

LAPI – Laboratorul de
Analiza și Prelucrarea
Imaginilor



Universitatea
POLITEHNICA din
București



Facultatea de Electronică,
Telecomunicații și
Tehnologia Informației

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 recunoaște un set de gesturi statice / dinamice sau care să rezolve problema de recunoaștere a feței:
 - Nu știm cum funcționează creierul uman pentru a clasifica gesturile;
 - Chiar dacă am ști 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 învățare va prelua aceste exemple și va “creea” un program care va face această clasificare în mod automat;

I. Introducere

Ce este Machine learning?

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

Învățare supervizată

- datele de antrenare conțin și ieșirea dorită;

Învățare nesupervizată

- datele de antrenare nu conțin ieșirea dorită (clusterizare);
- ideea de bază este de a se găsi șabloane și pattern-uri în date care să fie evidențiate în mod automat.

Învățare semi-supervizată

- doar o parte din datele de antrenare conțin ieșirea dorită;

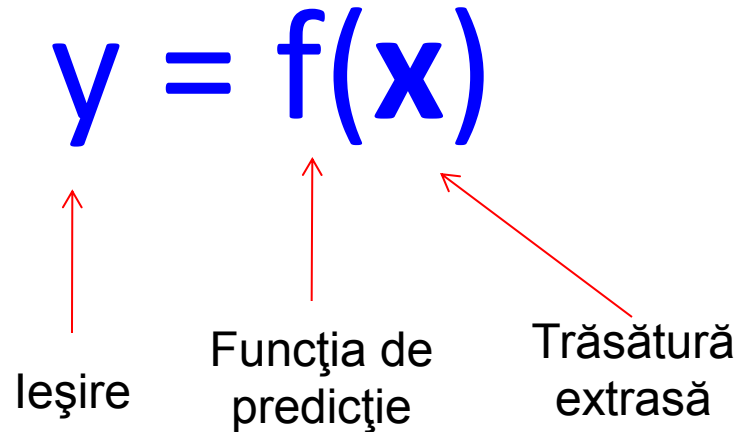
Reinforcement Learning

- se învață în funcție de feedback-ul primit după ce o decizie este luată.

I. Introducere

Învățare supervizată

Model de bază



Antrenare: fiind dată o mulțime de antrenare împreună cu răspunsul dorit $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, se estimează predicția funcției f prin minimizarea erorii de predicție pe mulțimea de antrenare;

Testare: se aplică funcția f pe un exemplu de test \mathbf{x} (care nu a fost folosit în procesul de antrenare) și prezintă ieșirea funcției $y = f(\mathbf{x})$.

