Tratamiento Inteligente de Datos

National Public Toilet Map

Autor

Ernesto Serrano Collado

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

Version 0.1 (8/1/2018)

Índice

Introducción

El *Mapa nacional de baños públicos* muestra la ubicación de más de 17.000 baños públicos y privados en toda Australia. Los detalles de las instalaciones sanitarias también se pueden encontrar a lo largo de las principales rutas de viaje y también para viajes más cortos. Se proporciona información útil sobre cada baño, como la ubicación, el horario de apertura, la disponibilidad de habitaciones para bebés, la accesibilidad para personas con discapacidades y los detalles de otros baños cercanos.

Es un *dataset* curioso y perfecto para practicar minería de datos para la asignatura *Tratamiento Inteligente de Datos*del *Master Profesional en Ingeniería Informática*. El *dataset* se puede extraer libremente desde la web: https://data.gov.au/dataset/national-public-toilet-map y tiene licencia Creative Commons.

Dataset

El *dataset* escogido contiene las estadísticas de los 17.000 baños públicos y privados en toda Australia. Las estadísticas cuentan entre otros con los siguientes datos:

1. toilet name.
2. address.
3. latitude and longitude.
4. general toilet features.
5. location.
6. accessibility.
7. opening hours.
8. additional features (e.g. showers, baby change facilities etc).
9. notes (e.g. coin operated showers etc).

Casi todos los datos son de tipo booleano, pasamos a continuación a mostrar la lista completa de los campos indicando los que hemos omitido así como los distintos tipos de datos que hemos extraido de ellos.

- ToiletID *(integer)*
- URL *(string)*
- Name *(string)*
- Address1 *(string)*
- Town *(string)*
- State *(string)*

- Postcode *(integer)*
- AddressNote *(string)*
- Male *(boolean)*
- Female *(boolean)*
- Unisex *(boolean)*
- DumpPoint *(boolean)*
- FacilityType *(string)*
- ToiletType *(string)*
- AccessLimited *(boolean)*
- PaymentRequired *(boolean)*
- KeyRequired *(boolean)*
- AccessNote *(string)*
- Parking *(boolean)*
- ParkingNote *(string)*
- AccessibleMale *(boolean)*
- AccessibleFemale *(boolean)*
- AccessibleUnisex *(boolean)*
- AccessibleNote *(boolean)*
- MLAK *(boolean)* MLA Key (acceso con código)
- ParkingAccessible
- AccessibleParkingNote *(string)*
- Ambulant *(boolean)* (baño portail)
- LHTransfer *(boolean)*
- RHTransfer *(boolean)*
- AdultChange *(boolean)*
- IsOpen *(string)*
- OpeningHoursSchedule *(string)*
- OpeningHoursNote *(string)*
- BabyChange *(boolean)*
- Showers *(boolean)*
- DrinkingWater *(boolean)*
- SharpsDisposal *(boolean)* Eliminación segura de agujas
- SanitaryDisposal *(boolean)* Eliminación segura de productos sanitarios
- IconURL *(string)*
- IconAltText *(string)*
- Notes *(string)*
- Status *(string)*
- Latitude *(double)*
- Longitude *(double)*

Datos adicionales

Para poder saber la proximidad de los baños publicos a los principales nucleos de població hemos tenido que hacer uso del problema del par de puntos mas cercanos así como extraer la lista de las principales ciudades de australia para calcular la distancia hacia estas.

El problema de los puntos mas cercanos lo vimos en el siguiente artículo:

- https://en.wikipedia.org/wiki/Closest_pair_of_points_problem

La lista de las principales ciudades de australia la hemos extraido de los siguientes sitios:

- https://en.wikipedia.org/wiki/List_of_cities_in_Australia_by_population
- http://www.geonames.org/AU/largest-cities-in-australia.html

Población de australia (2016)

Name	State	Population	Latitude	Longitude
Sydney	New South Wales	4,627,345	-33.868	151.207
Melbourne	Victoria	4,246,375	-37.814	144.963
Brisbane	Queensland	2,189,878	-27.468	153.028
Perth	Western Australia	1,896,548	-31.952	115.861
Adelaide	South Australia	1,225,235	-34.929	138.599
Gold Coast	Queensland	591,473	-28	153.431
Canberra	ACT	367,752	-35.283	149.128
Newcastle	New South Wales	308,308	-32.927	151.776
Wollongong	New South Wales	292,190	-34.424	150.893
Logan City	Queensland	282,673	-27.639	153.109

Para calcular la distancia de cada baño público hacia las principales ciudades se ha utilizado la *fórmula del haversine o semiverseno* que es una importante ecuación para la navegación astronómica, en cuanto al cálculo de la distancia de círculo máximo entre dos puntos de un globo sabiendo su longitud y su latitud.

Mas información: https://en.wikipedia.org/wiki/Haversine_formula

La fórmula dice que para cualquier par de puntos sobre una esfera:

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

donde:

- **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,
- lλ_1 λ_2 longitud del punto 1 y longitud del punto 2 en radianes.

