

IBM

Data Science Experience



Manual para Workshop
Febrero 2018

Índice

Introducción

Workshop 1. Comenzar con DSX

- 1.** Registrarse
- 2.** Añadir usuarios adicionales
- 3.** Comenzar con DSX

Workshop 2. Crea y organiza los recursos en un proyecto

- 1.** Crear un proyecto
- 2.** Añade colaboradores
- 3.** Aprende a gestionarlo: borrar assets, añadir, crear conexiones, ...

Workshop 3. Consigue y prepara los datos y analiza los datos de manera sencilla

- 1.** Añadir datos
- 2.** Crear un modelo automático
- 3.** Despliega el modelo en Watson Machine Learning
- 4.** Crear un modelo semi-automático o manual

Workshop 4. Analiza en profundidad los datos

- 1.** Notebooks
- 2.** Visualizaciones
- 3.** Algoritmos de analítica predictiva de SPSS
- 4.** RStudio
- 5.** Librerías de Deep Learning

Workshop 5. Recursos para aprender Data Science Experience en Local

Introducción

IBM Data Science Experience es un entorno que reúne todo lo que necesita un Data Scientist. Incluye las herramientas de código abierto más populares, además los equipos de data scientist de IBM han unido al código abierto una serie de funcionalidades que aportan un gran valor añadido, todo integrado a la perfección en esta única herramienta para que tanto el análisis como los usuarios sean más efectivos y eficaces.

Actualmente existen dos: IBM Data Science Experience en Local y en Cloud. En este workshop nos centraremos en IBM Data Science Experience en Cloud, que forma parte de Watson Data Platform, que es una plataforma híbrida que interconecta los datos con servicios analíticos, dando solución a los problemas típicos a los que se enfrentan las empresas. Como pueden ser: mala colaboración entre los componentes de un equipo, falta de confianza en el dato o falta de confianza en el resultado, problemas de seguridad, etc.

Data Science Experience se crea basándose en tres pilares fundamentales: **aprender**, **crear** y **colaborar**.



Learn

Get started or get better with built-in learning.



Create

Use the best of open source tooling with IBM innovation.



Collaborate

Work smarter using community, work faster with your team.

Aprender:

DSX cuenta con herramientas de aprendizaje incluidas, con numerosos tutoriales de niveles que van desde niveles básicos a avanzados para que cualquiera pueda empezar a disfrutar de la herramienta. Además, puedes complementar el aprendizaje con los cursos y clases gratuitos sobre Data Mining y machine learning uniéndote a los más de 400,000 usuarios registrados en **Cognitive Class**.

Utiliza los conjuntos de datos, código de ejemplo, tutoriales y artículos técnicos que están a disposición de los usuarios.

Crear:

Data Science Experience recomienda fusionar lo mejor del código abierto, con el valor añadido que aportan las herramientas de IBM para crear modelos de datos punteros. Además, DSX cuenta con una gran inversión en Spark, líder en la industria (posee más de 3500 desarrolladores e investigadores).

Gracias a DSX puedes usar el código abierto y las herramientas potentes de analítica avanzada de modo integrado, gobernado y seguro.

Colaborar:

Las características colaborativas proporcionan una ayuda importante para aumentar la productividad y el impacto en el negocio.

Con Data Science Experience puedes administrar los recursos del proyecto y la colaboración de los usuarios además de poder compartir, bifurcar y reutilizar assets con Github.

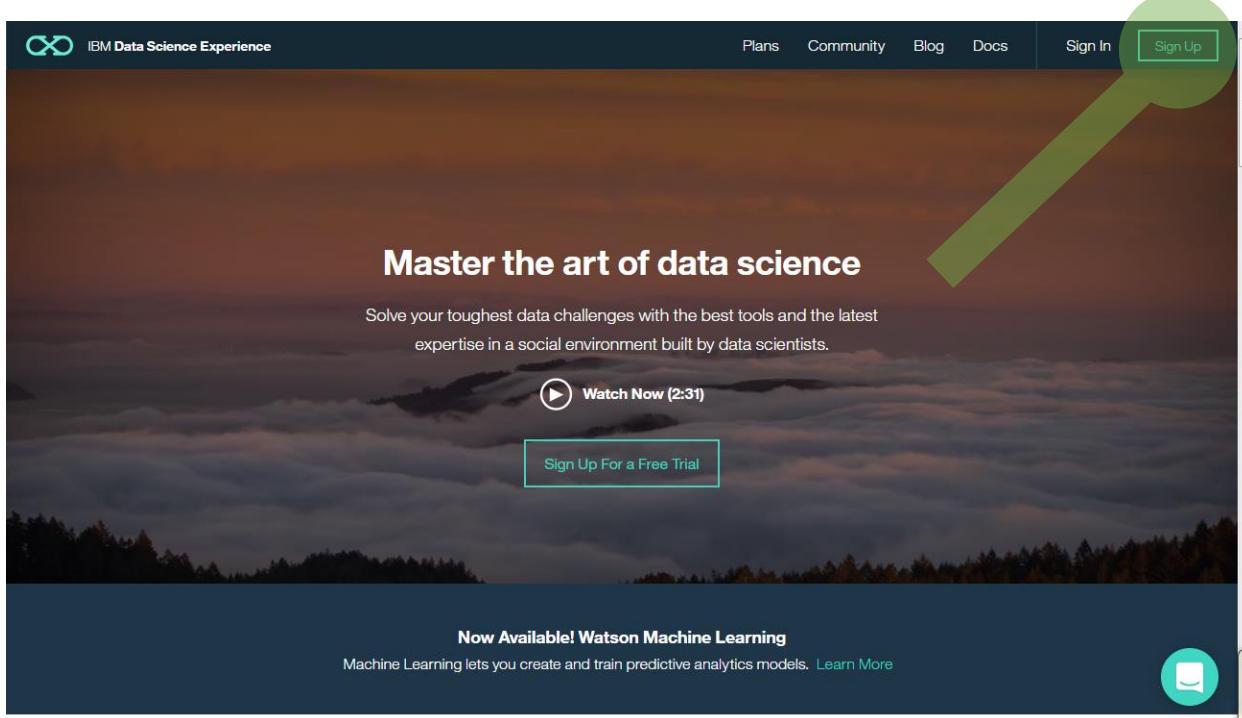
Workshop 1.

1. Registrarse

IBM Data Science Experience forma parte de Watson Data Platform. Watson Data Platform es una plataforma que permite a los equipos colaborar, compartir datos y modelos y poner en producción dichos modelos.

Para crearse una cuenta nueva, comienza por registrarte en Data Science Experience. Después de registrarte, puedes agregar otras aplicaciones de Watson Data Platform en cualquier momento desde la propia herramienta.

Para registrarse Entra en <https://datascience.ibm.com/> Haciendo **Sing Up**



Si aún no tienes una cuenta IBMid e IBM Cloud, se crearán durante el proceso de suscripción.

Nota: asegúrate de permanecer con tu navegador predeterminado durante el proceso de inicio de sesión e inicio de sesión. Si te encuentras en IBM Cloud Dashboard,

simplemente regresa a la página de registro y sigue el enlace que indica que ya tienes una cuenta.

Una vez que te registras, tu IBMID está vinculado a tu cuenta de IBM Cloud y a la cuenta de Watson Data Platform. Usa tu IBMID (la dirección de correo electrónico que proporcionaste) para iniciar sesión en DSX.

Si eres el único usuario en la cuenta, ¡ya está todo listo! Como propietario de la cuenta IBM Cloud que se suscribió a una aplicación Watson Data Platform, tienes los permisos necesarios para agregar servicios, proyectos, catálogos, etc.

2. Añadir usuarios adicionales

Las aplicaciones de Watson Data Platform, IBM Data Science Experience, Data Catalog e IBM Data Refinery están diseñadas para la colaboración entre muchos usuarios. Después de crear una cuenta, puedes agregar usuarios para que puedan compartir servicios y recursos que se aprovisionan para la cuenta.

Estos pasos describen las tareas de administración que debe seguir el propietario de la cuenta:

- Proporcionar un almacenamiento (object storage) en IBM Cloud para tu cuenta,
- agregar usuarios y asignar roles de usuario para que puedan acceder a los recursos de la cuenta.

Lo vemos detalladamente:

Paso 1: Proporciona un almacenamiento (object storage) en IBM Cloud para tu cuenta.

Inicia sesión en su cuenta de IBM Cloud.

Ve a **Catalog** y haz clic en **Storage** y luego click **Object Storage** y aprovisiona el Cloud Object Storage.

This is not just any cloud.
This is the IBM Cloud.

The IBM Cloud is the cloud for
smarter business.

Sign up Learn about IBM Cloud Private

IBM Bluemix is now IBM Cloud >

Cloud in the News | Pivotal simplify app development with Spring and IBM Software | Read more →

IBM: Defining bare metal since 2005. | Read more →

IBM gives clients control of their data in Europe. | Let's talk

<https://console.bluemix.net/catalog/>

All Categories

Infrastructure

Compute

Storage >

Network

Security

Containers

VMware

Platform

Boilerplates

APIs

Application Services

Blockchain

Cloud Foundry Apps

Data & Analytics

DevOps

Finance

Functions

Integrate

Internet of Things

Mobile

Network

Security

Watson

Block Storage

Persistent iSCSI based storage with high-powered performance and capacity up to 12TB.

File Storage

Fast and flexible NFS-based file storage with capacity options from 20GB to 12TB.

Object Storage

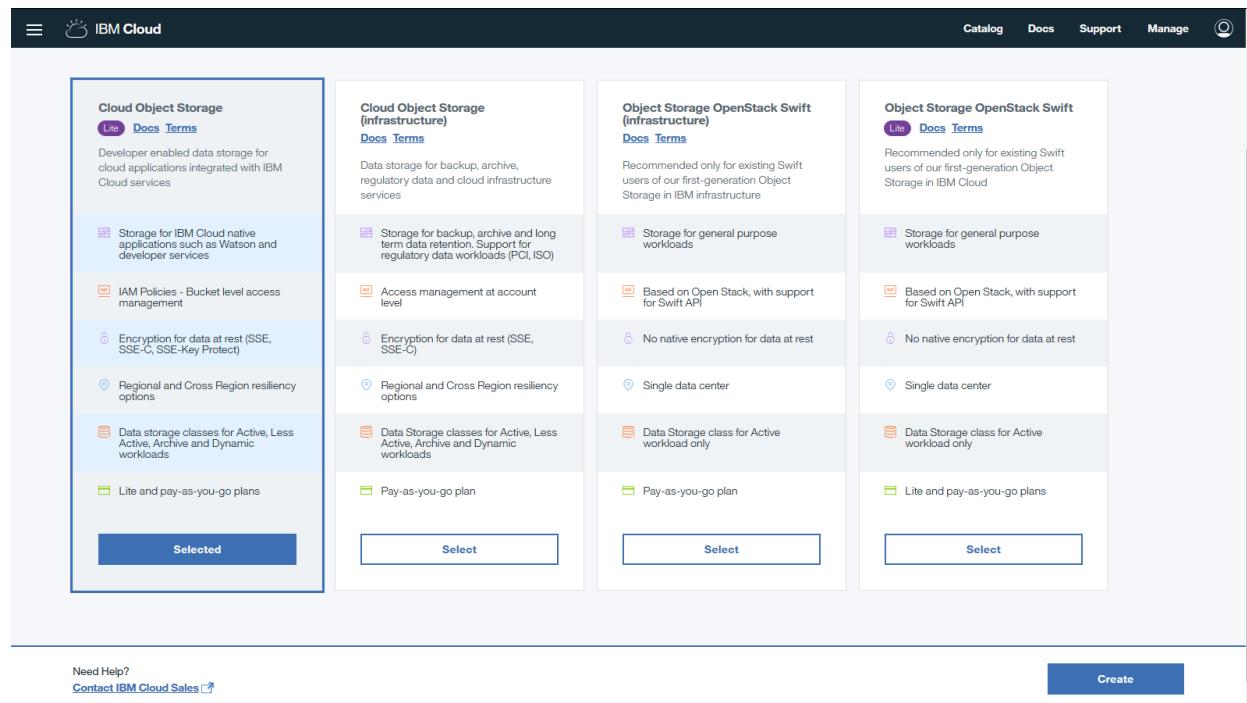
Provides flexible, cost-effective, and scalable cloud storage for unstructured data.

IBM Lite IBM

<https://console.bluemix.net/catalog/infrastructure/cloud-object-storage>

Si estás configurando usuarios para Data Catalog o IBM Data Refinery, selecciona un plan de acuerdo con sus necesidades.

Si tiene proyectos existentes (heredados) en DSX y deseas seguir creando proyectos heredados, haga clic en Comparar versiones y seleccione una de las opciones de Almacenamiento de objetos de OpenStack Swift, de acuerdo con tus necesidades.



The screenshot shows the IBM Cloud Catalog interface. At the top, there are navigation links: Catalog, Docs, Support, Manage, and a user icon. Below the navigation, there are two main sections for comparing storage options:

- Cloud Object Storage** (Selected): This option is highlighted with a blue border. It includes a 'Lite' button, a 'Docs' link, and a 'Terms' link. The description states: "Developer enabled data storage for cloud applications integrated with IBM Cloud services". Below this, there are several icons representing features: Storage for IBM Cloud native applications, IAM Policies - Bucket level access management, Encryption for data at rest (SSE, SSE-C, SSE-Key Protect), Regional and Cross Region resiliency options, Data storage classes for Active, Less Active, Archive and Dynamic workloads, and Lite and pay-as-you-go plans. A large blue 'Selected' button is at the bottom.
- Cloud Object Storage (Infrastructure)**: This option includes a 'Docs' link and a 'Terms' link. The description states: "Data storage for backup, archive, regulatory data and cloud infrastructure services". It lists features: Storage for backup, archive and long term data retention, Support for regulatory data workloads (PCI, ISO), Access management at account level, Encryption for data at rest (SSE, SSE-C), Regional and Cross Region resiliency options, Data Storage classes for Active, Less Active, Archive and Dynamic workloads, and Pay-as-you-go plan. A 'Select' button is at the bottom.
- Object Storage OpenStack Swift (Infrastructure)**: This option includes a 'Docs' link and a 'Terms' link. The description states: "Recommended only for existing Swift users of our first-generation Object Storage in IBM infrastructure". It lists features: Storage for general purpose workloads, Based on Open Stack, with support for Swift API, No native encryption for data at rest, Single data center, Data Storage class for Active workload only, and Pay-as-you-go plan. A 'Select' button is at the bottom.
- Object Storage OpenStack Swift**: This option includes a 'Docs' link and a 'Terms' link. The description states: "Recommended only for existing Swift users of our first-generation Object Storage in IBM Cloud". It lists features: Storage for general purpose workloads, Based on Open Stack, with support for Swift API, No native encryption for data at rest, Single data center, Data Storage class for Active workload only, and Lite and pay-as-you-go plans. A 'Select' button is at the bottom.

At the bottom of the catalog interface, there are links for 'Need Help?' and 'Contact IBM Cloud Sales', and a large blue 'Create' button.

Paso 2: agrega usuarios y asigna roles de usuario

Los usuarios que invites a la cuenta pueden compartir los servicios y recursos en la cuenta. Por ejemplo, los usuarios pueden crear proyectos o catálogos utilizando una instancia existente de IBM Cloud Object Storage de la cuenta. Estos usuarios también pueden agregarse como colaboradores en catálogos y proyectos restringidos. Si el usuario invitado aún no tiene una cuenta IBM Cloud, el usuario recibirá un correo electrónico para completar el proceso de registro.

Haz clic en **Manage**> **account**> **Users** para agregar usuarios autorizados a tu organización.

