**Projekt – Techniki Eksploracji Danych**

Projekt obejmuje wybór i wstępną analizę danych do problemów klasyfikacji binarnej, wieloklasowej i regresji. Kolejnym etapem jest użycie zbudowanych modeli KNN, drzew decyzyjnych i sieci neuronowych do predykcji. W tym celu wykonana została kroswalidacja modeli (9 razy), czyli wielokrotne wylosowanie danych uczących i walidacyjnych, co pozwoli na uśrednienie statystyk takich jak AUC, dokładność, MAE czy MAPE. Wykonane zostały również badania wpływu hiper-parametrów na wyniki danych modeli.

Dane:

* Klasyfikacja binarna - Breast Cancer https://archive.ics.uci.edu/ml/datasets/Breast+Cancer
* Klasyfikacja wieloklasowa - Abalone

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

* Regresja - Automobile  
  <https://archive.ics.uci.edu/ml/datasets/Automobile>

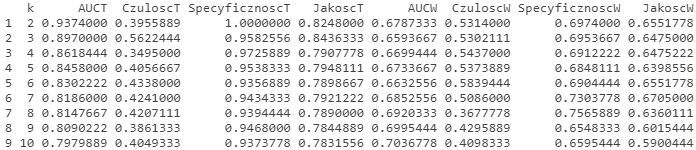
Dane do klasyfikacji binarnej to 286 wierszy i 9 kolumn z wartościami uczącymi. Kolumna nr 1 zawiera klasę “no-recurrence-events” i “recurrence-events”. Rozkład klas to 201:85. W danych pojawiały się dane ze znakami zapytania – wiersze te zostały usunięte przez co ostateczny rozkład klas zmienił się – 196:81. Kolumny z informacjami uczącymi zawierały różne wartości numeryczne i faktor. Kolumny z faktorem zostały zmienione na odpowiednie wartości numeryczne odpowiadające poziomom.

Zbiór dla klasyfikacji wieloklasowej został wybrany przez prowadzącego. Jest to zbiór 4177 wierszy i 8 kolumnami, z dodatkową kolumną klas. W zbiorze nie występowały żadne dane typu NA lub uniemożliwiające uczenie modeli. W tym przypadku tylko pierwsza kolumna zawierająca 3 poziomy została zamieniona na wartości numeryczne. Docelowe klasy wstępnie były opisane przy pomocy wartości numerycznych, ale ze względu na problem klasyfikacji zamieniono tę kolumnę na factor.

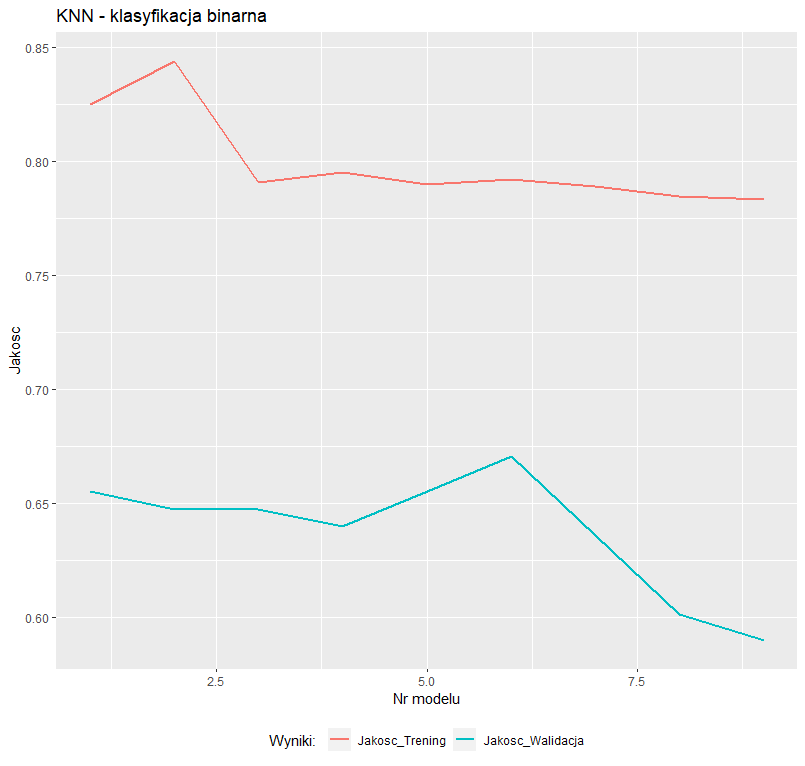
Do problem regresji wybrano mały zbiór z 205 wierszami i 25 kolumnami, plus dodatkowa kolumna z wartościami docelowymi. Kilka wierszy z wartością docelową “?” usunięto (4 wiersze). Wszystkie kolumny z wartosciami uczącymi zostały zamienione na wartości numeryczne.

1. **Klasyfikacja Binarna**
   1. **KNN**

- parametr k = 2:10

****

**Porównanie wyników dokładności (jakości) modeli KNN**



W przypadku modelu KNN dla wybranego zbioru, do problemu klasyfikacji binarnej, zwiększanie parametru k powoduje spadek dokładności predykcji. Najlepszy model, dla k = 7, osiągnął dokładność na zbiorze walidacyjnym na poziomie 0.67 .

