

Predviđjanje kvaliteta vina na osnovu hemijskih atributa

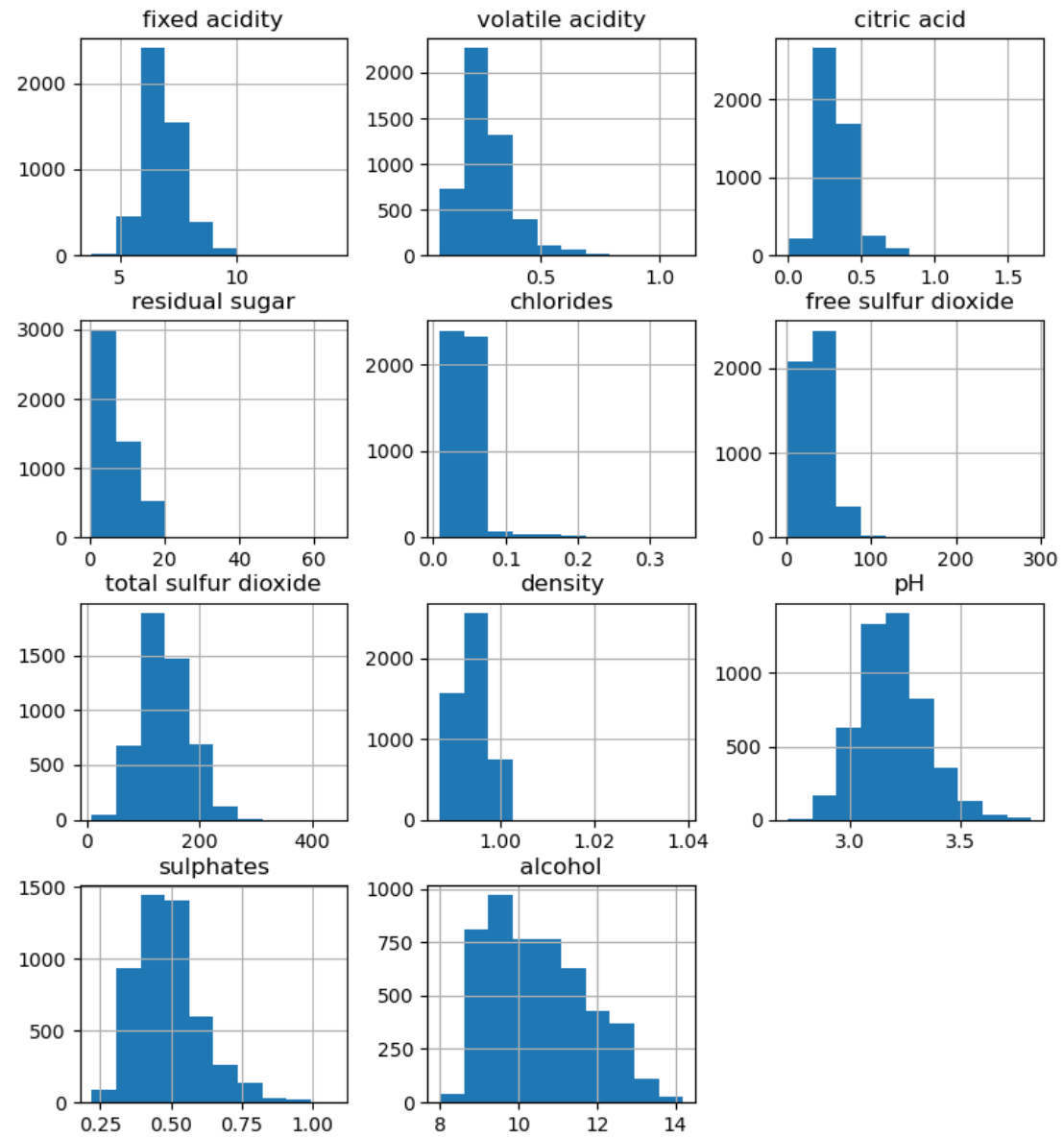
Autor: Aleksandra Petrovic 4008/21

Uvod

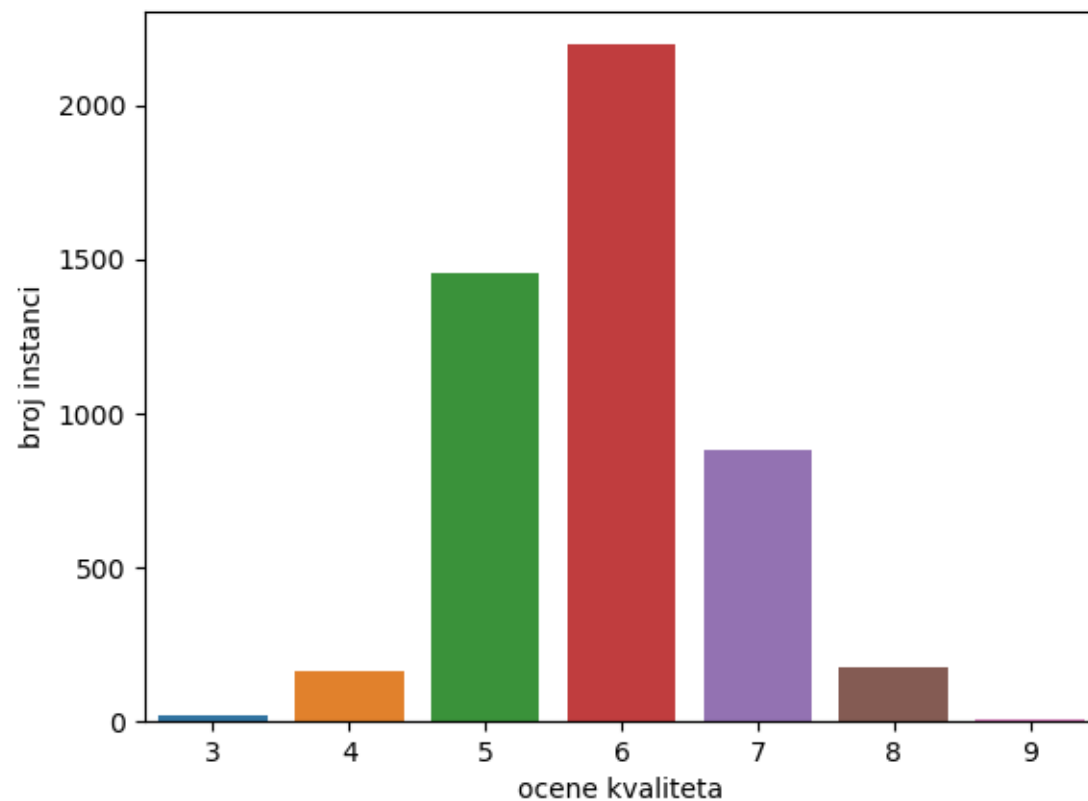
- Cilj ovog projekta je da se isprobaju razni modeli za predviđanje kvaliteta vina na osnovu hemijskih atributa koji su dostupni u bazi podataka
- Baza podataka koja se koristi je na [linku](#)
- Sastoji se iz dva skupa podataka, belog i crvenog vina “Vihno Verde” koje potice iz Portugaliije
- U ovom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

