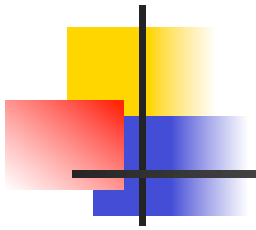


Retrieving and Working with Datasets

Prof. Pietro Ducange

Where to retrieve interesting datasets

- UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets.html

- Keel Dataset Repository

http://sci2s.ugr.es/keel/datasets.php

- WEKA

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

- ChemDB

  http://www.cs.ox.ac.uk/activities/machinelearning/applications.html

Where to retrieve interesting datasets: challenges

- Kaggle

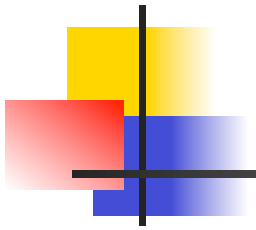
https://www.kaggle.com/datasets

- TIM big data Challenge

https://dandelion.eu/datamine/open-big-data/

- TunedIT

http://tunedit.org/challenges/



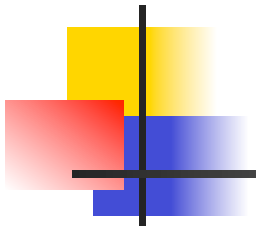
PIMA Indian Diabetes

- From the UCI repository
- The class attribute specifiy wheather patient shows or not signs of diabetes according to World Health Organization criteria
- 2 classes, 8 attributes, 768 instances, 500 (65.1%) negative , and 268 (34.9%) positive tests for diabetes
- All patients were females at least 21 years old of Pima Indian heritage

PIMA Indian Diabetes

Attributes:

1. Number of times pregnant
2. Plasma glucose concentration
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function (a sort of ancestor's history)
8. Age (years)