* 1. **Drzewa decyzyjne**

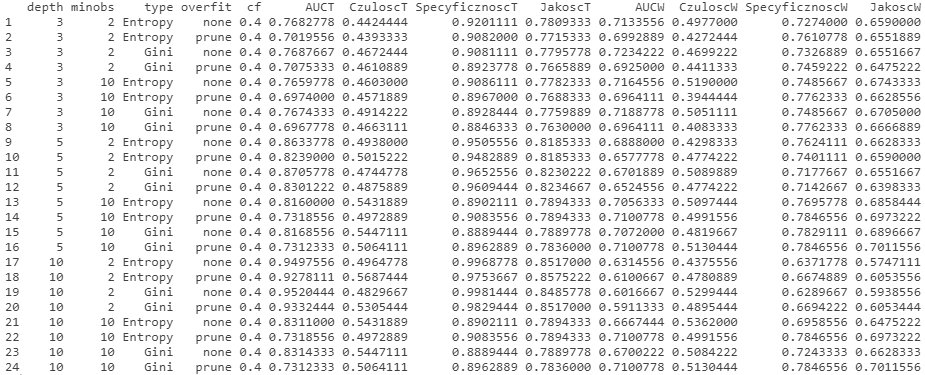
- głębokość drzewa = 3, 5, 10

- min obserwacji = 2, 10

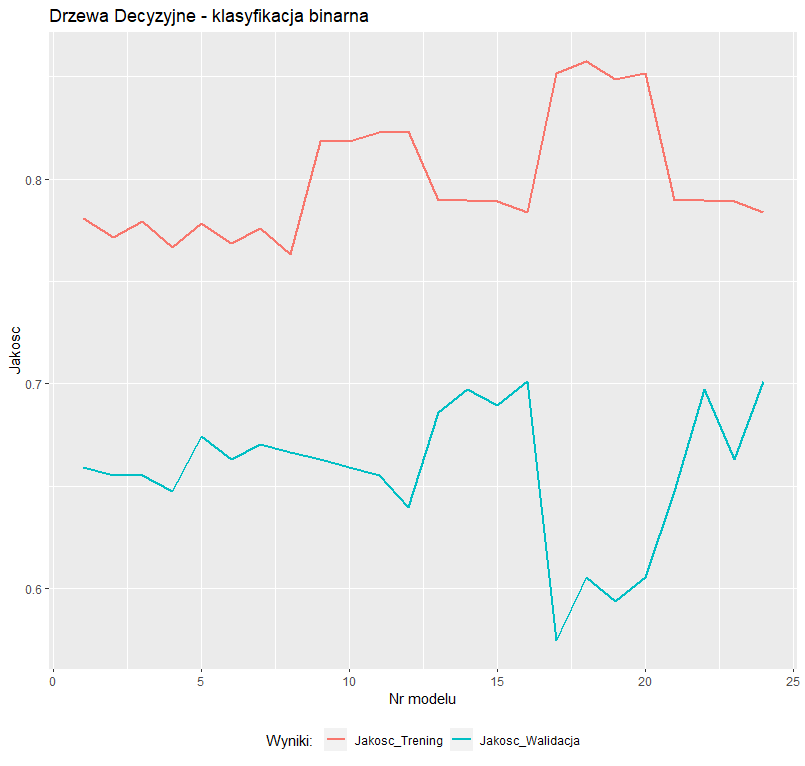
- typ = Entropia i Gini

- przycięcie drzewa = ‘none’ i ‘prune’

- jeśli przycięcie, cf = 0.4

****

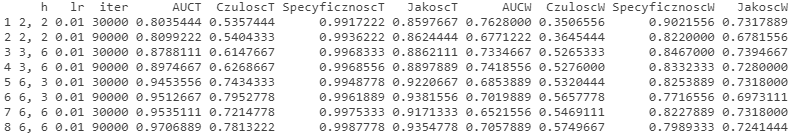
**Porównanie wyników dokładności (jakości) modeli Drzew Decyzyjnych**



