Predvidjanje kvaliteta vina na osnovu hemijskih aributa

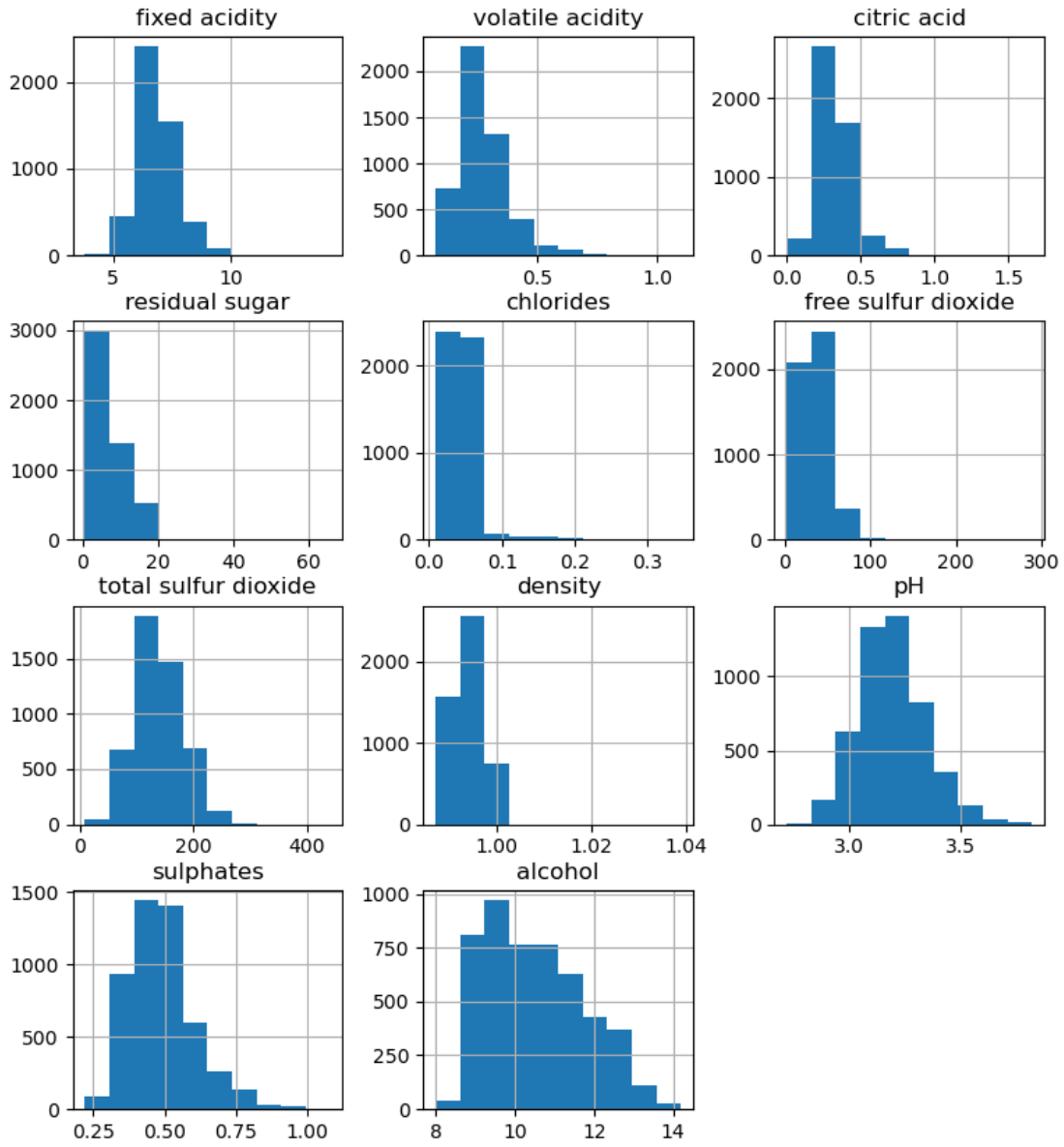
Autor: Aleksandra Petrovic 4008/21

Uvod

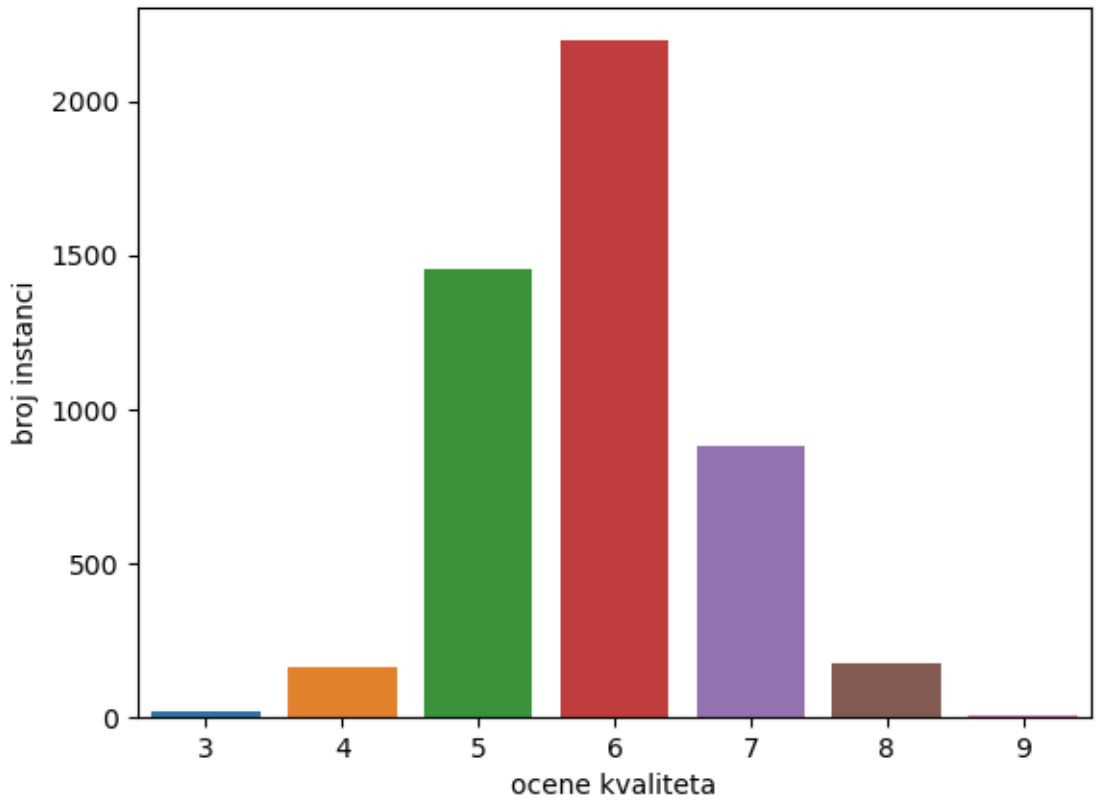
- Cilj ovog projekta je da se isprobaju razni modeli za predvidjanje kvaliteta vina na osnovu hemijskih atributa koji su dostupni u bazi podataka
- Baza podataka koja se koristi je na linku
- Sastoji se iz dva skupa podataka, belog i crvenog vina “Vihno Verde” koje potice iz Portugalije
- U ovom projektu analiziran je skup belog vina
- Ima 11 atributa i 4898 instanci
- Kolona “quality” govori o kvalitetu vina i uzima vrednosti od 0-10

Pretprocesiranje

