PRIRODNO-MATEMATIČKI FAKULTET AKADEMSKA 2021/22. GODINA

KLASIFIKACIJA SKUPA PODATAKA O CRVENOM VINU

SEMINARSKI RAD IZ PREDMETA OSNOVE RAČUNARSKIH SISTEMA 2

Student: Irena Subotić

Banja Luka, februar 2022.godina

SADRŽAJ

UVOD

Istraživanje podataka(Data Mining) je proces prikupljanja,čišćenja i obrade podataka, odnosno njihove analize kako bi se izvukle prethodno nepoznate i zanimljive informacije ili šabloni.

Metode istraživanja podataka:

- metode predviđanja (cilj je na osnovu dostupnih podataka predvidjeti buduće stanje)
- metode opisa (pronalaženje obrazaca koji mogu da se koriste za razumijevanje podataka)

Klasifikacija je metoda predikcije (predviđanja) i njen cilj je predvidjeti vrijednost drugog atributa na osnovu postojećih, odnosno potrebno je odrediti kojoj kategoriji (klasi) ciljnog atributa pripada određeni primjer iz skupa podataka.

Tema ovog seminarskog rada je klasifikacija podataka o crvenom vinu.

U ovom seminarskom radu će biti korištene sljedeće metode klasifikacije:

1. STABLO ODLUČIVANJA

2. KNN ALOGORITAM

3. NAIVE BAYES ALGORITAM

Podaci koji će biti obrađeni su podaci o vinu i cilj će biti odrediti kvalitet vina na osnovu njegovih osobina. Kvalitet vina se mjeri ocjenom od 1 do 10 ali u ovom skupu podataka se nalaze samo podaci čije su ocjene kvaliteta 3,4,5,6,7,8.

Skup podataka koji će biti obrađen se može pronaći na sljedećem linku:

https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

ANALIZA SKUPA PODATAKA

Ovaj skup podataka sadrži 12 kolona i 1599 redova. Sljedeća slika pokazuje učitani skup podataka.

```
In [120]: data=pd.read_csv('winequality-red.csv')
```

```
In [121]: data
```

Out[121]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

1599 rows × 12 columns

Skup podataka nema nedostajuće vrijednosti.

```
In [153]: data.isnull().sum()
```

```
Out[153]: fixed acidity          0
          volatile acidity       0
          citric acid            0
          residual sugar         0
          chlorides              0
          free sulfur dioxide    0
          total sulfur dioxide   0
          density                0
          pH                     0
          sulphates              0
          alcohol                0
          quality                0
          dtype: int64
```

Sljedeća slika pokazuje nazive kolona i kog tipa su vrijednosti u kolonama.

```
In [154]: data.columns

Out[154]: Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
                 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
                 'pH', 'sulphates', 'alcohol', 'quality'],
                dtype='object')

In [155]: data.dtypes

Out[155]: fixed acidity           float64
          volatile acidity        float64
          citric acid             float64
          residual sugar          float64
          chlorides               float64
          free sulfur dioxide     float64
          total sulfur dioxide    float64
          density                 float64
          pH                      float64
          sulphates               float64
          alcohol                 float64
          quality                   int64
          dtype: object
```

1 - fiksna kiselost: većina kiselina uključenih u vino (fiksne ili neisparljive kiseline su kiseline koje ne isparavaju lako)

2 - isparljiva kiselost: količina sirćetne kiseline u vinu koja na previsokim nivoima može dovesti do neprijatnog ukusa

3 - limunska kiselina: nalazi se u malim količinama, limunska kiselina može dodati "svježinu" i ukus vinima

4 - rezidualni šećer: količina šećera preostala nakon prestanka fermentacije, rijetko je naći vina sa manje od 1 gram/litar, vina sa više od 45 grama/litar se smatraju slatkim

5 - hloridi: količina soli u vinu

6 - slobodni sumpor dioksid: slobodni oblik SO2 postoji u ravnoteži između molekulskog SO2 (kao rastvorenog gasa) i bisulfitnog jona; usporava rast mikroba i oksidaciju vina

7 - ukupni sumpor dioksid: količina slobodnih i vezanih oblika S02; u niskim koncentracijama SO2 se uglavnom ne može detektovati u vinu, ali pri koncentracijama slobodnog SO2 preko 50 ppm, SO2 postaje očigledan po mirisu i ukusu vina