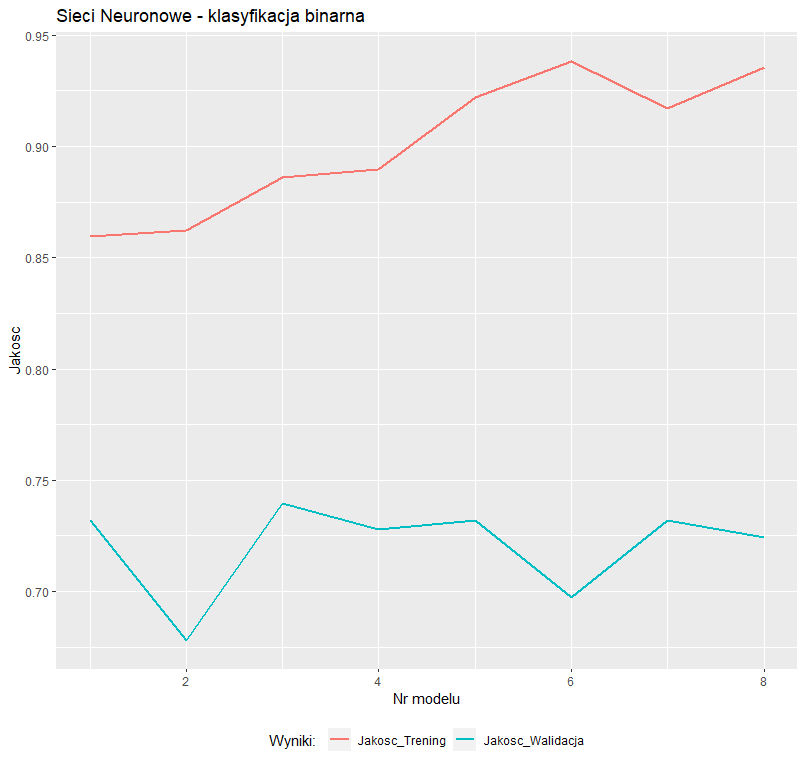
Drzewa decyzyjne osiągnęły lepszy wynik w porównaniu do KNN – 0.70 (dla zbioru walidacyjnego. W tym przypadku dużo dało zwiększenie głębokości drzewa oraz wartości minimalnej liczby obserwacji do podziału drzewa. Marginalnie lepsze wyniki daje też typ obliczeń – z Gini wyniki są marginalnie lepsze niż liczone Entropią. Kolejnym parametrem, który daje kolejne procenty dokładności jest ucinanie drzewa. Dla najlepszego modelu ucięcie drzewa daje skok z 0.689 na 0.701, ale inny model miał przeskok z 0.663 na 0.701, tylko przez przycięcie drzewa.

* 1. **Sieci Neuronowe**

- warstwy neuronów = (2,2), (3,6), (6,3), (6,6)  
- współczynnik uczenia (lr) = 0.01  
- iteracje = 30000 i 90000

****

**Porównanie wyników dokładności (jakości) modeli Sieci Neuronowych**



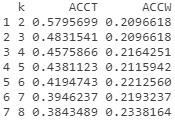
Sieci Neuronowe w przypadku klasyfikacji binarnej osiągnęły najlepszą dokładność równą 0.74 . Najlepszy model to 2 warstwy z liczbą neuronów odpowiednio 3, w pierwszej warstwie, i 6 w drugiej. Co dziwne model z mniejszą liczbą iteracji osiągnął lepszy wynik niż model z 3 razy większą ilością iteracji. Modele, które uzyskiwały dobre statystyki AUC, czułości czy specyficzności na danych treningowych osiągały bardzo podobne wyniki na danych walidacyjnych w porównaniu do początkowych modeli.

**Tabela porównawcza najlepszych modeli  
Klasyfikacja Binarna**

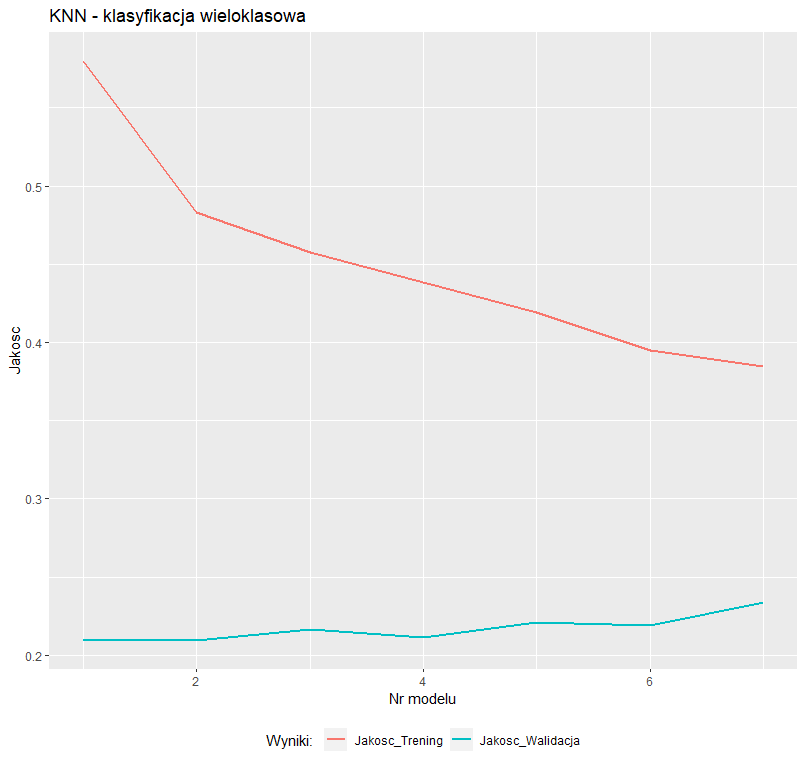
|  |  |  |  |
| --- | --- | --- | --- |
| Model | Implementacja | Hiper-parametry | Dokładność |
| KNN | Własna | k = 7 | 0.67 |
| Drzewa decyzyjne | Własna | depth = 5, minobs = 10, Gini, overfit = prune, cf = 0.4 | 0.70 |
| Sieci NN | Własna | h = (3,6), lr = 0.01,  *iter =* 30000 | 0.74 |
| KNN | Biblioteka R | k = 14 | 0.74 |
| Drzewa decyzyjne | Biblioteka R | maxdepth = 2 | 0.76 |
| Sieci NN | Biblioteka R | size = 2, decay = 0.0002 | 0.77 |

