

PCT Observer

Tablero de Parques Científicos/Tecnológicos

Contenido

Resumen	2
Summary	2
Descripción	2
Arquitectura del sistema	3
Estructura de archivos estática	4
Estructura de archivos dinámica	4
Instrucciones para el despliegue	5
Estructura de archivos despliegue Docker	5
Configuración y despliegue de PCT Observer	5
Referencias	5
Anexo 1	6
Módulo Estadística Descriptiva	6
Módulo Análisis Bivariado	7
Módulo Series Temporales	8
Módulo Análisis Tridimensional	9
Módulo Árboles de Decisión	10
Módulo Análisis de Agrupamientos	11

Resumen

PCT Observer es una aplicación web para visualizar y analizar datos relacionados con los parques científicos/tecnológicos. Permite descubrir la existencia de diferencias significativas o relaciones entre los indicadores clave, comparar su evolución en el tiempo, y determinar los indicadores más relevantes para caracterizar los diferentes tipos de parques.

Summary

PCT Observer is a web application that allows to analyze and visualize key indicators of the scientific/technological parks. It facilitates the user the discovery of statistically significant differences or relationships between indicators, comparing their time series, and exploring the features that better characterize the different park types.

Descripción

PCT Observer es una aplicación web que permite realizar tareas de análisis de datos relacionados con los parques científicos/tecnológicos. En su versión actual, comprende los siguientes módulos:

- Estadística Descriptiva: permite filtrar los indicadores clave que son de interés, mostrando estadísticas como medias, desviaciones estándar, porcentaje de valores faltantes, etc. para cada tipo de parque o general.
- Análisis Bivariado: permite explorar diferencias significativas entre los valores medios de dos indicadores para cada tipo de parque. Implementa la prueba Brunner-Munzel [1, 2] teniendo en cuenta el tamaño de las muestras.
- Series Temporales: permite analizar visualmente la evolución temporal de un indicador para los diferentes parques o tipos.
- Análisis Tridimensional: diseñado para explorar visualmente patrones entre los valores de tres indicadores simultáneamente para el año, tipos o parques seleccionados.
- Árboles de Decisión: con el módulo es posible entrenar y explorar visualmente diferentes árboles de decisión para clasificación, ayudando a identificar un conjunto de indicadores que explique con precisión la taxonomía de parques científicos. Define el tipo de parque como variable objetivo y un conjunto de indicadores configurables como variables predictivas. Utiliza la versión optimizada del algoritmo CART [3] implementada en scikit-learn¹.
- Análisis de Agrupamientos: ayuda a develar información sobre la estructura de los datos, la importancia de los indicadores y posibles anomalías. Utiliza la versión del algoritmo K-Medias [4]

¹ [sklearn.tree.DecisionTreeClassifier — scikit-learn 1.0.2 documentation](https://scikit-learn.org/stable/modules/tree.html)

de scikit-learn² con el índice Silhouette para evaluar el agrupamiento. Cada variable se escala de acuerdo con los valores máximo y mínimo. Es posible configurar el número de grupos y los indicadores a considerar.

En el Anexo 1 se muestran algunas capturas de pantalla de la aplicación.

Arquitectura del sistema

El tablero implementa una arquitectura cliente-servidor basada en Voila³, Jupyter⁴ y Python.

La arquitectura (Figura 1) consta de dos componentes principales. El **frontend** es el componente con el que interactúa el usuario mediante la interfaz web construida mediante Jupyter y Voila.

El **backend** provee servicios al **frontend**. Gestiona los diferentes eventos y coordina los módulos de Acceso a Datos y Herramientas. El primero provee una capa de abstracción al datalake que alimenta a la aplicación, en su versión actual un archivo estático, pero que puede ser reemplazado fácilmente por una base de datos relacional u de otro tipo. Por otra parte, el módulo de Herramientas implementa las funcionalidades relativas a las herramientas estadísticas y de aprendizaje automático utilizadas.

