**Analiza**

W projekcie podjęłam temat klasyfikacji binarnej, wieloklasowej oraz regresję.

Natomiast do klasyfikacji wieloklasowej został wykorzystany zbiór car. Bezpośrednio wiąże CAR z sześcioma atrybutami wejściowymi: zakup, konserwacja, drzwi, osoby, bezpieczeństwo i bagażnik.

Zbiór Car

Wartości klas:  
  
unacc, acc, good, vgood  
  
Atrybuty

buying: vhigh, high, med, low.  
maint: vhigh, high, med, low.  
doors: 2, 3, 4, 5more.  
persons: 2, 4, more.  
lug\_boot: small, med, big.  
safety: low, med, high.

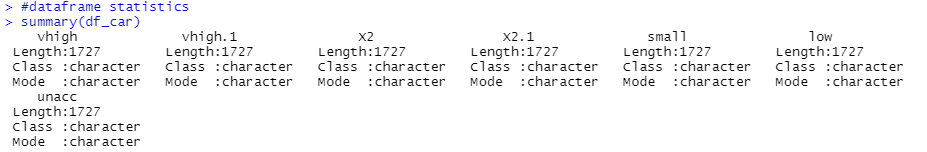
Rozwiązanie zadania klasyfikacji umożliwia szereg technik i algorytmów uczących. Takie algorytmy nazywamy klasyfikatorami, w analizie zostały użyte klasyfikatory oparte o : drzewa decyzyjne, regresje liniową, sieci neuronowe, KNN.

**Szczegóły**

W pierwszym etapie zaimportowano bazę oraz sprawdzono ewentualne braki. Zgodnie z opisem na <https://archive.ics.uci.edu/ml/machine-learning-databases/car/> dane są kompletne.

Następnie na bazie zbiorów zostanie stworzony model klasyfikujący oraz zostanie podzielony zbiór na uczący = 0,8 oraz testowy = 0,2. Następnie zostało oszacowane dopasowanie modelu KNN do zbioru danych treningowych.

**Tabela – zbiór Car**



**KLASYFIKACJA BINARNA**

Dane pochodzą z

<https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication>

1372 obserwacji oraz 5 zmiennych

Variance

Skewness

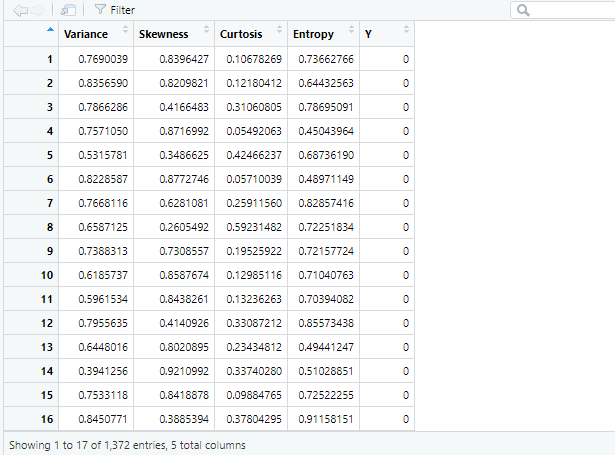
Curtosis

Entropy

Wszystkie zmienne są numeryczne.

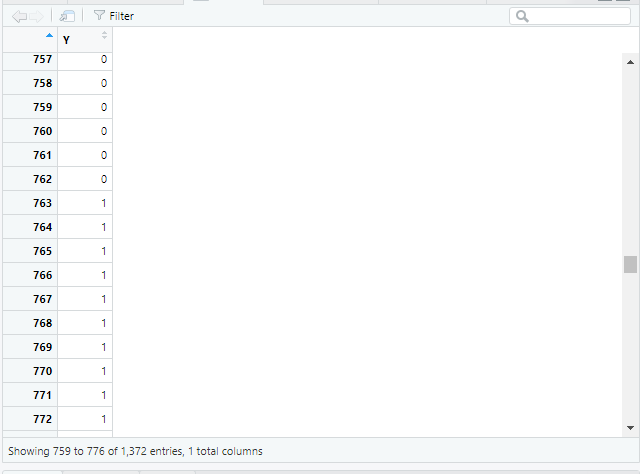
**NORMALIZACJA**

Normalizacja metodą Min-Max. Ta metoda przeprowadza liniową transformację pierwotnych danych najczęściej do przedziału [0, 1] według wzoru: V ′ = V − min max − min ∗ (new max − new min) + new min gdzie [min, max] jest oryginalnym przedziałem, w którym mieszczą się dane wejściowe, natomiast [new min, new max] jest nowym przedziałem danych.



Fragment tablicy wynikowej z przeprowadzonej normalizacji na zbiorze f2.

