Odlični v znanosti 2013

SLOVENSKI 9. FORUM INOVACIJ

INOVATIVNI POTENCIAL SLOVENIJE

Metoda *NoiseRank* za odkrivanje anomalij v podatkih

Borut Sluban, ***Institut „Jožef Stefan“***

SPIRIT SLOVENIJA JAVNA AGENCIJA | I FEEL SLOVENIA

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA GOSPODARSKI RAZVOJ IN TEHNOLOGIJO

JAVNA AGENCIJA ZA RAZISKOVALNO DEJAVNOST REPUBLIKE SLOVENIJE

Anomalije?

Anomalije?

Anomalije so …

- **Napake v podatkih - šum**

  živali bele barve:

Anomalije so …

- **Napake v podatkih - šum**

  živali bele barve:

- **Izjeme ali osamelci v podatkih**

  čreda ovc:

Namen dela

- **Razvoj metodologije za odkrivanje anomalij v podatkih**

Namen dela

- **Razvoj metodologije za odkrivanje anomalij v podatkih**
- **za potrebe:**
  - **čiščenje podatkov**
  - **razumevanje podatkov / domene**
  - **odkrivanje novih posebnih primerov v podatkih**

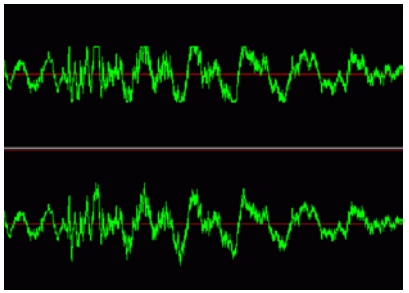
Podatki?

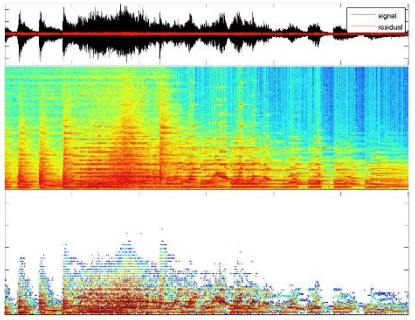
Podatki?

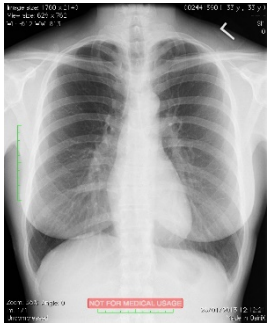
Podatki?


ur level of satisfaction
4
5
na
where 1 represents
te your level of sa
5
na
4
to 5 where 1 represent
ou rate your level of sati
na

Podatki?







Podatki

Ime	Starost	Poklic	Prosti čas	Najljubša jed
Janko	9	šolar/ka	Sprehodi po gozdu	sladkarije
Metka	7	šolar/ka	Sprehodi po gozdu	sladkarije
Sneguljčica	17	manekenka	Sprehodi po gozdu	rdeče jabolko
Volk	467	gozdar	Sprehodi po gozdu	Rdeča kapica
Špicparkelj	?	škrat	Sprehodi po gozdu	borovnice

Odkrivanje anomalij

Odkrivanje anomalij

- **Stvari v naravi:**
  - **so urejene po določenih vzorcih**

Odkrivanje anomalij

- **Stvari v naravi:**
  - **so urejene po določenih vzorcih**

  - **sledijo fizikalnim zakonom**

Odkrivanje anomalij

- **Stvari v naravi:**
  - **so urejene po določenih vzorcih**

  - **sledijo fizikalnim zakonom**

  - **niso naključne**

Odkrivanje anomalij

- **Napake ali izjeme zaznamo kot:**

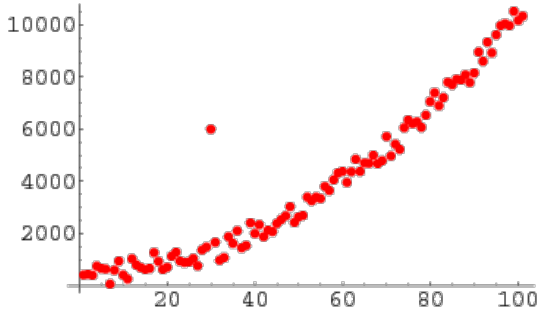
Odkrivanje anomalij

- **Napake ali izjeme zaznamo kot:**
  - **neskladja z ustaljenimi vzorci**

Odkrivanje anomalij

- **Napake ali izjeme zaznamo kot:**
  - **neskladja z ustaljenimi vzorci**

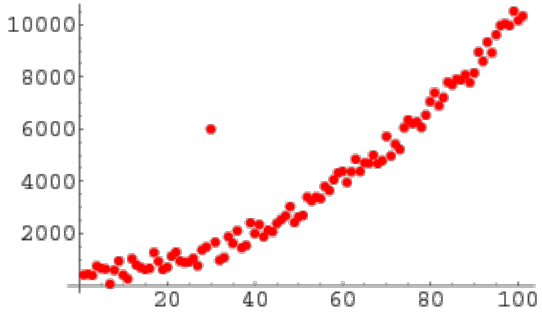
  - **velika odstopanja od pričakovanih vrednosti**