Cloud Object Storage
Lite Docs Terms

Developer enabled data storage for cloud applications integrated with IBM Cloud services

- Storage for IBM Cloud native applications such as Watson and developer services
- IAM Policies - Bucket level access management
- Encryption for data at rest (SSE, SSE-C, SSE-Key Protect)
- Regional and Cross Region resiliency options
- Data storage classes for Active, Less Active, Archive and Dynamic workloads
- Lite and pay-as-you-go plans

Selected

Cloud Object Storage (Infrastructure)
Docs Terms

Data storage for backup, archive, regulatory data and cloud infrastructure services

- Storage for backup, archive and long term data retention. Support for regulatory data workloads (PCI, ISO)
- Access management at account level
- Encryption for data at rest (SSE, SSE-C)
- Regional and Cross Region resiliency options
- Data Storage classes for Active, Less Active, Archive and Dynamic workloads
- Pay-as-you-go plan

Select

Object Storage OpenStack Swift (Infrastructure)
Docs Terms

Recommended only for existing Swift users of our first-generation Object Storage in IBM infrastructure

- Storage for general purpose workloads
- Based on Open Stack, with support for Swift API
- No native encryption for data at rest
- Single data center
- Data Storage class for Active workload only
- Pay-as-you-go plan

Select

Platform Notifications
Cloud Foundry Orgs
Resource Groups
Platform Notifications

Platform Notifications is recommended only for existing Swift users of our first-generation Object Storage in IBM Cloud

Storage for general purpose workloads

- Based on Open Stack, with support for Swift API
- No native encryption for data at rest
- Single data center
- Data Storage class for Active workload only
- Lite and pay-as-you-go plans

Select

Need Help?
Contact IBM Cloud Sales

<https://console.bluemix.net/iam/#/users>

Create

En la página **Users**, haz clic en Invitar usuarios.

Identity & Access

Users

Service IDs

Authorizations

Platform API Keys

Users

Use the View by filter to view specific organizations.

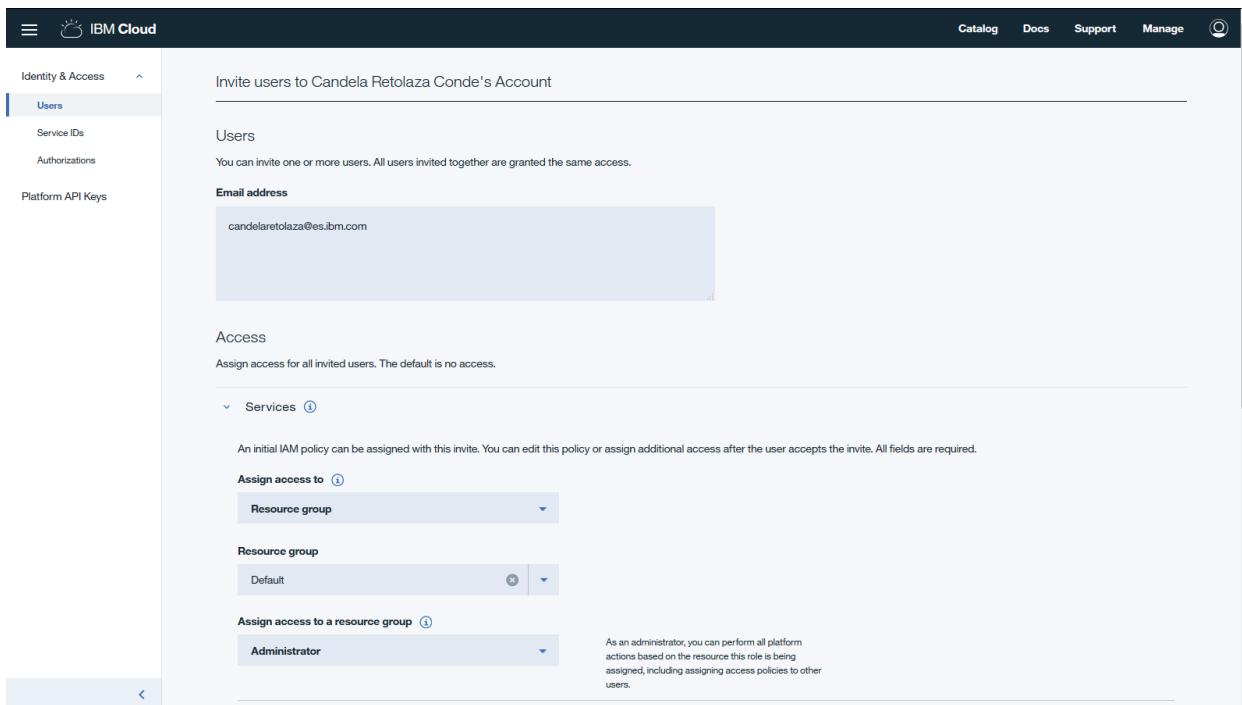
View by Account: Candela Retolaza Conde's Account

User	Email	Status
Candela Retolaza Conde	candela@ibm.com	ACTIVE

25 Users per page | 1-1 of 1 items

Invite users

Selecciona un usuario existente de IBMid. Puedes agregar múltiples usuarios y la configuración posterior se aplicará a todos ellos.



IBM Cloud

Identity & Access

Users

Service IDs

Authorizations

Platform API Keys

Invite users to Candela Retolaza Conde's Account

Users

You can invite one or more users. All users invited together are granted the same access.

Email address

candelaretolaza@es.ibm.com

Access

Assign access for all invited users. The default is no access.

Services

An initial IAM policy can be assigned with this invite. You can edit this policy or assign additional access after the user accepts the invite. All fields are required.

Assign access to

Resource group

Resource group

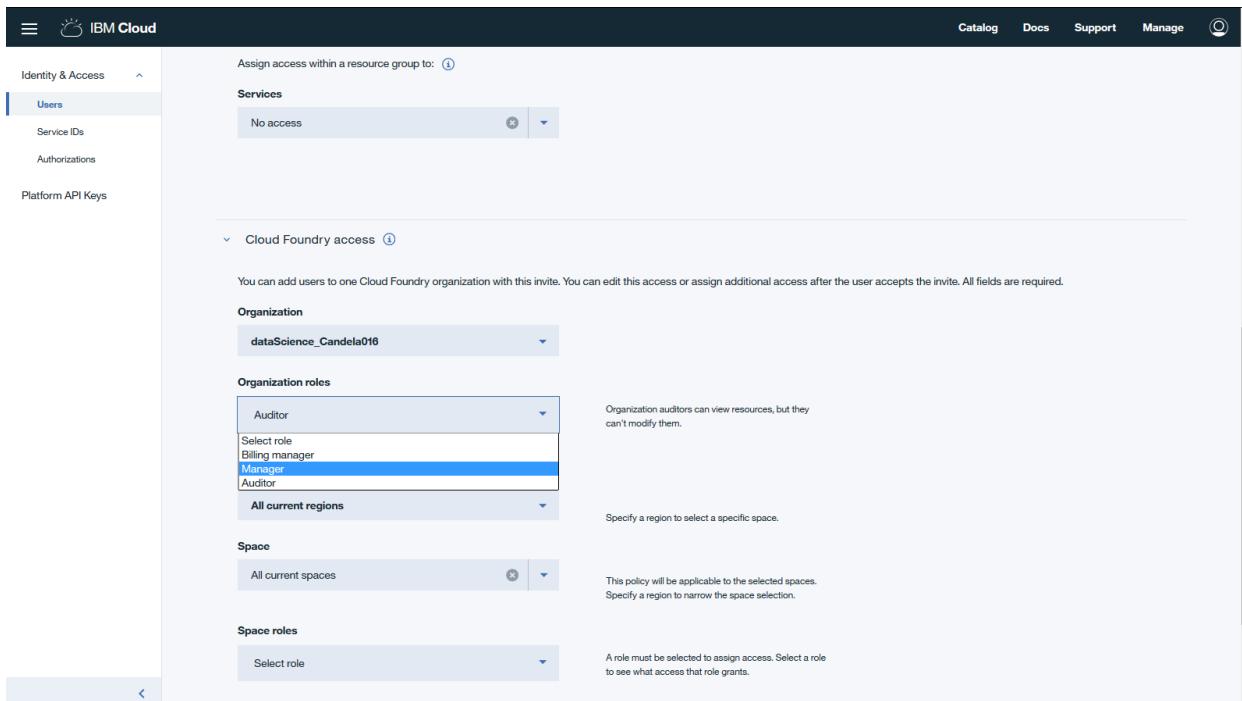
Default

Assign access to a resource group

Administrator

As an administrator, you can perform all platform actions based on the resource this role is being assigned, including assigning access policies to other users.

En la sección Acceso de la página Invitar usuarios, expanda el acceso a Cloud Foundry y seleccione la Organización a la que está agregando usuario.



IBM Cloud

Identity & Access

Users

Service IDs

Authorizations

Platform API Keys

Assign access within a resource group to:

Services

No access

Cloud Foundry access

You can add users to one Cloud Foundry organization with this invite. You can edit this access or assign additional access after the user accepts the invite. All fields are required.

Organization

dataScience_Candela016

Organization roles

Auditor

Select role

Billing manager

Manager

Auditor

All current regions

Organization auditors can view resources, but they can't modify them.

Space

All current spaces

Specify a region to select a specific space.

Space roles

Select role

This policy will be applicable to the selected spaces. Specify a region to narrow the space selection.

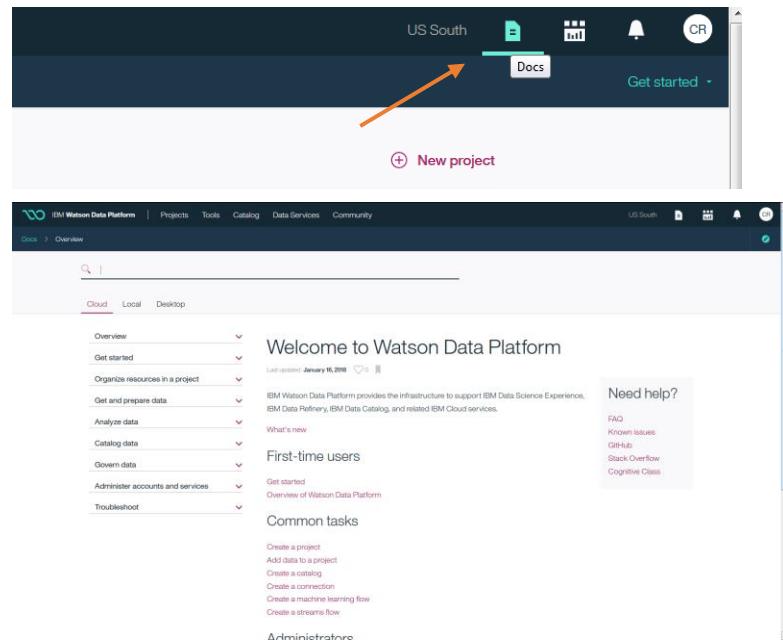
A role must be selected to assign access. Select a role to see what access that role grants.

- a. Asignar el rol del nuevo usuario en la organización.
- b. Secciona una Región y Espacio, o acepta los valores predeterminados.
- c. Para permitir que Watson Data Platform cree una instancia de Spark durante la creación del proyecto, asigne la función Desarrollador como la función de espacio.
- d. Para finalizar, haz clic en Invitar usuarios.

Para más información, visita la ayuda de DSX.

NOTA: Donde encontrar la ayuda y documentación de Data Science Experience

Click en Docs.



Sus usuarios ahora pueden iniciar sesión y pueden cambiar su cuenta y organización en la Configuración del perfil. Los usuarios asociados a tu cuenta ahora pueden trabajar juntos y usar las aplicaciones y servicios disponibles de la cuenta.

3. Comenzar con DSX

Entra en IBM Data Science Experience:

NAME	ROLE	COLLABORATORS	DATE CREATED	LAST UPDATED
Demos DSX	Admin		Sep 25, 2017	Jan 09, 2018
Machine Learning, Data Science	Admin	+2	Oct 08, 2017	Dec 22, 2017
Demo DSX	Admin		Nov 21, 2017	Dec 21, 2017

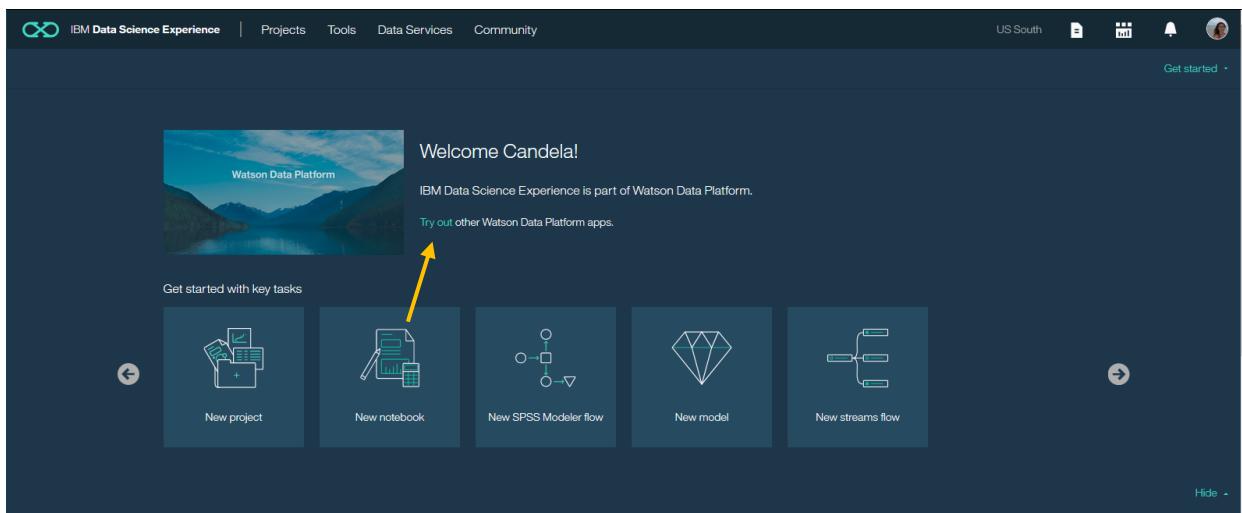
Para empezar a utilizar Data Science Experience puedes seguir los siguientes pasos:

1. Configura un proyecto para organizar sus recursos.
2. Agrega colaboradores a tu proyecto.
3. Agrega datos a tu proyecto.
4. Opcional: agrega servicios analíticos como IBM Streaming Analytics o Watson Machine Learning.
5. Comience a analizar datos. Por ejemplo, puede crear notebooks, usar RStudio, crear flujos de SPSS o modelos de aprendizaje automático.

¿Necesitas inspiración? Haz clic en el botón **Comunidad** en tu Data Science Experience para explorar los conjuntos de datos seleccionados, los Notebooks de ejemplo, los artículos y tutoriales, tanto para aprender de ellos como para utilizarlos como puntos de partida.

Antes de comenzar vamos a crear un catálogo de datos.

Data Catalog proporciona herramientas de administración de datos para indexar, clasificar y controlar el acceso a los assets. Los assets pueden incluir archivos y ficheros de datos, datos provenientes de conexiones a bases de datos y conexiones a fuentes de datos. Un catálogo contiene metadatos sobre los contenidos de los assets y cómo acceder a ellos y un conjunto de colaboradores que necesitarán usar los assets para el posterior análisis de datos.



Welcome Candela!

IBM Data Science Experience is part of Watson Data Platform.

Try out other Watson Data Platform apps.

