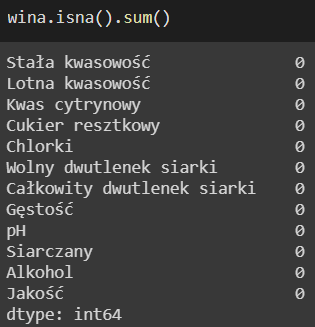
Projekt

Dawid Sikorski 291951

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

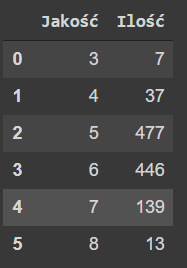
Celem zadania, było przewidzenie jakości wina szacując wartość biorąc pod uwagę model regresji oraz szacowania. Ostatnią częścią zadania było pogrupowanie. Do zrealizowania zadania wykorzystano dane zebrane w pliku ‘winequality-red.csv`, w którym znajduje się 1599 obserwacji z jedenastoma ilościowymi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



*Braki danych z pliku*



*Ilość win w danej jakości w całym zbiorze danych*



*Ilość win w danej jakości w zbiorze treningowym*

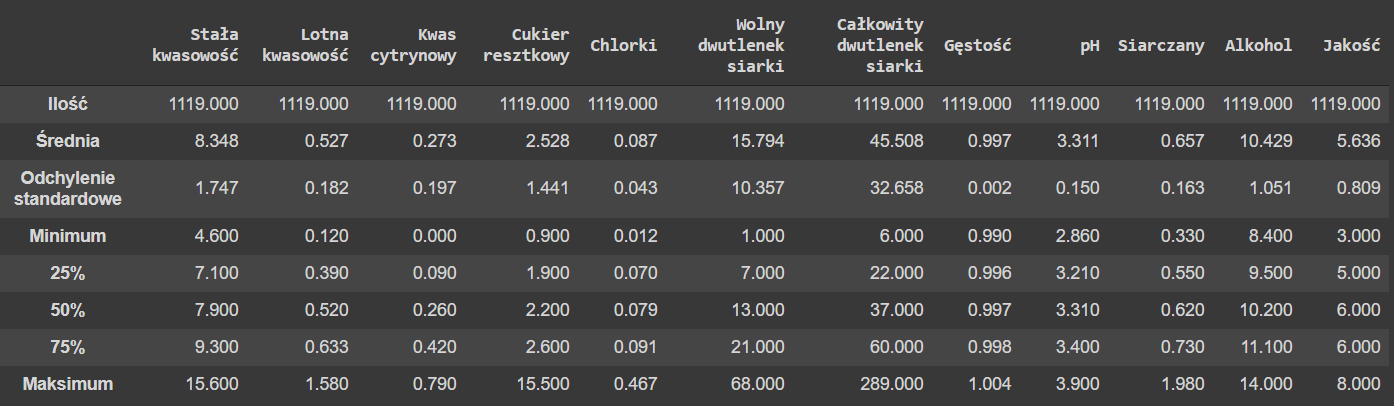
Plik nie zawierał braków danych. Wszystkie kolumny były zmiennymi ilościowymi, tak więc nie wymagały żadnych operacji początkowych. W celach estetycznych nazwy kolumn zostały zamienione na polskie nazwy. Warto zauważyć, iż wina oceniane były w skali od 0 do 10, a wszystkie zebrane obserwacje mieszczą 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.

