



**Universidade Federal do Sul e Sudeste do Pará**  
**Faculdade de Computação e Engenharia Elétrica**  
**Inteligência Artificial**

# **Aula 15 - Capítulo 5: Pré-Processamento de Dados**

**Prof. Dr. Elton Alves**

# Introdução

❑ O desempenho de um algoritmo de Aprendizagem de Máquina depende do estado dos dados.

❑ Estado do dado:

○ Ruidosos

- Contém erros, ou valores diferentes do esperado.

○ Incompletos

- Atributos faltosos

❑ Técnicas de pré-processamento de dados melhoram a qualidade dos dados.

# Limpeza dos Dados – Dados ruidosos

❑ **Dados incompletos** - ausência de valores.

Idade	Sexo	Peso	Manchas	temp	#Int	Diagnostico
—	M	79	—	38	—	doente
18	F	67	inexistente	39,5	4	doente
49	M	92	espalhadas	38	2	saúdavel
18	—	43	inexistente	38,5	8	doente
21	F	52	uniforme	37,6	1	saúdavel
22	F	72	inexistente	38	3	doente
—	F	87	espalhadas	39	6	doente
34	M	67	uniforme	38,2	2	saúdavel

❑ **Soluções:**

- Eliminar os objetos ausentes;
- Preencher manualmente.
- Etc.

# Limpeza dos Dados – Dados ruidosos

❑ **Dados inconsistentes** – valores conflitantes em seus atributos.

Idade	Sexo	Peso	Manchas	temp	#Int	Diagnostico
–	M	79	–	38	–	doente
18	F	67	inexistente	39,5	4	doente
49	M	92	espalhadas	38	2	saúdavel
18	–	43	inexistente	38,5	8	doente
21	F	52	uniforme	37,6	1	saúdavel
22	F	72	inexistente	38	3	doente
–	F	87	espalhadas	39	6	doente
22	F	72	uniforme	38	3	saúdavel

❑ **Solução:**

○ **Remoção**

# Limpeza dos Dados – Dados ruidosos

❑ **Dados redundantes** – um objeto é redundante quando ele é muito semelhante a um outro objeto do mesmo conjunto de dados.

Idade	Sexo	Peso	Manchas	temp	#Int	Diagnostico
28	M	79	–	38	–	doente
18	F	67	inexistente	39,5	4	doente
49	M	92	espalhadas	38	2	saúdavel
18	F	43	inexistente	39,5	4	doente
21	F	52	uniforme	37,6	1	saúdavel
22	F	72	inexistente	38	3	doente
–	F	87	espalhadas	39	6	doente
22	F	72	uniforme	38	3	saúdavel

❑ **Solução:**

○ **Eliminação**

# Limpeza dos Dados – Dados ruidosos

❑ **Dados com ruídos** – não pertencem á distribuição que gerou os dados analisados (*outliers*).

Idade	Sexo	Peso	Manchas	temp	#Int	Diagnostico
28	M	79	–	38	–	doente
18	F	300	inexistente	39,5	4	doente
49	M	92	espalhadas	38	2	saúdavel
18	F	43	inexistente	39,5	4	doente
21	F	52	uniforme	37,6	1	saúdavel
22	F	72	inexistente	38	3	doente
–	F	87	espalhadas	39	6	doente
22	F	72	uniforme	38	3	saúdavel

❑ **Solução: Remoção do ruído (suavização) e Identificação de valores discrepantes (clusterização)**

# Transformação de Atributos Numéricos

□ **Normalização dos dados** – grande diferença dos limites dos valores dos atributos distintos (**predominância de atributos**).

