



Data Warehouse and Data Mining

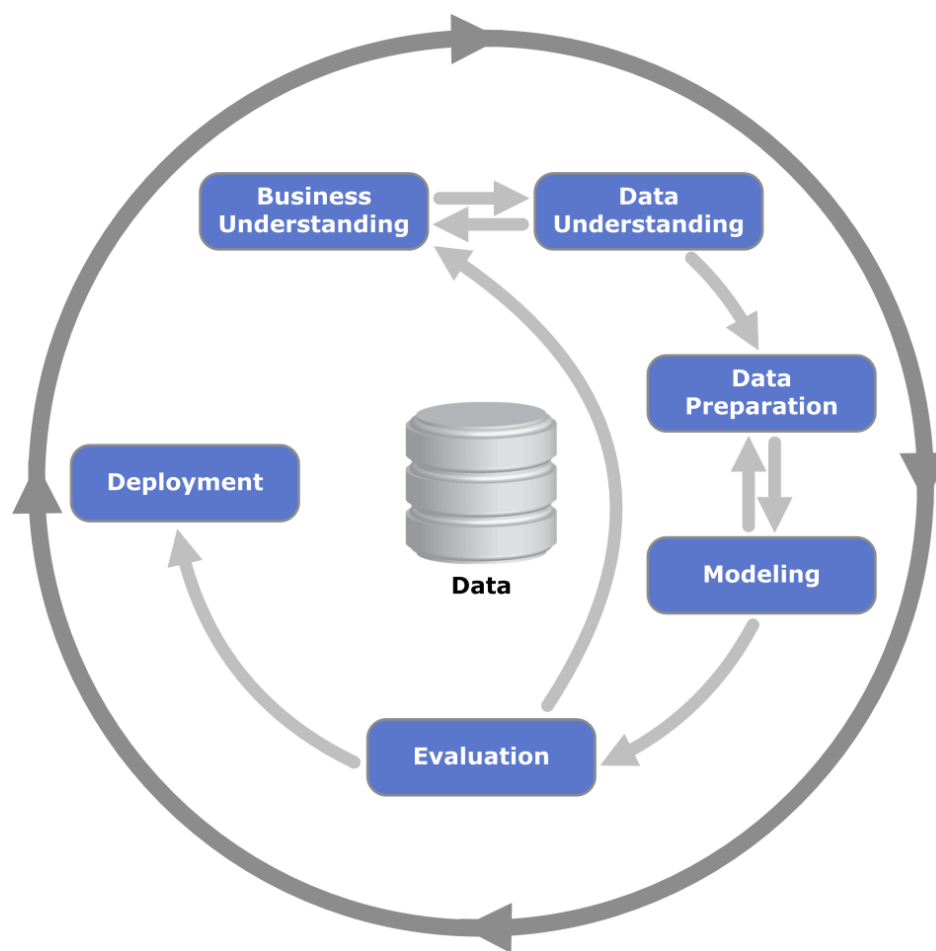
Module II – Data Mining

Data Preparation

Ph.D. Ettore Ritacco



The Knowledge Discovery Process (CRISP-DM)





Business Understanding

- You are a doctor of a medical division, who collected some data from your patients
- All of them contracted the same disease
- The therapy consists in 5 different and exclusive cures, according to the patient condition
- Does an automatic procedure, for the cure assignment, exist?



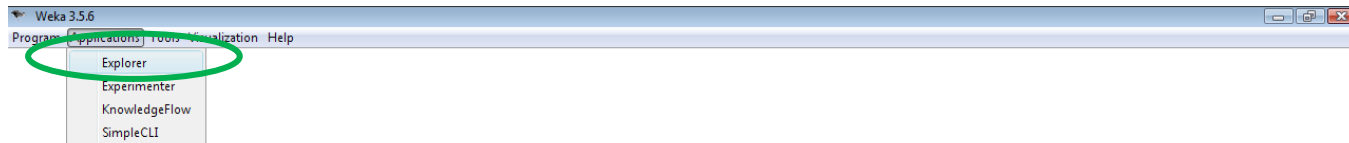
Data Understanding

Attribute	Description
Instance_number	Incremental tuple ID (INTEGER)
ID	Patient's ID (INTEGER)
Age	Patient's age (INTEGER)
Sex	Patient's gender: F or M
BP	Blood Pressure: HIGH, NORMAL or LOW
Cholesterol	Concentration of cholesterol in the blood: NORMAL or HIGH
Na	Concentration of sodium in the blood (REAL)
K	Concentration of potassium in the blood (REAL)
Drug	The chosen cure: drugY, drugC, drugX, drugA, drugB