Preprocesiranje

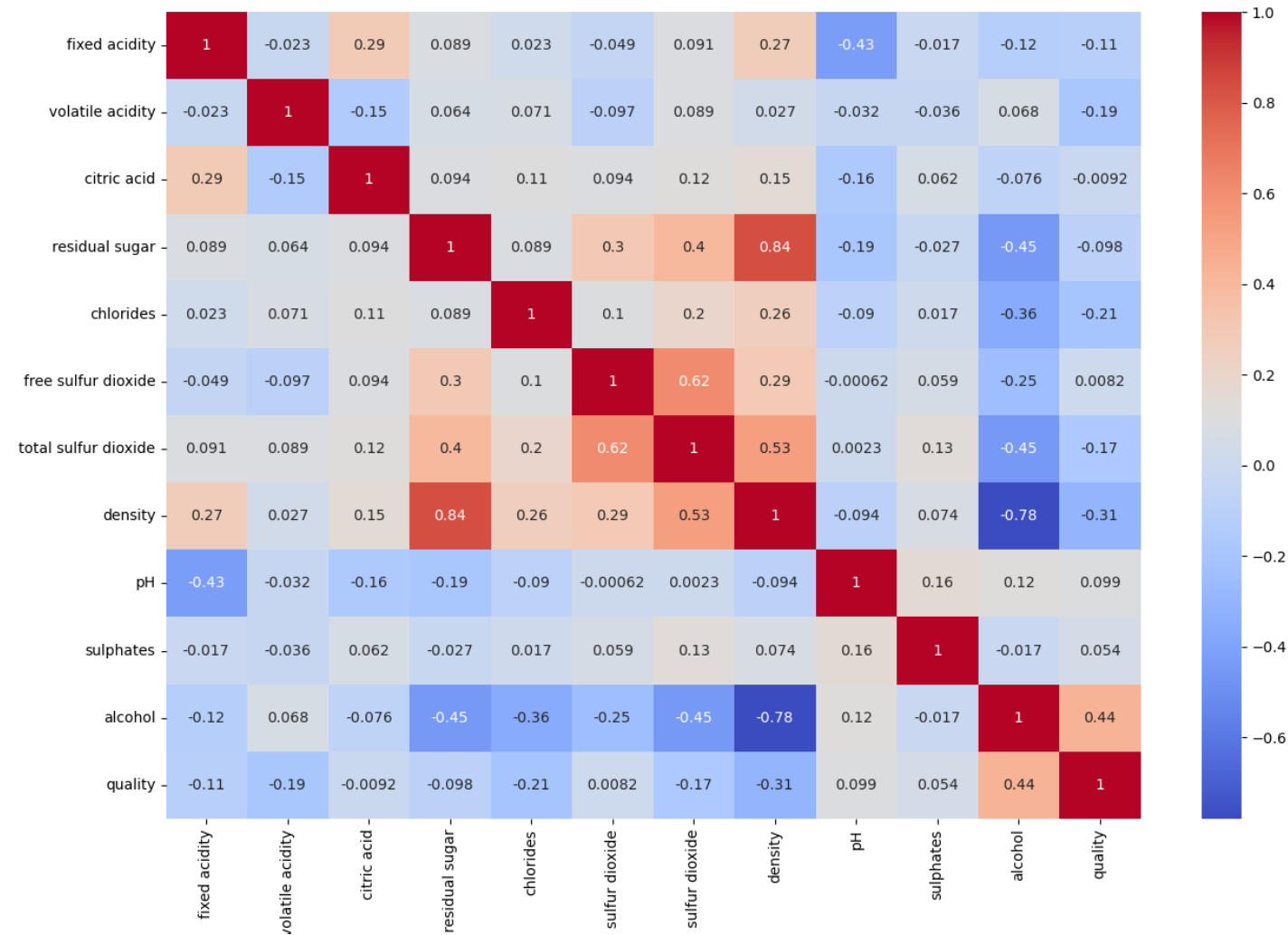
- Svi atributi su neprekidnog tipa
- Nema nedostajucih podataka
- Ulazne promenljive (bazirane na fizičko-hemijskim testovima):
 - 1 - fiksna kiselost (fixed acidity)
 - 2 - fluktuirajuća kiselost (volatile acidity)
 - 3 - limunska kiselina (citric acid)
 - 4 - ostatak šećera (residual sugar)
 - 5 - 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)



