

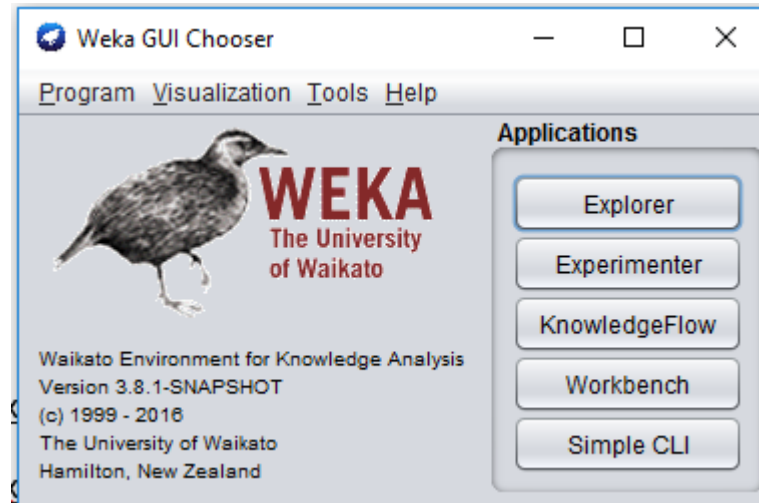


Preparación de Datos

PhD Jose Ricardo Zapata
< joser.zapata@upb.edu.co >

UNIVERSIDAD PONTIFICIA BOLIVARIANA
FACULTAD TIC

Weka



Explorer: Realiza operaciones sobre un sólo conjunto de datos.

Experimenter: Realiza contrastes estadísticos entre métodos.

KnowledgeFlow: Muestra el funcionamiento interno del proceso.

Simple CLI: Consola para ejecución manual de análisis.

Representación de Datos en Weka

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize Forecast

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

Filter: Choose None Apply

Current relation: Relation: weather Instances: 14 Attributes: 5 Sum of weights: 14

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

Status: OK

Selected attribute: Name: outlook Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom) Visualize All

Log x 0

Representación de Datos en Weka

EJERCICIO PREPARACIÓN DE DATOS:

1. Abrir el archivo “Preparacion de datos.xlsx”
2. Guardar la pestaña “formato arff” en un archivo separado por comas CSV
3. Abrir el archivo CSV con bloc de notas (verificar que esté separado por comas)

```
Id,Estrato,Sexo,Enfermedad,Colegio_U,Activo_Web,Asistencia,Entregas_Completas,Trabaja,Examen_Final
1,3,F,SI,SI,BAJA,0.1,0.45,NO,APROBADO
2,2,M,NO,NO,MEDIA,0.45,0.6,NO,APROBADO
3,4,M,NO,NO,ALTA,0.5,0.75,SI,DESAPROBADO
4,3,F,NO,NO,ALTA,0,0.5,SI,DESAPROBADO
5,4,F,SI,NO,MEDIA,0.65,0.85,NO,APROBADO
6,4,M,NO,NO,BAJA,0.1,0,NO,DESAPROBADO
7,3,M,NO,NO,MEDIA,0.2,0.9,NO,APROBADO
8,2,M,NO,NO,MEDIA,0.3,0.8,SI,DESAPROBADO
9,2,F,NO,SI,BAJA,0.35,0.7,NO,APROBADO
10,2,M,SI,NO,BAJA,0.75,0.5,SI,DESAPROBADO
11,4,M,NO,SI,ALTA,0.7,0.6,NO,APROBADO
12,3,F,NO,NO,MEDIA,0.9,0.8,NO,APROBADO
13,2,F,NO,NO,ALTA,0.2,0.25,NO,DESAPROBADO
14,4,F,SI,NO,ALTA,0.2,0.2,NO,DESAPROBADO
15,3,F,NO,NO,BAJA,0.9,0.8,NO,APROBADO
16,2,M,NO,NO,MEDIA,1,1,NO,APROBADO
```

File: aprobacion_curso.arff

Representación de Datos en Weka

EJERCICIO PREPARACIÓN DE DATOS:

