Projekt

Dawid Sikorski 291951

https://colab.research.google.com/drive/1nDziM5REQ1sOFnou5sCWh2qdXBsayJlu?usp=sharing

Celem zadania, było przewidzenie jakości wina szacując wartość biorąc pod uwagę model regresji oraz szacowania. Ostatnią częścią zadania było pogrupowanie. Do zrealizowania zadania wykorzystano dane zebrane w pliku ‘winequality-red.csv`, w którym znajduje się 1599 obserwacji z jedenastoma ilościowymi oraz zmienną celu.

```
wina.isna().sum()
```

```
Stała kwasowość              0
Lotna kwasowość              0
Kwas cytrynowy               0
Cukier resztkowy             0
Chlorki                      0
Wolny dwutlenek siarki       0
Całkowity dwutlenek siarki   0
Gęstość                      0
pH                           0
Siarczany                    0
Alkohol                      0
Jakość                       0
dtype: int64
```

Braki danych z pliku

	Jakość	Ilość
0	3	10
1	4	53
2	5	577
3	6	535
4	7	167
5	8	17

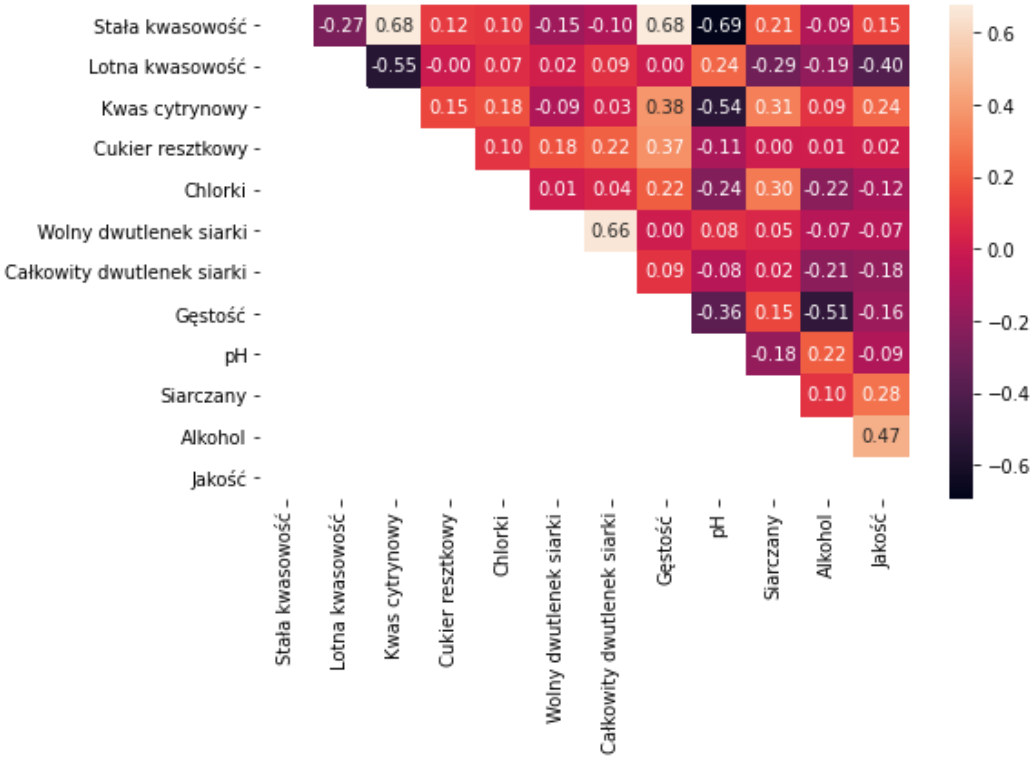
Ilość win w danej jakości w całym zbiorze danych

	Jakość	Ilość
0	3	7
1	4	37
2	5	477
3	6	446
4	7	139
5	8	13

Ilość win w danej jakości w zbiorze treningowym

Plik nie zawierał braków danych. Wszystkie kolumny były zmiennymi ilościowymi, tak więc nie wymagały żadnych operacji początkowych. W celach estetycznych nazwy kolumn zostały zamienione na polskie nazwy. Warto zauważyć, iż wina oceniane były w skali od 0 do 10, a wszystkie zebrane obserwacje mieszczą się w przedziale od 3 do 8, przy czym posiadamy bardzo mały odsetek win należących do klasy jakości 3, 4 oraz 8. Może to wpłynąć na proces uczenia, tak że do grupy 5 i 6, gdzie jest najwięcej obserwacji będą najczęściej klasyfikowane nowe obserwacje.