8 - gustina: gustina u zavisnosti od procenta alkohola i šećera

9 - pH: opisuje koliko je vino kiselo ili bazno na skali od 0 (veoma kiselo) do 14 (veoma bazno); većina vina je između 3-4 na pH skali

10 - sulfati: aditiv za vino koji može doprinijeti nivou gasa sumpor-dioksida (S02), koji djeluje kao antimikrobno i antioksidativno

11 - alkohol: procenat alkohola u vinu

12 - kvalitet (0-10)

Pomoću metode describe() se ispišu statistički podaci kolona.

```
In [191]: data.describe().T
```

Out[191]:

	count	mean	std	min	25%	50%	75%	max
fixed acidity	1599.0	8.319637	1.741096	4.60000	7.1000	7.90000	9.200000	15.90000
volatile acidity	1599.0	0.527821	0.179060	0.12000	0.3900	0.52000	0.640000	1.58000
citric acid	1599.0	0.270976	0.194801	0.00000	0.0900	0.26000	0.420000	1.00000
residual sugar	1599.0	2.538806	1.409928	0.90000	1.9000	2.20000	2.600000	15.50000
chlorides	1599.0	0.087467	0.047065	0.01200	0.0700	0.07900	0.090000	0.61100
free sulfur dioxide	1599.0	15.874922	10.460157	1.00000	7.0000	14.00000	21.000000	72.00000
total sulfur dioxide	1599.0	46.467792	32.895324	6.00000	22.0000	38.00000	62.000000	289.00000
density	1599.0	0.996747	0.001887	0.99007	0.9956	0.99675	0.997835	1.00369
pH	1599.0	3.311113	0.154386	2.74000	3.2100	3.31000	3.400000	4.01000
sulphates	1599.0	0.658149	0.169507	0.33000	0.5500	0.62000	0.730000	2.00000
alcohol	1599.0	10.422983	1.065668	8.40000	9.5000	10.20000	11.100000	14.90000
quality	1599.0	5.636023	0.807569	3.00000	5.0000	6.00000	6.000000	8.00000

Matrica korelacije pokazuje međusobnu zavisnost atributa, kao i zavisnost atributa i klase.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	0.256131	0.671703	0.114777	0.093705	0.153794	0.113181	0.668047	0.682978	0.183006	0.061668	0.124052
volatile acidity	0.256131	1.000000	0.552496	0.001918	0.061298	0.010504	0.076470	0.022026	0.234937	0.260987	0.202288	0.390558
citric acid	0.671703	0.552496	1.000000	0.143577	0.203823	0.060978	0.035533	0.364947	0.541904	0.312770	0.109903	0.226373
residual sugar	0.114777	0.001918	0.143577	1.000000	0.055610	0.187049	0.203028	0.355283	0.085652	0.005527	0.042075	0.013732
chlorides	0.093705	0.061298	0.203823	0.055610	1.000000	0.005562	0.047400	0.200632	0.265026	0.371260	0.221141	0.128907
free sulfur dioxide	0.153794	0.010504	0.060978	0.187049	0.005562	1.000000	0.667666	0.021946	0.070377	0.051658	0.069408	0.050656
total sulfur dioxide	0.113181	0.076470	0.035533	0.203028	0.047400	0.667666	1.000000	0.071269	0.066495	0.042947	0.205654	0.185100
density	0.668047	0.022026	0.364947	0.355283	0.200632	0.021946	0.071269	1.000000	0.341699	0.148506	0.496180	0.174919
pH	0.682978	0.234937	0.541904	0.085652	0.265026	0.070377	0.066495	0.341699	1.000000	0.196648	0.205633	0.057731
sulphates	0.183006	0.260987	0.312770	0.005527	0.371260	0.051658	0.042947	0.148506	0.196648	1.000000	0.093595	0.251397
alcohol	0.061668	0.202288	0.109903	0.042075	0.221141	0.069408	0.205654	0.496180	0.205633	0.093595	1.000000	0.476166
quality	0.124052	0.390558	0.226373	0.013732	0.128907	0.050656	0.185100	0.174919	0.057731	0.251397	0.476166	1.000000

