (логистичка регресија – без 4. и 5. обележја)

Метрике изнад приказане су када из употребе искључимо 4. и 5. обележје. За изостављање само 4. обележја, метрике на тест скупу дате су у табелама 8 и 9.

	$\widehat{N}$	$\widehat{P}$
N	181268	1251
P	7039	10442

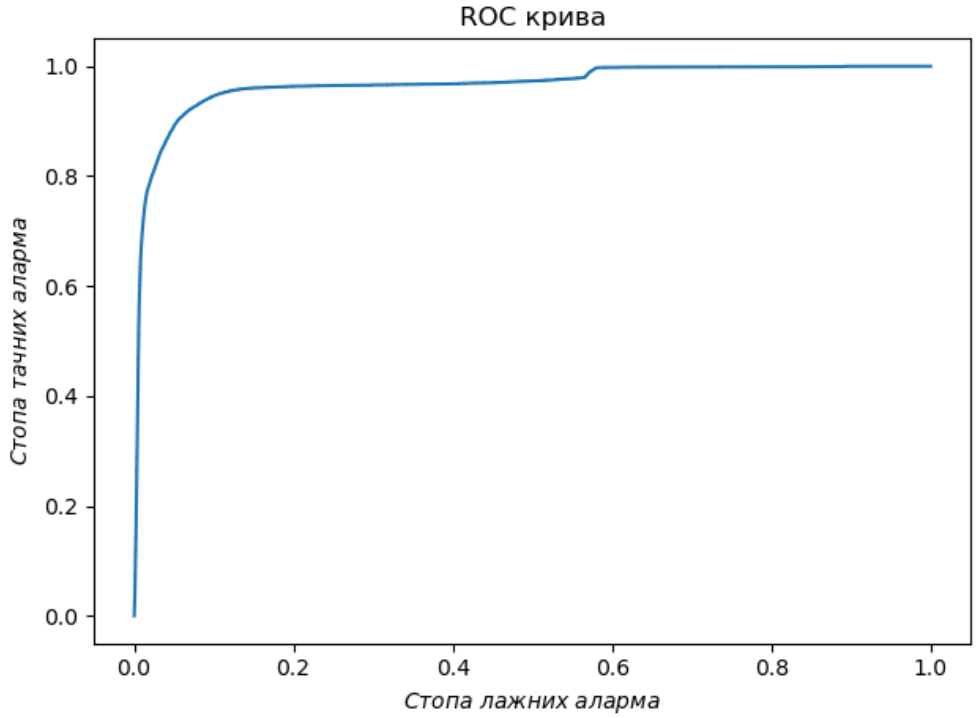
Табела 8 – Матрица конфузије (логистичка регресија – без 4. обележја)

Тачност	0,96
Прецизност	0,89
Осетљивост	0,60
F1-скор	0,72

Табела 9 – Вредности метрика (логистичка регресија – без 4. обележја)

У овом случају *F1*-скор има исту вредност као и када не изоставимо 4. обележје.

График ROC (*Receiver Operating Characteristics*) криве дат је на слици 11.



Слика 11 – ROC крива за логистичку регресију

Гаусовски наивни Бејз

Термин наиван односи се на то да је на самом почетку претпостављена условна независност предиктора (2).

$$P(x_1, x_2, \dots, x_n|y) = P(x_1|y)P(x_2|y)\cdots P(x_n|y) \tag{2}$$

Обучавањем овог модела, на тестирајућем скупу добијамо следеће метрике.

$\widehat{N}$ $\widehat{P}$

N	179752	2767
P	7197	10284

Табела 10 – Матрица конфузије (Гаусовски наивни Бејз)

Тачност	0,95
Прецизност	0,79
Осетљивост	0,59
F1-скор	0,67

Табела 11 – Вредности метрика (Гаусовски наивни Беј)

Метода носећих вектора

Иако метода носећих вектора није погодна за проблеме са великим скупом података, ипак сам је применио. У табелама 12 и 13 дате су метрике добијене на тестирајућем скупу применом ове методе.

	$\widehat{N}$	$\widehat{P}$
N	182405	114
P	264	17217

Табела 12 – Матрица конфузије (Метода носећих вектора)

Тачност	0,99
Прецизност	0,99
Осетљивост	0,98
F1-скор	0,99

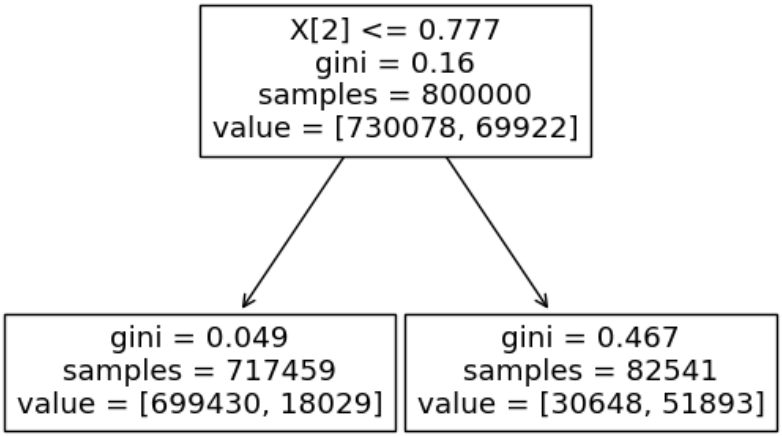
Табела 13 – Вредности метрика (Метода носећих вектора)

Овај алгоритам је заиста непогодан за примену на овом скупу података. Време потребно да обучи на скупу од 800000 података изнело је приближно 1398 секунди.

