

Previsione delle portate con Decision tree

*Dall'Elaborato di
Giulio Carreras e Marco Zei
A.A. 2009 - 10*



Previsione qualitativa delle portate

👉 Elaborato svolto da *Giulio Carreras e Marco Zei*, A.A. 2009-10

👉 Previsione delle portate nel fiume Arno basata su alcuni parametri idrologici rilevati alla stazione di Subbiano (AR), in Casentino.

👉 In questa stazione si è recentemente registrata una forte riduzione di portata

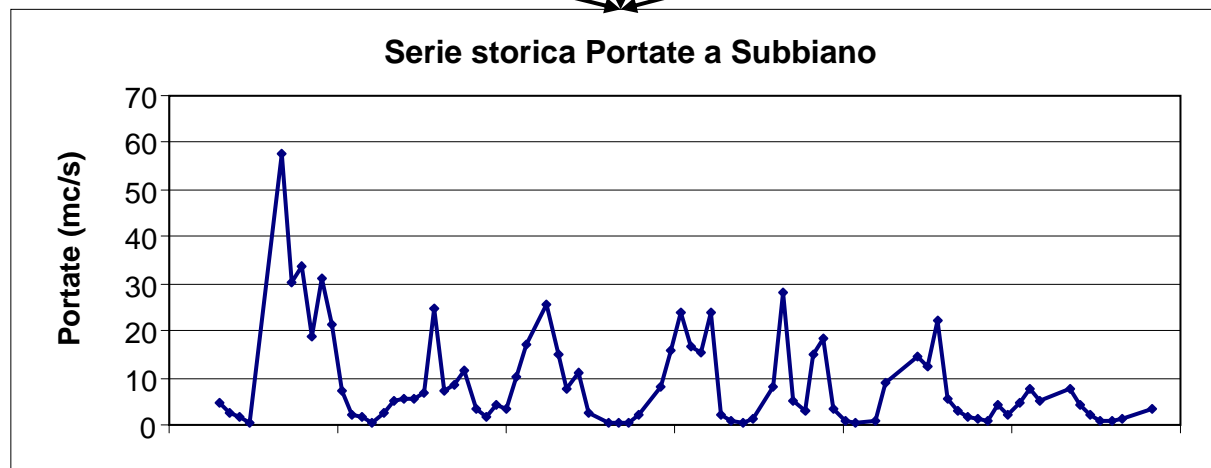
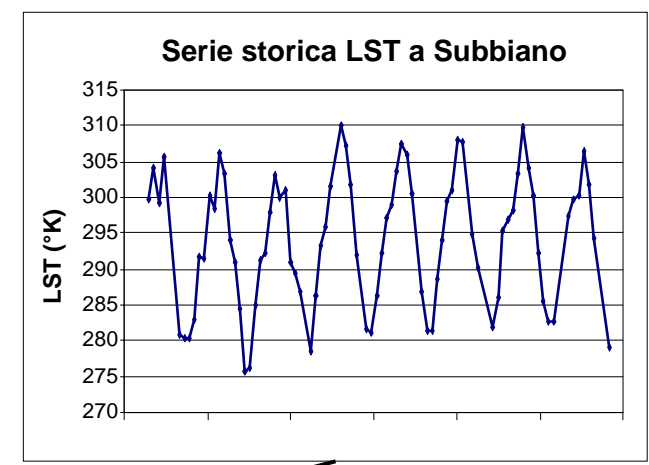
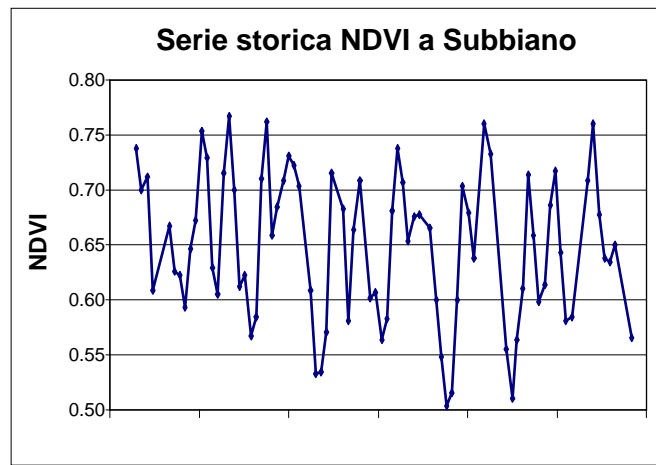
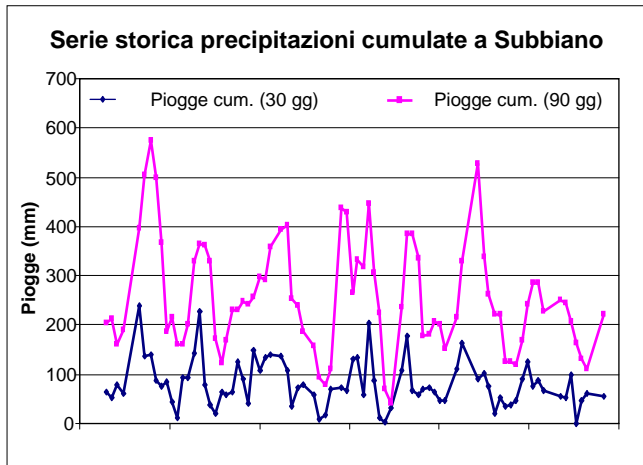