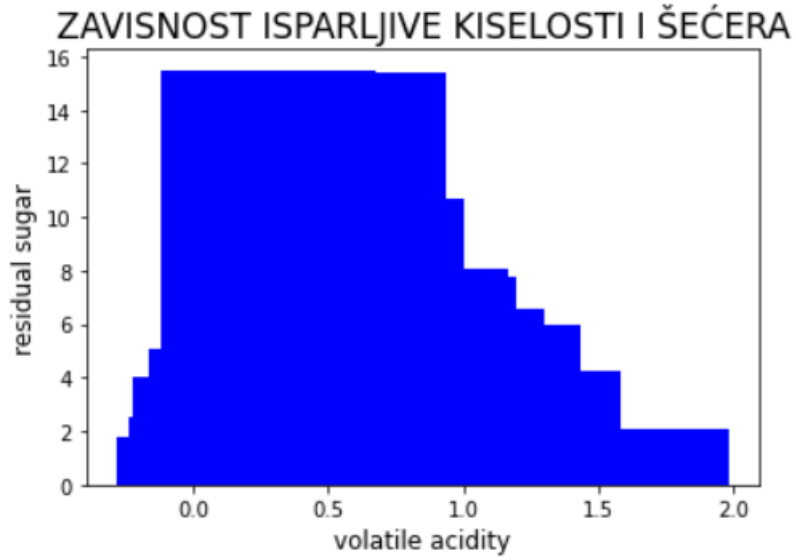
```
In [158]: data.corr().abs().unstack().sort_values().drop_duplicates()

Out[158]: volatile acidity      residual sugar         0.001918
          residual sugar        sulphates              0.005527
          free sulfur dioxide   chlorides              0.005562
          volatile acidity      free sulfur dioxide    0.010504
          quality               residual sugar         0.013732
                                                         ...
          total sulfur dioxide  free sulfur dioxide    0.667666
          density               fixed acidity          0.668047
          fixed acidity         citric acid            0.671703
          pH                    fixed acidity          0.682978
          fixed acidity         fixed acidity          1.000000
          Length: 67, dtype: float64
```

Na prethodnim slikama se može vidjeti da su atributi 'volatile acidity' i 'residual sugar' u najmanjoj korelaciji koja iznosi 0.001918, dok su u najvećoj korelaciji atributi 'pH' i 'fixed acidity' koja je jednaka 0.682978 . Također se vidi da su na glavnoj dijagonali sve jedinice što znači da je svaki atribut u korelaciji sam sa sobom. Najmanju zavisnost sa klasom 'quality' ima atribut 'residual sugar',korelacija jednaka 0.013732, dok najveću zavisnost sa klasom 'quality' ima atribut 'alcohol' ,vrijednost korelacije je 0.476166 što pokazuje da korelacija i nije baš jaka ali je najveća u odnosu korelacije drugih atributa sa klasom 'quality'.

Međusobnu zavisnost navedenih atributa i klase sa atributima pokazuju sljedeće slike.

Kao što je već pomenuto, korelacija između kvaliteta i šećera je jednaka 0.013732 (slaba korelacija).Sa slike se može uočiti da vino kvaliteta 5 i 6 sadrži najviše šećera.



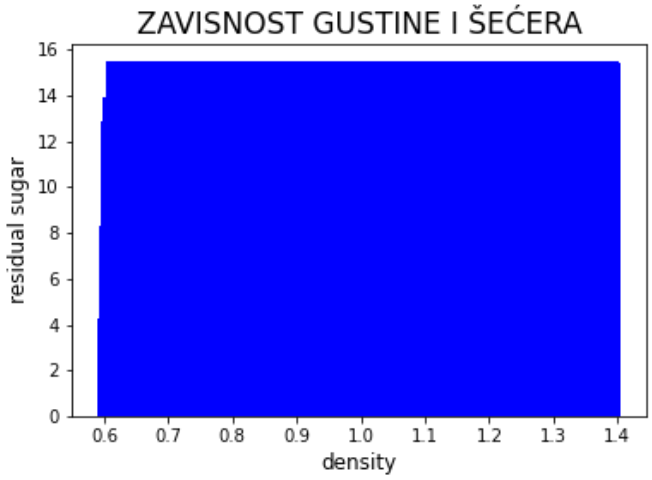
Atributi 'volatile acidity' i 'residual sugar' su u najmanjoj korelaciji.

