

Dominik Wróbel

Inżynieria oprogramowania i systemów

Informatyka, II stopień, 2018/19

Metody eksploracji danych

Laboratorium 3 – 26.03.2019

Regresja logistyczna

### 3.1 Przygotowanie danych

Dane zawarte w plikach egzamin-cpp.csv, egzamin-cpp-train.csv, egzamin-cpp-test.csv przekształcono do plików .arff przy pomocy komend i opcji:

- `java -cp /opt/weka/weka.jar weka.core.converters.CSVLoader egzamin-cpp.csv -F ';' -S 1 -D 3 -format "yyyy-mm-dd" > egzamin-cpp.arff`
- `java -cp /opt/weka/weka.jar weka.core.converters.CSVLoader egzamin-cpp-test.csv -F ';' -S 1 -D 3 -format "yyyy-mm-dd" -N 5 > egzamin-cpp-test.arff`
- `java -cp /opt/weka/weka.jar weka.core.converters.CSVLoader egzamin-cpp-train.csv -F ';' -S 1 -D 3 -format "yyyy-mm-dd" > egzamin-cpp-train.arff`

Fragmenty uzyskanych plików egzamin-cpp.arff oraz egzamin-cpp-test.arff przedstawione są poniżej.

egzamin-cpp.arff	egzamin-cpp-test.arff
1 @relation egzamin-cpp	1 @relation egzamin-cpp-test
2	2
3 @attribute ImieNazwisko string	3 @attribute ImieNazwisko string
4 @attribute OcenaC numeric	4 @attribute OcenaC numeric
5 @attribute DataC date yyyy-mm-dd	5 @attribute DataC date yyyy-mm-dd
6 @attribute OcenaCpp numeric	6 @attribute OcenaCpp numeric
7 @attribute Egzamin numeric	7 @attribute Egzamin {*unknown*}
8	8
9 @data	9 @data
10	10
11 'Dqhoil Dhxppluj',3.5,2016-01-14,4,3	11 'Ynjln Njjmbd',3.5,2016-01-20,3,?
12 'Bhnhgpxj Lwjmq',4.5,2016-01-14,4,3	12 'Cjpwbybuiq Hptkfnanlnspv',4.5,2016-01-22,4,?
13 'Wkgjnerme Djfbw',4,2016-01-20,3,2	13 'Ilfcoq Jumbwxhpjy',5,2016-01-19,5,?
14 'Sredvmuwt Tcimknl',4.5,2016-01-20,4.5,3.5	14 'Ealxab Kojmtw',4.5,2016-01-20,3.5,?
15 'Tiowe Bqoilngbrx',4,2016-01-14,4.5,3	15 'Ysbxkv Fwxo',3.5,2016-01-14,3,?
16 'Bvaysqv Wuyih',3.5,2016-01-14,5,3	16 'Yjfdsu Ogiaw',4,2016-01-20,5,?
17 'Jjoaxp Ktapcy',5,2016-01-20,4,3.5	17 'Wcmuixng Mlwrt',5,2016-01-20,4,?
18 'Mkengbtw Aainhh',3.5,2016-01-20,3,2	18 'Ihtuoileg Fwffgx',5,2016-01-14,4,?
19 'Fbfffj Muupwshu',4,2016-01-14,5,4	19 'Niwa Voovevy',5,2016-01-14,5,?
20 'Yahwfyp Bvnlsig',5,2016-01-14,4.5,4	20 'Reu Gxbqr',3.5,2016-01-20,3.5,?
21 'Ecwjmpr Krixbwvk',5,2016-01-20,3,3	21 'Yjtity Betvbfqcty',3.5,2016-01-14,3,?
22 'Ikoj Xvbskh',5,2016-01-20,3.5,4.5	22 'Ttctd Lkwms',4.5,2016-01-19,4.5,?
23 'Yimqki Fwchtt',4.5,2016-01-20,3.5,3	23 'Fkxlje Whrpnysg',4,2016-01-14,3.5,?
24 'Pnbnsft Fecloeimw',3.5,2016-01-14,3,2	24 'Bwcmey Ekngovpwc',4.5,2016-01-14,5,?
25 'Byieeoy Pxxhrx',5,2016-01-20,3,2	25 'Onbfo Yjvfwjooj',4,2016-01-20,4.5,?
26 'Kbhjpsps Yyaxmiqvx',4.5,2016-01-14,5,3	26 'Vnxyon Yxssjcg',4,2016-01-14,3.5,?
27 'Ddvictnni Gugtc',3,2016-01-20,3,2	27 'Ttxpoll Ylipratxi',4.5,2016-01-19,5,?
28 'Iftvf Benjw',4.5,2016-01-20,4,3	28 'Xkfhun Roam',5,2016-01-20,4.5,?