Стабло

Због мешовитости обележја (у смислу да има и континуалних и бинарних) требало би да је овај модел, а и сви који у себи садрже стабло, погодан за овакав порблем.

Уколико дубину стабла ограничим на један добијам следеће резултате. Изглед стабла приказан је на слици 12.



Слика 12 – Стабло дубине један

Оно што је занимљиво је да и алгоритам за обучавање стабла бира исто обележје и исти праг као када сам применио класификацију на основу информативног обележја и

погодно изабраног прага. Реч је трећем обележју, а праг је 4, односно 0,78 за стандардизоване податке.

	$\widehat{N}$	$\widehat{P}$
N	174960	7559
P	4423	13058

Табела 14 – Матрица конфузије (Стабло дубине један)

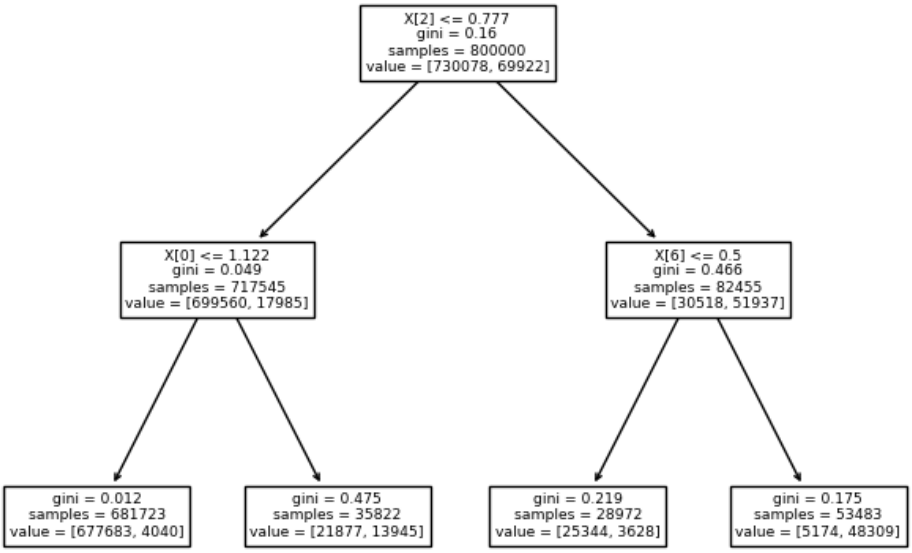
Тачност	0,94
Прецизност	0,63
Осетљивост	0,75
F1-скор	0,69

Табела 15 – Вредности метрика (Стабло дубине један)

Очекивано је да и метрике приказане у табели 15 изгледају исто као у случају класификатора са најинформативнијим обележјем.

Примећујем да се са повећањем дубине стабла грешка на тестирајућем скупу не повећава. Иначе се очекује да са пвећањем дубине расте и варијанса односно грешка на тестирајућем скупу, што у овом примеру није случај.

Даље су приказане вредности метрика за стабла веће дубине.



Слика 13 – Стабло дубине два

	$\widehat{N}$	$\widehat{P}$
N	181200	1319
P	5315	12166

Табела 16 – Матрица конфузије (Стабло дубине два)

Тачност	0,97
Прецизност	0,90
Осетљивост	0,70
F1-скор	0,79

Табела 17 – Вредности метрика (Стабло дубине два)

	$\widehat{N}$	$\widehat{P}$
N	180060	2459
P	1717	15764

Табела 18 – Матрица конфузије (Стабло дубине три)

Тачност	0,98
Прецизност	0,87
Осетљивост	0,90
F1-скор	0,88

Табела 19 – Вредности метрика (Стабло дубине три)

	$\widehat{N}$	$\widehat{P}$
N	182346	173
P	77	17404

Табела 20 – Матрица конфузије (Стабло дубине пет)

Тачност	0,99
Прецизност	0,99
Осетљивост	0,99
F1-скор	0,99

Табела 21 – Вредности метрика (Стабло дубине пет)