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



- Matrica konfuzije, korelacija izmedju kolona



Modeli

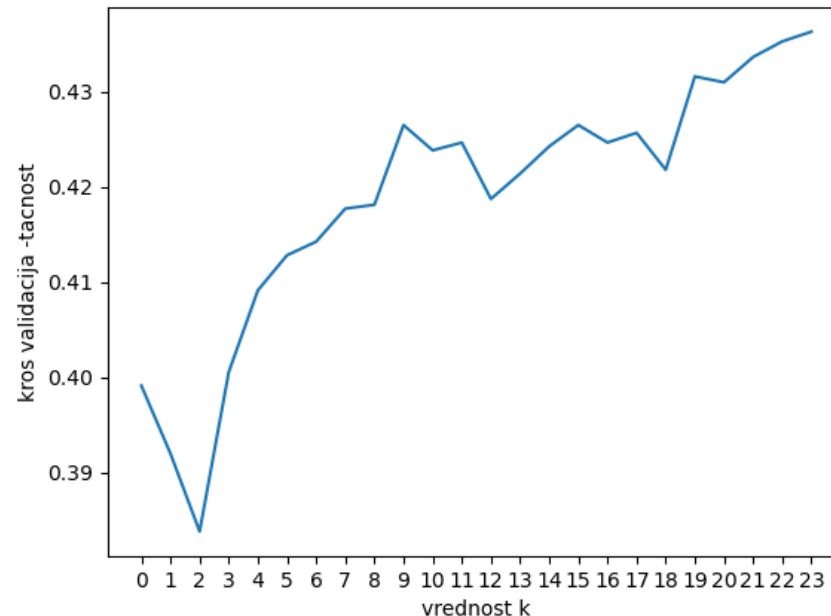
- Modeli koji su isprobani u ovom projektu su:
 - KNN – K najblizih suseda
 - Linearna regresija sa tezinama
 - AdaBoost
 - Multinomijalna logisticka regresija
 - Potpuno povezana neuronska mreza

Podela i standardizacija

- Izvršena je podela na test i trening skup
- Treening skup je 80%
- Standardizacija je izvršena 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
- $MSE=0.67$
- $MAPE=0.09$
- $R^2=0.15$

- Tacnost je 0.52
- 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 logistička regresija

- Statistički model koji se koristi za predviđanje više klasa
- Za probleme klasifikacije se koristi
- Generalizuje logističku regresiju za više od dve klase
- Softmax funkcija pretvara tezinu i ulaz u verovatnocu za svaku od klasa
- Tacnost je 0.54
- $MSE=0.64$
- $MAPE=0.09$
- $R^2=0.18$

- Matrica konfuzije

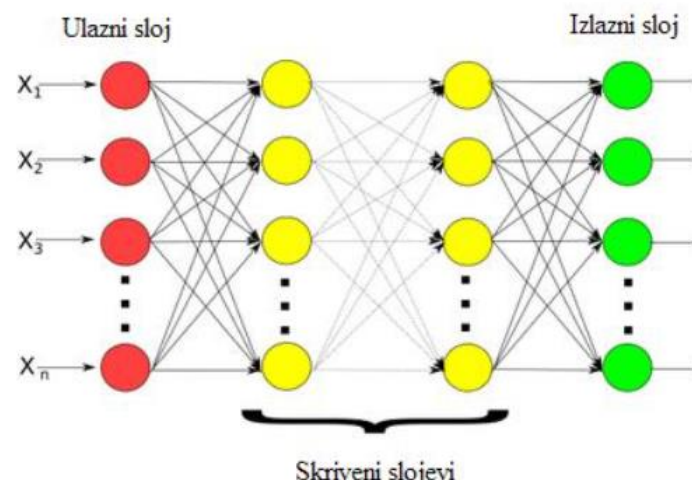
```
[[ 0  0  1  3  0  0  0]
 [ 0  3 19 10  1  0  0]
 [ 0  1 62 125  2  1  0]
 [ 0  2 79 324 35  0  0]
 [ 0  0  6 128 42  0  0]
 [ 0  0  1 23 11  0  0]
 [ 0  0  0  0  1  0  0]]
```

