WOJSKOWA AKADEMIA TECHNICZNA IM. JAROSŁAWA DĄBROWSKIEGO W WARSZAWIE

***Projekt z przedmiotu „Metody uczenia maszynowego II”***

**Autor:** *Jakub Kapusta*

**Grupa:** *I9B2S4*

**Prowadzący*:*** *dr Jarosław Olejniczak*

**Wybrany język:** R

Spis treści:

[1. Opis wybranego zbioru danych 3](#_Toc62757762)

[2. Analiza wstępna (wizualna) zbioru 4](#_Toc62757763)

[3. Liniowy model regresji 7](#_Toc62757764)

[4. Sieci neuronowe 8](#_Toc62757765)

[5. Drzewa klasyfikacyjne 11](#_Toc62757766)

[6. Bagging 13](#_Toc62757767)

[7. Random Forest 15](#_Toc62757768)

[8. SVM 17](#_Toc62757769)

[9. Podsumowanie utworzonych modeli 18](#_Toc62757770)

[10. Wnioski końcowe 19](#_Toc62757771)

# Opis wybranego zbioru danych

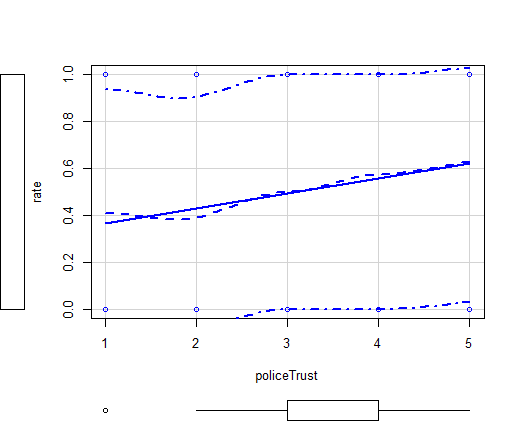
Jako zbiór danych wybrałem **Somerville Happiness Survey Data Set** pobrany ze strony <https://archive.ics.uci.edu/ml/datasets/Somerville+Happiness+Survey>. Zbiór ten przedstawia wyniki ankiety przeprowadzonej w 2015 roku wśród losowych mieszkańców Somerville. Pytania dotyczyły oceny mieszkańców ich szczęścia i satysfakcji ze służb miejskich.

Opis atrybutów podsumowujących wyniki ankiety przedstawia Tabela 1.

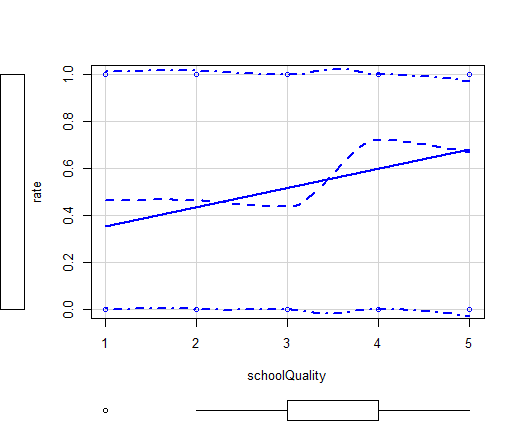
Tabela 1 Opis atrybutów opisujących zbiór danych

|  |  |  |  |
| --- | --- | --- | --- |
| **Oznaczenie atrybutu w pliku** | **Oznaczenie atrybutu w tworzonych modelach** | **Znaczenie atrybutu** | **Możliwe wartości** |
| D | rate | Ocena mieszkańca (czy jest szczęśliwy, czy nie) | 0 (nieszczęśliwy)  1(szczęśliwy) |
| X1 | cityServiceInfoAvailability | Dostępność informacji o służbach miejskich | 1  2  3  4  5 |
| X2 | housingCost | Koszty zakwaterowania |
| X3 | schoolQuality | Jakość szkół publicznych |
| X4 | policeTrust | Zaufanie do lokalnej policji |
| X5 | infrastructureMaintance | Utrzymanie ulic i chodników |
| X6 | eventsAvailability | Dostępność wydarzeń społecznościowych |