4. Añadimos la descripción de atributos

```
@relation aprobacion_curso

% Tipos de atributos: numeric, integer, date, string, enumerate{a,b,c}

@attribute Id integer
@attribute Estrato{1,2,3,4,5}
@attribute Sexo{F,M}
@attribute Enfermedad{SI,NO}
@attribute Colegio_U{SI, NO}
@attribute Activo_web{ALTA, MEDIA, BAJA}
@attribute Asistencia numeric
@attribute Entregas_completas numeric
@attribute Trabaja{SI, NO}
@attribute Examen_Final{APROBADO, DESAPROBADO}

@data
1,3,F,SI,SI,BAJA,0.1,0.45,NO,APROBADO
2,2,M,NO,NO,MEDIA,0.45,0.6,NO,APROBADO
3,4,M,NO,NO,ALTA,0.5,0.75,SI,DESAPROBADO
4,3,F,NO,NO,ALTA,0,0.5,SI,DESAPROBADO
5,4,F,SI,NO,MEDIA,0.65,0.85,NO,APROBADO
6,4,M,NO,NO,BAJA,0.1,0,NO,DESAPROBADO
7,3,M,NO,NO,MEDIA,0.2,0.9,NO,APROBADO
8,2,M,NO,NO,MEDIA,0.3,0.8,SI,DESAPROBADO
9,2,F,NO,SI,BAJA,0.35,0.7,NO,APROBADO
10,2,M,SI,NO,BAJA,0.75,0.5,SI,DESAPROBADO
11,4,M,NO,SI,ALTA,0.7,0.6,NO,APROBADO
```

Representación de Datos en Weka

EJERCICIO PREPARACIÓN DE DATOS:

4. Añadimos la descripción de atributos

```
@relation aprobacion_curso
% Tipos de atributos: numeric, integer, date, string, enumerate{a,b,c}
@attribute Id integer
@attribute Estrato{1,2,3,4,5}
@attribute Sexo{F,M}
@attribute Enfermedad{SI,NO}
@attribute Colegio_U{SI, NO}
@attribute Activo_web{ALTA, MEDIA, BAJA}
@attribute Asistencia numeric
@attribute Entregas_completas numeric
@attribute Trabaja{SI, NO}
@attribute Examen_Final{APROBADO, DESAPROBADO}

@data
```

Datos en Weka

Archivo
arff

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize Forecast

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

Filter: Choose None Apply

Current relation: Relation: aprobacion_curso Instances: 16 Attributes: 10 Sum of weights: 16

Attributes: All None Invert Pattern

No.	Name
1	Id
2	Estrato
3	Sexo
4	Enfermedad
5	Colegio_U
6	Activo_web
7	Asistencia
8	Entregas_completas
9	Trabaja
10	Examen_Final

Remove

Status: OK Log x 0

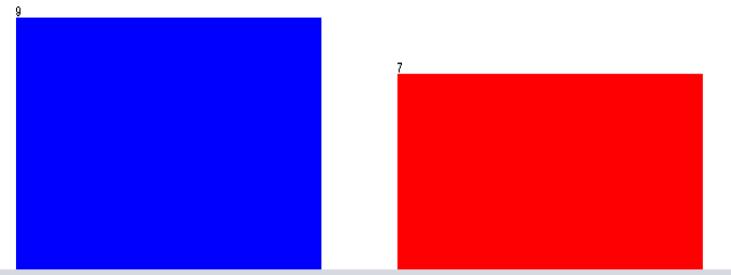
Selected attribute

Name: Examen_Final
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	APROBADO	9	9.0
2	DESAPROBADO	7	7.0

Class: Examen_Final (Nom)

Visualize All



Estadística
descriptiva

Ver
histogramas

Eliminar
variables

Preparación de los Datos

1. Integración de los Datos
2. Descripción de los Datos
3. Limpieza de Datos
 - Datos Ausentes/Nulos
 - Datos Atípicos
 - Registros duplicados
4. Transformaciones
 - Normalización (Filtro: unsupervised/attribute/normalize)
 - Discretización (Filtro: unsupervised/attribute/discretize)
 - Conversión categórica a numérica (automática en weka)

