WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA
UNIWERSYTET MIKOŁAJA KOPERNIKA W TORUNIU
DATA SCIENCE W BIZNESIE 2019/2020

Łukasz Marchlewicz
Łukasz Morawiec

OCENA JAKOŚCI WINA

Praca dyplomowa
napisana pod kierunkiem
dr Joanny Karłowskiej-Pik

Toruń 2020

Spis treści

1. Zrozumienie biznesu

W dobie rosnącej popularności wina i dużej konkurencji na tym rynku, wysoka jakość produktu, która spełni oczekiwanie klienta jest najważniejsza. Niestety, dyskryminacja win nie jest łatwym procesem, ze względu na ich złożoność i niejednorodność. Dlatego do oceny jakości wina stosuje się testy fizykochemiczne, jak i sensoryczne przeprowadzane przez człowieka. Taka klasyfikacja win zapewnienia odpowiednią ich jakość ułatwiając kontrolę etapów przetwarzania oraz chroni przed ich fałszowaniem. Ze względu na wysoki koszt oraz czasochłonność analizy wykonywanej przez kipera, można by, stosując metody machine learningu i w oparciu o łatwo dostępne testy analityczne (dostępne już na etapie certyfikacji), próbować przewidzieć taką ocenę. Wówczas, przewidywana wartość może być wykorzystana do projektowania nowych rodzajów wina, definiowania polityki cenowej lub wspomagania podejmowania decyzji w systemach doradczych. W naszej pracy do tego celu wykorzystaliśmy dwie najbardziej znane metody predykcji, regresję oraz klasyfikację.

2. Otoczenie biznesowe

Podmiotem zainteresowanym wykorzystaniem nowoczesnych metod i narzędzi machine learningu, jest firma pośrednicząca w procesie dystrybucji wina, która dysponuje zestawem danych dotyczących parametrów fizykochemicznych wybranych partii wina wraz z oceną jakości wina dokonaną przez człowieka. Wykorzystując te dane firma zamierza zrealizować następujące cele biznesowe:

- wskazanie najważniejszych parametrów wpływających na ocenę jakości wina;
- stworzenie mechanizmu automatycznej oceny jakości wina, który mógłby zastąpić ocenę dokonywaną przez człowieka.

3. Zrozumienie danych

Dane analizowane to zbiory danych dotyczące jakości wina czerwonego i białego. Zawierają one 1598 obserwacji dla wina czerwonego i 4897 obserwacji dla wina białego. Obserwacji dla wina białego jest o ok. 30% więcej niż dla wina czerwonego.

Oba rodzaje win opisane są 11 cechami fizykochemicznymi oraz oceną sensoryczną ekspertów.

Opis cech zbioru danych:

- fixed acidity - kwasowość pierwotna wina (stężenie kwasu winowego w g/L);
- volatile acidity - kwasowość powstała podczas procesu dojrzewania (stężenie kwasu octowego w g/L);
- citric acid - stężenie kwasu cytrynowego (g/L) ;
- residual sugar - zawartość cukru pozostała po procesie fermentacji (g/L);
- chlorides - zawartość chlorku sodu (g/L);
- free sulfur dioxide - zawartość wolnego dwutlenku siarki (mg/L);
- total sulfur dioxide - całkowita zawartość dwutlenku siarki (mg/L);
- density - gęstość wina (g/cm^3);
- pH - odczyn wina wyrażony przez poziom pH (powyżej 7 – zasadowość, poniżej 7 – kwasowość);
- sulphates - zawartość siarczanu potasu (g/L);
- alcohol - procentowa zawartość alkoholu;
- quality - jakość wina, ocena ekspercka, liczba naturalna z przedziału (0-10);

Dane wejściowe obejmują zatem obiektywne testy, jak np. kwasowość pierwotną (fixed acidity) czy zawartość alkoholu (alcohol), a dane wyjściowe są oparte na danych sensorycznych, gdzie każdy ekspert ocenił jakość wina w przedziale od 0 -bardzo zła do 10 -bardzo dobra.

3.1 Porównanie win czerwonych z białymi na podstawie statystyki opisowej

Tabela 1 Statystyki opisowe dla atrybutów dla wina czerwonego

Rodzaj cechy	min	25%	50%	75%	max	mean	std
fixed acidity	4,60	7,10	7,90	9,20	15,90	8,32	1,74
volatile acidity	0,12	0,39	0,52	0,64	1,58	0,53	0,18
citric acid	0,00	0,09	0,26	0,42	1,00	0,27	0,19
residual sugar	0,90	1,90	2,20	2,60	15,50	2,54	1,41
chlorides	0,01	0,07	0,08	0,09	0,61	0,09	0,05
free sulfur dioxide	1,00	7,00	14,00	21,00	72,00	15,88	10,46
total sulfur dioxide	6,00	22,00	38,00	62,00	289,00	46,48	32,90
density	0,99	1,00	1,00	1,00	1,00	1,00	0,00
pH	2,74	3,21	3,31	3,40	4,01	3,31	0,15
sulphates	0,33	0,55	0,62	0,73	2,00	0,66	0,17
alcohol	8,40	9,50	10,20	11,10	14,90	10,42	1,07
quality	3,00	5,00	6,00	6,00	8,00	5,64	0,81

Tabela 2 Statystyki opisowe dla atrybutów dla wina białego

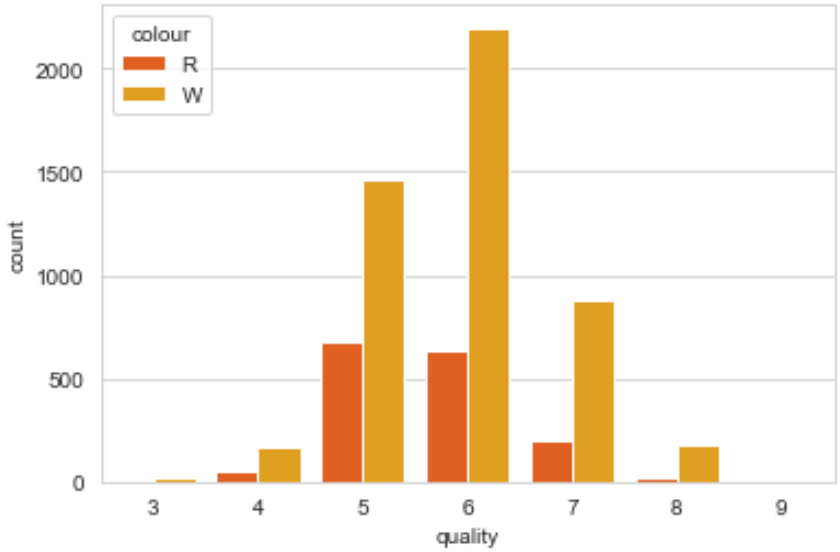
Rodzaj cechy	min	25%	50%	75%	max	mean	std
fixed acidity	3,80	6,30	6,80	7,30	14,20	6,85	0,84
volatile acidity	0,08	0,21	0,26	0,32	1,10	0,28	0,10
citric acid	0,00	0,27	0,32	0,39	1,66	0,33	0,12
residual sugar	0,60	1,70	5,20	9,90	65,80	6,39	5,07
chlorides	0,01	0,04	0,04	0,05	0,35	0,05	0,02
free sulfur dioxide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

Analizując powyższe statystyki można zauważyć, że pomiędzy winami czerwonymi a białymi dla wartości średnich (mean) największe różnice występują w przypadku zawartości wolnego dwutlenku siarki 19,43 (free sulfur dioxide), całkowitej zawartości dwutlenku siarki 91,87 (total sulfur dioxide) i zawartości cukru 3,85 (residual sugar) z przewagą dla win białych oraz kwasowości pierwotnej 1,47 (fixed acidity) dla win czerwonych. W przypadku oceny jakości

(quality), warto zaznaczyć, że średnia ocena dla win białych jest wyższa, niż dla win czerwonych (5,88 vs 5,64).