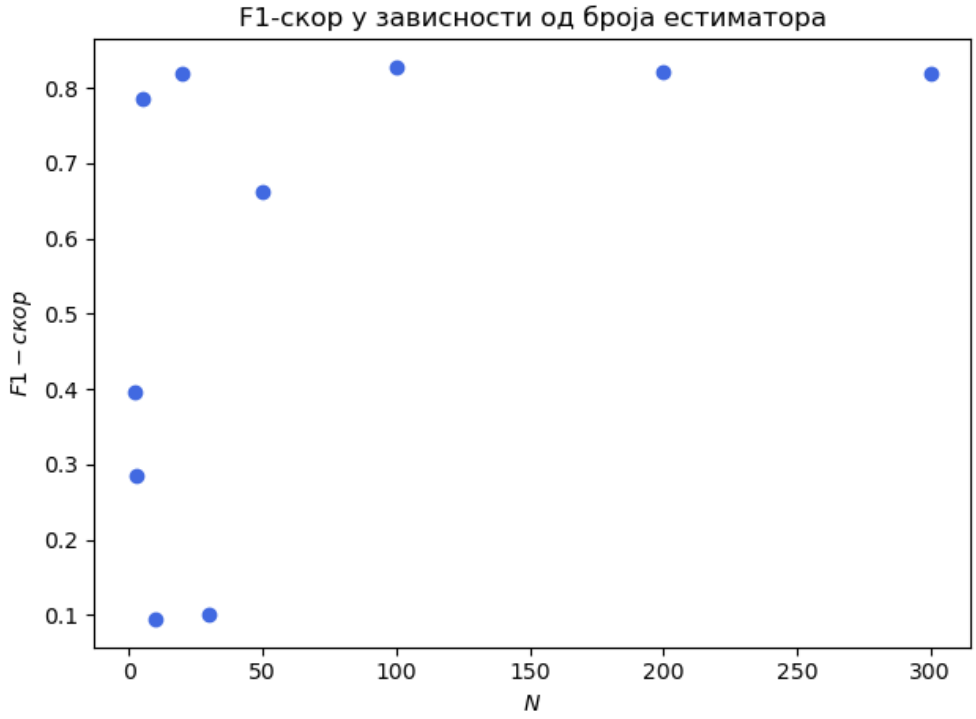
С обзиром на добре резултате примене самог стабла на класификацију, нема претеране потребе за коришћењем других модела, али сам ипак применио неке ансамбл методе.

Случајне шуме

Случајне шуме представљају модел у коме се користи скуп стабала за одлучивање. Свако од тих стабала се обучава на посебном скупу података. Сваки од тих скупова се формира тако што из оригиналног скупа насумично извучемо податак, упишемо у нови скуп, и потом вратимо у оригинални. Пракса каже да се новоформирани скуп састоји од 70% оригинала и 30% копија. Циљ је да направимо класификатор тако да уместо да имамо једног јаког ученика (једно добро обучено стабло које је због своје дубине склоно преобучавању), имамо више слабих ученика, тј. више плитких стабала. Очекује се да ће са повећањем броја чланова ансамбла грешка бити све мања, а с тиме и процена боља.

Одлуке се доносе на основу бројности излаза стабала.

Уколико за различите бројеве чланова ансамбла, при чему је број коришћених обележја фиксиран на два, а максимална дубина стабла износи три, *F1*-скор има ведности приказане на слици 14.



Слика 14 – Случајна шума (F1-скор у зависности од броја естиматора)

Очигледноје да се мора повећати број коришћених предиктора. То сам учинио и поставио тај број на три.

За случај када број естиматора износи десет, макс. дубина четири, а макс. број обележја три, метрике су приказане у табелама 22 и 23.

	$\widehat{N}$	$\widehat{P}$
N	182519	0
P	689	16792

Табела 22 – Матрица конфузије (Случајна шума; N=10, d_{max}=4, n_{max}=3)

Тачност	0,99
Прецизност	1,00
Осетљивост	0,96
F1-скор	0,98

Табела 23 – Вредности метрика (Случајна шума; N=10, d_{max}=4, n_{max}=3)

Време извршења дела кода за обучавање овакве случајне шуме износи 14,37 секунди.

	$\widehat{N}$	$\widehat{P}$
N	182364	155
P	153	17328

Табела 24 – Матрица конфузије (Случајна шума; N=10, d_{max}=5, n_{max}=3)

Тачност	0,99
Прецизност	0,99
Осетљивост	0,99
F1-скор	0,99

Табела 25 – Вредности метрика (Случајна шума; N=10, d_{max}=5, n_{max}=3)

За обучавање случајне шуме са максималном дубином једнаком 5 износило је 15,31 секунду.

XGBoost метода

Boosting алгоритми, као и случајне шуме раде са скупом стабала, с тим да се скупови података на којима се обучавају бирају на другачији начин. Наредно стабло у ансамблу се обучава на подацима на којима је претходно грешило.

	$\widehat{N}$	$\widehat{P}$
N	182352	167
P	79	17402

Табела 26 – Матрица конфузије (XGBoost; N=10, d_{max}=4, n_{max}=3)

Тачност	0,99
Прецизност	0,99
Осетљивост	0,99
F1-скор	0,99

Табела 27 – Вредности метрика (XGBoost; N=10, d_{max}=4, n_{max}=3)

За обучавање овог модела било је потребно 7,92 секунде. Осим што даје боље метрике од случајне шуме овај модел се и обучава значајно брже од случајне шуме.