TRATAMIENTO INTELIGENTE DE DATOS

Ernesto Serrano Collado

1.

INTRODUCCIÓN

Presentación
del dataset


España
Australia

18.865 W.C.

Whoa! That’s a big number, aren’t you proud?

NATIONAL PUBLIC TOILET MAP

Dataset con información detallada de más de 17.000 baños públicos y privados en toda Australia.

Algunos de los datos

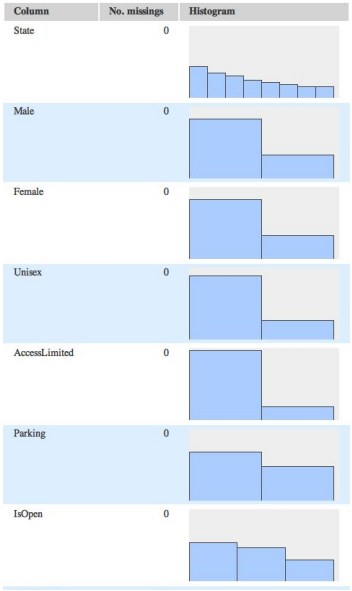
- Nombre del baño.
- Dirección.
- Latitud y longitud.
- Características generales del baño.
- Ubicación.
- Accesibilidad.
- Horas de apertura.
- Características adicionales (por ejemplo, duchas, instalaciones para cambiar bebés, etc.).
- Notas (por ejemplo, duchas accionadas por monedas, etc.).

https://data.gov.au/dataset/national-public-toilet-map

2.

PRE-PROCESAMIENTO

ESTADÍSTICAS



Con el nodo de estadísticas de KNIME podemos hacernos una primera aproximación de los datos de nuestro dataset

State	Male	Female	Unisex	AccessLimited	Parking	IsOpen	BabyChange	Showers	DrinkingWater	SharpsDisposal	SanitaryDisposal	Status
No. missings: 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0	**No. missings:** 0
Top 20: New South Wales : 6244 Victoria : 4233 Queensland : 3546 Western Australia : 2112 South Australia : 1523 Tasmania : 823 Northern Territory : 206 Australian Capital Territory : 178	**Top 20:** true : 15013 false : 3852	**Top 20:** true : 14998 false : 3867	**Top 20:** false : 16252 true : 2613	**Top 20:** false : 18182 true : 683	**Top 20:** false : 11606 true : 7259	**Top 20:** AllHours : 8616 Variable : 6946 DaylightHours : 3303	**Top 20:** false : 16642 true : 2223	**Top 20:** false : 17762 true : 1103	**Top 20:** false : 17327 true : 1538	**Top 20:** false : 16045 true : 2820	**Top 20:** false : 15646 true : 3219	**Top 20:** Verified : 14614 Unverified : 4251
Bottom 20:	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**	**Bottom 20:**

CÁLCULO DE DISTANCIAS

Fórmula del haversine o semi verséneo

$$\operatorname{hav}\left(\frac{d}{r}\right) = \operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)$$

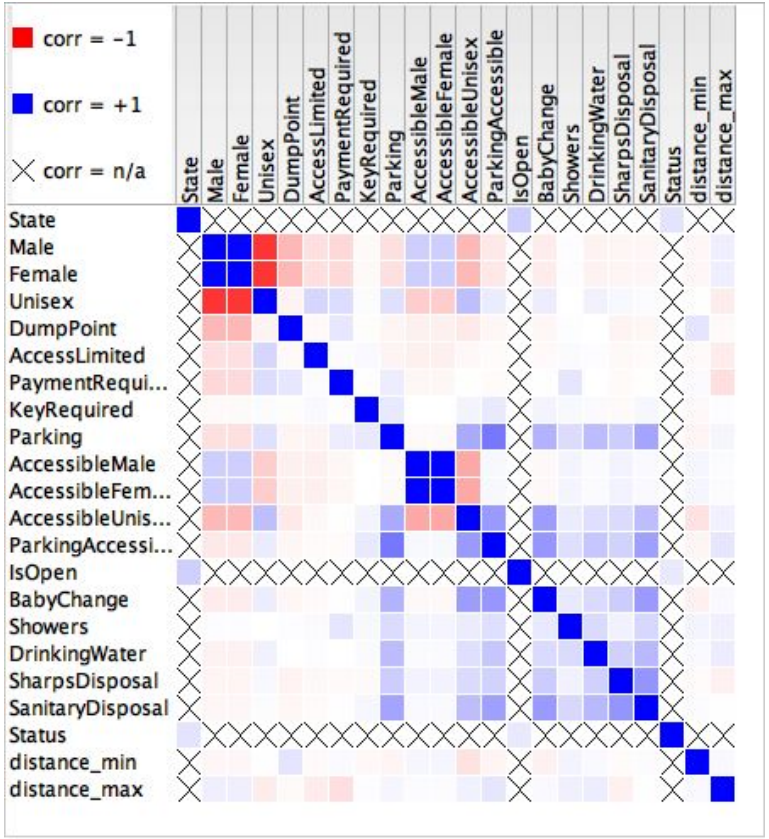
- **d** es la distancia entre dos puntos (sobre un círculo máximo de la esfera, véase distancia esférica),
- **R** es el radio de la esfera, en este caso **6371** que es el radio en kilómetros de la tierra,
- φ_1 φ_2 latitud del punto 1 y latitud del punto 2 en radianes,
- λ_1 λ_2 longitud del punto 1 y longitud del punto 2 en radianes.