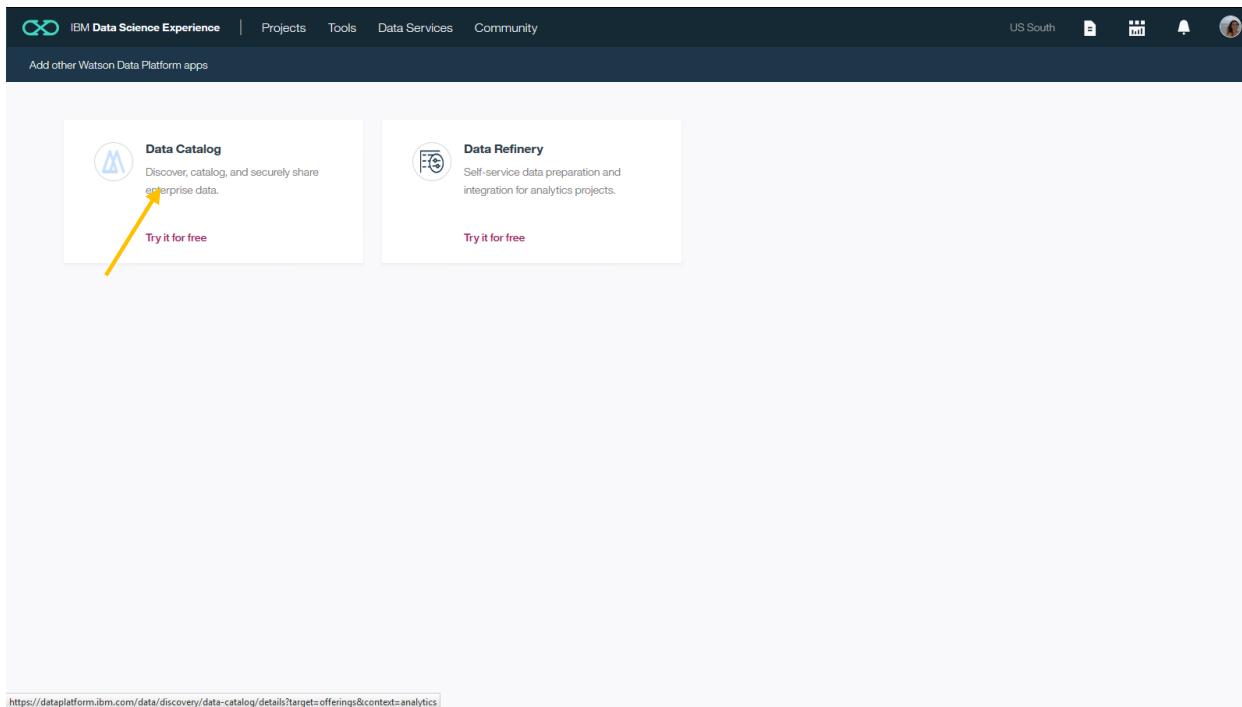
Get started with key tasks

New project	New notebook	New SPSS Modeler flow	New model	New streams flow
-----------------------------	------------------------------	---------------------------------------	---------------------------	----------------------------------

Recently updated projects [View all \(15\)](#) [+ New project](#)

NAME	ROLE	COLLABORATORS	DATE CREATED	LAST UPDATED
Demos DSX	Admin		Sep 25, 2017	Jan 09, 2018
Machine Learning, Data Science	Admin	+2	Oct 08, 2017	Dec 22, 2017
Demo DSX	Admin		Nov 21, 2017	Dec 21, 2017

<https://dataplatform.ibm.com/data/discovery?target=offerings&context=analytics>



Add other Watson Data Platform apps

Data Catalog
Discover, catalog, and securely share enterprise data.

[Try it for free](#)

Data Refinery
Self-service data preparation and integration for analytics projects.

[Try it for free](#)

<https://dataplatform.ibm.com/data/discovery/data-catalog/details?target=offerings&context=analytics>

Créate un catálogo.

Data Catalog

Features

Catalog
Create a 360-degree view of your data, no matter where (or in what format) it is stored.

Find
With the right tools, discover the data you need, and collaborate to discover fresh insights.

Govern (Data Catalog Professional)
Control data access by defining policies and monitoring enforcement.

Pricing Plan: Monthly Process shown above reflect the: [United States](#)

Plan	Features	Pricing
<input checked="" type="radio"/> Lite	Lite - Free Free usage Limited to 1 catalog Limited to 5 free discovery connections	Free
<input type="radio"/> Professional	Professional - 500 users included Unlimited catalogs Unlimited discovery connections Policy Authoring & Business Glossary Policy Enforcement Classification & Profiling	\$5,000 USD/Instance \$500 USD/250 Authorized Users

The Lite plan for Data Catalog offers everything you need to begin your journey to becoming a data-centric organization.

[Create](#)

Confirm Creation

Organization: candelaretolaza@es.ibm.com

Plan: Lite

Space: DataSciX

Service name: data-catalog-jz

[Cancel](#) [Confirm](#)

Workshop 2.

Crea y organiza los recursos en un proyecto

1. Crear un proyecto

NAME	ROLE	COLLABORATORS	DATE CREATED	LAST UPDATED
Demos DSX	Admin	1	Sep 29, 2017	Jan 09, 2018
Machine Learning, Data Science	Admin	2	Oct 06, 2017	Dec 22, 2017
Demo DSX	Admin	1	Nov 21, 2017	Dec 21, 2017

Creamos un proyecto nuevo, por ejemplo: Workshop, añadimos una descripción opcional: por ejemplo, Workshop DSX. Seleccionamos el servicio de spark, y dónde lo almacenamos y creamos el proyecto dando al crear.

Un proyecto sirve para organizar tus recursos para trabajar y hacer minería de datos. Los recursos de su proyecto pueden incluir:

Notebooks, Modelos, Flows de SPSS, Streams flows, deployments, archivos de assets de datos y conexiones, colaboradores, marcadores a los recursos de la comunidad, tokens de acceso, enlaces a repositorios de GitHub para publicar notebooks, servicios de spark u otros motores, y otros servicios asociados, como Watson Machine Learning o IBM Streaming.

Si tienes permisos de administrador en un proyecto, tienes control total sobre él. Si tiene permisos de Editor, puedes agregar activos y colaboradores a un proyecto. La página Overview proporciona un resumen del estado actual del proyecto, incluida información

sobre el uso del almacenamiento, la actividad reciente, los colaboradores, los marcadores y los activos.

Workshop

Last Updated: Jan 23 2018

0 Assets

0 Bookmarks

1 Collaborators

Date created: Jan 23 2018

Description: No description available

Storage

Collaborators: View all (1)

Bookmarks: View all (0)

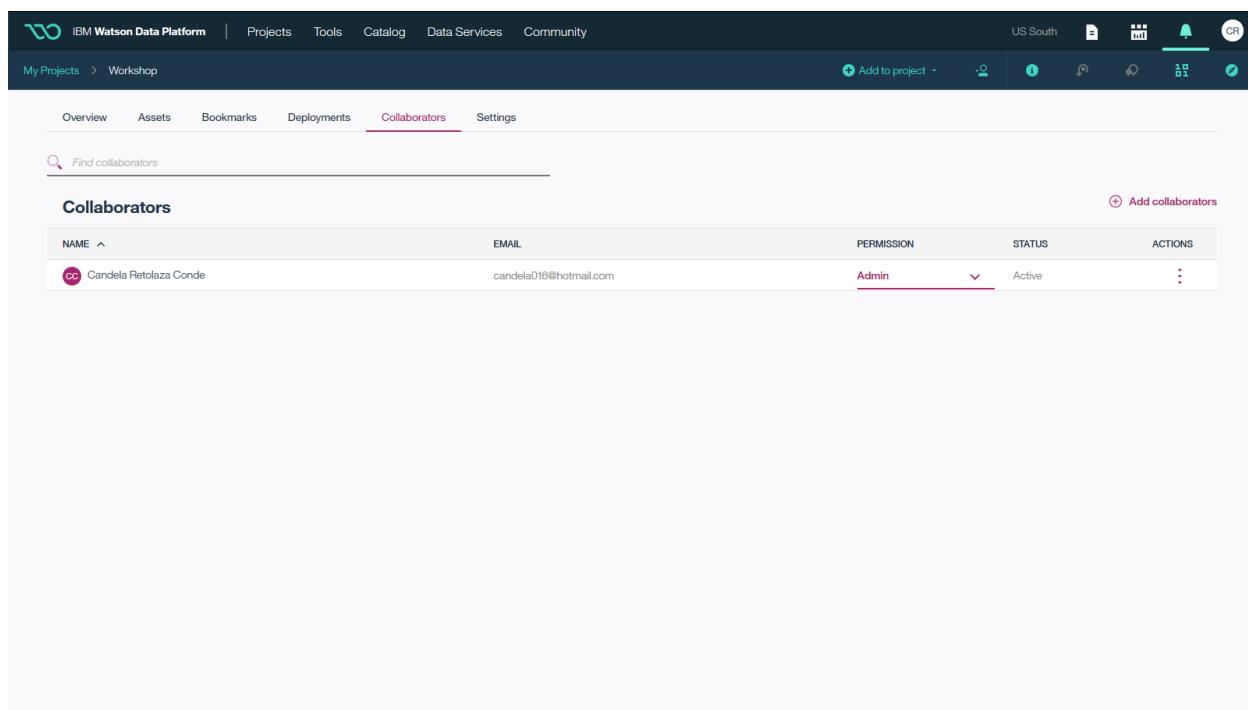
You currently have 0 bookmarks

Transferring data from dataplatform.ibm.com...

Ahora, ya tenemos un proyecto nuevo. Lo primero que nos muestra en el resumen es que no tenemos nada en el proyecto y que solo tiene un colaborador.

2. Añade colaboradores

Lo primero que vamos a hacer en nuestro proyecto nuevo es añadir un nuevo colaborador al proyecto. Pincho en colaboradores y en añadir uno nuevo.

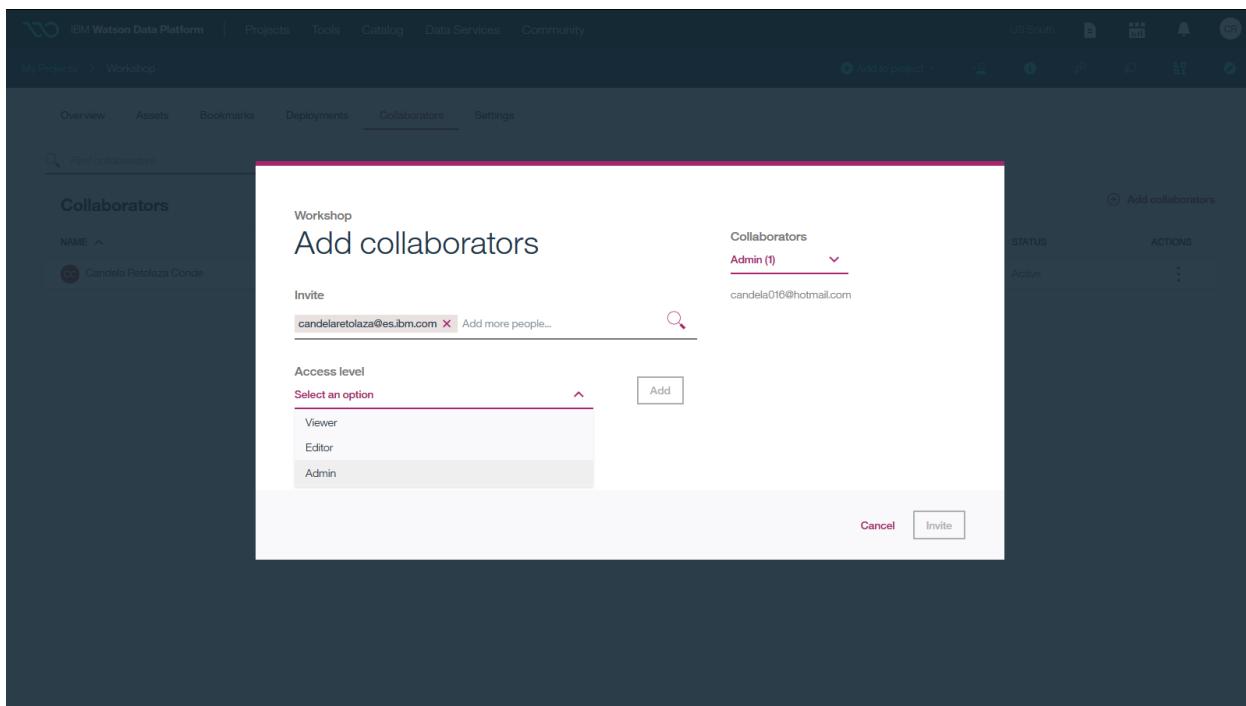


The screenshot shows the IBM Watson Data Platform interface. At the top, there is a navigation bar with links for Projects, Tools, Catalog, Data Services, and Community. On the right side of the top bar, there are icons for US South, a file, a bar chart, a bell, and a user profile. Below the top bar, the main navigation bar for the project shows 'My Projects > Workshop'. The 'Collaborators' tab is selected. A search bar labeled 'Find collaborators' is present. A table lists one collaborator: 'Candela Retolaza Conde' with email 'candela016@hotmail.com', permission 'Admin', and status 'Active'. There is a link to 'Add collaborators'.

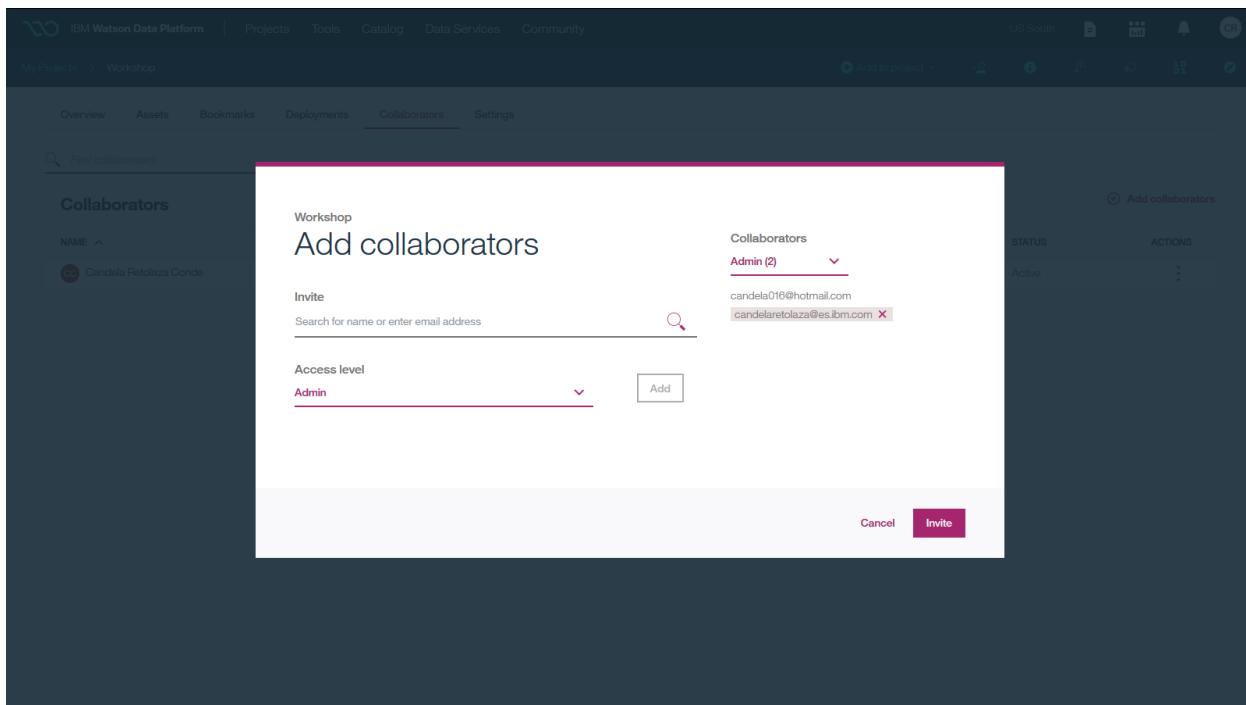
NAME	EMAIL	PERMISSION	STATUS	ACTIONS
Candela Retolaza Conde	candela016@hotmail.com	Admin	Active	⋮

Para poder hacer esto, solo necesito que la persona a la que quiero añadir tenga una cuenta en mi IBM Cloud. Puedes crear cuentas asociadas a tu IBM Cloud tal y como hemos contado en el Workshop 1.2.

Añadimos el correo, seleccionamos el tipo de acceso que queremos dar a esa persona y enviamos la invitación. Y ya estará en nuestro proyecto.



The screenshot shows the 'Add collaborators' dialog for a 'Workshop' project. The dialog has a title 'Add collaborators' and a sub-section 'Workshop'. It includes a search bar 'Invite' with the email 'candelaretolaza@es.ibm.com' and a 'Find' button. Below the search bar is a dropdown for 'Access level' with options 'Viewer', 'Editor', and 'Admin', with 'Admin' selected. There is a 'Select an option' button and an 'Add' button. On the right, a list of collaborators shows 'Admin (1)' with 'candelaretolaza@es.ibm.com'. Below the list are 'STATUS' and 'ACTIONS' columns. At the bottom are 'Cancel' and 'Invite' buttons.