- 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

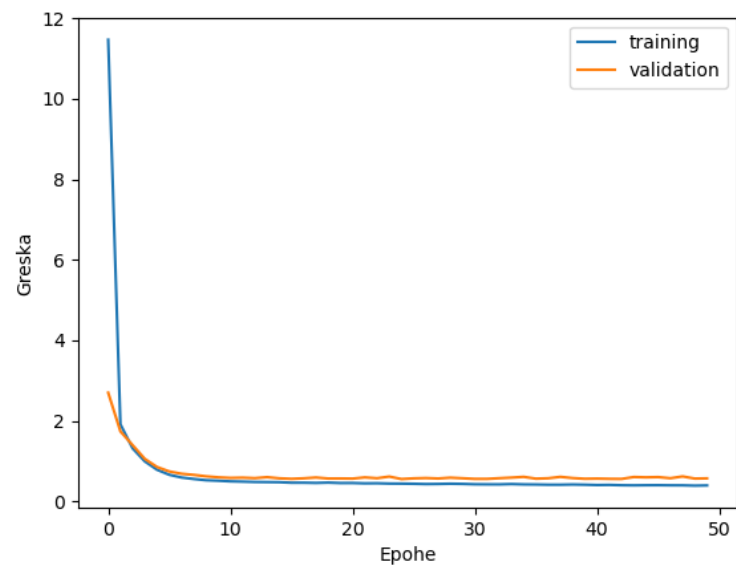
Model: "sequential"

| Layer (type)    | Output Shape | Param # |
|-----------------|--------------|---------|
| dense (Dense)   | (None, 64)   | 768     |
| dense_1 (Dense) | (None, 32)   | 2080    |
| dense_2 (Dense) | (None, 10)   | 330     |

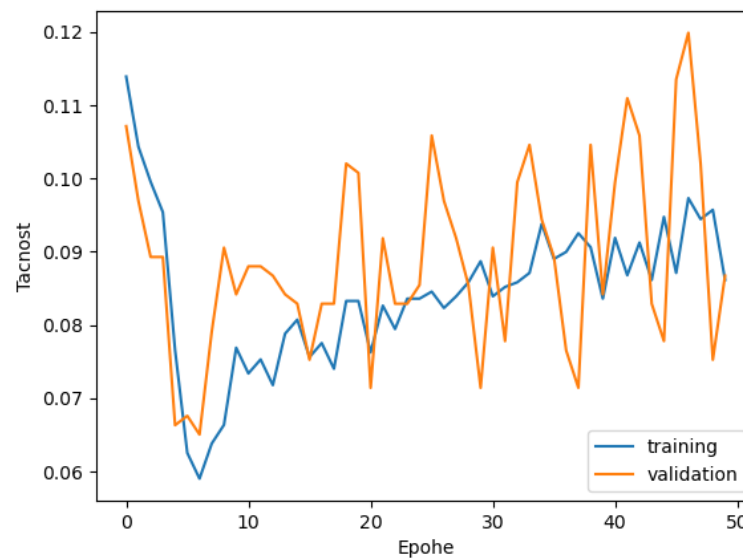
=====  
Total params: 3178 (12.41 KB)  
Trainable params: 3178 (12.41 KB)  
Non-trainable params: 0 (0.00 Byte)  
=====