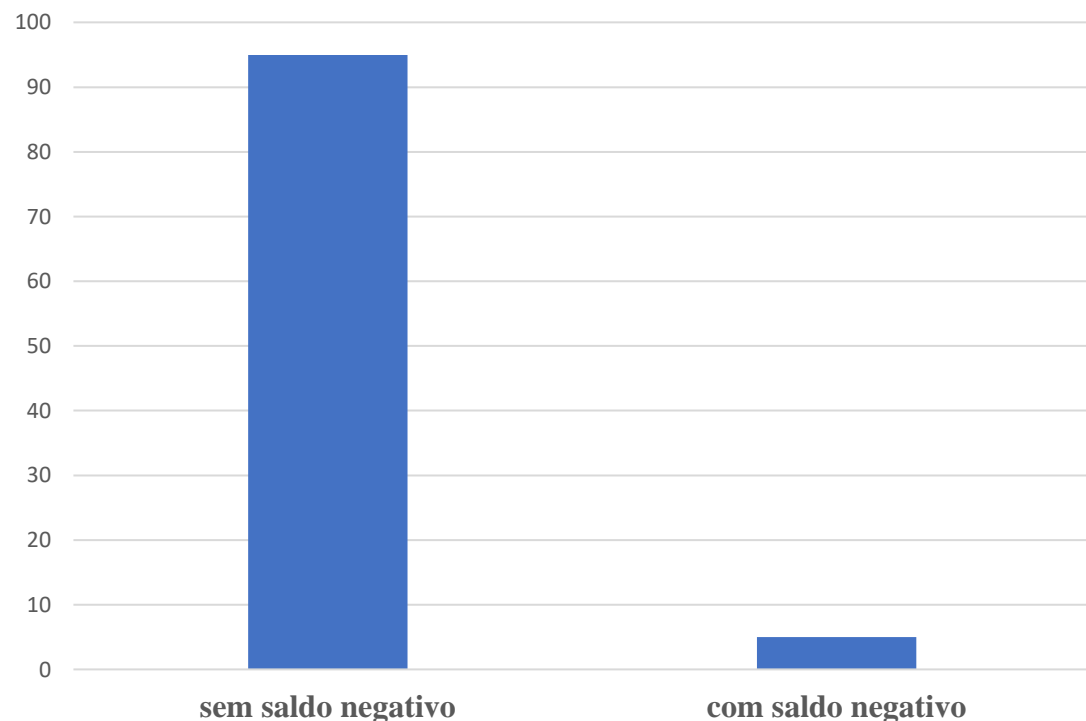
$$v_{novo} = \min + \frac{v_{atual} - \min}{\max - \min} (\max - \min)$$

# Dados desbalanceados

❑ **Exemplo:** conjunto de dados de clientes de um banco.

○ **Classe majoritária:** 95% sem saldo negativo.

○ **Classe minoritária:** 5% com saldo negativo.

