LAPI – Laboratorul de Analiza şi Prelucrarea Imaginilor

Universitatea POLITEHNICA din Bucureşti

Facultatea de Electronică, Telecomunicaţii şi Tehnologia Informaţiei

TACAI - Tehnici de Analiză și Clasificare Automată a Informației

Note de laborator

Dr.ing. Ionuț Mironică

Conf.dr.ing. Bogdan Ionescu

Laborator 1

Cuprins:

- Introducere în algoritmi de clasificare
- Introducere Weka
- Evaluarea performanței de clasificare
- Exerciții

I. Introducere

Ce este Machine learning?

- **Exemplu aplicație:** Este foarte greu de scris un algoritm care recunoaşte un set de gesturi statice / dinamice sau care să rezolve problema de recunoaştere a feţei:
  - Nu ştim cum funcţionează creierul uman pentru a clasifica gesturile;
  - Chiar dacă am şti nu am avea idee cum să programăm deoarece ar fi foarte complicat;
  - Ar trebui să scriem o funcție diferită pentru fiecare gest.

- În loc să scriem programe foarte multe, putem colecta exemple care specifică fiecare gest;

- Un algoritm de învăţare va prelua aceste exemple şi va “creea” un program care va face această clasificare în mod automat;

I. Introducere

Ce este Machine learning?

- Există mii de algoritmi de învăţare / sute dintre ei apar anual;
- Există mai multe tipuri de învăţare:

Învățare supervizată

- datele de antrenare conţin şi ieşirea dorită;

Învățare nesupervizată

- datele de antrenare nu conţin ieşirea dorită (clusterizare);
- ideea de bază este de a se găsi şabloane şi pattern-uri în date care să fie evidenţiate în mod automat.

Învăţare semi-supervizată

- doar o parte din datele de antrenare conţin ieşirea dorită;

Reinforcement Learning

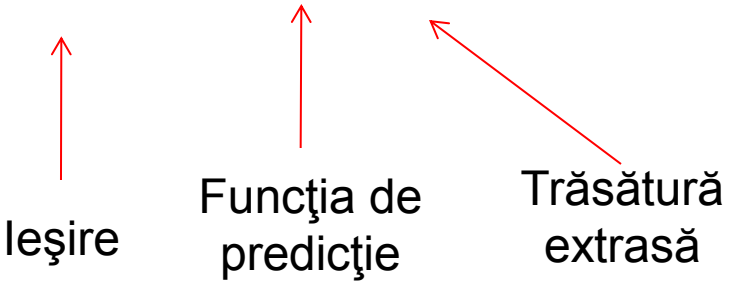
- se învaţă în funcţie de feedback-ul primit după ce o decizie este luată.

I. Introducere

Învăţare supervizată

Model de bază

$$y = f(\mathbf{x})$$



Antrenare: fiind dată o mulţime de antrenare împreună cu răspunsul dorit $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, se estimează predicţia funcţiei f prin minimizarea erorii de predicţie pe mulţimea de antrenare;

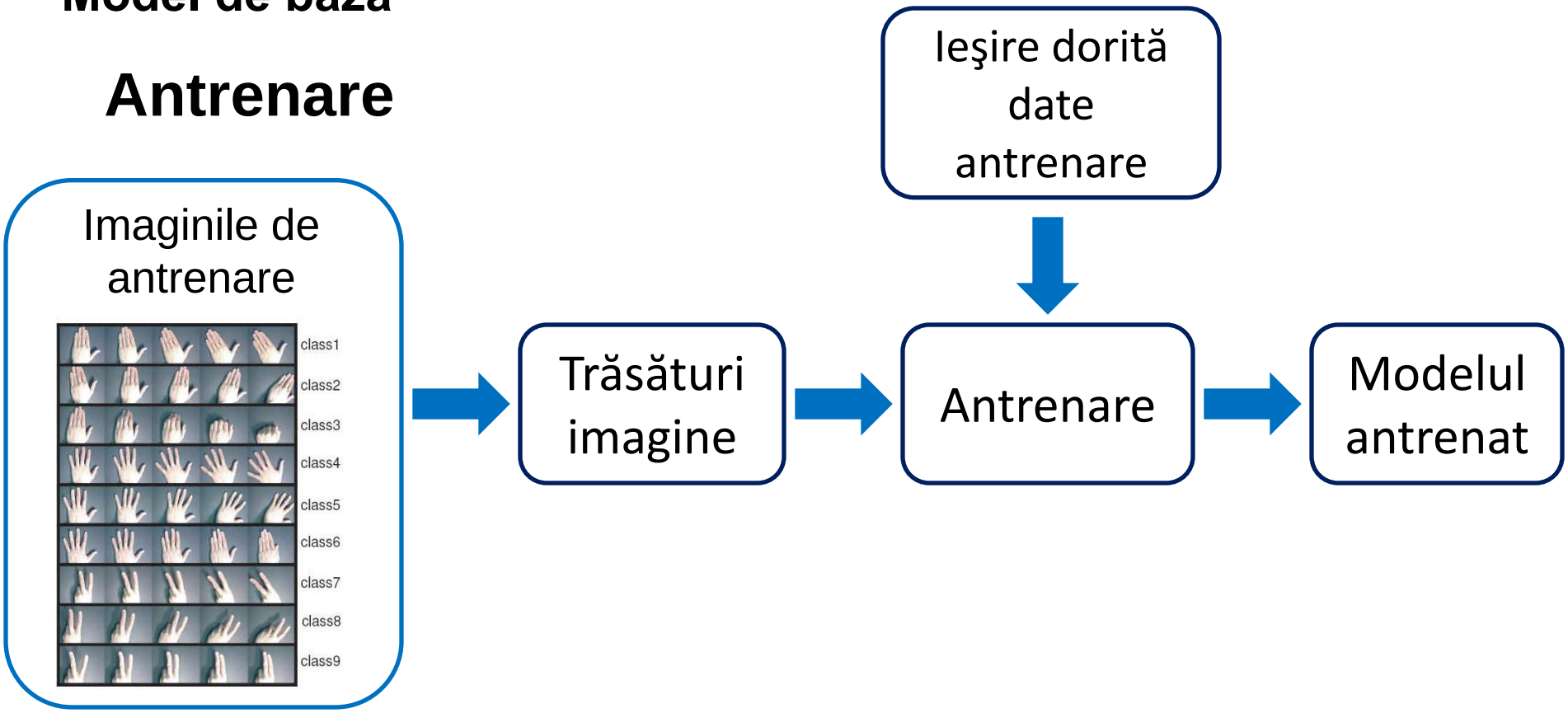
Testare: se aplică funcţia f pe un exemplu de test $\mathbf{x}$(care nu a fost folosit în procesul de antrenare) şi prezintă ieşirea funcţiei $y = f(\mathbf{x})$.