1. **Klasyfikacja Wieloklasowa**
   1. **KNN**

- parametr k = 2:8

****

**Porównanie wyników dokładności (jakości) modeli KNN**



W przypadku modelu KNN dla wybranego zbioru, do problemu klasyfikacji wieloklasowej, zwiększanie parametru k powoduje spadek dokładności predykcji na zbiorze treningowym, ale zwiększana jest predykcja na zbiorze walidacyjnym. Dla k = 8, KNN osiągnął dokładność na zbiorze walidacyjnym na poziomie 0.23, gdy na zbiorze treningowym było to 0.38 . Różnica w wynikach na zbiorze treningowym jest drość drastyczna. Możliwe, że dalsze zwiększanie parametru k spowodowałoby dalszy spadek / delikatny wzrost dokładności w zależności od zbioru.

* 1. **Drzewa decyzyjne**

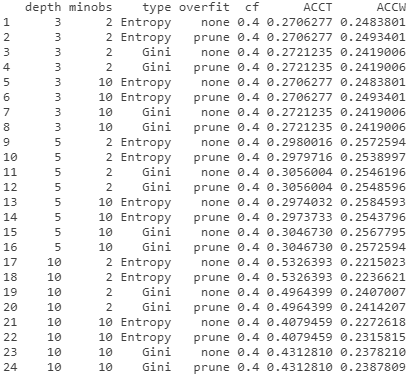
- głębokość drzewa = 3, 5, 10

- min obserwacji = 2, 10

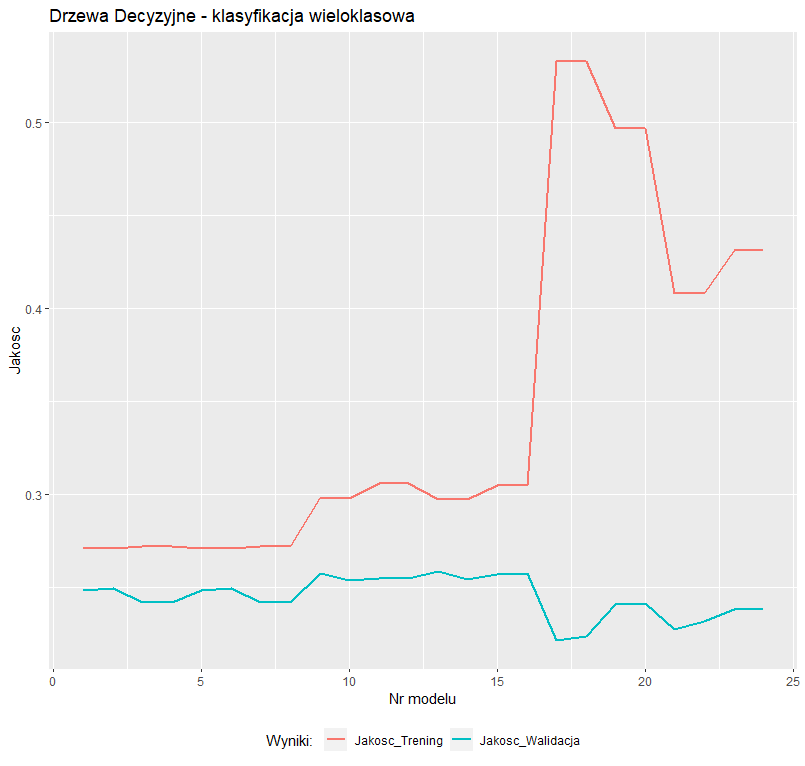
- typ = Entropia i Gini

- przycięcie drzewa = ‘none’ i ‘prune’

- jeśli przycięcie, cf = 0.4

****

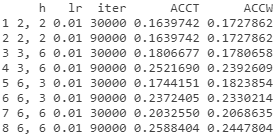
**Porównanie wyników dokładności (jakości) modeli Drzew Decyzyjnych**



