Magdalena Thomas
Politechnika Gdańska,
Fizyka Techniczna, Informatyka Stosowana sem. 3
nr 155998

Analiza i modelowanie danych z wykorzystaniem modelu regresji liniowej oraz modelu wielomianowego

Obiektem przeprowadzonych badań są dane otrzymane z analizy chemicznej win uprawianych w tym samym rejonu Włoch, ale pochodzących z trzech różnych odmian. Analiza wyodrębniła 13 odmiennych składników, których wartości zapisane zostały w pliku z rozszerzeniem csv. Analizę oraz modelowanie danych wykonano przy użyciu programów napisanych w języku Python.

Cel badań:

- ogólna analiza podanych wartości tzn. wyznaczenie wartości minimalnej, maksymalnej, średniej, odchylenia standardowego dla każdego składnika
- obliczenie oraz zobrazowanie regresji liniowej i wielomianowej czwartego stopnia dla badanych komponentów
- zastosowanie modelu regresji liniowej oraz modelu wielomianowego zarówno dla całego zbioru danych jak i dla wybranych z niego wartości

Omawiane dane zostały zaczerpnięte ze strony internetowej dnia 10.06.2019r. *https://archive.ics.uci.edu/ml/datasets/Wine*. Składają się z 14 kolumn oraz 178 wierszy. Kolumny reprezentują wartości poszczególnych składników wykrytych podczas analizy chemicznej w winach m.in.: alkohol, kwas jabłkowy, magnez, fenole, flawaonidy. Zbadano również odcień oraz intensywność koloru. Kolejne wiersze to wyniki analiz win o jednej z trzech odmian.

Podczas analizy omawianych zbiorów powstały trzy pliki z programami, w których realizowano założone cele. Plik *wine.py* zawiera analizę ogólną, na początku której wczytano dane do przygotowanych wcześniej tablic. Następnie wykonano obliczenia, których przykładowe wyniki zaprezentowano w tabeli poniżej.

Zmienna: *alcohol* **MIN:** *11.03* **MAX:** *14.83* **ŚREDNIA:** *13.00061797752809* **MEDIANA:** *13.05* **ODCHYLENIE STANDARDOWE:** *0.8095429145285168*	**Zmienna:** *malic_acid* **MIN:** *0.74* **MAX:** *5.8* **ŚREDNIA:** *2.3363483146067416* **MEDIANA:** *1.8650000000000002* **ODCHYLENIE STANDARDOWE:** *1.1140036269797893*	**Zmienna:** *magnesium* **MIN:** *70.0* **MAX:** *162.0* **ŚREDNIA:** *99.74157303370787* **MEDIANA:** *98.0* **ODCHYLENIE STANDARDOWE:** *14.242307673359806*

Tabela 1. Przykładowe wyniki analizy ogólnej z programu *wine.py*

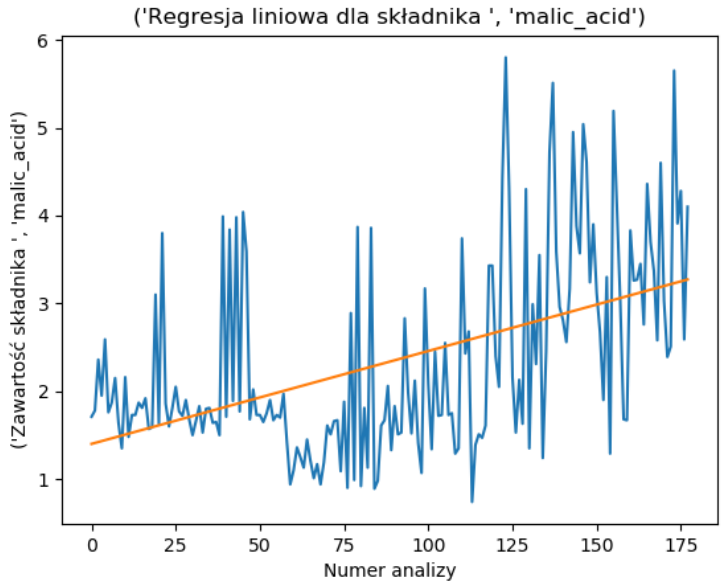
Zauważono, iż analizie chemicznej podawano wina zgodnie z ich odmianą, dlatego w kolejnym kroku wartości każdego ze składników podzielono na 3 części. Obliczono wartości średnie w każdym

podzbiorze, a otrzymaną liczbę porównano z wartością średnią uzyskaną na podstawie całego zbioru danej kolumny. Wynik programu przedstawiono na obrazie nr 1.