Exercise

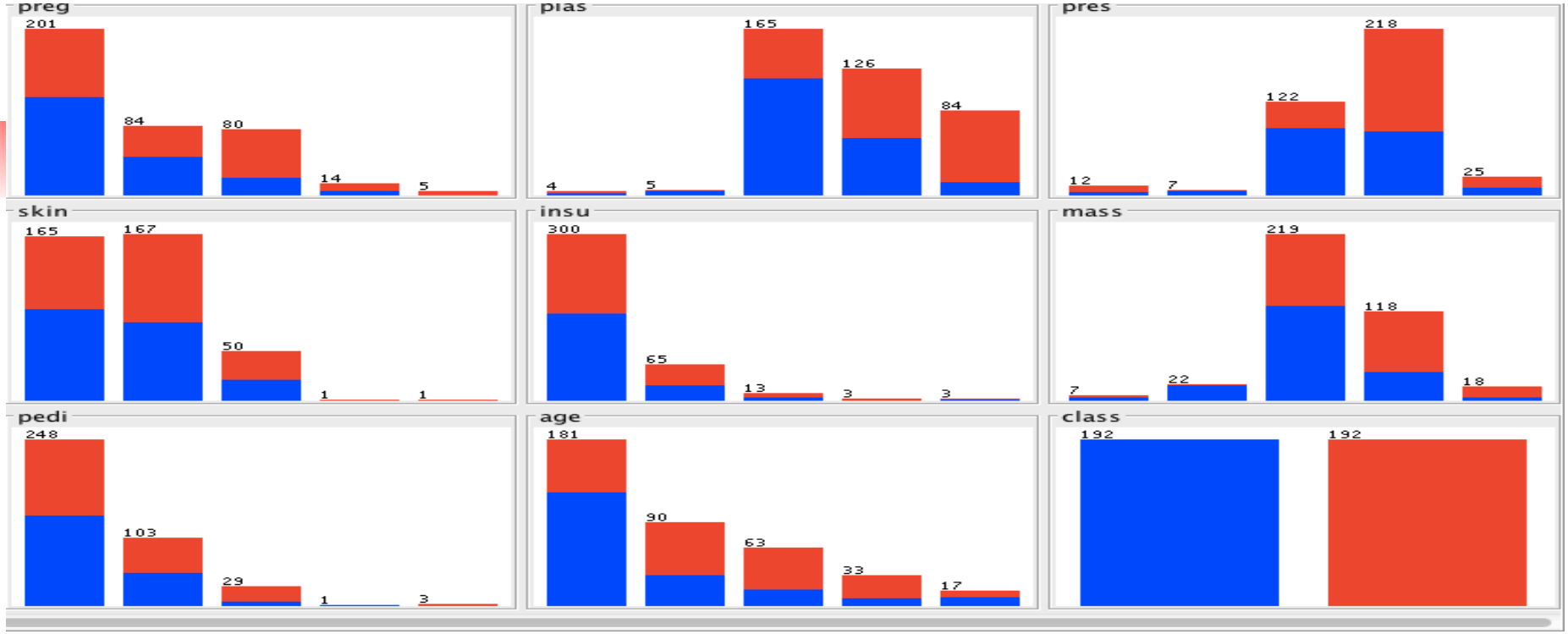
1. Load the diabetes.arff dataset

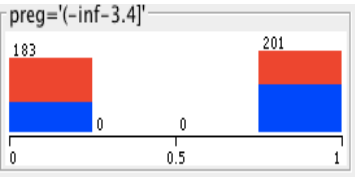
2. Apply a supervised resampling considering just 50% of the original instances (with default parameters)

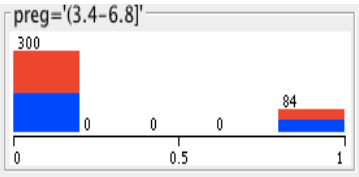
3. Apply a supervised resampling considering just 50% of the original instances (with the first parameter set to 1), save the dataset and analyze the differences with the results of 2)

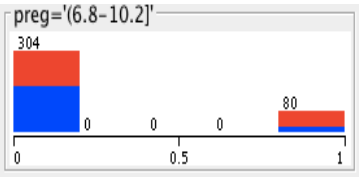
4. Discretize the saved dataset by using 5 bins (unsupervised filter)

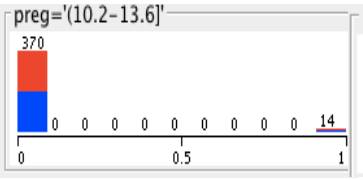
5. Apply the NominalToBinary Filter to the discretized dataset and comment the results

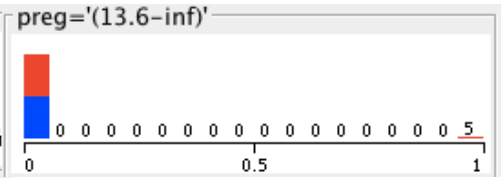

preg
201
84
80
14
5
plas
4
5
165
126
84
pres
12
7
122
218
25
skin
165
167
50
1
1
insu
300
65
13
3
3
mass
7
22
219
118
18
pedi
248
103
29
1
3
age
181
90
63
33
17
class
192
192


preg='(-inf-3.4]'
183
0
0
201
0
0.5
1


preg='(3.4-6.8]'
300
0
0
0
84
0
0.5
1


preg='(6.8-10.2]'
304
0
0
0
80
0
0.5
1


preg='(10.2-13.6]'
370
0 0 0 0 0 0 0 0 0 14
0
0.5
1


preg='(13.6-inf)'
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5
0
0.5
1

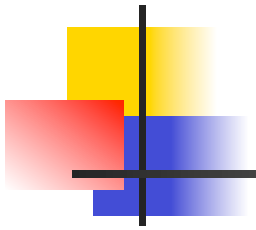
Hepatitis Data Set

Type Classification
Origin Real world **Features** 19
(Real / Integer / Nominal) (2 / 17 / 0)
Classes 2 **Missing values?** Yes
Total instances 155

UCI URL: https://archive.ics.uci.edu/ml/datasets/Hepatitis

The task is to predict if patients will die (1) or survive (2).

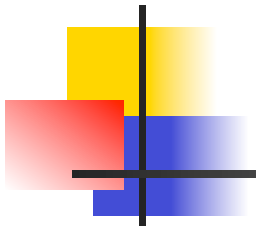
Class Distribution: DIE: 32 LIVE: 123



Hepatitis Data Set

Attribute Information:

1. Class: DIE, LIVE
2. AGE: 10, 20, 30, 40, 50, 60, 70, 80
3. SEX: male, female
4. STEROID: no, yes
5. ANTIVIRALS: no, yes
6. FATIGUE: no, yes
7. MALAISE: no, yes
8. ANOREXIA: no, yes
9. LIVER BIG: no, yes
10. LIVER FIRM: no, yes
11. SPLEEN PALPABLE: no, yes
12. SPIDERS: no, yes
13. ASCITES: no, yes
14. VARICES: no, yes
15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
17. SGOT: 13, 100, 200, 300, 400, 500,
18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
20. HISTOLOGY: no, yes