Następnie zbiór danych został podzielony na treningowy oraz testowy w stosunku 70:30, przy pomocy numeru indeksu jako ziarno generatora liczb losowych.



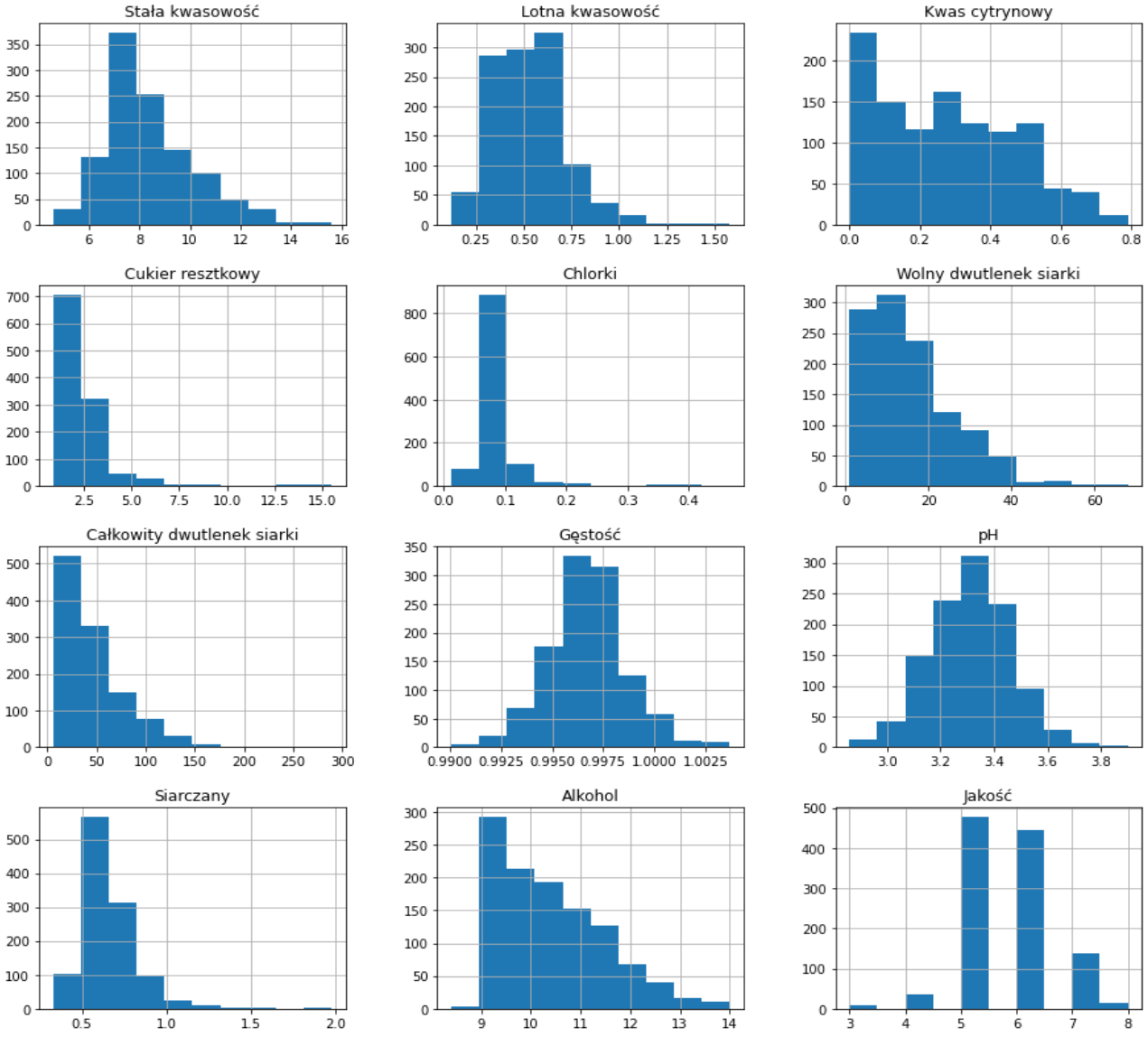
Macierz korelacji dla zbioru treningowego