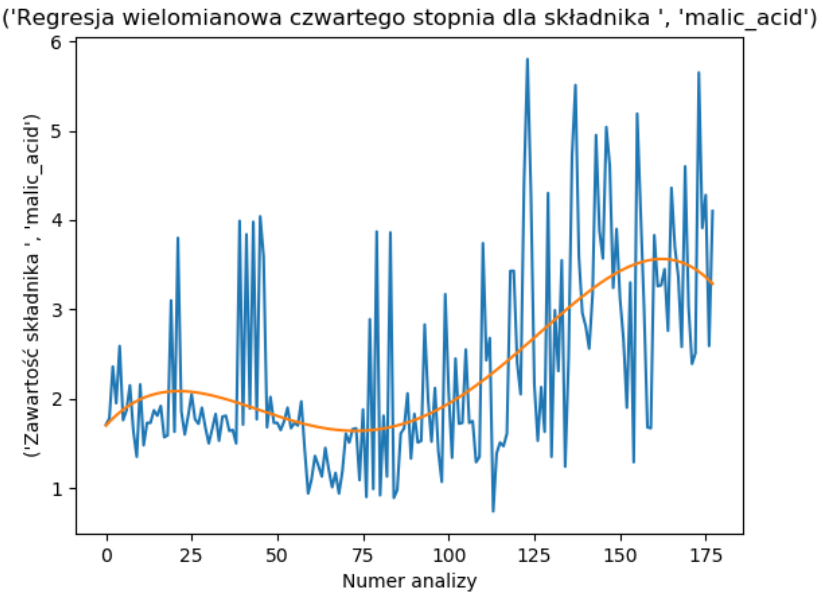
```
Składnik: alcohol
Dla: alcohol najbardziej zbliżoną do obliczonej wartość średnią ma gatunek pierwszy

Składnik: malic_acid
Dla: malic_acid najbardziej zbliżoną do obliczonej wartość średnią ma gatunek trzeci

Składnik: magnesium
Dla: magnesium najbardziej zbliżoną do obliczonej wartość średnią ma gatunek pierwszy

Składnik: total_phenols
Dla: total_phenols najbardziej zbliżoną do obliczonej wartość średnią ma gatunek pierwszy

Składnik: flavanoids
Dla: flavanoids najbardziej zbliżoną do obliczonej wartość średnią ma gatunek pierwszy

Składnik: nonflavanoid_phenols
Dla: nonflavanoid_phenols najbardziej zbliżoną do obliczonej wartość średnią ma gatunek trzeci

Składnik: color_intensity
Dla: color_intensity najbardziej zbliżoną do obliczonej wartość średnią ma gatunek trzeci

Składnik: hue
Dla: hue najbardziej zbliżoną do obliczonej wartość średnią ma gatunek drugi

Składnik: diluted_wines
Dla: diluted_wines najbardziej zbliżoną do obliczonej wartość średnią ma gatunek pierwszy
```

Obraz nr 1. Obliczanie wartości średniej w programie wine.py

Na końcu analizy ogólnej zbioru obliczono regresję liniową i wielomianową czwartego stopnia dla każdego składnika. Wyniki przedstawiono na wykresach, z których można wnioskować, iż wyższy stopień regresji jest równoznaczny z lepszym dopasowaniem. Poniżej przykładowe wykresy dla zmiennej malic_acid (kwas jabłkowy).



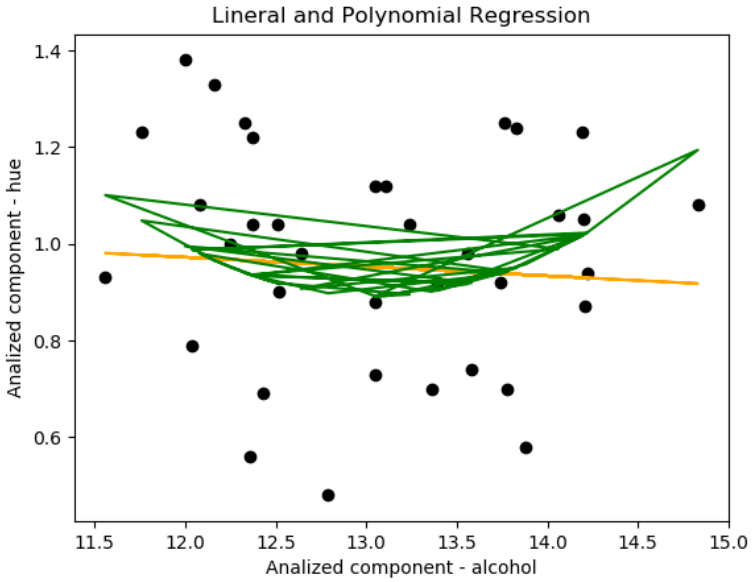
Obraz nr 2. Wykres regresji liniowej kwasu jabłkowego otrzymany w wyniku programu *wine.py*
Wzór prostej: y(x) = 0.01057193681650422 * x + 1.400731906346118



