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
2. Analiza wstępna (wizualna) zbioru 4
3. Liniowy model regresji 7
4. Sieci neuronowe 8
5. Drzewa klasyfikacyjne 11
6. Bagging 13
7. Random Forest 15
8. SVM 17
9. Podsumowanie utworzonych modeli 18
10. Wnioski końcowe 19

1. Opis wybranego zbioru danych

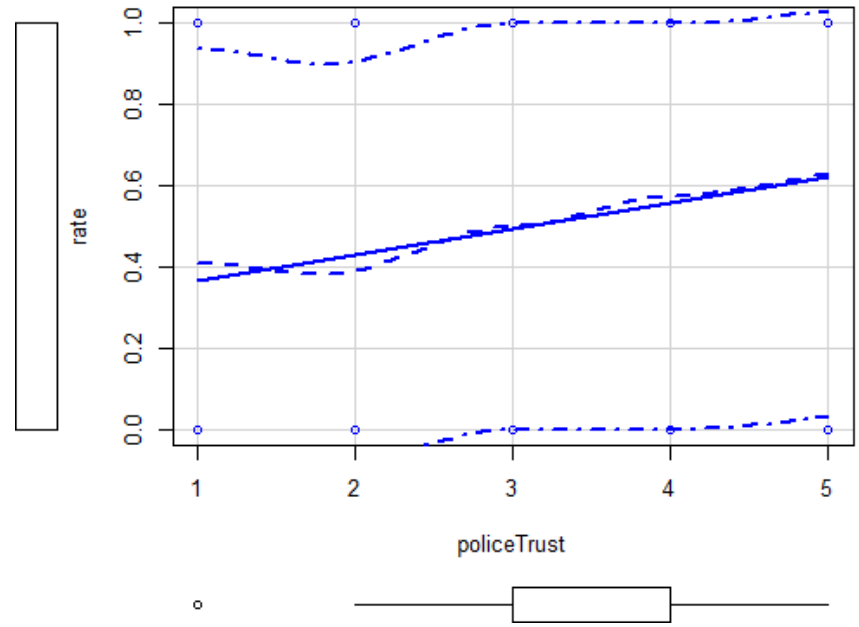
Jako zbiór danych wybrałem **Somerville Happiness Survey Data Set** pobrany ze strony https://archive.ics.uci.edu/ml/datasets/Somerville+Happiness+Survey. Zbiór ten przedstawia wyniki ankiety przeprowadzonej w 2015 roku wśród losowych mieszkańców Somerville. Pytania dotyczyły oceny mieszkańców ich szczęścia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

Opis atrybutów podsumowujących wyniki ankiety przedstawia Tabela 1.

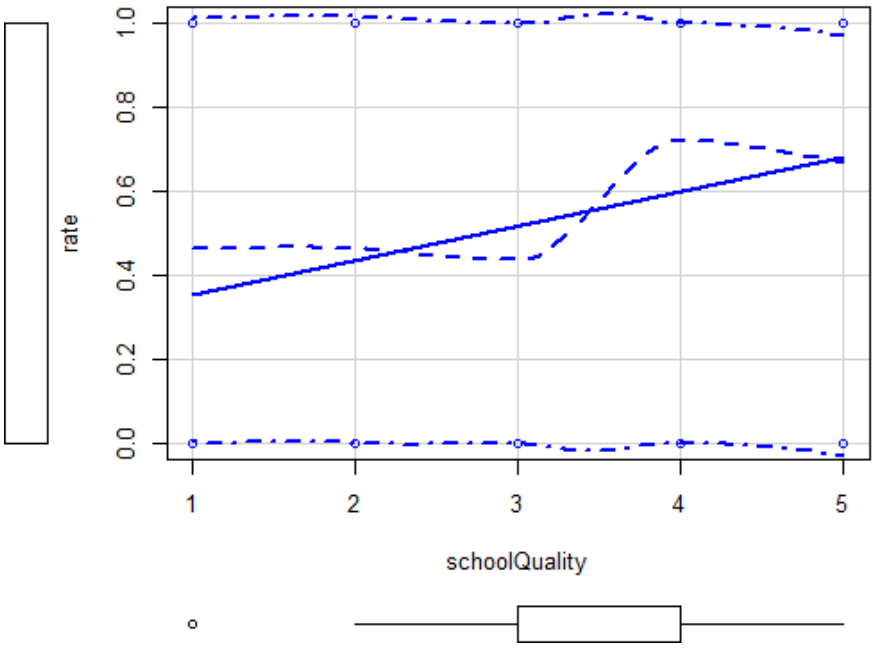
Tabela 1 Opis atrybutów opisujących zbiór danych

