NAIVNI BAYESOV KLASIFIKATOR

Fakulteta za matematiko in fiziko
Matematika z računalnikom

Avtor: **Zala Jamšek**
Mentor: **Vid Podpečan**,
Andrej Bauer

Ljubljana, 17.1.2013

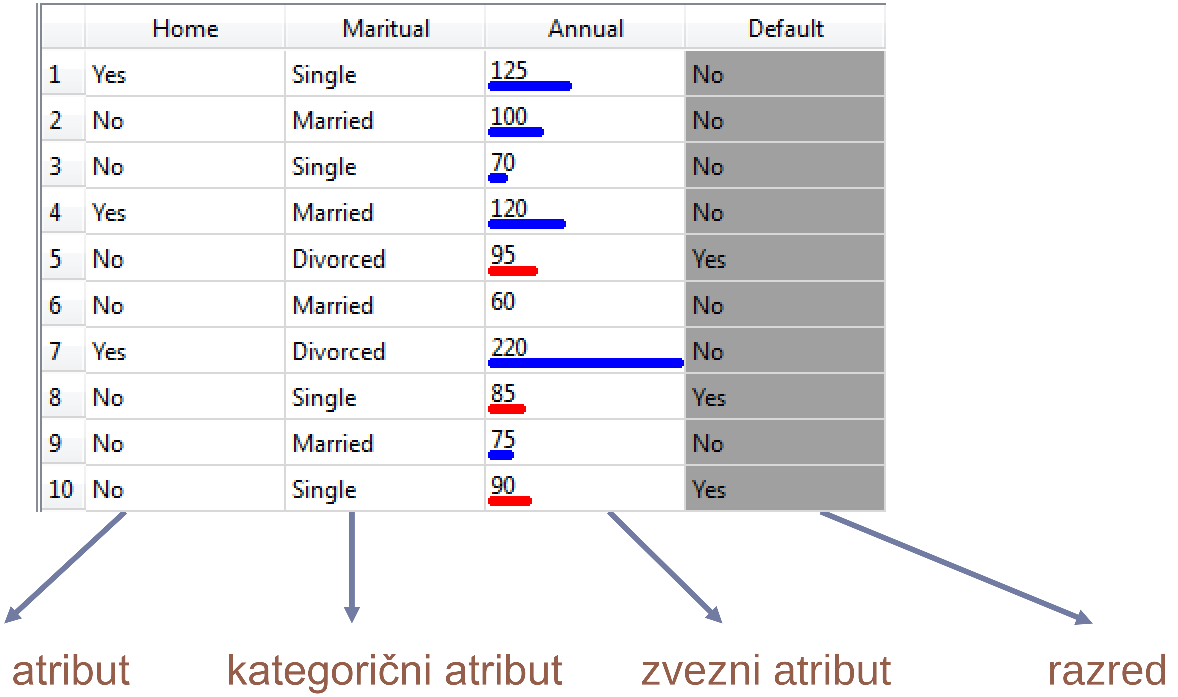
Naivni Bayesov klasifikator

- algoritem strojnega učenja

- algoritem:
  - Bayesovo pravilo
  - predpostavko o pogojni neodvisnosti atributov pri danem razredu

- **CILJ**: s pomočjo učnega algoritma določiti klasifikator za učno množico podatkov

Primer:



	Home	Maritual	Annual	Default
1	Yes	Single	125	No
2	No	Married	100	No
3	No	Single	70	No
4	Yes	Married	120	No
5	No	Divorced	95	Yes
6	No	Married	60	No
7	Yes	Divorced	220	No
8	No	Single	85	Yes
9	No	Married	75	No
10	No	Single	90	Yes