Se ha aplicado esa fórmula utilizando una hoja de cálculo para poder aplicarlo a cada baño público cruzándolo con los datos de las principales ciudades

```
= ACOS(COS(RADIANS(90-Latitude_1))
- COS(RADIANS(90-Latitude_2))
+ SIN(RADIANS(90-Latitude1)) SIN(RADIANS(90-Latitude_2))
- COS(RADIANS(Longitude1-Longitude2))
* 6371
```

Una vez obtenidos estos datos adicionales se han agregado al dataset indicando tanto la mínima distancia como la máxima a cualquiera de las 10 principales ciudades de Australia con lo que tenemos las siguientes columnas adicionales:

- Sidney *(integer)*
- Melbourne *(integer)*
- Brisbane *(integer)*
- Perth *(integer)*
- Adelaide *(integer)*
- Gold Coast *(integer)*
- Canberra *(integer)*
- Newcastle *(integer)*
- Wollongong *(integer)*
- Logan City *(integer)*
- distance_min *(integer)*
- distance_max *(integer)*

Objetivos

El objetivo principal es ver si hay mayor cantidad de urinarios públicos cerca de los núcleos urbanos, y si los servicios que ofrecen los mismos se ven incrementados por la proximidad.

Minería de datos

Se ha decidido realizar el trabajo utilizando la herramienta KNIME debido a la sencillez de uso al ser una herramienta visual que hace muy sencillo el poder ir probando las distintas técnicas.

Pre-procesamiento de datos

Antes de aplicar las técnicas aprendidas en la asignatura realizaremos un pre-procesado de los datos mediante técnicas de estadística descriptiva con el objetivo de conocer nuestro *dataset* para poder utilizar posteriormente las distintas técnicas.

Lo primero de todo hemos agregado un nodo *FileReader* que hemos configurado para leer nuestro dataset, en el mismo se han configurado los tipos de columna ya que por defecto *KNIME* lo lee todo automáticamente como cadenas y muchos de nuestros datos son de tipo booleano y de tipo numérico. Además se han omitido desde el propio *FileReader* algunas columnas que contenían textos que no nos resultan útiles para el procesamientos

Se ha agregado un nodo de estadísticas para ver una primera aproximación visual de los datos que tenemos.

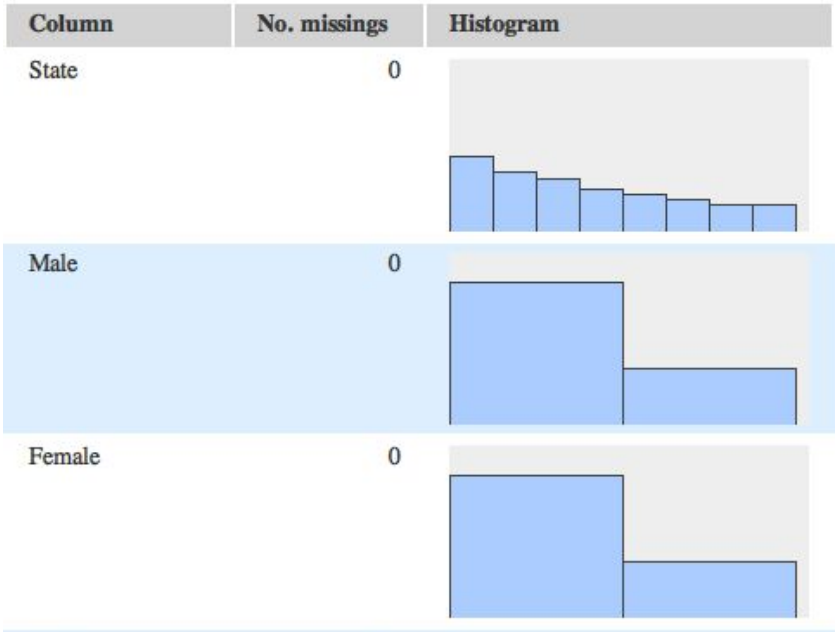


Figura: Estadísticas Nominales

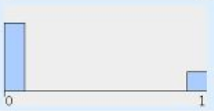
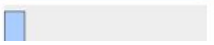
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Male	0.0	0.7958	1	1	0.4031	-1.4678	0.1544	0	0	0	
Female	0.0	0.795	1	1	0.4037	-1.4617	0.1366	0	0	0	
Unisex	0.0	0.1385	0.0	1	0.3454	2.0931	2.3814	0	0	0	
DumpPoint	0.0	0.028	0.0	1	0.1649	5.7239	30.7664	0	0	0	
AccessLimited	0.0	0.0362	0.0	1	0.1868	4.9661	22.6647	0	0	0	
PaymentRequired	0.0	0.0086	0.0	1	0.0923	10.6526	111.4891	0	0	0	
KeyRequired	0.0	0.022	0.0	1	0.1467	6.5182	40.4914	0	0	0	
Parking	0.0	0.3848	0.0	1	0.4866	0.4736	-1.7759	0	0	0	
AccessibleMale	0.0	0.22	0.0	1	0.4142	1.3521	-0.1719	0	0	0	
AccessibleFemale	0.0	0.2205	0.0	1	0.4146	1.3487	-0.181	0	0	0	
AccessibleUnisex	0.0	0.2915	0.0	1	0.4545	0.9177	-1.1579	0	0	0	
ParkingAccessible	0.0	0.2084	0.0	1	0.4062	1.4362	0.0626	0	0	0	
BabyChange	0.0	0.1178	0.0	1	0.3224	2.3708	3.6211	0	0	0	
Showers	0.0	0.0585	0.0	1	0.2346	3.764	12.169	0	0	0	
DrinkingWater	0.0	0.0815	0.0	1	0.2736	3.0588	7.357	0	0	0	