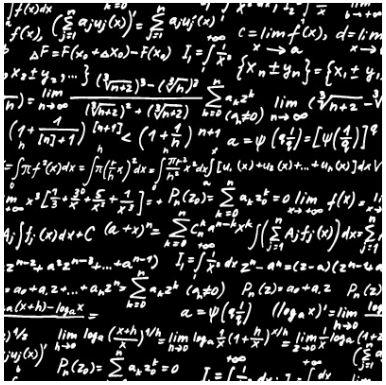
Odkrivanje anomalij

- **Napake ali izjeme zaznamo kot:**
  - **neskladja z ustaljenimi vzorci**

  - **velika odstopanja od pričakovanih vrednosti**

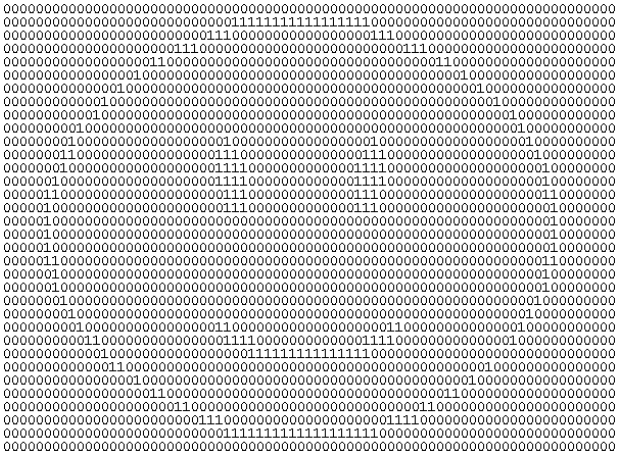


  - **težko opisljive**

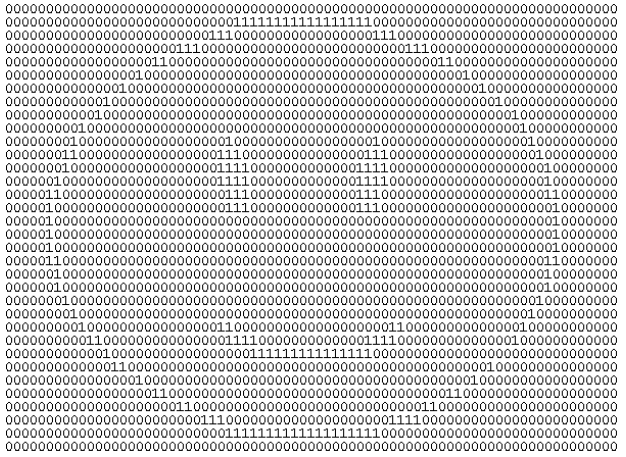


Odkrivanje anomalij

- v podatkih želimo prepoznati „zakone“

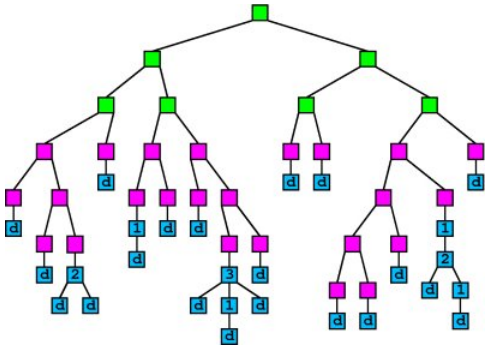


Odkrivanje anomalij

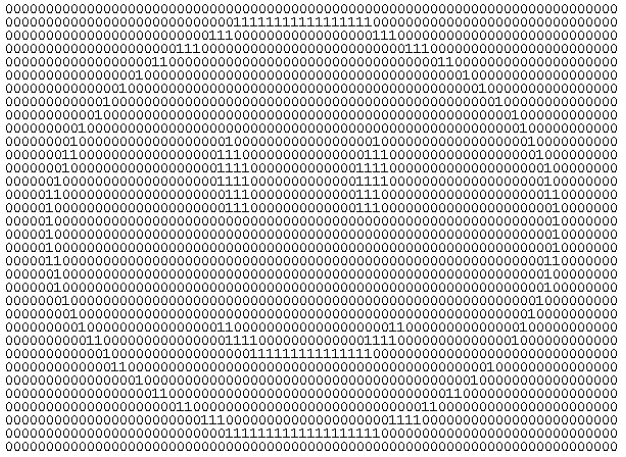


- v podatkih želimo prepoznati „zakone“

- gradimo modele
  – vzorci in pravila = „zakoni“ podatkov

