

UNIwersytet Ekonomiczny w Katowicach

Przedmiot:

Inżynieria procesu odkrywania wiedzy

Temat:

Projekt zaliczeniowy – SAS Enterprise Miner

Prowadzący:

dr Mariusz Żytniewski

Anna Krzyżowska – 139503

Informatyka i Ekonometria,

Analityka Danych,

Rok 2, semestr 4

Spis treści

Wstęp.....	3
Metody prognozowania.....	4
Dane	4
Przygotowanie	5
Wstępna eksploracja danych	6
Korelacja zmiennych.....	9
Partycjonowanie danych.....	11
Drzewo decyzyjne	13
Regresja logistyczna.....	16
Porównanie modeli.....	19
Metody grupowania.....	23
Dane	23
Analiza skupień	25
Sieci neuronowe Kohonena.....	34
Podsumowanie	38
Spis ilustracji	39
Bibliografia.....	40

Wstęp

Praca została przygotowana na przedmiot „Inżynieria procesu odkrywania wiedzy”. Podzieliłam ją na dwa główne rozdziały: prognozowanie i grupowanie. Do stworzenia tej pracy posłużyłam się programami: SAS Base – w celu zaimportowania bazy do bibliotek Sas-owych oraz SAS Enterprise Miner Workstation w celu porównania drzewa decyzyjnego oraz regresji logistycznej. Która z metod okaże się skuteczniejsza dla zadanych danych? Drugą część pracy przeznaczyłam na pokazanie metod grupowania danych i jakie wnioski można z nich wyciągnąć to jest: analiza skupień i sieci neuronowych Kohonena.



Rysunek 1 Logo SAS

Metody prognozowania

Modele predykcyjne dają możliwość oszacowania prawdopodobieństwa wystąpienia danego zjawiska dla określonych zmiennych na podstawie badania podmiotów występujących w próbie. Wskazane jest budowanie i porównywanie różnych modeli predykcyjnych, w tym: modele regresji logistycznej, sieci neuronowe i drzewa decyzyjne. Każdy z tych wymienionych modeli będzie prawdopodobnie figurować w kilku przypadkach różniąc się metodą estymacji i parametrami.

Dane

Zbiór danych pochodzi ze strony <https://www.kaggle.com/uciml/adult-census-income>. Dane te zostały pobrane z bazy danych Biura Spisu Powszechnego z 1994 r. przez Ronny'ego Kohaviego i Barry'ego Beckera (Data Mining and Visualization, Silicon Graphics).

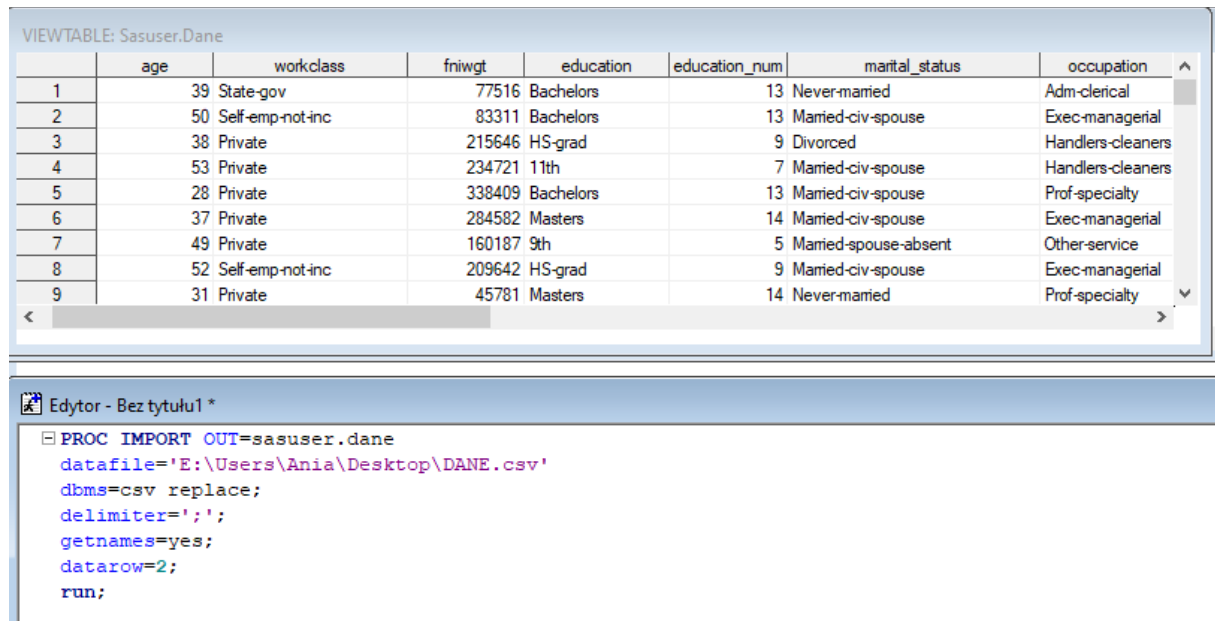
Celem tego rozdziału jest znalezienie jak najlepszego modelu prognozowania, wybierając pomiędzy drzewem decyzyjnym, a regresją logistyczną, która ma za zadanie ustalić czy roczny dochód danej osoby przekroczy 50 000 dolarów w oparciu o zadane zmienne.

Zmienne znajdujące się w bazie:

- Target – cel, 1-jeśli przekracza 50000 USD, 0-nie przekracza 50000USD;
- Age – wiek podmiotu;
- Workclass – klasa zawodowa;
- Fnlwgt – ID podmiotu;
- Education – poziom wykształcenia;
- Education-num – ilość lat kształcenia;
- Marital status – stan cywilny;
- Occupation – zawód;
- Relationship – relacja w rodzinie;
- Race – pochodzenie etniczne;
- Sex – płeć;
- Capital-gain – zyski kapitałowe;
- Capital-loss – strata kapitałowa;
- Hours-peer-week – godziny tygodniowo;
- Native-country – kraj pochodzenia;

Przygotowanie

Dane są w formacie csv. Zostały załadowane do programu SAS Enterprise Miner poprzez program SAS Base:



VIEWTABLE: Sasuser.Dane

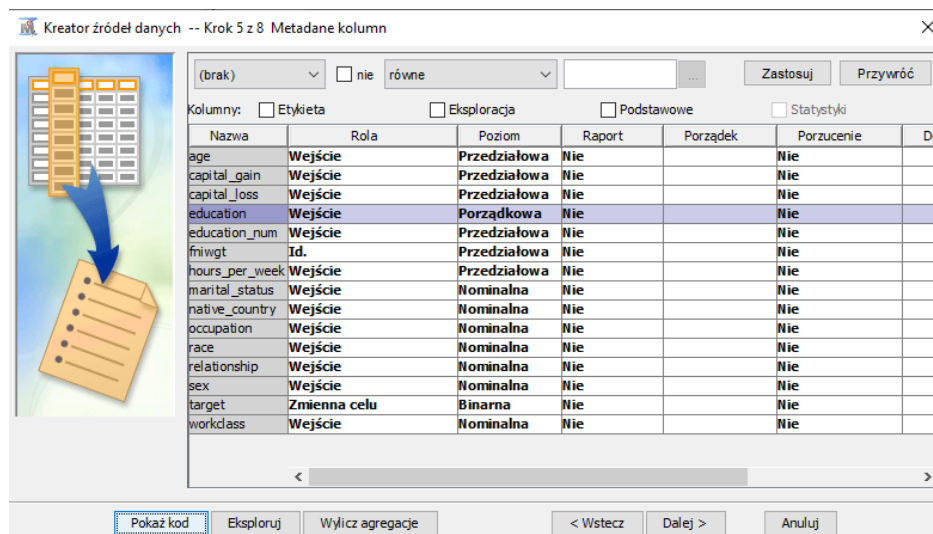
	age	workclass	fniwgt	education	education_num	marital_status	occupation
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial
7	49	Private	160187	9th	5	Married-spouse-absent	Other-service
8	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial
9	31	Private	45781	Masters	14	Never-married	Prof-specialty

Edytor - Bez tytułu *

```
PROC IMPORT OUT=sasuser.dane
datafile='E:\Users\Ania\Desktop\DA NE.csv'
dbms=csv replace;
delimiter=';';
getnames=yes;
datarow=2;
run;
```

Rysunek 2 Import bazy do tabel Sas-owych

Zmienną celu została zmienna binarna ‘target’, która określa czy dana osoba przekroczyła próg zarobkowy 50 tysięcy dolarów. Edukacja została wybrana jako porządkowa, ponieważ zakładamy, że różne etapy edukacji mają różną rangę (szkoła, liceum, college, itp.). Jako nominalne wybrano klasę zawodową, pochodzenie, kraj, stan cywilny, zawód i relacje międzyludzkie, aby reprezentować kategoriyczny charakter zmiennej, a pozostałe zmienne zostały wybrane jako przedziałowe. Id została zmienna ‘fniwgt’.



Kreator źródeł danych -- Krok 5 z 8 Metadane kolumn

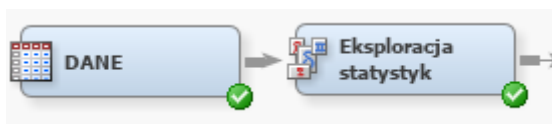
(brak) ☐ nie ☐ równe ☐ Zastosuj ☐ Przywróć

Kolumny: ☐ Etykieta ☐ Eksploracja ☐ Podstawowe ☐ Statystyki

Nazwa	Rola	Poziom	Raport	Porządek	Porzucenie	Do
age	Wejście	Przedziałowa	Nie		Nie	
capital_gain	Wejście	Przedziałowa	Nie		Nie	
capital_loss	Wejście	Przedziałowa	Nie		Nie	
education	Wejście	Porządkowa	Nie		Nie	
education_num	Wejście	Przedziałowa	Nie		Nie	
fniwgt	Id.	Przedziałowa	Nie		Nie	
hours_per_week	Wejście	Przedziałowa	Nie		Nie	
marital_status	Wejście	Nominalna	Nie		Nie	
native_country	Wejście	Nominalna	Nie		Nie	
occupation	Wejście	Nominalna	Nie		Nie	
race	Wejście	Nominalna	Nie		Nie	
relationship	Wejście	Nominalna	Nie		Nie	
sex	Wejście	Nominalna	Nie		Nie	
target	Zmienna celu	Binarna	Nie		Nie	
workclass	Wejście	Nominalna	Nie		Nie	

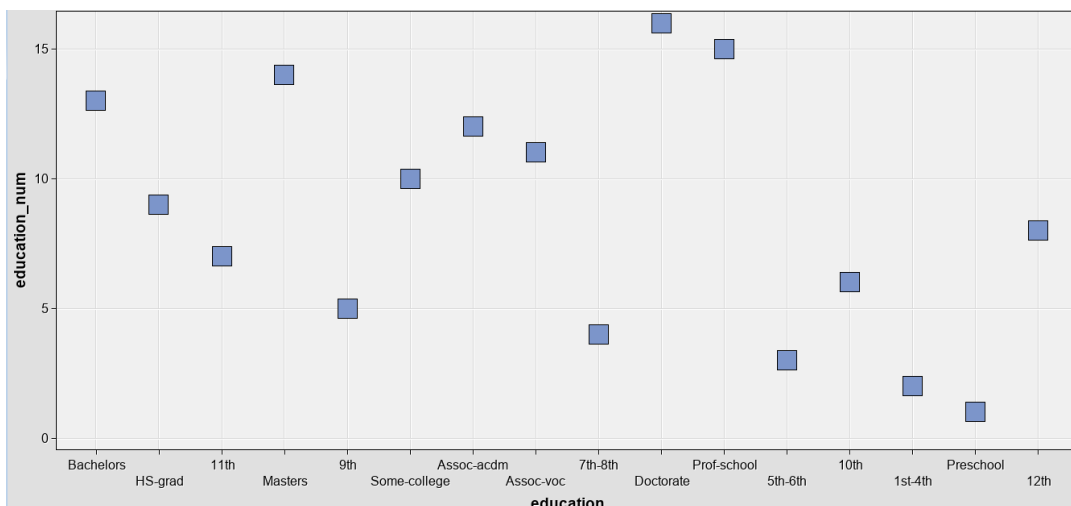
Rysunek 3 Rola i poziomy dla zmiennych

Wstępna eksploracja danych



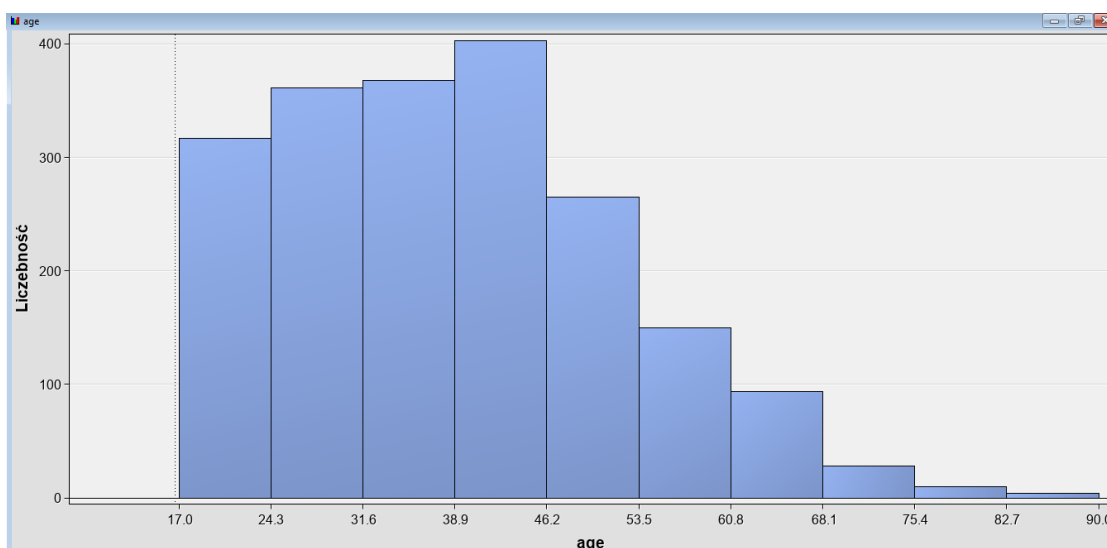
Rysunek 4 Węzeł - eksploracja statystyk

Wstępna eksploracja danych została przeprowadzona w celu wstępnego zrozumienia zbioru danych. Będę badać korelacje pomiędzy zmiennymi, aby lepiej zrozumieć zbiór danych.



