**Projekt zaliczeniowy**

**Techniki eksploracji danych**

Paulina Górska

Informatyka i Ekonometria

Nr albumu: 185203

**Zakres projektu**

Założeniem projektu było rozwiązanie trzech problemów:

1. klasyfikacji binarnej
2. klasyfikacji wieloklasowej
3. regresji.

Rozwiązania powyższych problemów należało dokonać za pomocą własnych implementacji algorytmów:

* k-najbliższych sąsiadów
* drzewa decyzyjne
* maszyna wektorów nośnych (tylko dla problemu klasyfikacji binarnej)
* sieci neuronowe.

Otrzymane wyniki dla własnych implementacji algorytmów należało porównać z wynikami otrzymanymi dla algorytmów opracowanych na podstawie wbudowanych pakietów języka R.

Analiza powinna obejmować wpływ hiper-parametrów na jakość opracowanych modeli oraz porównanie wyników najlepszych modeli , własnych jak i tych wbudowanych.

**Zbiory danych**

1. Klasyfikacja binarna:

<https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset>

* Liczba obserwacji: 80
* Liczba atrybutów: 5
* Zmienna celu: zmienna binarna (0 – nie było cesarskiego cięcia(nie); 1 – było cesarskie cięcie (tak))

1. Klasyfikacja wieloklasowa:

<https://archive.ics.uci.edu/ml/datasets/Balance+Scale>

* Liczba obserwacji: 625
* Liczba atrybutów: 4
* Zmienna celu: trzy klasy (L, B, R)

1. Regresja:

https://archive.ics.uci.edu/ml/datasets/Servo

* Liczba obserwacji: 167
* Liczba atrybutów: 4
* Zmienna celu: zmienna numeryczna

Wszystkie zbiory danych nie zawierały braków danych.

Przekształcenia danych:

* zmienna celu w problemie klasyfikacji wieloklasowej została przekształcona z oznaczeń literowych na wartości liczbowe: 1-B; 2-L; 3-R
* w problemie regresji dwie zmienne: „motor” oraz „screw” zostały przekształcone na wartości liczbowe (1,2,3,4,5)
* dokonano normalizacji danych dla trzech algorytmów: k-najbliższych sąsiadów, maszyna wektorów nośnych oraz sieci neuronowe
* dla algorytmu maszyny wektorów nośnych przekształcono zmienną celu do wartości {-1;1}

**Najlepsze modele dla poszczególnych algorytmów**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Algorytm** | **Implementacja** | **Parametry** | **Ocena jakości dla zbioru Testowego**  **(Trafność)** |
| **Klasyfikacja Binarna** | **knn** | **Własna** | **k = 11** | **0.8125** |
|  | SVM | Własna | C = 30 | 0.7625 |
|  | Sieci NN | Własna | h = (4,4), *iter =* 10000 | 0.6625 |
|  | knn | Biblioteka R | k = 11 | 0.7321 |
|  | SVM | Biblioteka R | C = 7 | 0.6519 |
|  | Sieci NN | Biblioteka R | h = 5 | 0.6149 |
|  | Drzewa decyzyjne | Biblioteka R | max depth = 3 | 0.6545 |
| **Klasyfikacja Wieloklasowa** | knn | Własna | k = 42 | 0.9194 |
|  | Sieci NN | Własna | h = (6,6), *iter =* 80000 | 0.4736 |
|  | knn | Biblioteka R | k = 9 | 0.9089 |
|  | **Sieci NN** | **Biblioteka R** | **h = 9** | **0.9679** |
|  | Drzewa decyzyjne | Biblioteka R | max depth = 12 | 0.7903 |
|  |  |  |  | **( MAE )** |
| **Regresja** | **knn** | **Własna** | **k = 2** | **0.2483** |
|  | Sieci NN | Własna | h = (8,8), *iter =* 100000 | 0.4776 |
|  | knn | Biblioteka R | k = 2 | 0.2658 |
|  | Sieci NN | Biblioteka R | h = 6 | 0.9705 |
|  | Drzewa decyzyjne | Biblioteka R | max depth = 3 | 0.4630 |