UNIWERSYTET WSB MERITO WROCŁAW	
Laboratorium z Podstaw Uczenia Maszynowego	
Ćwiczenie nr 2	DATA: 06.06.2023
Temat: Regresja liniowa i logistyczna	
Nazwisko i imię	Kacper Studenny
Nr indeksu	81573
Grupa:	2
Prowadzący	*Prof. dr hab. Urszula Markowska-Kaczmar*

Regresja Logistyczna

1 Opis danych

Zbiór posiada dziesięć klas, zbiór liczb (0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
Ogólna liczba próbek w zbiorze wynosi 70000
Liczba próbek w każdej klasie:

1.0	7877
7.0	7293
3.0	7141
2.0	6990
9.0	6958
0.0	6903
6.0	6876
8.0	6825
4.0	6824
5.0	6313

Jedną z istotnych informacji wynikających z analizy zbioru jest fakt, że zbiór jest dość dobrze zbalansowany, oznacza to że liczność przykładów w poszczególnych klasach jest podobna lub zbliżona. Istnieje kilka powodów, dla których zbalansowany zbiór danych jest istotny:

-Uniknięcie przewagi jednej klasy.

-Błędne estymacje.

-Efektywność algorytmów.

-Ocena modelu.

Oprócz zbalansowania danych zbioru uczącego, istnieje kilka innych ważnych czynników, które można uwzględnić przy ocenie zbioru MNIST:

-Rozmiar zbioru danych: Zbiór MNIST zawiera duży zestaw obrazów o rozmiarze 28x28 pikseli, co daje łącznie 784 cechy dla każdego obrazu. Duży rozmiar zbioru danych może

zapewnić bardziej reprezentatywne próbki, co może przyczynić się do lepszej wydajności modelu.

-Reprezentacja klas: Zbiór MNIST zawiera 10 różnych klas,które odpowiadają cyfrom od 0 do 9 .Każda klasa ma taką samą liczebność (ok 6000 przykładów), co oznacza, że zbiór jest zbalansowany pod względem liczności klas.

-Warianty pisma: Zbiór MNIST zawiera różne warianty pisanych cyfr, co oznacza, że obrazy mogą mieć pewne różnice w stylu pisania. Przykłady z różnymi wariantami mogą pomóc w treningu modelu na bardziej ogólnych wzorcach.

-Jednoznaczność etykiet: W przypadku zbioru MNIST, każdemu obrazowi jest przypisana jednoznaczna etykieta reprezentująca cyfrę od 0 do 9. To sprawia, że zbiór danych jest odpowiedni do problemów klasyfikacji, gdzie celem jest przewidzenie jednej z 10 możliwych klas.

-Wielkość próbek: Każdy obraz w zbiorze MNIST ma rozmiar 28x28 pikseli, co oznacza, że jest to stosunkowo niewielka liczba cech. Małe rozmiary próbek mogą pomóc w szybszym przetwarzaniu danych i uczeniu modelu.

-Jakość danych: Zbiór MNIST jest szeroko używany w dziedzinie uczenia maszynowego i został starannie opracowany i sprawdzony. Obrazy w zbiorze są czyste i dobrze jakościowe, co ułatwia analizę i ekstrakcję cech.

Podsumowując, zbiór MNIST ma kilka wartościowych cech, takich jak zbalansowanie klas, duży rozmiar, różnorodność wariantów pisma i jednoznaczne etykiety. Wszystkie te czynniki przyczyniają się do jego popularności i szerokiego wykorzystania w zadaniach związanych z klasyfikacją cyfr pisanych.