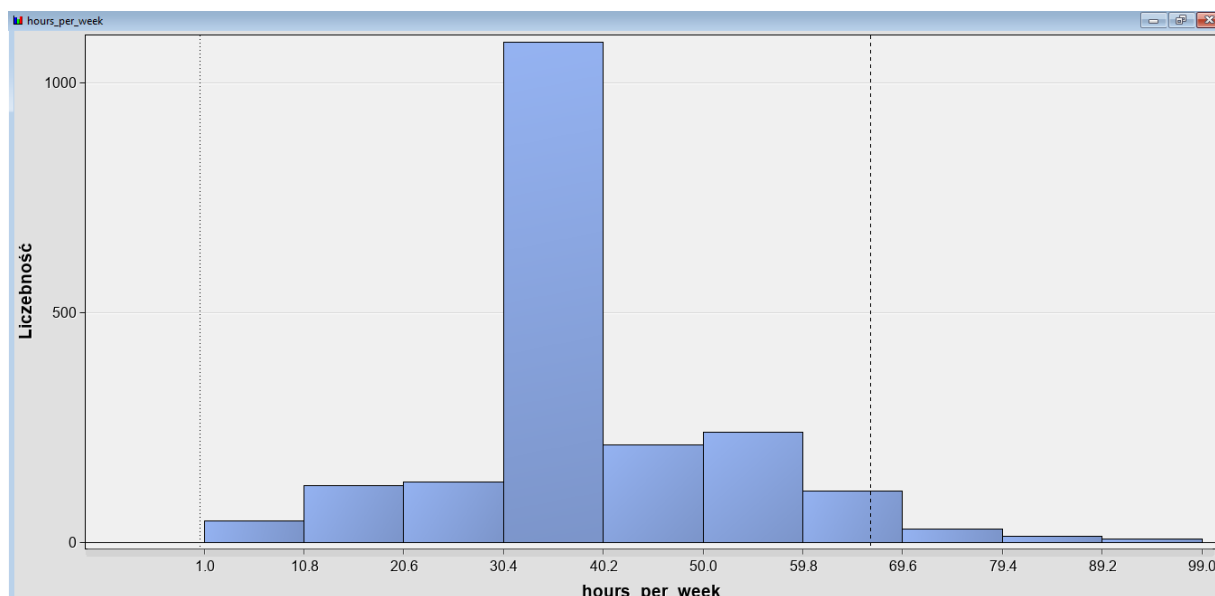
Rysunek 5 Wykres rozrzutu dla poziomu wykształcenia i lat nauki

Wykres rozproszenia pokazał, że ilość lat nauki jest powiązana z poziomem wykształcenia. Najwyższym poziomem wykształcenia jest stopień doktora, następnie profesora, magistra, a następnie licencjata, zgodnie z najwyższą liczbą lat nauki. Wykazało to pozytywną korelację pomiędzy tymi dwoma zmiennymi.



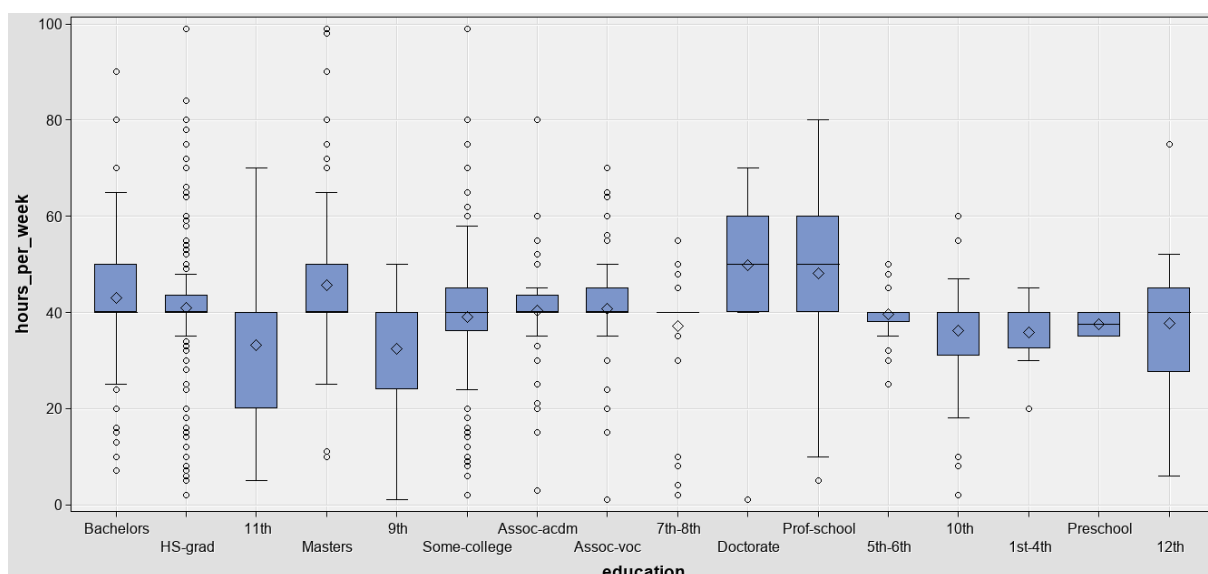
Rysunek 6 Histogram zmiennej wiek

Na podstawie wyników badań histogramu wieku: najmłodszy wiek osoby badanej wynosił 17 lat, a najwyższy wiek około 90 lat. Większość populacji w tym wieku przypada na przedział wiekowy 38,9-46,2 roku – ponad 400 osób, a następnie 31,6-38,9, 24,3-31,6 i 17-24,3 roku. Te trzy grupy wiekowe liczą ponad 300 osób. Najniższa liczba ludności w tym przedziale wiekowym wynosiła 82,7-90 lat.



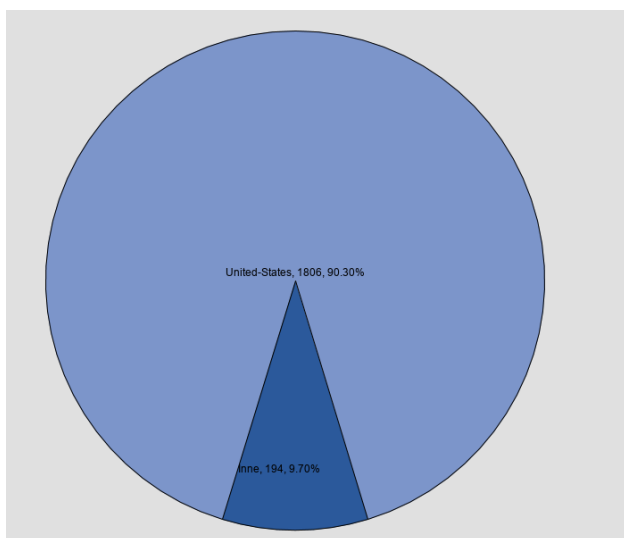
Rysunek 7 Histogram pracujących zmiennych w tygodniu

W oparciu o ‘godziny tygodniowo’, większość badanych poświęca około 30-40 godzin tygodniowo na pracę, co jest uważane za powszechnie spędzany czas w pracy. A ta próbka godzin pracy stanowiła ponad 50% całej zawartej tutaj populacji.



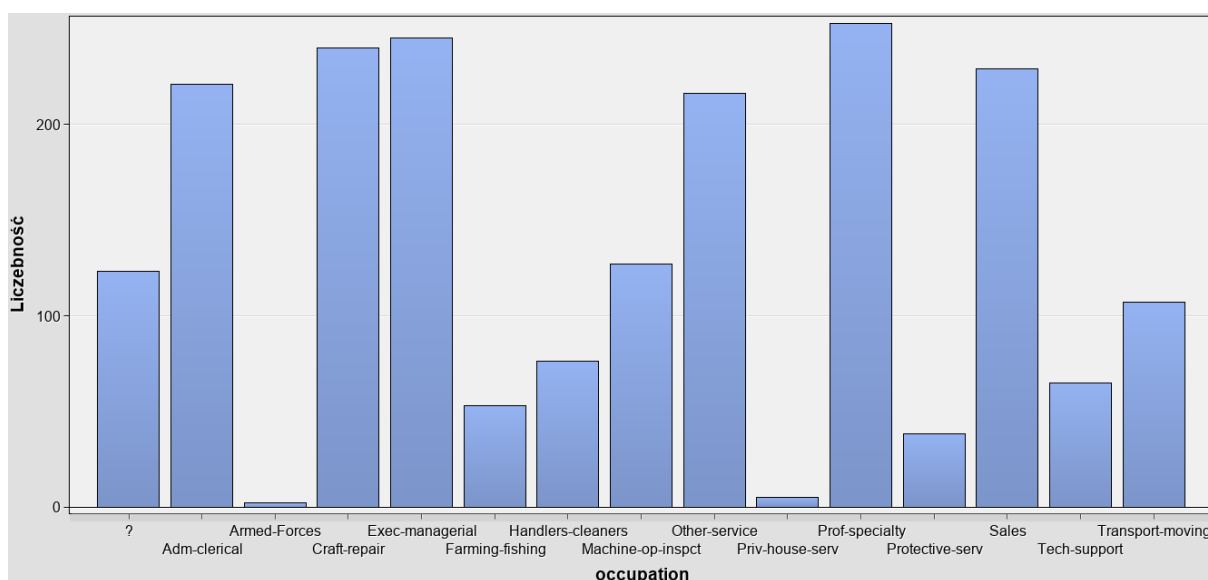
Rysunek 8 Boxplot dla edukacji i godzin spędzonych tygodniowo w pracy

Używam boxplota do zbadania zależności, ile godzin tygodniowo zostaje poświęcone pracy na podstawie poziomu wykształcenia. Z wykresu wynika, że pracownicy z wyższym wykształceniem (licencjat, magister, profesor i doktorant) mają tendencję do spędzania ponad 40 godzin tygodniowo na pracy w porównaniu z innymi rodzajami edukacji. Potwierdza to założenie, że osoby z wyższym wykształceniem mają tendencję do zarabiania 50 000 USD rocznie z powodu dłuższej godziny pracy.



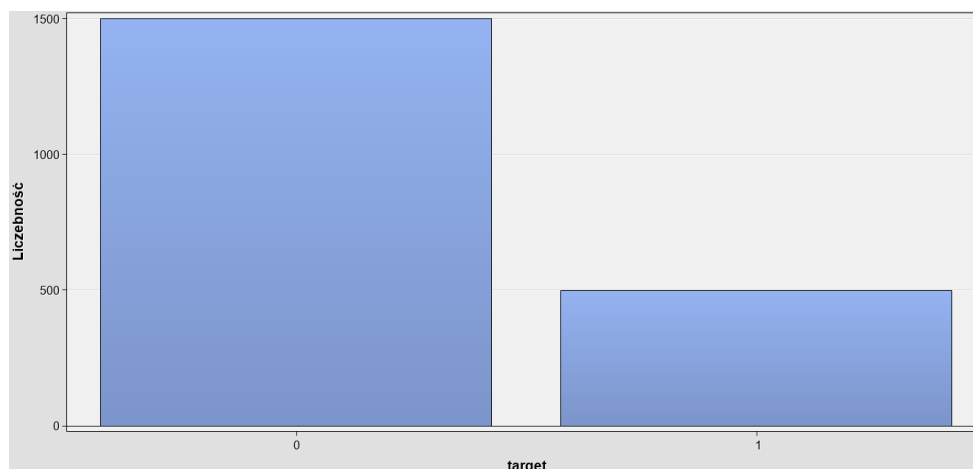
Rysunek 9 Wykres kołowy dla krajów

Na podstawie wykresu kołowego można stwierdzić, że ponad 90% badanych pochodziło z USA.



Rysunek 10 Wykres liczebności dla wykonywanych zawodów

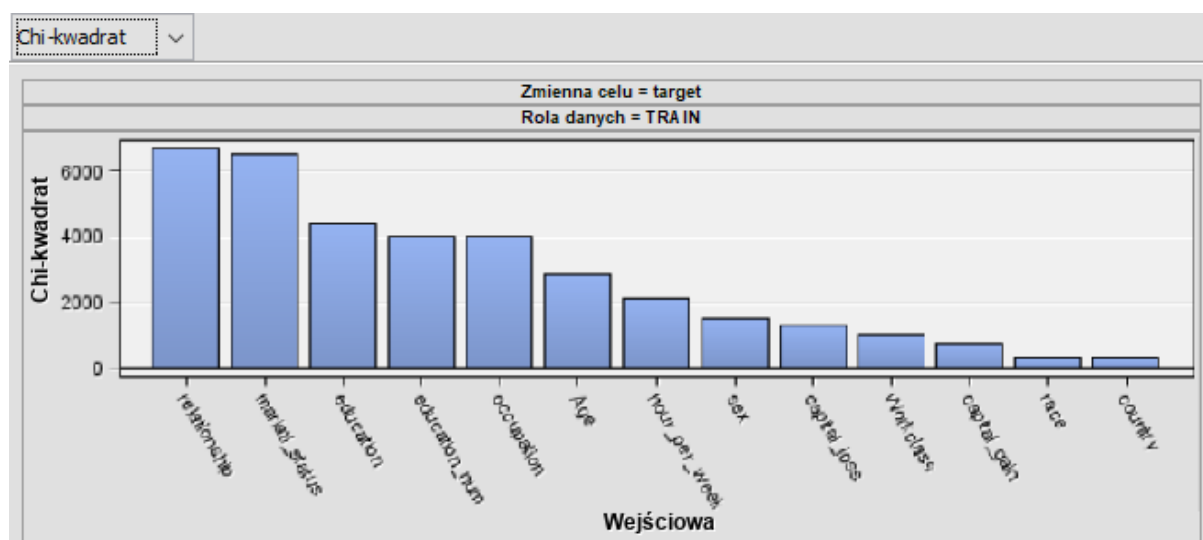
Bazując na wykresie słupkowym zmiennej ‘zawód’, że największej ludzi jest zawodowymi specjalistami, a najniższą grupą ludzi są siły zbrojne.



Rysunek 11 Podział zmiennej celu

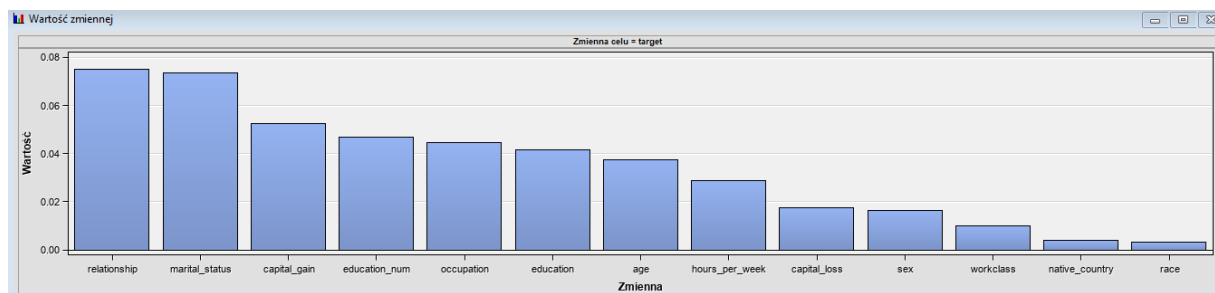
Na podstawie wykresu słupkowego dla naszej zmiennej docelowej można stwierdzić, że ponad 75% populacji osiągało dochód poniżej 50 000 rocznie, a około 25% osiągało roczny dochód powyżej 50 000 USD.

Korelacja zmiennych



Rysunek 12 Chi-kwadrat dla zmiennych

Węzeł eksploracja statystyk generuje statystyki opisowe. Wykres Chi-kwadrat i Wartość zmiennej jest tworzony w celu określenia, które zmienne najbardziej wpływają na zmienną celu. W tym przypadku oznacza to, że zmienne ‘relationship’ i ‘marital status’ są najbardziej istotne w celu ustalenia czy dochód przekraczał 50 000 USD. Pochodzenie etniczne i kraj mają najsłabszy wpływ na zmienną celu.



Rysunek 13 Wykres istotności zmiennych

Statystyki opisowe zmiennych klasyfikujących
(maksymalnie 500 obserwacji)

Rola danych=TRAIN

Rola danych	Nazwa zmiennej	Rola	Liczba poziomów	Moda	Procent mody	Moda2	Procent mody2
TRAIN	education	INPUT	16	HS-grad	32.25	Some-college	22.39
TRAIN	marital_status	INPUT	7	Married-civ-spouse	45.99	Never-married	32.81
TRAIN	native_country	INPUT	42	United-States	89.59	Mexico	1.97
TRAIN	occupation	INPUT	15	Prof-specialty	12.71	Craft-repair	12.59
TRAIN	race	INPUT	5	White	85.43	Black	9.59
TRAIN	relationship	INPUT	6	Husband	40.52	Not-in-family	25.51
TRAIN	sex	INPUT	2	Male	66.92	Female	33.08
TRAIN	workclass	INPUT	9	Private	69.70	Self-emp-not-inc	7.80
TRAIN	target	TARGET	2	0	75.92	1	24.08

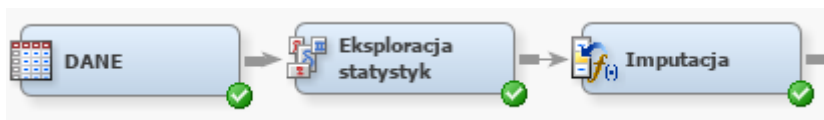
Rysunek 14 Podsumowanie statystyczne dla zmiennych

Bazując na statystykach podsumowujących zmienne, dalsze badania wykazały, że procentowy udział dla kraju (Stany Zjednoczone) i rasy (biała) przyczynia się do najwyższego odsetka procentu mody (89,59% i 85,43%), co jest uważane za główną przyczynę najniższego znaczenia zarówno dla wykresu Chi-Square, jak i Variable Worth.