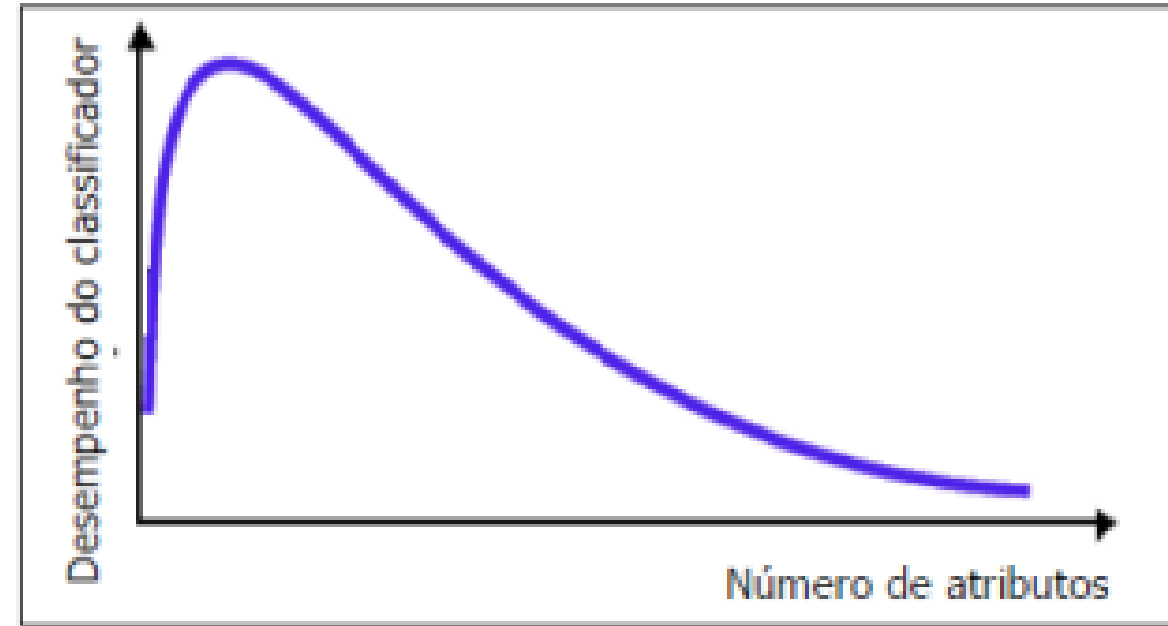


**Solução:** Técnicas de balanceamento de dados sintéticos - SMOTE.



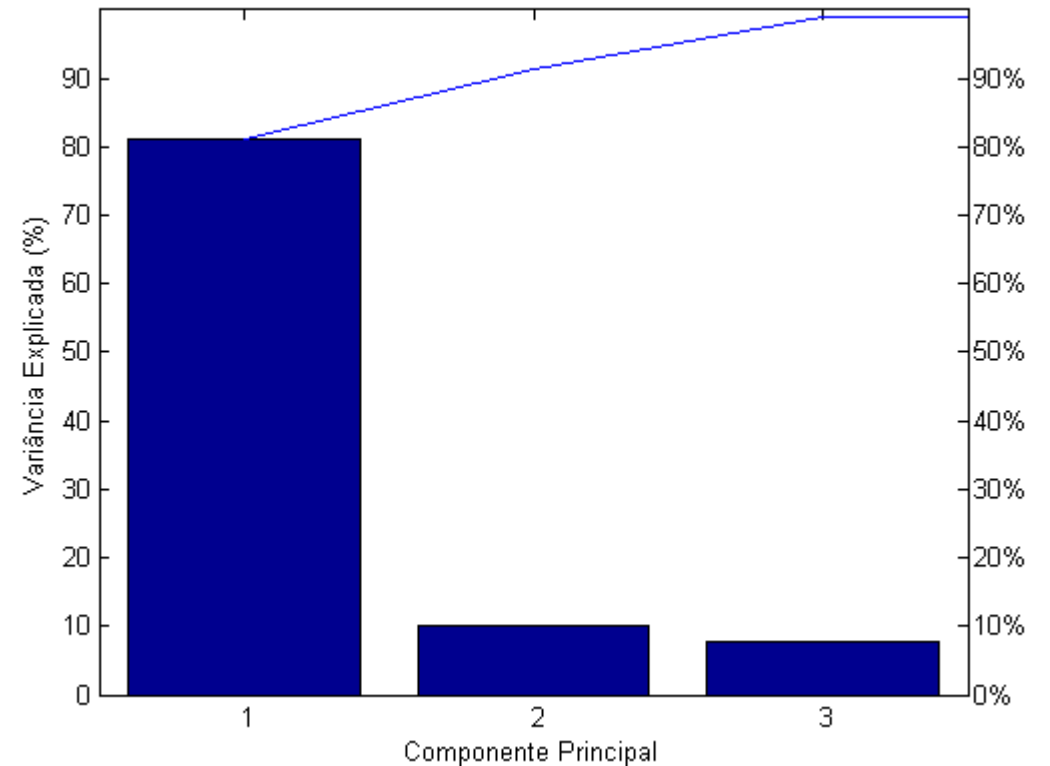
# Redução de Dimensionalidade

- ❑ Problemas com elevado número de atributos.
- ❑ Redução de atributos podem melhorar o desempenho do modelo induzido.
- ❑ **Análise de Componentes Principais (PCA).**