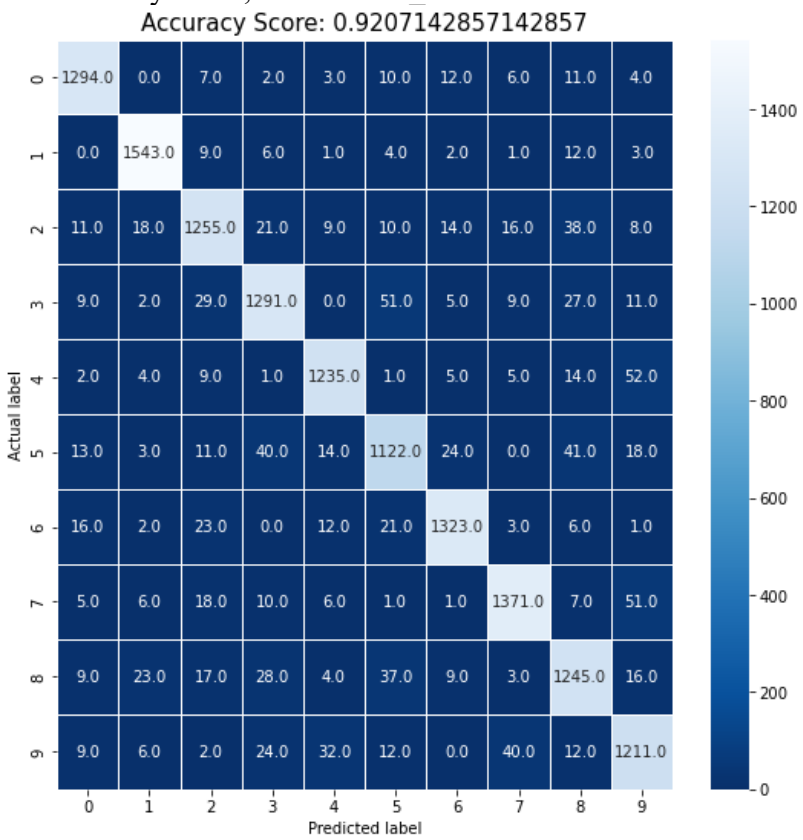
Za pomocą tego fragmentu kodu :

```
# Znajdź indeksy przykładów, dla których predykcja różni się od prawdziwej etykiety
bledne = np.where(y_pred != y_test['column_0'])[0]

# Zapisz pierwszych pięć indeksów przykładów, dla których rozpoznawanie jest błędne
pierwsze_piec = bledne[:5]

# Wyświetl pierwsze pięć indeksów błędnych rozpoznań
print('Pierwsze pięć indeksów błędnych rozpoznań:')
for i, idx in enumerate(pierwsze_piec):
    print('Numer wzorca:', idx)
    print('Rzeczywista etykieta:', y_test['column_0'].iloc[idx])
    print('Przewidziana etykieta:', y_pred[idx])
    print('---')
```

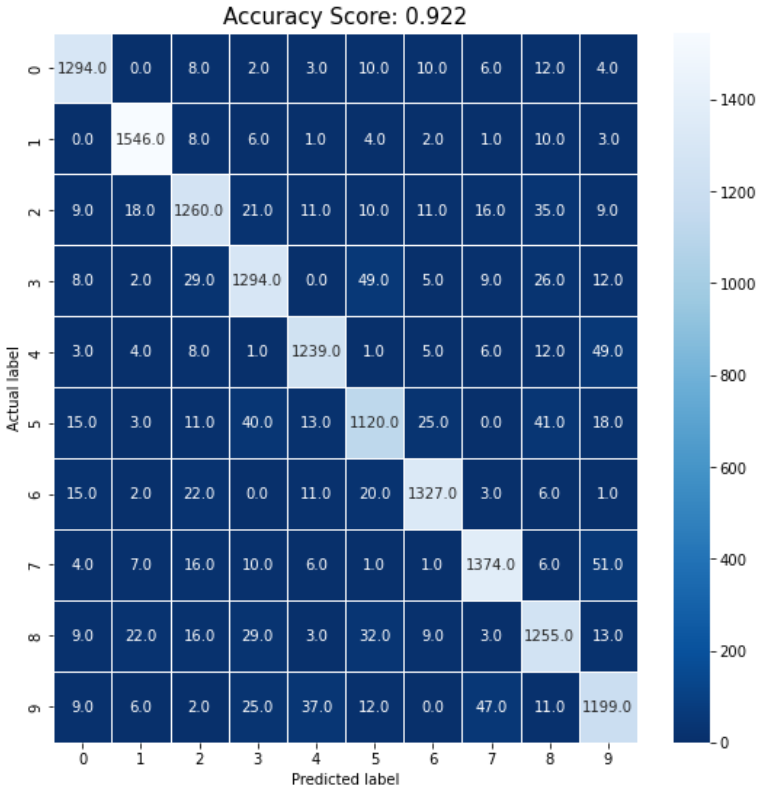
```
Pierwsze pięć indeksów błędnych rozpoznań:
Numer wzorca: 12
Rzeczywista etykieta: 0.0
Przewidziana etykieta: 4.0
---
Numer wzorca: 13
Rzeczywista etykieta: 2.0
Przewidziana etykieta: 4.0
---
Numer wzorca: 25
Rzeczywista etykieta: 8.0
Przewidziana etykieta: 9.0
---
Numer wzorca: 46
Rzeczywista etykieta: 5.0
Przewidziana etykieta: 9.0
---
Numer wzorca: 55
Rzeczywista etykieta: 6.0
Przewidziana etykieta: 8.0
---
```

