

Celem projektu była eksploracja i analiza danych dotyczących składu białego wina, znajdujących się w pliku „winequality-white” w formacie .csv, a następnie wykorzystanie różnego rodzaju algorytmów do klasyfikacji, szacowania oraz grupowania danych. Zmienną celu jest zmienna przedstawiająca jakość wina (w skali od 3 do 9), natomiast pozostałe zmienne przedstawiają się następująco:

1. stała kwasowość,
2. lotna kwasowość,
3. kwas cytrynowy,
4. cukier resztkowy,
5. chlorki,
6. wolny dwutlenek siarki,
7. całkowity dwutlenek siarki,
8. gęstość,
9. pH,
10. siarczany,
11. alkohol.

Wszystkie predyktory to zmienne ilościowe. Do celów projektu wykorzystany został język programowania Python.

1. Analiza oraz ocena jakości danych

Pierwszym krokiem analizy było sprawdzenie jakości i rozkładu danych. Oznacza to, że sprawdzone zostało czy dane nie posiadają baraków, jeśli posiadają to w jakiej postaci, oraz ewentualne ich przekształcenie, a także sprawdzenie jak wyglądają i jakie przeciętne wartości przyjmują dane zmienne. W tym celu w zastosowane zostały specjalne funkcje *head* oraz *describe* a także *isnull*. Pierwsza z nich służy do wyświetlenia kilku pierwszych rekordów danych, kolejna pokazuje opisy statystyczne każdej zmiennej, takie jak np. średnia, odchylenie standardowe, czy kwartyle. Wartości te przedstawione zostały poniżej w Tabeli 1. Analizując ją można zauważyć, że zmienne przyjmują bardzo różne wartości. Ostatnia z użytych tutaj funkcji *isnull*, sprawdza czy występują brakujące rekordy. W tym wypadku okazało się, że takich nie ma.

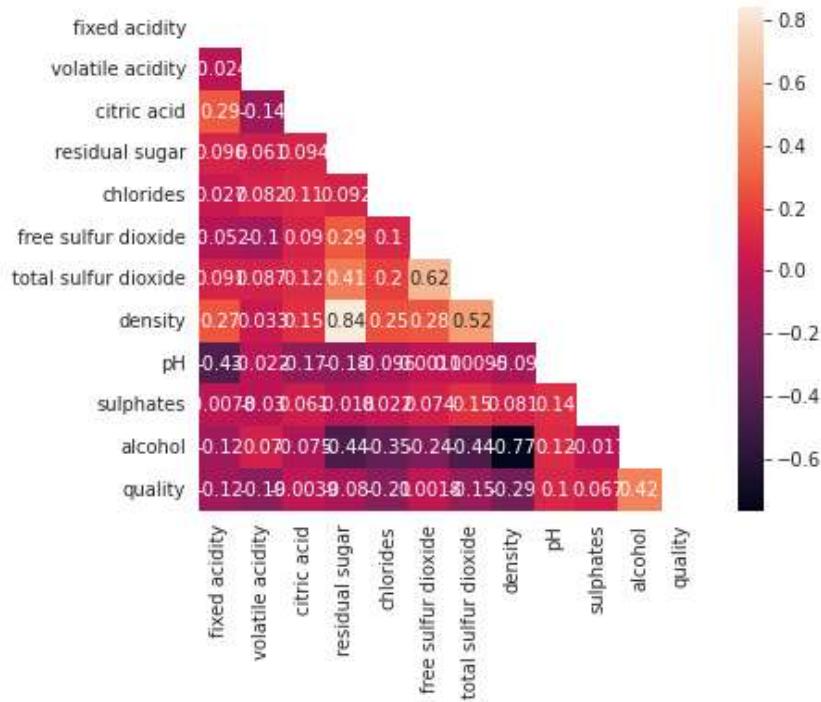
Tabela 1 Ogólny opis zbioru danych.

	ilość	średnia	odchylenie standardowe	min	25%	50%	75%	max
fixed acidity	4898	6,85	0,84	3,80	6,30	6,80	7,30	14,20
volatile acidity	4898	0,28	0,10	0,08	0,21	0,26	0,32	1,10
citric acid	4898	0,33	0,12	0,00	0,27	0,32	0,39	1,66
residual sugar	4898	6,39	5,07	0,60	1,70	5,20	9,90	65,80
chlorides	4898	0,05	0,02	0,01	0,04	0,04	0,05	0,35
free sulfur dioxide	4898	35,31	17,01	2,00	23,00	34,00	46,00	289,00
total sulfur dioxide	4898	138,36	42,50	9,00	108,00	134,00	167,00	440,00
density	4898	0,99	0,00	0,99	0,99	0,99	1,00	1,04
pH	4898	3,19	0,15	2,72	3,09	3,18	3,28	3,82
sulphates	4898	0,49	0,11	0,22	0,41	0,47	0,55	1,08
alcohol	4898	10,51	1,23	8,00	9,50	10,40	11,40	14,20
quality	4898	5,88	0,89	3,00	5,00	6,00	6,00	9,00

Po przeprowadzeniu ogólnej analizy całego zbioru, kolejnym krokiem było podzielenie go na uczący i testowy, przyjmując ziarno pseudolosowe podziału **313770**.

Korelacje

Następnie sprawdzono czy zmienne są ze sobą skorelowane i jeśli tak to w jakim stopniu. Macierz korelacji przedstawiona została na poniższym Rys.2.

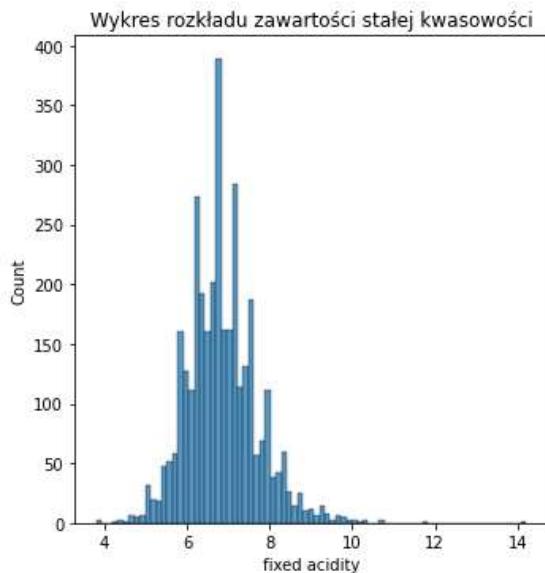


Rys. 1 Macierz korelacji między zmiennymi.

Można zauważyć, że najbardziej skorelowanymi dodatnio zmiennymi są cukier resztkowy oraz gęstość na poziomie 0.84, natomiast najbardziej skorelowane ujemnie jest zawartość alkoholu z gęstością na poziomie -0.77. Dodatkowo dość wysoką korelację wynoszącą 0.62 można zauważyć między wolnym i całkowitym dwutlenkiem siarki.