2.

ANÁLISIS DESCRIPTIVO

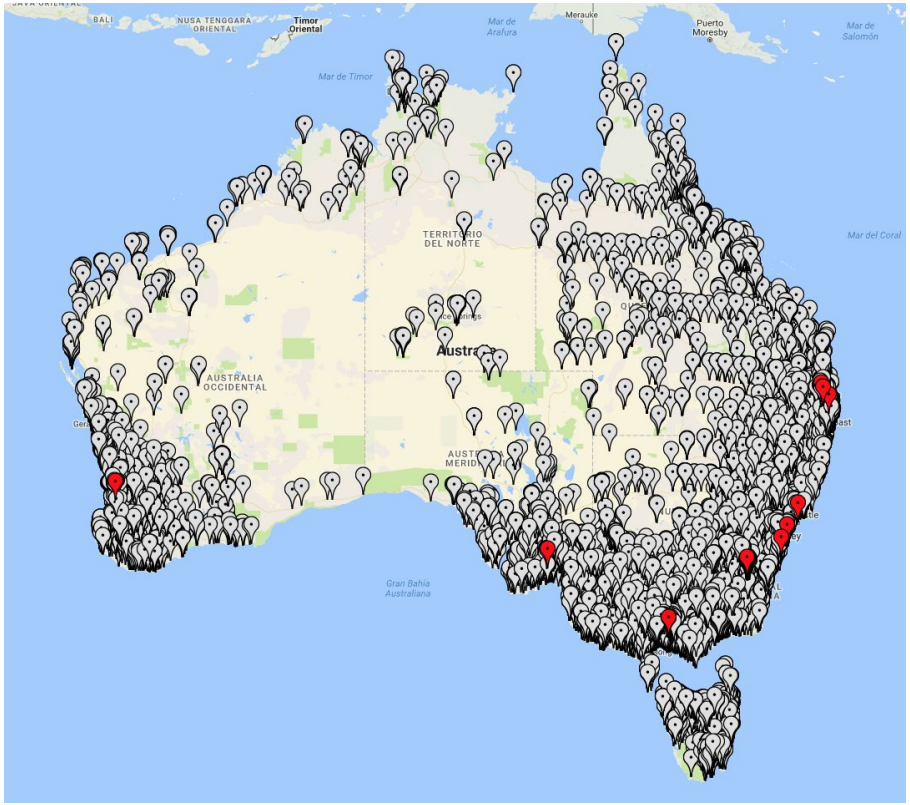
Análisis descriptivo

CORRELACIÓN



- Correlación entre Male y Female.
- Correlación entre AccesibleMale y AccessibleFemale.
- Correlación entre Status y IsOpen.
- Correlación negativa entre Unisex y Male/Female.
- Correlación negativa entre AccesibleUnisex y AccesibleMale/AccessibleFemale

MAPA GENERAL



Podemos ver como los baños se distribuyen sobre los grandes núcleos de población con grandes zonas despobladas por el centro del país.

MAPA POR ESTADOS

Se aprecia la diferencia entre las distintas provincias conteniendo Sidney casi 11.000 de un total de 17.000 y contando los Territorios del norte con tan solo 211

MAPA POR ESTADOS - ADULT CHANGE

Vemos la poca cantidad de vestidores para adultos, de nuevo se aprecia mucha diferencia entre los distintos estados.