Preparación de los Datos

5. Selección de Variables

- Variables irrelevantes
- Variables redundantes

6. Análisis de Correlaciones

- Correlación entre variables
- Correlación con la variable a predecir (predicción)

7. Reducción de Variables

- PCA(Select attributes/Principal Components)

8. Balanceo de Datos

- Selección aleatoria (Filtro: supervised/instance/Resample)
- Adicionar registros (Filtro: supervised/instance/Smote)

Preparación de Datos en Weka

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize Forecast

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

Filter

Choose None Apply

Current relation

Relation: aprobacion_curso
Instances: 16

Attributes: 10
Sum of weights: 16

Attributes

All None Invert Pattern

No.	Name
1	Id
2	Estrato
3	Sexo
4	Enfermedad
5	Colegio_U
6	Activo_web
7	Asistencia
8	Entregas_completas
9	Trabaja
10	Examen_Final

Remove

Status

OK

Log x 0

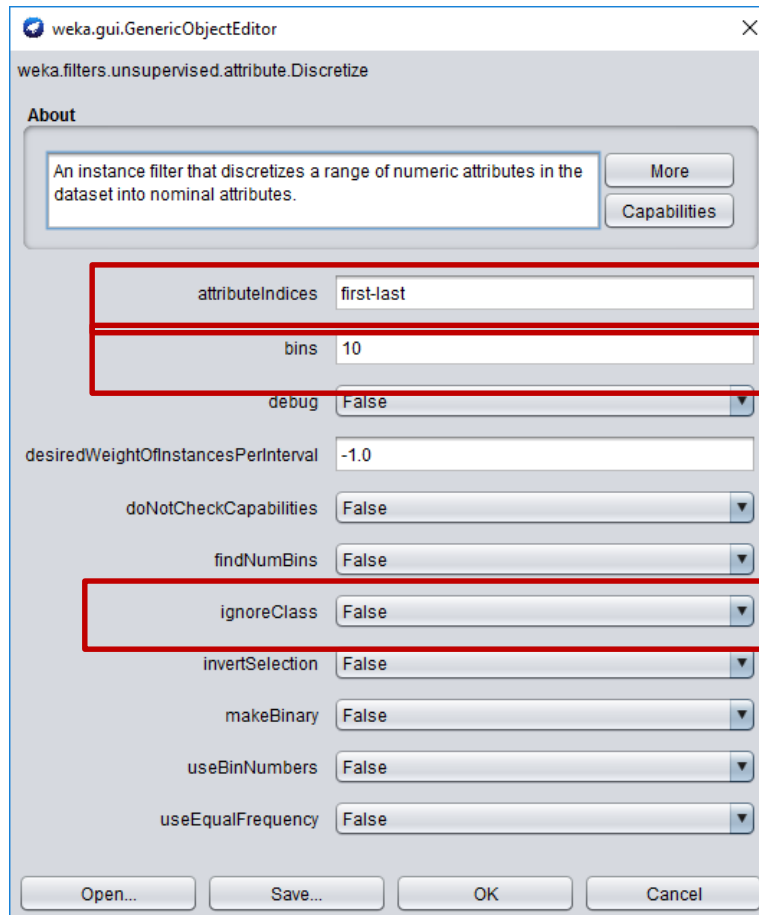
Filtros:

- Discretize
- Merge two values
- Normalize
- NumericToBinary
- Remove
- StringToWordVector
- Smote

Transformación: Discretize

Conversión de variable numérica a categórica.

Filtro: unsupervised/attribute/discretize



The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.filters.unsupervised.attribute.Discretize' filter. The 'About' section describes it as an instance filter that discretizes a range of numeric attributes. The settings are as follows:

Property	Value
attributeIndices	first-last
bins	10
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
useBinNumbers	False
useEqualFrequency	False