Z macierzy korelacji możemy zaobserwować, iż Stała kwasowość jest skorelowana z Kwasem cytrynowym, Gęstością oraz pH. Lotna kwasowość jest skorelowana z Kwasem cytrynowym, który z kolei jest skorelowany z pH. Wolny dwutlenek siarki jest skorelowany z Całkowitym dwutlenkiem siarki. Ostatnią korelacją obserwujemy pomiędzy Gęstością a zawartością Alkoholu. Patrząc na korelacje związane ze zmienną celu „Jakość" możemy zaobserwować, iż nie widać korelacji z Cukrem resztkowym. Z kolei im większa zawartość Alkoholu, Siarczanów, Kwasu cytrynowego oraz Stała kwasowość tym jakość wina jest lepsza. Pozostałe zmienne powodują pogorszenie jakości, przy czym Lotna kwasowość najbardziej. W związku z tymi spostrzeżeniami postanowiłem dogłębniej przeanalizować dane w każdej z kolumn, tak aby upewnić się, które zmienne można wyrzucić ze zbioru tak aby modele działały poprawnie oraz były prostsze.

	Stała kwasowość	Lotna kwasowość	Kwas cytrynowy	Cukier resztkowy	Chlorki	Wolny dwutlenek siarki	Całkowity dwutlenek siarki	Gęstość	pH	Siarczany	Alkohol	Jakość
Ilość	1119.000	1119.000	1119.000	1119.000	1119.000	1119.000	1119.000	1119.000	1119.000	1119.000	1119.000	1119.000
Średnia	8.348	0.527	0.273	2.528	0.087	15.794	45.508	0.997	3.311	0.657	10.429	5.636
Odchylenie standardowe	1.747	0.182	0.197	1.441	0.043	10.357	32.658	0.002	0.150	0.163	1.051	0.809
Minimum	4.600	0.120	0.000	0.900	0.012	1.000	6.000	0.990	2.860	0.330	8.400	3.000
25%	7.100	0.390	0.090	1.900	0.070	7.000	22.000	0.996	3.210	0.550	9.500	5.000
50%	7.900	0.520	0.260	2.200	0.079	13.000	37.000	0.997	3.310	0.620	10.200	6.000
75%	9.300	0.633	0.420	2.600	0.091	21.000	60.000	0.998	3.400	0.730	11.100	6.000
Maksimum	15.600	1.580	0.790	15.500	0.467	68.000	289.000	1.004	3.900	1.980	14.000	8.000

Analiza danych treningowych

W dalszym procesie, przeszedłem do zebrania podstawowych informacji dla każdej z kolumn. W powyższej tabeli zostały zebrane informacje o ilości, średniej, odchyleniu standardowym, minimum i maksimum dla każdej zmiennej. Dodatkowo wyeksportowałem histogramy dla każdego predyktora.



Histogramy danych treningowych

Jakość	3.000000	4.000000	5.000000	6.000000	7.000000	8.000000
Stała kwasowość	8.385714	7.632432	8.178407	8.372646	9.022302	8.523077
Lotna kwasowość	0.882143	0.721081	0.577453	0.491760	0.401115	0.454615
Kwas cytrynowy	0.167143	0.154324	0.242956	0.279170	0.382086	0.383077
Cukier resztkowy	2.042857	2.602703	2.531971	2.469843	2.715108	2.438462
Chlorki	0.120857	0.082973	0.092055	0.084558	0.078137	0.067077
Wolny dwutlenek siarki	13.857143	12.027027	16.979036	15.738789	13.395683	11.615385
Całkowity dwutlenek siarki	29.285714	37.864865	54.861635	40.233184	34.381295	32.692308
Gęstość	0.997223	0.996429	0.997103	0.996610	0.996156	0.995278
pH	3.398571	3.405676	3.307757	3.316704	3.279065	3.268462
Siarczany	0.572857	0.556216	0.621551	0.674283	0.738489	0.790769
Alkohol	10.035714	10.316216	9.920335	10.624066	11.443285	12.069231