The screenshot shows the 'Add collaborators' dialog for a 'Workshop' project, similar to the first one but with a different email entered. The 'Invite' search bar now contains 'candelaretolaza@es.ibm.com' with a red 'X' button. The 'Access level' dropdown shows 'Admin' selected. The list of collaborators on the right shows 'Admin (2)' with two entries: 'candelaretolaza@es.ibm.com' and 'candelaretolaza@hotmail.com'. The 'STATUS' and 'ACTIONS' columns are present. At the bottom are 'Cancel' and 'Invite' buttons.

Una vez añadido:

NAME	EMAIL	PERMISSION	STATUS	ACTIONS
Candela Retolaza Conde	candela016@hotmail.com	Admin	Active	⋮
Candela Retolaza Conde	candelaretolaza@es.ibm.com	Admin	Active	⋮

Desde aquí puedo cambiar el tipo de permiso de cada colaborador.

3. Aprende a gestionarlo: borrar assets, añadir, crear conexiones...

Assets:

Si tiene permisos de administrador o editor en un proyecto, puedes agregar recursos.

Los tipos de activos enumerados están condicionados a las aplicaciones de Watson Data Platform que tiene. Para agregar assets a un proyecto, elija el tipo de asset en el menú Agregar al proyecto:

Conexiones, datos de fichero plano, datos de bases de datos, Notebooks, flujos de aprendizaje automático, modelos, modelos de SPSS, Streams flows.

My Projects > Workshop

Overview Assets Bookmarks Deployments Collaborators Settings

What assets are you looking for?

Data assets
0 assets selected.

<input type="checkbox"/> NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
you currently have no data assets					

Notebooks + New notebook

NAME	SHARED	SCHEDULED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
you currently have no notebooks							

Streams flows + New streams flow

NAME	MODIFIED BY	LAST MODIFIED	ACTIONS
you currently have no streams flows			

Models + New model

NAME	STATUS	RUNTIME	LAST MODIFIED	ACTIONS
you currently have no models				

Load Files Catalog

Find in storage

0 selected

No files found.

Para agregar flujos de datos a un proyecto, debe seleccionar la herramienta de Refinería de datos y comenzar a limpiar y dar forma a los datos en los activos de datos en su proyecto.

Si tiene permisos de administrador en un proyecto, puede eliminar activos. Para eliminar un activo, elija Eliminar en el menú ACCIONES al lado del nombre del activo.

Gestión de proyectos:

Podremos gestionar el almacenamiento, los servicios asociados, los tokens, ver a quién pertenece la cuenta y conectar el proyecto en github. Se propone al lector que explore por la herramienta antes de comenzar el siguiente Workshop.

Project information

Project name: Workshop

Description: Project description

Type: Cloud Object Storage (Beta) Bucket Name: workshop984b65fe9c4b427392cac90bb0c561aa

0 Byte Used 0% of 5 GB used

Associated services

NAME	SERVICE TYPE	PLAN	ACTIONS
Spark-ae	Spark	Personal	⋮

Access tokens

you currently have no access tokens

Connect to a GitHub repository

Repository URL: `https://github.com/owner/repository-name`

Connect

Storage

Type: Cloud Object Storage (Beta) Bucket Name: workshop984b65fe9c4b427392cac90bb0c561aa

0 Byte Used 0% of 5 GB used

Associated services

NAME	SERVICE TYPE	PLAN	ACTIONS
Spark-ae	Spark	Personal	⋮

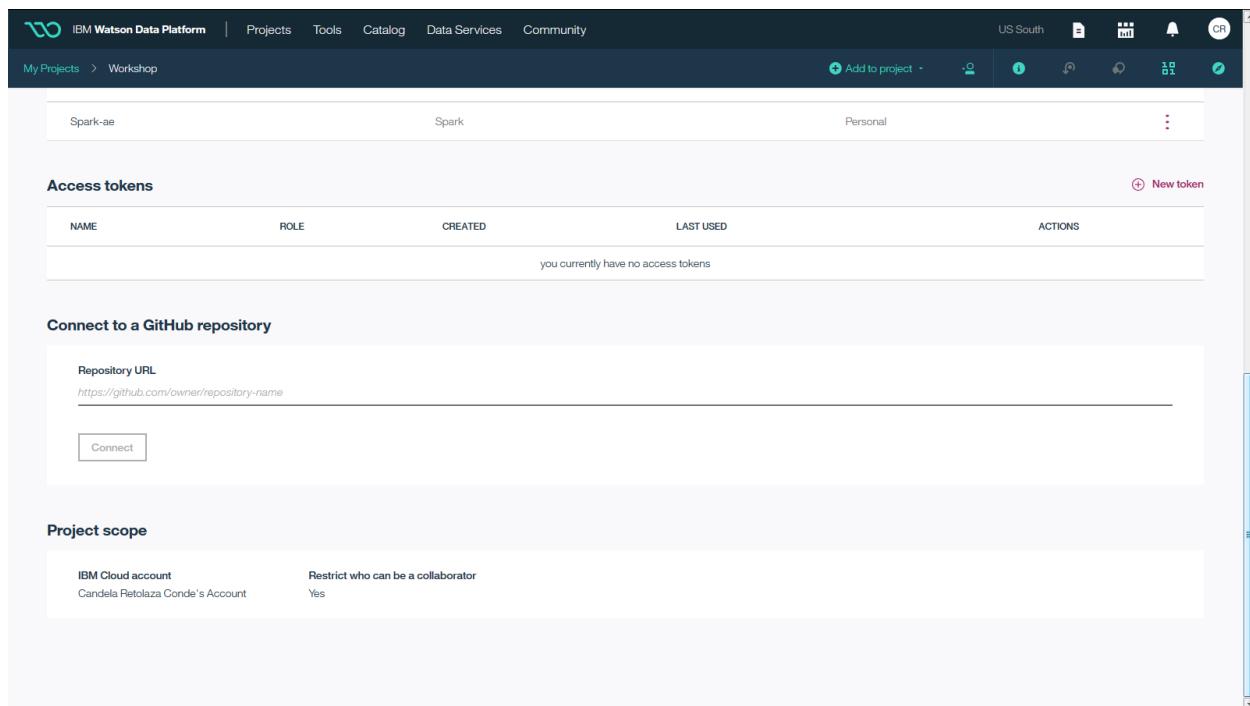
Access tokens

you currently have no access tokens

Connect to a GitHub repository

Repository URL: `https://github.com/owner/repository-name`

Connect



The screenshot shows the IBM Watson Data Platform interface. At the top, there is a navigation bar with links for Projects, Tools, Catalog, Data Services, and Community. On the right side of the top bar, there are icons for US South, a file, a bar chart, a bell, and a user profile. Below the top bar, the main content area shows a project named "Spark" under the "My Projects" section. The project details include "Spark" as the name, "Personal" as the owner, and a "Spark" icon. There is a "Spark" tab selected. Below this, there is a section titled "Access tokens" with a "New token" button. A table header for "Access tokens" is shown with columns for NAME, ROLE, CREATED, LAST USED, and ACTIONS. A message below the table states "you currently have no access tokens". Further down, there is a section titled "Connect to a GitHub repository" with a "Repository URL" input field containing "https://github.com/owner/repository-name" and a "Connect" button. At the bottom, there is a "Project scope" section with "IBM Cloud account" set to "Candela Retolaza Conde's Account" and "Restrict who can be a collaborator" set to "Yes".

Workshop 3.

Consigue y prepara los datos y analiza los datos de manera sencilla

1. Añadir datos

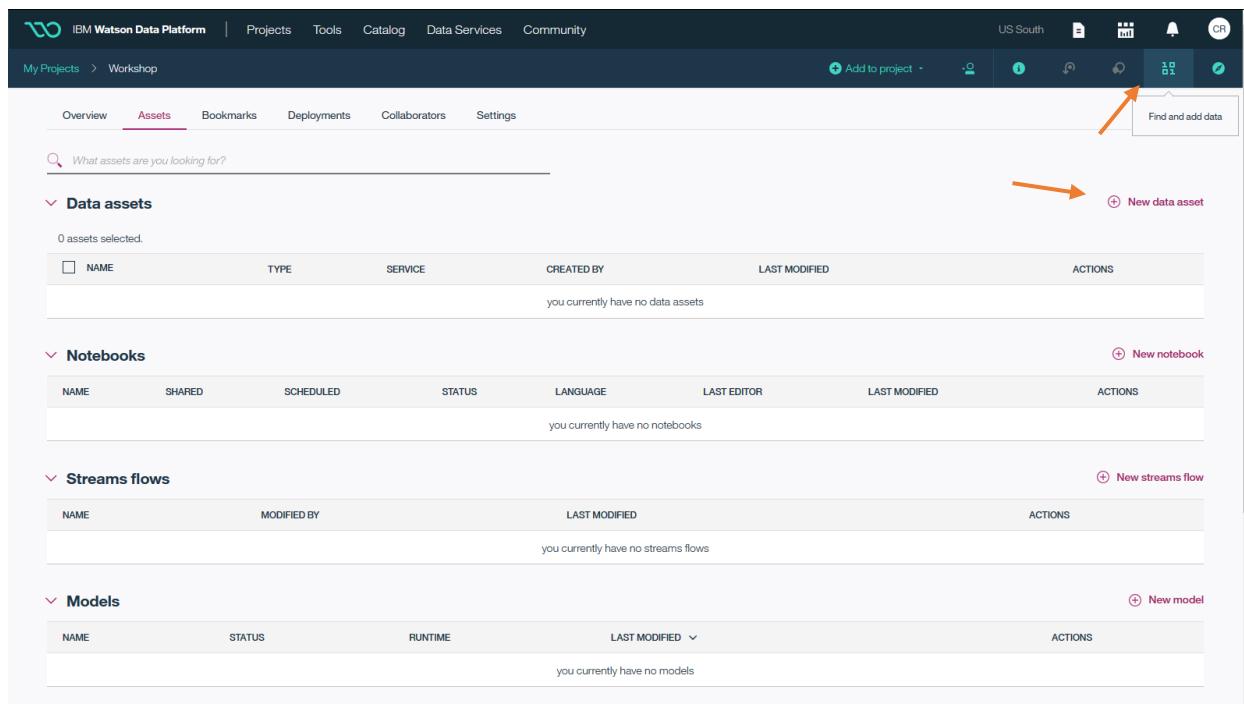
Después de crear un proyecto, tenemos que añadir datos para poder trabajar con los mismos. Todos los colaboradores en el proyecto están autorizados automáticamente para acceder a los datos en el proyecto.

Puede añadir assets de datos de estas fuentes a un proyecto:

- Archivos locales
- Comunidad
- Conexiones de base

Vamos a comenzar añadiendo un fichero local a nuestro proyecto. Para agregar archivos de datos a un proyecto:

Desde la página **Assets** de su proyecto, puedes añadir datos de dos maneras, en el icono  arriba a la derecha, o clickando en **New data asset**.



IBM Watson Data Platform | Projects Tools Catalog Data Services Community

My Projects > Workshop

Overview Assets Bookmarks Deployments Collaborators Settings

What assets are you looking for?

▼ Data assets

0 assets selected.

<input type="checkbox"/> NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
you currently have no data assets					

▼ Notebooks

0 assets selected.

NAME	SHARED	SCHEDULED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
you currently have no notebooks							

▼ Streams flows

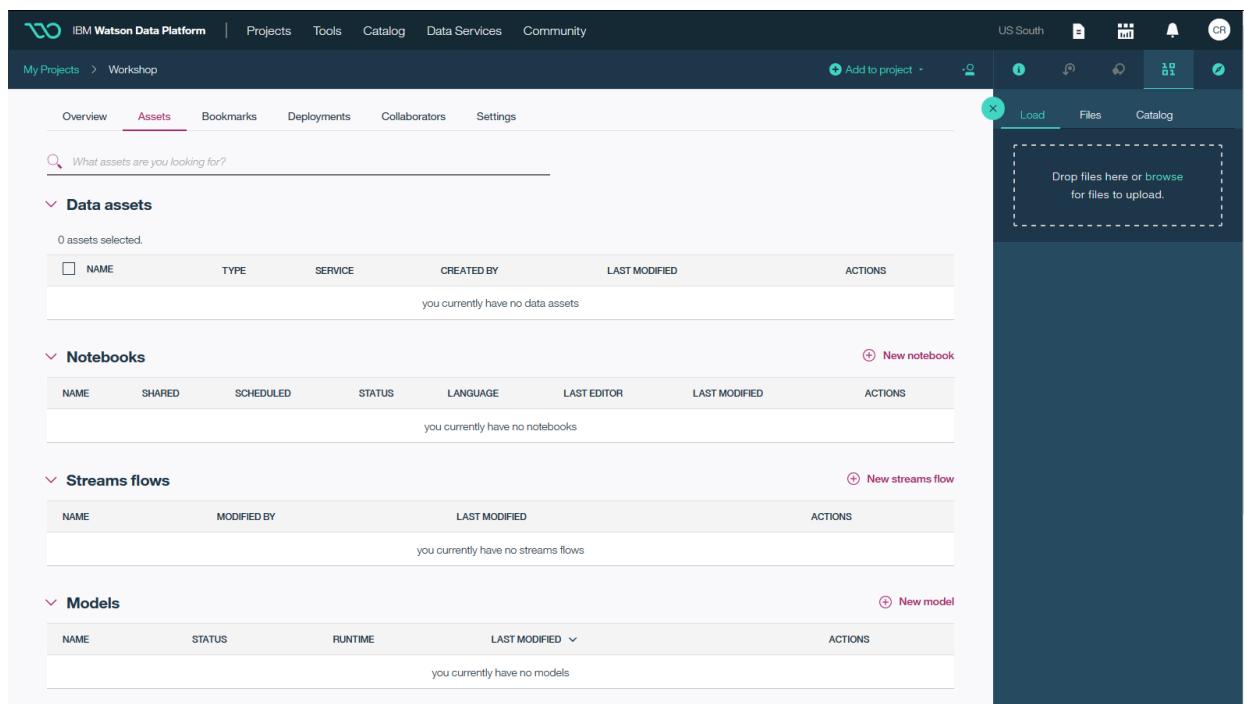
0 assets selected.

NAME	MODIFIED BY	LAST MODIFIED	ACTIONS
you currently have no streams flows			

▼ Models

0 assets selected.

NAME	STATUS	RUNTIME	LAST MODIFIED	ACTIONS
you currently have no models				



IBM Watson Data Platform | Projects Tools Catalog Data Services Community

My Projects > Workshop

Overview Assets Bookmarks Deployments Collaborators Settings

What assets are you looking for?

▼ Data assets

0 assets selected.

<input type="checkbox"/> NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
you currently have no data assets					

▼ Notebooks

0 assets selected.

NAME	SHARED	SCHEDULED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
you currently have no notebooks							

▼ Streams flows

0 assets selected.