Figura: Estadísticas numéricas

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:**

Figura: Resume estadísticas

Para comprobar las observaciones realizadas mediante los histogramas pasamos a realizar una correlación lineal entre las variables con el objetivo de encontrar que las variables observadas tienen cierta correlación con las distancias agregadas y además comprobar si hay algunas variables con una correlación muy alta lo que puede indicar que se derivan unas de otras y se pueden eliminar del dataset al aportar la misma información.

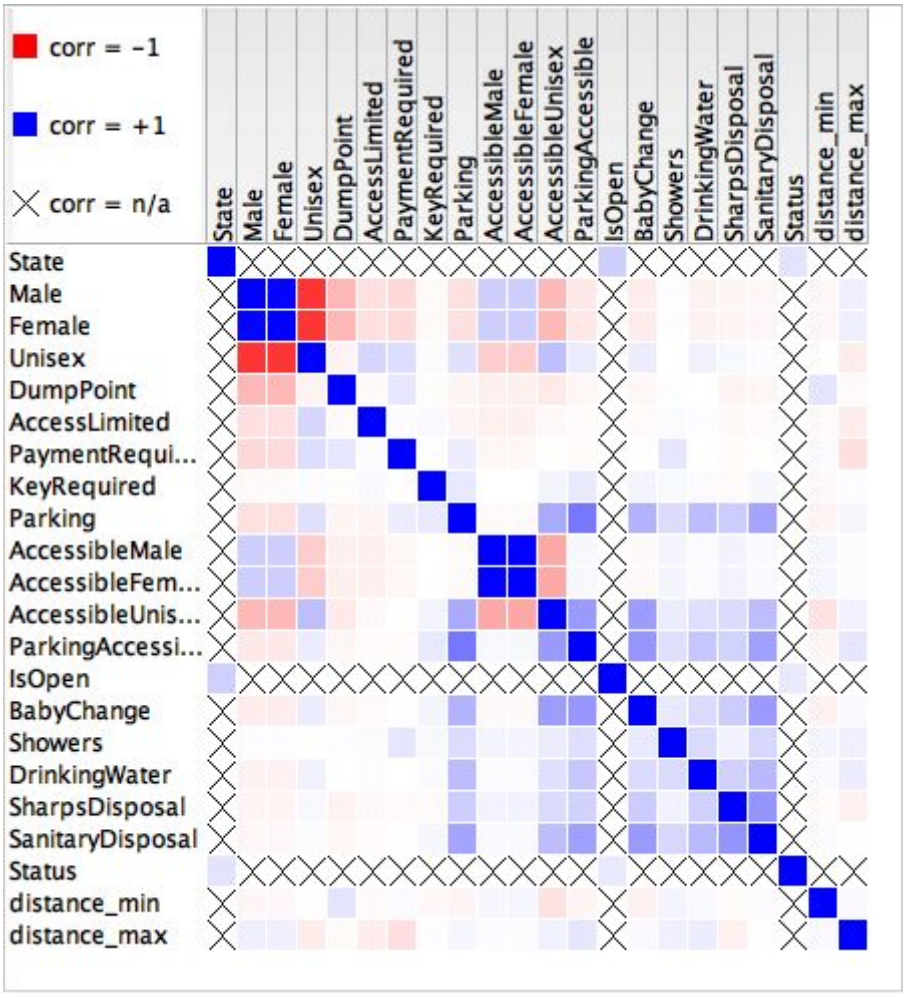


Figura: Correlación

Como se puede apreciar no existen muchas X en esta matriz lo que indica que hay correlación entre las variables.

Las correlaciones más importantes que se observan son:

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

Análisis descriptivo

Una vez corregidos los datos hemos visualizado los distintos puntos en el mapa para hacernos una idea de la localización de los mismos, los distintos baños publicos aparecen en gris, y las principales ciudades aparecen marcadas en rojo.