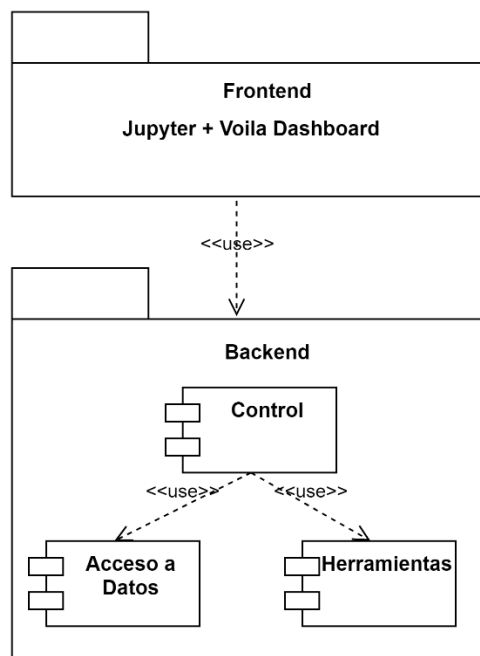


Figura 1: Arquitectura general de PCT Observer

² [sklearn.cluster.KMeans — scikit-learn 1.0.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)

³ Voila: <https://voila.readthedocs.io/en/stable/>

⁴ Jupyter: <https://jupyter.org/>

Estructura de archivos estática

El sistema se distribuye con la siguiente estructura de archivos, requerida para su adecuado funcionamiento.

- 📁 Directorio raíz
 - 📁 app -> fuentes del sistema (frontend y backend)
 - 📄 ExploratoryDataAnalysis.ipynb -> cuaderno Jupyter
 - 📄 cfg.xlsx -> archivo de configuración (indicadores a considerar, alias para los parques científicos, etc.)
 - 📁 src -> fuentes del sistema (frontend y backend)
 - 📁 dao -> código de acceso a datos
 - 📄 __init__.py -> inicialización de módulo Python
 - 📄 data.py -> rutinas de acceso a datos
 - 📁 datatools -> módulo Herramientas
 - 📄 __init__.py -> inicialización de módulo Python
 - 📄 cluster.py -> funcionalidades relacionadas con el análisis de agrupamientos
 - 📄 stat_desc.py -> funcionalidades relacionadas con estadística descriptiva
 - 📄 stat_biv.py -> funcionalidades relacionadas con el análisis bivariado
 - 📄 tree.py -> funcionalidades relacionadas con árboles de decisión
 - 📁 front -> componente frontend del sistema
 - 📁 controllers
 - 📄 __init__.py -> inicialización de módulo Python
 - 📄 app_cluster.py -> controladores módulo de análisis de agrupamiento
 - 📄 app_stat_3d.py -> controladores módulo de análisis tridimensional
 - 📄 app_stat_biv.py -> controladores módulo de análisis bivariado
 - 📄 app_stat_desc.py -> controladores módulo de estadística descriptiva
 - 📄 app_stat_tseries.py -> controladores módulo de series temporales
 - 📄 app_tree.py -> controladores módulo de árboles de decisión
 - 📄 utils.py -> utilidades relativas al frontend
 - 📄 widget_factory.py -> widgets
 - 📁 resources -> iconos u otros elementos multimedia
 - 📁 images -> iconos, imágenes, etc.
 - 📁 utils -> utilidades comunes
 - 📄 __init__.py -> inicialización de módulo Python
 - 📄 utils.py -> utilidades

Estructura de archivos dinámica

Durante la ejecución, se crean varios archivos temporales relacionados con la descarga de datos desde los módulos de estadística descriptiva y árboles de decisión. La carpeta para estos archivos se especifica mediante la variable de entorno `STATIC_CONTENT_PATH`, que puede establecer en el archivo `docker-compose.yml`. Por restricciones de seguridad de Voila, debe situarse en el mismo sub-árbol del sistema de archivos que `ExploratoryDataAnalysis.ipynb`.