NAME	MODIFIED BY	LAST MODIFIED	ACTIONS
you currently have no streams flows			

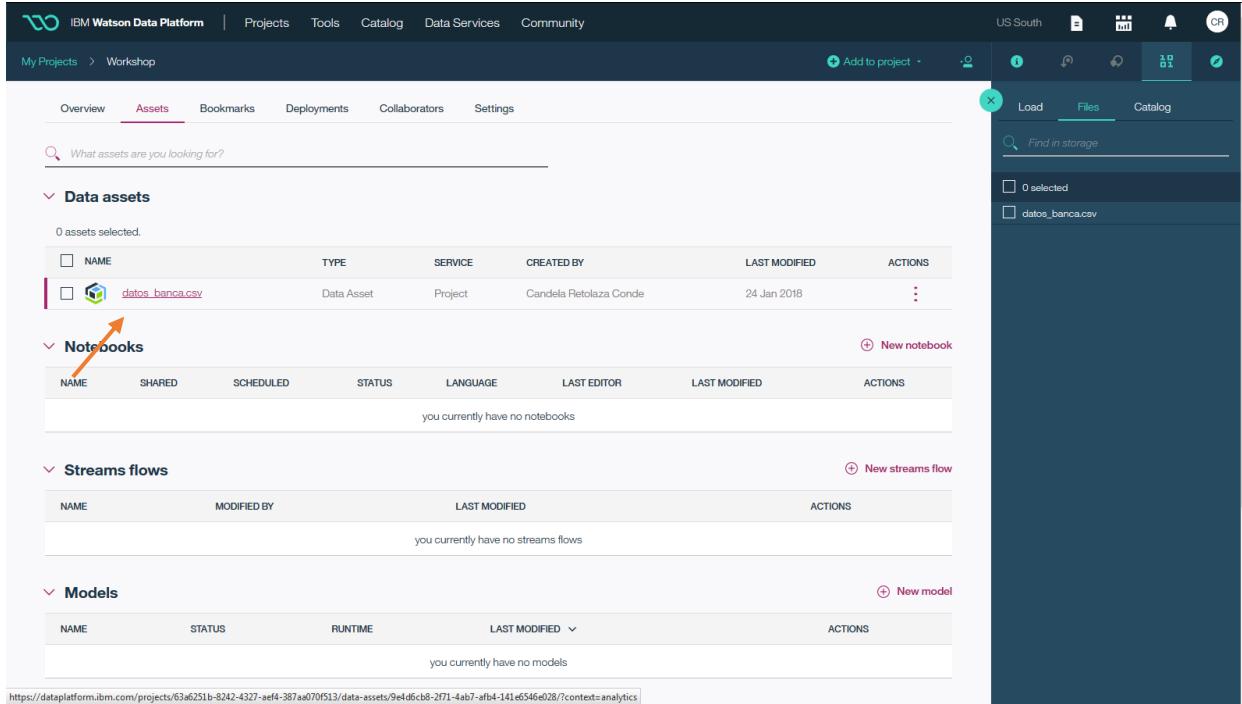
▼ Models

0 assets selected.

NAME	STATUS	RUNTIME	LAST MODIFIED	ACTIONS
you currently have no models				

Haga clic en **Load** y luego busque los archivos en el PC o arrástrelos.

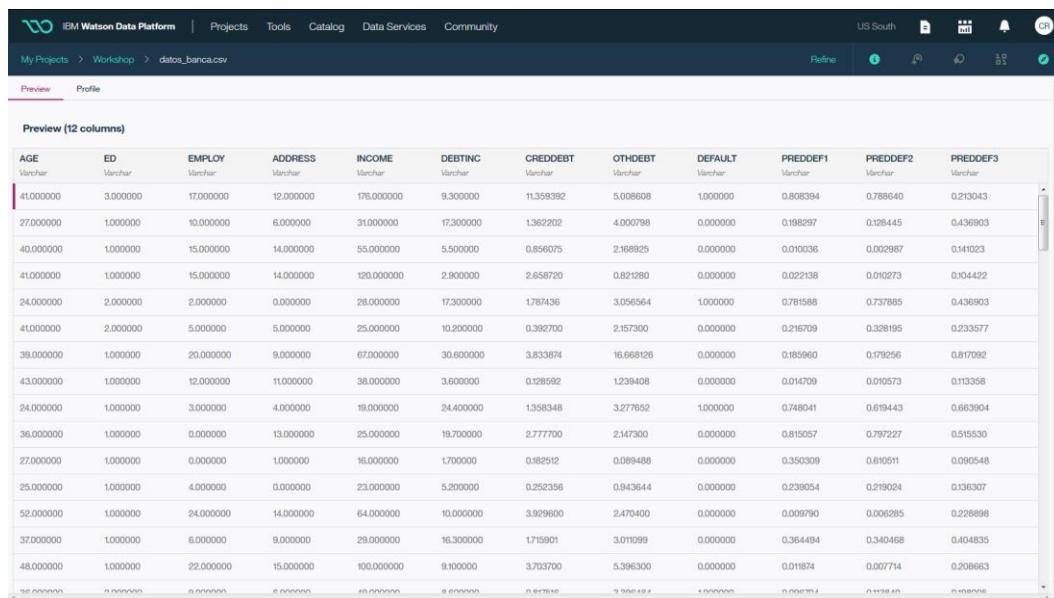
Subimos el fichero `datos_banca.csv`. Debes permanecer en la página hasta que la carga esté completa. Puede cancelar un proceso de carga en curso si desea dejar de cargar un archivo.



The screenshot shows the 'Assets' tab in the IBM Watson Data Platform. The left sidebar lists categories: Data assets, Notebooks, Streams flows, and Models. The 'Data assets' section contains a table with one row for 'datos_banca.csv'. The 'Notebooks' section is empty. The 'Streams flows' and 'Models' sections are also empty. The right sidebar shows a file browser with '0 selected' and 'datos_banca.csv' listed. The top navigation bar includes 'My Projects', 'Workshop', 'Add to project', and various project status indicators.

Los archivos se guardan en el object storage que está asociado con su proyecto y se enumeran como assets de datos en la página de **Assets** de su proyecto.

Haciendo click en el fichero, podemos ver cómo son nuestros datos.



The screenshot shows the preview of the 'datos_banca.csv' file. The top navigation bar includes 'My Projects', 'Workshop', 'Refine', and project status indicators. The preview table has 12 columns: AGE, ED, EMPLOY, ADDRESS, INCOME, DEBTINC, CREDDEBT, OTHDEBT, DEFAULT, PREDEF1, PREDEF2, and PREDEF3. The table contains 20 rows of data. The first row is highlighted with a red border. The columns are defined as Varchar, and Varchar respectively.

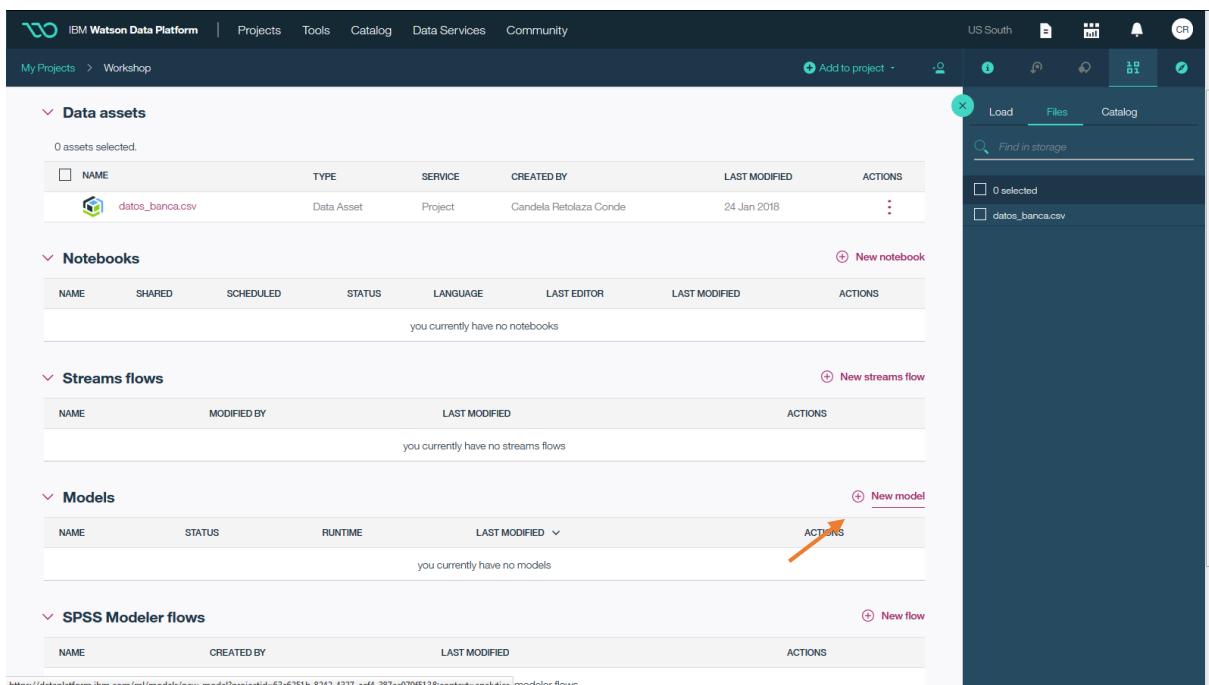
El fichero que acabamos de subir a nuestro proyecto contiene información de demográfica de clientes de un banco ficticio, con datos como: edad, nivel educativo, años en el trabajo actual, años en la misma vivienda, salario... en el que, además, tenemos información sobre los créditos que tiene cada cliente y un histórico de datos de clientes sobre si han hecho impago.

Por tanto, vamos a suponer que un banco está preocupado por el posible impago de sus créditos. Vamos a utilizar datos de créditos anteriores para predecir los clientes potenciales que tendrán problemas para pagar sus créditos, a estos clientes de alto riesgo se les puede negar un crédito u ofrecer otros productos.

Podremos refinar los datos desde Data Renifery o haciendo click en **Refine** (aún en BETA).

2. Crear un modelo automático

Vamos a crear un modelo, en este caso, creamos uno automático o semi-automático, con el fichero datos_banca.csv que acabamos de subir y entender, y queremos intentar predecir qué variables producen impago. Clicamos en **New Model**.



The screenshot shows the IBM Watson Data Platform interface. The top navigation bar includes 'IBM Watson Data Platform', 'Projects', 'Tools', 'Catalog', 'Data Services', and 'Community'. The 'Projects' tab is selected, showing 'My Projects' and 'Workshop'. The main content area is divided into sections: 'Data assets', 'Notebooks', 'Streams flows', 'Models', and 'SPSS Modeler flows'. The 'Models' section is currently active, displaying a table with columns: NAME, STATUS, RUNTIME, LAST MODIFIED, and ACTIONS. A red arrow points to the '+ New model' button in the ACTIONS column. The table shows '0 selected' and '0 models'.

Definimos el nombre del modelo, y debemos de tener un servicio de machine learning asociado a nuestra cuenta. Nos creamos para comenzar una versión gratuita.

New model BETA

Define model details

Name
Modelo Predictivo Impago

Description
Model description

Machine Learning Service
No Machine Learning service instances associated with your project.
Associate a Machine Learning service instance with your project on the project settings page, then click the reload button below to refresh the instances available for association with your new model builder instance.

Reload

Select model type

Model builder From sample

Spark Service
Spark-ae

Automatic
Prepare my data and create a model automatically

Manual
Let me prepare my data and select which models to train

Need something more flexible? Create a [notebook](#) or design an [SPSS Modeler flow](#).

Create

Machine Learning

Existing New

Machine Learning

IBM Watson Machine Learning is a full-service Bluemix offering that makes it easy for developers and data scientists to work together to integrate predictive capabilities with their applications. The Machine Learning service is a set of REST APIs that you can call from any programming language to develop applications that make smarter decisions, solve tough problems, and improve user outcomes.

Features

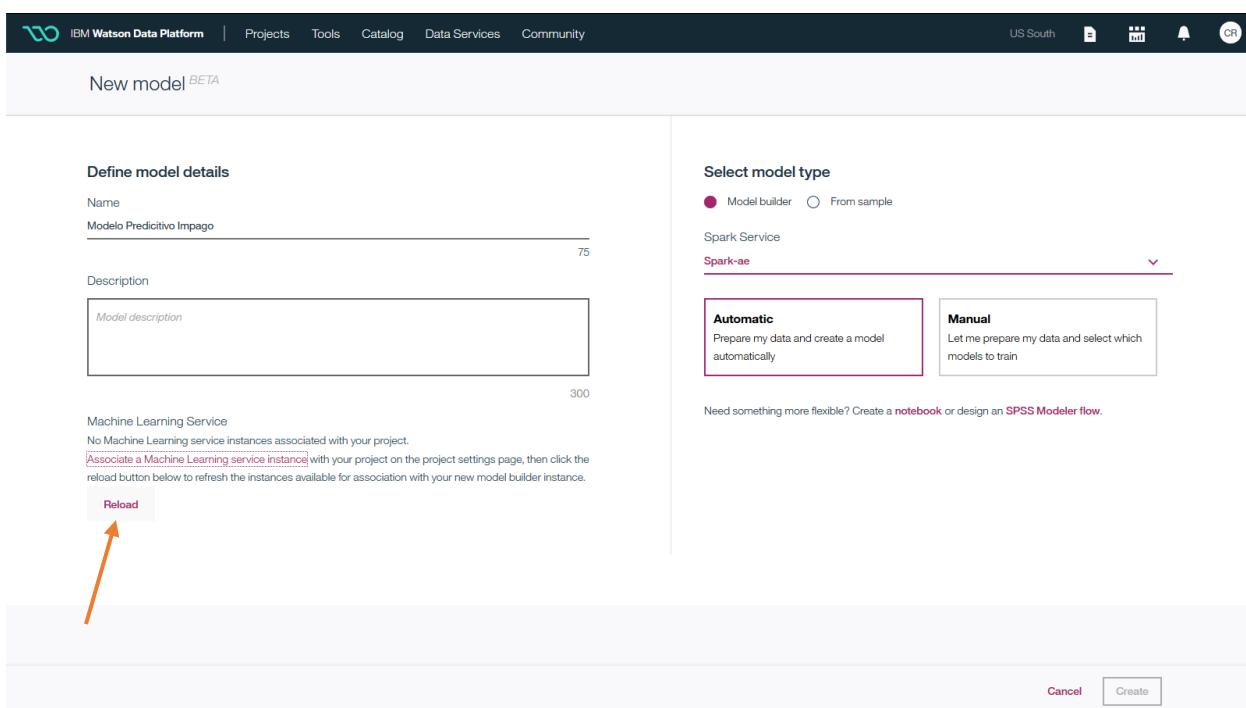
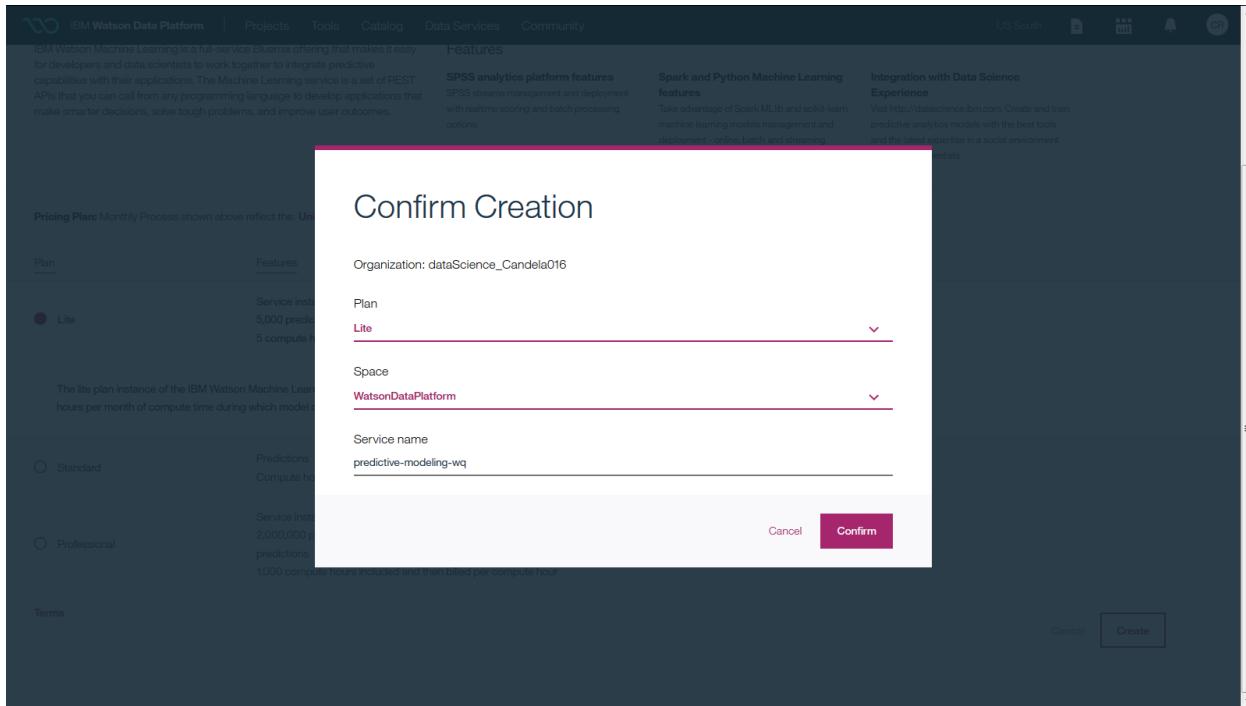
SPSS analytics platform features
SPSS streams management and deployment with realtime scoring and batch processing options.