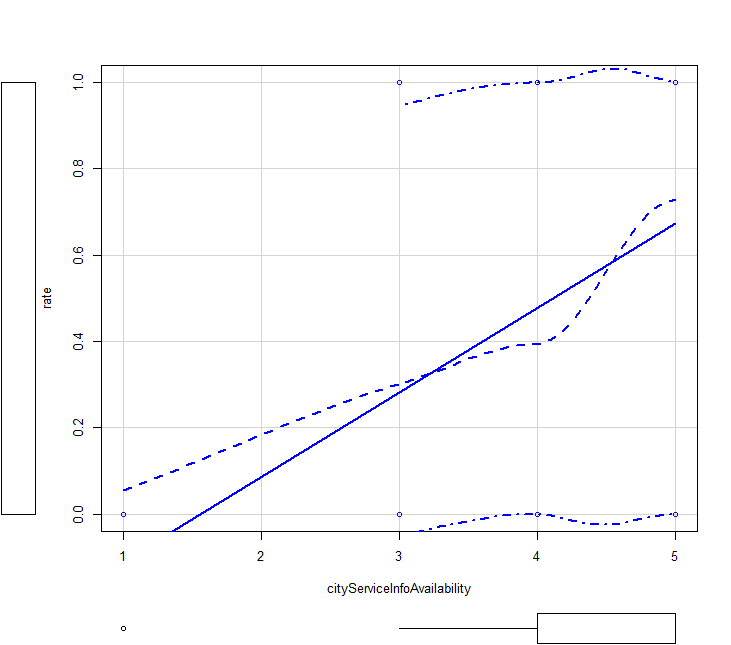
# Analiza wstępna (wizualna) zbioru



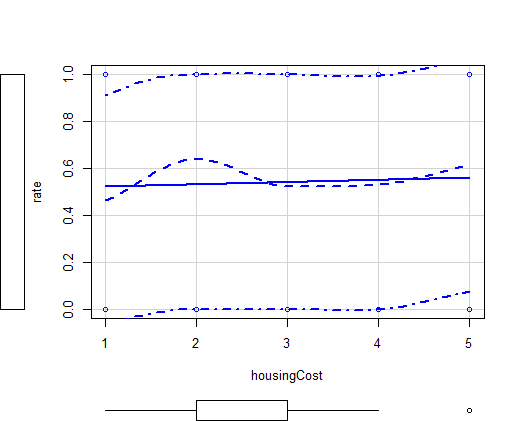
Rysunek 1 Wykres zależności rate od policeTrust. Z wykresu wynika, że wraz ze wzrostem policeTrust wzrasta również rate.



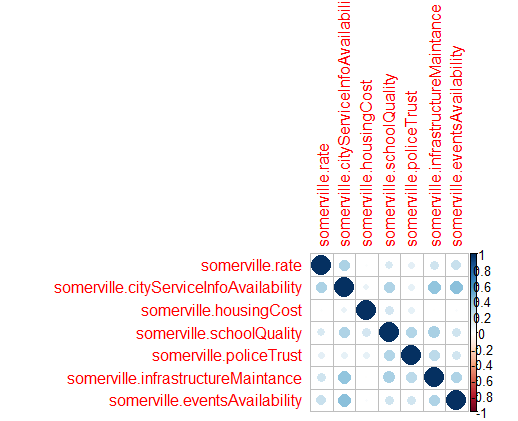
Rysunek 2 Wykres zależności rate od schoolQuality. Z wykresu wynika, że wraz ze wzrostem schoolQuality wzrasta również rate.



Rysunek 3 Wykres zależności rate od cityServiceInfoAvailability. Z wykresu wynika, że wraz ze wzrostem cityServiceInfoAvailability wzrasta również rate.

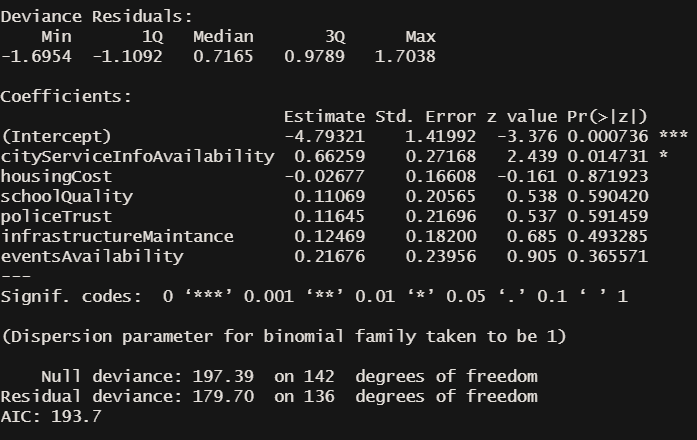


Rysunek 4 Wykres zależności rate od housingCost. Z wykresu wynika, że housingCost nie wpływa na rate.