TRAIN	education	INPUT	16	2
TRAIN	marital_status	INPUT	7	1
TRAIN	native_country	INPUT	42	1
TRAIN	occupation	INPUT	15	0
TRAIN	race	INPUT	5	1
TRAIN	relationship	INPUT	6	2
TRAIN	sex	INPUT	2	1
TRAIN	workclass	INPUT	9	1
TRAIN	target	TARGET	2	0

Rysunek 15 Brakujące wartości

Na podstawie wyników można też stwierdzić, że brakujące wartości są niewielkie. Zamiast usuwać brakujące wartości, wybrałam metodę imputacji, aby uzupełnić wszystkie brakujące wartości.



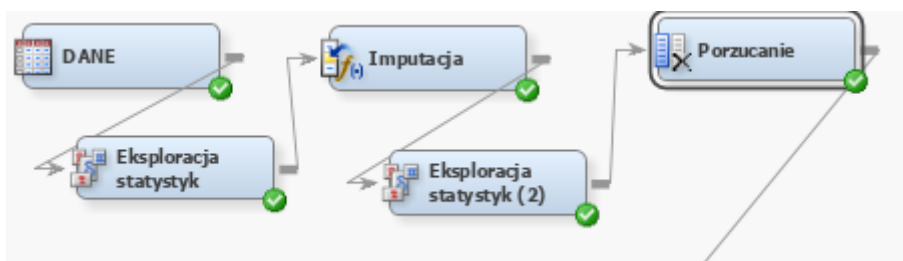
Rysunek 16 Węzeł 3 Imputacja

Przeprowadzam jeszcze raz eksplorację statystyk, w celu sprawdzenia, czy imputacja zmiennych wprowadziła jakieś zmiany.



Rysunek 17 Węzeł 4 Eksploracja statystyk 2

Statystyki praktycznie niczym się nie różnią od powyżej opisanych, ponieważ ilość braków była niewielka i nie miała wpływu na wyniki.

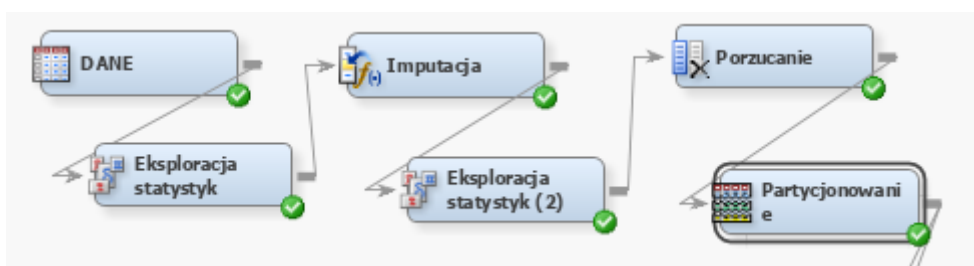


Rysunek 18 Węzeł 5 Porzucenie

Zmienna fniwgt składa się z ponad 21000 unikalnych wartości i nie przyczyniła się do niczego znaczącego w ramach wykresu Chi-kwadrat i Wartości zmiennych. Zmienna fnlwight usunęłam w celu zmniejszenia złożoności końcowego modelowania.

Partycjonowanie danych

Kolejnym krokiem w przygotowaniu modelu jest określenie jaki zbiór danych będzie używany do uczenia, walidacji i testów. Do tego celu używam partycjonowania.



Rysunek 19 Węzeł 6 Partycjonowanie

Dane zostały podzielone na zestaw do uczenia 40%, zestaw do walidacji 30% i zestaw do testów 20%.

Uczenie	
Zmienne	<input type="text" value="Dane"/>
Typ wyniku	Dane
Metoda partycjonowania	Domyślna
Ziarno losowe	12345
<input type="checkbox"/> Alokacje do zbiorów	
Uczenie	40.0
Walidacja	30.0
Test	20.0
Raport	

Rysunek 20 Ustawienia dla partycjonowania

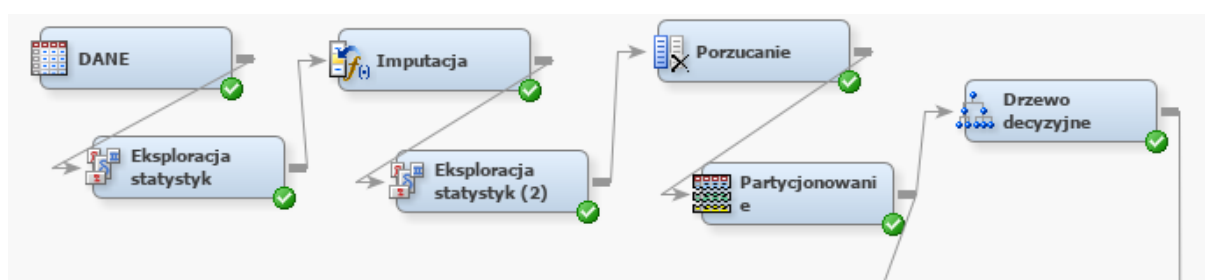
Podsumowanie partycjonowania

Typ	Zbiór	Liczba obserwacji
DATA	EMWS1.Drop_TRAIN	32561
TRAIN	EMWS1.Part_TRAIN	14471
VALIDATE	EMWS1.Part_VALIDATE	10854
TEST	EMWS1.Part_TEST	7236

Rysunek 21 Wynik partycjonowania

Procentowy udział zmiennej celu dla którego wartość liczbowa wynosi 1 w próbie jest równy 24,081%, następnie w podzbiorach: testowy – 24,0741, uczący – 24,0826, walidacyjny – 24,0833.

Drzewo decyzyjne



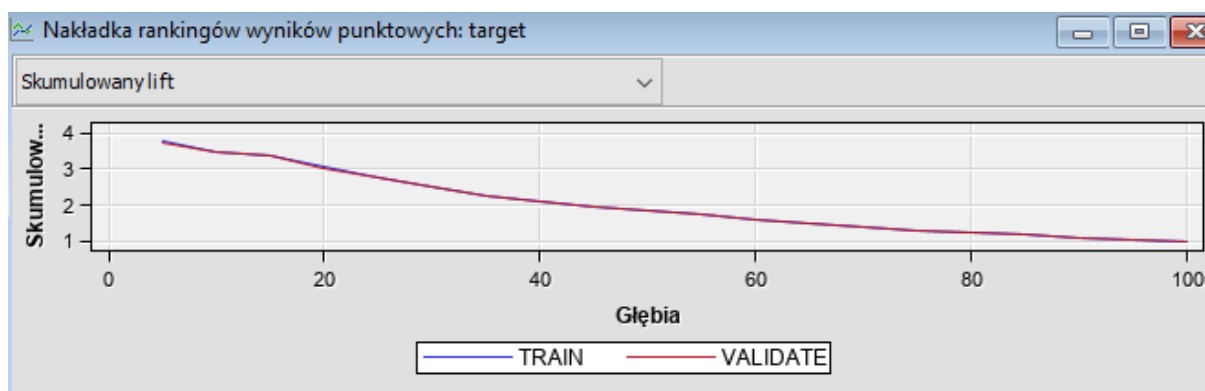
Rysunek 22 Dodanie 7 węzła Drzewo decyzyjne

Jednym z głównych celów tej pracy jest porównanie modelu pomiędzy regresją logistyczną a drzewem decyzyjnym. Po węźle partycjonowania danych linia łącząca została podzielona na 2 modele, które są regresją logistyczną i drzewo decyzyjne. Zastosowano domyślne ustawienie dla drzewa decyzyjnego, wszystkie parametry wymagane dla drzewa decyzyjnego zostały ustalone w następujący sposób:

Węzeł	
Wielkość liścia	5
Liczba reguł	5
Liczba reguł zastępczych	0
Wielkość podziału	.
Poszukiwanie podziału	
Użyj decyzji	Nie
Użyj prawdopodobieństw a priori	Nie
Wyczerpujące	5000
Próba węzła	20000
Poddrzewo	
Metoda	Ocena
Liczba liści	1
Miara oceny	Decyzja
Ułamek ocen	0.25
Walidacja krzyżowa	
Wykonuj walidację krzyżową	Nie
Liczba podzbiorów	10
Liczba powtórzeń	1
Ziarno	12345

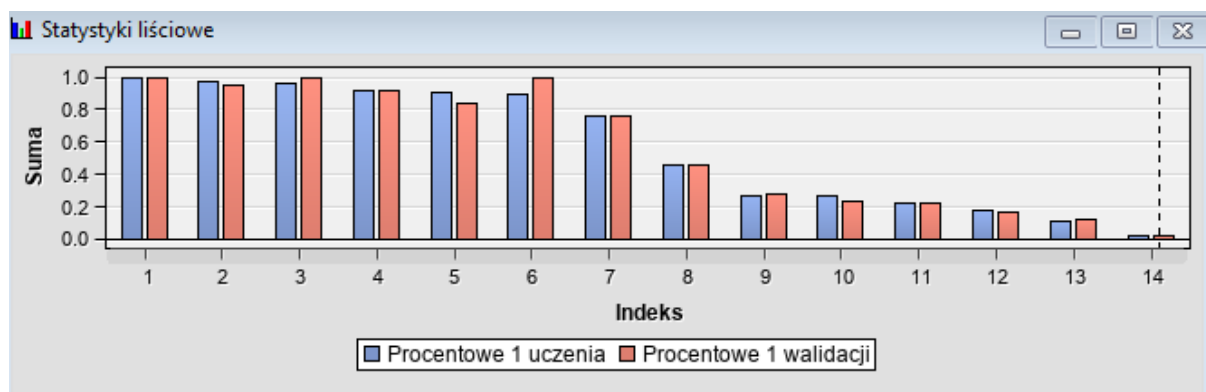
Rysunek 23 Ustawienia dla drzewa decyzyjnego

Rezultaty:



Rysunek 24 Nakładka rankingów wyników punktowych: target

Skumulowany wykres rozrzutu wskazuje na zgodność rozkładu dla zbioru treningowego i walidacyjnego, co świadczy o poprawnym dopasowaniu modelu. Jeżeli by się różniło, świadczyłoby to o przetrenowaniu lub niedotrenowaniu modelu.



Rysunek 25 Statystyki liściowe

Powyższy wykres umożliwia porównanie udziału procentowego wartości zmiennej celu =1 we wszystkich liściach dla danych treningowych i walidacyjnych. Należy przyciąć taki liść, dla którego różnice w wysokościach słupków są duże. Tutaj nie ma takiej potrzeby.

Statystyki dopasowania						
Zmienna celu	Etykieta zmiennej celu	Statystyki dopasowania	Etykieta statystyk	Uczenie	Walidacja	Test
target		_NOBS_	Suma liczebności	14471	10854	7236
target		_MISC_	Odsetek błędnych klasyfikacji	0.145118	0.145661	0.144417
target		_MAX_	Maksymalny błąd bezwzględny	0.992481	0.992481	0.992481
target		_SSE_	Suma błędów kwadratowych	3064.385	2307.346	1540.075
target		_ASE_	Przeciętny błąd kwadratowy	0.10588	0.10629	0.106418
target		_RASE_	Pierwiastek ze średniego błędu k...	0.325392	0.326022	0.326217
target		_DIV_	Dzielnik ASE	28942	21708	14472
target		_DFT_	Łączne stopnie swobody	14471		

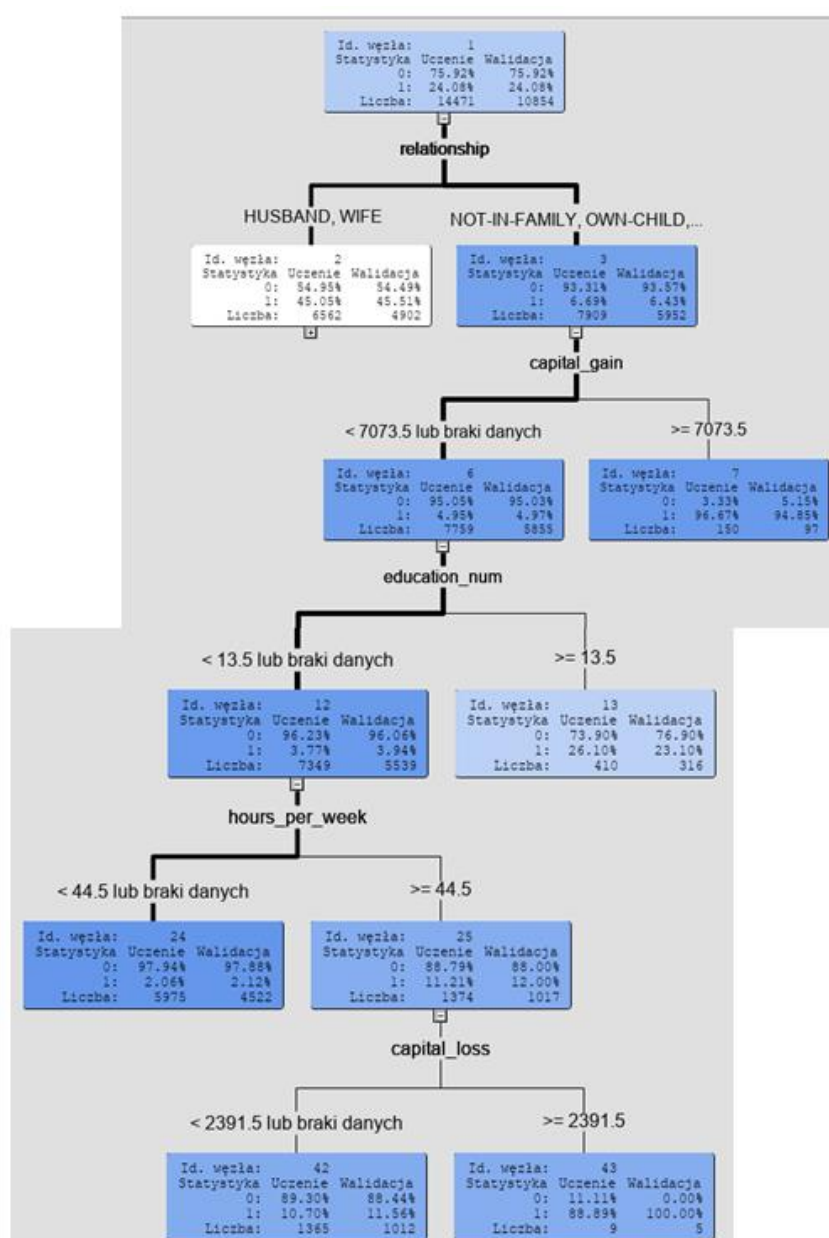
Rysunek 26 Statystyki dopasowania

Powyższe okno zawiera statystyki dopasowania modelu dla danych testowych (uczeniowych), walidacyjnych i testowych. Duże różnice wartości statystyk dopasowania mogą wskazywać na przetrenowanie lub niedotrenowanie modelu. Odsetek błędnych klasyfikacji wynosi około 14% i jest najmniejszy dla zbioru testowego.

Drzewo decyzyjne w tradycyjnej postaci prezentuje się następująco: (Ryc. 27 na następnej stronie).

