

Ciencia de los Datos

Maestría Virtual en Ingeniería de Sistemas y Computación

Ambiente de configuración

Introducción

En esta sección se realizan las configuraciones de los ambientes para trabajar en ciencia de datos. Se deben instalar y configurar 3 herramientas: PostgreSQL, PgAdmin 4 y Jupyter Lab. Estas herramientas, nos permitan, manejar y administrar base de datos relacionales, y procesar datos usando el lenguaje de programación Python.

Antes de empezar:

Instalar Git

<https://git-scm.com/book/en/v2/Getting-Started-Installing-Git>

Instalar Docker

Para Windows, ir a <https://docs.docker.com/desktop/windows/install/> y dar clic.

Luego instale el archivo .exe. Se instalara la aplicación de manera típica, dando sobre el botón siguiente, siguiente, hasta finalizar.

Para mac, ir a <https://docs.docker.com/desktop/mac/install/> y dar clic en , Haga doble clic en Docker.dmg para abrir el instalador, luego arrastre Docker a la carpeta Aplicaciones. Luego siga los pasos.

Para Ubuntu-Linux, siga estos pasos:

Ciencia de los Datos

<https://www.digitalocean.com/community/tutorials/how-to-install-and-use-docker-on-ubuntu-20-04-es>

Paso 1 - Descarga configuracion yml

Descargar el repositorio con el comando: git clone

<https://github.com/cghidalgos/dataScienceCourse>

configura las rutas de tu maquina para guardar los notebooks que vayas a utilizar.

- Abre la carpeta configuración
- Abre el archivo **docker-compose.yml**
- Vaya a la linea 30
- Modifica la ruta, ejemplo:
 - /Users/user1/Documents/MiCarpeta:/home/jovyan :/home/jovyan se quedan

NOTA: recuerde dar permisos de lectura y escritura a su carpeta

Paso 2 - Iniciar los servicios

Abre una terminal/consola y ve a la ruta **dataScience-/configuración/** y usa el comando **docker-compose up -d** este sirve para crear e iniciar todos los servicios de la configuración del archivo previamente descargado. En el se encuentran las aplicaciones PostgreSQL, PgAdmin 4 y Jupyter Lab

Paso 3 - Configurar servicio PostgreSQL

En nuestra máquina local accedemos al enlace <http://localhost:5050/> desde cualquier navegador y se mostrará lo siguiente:

Ciencia de los Datos



Usuario: admin@admin.com

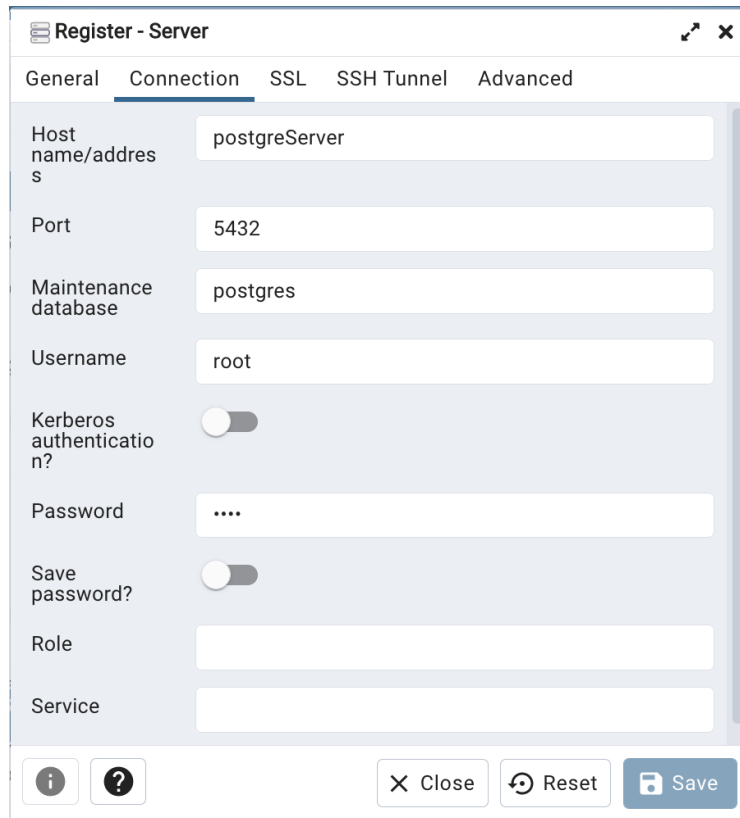
Contraseña: root

Ahora, vamos a nuestro pgAdmin en el navegador. Ahora vamos a agregar nuevo servidor



ahora vamos a poner un nombre a nuestro servidor, puede ser el que usted quiera

Ciencia de los Datos



The screenshot shows a 'Register - Server' dialog box with the 'Connection' tab selected. The fields are filled with the following values:

Field	Value
Host name/addresses	postgreServer
Port	5432
Maintenance database	postgres
Username	root
Kerberos authentication?	<input type="checkbox"/>
Password
Save password?	<input type="checkbox"/>
Role	
Service	

At the bottom, there are buttons for 'Close', 'Reset', and 'Save', along with information and help icons.

y llenamos los siguientes datos:

nombre/direccion de servidor: postgreServer

Nombre de usuario: root

contraseña: root

luego salvar, y tenemos nuestro servidor listo para crear nuestras bases de datos.

Paso 3 - Configurar servicio Jupyter Lab


Ciencia de los Datos

Ahora ve a tu terminal y escribe `docker exec -it jupyter_notebook jupyter server list`

```
Currently running servers:  
http://4d4d3df2a029:8888/?token=9bb83758ceb4d7b62436f11c0b492bdc00f72aef468fec70 :: /home/jovyan
```

Copiar el token `9bb83758ceb4d7b62436f11c0b492bdc00f72aef468fec70`

En nuestra maquina local accedemos al enlace <http://localhost:8080/> desde cualquier navegador y se mostrará lo siguiente:

 jupyter

Password or token:

Token authentication is enabled

If no password has been configured, you need to open the server with its login token in the URL, or paste it above. This requirement will be lifted if you [enable a password](#).

The command:

```
jupyter server list
```

will show you the URLs of running servers with their tokens, which you can copy and paste into your browser. For example:

```
Currently running servers:  
http://localhost:8888/?token=c8de56fa... :: /Users/you/notebooks
```

Pegamos el token que copiamos y le damos clic en Log In. Ahora tendremos nuestro servidor de Jupyter Lab. Puede copiar o crear notebooks, que deberan estar en la carpeta que eligió en el [Paso 1](#).

Maestría Virtual en Ingeniería de Sistemas y Computación

Taller 1

Introducción

La calidad de aire es un problema crítico en las grandes ciudades del mundo. Varias afecciones respiratorias están relacionadas con la calidad del aire que respiramos y por tanto, es importante para las autoridades locales medir, reportar y predecir de manera constante y precisa los niveles de los diferentes contaminantes presentes en el aire de la ciudad. Por esta razón, la secretaría distrital de ambiente de Bogotá instaló 19 estaciones de monitoreo de aire en la ciudad y proporciona de manera libre estos datos para que cualquiera pueda hacer uso de esta información.

Uno de los contaminantes más peligrosos para la salud humana es el material particulado de tamaño menor a 2.5 micras (PM2.5) ya que se acumula en los pulmones y puede causar daños permanentes a quienes están expuestos a él por largos periodos de tiempo.

Al analizar los datos provenientes de las estaciones de monitoreo, se han identificado varios problemas asociados a la calidad de los datos y a la fiabilidad de la información. Por ejemplo se ha identificado que más del 20% de los datos correspondientes a las mediciones de dicho contaminante están perdidas. Esto es un fenómeno común, pues los sensores de las diferentes estaciones fallan por diversos motivos, como cortes de energía, periodos de mantenimiento preventivo y reparaciones efectuadas a las estaciones.

Adicionalmente, el área que cubren los sensores es muy pequeña comparada con el área de una ciudad como Bogotá. Entonces se necesitan modelos que permitan informar a la ciudadanía sobre los niveles de contaminación aún en áreas que no cuentan con sensores. Incluso, si los datos son de muy buena calidad podríamos crear modelos que predigan la calidad del aire en periodos de tiempo futuros.

El propósito del proyecto de este curso es crear modelos de aprendizaje de máquina que permitan curar y completar los datos de la red de monitoreo de Bogotá y generar aplicaciones que puedan ser usados por los ciudadanos para informarse sobre la calidad de aire que respiran. Para esto, usaremos los datos que la red de monitoreo de Bogotá publica en su plataforma de manera permanente. Para comenzar, vamos a crear un repositorio de datos que nos permitirá acceder a los datos de manera eficiente.