Data Understanding

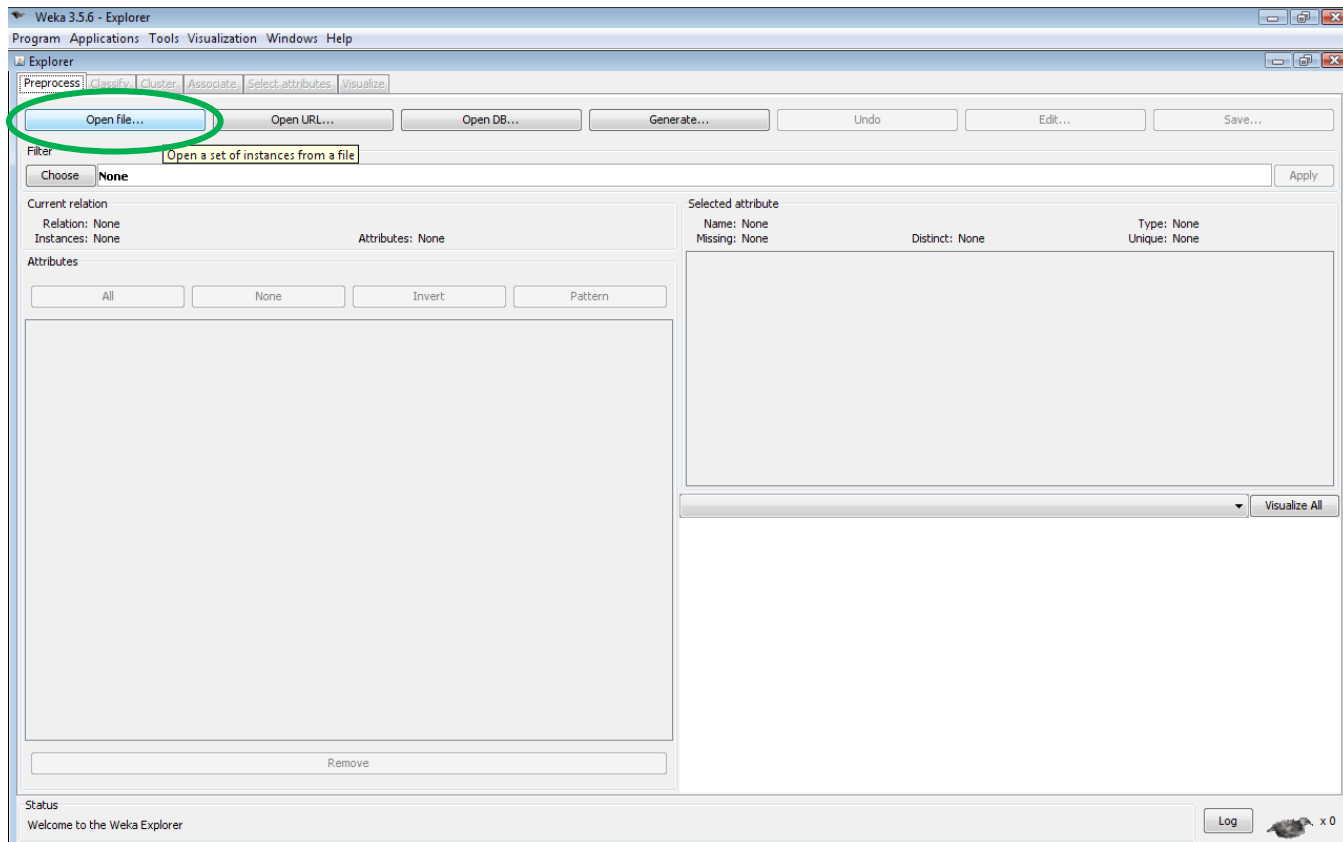
- Data acquisition in Weka:





Data Understanding

- Data acquisition in Weka:



Data Understanding

- Data acquisition in Weka:

The screenshot displays the Weka 3.5.6 Explorer window. The interface includes a menu bar (Program, Applications, Tools, Visualization, Windows, Help) and a toolbar with buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Generate...', 'Undo', 'Edit...', and 'Save...'. The 'Filter' section shows 'Choose None' and an 'Apply' button. The 'Current relation' section indicates 'Relation: DRUG1n' and 'Instances: 200'. The 'Attributes' section lists seven attributes: Age, Sex, BP, Cholesterol, Na, K, and Drug. The 'Selected attribute' section shows 'Name: Age', 'Missing: 0 (0%)', 'Distinct: 57', and 'Type: Numeric'. A table of statistics for 'Age' is shown below:

Statistic	Value
Minimum	15
Maximum	74
Mean	44.315
StdDev	16.544

The 'Class: Drug (Nom)' section shows a visualization of the data distribution. The x-axis represents the 'Age' attribute, ranging from 15 to 74. The y-axis represents the frequency of instances. The visualization shows a distribution of instances across different age groups, with a peak around age 36. A tooltip for the peak shows the value 36 and the range (34.667, 44.5].

Status: OK

Data Understanding

- Target attribute (the class attribute): Drug

The screenshot shows the Weka 3.5.6 Explorer interface. The 'Attributes' list on the left has 'Drug' selected and highlighted with a green oval. The 'Selected attribute' panel on the right shows the following data:

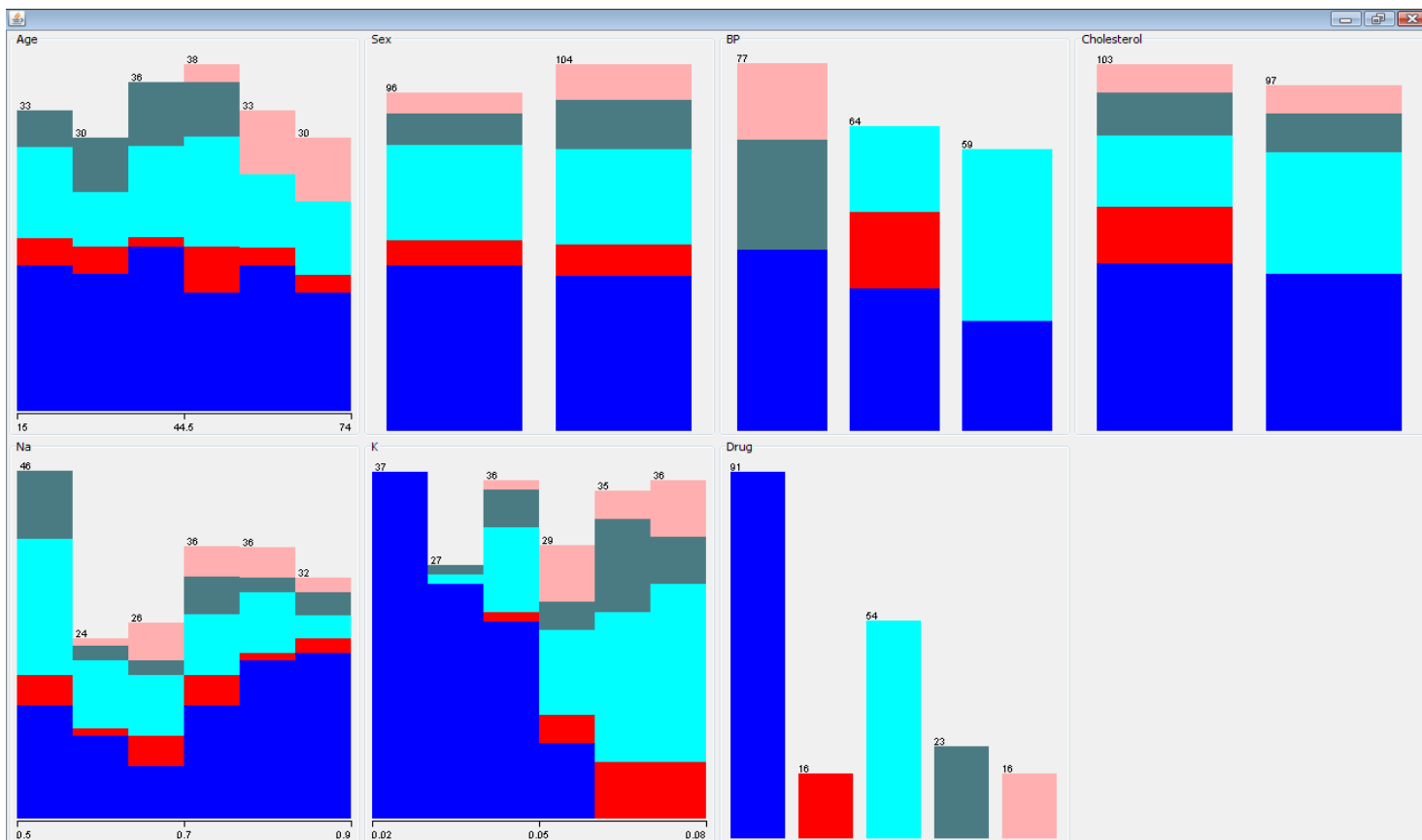
Label	Count
drugY	91
drugC	16
drugX	54
drugA	23
drugB	16