otrzymujemy pięć pierwszych wzorców dla których rozpoznawanie jest błędne.

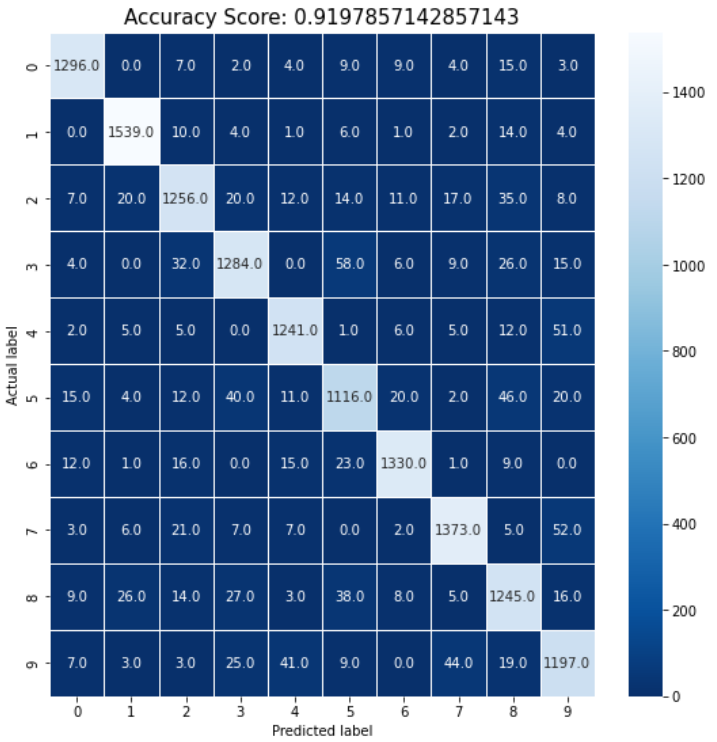
Dokładność modelu wynosi 0,915 dla max_iter=**100**