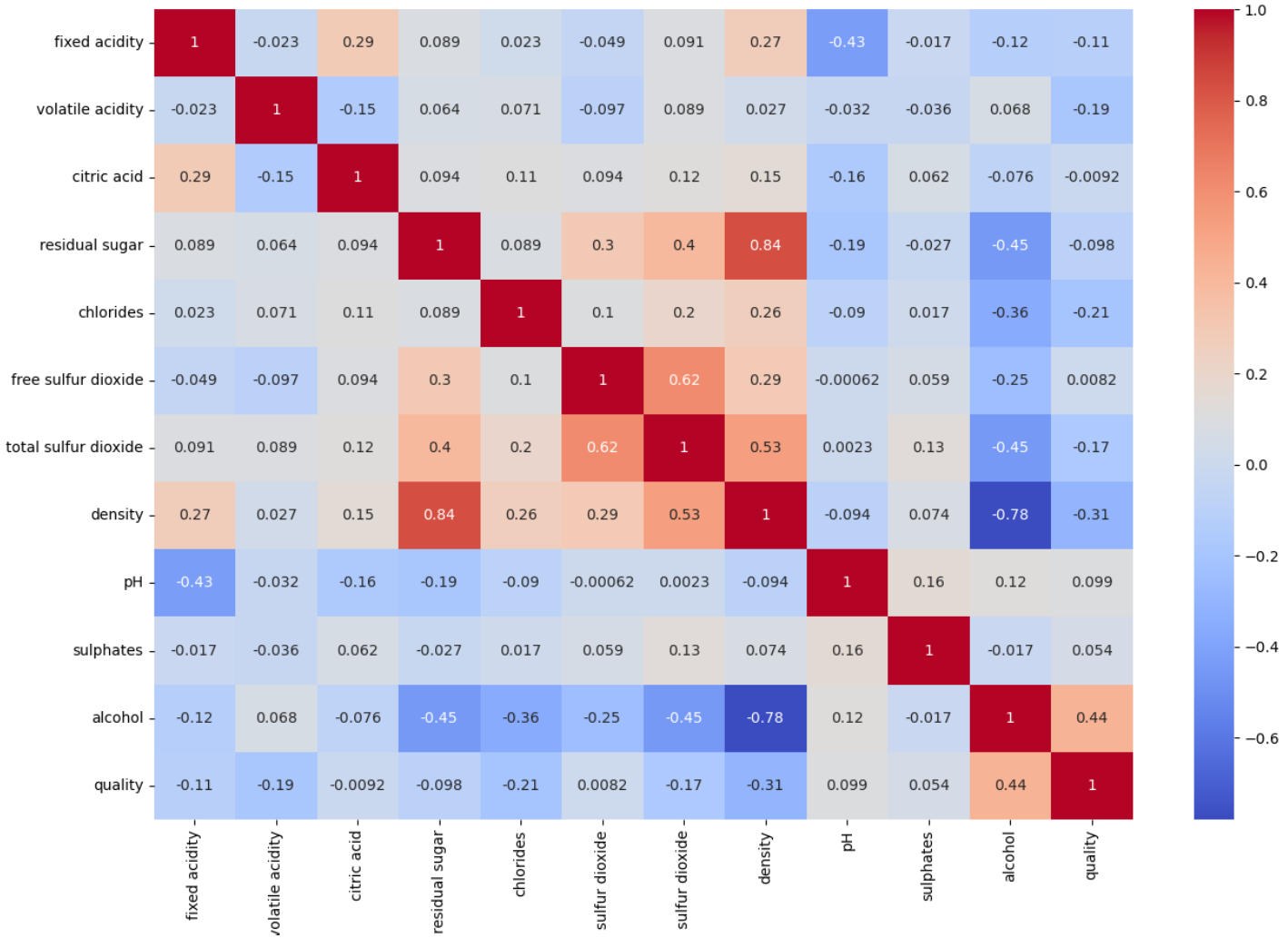
- Svi atributi su neprekidnog tipa
- Nema nedostajucih podataka
- Ulazne promenljive (bazirane na fizičko-hemijskim testovima):
  - 1 - fiksna kiselost (fixed acidity)
  - 2 - fluktuirajuća kiselost (volatile acidity)
  - 3 - limunska kiselina (citric acid)
  - 4 - ostatak šećera (residual sugar)
  - 5 - hloridi (chlorides)
  - 6 - slobodan sumpor-dioksid (free sulfur dioxide)
  - 7 - ukupan sumpor-dioksid (total sulfur dioxide)
  - 8 - gustina (density)
  - 9 - pH
  - 10 - sulfati (sulphates)
  - 11 - alkohol (alcohol)
- Izlazna promenljiva (bazirana na senzornim podacima):
  - 12 - kvalitet - ocena između 0 i 10 (quality)


fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol

- Neizbalansirana podela po ocenama kvaliteta na osnovu histograma
- Najvise srednje ocenjenih vina, malo sa visokom ocenom (odlicnih) kao i sa niskom ocenom (losih)



- Matrica konfuzije, korelacija izmedju kolona



Modeli

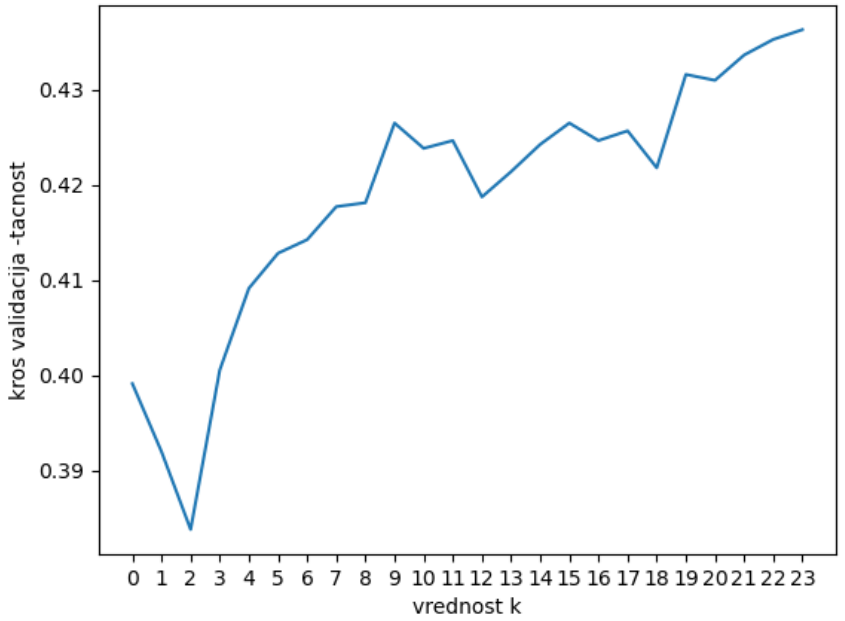
- Modeli koji su isprobani u ovom projektu su:
  - KNN – K najblizih suseda
  - Linearna regresija sa tezinama
  - AdaBoost
  - Multinomijalna logisticka regresija
  - Potpuno povezana neuronska mreza
  - Aditivna logisticka regresija
  - SMOTE algoritam

Podela i standardizacija

- Izvrsena je podela na test i trening skup
- Trening skup je 80%
- Standardizacija je izvrsena standardnim scaler-om

KNN

- K najblizih suseda je model koji klasifikuje novu instancu na osnovu k vrednosti kojih se nalaze u njegovom okruzenju
- Isprobano je 1-25 vrednosti za k, sa ocenom tacnoscu i koriscena je kros validacija za izbor modela
- Izabran je k=19 i treniran je model za tu vrednost



- Izracunata je tacnost modela na osnovu test skupa
- Tacnost je 0.57
- Izvestaj:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	4
4	0.00	0.00	0.00	33
5	0.62	0.62	0.62	291
6	0.58	0.71	0.64	440
7	0.43	0.35	0.38	176
8	0.00	0.00	0.00	35
9	0.00	0.00	0.00	1
accuracy			0.57	980
macro avg	0.23	0.24	0.23	980
weighted avg	0.52	0.57	0.54	980