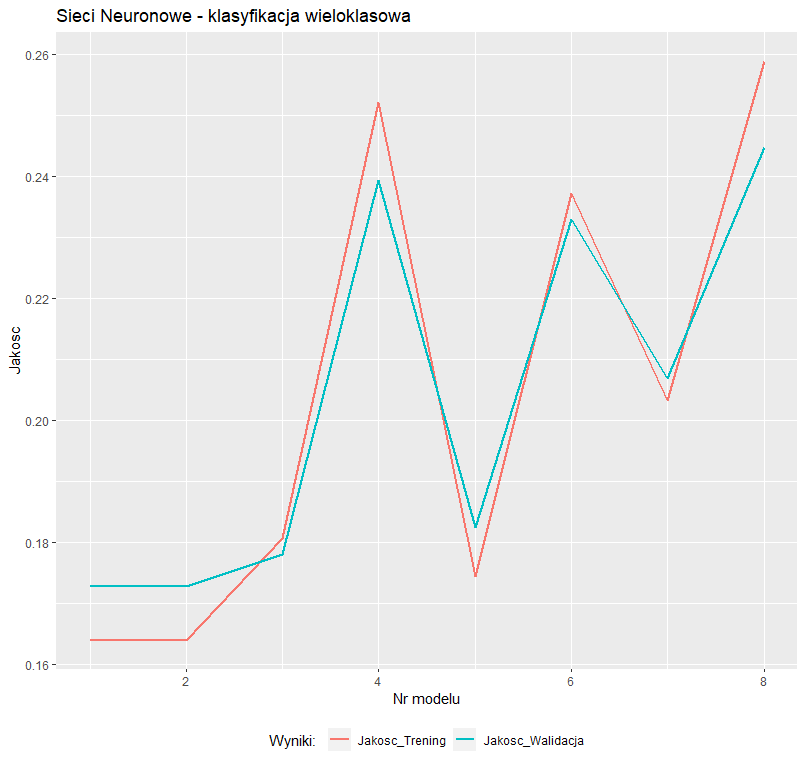
W przypadku drzew decyzyjnych zależność wyników od hiper-parametrów jest bardzo duża dla zbioru treningowego, lecz znikoma dla walidacji. Zmiana głębokości drzewa daje minimalny wzrost wyników dokładności, lecz przeskok wartości głębokości z 5 na 10 to ogromny wzrost predykcji – przynajmniej dla zbioru treningowego. W przypadku zbioru walidacyjnego wyniki są marginalnie gorsze. Najlepszy model osiągnął dokładność predykcji, na zbiorze walidacyjnym, równą 0.26 .

* 1. **Sieci Neuronowe**

- warstwy neuronów = (2,2), (3,6), (6,3), (6,6)  
- współczynnik uczenia (lr) = 0.01  
- iteracje = 30000 i 90000

****

**Porównanie wyników dokładności (jakości) modeli Sieci Neuronowych**



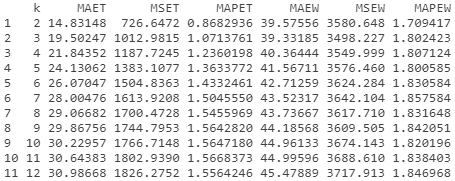
Sieci Neuronowe dla klasyfikacji wieloklasowej nie wykazują aż tak dużych różnic w predykcji w zależności od zbioru. Dobrze sobie radziły kombinacje warstw z niesymetryczną wartością neuronów – (3,6) oraz (6,3). Jednak najlepsza okazał się sieć z 2 warstwami po 6 neuronów każda. Duży wpływ na wyniki miała wartość iteracji sieci. Dzięki 3 razy większej ilości iteracji dokładność zwiększyła się z 0.21 na 0.25 (dla modelu o architekturze (6,6) i lr = 0.01).