Następnie podzielono oryginalny zbiór na testowy oraz uczący.



Fragment podziału, wartość „ 0 „ od 1 do 762 obserwacji. Natomiast wartość „1” od 763 do 1372.

**REGRESJA**

Dane pochodzą z <https://archive.ics.uci.edu/ml/machine-learning-databases/servo>

167 obserwacji oraz 5 zmiennych

Motor – character

Screw - character

Pgain - int

Vgain -int

Sprawdzono braki danych, zbiór nie posiada takich obserwacji.

Następnie stworzono kolumny zero-jedynkowych dla zmiennych jakościowych, oraz usunięto pierwszą i drugą kolumnę.

Podzielono zbiór oryginalny na zbiór uczący i testowy.

**KNN PAKIET – K- najbliższych sąsiadów**

k-Nearest Neighbors

1372 samples

4 predictor

2 classes: '0', '1'

Brak wstępnego przetwarzania

Ponowne próbkowanie: Cross-Validated (10-krotne)

Zestawienie wielkości próbek: 1234, 1235, 1235, 1235, 1235, 1235, ...

Resampling results across tuning parameters:

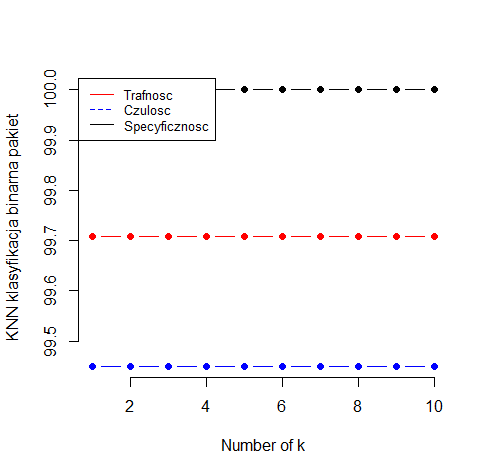
|  |  |
| --- | --- |
| k | Accuracy Kappa |
| 1 | 0.9985401 0.9970496 |
| 2 | 0.9978102 0.9955792 |
| 3 | 0.9985401 0.9970496 |
| 4 | 0.9985401 0.9970496 |
| 5 | 0.9985401 0.9970496 |
| 6 | 0.9985401 0.9970496 |
| 7 | 0.9985401 0.9970496 |
| 8 | 0.9985401 0.9970496 |
| 9 | 0.9985401 0.9970496 |
| 10 | 0.9985401 0.9970496 |

Dokładność została wykorzystana do wyboru optymalnego modelu za pomocą

największa wartość.

Ostateczna wartość zastosowana w modelu wyniosła k = 10.

|  |  |  |  |
| --- | --- | --- | --- |
| K | Trafnosc | Czulosc | Specyficznosc |
| [1,] | 1 99.70845 | 99.45055 | 100 |
| [2,] | 2 99.70845 | 99.45055 | 100 |
| [3,] | 3 99.70845 | 99.45055 | 100 |
| [4,] | 4 99.70845 | 99.45055 | 100 |
| [5,] | 5 99.70845 | 99.45055 | 100 |
| [6,] | 6 99.70845 | 99.45055 | 100 |
| [7,] | 7 99.70845 | 99.45055 | 100 |
| [8,] | 8 99.70845 | 99.45055 | 100 |
| [9,] | 9 99.70845 | 99.45055 | 100 |
| [10,] | 10 99.70845 | 99.45055 | 100 |