Parametri utilizzati

Antecedenti

- ☞ Stagionalità
 - Primavera, estate, autunno ed inverno
- ☞ Piogge cumulate negli ultimi 30 e 90 giorni
 - Serie storica dal 2000 al 2008 riferite alla stazione pluviometrica di Salutio
- ☞ NDVI (Normalized Difference Vegetation Index)
 - Parametro adimensionale telerilevato da MODIS che indica lo stato della copertura vegetale e perciò correlato all'umidità del suolo
- ☞ LST (Land Surface Temperature)
 - Altro parametro telerilevato da MODIS, definito come funzione lineare delle temperature di brillantezza nei canali a $11.03 \mu\text{m}$ e $12.02 \mu\text{m}$
- ☞ Portata alla stazione idrometrica di Subbiano
 - Dati forniti dall'Autorità di Bacino del fiume Arno

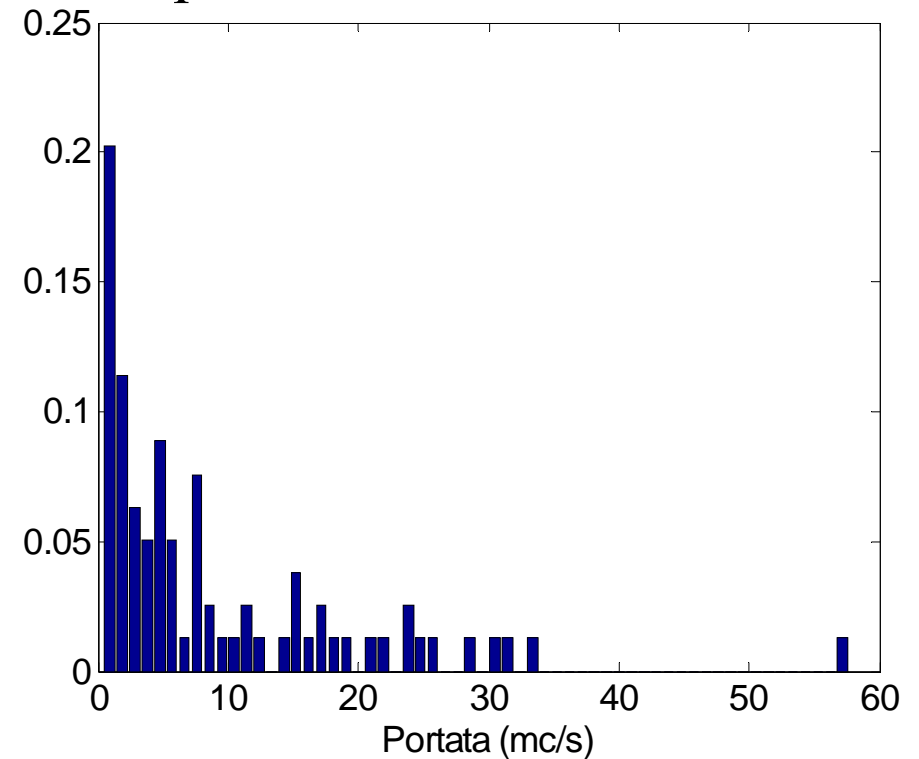
Serie storiche 2000 - 2008



Struttura dell'albero

☞ Per semplificarne la struttura si è scelto di raggruppare le uscite (portata) in un piccolo numero di classi “qualitative”

Classe	Portate (mc/s)
magra	$Q < 1.5$
bassa	$1.5 < Q < 4.5$
media	$4.5 < Q < 9$
alta	$9 < Q < 19$
piena	$Q > 19$



☞ Gli antecedenti sono misti

- La stagionalità è qualitativa (Primavera, Estate, Autunno, Inverno)
- Gli altri dati sono numerici (Piogge, NDVI, LST)

Struttura del file dati .arff

Nome della relazione @RELATION portate

Definizione degli attributi antecedenti

L'ultimo è il conseguente

Da qui iniziano i dati elencati nell'ordine in cui sono stati definiti gli attributi

```

@RELATION portate
@ATTRIBUTE stagione {inverno, primavera, estate, autunno}
@ATTRIBUTE piogge_cum_30 real
@ATTRIBUTE piogge_cum_90 real
@ATTRIBUTE NDVI real
@ATTRIBUTE Lst real
@ATTRIBUTE Portata {magra, bassa, media, alta, piena}

@DATA
primavera,64.5,203.5,0.74,299.8,bassa
primavera,51.8,212.0,0.70,304.1,bassa
estate,79.8,160.4,0.71,299.3,bassa
estate,61.7,189.5,0.61,305.7,magra
autunno,240.4,395.3,0.67,280.8,piena
.....
    
```

definizione delle classi

Pre-processing

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. The 'Current relation' is 'portate' with 6 attributes and 79 instances. The 'Selected attribute' is 'stagione', which is a nominal attribute with 4 distinct values and 0 missing values. The 'Attributes' list shows 'stagione' selected. The 'Class' is set to 'Portata (Nom)'. A bar chart visualizes the distribution of the 'stagione' attribute across the classes.