Below the table, a bar chart visualizes the distribution of the 'Drug' class. The bars are colored blue (91), red (16), cyan (54), dark green (23), and pink (16). A green arrow points to the bar chart area.



Data Understanding

- Distributions: comments?





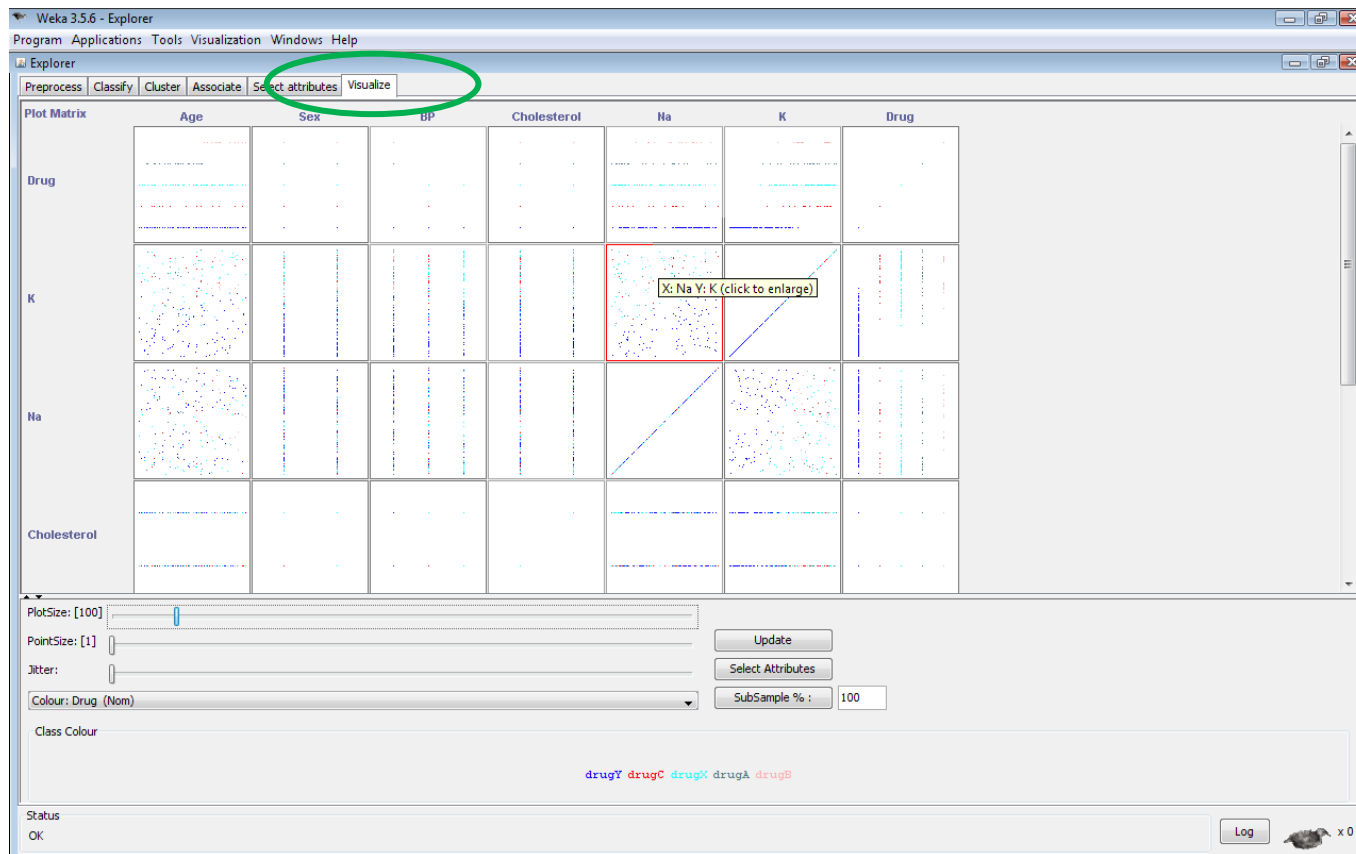
Data Understanding

- When age is greater than (about) 44 years, the class drugB starts to appear
- The sex and the concentration of sodium seem to not affect the class distribution
- Some cures are provided according to the values of blood pressure and cholesterol
- When the concentration of potassium is lower than (about) 0.05 the dominant class is drugY
- Data are unbalanced:
 - There is one dominant class: drugY (the blue one)