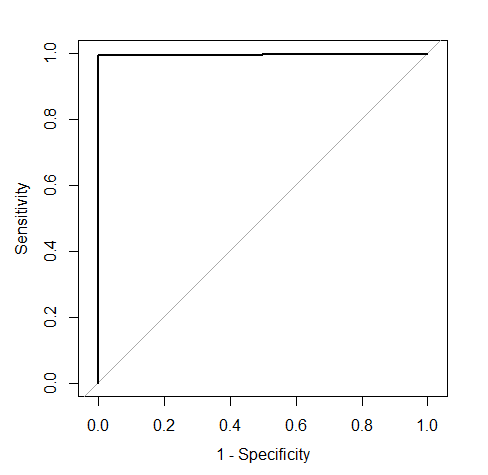


**Y\_kl\_b\_pred**

Faktyczne 0 Faktyczne 1

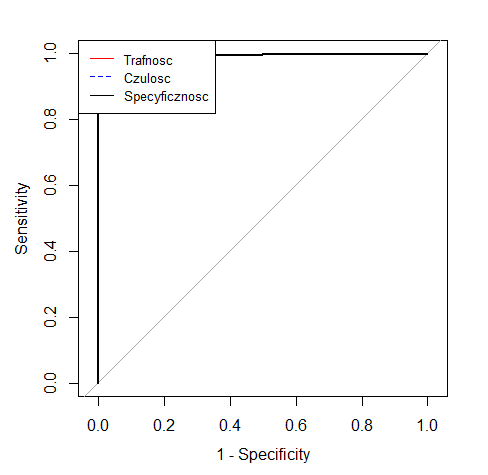
Przewidywane 0 181 0

Przewidywane 1 1 161



**Pole pod krzywą AUC = 0.9969**

**KNN kod własny – Klasyfikacja binarna**

****

**DRZEWA DECYZYJNE PAKIET KLASYFIKACJA BINARNA**

Dane : <https://cran.rstudio.com/bin/windows/contrib/4.0/gbm_2.1.8.zip>

1372 samples

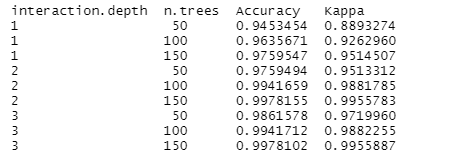
4 predictor

2 classes: „0” or „1”

Brak wstępnego przetwarzania.

Ponowne próbkowanie: Cross Validated ( 10 krotnie). Zestawienie próbek wynosi: 1234, 1235,135,1235,1235, …

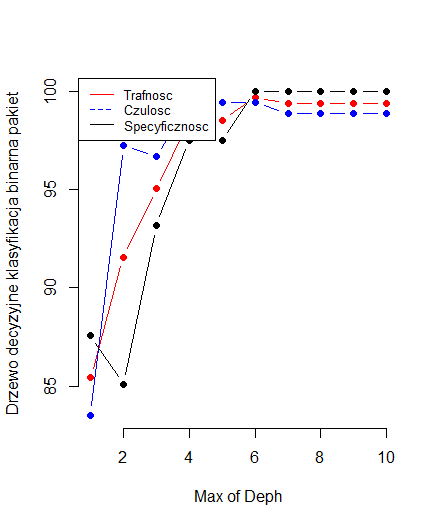
Ponowne próbkowanie wyników w różnych parametrach strojenia:



Parametr strojenia „srinkage” był utrzymywany na stałym poziomie 0,1.

Parametr strojenia „n.minobsinnode” był utrzymywany na stałym poziomie value =10.

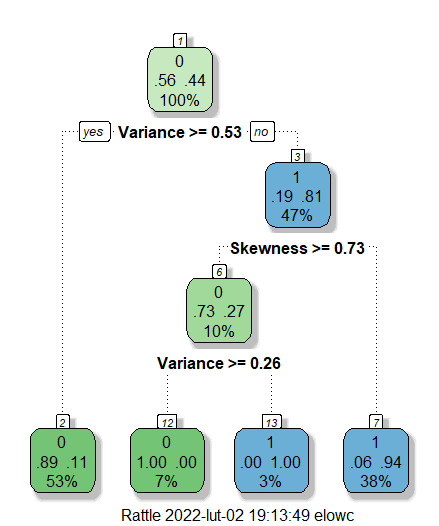
Dokładność została wykorzystana do wyboru optymalnego modelu za pomocą największej wartości. Ostateczne wartości zastosowane w modelu: to n.trees =150, interaction.depth = 2, shrinkage = 0,1 i n.minobsinnode = 10.

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MaxDeph** | **Trafność** | **Czułość** | **Specyficzność** |
| [1,] | 1 | 85.42274 | 83.51648 | 87.57764 |
| [2,] | 2 | 88.92128 | 97.25275 | 79.50311 |
| [3,] | 3 | 92.41983 | 96.70330 | 87.57764 |
| [4,] | 4 | 92.41983 | 96.70330 | 87.57764 |
| [5,] | 5 | 92.41983 | 96.70330 | 87.57764 |
| [6,] | 6 | 92.41983 | 96.70330 | 87.57764 |
| [7,] | 7 | 92.41983 | 96.70330 | 87.57764 |
| [8,] | 8 | 92.41983 | 96.70330 | 87.57764 |
| [9,] | 9 | 92.41983 | 96.70330 | 87.57764 |
| [10,] | 10 | 92.41983 | 96.70330 | 87.57764 |

Y\_kl\_b\_pred – 1372 obserwacji z 1 zmienną

|  |  |  |
| --- | --- | --- |
|  | Faktyczne 0 | Faktyczne 1 |
| Przewidywane 0 | 176 | 20 |
| Przewidywane 1 | 6 | 141 |