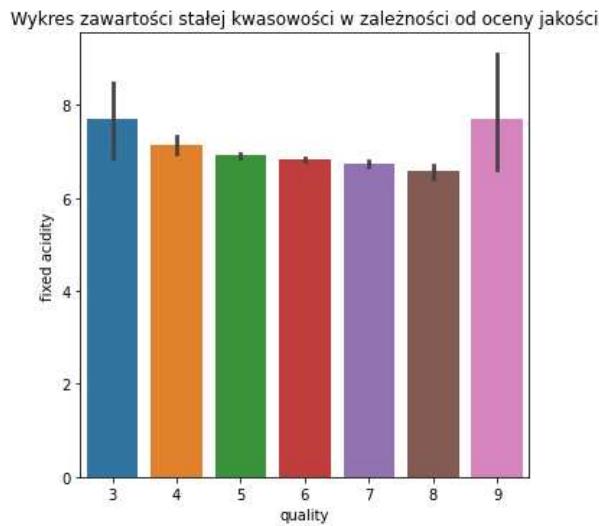
Najwyższą korelację w stosunku do zmiennej celu, czyli jakości wina można zaobserwować przy zmiennej mówiącej o zawartości alkoholu. Może to już na tym etapie sugerować, że jest to istotna zmienna.

Szczegółowa analiza zmiennych zbioru uczącego



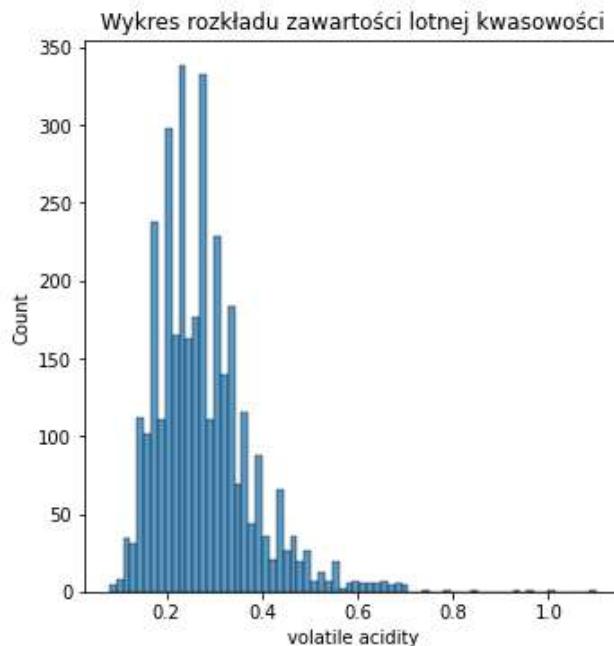
Rys. 2

Na powyższym histogramie stałej kwasowości (Rys.2) można zauważać, że najwięcej win przyjmuje wartość jej w okolicach 7 jednostek. Da się także dostrzec, że wykres ten jest prawostronnie skośny co może oznaczać, że w tej zmiennej występują obserwacje odstające.



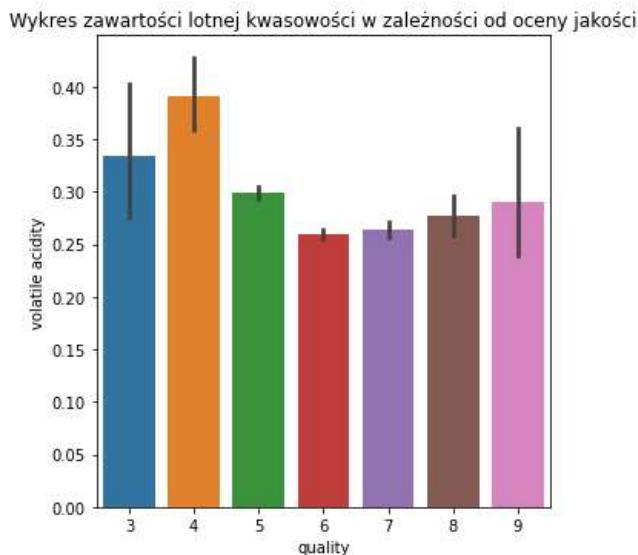
Rys. 3

Patrząc na wykres słupkowy zmiennej stałej kwasowości (Rys.3) w zależności od jakości, można zauważać, że najwięcej ocen najlepszych, czyli w tym wypadku 9, jest przy wysokiej zawartości tego czynnika. Jednakże, nie jest to prawdopodobnie żadna reguła ponieważ, wina o ocenie 3, również mają wysoką jego zawartość. Można powiedzieć, że wykres ma charakter spadkowy, czyli im mniejsza zawartość tego składnika tym lepsze wino, jednakże ostatni słupek wyłamuje się z tej reguły.



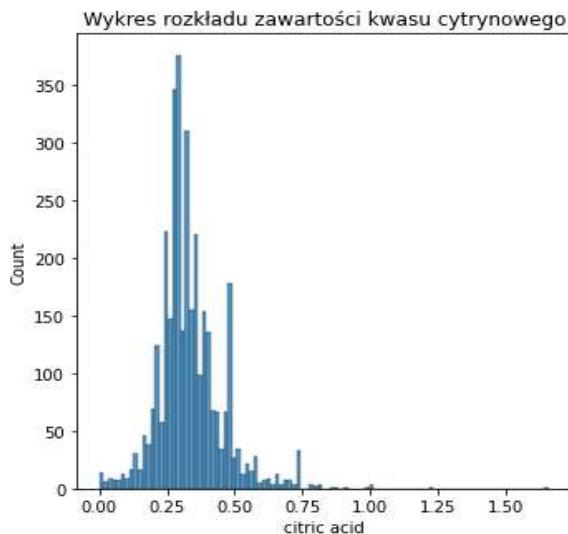
Rys. 4

Na histogramie zawartości lotnej kwasowości można zaobserwować, że najwięcej win ma jej zawartość w przedziale od 0.2 do 0.3. Wykres ten, tak jak poprzedni również jest prawostronnie skośny, czyli prawdopodobnie zawiera trochę obserwacji odstających.



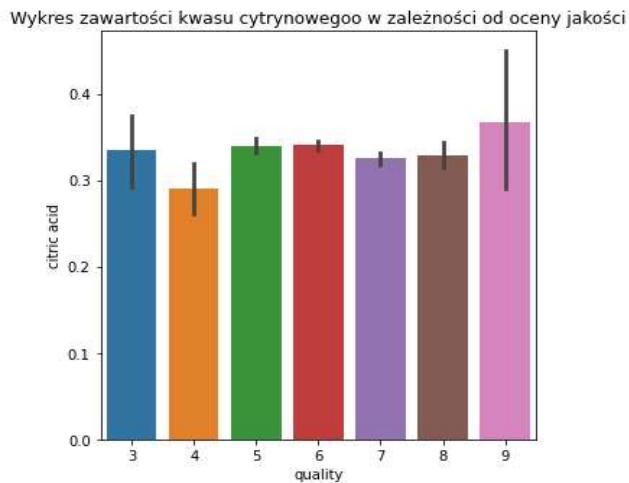
Rys. 5

Na wykresie zależności zawartości lotnej kwasowości od oceny można zauważać, że najgorsze oceny (3 i 4) mają wina z dużą zawartością tej substancji. Widać również, że przy ocenach 3, 4 oraz 9 występują obserwacje odstające, na które średnia arytmetyczna jest wrażliwa. Tzn, że jeśli histogram zaobserwowany na rys. jest prawostronnie skośny to średnia zawartość tej substancji w zależności od oceny może być zawyżona. Jednak, różnice w wysokości w słupkach są na tyle duże, że można zaryzykować stwierdzeniem, że substancja ta na jakość wina wpływa negatywnie.