Data Understanding

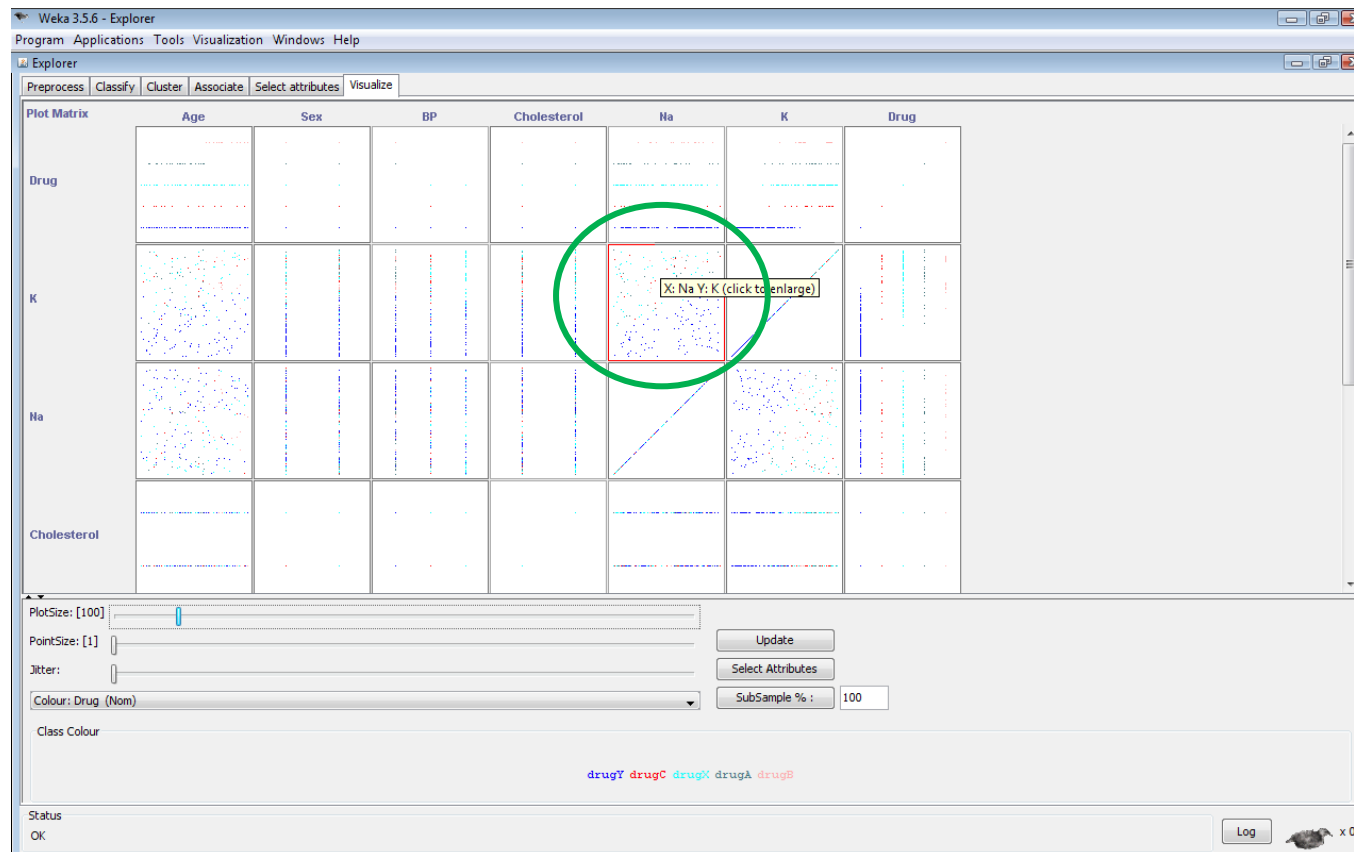
- Scatter plot





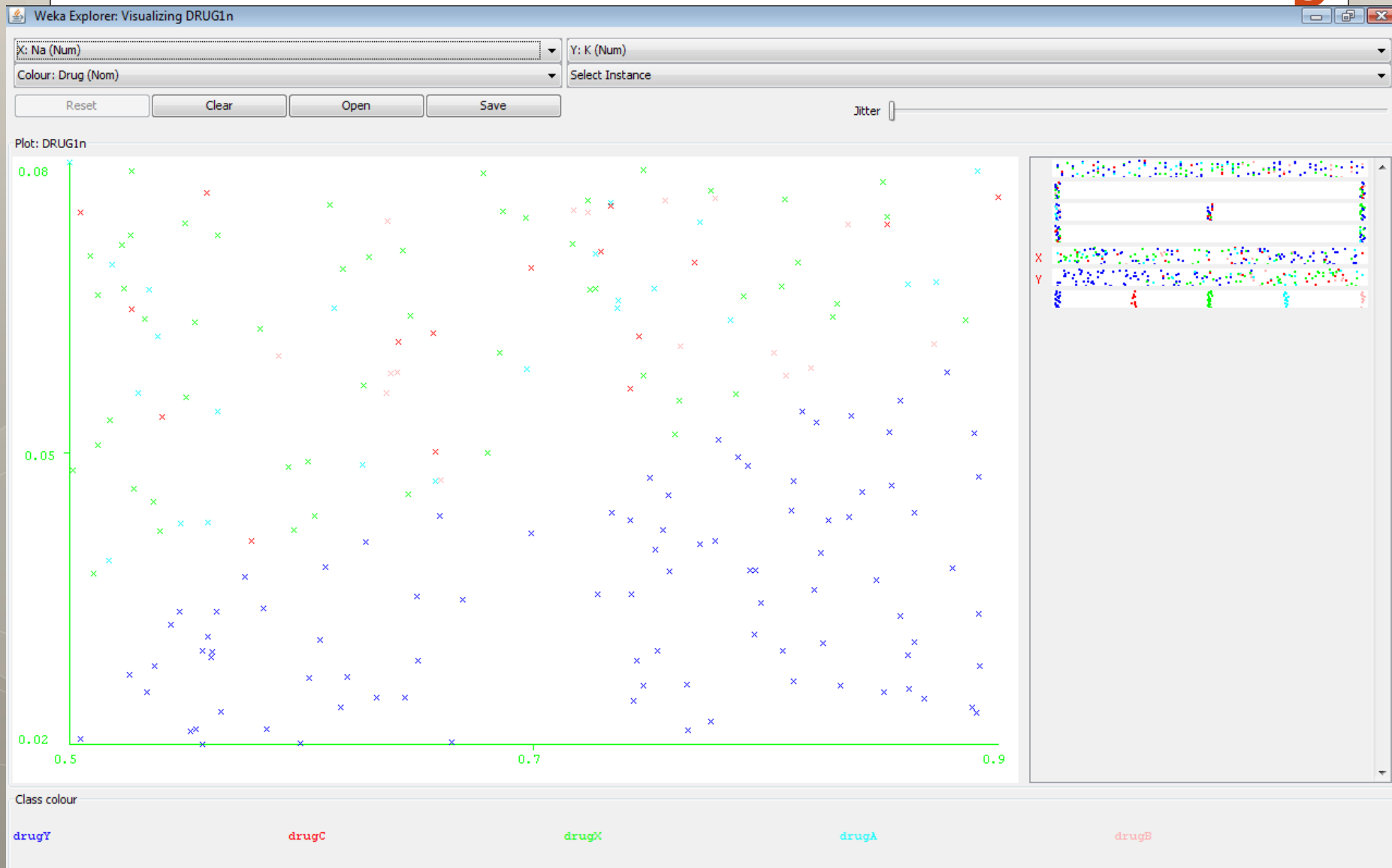
Data Understanding

Scatter plot





Data Understanding





Data Understanding

- Comments?