Najviše alkohola ima vino kvaliteta 5.

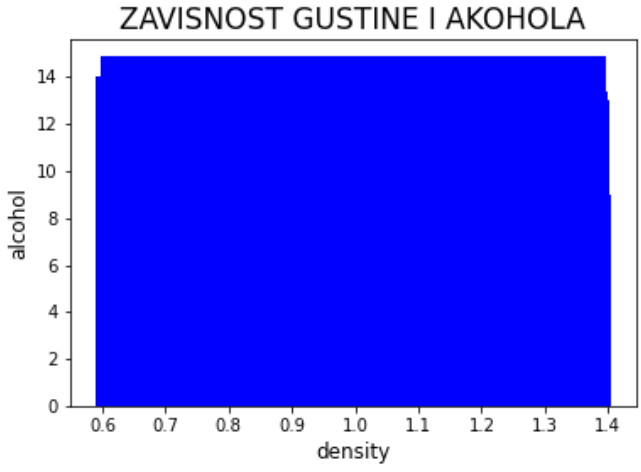
Pokazano je u matrici korelacije da od svih atributa najveću zavisnost sa klasom 'quality' ima atribut 'alcohol'.

Prethodna slika pokazuje dva zavisna atrubuta 'pH' i 'fixed acidity' čija korelacija iznosi 0.682978 (pH opisuje koliko je vino kiselo ili bazno na skali od 0 (veoma kiselo) do 14 (veoma bazno)).

Sljedeće slike pokazuju zavisnost gustine sa šećerom i sa alkoholom.



(korelacija=0.35)



(korelacija=0.49)

Klasa 'quality'

Klasa 'quality' pokazuje kvalitet vina koji je označen brojevima 3,4,5,6,7,8. Najviše ima instanci čiji je kvalitet 5 a najmanje ima instanci koje predstavljaju vino kvaliteta 3.

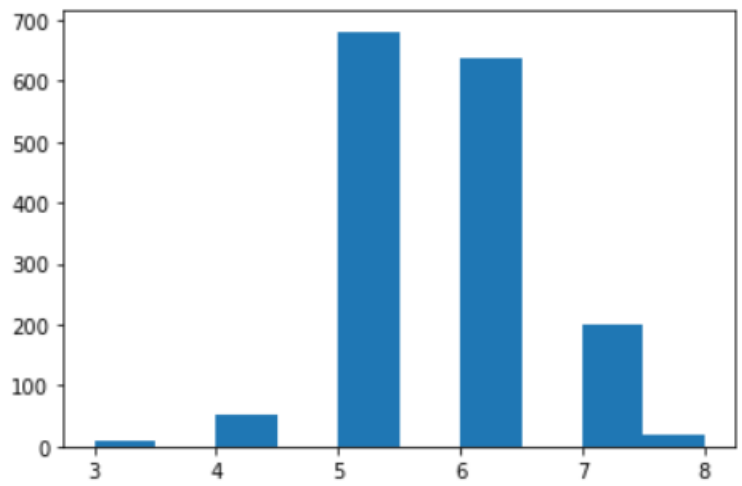
```
In [159]: data['quality'].value_counts()
```

```
Out[159]: 5    681
          6    638
          7    199
          4     53
          8     18
          3     10
          Name: quality, dtype: int64
```

```
In [160]: procenat=data['quality'].value_counts(normalize=True)*100
          procenat_=round(procenat,2)
          procenat_
```

```
Out[160]: 5    42.59
          6    39.90
          7    12.45
          4     3.31
          8     1.13
          3     0.63
          Name: quality, dtype: float64
```

Grafički se to može prikazati na sljedeći način, gdje je očigledno da ima najviše instaci koje predstavljaju vino kvaliteta 5 i 6.



MODELI KLASIFIKACIJE

Kreirani su modeli klasifikacije pomoću:

1. stabla odlučivanja
2. KNN algoritama
3. Naive Bayes algoritama

Podaci su podijeljeni na trening i test skup u odnosu 70%:30% što se može vidjeti na sljedećoj slici.