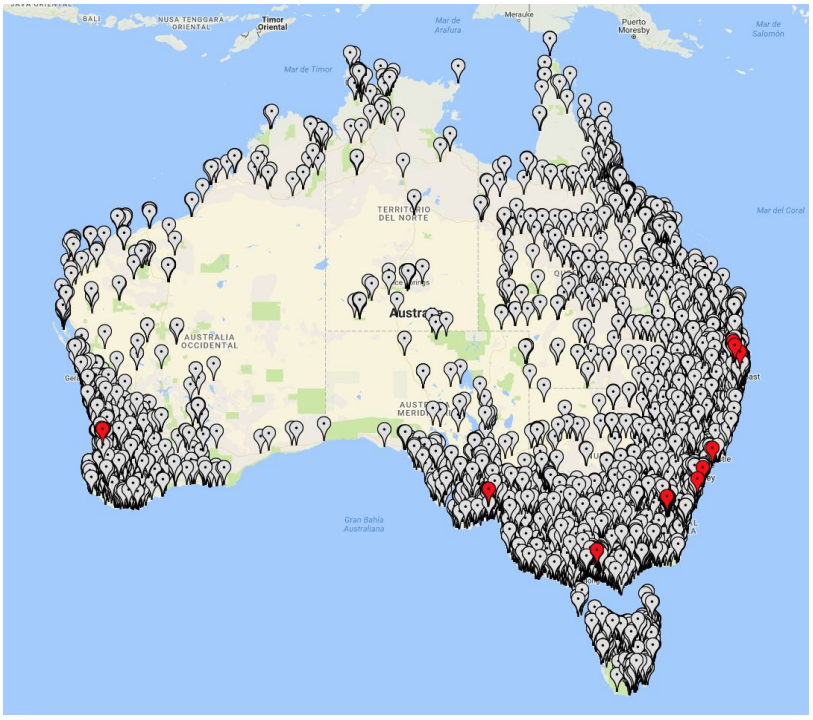


Figura: Mapa general

También hemos visto interesante diferenciar los baños públicos dependiendo del estado, como podemos ver en el siguiente mapa.

Figura: Mapa por estados

Se aprecia como Nueva Gales del Sur, que es donde está Sidney, es la que mayor cantidad de baños tiene con `10591`, y por otro lado los Territorios del Norte solo poseen `211`.

Otro dato curioso es la poca cantidad de baños que cuentan con un cambiador de ropa para adultos. Dato que contrasta con la gran cantidad de baños que cuentan con punto de eliminación segura de agujas, que además es mayor que la cantidad de baños que poseen agua potable.

Figura: Cambiador para adultos (por estados)

Figura: Puntos de eliminación segura de agujas (por estados)

Figura: Agua potable (por estados)

También podemos ver como la representación de las variables es uniforme a lo largo de todo el país en el histograma siguiente:

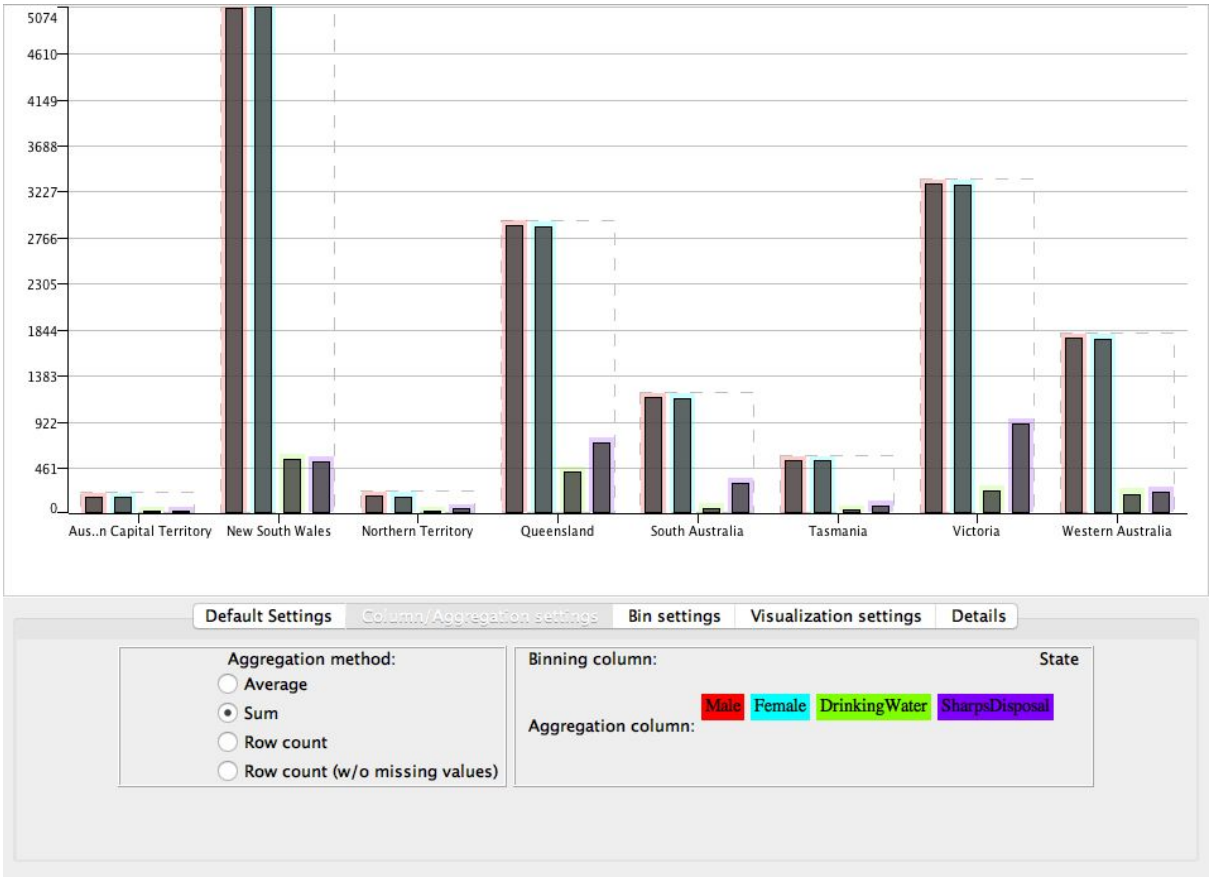


Figura: Histograma por estados

Para el mismo se han usado las columnas *Male*, *Female*, *DrinkingWater* y *SharpsDisposal*.

Podemos ver como el promedio de distribución de aseos para mujeres es uniforme en todo el territorio así como el agua potable, pero Tasmania es el el que tiene un mayor número de puntos de eliminación de residuos para caravanas, esto es porque es un destino común para gente que viaja en caravana. En los territorios del sur es donde mayor número de aseos de pago encontraremos. a continuación se muestran los gráficos donde se aprecian esos detalles