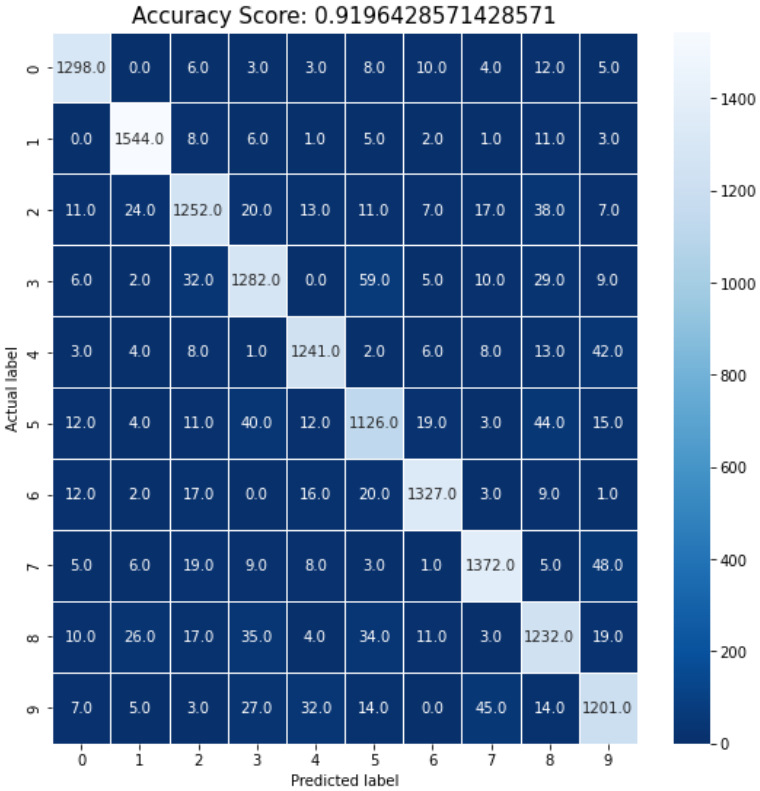
fit_intercept=**False**, max_iter = 100



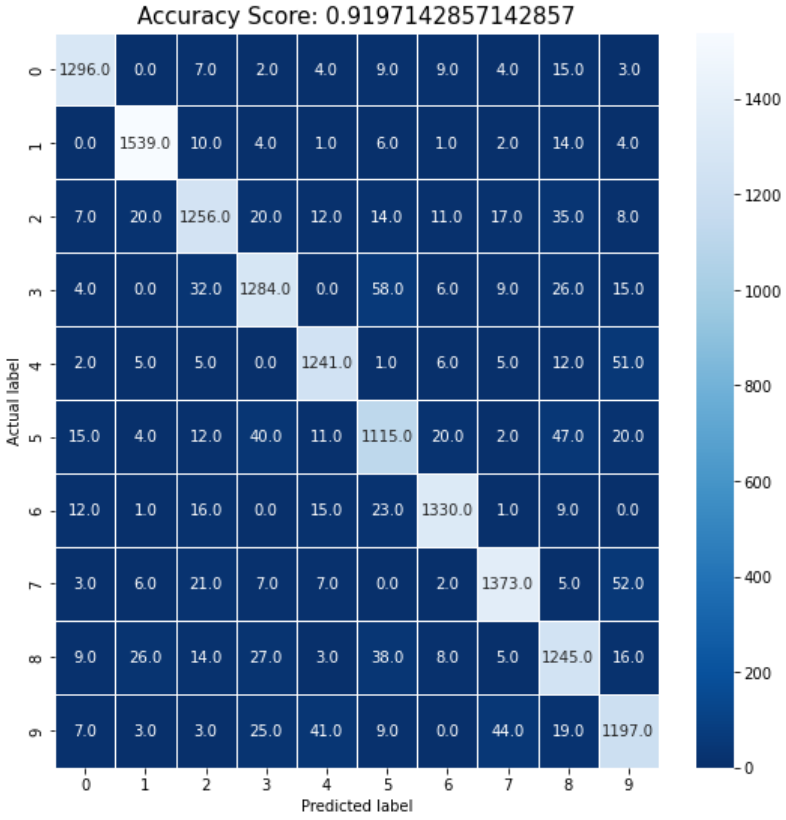
fit_intercept=True, max_iter = **50**



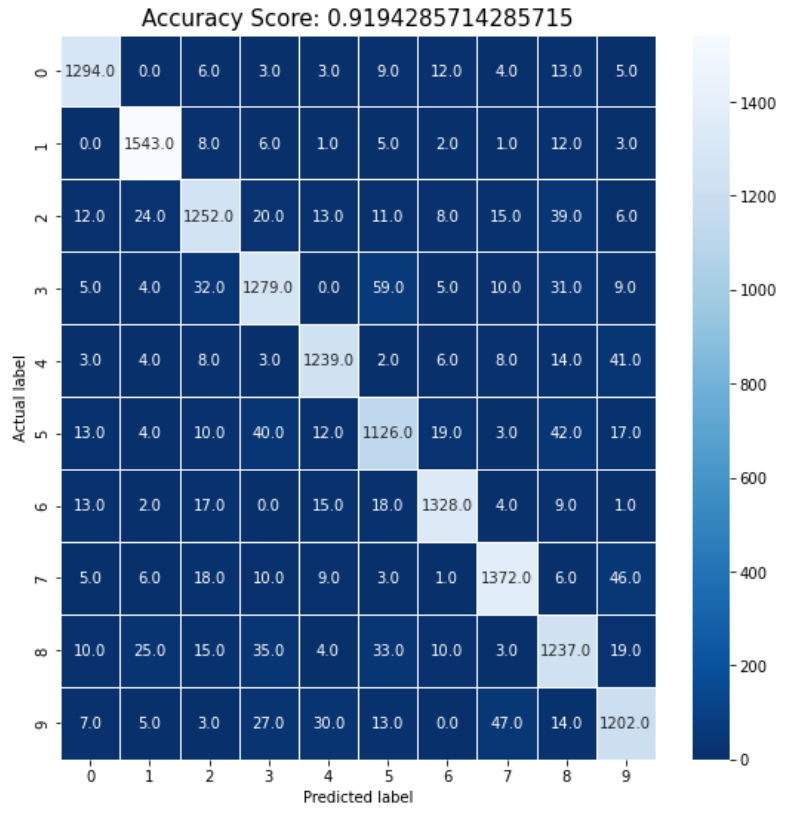
fit_intercept=True, max_iter = **200**



fit_intercept=**False**, max_iter = **50**



fit_intercept=**False**, max_iter = **200**



Porównanie

Nr próby	fit_intercept=	max_iter=	Dokładność modelu
1	True	100	0.92
2	**False**	100	0,922
3	**false**	50	0,9197
4	**false**	200	0,9194
5	True	50	0,9198
6	True	200	0,9196

Modyfikacja tych kilku właściwości zmienna dokładność modelu. Przeważnie spodziewamy się, że zwiększenie ilości iteracji pozwoli nam na zbudowanie dokładniejszego modelu. Ustawienie fit_intercept, aby przewidywana regresja zaczynała się w punkcie y=0 wpływać może różnie, albo na poprawę, albo na pogorszenie. Zmiana fit_intercept na True okazała się w tym przypadku nieznacznie pogarszać model. Ustawianie iteracji do jeszcze mniejszych ilość powtórzeń jak np. 10 sprawiało, że układ znacząco stracił na dokładności spadając do rzędu 80%. Natomiast 1000 iteracji pogorszyło model do dokładności 0.915.

WNIOSKI

Podsumowując ćwiczenie dotyczące regresji logistycznej na zbiorze danych MNIST, można wyciągnąć następujące wnioski:

1.Zbiór danych Mnist zawiera obrazy ręcznie pisanych cyfr, składa się z 10 klas reprezentujących cyfry od 0 do 9. Każda klasa ma zbliżoną liczbę próbek. Zbiór ten jest często stosowany jako przykład w zadaniach rozpoznawania obrazów.

2.Bardzo ważne i istotne jest zbalansowanie zbioru uczącego, czyli zapewnienie, że liczność próbek w poszczególnych klasach jest podobna. W tym przypadku, warunek ten jest spełniony.

3.Zastosowano model regresji logistycznej do klasyfikacji cyfr na podstawie obrazów. Model ten jest popularnym algorytmem do problemów binarnej i wieloklasowej klasyfikacji.

4.uzyskano dokładność modelu na poziomie ok 92% co oznacza, że 92% próbek ze zbioru testowego zostało poprawnie zidentyfikowanych.

5.Przy użyciu macierzy pomyłek możemy analizować wyniki klasyfikacji dla poszczególnych klas, identyfikować błędy i oceniać skuteczność modelu.

