Projekt Zespołowy	
Kierunek *Automatyka i Robortyka*	**Termin** *Środa 13:00-16:00*
Wykonał *241165 Daniel Jablonski*	**Temat** *Identyfikacja win*
Prowadzący *Dr inż. Krzysztof Halawa*	**data** *16 czerwca 2020*

Raport

1 Identyfikacja wina na podstawie parametrów

1.1 Cel

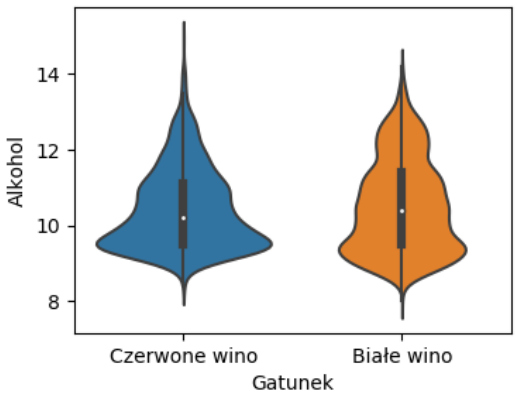
W ramach projektu podjęto się próby stworzenia programu wspomagającego identyfikację win na wino czerwone i wino białe na podstawie danych zapewnionych przez UCI.

1.2 Analiza zestawu danych

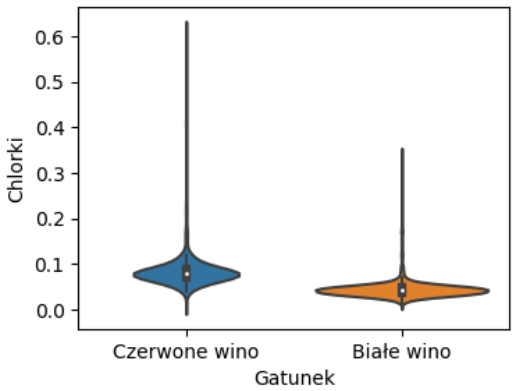
Dane były podzielone na osobne pliki csv dla win czerwonych oraz białych. Dane były docelowo przygotowane dla nauki klasyfikacji na podstawie oceny wina. Jednak dla naszych celów kolumnę z jakością zastąpiono kolumną z gatunkiem wina. Każde wino składało się z 11 niezależnych własności takich jak na przykład: zawartość cukru, siarczanów, gęstość czy poziom alkoholu. Dla żadnego wina jakakoliwek cecha była pusta.

1.3 Badanie eksploracyjne zestawu danych

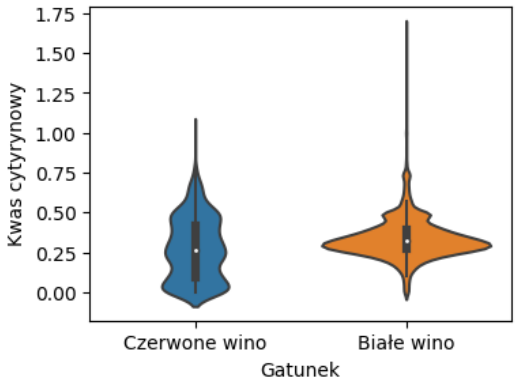
Na początku zbadano analitycznie czy występuje cecha ze zbioru danych, która kategorycznie odróżnia wino czerwone od białego.Do prównania cech wina dla wina czerwonego i wina białego zastosowano wykres skrzypcowy. Który przedstawia zakres wartości dla danej cechy gatunku wina jak również kwantyle co jest przydatne przy wstępnej analizie.