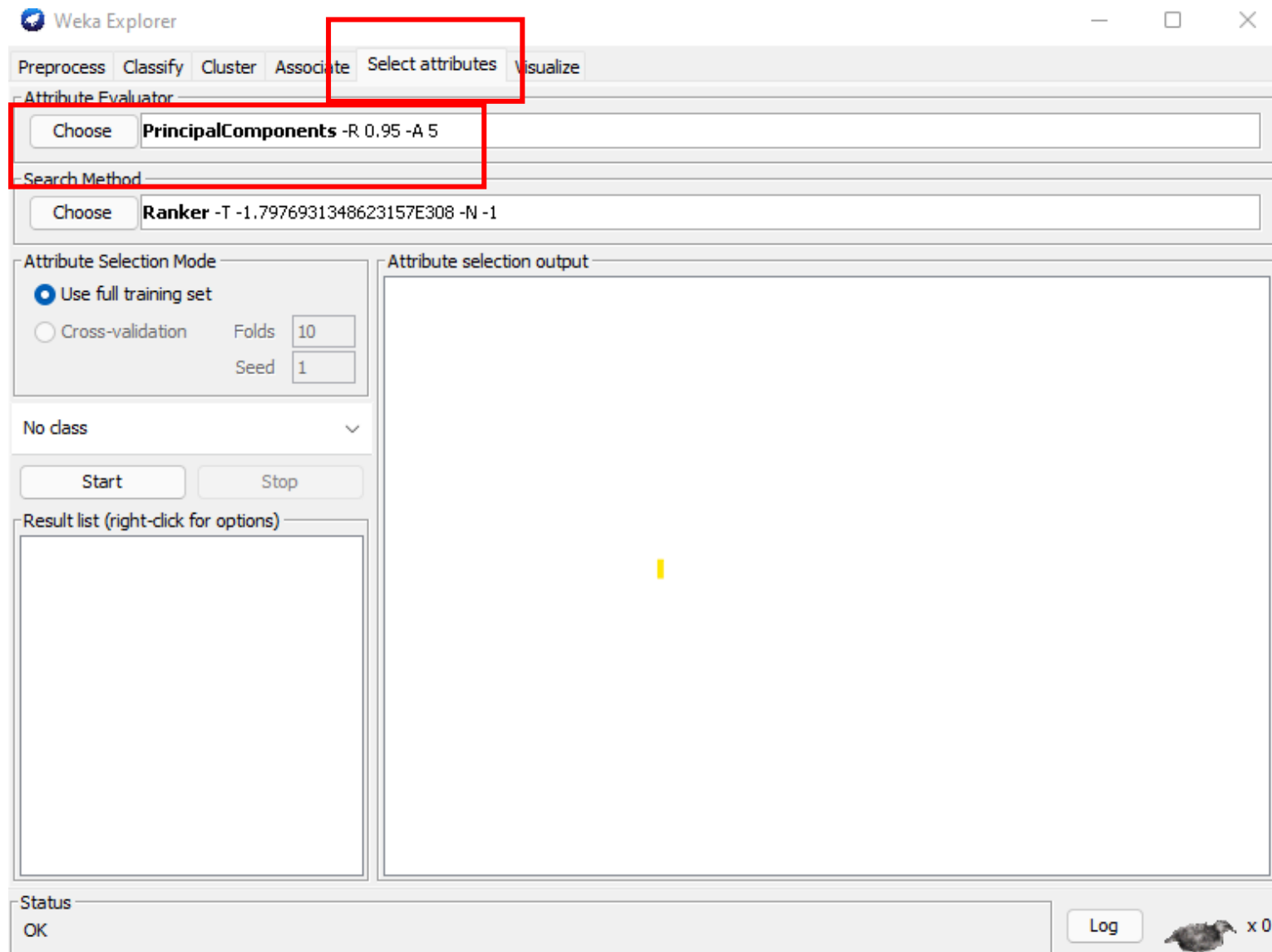


# Análise de Componentes Principais (PCA)

- Consiste na **transformação** dos dados para um novo espaço com dimensão inferior ao original.
- O novo espaço fica caracterizado por um novo conjunto de eixos, ortonormais entre si, ordenados em ordem **decrescente de variância**.