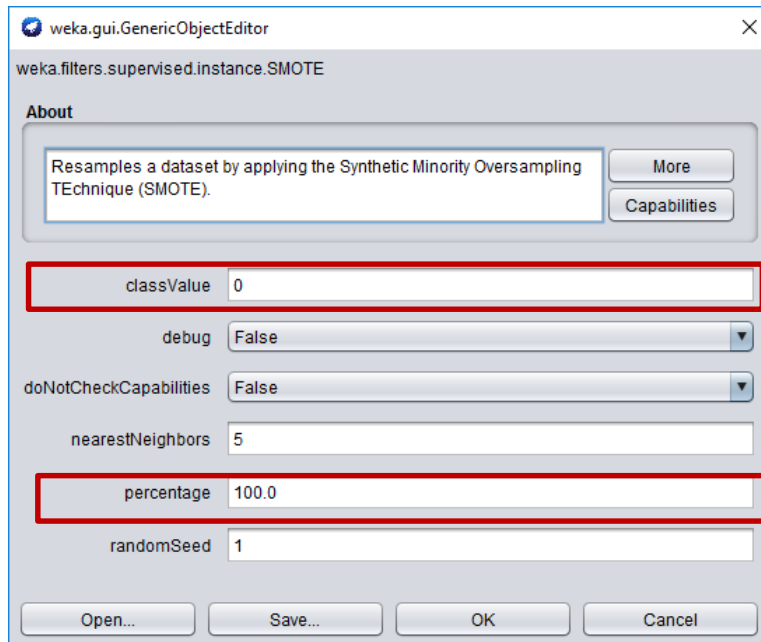
Buttons at the bottom: Open..., Save..., OK, Cancel.

- Número de atributo a discretizar
- Cantidad de categorías
- Ignorar si el atributo es la clase a predecir

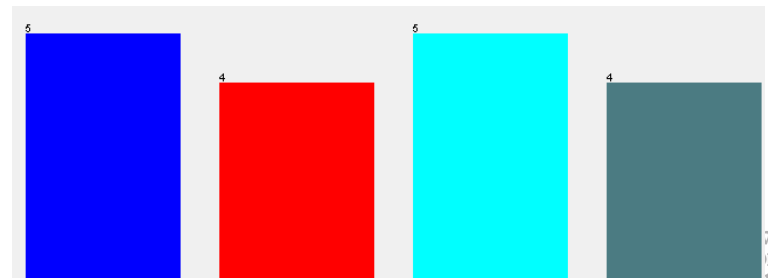
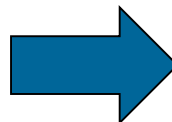
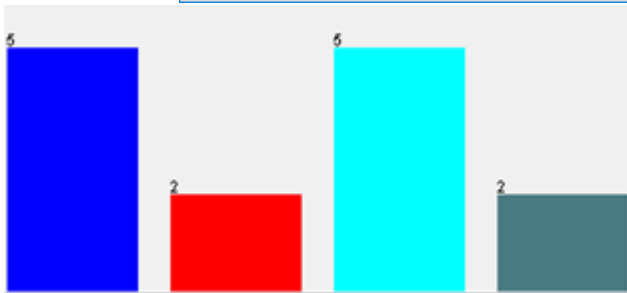
Balanceo: Smote

Balanceo de la variable a predecir.