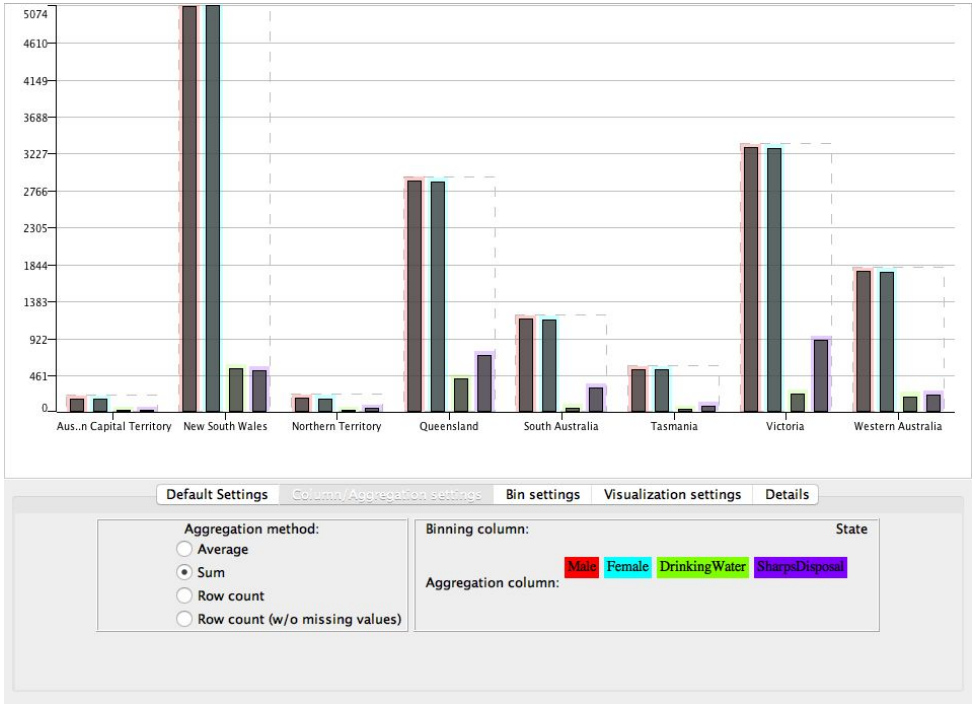
MAPA POR ESTADOS - SHARP DISPOSAL

Curiosamente hay un mayor número de baños públicos con punto de eliminación segura de agujas que cambiadores para adultos.

MAPA POR ESTADOS - DRINKABLE

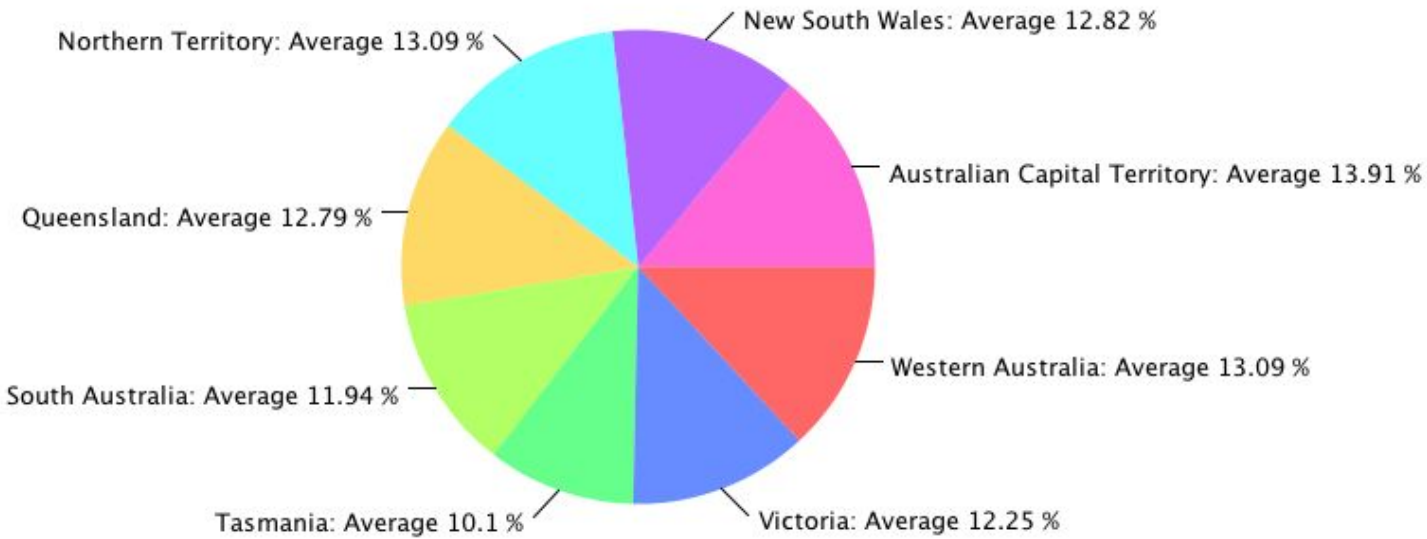
También resulta curioso ver como los baños con agua para consumo humano son menores que los que permiten desechar jeringuillas