I. Introducere

Învăţare supervizată

Model de bază

Antrenare



I. Introducere

Învăţare supervizată

Model de bază

Testare

Extragere trăsături (X)

Aplicare funcţie de predicţie f(x)

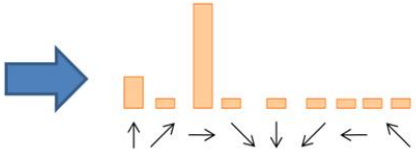
Predicţie

I. Introducere

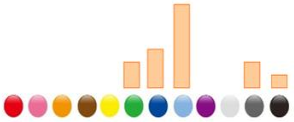
Învăţare supervizată

Ce sunt trăsăturile?

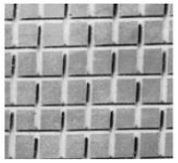
pixeli

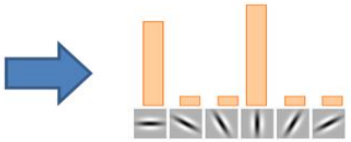


muchii



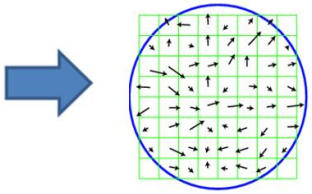
culoare





textură

formă

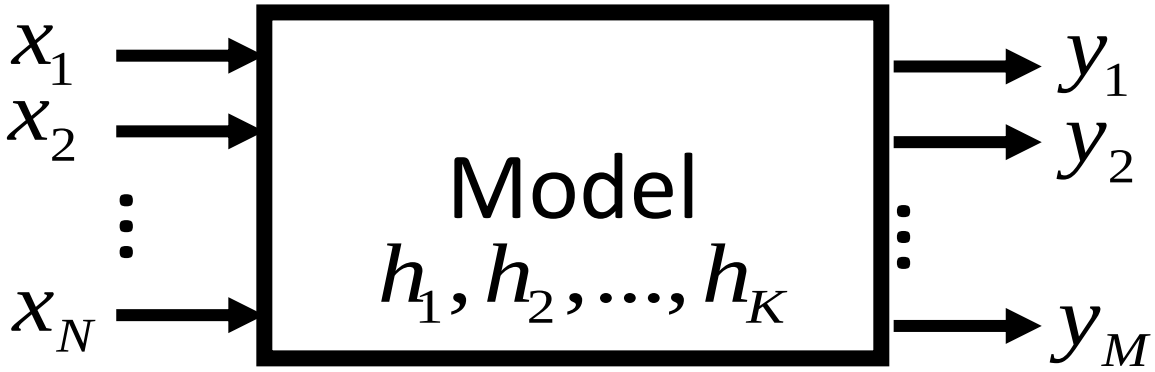


puncte de interes

I. Introducere

Învăţare supervizată

Învățare supervizată – schemă de bază



Variabile de intrare: $\mathbf{x} = (x_1, x_2, \ldots, x_N)$

Variabile ascunse: $\mathbf{h} = (h_1, h_2, \ldots, h_K)$

Variabile de ieşire: $\mathbf{y} = (y_1, y_2, \ldots, y_K)$

I. Introducere

Învăţare supervizată

Algoritmi existenţi

- Support vector machines (SVM),
- Reţele neurale,
- Naïve Bayes,
- Reţele bayesiene,
- Arbori aleatorii,
- K-nearest neighbor (k-NN),

Etc.

Care este cel mai bun algoritm?

I. Introducere

Învăţare supervizată

Teorema “No free lunch”

II. Weka

Informații generale

- Reprezintă o colecție de algoritmi de Machine Learning pentru diferite probleme de data-mining;
- Dezvoltat de către Universitea din Waikato, Noua Zeelandă;
- Open-Source dezvoltat in JAVA (GNU General Public License);
- Trăsături principale:
  - intrumente de preprocesare,
  - algoritmi de învățare,
  - algoritmi de clusterizare,
  - reguli de asociere,
  - metode de evaluare,
  - interfață grafică,
  - instrumente pentru comparația clasificatorilor.

II. Weka

Documentație

- **Site-ul Weka:**

http://www.cs.waikato.ac.nz/~ml/weka/

- **Documentație Weka:**

http://transact.dl.sourceforge.net/sourceforge/weka/WekaManual-3.6.0.pdf

http://www.cs.waikato.ac.nz/ml/weka/documentation.html

- **Tutoriale video:**

https://weka.waikato.ac.nz/explorer

II. Weka

Interfață

Exporer:

- reprezintă componenta principală vizuală Weka. Conține mai multe componente:
  - *Preprocess*: opțiuni de încărcare a bazelor de date (format arff / csv) / conectare Sql / procesări atașate datelor.
  - *Clasiffy*: algoritmi de clasificare;
  - Associate: algoritmi de asociere;
  - *Cluster*: algoritmi de învățare nesupervizată
  - *Select* attributes: algoritmi de detecție a atributelor
  - *Visualize* modalități de vizualizare a datelor.

Experimenter:

- permite configurarea unui sistem ce asigură comparația sistematică a perfomanțelor algorimilor de clasificare pe diferite colecții de baze de date.

KnowledgeFlow:

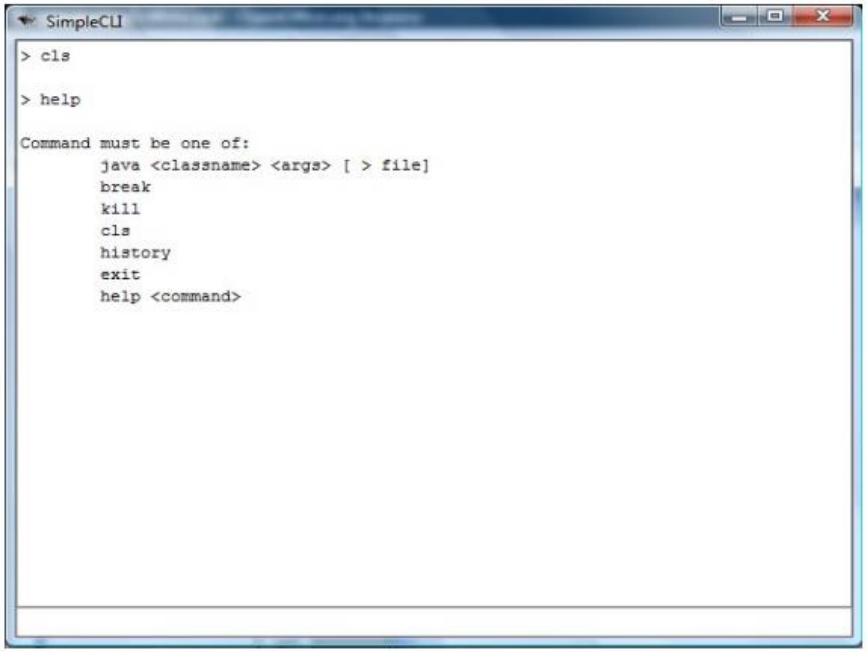
- permite configurarea de fluxuri automate.