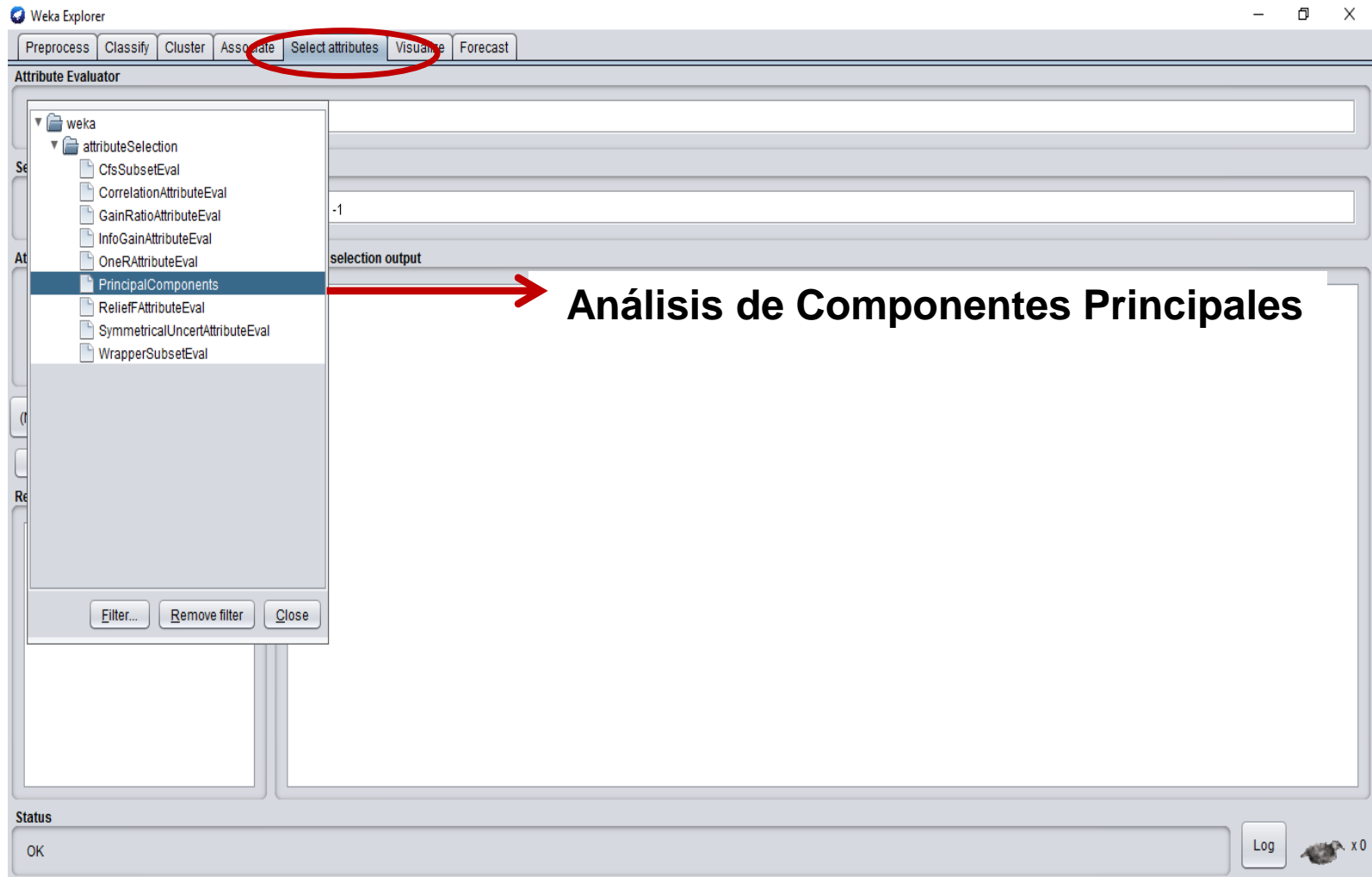
Filtro: supervised/instance/SMOTE



- Clase a balancear
- Cuantos datos quiero aumentar?