Spark and Python Machine Learning features
Take advantage of Spark MLlib and scikit-learn machine learning models management and deployment - online, batch and streaming.

Integration with Data Science Experience
Visit <http://datascience.ibm.com>. Create and train predictive analytics models with the best tools and the latest expertise in a social environment built by data scientists.

Pricing Plan: Monthly Process shown above reflect the: [United States](#)

Plan	Features	Pricing
<input checked="" type="radio"/> Lite	Service instance (5 models per instance) 5,000 predictions 5 compute hours	Free
<input type="radio"/> Standard	Predictions Compute hours Service instance	\$0.5 USD/1,000 predictions \$0.45 USD/hour \$1000 USD/instance



Si hacemos **Reload**, tendremos nuestro servicio de Machine Learning listo para usar:

New model BETA

Define model details

Name
Modelo Predictivo Impago

Description
Model description

Machine Learning Service
predictive-modeling-wq

Select model type

Model builder From sample

Spark Service
Spark+e

Automatic
Prepare my data and create a model automatically

Manual
Let me prepare my data and select which models to train

Need something more flexible? Create a [notebook](#) or design an [SPSS Modeler flow](#).

[Cancel](#) [Create](#)

Crearemos un modelo automático. Seleccionamos el fichero sobre el que queremos trabajar, en este caso solo tenemos uno:

My Projects > Workshop > Modelo Predictivo Impago

Select Data

Train

Evaluate

Select data asset

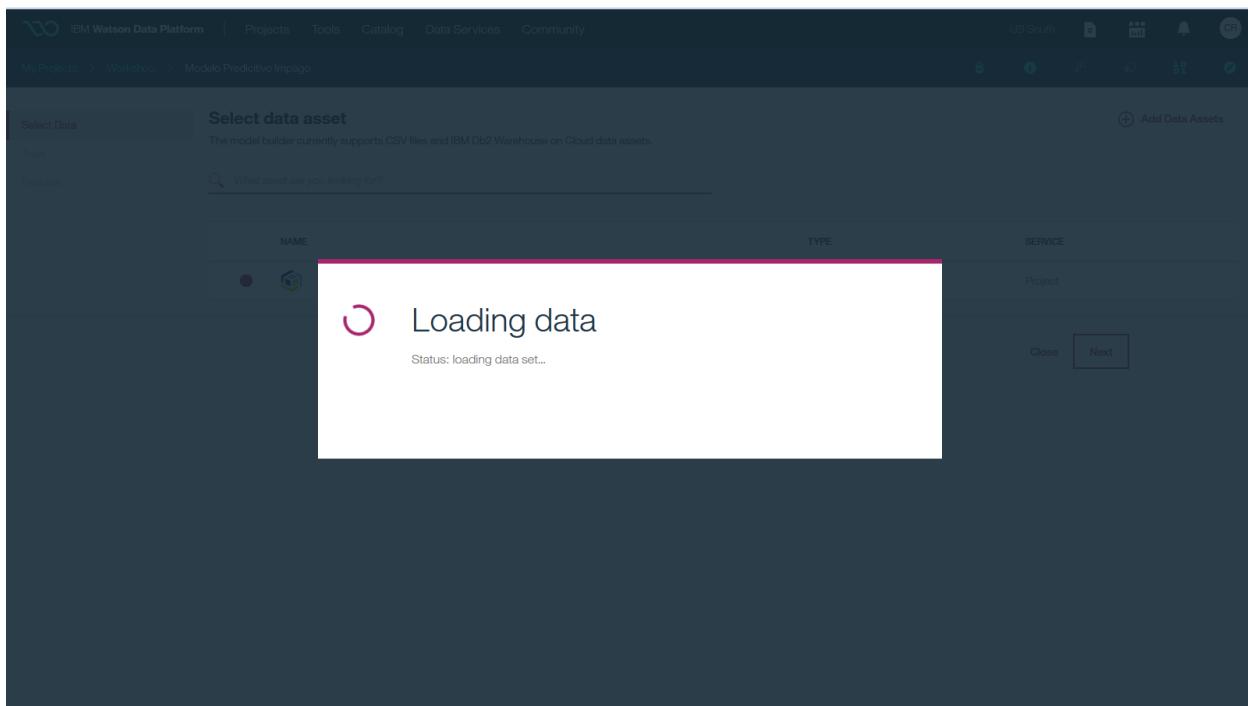
The model builder currently supports CSV files and IBM Db2 Warehouse on Cloud data assets.

What asset are you looking for?

NAME	TYPE	SERVICE
datos_banca.csv	Data Asset	Project

Click to preview data

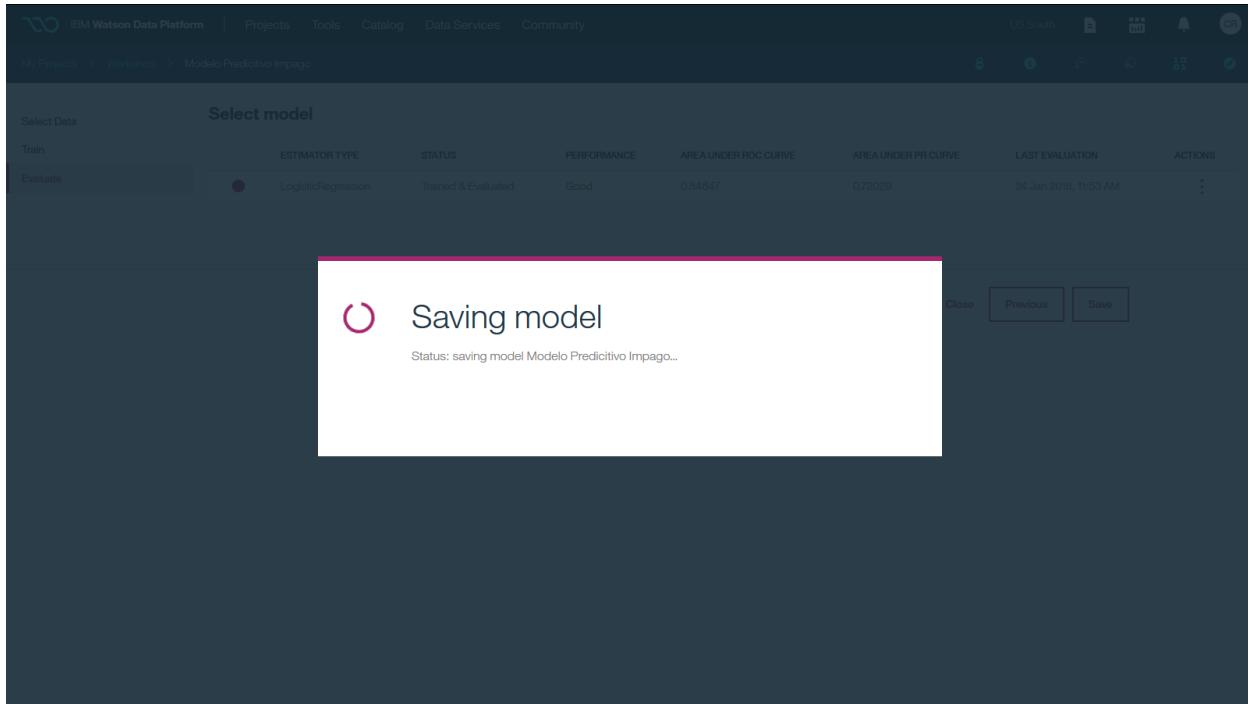
[Close](#) [Next](#)



Para hacer un primer modelo, vamos a utilizar la variable de impago que se llama **Default** para analizar qué factores hacen que un cliente sea más propenso a hacer impago que otro. Además, le decimos que utilice todas las demás variables para predecir.

Una vez hemos elegido que tipo de modelo de modelo queremos y seleccionadas quiénes son mis variables, vemos que también nos selecciona una parte de los datos para entrenamiento, y otra parte para testear. Ejecutamos.

Nos dice que tipo de estimador a utilizado (una regresión logística) y si es una buena predicción o no, y el área bajo la curva ROC y el área bajo la curva PR. Podemos guardar el modelo, o volver atrás y repetir con otro modelo.



The screenshot shows the IBM Watson Data Platform interface. At the top, there is a navigation bar with links for 'Projects', 'Tools', 'Catalog', 'Data Services', and 'Community'. Below the navigation bar, the path 'My Projects > Workshops > Modelo-Predictivo Impago' is visible. The main content area is titled 'Select model' and shows a table with two rows: 'Train' and 'Evaluate'. The 'Evaluate' row is selected, indicated by a blue dot. The table columns are 'ESTIMATOR TYPE', 'STATUS', 'PERFORMANCE', 'AREA UNDER ROC CURVE', 'AREA UNDER PR CURVE', 'LAST EVALUATION', and 'ACTIONS'. The 'Evaluate' row has values: 'LogisticRegression', 'Trained & Evaluated', 'Good', '0.64847', '0.72029', '24 Jan 2018, 11:53 AM', and a 'More' button. A modal dialog box is overlaid on the page, titled 'Saving model'. It contains the text 'Status: saving model Modelo Predictivo Impago...' and has three buttons: 'Close', 'Previous', and 'Save'.

3. Despliega el modelo en Watson Machine Learning

Cuando guardamos el modelo, nos muestra un resumen del modelo, nos deja evaluarlo y desplegarlo.

IBM Watson Data Platform | Projects Tools Catalog Data Services Community

My Projects > Workshop > Modelo Predictivo Impago

US South

Modelo Predictivo Impago

Overview Evaluation Deployments

Summary

Machine learning service	predictive-modeling-wq
Runtime environment	spark-2.0
Training date	24 Jan 2018, 11:55 AM
Label column	default
Latest version	b13c03db-7754-41f9-88ed-7ee99ebdfb33
Model builder details	View

Input Schema

COLUMN	TYPE
age	decimal(31,6)
ed	decimal(31,6)
employ	decimal(31,6)
address	decimal(31,6)

IBM Watson Data Platform | Projects Tools Catalog Data Services Community

My Projects > Workshop > Modelo Predictivo Impago

US South

Modelo Predictivo Impago

Overview Evaluation Deployments

Last Evaluation Result

Version	b13c03db-7754-41f9-88ed-7ee99ebdfb33
Phase	setup
AreaUnderPR	0.72
AreaUnderROC	0.846

Performance Monitoring

Configure performance monitoring to evaluate and retrain the model periodically to ensure the model performance is acceptable. You will need an existing IBM Db2 Warehouse on Cloud connection associated with your project to be used as your feedback data connection.

[Configure Performance Monitoring](#)

Versions

TIME	VERSION	DEPLOYED	ACTIONS
24 Jan 2018 11:58am	b13c03db-7754-41f9-88ed-7ee99ebdfb33		⋮

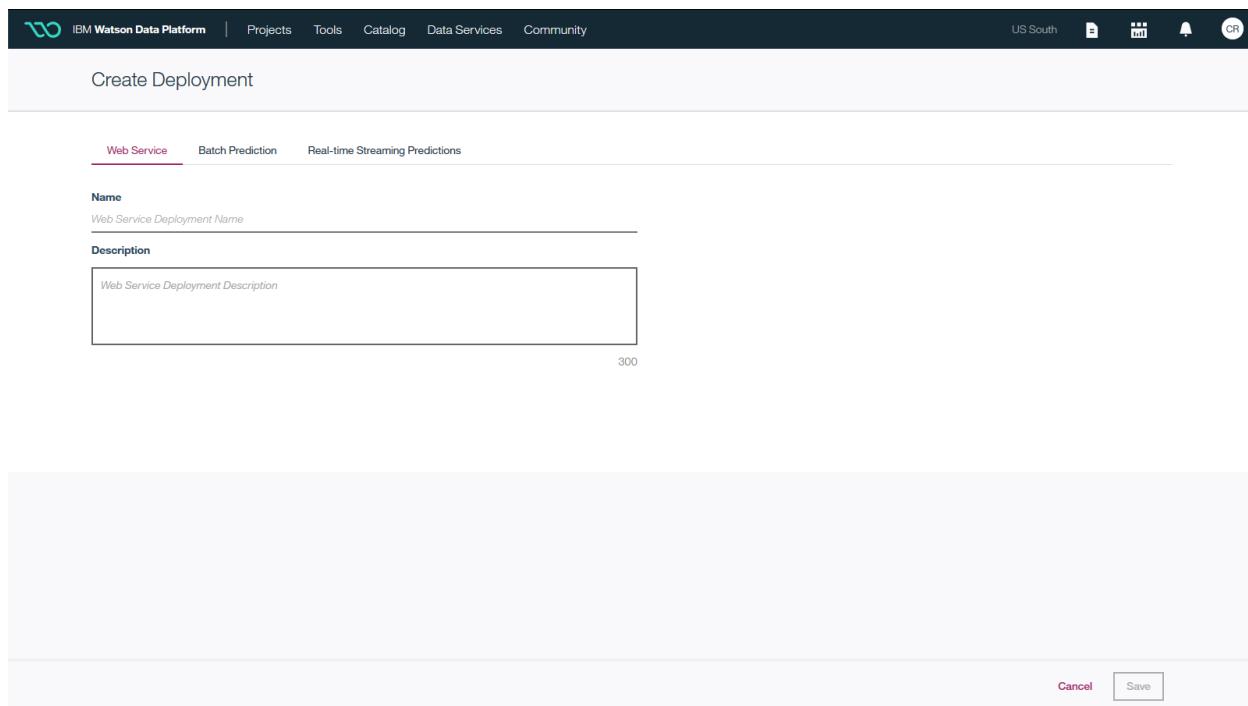
Ahora, podemos hacer un test y así, probar con otros datos nuestro modelo, y ver que output nos devuelve: por ejemplo 41 años, 1 educación que es nivel básico, 1 año en el mismo empleo y dejamos todo lo demás, le damos a probar y nos devuelve el resultado de la predicción.

Para poder hacer y configurar la supervisión del rendimiento para evaluar y volver a entrenar el modelo periódicamente para garantizar que el rendimiento del modelo sea aceptable, se necesita una conexión existente de IBM Db2 Warehouse en la nube asociada con su proyecto para utilizarla como su conexión de datos de retroalimentación.

También podemos ponerlo en producción, entrando a la pestaña de **deployments**. Añadimos un deployment nuevo.

The screenshot shows the IBM Watson Data Platform interface. At the top, there is a navigation bar with links for 'Projects', 'Tools', 'Catalog', 'Data Services', and 'Community'. Below the navigation bar, the path 'My Projects > Workshop > Modelo Predictivo Impago' is visible. The main content area is titled 'Modelo Predictivo Impago' and shows three tabs: 'Overview', 'Evaluation', and 'Deployments'. The 'Deployments' tab is currently selected. A table is present with columns: 'NAME', 'STATUS', 'DEPLOYMENT TYPE', and 'ACTIONS'. A red '+' button labeled 'Add Deployment' is located in the top right corner of the table area. The table displays the message 'Your model is not deployed.'

Podemos desplegar de tres maneras diferentes nuestros modelos: Web service, Batch Prediction y Real-time Streaming Predictions.