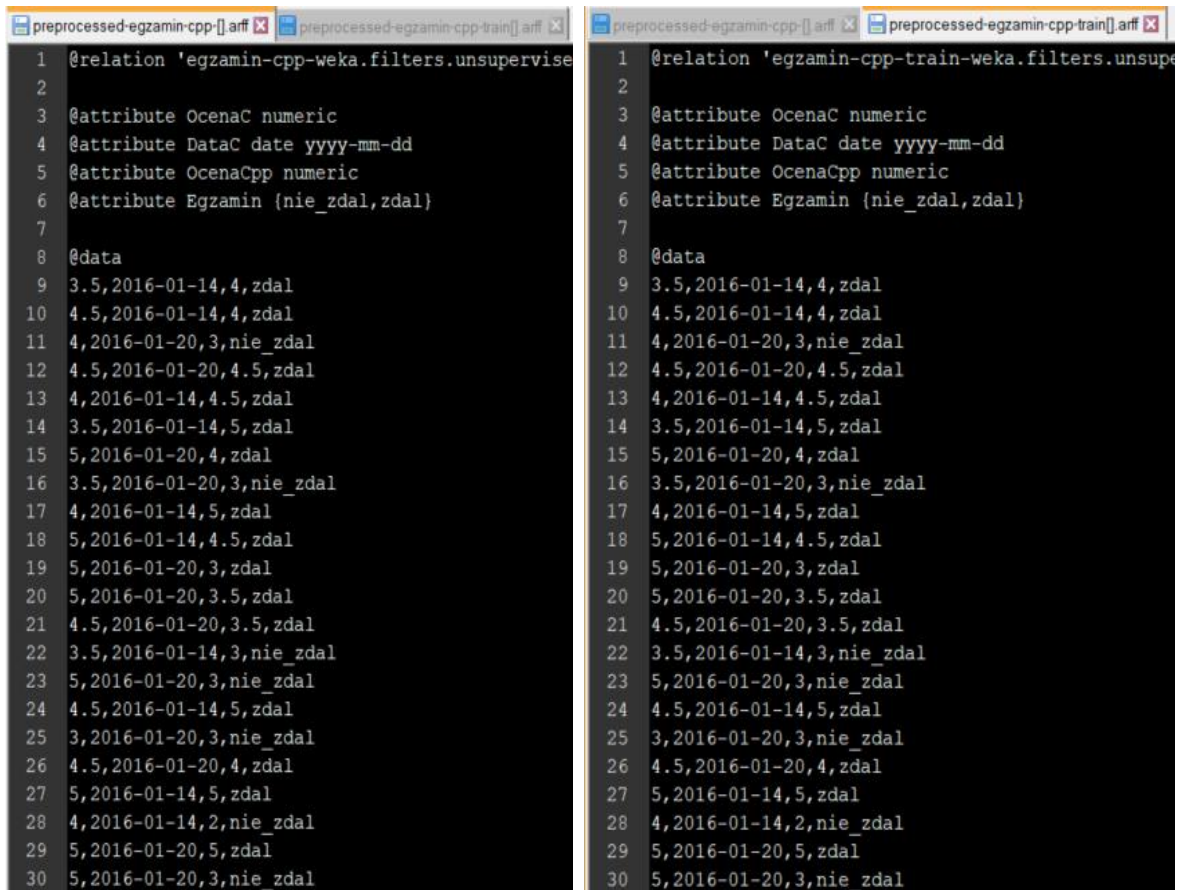
### 3.2 Konwersja danych

W tym zadaniu poddano konwersji pliki egzamin-cpp.arff oraz egzamin-cpp-train.arff. Konwersja miała na celu zmianę atrybutu reprezentującego wynik z egzaminu (liczbowego) na wartość nominalną reprezentowaną przez etykietę (zdal, nie\_zdal). Z danych usunięto również imiona i nazwiska jako dane, które są nieistotne z punktu widzenia analizy danych.