Exercise

- Download the rough Hepatitis dataset from UCI
- Prepare it for loading in WEKA (if needed)
- Load the dataset in WEKA
- Handle missing values
- Rebalance the dataset with the resampling filter

Some indications for relabeling Hepatitis dataset

Relabeled values in attribute SEX
% From: 2 To: male
% From: 1 To: female
% Relabeled values in attribute STEROID
% From: 1 To: no
% From: 2 To: yes
% Relabeled values in attribute ANTIVIRALS
% From: 2 To: no
% From: 1 To: yes
% Relabeled values in attribute FATIGUE
% From: 2 To: no
% From: 1 To: yes
% Relabeled values in attribute MALAISE
% From: 2 To: no
% From: 1 To: yes
% Relabeled values in attribute ANOREXIA
% From: 2 To: no
% From: 1 To: yes

% Relabeled values in attribute LIVER_BIG
% From: 1 To: no
% From: 2 To: yes
% Relabeled values in attribute LIVER_FIRM
% From: 2 To: no
% From: 1 To: yes
% Relabeled values in attribute SPLEEN_PALPABLE
% From: 2 To: no
% From: 1 To: yes
% Relabeled values in attribute SPIDERS
% From: 2 To: no
% From: 1 To: yes
% Relabeled values in attribute ASCITES
% From: 2 To: no
% From: 1 To: yes
% Relabeled values in attribute VARICES
% From: 2 To: no
% From: 1 To: yes
% Relabeled values in attribute HISTOLOGY
% From: 1 To: no
% From: 2 To: yes

Rought CSV file (1)

2,30,2,1,2,2,2,2,1,2,2,2,2,2,1.00,85,18,4.0,?,1
....
2,50,1,1,2,1,2,2,1,2,2,2,2,2,0.90,135,42,3.5,?,1
2,31,1,?,1,2,2,2,2,2,2,2,2,2,0.70,46,52,4.0,80,1
2,34,1,2,2,2,2,2,2,2,2,2,2,2,1.00,?,200,4.0,?,1
2,34,1,2,2,2,2,2,2,2,2,2,2,2,0.90,95,28,4.0,75,1
1,51,1,1,2,1,2,1,2,2,1,1,2,2,?,?,?,?,?,1
....

To load in WEKA the dataset, just open the file paste the string of the name of each attribute, separeted by a comma, as the first line of the file and save it as a .CSV file->

Rought CSV file (2)

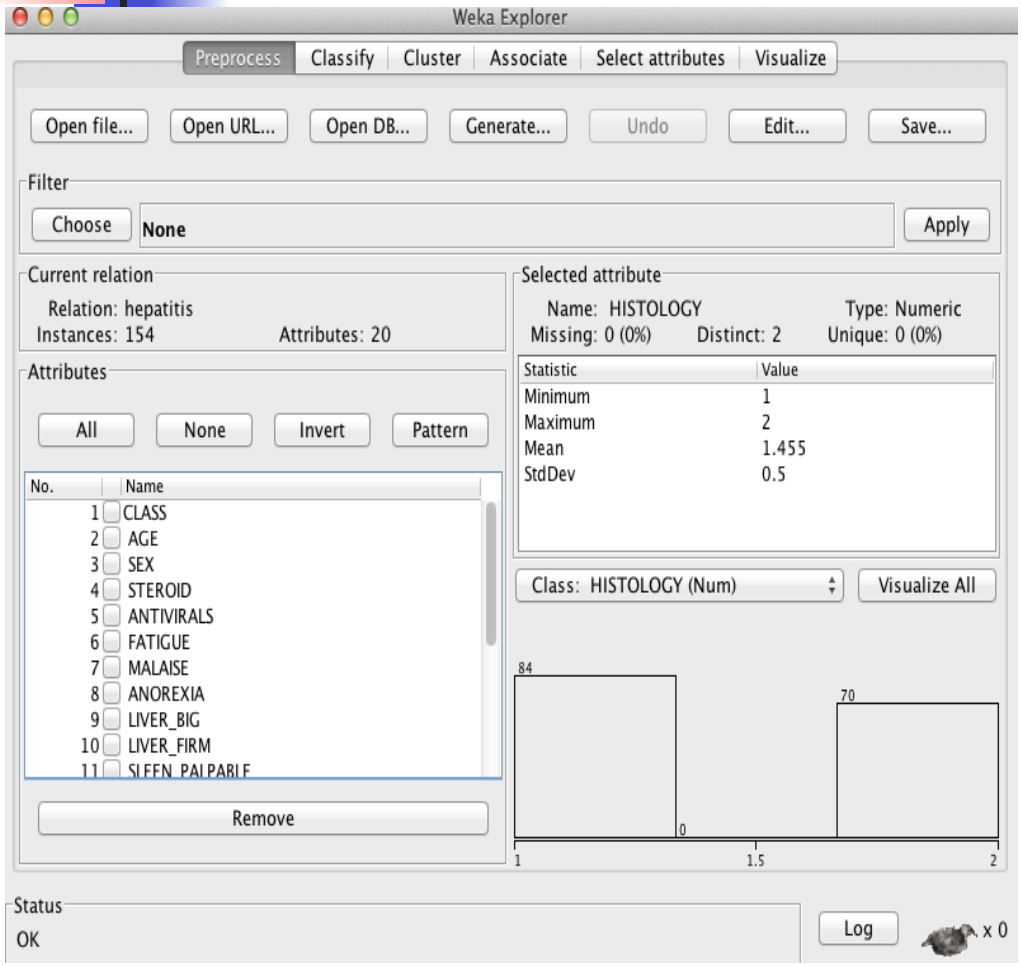
CLASS, AGE, SEX, STEROID, ANTIVIRALS, FATIGUE, MALAISE, ANOREXIA, LIVER_BIG, LIVER_FIRM, SLEEN_PALPABLE, SPIDERS, ASCITES, VARICES, BILIRUBIN, ALK_PHOSPHATE, SGOT, ALBUMIN, PROTIME, HISTOLOGY