3.2 Porównanie win czerwonych z białymi na podstawie oceny jakościowej (quality)

Rysunek 1 Porównanie oceny jakości win



Z powyższego rysunku można wywnioskować, że niezależnie od koloru wina większość win otrzymała notę w granicach 5-7. Najniżej ocenione wino posiada ocenę 3, a najwyżej – ocenę 9. W zbiorze danych znajduje się więc wiele win przeciętnych i tylko kilka ocenianych wysoko. Brak jest win z oceną skrajną - 0 lub 10.

3.3 Korelacja pomiędzy cechami w odniesieniu do oceny jakościowej wina – (współczynnik korelacji r Pearsona)

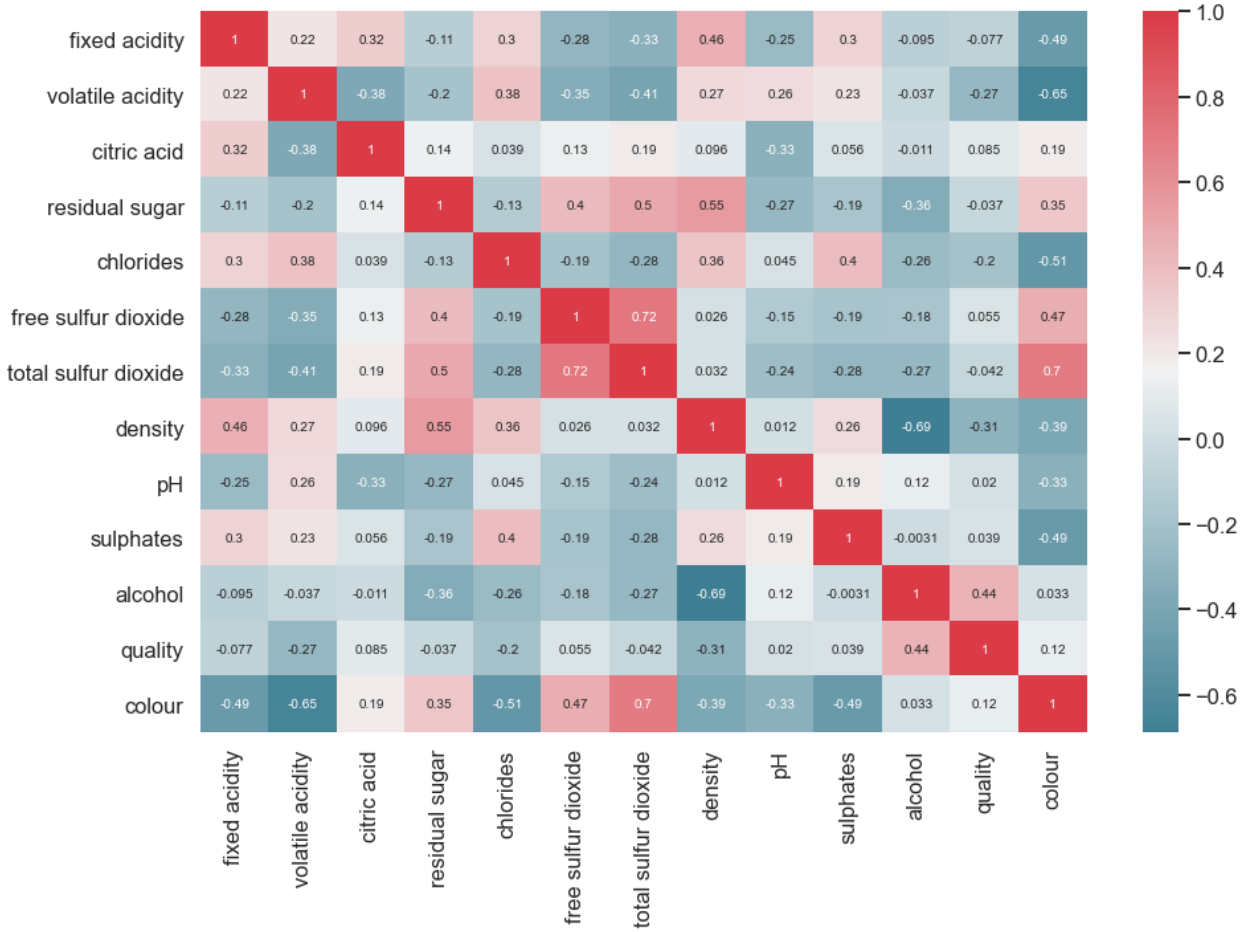
Współczynnik korelacji liniowej Pearsona określa poziom zależności liniowej między zmiennymi losowymi i mieści się w przedziale domkniętym [−1, 1]. [1]Gdy zbliża się do 1, oznacza to, że istnieje silna dodatnia korelacja, i tak, dla wartości jakość (quality) ma ona tendencję rosnącą wraz ze wzrostem zawartości alkoholu (alcohol). Gdy współczynnik jest bliski –1, oznacza, że istnieje silna

[1] https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/3033,pojecie.html

korelacja ujemna, na przykład, kiedy spada gęstość wina (density) lub zmniejsza się zawartość stężenia kwasu octowego (volatile acidity), wówczas ocena jakości wina rośnie.