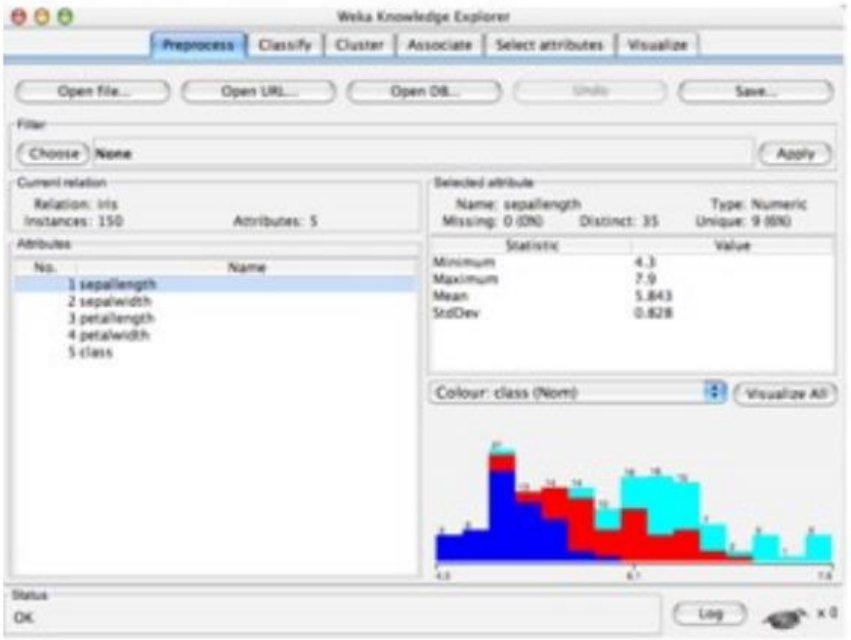
SimpleCLI:

- consolă de apelare a funcțiilor de clasificare.

II. Weka

CLI vs GUI





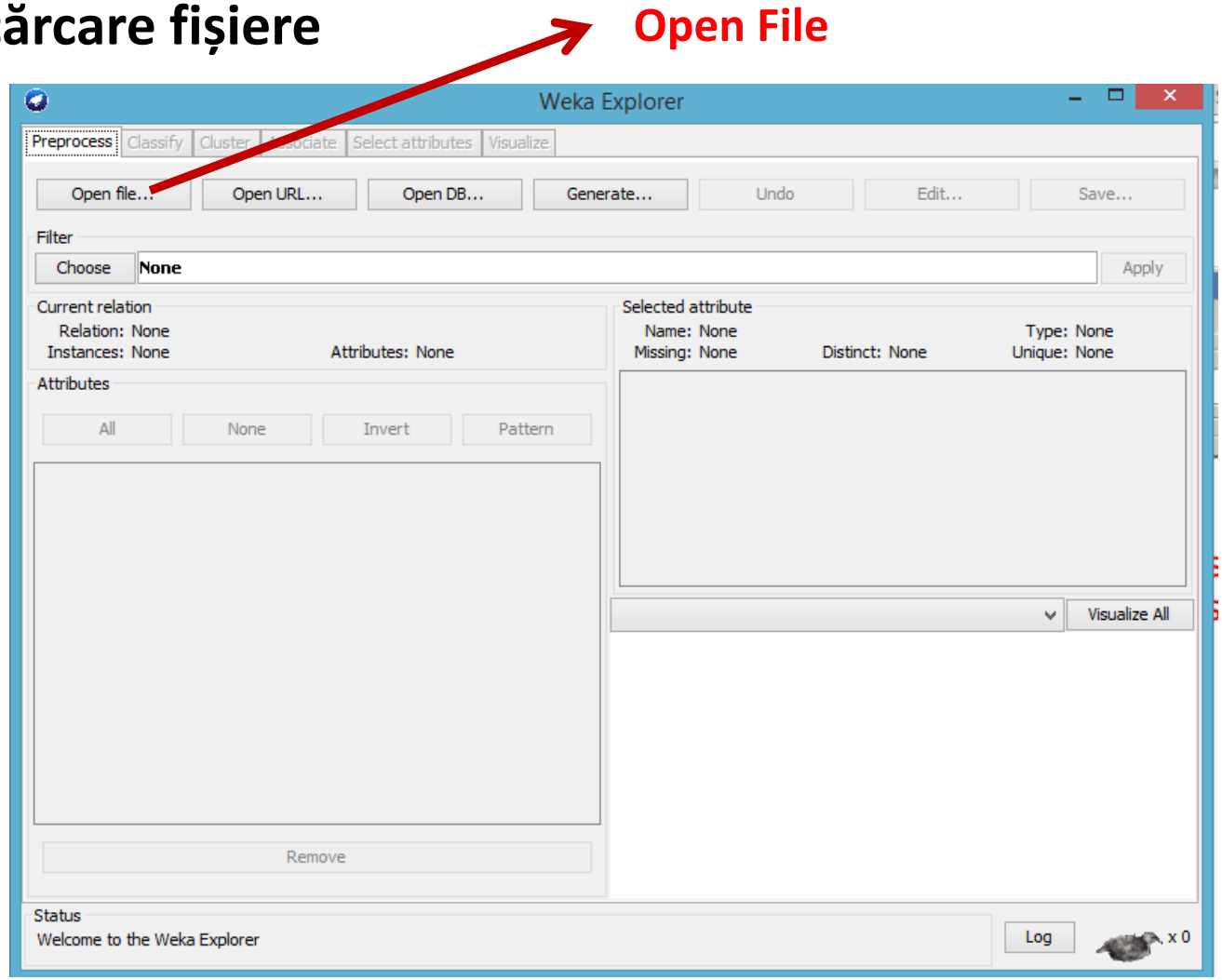
- Recomandat pentru utilizarea în profunzime a algoritmilor;
- Oferă funcționalități indisponibile în interfața grafică.

- Mai ușor de utilizat;
- Oferă funcționalități intuitive: Explorer, Experimenter și KnowledgeFlow.