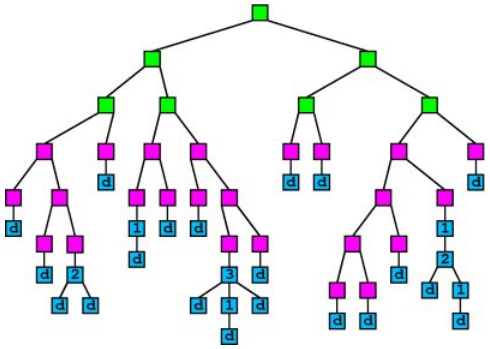


Odkrivanje anomalij

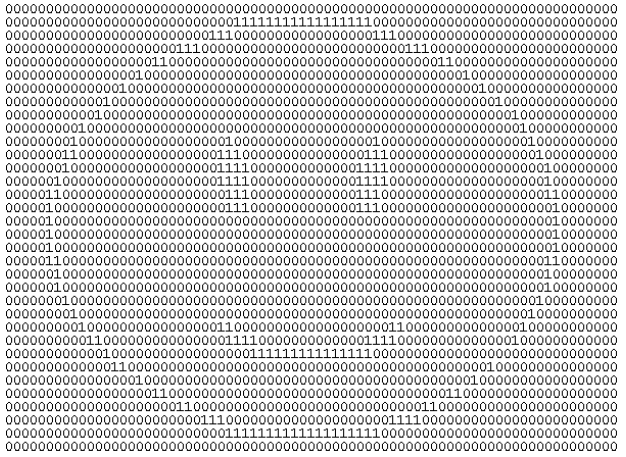


- v podatkih želimo prepoznati „zakone“

- gradimo modele
  – vzorci in pravila = „zakoni“ podatkov
- napake ali izjeme
  – ne sledijo zakonom (modelom)

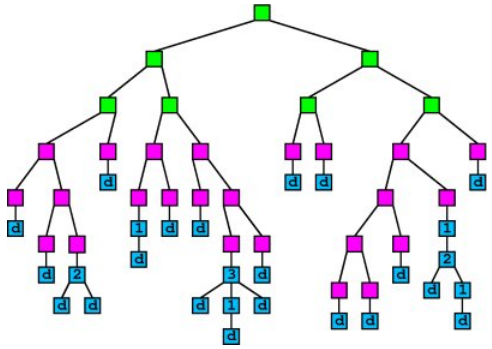


Odkrivanje anomalij

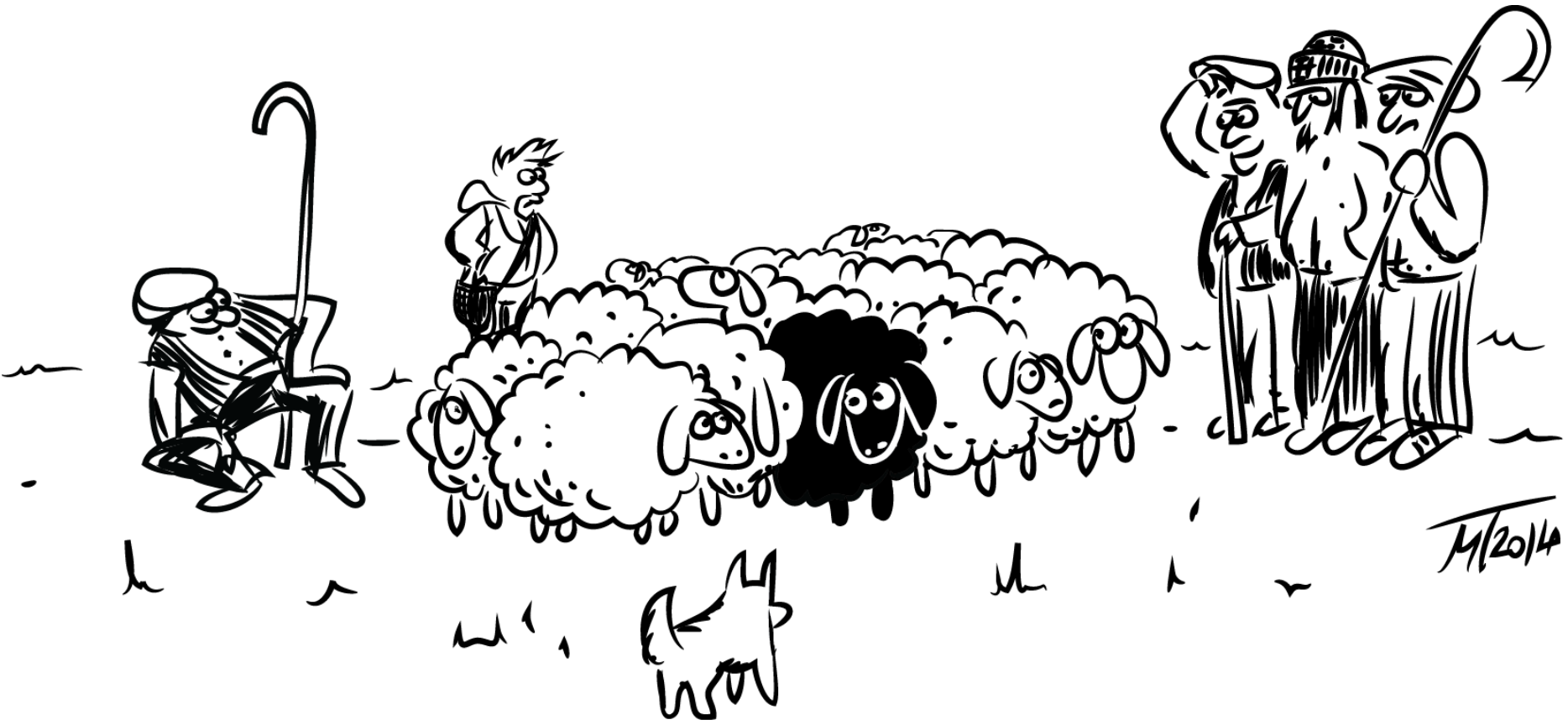


- v podatkih želimo prepoznati „zakone“

- gradimo modele
  – vzorci in pravila = „zakoni“ podatkov
- napake ali izjeme
  – ne sledijo zakonom (modelom)
- „Več glav več ve“
  – uporaba več modelov



