Explaining Classifications for Individual Instances

3 marca 2019

Wpływ atrybutu A_i na predykcję dla danej obserwacji:

- x- obserwacja
- A_i - atrybuty
- Model klasyfikacji:

$$f : x \mapsto f(x) \tag{1}$$

Wpływ atrybutu A_i na predykcję dla danej obserwacji:

- x- obserwacja
- A_i - atrybuty
- Model klasyfikacji:

$$f : x \mapsto f(x) \tag{1}$$

-

$$predDiff_i(x) = f(x) - f(x \backslash A_i) \tag{2}$$

Sposoby ewaluacji predDiff dla problemu klasyfikacji;

- information difference

Sposoby ewaluacji predDiff dla problemu klasyfikacji;

- information difference
- weight of evidence

Sposoby ewaluacji predDiff dla problemu klasyfikacji;

- information difference
- weight of evidence
- difference of probabilities

Sposoby ewaluacji predDiff dla problemu klasyfikacji;

- information difference
- $$infDiff_i(y|x) = \log_2 p(y|x) - \log_2 p(y|x \backslash A_i) \qquad (3)$$
- weight of evidence
- difference of probabilities

Sposoby ewaluacji predDiff dla problemu klasyfikacji;

- information difference
- $$infDiff_i(y|x) = \log_2 p(y|x) - \log_2 p(y|x \backslash A_i) \quad (3)$$
- weight of evidence
- $$WE_i(y|x) = \log_2(odds(y|x)) - \log_2(odds(y|x \backslash A_i)) \quad (4)$$
- difference of probabilities

Sposoby ewaluacji predDiff dla problemu klasyfikacji;

- information difference
- $$infDiff_i(y|x) = \log_2 p(y|x) - \log_2 p(y|x \backslash A_i) \quad (3)$$
- weight of evidence
- $$WE_i(y|x) = \log_2(odds(y|x)) - \log_2(odds(y|x \backslash A_i)) \quad (4)$$
- difference of probabilities
- $$probDiff_i(y|x) = p(y|x) - p(y|x \backslash A_i) \quad (5)$$