Przeprowadzę teraz analizę uzyskanego drzewa pod względem najliczniejszej uzyskanej grupy. Ile osób spełnia warunek dochodu powyżej 50000USD w najliczniejszej grupie? Pierwszą zmienną, która spowodowała podział populacji jest zmienna relationship. W populacji jest około 24% dochodów, które przekraczają 50000 USD (target=1). Natomiast dla relacji typu mąż – żona jest ich już 45%, a dla pozostałych relacji tylko niecałe 7%. Następnie w tym węźle jest wykonywany podział zbioru obiektów według zmiennej capital_gain. Gdy <7073,5 lub

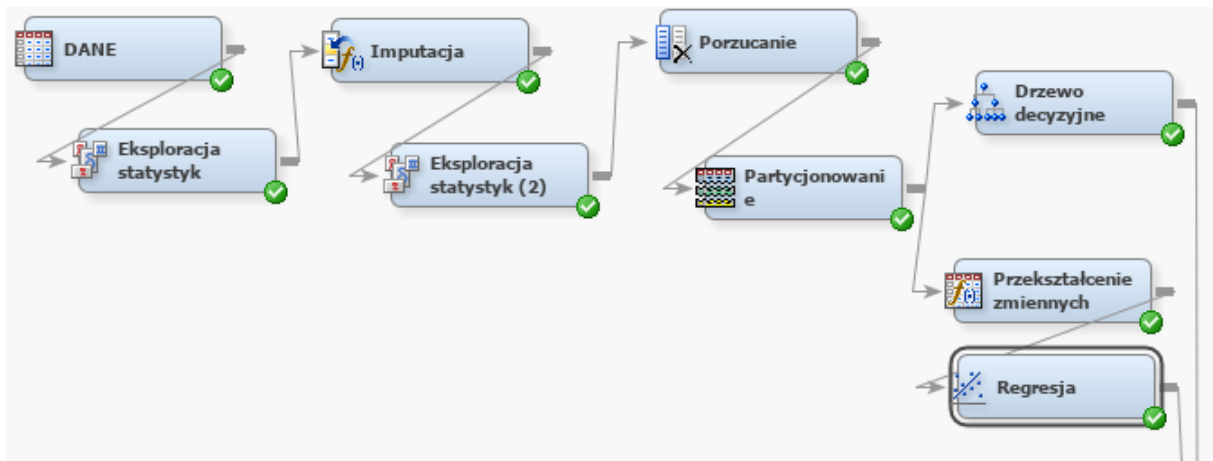
występują braki danych otrzymuję tylko 5% dla celu=1. Ale jest to najliczniejsza grupa. Następnie ten węzeł jest rozpatrywany pod kątem zmiennej ilości lat edukacji. W przypadku gdy liczba lat nie przekracza 13, zmienna celu obniża się do 3,77%. Dalej rozpatrywany jest węzeł ilości godzin pracy w tygodnie, dla <44,5 lub braków danych procent osób, których dochód nie przekracza 50 tysięcy USD wynosi już tylko 2,06% i jest to najliczniejsza grupa licząca w uczeniu 5975, a w walidacji 4522 osób.



Rysunek 27 Drzewo decyzyjne

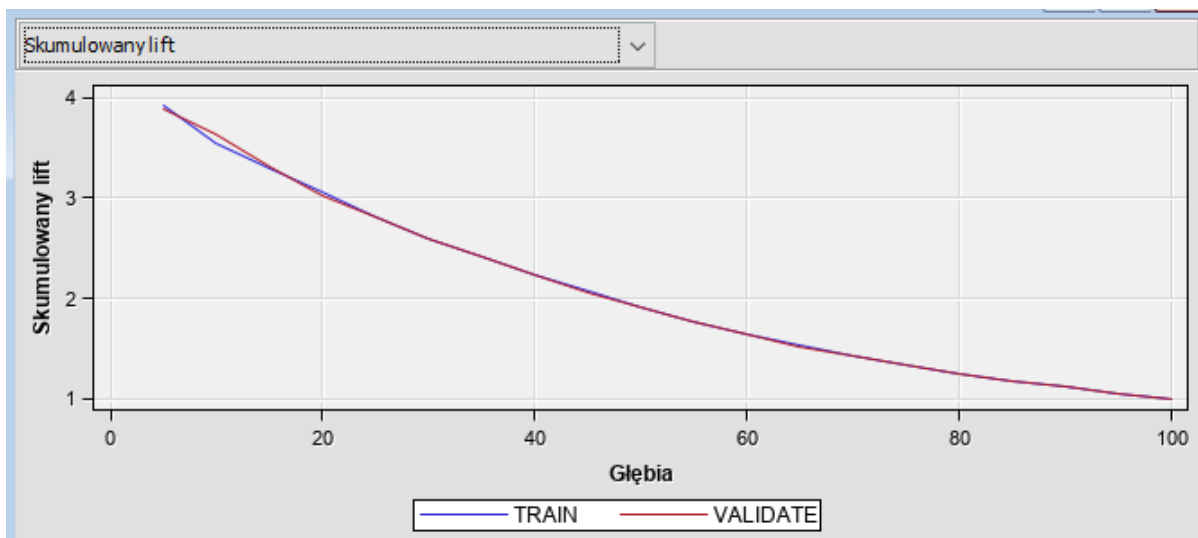
Regresja logistyczna

Dla regresji logistycznej został wprowadzony węzeł przekształcenia przed węzłem regresji logistycznej. Węzeł przekształcenia został użyty do stabilizacji wariancji, usunięcia nieliniowości, poprawy addytywności i dążeniu do normalności rozkładu.



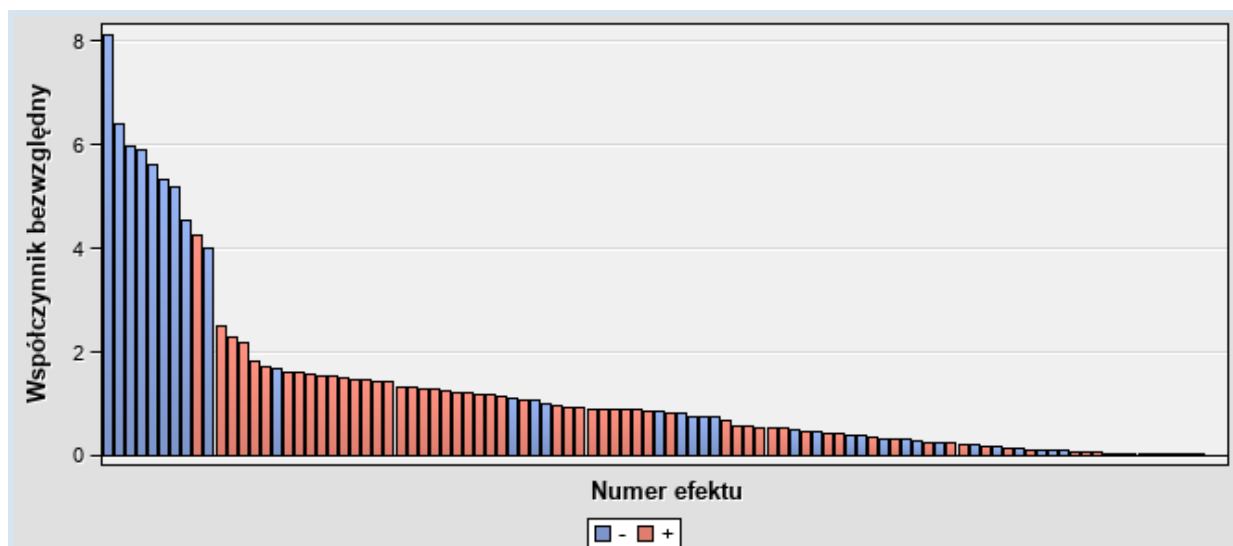
Rysunek 28 Węzeł 9 Regresja

Rezultat:



Rysunek 29 Nakładka rankingów wyników punktowych: target 2

Nakładka rankingów wyników punktowych: target, dla skumulowanego liftu wskazuje na dopasowanie dla zbioru treningowego i walidacyjnego. Oznacza to dobre dopasowanie modelu.



Rysunek 30 Wykres efektów

W oknie wykresów efektów, pokazane są współczynniki regresji w kolejności od największych do najmniejszych wartości bezwzględnych. Kolor czerwony oznacza liczbę dodatnią, kolor niebieski ujemną. Na przykład zmienna workclass ma współczynnik 4,27.

Zmienna celu	Etykieta zmiennej celu	Statystyki dopasowania	Etykieta statystyk	Uczenie	Walidacja	Test
target		_AIC_	Kryterium informacyjne Akaikego	9350.516		
target		_ASE_	Przeciętny błąd kwadratowy	0.101719	0.101492	0.10264
target		_AVERR_	Funkcja błędu przeciętnego	0.316306	0.31823	0.324261
target		_DFE_	Stopnie swobody błędu	14373		
target		_DFM_	Stopnie swobody modelu	98		
target		_DFT_	Stopnie swobody razem	14471		
target		_DIV_	Mianownik dla ASE	28942	21708	14472
target		_ERR_	Funkcja błędu	9154.516	6908.142	4692.709
target		_FPE_	Końcowy błąd prognozy	0.103108		
target		_MAX_	Największy błąd bezwzględny	0.999905	0.999996	0.999997
target		_MSE_	Średni błąd kwadratowy	0.102413	0.101492	0.10264
target		_NOBS_	Suma liczebności	14471	10854	7236
target		_NW_	Liczba wag ocen	98		
target		_RASE_	Przeciętna suma kwadratów	0.318934	0.318578	0.320375
target		_RFPE_	Końcowy błąd prognozy	0.321102		
target		_RMSE_	Pierwiastek z błędu średniokwadratowego	0.32002	0.318578	0.320375
target		_SBC_	Kryterium bayesowskie Schwarza	10093.35		
target		_SSE_	Suma błędów kwadratowych	2943.957	2203.192	1485.411
target		_SUMW_	Suma wag przypadków razy liczbę	28942	21708	14472
target		_MISC_	Odsetek błędnych klasyfikacji	0.147122	0.149807	0.149392

Rysunek 31 Statystyki dopasowania

Okno wynikowe statystyk dopasowania zawiera statystyki dopasowania dla danych testowych, walidacyjnych i treningowych. Odsetek błędnych klasyfikacji wynosi prawie 15% i jest najmniejszy dla zbioru testowego.

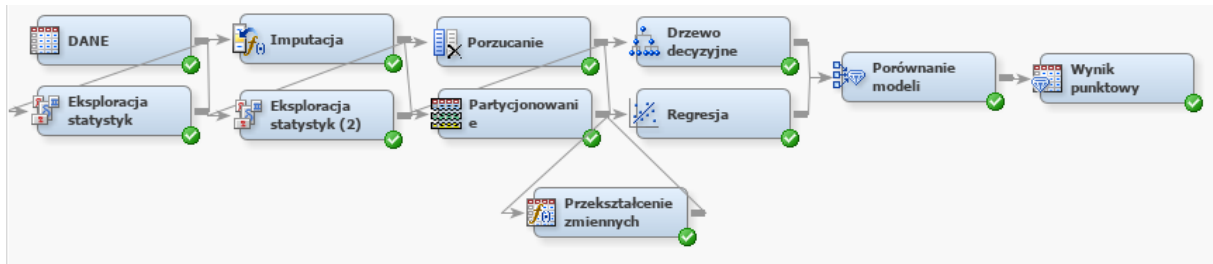
Analiza ocen maksymalnej wiarygodności							
Parametr	DF	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.	Ocena standaryzowana	Exp(oszacowanie)
Intercept	1	-8.1418	3.9941	4.16	0.0415		0.000
age	1	0.0265	0.00249	112.97	<.0001	0.1973	1.027
capital_gain	1	0.000322	0.000015	453.27	<.0001	1.2127	1.000
capital_loss	1	0.000656	0.000056	136.93	<.0001	0.1437	1.001
education 10th	1	-0.00628	1.5457	0.00	0.9968		0.994
education 11th	1	-0.1278	1.5468	0.01	0.9341		0.880
education 12th	1	0.4342	1.5637	0.08	0.7813		1.544
education 1st-4th	1	-6.3908	10.0736	0.40	0.5258		0.002
education 5th-6th	1	-0.3216	1.5819	0.04	0.8389		0.725
education 7th-8th	1	-0.7944	1.5535	0.26	0.6091		0.452
education 9th	1	-0.1490	1.5657	0.01	0.9242		0.862
education Assoc-acdm	1	1.2872	1.5376	0.70	0.4025		3.623
education Assoc-voc	1	1.2492	1.5363	0.66	0.4162		3.488
education Bachelors	1	1.7074	1.5327	1.27	0.2412		6.034

Rysunek 32 Raport - wynik

W oknie wyników można odczytać ocenę, błąd standardowy, statystyką Walda, istotność zmiennych i współczynniki. Dla zmiennych ciągłych są podane oceny standaryzowane. Przykładowo można odczytać, że zmienna age przyjmuje wartość 0,0025 dla błędu standardowego, co oznacza, że ten parametr zmienia się o 0,0025 lat w różnych badaniach tego samego zjawiska. Dodatkowo dla zmiennej age został zaprezentowany test Chi-kwadrat Walda i wartość p dla niego. Oznacza, to że ta zmienna jest istotna, ponieważ wartość p jest wystarczająco niska.

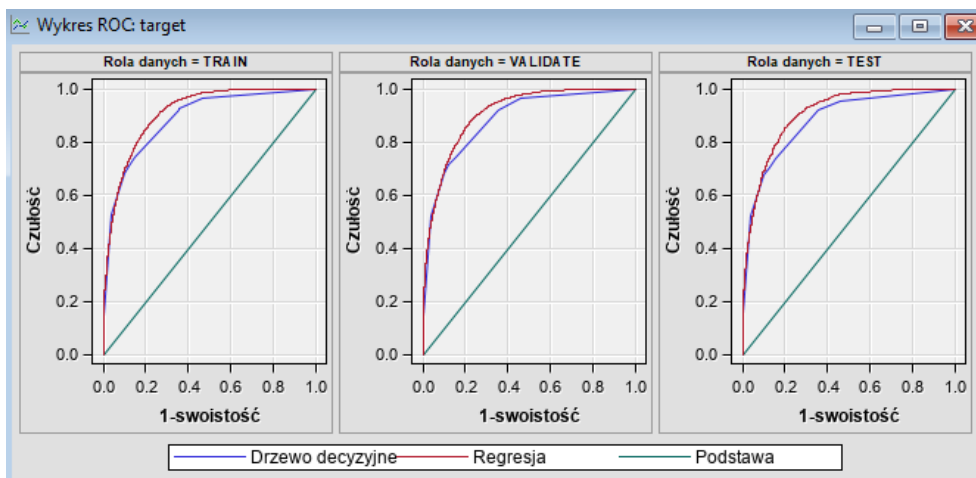
Porównanie modeli

Dodaję węzeł porównywania modeli i wynik punktowy.