Konwersję danych przeprowadzono przy pomocy Weka Explorer, kolejno użyte przekształcenia to:

- filters.unsupervised.attribute.NumericToNominal –R last
- filters.unsupervised.attribute.RenameNominalValues –R 5 –N 2:nie\_zdal
- filters.unsupervised.attribute.MergeManyValues –C last –L zdal –R 2,3,4,5,6 -unset-class-temporarily
- filters.unsupervised.attribute.Remove -R1

Fragmenty uzyskanych plików, które nazwano zgodnie z instrukcją preprocessed-egzamin-cpp-[] .arff oraz preprocessed-egzamin-cpp-train[] .arff przedstawione są poniżej.



```
1 @relation 'egzamin-cpp-weka.filters.unsupervised.attribute.NumericToNominal'
2
3 @attribute OcenaC numeric
4 @attribute DataC date yyyy-mm-dd
5 @attribute OcenaCpp numeric
6 @attribute Egzamin {nie_zdal,zdal}
7
8 @data
9 3.5,2016-01-14,4,zdal
10 4.5,2016-01-14,4,zdal
11 4,2016-01-20,3,nie_zdal
12 4.5,2016-01-20,4.5,zdal
13 4,2016-01-14,4.5,zdal
14 3.5,2016-01-14,5,zdal
15 5,2016-01-20,4,zdal
16 3.5,2016-01-20,3,nie_zdal
17 4,2016-01-14,5,zdal
18 5,2016-01-14,4.5,zdal
19 5,2016-01-20,3,zdal
20 5,2016-01-20,3.5,zdal
21 4.5,2016-01-20,3.5,zdal
22 3.5,2016-01-14,3,nie_zdal
23 5,2016-01-20,3,nie_zdal
24 4.5,2016-01-14,5,zdal
25 3,2016-01-20,3,nie_zdal
26 4.5,2016-01-20,4,zdal
27 5,2016-01-14,5,zdal
28 4,2016-01-14,2,nie_zdal
29 5,2016-01-20,5,zdal
30 5,2016-01-20,3,nie_zdal
```

```
1 @relation 'egzamin-cpp-train-weka.filters.unsupervised.attribute.NumericToNominal'
2
3 @attribute OcenaC numeric
4 @attribute DataC date yyyy-mm-dd
5 @attribute OcenaCpp numeric
6 @attribute Egzamin {nie_zdal,zdal}
7
8 @data
9 3.5,2016-01-14,4,zdal
10 4.5,2016-01-14,4,zdal
11 4,2016-01-20,3,nie_zdal
12 4.5,2016-01-20,4.5,zdal
13 4,2016-01-14,4.5,zdal
14 3.5,2016-01-14,5,zdal
15 5,2016-01-20,4,zdal
16 3.5,2016-01-20,3,nie_zdal
17 4,2016-01-14,5,zdal
18 5,2016-01-14,4.5,zdal
19 5,2016-01-20,3,zdal
20 5,2016-01-20,3.5,zdal
21 4.5,2016-01-20,3.5,zdal
22 3.5,2016-01-14,3,nie_zdal
23 5,2016-01-20,3,nie_zdal
24 4.5,2016-01-14,5,zdal
25 3,2016-01-20,3,nie_zdal
26 4.5,2016-01-20,4,zdal
27 5,2016-01-14,5,zdal
28 4,2016-01-14,2,nie_zdal
29 5,2016-01-20,5,zdal
30 5,2016-01-20,3,nie_zdal
```

### 3.2 Klasyfikator

W tym zadaniu przeprowadzono klasyfikację przy użyciu regresji logistycznej. Zadanie wykonano w Weka Explorer.

Dane uzyskane z regresji **korzystając z walidacji krzyżowej**:

=== Run information ===

```
Scheme:      weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation:    egzamin-cpp-weka.filters.unsupervised.attribute.NumericToNominal-Rlast-wek
Instances:   103
Attributes:  4
              OcenaC
              DataC
              OcenaCpp
              Egzamin
Test mode:   10-fold cross-validation
```

=== Classifier model (full training set) ===

Logistic Regression with ridge parameter of 1.0E-8  
Coefficients...

Variable	Class nie_zdal
OcenaC	-0.9985
DataC	0
OcenaCpp	-2.0411
Intercept	-364.4042

Odds Ratios...