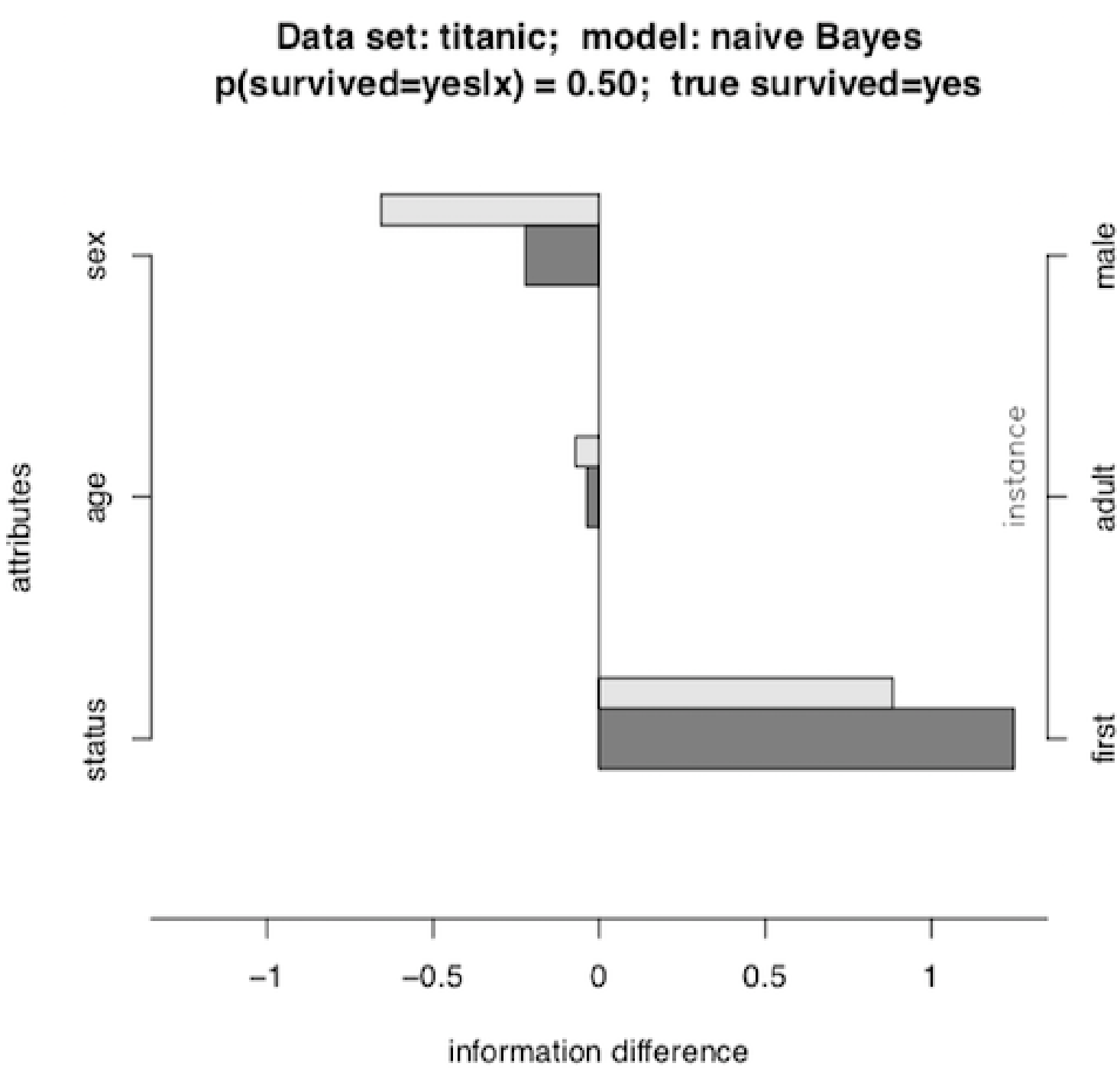
Przykład wizualizacji

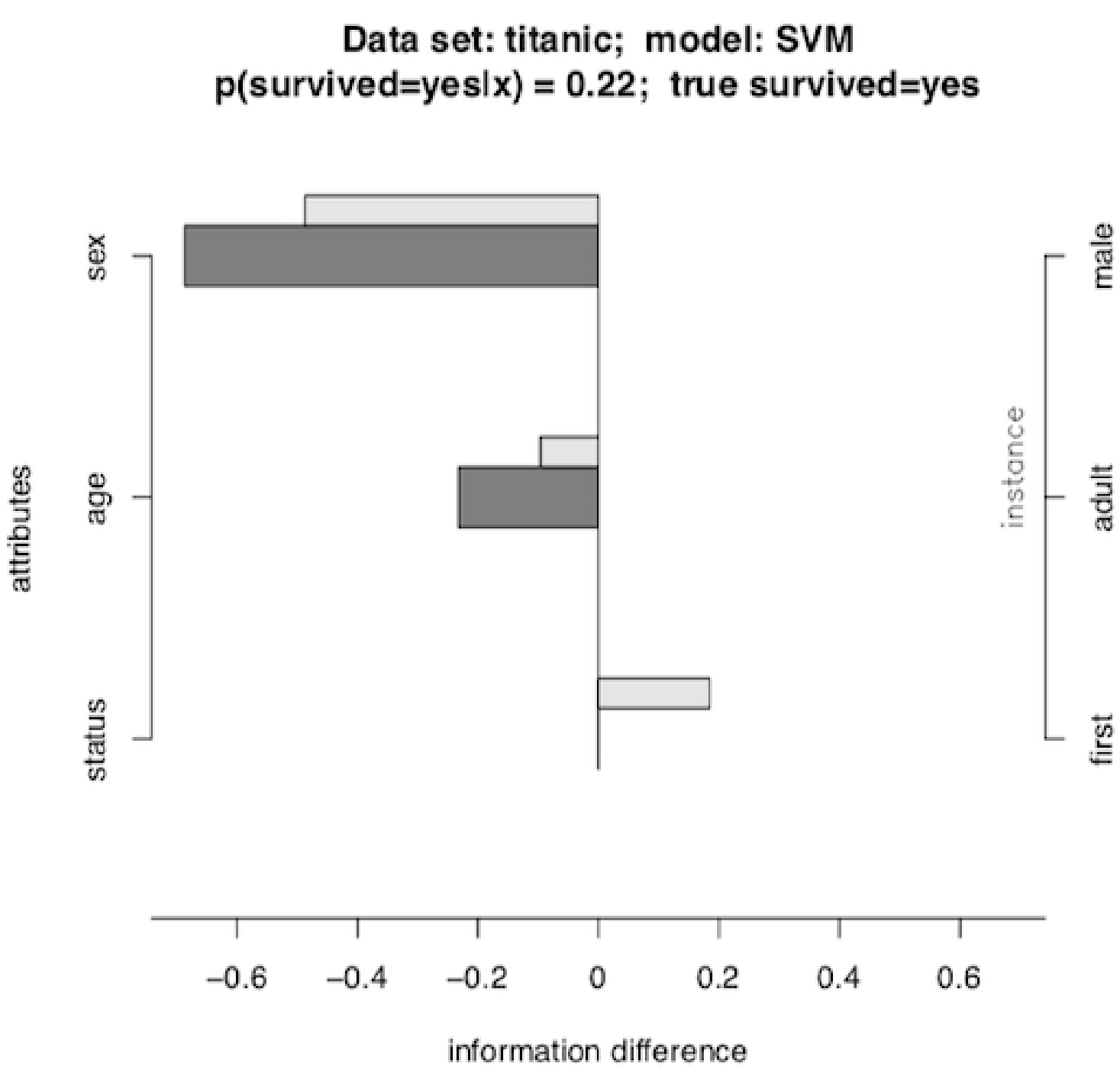
```
package: ExplainPrediction

explainVis(model, trainData, testData,
           method=c("EXPLAIN", "IME"), classValue=1, ...)
```

