PCT Observer

Tablero de Parques Científicos/Tecnológicos

Contenido

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, porciento 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-learn[1].

- 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

de scikit-learn[2] 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 Voila[3], Jupyter[4] 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.

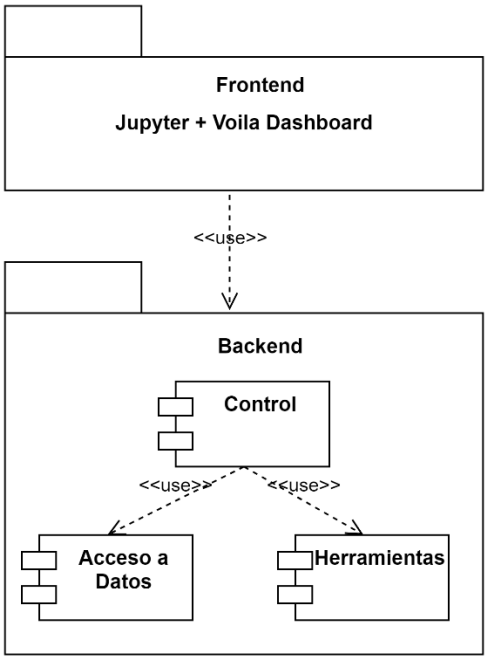


Figura 1: Arquitectura general de PCT Observer

[2] sklearn.cluster.KMeans — scikit-learn 1.0.2 documentation
[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 -> qgrid[5], 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 ≥ 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.

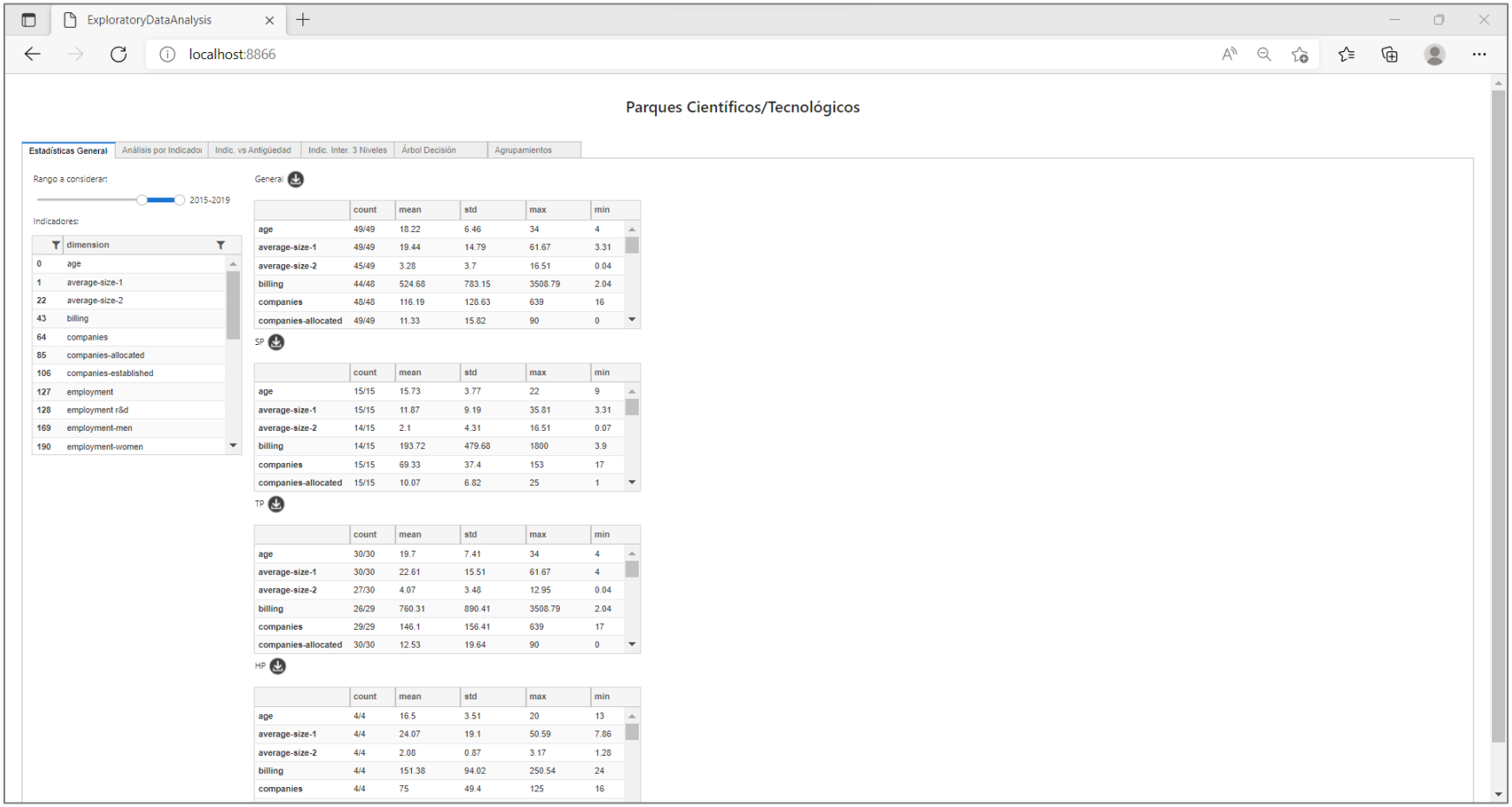
[4] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