II. Weka

Explorer: Preprocess

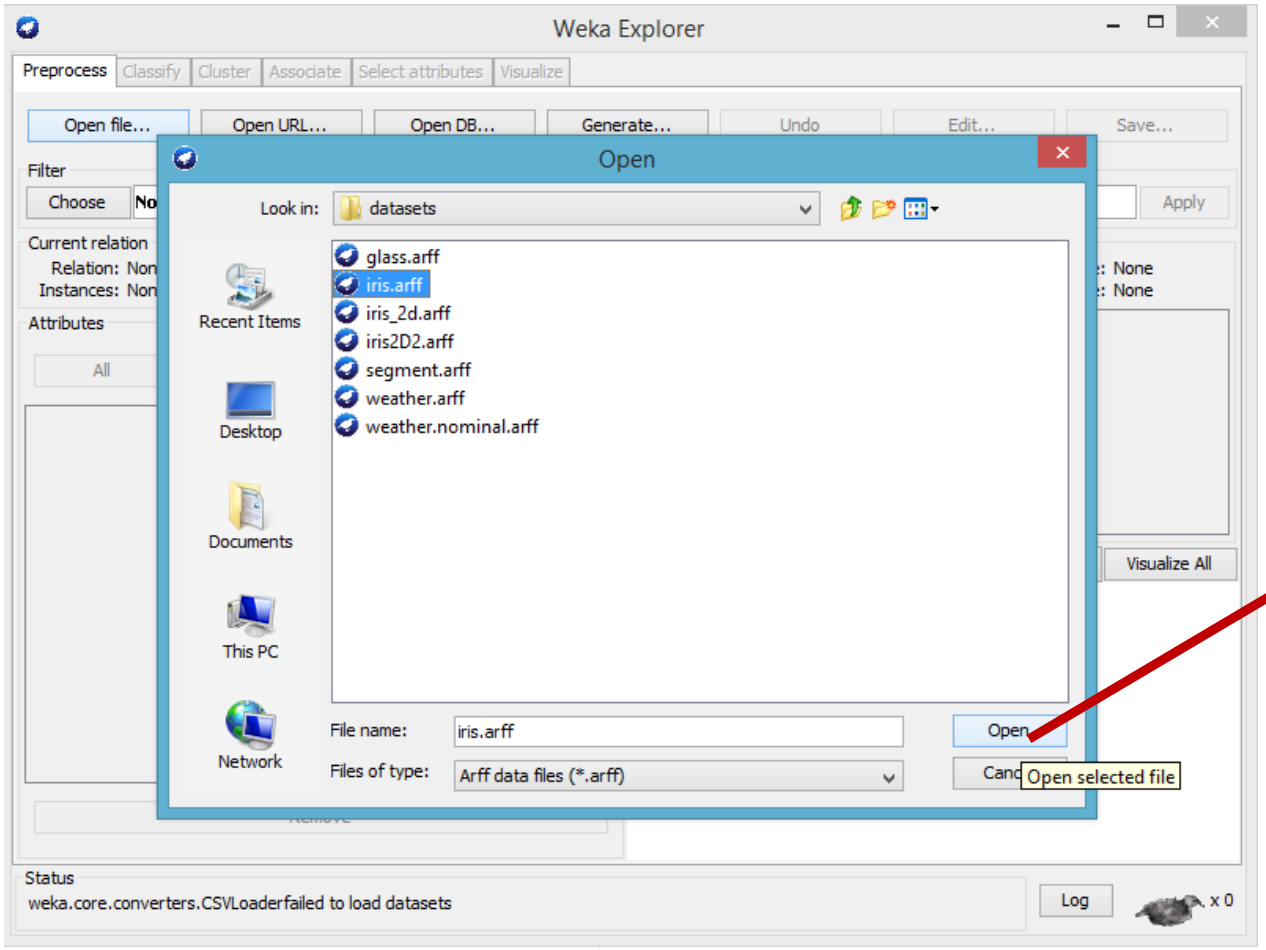
Încărcare fișiere



II. Weka

Explorer: Preprocess

Încărcare fișiere



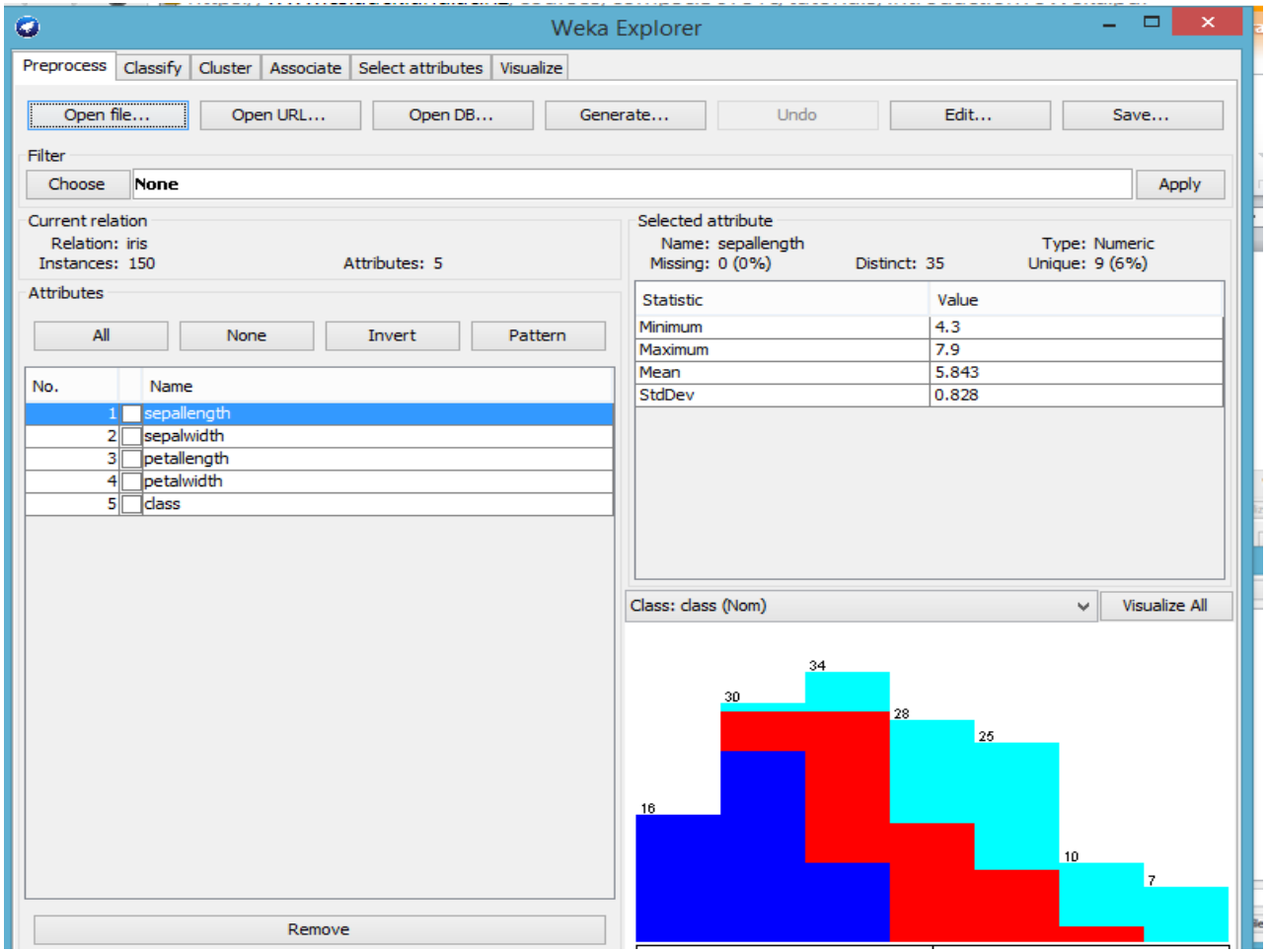
Open:

Selecție fișier (arff / csv)

II. Weka

Explorer: Preprocess

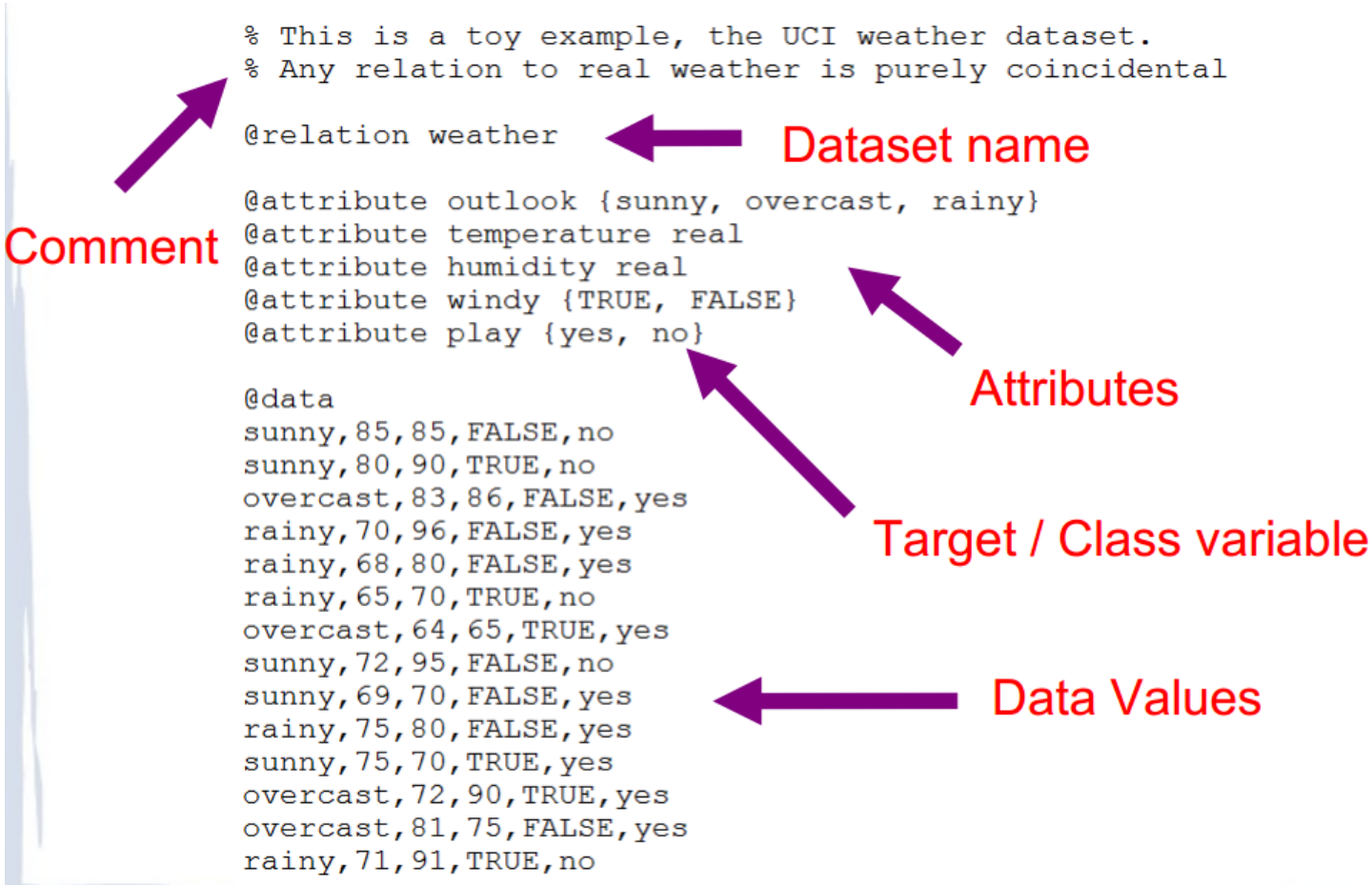
Încărcare fișiere



II. Weka

Explorer: Preprocess

Structură fișier arff



II. Weka

Explorer: Preprocess

Atribute fișier arff

- **Atribute nominale**: valorile sunt selectate dintr-o listă predefinită;
- **Atribute numerice:** valorile sunt întregi sau reale;
- **Șiruri de caractere (string):** sunt încadrate între ghilimele;
- **Relaționale:** pentru date care aparțin în mai multe tipuri de instanțe.