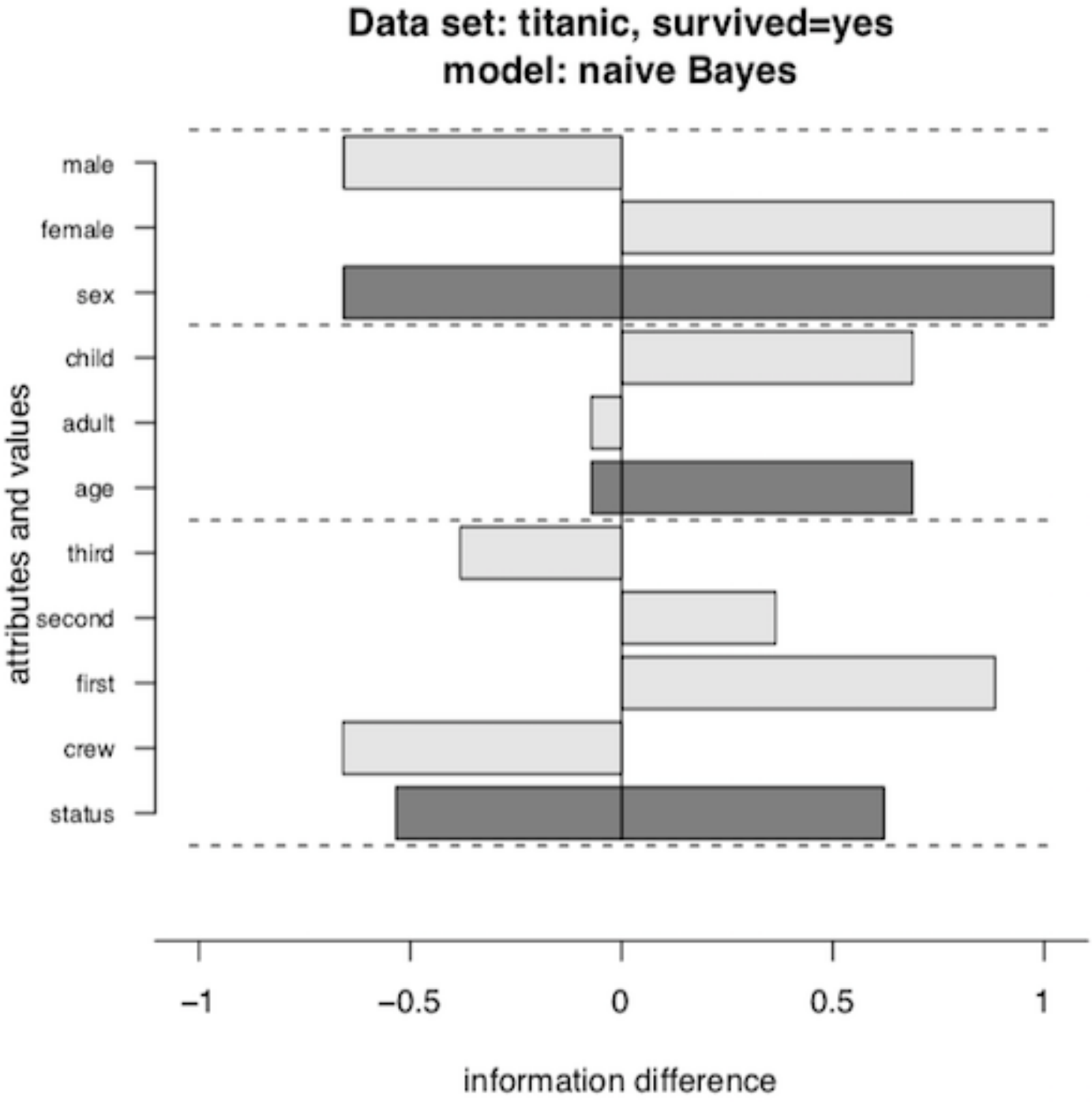
- Służy do obliczenia wyjaśnień i pokazania jej wizualizacji
- agregacja wyjaśnień dla obserwacji daje wyjaśnienia dla modelu