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

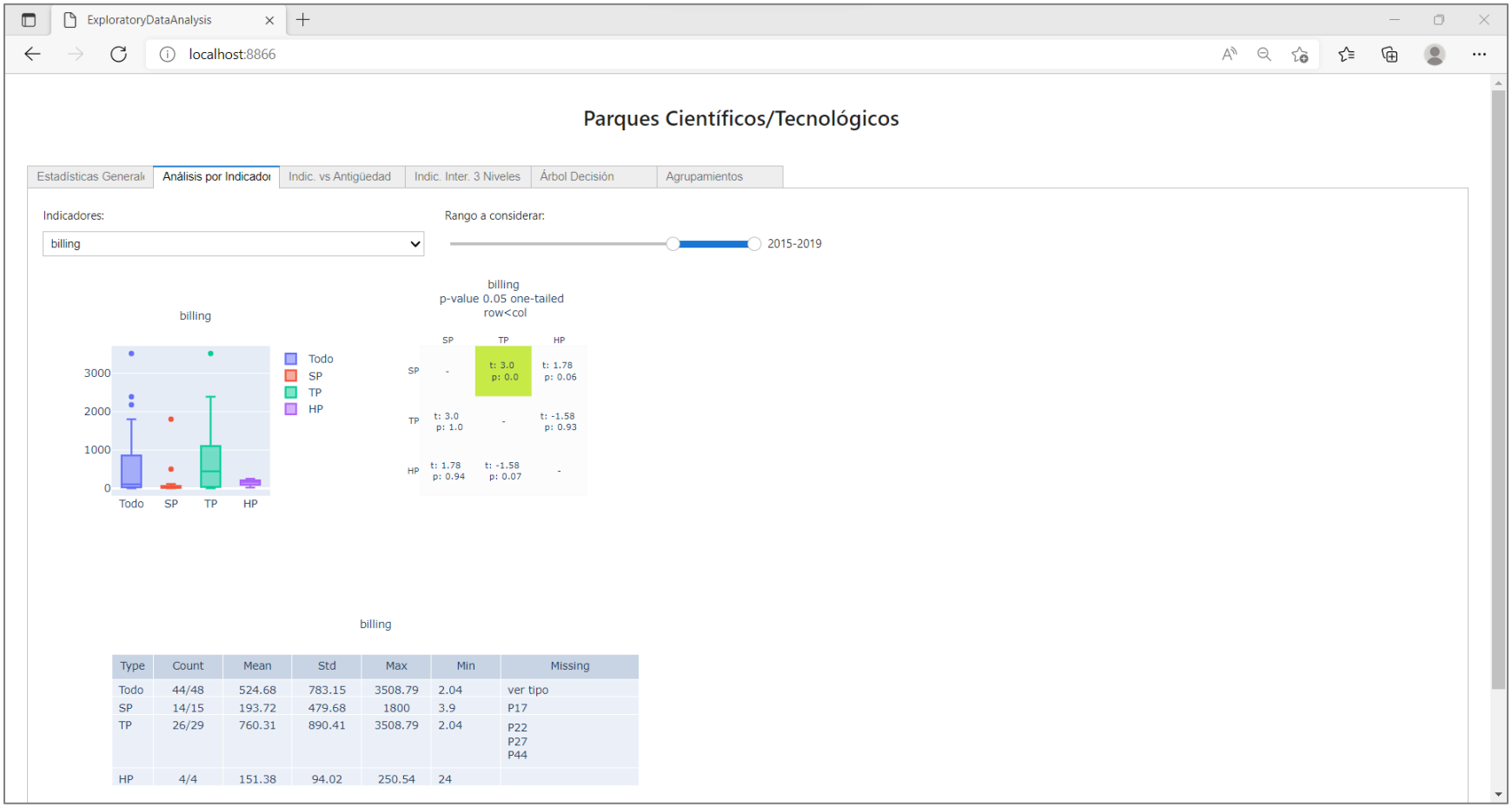
Anexo 1

Pantallas PCT Observer

Módulo Estadística Descriptiva