AUC = 0,9286

**DRZEWA DECYZYJNE – KLASYFIKACJA BINARNA kod własny**

**levelName Count**

1 Root 1029

2 ¦--Variance <= 0.529527868521443 485

3 ¦ ¦--Skewness <= 0.733643408532182 387

4 ¦ ¦ ¦--Curtosis <= 0.357559178926056 232

5 ¦ ¦ °--Curtosis > 0.357559178926056 155

6 ¦ °--Skewness > 0.733643408532182 98

7 ¦ ¦--Variance <= 0.208712834159041 26

8 ¦ °--Variance > 0.208712834159041 72

9 °--Variance > 0.529527868521443 544

10 ¦--Variance <= 0.636739285637021 189

11 ¦ ¦--Curtosis <= 0.0959657096086329 35

12 ¦ °--Curtosis > 0.0959657096086329 154

13 °--Variance > 0.636739285637021 355

14 ¦--Variance <= 0.680310667849339 61

15 °--Variance > 0.680310667849339 294

**Prob Leaf**

1 0.563654033041788, 0.436345966958212

2 0.193814432989691, 0.806185567010309

3 0.0568475452196382, 0.943152454780362

4 1 \*

5 0.141935483870968, 0.858064516129032 \*

6 0.73469387755102, 0.26530612244898

7 1 \*

8 1 \*

9 0.893382352941177, 0.106617647058824

10 0.714285714285714, 0.285714285714286

11 1 \*

12 0.876623376623377, 0.123376623376623 \*

13 0.988732394366197, 0.0112676056338028

14 0.934426229508197, 0.0655737704918033 \*

15 1 \*

**TRAFNOŚĆ = 95,62682**

**levelName Count**

1 Root 1029

2 ¦--Variance <= 0.529527868521443 485

3 ¦ ¦--Skewness <= 0.733643408532182 387

4 ¦ °--Skewness > 0.733643408532182 98

5 ¦ ¦--Variance <= 0.208712834159041 26

6 ¦ °--Variance > 0.208712834159041 72

7 °--Variance > 0.529527868521443 544

8 ¦--Variance <= 0.636739285637021 189

9 ¦ ¦--Curtosis <= 0.0959657096086329 35

10 ¦ °--Curtosis > 0.0959657096086329 154

11 °--Variance > 0.636739285637021 355

Prob Leaf

1 0.563654033041788, 0.436345966958212

2 0.193814432989691, 0.806185567010309

3 0.0568475452196382, 0.943152454780362 \*

4 0.73469387755102, 0.26530612244898

5 1 \*

6 1 \*

7 0.893382352941177, 0.106617647058824

8 0.714285714285714, 0.285714285714286

9 1 \*

10 0.876623376623377, 0.123376623376623 \*

11 0.988732394366197, 0.0112676056338028 \*

**TRAFNOŚĆ\_CP = 95,62682**

**MASZYNA WEKTORÓW NOŚNYCH PAKIET KLASYFIKACJA BINARNA**

Instalacja pakietu : <https://cran.rstudio.com/bin/windows/contrib/4.0/kernlab_0.9-29.zip>

ed\_packages

**Support Vector Machines with Polynomial Kernel**

1372 samples

4 predictor

2 classes: '0', '1'

**No pre-processing**

Resampling: Cross-Validated (10 fold)

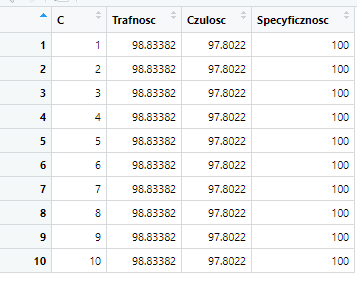
Summary of sample sizes: 1234, 1235, 1235, 1235, 1235, 1235, ...