2,30,2,1,2,2,2,2,1,2,2,2,2,2,1.00,85,18,4.0,?,1

....

2,50,1,1,2,1,2,2,1,2,2,2,2,2,0.90,135,42,3.5,?,1
2,31,1,?,1,2,2,2,2,2,2,2,2,2,0.70,46,52,4.0,80,1
2,34,1,2,2,2,2,2,2,2,2,2,2,2,1.00,?,200,4.0,?,1
2,34,1,2,2,2,2,2,2,2,2,2,2,2,0.90,95,28,4.0,75,1
1,51,1,1,2,1,2,1,2,2,1,1,2,2,?,?,?,?,?,1

Loading the CSV file for the dataset in WEKA



Open the dataset editing (edit button) and set the CLASS attribute as a class.

Using the save button, you can save the dataset in .arff format.

If you open the arff file with a text editor you will find:

@relation hepatitis

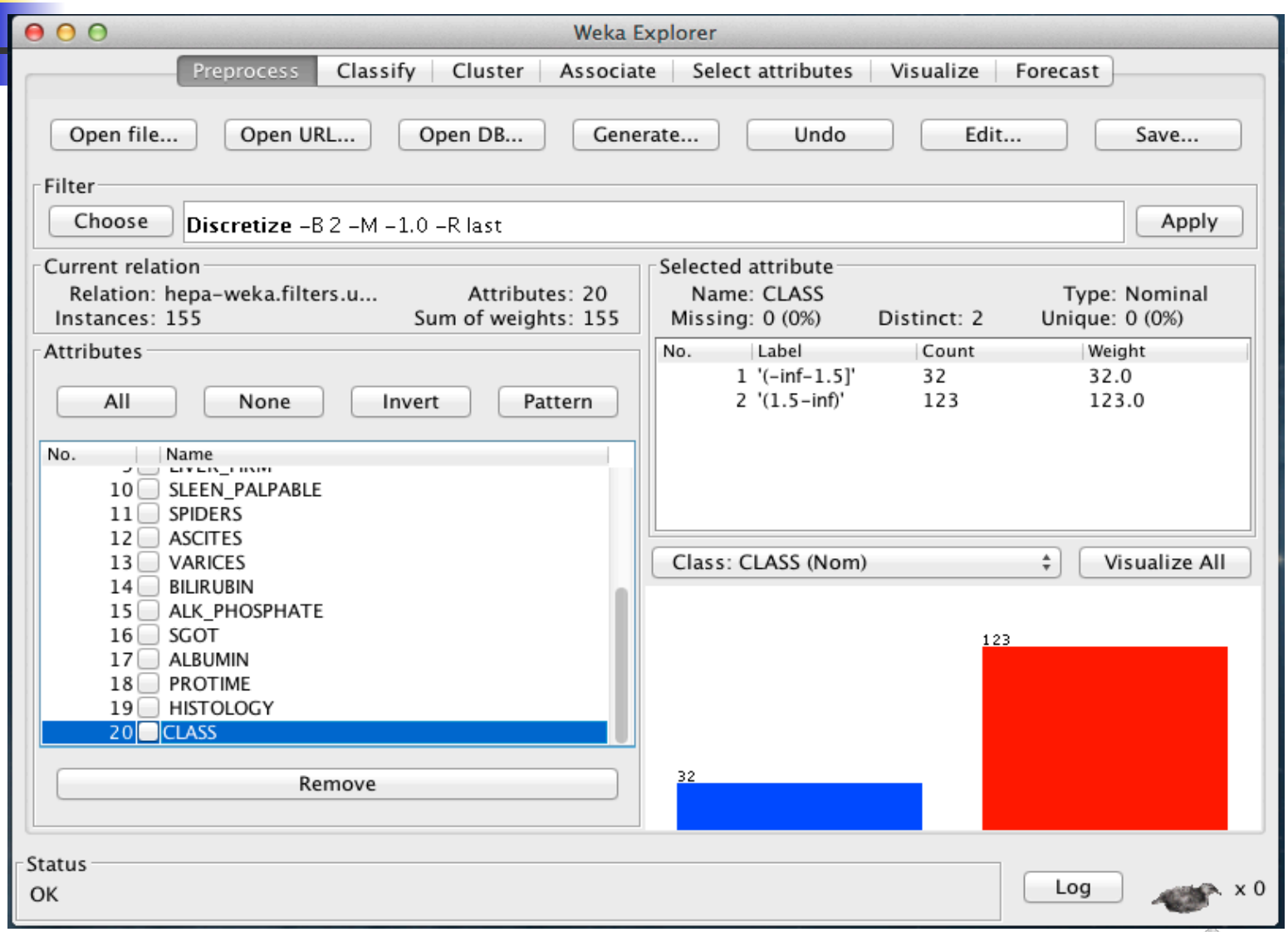
@attribute ' AGE' numeric
@attribute ' SEX' numeric
@attribute ' STEROID' numeric
@attribute ' ANTIVIRALS' numeric
@attribute ' FATIGUE' numeric
@attribute ' MALAISE' numeric
@attribute ' ANOREXIA' numeric
@attribute ' LIVER_BIG' numeric
@attribute ' LIVER_FIRM' numeric
@attribute ' SLEEN_PALPABLE' numeric
@attribute ' SPIDERS' numeric
@attribute ' ASCITES' numeric
@attribute ' VARICES' numeric
@attribute ' BILIRUBIN' numeric
@attribute ' ALK_PHOSPHATE' numeric
@attribute ' SGOT' numeric
@attribute ' ALBUMIN' numeric