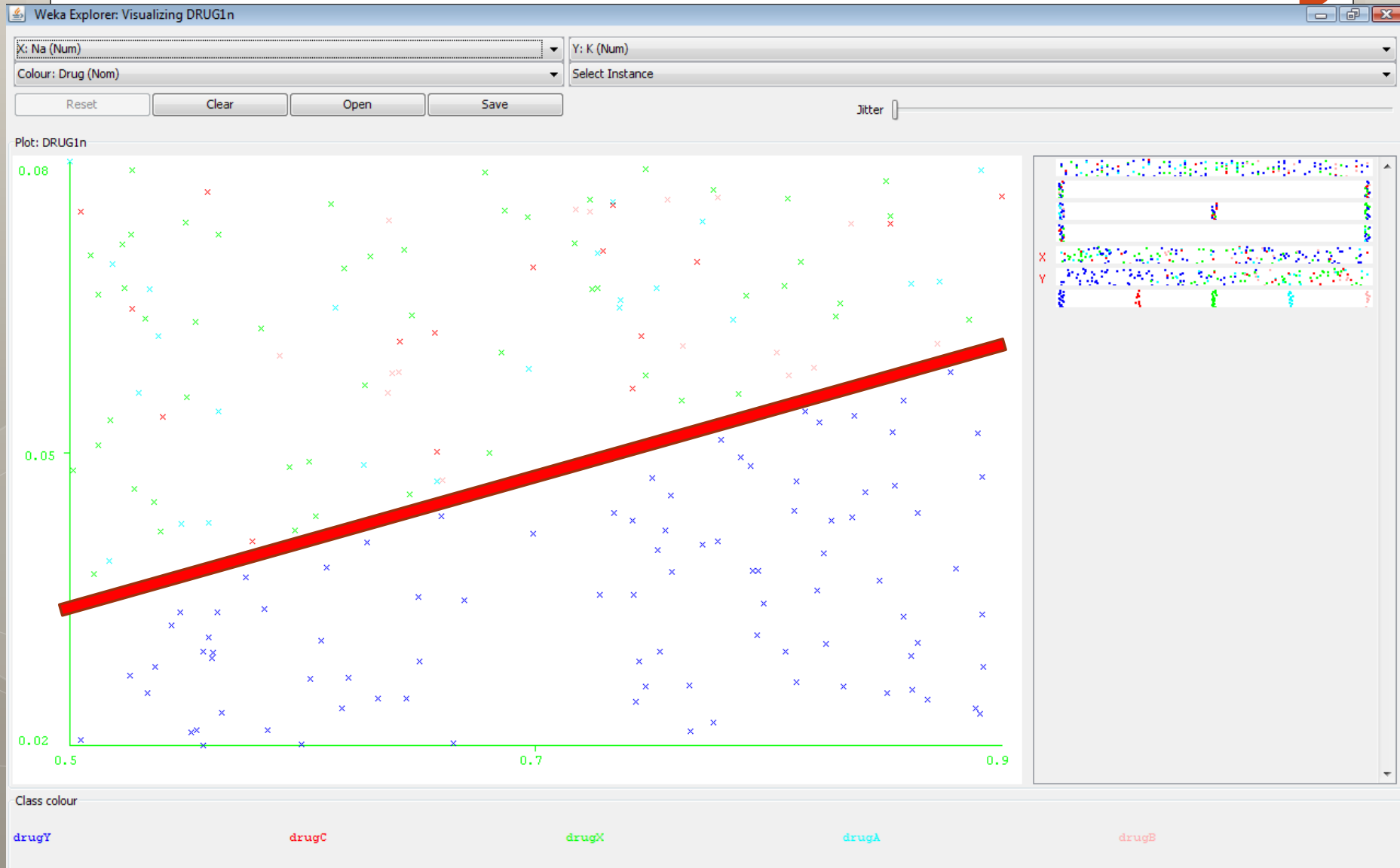


Data Understanding

- A hyperplane can clearly separate one class (drugY, the blue one) from the others!



Data Understanding





Data Preparation

- Hyperplane:
 - $K = m * Na + q$
- Let's ignore the constant q , then
 - $m = Na / K$
- We can create a new attribute Na_su_K (Na out of K, in *italian*)

Data Preparation

- Choose the AddExpression filter

The screenshot shows the Weka 3.5.6 Explorer interface. The 'Filter' list on the left contains various filters, with 'AddExpression' highlighted. The 'Selected attribute' section shows 'Age' with 57 distinct values. The 'Class: Drug (Nom)' section shows a bar chart with values 33, 30, 38, 38, 33, 30.