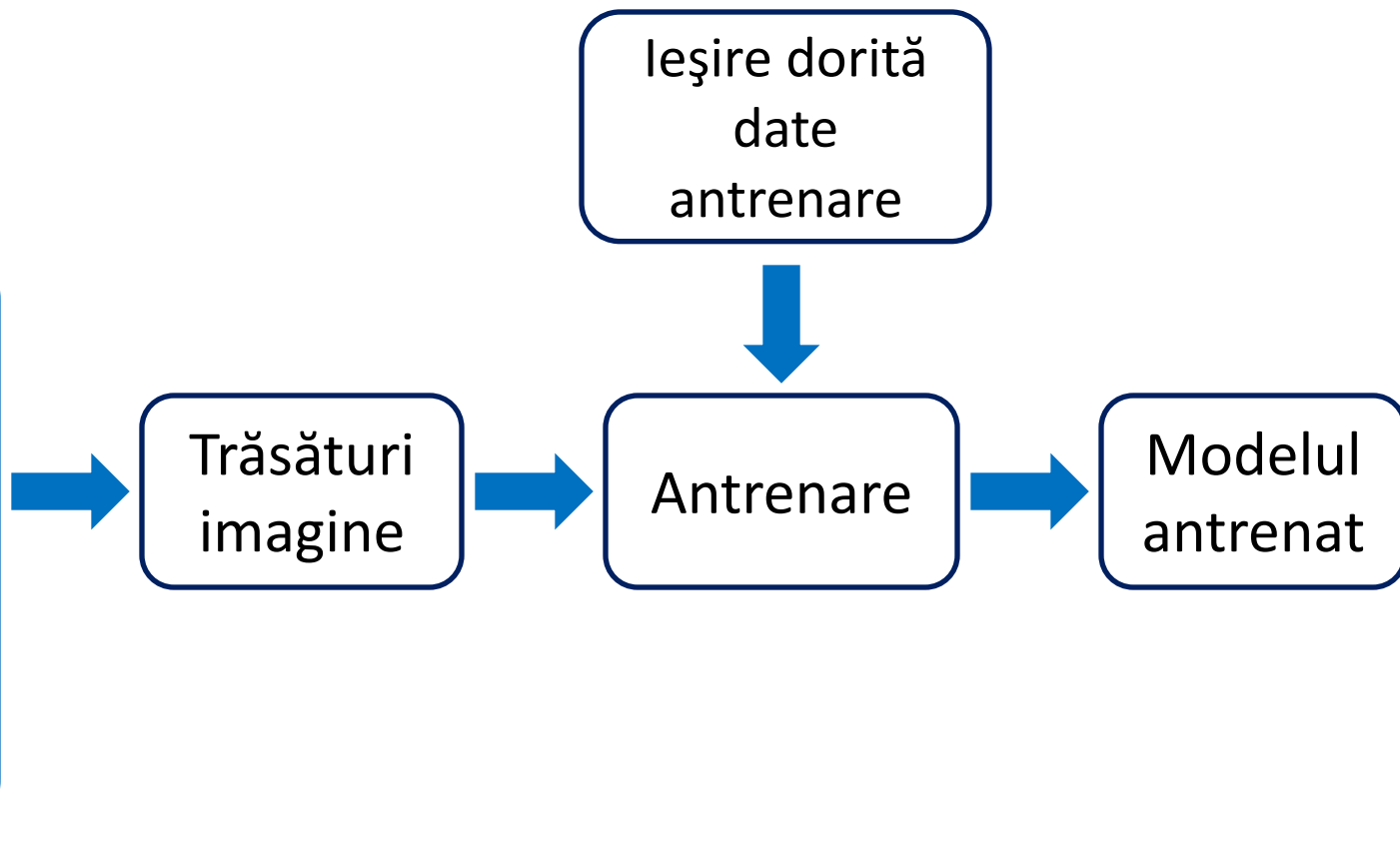
I. Introducere

Învățare supervizată

Model de bază

Antrenare

Imaginile de
antrenare

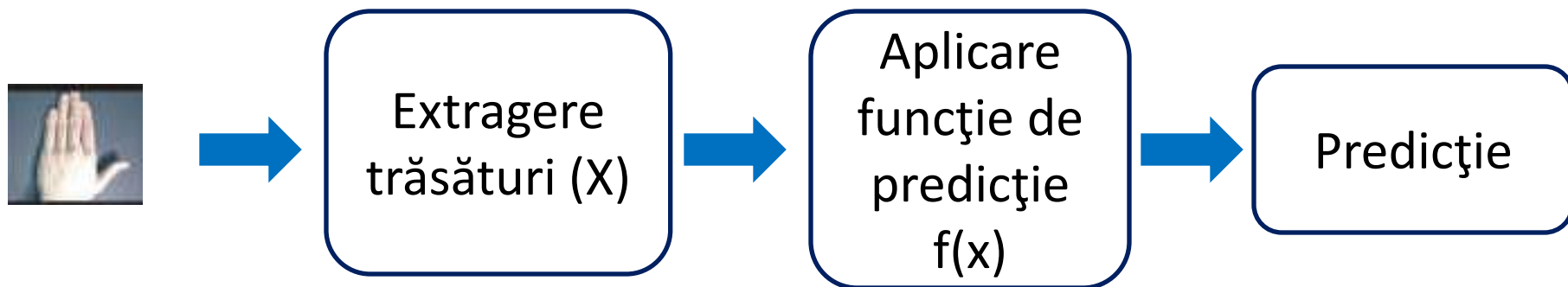


I. Introducere

Învățare supervizată

Model de bază

Testare



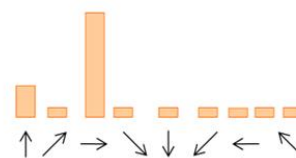
I. Introducere

Învățare supervizată

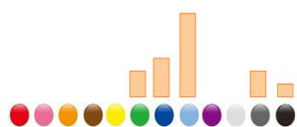
Ce sunt trăsăturile?



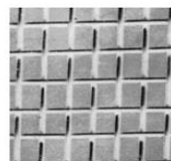
pixeli



muchii



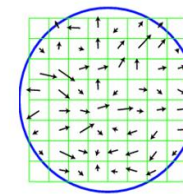
culoare



textură



formă

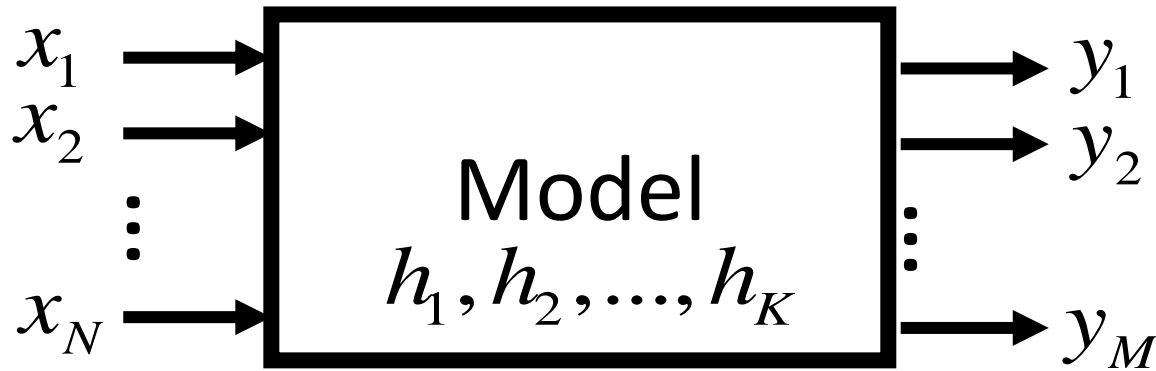


puncte de interes

I. Introducere

Învățare supervizată

Învățare supervizată – schemă de bază



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

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

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

I. Introducere

Învățare supervizată

Algoritmi existenți

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

Etc.