II. Weka

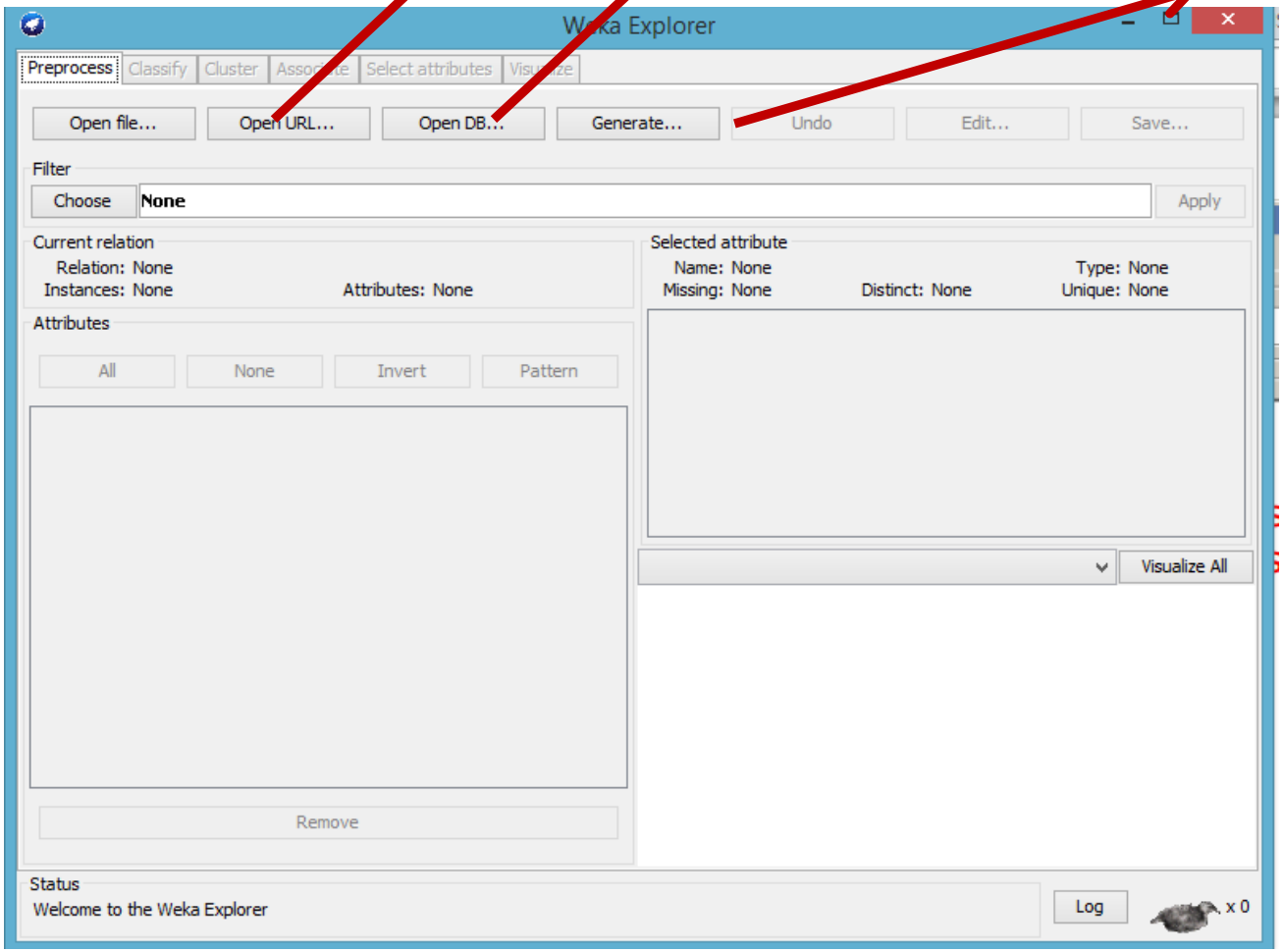
Explorer: Preprocess

Încărcare fișiere

Open URL: încărcare fișiere aflate la un link;

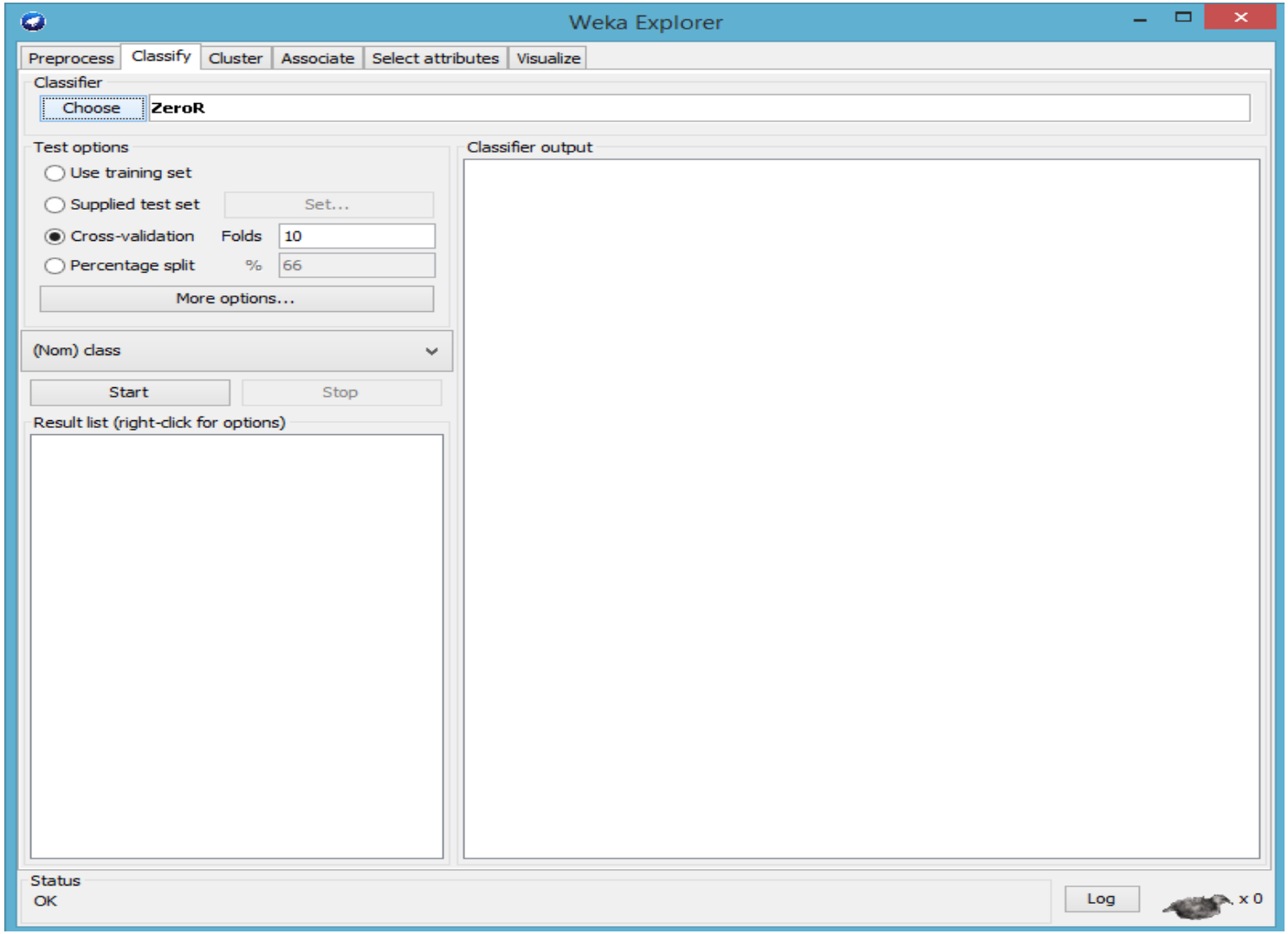
Open DB: conectare la o baza de date relațională (SQL) prin driver JBDC;

Generate: generare baze de date.