Weka 3.5.6 - Explorer
Program Applications Tools Visualization Windows Help

Explorer
Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

- weka
 - filters
 - AllFilter
 - MultiFilter
 - supervised
 - attribute
 - Add
 - AddCluster
 - AddExpression**
 - AddID
 - AddNoise
 - AddValues
 - Center
 - ChangeDateFormat
 - ClassAssigner
 - ClusterMembership
 - Copy
 - Discretize
 - FirstOrder
 - InterquartileRange
 - KernelFilter
 - MakeIndicator
 - MathExpression
 - MergeTwoValues

Attributes: 7

Invert Pattern

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 57
Type: Numeric
Unique: 7 (4%)

Statistic	Value
Minimum	15
Maximum	74
Mean	44.315
StdDev	16.544

Class: Drug (Nom) Visualize All

33 30 38 38 33 30

15 44.5 74

Status OK Log x 0



Data Preparation

- Configure the filter

The screenshot shows the Weka 3.5.6 Explorer interface. The main window displays the 'AddExpression' filter configuration. The 'Current relation' is 'DRUG1n' with 200 instances and 7 attributes. The 'Selected attribute' is 'Age' (Type: Numeric, Unique: 7 (4%), Distinct: 57). The 'Attributes' list includes Age, Sex, BP, Cholesterol, Na, K, and Drug. A dialog box titled 'weka.gui.GenericObjectEditor' is open, showing the filter's name 'Na_su_K' and the expression 'a5/a6'. The dialog also includes a 'debug' dropdown set to 'False' and buttons for 'Open...', 'Save...', 'OK', and 'Cancel'. The background shows a visualization of the data with a color-coded bar chart.

Value
15
74
44.315
16.544



Data Preparation

- This filter adds a new attribute to the table, applying a mathematical expression to the existing attributes (field *expression*). The new attribute will take name from the field *name*.
- Click on *more* and *capabilities* for further information about the filter



Data Preparation

Weka 3.5.6 - Explorer

Program Applications Tools Visualization Windows Help

Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **AddExpression -E a5/a6 -N Na_su_K** Apply

Current relation
Relation: DRUG1n-weka.filters.unsupervised.attribute.AddExpression-Ea5/a6-NNa_su_K
Instances: 200 Attributes: 8

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> Age
2	<input type="checkbox"/> Sex
3	<input type="checkbox"/> BP
4	<input type="checkbox"/> Cholesterol
5	<input type="checkbox"/> Na
6	<input type="checkbox"/> K
7	<input type="checkbox"/> Drug
8	<input checked="" type="checkbox"/> Na_su_K

Remove

Status: OK

Log x 0

Selected attribute

Name: Na_su_K Type: Numeric
Missing: 0 (0%) Distinct: 200 Unique: 200 (100%)

Statistic	Value
Minimum	6.269
Maximum	38.247
Mean	16.084
StdDev	7.224

Class: Drug (Nom) Visualize All

Bin Range	Count
6.27 - 10.54	54
10.54 - 14.81	60
14.81 - 19.08	39
19.08 - 23.35	15
23.35 - 27.62	17
27.62 - 31.89	10
31.89 - 36.16	5



Data Preparation

- Comments?



Data Preparation

- The attribute Na and K now are redundant. Let's remove them



Data Preparation

Weka 3.5.6 - Explorer

Program Applications Tools Visualization Windows Help

Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Open File... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **AddExpression -E a5/a6 -N Na_su_K** Apply

Current relation
Relation: DRUG1n-weka.filters.unsupervised.attribute.AddExpression-Ea5/a6-NNa_su_K
Instances: 200 Attributes: 8

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> Age
2	<input type="checkbox"/> Sex
3	<input type="checkbox"/> BP
4	<input type="checkbox"/> Cholesterol
5	<input checked="" type="checkbox"/> Na
6	<input checked="" type="checkbox"/> K
7	<input type="checkbox"/> Drug
8	<input type="checkbox"/> Na_su_K

Selected attribute
Name: Na
Missing: 0 (0%) Distinct: 200 Type: Numeric
Unique: 200 (100%)

Statistic	Value
Minimum	0.5
Maximum	0.896
Mean	0.697
StdDev	0.119

Class: Drug (Nom) Visualize All

Remove selected attributes.

Status OK Log x 0



Modeling

- Data seems ready to be used in the modeling phase