Variable	Class nie_zdal
OcenaC	0.3684
DataC	1
OcenaCpp	0.1299

Time taken to build model: 0.02 seconds

(Cz.2 wyników na kolejnej stronie)

1. Wzór na hiperpłaszczyznę separującą:

$$-364,4042 - 0,9985 \cdot OcenaC - 0 \cdot DataC - 2,0411 \cdot OcenaCpp = 0$$

2. Wpływ wzrostu/spadku ocen na szanse zdania/niezdania egzaminu

Ocena	Wzrost o 1	Wzrost o 1
OcenaC	$szansa(nie\_zdanie) = e^{-0,9985} = 0,3684$	$szansa(zdanie) = e^{0,9985} = 2.7142$
OcenaCpp	$szansa(nie\_zdanie) = e^{-2,0411} = 0,1299$	$szansa(zdanie) = e^{2,0411} = 7.6991$

3. Data nie ma wpływu na uzyskaną ocenę z egzaminu, ponieważ

$$e^0 = 1$$

```
=== Stratified cross-validation ===
=== Summary ===
```

Correctly Classified Instances	81	78.6408 %
Incorrectly Classified Instances	22	21.3592 %
Kappa statistic	0.4343	
Mean absolute error	0.2455	
Root mean squared error	0.3631	
Relative absolute error	63.015 %	
Root relative squared error	82.4758 %	
Total Number of Instances	103	

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,556	0,132	0,600	0,556	0,577	0,435	0,860	0,693	nie_zdal
	0,868	0,444	0,846	0,868	0,857	0,435	0,860	0,946	zdal
Weighted Avg.	0,786	0,362	0,782	0,786	0,784	0,435	0,860	0,880	

```
=== Confusion Matrix ===
```

```

a  b  <-- classified as
15 12 |  a = nie_zdal
10 66 |  b = zdal
```

4. Z wyników widoczne jest, że więcej poprawnych klasyfikacji uzyskano dla klasy 'zdal' (TP = 0,868) niż dla klasy 'nie\_zdal' (TP = 0,556). Z 25 osób, które nie zdały, 10 zostało zakwalifikowane jako te, które zdały egzamin. Z 78 osób, które zdały egzamin, 12 zostało zakwalifikowane jako te, które nie zdały egzaminu.
5. W wyniku zastosowania zbioru uczącego otrzymano minimalne zmodyfikowane dano, różnice zaznaczono kolorem niebieskim poniżej. Zastosowania zbioru uczącego pozwoliła na poprawne przypisanie do klas większej liczby instancji.

```
=== Run information ===
```

```

Scheme:      weka.classifiers.functions.Logistic -R 1.0
Relation:    egzamin-cpp-weka.filters.unsupervised.attr
Instances:   103
Attributes:  4
              OcenaC
              DataC
              OcenaCpp
              Egzamin
Test mode:   evaluate on training data
```

```
=== Classifier model (full training set) ===
```

```
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
```

Variable	Class
nie_zdal	
OcenaC	-0.9985
DataC	0
OcenaCpp	-2.0411
Intercept	-364.4042

```
Odds Ratios...
```

Variable	Class
nie_zdal	
OcenaC	0.3684
DataC	1
OcenaCpp	0.1299

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	83	80.5825 %
Incorrectly Classified Instances	20	19.4175 %
Kappa statistic	0.4729	
Mean absolute error	0.2357	
Root mean squared error	0.3461	
Relative absolute error	60.5966 %	
Root relative squared error	78.6819 %	
Total Number of Instances	103	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,556	0,105	0,652	0,556	0,600	0,476	0,883	0,734	nie_zdal
	0,895	0,444	0,850	0,895	0,872	0,476	0,883	0,956	zdal
Weighted Avg.	0,806	0,356	0,798	0,806	0,801	0,476	0,883	0,898	