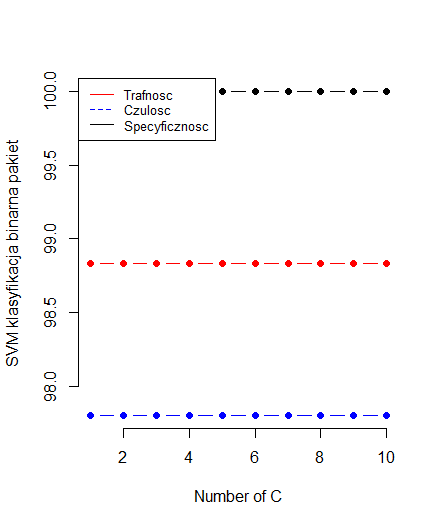
Resampling results across tuning parameters:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **degree** | **scale** | **C** | **Accuracy** | **Kappa** |
| 1 | 0.001 | 0.25 | 0.5553898 | 0.0000000 |
| 1 | 0.001 | 0.50 | 0.7551359 | 0.4753742 |
| 1 | 0.001 | 1.00 | 0.8906802 | 0.7745771 |
| 1 | 0.010 | 0.25 | 0.9489897 | 0.8963005 |
| 1 | 0.010 | 0.50 | 0.9708399 | 0.9411438 |
| 1 | 0.010 | 1.00 | 0.9766741 | 0.9529939 |
| 1 | 0.100 | 0.25 | 0.9795832 | 0.9588818 |
| 1 | 0.100 | 0.50 | 0.9817677 | 0.9632799 |
| 1 | 0.100 | 1.00 | 0.9825029 | 0.9647446 |
| 2 | 0.001 | 0.25 | 0.7544060 | 0.4737393 |
| 2 | 0.001 | 0.50 | 0.8906802 | 0.7745771 |
| 2 | 0.001 | 1.00 | 0.9431609 | 0.8842207 |
| 2 | 0.010 | 0.25 | 0.9730244 | 0.9455848 |
| 2 | 0.010 | 0.50 | 0.9781286 | 0.9559543 |
| 2 | 0.010 | 1.00 | 0.9781181 | 0.9559750 |
| 2 | 0.100 | 0.25 | 0.9832275 | 0.9662161 |
| 2 | 0.100 | 0.50 | 0.9868719 | 0.9735260 |
| 2 | 0.100 | 1.00 | 0.9956257 | 0.9911623 |
| 3 | 0.001 | 0.25 | 0.8498625 | 0.6870252 |
| 3 | 0.001 | 0.50 | 0.9256638 | 0.8483981 |
| 3 | 0.001 | 1.00 | 0.9548186 | 0.9084819 |
| 3 | 0.010 | 0.25 | 0.9781339 | 0.9559537 |
| 3 | 0.010 | 0.50 | 0.9773881 | 0.9545187 |
| 3 | 0.010 | 1.00 | 0.9788480 | 0.9574454 |
| 3 | 0.100 | 0.25 | 0.9905215 | 0.9808782 |
| 3 | 0.100 | 0.50 | 0.9963557 | 0.9926375 |
| 3 | 0.100 | 1.00 | 0.9985401 | 0.9970544 |

Dokładność została wykorzystana do wyboru optymalnego modelu za pomocą największej wartości. Ostateczne wartości zastosowane do modelu to stopień 3, skala =0.1 , C = 1.

**svm\_kl\_b\_pakiet**

****

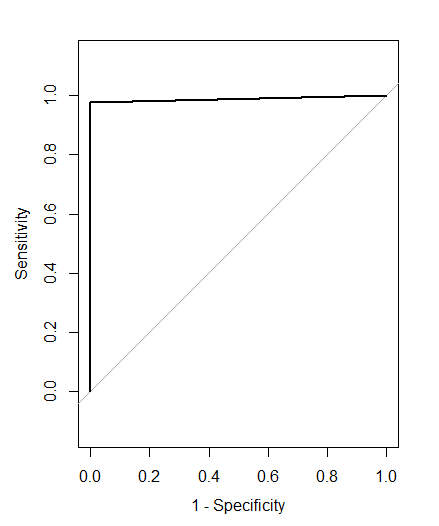
****

**y\_mwn\_kl\_b\_Pred**

Faktyczne 0 Faktyczne 1

Przewidywane 0 178 0

Przewidywane 1 4 161

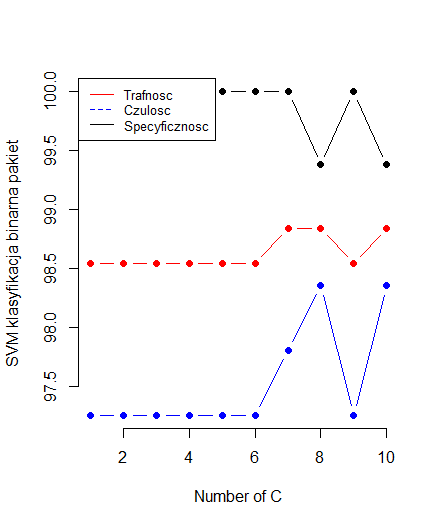


**Area under the curve: 0.9879**

**MASZYNA WEKTORÓW NOŚNYCH KLASYFIKACJA BINARNA – kod własny**

**svm\_kl\_b\_kod**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **C** | **Trafność** | **Czułość** | **Specyficzność** |
| [1,] | 1 | 98.54227 | 97.25275 | 100.00000 |
| [2,] | 2 | 98.54227 | 97.25275 | 100.00000 |
| [3,] | 3 | 98.54227 | 97.25275 | 100.00000 |
| [4,] | 4 | 98.54227 | 97.25275 | 100.00000 |
| [5,] | 5 | 98.54227 | 97.25275 | 100.00000 |
| [6,] | 6 | 98.54227 | 97.25275 | 100.00000 |
| [7,] | 7 | 98.83382 | 97.80220 | 100.00000 |
| [8,] | 8 | 98.83382 | 98.35165 | 99.37888 |
| [9,] | 9 | 98.54227 | 97.25275 | 100.00000 |
| [10,] | 10 | 98.83382 | 98.35165 | 99.37888 |