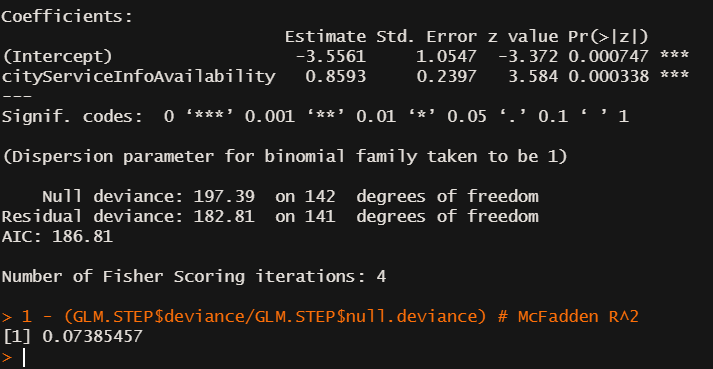


Rysunek 5 Korelacja zmiennych. Z rysunku wynika, że zmienne nie są ze sobą silnie skorelowane. Najsilniej skorelowane ze zmienną objaśnianą jest cityServiceInfoAvailibility.

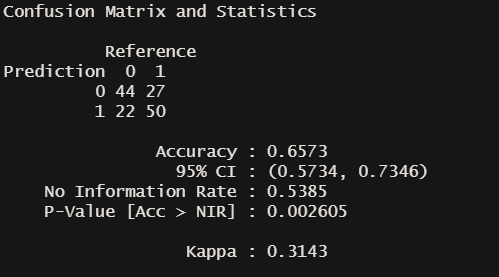
# Liniowy model regresji



Rysunek 6 Podsumowanie liniowego modelu regresji.

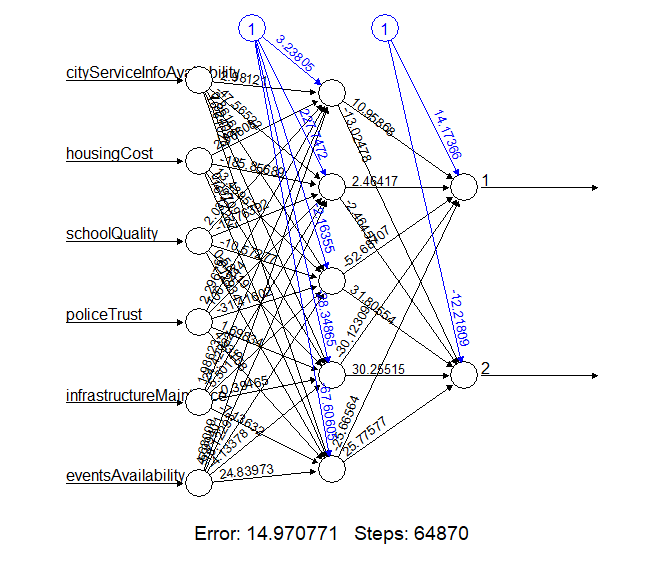


Rysunek 7 Linowy model regresji po redukcji zmiennych nieistotnych.

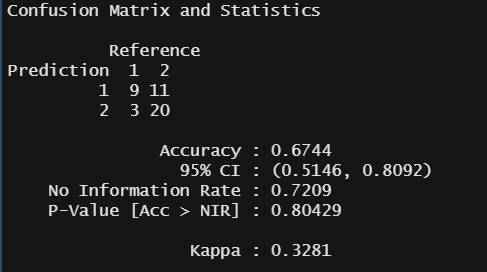


Rysunek 8 Macierz pomyłek dla modelu regresji liniowej (po redukcji zmiennych nieistotnych).