Current relation
Relation: portate
Instances: 79
Attributes: 6
Sum of weights: 79

Selected attribute
Name: stagione
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)


No.	Label	Count	Weight
1	inverno	19	19.0
2	primavera	22	22.0
3	estate	22	22.0
4	autunno	16	16.0

Attributes
All None Invert Pattern

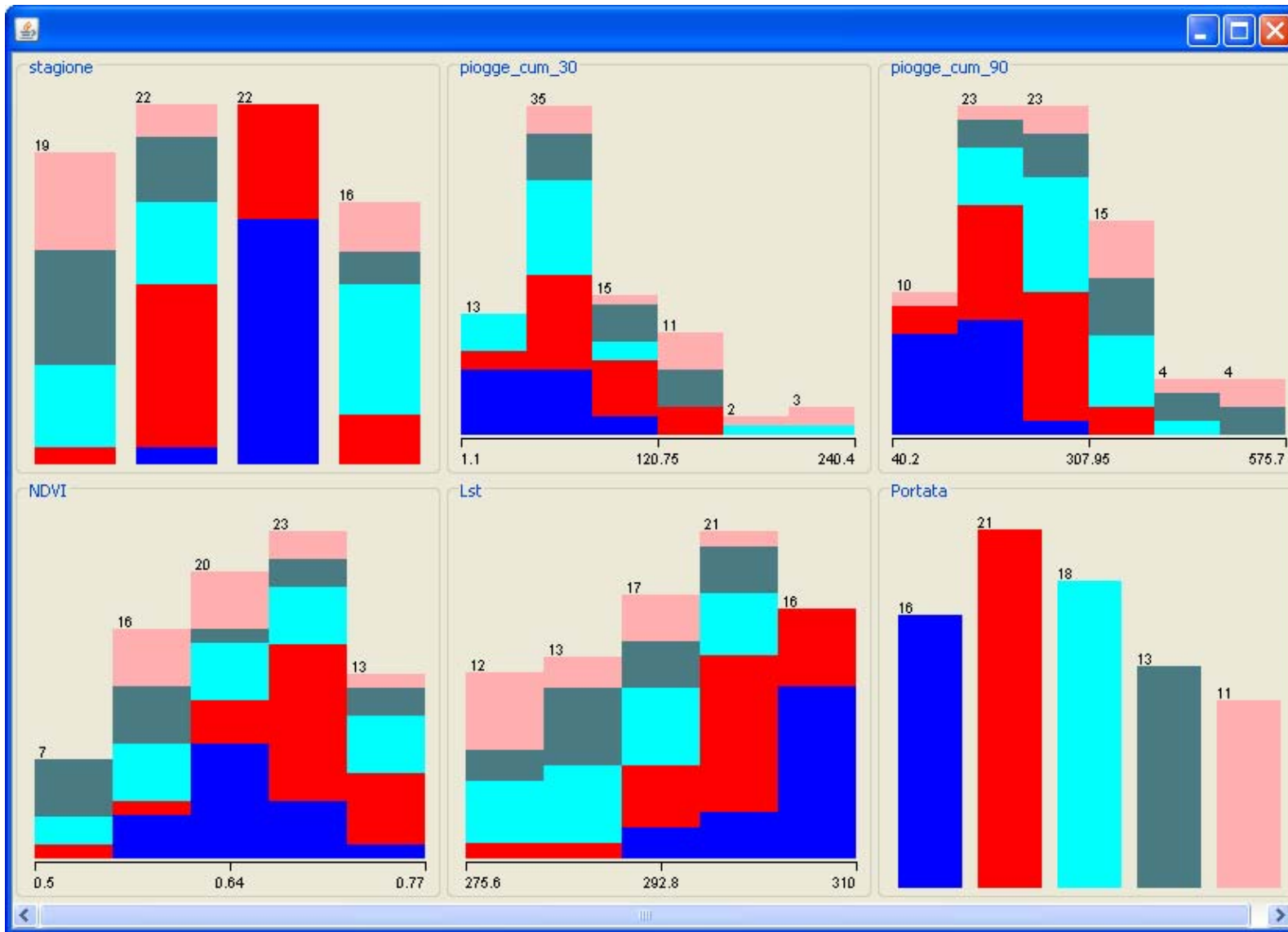
No.	Name
1	<input checked="" type="checkbox"/> stagione
2	<input type="checkbox"/> piogge_cum_30
3	<input type="checkbox"/> piogge_cum_90
4	<input type="checkbox"/> NDVI
5	<input type="checkbox"/> Lst
6	<input type="checkbox"/> Portata