HISTOGRAMA POR ESTADOS

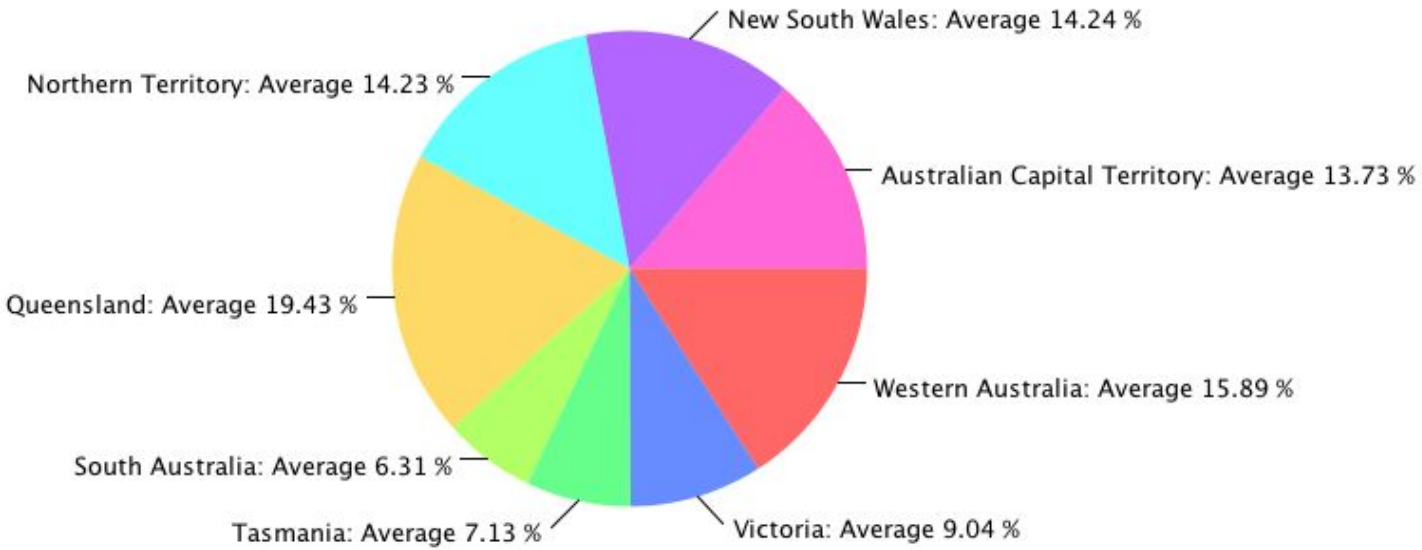


Como podemos ver la distribución de los elementos es uniforme por cada estado.