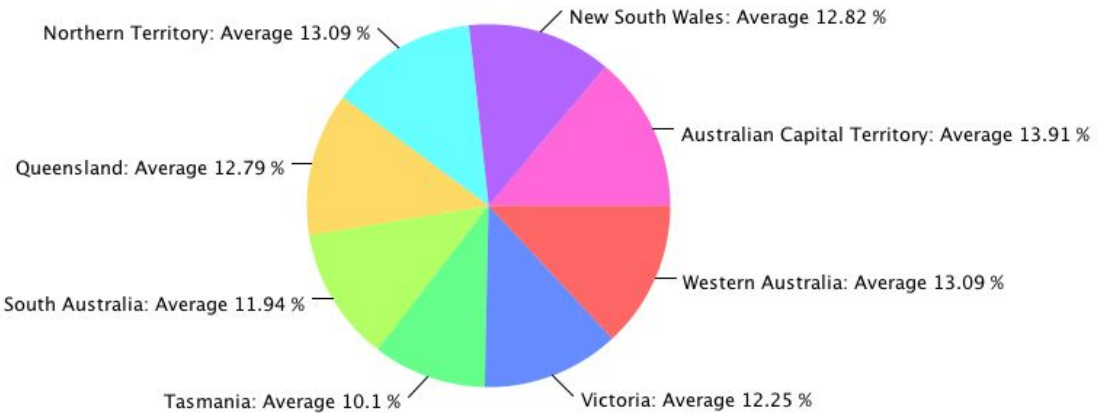


Figura: Aseo para mujeres por estados

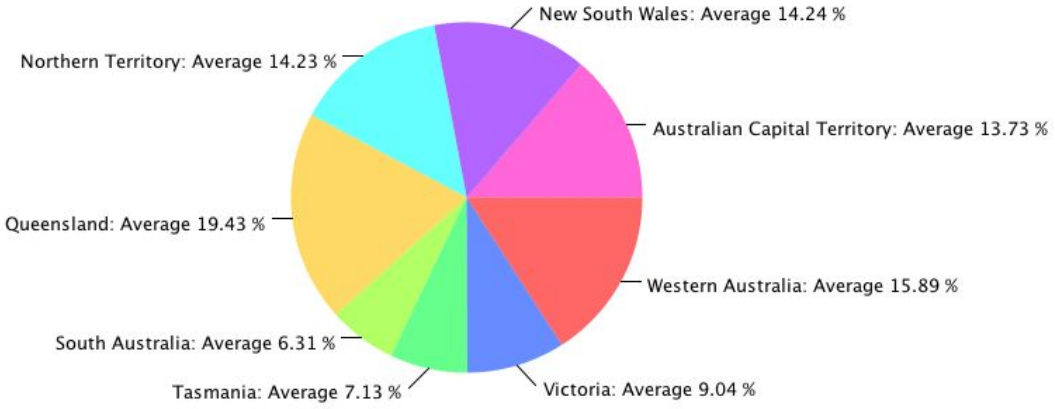


Figura: Agua potable por estados

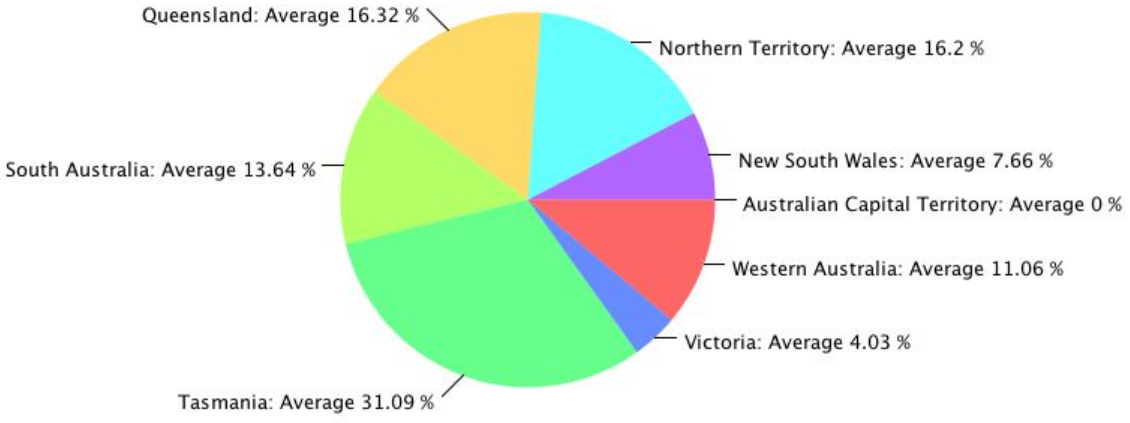


Figura: Papelera por estados

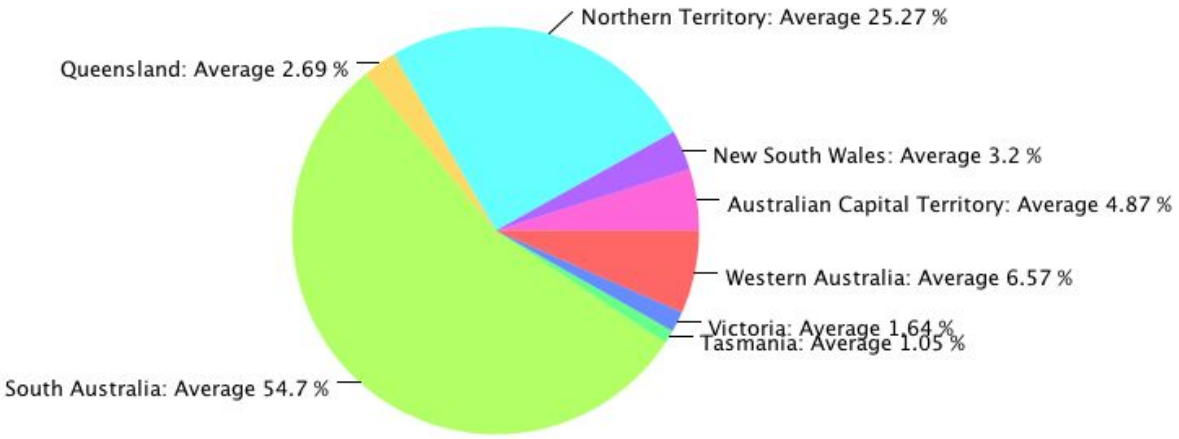


Figura: Aseo de pago por estados

Clustering

Como la mayor parte de las variables son boolean no podemos usar el 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. También se ha realizado el método de las K-Medias con los campos *distance_min* y *distance_max*.