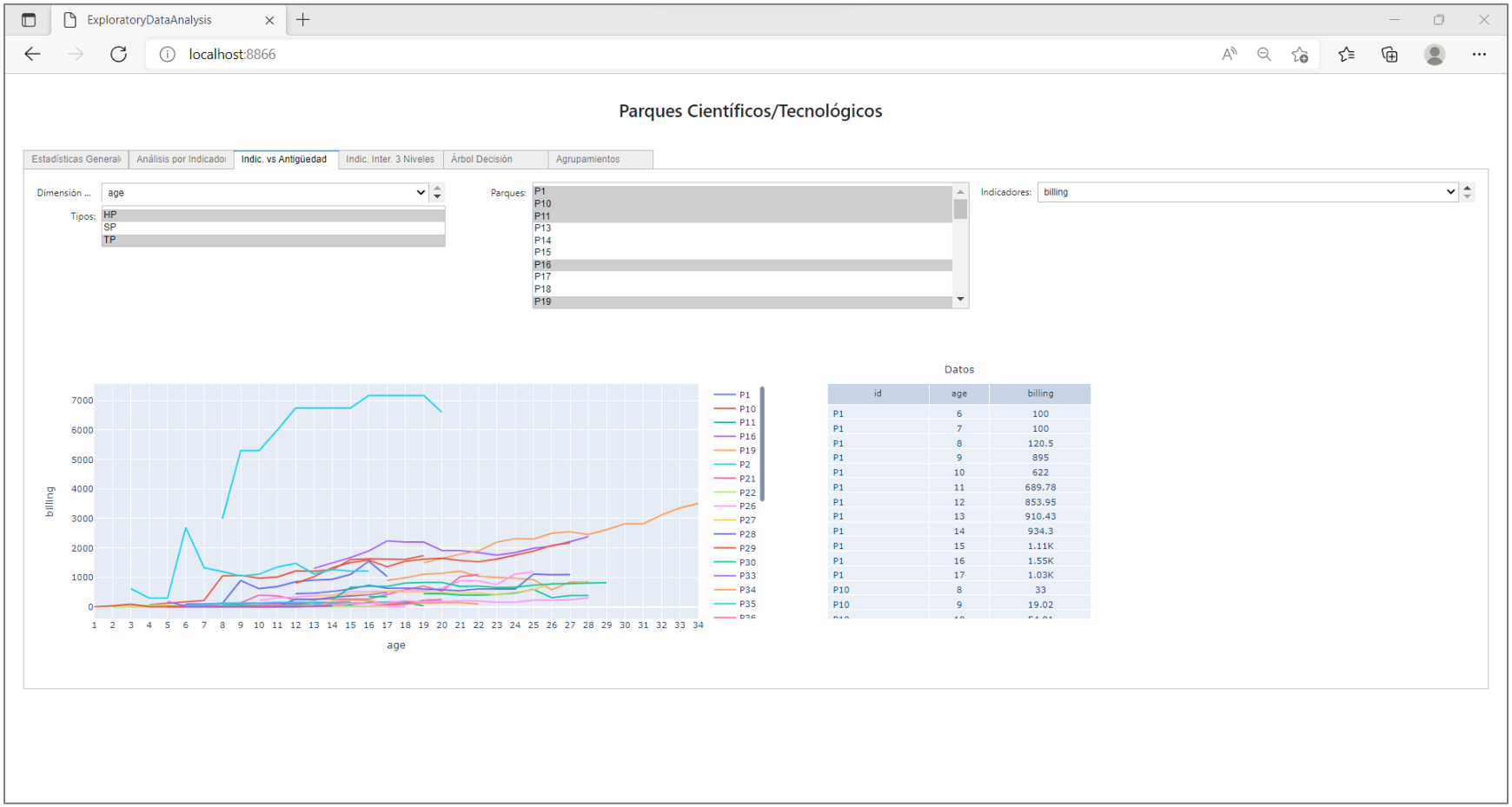


Módulo Análisis Bivariado

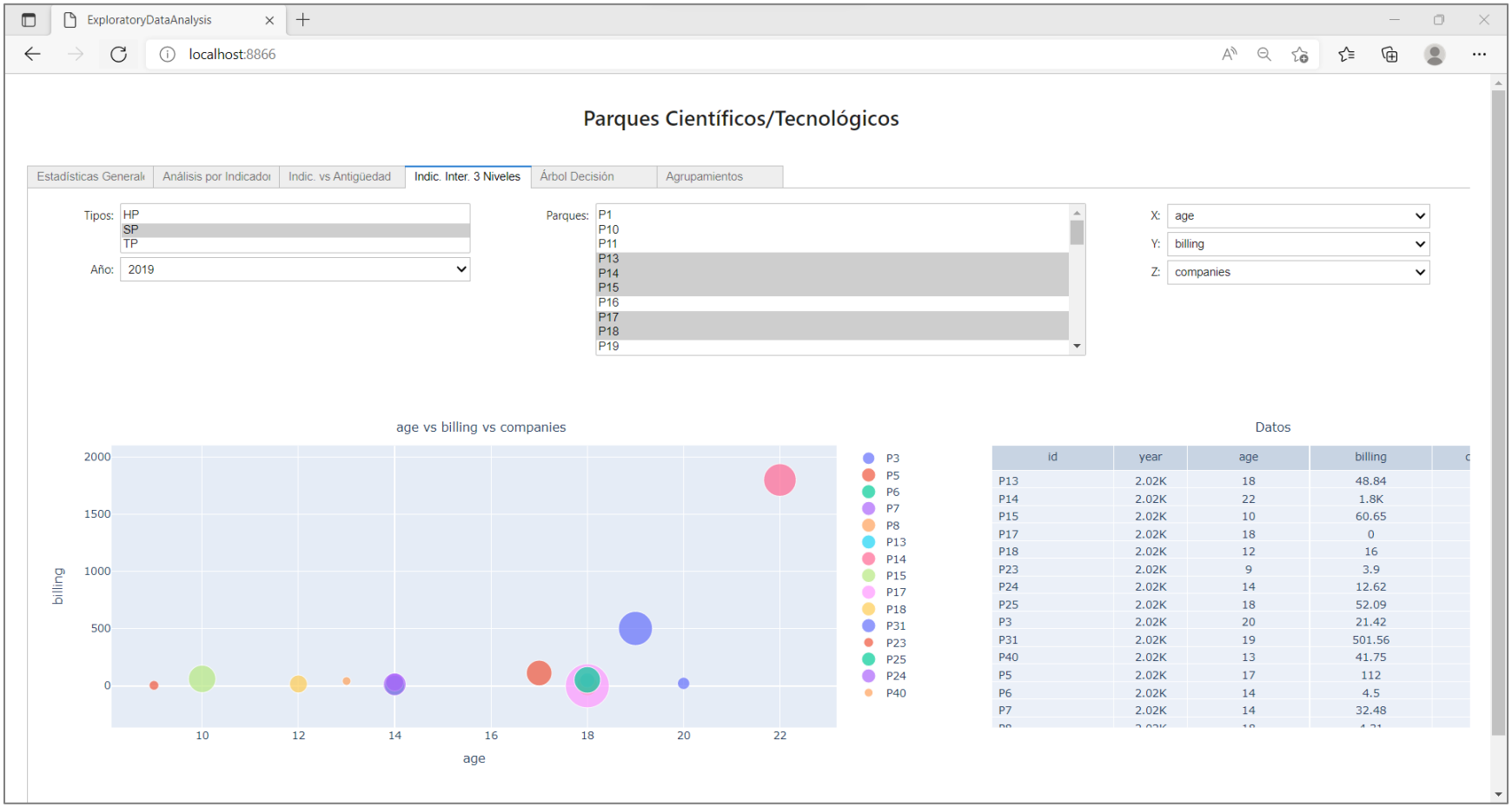


Type	Count	Mean	Std	Max	Min	Missing
Todo	44/48	524.68	783.15	3508.79	2.04	ver tipo
SP	14/15	193.72	479.68	1800	3.9	P17
TP	26/29	760.31	890.41	3508.79	2.04	P22 P27 P44
HP	4/4	151.38	94.02	250.54	24	