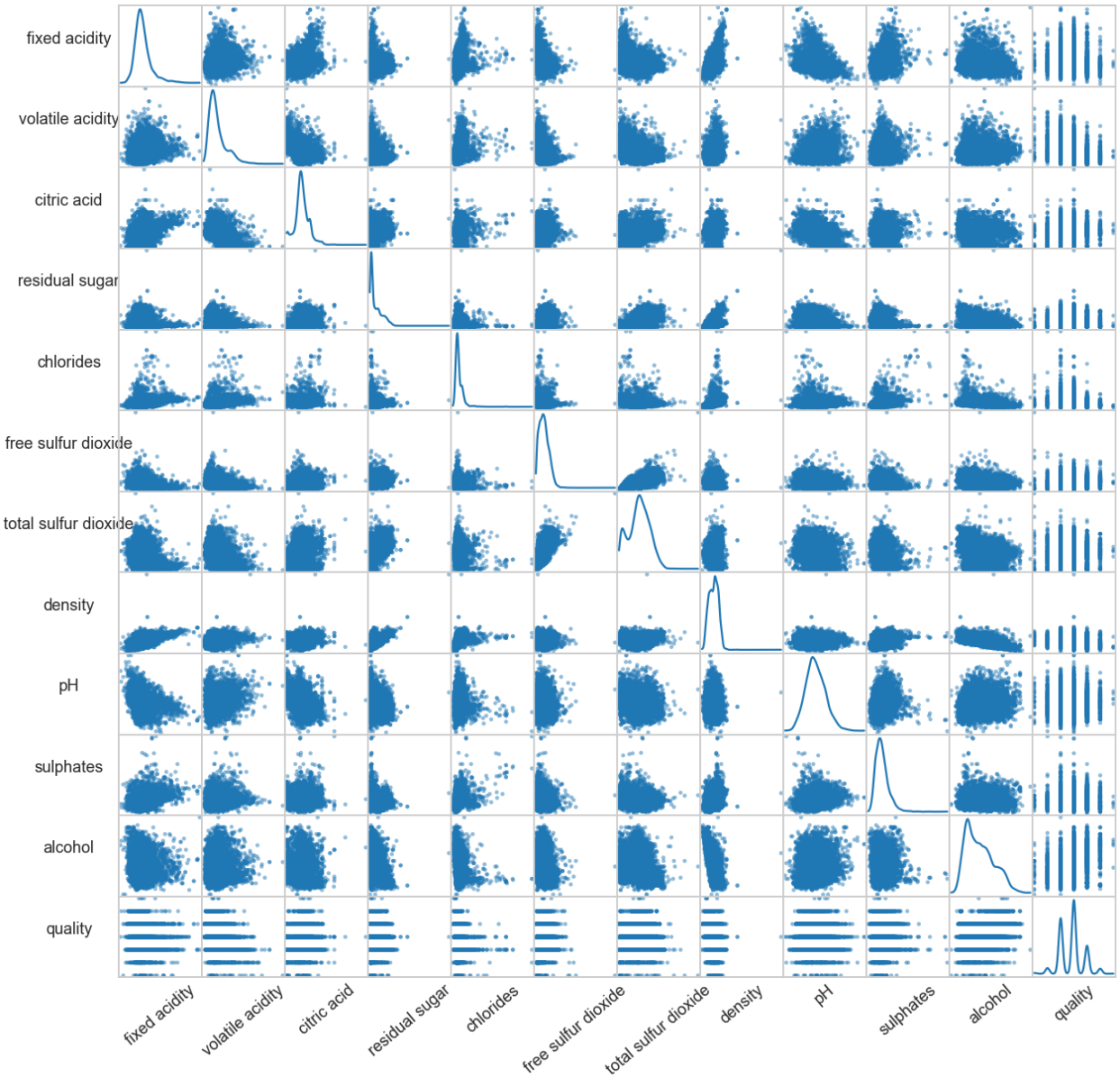
Najwyższa korelacja 0,72 występuje pomiędzy zawartością wolnego dwutlenku siarki (free sulfur dioxide), a jego całkowitą zawartością (total sulfur dioxide), jakkolwiek, te dwie wartości bardzo słabo korelują z oceną jakości (quality).

Rysunek 2 Mapa korelacji pomiędzy cechami w odniesieniu do oceny jakościowej wina (quality)



By jeszcze lepiej zobrazować zależności pomiędzy parami cech możemy przedstawić każdą cechę względem każdej innej stosując macierz rozproszoną. Wówczas można zaobserwować, że na przykład, dla pary: gęstość wina (density) i kwasowość pierwotna wina (fixed acied) zachodzi dodatnia korelacja liniowa, a co potwierdza również powyższa mapa korelacji (współczynnik korelacji wynosi 0,46). Widać również wyraźnie trend wzrostowy, a punkty nie są zbyt rozproszone między tymi cechami.

Rysunek 3 Badanie zależności par cech za pomocą macierzy rozproszonej



4. Przygotowanie danych

Oba badane zbiory danych składające się z win czerwonych i białych posiadają kompletne dane i mogą być zastosowane w dalszej analizie. Ze względu na taką samą strukturę obu zestawów danych, możemy połączyć je w jeden zbiór.

Tabela 3 Ocena jakości i kompletności danych

	types	counts	distincts	nulls	uniques
density	float64	6495	998	0	[0.9968, 0.997, 0.998, 0.9978, 0.9964, 0.9946,...
residual sugar	float64	6495	316	0	[2.6, 2.3, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1, 3.8, ...
total sulfur dioxide	float64	6495	276	0	[67.0, 54.0, 60.0, 34.0, 40.0, 59.0, 21.0, 18....
chlorides	float64	6495	214	0	[0.098, 0.092, 0.075, 0.076, 0.069, 0.065, 0.0...
volatile acidity	float64	6495	187	0	[0.88, 0.76, 0.28, 0.7, 0.66, 0.6, 0.65, 0.58,...
free sulfur dioxide	float64	6495	135	0	[25.0, 15.0, 17.0, 11.0, 13.0, 9.0, 16.0, 52.0...
sulphates	float64	6495	111	0	[0.68, 0.65, 0.58, 0.56, 0.46, 0.47, 0.57, 0.8...
alcohol	float64	6495	111	0	[9.8, 9.4, 10.0, 9.5, 10.5, 9.2, 9.9, 9.1, 9.3...
pH	float64	6495	108	0	[3.2, 3.26, 3.16, 3.51, 3.3, 3.39, 3.36, 3.35,...
fixed acidity	float64	6495	106	0	[7.8, 11.2, 7.4, 7.9, 7.3, 7.5, 6.7, 5.6, 8.9,...
citric acid	float64	6495	89	0	[0.0, 0.04, 0.56, 0.06, 0.02, 0.36, 0.08, 0.29...
quality	int64	6495	7	0	[5, 6, 7, 4, 8, 3, 9]
colour	object	6495	2	0	[R, W]

Jednocześnie zmieniamy dane kategoryczne dla koloru wina na zmienną numeryczną R=0 i W=1 i oddzielamy zmienną objaśniającą od zmiennej celu.

4.1 Podział zbioru danych

Podczas budowy modelu, którego celem jest predykcja pewnych wartości na podstawie zbioru danych uczących poważnym problemem jest ocena jakości uczenia i zdolności poprawnego przewidywania. Przeprowadzanie testów na tym samym zbiorze, na którym model był uczony, może prowadzić do przekłamania wyników. Stąd przed modelowaniem, dane dzieli się na dwa podzbiory: uczący i testowy. Podzbiór uczący będzie używany do uczenia modelu, zaś dzięki zastosowaniu podzbioru testowego sprawdzimy jak dobrze model działa. W naszym przypadku dane podzielono w proporcji 80/20: 80% to część zbioru uczącego i 20% to dane testowe.

4.2 Standaryzacja zmiennych cech

Standaryzacja jest rodzajem normalizacji danych. Proces standaryzacji jest ważny, dlatego, że analizowane dane wyrażone są w różnych wymiarach i skalach, co może prowadzić do stronniczego wyniku predykcji pod względem błędnej klasyfikacji i dokładności, stąd przed przystąpieniem do modelowania konieczne jest wyskalowanie danych.

Standaryzacja to technika skalowania, w której dane są pozbawione skali, poprzez konwersję rozkładu statystycznego danych i oznacza, że należy przekształcić dane w taki sposób, aby ich rozkład miał średnią równą 0 i odchylenie standardowe równe 1. W ten sposób cały zbiór danych skaluje się łącznie z zerową średnią i jednostkową wariancją.[2]

Standardowy wynik próbki x oblicza się jako:

$z = (x - u) / s$

gdzie u jest średnią z próbek uczących lub zerem, a s jest odchyleniem standardowym próbek uczących lub jedynką.