@attribute ' PROTIME' numeric
@attribute ' HISTOLOGY' numeric
@attribute CLASS numeric

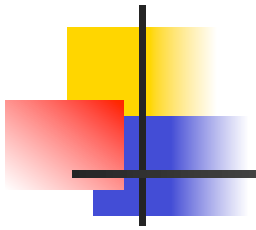
You can relabel by hand the arff file (pay attention to the class attribute).

@relation hepatitis

@attribute ' AGE' numeric
@attribute ' SEX' numeric
@attribute ' STEROID' numeric
@attribute ' ANTIVIRALS' numeric
@attribute ' FATIGUE' numeric
@attribute ' MALAISE' numeric
@attribute ' ANOREXIA' numeric
@attribute ' LIVER_BIG' numeric
@attribute ' LIVER_FIRM' numeric
@attribute ' SLEEN_PALPABLE' numeric
@attribute ' SPIDERS' numeric
@attribute ' ASCITES' numeric
@attribute ' VARICES' numeric
@attribute ' BILIRUBIN' numeric
@attribute ' ALK_PHOSPHATE' numeric
@attribute ' SGOT' numeric
@attribute ' ALBUMIN' numeric
@attribute ' PROTIME' numeric
@attribute ' HISTOLOGY' numeric
@attribute 'class' { 1, 2}

If you load the new arff dataset in WEKA





After applying the ReplaceMissingValues filter (weka.filters.unsupervised.attribute.ReplaceMissingValues), you can apply the supervised resample filter with 100% and the first parameter set to 1

The rebalanced Hepatitis Dataset in WEKA