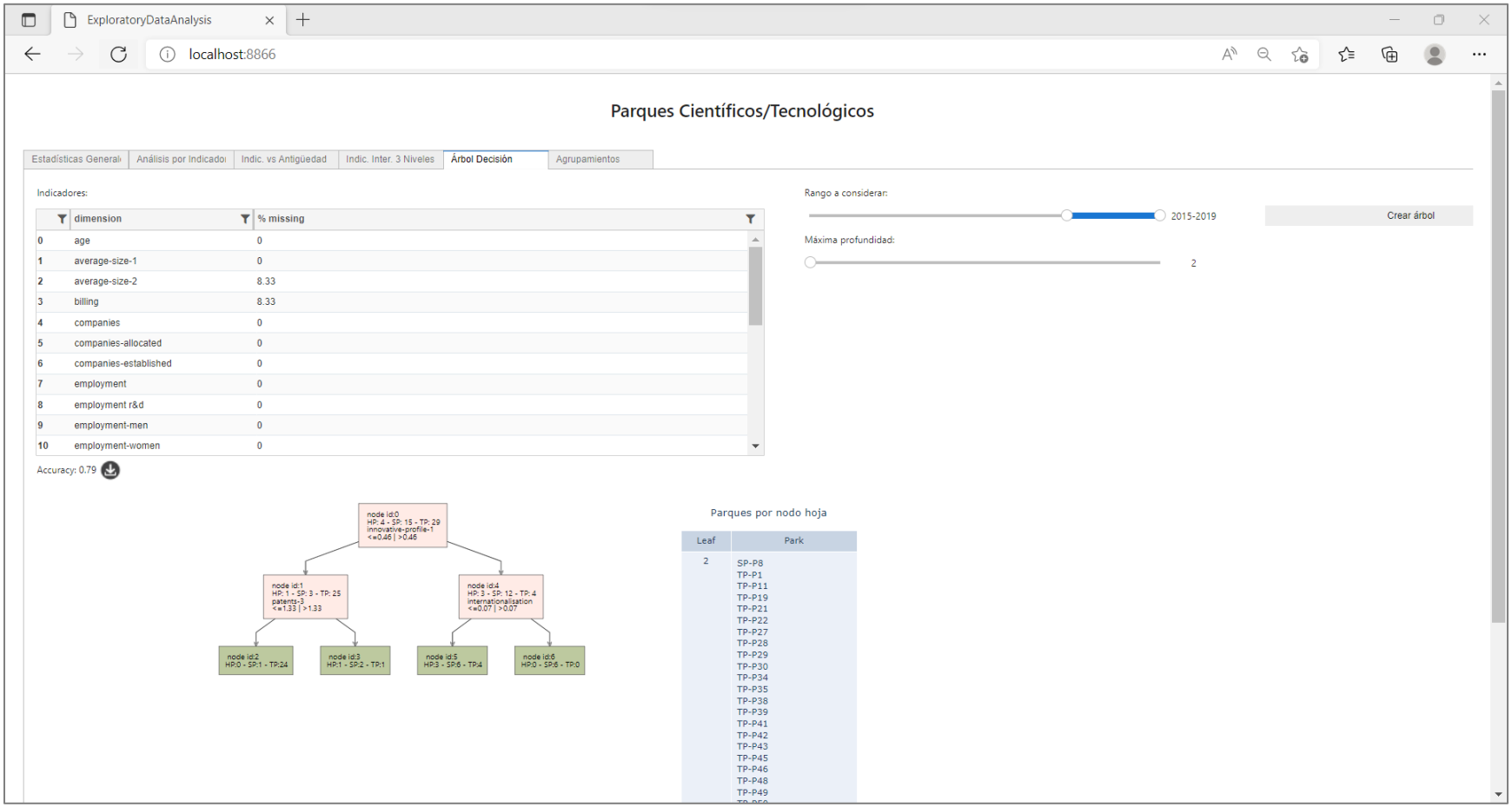
Módulo Series Temporales



Módulo Análisis Tridimensional



Módulo Árboles de Decisión



Módulo Análisis de Agrupamientos