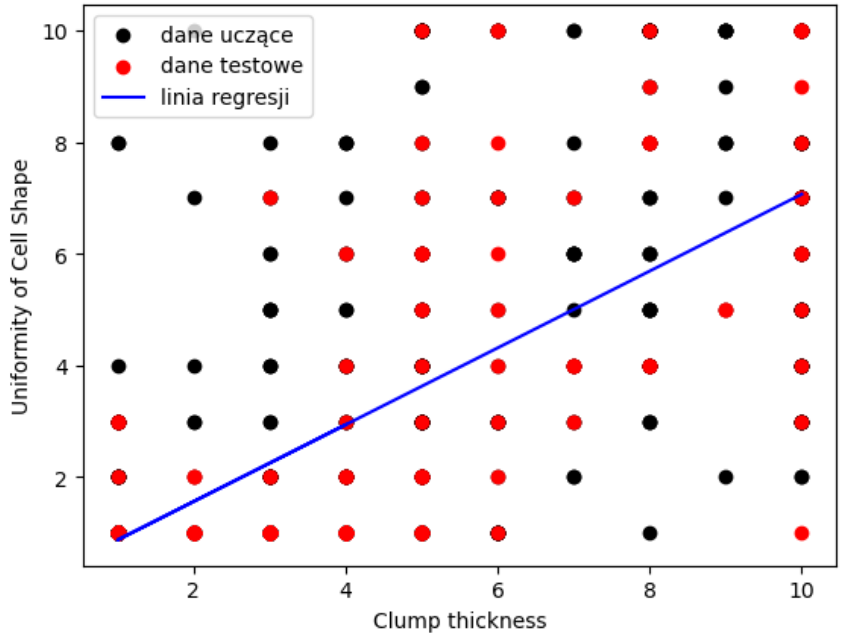
6.W identyfikowaniu błędów klasyfikacji możemy pomóc sobie, przeglądając kilka wzorców , dla których rozpoznanie było błędne. W naszym przypadku jasno widać, że błędy występują często w rozpoznaniu podobnych do siebie znaków jak 8 i 9 , 6 i 8.

7.Eksperymentowanie z hiperparametrami modelu, takimi jak ‘fit_intercept’ i ‘max_iter’, może wpływać na wyniki klasyfikacji. Możemy porównać macierze pomyłek i dokładność dla różnych wartości tych hiperparametrów.

8.Dobór optymalnych parametrów ma kluczowe znaczenie dla uzyskania jak najwyższej dokładności modelu. Warto więc eksperymentować i dostosowywać hiperparametry , aby znaleźć optymalne ustawienia.

9.Regresja logistyczna na zbiorze danych MNIST jest przykładem kasyfikacji obrazów, która może być stosowana w różnych dziedzinach, takich jak rozpoznawanie pisma odręcznego, rozpoznawanie znaków, czy identyfikacji obrazów medycznych.

Regresja Liniowa



Badanych danych jest 682 po usunięciu niepełnych rekordów.

Zbiór danych dotyczy zestawu badań nowotworów, gdzie określono końcową diagnozę jako ‘Niegroźny’ lub “Złośliwy”.Zbiór posiada dwie klasy: benign nazwany Niegroźnym (mającą 457 próbek) i malignant nazwany Złośliwym (mającą 241 próbek). W zbiorze danych klasa Niegroźny reprezentowana jest liczbą 2, a złośliwy liczbą 4.
Badane cechy to Uniformity of Cell Shape (Jednorodność kształtu komórki) i Clump thickness (grubość grudek). Obie cechy opisane są wartością do 1 do 10.

Wnioski:

Z regresji liniowej można wyciągnąć następujące wnioski:

1. Zależność liniowa: Regresja liniowa pozwala modelować zależność pomiędzy zmiennymi niezależnymi a zmienną zależną przy założeniu, że istnieje liniowa relacja pomiędzy nimi. Wnioskiem z tego jest, że można stosować regresję liniową do modelowania prostych zależności między zmiennymi.
2. Dopuszczalność założeń: Regresja liniowa zakłada pewne założenia, takie jak liniowość zależności,. Analiza dopuszczalności tych założeń może dostarczyć wniosków na temat adekwatności modelu i konieczności dalszej analizy lub transformacji danych.
3. Prognozowanie: Regresja liniowa może być wykorzystana do prognozowania wartości zmiennej zależnej na podstawie wartości zmiennych niezależnych. Wnioskiem jest, że regresja liniowa może być przydatna do prognozowania wartości na podstawie dostępnych danych.
4. Regresja liniowa może być wykorzystana do prognozowania wartości zmiennej na podstawie dostępnych danych