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	

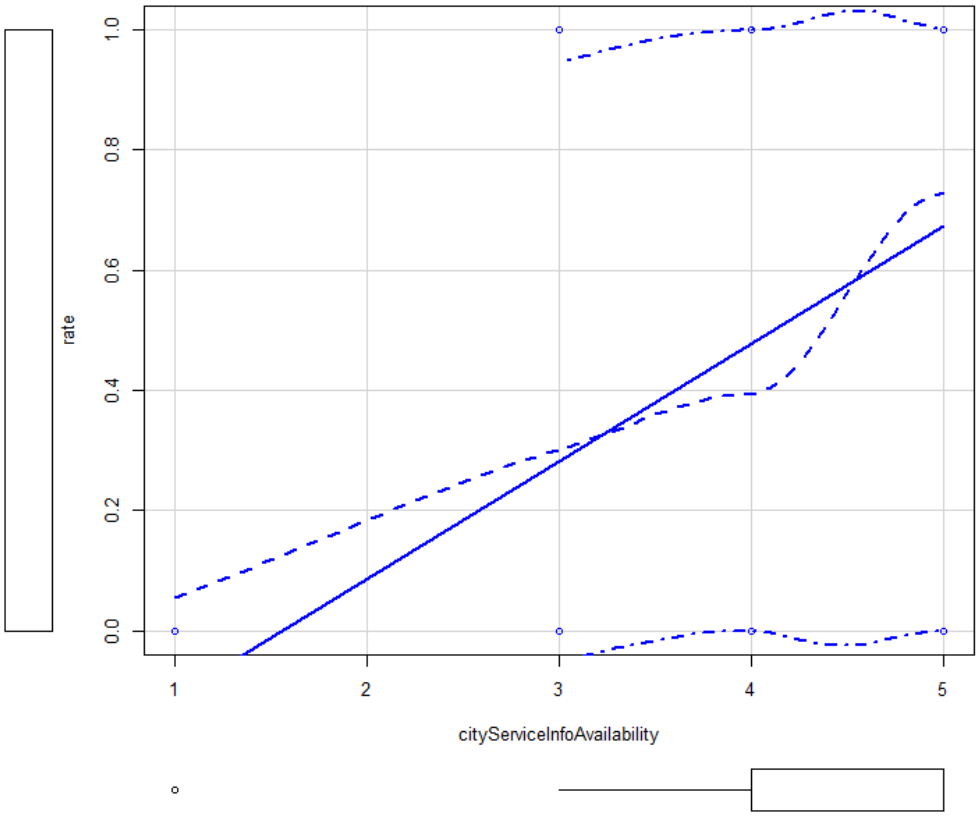
2. Analiza wstępna (wizualna) zbioru



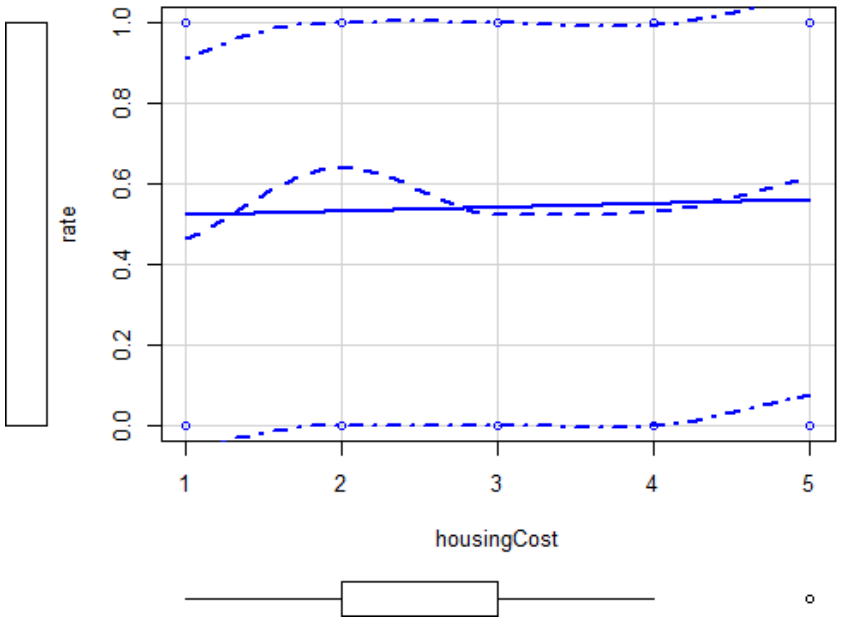
Rysunek 1 Wykres zależności rate od policeTrust. Z wykresu wynika, że wraz ze wzrostem policeTrust wzrasta również rate.



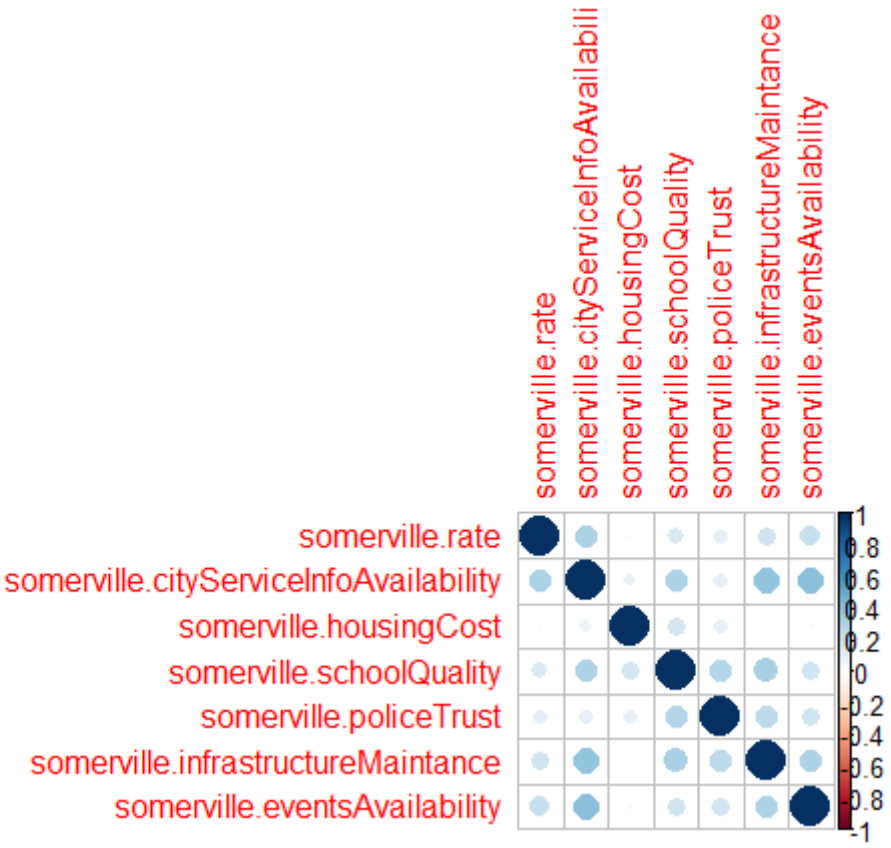
Rysunek 2 Wykres zależności rate od schoolQuality. Z wykresu wynika, że wraz ze wzrostem schoolQuality wzrasta również rate.