- 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



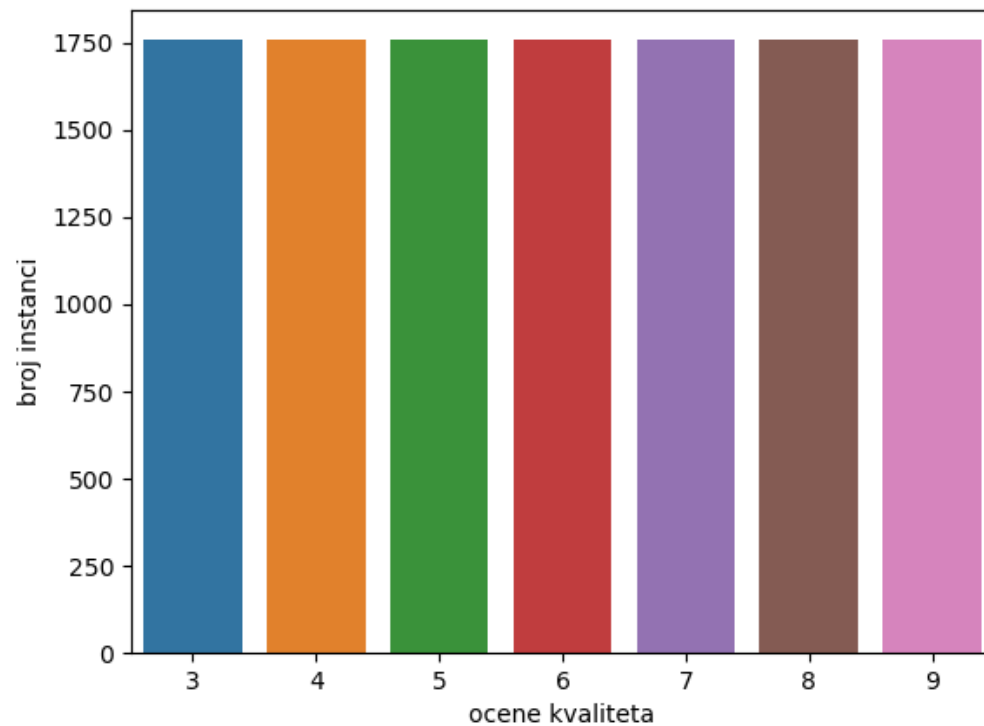
- Grafik promene tacnosti



- 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



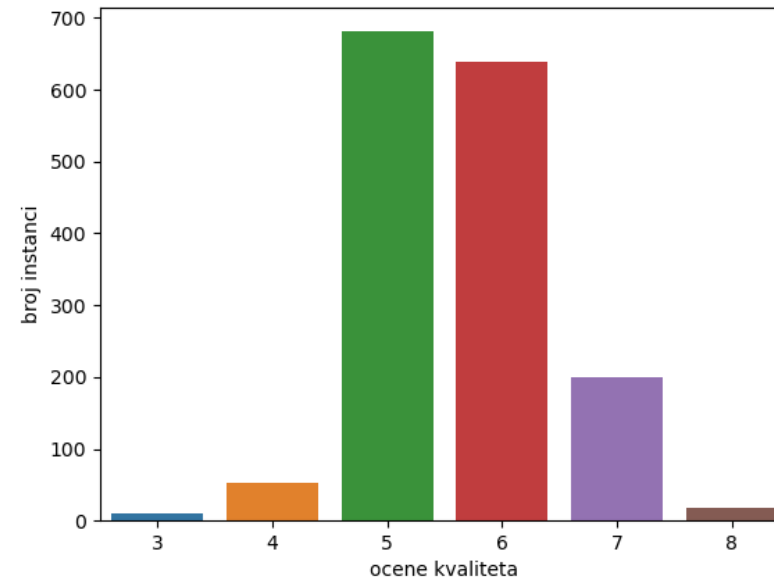
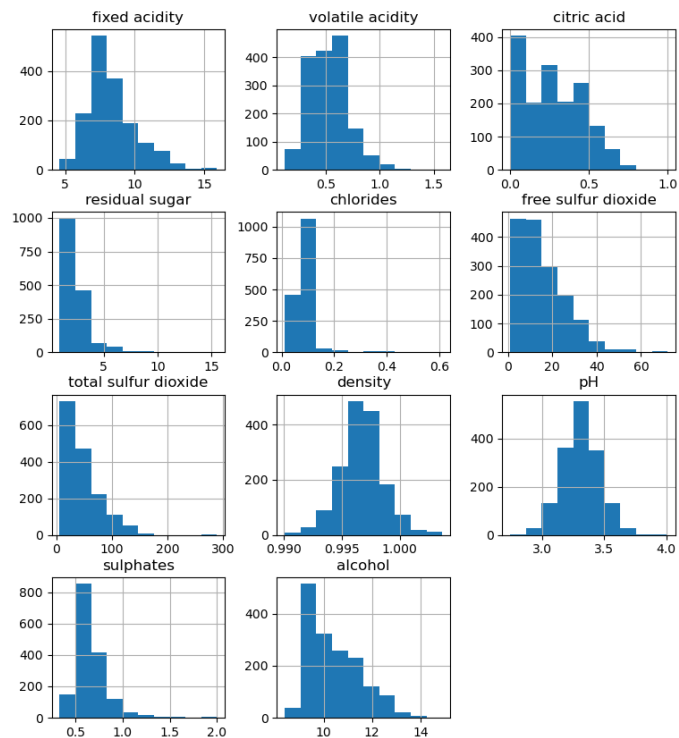
# Rezultati