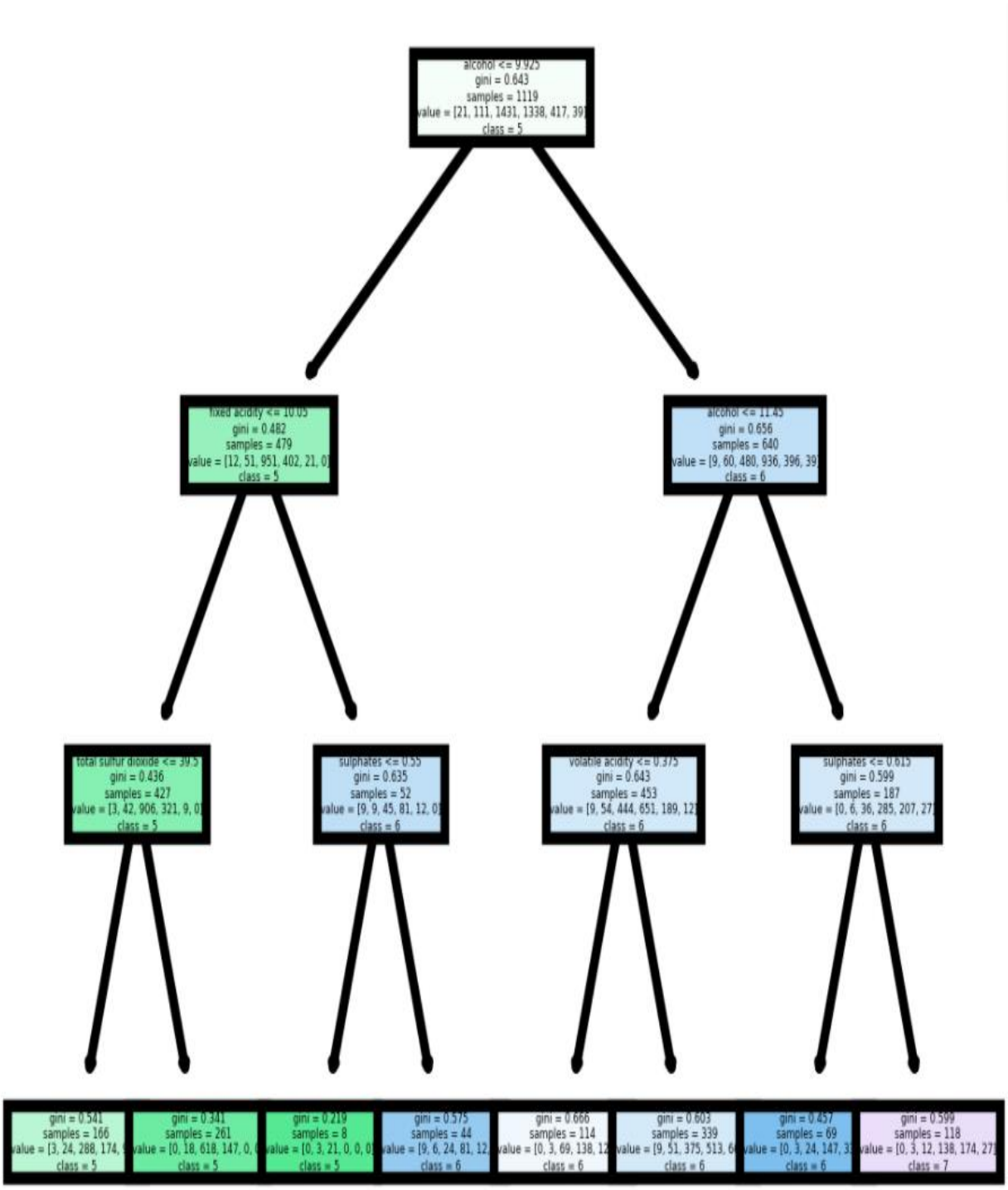
```
X_train,X_test,y_train,y_test=train_test_split(data[atributi],data[klasa], test_size=0.30, stratify=data[klasa], random_state=54)
```

Koristeći testni skup podataka izvršena je ocjena kvaliteta istreniranih modela pri čemu je korištena tačnost kao mjera. Prikazane su matrice korelacije za istrenirane modele, pri čemu je kod njih cilj da što više elemenata bude na glavnoj dijagonali, jer upravo oni pokazuju tačno predviđene elemente.