- 📁 Directorio raíz
 - 📁 tmp -> archivos temporales para descarga (hojas de cálculo, gráficas, etc.)

Instrucciones para el despliegue

Para facilitar su despliegue, el software se distribuye mediante contenedores Docker gestionado con docker-compose, no obstante, los requerimientos pueden replicarse nativamente.

Estructura de archivos despliegue Docker

📁 Directorio raíz

📁 docker -> archivos relacionados con proyecto Docker

📁 dependencies -> dependencias de software

📁 qgrid -> qgrid⁵, librería para mostrar dataframes de Pandas

📄 index.js-> versión modificada para PCT Observer

📄 .env -> configuración de entorno para docker-compose

📄 docker-compose.yml -> instrucciones de despliegue para docker-compose

📄 dockerfile -> archivo docker para la creación de la imagen

Configuración y despliegue de PCT Observer

Se describe el despliegue mediante docker-compose

- i. se asume ordenador con docker \geq 4.3.0. Opcionalmente configurar puerto (defecto 8866), carpeta de archivos temporales, etc.
 - navegar a directorio docker (ver Estructura de archivos despliegue Docker)
 - docker-compose up -d
- ii. acceder al tablero navegando a <http://localhost:8866>

Referencias

[1] Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 42(1), 17-25.

[2] Neubert, K., & Brunner, E. (2007). A studentized permutation test for the non-parametric Behrens-Fisher problem. *Computational Statistics & Data Analysis*, 51(10), 5192-5204.

[3] Breiman, L., Friedman, G. H., & Olshen, R. A. (2017). *Classification and regression trees* Routledge.

[4] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

⁵ <https://github.com/quantopian/qgrid/>

Anexo 1

Pantallas PCT Observer

Módulo Estadística Descriptiva

The screenshot shows a web browser window with the URL localhost:8866. The page title is "Parques Científicos/Tecnológicos". The interface includes a navigation menu with options like "Estadísticas General", "Análisis por Indicador", "Indic. vs Antigüedad", "Indic. Inter. 3 Niveles", "Árbol Decisión", and "Agrupamientos". A "Rango a considerar:" slider is set to "2015-2019". A list of indicators is shown on the left, with "dimension" selected. The main content area displays four summary tables for different categories: "General", "SP", "TP", and "HP". Each table lists variables such as age, average-size-1, average-size-2, billing, companies, and companies-allocated, along with their respective count, mean, standard deviation, maximum, and minimum values.