En esta sección se realiza una bodega de datos que contine 68.000 registros de la calidad del aire de la ciudad de Bogotá. La bodega de datos, se conecta con el lenguaje de programación Python para hacer procesamiento de datos y luego mostrar la información usando librerías de datos. Para iniciar realice la siguientes actividades:

Descargar el repositorio con el comando:

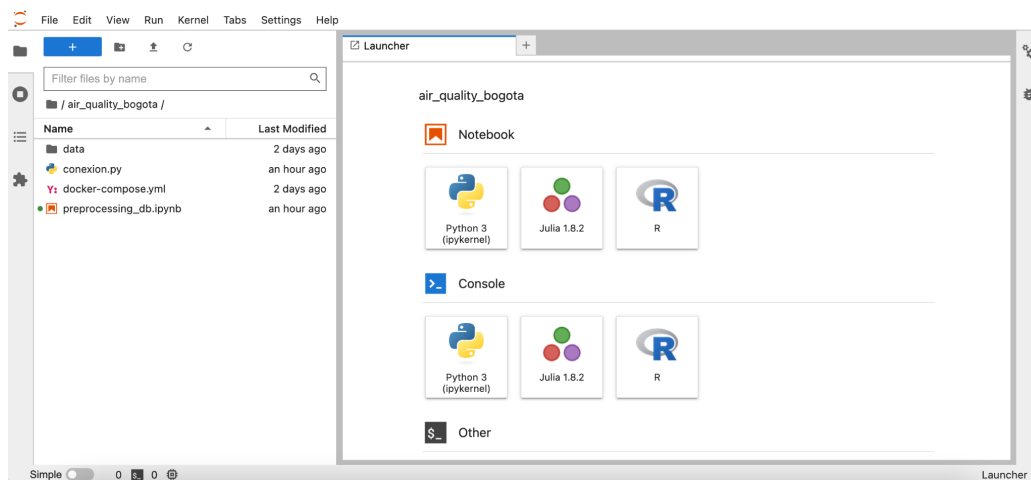
```
git clone https://github.com/cghidalgos/dataScienceCourse
```

Abra la carpeta dataScience/ y copie la carpeta Taller1, peguela la en la carpeta que eligió en el Paso 1.

Parte 1

Ciencia de los Datos

Ahora, vaya a su JupyterLab (<http://localhost:8888/>) y ejecute el notebook **CienciaDatosTaller1.ipynb** usando la pestaña Run, y luego, Run all cells. Esto creara la bodega de datos.



Finalmente, vaya a su PgAdmin (<http://localhost:5050/>) para ver su bodega de datos creada.

Parte 2

Ahora, vaya a su JupyterLab (<http://localhost:8888/>) y ejecute el notebook **DW_visualization.ipynb** usando la pestaña Run, y luego, Run all cells. Esto tomará la información de la bodega de datos, y hará algunas consultas y visualizaciones.

Taller 2

Introducción

En el taller anterior realizamos los pasos necesarios para consolidar una bodega de datos usando la información publicada por la Secretaría Distrital de Ambiente de Bogotá sobre calidad del aire.

En esta sección se usarán los datos de la misma bodega y se prepararán para ser usados en los modelos de aprendizaje de máquina que nos permitirán predecir los datos faltantes en los registros. Esta práctica se hará en los cuadernos (notebooks) de Python usando la herramienta JupyterLab.

Parte 1

Descargar el repositorio con el comando:

```
git clone https://github.com/cghidalgos/dataScienceCourse
```

Abrir la carpeta *dataScience/* y copiar la carpeta *CienciaDatosTaller2*, luego pegarla en la carpeta que eligió en el Paso 1.

Después, usando Jupyter Lab, abra el archivo *"CienciaDatosTaller2.ipynb"*

Parte 2

Desarrolle cada uno de los puntos que se muestran en las celdas “Desarrollar”

Taller 3

Introducción

En la sección anterior se limpió el conjunto de datos, usando un conjunto de herramientas estadísticas. Ahora, vamos a crear varios modelos predictivos, para completar los datos faltantes del **contaminante PM2.5**. Cabe anotar que hemos utilizado este contaminante durante esta práctica, pero la metodología puede ser extendida para cualquier otro contaminante. Por lo tanto, es una práctica opcional implementar un modelo para predicción CO2.

En este taller usaremos 2 modelos de ajuste de curvas diferentes:

- Árboles de regresión
- Redes neuronales

Se evaluará el desempeño de ambos modelos y se usará el mejor para imputar los datos faltantes en el dataset.

Parte 1

Descargar el repositorio con el comando:

```
git clone https://github.com/cghidalgos/dataScienceCourse
```



Ciencia de los Datos

Abrir la carpeta `dataScience/` y copie la carpeta **CienciaDatosTaller3**, luego pegarla en la carpeta que eligió en el Paso 1.

Ahora, usando JupyterLab abrir el archivo “CienciaDatosTaller3.ipynb”

Parte 2

Desarrolle los puntos que se muestran en las celdas “Desarrollar”