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

Figura: Clustering SharpDisposal

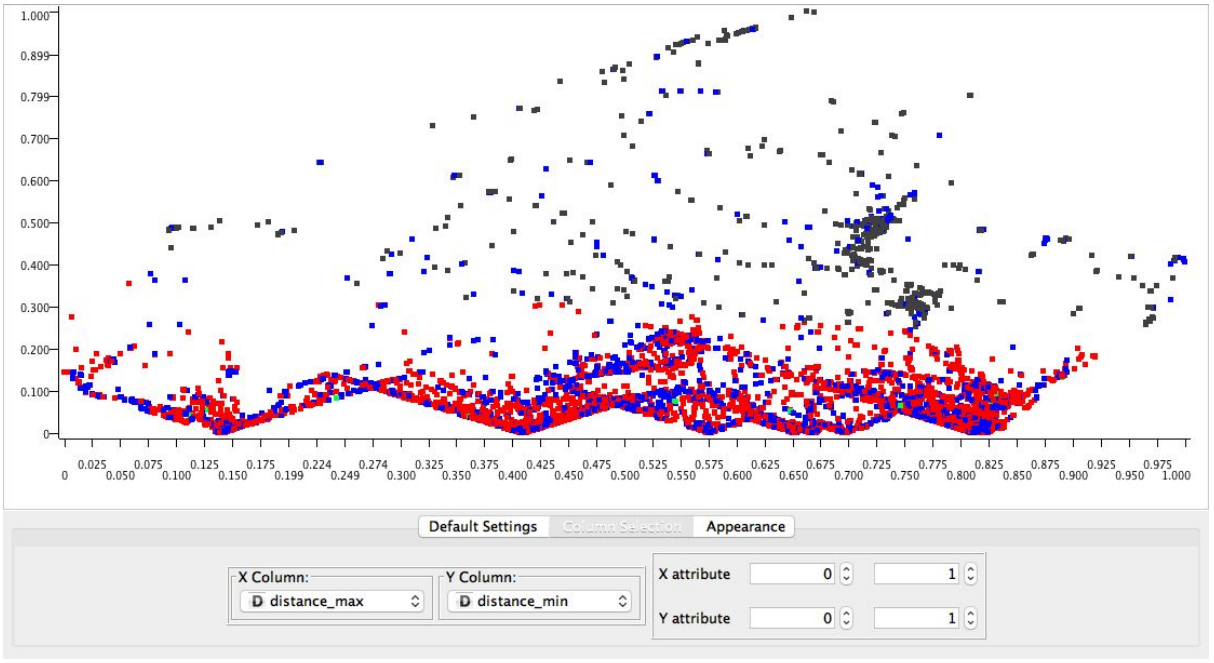


Figura: Clustering distancias

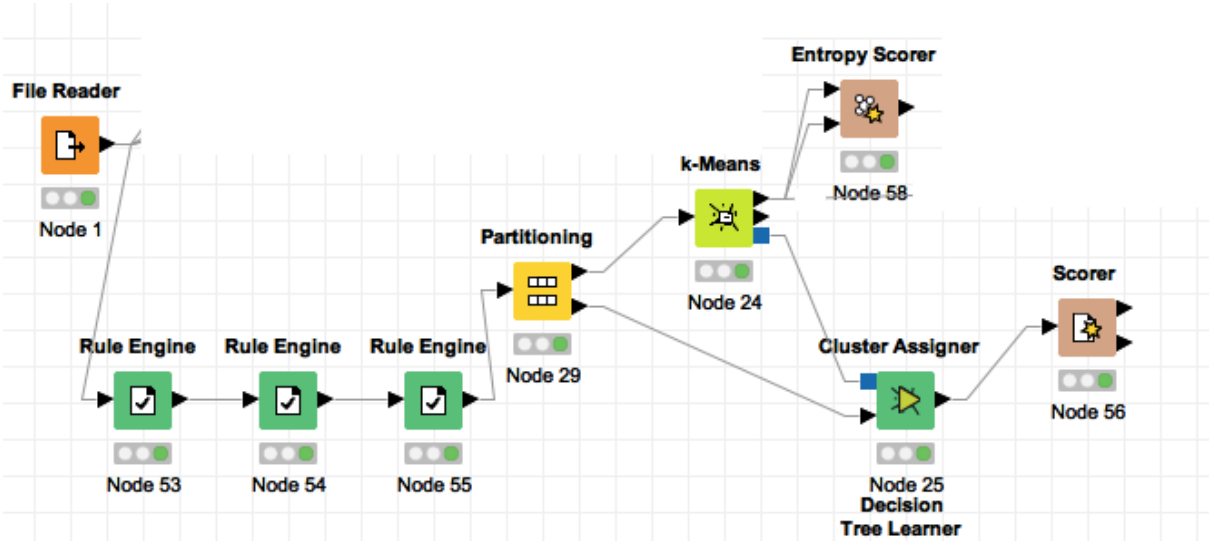


Figura: Configuración Clustering en KNIME

NOTA: No se ha podido realizar un *HierarchicalClustering* por problemas de memoria.

Análisis predictivo

Clasificación

Hemos configurado KNIME para realizar una clasificación básica agregando un *Decision Tree Learner* y viendo el árbol de decisión que ha creado, a continuación se puede como se ha configurado KNIME.

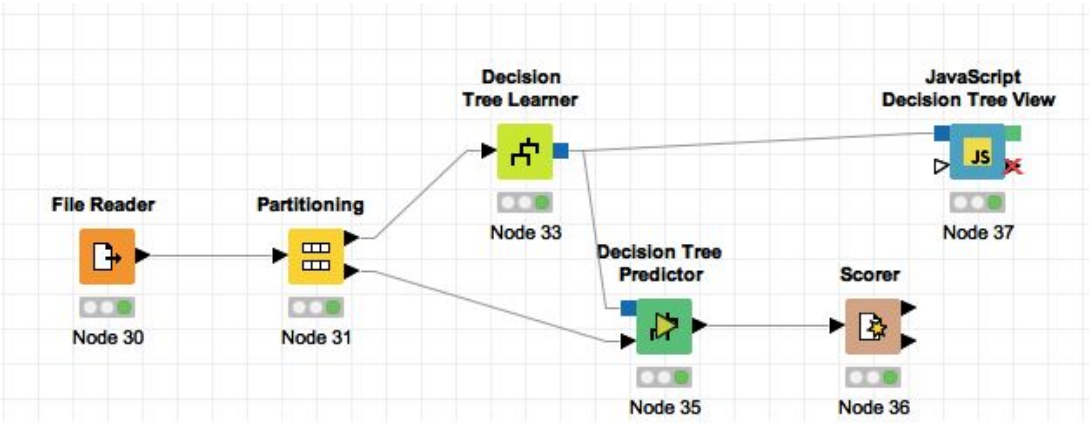


Figura: Configuración para arbol de decisión

Le hemos dicho que aprenda en base al estado y el campos IsOpen que como ya vimos tenían una alta correlación y como podemos ver ha clasificado esos parametros del dataset.