II. Weka

Utilizare algoritmi de clasificare



II. Weka

Explorer: Preprocess

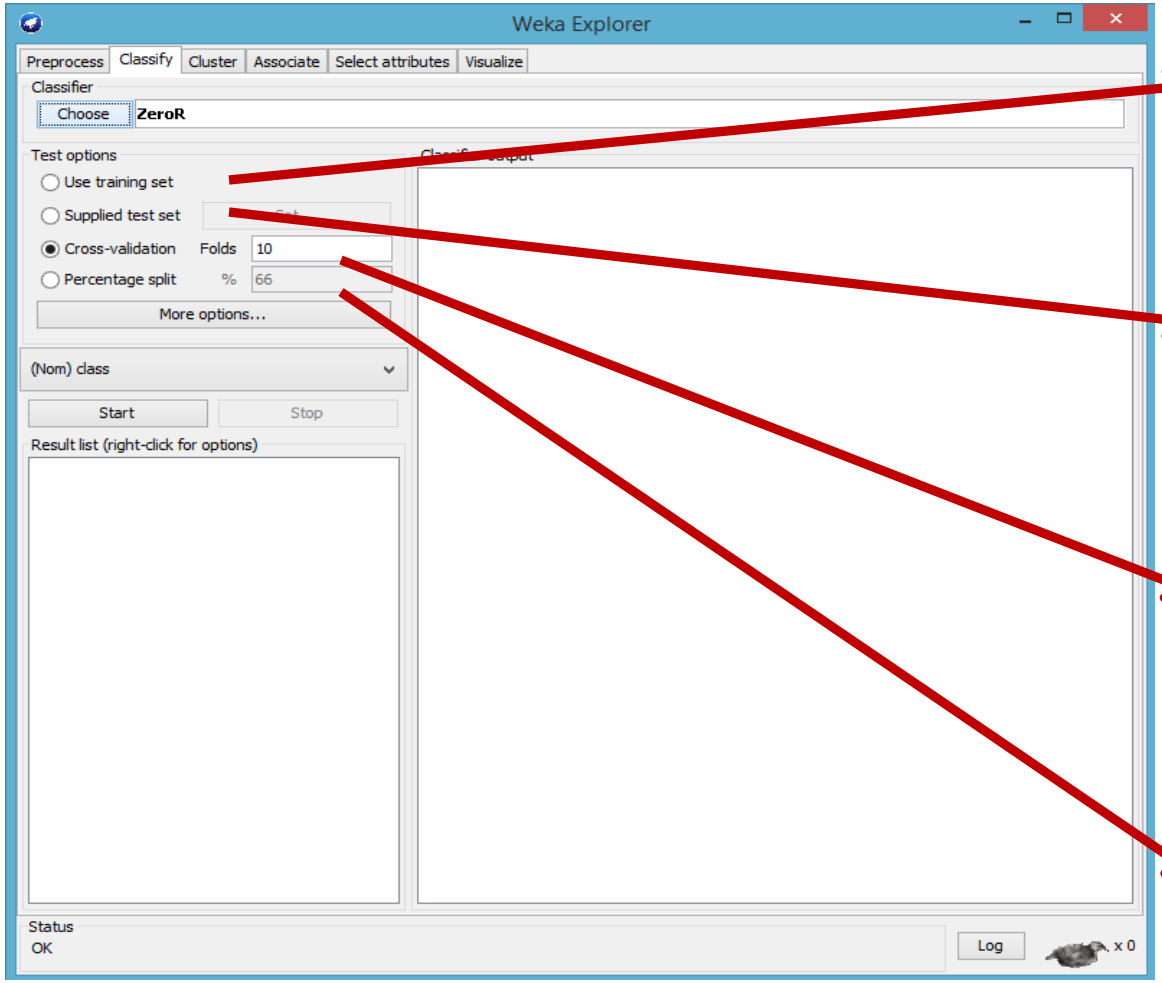
Încărcare fișiere

1. Deschide weka GUI
2. Click 'Explorer'
3. 'Open file...'
4. Selectează tabul 'Classify' Alege un clasificator
5. Selectează parametrii de clasificar
7. Click 'Start'
8. Wait...
9. Vizualizează rezultate
    a. ‘Salvează rezultat'
    b. 'Salvează clasificator'

II. Weka

Utilizare algoritmi de clasificare

Opțiuni de împărțire a bazei de date (Test Options)



Use training set: utilizarea setului de antrenare și pentru testare;

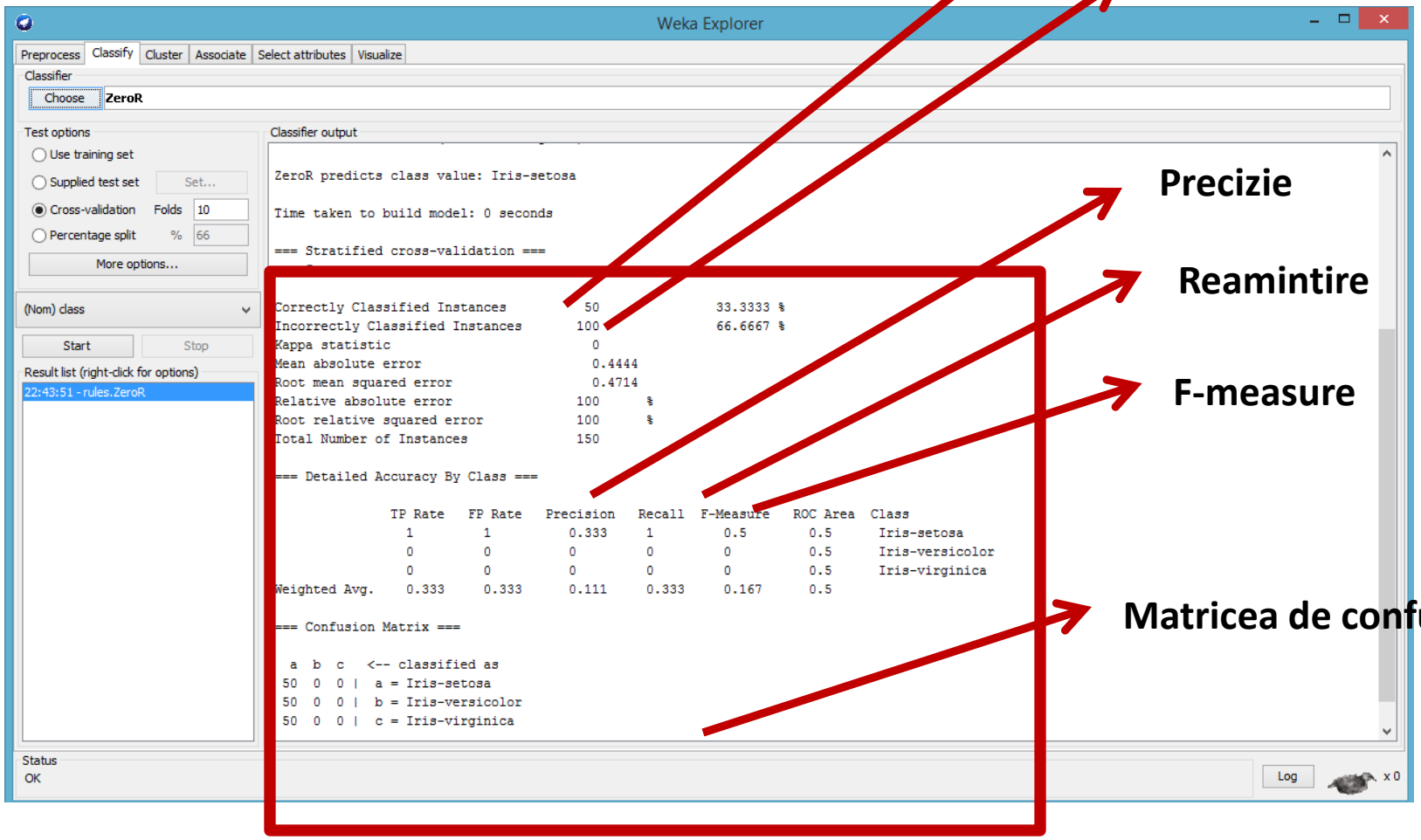
Supplied testset: baza de date de test încărcată separat;

Cross-validation: împărțirea setului de antrenare în mai multe seturi succesive de antrenare / testare;

Percentage split: împărțirea setului de antrenare într-un set de de antrenare și testare.