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.

3. Liniowy model regresji

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6954  -1.1092   0.7165   0.9789   1.7038

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -4.79321    1.41992  -3.376 0.000736 ***
cityServiceInfoAvailability  0.66259    0.27168   2.439 0.014731 *
housingCost                -0.02677    0.16608  -0.161 0.871923
schoolQuality               0.11069    0.20565   0.538 0.590420
policeTrust                 0.11645    0.21696   0.537 0.591459
infrastructureMaintance     0.12469    0.18200   0.685 0.493285
eventsAvailability          0.21676    0.23956   0.905 0.365571
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 197.39  on 142  degrees of freedom
Residual deviance: 179.70  on 136  degrees of freedom
AIC: 193.7
```

Rysunek 6 Podsumowanie liniowego modelu regresji.

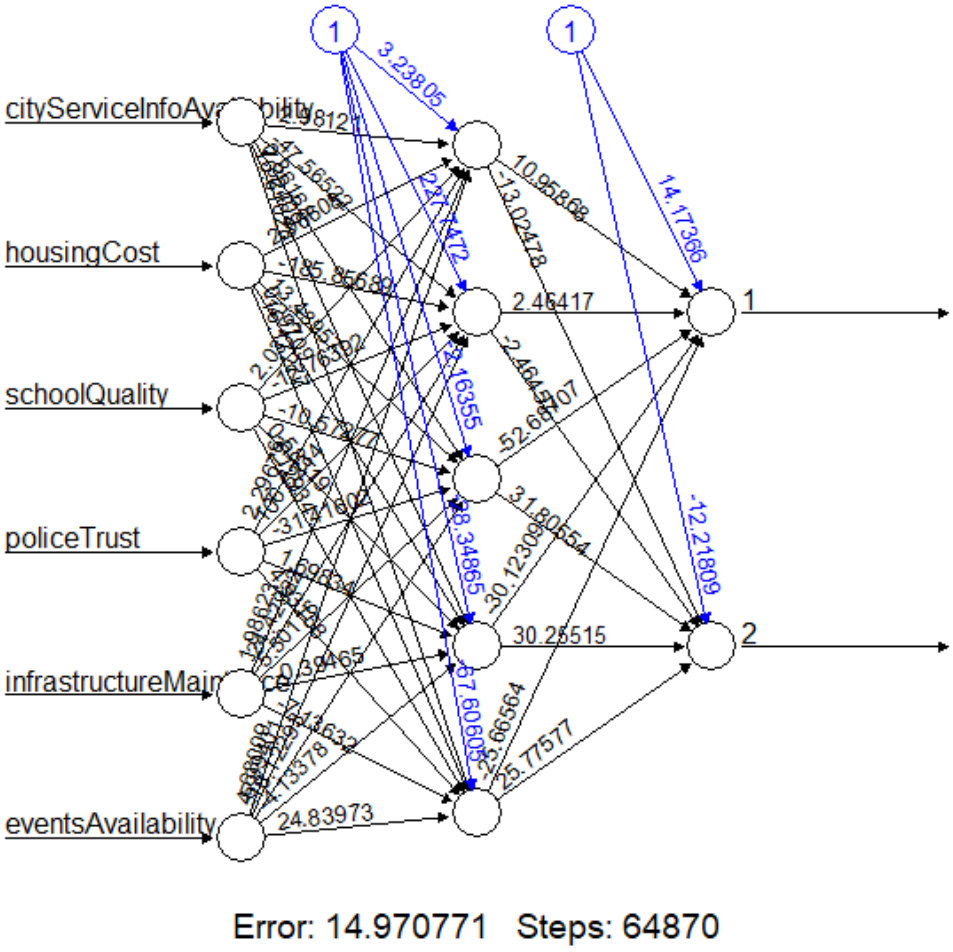
```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.5561     1.0547  -3.372 0.000747 ***
cityServiceInfoAvailability   0.8593     0.2397   3.584 0.000338 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 197.39  on 142  degrees of freedom
Residual deviance: 182.81  on 141  degrees of freedom
AIC: 186.81

Number of Fisher Scoring iterations: 4

> 1 - (GLM.STEP$deviance/GLM.STEP$null.deviance) # McFadden R^2
[1] 0.07385457
> |
```

Rysunek 7 Linowy model regresji po redukcji zmiennych nieistotnych.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 44 27
         1 22 50

               Accuracy : 0.6573
                 95% CI : (0.5734, 0.7346)
    No Information Rate : 0.5385
    P-Value [Acc > NIR] : 0.002605

                  Kappa : 0.3143
```