Centrowanie i skalowanie odbywa się niezależnie dla każdej cechy poprzez obliczanie odpowiednich statystyk na próbkach w zbiorze uczącym. Następnie zapisuje się średnią i odchylenie standardowe do wykorzystania w późniejszych danych przy użyciu transformacji.[3]

[2] https://pl.wikipedia.org/wiki/Standaryzacja_(statystyka)

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Tabela 4 Standaryzacja danych przy użyciu funkcji StandardScaler()

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	colour
2458	-0,48	-0,06	-0,54	2,22	-0,31	0,03	0,55	1,16	-0,86	-0,08	-1,07	0,57
3820	-0,25	-0,36	-0,27	1,42	-0,40	1,23	0,41	0,17	-1,17	0,18	-0,06	0,57
377	3,20	1,73	2,35	0,16	0,89	-1,39	-1,62	1,36	-0,67	3,04	2,38	-1,75
474	1,82	2,06	-0,54	-0,68	0,86	-1,45	-1,55	1,36	-0,49	0,45	-0,23	-1,75
428	-0,10	-0,18	1,04	1,71	-0,01	1,34	1,44	1,02	-0,74	-0,28	-0,99	0,57
...	,,,	,,,	,,,	,,,	,,,	,,,	,,,	,,,	,,,	,,,	,,,	,,,
4101	-0,25	0,18	-0,61	0,85	0,02	1,34	0,89	0,32	0,07	-0,75	-0,40	0,57
952	0,75	-0,66	0,35	-0,86	-0,40	0,49	0,02	-0,64	-1,42	-1,15	-0,40	0,57
537	4,36	0,06	1,18	0,08	0,27	-1,45	-1,43	2,22	-0,11	0,85	1,29	-1,75
1220	2,82	-0,12	1,39	-0,76	2,12	-0,77	-1,27	0,87	0,39	1,58	0,86	-1,75
2673	-0,02	0,06	-0,47	0,04	-0,68	-0,42	0,08	-0,46	-1,80	0,85	-0,15	0,57

5. Modelowanie

W oparciu o eksploracyjną analizę danych i analizę korelacji w części modelowania przygotowujemy modele regresji i modele klasyfikacyjne.

5.1 Regresja liniowa wielu zmiennych (wieloraka)

Regresja liniowa jest najprostszym wariantem regresji. Analiza regresji jest bardzo popularną i często stosowaną techniką statystyczną pozwalającą opisywać związki zachodzące pomiędzy zmiennymi wejściowymi (objaśniającymi) a wyjściowymi (objaśnianymi). Zakłada się tutaj, że istnieje liniowa relacja pomiędzy cechami, a zmienną którą przewidujemy.

W przypadku, kiedy chcemy zbadać wpływ wielu niezależnych $(X_1,X_2,...,X_k)$ zmiennych na jedną zmienną zależną (Y) (w naszym przypadku jest to jakość wina (quality), wykorzystuje się model liniowej regresji wielorakiej. Jest ona rozszerzeniem modeli regresji liniowej opartej o współczynnik korelacji liniowej Pearsona. Zakłada ona występowanie liniowego związku pomiędzy badanymi zmiennymi. Liniowy model regresji wielorakiej przyjmuje postać:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

gdzie,

Y - zmienna zależna, objaśniana przez model;

$X_1, X_2, \ldots, X_k$ - zmienne niezależne, objaśniające;

$\beta_0, \beta_1, \beta_2, \ldots \beta_k$ – parametry;

ε- składnik losowy (reszta modelu).

Istnieje kilka metod pomiaru jakości stworzonego modelu regresji, które opierają się na pomiarze błędu, jaki średnio popełnia dany model. W badanym przez nas modelu przyjęliśmy MSE (Mean Square Error) czyli błąd średniokwadratowy, który jest wartością oczekiwaną kwadratu błędu, czyli różnicy między estymatorem a wartością estymowaną:

$$MSE = \frac{\sum_{i=0}^{N} (\overline{y_i} - y_i)^2}{N}$$

gdzie,

N - ilość liczb;

ỹi – estymator;

yi – wartość estymowana.[4]

MSE jest miarą jakości estymatora - jest zawsze nieujemna, a wartości bliższe zeru są lepsze. Dwa lub więcej modeli można porównać przy użyciu ich MSE - jako miary tego, jak dobrze wyjaśniają one dany zbiór obserwacji. [5]

Do walidacji modelu będziemy stosować walidację krzyżową. Celem walidacji krzyżowej jest przetestowanie zdolności modelu do przewidywania nowych danych, które nie zostały użyte do jego oszacowania, aby wskazać problemy, takie jak nadmierne dopasowanie lub błąd selekcji, oraz sprawdzić, jak model uogólni się na niezależny zbiór danych (spoza testu). Jedna runda sprawdzianu krzyżowego obejmuje podział na próbki - podzbiory, wykonując analizy w jednej podgrupie (zwany zestawem szkoleniowym) i walidacji na innej podgrupie (zwanego zestawem do walidacji lub zestawem testowym). Aby zmniejszyć zmienność, w większości metod przeprowadza się wiele rund walidacji krzyżowej przy użyciu różnych partycji, a wyniki walidacji

[4] https://www.naukowiec.org/wiedza/statystyka/regresja-liniowa_765.html

[5] https://pl.qaz.wiki/wiki/Mean_squared_error

są łączone (np. uśredniane) w rundach, aby uzyskać oszacowanie wydajności predykcyjnej modelu. [6]

W opracowanym modelu przyjęliśmy, że próba dzielona jest na 10 podzbiorów.

W pierwszym badaniu dla modelu 0 przyjęliśmy, że zbadamy regresję liniową dla wszystkich cech. W wyniku uzyskaliśmy MSE na poziomie 0,543. Ocena naszego modelu pokazuje, że relacja między naszymi zmiennymi jest nieliniowa.

Tabela 5 Wynik MSE dla regresji liniowej - wszystkie cechy

Model	Szczegóły	MSE_śr_CV10	MSE_odch_std
Regresja liniowa	wszystkie cechy	0.543169	0.029866

Ponieważ analiza korelacji pokazała, że mamy zmienne niezależne silnie ze sobą skorelowane, usunęliśmy zmienne o najwyższym współczynniku (patrz Rysunek2), takie jak zawartość wolnego dwutlenku siarki (Free Sulfur Dioxide), a jego całkowita zawartość w winie (Total Sulfur Dioxide). Wartość MSE pogorszyła się i wyniosła 0,548.