Linearna regresija sa tezinama

- Linearna regresija sa tezinama je neprarametarski model
- Koristi se kada zelimo da dodamo tezinu nekim instancama, nesto sto se smatra bitnijim ili manje bitnim
- Isprobane tezine
  - w1=1/y_train
  - w2=np.abs((1-np.sum(y_train))/y_train)
  - w3= np.abs(np.random.randn(3918))
- Izabrana je prva tezina za treniranje modela, zbog najmanje greske
- MSE=0.67
- MAPE=0.09
- R2=0.15

- Tacnost je 0.53
- Izvestaj

	precision	recall	f1-score	support
3	0.00	0.00	0.00	4
4	0.14	0.03	0.05	33
5	0.58	0.53	0.56	291
6	0.52	0.75	0.61	440
7	0.43	0.17	0.24	176
8	0.00	0.00	0.00	35
9	0.00	0.00	0.00	1
accuracy			0.53	980
macro avg	0.24	0.21	0.21	980
weighted avg	0.49	0.53	0.49	980

AdaBoost

- AdaBoost (Adaptive Boosting) je algoritam ansambla (ensemble) koji se koristi za poboljšanje performansi klasifikacionih modela
- Zasniva se na promenama tezina instanci
- Neophodno je prvo da se napravi bazni model, dodele mu se parametri
- Zatim se pravi ansambl i trenira se

- Tacnost je 0.7
- Izvestaj:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	4
4	0.82	0.27	0.41	33
5	0.73	0.70	0.72	291
6	0.67	0.81	0.73	440
7	0.69	0.59	0.63	176
8	1.00	0.37	0.54	35
9	0.00	0.00	0.00	1
accuracy			0.70	980
macro avg	0.56	0.39	0.43	980
weighted avg	0.71	0.70	0.69	980

- Matrica konfuzije

```
[[  0   0   1   3   0   0   0]
 [  0   9  16   7   1   0   0]
 [  0   2 204  84   1   0   0]
 [  0   0  55 355  30   0   0]
 [  0   0   3  70 103   0   0]
 [  0   0   0   9  13  13   0]
 [  0   0   0   0   1   0   0]]
```

- MSE=0.4
- R2=0.49
- MAPE=0.6

Multinomijalna logisticka regresija

- Statisticki model koji se koristi za predvidjanje vise klasa
- Za probleme klasifikacije se koristi
- Generalizuje logisticku regresiju za vise od dve klase
- Softmax funkcija pretvara tezinu i ulaz u verovatnocu za svaku od klasa
- Tacnost je 0.54
- MSE=0.64
- MAPE=0.09
- R2=0.18

- Matrica konfuzije

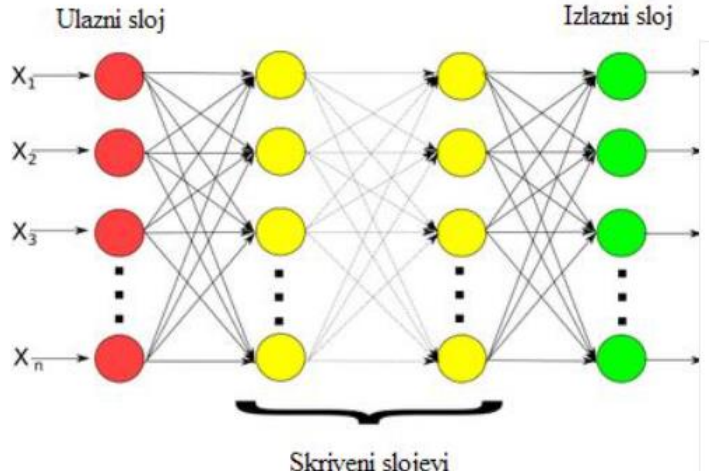
```
[[  0   0   1   3   0   0   0]
 [  0   3  19  10   1   0   0]
 [  0   1 162 125   2   1   0]
 [  0   2  79 324  35   0   0]
 [  0   0   6 128  42   0   0]
 [  0   0   1  23  11   0   0]
 [  0   0   0   0   1   0   0]]
```

- Izvestaj

	precision	recall	f1-score	support
3	0.00	0.00	0.00	4
4	0.50	0.09	0.15	33
5	0.60	0.56	0.58	291
6	0.53	0.74	0.62	440
7	0.46	0.24	0.31	176
8	0.00	0.00	0.00	35
9	0.00	0.00	0.00	1
accuracy			0.54	980
macro avg	0.30	0.23	0.24	980
weighted avg	0.52	0.54	0.51	980

Potpuno povezana neuronska mreza

- Potpuno povezana neuronska mreza je tip mreza gde je svaki neuron povezan sa svim neuronima iz prethodnog i narednog sloja
- Sadrzi ulazni sloj, skrivene slojeve i izlazni sloj
- Koristi se Keras i Tensorflow biblioteka
- Ulaz je broj atributa
- Izlaz je vrednost ocena kvaliteta (0-10)
- Isprobano je vise modela sa razlicitim aktivacionim funkcijama, razlicitim brojevima slojeva i neurona
- model = Sequential([Input(shape=(number_of_features,)),
- Dense(units=64, activation='relu'),
- Dense(units=32, activation='relu'),
- Dense(units=output_size,activation='linear')
-])