Średnia zawartość w danych treningowych pogrupowanych względem jakości

Przed rozpoczęciem budowania modeli postanowiłem pogrupować wina ze zbioru treningowego oraz wyliczyć średnią zawartość każdego predyktora w danej grupie. Można tu zaobserwować, iż wina należące do jakości 3 mają największą średnią zawartość Chlorków oraz najniższą zawartość Siarczanów i Kwasu cytrynowego. Wina z jakością 7 i 8 posiadają średnio najwięcej Kwasu cytrynowego, Siarczanów i Alkoholu oraz posiadają najmniejszą średnią ilość Chlorków. Wina z największą zawartością Siarki znajdują się w 5 i 6 grupie.

Z wiedzy ogólnej wiem, iż największy wpływ w podziale wina na słodkie, półsłodkie, półwytrawne oraz wyprawne ma wpływ zawartość alkoholu, pH, cukier oraz kwasowość lotna.

Patrząc więc na wszystkie powyższe obserwacje do przeprowadzania klasyfikacji i regresji postanowiłem wybrać następujące **predyktory**:

Lotna kwasowość, Cukier resztkowy, Chlorki, Całkowity dwutlenek siarki, Siarczany oraz Alkohol.

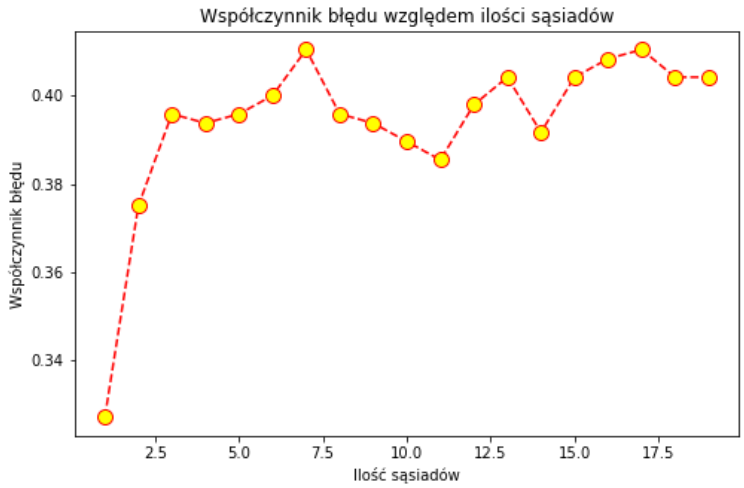
W celu przewidzenia jakości wina, używając modelu szacowania postanowiłem użyć algorytmu „k najbliższych sąsiadów” oraz dla modelu regresji „sieci neuronowej”. Grupowanie postanowiłem wykonać za pomocą sieci Kohonena „SOM”.

KNN

Parametry:

- N_neighbors = 11
- Metric = euclidean
- Weight = ‘uniform’

Zgodnie z założeniami w pierwszym kroku wszystkie wartości ilościowe zostały zestandaryzowane. Na początku postanowiłem wywołać algorytm dla zmiennej ilości sąsiadów od 1 do 20, tak aby zobaczyć jaka ilość mogłaby najlepiej pomóc w problemie klasyfikacji.