Care este cel mai bun algoritm?

I. Introducere

Învățare 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 Universitatea din Waikato, Noua Zeelandă;
- Open-Source dezvoltat in JAVA (GNU General Public License);
- Trăsături principale:
 - instrumente 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 performanțelor algorimilor de clasificare pe diferite colecții de baze de date.

KnowledgeFlow:

- permite configurarea de fluxuri automate.

SimpleCLI:

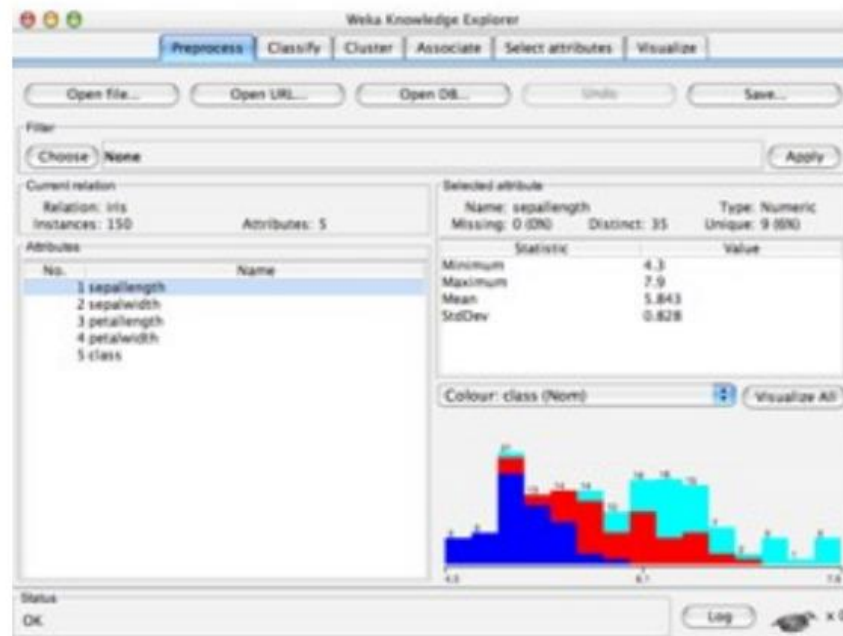
- consolă de apelare a funcțiilor de clasificare.



II. Weka

CLI vs GUI

```
SimpleCLI
> cls
> help
Command must be one of:
  java <classname> <args> [ > file]
  break
  kill
  cls
  history
  exit
  help <command>
```



