


Informatyka Stosowana		
Laboratorium 5	Implementacja normalizacji danych i standaryzacji	 POLITECHNIKA BYDGOSKA Wydział Telekomunikacji, Informatyki i Elektrotechniki
Przedmiot	Eksploracyjna analiza danych	
Prowadzący	mgr inż. Gracjan Kątek	

1. Wprowadzenie

Normalizacja i standaryzacja to kluczowe techniki w przetwarzaniu danych, szczególnie w przypadku analizy danych i modelowania w uczeniu maszynowym. Obie metody pozwalają przekształcić dane, aby ich wartości były bardziej jednorodne, co wpływa na poprawę efektywności algorytmów i redukcję błędów obliczeniowych.

- **Normalizacja** przekształca dane tak, aby mieściły się w określonym zakresie, zwykle od 0 do 1. Jest przydatna w przypadku algorytmów opartych na odległościach, takich jak KNN czy SVM.
- **Standaryzacja** przekształca dane, aby miały średnią równą 0 i odchylenie standardowe równe 1. Jest szczególnie istotna w przypadku algorytmów, które zakładają normalny rozkład danych, takich jak regresja liniowa czy metody PCA.

2. Zadania do samodzielnego wykonania

Zadanie 1: Analiza danych przed normalizacją i standaryzacją

Wczytaj zestaw danych zawierający cechy numeryczne (np. wynagrodzenia, powierzchnie mieszkań, ceny produktów). Za pomocą funkcji `describe()` wyświetl statystyki opisowe dla wszystkich kolumn, aby przeanalizować zakres, średnią, odchylenie standardowe i kwartyle. Odpowiedz na pytanie: które kolumny wymagają przekształcenia?

Zadanie 2: Ręczna implementacja normalizacji

Zaimplementuj normalizację ręcznie, używając wzoru:

$$x' = \frac{x_{\max} - x_{\min}}{x - x_{\min}}$$

Wybierz jedną kolumnę z danych (np. `price`), a następnie przekształć jej wartości tak, aby mieściły się w zakresie od 0 do 1. Porównaj przekształcone dane z oryginalnymi, wizualizując wyniki za pomocą histogramu.

Zadanie 3: Ręczna implementacja standaryzacji

Zaimplementuj standaryzację ręcznie, korzystając ze wzoru:

$$z = \frac{x - \mu}{\sigma}$$

Gdzie:

- μ to średnia,
- σ to odchylenie standardowe.

Przekształć wartości wybranej kolumny (np. `salary`) i wyświetl histogram dla danych przed i po standaryzacji.

Zadanie 4: Normalizacja danych przy użyciu scikit-learn

Skorzystaj z klasy `MinMaxScaler` z biblioteki scikit-learn, aby znormalizować wszystkie kolumny numeryczne w zestawie danych. Wyświetl pierwsze pięć wierszy przekształconych danych oraz sprawdź, czy wszystkie wartości mieszczą się w zakresie od 0 do 1.

Zadanie 5: Standaryzacja danych przy użyciu scikit-learn

Wykorzystaj klasę `StandardScaler` z biblioteki scikit-learn, aby zestandaryzować wszystkie kolumny numeryczne w danych. Porównaj statystyki opisowe dla danych przed i po standaryzacji. Czy średnia i odchylenie standardowe odpowiadają wartościom 0 i 1?

Zadanie 6: Wpływ normalizacji i standaryzacji na algorytmy

Podziel dane na cechy (`X`) i etykiety (`y`), a następnie zastosuj normalizację oraz standaryzację do cech `X`. Porównaj wyniki działania algorytmu KNN (np. dokładność klasyfikacji) dla danych oryginalnych, znormalizowanych i zestandaryzowanych. Wyciągnij wnioski.