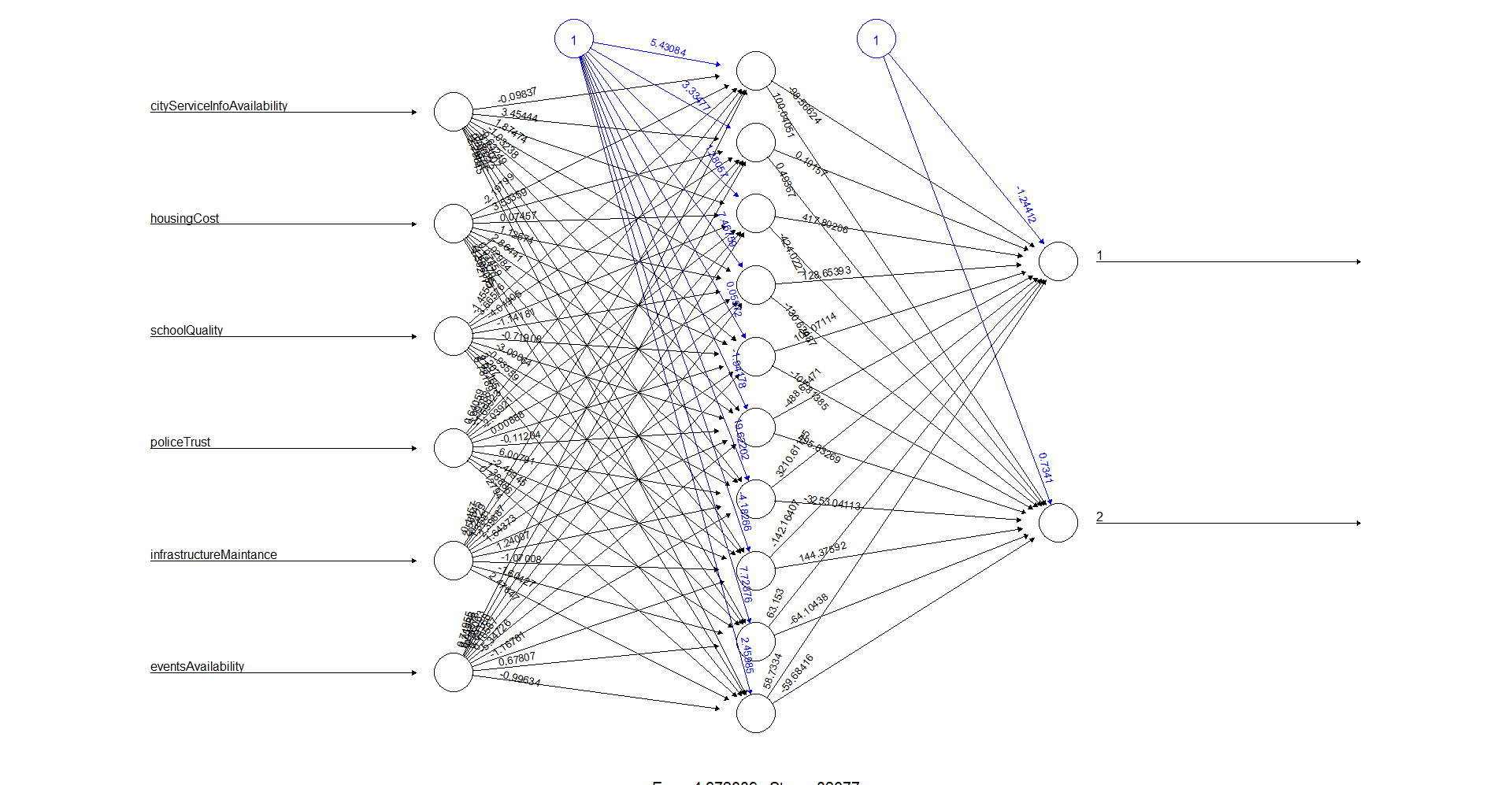
# Sieci neuronowe



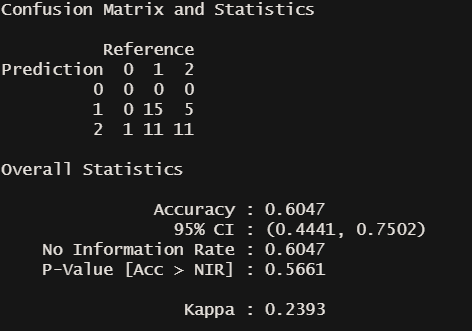
Rysunek 9 Sieć o pięcioelementowej warstwie ukrytej dla pakietu neuralnet



Rysunek 10 Macierz pomyłek dla sieci o strukturze 6-5-2 (dla pakietu neuralnet)