Przykład wizualizacji

Dane:	Titanic
Liczba obserwacji:	2201
Dane uczące:	50%
Modele:	NB ; SVM
Wyjaśnienie:	InfDiff


Data set: titanic; model: naive Bayes
p(survived=yes|x) = 0.50; true survived=yes
attributes
sex
age
status
instance
male
adult
first
−1
−0.5
0
0.5
1
information difference


Data set: titanic; model: SVM
p(survived=yes|x) = 0.22; true survived=yes
attributes
sex
age
status
instance
male
adult
first
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
information difference


Data set: titanic, survived=yes
model: naive Bayes
male
female
sex
child
adult
age
third
second
first
crew
status
attributes and values
−1
−0.5
0
0.5
1
information difference

Jakość modelu a wyjaśnienia

Im lepiej model odzwierciedla problem,
tym bliższe są wyjaśnienia modelu - wyjaśnieniom prawdziwym.

Table 3: Performance and average distances to the true explanation for five classification methods on five data sets.

method		condInd	xor	group	cross	chess
NB	acc	0.90	0.51	0.35	0.50	0.50
	AUC	0.96	0.51	0.50	0.50	0.50
	$\overline{d_{exp}}$	**0.06**	0.39	0.46	0.45	0.47
DT	acc	0.89	0.90	0.33	0.52	0.52
	AUC	0.95	0.90	0.50	0.56	0.50
	$\overline{d_{exp}}$	0.17	**0.01**	0.35	0.33	0.35
kNN	acc	0.86	0.90	0.99	0.55	0.71
	AUC	0.93	0.90	0.83	0.59	0.78
	$\overline{d_{exp}}$	0.16	0.10	**0.08**	0.40	0.33
SVM	acc	0.89	0.58	0.66	0.98	0.53
	AUC	0.95	0.52	0.76	0.99	0.52
	$\overline{d_{exp}}$	0.12	0.39	0.22	**0.04**	0.42
ANN	acc	0.89	0.90	0.98	0.95	0.84
	AUC	0.92	0.90	0.82	0.98	0.90
	$\overline{d_{exp}}$	0.27	0.09	0.09	0.08	**0.16**

Dane: 'Groups'