# PCA no Weka



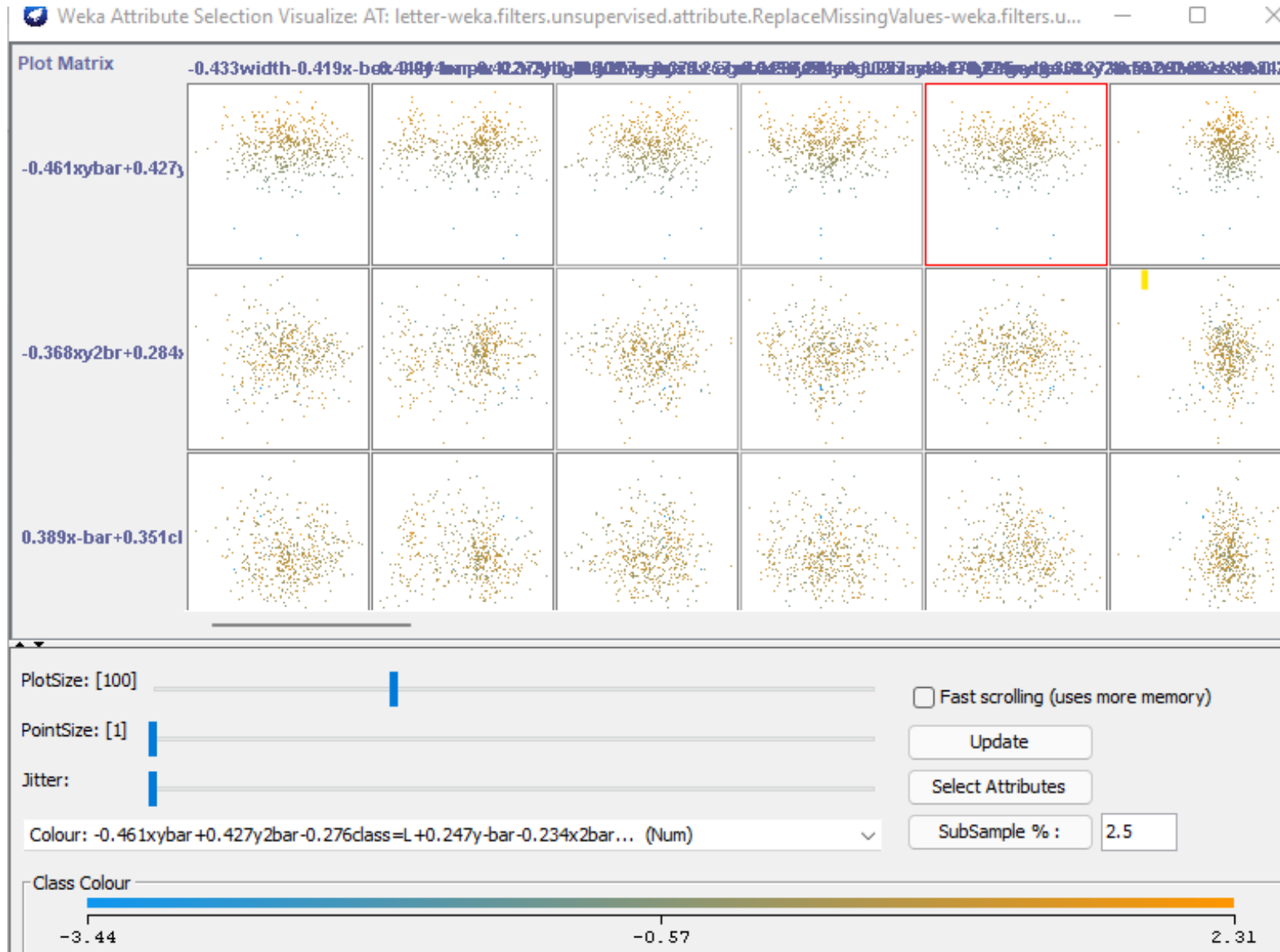
# PCA no Weka

Ranked attributes:

```
0.8949    1 -0.433width-0.419x-box-0.414onpix-0.379high-0.377y-box...
0.8143    2 -0.449y-bar-0.422x2ybr-0.409xegvy+0.257x-bar+0.244yegvx...
0.7574    3 0.49 y2bar-0.374x-ege-0.287x2bar+0.277xybar+0.271y-ege...
0.7088    4 0.445xy2br+0.302class=C-0.295xybar-0.273x-bar+0.262x2bar...
0.6682    5 -0.372y-ege+0.368xy2br-0.292class=P-0.284class=B+0.245class=J...
0.631     6 -0.537class=L+0.347xybar+0.34 x-bar+0.304yegvx+0.3 class=J...
0.5971    7 0.484x2bar+0.31 class=U-0.299class=W+0.239class=J+0.239xybar...
0.5651    8 0.386class=M-0.297class=L+0.292class=Z-0.288class=P+0.24 y2bar...
0.534     9 0.337class=K+0.327class=F-0.326class=Y-0.317x-bar+0.281class=R...
0.5067   10 0.348class=R+0.343class=I+0.324class=V-0.295class=W+0.287class=T...
0.4802   11 -0.465class=W+0.385class=F+0.317class=A-0.266class=D-0.247class=N...
0.4541   12 0.509class=T+0.394class=N-0.333class=U-0.307class=B+0.264class=R...
0.4287   13 0.443class=Q-0.405class=X+0.403class=I-0.291class=V+0.221class=F...
0.4037   14 0.457class=H-0.454class=S-0.325class=I-0.317class=N-0.256class=O...
0.3788   15 0.379class=X+0.365class=Q-0.361class=T+0.325class=Z+0.278class=N...
0.354    16 0.651class=D-0.353class=U-0.348class=P+0.25 class=V+0.225class=F...
0.3292   17 -0.416class=U+0.408class=Y+0.394class=M-0.331class=Q-0.254class=F...
0.3045   18 -0.442class=P-0.385class=E+0.345class=B+0.271class=F+0.236class=J...
0.2797   19 -0.57class=G+0.401class=E+0.319class=N-0.242class=T+0.232class=Y...
0.2549   20 0.505class=R-0.419class=B+0.36 class=Y-0.31class=V+0.212class=U...
0.2302   21 0.507class=O+0.43 class=X-0.371class=N-0.35class=G-0.284class=Z...
0.2054   22 0.526class=Z-0.348class=X-0.347class=G+0.333class=V-0.286class=E...
0.1807   23 -0.409class=C+0.354class=E-0.33class=B-0.306class=K+0.303class=V...
```