**

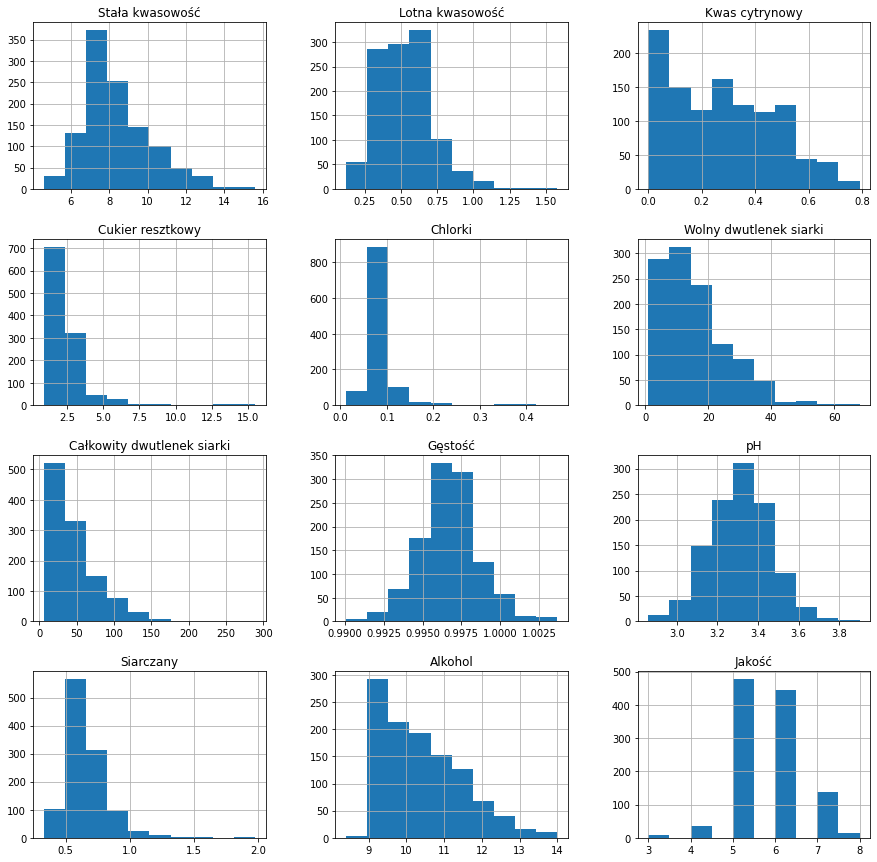
*Macierz korelacji dla zbioru treningowego*

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.

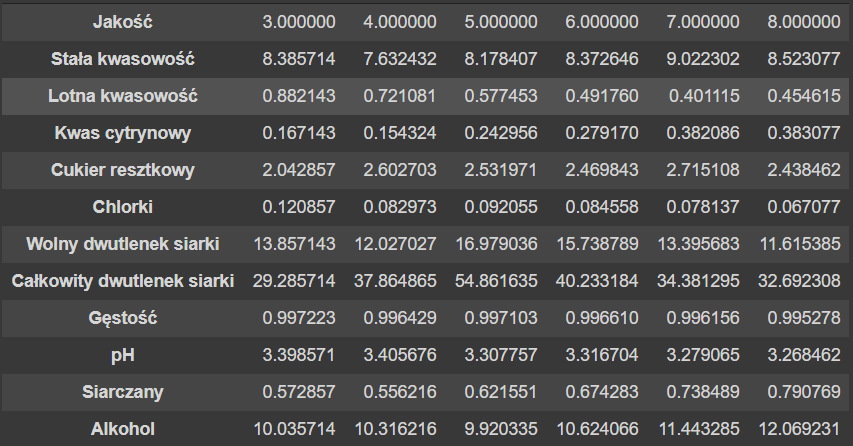
**

*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*

**

*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.**

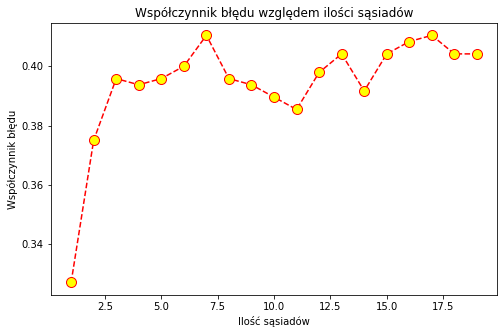
W celu przewidzenia jakości wina, używając modelu szacowania postanowiłem użyć algorytmu „k najbliższych sąsiadów” oraz dla modelu regresji „sieci neuronowej”. Grupowanie postanowiłem wykonać za pomocą sieci Kohonena „SOM”.