Obraz nr 3. Wykres regresji wielomianowej czwartego stopnia kwasu jabłkowego, otrzymany w wyniku programu *wine.py*
Wzór prostej: y(x) = -4.20062406220545e-08 * x + 1.4348138930077392e-05

W plikach *regression_1.py* oraz *regression_2.py* do analizy danych zastosowano model regresji liniowej oraz model wielomianowy. Narzędzi tych używa się do modelowania związku pomiędzy zmiennymi i przewidywania nieznanych wartości jednej zmiennej na podstawie znajomości innych. W regresji liniowej zależności te opisywane są za pomocą funkcji liniowej, natomiast w modelu wielomianowym przy użyciu wielomianu. Regresje wielomianowe wykorzystuje przy pracy z komponentami, których zależności są nieliniowe. W zwykłej wielokrotnej analizie regresji liniowej zakłada się, że wszystkie zmienne niezależne są niezależne. Natomiast w modelu regresji wielomianowej założenie to nie jest spełnione.

Programy różnią się przygotowaniem danych do analizy. W *regression_1.py* po wczytaniu, dane zostały podzielone na próbę uczącą się (80%) i testową (20%). Nierówny podział wynika z faktu, iż podczas analizy istotne było, aby model posiadał dobre zdolności przewidywana wartości. Badaniu poddane zostały wartości z kolumny ‘alcohol’ oraz ‘hue’. Otrzymane wartości zaprezentowano na wykresach zapisanych w formacie PDF oraz obliczono miary błędów dopasowania.
Zaprezentowano MAE (Mean Absolute Error) – średni błąd bezwzględny, informujący o ile średnio będzie wynosić odchylenie przewidywanej wartości od rzeczywistej oraz współczynnik determinacji R2 określający jak bardzo zmiany jakiejś wartości są determinowane zmianami w zakresie innej cechy.



Obraz nr 4. Porównanie modelu regresji liniowej oraz wielomianowej drugiego stopnia dla zależności alkoholu od odcienia wina. Wynik programu *regression_1.py*

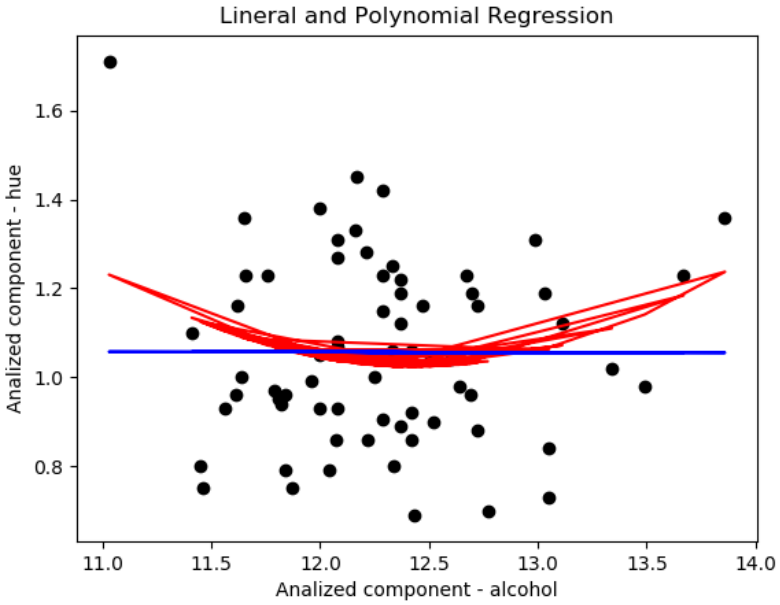
Na obrazie nr 4. widać zależności między alkoholem a jego odcieniem w całym zbiorze danych (dla wszystkich gatunków). Kolorem pomarańczowym zaznaczona jest regresja liniowa, natomiast zielonym wielomianowa.