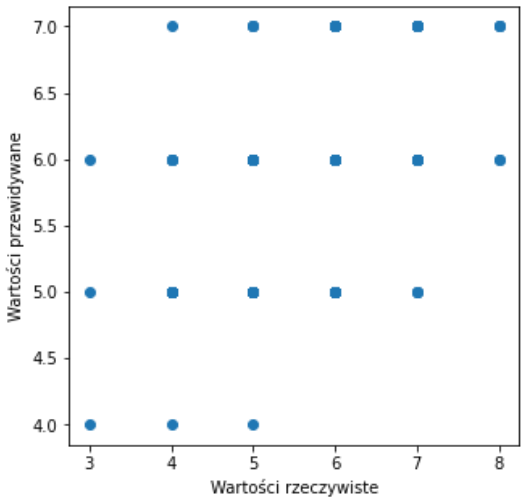


Średni błąd klasyfikacji względem ilości sąsiadów dla danych testowych

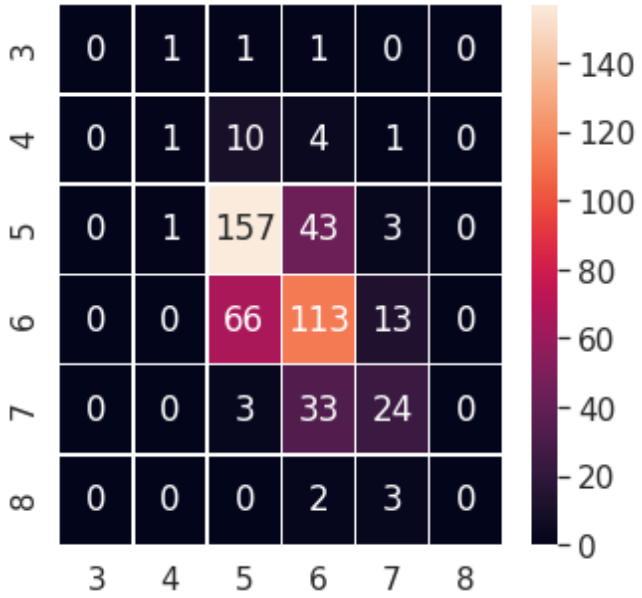
```
MSE dla uczacej                0.4718498659517426
MAE dla uczacej                0.38427167113494193
RMSE dla uczacej               0.6869132885246453
Trafność dla uczącej:          0.6577301161751564
Trafność dla uczącej[+/-1]:    0.9597855227882037

MSE dla testowej               0.5
MAE dla testowej               0.42083333333333334
RMSE dla testowej              0.7071067811865476
Trafność dla testowej:         0.6145833333333334
Trafność dla testowej[+/-1]:   0.96875
```

Jakość modelu - KNN



Zestawienie wartości rzeczywistych z przewidywanymi – KNN



Macierz pomyłek – KNN

Na podstawie analizy wyników otrzymanych podczas liczenia błędów ‘MAE’, ‘MSE’ oraz ‘RMSE’ dla próby treningowej i testowej możemy stwierdzić iż model nie uległ przeuczeniu, a na zestawieniu otrzymano „chmurkę”, tak więc możemy stwierdzić iż model przewiduje poprawnie. Patrząc z kolei na zwykłą trafność dla danych testowych otrzymaliśmy wynik równy około 61%, co nie jest najgorszą jakością, lecz stosując trafność z odstępstwem o 1 otrzymaliśmy prawie 97% skuteczność co jest już bardzo dobrym wynikiem. Warto zauważyć, że model nie klasyfikuje poprawnie win o jakości 3, 4 oraz 8. Może być to spowodowane wcześniejszymi obserwacjami, iż tych win mamy najmniej w zbiorze uczącym.

Sieć neuronowa

Parametry:

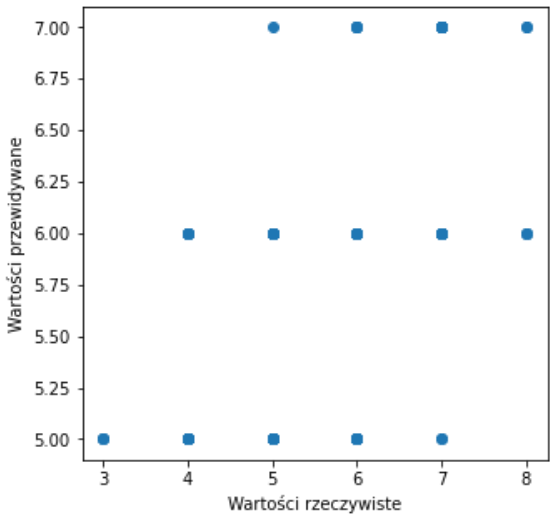
- Hidden_layer_size = (10,)
- Activation = tanh
- Solver = lbfgs
- Alpha = 0.0001
- Max_iter = 10000
- Random_state = 291951