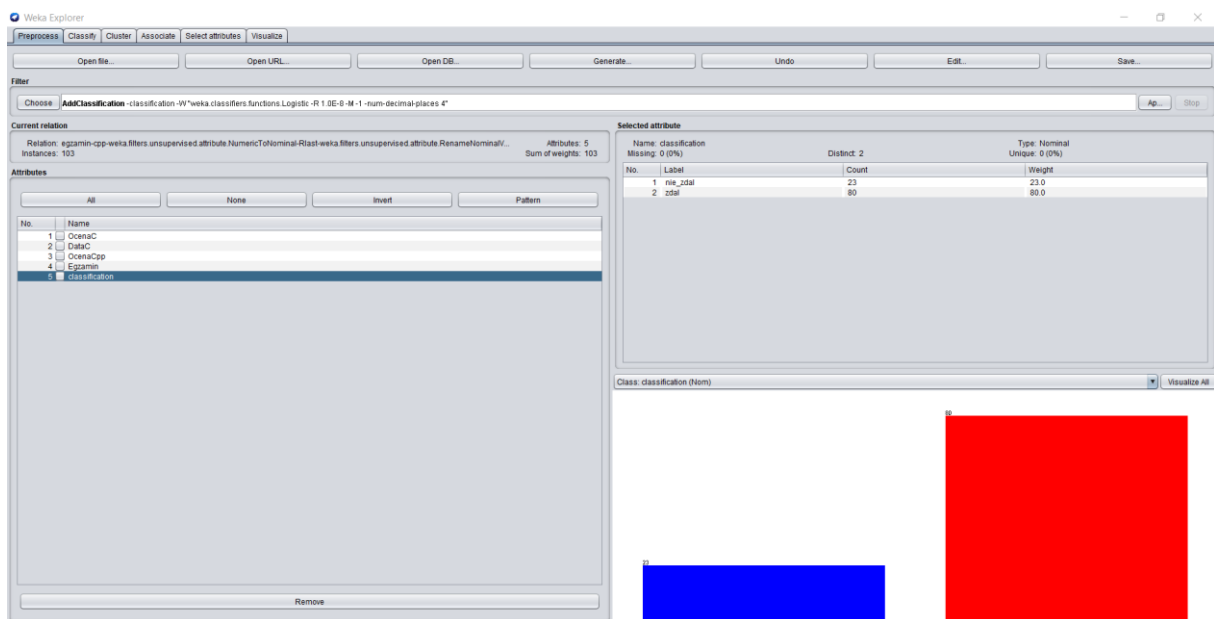
=== Confusion Matrix ===

```
a b <-- classified as
15 12 | a = nie_zdal
 8 68 | b = zdal
```

### 3.3 Porównaj jawnie wyniki klasyfikacji

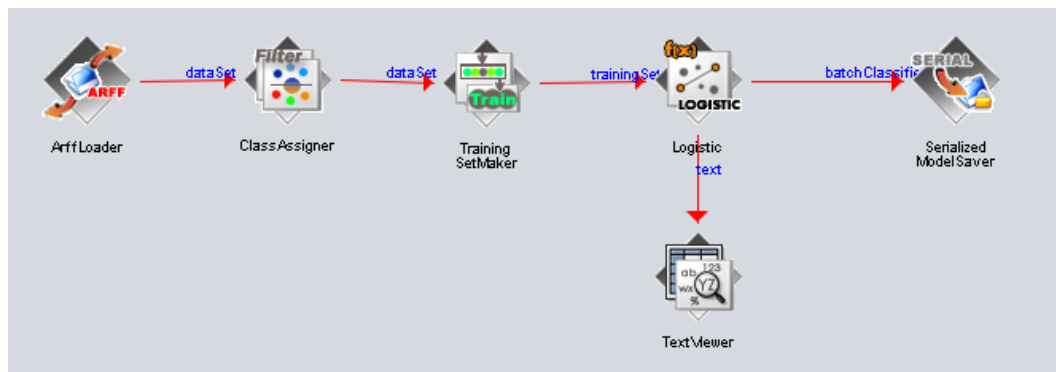
W wyniku zastosowania AddClassification otrzymano 23 osoby, które nie zdały egzaminu oraz 80, które zdały egzamin.

Selected attribute			
Name: classification		Type: Nominal	
Missing: 0 (0%)		Unique: 0 (0%)	
Distinct: 2			
No.	Label	Count	Weight
1	nie_zdal	23	23.0
2	zdal	80	80.0



### 3.4 Klasyfikacja w Knowledge Flow

W pierwszej części tego zadania zbudowano model regresji logistycznej na podstawie pliku preprocessed-egzamin-cpp-train.arff i zapisano go do pliku.