binarni atribut

kategorični atribut

zvezni atribut

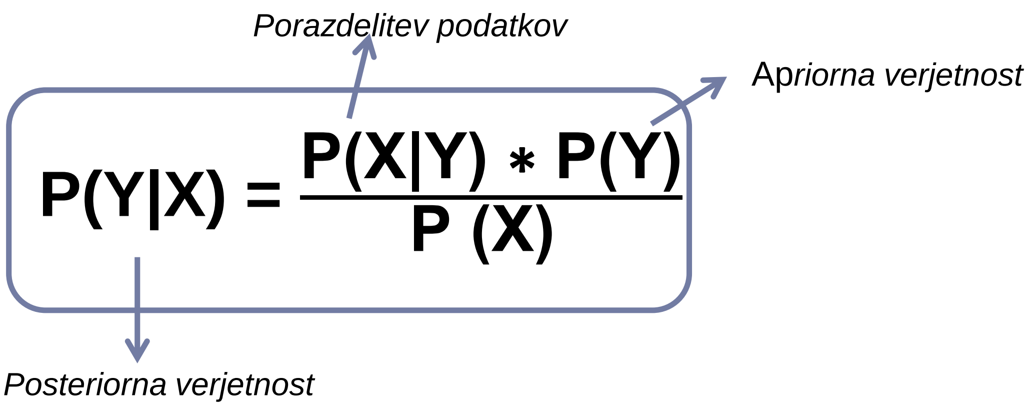
razred

- $X = (X_1, X_2,...,X_d)$
- d je število atributov
- x_i je vrednost atributa X_i
- Y je razred

Bayesovo pravilo

$$P(X|Y)= \frac{P(X,Y)}{P(Y)} \qquad P(Y|X) = \frac{P(X,Y)}{P(X)}$$

$$P(X,Y) = P(X|Y) * P(Y) = P(Y|X) * P(X)$$

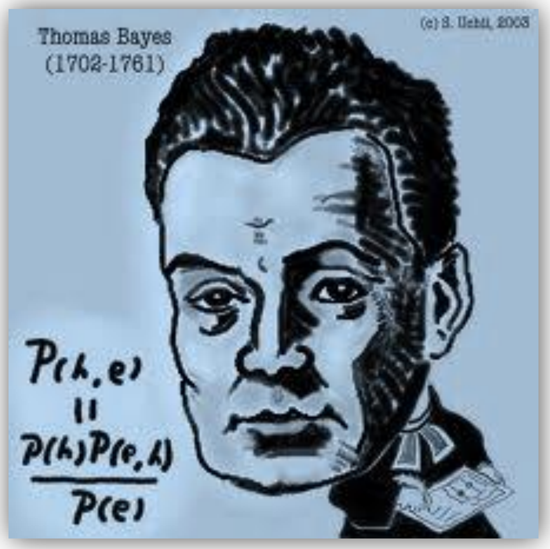


Porazdelitev podatkov

Apriorna verjetnost

$$\boxed{P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}}$$

Posteriorna verjetnost



Naivni Bayesov klasifikator

- predpostavka pogojne neodvisnosti atributov pri danem razredu:

$$\mathbf{P(X|\ Y{=}y) = \prod_{i=1}^{d} P(X_i|Y = y)}$$

- Posteriorna verjetnost:

$$\mathbf{P(Y|X)} = \frac{\prod_{i=1}^{d} P(X_i|Y{=}y) * P(Y)}{P\ (X)}$$

ker P(X) konstantna za vsak Y, jo lahko izpustimo

Naivni Bayesov klasifikator

Posteriorna verjetnost:

$$\mathbf{P(Y|X)} = \prod_{i=1}^{d} \boldsymbol{P(X_i|Y = y)} * \mathbf{P(Y)}$$

Naivni Bayesov klasifikator:

$$\text{classify}(x_1, x_2,\ldots,x_d) = \underset{y \in Y}{\text{argmax}}\ \mathbf{P(Y} = \boldsymbol{y}\mathbf{)} \prod_{i=1}^{d} \boldsymbol{P(X_i|Y = y)}$$