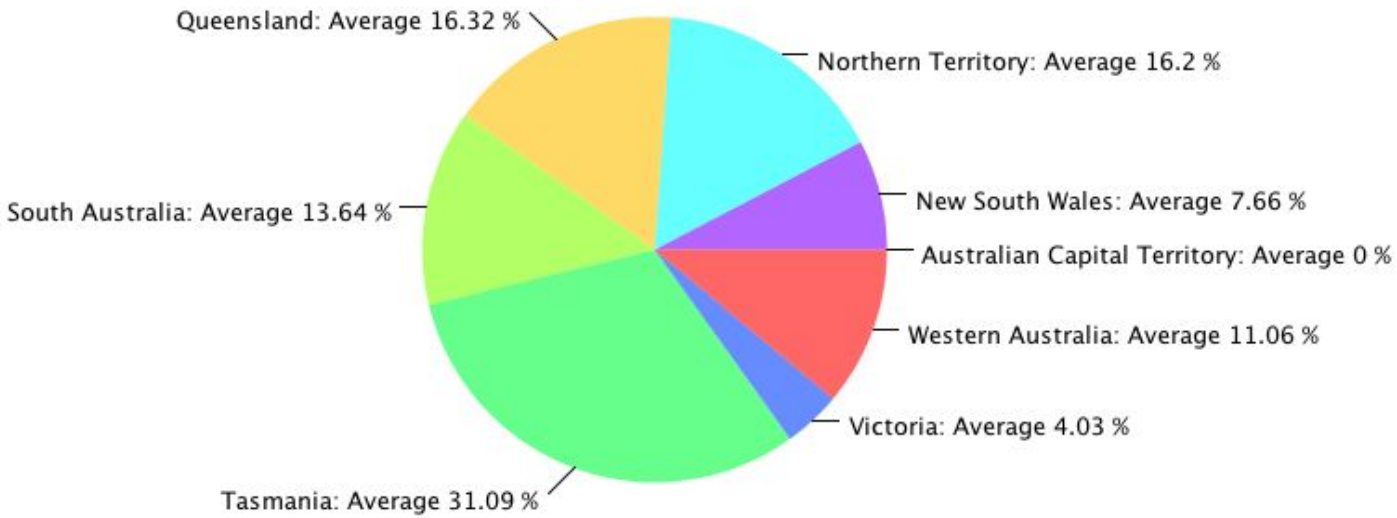
ASEO PARA MUJERES POR ESTADOS



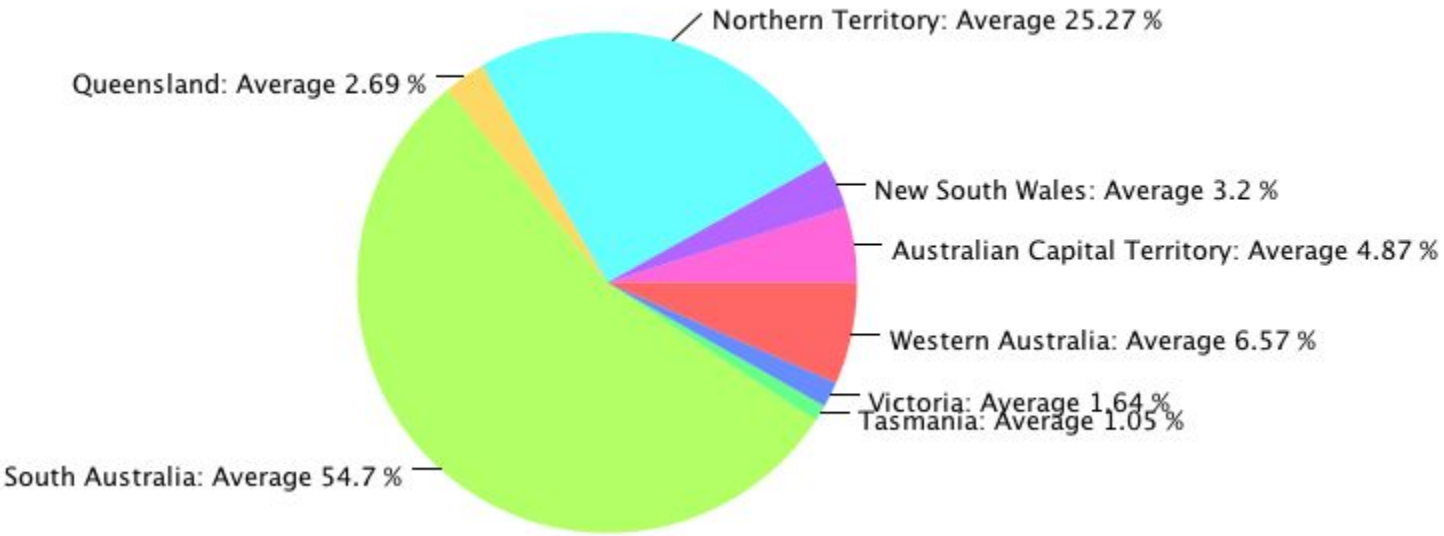
AGUA POTABLE POR ESTADOS



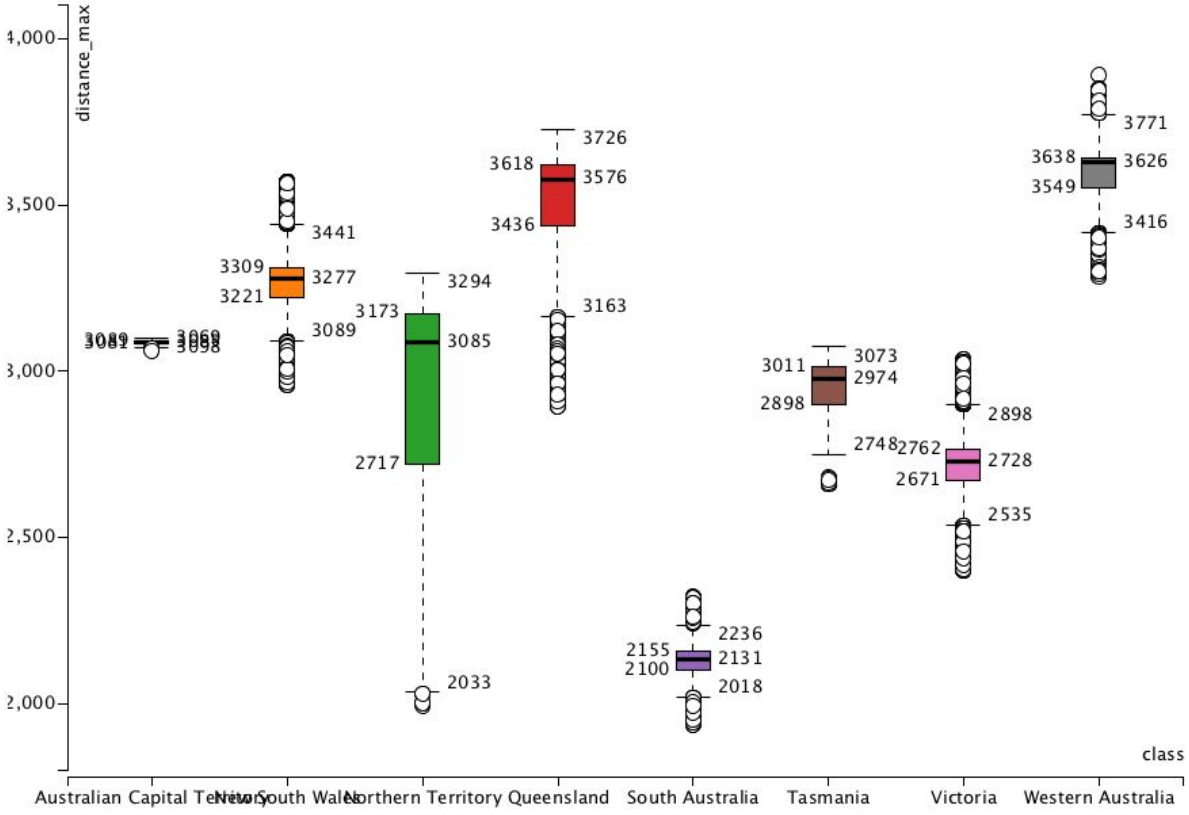
ELIMINACION RESIDUOS POR ESTADOS



ASEO DE PAGO POR ESTADOS



DISTANCIA MAXIMA POR ESTADOS



CLUSTERING

