

### 3.4, 228884, Mateusz Guściora

#### 3.4.1

a) przykładowe procesy, dla zbioru danych Labor-Negotiations, Titanic.

Zapoznajemy się z podanymi blokami `replace missing values`- zastępujący wartości wartościami podanymi w parametrach, operator `impute missing values`- operator szacujący brakujące wartości na podstawie nadanego modelu, zawierający podproces. `Replace all missings`- Który automatycznie uzupełnia brakujące dane w całym zbiorze.

Uruchomiono procesy przykładowe dla bloku `impute missing values` dla `Labour-Negotiations`, dla `replace all missing values` dla `Labor_Negotiations` oraz dla `replace all missings` dla `Titanic`.

Stworzono proces z wykorzystaniem `multiply` dla dwóch strumieni z użyciem `Replace Missing Values` oraz `Impute missing values`. Poniżej porównanie `performance vector` dla algorytmu `rule induction` z operatorem `Replace missing values` i bez („z wykorzystaniem danych oryginalnych i zbiorów z uzupełnionymi brakami danych”)

Poniżej z `Replace Missing values Labor Negotiations`

**accuracy: 75.00% +/- 20.41% (micro average: 75.00%)**

	true bad	true good	class precision
pred. bad	10	6	62.50%
pred. good	4	20	83.33%
class recall	71.43%	76.92%	

**kappa: 0.420 +/- 0.377 (micro average: 0.385)**

	true bad	true good	class precision
pred. bad	8	5	61.54%
pred. good	6	21	77.78%
class recall	57.14%	80.77%	

Po próbie uruchomienia bez operatora `Replace missing values` stwierdzono, że proces trwa zbyt długo i nie uzyskano wyników.

b)

Został stworzony plik `csv bank6missingvalues` który różni się od `bank6` tym że nie ma wartości w miejscach `unknown` i `999.0` w `pdays`.

Proces dla danych bankowych-`bank6missing values`

**accuracy: 89.21% +/- 0.21% (micro average: 89.21%)**

	true no	true yes	class precision
pred. no	36741	3891	90.42%
pred. yes	771	1785	69.84%
class recall	97.94%	31.45%	

**kappa: 0.383 +/- 0.017 (micro average: 0.383)**

	true no	true yes	class precision
pred. no	36741	3891	90.42%
pred. yes	771	1785	69.84%
class recall	97.94%	31.45%	

Wyniki dla danych bank6 tzn. danych gdzie znajduje się unknown i 999 w pdays

**accuracy: 89.08% +/- 0.25% (micro average: 89.08%)**

	true no	true yes	class precision
pred. no	36595	3797	90.60%
pred. yes	917	1879	67.20%
class recall	97.56%	33.10%	

**kappa: 0.391 +/- 0.016 (micro average: 0.391)**

	true no	true yes	class precision
pred. no	36595	3797	90.60%
pred. yes	917	1879	67.20%
class recall	97.56%	33.10%	

c)Proces przykładowy - churn modeling

Dla rule induction z użyciem operatora Replace missing values gdzie jako w parametrach w liście wybrano unknown i pdays(w obu jako average) performance vector pokazano poniżej gdzie 89,18

Accuracy daje zadowalający wynik, kappa natomiast wynosi 0,382.

**kappa: 0.382 +/- 0.021 (micro average: 0.382)**

	true no	true yes	class precision
pred. no	36734	3893	90.42%
pred. yes	778	1783	69.62%
class recall	97.93%	31.41%	

**accuracy: 89.18% +/- 0.32% (micro average: 89.18%)**

	true no	true yes	class precision
pred. no	36734	3893	90.42%
pred. yes	778	1783	69.62%
class recall	97.93%	31.41%	

d) Proces przykładowy Churn modeling z operatorem Cost-Sensitive Scoring i operatorem SMOTE Upsampling.

Cost-Sensitive Scoring- jest to operator/algorytm wykorzystujący jeden model uczenia maszynowego. Tylko jeden model jest uczony. Operator SMOTE Upsampling, najpierw trzeba było pobrać poprzez dodanie toolbox do extensions. Jest to operator służący dla niezbalansowanych zbiorów, korzysta także z algorytmu k-najbliższych sąsiadów.

Dla danych bank6

Przy strumieniu z cross validation zawierającej algorytm decision tree i operator Cost-sensitive Scoring

**accuracy: 86.86% +/- 0.01% (micro average: 86.86%)**

	true no	true yes	class precision
pred. no	37512	5676	86.86%
pred. yes	0	0	0.00%
class recall	100.00%	0.00%	

Kappa 0?(nie wiem dlaczego)

e)

Dla danych niekompletnych

Używanie operatorów `replace missing` czy `impute missing values` nie zmieniało zbytnio dokładności predykcji oraz wartości kappa wyniki oscylowały w granicach 89% a w przypadku danych z unknown i z 999 także 89% dokładności.