Odkrivanje anomalij



NoiseRank

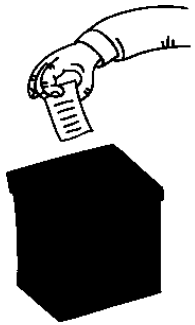
- **Ansambelska metoda za odkrivanje šuma in osamelcev v podatkih**

NoiseRank

- **Ansambelska metoda za odkrivanje šuma in osamelcev v podatkih**
  - **uporaba poljubnega števila modelov**

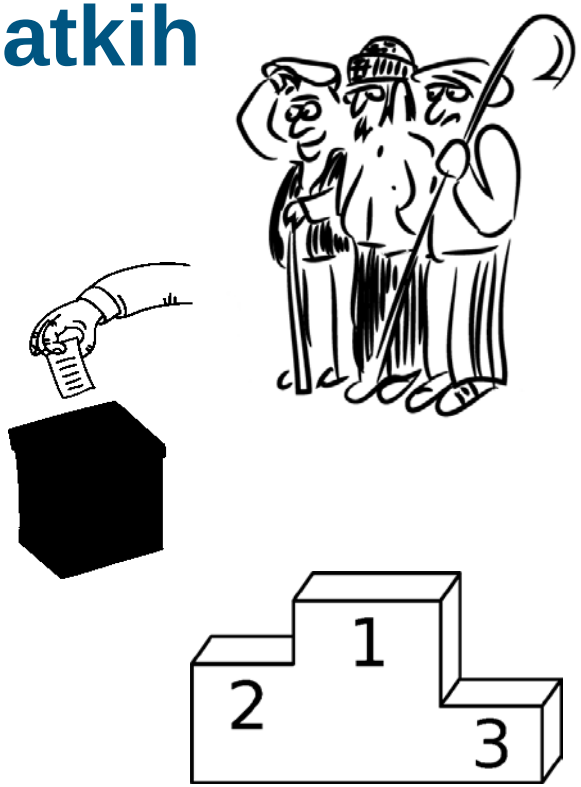
NoiseRank

- **Ansambelska metoda za odkrivanje šuma in osamelcev v podatkih**
  - **uporaba poljubnega števila modelov**
  - **združevanje njihovih rezultatov**



NoiseRank

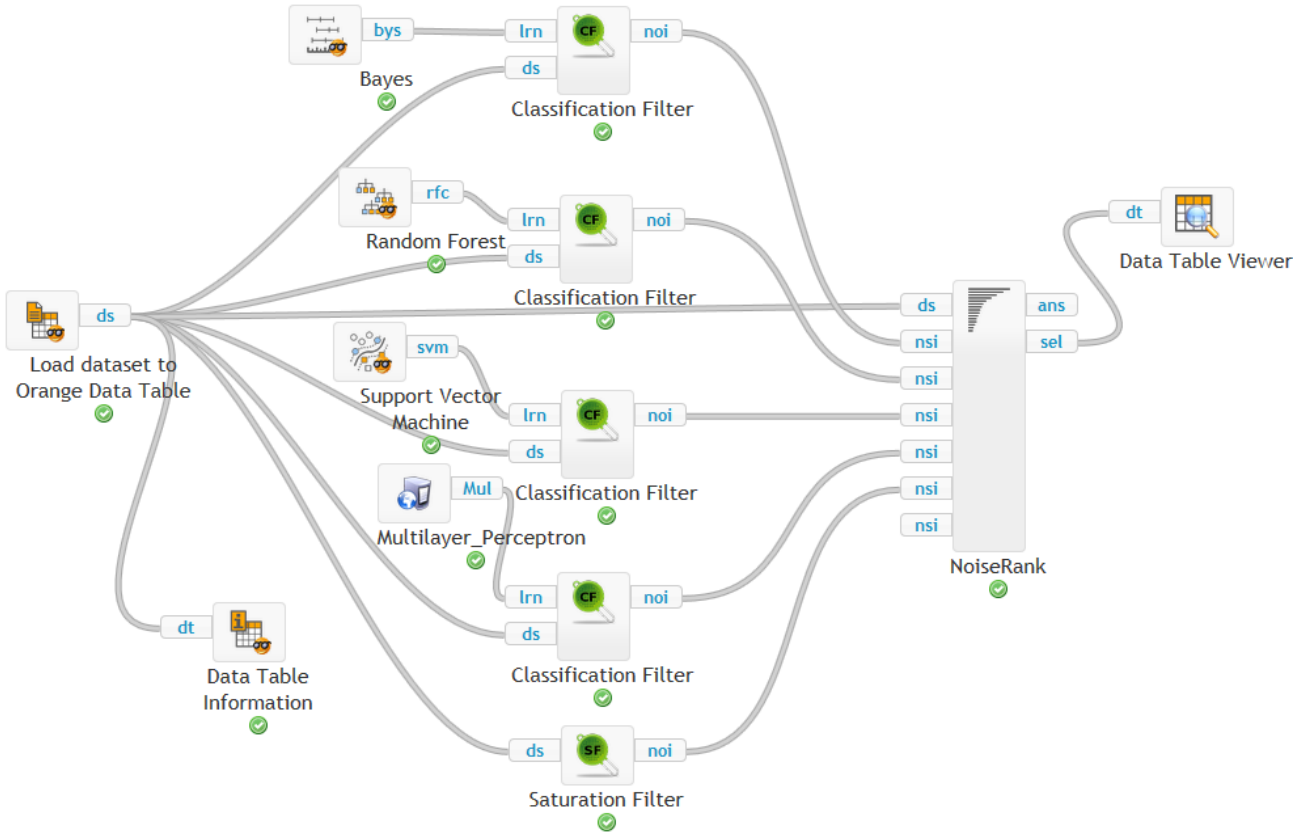
- **Ansambelska metoda za odkrivanje šuma in osamelcev v podatkih**
  - **uporaba poljubnega števila modelov**
  - **združevanje njihovih rezultatov**
  - **rangiranje primerov po „šumnosti“**