**KNN**

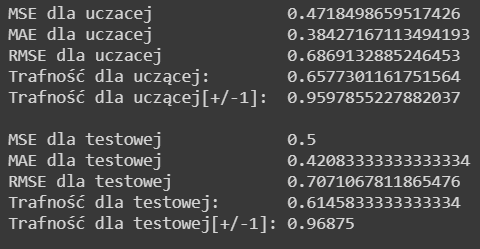
Parametry:

* N\_neighbors = 11
* Metric = euclidean
* Weight = ‘uniform’

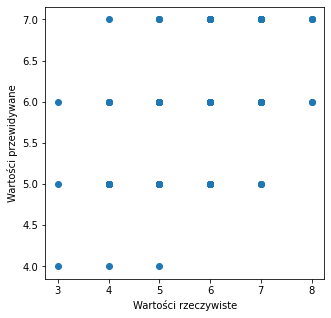
Zgodnie z założeniami w pierwszym kroku wszystkie wartości ilościowe zostały zestandaryzowane. Na początku postanowiłem wywołać algorytm dla zmiennej ilości sąsiadów od 1 do 20, tak aby zobaczyć jaka ilość mogłaby najlepiej pomóc w problemie klasyfikacji.



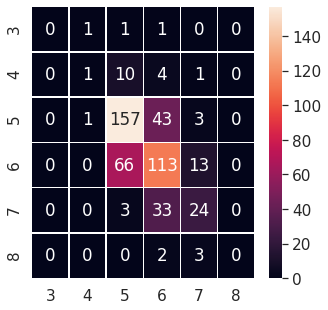
*Średni błąd klasyfikacji względem ilości sąsiadów dla danych testowych*



*Jakość modelu - KNN*



*Zestawienie wartości rzeczywistych z przewidywanymi – KNN*

**

*Macierz pomyłek – KNN*

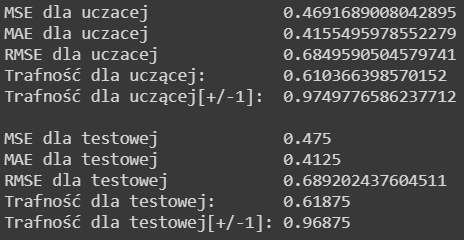
Na podstawie analizy wyników otrzymanych podczas liczenia błędów ‘MAE’, ‘MSE’ oraz ‘RMSE’ dla próby treningowej i testowej możemy stwierdzić iż model nie uległ przeuczeniu, a na zestawieniu otrzymano „chmurkę”, tak więc możemy stwierdzić iż model przewiduje poprawnie. Patrząc z kolei na zwykłą trafność 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**

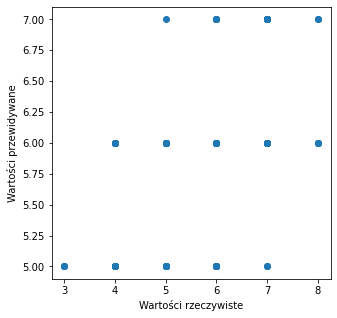
Parametry:

* Hidden\_layer\_size = (10, )
* Activation = tanh
* Solver = lbfgs
* Alpha = 0.0001
* Max\_iter = 10000
* Random\_state = 291951

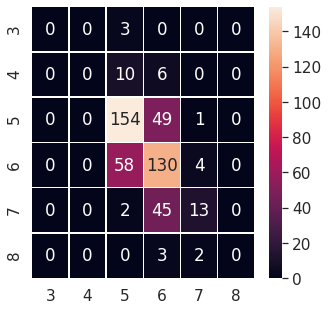
Zgodnie z założeniami wszystkie predyktory oraz zmienna celu zostały znormalizowane. Metodą prób i błędów, określenie większej ilości warstw ukrytych bądź też zwiększanie ilości neuronów nie wpływało na wynik, lub powodowało przeuczenie się sieci.