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

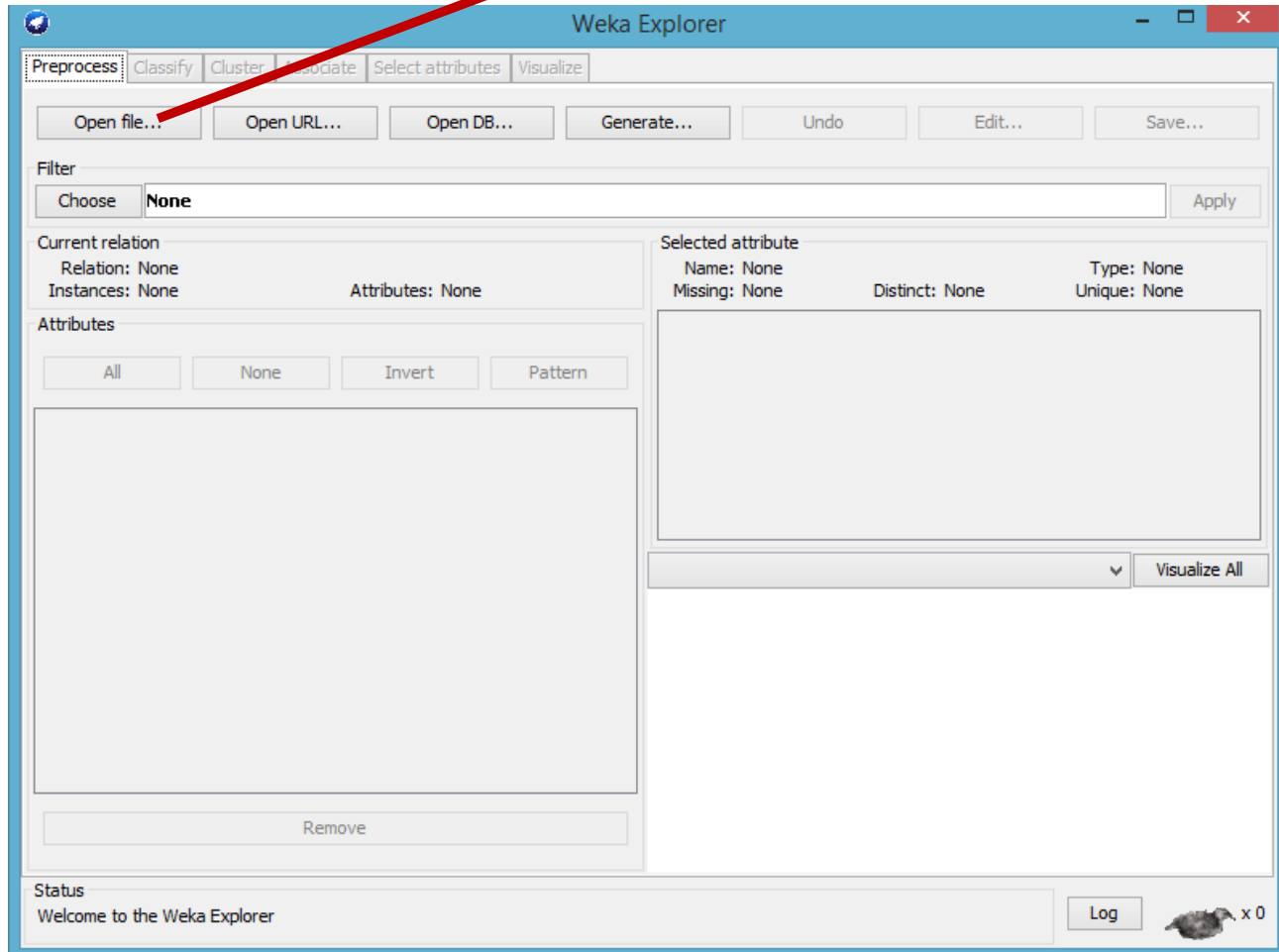
- Mai ușor de utilizat;
- Oferă funcționalități intuitive: Explorer, Experimenter și KnowledgeFlow.