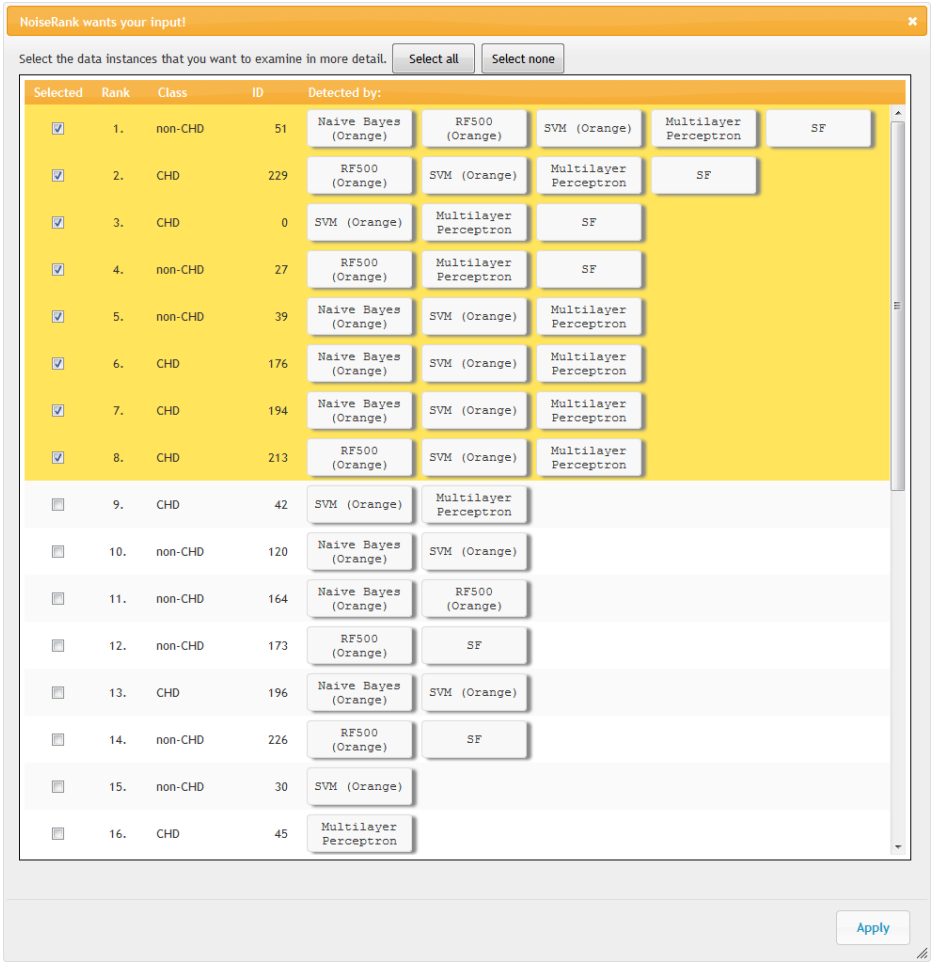
Uporaba

http://www.clowdflows.org

Uporaba



Uporaba


NoiseRank wants your input!
Select the data instances that you want to examine in more detail.
Select all
Select none
Selected
Rank
Class
ID
Detected by:
1. non-CHD 51 Naive Bayes (Orange) RF500 (Orange) SVM (Orange) Multilayer Perceptron SF
2. CHD 229 RF500 (Orange) SVM (Orange) Multilayer Perceptron SF
3. CHD 0 SVM (Orange) Multilayer Perceptron SF
4. non-CHD 27 RF500 (Orange) Multilayer Perceptron SF
5. non-CHD 39 Naive Bayes (Orange) SVM (Orange) Multilayer Perceptron
6. CHD 176 Naive Bayes (Orange) SVM (Orange) Multilayer Perceptron
7. CHD 194 Naive Bayes (Orange) SVM (Orange) Multilayer Perceptron
8. CHD 213 RF500 (Orange) SVM (Orange) Multilayer Perceptron
9. CHD 42 SVM (Orange) Multilayer Perceptron
10. non-CHD 120 Naive Bayes (Orange) SVM (Orange)
11. non-CHD 164 Naive Bayes (Orange) RF500 (Orange)
12. non-CHD 173 RF500 (Orange) SF
13. CHD 196 Naive Bayes (Orange) SVM (Orange)
14. non-CHD 226 RF500 (Orange) SF
15. non-CHD 30 SVM (Orange)
16. CHD 45 Multilayer Perceptron
Apply

Rezultati

Rezultati

- **Medicinska domena***

* Institute of Cardiovascular Prevention and Rehabilitation, Zagreb, Hrvaška

Rezultati

- **Medicinska domena***
- **Bolezni srca (CHD)**
  - **1. Napačna diagnoza**
  - **2. Zapleten primer**
  - **3. Športnik**

Rank	Class	ID	Detected by:
1.	non-CHD	51	__Bayes_____RF_____SVM_____NN_____SF___
2.	CHD	229	____RF_____SVM_____NN_____SF___
3.	CHD	0	___SVM_____NN_____SF___
3.	non-CHD	27	____RF_____NN_____SF___
3.	non-CHD	39	__Bayes_____SVM_____NN___
3.	CHD	176	__Bayes_____SVM_____NN___
3.	CHD	194	__Bayes_____SVM_____NN___
3.	CHD	213	____RF_____SVM_____NN___
4.	CHD	42	___SVM_____NN___
4.	non-CHD	120	__Bayes_____SVM___
4.	non-CHD	164	__Bayes_____RF___
4.	non-CHD	173	____RF_____SF___
4.	CHD	196	__Bayes_____SVM___
4.	non-CHD	226	____RF_____SF___
5.	non-CHD	30	___SVM___
5.	CHD	45	____NN___
5.	non-CHD	59	____RF___
5.	non-CHD	62	__Bayes__
5.	non-CHD	63	____SF___
5.	CHD	67	___SVM___
5.	non-CHD	72	___SVM___
5.	CHD	97	____SF___
5.	non-CHD	135	___SVM___
5.	non-CHD	177	____NN___
5.	CHD	181	__Bayes__
5.	non-CHD	189	___SVM___
5.	CHD	195	___SVM___
5.	CHD	200	____NN___
5.	CHD	205	___SVM___
5.	CHD	231	__Bayes__

* Institute of Cardiovascular Prevention and Rehabilitation, Zagreb, Hrvaška

Rezultati

- **Novice**
- **Izgredi po Kenijskih volitvah 2009***

* IPrA Research Center, University of Antwerp, Netherlands

Rezultati

- **Novice**
- **Izgredi po Kenijskih volitvah 2009**
- **Primerjava lokalnih in „zahodnih“ novic**