**Variância Explicada**

# PCA no Weka



**Atributos seleccionados**

# PCA no Weka

The screenshot shows the Weka Explorer interface with the 'Attribute Evaluator' set to 'PrincipalComponents -R 0.95 -A 5' and the 'Search Method' set to 'Ranker -T -1.797693'. The 'Attribute Selection Mode' is set to 'Use full training set'. A dialog box titled 'weka.gui.GenericObjectEditor' is open, showing the configuration for the 'Ranker' attribute selection method. The 'numToSelect' field is highlighted with a red box and contains the value '-1'. The 'threshold' field contains the value '-1.7976931348623157E308'. The dialog box also includes an 'About' section with the text 'Ranker : Ranks attributes by their individual evaluations.' and buttons for 'Open...', 'Save...', 'OK', and 'Cancel'. The background shows a list of attributes and their values, including '10:07:00 - Ranker + PrincipalCompon'.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Attribute Evaluator  
Choose **PrincipalComponents** -R 0.95 -A 5

Search Method  
Choose **Ranker** -T -1.797693

Attribute Selection Mode  
☒ Use full training set  
☐ Cross-validation Folds 10 Seed 1

No class

Start Stop

Result list (right-click for options)  
10:07:00 - Ranker + PrincipalCompon

Status  
OK

Log x 0

weka.gui.GenericObjectEditor

weka.attributeSelection.Ranker

About

Ranker :  
Ranks attributes by their individual evaluations.

generateRanking True

numToSelect -1

startSet

threshold -1.7976931348623157E308

Open... Save... OK Cancel

Seleção automática dos atributos

# PCA no Weka

The screenshot shows the Weka Explorer interface with the 'Attribute Evaluator' set to 'PrincipalComponents' and the 'Search Method' set to 'Ranker'. A 'GenericObjectEditor' dialog is open, showing the configuration for the 'Ranker' attribute selection method. The dialog has a red box around the 'generateRanking' (True) and 'numToSelect' (6) fields. The 'threshold' field is set to '-1.7976931348623157E308'. The 'About' section describes the Ranker method: 'Ranks attributes by their individual evaluations.' The 'Result list' on the left shows two entries: '10:07:00 - Ranker + PrincipalComponent' and '10:23:21 - Ranker + PrincipalComponent', with the latter selected. The 'Result list' on the right shows a list of attributes and their values, including 'x-bar', 'yegvx', and 'class'.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Attribute Evaluator: Choose **PrincipalComponents** -R 0.95 -A 5

Search Method: Choose **Ranker** -T -1.7976931348623157E308

Attribute Selection Mode:  
☒ Use full training set  
☐ Cross-validation Folds: 10 Seed: 1

No class

Start Stop

Result list (right-click for options):  
10:07:00 - Ranker + PrincipalComponent  
10:23:21 - Ranker + PrincipalComponent

weka.gui.GenericObjectEditor

weka.attributeSelection.Ranker

About  
Ranker : More  
Ranks attributes by their individual evaluations.

generateRanking True  
numToSelect 6

startSet  
threshold -1.7976931348623157E308

Open... Save... OK Cancel

Selected attributes: 1,2,3,4,5,6 : 6

7 -0.161 -0.1  
5 -0.0608 0.2  
1 -0.204 0.0  
7 -0.1013 0.3  
8 0.0938 -0.2  
7 -0.0081 0.0  
1 -0.3264 -0.0  
4 -0.0645 -0.1

-box...  
yegvx...  
-ege...  
2x2bar...

0.668 5 -0.372y-ege+0.368xy2br-0.292class=P-0.284class=B+0.245class=J...  
0.631 6 -0.537class=L+0.347xybar+0.34 x-bar+0.304yegvx+0.3 class=J...