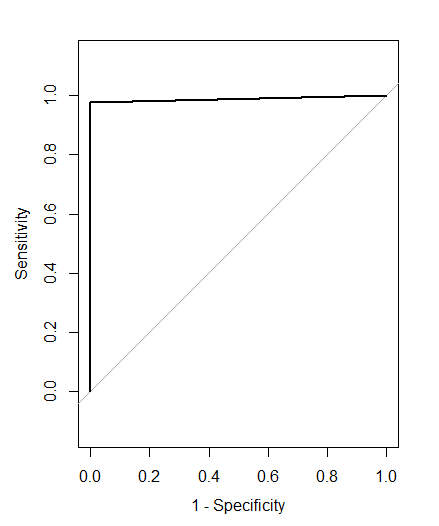


**y\_mwn\_kl\_b\_Pred**

svm\_kl\_b\_kodC\_tab Faktyczne 0 Faktyczne 1

Przewidywane 0 178 0

Przewidywane 1 4 161

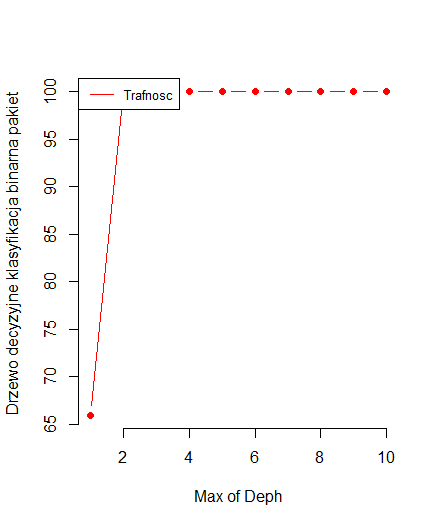


**Area under the curve: 0.9879**

**DRZEWO DECYZYJNE KLASYFIKACJA WIELOKLASOWA PAKIET**

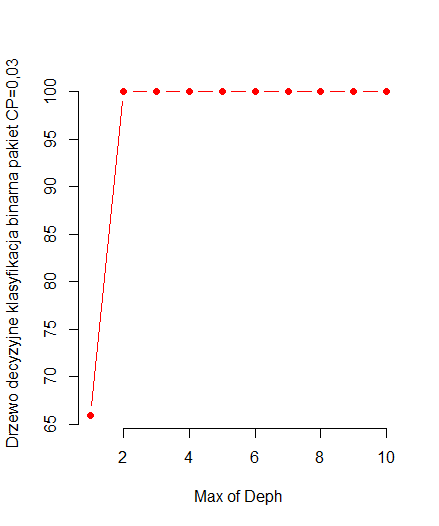
**drzewko\_klw\_pakiet**

|  |  |  |
| --- | --- | --- |
|  | **MaxDeph** | **Trafnosc** |
| [1,] | 1 | 65.97222 |
| [2,] | 2 | 100.00000 |
| [3,] | 3 | 100.00000 |
| [4,] | 4 | 100.00000 |
| [5,] | 5 | 100.00000 |
| [6,] | 6 | 100.00000 |
| [7,] | 7 | 100.00000 |
| [8,] | 8 | 100.00000 |
| [9,] | 9 | 100.00000 |
| [10,] | 10 | 100.00000 |

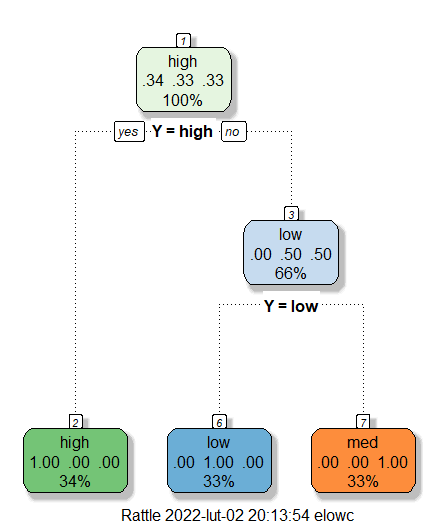
****

**Drzewko\_klw\_pakiet\_cp**

|  |  |  |
| --- | --- | --- |
|  | **MaxDeph** | **Trafnosc** |
| [1,] | 1 | 65.97222 |
| [2,] | 2 | 100.00000 |
| [3,] | 3 | 100.00000 |
| [4,] | 4 | 100.00000 |
| [5,] | 5 | 100.00000 |
| [6,] | 6 | 100.00000 |
| [7,] | 7 | 100.00000 |
| [8,] | 8 | 100.00000 |
| [9,] | 9 | 100.00000 |
| [10,] | 10 | 100.00000 |