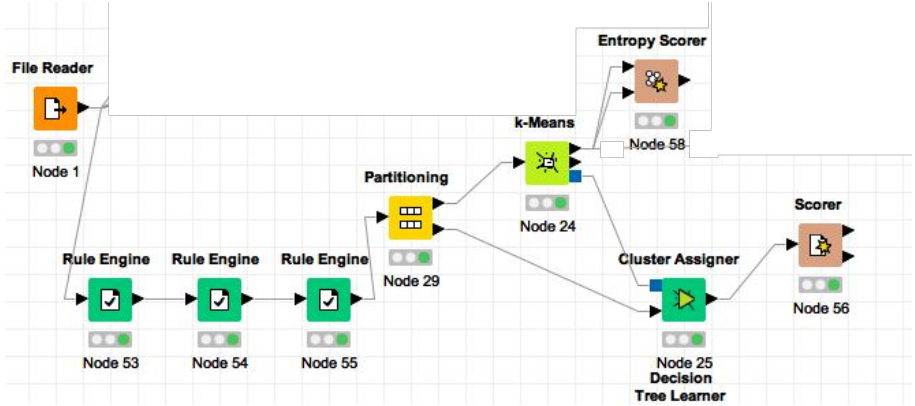
Método de las **K-medias** directamente por lo que hay que agregar algunos nodos **RuleEngine** para poder convertir los valores a enteros. Usando dos campos que tiene correlación como son **SharpsDisposal** y **SanitaryDisposal** y vemos como los clasifica correctamente.