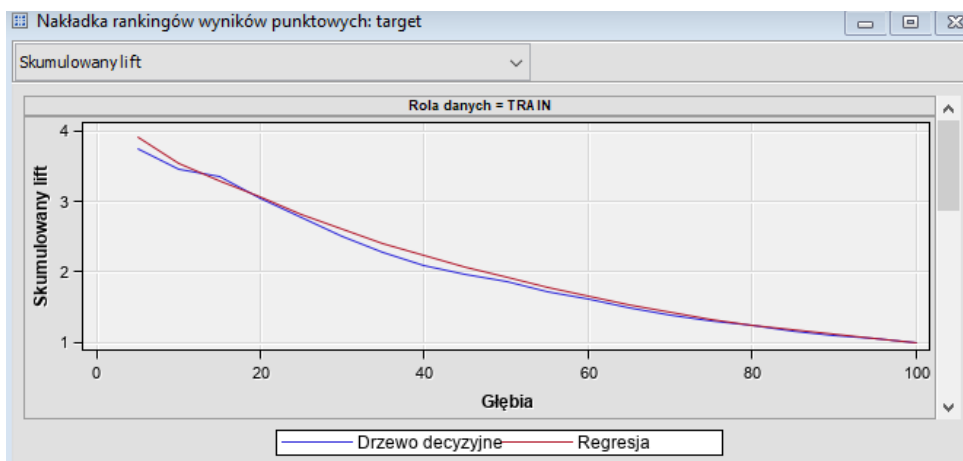
Rysunek 33 Węzeł 10 Porównanie modeli

Rezultat jest następujący:



Rysunek 34 Wykres ROC: target

Na podstawie krzywej wydajności ROC dla obu modeli, dokładność obu modeli wynosi około 0,90, co jest uważane za dobre (powyżej 0,90 jest uważane za doskonałe), a regresja logistyczna działa nieco lepiej – czerwona linia wysunięta bardziej w kierunku lewego górnego rogu. Na przekątnej zamieszczona jest linia przedstawiająca model podstawowy.



Rysunek 35 Nakładka rankingów wyników punktowych: target

Z porównania krzywych wzrostu wynika, że regresja jest odrobinę lepsza niż drzewo decyzyjne. Decyduje o tym dla początkowych i środkowych wartości linia czerwona – oznaczająca regresję logistyczną – położona tuż nad linią niebieską.

Wzrost poprzedzający	Wzrost modelu	Opis modelu	Zmienna celu	Etykieta zmiennej celu	Kryterium wyboru: Walidacja: Odsetek błędnych klasyfikacji	Uczenie: Suma liczebności	Uczenie: Odsetek błędnych klasyfikacji	Uczenie: Maksymalny błąd bezwzględny	Uczenie: Suma błędów kwadratowych	Uczenie: Przeciętny błąd kwadratowy	Uczenie: Pierwiastek ze średniego błędu kwadratowego	Uczenie: Dzielnik ASE	Uczenie: Łączne stopnie swobody	Walidacja: Suma liczebności	Walidacja: Odsetek błędnych klasyfikacji	Walidacja: Maksymalny błąd bezwzględny
Tree	Tree	Drzewo dec...target			0.145661	14471	0.145118	0.992481	3064.385	0.10588	0.325392	28942	14471	10854	0.145661	0.992481
Reg	Reg	Regresja target			0.149807	14471	0.147122	0.999905	2943.957	0.101719	0.318934	28942	14471	10854	0.149807	0.999898

Rysunek 36 Statystyki dopasowania

Walidacja odsetek błędnych klasyfikacji dla drzewa decyzyjnego wynosił 0,146, a dla regresji logistycznej 0,150. Możemy powiedzieć, że drzewo decyzyjne wykonuje o 2,4% lepsze dopasowanie niż regresja logistyczna.

Statystyki	Tree	Reg
Uczenie: Obcięcie prawdopodobieństwa klasowego dwuczynnikowego Kołmogorowa-Smirnowa	0.24	0.28
Uczenie: Statystyka Kołmogorowa-Smirnowa	0.59	0.65
Uczenie: Kryterium informacyjne Akaikiego	.	9350.52
Uczenie: Przeciętny błąd kwadratowy	0.11	0.10
Uczenie: Indeks ROC	0.88	0.91
Uczenie: Funkcja błędu przeciętnego	.	0.32
Uczenie: Skumulowana odpowiedź procentowa otrzymana	34.54	35.44
Uczenie: Odpowiedź procentowa otrzymana	15.80	15.84
Kryterium wyboru: Walidacja: Odsetek błędnych klasyfikacji	0.15	0.15
Uczenie: Stopnie swobody błędu	.	14373.00
Uczenie: Stopnie swobody modelu	.	98.00
Uczenie: Łączne stopnie swobody	14471.00	14471.00
Uczenie: Dzielnik ASE	28942.00	28942.00
Uczenie: Funkcja błędu	.	9154.52
Uczenie: Końcowy błąd prognozy	.	0.10
Uczenie: Zysk	245.21	254.16
Uczenie: Współczynnik Giniego	0.77	0.82
Uczenie: Klasowa dwuczynnikowa statystyka Kołmogorowa-Smirnowa	0.59	0.65
Uczenie: Odcięcie prawdopodobieństwa Kołmogorowa-Smirnowa	0.22	0.22
Uczenie: Skumulowany lift	3.45	3.54
Uczenie: Lift	3.16	3.17
Uczenie: Maksymalny błąd bezwzględny	0.99	1.00
Uczenie: Odsetek błędnych klasyfikacji	0.15	0.15
Uczenie: Średni błąd kwadratowy	.	0.10
Uczenie: Suma liczebności	14471.00	14471.00
Uczenie: Liczba wag ocen	.	98.00
Uczenie: Pierwiastek ze średniego błędu kwadratowego	0.33	0.32
Uczenie: Skumulowana odpowiedź procentowa	83.14	85.29
Uczenie: Odpowiedź procentowa	76.03	76.24
Uczenie: Końcowy błąd prognozy	.	0.32
Uczenie: Pierwiastek z błędu średniokwadratowego	.	0.32
Uczenie: Kryterium bayesowskie Schwarza	.	10093.35
Uczenie: Suma błędów kwadratowych	3064.39	2943.96
Uczenie: Suma wag przypadków razy liczeb.	.	28942.00

Rysunek 37 Tabela dopasowania statystyk dla zbioru treningowego

Rola danych=Valid

Statystyki	Tree	Reg
Walidacja: Statystyka Kołmogorowa-Smirnowa	0.59	0.65
Walidacja: Przeciętny błąd kwadratowy	0.11	0.10
Walidacja: Indeks ROC	0.88	0.91
Walidacja: Funkcja błędu przeciętnego	.	0.32
Walidacja: Obcięcie prawdopodobieństwa klasowego dwuczynnikowego Kołmogorowa-Smirnowa	0.24	0.22
Walidacja: Skumulowana odpowiedź procentowa otrzymana	34.51	36.42
Walidacja: Odpowiedź procentowa otrzymana	15.83	16.99
Walidacja: Dzielnik VASE	21708.00	21708.00
Walidacja: Funkcja błędu	.	6908.14
Walidacja: Zysk	244.86	263.99
Walidacja: Współczynnik Giniego	0.77	0.81
Walidacja: Klasowa dwuczynnikowa statystyka Kołmogorowa-Smirnowa	0.59	0.65
Walidacja: Odciecie prawdopodobieństwa Kołmogorowa-Smirnowa	0.26	0.20
Walidacja: Skumulowany lift	3.45	3.64
Walidacja: Lift	3.16	3.40
Walidacja: Maksymalny błąd bezwzględny	0.99	1.00
Walidacja: Odsetek błędnych klasyfikacji	0.15	0.15
Walidacja: Średni błąd kwadratowy	.	0.10
Walidacja: Suma liczebności	10854.00	10854.00
Walidacja: Pierwiastek ze średniego błędu kwadratowego	0.33	0.32
Walidacja: Skumulowana odpowiedź procentowa	83.05	87.66
Walidacja: Odpowiedź procentowa	76.21	81.77
Walidacja: Pierwiastek z błędu średniokwadratowego	.	0.32
Walidacja: Suma błędów kwadratowych	2307.35	2203.19
Walidacja: Suma wag przypadków razy liczebn.	.	21708.00

Rysunek 38 Tabela dopasowania statystyk dla zbioru walidacyjnego

Rola danych=Test

Statystyki	Tree	Reg
Testowanie: Statystyka Kołmogorowa-Smirnowa	0.58	0.65
Testowanie: Przeciętny błąd kwadratowy	0.11	0.10
Testowanie: Indeks ROC	0.88	0.91
Testowanie: Funkcja błędu przeciętnego	.	0.32
Testowanie: Obcięcie prawdopodobieństwa klasowego dwuczynnikowego Kołmogorowa-Smirnowa	0.24	0.28
Testowanie: Skumulowana odpowiedź procentowa otrzymana	34.88	35.71
Testowanie: Odpowiedź procentowa otrzymana	15.83	15.90
Testowanie: Dzielnik TASE	14472.00	14472.00
Testowanie: Funkcja błędu	.	4692.71
Testowanie: Zysk	248.65	256.86
Testowanie: Współczynnik Giniego	0.76	0.81
Testowanie: Klasowa dwuczynnikowa statystyka Kołmogorowa-Smirnowa	0.58	0.64
Testowanie: Odciecie prawdopodobieństwa Kołmogorowa-Smirnowa	0.22	0.24
Testowanie: Skumulowany lift	3.49	3.57
Testowanie: Lift	3.16	3.18
Testowanie: Największy błąd bezwzględny	0.99	1.00
Testowanie: Odsetek błędnych klasyfikacji	0.14	0.15
Testowanie: Dolna gr. prz. ufn. 95% dla TMISC	.	0.14
Testowanie: Górna gr. prz. ufn. 95% dla TMISC	.	0.16
Testowanie: Średni błąd kwadratowy	.	0.10
Testowanie: Suma liczebności	7236.00	7236.00
Testowanie: Pierwiastek z przeciętnego błędu kwadratowego	0.33	0.32
Testowanie: Skumulowana odpowiedź procentowa	83.93	85.91
Testowanie: Odpowiedź procentowa	76.16	76.52
Testowanie: Pierwiastek z błędu średniokwadratowego	.	0.32
Testowanie: Suma błędów kwadratowych	1540.07	1485.41
Testowanie: Suma iloczynów wagi i liczebności	14472.00	14472.00

Rysunek 39 Tabela dopasowania statystyk dla zbioru testowego

	Drzewo decyzyjne	Regresja logistyczna
Rola danych:Valid		
Statystyka Kołmogorowa-Smirnowa ↑	0,59	0,65
Indeks ROC ↑	0,88	0,91
Współczynnik Giniego ↑	0,77	0,81
Średni błąd kwadratowy ↓	.	0,1
Rola danych:Test		
Statystyka Kołmogorowa-Smirnowa ↑	0,58	0,65
Indeks ROC ↑	0,88	0,91
Współczynnik Giniego ↑	0,76	0,81
Średni błąd kwadratowy ↓	.	0,1

Rysunek 40 Tabela porównania statystyk modelu dla drzewa decyzyjnego i regresji logistycznej

Statystyka Kołmogorowa-Smirnowa – liczy zgodność empiryczną rozkładu z teoretycznym normalnym – im wyższy tym lepszy;

Indeks ROC – najwyższa wartość, ocenia jakość dopasowania modelu;

Współczynnik Giniego – miara koncentracji (nierównomierności) rozkładu zmiennej losowej;

Średni błąd kwadratowy – ocena błędu prognozy ex-post, im mniejszy tym lepszy;

W oparciu o powyższe wskaźniki można stwierdzić, że regresja liniowa i drzewo decyzyjne mają dobrą wydajność. Jednak regresja liniowa jest nieco lepsza niż drzewo decyzyjne.

Metody grupowania

Węzeł przetwarzania grupowego w Enterprise Miner jest przeznaczony do wykonywania osobnej analizy na każdym poziomie klasy jakiejś kategorycznej zmiennej grupowania, analizowania więcej niż jednej zmiennej docelowej, kontrolowania ilości razy, kiedy kolejne węzły będą zapętlać się w diagramach przepływu procesu, które są podłączone do węzła, oraz wykonywania operacji na nich poprzez ponowne próbkowanie zestawu danych wejściowych z wymienianymi zmiennymi. Grupowanie zmiennych stosuje się w celu ułatwienia przeprowadzenia analizy dla bardzo dużej ilości danych, ponieważ część z nich powoduje nadmiarowość informacji, opisując te same lub zbliżone właściwości obiektów. Chciałam w tej części pracy pokazać dwie metody grupowania: poprzez analizę skupień i sieci neuronowych Kohonena oraz jakie informacje można dzięki temu uzyskać.

Dane

Zbiór danych, który został użyty znajduje się na stronie: <https://www.openml.org/d/40701>. Zestaw danych zawiera informacje o tym jak klienci korzystali ze swojego konta telefonicznego, jakie wybrali usługi oraz czy klient zrezygnował czy przedłużył umowę. Zbiór składa się z 21 zmiennych i dane o 5000 klientach.

Zmienne znajdujące się w bazie:

- class (target) – czy klient przedłużył umowę;
- state – kod stanu;
- account_length – jak długo, klient posiada konto;
- area_code – numer kierunkowy;
- phone_number – numer telefonu – traktowane jako ID;
- international_plan – czy klient posiada plan międzynarodowy;
- voice_mail_plan – czy klient posiada pocztę głosową;
- numer_vmail_messages – liczba nagrań na pocztę głosową;
- total_day_minutes – całkowita liczba minut rozmowy w dzień;
- total_day_calls – całkowita liczba połączeń w dzień;
- total_day_charge – całkowita opłata za rozmowy w dzień;
- total_eve_minutes – całkowita liczba minut rozmowy wieczorem;
- total_eve_calls – całkowita liczba połączeń wieczorem;
- total_eve_charge – całkowita opłata za rozmowy wieczorem;
- total_night_minutes – całkowita liczba minut rozmowy w nocy;
- total_night_calls – całkowita liczba połączeń w nocy;

- total_night_charge – całkowita opłata za rozmowy w nocy;
- total_intl_minutes – całkowita liczba minut rozmów zagranicznych;
- total_intl_calls – całkowita liczba połączeń zagranicznych;
- total_intl_charge – całkowita opłata za połączenia zagraniczne;
- numer_customer_service_calls – liczba połączeń z biurem obsługi klienta;

Po załadowaniu bazy do programu SAS Base, tabela wygląda tak:

VIEWTABLE: Sasuser.Baza2												
	state	account_length	area_code	phone_number	international_plan	voice_mail_plan	number_vmail_messages	total_day_minutes	total_day_calls	total_day_charge	total_eve_minutes	total_eve
1	16	128	415	2845	0	1	25	265.1	110	45.07	197.4	99
2	35	107	415	2301	0	1	26	161.6	123	27.47	195.5	103
3	31	137	415	1616	0	0	0	243.4	114	41.38	121.2	110
4	35	84	408	2510	1	0	0	299.4	71	50.9	61.9	88
5	36	75	415	155	1	0	0	166.7	113	28.34	148.3	122
6	1	118	510	3355	1	0	0	223.4	98	37.98	220.6	101
7	19	121	510	1516	0	1	24	218.2	88	37.09	348.5	108
8	24	147	415	116	1	0	0	157	79	26.69	103.1	94
9	18	117	408	425	0	0	0	184.5	97	31.37	351.6	80
10	49	141	415	163	1	1	37	258.6	84	43.96	222	111
11	15	65	415	100	0	0	0	129.1	137	21.95	228.5	83
12	39	74	415	916	0	0	0	187.7	127	31.91	163.4	148
13	12	168	408	1854	0	0	0	128.8	96	21.9	104.9	71

Rysunek 41 Baza dla metod grupowania

Analiza skupień

Tworzę zbiór danych źródłowych na potrzeby tego projektu. Nie jest potrzebne tworzenie zbioru walidacyjnego ani testowego, ponieważ nie ma potrzeb stworzenia modelu predykcyjnego. Zadaniem jest podzielenie tego zbioru na jak najbardziej jednorodne grupy.



Rysunek 42 Węzeł 2 Klasteryzacja

Niektóre zmienne są ze sobą silnie skorelowane, a takich należy unikać do wprowadzania modelu – mogło spowodować to niestabilność modelu. Można wnioskować, że zmienne minuty, rozmowy i opłata są skorelowane. Zdecydowałam, że zostawię po jednej zmiennej związanej z minutami.