IBM Watson Data Platform | Projects Tools Catalog Data Services Community US South

Create Deployment

Web Service Batch Prediction Real-time Streaming Predictions

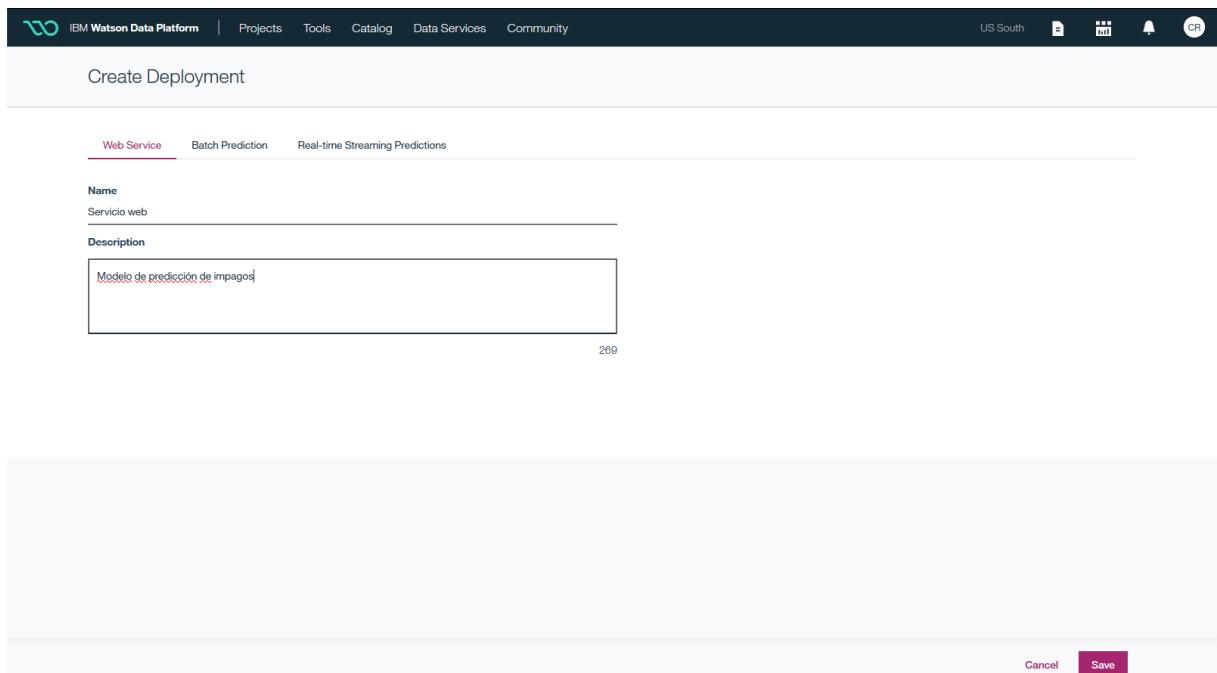
Name
Web Service Deployment Name

Description
Web Service Deployment Description

300

Cancel Save

Una vez creado el servicio, está listo para utilizarlo. Es decir, creamos por ejemplo un deployment ONLINE que nos servirá para crear una página web o una aplicación móvil, Podemos hacer un deployment en batch, para que se ejecute cada cierto tiempo, o utilizar el Streaming para hacer el deployment en tiempo real.



IBM Watson Data Platform | Projects Tools Catalog Data Services Community US South

Create Deployment

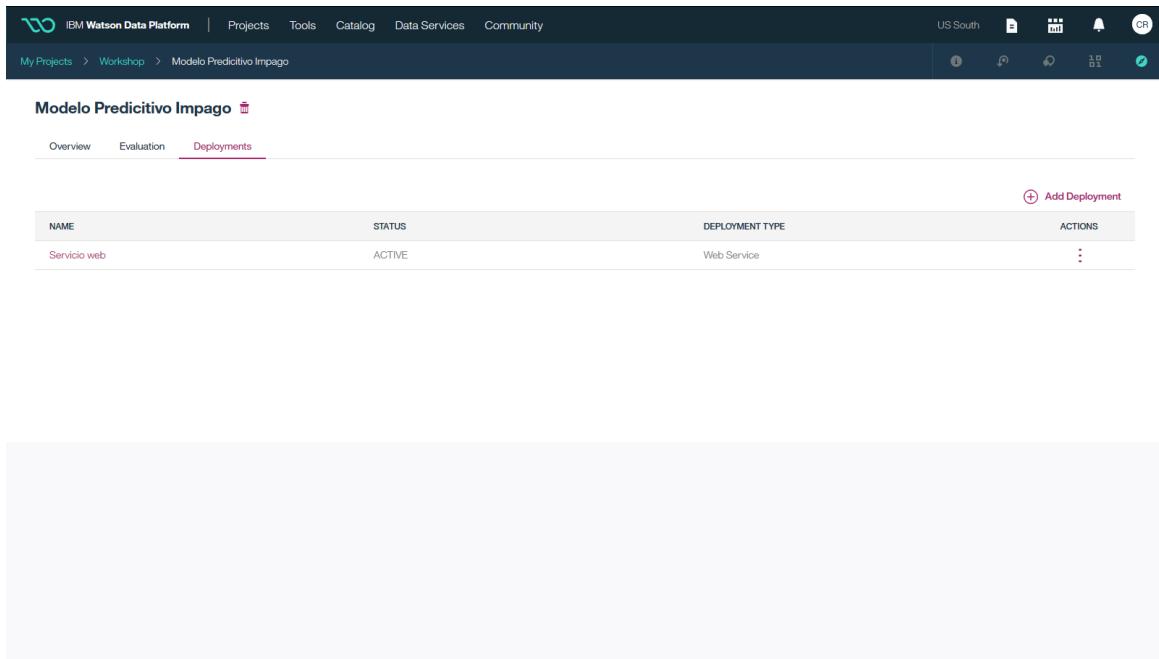
Web Service Batch Prediction Real-time Streaming Predictions

Name
Servicio web

Description
Modelo de predicción de imágenes

200

Cancel Save



Modelo Predictivo Impago

Overview Evaluation Deployments

Add Deployment

NAME	STATUS	DEPLOYMENT TYPE	ACTIONS
Servicio web	ACTIVE	Web Service	⋮

Se propone al lector que cree algún despliegue del modelo obtenido.

4. Crear un modelo semi-automático o manual.

Para finalizar el Workshop 2, vamos a mostrar cómo hacer un modelo, en lugar de automático, manual. Seguimos los mismos pasos que en el apartado 2.2. Pinchamos en **New Model**.

My Projects > Workshop

Data assets
0 assets selected.

NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
datos_banca.csv	Data Asset	Project	Candela Retolaza Conde	24 Jan 2018	⋮

Notebooks

NAME	SHARED	SCHEDULED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
you currently have no notebooks							

Streams flows

NAME	MODIFIED BY	LAST MODIFIED	ACTIONS
you currently have no streams flows			

Models

NAME	STATUS	RUNTIME	LAST MODIFIED	ACTIONS
Modelo Predictivo Impago	trained	spark-2.0	24 Jan 2018	⋮

SPSS Modeler flows

NAME	CREATED BY	LAST MODIFIED	ACTIONS
https://dataplatform.ibm.com/ml/models/new-model?projectId=63a6251b-8242-4327-ae4-387a070f513&context=data			

Igual que antes, definimos el modelo, y ahora seleccionamos modelo **MANUAL**

New model BETA

Define model details

Name
Modelo predictivo automático

Description
Model description

Machine Learning Service
predictive-modeling-wq

Select model type

Model builder From sample

Spark Service
Spark-ae

Automatic
Prepare my data and create a model automatically

Manual
Let me prepare my data and select which models to train

Need something more flexible? Create a [notebook](#) or design an [SPSS Modeler flow](#).

Cancel **Create**

Igual que antes, seleccionamos el fichero de datos de banca.

La diferencia es que ahora nos sugiere una de las técnicas y podemos añadir estimadores (que en la manera automática elegía por nosotros).

IBM Watson Data Platform | Projects Tools Catalog Data Services Community

My Projects > Workshop > Modelo predictivo automático

US South

Train Evaluate

Select Data

Column value to predict (Label Col): default (Decimal)

Feature columns: All (default)

Suggested technique: **Binary Classification**

Classify new data into defined categories based on existing data. Choose if your label column contains two distinct categories.

Multiclass Classification

Classify new data into defined categories based on existing data. Choose if your label column contains a discrete number of categories.

Regression

Predict values from a continuous set of values. Choose if your label column contains a large number of values.

Validation Split: Train: 80 Test: 20 Holdout: 20

Add Estimators

Configured estimators

Close Previous Next

Podemos seleccionar uno o varios estimadores. Añadimos y ejecutamos.

IBM Watson Data Platform | Projects Tools Catalog Data Services Community

My Projects > Workshop > Modelo predictivo automático

Select Data

Set

Column definition

Training

Validation

Add

Cancel Add Previous Next

Select estimator(s)

What type of estimator are you looking for?

Logistic Regression

Analyses a data set in which there are one or more independent variables that determine one of two outcomes. Only binary.

Decision Tree Classifier

Maps observations about an item represented in the branches to conclusions about the item's target value (represented in).

Random Forest Classifier

Constructs multiple decision trees to produce the label that is a mode of each decision tree. It supports both binary and...

Gradient Boosted Tree Classifier

Produces a classification prediction model in the form of an ensemble of decision trees. It only supports binary labels, a...

Estimators

ESTIMATOR TYPE	STATUS	PERFORMANCE	AREA UNDER ROC CURVE	AREA UNDER PR CURVE	LAST EVALUATION	ACTIONS
RandomForestClassifier	Trained & Evaluated	Good	0.84825	0.634	24 Jan 2018, 12:39 PM	⋮
LogisticRegression	Trained & Evaluated	Good	0.83114	0.61084	24 Jan 2018, 12:39 PM	⋮
DecisionTreeClassifier	Trained & Evaluated	Poor	0.62931	0.43677	24 Jan 2018, 12:39 PM	⋮

Ahora de los tres estimadores, dos son buenos, y podemos guardarlos y desplegarlos tal y como se explicó en el apartado 2.3.

Workshop 4.

Parte predictiva

1. Notebooks

Para aquellos que no están familiarizados con los Notebooks de Jupiter, los Notebooks permiten a los equipos combinar documentación y código, ejecutar programas línea por línea y combinar los resultados en convincentes visualizaciones. Se trata de un entorno unificado para la colaboración y totalmente accesible por profesionales no técnicos a través de un navegador web.

Para crear un Notebook en IBM Data Science Experience (DSX):

Paso 1. Entrar en el proyecto, y desde la vista de Assets del proyecto, haga clic en el enlace Nuevo Notebook.

The screenshot shows the IBM Watson Data Platform interface. The top navigation bar includes 'IBM Watson Data Platform', 'Projects', 'Tools', 'Catalog', 'Data Services', and 'Community'. The top right corner shows 'US South' and various status icons. The main content area is titled 'My Projects > Workshop'. The 'Assets' tab is selected, showing the 'Data assets' section with a single entry: 'datos_banca.csv' (Data Asset, Project, Created by Candelaria Rielolaza Conde, Last modified 24 Jan 2018). Below this is the 'Notebooks' section, which displays the message 'you currently have no notebooks'. To the right of the main content area is a sidebar with 'Load', 'Files', and 'Catalog' tabs, and a search bar for 'Find in storage'.

En la ventana Crear Notebook, especifique el método a usar para crear su notebook.

New notebook

Blank From file From URL

Name*
Notebook de ejemplo

Description
Type your Description here

Language*
Python 2 R Scala Python 3.5 Experimental

Spark version*
2.1 2.0

Spark service*
Spark-ae

Cancel Create Notebook

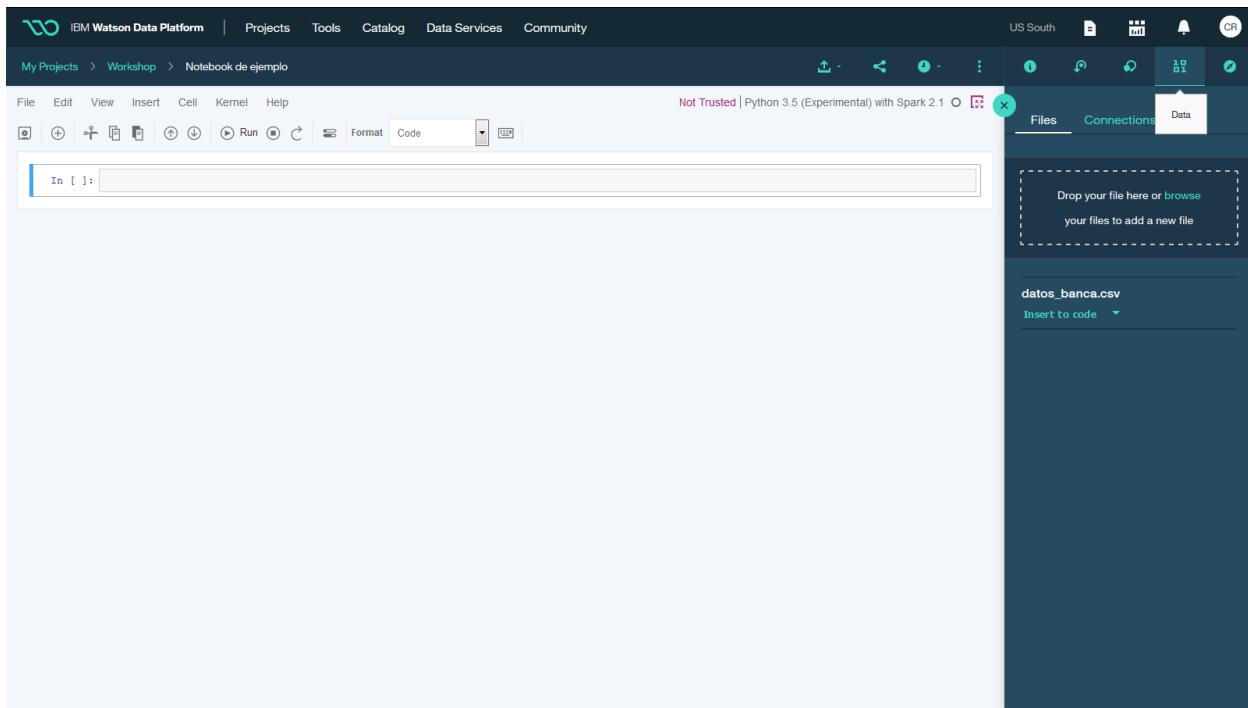
Puedes crear un notebook en blanco, cargar un archivo de notebook desde su sistema de archivos o cargar un archivo de notebook desde una URL. El notebook que crea o selecciona debe ser un archivo.ipynb.

Después de crear un Notebook, estás listo para comenzar a escribir y ejecutar código para analizar datos. Antes de comenzar a codificar, deberá familiarizarse con la interfaz del notebook y cómo codificar en Markdown para escribir el código. Los notebooks se ejecutan en un Kernel de Jupyter en un clúster Spark.

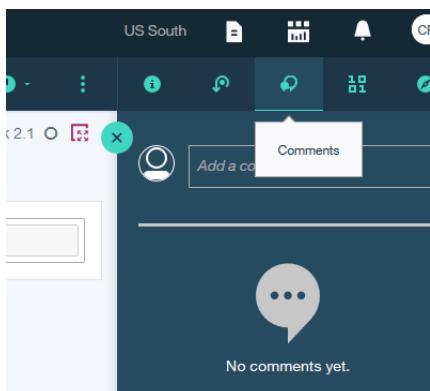
Para desarrollar aplicaciones analíticas en un Notebook, siga estos pasos generales:

- i. Importa bibliotecas preinstaladas para Python y R o instale sus propias bibliotecas.
- ii. Instale bibliotecas personalizadas o de terceros para cualquier idioma. Para Scala, no hay bibliotecas preinstaladas en el servicio Spark. Se almacenan en caché cuando los descarga y solo están disponibles durante el tiempo que se ejecuta el notebook.

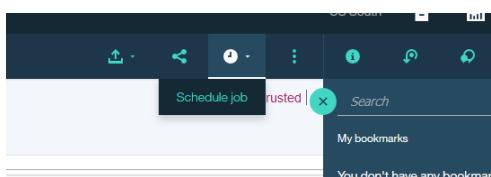
- iii. Cargar y acceder a los datos. Vemos que podemos añadir a nuestro notebook un dataset o un fichero de datos, además de conexiones de datos.