# PCA no Weka

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds

☐ Percentage split %

More options...

(Nom) class ▾

Start Stop

Result list (right-click for options)

- 10:27:23 - bayes.BayesNet
- 10:29:05 - bayes.NaiveBayes

Classifier output

=== Stratified cross-validation ===

=== Summary ===


Correctly Classified Instances	12823	64.115 %
Incorrectly Classified Instances	7177	35.885 %
Kappa statistic	0.6268	
Mean absolute error	0.0323	
Root mean squared error	0.1391	
Relative absolute error	43.6524 %	
Root relative squared error	72.3267 %	
Total Number of Instances	20000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
	0,872	0,007	0,839	0,872	0,855	0,849
	0,702	0,030	0,479	0,702	0,570	0,560
	0,750	0,009	0,768	0,750	0,759	0,750
	0,704	0,019	0,606	0,704	0,651	0,638
	0,350	0,009	0,603	0,350	0,443	0,444
	0,733	0,013	0,700	0,733	0,716	0,704
	0,538	0,019	0,537	0,538	0,537	0,519
	0,304	0,010	0,535	0,304	0,387	0,387

Status

OK

Log  x 0



# Naive Bayes + PCA no Weka

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter  
Choose **PrincipalComponents -R 0.95 -A 5 -M -1** Apply Stop

Current relation  
Relation: letter Attributes: 17  
Instances: 20000 Sum of weights: 20000

Attributes  
All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> x-box
2	<input type="checkbox"/> y-box
3	<input type="checkbox"/> width
4	<input type="checkbox"/> high
5	<input type="checkbox"/> onpix
6	<input type="checkbox"/> x-bar
7	<input type="checkbox"/> y-bar
8	<input type="checkbox"/> x2bar
9	<input type="checkbox"/> y2bar
10	<input type="checkbox"/> xybar
11	<input type="checkbox"/> x2ybr
12	<input type="checkbox"/> xy2br
13	<input type="checkbox"/> x-ege
14	<input type="checkbox"/> y-ege

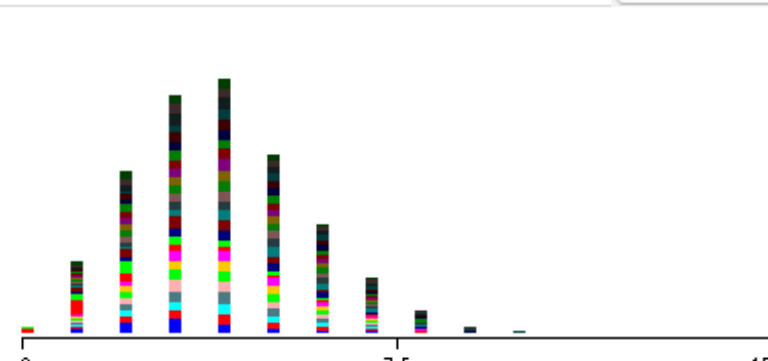
Remove

Status  
OK

Selected attribute  
Name: x-box  
Missing: 0 (0%)  
Distinct: 16  
Type: Numeric  
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	15
Mean	4.024
StdDev	1.913

Class: class (Nom) Visualize All



Log x 0

# Naive Bayes + PCA no Weka

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter  
Choose **PrincipalComponents** -R 0.95 -A 5 -M -1 Apply Stop

Current relation  
Relation: letter\_principal components-wek... Attributes: 13  
Instances: 20000 Sum of weights: 20000