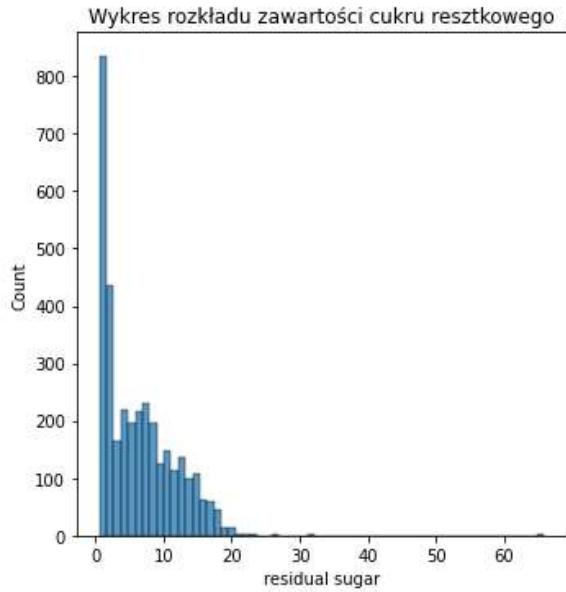
Rys. 6

Kwas cytrynowy jest substancją regulującą kwasowość wina. Można zauważyć, że najczęściej jest na poziomie około 0.3, jego rozkład również jest prawostronnie skośny z występującymi obserwacjami odstającymi.



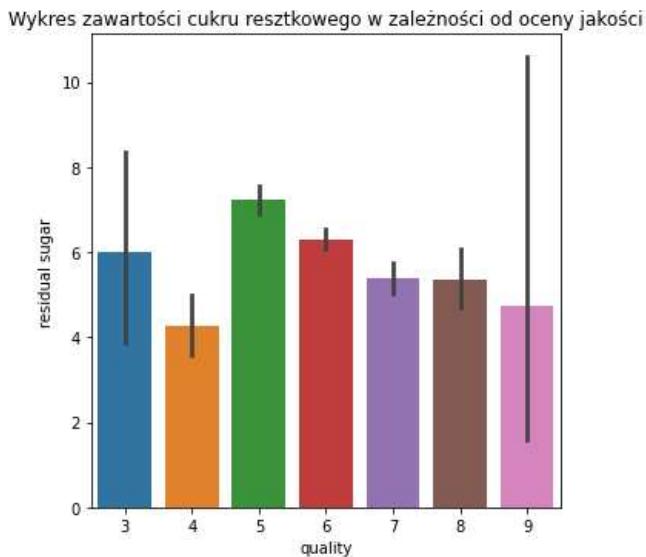
Rys. 7

Na powyższym wykresie można zaobserwować, że wszystkie słupki są mniej więcej na podobnym poziomie. Najbardziej wyróżnia się słupek z oceną 9, może to oznaczać, że kwaśne wina są lepiej oceniane, aczkolwiek jego wysokość może być też spowodowana poprzez występujące tam obserwacje odstające.



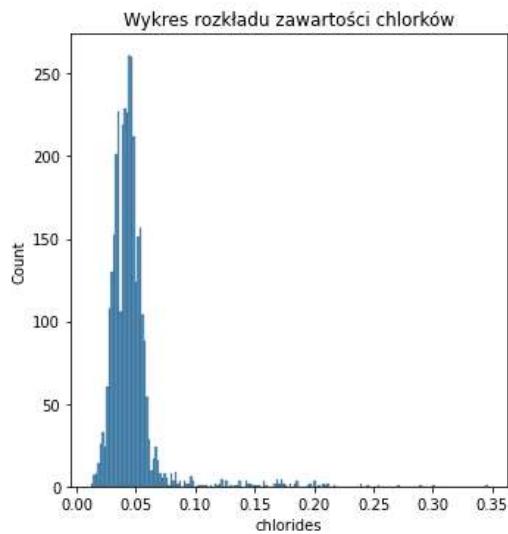
Rys. 8

Cukier resztowy odpowiada za poziom słodkości wina. Na powyższym histogramie można zaobserwować, że najczęściej win jest wytrawnych, posiadających bardzo mało tej substancji, a dalej wykres jest prawostronnie skośny, czyli prawdopodobnie jest średnia ilość win półwytrawnych i półsłodkich, a słodkich znacznie mniej, co pokrywa się z rzeczywistymi obserwacjami w sklepach.



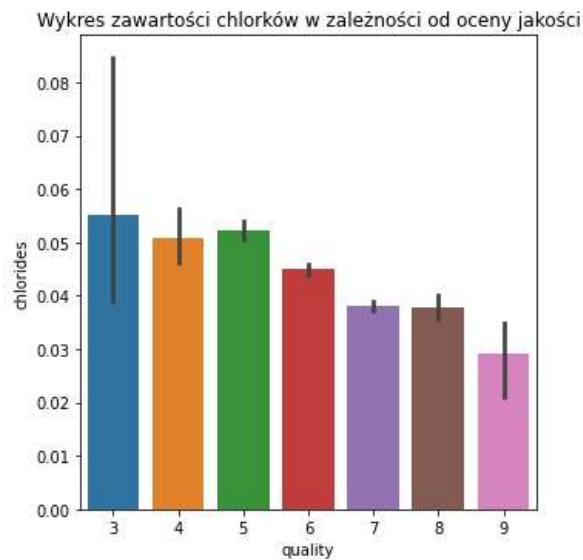
Rys. 9

Na powyższym wykresie zależności zawartości cukru resztowego od jakości można zauważyć, że wina z największą jego zawartością są oceniane jako średnie (5 i 6), natomiast wina oceniane jako bardzo dobre jakościowo raczej nie są bardzo słodkie, chociaż występuje tu duża liczba obserwacji odstających, więc prawdopodobnie zdarzają się odstępstwa od normy.