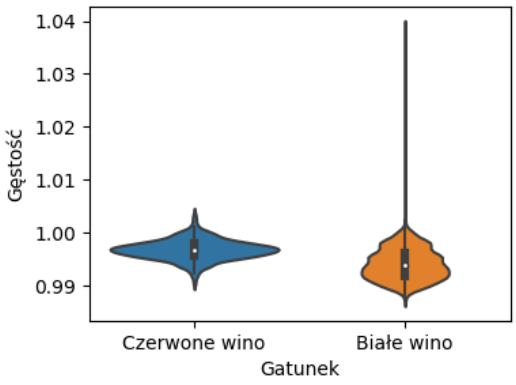
(a) Alkohol



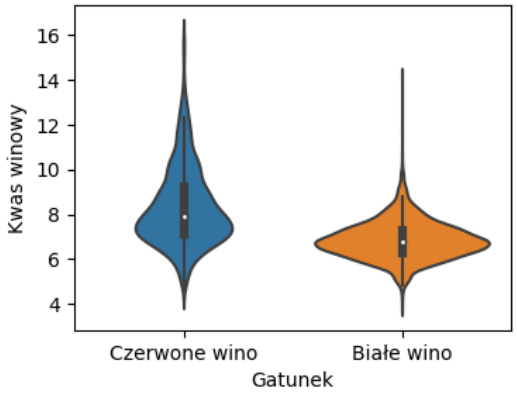
(b) Chlorki



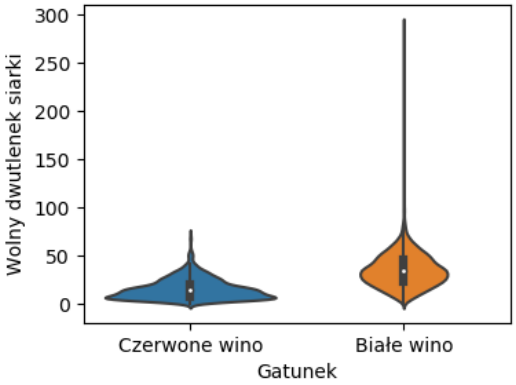
(c) Kwas cytrynowy



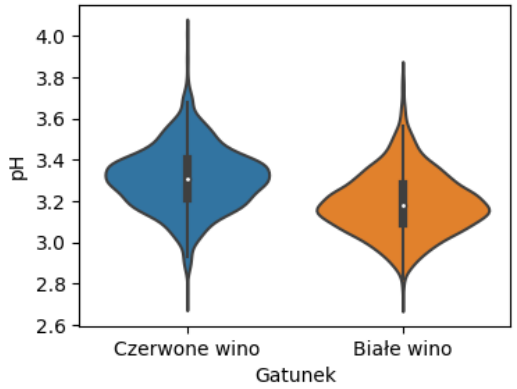
(d) Gęstość



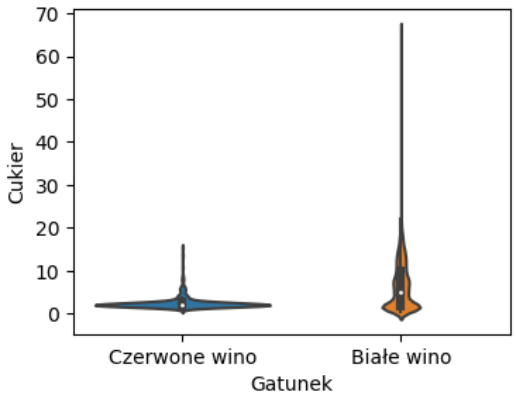
(e) Kwas winowy



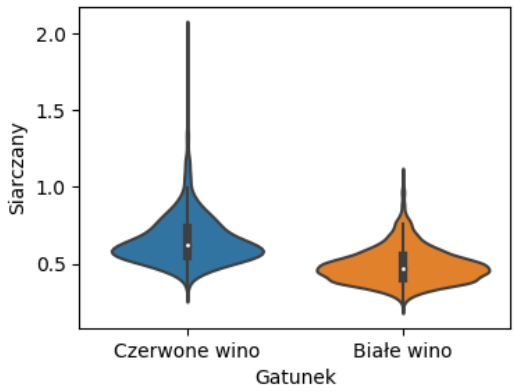
(f) Wolny dwutlenek siarki



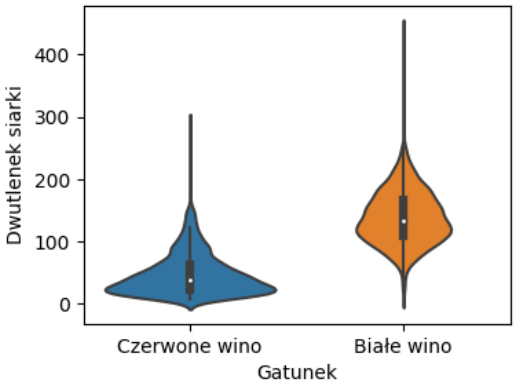
(g) pH



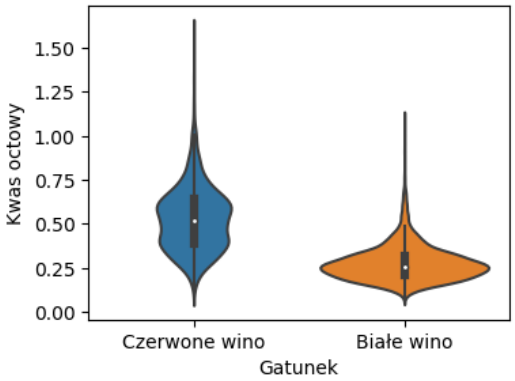
(h) Cukier



(i) Siarczany



(j) Dwutlenek siarki



(k) Kwas octowy

Rysunek 1: Porównanie cech dla win czerwonych i białych

1.4 Wstępna analiza

Po przeanalizowaniu wykresów porównawczych można zauważyć cechy, które różnią wino czerwonego od wina białego. Wino czerwone ma mniejszą zawartość dwutlenka siarki, większą zawartość chlorków. Wina czerwone są najczęściej winami wytawnymi(mniejsza zawartość cukru) co pokrywa się z