U narednim poglavljima će biti prikazani rezultati navedenih modela klasifikacije kao i njihova tačnost.

STABLO ODLUČIVANJA

Na slici se može vidjeti stablo odlučivanja za ovaj skup podataka. Ovaj algoritam kao mjeru sličnosti podataka koristi Ginijev indeks, dok je maksimalna dubina stabla jednaka 3.



Korijen stabla pokazuje da je najbolje izvršiti podjelu po atributu 'alcohol' i tu se vidi da je vrijednost Ginijevog indeksa jednaka 0.643. Ako je vrijednost atributa 'alcohol' manja ili jednaka 9.925 onda se posmatra lijevi dio stabla gdje se uočava da je u tom slučaju atribut 'fixed acidity' pogodan za dalju podjelu.Zatim se podjela vrši u odnosu na vrijednost Ginijevog indeksa od 'fixed acidity'. Ukoliko je vrijednost atributa 'alcohol' veća od 9.925 onda se posmatra desni dio stabla.

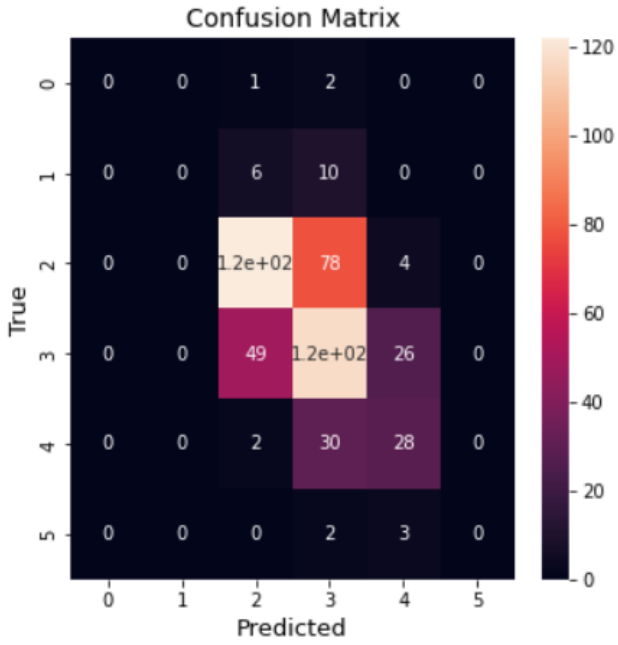
Koristeći testni skup podataka izvršena je ocjena kvaliteta i tačnost iznosi 0.55625.

```
predS=clf.predict(X_test)
accuracyS=accuracy_score(Y_test,predS)
accuracyS
```

```
0.55625
```

```
confusion_matrix(Y_test,predS)
```

```
array([[  0,   0,   1,   2,   0,   0],
       [  0,   0,   6,  10,   0,   0],
       [  0,   0, 122,  78,   4,   0],
       [  0,   0,  49, 117,  26,   0],
       [  0,   0,   2,  30,  28,   0],
       [  0,   0,   0,   2,   3,   0]], dtype=int64)
```



Prethodne slike pokazuju matricu konfuzije gdje se lako može uočiti da elementi na glavnoj dijagonali predstavljaju tačno predviđene podatke. Da je to tako može se matematički dokazati tako što se pokaže da kada se suma elemenata matrice pomnoži sa brojnom vrijednosti tačnosti dobije se suma elemenata na glavnoj dijagonali.

Suma elemenata matrice=480

Tačnost=0.55625

Suma elemenata na glavnoj dijagonali=267

480*0.55625=267

Analogno, na osnovu matrice konfuzije se može odrediti tačnost ukoliko nije poznata: tačnost=267/480=0.55625.