Zgodnie z założeniami wszystkie predyktory oraz zmienna celu zostały znormalizowane. Metodą prób i błędów, określenie większej ilości warstw ukrytych bądź też zwiększanie ilości neuronów nie wpływało na wynik, lub powodowało przeuczenie się sieci.

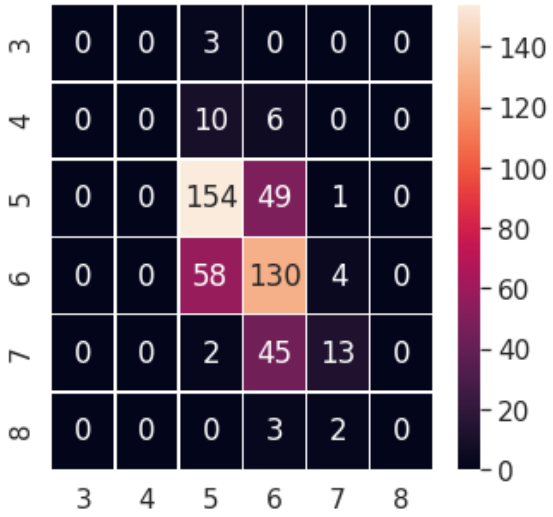
```
MSE dla uczacej                0.4691689008042895
MAE dla uczacej                0.4155495978552279
RMSE dla uczacej               0.6849590504579741
Trafność dla uczącej:          0.610366398570152
Trafność dla uczącej[+/-1]:    0.9749776586237712

MSE dla testowej               0.475
MAE dla testowej               0.4125
RMSE dla testowej              0.689202437604511
Trafność dla testowej:         0.61875
Trafność dla testowej[+/-1]:   0.96875
```

Jakość modelu – sieć neuronowa



Zestawienie wartości rzeczywistych z przewidywanymi – sieć neuronowa



Macierz pomyłek – sieć neuronowa

Na podstawie analizy wyników otrzymanych podczas liczenia błędów 'MAE', 'MSE' oraz 'RMSE' dla próby treningowej i testowej możemy stwierdzić iż model nie uległ przeuczeniu, a na zestawieniu otrzymano „chmurkę", tak więc możemy stwierdzić iż model przewiduje poprawnie. Patrząc z kolei na zwykłą trafność dla danych testowych otrzymaliśmy wynik równy około 62%, co nie jest najgorszą jakością, lecz stosując trafność z odstępstwem o 1 otrzymaliśmy prawie 97% skuteczność co jest już bardzo dobrym wynikiem. Warto zauważyć, że model nie klasyfikuje poprawnie win o jakości 3, 4 oraz 8. Może być to spowodowane wcześniejszymi obserwacjami, iż tych win mamy najmniej w zbiorze uczącym.

Wnioski

MSE dla uczacej	0.4718498659517426
MAE dla uczacej	0.38427167113494193
RMSE dla uczacej	0.6869132885246453
Trafność dla uczącej:	0.6577301161751564
Trafność dla uczącej[+/-1]:	0.9597855227882037
MSE dla testowej	0.5
MAE dla testowej	0.42083333333333334
RMSE dla testowej	0.7071067811865476
Trafność dla testowej:	0.6145833333333334
Trafność dla testowej[+/-1]:	0.96875

MSE dla uczacej	0.4691689008042895
MAE dla uczacej	0.4155495978552279
RMSE dla uczacej	0.6849590504579741
Trafność dla uczącej:	0.610366398570152
Trafność dla uczącej[+/-1]:	0.9749776586237712
MSE dla testowej	0.475
MAE dla testowej	0.4125
RMSE dla testowej	0.689202437604511
Trafność dla testowej:	0.61875
Trafność dla testowej[+/-1]:	0.96875