II. Weka

Explorer: Preprocess

Încărcare fișiere

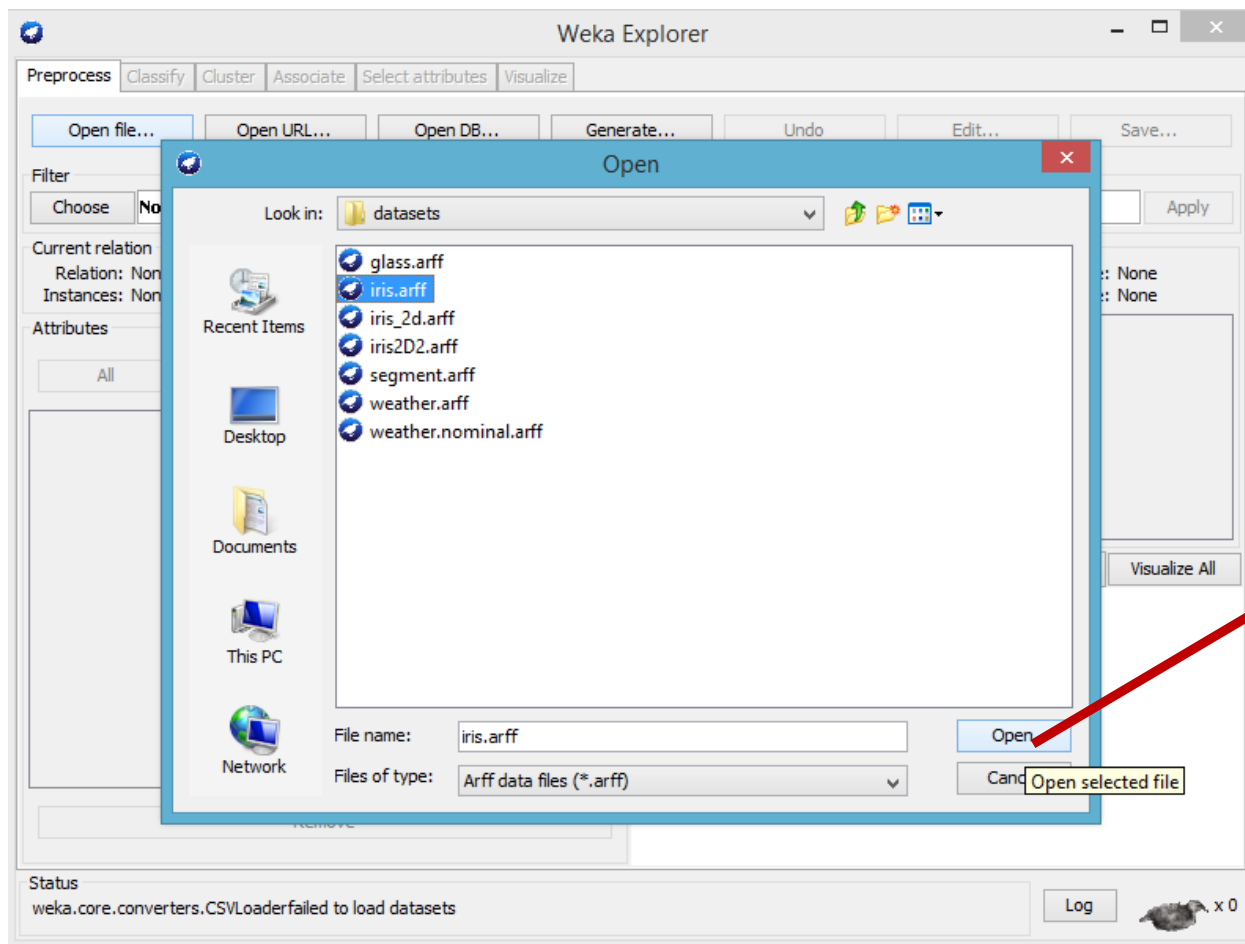
Open File



II. Weka

Explorer: Preprocess

Încărcare fișiere



Open:
Selecție fișier
(arff / csv)

II. Weka

Explorer: Preprocess

Încărcare fișiere

The screenshot shows the Weka Explorer interface in the Preprocess tab. The 'Current relation' is 'iris' with 150 instances and 5 attributes. The 'Selected attribute' is 'sepalength', which is numeric with 35 distinct values and 9 unique values (6%). A table of statistics for 'sepalength' is shown below:

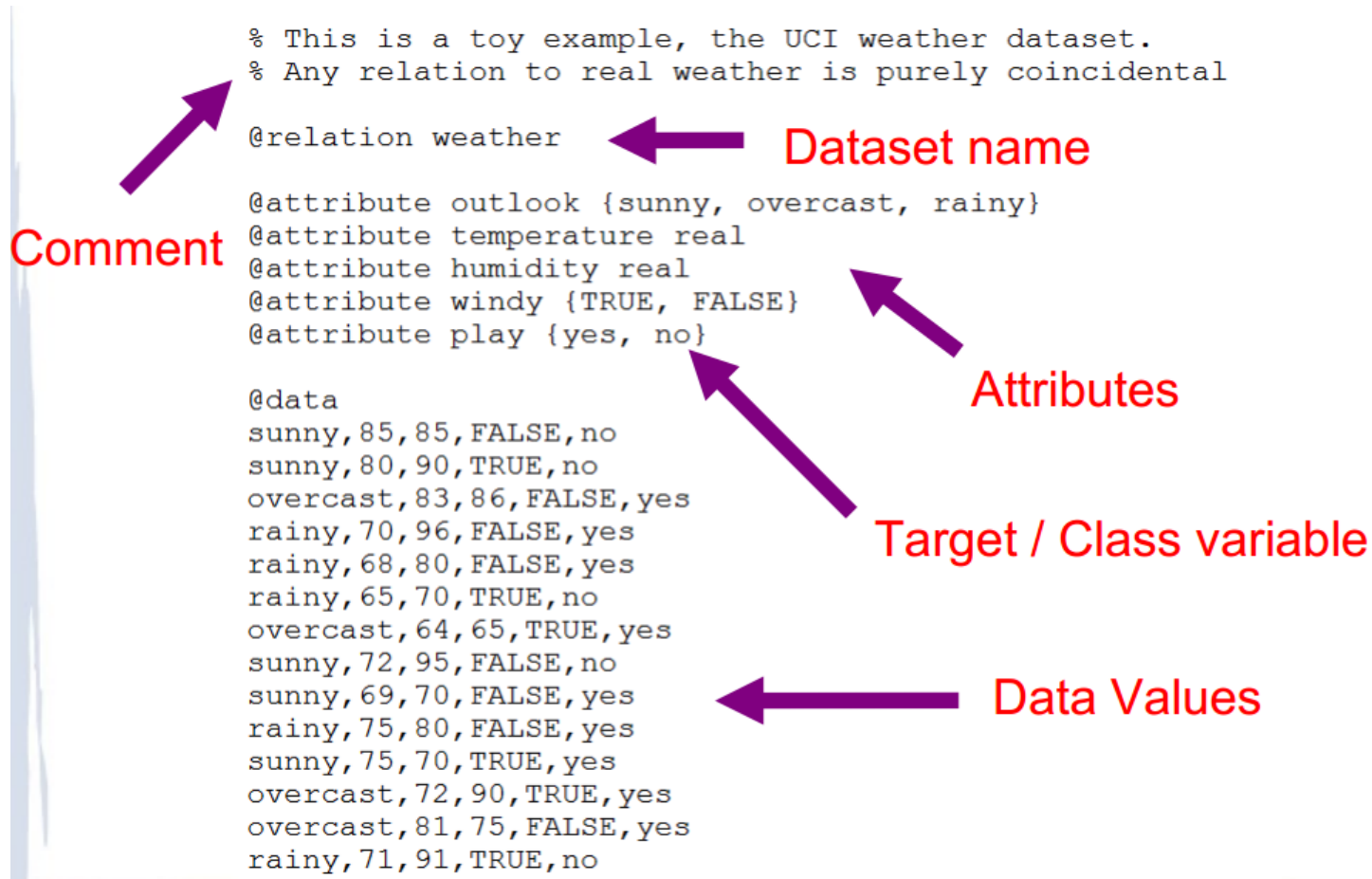
Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