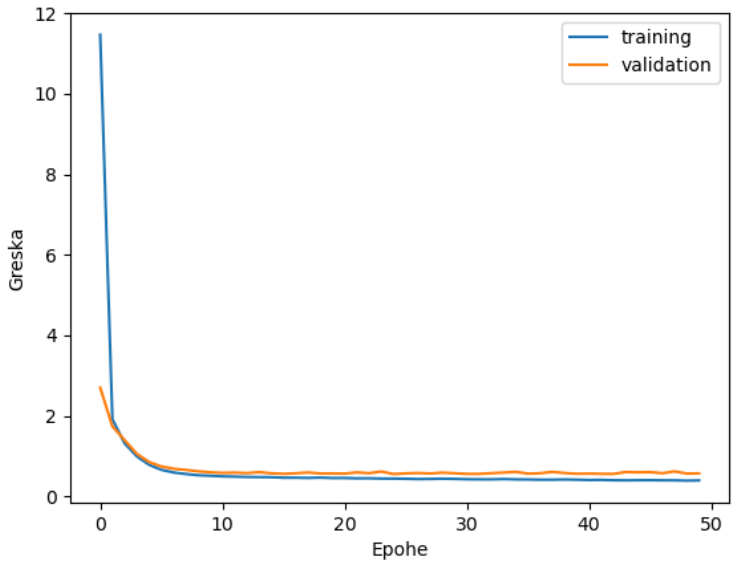


- Model

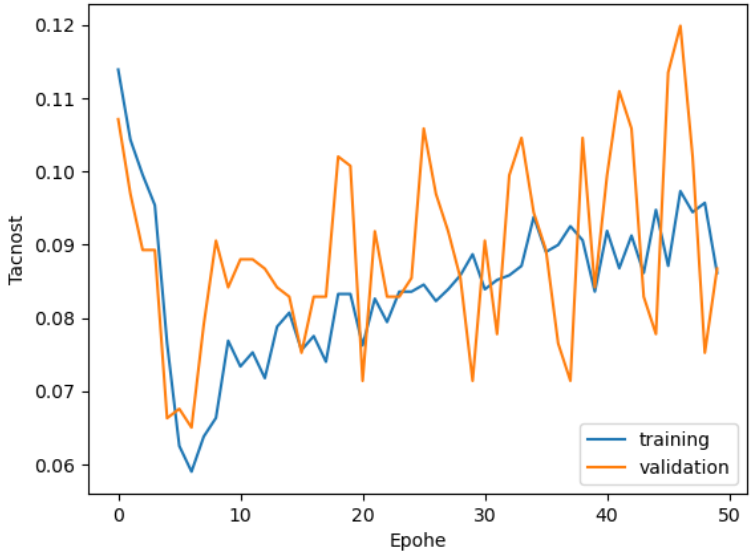
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                768

 dense_1 (Dense)             (None, 32)                2080

 dense_2 (Dense)             (None, 10)                330

=================================================================
Total params: 3178 (12.41 KB)
Trainable params: 3178 (12.41 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

- Model se kompajlira sa optimizatorom ADAM, greskom srednjom kvadratnom i metrikom tacnoscu
- Model se trenira za 50 epoha, gde je batch_size 16, validacioni skup je 20%