III. Evaluarea performanței de clasificare

Parametri de evaluare



Acuratețe de clasificare

Procent clasificări incorecte

Precizie

Reamintire

F-measure

Matricea de confuzie

III. Evaluarea performanței de clasificare

Acuratețea de clasificare

- Dat fiind un sistem de clasificare, fiecare document este clasificat ca relevant / nerelevant.
- Acuratețea de clasificare: procentul de documentele care sunt clasificate corect:
- $Acuratețea = \frac{TP+TN}{TP+TN+FN+FP}$
- $Procentul\ de\ clasificări\ incorecte = \frac{FN+FP}{TP+TN+FN+FP}$

= 1 – acuratețe

	Relevant	Nerelevant
Regăsit	True positive (TP)	False pozitive (FP)
Neregăsit	False negative (FN)	True negative (TN)

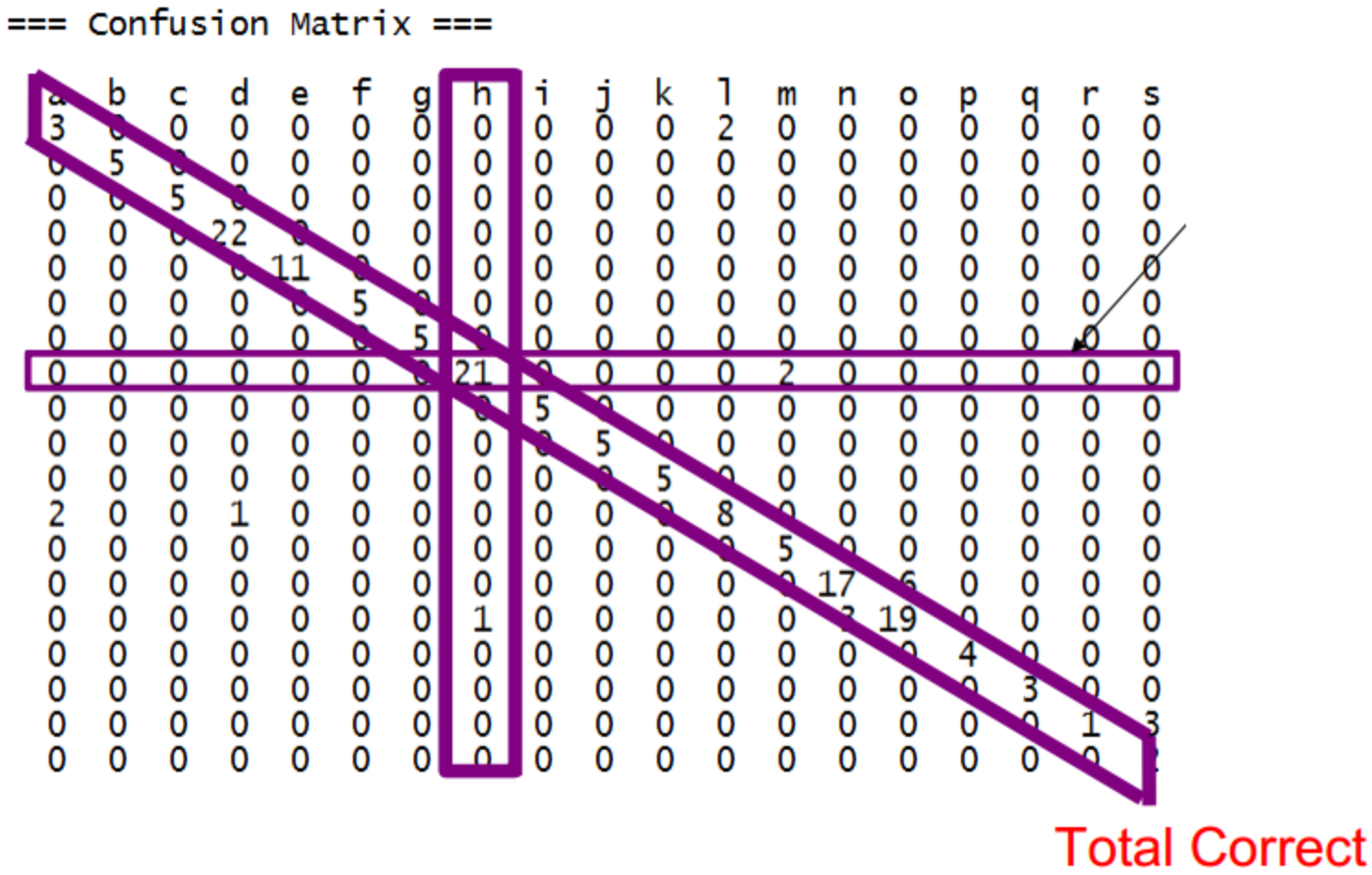
III. Evaluarea performanței de clasificare

Matricea de confuzie

		Predicted Class		
		Class 1	Class 2	Class 3
Actual Class	Class 1	2	1	1
	Class 2	1	2	1
	Class 3	1	2	3

III. Evaluarea performanței de clasificare

Matricea de confuzie



IV. Exerciții

- Încărcați rând pe rând bazele de date:

Weather.arff, Iris.arff și Glass.arff;

- Utilizați ca și clasificatori: ZeroR, OneR, arbori de decizie (J48), Nearest Neighbor (IBK) și Naive Bayes. Verificați care este cel mai bun clasificator. Rețineți acuratețea de clasificare a fiecărui clasificator în parte;

- Eliminați rând pe rând câte o coloană din bazele de date și comparați acuratețea cu cea obținută ulterior;

IV. Exerciții

- Adăugați zgomot un filtru de preprocesare (Swap Values / Add noise) și verificați efectul acestora asupra acurateței de preprocesare.

- Modificați opțiunea “Cross-Validation Folds” cu “Percentage split”. Care este diferența dintre cele două opțiuni?