The 'Attributes' list shows 'sepalength' selected. A histogram at the bottom right displays the distribution of 'sepalength' values, with bars colored blue, red, and cyan. The counts for each bar are 16, 30, 34, 28, 25, 10, and 7.

II. Weka

Explorer: Preprocess

Structură fișier arff



II. Weka

Explorer: Preprocess

Attribute fișier arff

- **Attribute nominale:** valorile sunt selectate dintr-o listă predefinită;
- **Attribute 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

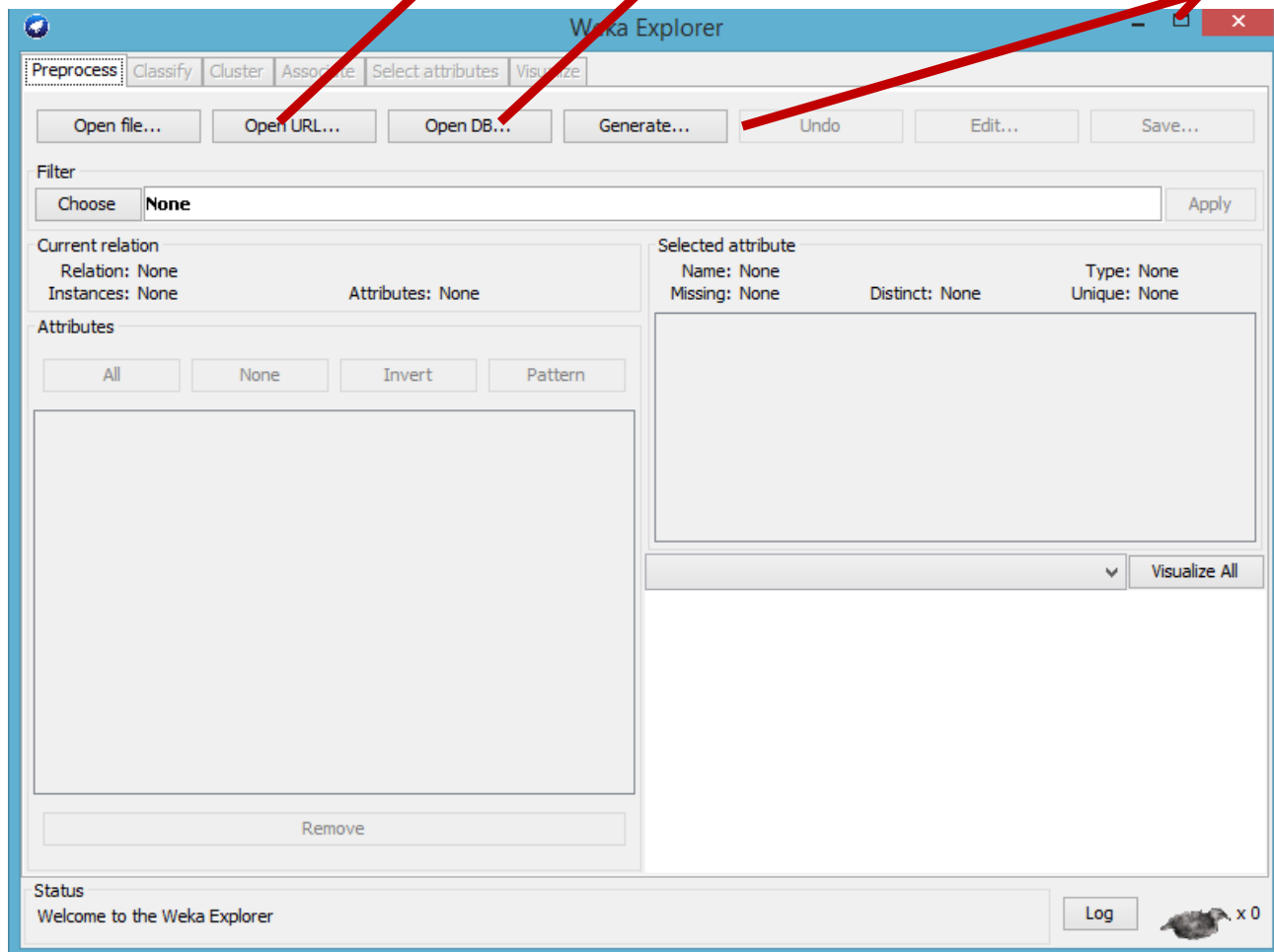
Explorer: Preprocess

Încărcare fișiere

Open URL: încărcare fișiere aflate la un link;