Attributes  
All None Invert Pattern

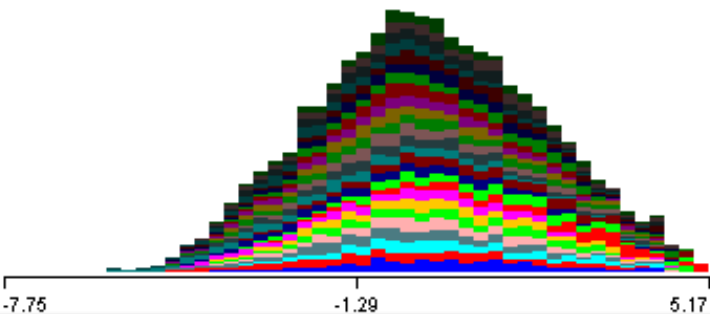
No.	Name
1	<input checked="" type="checkbox"/> -0.438width-0.427x-box-0.417onpix-0.4high-0.4y-box...
2	<input type="checkbox"/> 0.509y-bar +0.472x2ybr +0.457xegvy-0.33x-bar-0.309ye...
3	<input type="checkbox"/> -0.551y2bar +0.452x2bar-0.443xybar +0.431x-eg +0.169...
4	<input type="checkbox"/> -0.496xy2br +0.416x-bar-0.394y2bar +0.317xybar-0.313x...
5	<input type="checkbox"/> -0.619y-eg +0.455xy2br +0.358xybar +0.239y-box-0.233...
6	<input type="checkbox"/> 0.665yegvx +0.414x-bar +0.287xybar +0.287xegvy +0.26...
7	<input type="checkbox"/> 0.527xy2br-0.522x2bar +0.351x-eg-0.279high-0.275y-b...
8	<input type="checkbox"/> 0.539x2bar +0.519xybar-0.29xegvy +0.288y-eg-0.248hi...
9	<input type="checkbox"/> -0.657x-bar +0.414yegvx-0.291xegvy-0.263x2ybr +0.203...
10	<input type="checkbox"/> 0.542y2bar-0.35y-eg-0.316high +0.296width +0.289yegv...
11	<input type="checkbox"/> 0.646xegvy-0.417y-bar-0.35x2ybr +0.275xybar-0.196y2b...
12	<input type="checkbox"/> -0.581onpix +0.514x-box +0.314y-eg-0.264high +0.233y-...
13	<input type="checkbox"/> class

Remove

Selected attribute  
Name: -0.438width-0.427x-box-0.417onpix-0.4hi... Type: Num...  
Missing: 0 (0%) Distinct: 18668 Unique: 178...

Statistic	Value
Minimum	-7.75
Maximum	5.171
Mean	0
StdDev	2.073

Class: class (Nom) Visualize All



-7.75 -1.29 5.17

Status  
OK Log x 0

# Naive Bayes + PCA no Weka

Choose

NaiveBayes

Test options

☐ Use training set

☐ Supplied test set 

Set...

☒ Cross-validation 

Folds 

10

☐ Percentage split 

% 

66

More options...

(Nom) class

▼

Start

Stop

Result list (right-click for options)

10:27:23 - bayes.BayesNet

10:29:05 - bayes.NaiveBayes

10:36:24 - bayes.NaiveBayes

Classifier output

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	12963	64.815 %
Incorrectly Classified Instances	7037	35.185 %
Kappa statistic	0.6341	
Mean absolute error	0.0345	
Root mean squared error	0.1369	
Relative absolute error	46.6611 %	
Root relative squared error	71.181 %	
Total Number of Instances	20000	


=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
	0,845	0,004	0,889	0,845	0,867	0,862
	0,757	0,026	0,538	0,757	0,629	0,621
	0,709	0,012	0,701	0,709	0,705	0,694
	0,733	0,015	0,666	0,733	0,698	0,685
	0,523	0,013	0,617	0,523	0,566	0,552
	0,688	0,018	0,607	0,688	0,645	0,631
	0,436	0,019	0,481	0,436	0,457	0,437
	0,324	0,021	0,374	0,324	0,347	0,325

Status

OK

Log

 x 0

# Naive Bayes + PCA no Weka

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

Test options:

- ☐ Use training set
- ☐ Supplied test set Set...
- ☒ Cross-validation Folds
- ☐ Percentage split %

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 10:27:23 - bayes.BayesNet
- 10:29:05 - bayes.NaiveBayes
- 10:36:24 - bayes.NaiveBayes
- 11:41:22 - bayes.NaiveBayes**

Classifier output

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===


=== Summary ===

Correctly Classified Instances	3388	<b>16.94</b>	%
Incorrectly Classified Instances	16612	83.06	%
Kappa statistic	0.1361		
Mean absolute error	0.0695		
Root mean squared error	0.1862		
Relative absolute error	93.9588	%	
Root relative squared error	96.8416	%	
Total Number of Instances	20000		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
	0,575	0,033	0,418	0,575	0,484	0,466
	0,307	0,091	0,118	0,307	0,171	0,138
	0,003	0,001	0,143	0,003	0,005	0,015
	0,010	0,005	0,073	0,010	0,018	0,012
	0,000	0,001	0,000	0,000	0,000	-0,005

Status: OK

Log  x 0