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

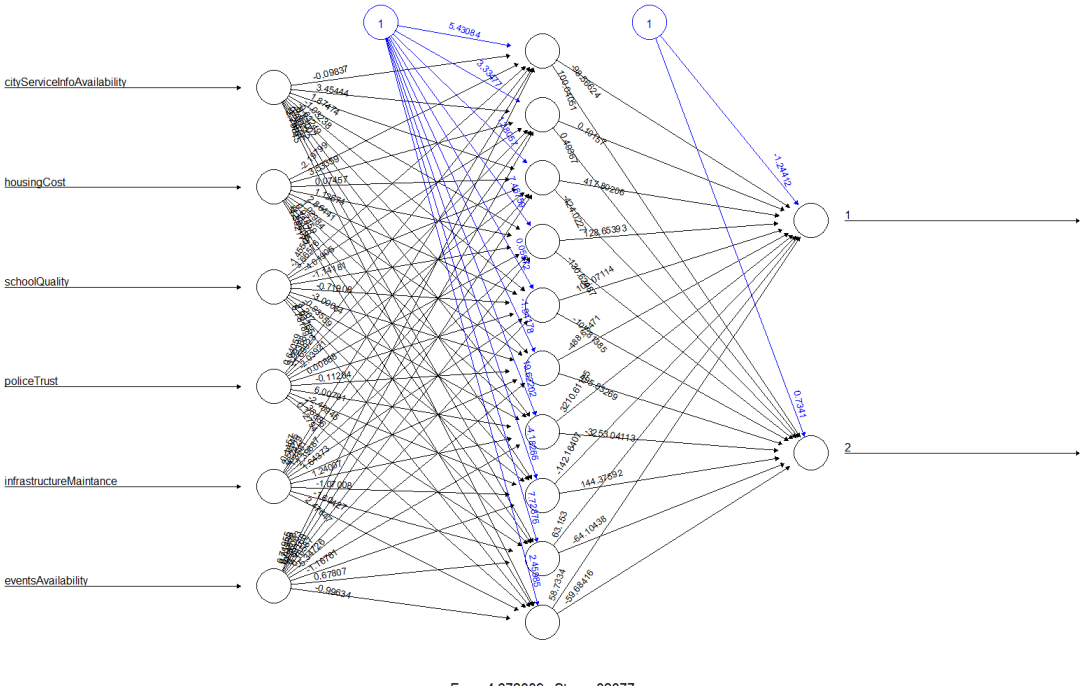
4. Sieci neuronowe



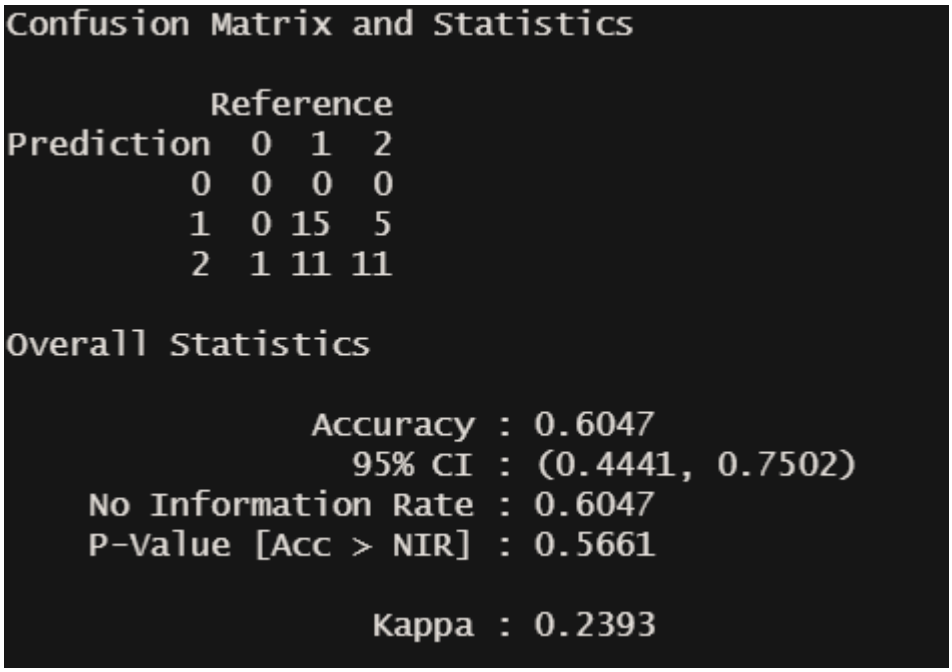
Rysunek 9 Sieć o pięcioelementowej warstwie ukrytej dla pakietu neuralnet

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1  9 11
         2  3 20

               Accuracy : 0.6744
                 95% CI : (0.5146, 0.8092)
    No Information Rate : 0.7209
    P-Value [Acc > NIR] : 0.80429