Tabela 6 Wynik MSE dla regresji liniowej - bez FSD i TSD

Model	Szczegóły	MSE_śr_CV10	MSE_odch_std
Regresja liniowa	wszystkie cechy	0.543169	0.029866
Regresja liniowa	bez FSD , TSD	0.547772	0.031271

W modelu 3 zastosowaliśmy regresję metodą LASSO dla wszystkich cech. Regresja metodą LASSO (operator najmniejszej bezwzględnej redukcji i wyboru) stanowi regularyzowaną odmianę regresji linowej. Własnością tej regresji jest fakt, że dąży ona do całkowitej eliminacji wag o mniej istotnych cechach (np. poprzez zmianę ich wartości na 0). Innymi słowy, regresja tą metodą automatycznie przeprowadza dobór cech i generuje model rzadki (tj. niezawierający niewiele niezerowych wag cech). [7]

[6] Walidacja krzyżowa (statystyki) - https://pl.qaz.wiki/wiki/Cross-validation_(statistics)

[7] Aurélien Géron: "Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow", Helion 2018

W regresji tej chcielibyśmy niejako zmusić algorytm do preferowania jak najniższej wagi dla zmiennej niezależnej α dodając jej wartość absolutną, dzięki czemu model będzie starał się przy okazji minimalizowania błędu, zmniejszyć także absolutną wartość wagi.

$$\sum_{i=1}^{N} \varepsilon_i^2 + \alpha \sum_{j=1}^{P} |\beta_j| = \sum_{i=1}^{N} (\overline{y_i} - \beta_0 - \sum_{j=1}^{P} \beta_j x_{ij})^2 + \alpha \sum_{j=1}^{P} |\beta_j| = RSS + \alpha \sum_{j=1}^{P} |\beta_j|$$

W P-wymiarowym przypadku zapiszemy wektor wag jako β=(β0,β1,β2,...,βP).

$\|\beta\|_1$ to norma L1, od której bierze nazwę sama metoda regularyzacji. Może być ona rozumiana jako długość wektora i obliczamy ją za pomocą:

$$\|\beta\|_1 = \sum |\beta_i|$$

Parametr α dobiera się eksperymentalnie i na im większą wartość go ustawimy, tym mniejsze będą wagi zmiennych, ponieważ suma ich wartości absolutnych będzie coraz bardziej wpływać na wynik całego wyrażenia. [8]W naszym przypadku parametr na poziomie 0,01.

Przy wykorzystaniu tej metody wartość MSE to 0,549.

Tabela 7 Wynik MSE dla regresji liniowej - regularyzacja LASSO

Model	Szczegóły	MSE_śr_CV10	MSE_odch_std
Regresja liniowa	wszystkie cechy	0.543169	0.029866
Regresja liniowa	bez FSD , TSD	0.547772	0.031271
Regresja liniowa (Lasso)	wszystkie cechy	0.548952	0.031405

W badaniu modelu 4 została wykorzystana regresja oparta o drzewa decyzyjne.

Drzewo decyzyjne buduje modele regresji lub klasyfikacji w postaci struktury drzewiastej. Model dzieli zbiór danych na mniejsze i mniejsze podzbiory, a jednocześnie przyrostowo opracowywane jest powiązane drzewo decyzyjne. Końcowym wynikiem jest drzewo z węzłami decyzyjnymi i liśćmi. Węzeł decyzyjny ma dwie lub więcej gałęzi, z których każda reprezentuje wartości testowanej cechy. Węzeł liścia reprezentuje decyzję dotyczącą celu liczbowego. Najwyższy węzeł decyzyjny w drzewie, który odpowiada najlepszemu predykatorowi zwanemu węzłem głównym.

[8] Kacper Łukawski: „Machine Learning w Python - wprowadzenie do sztucznej inteligencji", 2020

Drzewa decyzyjne mogą obsługiwać zarówno dane jakościowe, jak i liczbowe.[9] Drzewa decyzyjne mają tendencję do nadmiernego dopasowywania się do danych o dużej liczbie cech. Uzyskanie odpowiedniego stosunku próbek do liczby cech jest ważne, ponieważ drzewo z kilkoma próbkami w dużej przestrzeni wymiarowej z dużym prawdopodobieństwem będzie nadmiernie dopasowane.[10]

W opracowanym modelu przyjęliśmy, że głębokość drzewa będzie równa 5.

Wynik MSE dla badanej metody to 0,551.

Tabela 8 Wynik MSE dla drzew decyzyjnych

Model	Szczegóły	MSE_śr_CV10	MSE_odch_std
Regresja liniowa	wszystkie cechy	0.543169	0.029866
Regresja liniowa	bez FSD , TSD	0.547772	0.031271
Regresja liniowa (Lasso)	wszystkie cechy	0.548952	0.031405
Drzewa decyzyjne	głębokość = 5	0.550576	0.033236

W ostatnim badaniu dla modelu 5 uruchamiamy lasy losowe, jako algorytm drzewa regresji używany w procesie modelowania. Pomaga to w utworzeniu losowej próby drzew decyzyjnych regresji wielorakiej i połączeniu ich w celu uzyskania bardziej stabilnej i dokładnej prognozy poprzez walidację krzyżową.

W lasach losowych każde drzewo w zespole jest budowane na podstawie próbki narysowanej z wymianą z zestawu uczącego. Co więcej, podczas dzielenia każdego węzła podczas budowy drzewa, najlepszy podział można znaleźć albo ze wszystkich funkcji wejściowych, albo z losowego podzbioru. Celem tych dwóch źródeł losowości jest zmniejszenie wariancji estymatora lasu. Rzeczywiście, poszczególne drzewa decyzyjne zazwyczaj wykazują dużą zmienność i mają tendencję do nadmiernego dopasowania. Dodatkowa losowość w lasach daje drzewa decyzyjne z nieco oddzielonymi błędami prognoz. Przyjmując średnią z tych prognoz, niektóre błędy można skasować. Losowe lasy osiągają zmniejszoną wariancję poprzez łączenie różnych drzew, czasami

[9] https://www.saedsayad.com/decision_tree_reg.htm

[10] https://scikit-learn.org/stable/modules/tree.html#tree

kosztem niewielkiego wzrostu odchylenia. W praktyce redukcja wariancji jest często znacząca, stąd daje ogólnie lepszy model.[11]

W przypadku badanego modelu została przyjęta głębokość drzewa na poziomie 5.

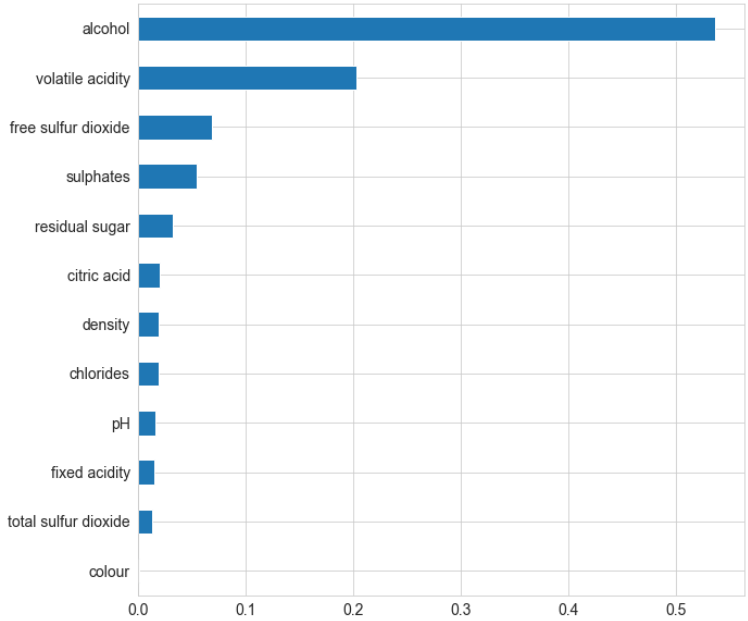
Wynik dla lasów losowych to MSE to 0,505.