DAILY NATION

THE STANDARD

THE TIMES

The INDEPENDENT

The New York Times

The Washington Post

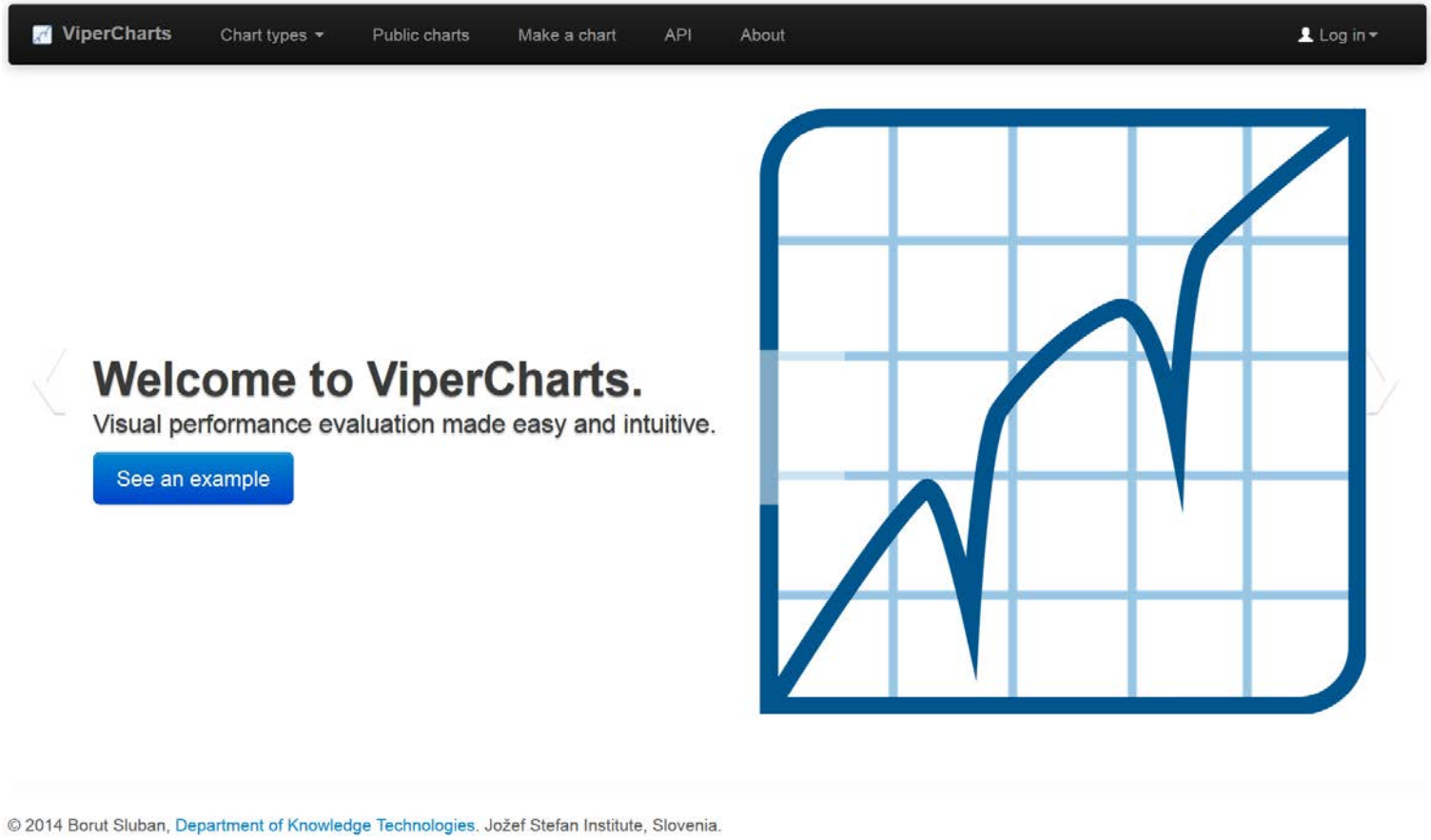
* IPrA Research Center,
University of Antwerp, Belgija

Rezultati

- 1. Napake pri zajemu
- 2. Gostujoči novinarji,
  druge tematike,
  drugi žanr

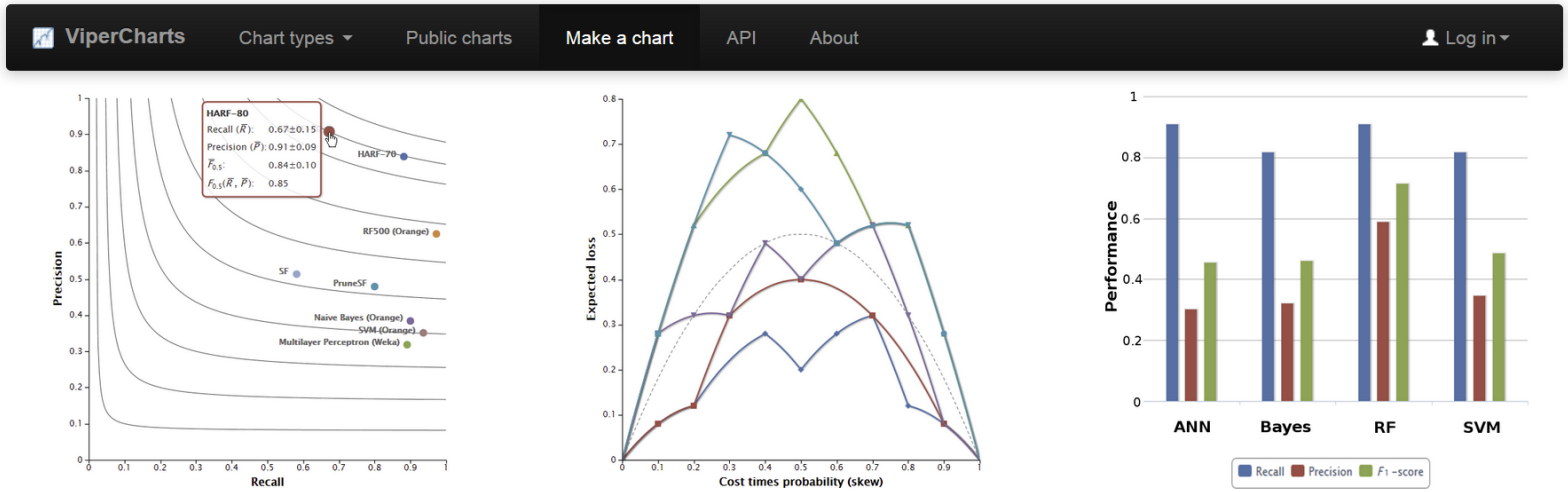
Rank	Class	ID	Detected by:
1.	WE	351	__Bayes_____RF_____SVM_____NN_____SF___
1.	WE	356	__Bayes_____RF_____SVM____PruneSF_____SF___
1.	WE	357	__Bayes_____RF_____SVM____PruneSF_____SF___
2.	LO	3	__Bayes_____RF_____SVM____PruneSF_
2.	LO	24	__Bayes_____RF_____SVM_____NN___
2.	LO	100	__Bayes_____RF_____SVM_____NN___
2.	LO	161	__Bayes_____RF_____SVM_____NN___
2.	LO	172	__Bayes_____RF_____SVM_____NN___
2.	WE	325	__Bayes_____RF_____SVM_____NN___
2.	WE	409	__Bayes_____RF_____SVM____PruneSF_
3.	LO	60	__Bayes___PruneSF_____SF___
3.	LO	152	__Bayes_____RF_____SVM___
3.	LO	347	__Bayes_____RF_____SVM___