                  Kappa : 0.3281
```

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

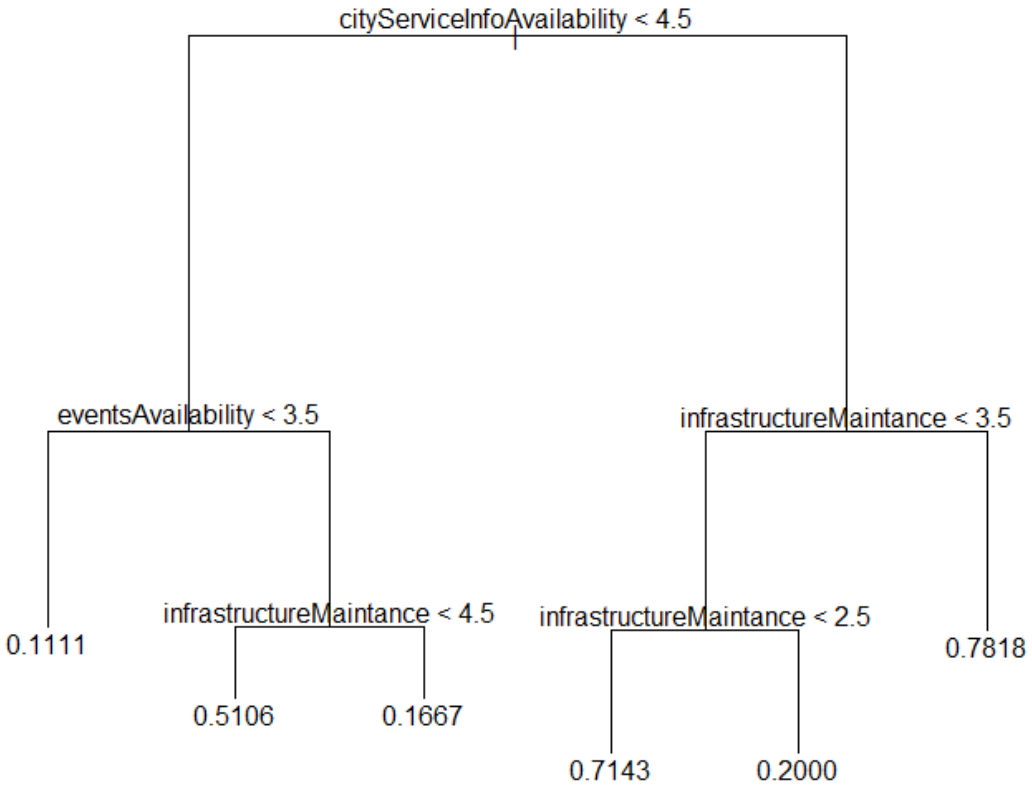


Rysunek 11 Sieć o 10-elementowej warstwie ukrytej dla pakietu neuralnet



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

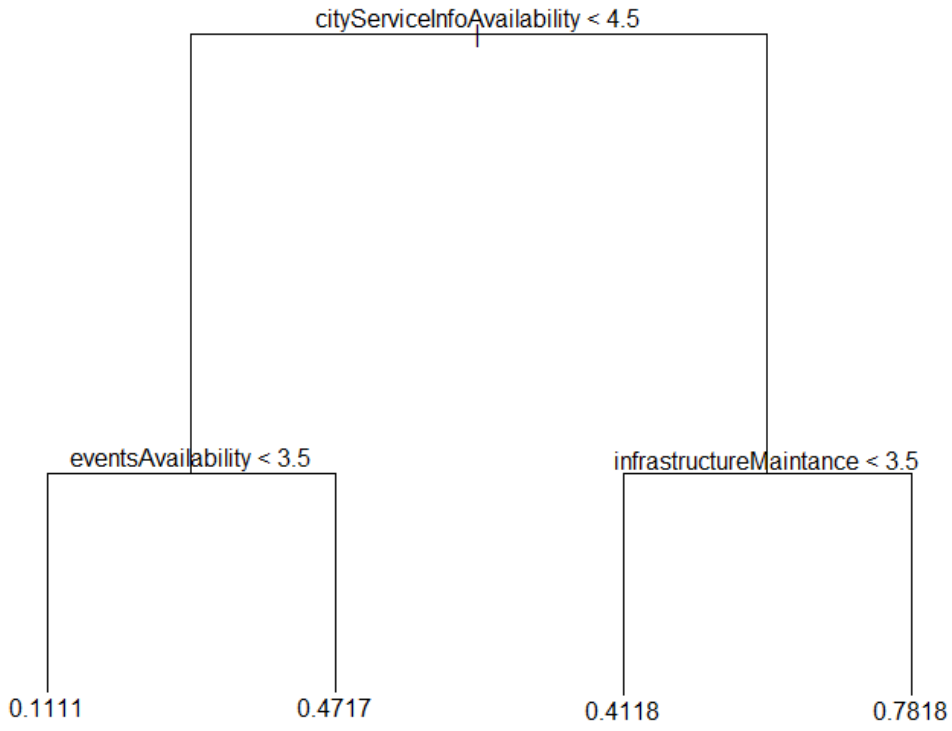
5. Drzewa klasyfikacyjne



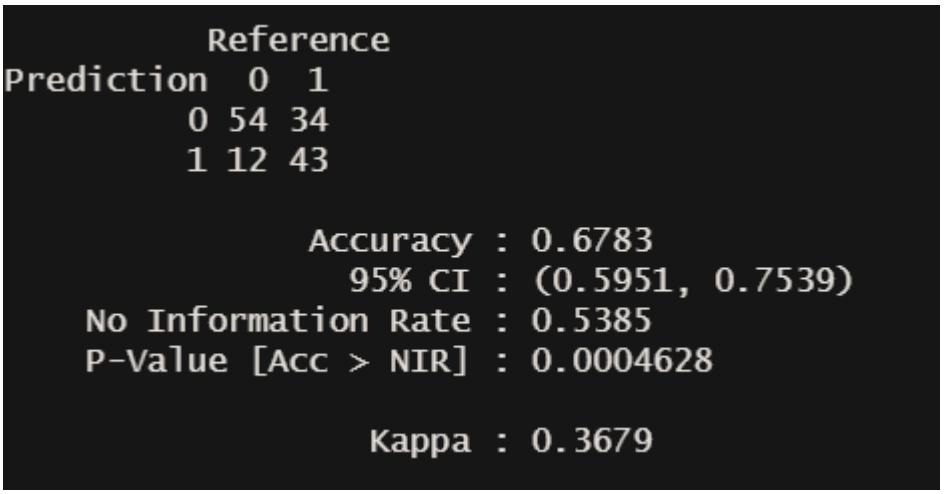
Rysunek 13 Drzewo klasyfikacyjne utworzone za pomocą pakietu tree

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 29  5
         1 37 72

               Accuracy : 0.7063
                 95% CI : (0.6244, 0.7794)
    No Information Rate : 0.5385
    P-Value [Acc > NIR] : 3.025e-05

                  Kappa : 0.3879
```