Uzyskany model, który zapisano do pliku:

```
=== Classifier model ===

Scheme: Logistic
Relation: egzamin-cpp-train-weka.filters.unsupervised.a

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
      Class
Variable  nie_zdal
=====
OcenaC    -0.9198
DataC      0
OcenaCpp   -2.5471
Intercept -1952.6739

Odds Ratios...
      Class
Variable  nie_zdal
=====
OcenaC    0.3986
DataC      1
OcenaCpp   0.0783
```

Następnie, zbudowano model, który przetwarza plik z danymi testowymi egzamin-cpp-test.arff przy użyciu modelu regresji logistycznej, który został wcześniej stworzony. W wyniku uzyskano plik, którego fragment przedstawiono poniżej:

 Text Viewer

Result list

18:01:27.791 - testSet:

Text

```
@relation 'egzamin-cpp-test-weka.filters.unsupervised.attribute.Remove-R:


@attribute OcenaC numeric
@attribute DataC date yyyy-mm-dd
@attribute OcenaCpp numeric
@attribute classification {nie_zdal,zdal}
@attribute distribution_nie_zdal numeric
@attribute distribution_zdal numeric

@data
3.5,2016-01-20,3,nie_zdal,0.808499,0.191501
4.5,2016-01-22,4,zdal,0.142711,0.857289
5,2016-01-19,5,zdal,0.005764,0.994236
4.5,2016-01-20,3.5,zdal,0.320153,0.679847
3.5,2016-01-14,3,nie_zdal,0.676852,0.323148
4,2016-01-20,5,zdal,0.016083,0.983917
5,2016-01-20,4,zdal,0.076809,0.923191
5,2016-01-14,4,zdal,0.03964,0.96036
5,2016-01-14,5,zdal,0.003222,0.996778
3.5,2016-01-20,3.5,nie_zdal,0.541591,0.458409
3.5,2016-01-14,3,nie_zdal,0.676852,0.323148
4.5,2016-01-19,4.5,zdal,0.031769,0.968231
4,2016-01-14,3.5,zdal,0.270102,0.729898
4.5,2016-01-14,5,zdal,0.005094,0.994906
4,2016-01-20,4.5,zdal,0.055188,0.944812
4,2016-01-14,3.5,zdal,0.270102,0.729898
4.5,2016-01-19,5,zdal,0.009099,0.990901
5,2016-01-20,4.5,zdal,0.022753,0.977247
3,2016-01-14,3.5,zdal,0.481436,0.518564
3.5,2016-01-14,2,nie_zdal,0.96396,0.03604
3.5,2016-01-19,4.5,zdal,0.076058,0.923942
5,2016-01-20,5,zdal,0.006473,0.993527
4,2016-01-14,3,nie_zdal,0.569407,0.430593
4,2016-01-14,3.5,zdal,0.270102,0.729898
4,2016-01-14,4,zdal,0.093839,0.906161
4,2016-01-20,5,zdal,0.016083,0.983917
4,2016-01-20,4.5,zdal,0.055188,0.944812
5,2016-01-20,5,zdal,0.006473,0.993527
4,2016-01-19,3.5,zdal,0.398916,0.601084
4.5,2016-01-20,4.5,zdal,0.035566,0.964434
4.5,2016-01-14,4,zdal,0.061367,0.938633
4.5,2016-01-14,5,zdal,0.005094,0.994906
4,2016-01-20,5,zdal,0.016083,0.983917
5,2016-01-14,5,zdal,0.003222,0.996778
3.5,2016-01-22,4,zdal,0.294601,0.705399
4,2016-01-14,4.5,zdal,0.028163,0.971837
4.5,2016-01-14,3,zdal,0.455003,0.544997
4.5,2016-01-14,5,zdal,0.005094,0.994906
3,2016-01-19,3,nie_zdal,0.856112,0.143888
4.5,2016-01-20,4,zdal,0.116438,0.883562
3.5,2016-01-14,3,nie_zdal,0.676852,0.323148
```



### 3.5 Test modelu na podstawie zbioru egzamin-cpp-train

W tym zadaniu skonfigurowano workflow w taki sposób aby czytać dane zawierające wszystkie możliwe kombinacje ocen z języka C i C++. W wyniku uzyskano plik przedstawiony poniżej.