Otrzymane wartości miary dopasowania dla modelu regresji liniowej:
MAE of linear regression: 0.18240669392776043
R2 of linear regression: -0.0009768217331300733

Otrzymane wartości miary dopasowania dla modelu wielomianowego:
MAE of polynomial regression: 0.18130160035880222
R2 of polynomial regression: 0.05736976918650161

Analizując powyższe wartości zaobserwowano, iż średni bezwzględny błąd ma wartości porównywalne dla obu modeli, natomiast model wielomianowy charakteryzuje się wyższym współczynnikiem determinacji.

Jak wspomniano wcześniej, plik *regression_2.py* również zawiera modelowanie danych za pomocą regresji liniowej oraz wielomianowej drugiego stopnia. Po wczytaniu, na zbiorze danych została wywołana funkcja *head()* oraz *info()* w celu poznania dokładnej struktury oraz sposobu przedstawienia danych. Dalej zmienną *'Cultivar'* zamieniono na zmienną kategoryczną na podstawie której zliczono ilość wierszy danych (przeprowadzanych analiz) według numeru odmiany. Największą grupę, która została poddana analizie stanowiły wina o odmianie numer 2, dlatego na jego podstawie przygotowano zbiory do zbadania zależności między alkoholem a odcieniem wina.
Wyniki obliczeń przedstawiono na wykresach, natomiast zestawienie modelu regresji liniowej oraz wielomianowej zaprezentowano na wykresie zapisanym w formacie PDF i przedstawionym poniżej. Wyznaczono błędy miary dopasowania.



Obraz nr 5. Porównanie regresji liniowej oraz wielomianowej drugiego stopnia dla zależności alkoholu i odcienia win odmiany numer 2. Wynik programu *regression_2.py*

Powyższy wykres prezentuje regresję liniową oraz wielomianową dla win o odmianie numer 2 w funkcji alkoholu i odcienia wina. Prosta niebieska została utworzona podczas modelowania przy użyciu regresji liniowej, natomiast krzywa czerwona podczas wielomianowej.

Otrzymane wartości miary dopasowania dla modelu regresji liniowej:
MAE of linear regression: 0.1636626543288099
R2 of linear regression: 4.153478273982714e-06

Otrzymane wartości miary dopasowania dla modelu wielomianowego:
MAE of polynomial regression: 0.16284514485002538
R2 of polynomial regression: 4.153478273982714e-06

Przedstawione powyżej wartości odnoszą się do omawianego zbioru danych. Średni bezwzględny błąd modelu wielomianowego ma wartość niewiele mniejszą od modelu regresji liniowej. Natomiast współczynniki determinacji mają identyczną wartość. Wydawać się może to mało prawdopodobne, jednak jest prawdziwe.

W raporcie przedstawiono analizę oraz modelowanie danych z analizy chemicznej włoskich win. Obliczenia oraz trzy pliki programu (*wine.py*, *regression_1.py*, *regression_2.py*) napisane zostały w języku Python. Dane, które poddano analizie zapisane były w pliku z rozszerzeniem csv. W pierwszym programie wykonano podstawowe obliczenia oraz obliczono regresję dla poszczególnych składników. W pozostałych programach odpowiednio przygotowane dane (m.in. podział na zbiory) zostały poddane modelowaniu. W obu programach większym średnim błędem bezwzględnym opatrzony jest model regresji liniowej, jednak różnica wartości jest minimalna. Modelowanie na całym zbiorze wartości (program *regression_1.py*) pozwoliło uzyskać większy współczynnik determinacji dla modelu

wielomianowego. Oznacza to, iż jest to model lepiej przewidujący wartości od modelu regresji liniowej. Wartości które poddane zostały analizie miały zbyt zbliżone wartości, dlatego nie można jednoznacznie stwierdzić który z nich został lepiej dopasowany.

Literatura:

https://archive.ics.uci.edu/ml/datasets/Wine
https://towardsdatascience.com/polynomial-regression-bbe8b9d97491
https://medium.com/coinmonks/polynomial-regression-11bec9262d64
http://visualmonsters.cba.pl/index.php/prognozowanie/blad-e-blad-procentowy-ep-sredni-blad-me-sredni-procentowy-blad-mpe-sredni-blad-bezwzgledny-mae-sredni-bezwzgledny-blad-procentowy-mape/
https://matematyka.poznan.pl/artykul/regresja-liniowa-czyli-o-zastosowaniu-funkcji-liniowej-w-analizie-statystycznej/
http://pogotowiestatystyczne.pl/slowniczek/wspolczynnik-determinacji/