- Grafik promene greske



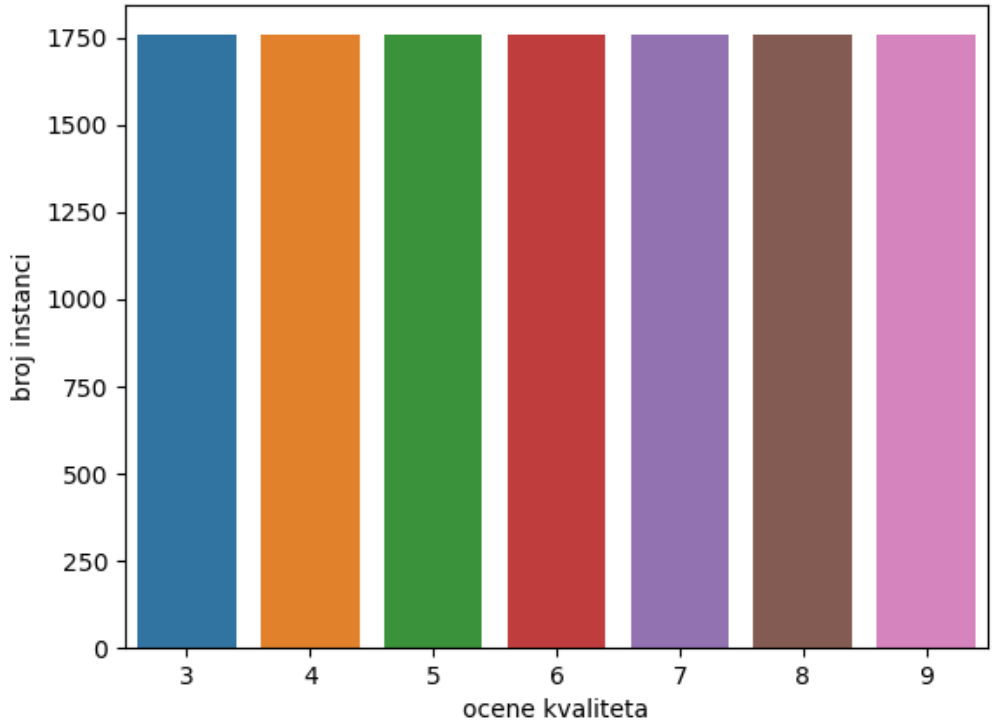
Grafik promene tacnosti



- Evaluiran model na test skupu
  - Tacnost 0.083, Greska 0.49
- Evaluiran model na trening skupu
  - Tacnost 0.081, Greska 0.41

SMOTE

- Prilikom koriscenja SMOTE algoritma, gde se postize balansiranje klasa, nije dobijen nikakav bolji rezultat



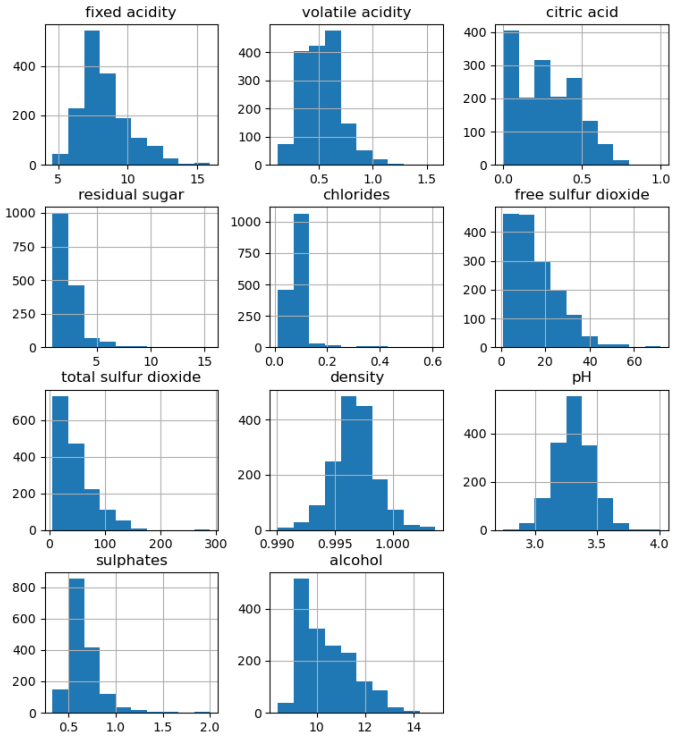
Rezultati

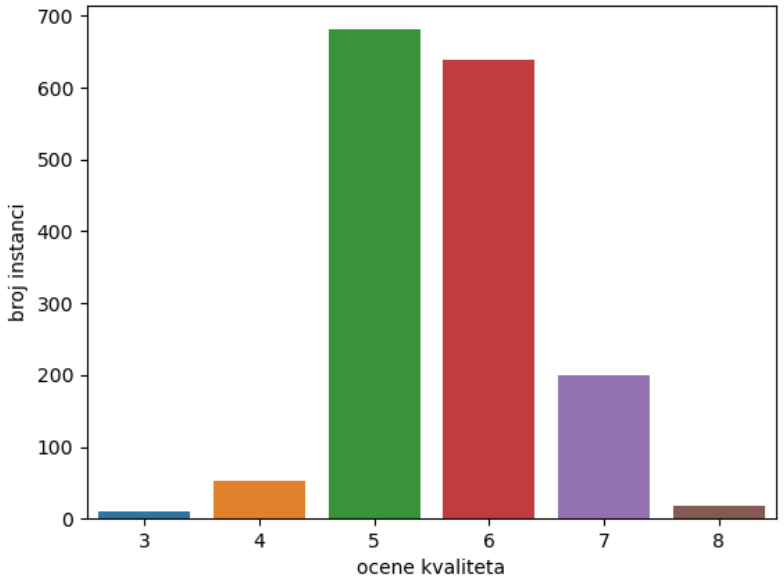
- Prema dobijenim podacima iz isprobanih modela, najbolju tacnost za predvidjanje ima model AdaBoost, gde tacnost iznosi 0.7