Rysunek 14 Macierz pomyłek dla drzewa utworzonego za pomocą pakietu tree

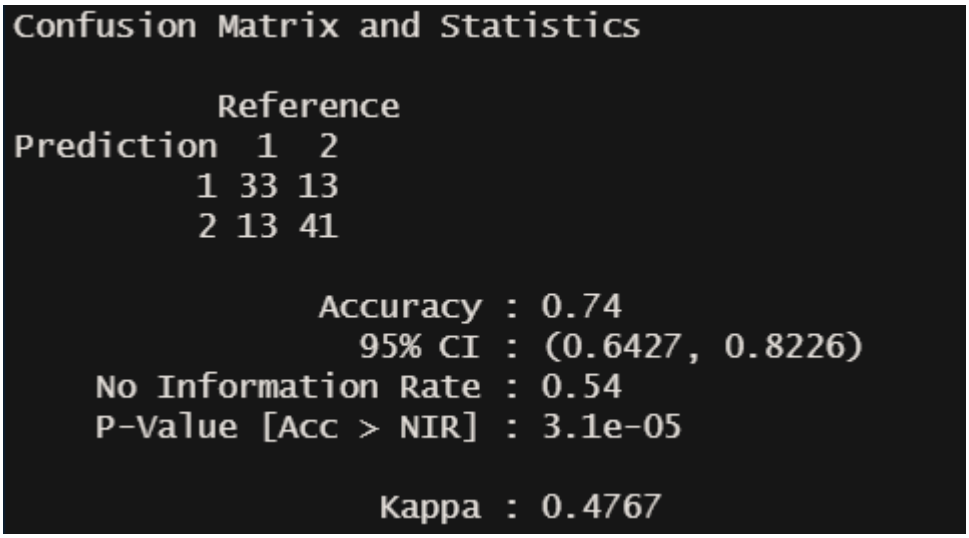


Rysunek 15 Drzewo klasyfikacyjne utworzone za pomocą pakietu tree (przycięte o jeden poziom)

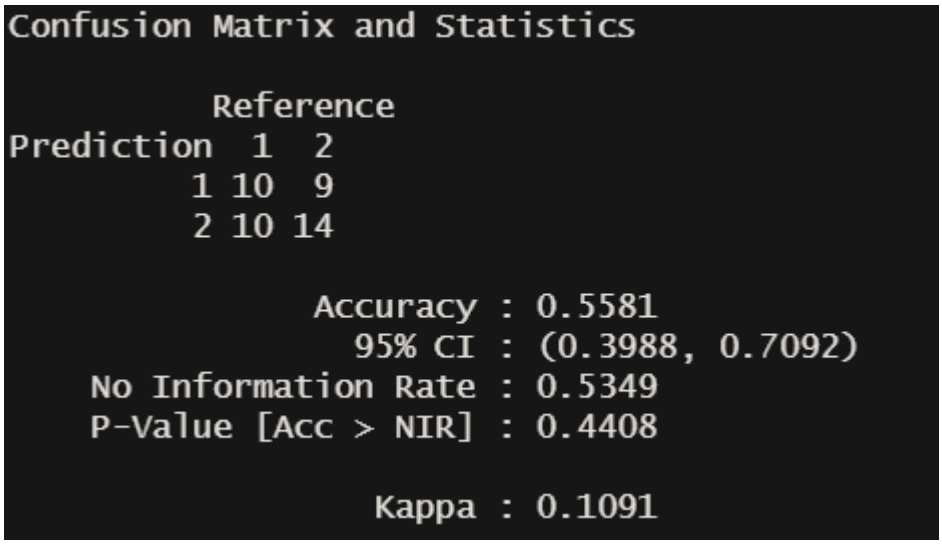


Rysunek 16 Macierz pomyłek dla drzewa utworzonego za pomocą pakietu tree (przyciętego o jeden poziom)

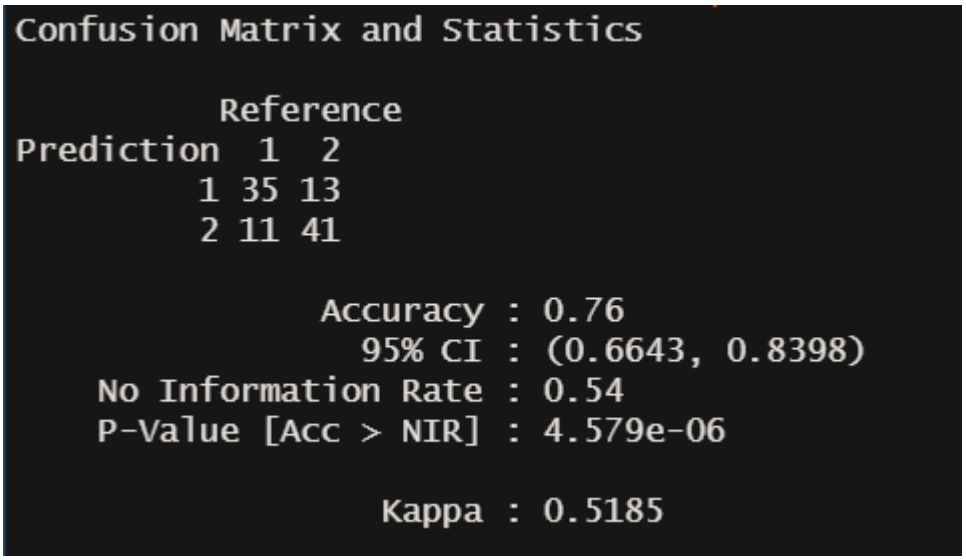
6. Bagging



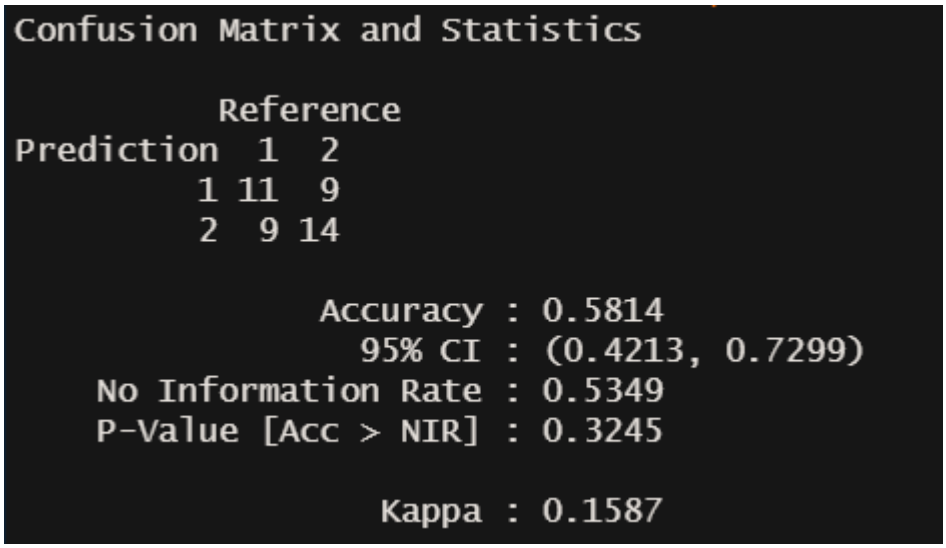
Rysunek 17 Macierz pomyłek dla metody Bagging (nbagg = 150, zbiór uczący)