Zestawienie jakość modelu – KNN z lewej oraz Sieć neuronowa z prawej

Obydwa modele uzyskały rewelacyjne wyniki. Otrzymaliśmy praktycznie identyczną trafność łącznie z odstępstwem o jeden dla obydwu modeli. Porównując dokładniej otrzymane rezultaty możemy zaobserwować drobną przewagę głównych miar jakości sieci neuronowej dla próby testowej. Może to byś spowodowane wyliczaniem wartości na podstawie danych wejściowych, w przeciwieństwie do KNN, który bada w tym przypadku 11 najbliższych sąsiadów i na ich podstawie wylicza końcową jakość.

SOM

Parametry:

- M = 4
- N = 1
- Dim = 11

Zgodnie z założeniami wszystkie predyktory bez zmiennej celu zostały znormalizowane. Dzieląc na więcej niż 4 grupy zaobserwowałem, iż każda kolejna posiada mało obserwacji skupionych wokół siebie lub była bardzo zbliżone do jednej z pozostałych.

Klaster	Ilość
0	407
1	203
2	250
3	259

Ilość obserwacji w grupie dla zbioru treningowego

Jak widać, zbiory są mniej więcej równoliczne, tak więc możemy założyć iż podział jest dobry. Następnie postanowiłem pogrupować obserwacje dla zbioru treningowego, oraz wyliczyć średnie wartości, tak aby móc określić czym wyróżnia się dana grupa.

Klaster	0.000000	1.000000	2.000000	3.000000
Stała kwasowość	7.379607	6.933005	8.786400	10.555212
Lotna kwasowość	0.665061	0.482759	0.403700	0.462104
Kwas cytrynowy	0.122531	0.181281	0.377400	0.480116
Cukier resztkowy	2.395086	2.134975	2.325400	3.241120
Chlorki	0.083966	0.070670	0.079636	0.111363
Wolny dwutlenek siarki	18.573710	14.985222	11.792000	15.922780
Całkowity dwutlenek siarki	53.592138	36.246305	34.032000	51.138996
Gęstość	0.996569	0.994712	0.996521	0.998838
pH	3.389042	3.393498	3.263040	3.170463
Siarczany	0.584619	0.642069	0.699440	0.739730
Alkohol	10.009582	11.234811	10.722000	10.172716
Jakość	5.294840	5.926108	5.904000	5.687259

Średnie wartości obserwacji w grupie dla zbioru treningowego

Możemy zauważyć iż główną wartością, dla której dzielą się grupy jest średnia zawartość **stałej kwasowości**, **lotnej kwasowości**, **wolnego dwutlenku siarki** oraz **całkowitego dwutlenku siarki**. Korzystając z wiedzy znajdującej się w Internecie postanowiłem podzielić klastry w następujący sposób:

Klaster 0 są to wina **słodkie**. Posiadają najwięcej lotnej kwasowości, wolnego dwutlenku siarki, całkowitego dwutlenku siarki, oraz najmniejszą ilość siarczanów, alkoholu i kwasu cytrynowego. Są to wina stosunkowo słabsze jakościowo, ich średnia wynosi około 5.3.

Klaster 1 są to wina **półwytrawne**. Zawierają najmniejszą ilość stałej kwasowości, chlorków i całkowitego dwutlenku siarki oraz największą ilość alkoholu. Są to wina stosunkowo wyższej jakości, ich średnia wynosi około 5.9.

Klaster 2 są to wina **wytrawne**. Zawierają najwięcej lotnej kwasowości i kwasu cytrynowego oraz najmniejszą zawartość wolnego dwutlenku siarki, chlorków i całkowitego dwutlenku siarki. Są to wina stosunkowo wyższej jakości, ich średnia wynosi około 5.9.