Remove

Status
OK

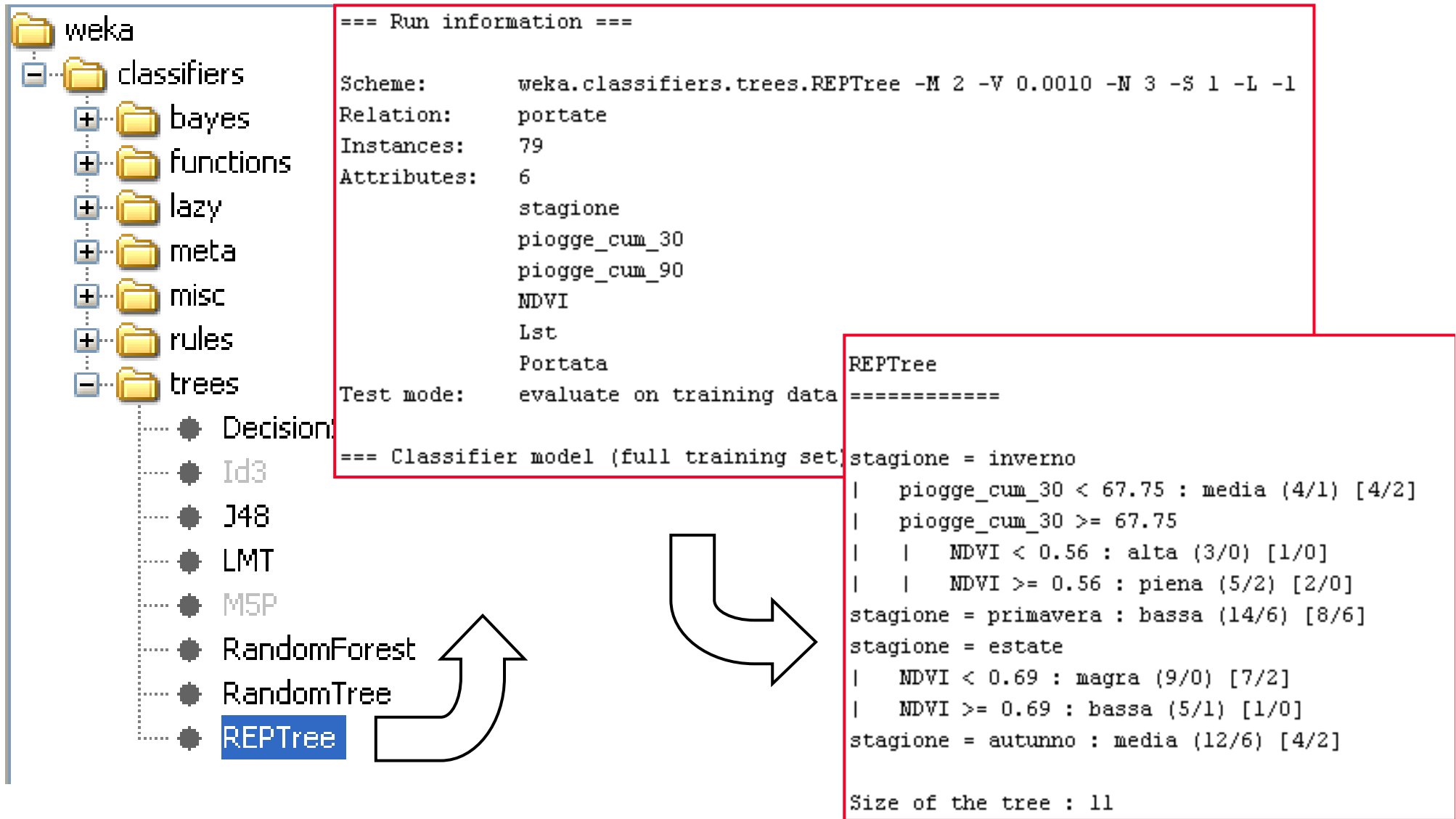
Log  x 0

Relazione di causalità di tutti gli antecedenti



i dati di tipo *real* sono suddivisi in 5 classi normalizzate nell'intervallo (0,1)