Primer

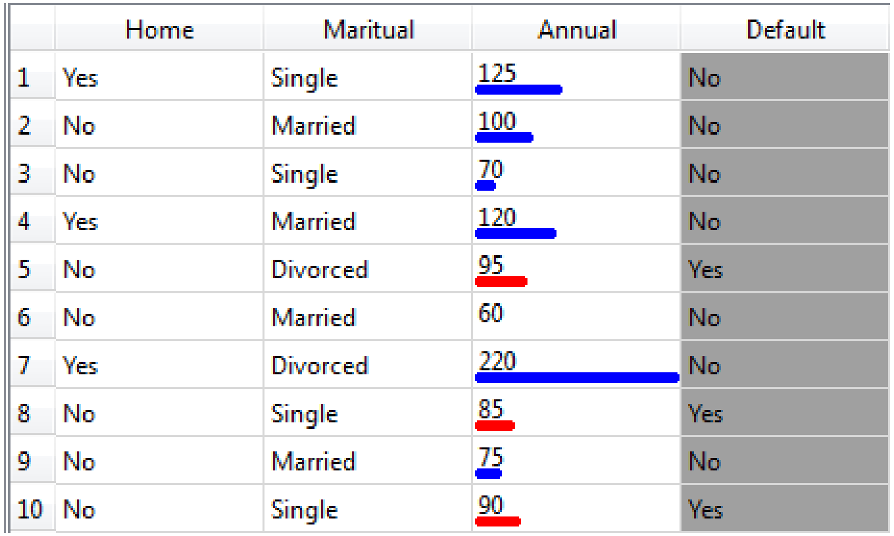
Ali bo stranka, ki ima lastnostmi:

Home = no

Maritual = single

Annual = 120

sklenila vezani depozit?



	Home	Maritual	Annual	Default
1	Yes	Single	125	No
2	No	Married	100	No
3	No	Single	70	No
4	Yes	Married	120	No
5	No	Divorced	95	Yes
6	No	Married	60	No
7	Yes	Divorced	220	No
8	No	Single	85	Yes
9	No	Married	75	No
10	No	Single	90	Yes

Apriorna verjetnost:

P(Y=yes) = 0,3

P(Y=no) = 0,7

Kategorični in binarni atributi:

P(Home=no | Y=yes) = $\frac{3}{3}$

P(Home=no| Y=no) = $\frac{4}{7}$

P(Maritual =single | Y=yes) = $\frac{2}{3}$

P(Maritual =single | Y=no) = $\frac{2}{7}$

Zvezni atributi:

1.)Diskretizacija:

2.)Privzamemo porazdelitev (npr. normlano):

$$P(X_i=x_i \mid Y=y_j) \approx \frac{1}{\sqrt{2\pi}\sigma_{ij}}\exp(-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2})$$

μ_{ij} - povprečje vzorca: $\bar{X} = \sum_{i=1}^{n}\frac{1}{n}X_i$

σ_{ij}^2 - varianco vzorca: $S^2 = \sum_{i=1}^{n}\frac{1}{n-1}(X_i - \bar{X})^2$

$$\bar{X} = \frac{125+100+ \ldots +75}{7} = 110$$

$S^2 = \frac{(125-110)^2+ \ldots +(75-110)^2}{6} = 2975$ S = 54,54

P (Balance = 120 | Y=no) = $\frac{1}{\sqrt{2\pi}\,(54.54)}\exp(-\frac{(120-110)^2}{2*2975})$ = 0,0072

Primer:

Verjetnost vzorca:

P (X | Y=yes) = $1 \ast \frac{2}{3} \ast 0{,}0072 = 0{,}0048$

P (X | Y=no) = $\frac{4}{7} \ast \frac{2}{7} \ast 10^{-9} = 0{,}16 \ast 10^{-9}$

Posteriorna verjetnost:

P(Y=no | X) = $0{,}7 \ast 0{,}16 \ast 10^{-9}$ = **$0{,}112 * 10^{-9}$**

P(Y=yes | X) = $0{,}3 \ast 0{,}0048$ = **$0{,}114 * 10^{-3}$**

P(yes | X) > P(no | X)