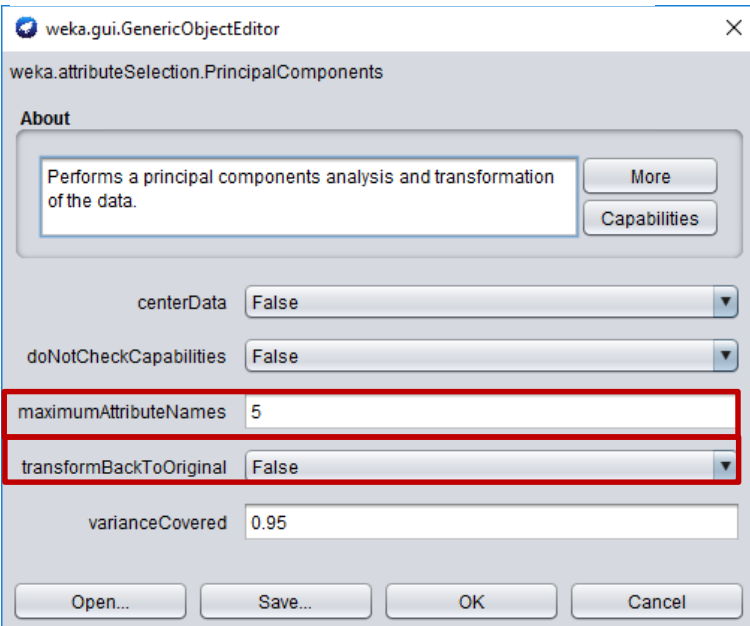


Reducción de Variables: PCA



Reducción de Variables: PCA

Análisis de Componentes Principales



The screenshot shows the 'weka.gui.GenericObjectEditor' window for 'weka.attributeSelection.PrincipalComponents'. It includes an 'About' section with a description and 'More'/'Capabilities' buttons. Below are several settings: 'centerData' (False), 'doNotCheckCapabilities' (False), 'maximumAttributeNames' (5, highlighted with a red box), 'transformBackToOriginal' (False, highlighted with a red box), and 'varianceCovered' (0.95). At the bottom are 'Open...', 'Save...', 'OK', and 'Cancel' buttons.

weka.gui.GenericObjectEditor

weka.attributeSelection.PrincipalComponents

About

Performs a principal components analysis and transformation of the data. More Capabilities

centerData False

doNotCheckCapabilities False

maximumAttributeNames 5

transformBackToOriginal False

varianceCovered 0.95

Open... Save... OK Cancel

centerData: False si se calcula matriz de correlaciones. True si se calcula la covarianza.

maximumAttributeNames: máxima cantidad de atributos a incluir en nuevas variables.

transformBackToOriginal: True si se evalúan los atributos iniciales. False si evalúan nuevas variables como combinación de atributos.

varianceCovered: se retienen los atributos necesarios para la varianza.

Reducción de Variables: PCA

Análisis de Componentes Principales

Correlation matrix

1	-0.52	-0.52	0.26	0.15	0.04	-0.24	0.2	0.03	0.13	0.1	-0.15	0.1	Estrato=2
-0.52	1	-0.45	-0.4	0.08	-0.02	-0.16	0.03	0.13	-0.08	0.21	0.08	-0.32	Estrato=3
-0.52	-0.45	1	0.13	-0.23	-0.02	0.42	-0.24	-0.16	-0.06	-0.32	0.08	0.22	Estrato=4
0.26	-0.4	0.13	1	0.29	0.16	-0.13	0.26	-0.13	0.14	0.14	-0.29	0.13	Sexo
0.15	0.08	-0.23	0.29	1	0.09	0.08	0.15	-0.23	0.06	0.23	0	-0.07	Enfermedad
0.04	-0.02	-0.02	0.16	0.09	1	-0.02	0.37	-0.37	0.11	0.04	-0.28	0.42	Colegio_U
-0.24	-0.16	0.42	-0.13	0.08	-0.02	1	-0.52	-0.45	-0.29	-0.37	-0.23	0.49	Activo_web=ALTA
0.2	0.03	-0.24	0.26	0.15	0.37	-0.52	1	-0.52	0.31	0.63	0.15	-0.42	Activo_web=MEDIA
0.03	0.13	-0.16	-0.13	-0.23	-0.37	-0.45	-0.52	1	-0.03	-0.29	0.08	-0.05	Activo_web=BAJA
0.13	-0.08	-0.06	0.14	0.06	0.11	-0.29	0.31	-0.03	1	0.6	0.13	-0.46	Asistencia
0.1	0.21	-0.32	0.14	0.23	0.04	-0.37	0.63	-0.29	0.6	1	-0.07	-0.58	Entregas_completas
-0.15	0.08	0.08	-0.29	0	-0.28	-0.23	0.15	0.08	0.13	-0.07	1	-0.65	Trabaja
0.1	-0.32	0.22	0.13	-0.07	0.42	0.49	-0.42	-0.05	-0.46	-0.58	-0.65	1	Examen_Final

Reducción de Variables: PCA

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize Forecast

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

Search Method
Choose Ranker -T-1.7976931348623157E308 -N -1

Attribute Selection Mode
☒ Use full training set
☐ Cross-validation Folds 10
Seed 1
(Nom) Examen_Final
Start Stop

Result list (right-click for options)
21:41:06 - Ranker + PrincipalComponent
21:41:20 - Ranker + PrincipalComponent