Tabela 9 Wynik MSE dla lasów losowych

Model	Szczegóły	MSE_śr_CV10	MSE_odch_std
Regresja liniowa	wszystkie cechy	0.543169	0.029866
Regresja liniowa	bez FSD , TSD	0.547772	0.031271
Regresja liniowa (Lasso)	wszystkie cechy	0.548952	0.031405
Drzewa decyzyjne	głębokość = 5	0.550576	0.033236
Lasy losowe	głębokość = 5	0.505044	0.028810

Jako 3 najważniejsze cechy w modelu zostały wskazane alkohol (alcohol), kwasowość (volatile acidity) oraz zawartość wolnego dwutlenku siarki (free sulfur dioxide).

Rysunek 4 Ważności cech dla lasów losowych



[11] https://scikit-learn.org/stable/modules/ensemble.html#random-forest-parameters

5.2 Klasyfikacja

Klasyfikacja to proces przewidywania klasy danych punktów. Klasy są czasami nazywane jako cele, etykiety lub kategorie. Modelowanie predykcyjne klasyfikacji jest zadaniem aproksymacji funkcji odwzorowania (f) ze zmiennych wejściowych (X) do dyskretnych zmiennych wyjściowych (y). Klasyfikacja należy do kategorii uczenia się nadzorowanego, w której cele również mają dane wejściowe. [12] Obecnie dostępnych jest wiele algorytmów klasyfikacji, ale nie można stwierdzić, który z nich jest lepszy od drugiego. Zależy to od zastosowania i charakteru dostępnego zbioru danych. W niniejszej pracy zostaną zastosowane: drzewa decyzyjne, lasy losowe oraz maszyny wektorów nośnych (SVC).

Hiperparametr to parametr, którego wartość służy do sterowania procesem uczenia. Natomiast wartości innych parametrów (zazwyczaj wagi węzłów) są wyprowadzane w wyniku uczenia. Hiperparametry można zaklasyfikować jako hiperparametry modelu, których nie można wywnioskować podczas dopasowywania modelu do zbioru uczącego, ponieważ odnoszą się do zadania wyboru modelu lub hiperparametrów algorytmu, które w zasadzie nie mają wpływu na wydajność modelu, ale wpływają na jakość procesu uczenia się. [13] Zwykle te hiperparametry reprezentują pewne koncepcje lub „pokrętła" wysokiego poziomu, które można wykorzystać do dostrojenia modelu podczas treningu, aby poprawić jego wydajność. Najlepiej dopasowane hiperparametry dla modelu zostały wyszukane przy użyciu siatki GridSearchCV, gdzie podczas dopasowywania do zbioru danych wszystkie możliwe kombinacje wartości parametrów są oceniane i zachowywana jest najlepsza kombinacja.[14]

Macierz pomyłek (Confusion Matrix) to macierz N×N, gdzie wiersze odpowiadają poprawnym klasom decyzyjnym, a kolumny decyzjom przewidywanym przez klasyfikator. Liczba n-ij na przecięciu wiersza i oraz kolumny j to liczba przykładów z klasy i-tej, które zostały zaklasyfikowane do klasy j-tej.[15]

[12] https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[13] https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)

[14] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[15] https://algolytics.pl/jak-ocenic-jakosc-i-poprawnosc-modeli-klasyfikacyjnych-czesc-3-confusion-matrix/

Krzywa AUC (Area Under The Curve) – ROC (Receiver Operating Characteristics) jest pomiarem wydajności dla problemu klasyfikacji przy różnych ustawieniach progów. ROC jest krzywą prawdopodobieństwa, a AUC reprezentuje stopień lub miarę rozdzielności. Mówi, na ile model jest w stanie rozróżnić klasy. Im wyższa wartość AUC, tym lepszy model w przewidywaniu zera jako 0, a 1 jako 1. A kiedy AUC wynosi 0,5, oznacza to, że model nie ma żadnej klasy separacji.[16]

Wyniki klasyfikacji trzech poniższych algorytmów są oceniane w obu trybach testowych, które są poddawane 5-krotnej walidacji krzyżowej i 80% podziału zbioru danych. Ponadto niektóre standardowe miary wydajności są obliczane w celu oceny wydajności algorytmów, jak czułość (recall), precyzja (precision) , miara F1 (F1-score) i dokładność (accuracy).

W modelu opartym o drzewa decyzyjne zostały wybrane następujące parametry: głębokość drzewa – 17, funkcje podziału – 2 oraz typ klasy zbalansowanej, tzn. wykorzystuje wartości y do automatycznego dostosowywania wag odwrotnie proporcjonalnych do częstotliwości klas w danych wejściowych[17], w wyniku badania otrzymaliśmy miarę dokładności na poziomie 0,46.

Tabela 10 Miary wydajności dla drzew decyzyjnych

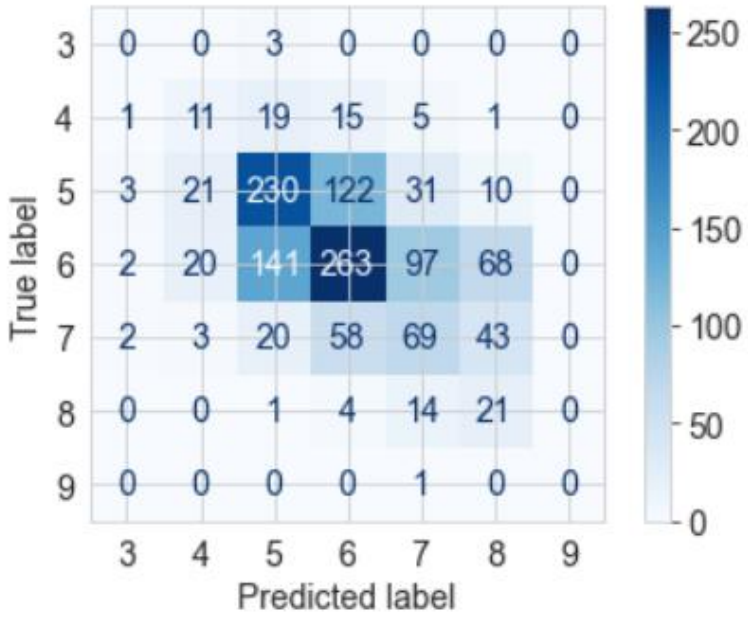
	precision	recall	f1-score	support
3	0.00	0.00	0.00	3
4	0.20	0.21	0.21	52
5	0.56	0.55	0.55	417
6	0.57	0.45	0.50	591
7	0.32	0.35	0.33	195
8	0.15	0.53	0.23	40
9	0.00	0.00	0.00	1
accuracy			0.46	1299
macro avg	0.26	0.30	0.26	1299
weighted avg	0.50	0.46	0.47	1299

Najlepiej sklasyfikowane zostały wina z oceną jakości 6 (263) oraz 5 (230), ale model ma tendencję do zawyżania oceny wina (pojawiają się oceny 7 i 8 dla oceny 6).