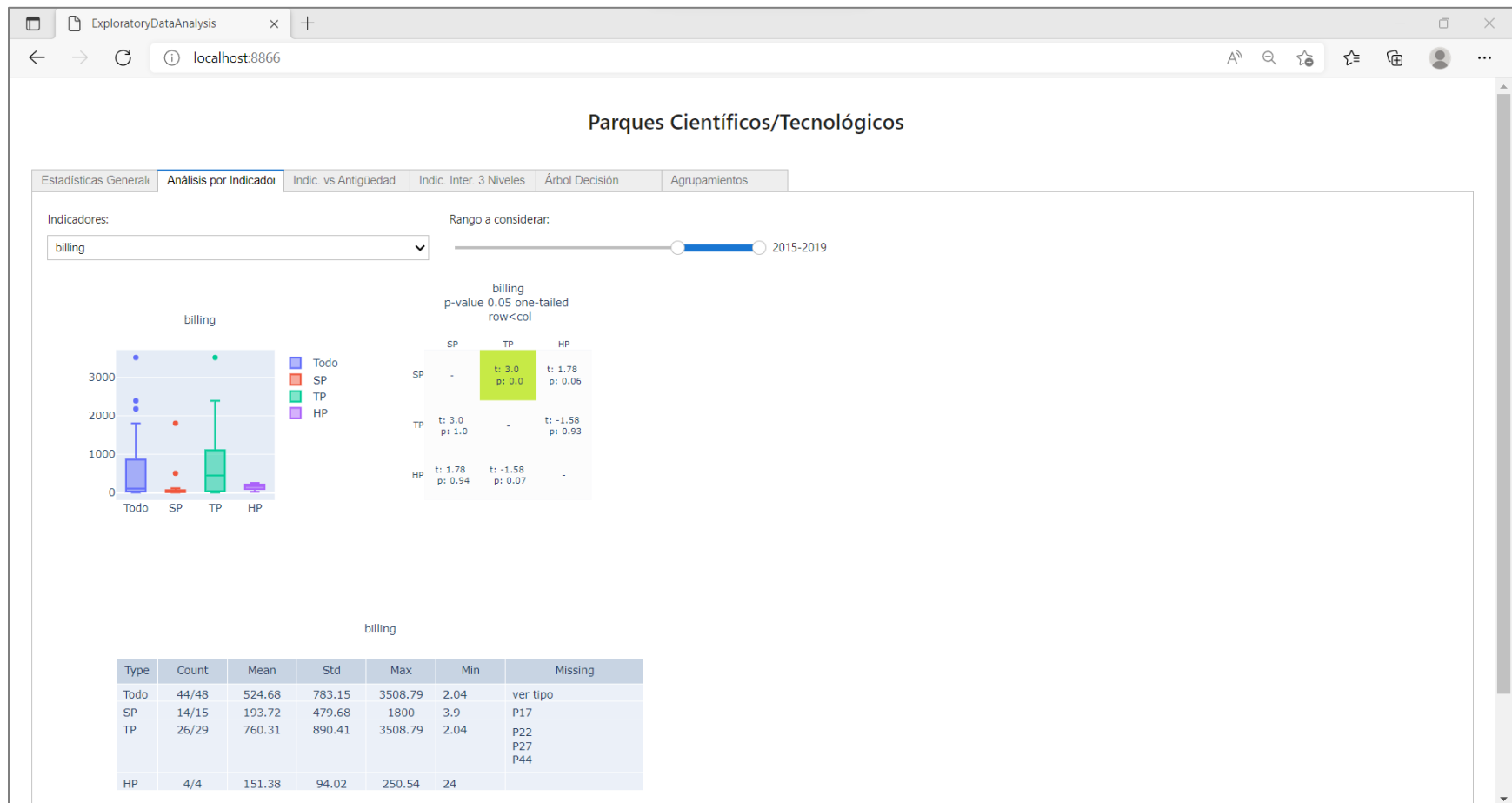
	count	mean	std	max	min
age	49/49	16.22	6.46	34	4
average-size-1	49/49	19.44	14.79	61.67	3.31
average-size-2	45/49	3.28	3.7	16.51	0.04
billing	44/48	524.68	783.15	3508.79	2.04
companies	48/48	116.19	128.63	639	16
companies-allocated	49/49	11.33	15.82	90	0

	count	mean	std	max	min
age	15/15	15.73	3.77	22	9
average-size-1	15/15	11.87	9.19	35.81	3.31
average-size-2	14/15	2.1	4.31	16.51	0.07
billing	14/15	193.72	479.68	1800	3.9
companies	15/15	69.33	37.4	153	17
companies-allocated	15/15	10.07	6.82	25	1

