Wprowadzenie do uczenia maszynowego

Lista 1

Janusz Szwabiński

1. **Instalacja i konfiguracja środowiska:** Utwórz wirtualne środowisko dla kursu. Zainstaluj niezbędne biblioteki: `pandas`, `numpy`, `scikit-learn`, `matplotlib`.

2. **Wczytanie i eksploracja danych:** Wczytaj publicznie dostępny zbiór danych `Iris` z biblioteki `scikit-learn` lub bezpośrednio z adresu URL. Wyświetl pierwsze kilka wierszy i podstawowe informacje o zbiorze (`df.info()`, `df.describe()`).

3. **Obsługa brakujących wartości:** Wczytaj zbiór danych `Titanic` (dostępny na Kaggle). Sprawdź, które kolumny zawierają brakujące wartości. Zaproponuj i zaimplementuj dwie różne strategie ich uzupełniania.

4. **Przetwarzanie danych kategorialnych:** Wykorzystaj kolumny kategorialne (np. płeć, port zaokrętowania) ze zbioru `Titanic`. Przekształć je na format numeryczny, używając techniki *One-Hot Encoding* z biblioteki `pandas` (`pd.get_dummies()`) oraz `scikit-learn`.

5. **Skalowanie danych numerycznych:** Wybierz kolumny numeryczne ze zbioru `Iris` (np. długość i szerokość płatka). Zastosuj dwie różne metody skalowania:

   - **Standaryzacja (StandardScaler):** Przekształć dane tak, aby miały średnią 0 i odchylenie standardowe 1.
   - **Normalizacja (MinMaxScaler):** Przekształć dane tak, aby ich wartości mieściły się w zakresie od 0 do 1.

   Zwizualizuj dane przed i po skalowaniu za pomocą wykresu punktowego.

6. **Podział zbioru danych:** Dokonaj podziału zbioru `Iris` na zbiór treningowy (70%) i testowy (30%) za pomocą funkcji `train_test_split` z biblioteki `scikit-learn`. Sprawdź wymiary powstałych zbiorów.

7. **Przygotowanie danych do modelowania:** Wybierz dowolny zbiór danych i przeprowadź pełny proces przygotowania danych: wczytanie, obsługa brakujących wartości, przetwarzanie danych kategorialnych i skalowanie danych numerycznych. Zapisz przygotowany zbiór do pliku.