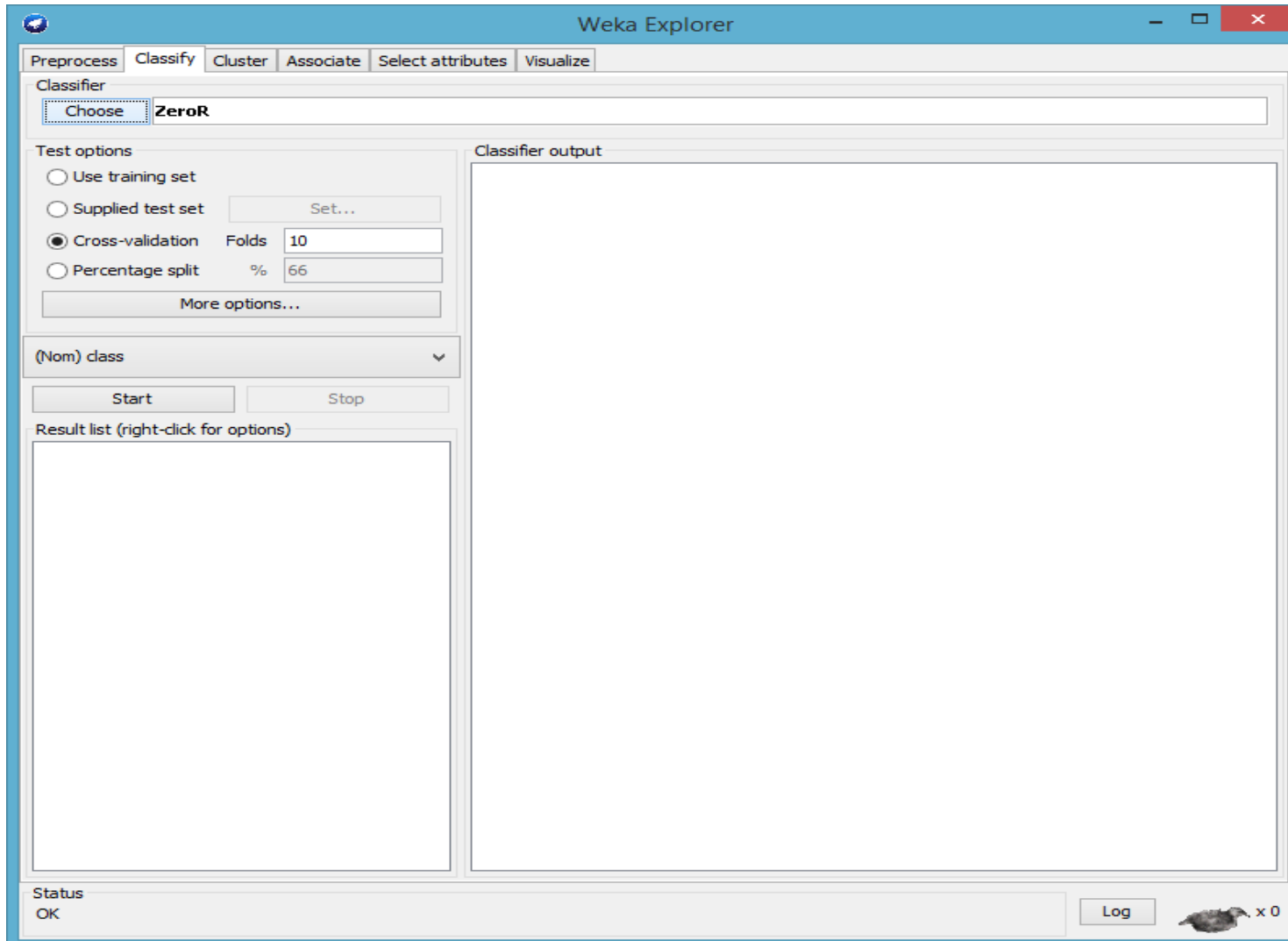
Open DB: conectare la o baza de date relațională (SQL) prin driver JDBC;

Generate: generare baze de date.