Classificazione con RepTree



The screenshot shows the Weka GUI with the 'trees' folder expanded and 'REPTree' selected. The 'Run information' window displays the following details:

```
=== Run information ===  
Scheme:      weka.classifiers.trees.REPTree -M 2 -V 0.0010 -N 3 -S 1 -L -1  
Relation:    portate  
Instances:   79  
Attributes:  6  
             stagione  
             piogge_cum_30  
             piogge_cum_90  
             NDVI  
             Lst  
             Portata  
Test mode:   evaluate on training data  
=== Classifier model (full training set) ===  
REPTree  
=====  
stagione = inverno  
|  piogge_cum_30 < 67.75 : media (4/1) [4/2]  
|  piogge_cum_30 >= 67.75  
|  |  NDVI < 0.56 : alta (3/0) [1/0]  
|  |  NDVI >= 0.56 : piena (5/2) [2/0]  
stagione = primavera : bassa (14/6) [8/6]  
stagione = estate  
|  NDVI < 0.69 : magra (9/0) [7/2]  
|  NDVI >= 0.69 : bassa (5/1) [1/0]  
stagione = autunno : media (12/6) [4/2]  
  
Size of the tree : 11
```

Valutazione dell'albero

Time taken to build model: 0 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	51
Incorrectly Classified Instances	28
Kappa statistic	0.5426
Mean absolute error	0.1876
Root mean squared error	0.3165
Relative absolute error	59.319 %
Root relative squared error	79.6343 %
Coverage of cases (0.95 level)	96.2025 %
Mean rel. region size (0.95 level)	61.7722 %
Total Number of Instances	79

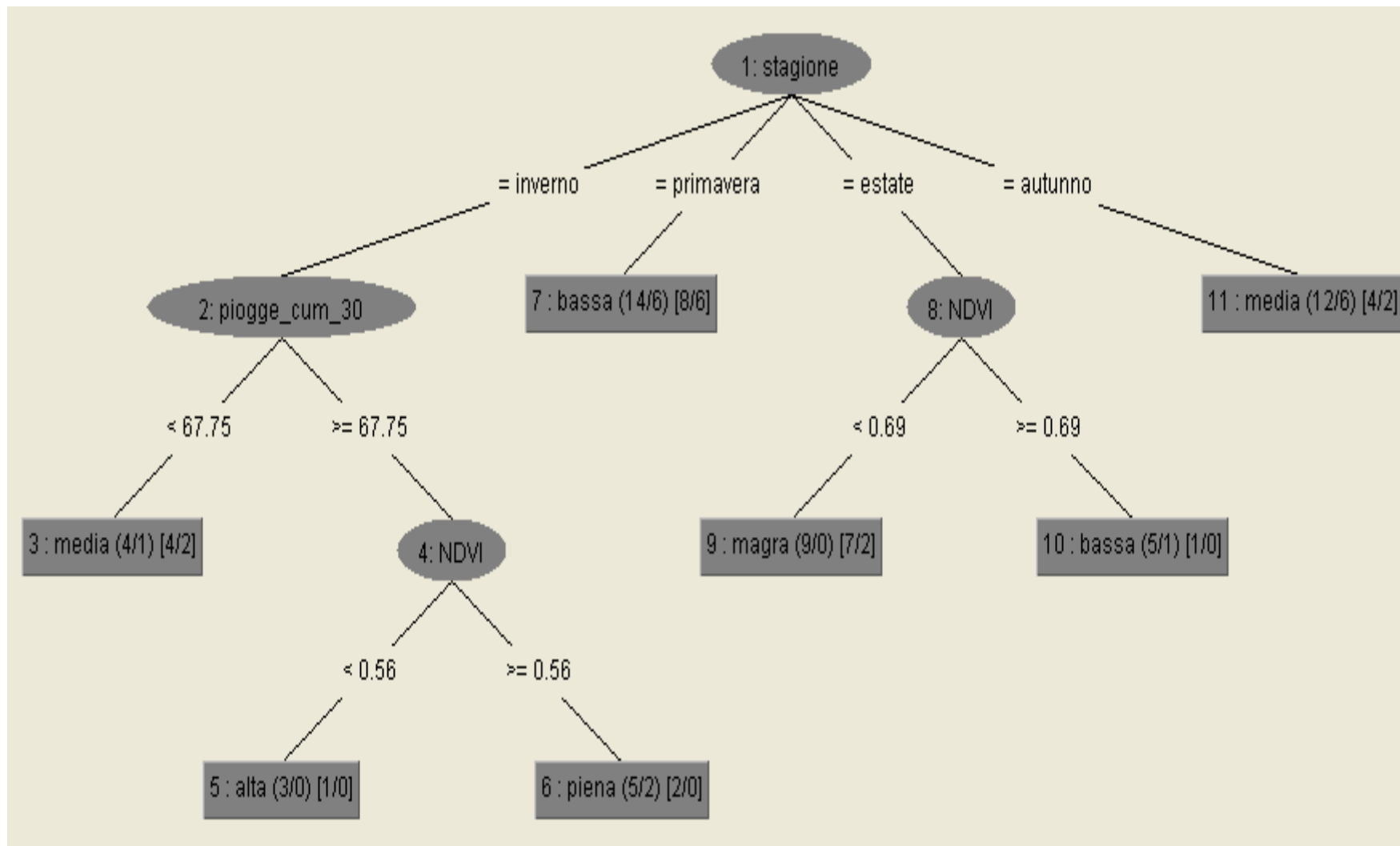
L'albero è stato ridotto
(Opzione "pruning" attiva)