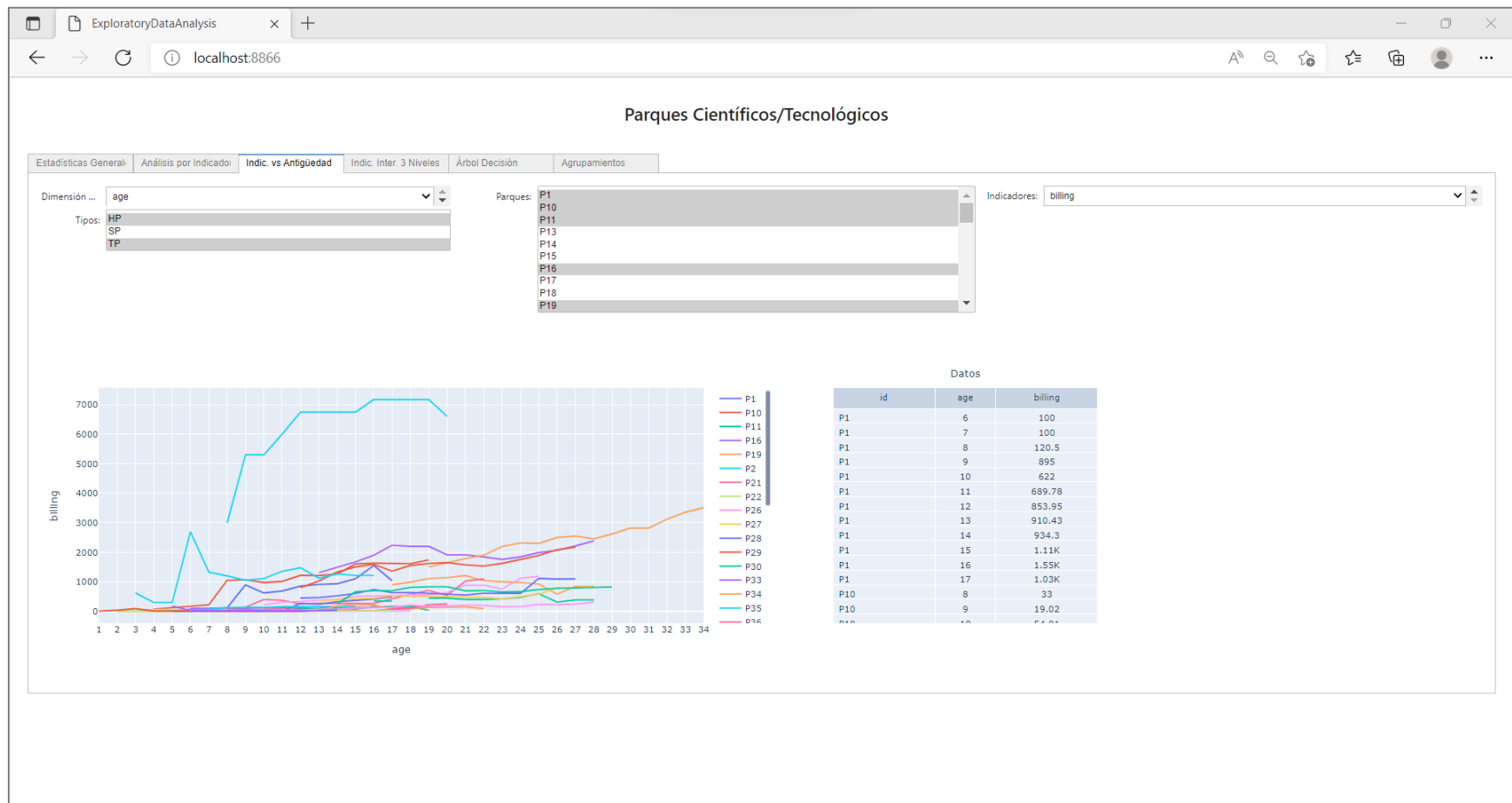
	count	mean	std	max	min
age	30/30	19.7	7.41	34	4
average-size-1	30/30	22.61	15.51	61.67	4
average-size-2	27/30	4.07	3.48	12.95	0.04
billing	26/29	760.31	890.41	3508.79	2.04
companies	29/29	146.1	156.41	639	17
companies-allocated	30/30	12.53	19.64	90	0

	count	mean	std	max	min
age	4/4	16.5	3.51	20	13
average-size-1	4/4	24.07	19.1	50.59	7.86
average-size-2	4/4	2.08	0.87	3.17	1.28
billing	4/4	151.38	94.02	250.54	24
companies	4/4	75	49.4	125	16