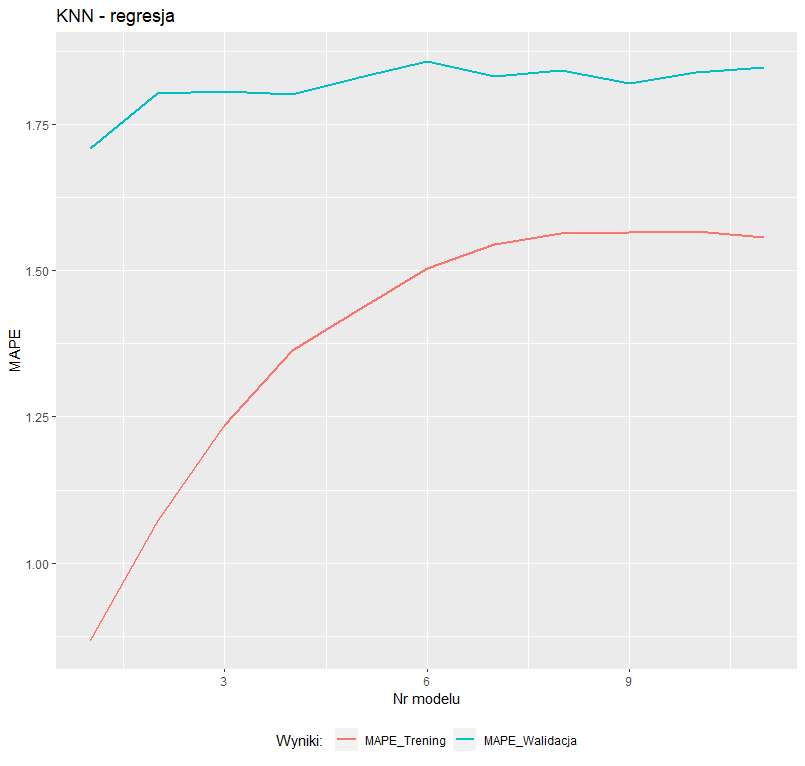
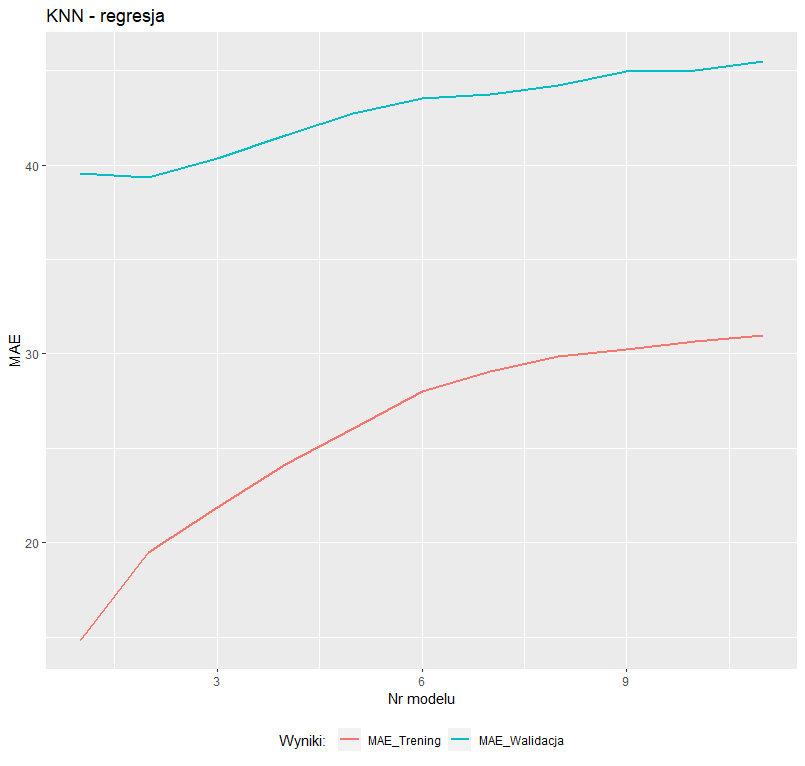
**Tabela porównawcza najlepszych modeli  
Klasyfikacja Wieloklasowa**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Implementacja | Hiper-parametry | Dokładność |
| KNN | Własna | k = 8 | 0.23 |
| Drzewa decyzyjne | Własna | depth = 5, minobs = 10, Entropia, overfit = prune | 0.26 |
| Sieci NN | Własna | h = (6,6), lr = 0.01,  *iter =* 90000 | 0.25 |
| KNN | Biblioteka R | k = 7 | 0.54 |
| Drzewa decyzyjne | Biblioteka R | maxdepth = 2 | 0.53 |
| Sieci NN | Biblioteka R | size = 9, decay = 0.02 / 0.0002 | 0.27 |

1. **Regresja**
   1. **KNN**

- parametr k = 2:12

****



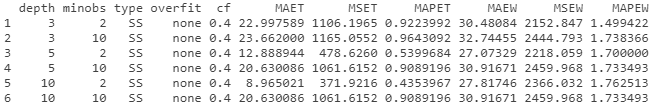
W przypadku regresji i wybranego zbioru, wyniki MAE czy MAPE są prawie 2 gorsze, porównując dane treningowe do walidacyjnych. Najlepsza predykcja dla KNN (k = 3, zbiór walidacyjny) to MAE = 39.33 (MAPE = 1.80). W pewnym momencie zwiększanie parametru k nie powoduje pogarszania czy polepszania predykcji.