wykresem. Jednak, żadna z cech definitywnie określa czy wino jest białe czy czerwone. Występują różnice w dystrybucji jednak zakres wartości jest zbliżony.

1.5 Klasyfikacja win

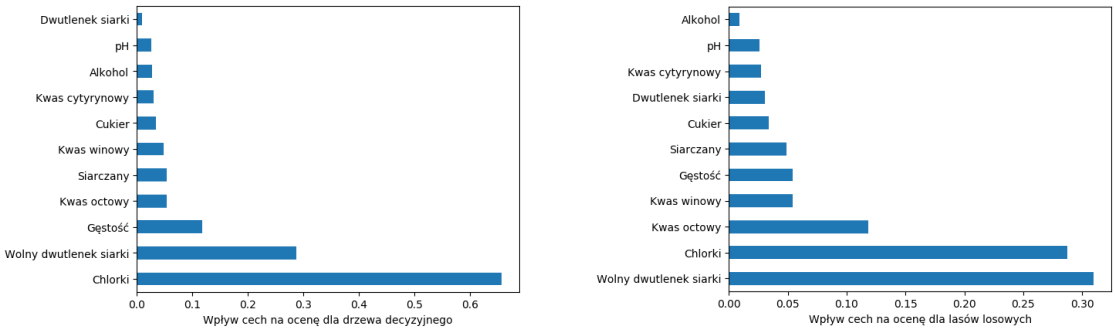
Metoda klasyfikacji polega na znajdowaniu odwzorowania danych w zbiórze predefiniowanych klas. Na podstawie zawartości bazy danych budowany jest model(np.drzewo decyzyjne), który służy do klasyfikowania nowych obiektów w zbiorze danych lub głębszego zrozumienia istniejącego podziału obiektów na predefiniowane klasy. U nas predefiniowane klasy to: wino białe i czerwone. Zestaw danych uczących i testowych osiągnięto za pomocą funkcji z biblioteki sklearn(train_test_split), dla której można zdefiniować proporcję zbioru testowego a uczącego oraz poziom przetasowania danych.
Dla klasyfikacji win wykorzystano bibliotekę keraz oraz scikit-learn.
Klasyfikatory z biblioteki scikit-learn:

- KNeighborsClassifier - jest typowym przykładem leniwego klasyfikatora. Nie uczy się on funkcji dyskryminacyjnej, lecz stara się "zapamiętać" cały zbiór próbek. Podejmuje decyzję na podstawie wybranej metryki odległości.
- DecisionTreeClassifier - Model drzew decyzyjnych, klasyfikator danych podejmujący decyzję na podstawie szeregu odpowiedzi.
- RandomForestClassifier - lasy losowe można rozumieć jako zespół drzew decyzyjnych. Metoda polega na uśrednianiu wielu(wysokich) drzew decyzyjnych, które osobno cechują się znaczną wariancją i łączeniu ich w jeden skuteczny model.
- SGDClassifier - klasyfikator liniowy zoptymalizowany przez SGD czyli przez stochastyczny spadek wzdłuż gradientu.
- MLPClassifier - Percepton wielowartstwowy jest to najpopularniejszy rodzaj sztucznych sieci neuronowych. Można w nim manipulować ilością warstw ukrytych oraz liczbą neuronów w każdej warstwie. Jednym z danych niezbędnych do dostarczenia podczas tworzenia instancji MLPClassifier jest optymalizator - użyto najczęściej stosowanego i najbardziej efektywnego dla większości zastosowań stochastycznego optymalizatora gradientowego ADAM . Oblicza on współczynniki uczenia się dla różnych parametrów.

Klasyfikatory bardzo dobrze sprawdziły się do identyfikacji gatunku wina.

Klasyfikator	Dokładność
K najbliższych sąsiadów	93%
Lasy losowe	99%
Drzewo decyzyjne	98%
Stochastyczny spadek wzdłuż gradientu	95%
Percepton wielowarstwowy	98%

Dla klasyfikatorów: lasy losowe oraz drzewo decyzyjne zbadano, która cecha najbardziej wpłyneła na decyzję.



(a) Drzewo decyzyjne

(b) Lasy losowe

Rysunek 2: Wpływ cech na decyzję

Biblioteki keras użyto do stworzenia własnej sekwencyjnej sieci wielowarstwowej.

Listing 1: Tworzenie sieci

```
model = Sequential()
model.add(Dense(8, input_dim=11, activation='relu'))
model.add(Dense(6, activation='relu'))
model.add(Dense(4, activation='relu'))
model.add(Dense(classifications, activation='sigmoid'))
```

Dana sieć po 200 epokach osiągneła dokładność 98.61% .

1.6 Wnioski

Problem nie był wyzwaniem dla sieci. Klasyfikacja polegała na binarnym przypisaniu wina na podstawie cech do jednej z dwóch kategorii(0-wino białe, 1-wino czerwone). Jest to analogiczny problem do klasyfikacji irysów. Jednak w naszym przypadku mieliśmy większą ilość niezależnych cech a dwie klasy klasyfikacji, w porównaniu do 4 cech i 3 grup klasyfikacji dla irysów. Dla naszego zastosowania skuteczne okazały się klasyfikatory oparte na drzewach decyzyjnych, dla których przeprowadzono dodatkową analizę która z cech najbardziej wpływa na identyfikację wina.Najgorzej sprawdził się model korzystający z algorytmu K najbliższych sąsiadów czyli algorytm leniwy. Najmniej znaczącymi cechami dla identyfikacji był wskaźnik pH i zawartość alkoholu. Co było by zgodne z definicją wina jako napoju alkoholowego otrzymanego w wyniku fermentacji winogron. Wino białe i czerwone nie różni się pod tym względem. Jednak w wyniku innego procesu produkcji cechami wyróżniającymi są: zawartość chlorków czy wolnych dwutlenków siarki.

2 Przetwarzanie języka naturalnego.

2.1 Cel

Po zapoznaniu się z podstawami uczenia maszynowego, zwrócono się do identyfikacji win a dokładnie szczepu na podstawie recenzji. W zadaniu wykorzystano zbiór danych: Kaggle/wine-reviews Jednakże, w wyniku dużej ilości unikalnych szczepów, zadanie zmieniono na przetwarzanie recenzji i analiza win na wina: dobre, neutralne, złe na podstawie recenzji.