KNN algoritam

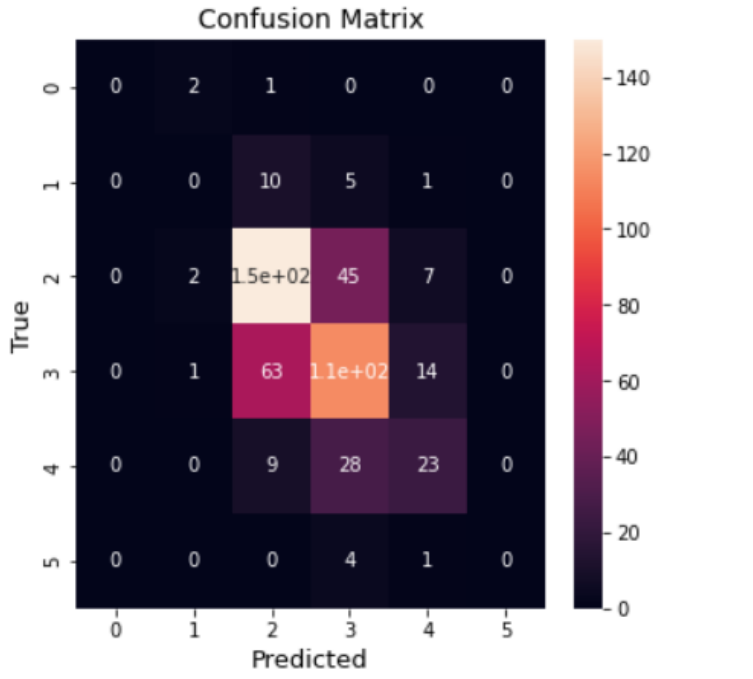
U KNN algoritmu je korištena euklidska metrika, dok je k=8 ,odnosno broj najbližih susjeda je 8. Tokom puštanja koda u Jupyteru prvobitno je bilo pokušano i za k<8 ali za k=8 se pokazala veća tačnost,ali ne i znatno veća jer je tačnost za k=8 jednaka 0.5979166666666667. Za ostale vrijednosti k je vrijedilo sljedeće:
k=3, tačnost=0.5604166666666667
k=5, tačnost=0.5520833333333334
k=7, tačnost=0.5770833333333333
k=9, tačnost=0.58125.

```
predKNN=model.predict(X_test1)
accuracyKNN=accuracy_score(Y_test1,predKNN)
accuracyKNN
```

```
0.5979166666666667
```

```
confusion_matrix(Y_test1,predKNN)
```

```
array([[  0,   2,   1,   0,   0,   0],
       [  0,   0,  10,   5,   1,   0],
       [  0,   2, 150,  45,   7,   0],
       [  0,   1,  63, 114,  14,   0],
       [  0,   0,   9,  28,  23,   0],
       [  0,   0,   0,   4,   1,   0]], dtype=int64)
```



Matrica konfuzije pokazuje tačnost, odnosno elementi glavne dijagonale su elementi koji su tačno predviđeni.

NAIVE BAYES

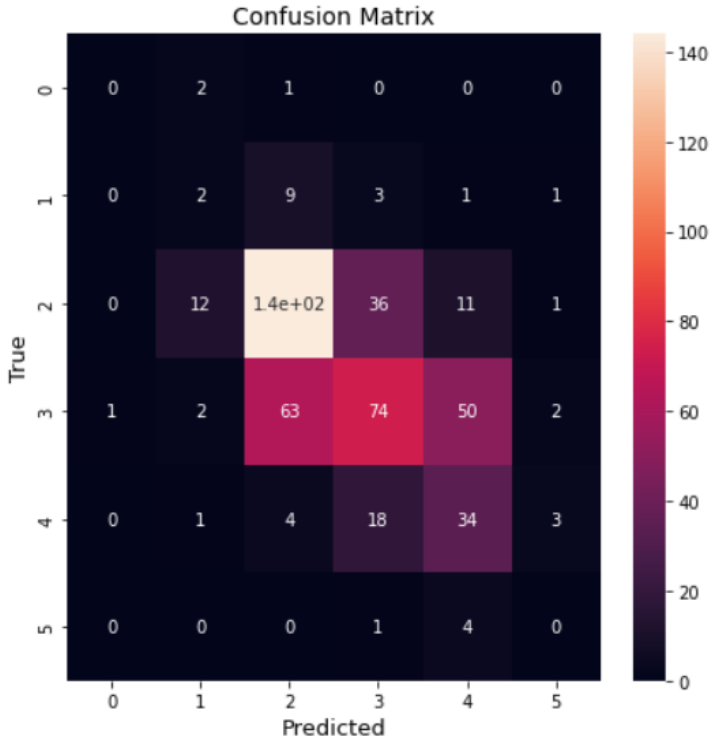
Model kreiran pomoću Naive Bayesovog algoritma daje tačnost 0.5291666666666667. Sljedeće slike to pokazuju, kao i matricu konfuzije.

```
predNB=model1.predict(X_test)
accuracyNB=accuracy_score(Y_test,predNB)
accuracyNB
```

```
0.5291666666666667
```

```
confusion_matrix(Y_test,predNB)
```

```
array([[  0,   2,   1,   0,   0,   0],
       [  0,   2,   9,   3,   1,   1],
       [  0,  12, 144,  36,  11,   1],
       [  1,   2,  63,  74,  50,   2],
       [  0,   1,   4,  18,  34,   3],
       [  0,   0,   0,   1,   4,   0]], dtype=int64)
```