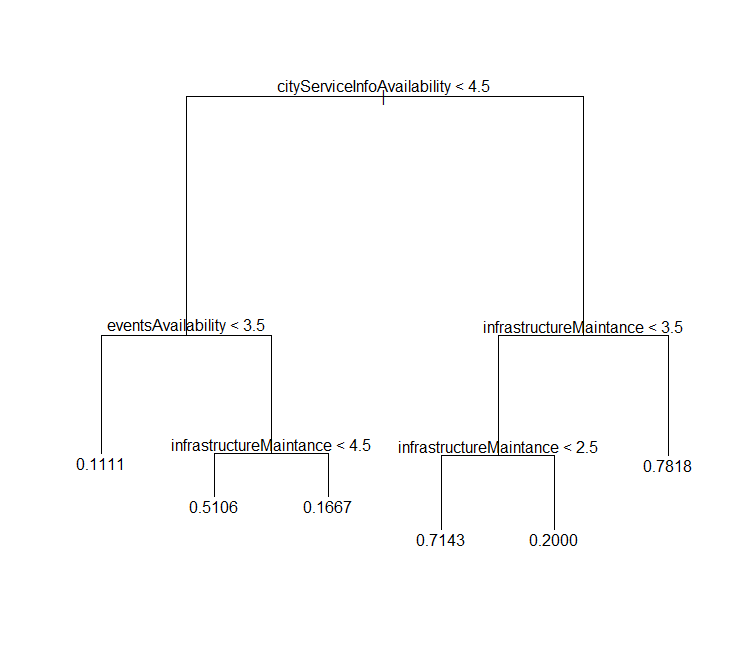


Rysunek 11 Sieć o 10-elementowej warstwie ukrytej dla pakietu neuralnet

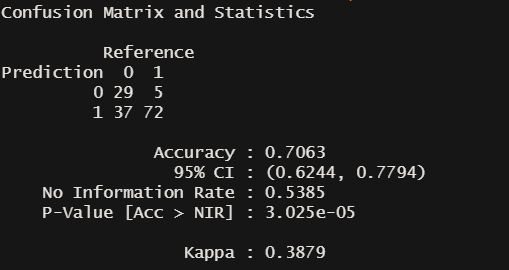


Rysunek 12 Macierz pomyłek dla sieci o strukturze 6-10-2 (dla pakietu neuralnet)

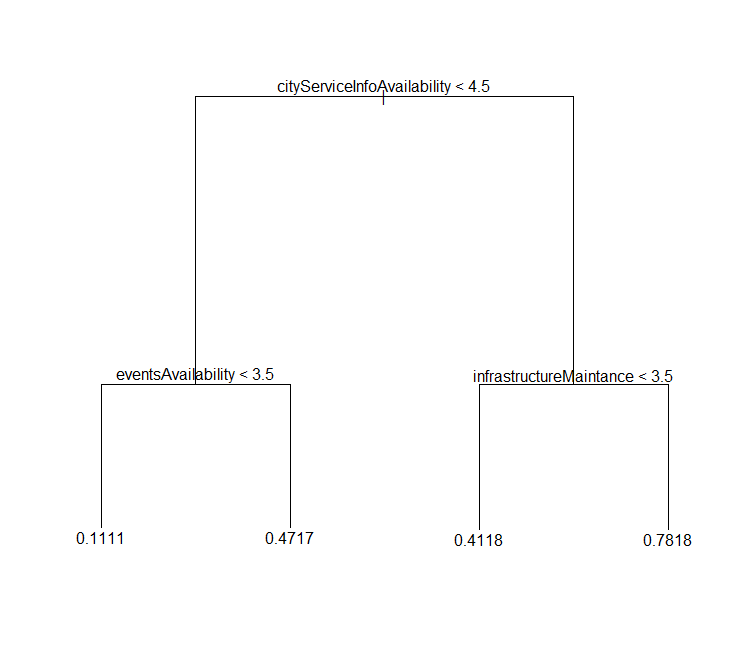
# Drzewa klasyfikacyjne



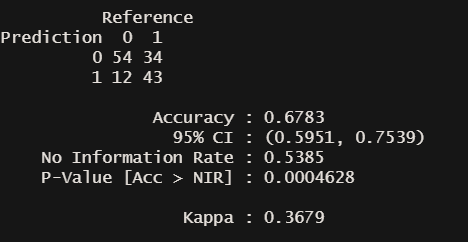
Rysunek 13 Drzewo klasyfikacyjne utworzone za pomocą pakietu tree