****

|  |  |  |  |
| --- | --- | --- | --- |
| **Y\_kl\_w\_pred** | |  |  |
|  | **high** | **low** | **med** |
| high | 141 | 0 | 0 |
| low | 0 | 144 | 0 |
| med | 0 | 0 | 147 |

****

**KNN REGRESJA PAKIET**

**k-Nearest Neighbors**

167 samples

10 predictor

**No pre-processing**

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 149, 150, 151, 151, 149, 151, ...

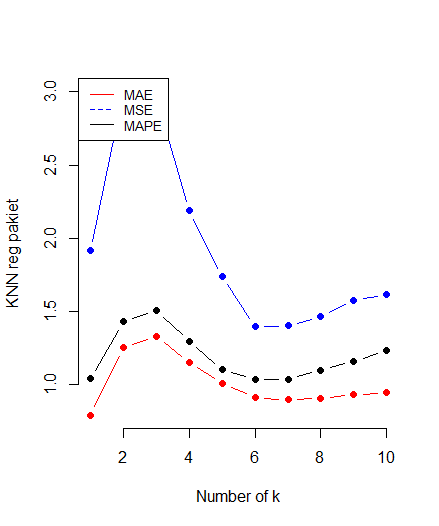
Resampling results across tuning parameters:

|  |  |  |  |
| --- | --- | --- | --- |
| k | **RMSE** | **Rsquared** | **MAE** |
| 1 | **0.8286041** | **0.69020013** | **0.4867758** |
| 2 | 1.3294610 | 0.33107560 | 0.9182543 |
| 3 | 1.6063257 | 0.07380455 | 1.1540598 |
| 4 | 1.5632901 | 0.09023407 | 1.1344123 |
| 5 | 1.4439653 | 0.19493357 | 1.0765526 |
| 6 | 1.2384203 | 0.40779961 | 0.9201752 |
| 7 | 1.1877508 | 0.47150378 | 0.8850425 |
| 8 | 1.1391374 | 0.54449395 | 0.8601151 |
| 9 | 1.1491034 | 0.54499957 | 0.8780776 |
| 10 | 1.1671734 | 0.53573000 | 0.8942730 |

Do wyboru optymalnego modelu wykorzystano RMSE za pomocą najmniejszej wartości. Ostateczna wartość zastosowana w modelu wynosiła k=1.

**knn\_reg\_pakiet**

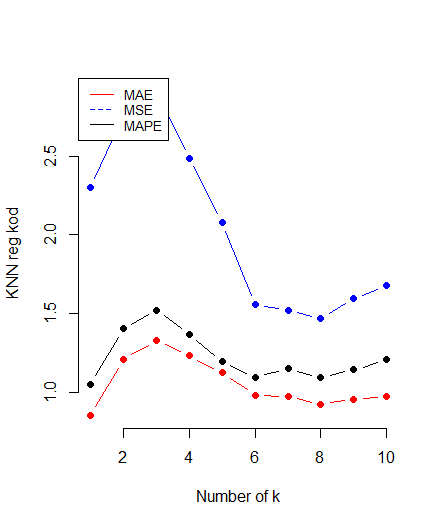
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **K** | **MAE** | **MSE** | **MAPE** |
| [1,] | 1 | 0.7934522 | 1.917881 | 1.041860 |
| [2,] | 2 | 1.2539531 | 3.002105 | 1.429552 |
| [3,] | 3 | 1.3279180 | 2.958937 | 1.507078 |
| [4,] | 4 | 1.1549905 | 2.187888 | 1.298095 |
| [5,] | 5 | 1.0079832 | 1.737547 | 1.105191 |
| [6,] | 6 | 0.9125771 | 1.399922 | 1.039713 |
| [7,] | 7 | 0.8988844 | 1.402185 | 1.039326 |
| [8,] | 8 | 0.9083303 | 1.467117 | 1.101616 |
| [9,] | 9 | 0.9313199 | 1.574483 | 1.162724 |
| [10,] | 10 | 0.9466561 | 1.616924 | 1.232471 |

****

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **MAE** | **MSE** | **MAPE** |
| 6.0000000 | 0.9125771 | 1.3999223 | 1.0397128 |

**KNN REGRESJA – kod własny**

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **MAE** | **MSE** | **MAPE** |
| 8.0000000 | 0.9224332 | 1.4672115 | 1.0919585 |

****

**DRZEWA DECYZYJNE PAKIET REGRESJA**

**Stochastic Gradient Boosting**

167 samples

10 predictor

**No pre-processing**

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 149, 150, 151, 151, 149, 151, ...