[16] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

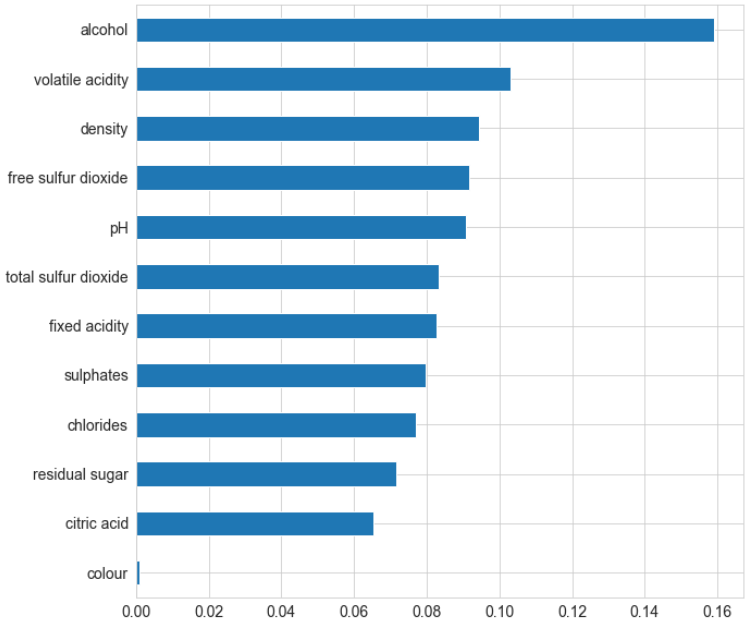
[17] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Rysunek 5 Macierz pomyłek dla drzew decyzyjnych



Jako najważniejsze cechy zostały wskazane: alkohol (alcohol), kwasowość (volatile acidity) oraz gęstość (density).

Rysunek 6 Ważności cech dla drzew decyzyjnych



Ocena ROC AUC dla drzew decyzyjnych to: 0,595.

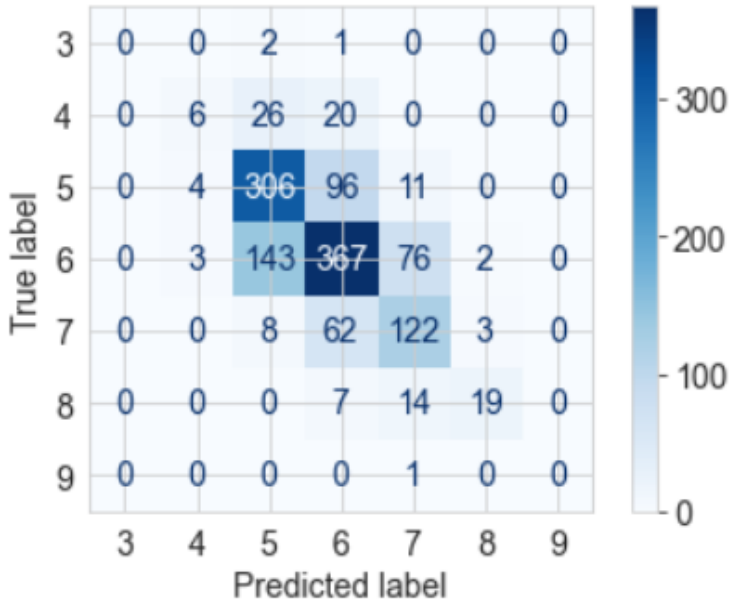
W przypadku lasów losowych zostały wybrane następujące parametry: głębokość drzewa – 12, liczba próbek wymagana do podzielenia węzła wewnętrznego – 1, liczba drzew – 140 oraz typ klasy zbalansowanej (jw.). W wyniku badania otrzymaliśmy miarę dokładności na poziomie 0,63.

Tabela 11 Miary wydajności dla lasów losowych

	precision	recall	f1-score	support
3	0.00	0.00	0.00	3
4	0.46	0.12	0.18	52
5	0.63	0.73	0.68	417
6	0.66	0.62	0.64	591
7	0.54	0.63	0.58	195
8	0.79	0.47	0.59	40
9	0.00	0.00	0.00	1
accuracy			0.63	1299
macro avg	0.44	0.37	0.38	1299
weighted avg	0.63	0.63	0.62	1299

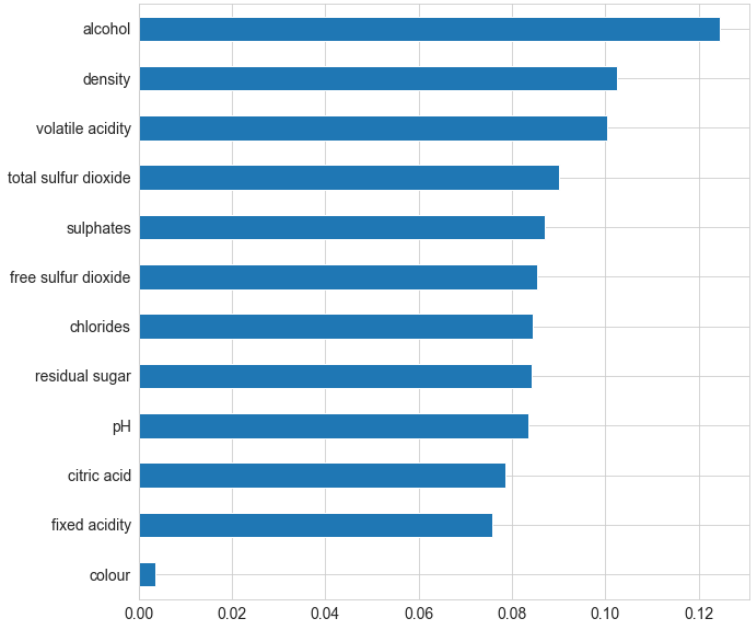
Podobnie jak w przypadku drzew decyzyjnych najlepiej sklasyfikowane zostały wina z oceną jakości 6 (367), 5 (306) oraz z wyższą skutecznością niż w przypadku drzew decyzyjnych rozpoznane zostały wina z oceną 7 (122).

Rysunek 7 Macierz pomyłek dla lasów losowych



Jako najważniejsze cechy zostały wskazane: alkohol (alcohol), gęstość (density) oraz kwasowość (volatile acidity).

Rysunek 8 Ważności cech dla lasów losowych



Ocena ROC AUC dla badanego modelu to: 0,643.

Ostatnim wykorzystanym w naszej pracy algorytmem są Maszyny wektorów nośnych (SVM) z klasyfikatorem C-Support Vector Classification (SVC).

SVM to elastyczne metody nadzorowanego uczenia maszynowego używane do klasyfikacji, regresji i wykrywania wartości odstających. SVM są bardzo wydajne w dużych przestrzeniach wymiarowych, wykorzystują podzbiór punktów szkoleniowych w funkcji decyzyjnej. Głównym celem maszyn SVM jest podzielenie zbiorów danych na klasy w celu znalezienia maksymalnej marginalnej hiperpłaszczyzny (MMH), co można wykonać w dwóch następujących krokach:

- maszyny wektorów pomocniczych będą najpierw iteracyjnie generować hiperpłaszczyzny, które w najlepszy sposób oddziela klasy;
- następnie wybierze hiperpłaszczyznę, która prawidłowo segreguje klasy.[18]

[18] https://www.tutorialspoint.com/scikit_learn/scikit_learn_support_vector_machines.htm

Klasyfikator SVC implementuje podejście „jeden na jeden” dla klasyfikacji wieloklasowej. Konstruowane są klasyfikatory n_classes * (n_classes - 1) / 2 i każdy z nich trenuje dane z dwóch klas.[19]

Przy klasyfikacji modelu za pomocą SVC zostały wybrane nastepujące parametry: parametr regularyzacji (karania) – 1,2, kernel rbf ze współczynnikiem 1,4.