**

*Jakość modelu – sieć neuronowa*

**

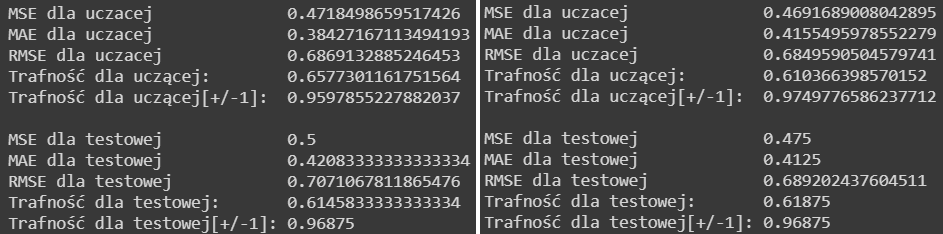
*Zestawienie wartości rzeczywistych z przewidywanymi – sieć neuronowa*

**

*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**

**

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

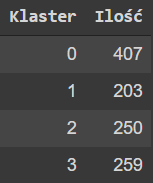
Obydwa modele uzyskały rewelacyjne wyniki. Otrzymaliśmy praktycznie identyczną trafność łącznie z odstępstwem o jeden dla obydwu modeli. Porównując dokładniej otrzymane rezultaty możemy zaobserwować drobną przewagę głównych miar jakości sieci neuronowej dla próby testowej. Może to byś spowodowane wyliczaniem 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:

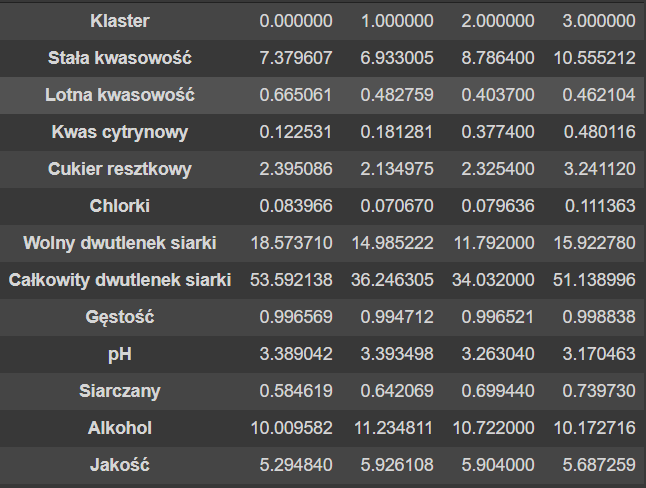
* M = 4
* N = 1
* Dim = 11

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.

**

*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.

**

*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:

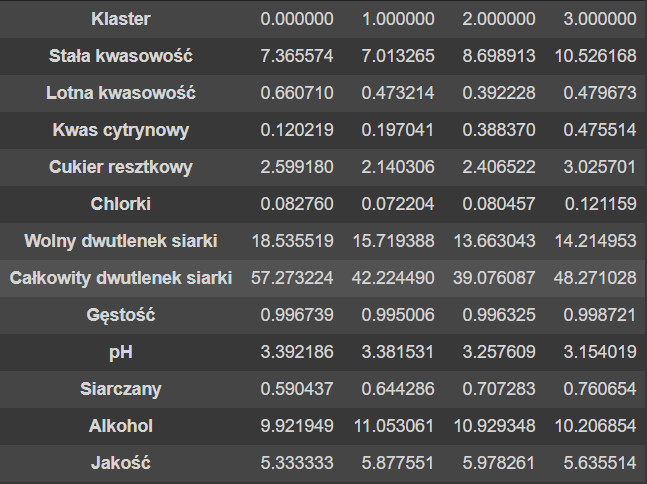
**Klaster 0** są to wina **słodkie**. Posiadają najwięcej lotnej kwasowości, wolnego dwutlenku siarki, całkowitego dwutlenku siarki, oraz najmniejszą ilość siarczanów, alkoholu i kwasu cytrynowego. Są to wina stosunkowo słabsze jakościowo, ich średnia wynosi około 5.3.

**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.



*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.

**

*Porównanie średnich wartości obserwacji w grupach dla danych treningowych i testowych*