Rys. 10

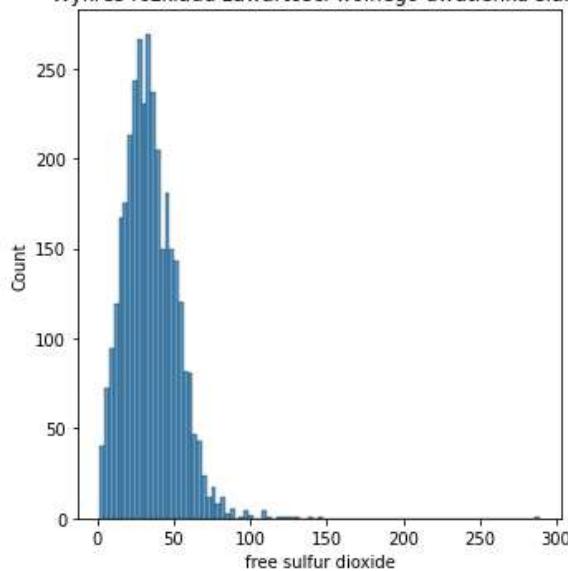
Na powyższym histogramie można zauważyc, że chlorki przyjmują bardzo niewielkie wartości i w większości win nie przekracza poziomu 0.05 jednostki. Chociaż ten histogram również jest prawostronnie skośny i posiada dużo obserwacji odstających.



Rys. 11

Na wykresie zależności zawartości chlorków od jakości jest bardzo wyraźna tendencja spadkowa, co oznacza, że wina z wysoką zawartością chlorków, są najgorzej oceniane, a im ich mniej tym wyższe dostają oceny. Można stwierdzić, że ta substancja nie wpływa korzystnie na jakość wina.

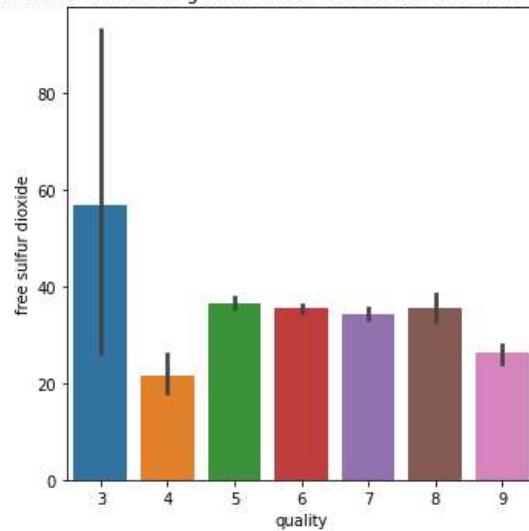
Wykres rozkładu zawartości wolnego dwutlenku siarki



Rys. 12

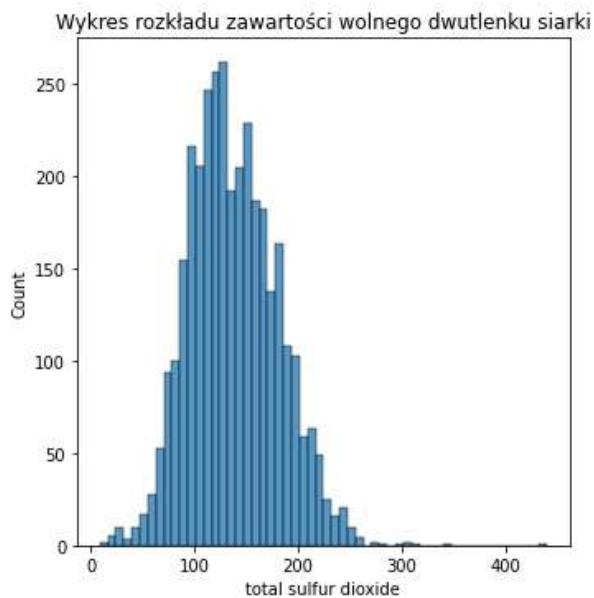
Zawartość wolnego dwutlenku siarki przyjmuje najczęściej wartości w przedziale od 25 do 50 jednostek. Wykres ten również jest prawostronnie skośny i występują w tej zmiennej obserwacje odstające.

Wykres zawartości wolnego dwutlenku siarki w zależności od oceny jakości



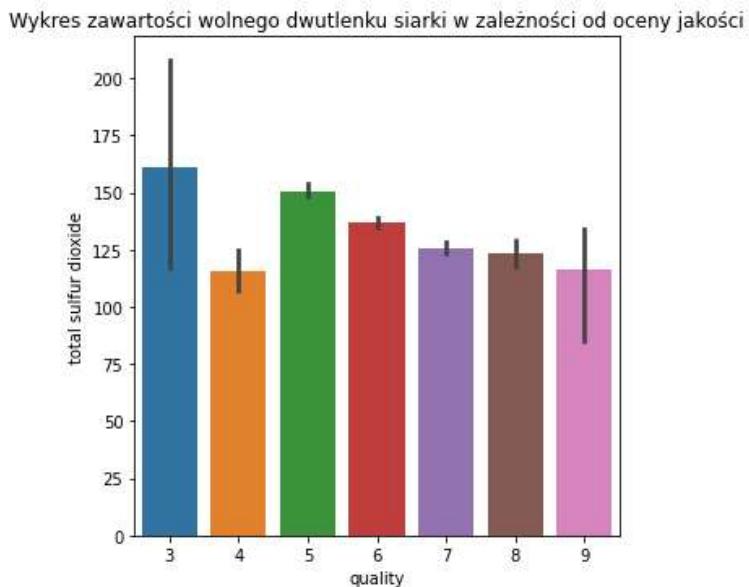
Rys. 13

Najgorsze oceny mają wina z najwyższą zawartością wolnego dwutlenku siarki, natomiast pozostałe zmienne mają mniej więcej ten sam poziom, oprócz 9, która ma wyraźnie mniejszą średnią zawartość tej substancji, oraz 4, która również znacznie odstaje w stosunku do pozostałych ocen.



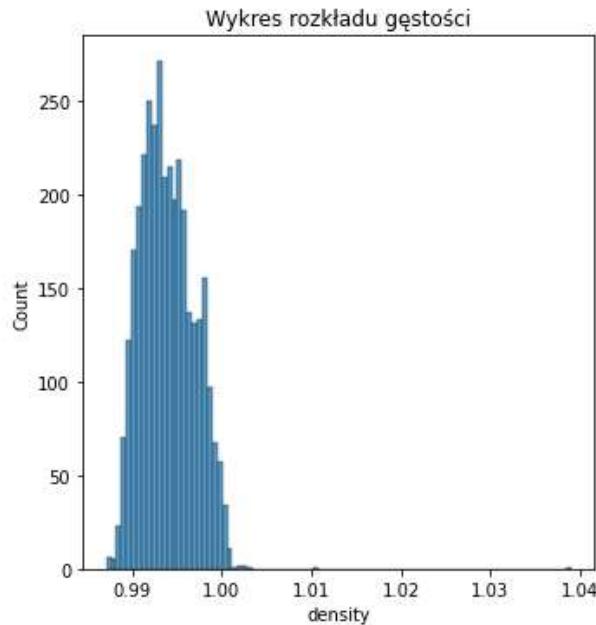
Rys. 14

Wolny dwutlenek siarki przyjmuje najczęściej wartości między 100 a 150 jednostek i tak jak poprzednie histogramy jest prawostronnie skośny.