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

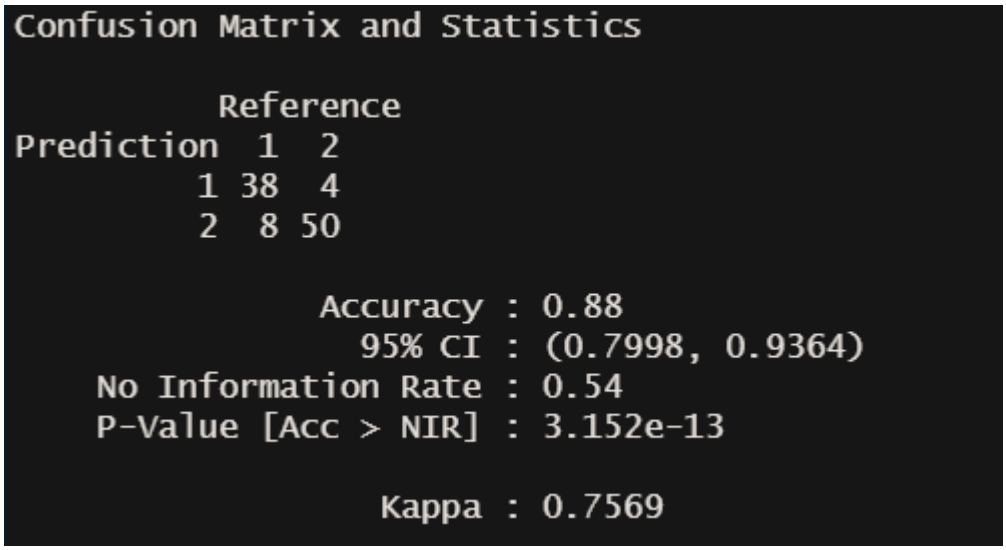


Rysunek 19 Macierz pomyłek dla metody Bagging (nbagg = 50, zbiór uczący)

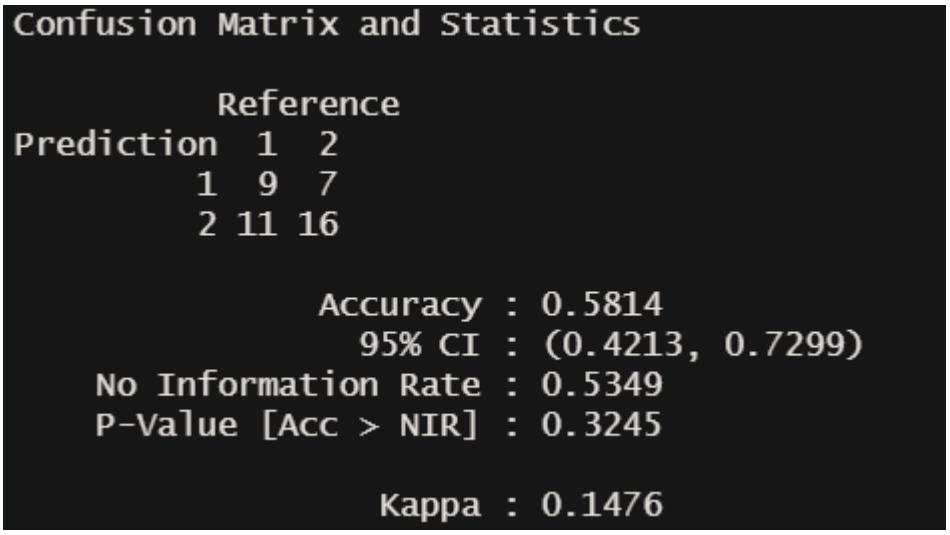


Rysunek 20 Macierz pomyłek dla metody Bagging (nbagg = 50, zbiór walidacyjny)