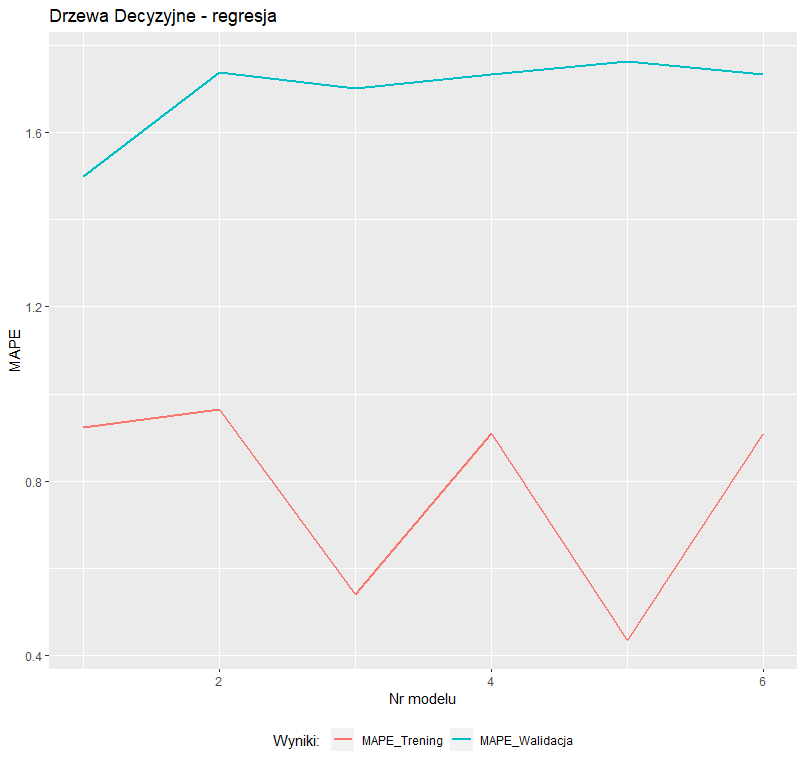
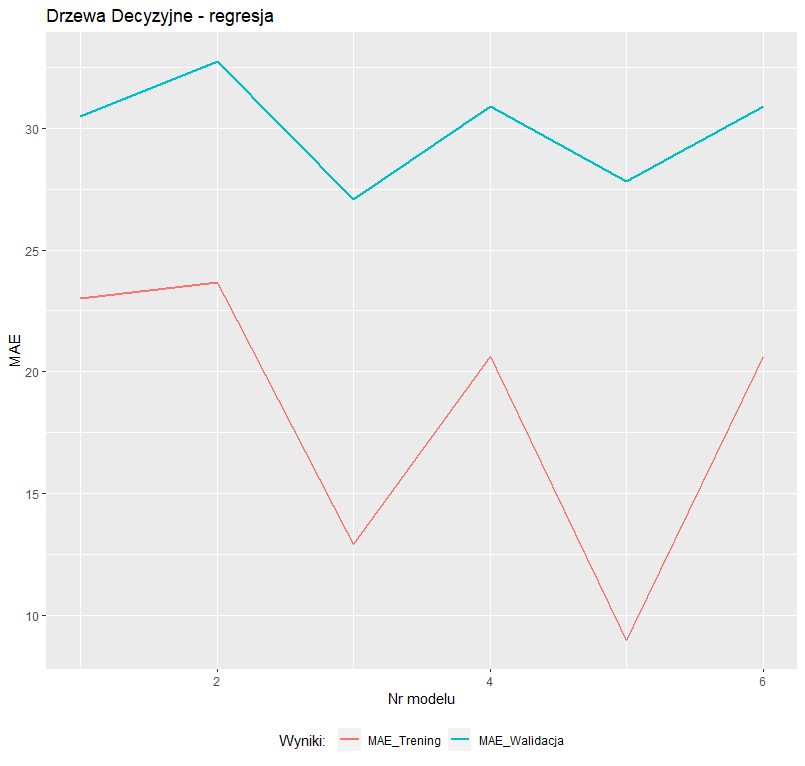
* 1. **Drzewa decyzyjne**

- głębokość drzewa = 3, 5, 10

- min obserwacji = 2, 10

- typ = SS

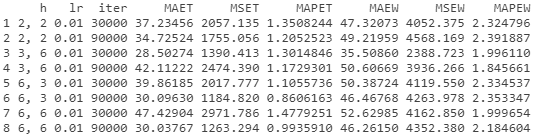
****

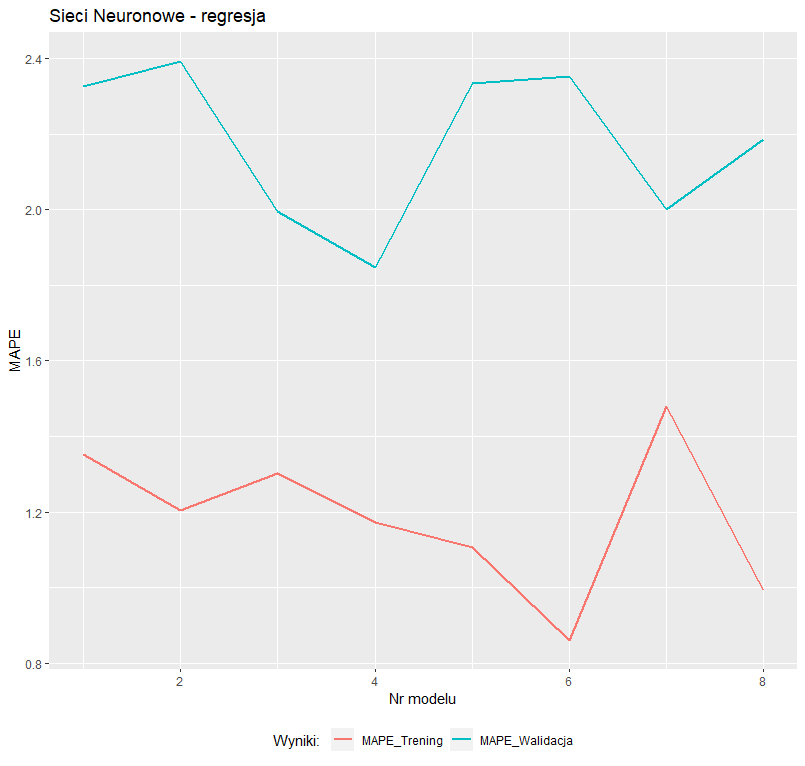
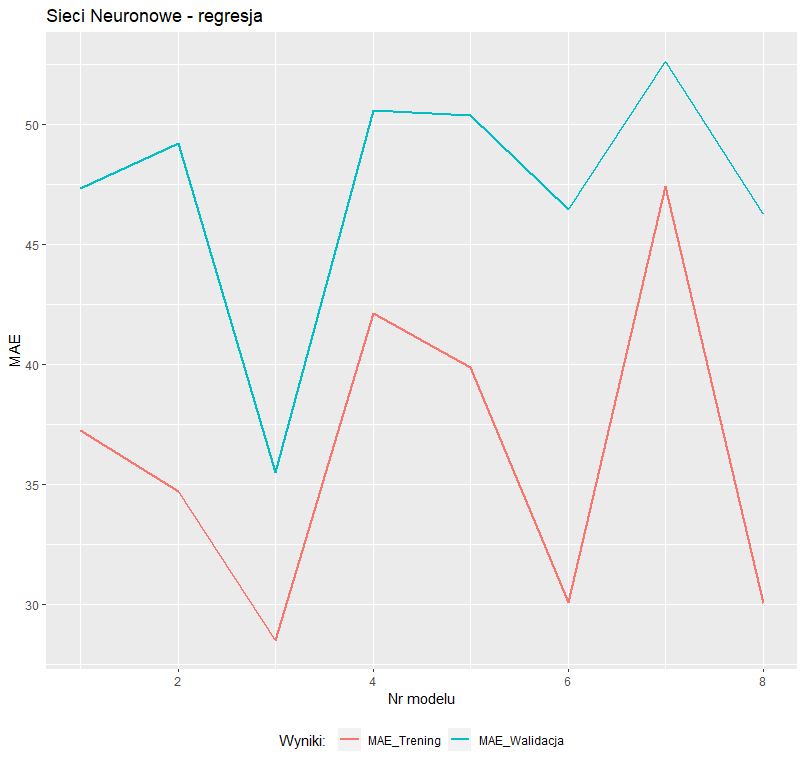