 Text Viewer

Result list

18:04:59.283 - testSet:

Text

```
@relation 'grid-weka.filters.unsupervised.attribute.Remove-Rfirst

@attribute OcenaC numeric
@attribute DataC date yyyy-mm-dd
@attribute OcenaCpp numeric
@attribute classification {nie_zdal,zdal}
@attribute distribution_nie_zdal numeric
@attribute distribution_zdal numeric

@data
3,2016-09-01,2,nie_zdal,0.902759,0.097241
3,2016-09-01,3,zdal,0.420965,0.579035
3,2016-09-01,3.5,zdal,0.169053,0.830947
3,2016-09-01,4,zdal,0.053866,0.946134
3,2016-09-01,4.5,zdal,0.015682,0.984318
3,2016-09-01,5,zdal,0.004439,0.995561
3.5,2016-09-01,2,nie_zdal,0.854252,0.145748
3.5,2016-09-01,3,zdal,0.314595,0.685405
3.5,2016-09-01,3.5,zdal,0.113824,0.886176
3.5,2016-09-01,4,zdal,0.034697,0.965303
3.5,2016-09-01,4.5,zdal,0.009958,0.990042
3.5,2016-09-01,5,zdal,0.002807,0.997193
4,2016-09-01,2,nie_zdal,0.787252,0.212748
4,2016-09-01,3,zdal,0.224674,0.775326
4,2016-09-01,3.5,zdal,0.07501,0.92499
4,2016-09-01,4,zdal,0.022189,0.977811
4,2016-09-01,4.5,zdal,0.00631,0.99369
4,2016-09-01,5,zdal,0.001774,0.998226
4.5,2016-09-01,2,nie_zdal,0.70026,0.29974
4.5,2016-09-01,3,zdal,0.154656,0.845344
4.5,2016-09-01,3.5,zdal,0.048703,0.951297
4.5,2016-09-01,4,zdal,0.014125,0.985875
4.5,2016-09-01,4.5,zdal,0.003993,0.996007
4.5,2016-09-01,5,zdal,0.001121,0.998879
5,2016-09-01,2,nie_zdal,0.595952,0.404048
5,2016-09-01,3,zdal,0.103545,0.896455
5,2016-09-01,3.5,zdal,0.031311,0.968689
5,2016-09-01,4,zdal,0.008964,0.991036
5,2016-09-01,4.5,zdal,0.002525,0.997475
5,2016-09-01,5,zdal,0.000708,0.999292
```

Wyznaczone w tym modelu prawdopodobieństwa mają w większości skrajne wartości (bardzo małe prawdopodobieństwo nie zdania przy dużym prawdopodobieństwie zdania).



### 3.6 Test modelu na podstawie zbioru egzamin-cpp

W tym zadaniu zmieniono sposób generacji modelu regresji. Model generowany jest na podstawie pliku preprocessed-egzamin-cpp.arff. Uzyskany model

```
=== Classifier model ===

Scheme: Logistic
Relation: egzamin-cpp-weka.filters.unsupervised.attribute.NumericToNominal-Rlas

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
      Class
Variable  nie_zdal
=====
OcenaC    -0.9985
DataC      0
OcenaCpp   -2.0411
Intercept -364.4042

Odds Ratios...
      Class
Variable  nie_zdal
=====
OcenaC     0.3684
DataC       1
OcenaCpp    0.1299
```

Następnie na podstawie nowego modelu ponownie przeprowadzono klasyfikację. Uzyskane dane zestawiono w tabeli zgodnie z poleceniem. Porównując te wyniki do tych wyznaczonych w 3.5 widać bardzo duży wpływ zbioru uczącego na wynik klasyfikacji.

OcenaC	OcenaCpp	Egzamin	Prawdopodobieństwo
3	2	nie_zdal	0.961375
3	3	nie_zdal	0.763757
3	3.5	nie_zdal	0.538138
3	4	zdal	0.704266
3	4.5	zdal	0.868555
3	5	zdal	0.948279
3.5	2	nie_zdal	0.93792
3.5	3	nie_zdal	0.662431
3.5	3.5	zdal	0.585744
3.5	4	zdal	0.796885
3.5	4.5	zdal	0.915868
3.5	5	zdal	0.967954
4	2	nie_zdal	0.901676
4	3	nie_zdal	0.543614
4	3.5	zdal	0.699652
4	4	zdal	0.866016
4	4.5	zdal	0.947186
4	5	zdal	0.980301
4.5	2	nie_zdal	0.847709
4.5	3	zdal	0.580382
4.5	3.5	zdal	0.793292
4.5	4	zdal	0.914153
4.5	4.5	zdal	0.967263
4.5	5	zdal	0.987949
5	2	nie_zdal	0.771623
5	3	zdal	0.694997
5	3.5	zdal	0.863436
5	4	zdal	0.946072
5	4.5	zdal	0.97987
5	5	zdal	0.992651