Attribute selection output

```
-0.4566 -0.079 -0.2024 -0.0313 0.0999 0.1164 -0.3265 0.3699 Entregas_completas  
-0.0092 -0.295 -0.1241 0.4425 0.1739 0.345 0.5241 -0.1868 Trabaja=NO  
  
PC space transformed back to original space.  
(Note: Can't evaluate attributes in the original space)  
  
Ranked attributes:  
1 13 Trabaja=NO  
1 6 Enfermedad=NO  
1 4 Estrato=4  
1 3 Estrato=3  
1 2 Estrato=2  
1 5 Sexo=M  
1 7 Colegio_U=NO  
1 12 Entregas_completas  
1 8 Activo_web=ALTA  
1 11 Asistencia  
1 10 Activo_web=BAJA  
1 9 Activo_web=MEDIA  
1 1 Id  
  
Selected attributes: 13,6,4,3,2,5,7,12,8,11,10,9,1 : 13
```

Status
OK Log x 0

Reducción de Variables: PCA

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize Forecast

Attribute Evaluator

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

Search Method

Choose **Ranker -T -1.7976931348623157E308 -N -1**

Attribute Selection Mode

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

(Nom) Examen_Final

Start Stop

Result list (right-click for options)

21:41:06 - Ranker + PrincipalComponent
21:41:20 - Ranker + PrincipalComponent
21:42:06 - Ranker + PrincipalComponent

Attribute selection output

```
0.288 0.345 -0.1596 0.2939 0.4481 0.0758 -0.1223 -0.1158 Estrato=4  
-0.1997 0.3787 0.2652 -0.1501 0.247 0.2699 -0.245 -0.3392 Sexo=M  
-0.2095 0.1319 -0.0699 -0.0968 -0.4359 0.6142 -0.0566 -0.3862 Enfermedad=NO  
-0.2002 0.2826 -0.1898 -0.0956 -0.102 -0.6002 0.1356 -0.5111 Colegio_U=NO  
0.3197 0.3774 -0.2315 0.1169 -0.3826 0.0781 -0.0888 0.2673 Activo_web=ALTA  
-0.4729 0.056 -0.1949 -0.1105 0.285 -0.0129 0.3546 0.0519 Activo_web=MEDIA  
0.1743 -0.4359 0.435 -0.0014 0.0849 -0.0646 -0.2815 -0.3215 Activo_web=BAJA  
-0.3632 -0.0522 -0.0198 0.4412 0.0976 -0.0989 -0.4539 0.0389 Asistencia  
-0.4566 -0.079 -0.2024 -0.0313 0.0999 0.1164 -0.3265 0.3699 Entregas_completas  
-0.0092 -0.295 -0.1241 0.4425 0.1739 0.345 0.5241 -0.1868 Trabaja=NO
```

Ranked attributes:

```
0.7701 1 -0.473Activo_web=MEDIA-0.457Entregas_completas-0.363Asistencia+0.32 Activo_web=ALTA+0.288Estrato=4...  
0.6065 2 -0.446Estrato=3-0.436Activo_web=BAJA+0.379Sexo=M+0.377Activo_web=ALTA+0.345Estrato=4...  
0.4771 3 0.575Estrato=2-0.441Estrato=3+0.435Activo_web=BAJA+0.265Sexo=M-0.231Activo_web=ALTA...  
0.3553 4 0.611Id+0.443Trabaja=NO+0.441Asistencia+0.294Estrato=4-0.286Estrato=3...  
0.2567 5 0.448Estrato=4-0.436Enfermedad=NO-0.394Id-0.383Activo_web=ALTA+0.285Activo_web=MEDIA...  
0.1779 6 0.614Enfermedad=NO-0.6Colegio_U=NO+0.345Trabaja=NO+0.27 Sexo=M-0.167Id...  
0.105 7 0.524Trabaja=NO-0.454Asistencia+0.355Activo_web=MEDIA-0.326Entregas_completas-0.282Activo_web=BAJA...  
0.044 8 -0.511Colegio_U=NO-0.386Enfermedad=NO+0.37 Entregas_completas-0.339Sexo=M-0.321Activo_web=BAJA...
```

Selected attributes: 1,2,3,4,5,6,7,8 : 8

Status

OK Log x0