Broj tačno predviđenih podataka se očitava na glavnoj dijagonali matrice konfuzije. Matrica konfuzije omogućava lakše uočavanje tačnosti dok slika prije matrice pokazuje brojnu vrijednost tačnosti. Tačnost se može dobiti iz matrice konfuzije na način kao što je to objašnjeno kod stabla odlučivanja.

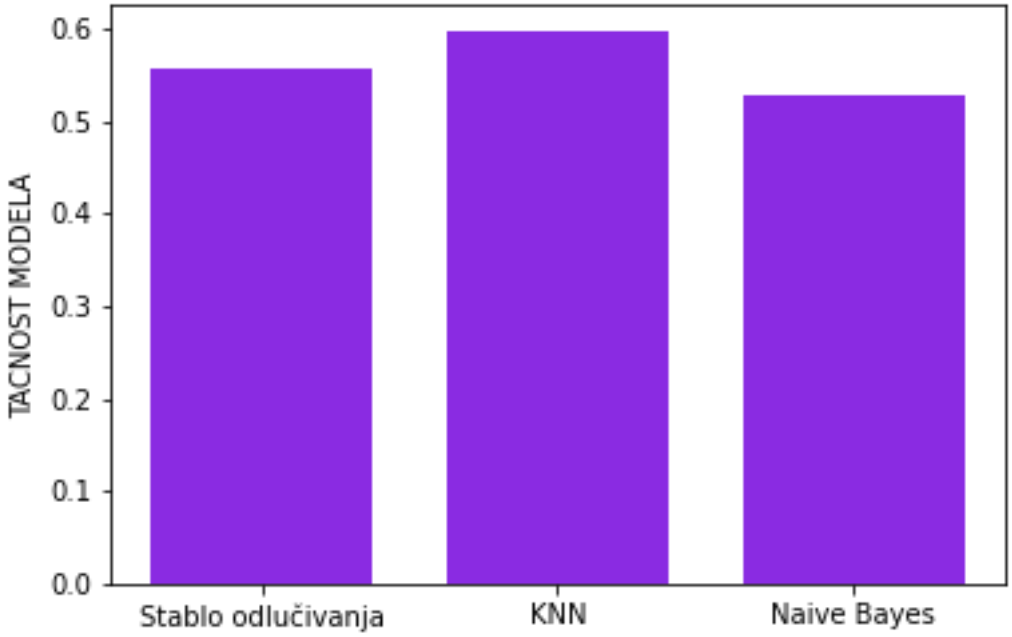
ZAKLJUČAK

Tačnost za stablo odlučivanja: 0.55625

Tačnost za KNN algoritam: 0.5979166666666667

Tačnost za Naive Bayes algoritam: 0.5291666666666667

Uočava se da je najveća tačnost dobijena modelom klasifikacije pomoću KNN algoritma . Grafički to pokazuje naredna slika.



Kod KNN algoritma je upoređena tačnost u odnosu na broj k, gdje je pokazano da je tačnost najveća bila za k=8, pa se iz tog razloga i posmatrao KNN za k=8. Kod stabla odlučivanja se moglo jasno grafički uočiti po kojem atributu je najbolje izvršiti podjelu, u odnosu na Ginijev indeks(to je bio atribut 'alcohol' kao što je već objašnjeno).