Wpływ hiper-parametrów na wyniki drzew decyzyjnych, w tym przypadku, jest bardzo znikomy, a przynajmniej dla zbioru walidacyjnego. Statystyka predykcji na zbiorze treningowym bardzo uwidacznia wpływ parametrów, np. minimalnej wartości obserwacji (minobs) – najniższe wartości MAE czy MAPE są dla parametru minobs = 2. Najlepszy model drzewa decyzyjnego osiągnął wartość MAE = 27.81 oraz MAPE = 1.70.

* 1. **Sieci Neuronowe**

- warstwy neuronów = (2,2), (3,6), (6,3), (6,6)  
- współczynnik uczenia (lr) = 0.01  
- iteracje = 30000 i 90000

****



Sieci Neuronowe poradziły sobie z problemem regresji nie tak źle jak na tak małą liczbę danych – lekko ponad 200 wierszy. Lepiej sobie radziły sieci z większą ilością neuronów - patrząc na MAE, czego nie widać w przypadku MAPE. Najlepszy model z 2 warstwami neurownów - (3,6) i 30000 iteracjami osiągnął wyniki MAE = 35.51 oraz MAPE = 1.99.

**Tabela porównawcza najlepszych modeli - Regresja**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Implementacja | Hiper-parametry | Dokładność (MAE / MAPE) |
| KNN | Własna | k = 3 | 39.33 / 1.80 |
| Drzewa decyzyjne | Własna | depth = 5, minobs = 2, SS | 27.81 / 1.70 |
| Sieci NN | Własna | h = (3,6), lr = 0.01,  *iter =* 30000 | 35.51 / 1.99 |
| KNN | Biblioteka R | k = 2 | 27.59 / - |
| Drzewa decyzyjne | Biblioteka R | maxdepth = 7 | 31.18 / - |
| Sieci NN | Biblioteka R | size = 12, decay = 0.02 | 0.018 / - |

**Podsumowanie**

W projekcie zbadano algorytmy KNN, Drzew Decyzyjnych i Sieci Neuronowych oraz wpływ hiper-parametrów. W zależności od problemu i zbioru danych różne algorytmy osiągały lepsze wyniki predykcji. Aby uśrednić wyniki dla poszczególnych modeli wykonano kroswalidację 9, a sama funkcja do niej uwzględniała zrównoleglenie, co pozwoliło wykonać badania większej ilości modeli. Najdłuższymi obliczeniami wykazał się algorytm KNN, który jest zrobiony w R na bazie podstawowej idei. Dodatkowo zbiór danych „Abalone” posiadał ponad 4000 wierszy co przy wolniejszej implementacji, jak np. KNN, powodowało bardzo długie obliczenia.

Algorytmy uzyskiwały bardzo podobną dokładność predykcji dla problemu klasyfikacji binarnej. Modele z własną implementacją algorytmu osiągnęły delikatnie gorsze wyniki dokładności (0.67 – 0.74) w porównaniu do modeli z bibliotek R (0.74 – 0.77). W przypadku klasyfikacji wieloklasowej ręcznie implementacje to wyniki na poziomie 0.23 – 0.26, w porównaniu do 0.53 – 0.54 dla KNN i drzew decyzyjnych oraz 0.27 dla sieci neuronowych z bibliotek R. Regresja to zdecydowana wygrana sieci neuronowych z biblioteki ‘nnet’. Natomiast reszta algorytmów, własnych i wbudowanych, osiała bardzo podobne wyniki MAE (27-40) i MAPE (dla własnych = 1.70 – 1.99).