Resampling results across tuning parameters:

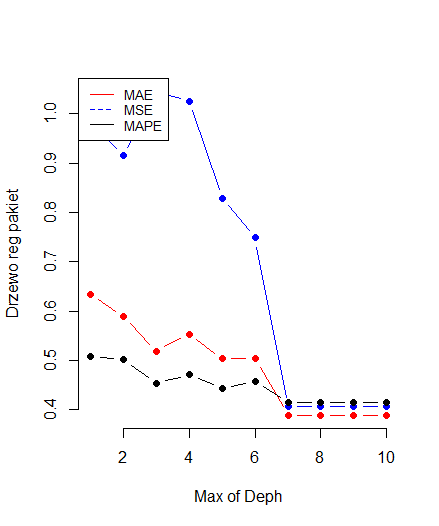
|  |  |  |  |
| --- | --- | --- | --- |
| **interaction.depth** | **n.trees** | **RMSE** | **Rsquared** |
| 1 | 50 | 0.9327305 | 0.6905924 |
| 1 | 100 | 0.8810335 | 0.7147317 |
| 1 | 150 | 0.8489496 | 0.7350909 |
| 2 | 50 | 0.8775027 | 0.7199204 |
| 2 | 100 | 0.8231920 | 0.7522402 |
| 2 | 150 | 0.7901975 | 0.7720234 |
| 3 | 50 | 0.8560103 | 0.7333112 |
| 3 | 100 | 0.7807622 | 0.7764553 |
| 3 | 150 | 0.7332382 | 0.7989652 |

|  |
| --- |
| **MAE** |
| 0.6261856 |
| 0.6210365 |
| 0.6119571 |
| 0.5584920 |
| 0.5475455 |
| 0.5285745 |
| 0.5156108 |
| 0.4886114 |
| 0.4602811 |

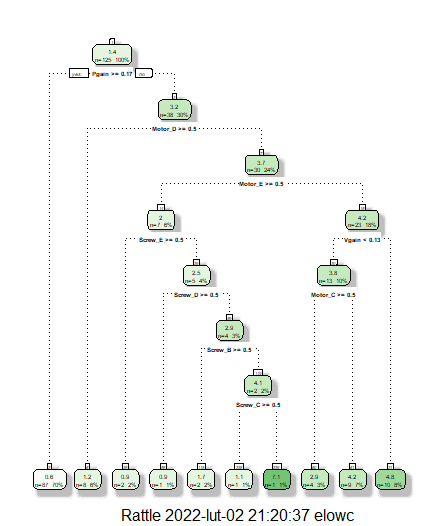
Tunnowany parametr „shrinkage” był utrzymywany na stałym poziomie, a wartość =0,1. Tunnowany paramter „n.minobsinnode” był utrzymywany na poziomie 10. Do wyboru optymalnego modelu wykorzystano RMSE za pomocą najmniejszej wartości. Ostatecznie wartości zastosowane w modelu to n.tree = 150, interaction.depth= 3, shrinkage = 0,1 i n.minobsinnode = 10.

**drzewo\_reg\_pakiet**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Deph** | **MAE** | **MSE** | **MAPE** |
| [1,] | 1 | 0.6341259 | 0.9825113 | 0.5076865 |
| [2,] | 2 | 0.5882523 | 0.9151545 | 0.5005857 |
| [3,] | 3 | 0.5179633 | 1.0452723 | 0.4529302 |
| [4,] | 4 | 0.5531764 | 1.0260530 | 0.4711010 |
| [5,] | 5 | 0.5029114 | 0.8287063 | 0.4429098 |
| [6,] | 6 | 0.5029114 | 0.7487033 | 0.4578171 |
| [7,] | 7 | 0.3886233 | 0.4058329 | 0.4139237 |
| [8,] | 8 | 0.3886233 | 0.4058329 | 0.4139237 |
| [9,] | 9 | 0.3886233 | 0.4058329 | 0.4139237 |
| [10,] | 10 | 0.3886233 | 0.4058329 | 0.4139237 |



|  |  |  |  |
| --- | --- | --- | --- |
| **Deph** | **MAE** | **MSE** | **MAPE** |
| 7.0000000 | 0.3886233 | 0.4058329 | 0.4139237 |



**DRZEWA DECYZYJNE REGRESJA – kod własny**

levelName Count Leaf

1 Root 125

2 ¦--Vgain <= 0.25 68 \*

3 °--Vgain > 0.25 57 \*