64.557 %
35.443 %

=== Confusion Matrix ===

	a	b	c	d	e	<-- classified as
14	2	0	0	0	0	a = magra
2	15	4	0	0	0	b = bassa
0	5	13	0	0	0	c = media
0	4	3	4	2	0	d = alta
0	2	4	0	5	0	e = piena

Albero ridotto

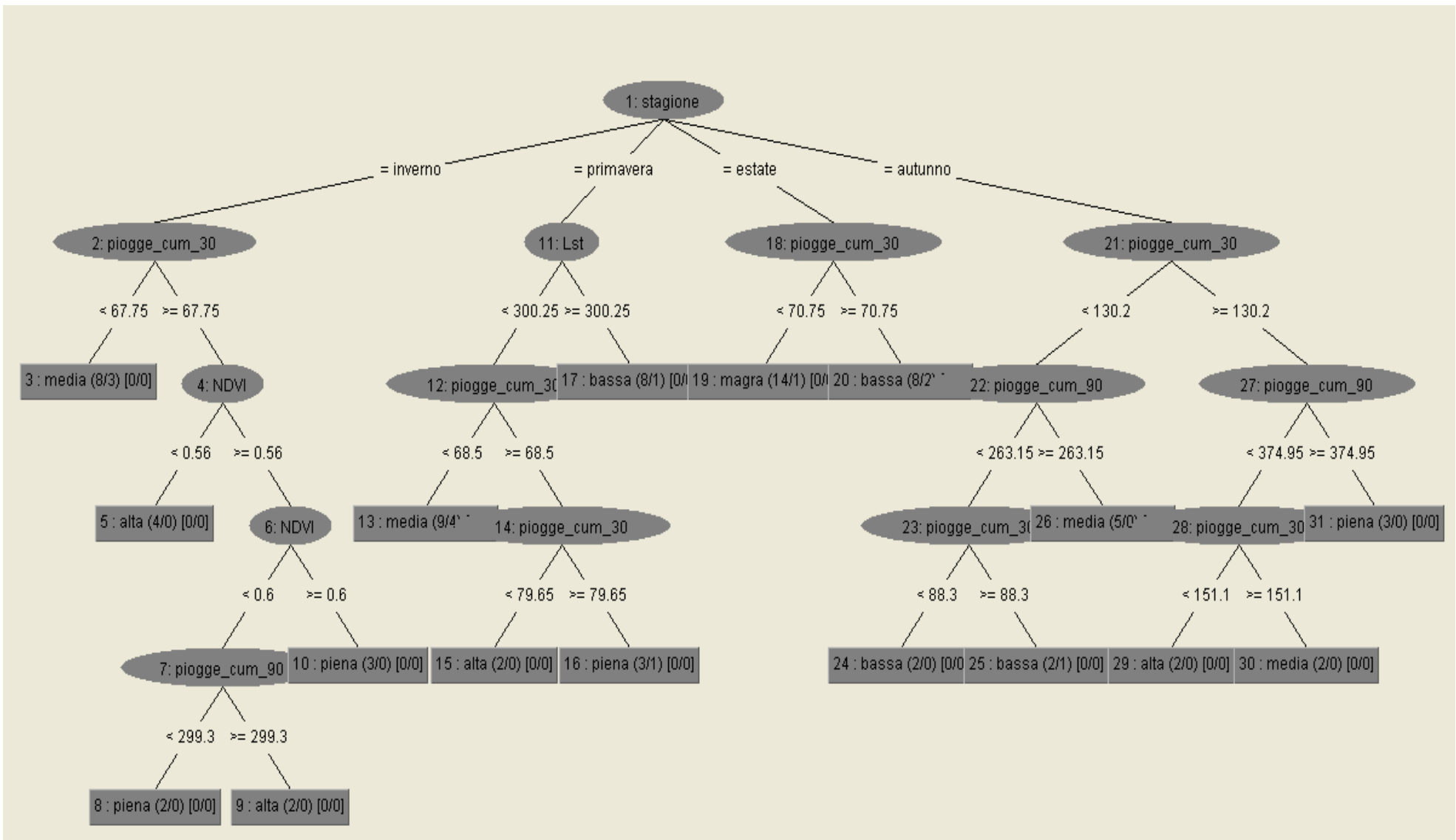


Albero completo (no Pruning)

```
stagione = inverno
|  piogge_cum_30 < 67.75 : media (8/3) [0/0]
|  piogge_cum_30 >= 67.75
|  |  NDVI < 0.56 : alta (4/0) [0/0]
|  |  NDVI >= 0.56
|  |  |  NDVI < 0.6
|  |  |  |  piogge_cum_90 < 299.3 : piena (2/0) [0/0]
|  |  |  |  piogge_cum_90 >= 299.3 : alta (2/0) [0/0]
|  |  |  NDVI >= 0.6 : piena (3/0) [0/0]
stagione = primavera
|  Lst < 300.25
|  |  piogge_cum_30 < 68.5 : media (9/4) [0/0]
|  |  piogge_cum_30 >= 68.5
|  |  |  piogge_cum_30 < 79.65 : alta (2/0) [0/0]
|  |  |  piogge_cum_30 >= 79.65 : piena (3/1) [0/0]
|  Lst >= 300.25 : bassa (8/1) [0/0]
stagione = estate
|  piogge_cum_30 < 70.75 : magra (14/1) [0/0]
|  piogge_cum_30 >= 70.75 : bassa (8/2) [0/0]
stagione = autunno
|  piogge_cum_30 < 130.2
|  |  piogge_cum_90 < 263.15
|  |  |  piogge_cum_30 < 88.3 : bassa (2/0) [0/0]
|  |  |  piogge_cum_30 >= 88.3 : bassa (2/1) [0/0]
|  |  piogge_cum_90 >= 263.15 : media (5/0) [0/0]
|  piogge_cum_30 >= 130.2
|  |  piogge_cum_90 < 374.95
|  |  |  piogge_cum_30 < 151.1 : alta (2/0) [0/0]
|  |  |  piogge_cum_30 >= 151.1 : media (2/0) [0/0]
|  |  piogge_cum_90 >= 374.95 : piena (3/0) [0/0]

Size of the tree : 31
```

Albero completo (no Pruning)



Valutazione dell'albero intero (no Pruning)

```
=== Evaluation on training set ===
```

```
=== Summary ===
```

```
Correctly Classified Instances      66