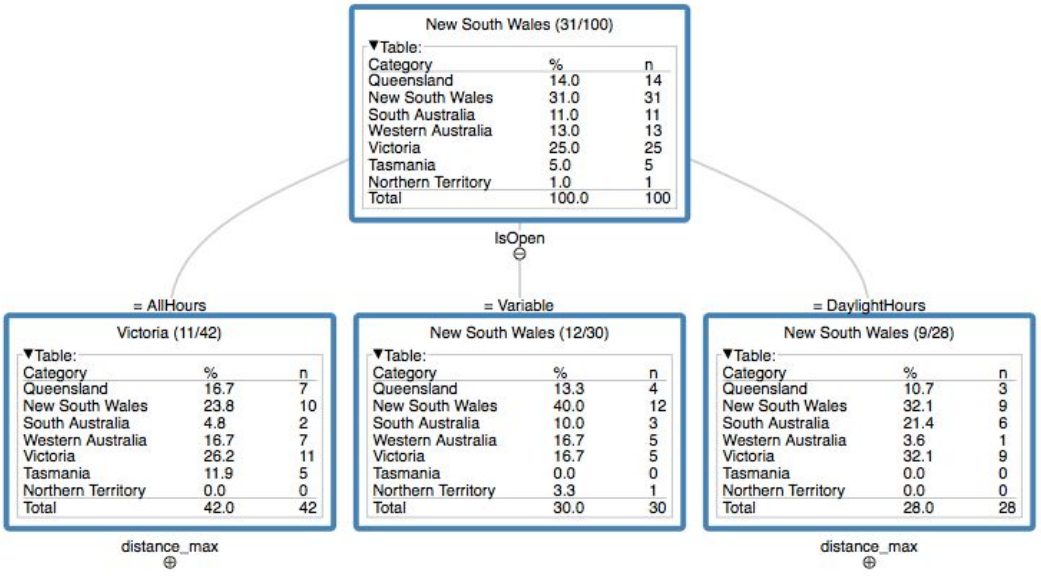


Figura: Arbol de decisión IsOpen por estados

Regresión

En cuanto a la regresión, se ha utilizado un *RegressionTreeLearner* junto a su correspondiente predictor y nos ha dado los resultado que se pueden ver en las imagenes adjuntas. Se ha buscando la regresión en base al campo SharpDisposal.

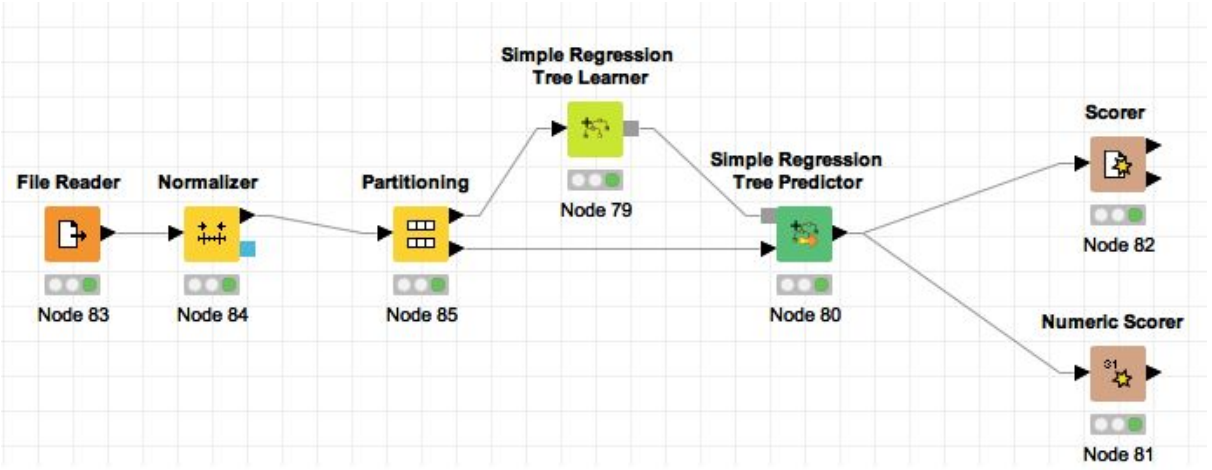


Figura: Configuración para la regresión

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

Figura: Puntuacion de la 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

Figura: Puntuacion numérica de la regresión

Conclusiones

A nivel técnico sobre las herramientas empleadas, he llegado a la conclusión de que aunque KNIME consuma muchos recursos es una herramienta muy util y muy sencilla de utilizar, en cambio R tiene una curva de aprendizaje mas alta pero permite mucha versatilidad en los datos aparte de ser rapidísima.

Respecto a los datos, tras trabajar en profundidad con ellos hemos visto que el dataset elegido no permitía mucho juego al estar basado en variables booleanas sin apenas correlación, pero agregando los datos de distancia he conseguido darle un valor añadido a los datos originales.

En resumen, 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.