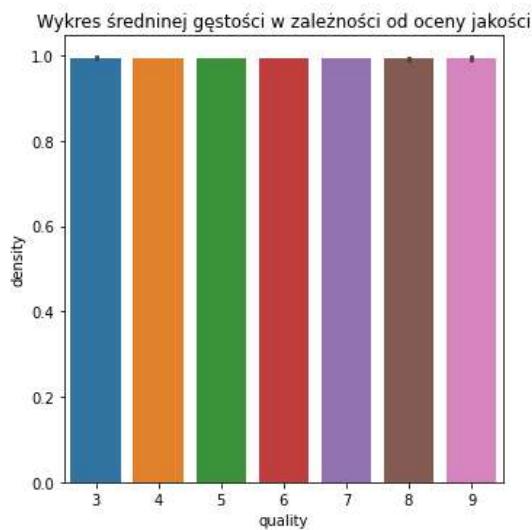
Rys. 15

Tutaj można zauważyć, że im mniejsza zawartość całkowitego dwutlenku siarki tym lepsze opinie. Ponownie tak jak na Rys.X wyjątkiem jest 4, która jest odstępstwem od tej reguły, jednak może to wynikać z tego, że opinia 4 jest rzadko stosowana.



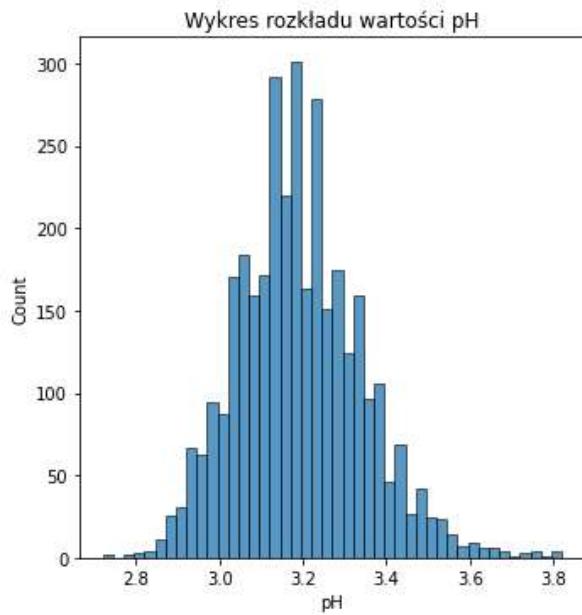
Rys. 16

Na powyższym histogramie można zauważyc, że gęstość wina wynosi między 0.99 a 1.00 jednostki. Oznacza to że wszystkie wina mają prawie taką samą gęstość, chociaż można zaobserwować również nieliczne obserwacje odstające.



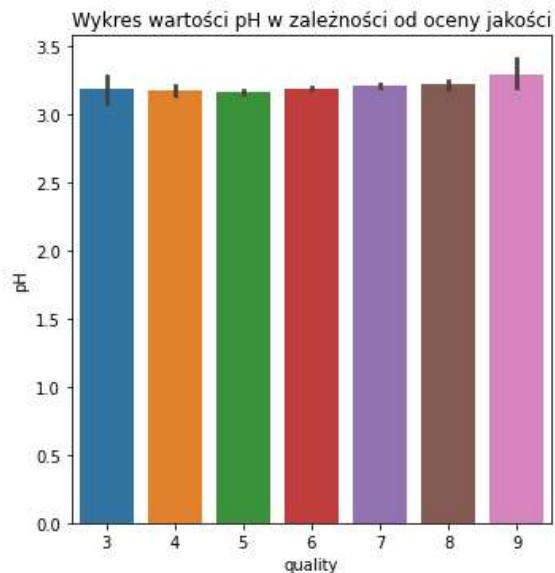
Rys. 17

Potwierdza się tutaj hipoteza z poprzedniego histogramu, że wszystkie wina mają podobną gęstość i jest ona zupełnie nieznacząca dla oceny.



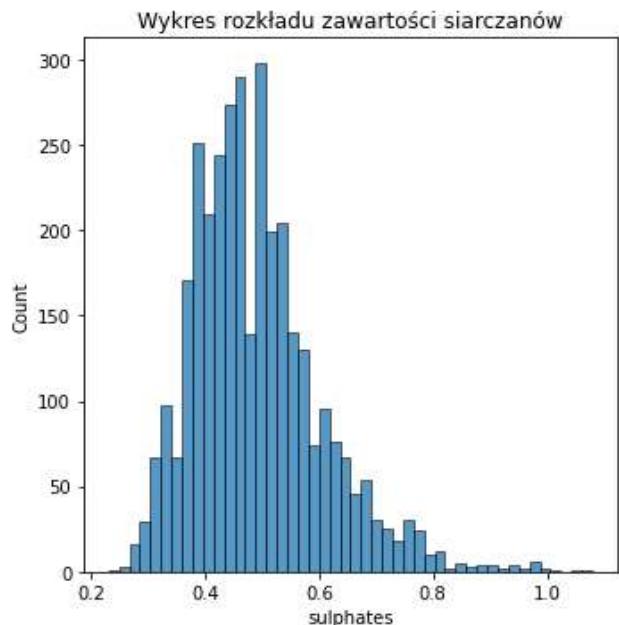
Rys. 18

Powyższy histogram przedstawia rozkład wartości pH w winach, najczęściej win ma tę wartość w okolicach 3.2 jednostki.



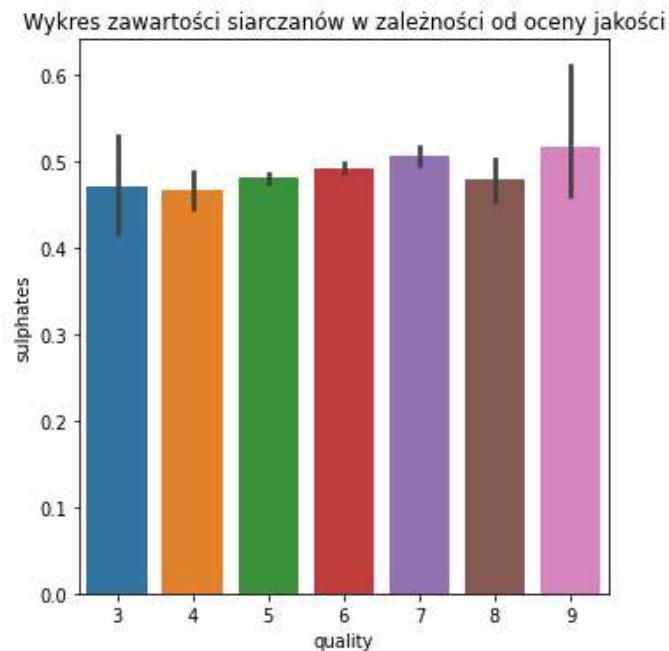
Rys. 19

Patrząc na rozkład wartości ph w zależności od oceny, można zauważyc, że im wyższa jego wartość tym wyższa ocena. Chociaż różnice pomiędzy wysokością słupków są raczej niewielkie, więc można zasugerować, że zmienna ta nie jest bardzo istotna.