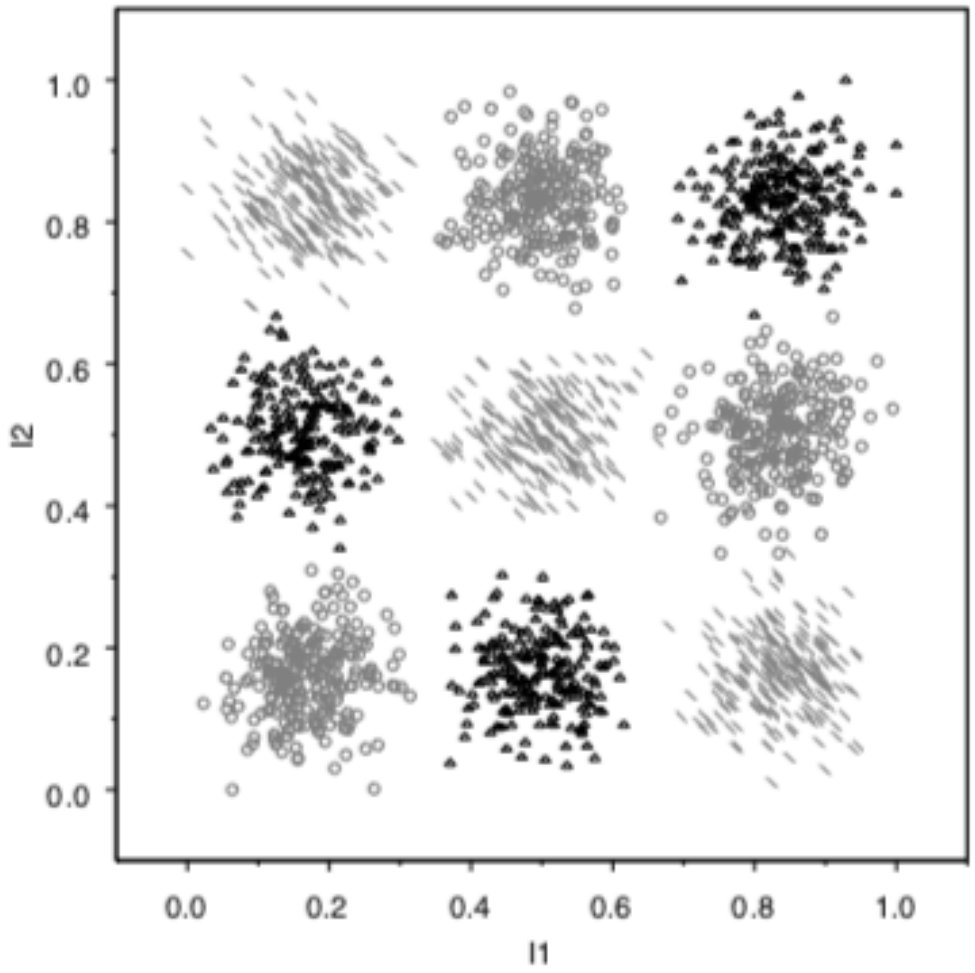
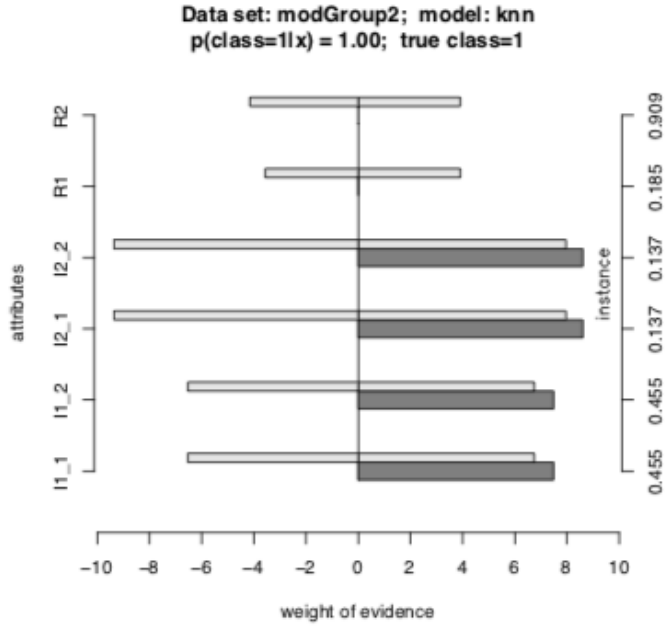
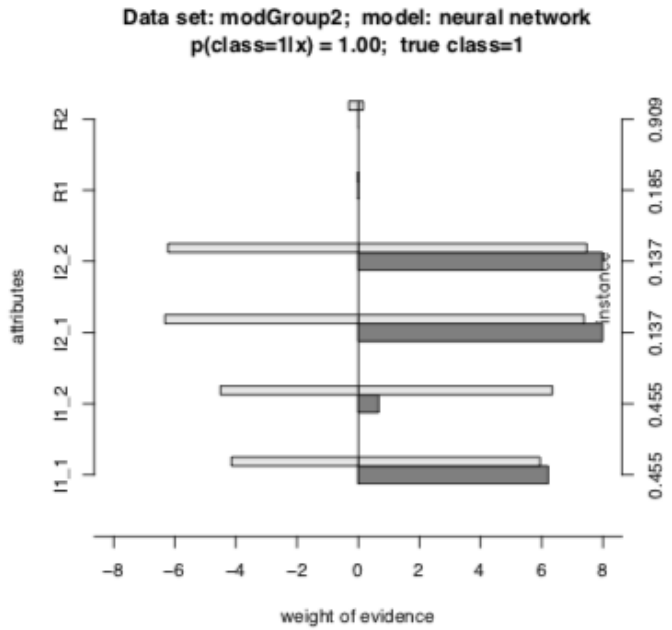


Figure 4: Visualization of two important attributes in the *groups* data set. Circles, triangles, and lines represent class values 0, 1, and 2.

Zbędne zmienne





$$p(y|x\backslash A_i) = \sum_{s=1}^{m_i} p(A_i = a_s|x\backslash A_i)p(y|x \leftarrow A_i = a_s) \tag{6}$$

$$p(y|x\backslash A_i) \doteq \sum_{s=1}^{m_i} p(A_i = a_s)p(y|x \leftarrow A_i = a_s) \tag{7}$$