Módulo Análisis Bivariado



Módulo Series Temporales



Módulo Análisis Tridimensional

ExploratoryDataAnalysis
localhost:8866

Parques Científicos/Tecnológicos

Tipos:

Año:

Parques:

X:

Y:

Z:

age vs billing vs companies

Datos

id	year	age	billing	c
P13	2.02K	18	48.84	
P14	2.02K	22	1.8K	
P15	2.02K	10	60.65	
P17	2.02K	18	0	
P18	2.02K	12	16	
P23	2.02K	9	3.9	
P24	2.02K	14	12.62	
P25	2.02K	18	52.09	
P3	2.02K	20	21.42	
P31	2.02K	19	501.56	
P40	2.02K	13	41.75	
P5	2.02K	17	112	
P6	2.02K	14	4.5	
P7	2.02K	14	32.48	
P8	2.02K	10	1.71	

Módulo Árboles de Decisión

ExploratoryDataAnalysis | localhost:8866

Parques Científicos/Tecnológicos

Estadísticas General
Análisis por Indicador
Indic. vs Antigüedad
Indic. Inter. 3 Niveles
Árbol Decisión
Agrupamientos

Indicadores:

dimension	% missing
0 age	0
1 average-size-1	0
2 average-size-2	8.33
3 billing	8.33
4 companies	0
5 companies-allocated	0
6 companies-established	0
7 employment	0
8 employment r&d	0
9 employment-men	0
10 employment-women	0

Accuracy: 0.79

Rango a considerar: Crear árbol

Máxima profundidad:

```

graph TD
    node0["node id0  
HP_4 < SP_15 - TP_29  
innovative-profile-1 <= 0.46 | > 0.46"]
    node1["node id1  
HP_1 < SP_3 - TP_25  
patents-3 <= 1.33 | > 1.33"]
    node4["node id4  
HP_3 < SP_12 - TP_4  
internationalisation <= 0.07 | > 0.07"]
    node2["node id2  
HP_0 < SP_1 - TP_24"]
    node3["node id3  
HP_1 < SP_2 - TP_1"]
    node5["node id5  
HP_3 < SP_6 - TP_4"]
    node6["node id6  
HP_0 < SP_6 - TP_0"]

    node0 --> node1
    node0 --> node4
    node1 --> node2
    node1 --> node3
    node4 --> node5
    node4 --> node6
        
```

Parques por nodo hoja

Leaf	Park
2	SP-P8
	TP-P1
	TP-P11
	TP-P19
	TP-P21
	TP-P22
	TP-P27
	TP-P28
	TP-P29
	TP-P30
	TP-P34
	TP-P35
	TP-P38
	TP-P39
	TP-P41
	TP-P42
	TP-P45
	TP-P45
	TP-P46
	TP-P48
	TP-P49

Módulo Análisis de Agrupamientos

ExploratoryDataAnalysis | localhost:8866

Parques Científicos/Tecnológicos

Estadísticas General | Análisis por Indicador | Indic. vs Antigüedad | Indic. Inter. 3 Niveles | Árbol Decisión | **Agrupamientos**

Indicadores:

dimension	% missing
0 age	0
1 average-size-1	0
2 average-size-2	8.33
3 billing	8.33
4 companies	0
5 employment	0
6 employment r&d	0
7 incubated companies	2.08
8 innovative-profile-1	0
9 international companies	0
10 internationalisation	0

Rango a considerar: 2015-2019

Número de grupos: 3

Extraer grupos

Parques por clúster
Silhouette Score: 0.26

Cluster	Park
0	TP-P33
0	TP-P34
0	TP-P42
1	HP-P47
1	TP-P1
1	TP-P11
1	TP-P19
1	TP-P21
1	TP-P22
1	TP-P26
1	TP-P27
1	TP-P28
1	TP-P29
1	TP-P30
1	TP-P35
1	TP-P36
1	TP-P37
1	TP-P38
1	TP-P39
1	TP-P41
1	TP-P43
1	TP-P44
1	TP-P45
1	TP-P46