3.	WE	363	__Bayes_____RF_____SVM___
3.	WE	369	__Bayes_____RF_____SVM___
3.	WE	463	____RF_____SVM_____NN___
4.	LO	20	____RF_____SVM___
4.	LO	67	____RF_____SVM___
4.	LO	119	__Bayes_____SF___
4.	LO	157	__Bayes___PruneSF_
4.	LO	158	_PruneSF_____SF___
4.	LO	184	__Bayes___PruneSF_
4.	LO	200	_PruneSF_____SF___
4.	LO	214	__Bayes_____NN___
4.	LO	220	_PruneSF_____SF___
4.	LO	221	__Bayes___PruneSF_
4.	WE	237	____RF_____SVM___
4.	WE	246	__Bayes_____SVM___
4.	WE	366	__Bayes___PruneSF_
4.	WE	374	__Bayes___PruneSF_
4.	WE	384	__Bayes___PruneSF_
4.	WE	386	___SVM_____NN___
4.	WE	407	__Bayes___PruneSF_
4.	WE	412	____RF_____NN___
4.	WE	415	_PruneSF_____SF___
4.	WE	420	__Bayes_____SF___
5.	LO	7	_PruneSF_
5.	LO	17	__Bayes__
5.	LO	70	_PruneSF_
5.	LO	83	___SVM___
5.	LO	129	____NN___
5.	LO	131	_PruneSF_
5.	LO	160	__Bayes__
5.	LO	180	___SVM___
5.	LO	186	__Bayes__
5.	LO	213	____NN___
5.	WE	245	_PruneSF_
5.	WE	311	____RF___
5.	WE	321	_PruneSF_
5.	WE	327	____RF___
...			