Incorrectly Classified Instances    13
Kappa statistic                     0.791
Mean absolute error                 0.0939
Root mean squared error            0.2167
Relative absolute error            29.7052 %
Root relative squared error        54.5219 %
Coverage of cases (0.95 level)     100 %
Mean rel. region size (0.95 level) 39.4937 %
Total Number of Instances          79
```

L'albero non è stato ridotto
(Opzione "pruning" disattivata)

83.5443 %

16.4557 %

```
=== Confusion Matrix ===
```

	a	b	c	d	e	<-- classified as
13	3	0	0	0	0	a = magra
1	16	4	0	0	0	b = bassa
0	1	17	0	0	0	c = media
0	0	2	10	1	0	d = alta
0	0	1	0	10	0	e = piena

Logica dell'albero

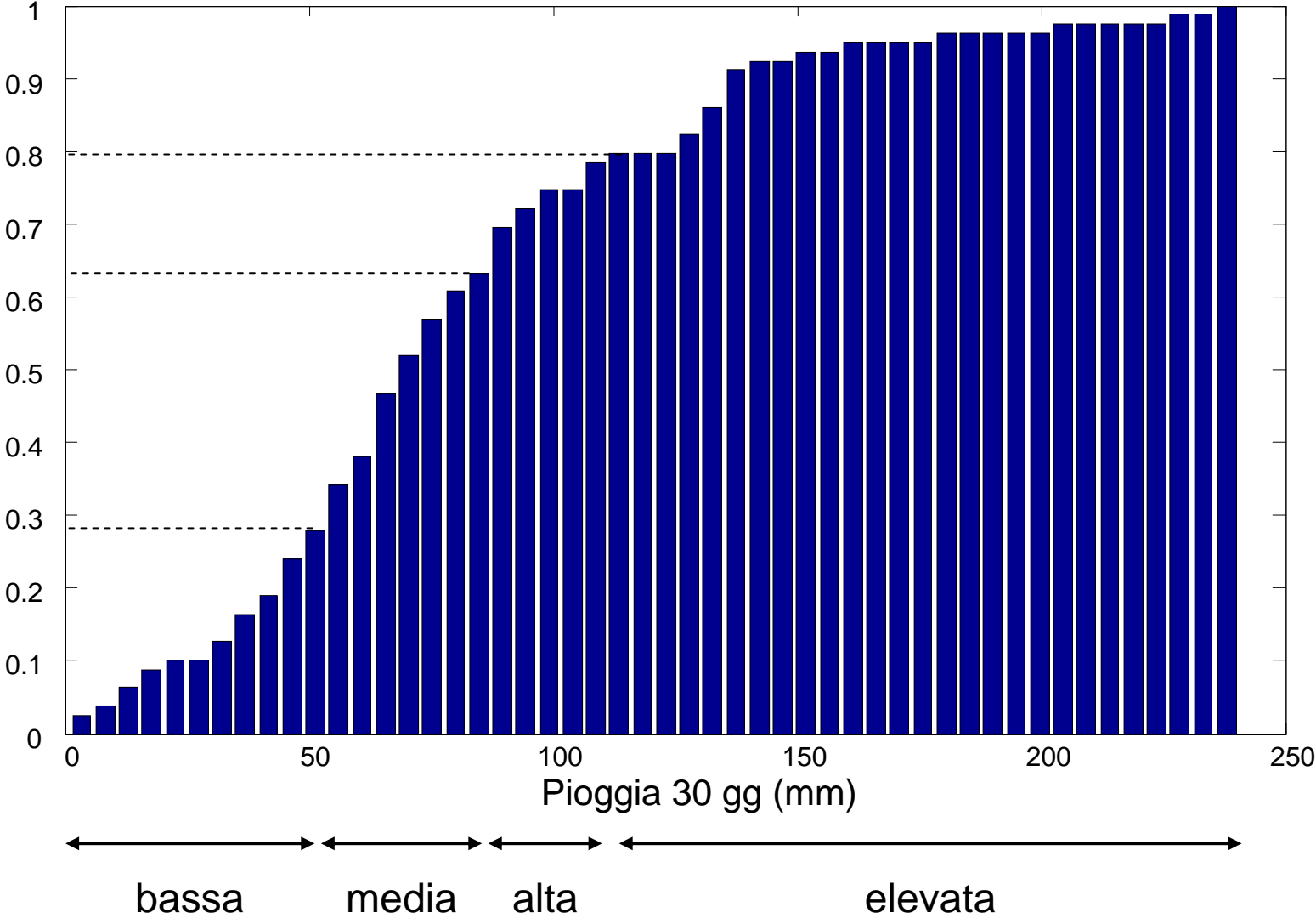
- ☞ WEKA individua la stagione come il primo nodo dell'albero
 - I secondi attributi sono LST e Piogge cumulate a 30 gg
- ☞ Questo risultato è plausibile perché la stagionalità e la LST sono due parametri molto correlati ed influiscono fortemente sulla portata
- ☞ In *estate* si hanno solo portate “magra” o “bassa” e non si supera mai il valore di 4.5 mc/s, limite superiore del qualificatore ‘bassa’
- ☞ In *autunno* per valori molto alti di piogge cumulate si possono avere valori di portata di piena, mentre per piovosità basse la portata si attesta su valori medi o bassi
- ☞ In *inverno*, invece, entra in gioco anche il parametro NDVI, e si hanno valori di portata ‘alta’ o ‘piena’
- ☞ In *primavera* per LST elevate la portata è bassa, mentre per valori bassi la portata in uscita dipende dalle piogge, e registra (per valori elevati di piogge) anche portate di piena