Klaster 3 są to wina **półsłodkie**. Zawierają najmniejszą ilość alkoholu oraz największą ilość stałej kwasowości, kwasu cytrynowego, cukru resztkowego, chlorków, całkowitego dwutlenku siarki i siarczanów. Są to wina stosunkowo średniej jakości, ich średnia wynosi 5.7.

Szukając informacji dlaczego kwasowość oraz zawartość siarki może wpływać na taki podział, dowiedziałem się, iż winogrona podaje się procesowi siarkowania, tak aby uchronić je przed utlenianiem i jednocześnie zabić bakterie, które mogłyby popsuć wino. Proces ten znacząco wpływa na przebieg fermentacji, co z kolei idzie na smaku czyli też jakości wina. Idąc tym krokiem, możemy również zaobserwować, iż wina posiadające średnio wyższą kwasowość oraz najmniejszą ilość siarki należą do win o lepszej jakości.

Klaster	0.000000	1.000000	2.000000	3.000000
Stała kwasowość	7.365574	7.013265	8.698913	10.526168
Lotna kwasowość	0.660710	0.473214	0.392228	0.479673
Kwas cytrynowy	0.120219	0.197041	0.388370	0.475514
Cukier resztkowy	2.599180	2.140306	2.406522	3.025701
Chlorki	0.082760	0.072204	0.080457	0.121159
Wolny dwutlenek siarki	18.535519	15.719388	13.663043	14.214953
Całkowity dwutlenek siarki	57.273224	42.224490	39.076087	48.271028
Gęstość	0.996739	0.995006	0.996325	0.998721
pH	3.392186	3.381531	3.257609	3.154019
Siarczany	0.590437	0.644286	0.707283	0.760654
Alkohol	9.921949	11.053061	10.929348	10.206854
Jakość	5.333333	5.877551	5.978261	5.635514

Średnie wartości obserwacji w grupie dla zbioru testowego

Ostatnim krokiem będzie upewnienie się w tych obserwacjach, badając zbiór testowy na tej samej sieci Kohonena. Patrząc na otrzymane średnie wartości grupując obserwacje względem klastrów, możemy zaobserwować, iż opisane powyżej zależności utożsamiają się z wynikami otrzymanymi dla zbiory testowego. Tak więc, im mniejsza średnia zawartość dwutlenku siarki tym wino jest lepszej jakości. Patrząc na poniższy rysunek, możemy zaobserwować, iż średnie wartości nie uległy większym zmianom.

Klaster	0	0	1	1	2	2	3	3
Stała kwasowość	7.37961	7.36557	6.933	7.01327	8.7864	8.69891	10.5552	10.5262
Lotna kwasowość	0.665061	0.66071	0.482759	0.473214	0.4037	0.392228	0.462104	0.479673
Kwas cytrynowy	0.122531	0.120219	0.181281	0.197041	0.3774	0.38837	0.480116	0.475514
Cukier resztkowy	2.39509	2.59918	2.13498	2.14031	2.3254	2.40652	3.24112	3.0257
Chlorki	0.0839656	0.0827596	0.07067	0.0722041	0.079636	0.0804565	0.111363	0.121159
Wolny dwutlenek siarki	18.5737	18.5355	14.9852	15.7194	11.792	13.663	15.9228	14.215
Całkowity dwutlenek siarki	53.5921	57.2732	36.2463	42.2245	34.032	39.0761	51.139	48.271
Gęstość	0.996569	0.996739	0.994712	0.995006	0.996521	0.996325	0.998838	0.998721
pH	3.38904	3.39219	3.3935	3.38153	3.26304	3.25761	3.17046	3.15402
Siarczany	0.584619	0.590437	0.642069	0.644286	0.69944	0.707283	0.73973	0.760654
Alkohol	10.0096	9.92195	11.2348	11.0531	10.722	10.9293	10.1727	10.2069
Jakość	5.29484	5.33333	5.92611	5.87755	5.904	5.97826	5.68726	5.63551
Zbiór	Treningowy	Testowy	Treningowy	Testowy	Treningowy	Testowy	Treningowy	Testowy
Ilość	407	183	203	98	250	92	259	107

Porównanie średnich wartości obserwacji w grupach dla danych treningowych i testowych