Nazwa	Rola	Poziom	Raport	Porządek	Porzucenie
account_length	Wejście	Przedziałowa	Nie		Nie
area_code	Odrzucona	Przedziałowa	Nie		Nie
class	Wejście	Binarna	Nie		Nie
international_pla	Wejście	Binarna	Nie		Nie
number_custome	Wejście	Przedziałowa	Nie		Nie
number_vmail_m	Wejście	Przedziałowa	Nie		Nie
phone_number	Id.	Nominalna	Nie		Nie
state	Wejście	Przedziałowa	Nie		Nie
total_day_calls	Odrzucona	Przedziałowa	Nie		Nie
total_day_charge	Odrzucona	Przedziałowa	Nie		Nie
total_day_minute	Wejście	Przedziałowa	Nie		Nie
total_eve_calls	Odrzucona	Przedziałowa	Nie		Nie
total_eve_charge	Odrzucona	Przedziałowa	Nie		Nie
total_eve_minute	Wejście	Przedziałowa	Nie		Nie
total_intl_calls	Odrzucona	Przedziałowa	Nie		Nie
total_intl_charge	Odrzucona	Przedziałowa	Nie		Nie
total_intl_minute	Wejście	Przedziałowa	Nie		Nie
total_night_calls	Odrzucona	Przedziałowa	Nie		Nie
total_night_cha	Odrzucona	Przedziałowa	Nie		Nie
total_night_minu	Wejście	Przedziałowa	Nie		Nie
voice_mail_plan	Wejście	Binarna	Nie		Nie

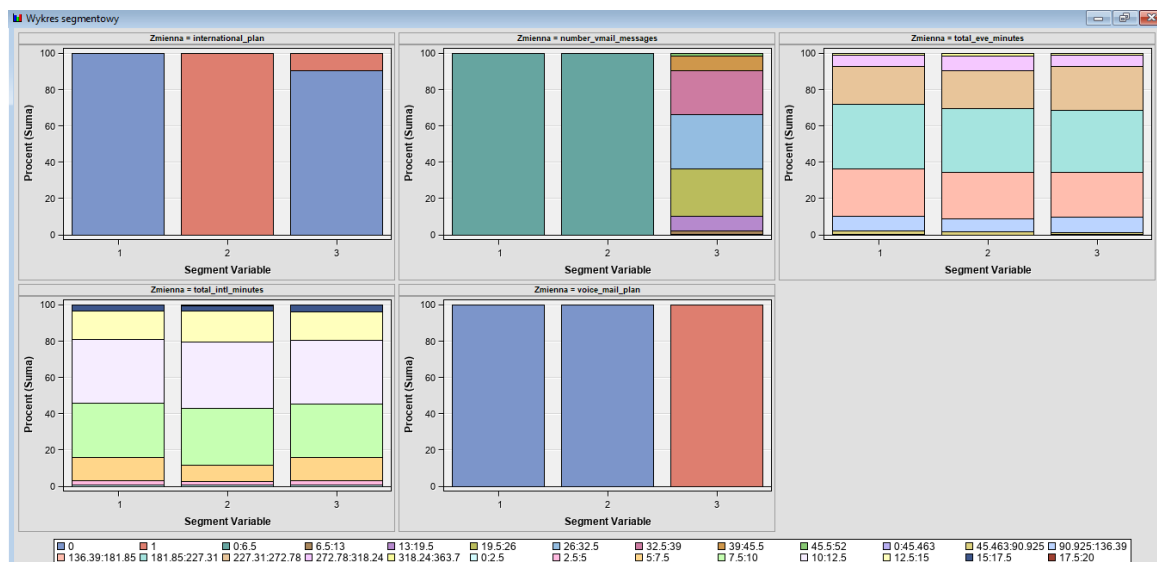
Rysunek 43 Ustalenie ról i poziomów dla zmiennych

Zmienna Class nie będzie brała udziału w analizie skupień, w węźle klasteryzacji przypisujemy jej wartość użycia ‘nie’.

Ogólne	
Id. węzła	Clus
Importowane dane	...
Eksportowane dane	...
Uwagi	...
Uczenie	
Zmienne	...
Standaryzacja wewnętrzna	Rozstęp
Liczba skupień	
Metoda podawania	Przez użytkownika
Maksymalna liczba skupień	3
Kryterium wyboru	
Metoda analizy skupień	Warda
Maksimum wstępne	50
Minimum	2
Maksimum końcowe	20
Obciążenie CCC	3
Kodowanie zmiennych klasyfikujących	
Kodowanie zmiennych porządkowych	Ranking
Kodowanie zmiennych nominalnych	GLM
Początkowe ziarna skupienia	

Rysunek 44 Ustawienia dla węzła klasteryzacji

Standaryzację wewnętrzną ustawiam na rozstęp, metodę podawania zmieniam na własną, w tym maksymalną liczbę skupień zmieniam na 3.



Rysunek 45 Wykres segmentowy

W oknie wykresów segmentowych wyświetla się rozkład każdej zmiennej, która miała rolę wejścia i w każdym skupieniu. Na osi y znajdują się zsumowane procenty. Zmienny odpowiadają poniżej zadane kolory. Można zauważyć tutaj dwie zmienne binarne, a mianowicie: international_plan i voice_mail_plan, czyli plan międzynarodowy i plan poczty głosowej. Ustawione wcześniej trzy segmenty odpowiadają teraz trzem słupkom na wykresie. Rozpatrując zmienną dotyczącą planu międzynarodowego, można stwierdzić, że przyjmuje wartość 1 dla wszystkich elementów skupienia nr 2, a wartość 0 dla wszystkich elementów skupienia nr 1. W przypadku skupienia nr 3 dla większości elementów wartościami są 0, a dla bardzo małej ilości 1.

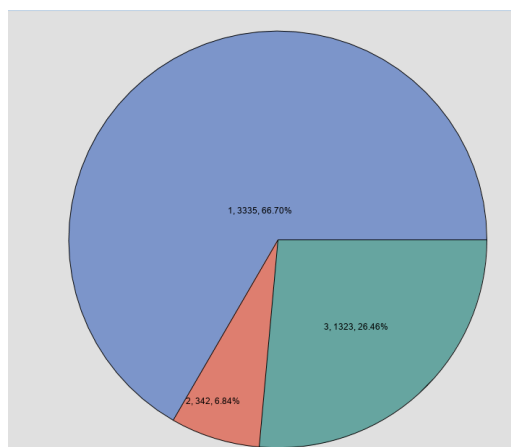
Kryterium skupienia	Największa względna zmiana ziaren skupień	Poprawa kryterium skupiania	Id. segmentu	Liczebność skupienia	Odchylenie standardowe średniej kwadratowej	Największa odległość od ziarna skupienia	Najbliższe skupienie	Odległość od najbliższego skupienia	account_length	number_customer_service_calls
0.149148	0	0	1	3335	0.133477	0.93792	2	1.414502	100.3193	1.586507
0.149148	0	0	2	342	0.134156	0.783249	1	1.414502	101.3684	1.535088
0.149148	0	0	3	1323	0.185905	1.581815	1	1.529	99.81859	1.538927

Rysunek 46 Tabela statystyki średnich 1

number_vm_messages	state	total_day_minutes	total_eve_minutes	total_intl_minutes	total_night_minutes	international_plan=0	international_plan=1	voice_mail_plan=0	voice_mail_plan=1
0	26.29205	179.683	199.5276	10.23253	200.7092	1	0	1	0
0	26.00585	185.4819	203.8465	10.47368	195.495	0	1	1	0
29.30915	25.25624	180.4739	202.6023	10.28073	200.8569	0.900983	0.099017	0	1

Rysunek 47 Tabela statystyki średnich 2

Z powyższej tabeli – Średnie statystyki – można odczytać, że international_plan=0 wynosi 90%, a zmienną 1 przyjmuje tylko 10% w segmencie 3. Kolumna ‘liczebność skupienia’ pokazuje ile elementów znajduje się w każdym skupieniu. Największą liczbę skupień posiada segment 1 – 3335, a najmniejszą segment drugi 342.



Rysunek 48 Wykres kołowy - rozmiar segmentu

Liczebność i wartości procentowe poszczególnych skupień ukazałam na powyższym wykresie kołowym.

SEGMENT	_1	_2	_3
1	0	9.094563	29.5377
2	9.094563	0	30.34844
3	29.5377	30.34844	0

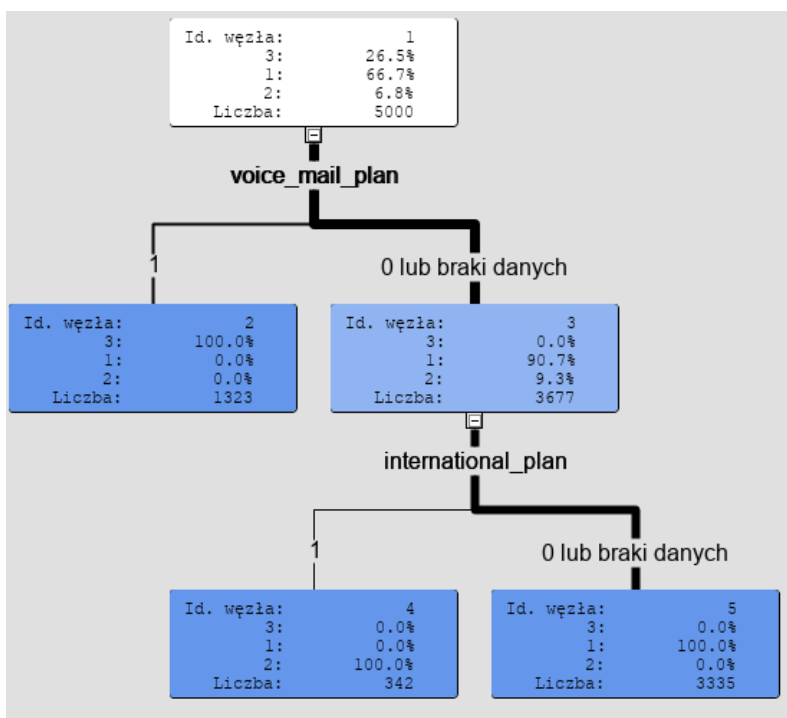
Rysunek 49 Tabela - odległość skupień

Powyższa tablica ukazuje odległości pomiędzy skupieniami. Tablica jest symetryczna. Można stwierdzić, że najbliższe skupienia to 1 i 2, a najdalsze 2 i 3.

Nazwa zmiennej	Etykieta	Liczba reguł podziału	Liczba reguł zastępczych	Istotność
voice_mail...		1	0	1
number_vm...		0	1	0.9999
total_intl_m...		0	1	0.857671
international...		1	0	0.590079
total_eve...		0	1	0.562052
number_cu...		0	0	0
total_day...		0	0	0
account_le...		0	0	0
total_night...		0	0	0

Rysunek 50 Tabela - istotność zmiennych

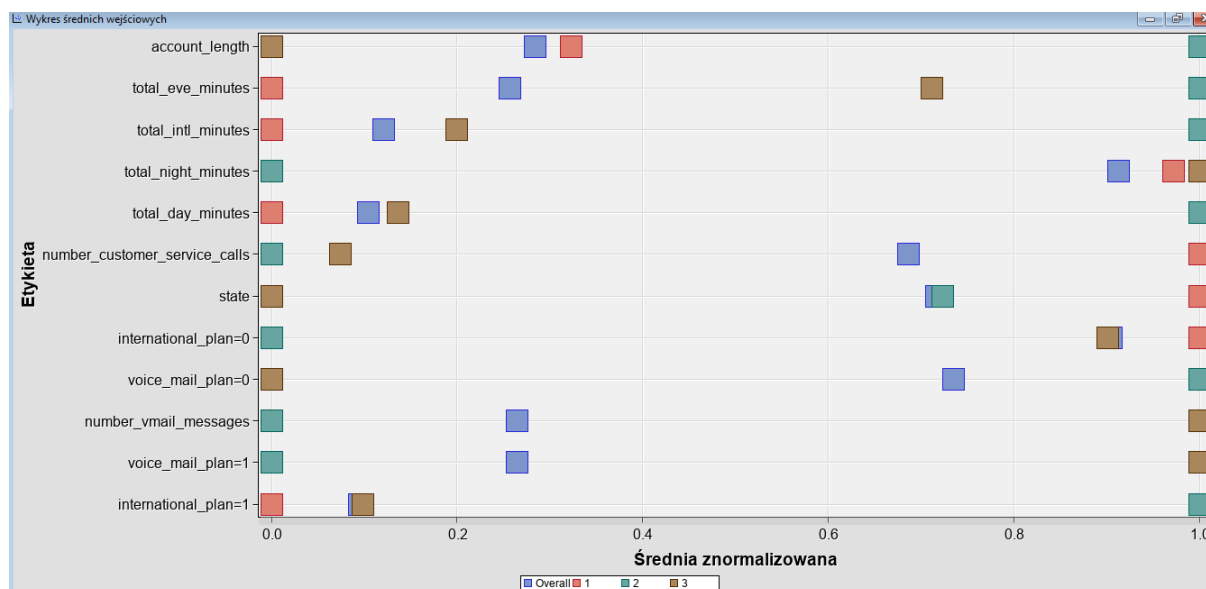
Powyższa tabela przedstawia istotność zmiennych. Wskaźnik informujący o jej znaczeniu zawiera się pomiędzy 0 i 1, określa jaką wagę posiadała zmienna podczas procesu skupienia. Dla czterech ostatnich zmiennych wartość istotności wynosi 0, co oznacza, że zmienna nie była wykorzystana jako zmienna decydująca o podziale skupienia. Najbardziej różnicującą zmienną dane była zmienna o tym, czy klient posiadał usługę poczty głosowej, a zaraz za nią liczba wiadomości na poczcie głosowej.



Rysunek 51 Drzewo przynależności obiektów do skupień

Na powyższym drzewie zmienną objaśnianą jest numer skupienia. Występuje podział na trzy skupienia. W tym przypadku o przyporządkowaniu obiektów do skupień decydują zmienne plan poczty głosowej i plan międzynarodowy. Początkowo zmienna voice_mail_plan rozdziela klientów na dwa węzły: dla wartości zmiennej 1 trafia 100% skupienia 3, a dla wartości 0 trafiają wartości ze skupienia 1 i 2, łącznie 3677 elementów. Na następnym poziomie

drzewa węzeł zawierający skupienia 1 i 2 zostaje rozdzielony na kolejne dwa węzły. Dla wartości 1 zmiennej international_plan zostaje utworzony węzeł zawierający elementy skupienia 2, to jest 342 elementy., natomiast dla wartości 0 lub brak danych węzeł z elementami skupienia 1 (2225 elementów).



Rysunek 52 Wykres średnich wejściowych

Powyższy wykres umożliwia porównanie znormalizowanych wartości średnich zmiennych ogółem i w skupieniach. Na przykład zmienna total_eve_minutes największe wartości w skupieniu 2.

segmentu	Liczba składowa	Odstępnosc standardowa	Największa odległość od ziarna skupienia	Najbliższe skupienie	Odstępnosc od najbliższego skupienia	account_length	number_customer_service_calls	number_vmail_messages	state	total_day_minutes	total_eve_minutes	total_intl_minutes	total_night_minutes	international_plan=0	international_plan=1	voice_mail_plan=0	voice_mail_plan=1
1	3335	0.133477	0.93792	2	1.414502	100.3193	1.586507	0	26.29205	179.683	199.5276	10.23253	200.7092	1	0	1	0
2	342	0.134156	0.783249	1	1.414502	101.3684	1.535088	0	26.00585	185.4819	203.8465	10.47368	195.495	0	1	1	0
3	1323	0.185905	1.581815	1	1.529	99.81859	1.538927	29.30915	25.25624	180.4739	202.6023	10.28073	200.8569	0.900983	0.099017	0	1

Rysunek 53 Statystyki średnich

Zmienna total_eve_minutes ma w skupieniach wartości średnie odpowiadające 1 skupieniu 199,53; 2 - 203,85; 3 – 202,6. Największa wartość średnia jest jednak w skupieniu 2, dlatego jej kwadracik jest najdalej na wykresie.