Napovedani razred je **yes.**

P(X_i=x_i | Y=y_j) = $\frac{n_c}{n}$

n_c $- št.primerov, ko$ X=x_i pri znani vrednosti Y

n – št. pojavitev y_i

če $n_c = 0$

P(Y|X) = 0

Laplacova metoda

P(X_i=x_i | Y=y_j) = $\frac{n_c+1}{n+k}$

k – št. razredov za razredov za atribut X_i

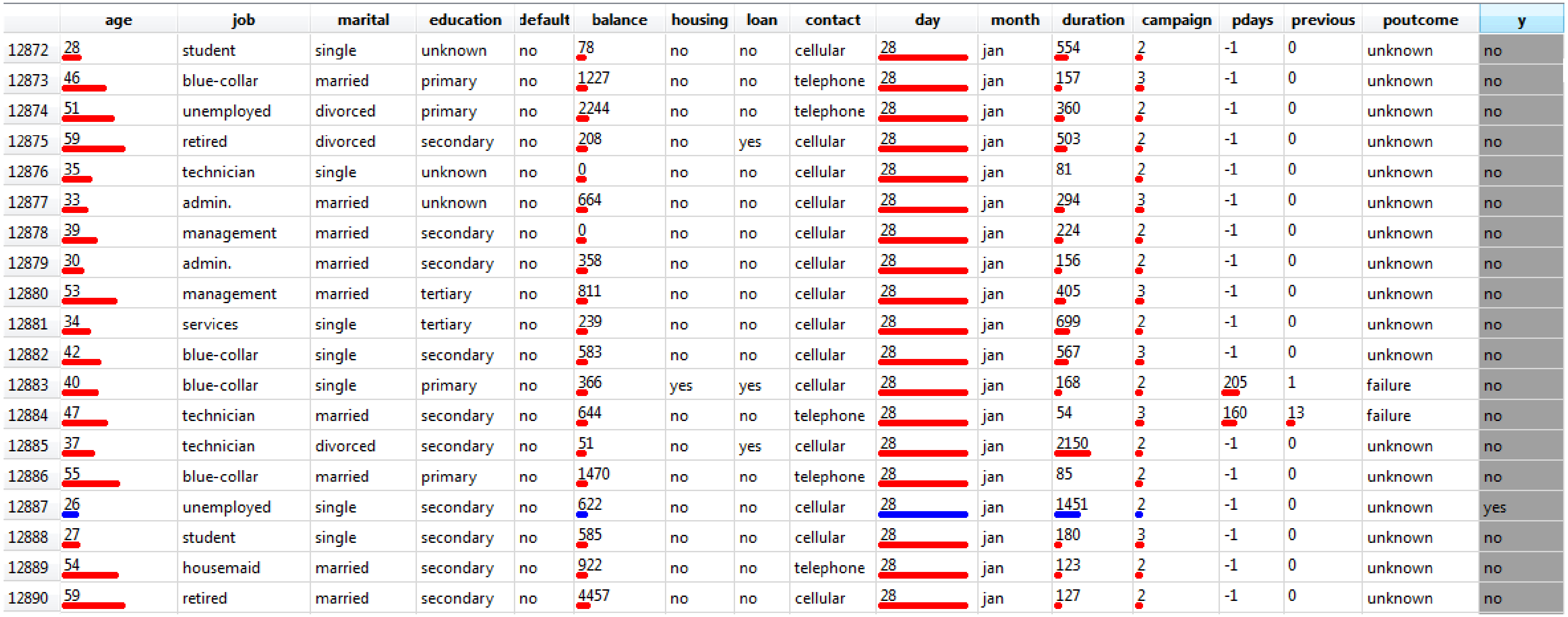
M – metoda

P(X_i=x_i | Y=y_j) = $\frac{n_c+mp}{n+m}$

m – parameter, ki pove, koliko zaupamo našim podatkom

p – apriorna verjetnost

Podatki



	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
12872	28	student	single	unknown	no	78	no	no	cellular	28	jan	554	2	-1	0	unknown	no
12873	46	blue-collar	married	primary	no	1227	no	no	telephone	28	jan	157	3	-1	0	unknown	no
12874	51	unemployed	divorced	primary	no	2244	no	no	telephone	28	jan	360	2	-1	0	unknown	no
12875	59	retired	divorced	secondary	no	208	no	yes	cellular	28	jan	503	2	-1	0	unknown	no
12876	35	technician	single	unknown	no	0	no	no	cellular	28	jan	81	2	-1	0	unknown	no
12877	33	admin.	married	unknown	no	664	no	no	cellular	28	jan	294	3	-1	0	unknown	no
12878	39	management	married	secondary	no	0	no	no	cellular	28	jan	224	2	-1	0	unknown	no
12879	30	admin.	married	secondary	no	358	no	no	cellular	28	jan	156	2	-1	0	unknown	no
12880	53	management	married	tertiary	no	811	no	no	cellular	28	jan	405	3	-1	0	unknown	no
12881	34	services	single	tertiary	no	239	no	no	cellular	28	jan	699	2	-1	0	unknown	no
12882	42	blue-collar	single	secondary	no	583	no	no	cellular	28	jan	567	3	-1	0	unknown	no
12883	40	blue-collar	single	primary	no	366	yes	yes	cellular	28	jan	168	2	205	1	failure	no
12884	47	technician	married	secondary	no	644	no	no	telephone	28	jan	54	3	160	13	failure	no
12885	37	technician	divorced	secondary	no	51	no	yes	cellular	28	jan	2150	2	-1	0	unknown	no
12886	55	blue-collar	married	primary	no	1470	no	no	telephone	28	jan	85	2	-1	0	unknown	no
12887	26	unemployed	single	secondary	no	622	no	no	cellular	28	jan	1451	2	-1	0	unknown	yes
12888	27	student	single	secondary	no	585	no	no	cellular	28	jan	180	3	-1	0	unknown	no
12889	54	housemaid	married	secondary	no	922	no	no	telephone	28	jan	123	2	-1	0	unknown	no
12890	59	retired	married	secondary	no	4457	no	no	cellular	28	jan	127	2	-1	0	unknown	no