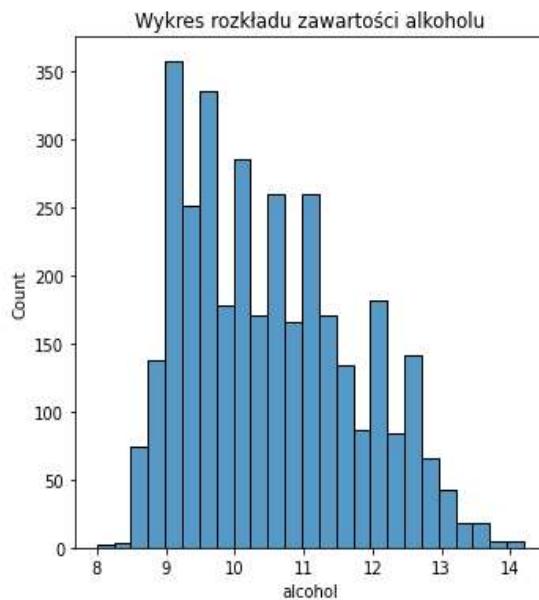
Rys. 20

Najwięcej win ma zawartość siarczanów w granicach 0.4-0.5 jednostki. Można zauważyć również delikatną prawoskoność wykresu sugerującą występowanie wartości odstających.



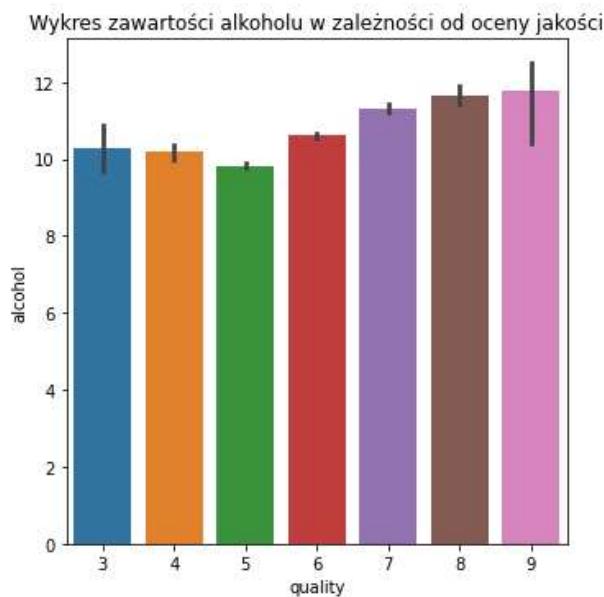
Rys. 21

Na powyższym wykresie widać, że zawartość siarczanów raczej pozytywnie wpływa na ocenę jakości wina. Mimo, że różnice w wysokościach słupków raczej są niewielkie, to dostrzegalny jest niewielki trend wzrostowy, z którego jednak wyłamują się wina ocenione na 8. Jednakże w przypadku tej zmiennej możliwym jest również, że taki rozkład słupków powstał przypadkowo i nie ma ona dużego wpływu na ocenę jakości.



Rys. 22

Na histogramie rozkładu alkoholu można zauważać, że większość win ma zawartość alkoholu na poziomie około 9%. W tej zmiennej również występuje prawoskośność rozkładu, jednakże jest znacznie mniejsza w stosunku do innych zmiennych i raczej nie występują tu wartości odstające.



Rys. 23

Na wykresie zależności zawartości alkoholu od jakości widać, że najlepsze oceny mają wina z wysoką zawartością alkoholu. Prawdopodobnie zawartość alkoholu może mieć wpływ na ocenę jakości.

Podczas powyższej analizy można było spostrzec, że pośród bardzo wielu zmiennych wystąpiły obserwacje odstające. Jednakże z uwagi na to, że usuwanie takich obserwacji jest bardzo szerokim zagadnieniem z wieloma różnymi podejściami, zdecydowano się je pozostawić.

Z danych usunięto natomiast zmienną gęstości, z tego względu, że we wszystkich winach przyjmowała bardzo zbliżoną wartość oraz ma bardzo wysoką korelację z cukrem resztowym, więc uznano, że nie będzie istotna dla modeli. Dodatkowo rozważano usunięcie zmiennej dotyczącej zawartości całkowitego dwutlenku siarki lub wolnego dwutlenku siarki, ze względu na ich dość dużą korelację, jednak ostatecznie uznano, że obie mogą być w pewien sposób istotne dla modeli i zdecydowano się je zostawić.

2. Stworzenie modeli

Klasyfikacja

Pierwszym wybranym modelem jest model K-najbliższych sąsiadów. Algorytm jest przykładem uczenia leniwego. Oznacza to, że zapamiętuje cały zbiór danych i porównuje nowe przypadki z rekordami zbioru. Służy on do znajdywania podobnych rekordów, oraz klasyfikacji i szacowania pewnej zmiennej dla nowych obserwacji.

Zmienna celu traktowana była w tym przypadku jako zmienna jakościowa. Początkowo konieczne było zastosowanie skorygowanej normalizacji zmiennych. Następnie dostosowując parametry, dającym największą trafność na zbiorze uczącym oraz testowym okazało się 16 sąsiadów, uwzględnienie wag w postaci odległości, a także wykorzystanie standardowej metryki euklidesowej. Trafności z tymi parametrami na zbiorze uczącym wynosi 1.00, a na zbiorze testowym 0.66. Jednakże, taka różnica między trafnością na obu zbiorach może oznaczać, że model nadmiernie dostosował się do danych zbioru uczącego, z tego względu zdecydowano się na zmianę parametrów kosztem ich pogorszenia.

Ostatecznie wybrano liczbę sąsiadów równą 3, miarę ‘uniform’, gdzie wagi wszystkich sąsiadów rozdzielane są po równo oraz również standardową miarę euklidesową. Trafność na zbiorze uczącym z takimi parametrami wynosi 78%, na zbiorze testowym 56%. Trafność z odstępstwem o 1 (w górę lub w dół) wynosi 95% na zbiorze uczącym, a 91% na zbiorze testowym. Błąd MAE na zbiorze uczącym wynosi 0.279, a na testowym 0.549. Oznacza to, że algorytm na zbiorze uczącym myli się średnio o 0.28 oceny, a na zbiorze testowym o około 0.55 oceny. Algorytm w tym wypadku również jest przeuczony, jednak nie aż tak drastycznie.

Tabela 2 Wyniki modelu klasyfikacji

	Zbior uczący	Zbior testowy
Trafność	0.777	0.556
Trafność z odstępstwem	0.952	0.912
MAE	0.279	0.549

Szacowanie

Drugim wybranym algorytmem jest MLP. Jest to algorytm sieci neuronowej i może służyć zarówno do szacowania jak i klasyfikacji. Jest to model wielowarstwowy i składa się z warstwy wejściowej, warstwy ukrytej oraz warstwy wyjściowej. Warstwy składają się z neuronów. W warstwie wejściowej każdy neurony odpowiadają zmiennym lub ich kategoriom (w przypadku zmiennych jakościowych), a

oprócz tego występuje neuron zwany stałym obciążeniem zewnętrznym. Dane w algorytmie, przekazywane są między warstwami do kolejnych neuronów, a ich połączenia mają pewne wagę, które początkowo przyjmują wartości z przedziału [0,1]. Do każdego neuronu warstwy ukrytej trafia suma wartości wszystkich neuronów z poprzedniej warstwy pomnożona przez odpowiednie wagę. Na tej sumie obliczana jest funkcja aktywacji, która zajmuje się wygaszaniem słabszych sygnałów, a wychwytywaniem silnych i przekazywaniem ich dalej. Wartości otrzymywane w warstwie wyjściowej są prawdopodobieństwami wartości lub przynależenia do pewnej kategorii zmiennej celu.