Versione “qualitativa” dell’albero

- ☞ L’albero precedente aveva *antecedenti misti (qualitativi e numerici) e conseguente nominale*
- ☞ Per estrarre delle regole decisionali è necessario che anche i dati in ingresso siano qualitativi, mantenendo invariata la classificazione di uscita.
- ☞ Si sono strutturati anche gli antecedenti in classi basandosi su una osservazione statistica degli istogrammi di ciascun parametro

Piogge_cum_30 (mm)	Classe	Piogge_cum_90 (mm)	Classe	NDVI	Classe	LST (°K)	Classe
P<50	basse	P<170	basse	N<0.61	basso	T < 286	bassa
50<P<70	medie	170<P<230	medie	0.61<N<0.7	medio	286<T<295	media
70<P<105	alte	230<P<330	alte	N>0.7	alto	295<T<301	alta
P>105	elevate	P>330	elevate			T>301	elevata

Distribuzione cumulativa delle piogge a 30 gg



Valutazione dell'albero qualitativo (no Pruning)

```
=== Evaluation on training set ===
```

```
=== Summary ===
```

Correctly Classified Instances	60	75.9494 %
Incorrectly Classified Instances	19	24.0506 %
Kappa statistic	0.6948	
Mean absolute error	0.1165	
Root mean squared error	0.2414	
Relative absolute error	36.8011 %	
Root relative squared error	60.6827 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	36.962 %	
Total Number of Instances	79	

```
=== Confusion Matrix ===
```

	a	b	c	d	e	<-- classified as
14	2	0	0	0	0	a = magra
2	16	1	2	0	0	b = bassa
0	3	12	2	0	0	c = media
0	2	1	11	0	0	d = alta
0	0	1	3	7	0	e = piena

Stima delle uscite

- ☞ Dall'albero così individuato si possono stimare i valori di portata in uscita, una volta assegnati i dati in ingresso.
- ☞ Per fare questo si usa la *cross-validation*
- ☞ Questo metodo suddivide i dati della serie storica in un numero di cartelle (*folds*) specificato
- ☞ Una frazione dei dati di una cartella viene utilizzata per la validazione del modello.
- ☞ Il procedimento viene ripetuto per tutte le cartelle, ottenendo così una previsione su tutti i singoli valori della serie storica.
- ☞ I dati sono stati suddivisi in 10 cartelle (default)
- ☞ Nel menu *more options* si imposta il formato di uscita dell'*Output Prediction* come file di testo (*Plain Text*)

Versioni dello stimatore

- 👉 Tutti i dati in forma numerica → previsione portata numerica
- 👉 Dati in ingresso numerici → previsione portata nominale
- 👉 Tutti i dati della serie in forma nominale

➡ Questa versione (albero tutto nominale) dà i migliori risultati