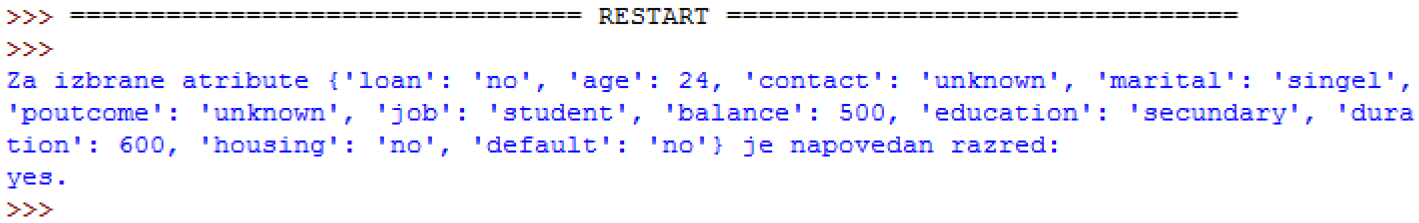
- *2012, Portugalska telefonska marketinška kampanija bančnih institucij*
- *45211 oseb*

***Razred y** – ali bo stranka sklenila vezani depozit*

Koda v Pythonu

- Osnovna metoda in Laplaceova matoda
- Binarni razred
- 10 % podatkov
- *Ali bo oseba najela posojilo, če ima naslednje lastnosti*:

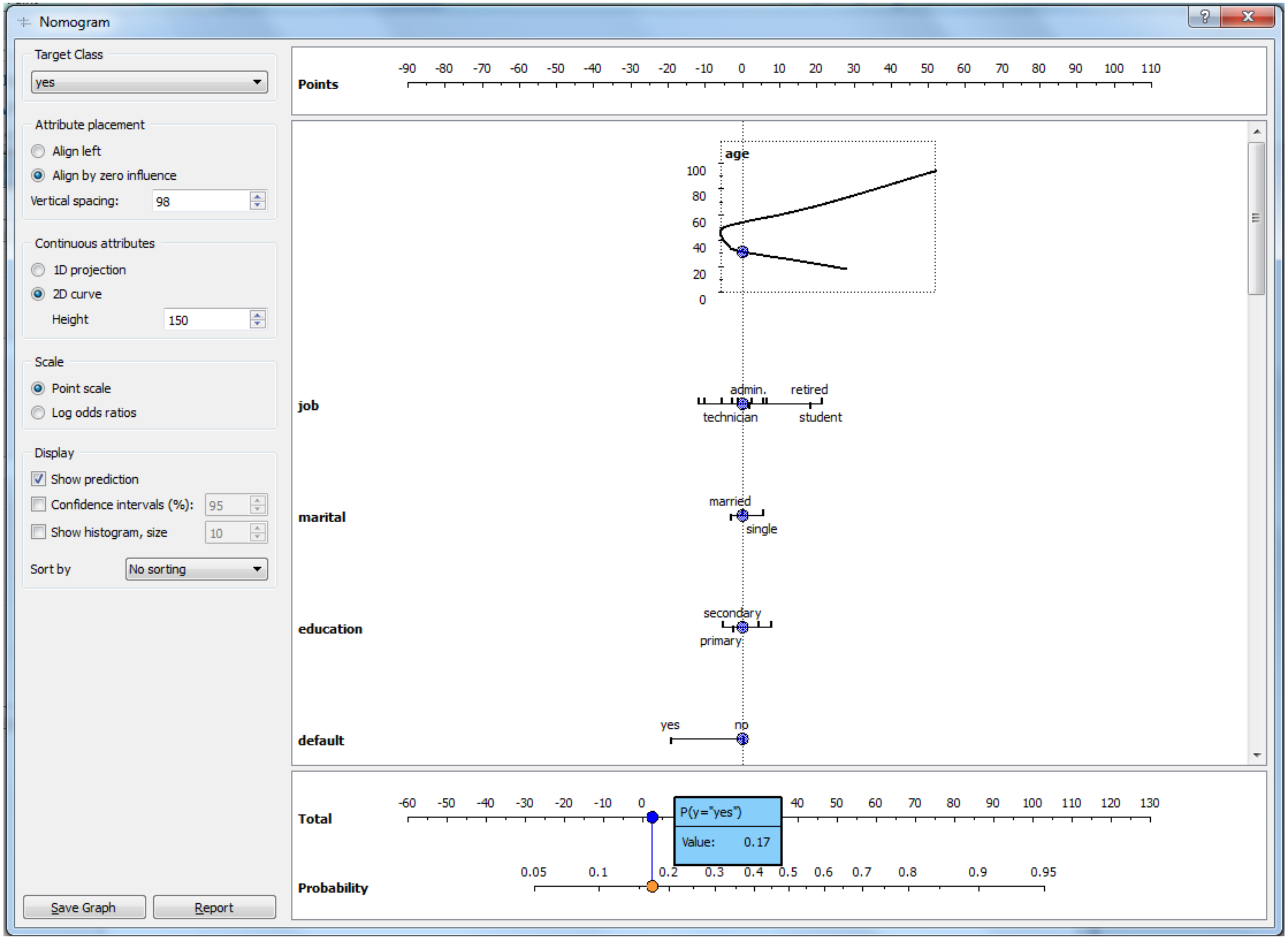
X=(age:24,job:student,marital:singel,
education:secundary,default:no,
balance:500,housing:no,loan:no,
contact:unknown,duration:600,
poutcome:unknown)



```
>>> ================================ RESTART ================================
>>>
Za izbrane atribute {'loan': 'no', 'age': 24, 'contact': 'unknown', 'marital': 'singel',
'poutcome': 'unknown', 'job': 'student', 'balance': 500, 'education': 'secundary', 'dura
tion': 600, 'housing': 'no', 'default': 'no'} je napovedan razred:
yes.
>>>
```