Kvantitativna analiza



http://viper.ijs.si

Kvantitativna analiza



Scatter Chart

Visual performance evaluation: compare the performance of your information retrieval, entity recognition or detection algorithms in the Precision-Recall or ROC space. More »

Make a Scatter Chart »

Curve Chart

Easily chart the performance of your binary classifier systems. Included performance visualizations for PR curves, Lift curves, ROC curves, Cost curves, Rate-driven curves and Kendall curves. More »

Make a Curve Chart »

Column Chart

Standard graphical presentation of algorithm performance. Vizualizes the values of one or more performance measures on the evaluated algorithms. More »

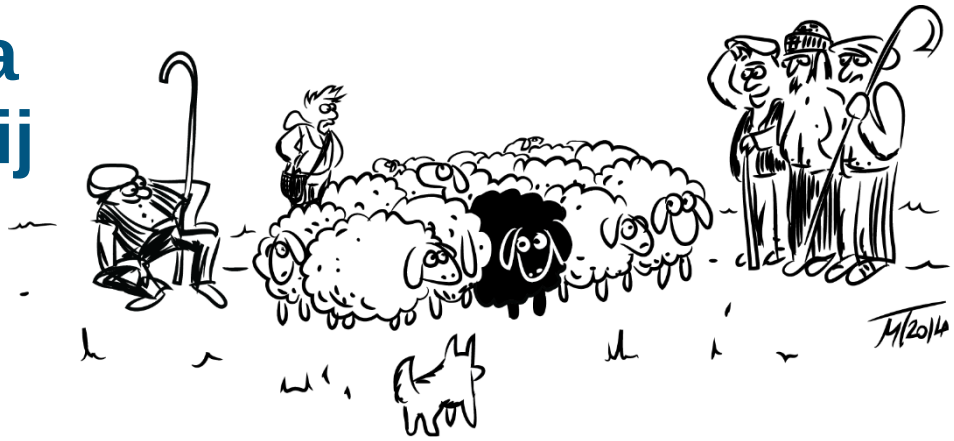
Make a Column Chart »

http://viper.ijs.si

JAVNA AGENCIJA ZA RAZISKOVALNO DEJAVNOST REPUBLIKE SLOVENIJE
SPIRIT SLOVENIJA I FEEL SLOVENIA
REPUBLIKA SLOVENIJA MINISTRSTVO ZA GOSPODARSKI RAZVOJ IN TEHNOLOGIJO

Povzetek

- **Ansambelska metoda za odkrivanje anomalij v podatkih**



Povzetek

- **Ansambelska metoda za odkrivanje anomalij v podatkih**

- **Podpora strokovnjakom pri analizi domenskih podatkov**

Povzetek

- **Ansambelska metoda za odkrivanje anomalij v podatkih**

- **Podpora strokovnjakom pri analizi domenskih podatkov**

- **Javno dostopno, razširljivo, nadgradljivo**

Hvala
za Vašo pozornost