W wyniku badania przy użyciu SVC otrzymaliśmy miarę dokładności na poziomie 0,66.

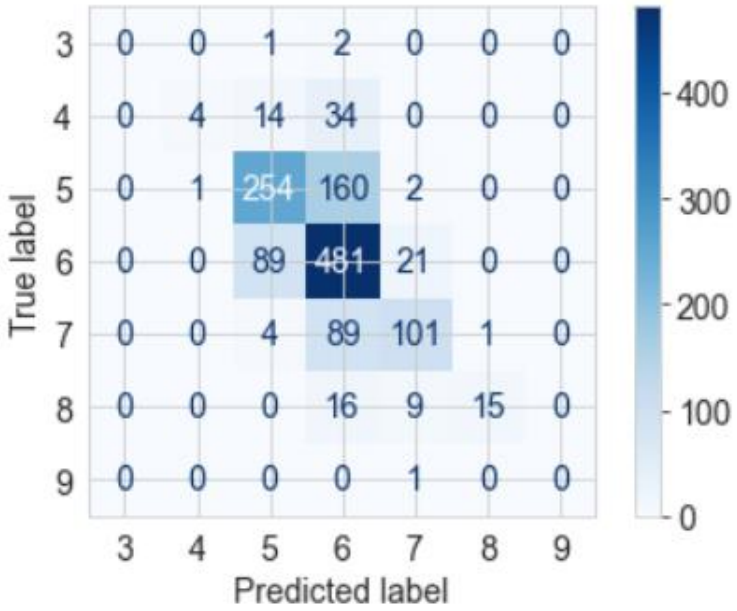
Tabela 12 Miary wydajności dla klasyfikacji SVC

	precision	recall	f1-score	support
3	0.00	0.00	0.00	3
4	0.80	0.08	0.14	52
5	0.70	0.61	0.65	417
6	0.62	0.81	0.70	591
7	0.75	0.52	0.61	195
8	0.94	0.38	0.54	40
9	0.00	0.00	0.00	1
accuracy			0.66	1299
macro avg	0.54	0.34	0.38	1299
weighted avg	0.68	0.66	0.64	1299

Po zastosowaniu tego modelu zmniejszyła się liczba błędnie sklasyfikowanych ocen dla win o ocenie 8 i 7. Najlepiej sklasyfikowane zostały wina z oceną jakości 6 (481) i 5 (254).

[19] https://scikit-learn.org/stable/modules/svm.html#mathematical-formulation

Rysunek 9 Macierz pomyłek dla klasyfikacji SVC



Ocena ROC AUC dla modelu opartego na maszyny wektorów nośnych (SVC) to: 0,630.

6. Ocena końcowa

Przechodząc przez wszystkie etapy przepływu pracy uczenia maszynowego CRISP-DM i analizując dane z testów fizykochemicznych staraliśmy się znaleźć najbardziej optymalne narzędzie wspomagające ocenę jakości wina przez ekspertów.

W tym celu na początku skorzystaliśmy z modeli najprostszych, łatwych w interpretacji. Do tego zadania zaangażowaliśmy modele regresji liniowej oraz bardziej złożone modele lasów losowych.

W przypadku regresji do oceny wydajności naszych modeli użyliśmy sprawdzianu krzyżowego (kroswalidacji) z metryką opierającą się na pomiarze błędu średniokwadratowego (MSE). Zgodnie z naszymi oczekiwaniami model, bardziej złożony, lasów losowych okazał się najlepszym z badanych modeli, z MSE na poziomie 0,51. Model regresji liniowej oraz model, których predyktory zostały skorygowane po przeprowadzonej analizy korelacji, jak również model oparty o techniki regularyzacji, nie odnotowują dużych różnic pod względem metryk wydajności. Rozsądne jest myśleć, że model lasów losowych daje nam najlepsze „prognozy". Jednak z perspektywy interpretacji, wydajności, uzyskane modele regresji liniowej możemy uznać za wartościowe. W kontekście, naszego pytania biznesowego, skupiającego się na przewidywaniu jakości win, najlepszym wyborem do oszacowania, zmiennej ilościowej będzie model lasów losowych.

W poszukiwanie najlepszego modelu w ujęciu klasyfikacji, dokonaliśmy podziału zmiennej celu na 7 klas. Uznaliśmy ten podział za naturalny z pewnymi ograniczeniami. Po pierwsze główny problem wynikał z faktu, że nasz zbiór danych jest nierównoważony (większość wartości zmiennej celu to klasa 5 i 6 co utrudniało identyfikację win wysokiej jakości i niskiej). Po drugie złożone modele mogą przez to nadmiernie dopasować dane, tracąc zdolność do generalizowania. Przyjęliśmy proces przewidywania klas budując modele drzew decyzyjnych, lasów losowych i maszyn wektorów nośnych. Testowane modele dokonywały mocnego przeuczania na danych treningowych, stąd ograniczyliśmy nasze modele wykorzystując metodę przeszukiwania siatki i wyboru najlepszego hiperparametu. Dodatkowo używaliśmy walidacji krzyżowej z oceną dokładności (accuracy).

Narzędziem, które posłużyło nam do oceny i porównania trafności klasyfikacji był wskaźnik pola krzywej ROC AUC. Najsłabszym zbadanym modelem był model oparty na drzewach decyzyjnych, który został oceniony na poziomie 0,60. Modelem najlepszym są lasy losowe z oceną 0,64, pomimo, niższego poziomu dokładności 0,63 niż ostatni model SVM, który miał najwyższą dokładność z pośród badanych modeli 0,66 i ROC AUC równym 0,63.

Tabela 13 Ocena modeli klasyfikacji

Model	Accuracy	Precision	Recall	F1 score	ROC AUC
Drzewa decyzyjne	0,45	0,26	0,30	0,26	0,60
Lasy losowe	0,64	0,44	0,37	0,38	0,64
SVM	0,66	0,54	0,34	0,38	0,63

W klasyfikacji oceny jakości zauważyliśmy, że proste modele nie byłyby w stanie uchwycić niuansów zmienności danych, zwłaszcza win wysokiej jakości, które mają niewiele rekordów w porównaniu z innymi klasami. Wnioskujemy, że można przewidzieć jakość wina na podstawie atrybutów fizykochemicznych, potrzebujemy jednak bardziej zrównoważonych danych. Warto również wspomnieć, że w zbiorze posiadamy tylko 12 cech, które mogą zawęzić dokładność naszego przewidywania. Rozwiązaniem jest rozszerzenie danych o dodatkowe cechy takie jak np.: rok zbioru, odmiana, rodzaj winogron.

Spis tabel

Spis rysunków