Analiza podatkov 1

Laplaceova metoda



Nomogram

Target Class

yes

Attribute placement

○ Align left

● Align by zero influence

Vertical spacing: 98

Continuous attributes

○ 1D projection

● 2D curve

Height 150

Scale

● Point scale

○ Log odds ratios

Display

☑ Show prediction

□ Confidence intervals (%): 95

□ Show histogram, size 10

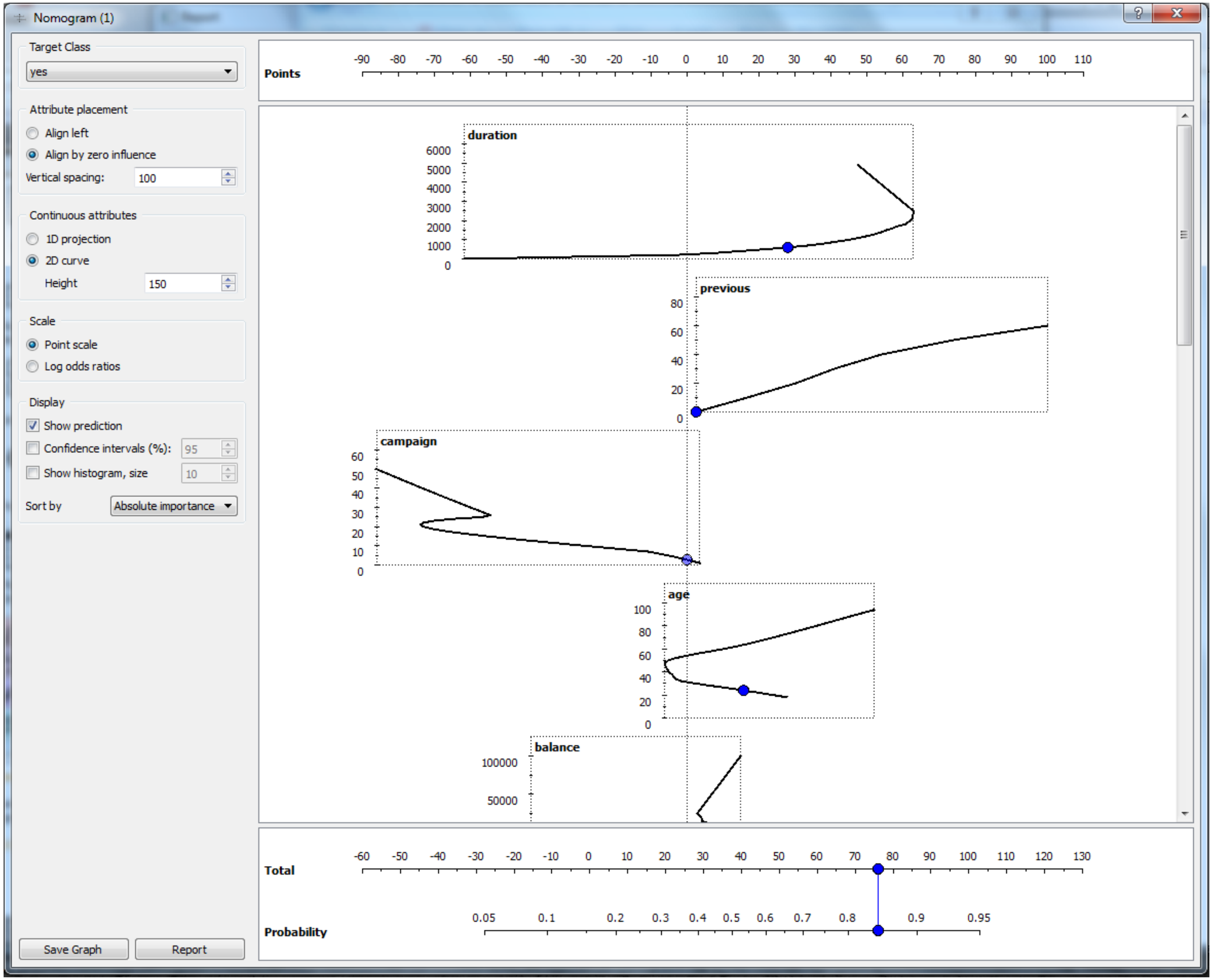
Sort by No sorting

P(y="yes")

Value: 0.17

Analiza podatkov 2

m-metoda (m= 2)
absloutna pomembnost atributov



Nomogram (1)

Target Class
yes

Attribute placement
○ Align left
● Align by zero influence
Vertical spacing: 100

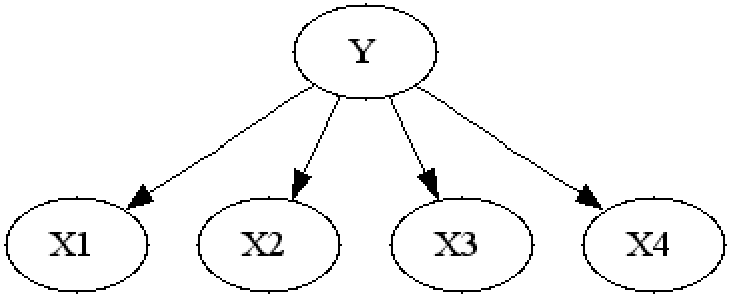
Continuous attributes
○ 1D projection
● 2D curve
Height 150

Scale
● Point scale
○ Log odds ratios

Display
☑ Show prediction
☐ Confidence intervals (%): 95
☐ Show histogram, size 10
Sort by Absolute importance

OPTIMALNOST BAYESOVEGA KLASIFIKATORJA

- pogojna neodvisnost pri danem razredu v realnosti redka



- ne kaznuje napačno izračunanih verjetnosti

Prave verjetnosti:	*P(yes|X) = 0,4*	*P(no|X)=0,6*
Izračunane verjetnosti:	*P(yes|X) = 0,1*	*P(no|X)=0,9*

- slabo oceni verjetnosti, dobro klasificira