W tym algorytmie zmienne także najpierw musiały zostać znormalizowane. Zdecydowano się na ustawienie parametru `hidden_layer_sizes` na (80,), funkcji aktywacyjnej 'relu' oraz solvera 'lbfgs'. Ustawienia takich parametrów dały wyniki 47% trafności na zbiorze uczącym, oraz 38% na zbiorze testowym. Nie są to zbyt dobre wyniki, ale w stosunku do próby wykorzystania regresji liniowej wielorakiej, na której trafności wychodziły w okolicach 30% ten model prezentuje się lepiej. Trafność z odstępstwem na zbiorze uczącym wynosi 98%, a na zbiorze testowym 96%. Błąd MAE na zbiorze uczącym wynosi 0.43, a na testowym 0.48. Oznacza to, że średnie odchylenia wynoszą 0.4 oceny na zbiorze uczącym a około 0.5 oceny na zbiorze testowym.

Tabela 3 Wyniki algorytmu szacowania.

	Zbiór uczący	Zbiór testowy
Trafność	0.472	0.380
Trafność z odstępstwem	0.975	0.960
MAE	0.430	0.480

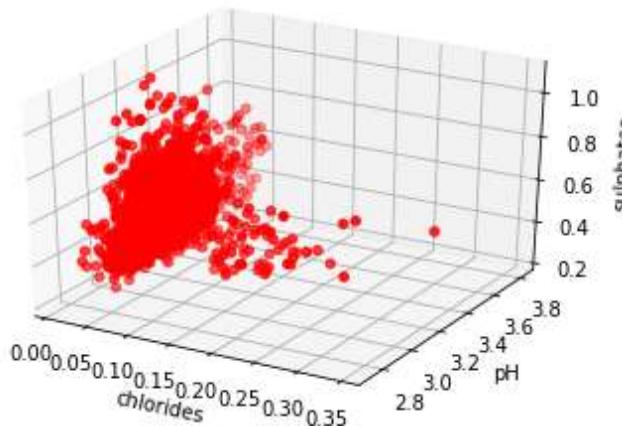
Niestety na żadnym z wybranych algorytmów nie ma możliwości przedstawienia najważniejszych predyktów.

Grupowanie

Ostatnim wykorzystanym algorytmem jest algorytm K-średnich. Jest to algorytm nienadzorowany, oznacza to, że nie występuje w nim zmienna celu. Celem grupowania jest znalezienie grup, które są jak najbardziej podobne do siebie oraz jak najbardziej różne od pozostałych grup. Algorytm ten działa na podstawie liczenia odległości między punktami. Jako początkowe centra grup wybiera się k obserwacji. Następnie przypisywane są obserwacje, które mają najbliższej do tych punktów centralnych. Kiedy już zostaną wybrane, centra wybiera się na nowo i cały proces jest powtarzany aż wszystkie obserwacje nie przestaną zmieniać grup.

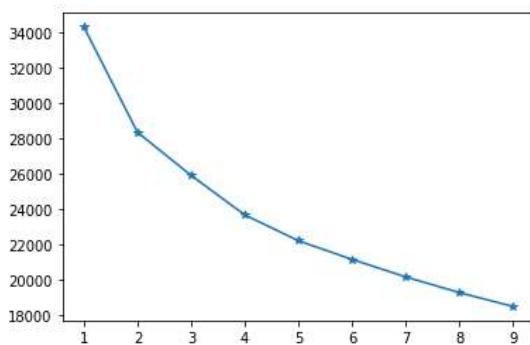
Przed podziałem na grupy dane należało zestandardyzować. Następnie wykonane zostały wykresy rozrzutu pomiędzy wszystkimi zmiennymi, żeby zaobserwować jakie są zależności i czy któraś wyraźnie rozdziela pozostałe na grupy. Wykres ten jednak jest zbyt duży i z tego względu nie został umieszczony w raporcie, aczkolwiek załączono go w folderze przesłanym razem z raportem. Plik nazywa się „wykresy rozrzutu.png”. Kolejnym krokiem było wykonanie wykresów trójwymiarowych pomiędzy zmiennymi, w których zauważono możliwe rozdiały. Przykładem jest poniższy Rys. 24 ,

gdzie przedstawiono zależności zmiennych chlorides, sulphates oraz pH. Widać na nim, że zarysuje się tu pewien podział na grupy, najbardziej widoczne są dwie.



Rys. 24 Wykres 3D zależności zmiennych chlorki, siarczany oraz pH.

Po wykonaniu i przeanalizowaniu wykresów kolejnym krokiem było zbudowanie modelu. W celu zbudowania modelu zmienne najpierw należało zestandardyzować. Następnie dopasowanie najlepszej liczby grup przeprowadzono za pomocą wyników miary Silhouette oraz analizy wykresu łokciowego. Miara Silhouette dla dwóch grup była najwyższa ponieważ osiągnęła wartość 0.19. Kolejną wartością wyróżniającą się na tle pozostałych była wartość 0.15 dla 4 grup. Miara ta przyjmuje, że wynik poniżej 0.2 jest słabym dopasowaniem, poprawne wartości znajdują się pomiędzy 0.2 a 0.5, a dopiero od 0.5 wartości są dobre. Oznacza to, że podział na grupy w przypadku wykorzystanych danych nie jest zbyt silny, jednakże w celach raportu zdecydowano się i tak go wykonać. Dodatkowo na poniższym Rys.25 można zauważać wyraźny punkt odcięcia w punkcie 2, natomiast przy 4 odcięcie to nie jest na tyle wyraźne, dlatego zdecydowano się wykonać grupowanie z parametrem 2.