Port	Tabela	Rola	Dane istnieją
TRAIN	EMWS2.Clus_TRAIN	Uczące	Tak
VALIDATE	EMWS2.Clus_VALIDATE	Walidacyjne	Nie
TEST	EMWS2.Clus_TEST	Testowe	Nie
CLUSSTAT	EMWS2.Clus_OUTSTAT	Statystyki skupień	Tak
CLUSMEAN	EMWS2.Clus_OUTMEAN	Średnie skupień	Tak
VARMAP	EMWS2.Clus_OUTVAR	Mapowanie zmiennych	Tak

Rysunek 54 Pliki tworzone i eksportowane przez węzeł Klasteryzacja

Na powyższej tabeli widać jakie pliki są tworzone i eksportowane przez Klasteryzację.

Eksploruj - EMWS2.Clus_TRAIN

Plik Widok Działania Odkno

Właściwości próby

Właściwość	Wartość
Wiersze	Nieznana
Kolumny	24
Biblioteka	EMWS2
Element	CLUS_TRAIN
Typ	VIEW
Metoda próbkowania	Góra
Wielkość pobrania	Domyslna
Pobrane wiersze	2000
Ziarno losowe	12345

Statystyki próby

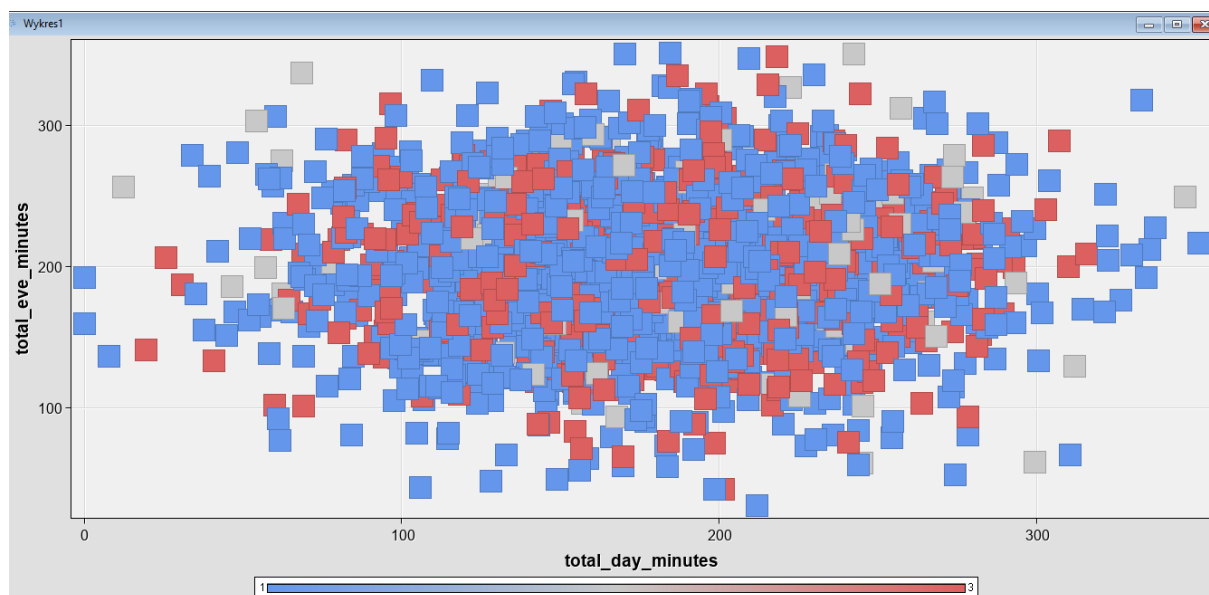
Nr obs.	Nazwa z...	Etykieta	Typ	Procent b...	Minimum	Maksimum	Średnia	Liczba
1	1_SEGMENT...	Segment D...	CLASS	0	0	1	0.48721	3
2	Distance	Odległość	VAR	0	0.079305	1.581815	0.48721	1.638
3	3_SEGMENT...	Id. segmentu	VAR	0	1	3	99.75	438.075
4	4account_le...		VAR	0	1	243	0.136	1.099
5	5area_code		VAR	0	408	510	1.547	8.4235
6	6class		VAR	0	0	1	25.9835	100.4875
7	7international...		VAR	0	0	1	30.62078	180.1189
8	8number_cu...		VAR	0	0	9	100.139	17.01906
9	9number_vm...		VAR	0	0	51		
10	10phone_nu...		VAR	0	1	4999		
11	11state		VAR	0	0	50		
12	12total_day_c...		VAR	0	0	165		
13	13total_day_c...		VAR	0	0	59.64		
14	14total_day_c...		VAR	0	0	350.8		
15	15total_eve_c...		VAR	0	12	168		
16	16total_eve_c...		VAR	0	2.65	29.89		

EMWS2.Clus_TRAIN

Nr obs.	state	account_l...	area_code	phone_nu...	internatio...	voice_ma...	number_v...	total_day...	total_day...	total_day...	total_eve...	total_eve...	total_eve...	total_nigh...	total_nigh...	total_nigh...
1	16	128	415	2845	0	1	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01
2	35	107	415	2301	0	1	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45
3	31	137	415	1616	0	0	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32
4	35	84	408	2510	1	0	0	299.4	71	50.9	51.9	88	5.26	196.9	89	8.86
5	36	75	415	155	1	0	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41
6	1	118	510	3355	1	0	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18
7	19	121	510	1516	0	1	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57
8	24	147	415	116	1	0	0	157	79	26.69	103.1	94	8.76	211.8	96	9.53
9	18	117	408	425	0	0	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71
10	49	141	415	163	1	1	37	258.6	84	43.96	222	111	18.87	326.4	97	14.69
11	15	65	415	100	0	0	0	129.1	137	21.95	228.5	83	19.42	208.8	111	9.4
12	39	74	415	916	0	0	0	187.7	127	31.91	163.4	148	13.89	196	94	8.82
13	12	168	408	1854	0	0	0	128.8	96	21.9	104.9	71	8.92	141.1	128	6.35
14	26	95	510	3509	0	0	0	156.6	88	26.62	247.6	75	21.05	192.3	115	8.65
15	12	62	415	2065	0	0	0	120.7	70	20.52	307.2	76	26.11	203	99	9.14
16	34	161	415	1268	0	0	0	332.9	67	56.59	317.8	97	27.01	160.6	128	7.23

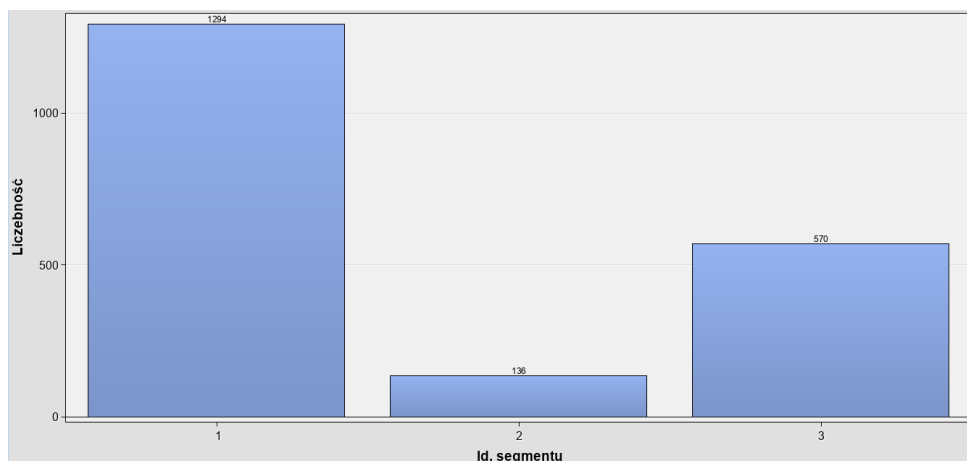
Rysunek 55 Okno eksploracji EMWS2.Clus_TRAIN

Na powyższym screenie można zobaczyć eksplorowanie EMWS2.Clus_TRAIN, które brały udział w Klasteryzacji.



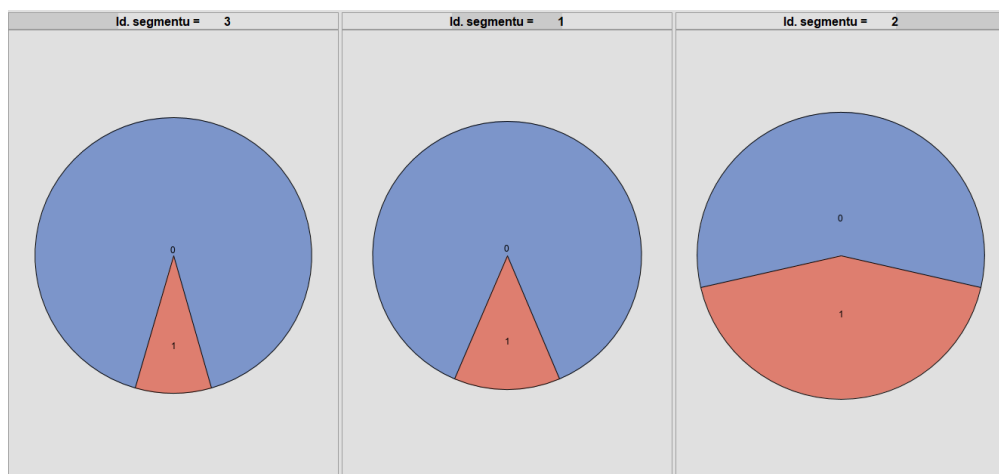
Rysunek 56 Wykres rozproszenia z zaznaczeniem przynależności obserwacji do skupień

Na tym wykresie kolory kwadracików odpowiadają odpowiedniemu skupieniu. Szary odpowiada 2 skupieniu, czerwony – 3, niebieski – 1. Można stwierdzić że najmniej skrajne wartości na wykresie dla całkowitych ilości minut w dzień i w nocy ma skupienie 3. Natomiast bardziej rozproszone jest skupienie 1 i 2.



Rysunek 57 Liczebność skupień na wykresie słupkowym

Na powyższym wykresie widać, jak rozkładały się elementy w zbiorze złożonym z 2000 obserwacji. Najwięcej trafiło do skupienia 1.



Rysunek 58 Wykres kołowy rozkładu zmiennej class

Powyższe wykresy kołowe przedstawiają liczebność zmiennej class w segmentach – to jest, czy klient przedłużył umowę (1). W segmencie 2 było najwięcej takich osób.



Rysunek 59 Węzeł 3 Segmentacja profilu

Postanowiłam dołączyć do diagramu węzeł ‘Segmentacja profilu’, który ułatwi porównanie rozkładu wartości zmiennej w danym skupieniu z rozkładem zmiennej w całym zbiorze danych.

Zmienna: _SEGMENT_ segment: 1 liczba: 1323
 Profile istotności drzewa decyzyjnego

Zmienna	Wartość	Ranking
voice_mail_plan	0.38917	1
number_vmail_messages	0.38877	2
class	0.00477	3
total_eve_minutes	0.00280	4
total_day_minutes	0.00222	5
total_night_minutes	0.00182	6
account_length	0.00155	7
total_intl_minutes	0.00122	8
number_customer_service_calls	0.00061	9
international_plan	0.00003	10

Rysunek 60 Wyniki Segmentacji - grupa 1

Grupa 1 – przeciętni użytkownicy. Grupa liczy 1323 klientów. Przystąpili do planu poczty głosowej, ale nikt nie przystąpił dla planu międzynarodowego.

Zmienna: _SEGMENT_ segment: 2 liczba: 3335
 Profile istotności drzewa decyzyjnego

Zmienna	Wartość	Ranking
voice_mail_plan	0.32015	1
number_vmail_messages	0.31982	2
international_plan	0.09297	3
total_eve_minutes	0.00262	4
total_night_minutes	0.00256	5
total_day_minutes	0.00224	6
account_length	0.00189	7
total_intl_minutes	0.00113	8
number_customer_service_calls	0.00054	9
class	0.00025	10

Rysunek 61 Wyniki Segmentacji - grupa 2

Grupa 2 – najbardziej liczna. Liczy 3335 osoby, którzy przystąpili do planu poczty głosowej, ale są też osoby, które zdecydowały się na plan międzynarodowy.

Zmienna: `_SEGMENT_` segment: 3 liczba: 342
 Profile istotności drzewa decyzyjnego

Zmienna	Wartość	Ranking
international_plan	0.089555	1
class	0.007215	2
voice_mail_plan	0.003367	3
number_vmail_messages	0.003363	4
total_day_minutes	0.001125	5
total_night_minutes	0.000973	6
total_eve_minutes	0.000825	7
total_intl_minutes	0.000527	8
account_length	0.000515	9
number_customer_service_calls	0.000091	10

Rysunek 62 Wyniki Segmentacji - grupa 3

Grupa 3 – elitarna. Liczy tylko 342 osoby, wśród których są uczestnicy planu międzynarodowego, ale nie ma praktycznie uczestników planu poczty głosowej.



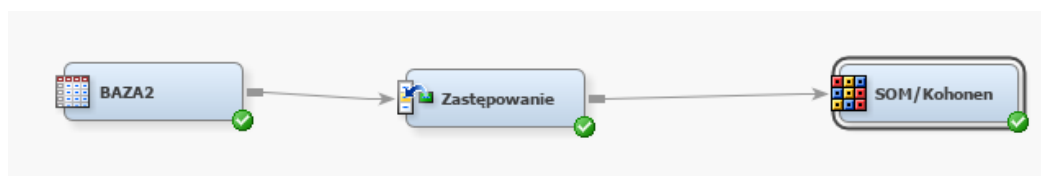
Rysunek 63 Wizualizacja segmentów

Powyżej znajduje się profil - wizualizacja wyodrębnionych grup, w celu lepszego zobrazowania podziału na segmenty.

Sieci neuronowe Kohonena

Sztuczną siecią neuronową o innej budowie, sposobie działa i metodzie uczenia niż tradycyjna sieć neuronowa jest sieć zwana os nazwiska twórcy koncepcji – sieć Kohonena. Najkrócej charakteryzuje się ją jako sieć samouczącą się z wbudowaną konkurencją i mechanizmem sąsiedztwa. Jest to sieć złożona z dwóch warstw neuronów: warstwy wejściowej i warstwy wyjściowej. Samouczenie polega na tym, że uczenie (trenowanie sieci odbywa się w trybie „bez nauczyciela”), co oznacza, że dla podawanych danych wejściowych do treningu nie jest przedstawiana prawidłowa odpowiedź. Sieć nie jest zapoznawana z tym, jakie sygnały wyjściowe powinny odpowiadać wprowadzanym sygnałom wejściowym.

Tworzę diagram, który zawiera trzy węzły: dane -BAZA2, Zastępowanie, SOM/Kohonen



Rysunek 64 Diagram Kohonen

Ustalam rolę i skalę pomiaru zmiennych:

(brak)		<input type="checkbox"/> nie równe					
Kolumny:		<input type="checkbox"/> Etykieta		<input type="checkbox"/> Eksploracja		<input type="checkbox"/> Podstawowe	
Nazwa	Rola	Poziom	Raport	Porządek	Porzucenie	Dolna granica	Górna granica
account_length	Wejście	Przedziałowa	Nie		Nie	.	.
area_code	Odrzucona	Przedziałowa	Nie		Nie	.	.
class	Odrzucona	Binarna	Nie		Nie	.	.
international_pla	Wejście	Binarna	Nie		Nie	.	.
number_custome	Wejście	Przedziałowa	Nie		Nie	.	.
number_vmil_m	Wejście	Przedziałowa	Nie		Nie	.	.
phone_number	Id.	Nominalna	Nie		Nie	.	.
state	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_day_calls	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_day_charge	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_day_minute	Wejście	Przedziałowa	Nie		Nie	.	.
total_eve_calls	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_eve_charge	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_eve_minute	Wejście	Przedziałowa	Nie		Nie	.	.
total_intl_calls	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_intl_charge	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_intl_minute	Wejście	Przedziałowa	Nie		Nie	.	.
total_night_calls	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_night_char	Odrzucona	Przedziałowa	Nie		Nie	.	.
total_night_minu	Wejście	Przedziałowa	Nie		Nie	.	.
voice_mail_plan	Wejście	Binarna	Nie		Nie	.	.