2.2 Analiza zbioru danych

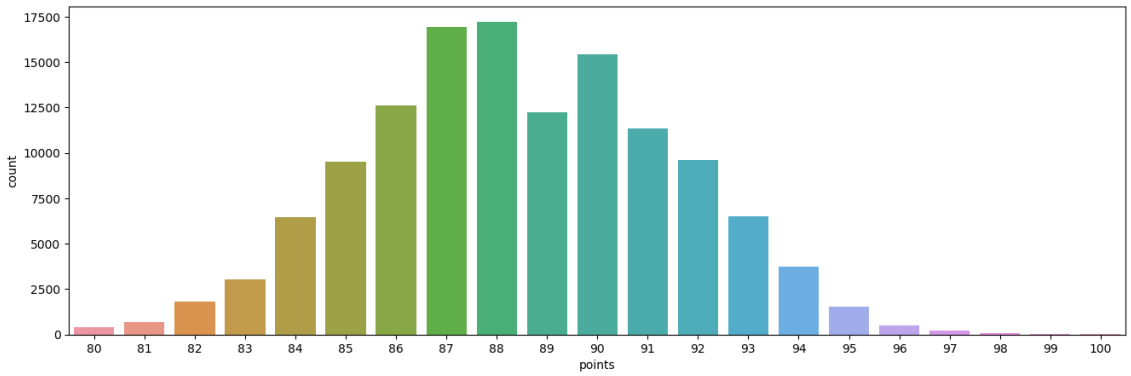
Do naszych celów wykorzystano ”winemag-data-130k-v2.csv” czyli zbiór danych wydobyty ze strony winemag.com. Przed przystąpieniem do dalszej analizy zbadano zawartość zbioru.

Dataset Shape: (129971, 14)

	Name	dtypes	Missing	Uniques
0	Unnamed: 0	int64	0	129971
1	country	object	63	43
2	description	object	0	119955
3	designation	object	37465	37979
4	points	int64	0	21
5	price	float64	8996	390
6	province	object	63	425
7	region_1	object	21247	1229
8	region_2	object	79460	17
9	taster_name	object	26244	19
10	taster_twitter_handle	object	31213	15
11	title	object	0	118840
12	variety	object	1	707
13	winery	object	0	16757

Rysunek 3: Podsumowanie zbioru danych

Jak widzimy liczba różnorodnych szczepów wynosi 707 co sprawiło spory problem, dlatego postanowiono wykorzystać do uczenia sieci opis('description') oraz punkty('points'), które zostały podzielone na 3 kategorie: wina dobre, wina neutralne, wina złe.