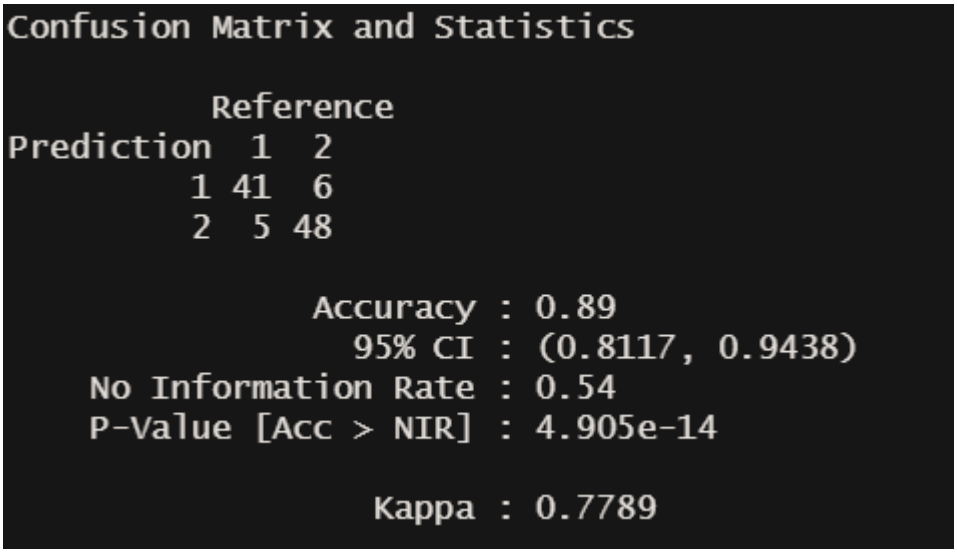
7. Random Forest



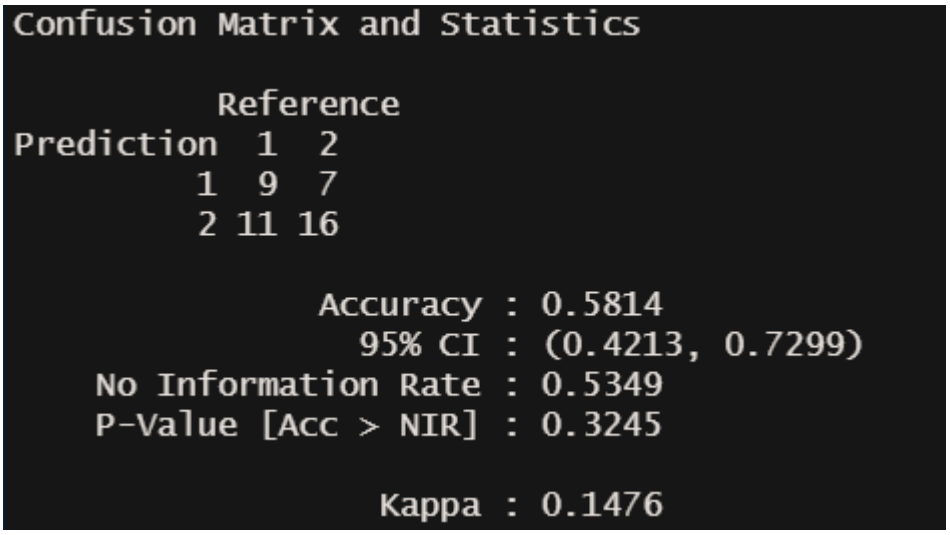
Rysunek 21 Macierz pomyłek dla metody Random Forest (ntrees = 150, zbiór uczący)



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

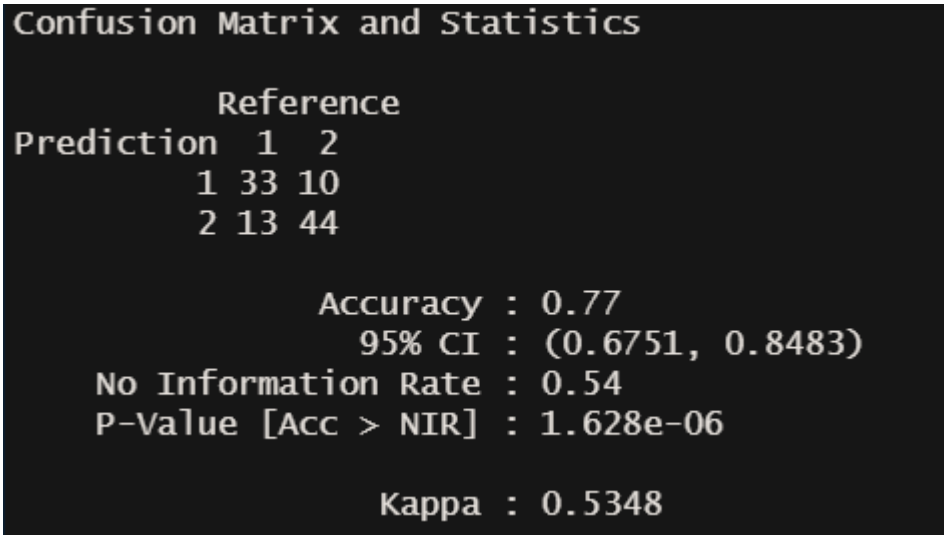


Rysunek 23 Macierz pomyłek dla metody Random Forest (ntrees = 10, zbiór uczący)

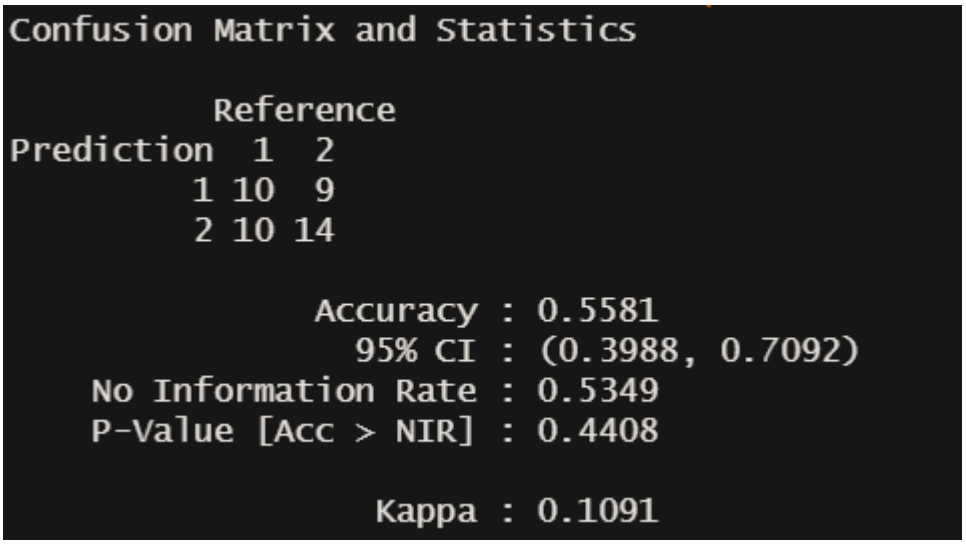


Rysunek 24 Macierz pomyłek dla metody Random Forest (ntrees = 10, zbiór walidacyjny)

8. SVM



Rysunek 25 Macierz pomyłek dla metody SVM (zbiór uczący)



Rysunek 26 Macierz pomyłek dla metody SVM (zbiór walidacyjny)

9. 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)	

10. 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.