SharpsDisposal \ SanitaryDisposal	1	0
1	1522	1283
0	1678	14282

Correct classified: 15,804 Wrong classified: 2,961

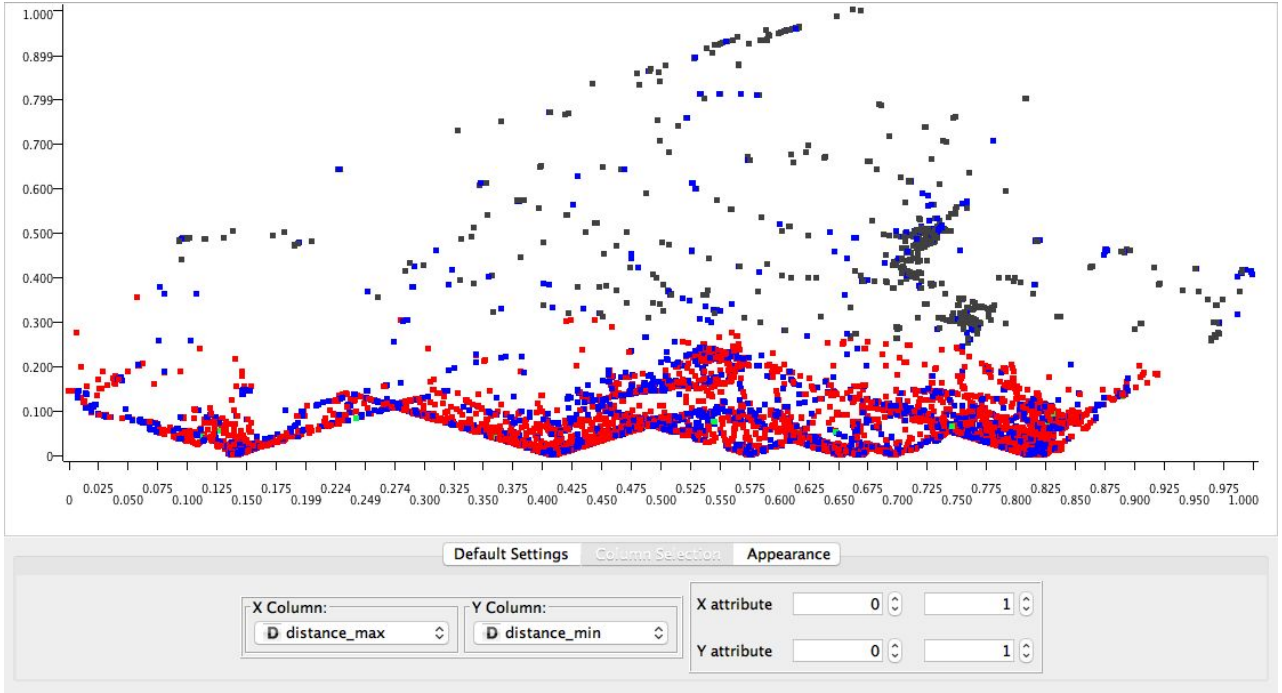
Accuracy: 84.221 % Error: 15.779 %

Cohen's kappa (κ) 0.413



CLUSTERING

K-means entre distancia máxima y mínima

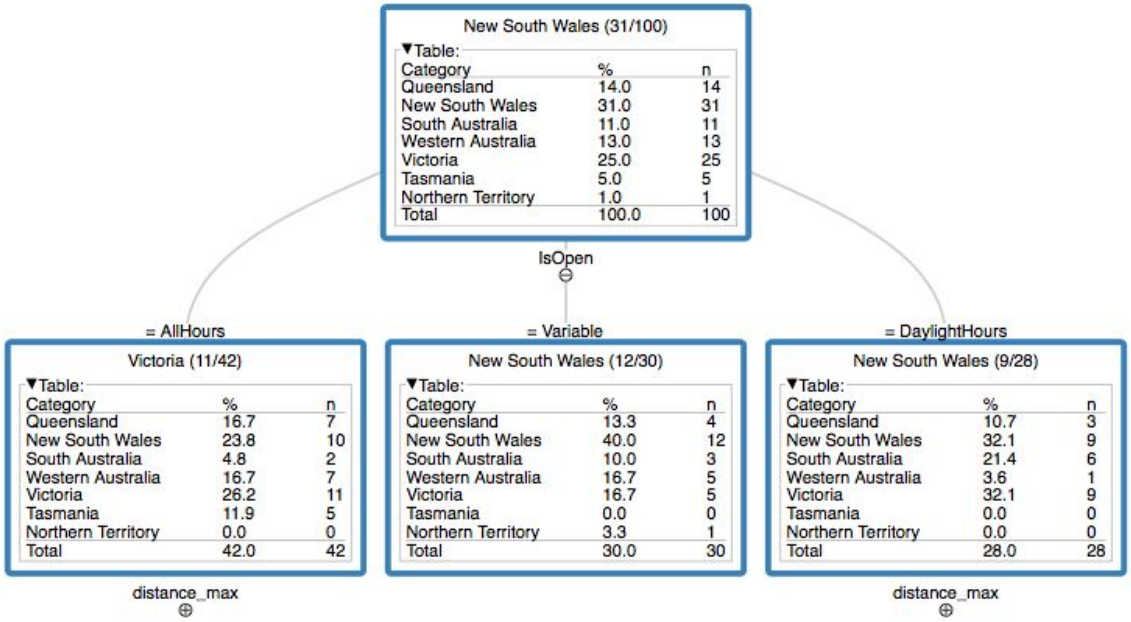


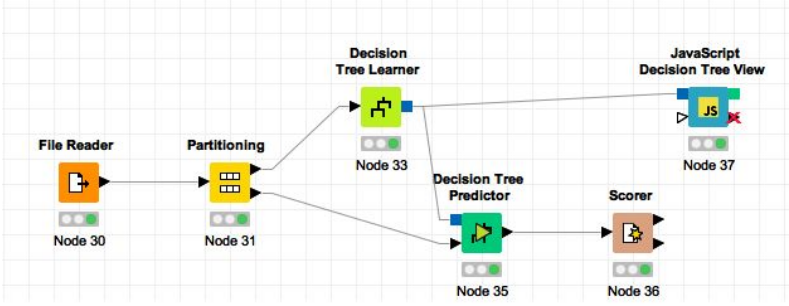
3.