Rysunek 65 Role i poziomy zmiennych

W węźle ‘zastępowanie’, zmieniamy dla zmiennych przedziałowych domyślną metodę granic na ‘brak’.

Ogólne	
Id. węzła	Repl
Importowane dane	...
Eksportowane dane	...
Uwagi	...
Uczenie	
<input type="checkbox"/> Zmienne przedziałowe	
Edytor zastąpień	...
Domyślna metoda granic	Brak
Wartości obciążenia	...
<input type="checkbox"/> Zmienne klasyfikujące	
Edytor zastąpień	...
Nierozpoznane poziomy	Zignoruj
Wynik punktowy	
Wartości zastępujące	Wyliczone
Ukryj	Nie
Raport	
Raport o zastąpieniach	Tak
Status	
Czas utworzenia	26.05.2015 15:26

Rysunek 66 Ustawienia dla węzła Zastępowanie

W edytorze zastąpień przyjmuje, że zmienna numer_vmail_messages będzie przyjmować wartości nie większe niż 20.

Nazwa	Użycie	Metoda ustalania granic	Dolna granica zastępowania	Górna granica zastępowania	Metoda zastępowania
account_length	Domyślne	Domyślne	.	.	Domyślne
area_code	Domyślne	Domyślne	.	.	Domyślne
number_customers	Domyślne	Domyślne	.	.	Domyślne
number_vmail_messages	Domyślne	Podana przez użytkownika	.	20	Domyślne
state	Domyślne	Domyślne	.	.	Domyślne
total_day_calls	Domyślne	Domyślne	.	.	Domyślne
total_day_charges	Domyślne	Domyślne	.	.	Domyślne
total_day_minutes	Domyślne	Domyślne	.	.	Domyślne

Rysunek 67 Okno edytor zastąpień

W edycji zmiennych w węźle SOM/Kohonen zmienna class nie będzie brała udziału.

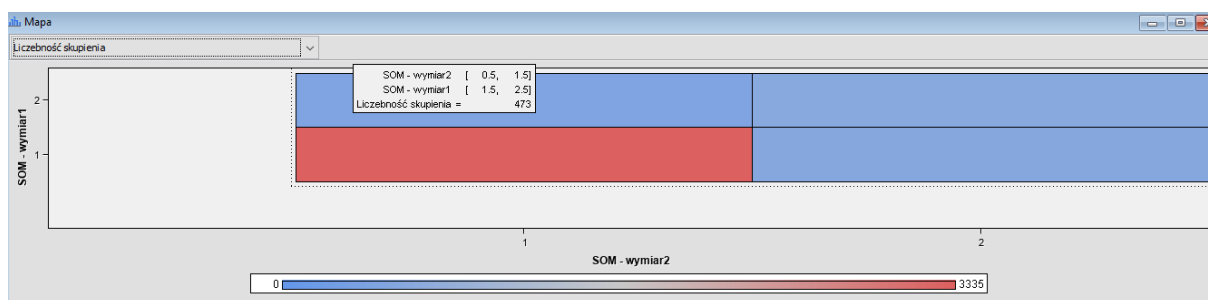
Nazwa	Użycie	Raport	
REP_number_vmail_messages	Domyślne	Nie	V
account_length	Domyślne	Nie	V
area_code	Domyślne	Nie	C
class	Nie	Nie	C
international_plans	Domyślne	Nie	V
number_customers	Domyślne	Nie	V
number_vmail_messages	Domyślne	Nie	C
phone_number	Domyślne	Nie	I

Rysunek 68 Ustawienia dla tabeli w węźle Kohonen

W węźle SOM/Kohonen ustalamy metodę SOM Kohonen, standaryzację wewnętrzną jako rozstęp, topologię sieci 2x2.

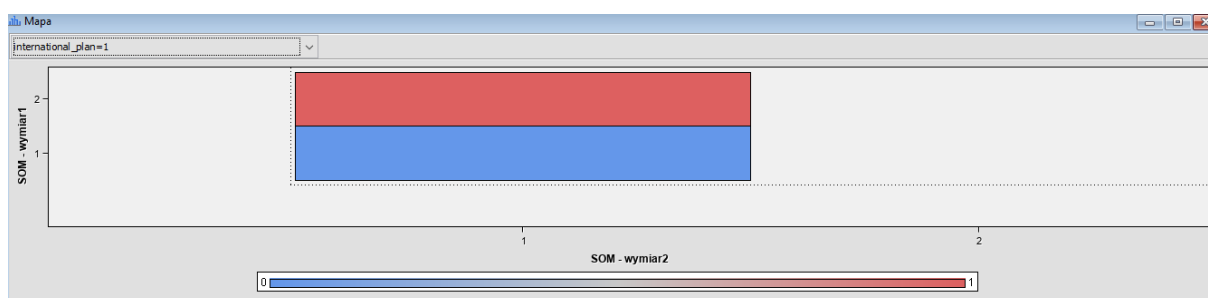
Ogólne	
Id. węzła	SOM
Importowane dane	***
Eksportowane dane	***
Uwagi	***
Uczenie	
Zmienne	***
Metoda	SOM Kohonena
Standaryzacja wewnętrzna	Rozstęp
<input type="checkbox"/> Segment	
- Wiersz	2
- Kolumna	2
<input type="checkbox"/> Opcje ziarna	
- Początkowa metoda	Domyslna
- Promień	0.0
<input type="checkbox"/> Uczenie wsadowe SOM	
- Użyj domyslnych	Tak
- Wygladzanie lokalne liniowe	Tak
- Wygladzanie Nadarayi-Watsona	Tak
<input type="checkbox"/> Opcje lokalne liniowe	
- Kryterium zbieżności	1.0E-4

Rysunek 69 Ustawienia dla węzła Kohonen



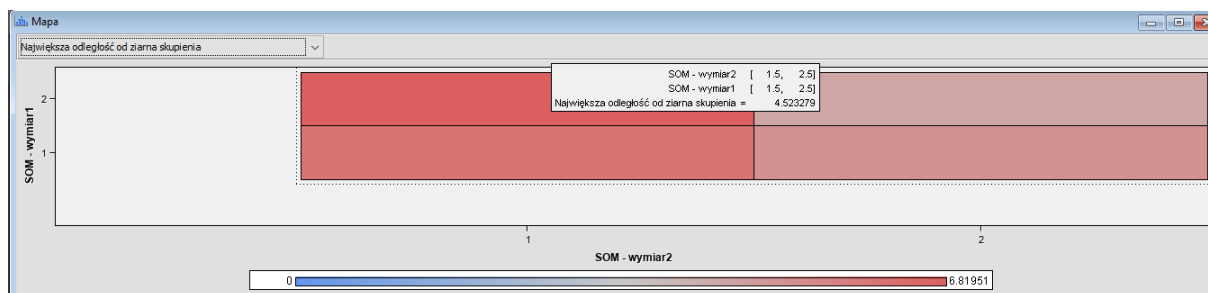
Rysunek 70 Liczebność skupień w poszczególnych grupach

Wyniki liczebności skupień (1,2) wynoszą 473, (2,2) – 615, (2,1)- 577, (1,1) – 3335.



Rysunek 71 Liczebność zmiennej plan międzynarodowy= 1

Na powyższym wykresie można zobaczyć udział w skupieniu zmiennej 'international_plan' dla wartości 1. W skupieniu (1,1) mają wartość 3,91E-15, natomiast w skupieniu (1,2) 1. Co oznacza, że wszystkie elementy w tym skupieniu dla planu międzynarodowego są równe 1.



Rysunek 72 Wykres jednorodności skupień - największa odległość od ziarna skupienia

Skupienie (2,2) jest najbardziej jednorodne, największa odległość od ziarna skupienia wynosi 4,52. Dla pozostałych skupień jest o wiele większa.

Kryterium skupienia	Największa względna zmiana ziaren skupień	Poprawa kryterium skupiania	SOM - id. segmentu	Liczebność skupienia
0.743846	0.09978	.	1	3335
0.743846	0.09978	.	2	577
0.743846	0.09978	.	3	473
0.743846	0.09978	.	4	615

Rysunek 73 Średnie statystyki

Największą liczebność skupienia miał segment 1 – 3335, a najmniejszą segment 3 – 473.

Podsumowanie

Szczegółowa analiza danych pozwala na wyciągnięcie istotnych wniosków, w celu przeprowadzania wyboru odpowiedniej metody. W celu prognozowania danych zmiennych SAS EM oferuje kilka możliwości. W moim przypadku rozpatrywałam różnice i dokładność stworzonego modelu pomiędzy drzewem decyzyjnym i regresją logistyczną. Zmienne 'relationship' i 'marital status' są najbardziej istotne w celu ustalenia czy dochód przekraczał 50 000 USD. Pochodzenie etniczne i kraj mają najsłabszy wpływ na zmienną celu. Najlepsza, przy aktualnych założeniach okazała się regresja logistyczna, ale tylko nieznacznie gorsze było drzewo decyzyjne. Wykonana metoda grupowania poprzez analizę skupień i sieciami neuronowymi Kohonena miała na celu eksplorację danych, typologię i redukcję danych. Analiza skupień obejmuje procedury, które pozwalają tworzyć grupy obiektów najmniej odległych od siebie lub najbardziej podobnych do siebie, które są traktowane jako punkty wielowymiarowej przestrzeni, gdzie wymiar przestrzeni jest określony liczbą zmiennych opisujących obiekty. Przykładowo podczas analizy skupień można się dowiedzieć, ile osób spełnia określone warunki (na podstawie zmiennych binarnych) na podstawie drzewa przynależności do skupień. Natomiast sieci neuronowe Kohonena pozwalają zobaczyć jak program poprzez uczenie sieci przyporządkowuje zmienne do skupień i tworzy określone kryteria.

Spis ilustracji

Rysunek 1 Logo SAS	3
Rysunek 2 Import bazy do tabel Sas-owych	5
Rysunek 3 Rola i poziomy dla zmiennych	5
Rysunek 4 Węzeł - eksploracja statystyk	6
Rysunek 5 Wykres rozrzutu dla poziomu wykształcenia i lat nauki	6
Rysunek 6 Histogram zmiennej wiek	6
Rysunek 7 Histogram przepracowanych zmiennych w tygodniu	7
Rysunek 8 Boxplot dla edukacji i godzin spędzonych tygodniowo w pracy	7
Rysunek 9 Wykres kołowy dla krajów	8
Rysunek 10 Wykres liczebności dla wykonywanych zawodów	8
Rysunek 11 Podział zmiennej celu	9
Rysunek 12 Chi-kwadrat dla zmiennych	9
Rysunek 13 Wykres istotności zmiennych	10
Rysunek 14 Podsumowanie statystyczne dla zmiennych	10
Rysunek 15 Brakujące wartości	10
Rysunek 16 Węzeł 3 Imputacja	11
Rysunek 17 Węzeł 4 Eksploracja statystyk 2	11
Rysunek 18 Węzeł 5 Porzucenie	11
Rysunek 19 Węzeł 6 Partycjonowanie	11
Rysunek 20 Ustawienia dla partycjonowania	12
Rysunek 21 Wynik partycjonowania	12
Rysunek 22 Dodanie 7 węzła Drzewo decyzyjne	13
Rysunek 23 Ustawienia dla drzewa decyzyjnego	13
Rysunek 24 Nakładka rankingów wyników punktowych: target	13
Rysunek 25 Statystyki liściowe	14
Rysunek 26 Statystyki dopasowania	14
Rysunek 27 Drzewo decyzyjne	15
Rysunek 28 Węzeł 9 Regresja	16
Rysunek 29 Nakładka rankingów wyników punktowych: target 2	16
Rysunek 30 Wykres efektów	17
Rysunek 31 Statystyki dopasowania	17
Rysunek 32 Raport - wynik	18
Rysunek 33 Węzeł 10 Porównanie modeli	19
Rysunek 34 Wykres ROC: target	19
Rysunek 35 Nakładka rankingów wyników punktowych: target	19
Rysunek 36 Statystyki dopasowania	20
Rysunek 37 Tabela dopasowania statystyk dla zbioru treningowego	20
Rysunek 38 Tabela dopasowania statystyk dla zbioru walidacyjnego	21
Rysunek 39 Tabela dopasowania statystyk dla zbioru testowego	21
Rysunek 40 Tabela porównania statystyk modelu dla drzewa decyzyjnego i regresji logistycznej	22
Rysunek 41 Baza dla metod grupowania	24
Rysunek 42 Węzeł 2 Klasteryzacja	25
Rysunek 43 Ustalenie ról i poziomów dla zmiennych	25
Rysunek 44 Ustawienia dla węzła klasteryzacji	26
Rysunek 45 Wykres segmentowy	26
Rysunek 46 Tabela statystyki średnich 1	27
Rysunek 47 Tabela statystyki średnich 2	27

Rysunek 48 Wykres kołowy - rozmiar segmentu	27
Rysunek 49 Tabela - odległość skupień	27
Rysunek 50 Tabela - istotność zmiennych.....	28
Rysunek 51 Drzewo przynależności obiektów do skupień	28
Rysunek 52 Wykreś średnich wejściowych	29
Rysunek 53 Statystyki średnich	29
Rysunek 54 Pliki tworzone i eksportowane przez węzeł Klasteryzacja	29
Rysunek 55 Okno eksploracji EMWS2.Clus_TRAIN	30
Rysunek 56 Wykres rozproszenia z zaznaczeniem przynależności obserwacji do skupień	30
Rysunek 57 Liczebność skupień na wykresie słupkowym	31
Rysunek 58 Wykres kołowy rozkładu zmiennej class.....	31
Rysunek 59 Węzeł 3 Segmentacja profilu	31
Rysunek 60 Wyniki Segmentacji - grupa 1.....	32
Rysunek 61 Wyniki Segmentacji - grupa 2	32
Rysunek 62 Wyniki Segmentacji - grupa 3	33
Rysunek 63 Wizualizacja segmentów.....	33
Rysunek 64 Diagram Kohonen	34
Rysunek 65 Role i poziomy zmiennych.....	34
Rysunek 66 Ustawienia dla węzła Zastępowanie	35
Rysunek 67 Okno edytor zastąpień	35
Rysunek 68 Ustawienia dla tabeli w węźle Kohonen	35
Rysunek 69 Ustawienia dla węzła Kohonen	36
Rysunek 70 Liczebność skupień w poszczególnych grupach.....	36
Rysunek 71 Liczebność zmiennej plan międzynarodowy= 1.....	36
Rysunek 72 Wykres jednorodności skupień - największa odległość od ziarna skupienia	37
Rysunek 73 Średnie statystyki	37

Bibliografia

- *Metody Data Mining w analizowaniu i prognozowaniu kondycji ekonomicznej przedsiębiorstw*, M. Lasek, Studia Informatyki Gospodarczej
- *Data Mining Using SAS Enterprise Miner: A Case Study Approach, Second Edition*. SAS Publishing
- *Data Mining Using SAS Enterprise Miner*, R. Matignon, A John Wiley & Sons, Inc., Publication
- *Decision Trees for Business Intelligence and Data Minign Using SAS Enterprise Miner*, Barry de Ville, SAS Press Series
- *Introduction to Data Minig Using SAS Enterprise Mining*, Patricia B. Cerrito