II. Weka

Utilizare algoritmi de clasificare



II. Weka

Explorer: Preprocess

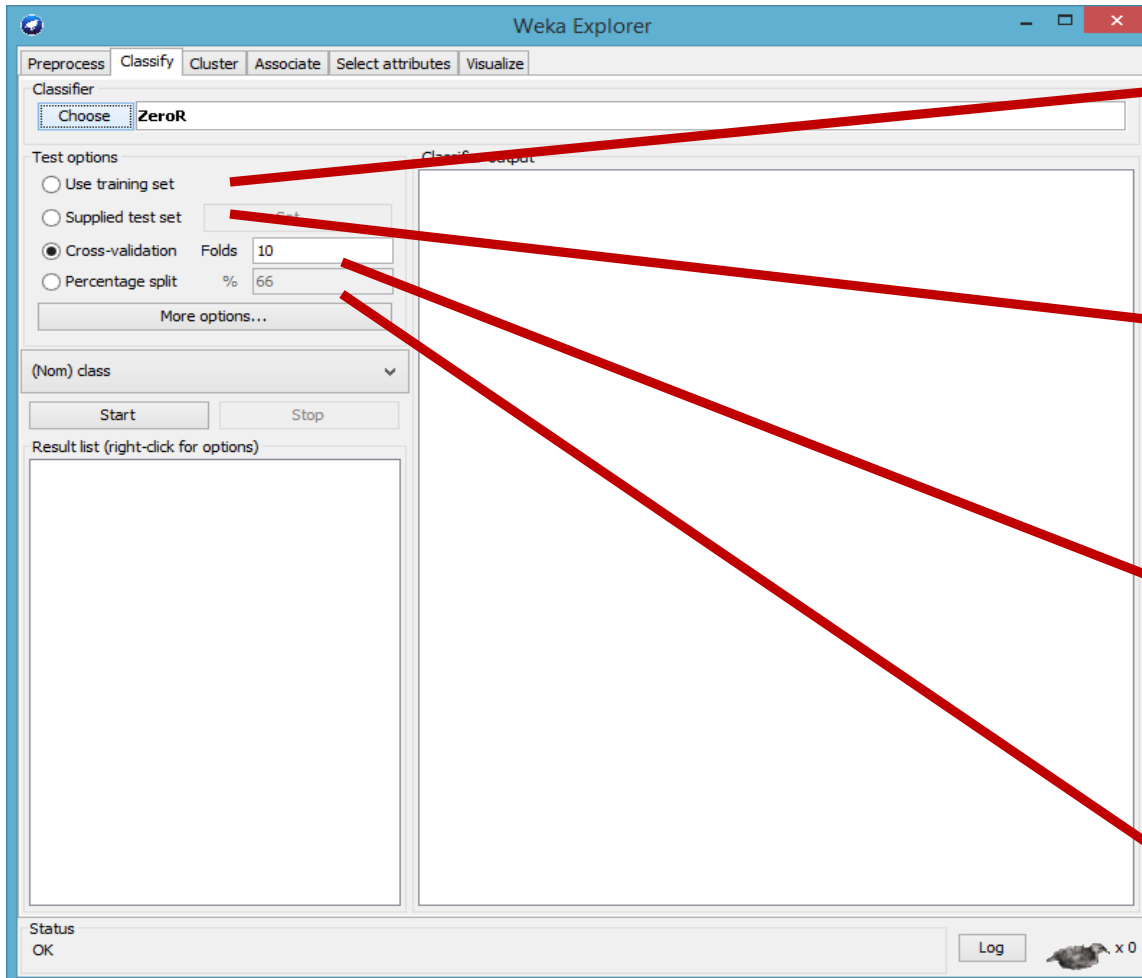
Încărcare fișiere

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

Weka Explorer

Classifier: ZeroR

Test options:
 Use training set
 Supplied test set
 Cross-validation Folds: 10
 Percentage split %: 66

Classifier output:

ZeroR predicts class value: Iris-setosa
Time taken to build model: 0 seconds
=== Stratified cross-validation ===

Correctly Classified Instances	50	33.3333 %
Incorrectly Classified Instances	100	66.6667 %
Kappa statistic	0	
Mean absolute error	0.4444	
Root mean squared error	0.4714	
Relative absolute error	100 %	
Root relative squared error	100 %	
Total Number of Instances	150	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.333	1	0.5	0.5	Iris-setosa
	0	0	0	0	0	0.5	Iris-versicolor
	0	0	0	0	0	0.5	Iris-virginica
Weighted Avg.	0.333	0.333	0.111	0.333	0.167	0.5	

=== Confusion Matrix ===

a	b	c	<-- classified as
50	0	0	a = Iris-setosa
50	0	0	b = Iris-versicolor
50	0	0	c = Iris-virginica

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:

- $Acurate\text{\u0219}ea = \frac{TP+TN}{TP+TN+FN+FP}$

- $Procentul\ de\ clasific\text{\u0219}ari\ incorecte = \frac{FN+FP}{TP+TN+FN+FP}$
 $= 1 - \text{acurate\text{\u0219}e}$

	Relevant	Nerelevant
Regăsit	True positive (TP)	False pozitiv (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

=== Confusion Matrix ===

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s
a	3	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
b	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	21	0	0	0	0	2	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0
j	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
l	2	0	0	1	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	17	6	0	0	0	0
o	0	0	0	0	0	0	0	1	0	0	0	0	0	7	19	0	0	0	0
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Total Correct

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