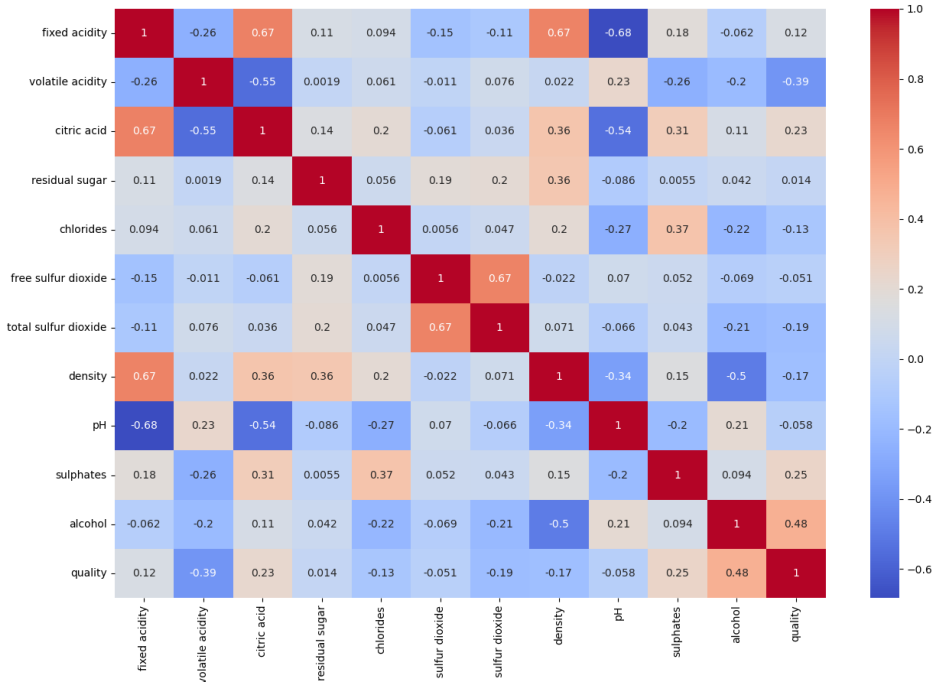
MODEL	TACNOST na test skupu
KNN	0.57
Linearna regresija sa tezinama	0.53
AdaBoost	0.7
Multinomijalna logisticka regresija	0.54
Potpuno povezana neuronska mreza	0.081

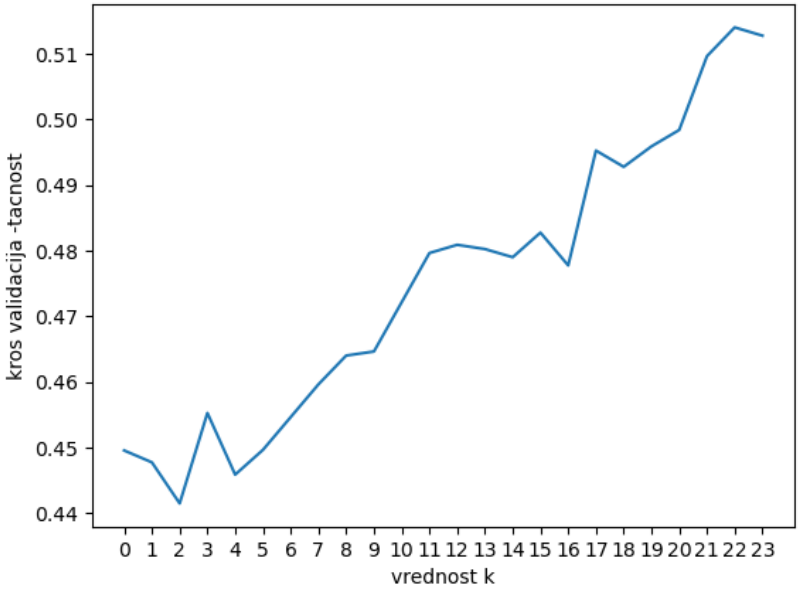
Skup crvenog vina

- Modeli KNN, Linearna regresija sa tezinama, AdaBoost i multinomijalna logisticka regresija isporbani su na podacima i crvenog vina