Rysunek 14 Macierz pomyłek dla drzewa utworzonego za pomocą pakietu tree

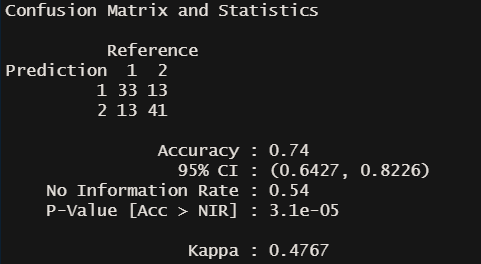


Rysunek 15 Drzewo klasyfikacyjne utworzone za pomocą pakietu tree (przycięte o jeden poziom)

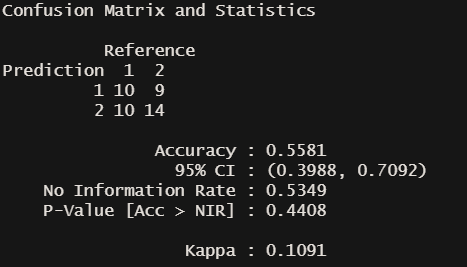


Rysunek 16 Macierz pomyłek dla drzewa utworzonego za pomocą pakietu tree (przyciętego o jeden poziom)

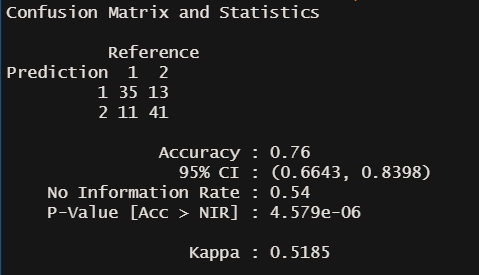
# Bagging



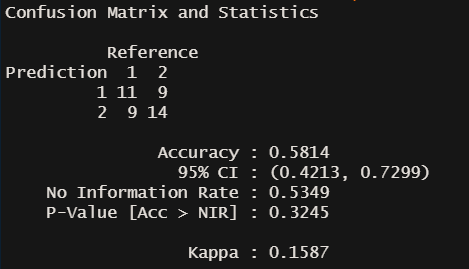
Rysunek 17 Macierz pomyłek dla metody Bagging (nbagg = 150, zbiór uczący)



Rysunek 18 Macierz pomyłek dla metody Bagging (nbagg = 150, zbiór walidacyjny)

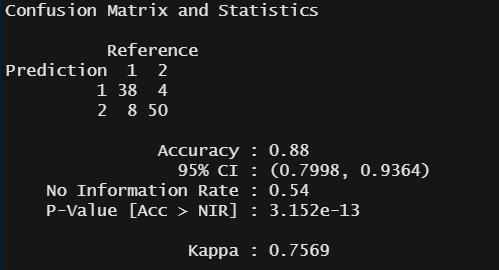


Rysunek 19 Macierz pomyłek dla metody Bagging (nbagg = 50, zbiór uczący)

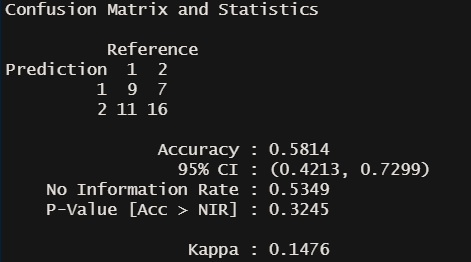


Rysunek 20 Macierz pomyłek dla metody Bagging (nbagg = 50, zbiór walidacyjny)

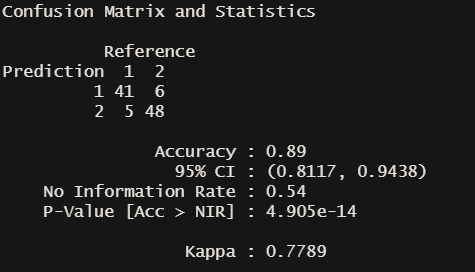
# Random Forest



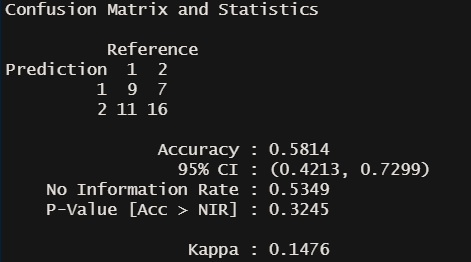
Rysunek 21 Macierz pomyłek dla metody Random Forest (ntrees = 150, zbiór uczący)