Rys. 25

Przy podziale danych na dwie grupy w opisie pierwszej przedstawionej na Rys.26 można zauważać, że wyróżnia się ona przede wszystkim wysoką zawartością cukrów resztkowych oraz niską zawartością alkoholu. Dodatkowo jest dość dużo kwasu cytrynowego, oraz bardzo duże ilości wolnego oraz

całkowitego dwutlenku siarki, a także większe ilości chlorków. Grupa ta ma również wyraźnie niższe pH. Po tych wartościach ogólnie można stwierdzić, że są to słodkie wina z niewielką ilością alkoholu. Raczej wina te nie są najwyższej jakości, co można zauważyć po uprzedniej analizie wykresów zależności ich zawartości od oceny. Są to raczej wina powszechnie i banalne.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol
count	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000
mean	0.178755	0.047481	0.243295	0.786525	0.391547	0.649474	0.815461	-0.222962	0.097448	-0.747192
std	0.923392	0.954353	1.137440	0.973720	1.338556	1.013332	0.800142	0.904102	0.933093	0.584577
min	-3.092590	-1.710396	-2.086977	-1.087846	-1.150121	-1.675259	-1.500635	-3.127867	-1.932398	-2.065820
25%	-0.408293	-0.571490	-0.535117	0.194613	-0.125411	-0.017551	0.301748	-0.793345	-0.523655	-1.162782
50%	0.058542	-0.150590	0.036621	0.802613	0.088070	0.611235	0.769899	-0.259739	0.004623	-0.916499
75%	0.642084	0.418863	0.853390	1.452458	0.386944	1.125696	1.308273	0.273866	0.532902	-0.423933
max	8.578267	7.202783	7.305860	11.712135	12.811555	14.501686	7.066534	3.675598	4.583038	2.120992

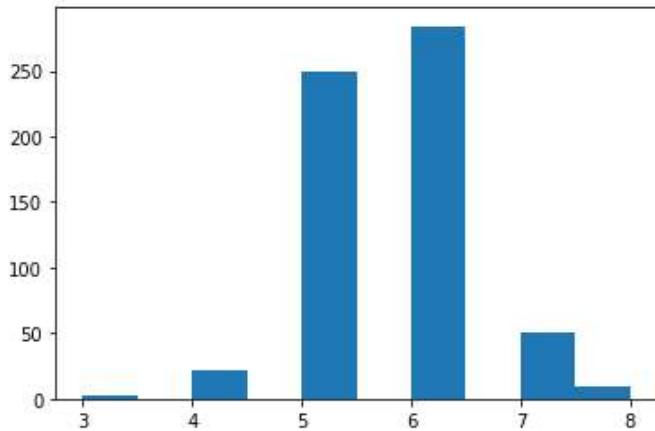
Rys. 26 Opis grupy 1

Wartości grupy drugiej (Rys.27) cechują się wysoką zawartością alkoholu, niską zawartością cukrów resztkowych, a także kwasu cytrynowego, możliwe że ze względu na to, że naturalnie mają kwaśny posmak, a cukier go za bardzo nie przytłumia. Wina te nie mają dużej ilości wolnego i całkowitego dwutlenku siarki oraz mają niską zawartość chlorków. Przez brak tych niepożądanych substancji, można powiedzieć, że są to wina bardziej ekskluzywne i mocne.

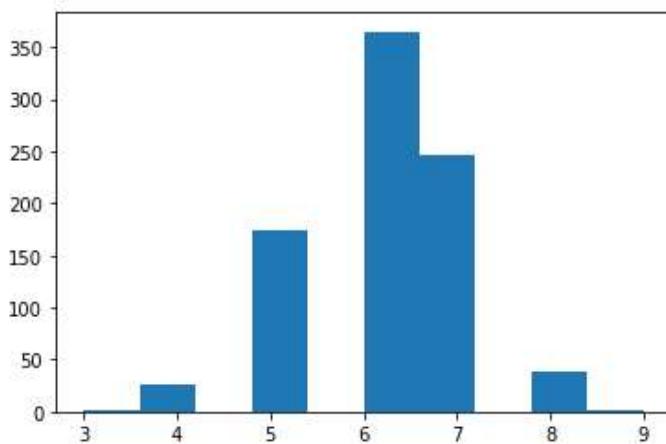
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol
count	2046.000000	2046.000000	2046.000000	2046.000000	2046.000000	2046.000000	2046.000000	2046.000000	2046.000000	2046.000000
mean	-0.120742	-0.032072	-0.164337	-0.531270	-0.264476	-0.438697	-0.550815	0.150603	-0.065823	0.504702
std	1.031651	1.028917	0.857654	0.579274	0.541045	0.710803	0.700715	1.033442	1.038060	0.901870
min	-3.559424	-1.957984	-2.740391	-1.127231	-1.448995	-1.903908	-3.022126	-2.994466	-2.284584	-1.737443
25%	-0.875127	-0.769560	-0.535117	-0.969693	-0.595070	-0.932148	-1.009076	-0.526542	-0.787795	-0.177650
50%	-0.174876	-0.175349	-0.208409	-0.812154	-0.338892	-0.474849	-0.564332	0.073764	-0.259516	0.479105
75%	0.408667	0.418863	0.199975	-0.223847	0.002678	0.039612	-0.119588	0.807471	0.444855	1.217954
max	4.026633	8.143618	10.817963	2.653687	5.681280	2.840567	1.858351	4.209203	5.199363	3.024030

Rys. 27 Opis grupy 2.

Poniższe histogramy (Rys. 28 i Rys. 29) przedstawiają jak rozkładały się oceny jakości z takim podziałem na grupy na zbiorze testowym. Można zauważyć, że w pierwszej grupie jest zdecydowanie więcej win przeciętnych i trochę więcej słabszych. Zdecydowanie przeważają oceny 5 i 6, ponieważ takich ocen było najwięcej, ale można zaobserwować różnicę patrząc na to, że w grupie drugiej jest więcej ocen 7 niż 5, oraz większa ilość 8 niż w grupie 1, a także pojawiają się 9.



Rys. 27 Histogram ocen jakości dla grupy 1 na zbiorze testowym.



Rys. 28 Histogram ocen dla grupy 2 na zbiorze testowym.

3. Podsumowanie

Ostatecznie model klasyfikacji okazał się najlepszym modelem, chociaż jego wyniki nie są wybitne, to trafność ma dużo wyższą niż algorytm szacujący. Może to wynikać z tego, że zmienna celu jest zmienną jakościową i bardziej prawidłowym podejściem wydaje się używanie algorytmu klasyfikującego w celu jej przewidzenia. Jednakże wyniki z odstępstwem o 1 osiągają wysokie wartości na obu modelach. Biorąc pod uwagę to, że każda osoba oceniająca wino jest w tym subiektywna i ma różne metody oceniania, to różnica jednego punktu oceny w górę lub w dół nie wydaje się dużą różnicą.

W wybranych modelach niestety nie można było ocenić ważności predyktorów, ponieważ modele te nie mają takich funkcji, więc niemożliwym było potwierdzenie obserwacji z analizy zmiennych. Aczkolwiek patrząc na grupy, które powstały po rozdzieleniu, można zauważać, że zmienne które zostały zaobserwowane jako ważne, mają większe różnice w wartościach pomiędzy grupami.

Dodatkowo dzięki algorytmowi k-średnich wyróżnić można było przede wszystkim jakie substancje zawierają wino oceniane jako dobre.

Ostatnim spostrzeżeniem jest to, że na pogorszenie jakości modeli mogło mieć wpływ nieusunięcie obserwacji odstających, których było dużo. Dodatkowo w zmiennej celu oceny nie były zbalansowane i znacznie przeważyły oceny 5 i 6, co prawdopodobnie też miało duży wpływ na zbudowane modele.