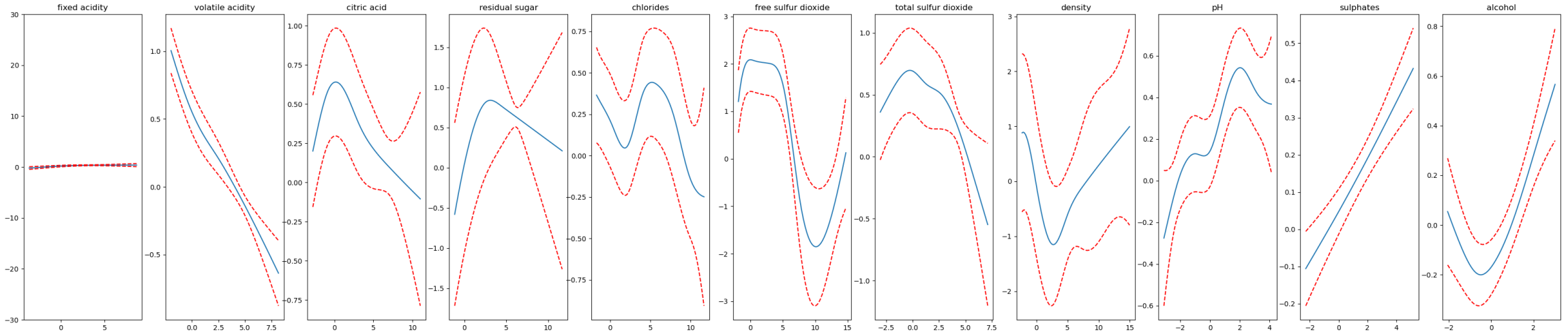



fixed acidity 1 -0.26 0.67 0.11 0.094 -0.15 -0.11 0.67 -0.68 0.18 -0.062 0.12
volatile acidity -0.26 1 -0.55 0.0019 0.061 -0.011 0.076 0.022 0.23 -0.26 -0.2 -0.39
citric acid 0.67 -0.55 1 0.14 0.2 -0.061 0.036 0.36 -0.54 0.31 0.11 0.23
residual sugar 0.11 0.0019 0.14 1 0.056 0.19 0.2 0.36 -0.086 0.0055 0.042 0.014
chlorides 0.094 0.061 0.2 0.056 1 0.0056 0.047 0.2 -0.27 0.37 -0.22 -0.13
free sulfur dioxide -0.15 -0.011 -0.061 0.19 0.0056 1 0.67 -0.022 0.07 0.052 -0.069 -0.051
total sulfur dioxide -0.11 0.076 0.036 0.2 0.047 0.67 1 0.071 -0.066 0.043 -0.21 -0.19
density 0.67 0.022 0.36 0.36 0.2 -0.022 0.071 1 -0.34 0.15 -0.5 -0.17
pH -0.68 0.23 -0.54 -0.086 -0.27 0.07 -0.066 -0.34 1 -0.2 0.21 -0.058
sulphates 0.18 -0.26 0.31 0.0055 0.37 0.052 0.043 0.15 -0.2 1 0.094 0.25
alcohol -0.062 -0.2 0.11 0.042 -0.22 -0.069 -0.21 -0.5 0.21 0.094 1 0.48
quality 0.12 -0.39 0.23 0.014 -0.13 -0.051 -0.19 -0.17 -0.058 0.25 0.48 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
sulfur dioxide
sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

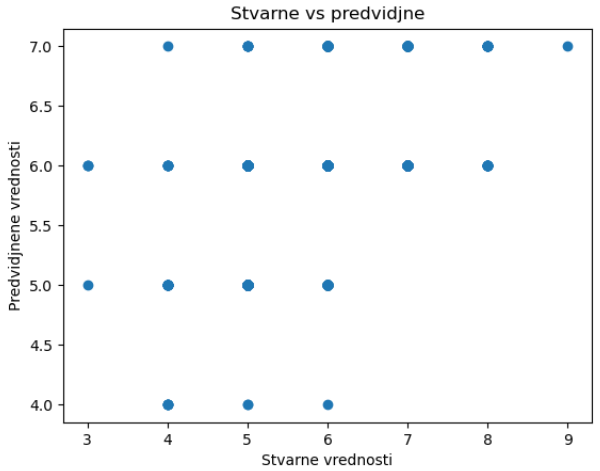

kros validacija -tacnost
0.51
0.50
0.49
0.48
0.47
0.46
0.45
0.44
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
vrednost k

Aditivna logisticka regresija

- Omogucava nelinearne odnose izmedju ciljne promenljive i atributa
- Splajnovi modeluju nelinearne odnose – ima ih 10, za svaku od klasa
- Aditivnost zbog glatkih funkcija
- Parametar lambda je zaduzen za glatkocu funkcije



- GridSearch parametara lambda, 100x11
- Dobijeni najbolji i korisceni za treniranje modela
- Zatim predvidjanje i evaluacija
- MSE=0.63
- MAPE=0.09
- R2=0.19



LinearGAM

Distribution:	NormalDist	Effective DoF:	38.554
Link Function:	IdentityLink	Log Likelihood:	-4764.6094
Number of Samples:	3918	AIC:	9608.3267
		AICc:	9609.1541
		GCV:	0.5137
		Scale:	0.5046
		Pseudo R-Squared:	0.3627

Feature Function	Lambda	Rank	EDoF	P > x	Sig. Code
s(0)	[134.6572]	10	2.9	3.65e-06	***
s(1)	[119.1806]	10	2.4	1.11e-16	***
s(2)	[7.1224]	10	3.5	8.95e-08	***
s(3)	[1.8874]	10	4.2	1.11e-16	***
s(4)	[0.4937]	10	5.4	5.75e-03	**
s(5)	[0.0833]	10	5.2	1.11e-16	***
s(6)	[6.3033]	10	3.3	8.51e-07	***
s(7)	[0.0513]	10	2.8	1.11e-16	***
s(8)	[2.4203]	10	4.9	1.83e-12	***
s(9)	[2108.3145]	10	1.2	6.28e-05	***
s(10)	[65.2]	10	2.7	7.46e-09	***
intercept		1	0.0	1.05e-10	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
terms=s(0) + s(1) + s(2) + s(3) + s(4) + s(5) + s(6) + s(7) + s(8) + s(9) + s(10) + intercept
```

Listing paketa

- Numpy
- Matplotlib
- Pandas
- Sklearn
- Tensorflow
- Keras
- Pygam

Literatura

- Skripta “Masinsko ucenje” , Mladen Nikolic, Andjelka Zecevic, Beograd 2019
- Predavanja i vezbe predmeta Masinsko ucenje na Masinskom fakultetu
- https://cs229.stanford.edu/proj2015/245_report.pdf