Rysunek 22 Macierz pomyłek dla metody Random Forest (ntrees = 150, zbiór walidacyjny)

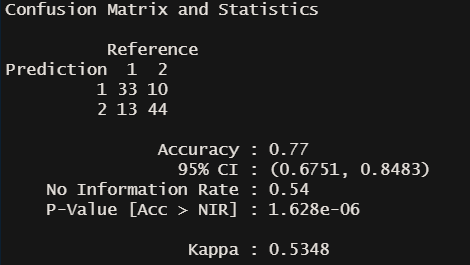


Rysunek 23 Macierz pomyłek dla metody Random Forest (ntrees = 10, zbiór uczący)

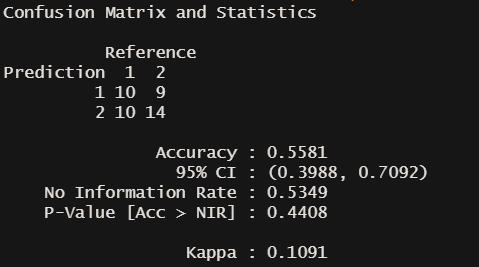


Rysunek 24 Macierz pomyłek dla metody Random Forest (ntrees = 10, zbiór walidacyjny)

# SVM



Rysunek 25 Macierz pomyłek dla metody SVM (zbiór uczący)



Rysunek 26 Macierz pomyłek dla metody SVM (zbiór walidacyjny)

# Podsumowanie utworzonych modeli

Tabela 2 Podsumowanie utworzonych modeli

|  |  |  |  |
| --- | --- | --- | --- |
| **Metoda** | **Pakiet** | **Współczynnik Accuracy** | **Uwagi** |
| Regresja liniowa | - | 0.6573 | Model po redukcji zmiennych nieistotnych przy użyciu algorytmu SVS |
| Sieci neuronowe | neuralnet | 0.6744 | Struktura sieci 6-5-2 |
| 0.6047 | Struktura sieci 6-10-2 |
| Drzewa klasyfikacyjne | tree | 0.7063 | - |
| 0.6783 | Drzewo zostało przycięte o jeden poziom |
| Bagging | ipred | 0.74 (zbiór uczący), 0.5581 (zbiór walidacyjny) | nbagg = 150 |
| 0.76 (zbiór uczący), 0.5814 (zbiór walidacyjny) | nbagg = 50 |
| Random Forest | randomForest | 0.88 (zbiór uczący), 0.5814 (zbiór walidacyjny) | ntrees = 150 |
| 0.89 (zbiór uczący),  0.5814 (zbiór walidacyjny) | ntrees = 10 |
| SVM | e1071 | 0.77 (zbiór uczący), 0.5581 (zbiór walidacyjny) | - |

# Wnioski końcowe

Na podstawie Tabela 2 można stwierdzić, że zgodnie z oczekiwaniami najlepsze wyniki uzyskano dla metody Random Forest. Metoda ta utworzyła model, dla którego Accuracy = 0.89 (dla zbioru uczącego) dla ntrees = 10.

Drugą z najlepszych metod była metoda SVM dla której Accuracy = 0.77, zaś trzecią Bagging (Accuracy = 0.76 dla zbioru uczącego).

Kolejną metodami są drzewa klasyfikacyjne oraz sieci neuronowe. Drzewo pozwoliło na uzyskanie Accuracy = 0.7063, zaś sieć 0.6744.

Na ostatnim miejscu znalazła się regresja liniowa (Accuracy = 0.6573 po redukcji zmiennych nieistotnych przy użyciu algorytmu SVS). Regresja jednak potwierdziła obserwację, którą można było wywnioskować z dwóch postaci drzewa klasyfikacyjnego (Rysunek 13 oraz Rysunek 15) – najbardziej istotną zmienną jest zmienna cityServiceInfoAvailability. Obserwację potwierdza również macierz korelacji (Rysunek 5) – najsilniej skorelowaną zmienną objaśniającą ze zmienną objaśnianą jest również cityServiceInfoAvailability.