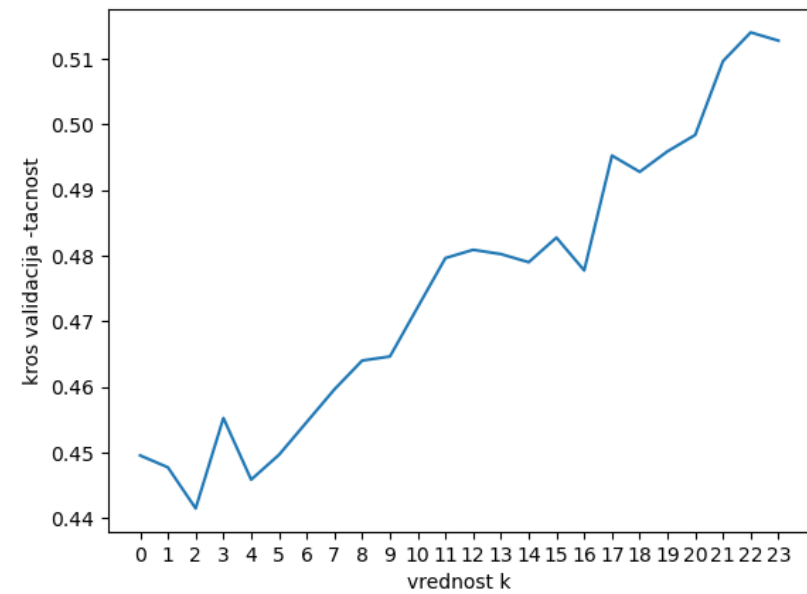
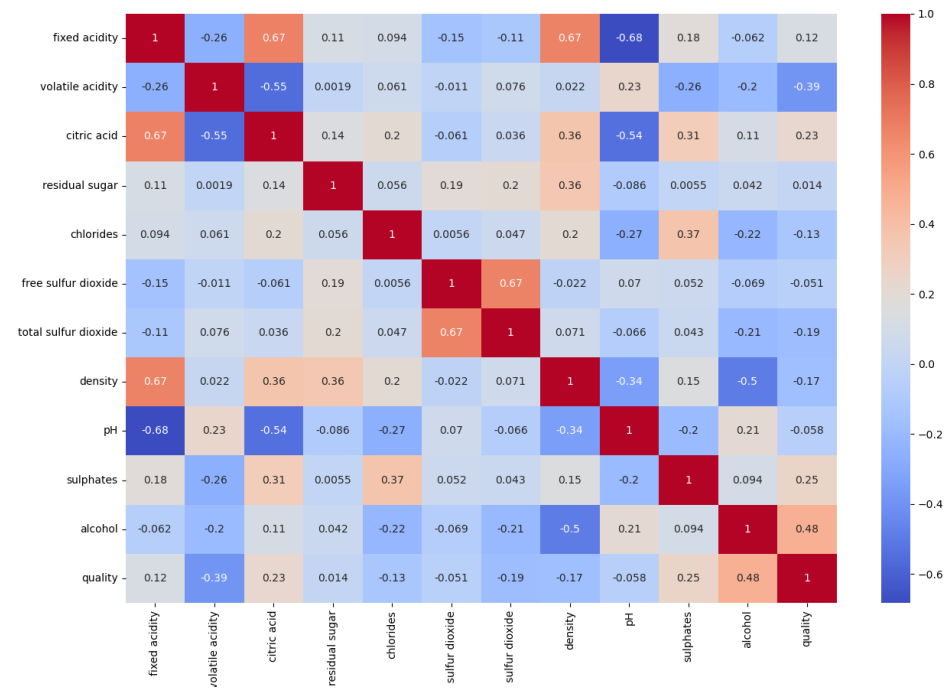
- Prema dobijenim podacima iz isprobanih modela, najbolju tacnost za predvidjanje ima model AdaBoost, gde tacnost iznosi 0.7

| 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







# Listing paketa

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

# Literatura

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