ANÁLISIS PREDICTIVO

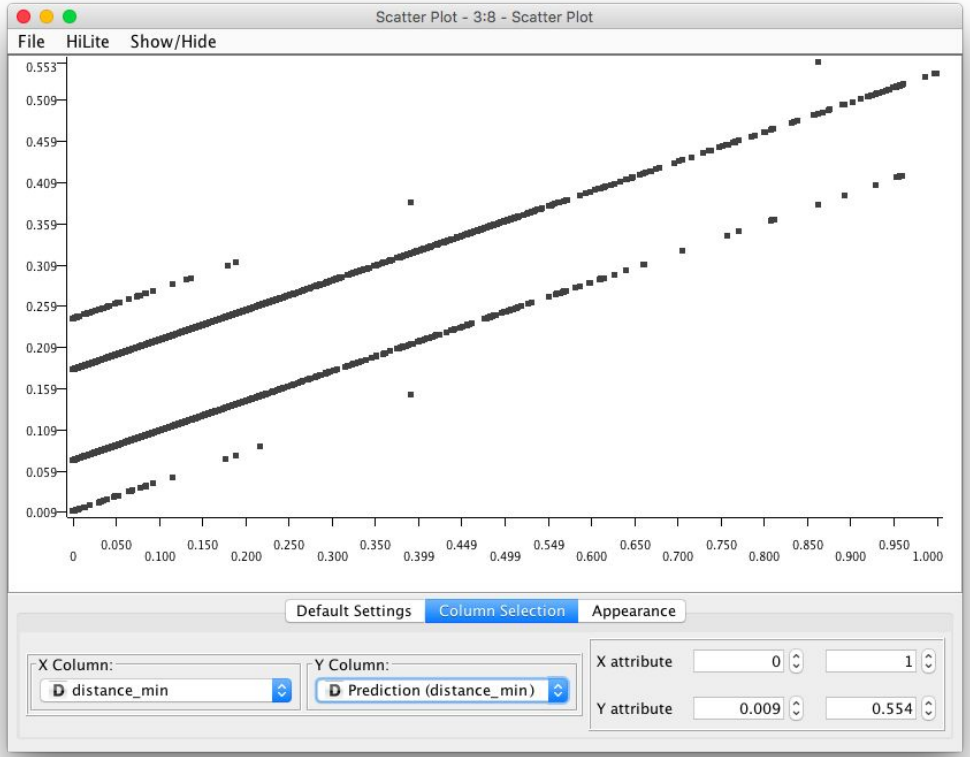
Análisis predictivo

CLASIFICACIÓN





CLASIFICACIÓN



REGRESIÓN

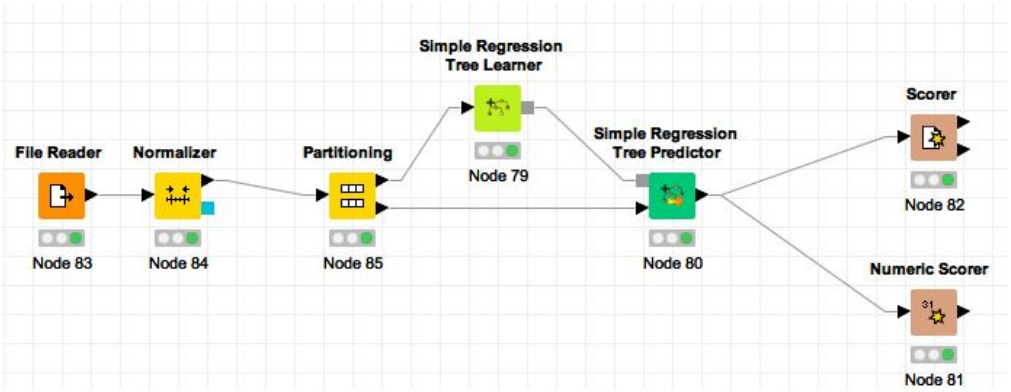
R^2:	-0.572
Mean absolute error:	0.2
Mean squared error:	0.2
Root mean squared error:	0.447
Mean signed difference:	0.031

SharpsDisposal \ Prediction (SharpsDisposal)	1.0	0.0
1.0	1222	1583
0.0	2167	13793

Correct classified: 15,015 Wrong classified: 3,750

Accuracy: 80.016 % Error: 19.984 %

Cohen's kappa (κ) 0.276



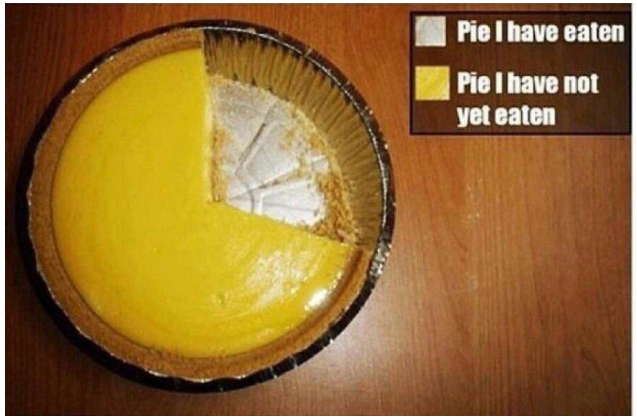
4.

CONCLUSIONES

Conclusiones
finales

CONCLUSIONES

Aunque el dataset parecía muy interesante de entrada, el uso de variables booleanas y la uniformidad de los datos en todos los aspectos han hecho que no se pueda extraer demasiada información con las técnicas aprendidas en la asignatura



MUCHAS GRACIAS!

Alguna pregunta?

Podeis contactarme en **erseco@correo.ugr.es**