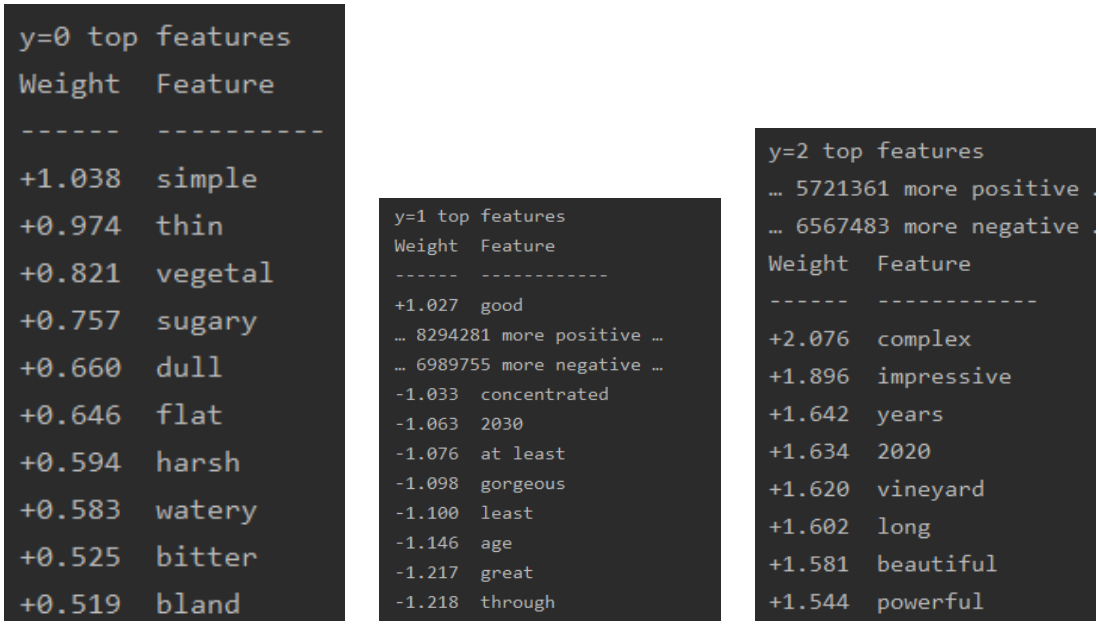
Rysunek 4: Histogram ocen

Na podstawie rozkładu ocen przyjęto:

- Wina dobre - Ocena powyżej 90 punktów
- Wina neutralne - Ocena 85-90 punktów
- Wina złe - Ocena poniżej 85 punktów

2.3 Analiza sentymentów

Dla każdego wina dodano kolumnę kategoria('category'), która no podstawie ilości punktów przypisywała do kategorii ustalonych wcześniej na podstawie histogramu. Następnie przekształcono recenzje w wektor cech oraz wyznaczenie tf-idf czyli ważenie częstości terminów - odwrotna częstość w tekście. Oraz użyto klasyfikatora liniowego z wykorzystującego SGD. Uzyskana przez nas dokładność wyniosła 76-78%



y=0 top features

Weight	Feature
+1.038	simple
+0.974	thin
+0.821	vegetal
+0.757	sugary
+0.660	dull
+0.646	flat
+0.594	harsh
+0.583	watery
+0.525	bitter
+0.519	bland

y=1 top features

Weight	Feature
+1.027	good
… 8294281 more positive …	
… 6989755 more negative …	
-1.033	concentrated
-1.063	2030
-1.076	at least
-1.098	gorgeous
-1.100	least
-1.146	age
-1.217	great
-1.218	through

y=2 top features

… 5721361 more positive

… 6567483 more negative

Weight	Feature
+2.076	complex
+1.896	impressive
+1.642	years
+1.634	2020
+1.620	vineyard
+1.602	long
+1.581	beautiful
+1.544	powerful

(a) Wino złe (b) Wino neutralne (c) Wino dobre

Rysunek 5: Waga słów dla opisów win

2.4 Wnioski

Uzyskana przez nas dokładność 77.3% jest zadowalająca, przy prawie 130 tysiącach recenzji win, ponad 100 tysięcy win zostało poprawie skatalogowane jako wino: dobre, neutralne, złe. Również analiza wektora wag dla słów potwierdza skuteczne działanie. Wina złe opisano jako: proste, mdłe, wodniste, gorzkie. Stwierdzenia, które są powszechnie używane do opisu win miernej jakości. Wina dobre opisano jako: złożone, potężne, zachwycające, piękne. Co pokrywa się z opisami stosowanymi do win o wybitnej jakości. Możliwym rozwinięciem zadania analizy sentymentów mogło by być wykorzystanie z biblioteki nltk i funkcji SentimentAnalyzer. Możliwym sposobem na identyfikację szczepu na podstawie recenzji mogłoby być wykorzystanie LDA - Latent Dirichlet Allocation, alokacja ukrytej zmiennej Dirichleta, która została poruszona w książce "Python. Uczenie maszynowe, rozdział 8".

3 Materiały pomocnicze

- Rashka S., Mirjalili "Python. Uczenie Maszynowe"
- Keras API documentation
- Scikit-learn User Guide
- ELI5 documentation
- https://www.kaggle.com/zynicide/wine-reviews
- https://archive.ics.uci.edu/ml/datasets/wine+quality
- Stack Overflow
- Repozytorium