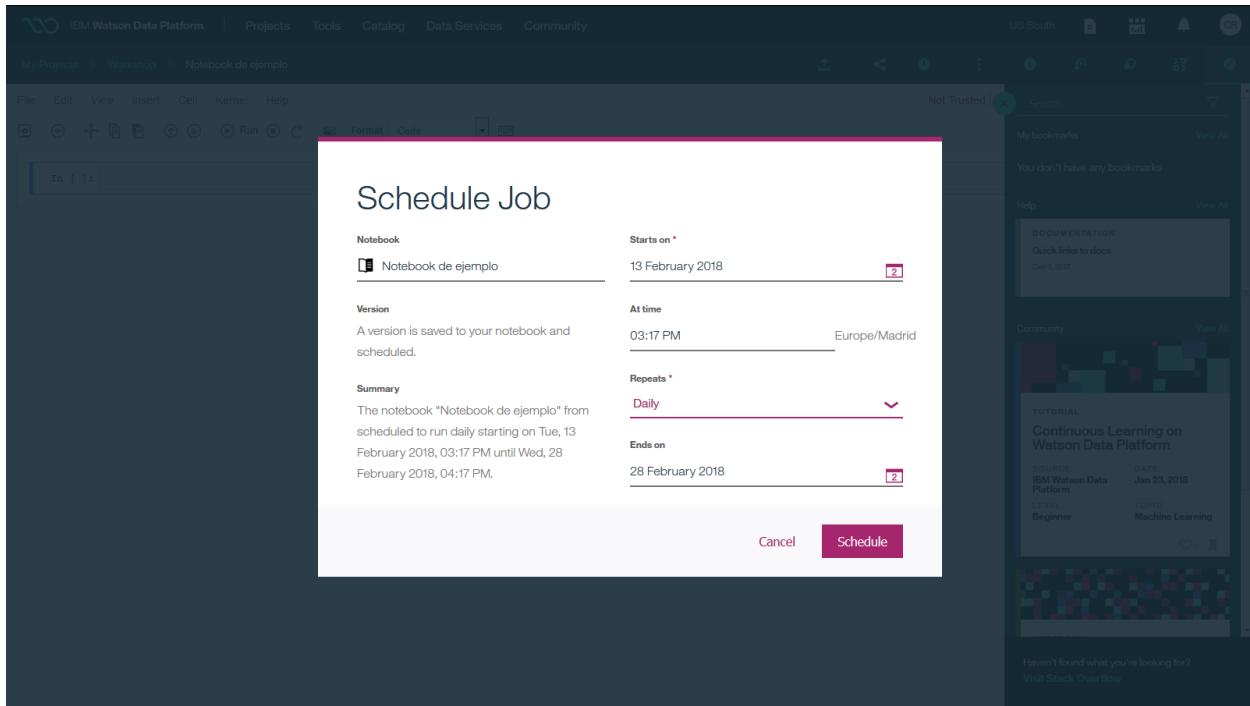


Colabora con otros miembros del proyecto. Puede agregar comentarios a los cuadernos haciendo clic en el ícono de comentario (ícono **Comentario**).

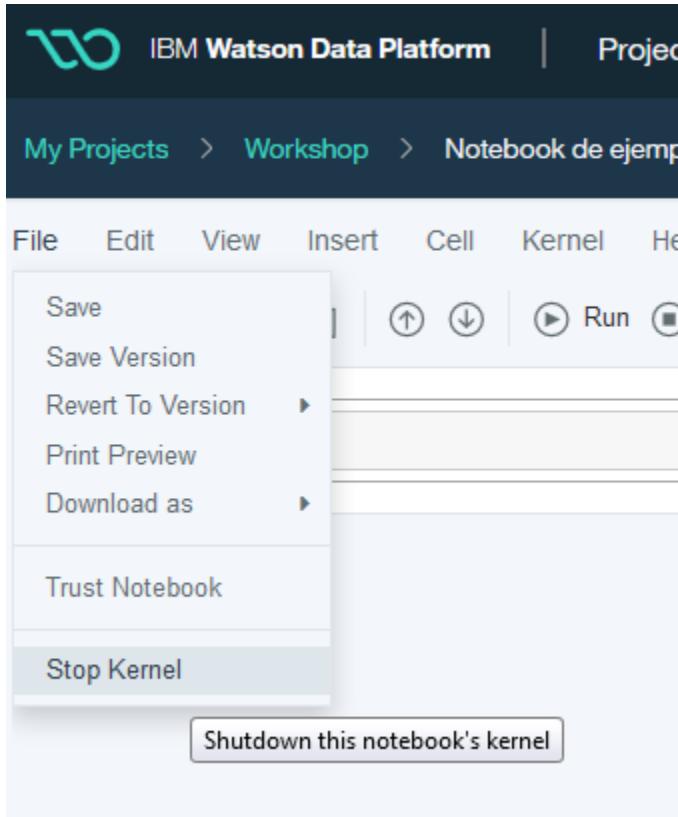


Si es necesario, programa el notebook para que se ejecute en otro momento.

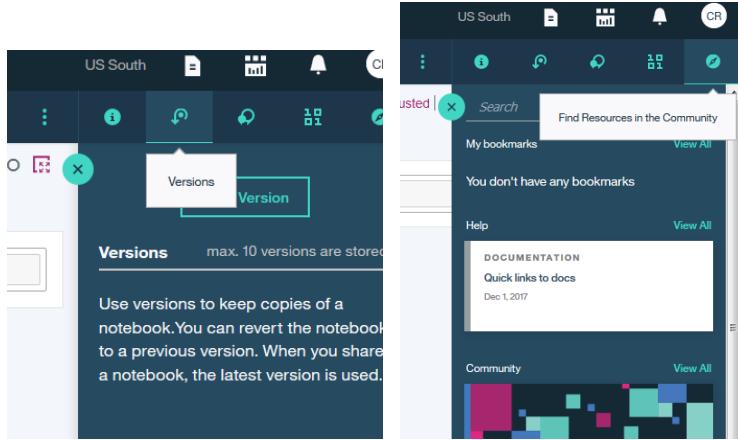




Cuando no estés trabajando activamente en el notebook, haz clic en **File> Stop kernel** para detener el kernel del notebook.

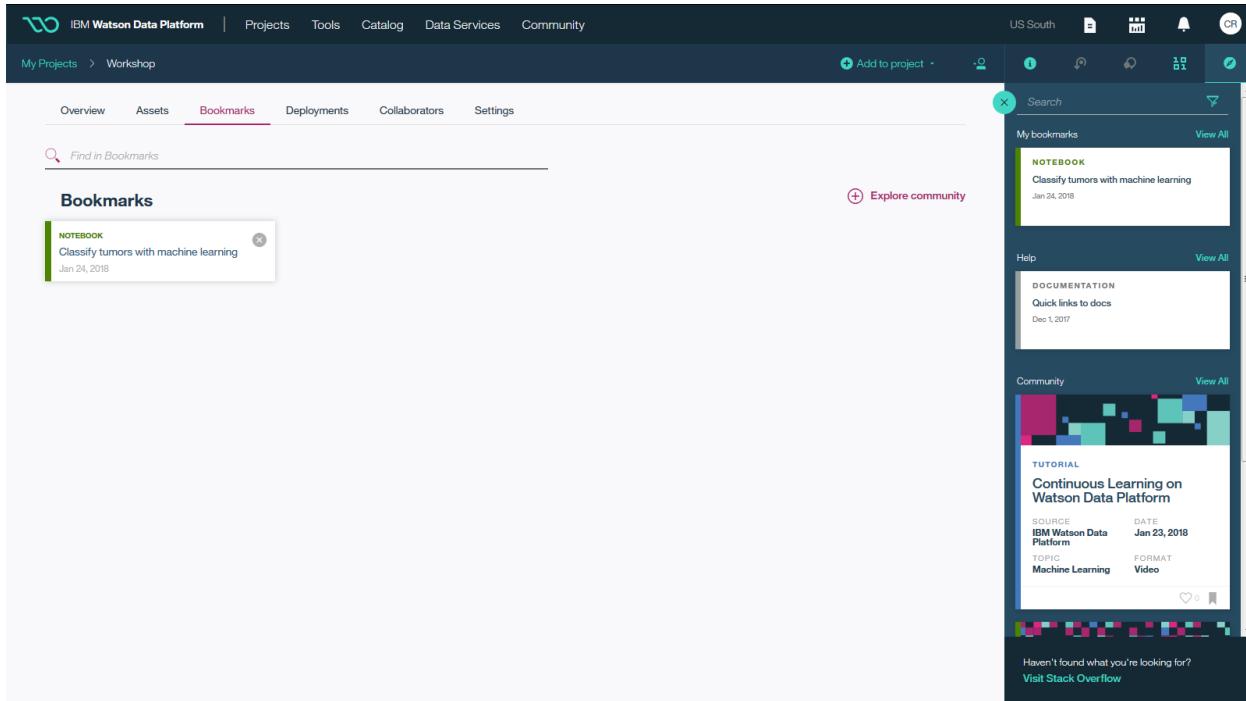


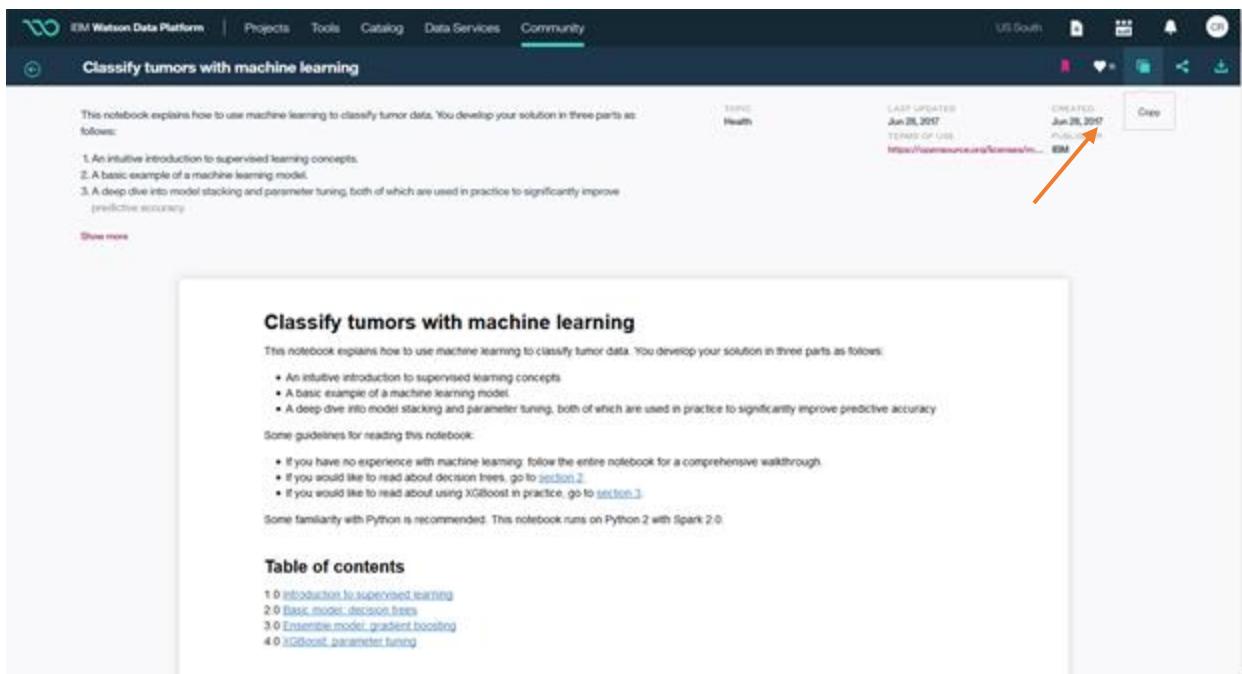
Además, podemos buscar recursos en la comunidad, y aprovechar esos recursos para enriquecer o comenzar un proyecto.



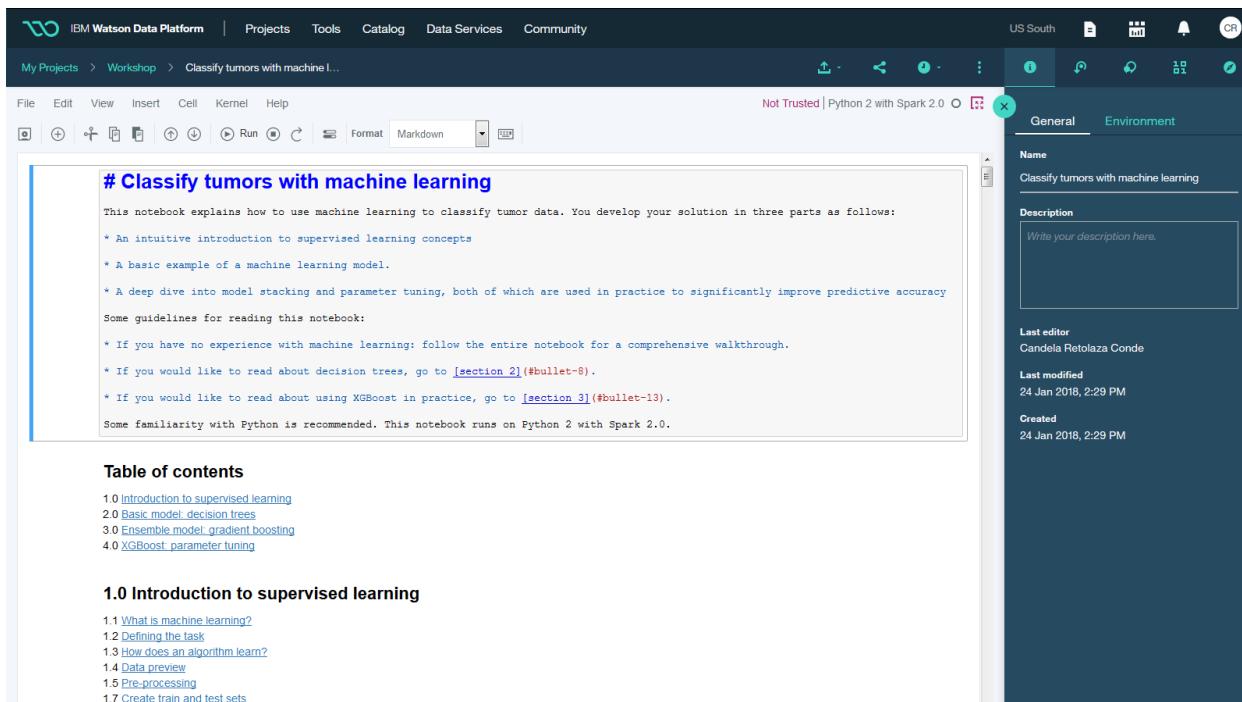
Para ello, debemos buscar en la comunidad algún notebook que nos pueda interesar, guardarlo en el proyecto, y copiarlo para utilizarlo o reutilizar ciertas partes de él.

A continuación, vamos a probarlo. Buscamos un notebook que nos interese: por ejemplo, buscamos por machine learning y escogemos uno que nos resulte interesante. Guardamos, y podemos copiarlo a nuestro proyecto y lo tendremos listo para utilizar.





This screenshot shows the IBM Watson Data Platform interface. At the top, there are navigation links: Projects, Tools, Catalog, Data Services, and Community. On the right side, there are icons for Health, US South, and a bell. Below the navigation, the title 'Classify tumors with machine learning' is displayed. The notebook content includes a brief introduction, a table of contents, and some guidelines. In the top right corner of the notebook area, there is a 'Copilot' button. An orange arrow points to this button, indicating its function.



This screenshot shows the same notebook in a different view within the IBM Watson Data Platform. The interface includes a top navigation bar with Project, Tools, Catalog, Data Services, and Community. Below that is a toolbar with File, Edit, View, Insert, Cell, Kernel, and Help. The notebook content is visible on the left, and on the right, there is a sidebar with tabs for 'General' and 'Environment'. The 'General' tab displays the notebook's name, a description field with a placeholder 'Write your description here.', and sections for 'Last editor' (Candela Retolaza Conde) and 'Last modified' (24 Jan 2018, 2:29 PM). The 'Environment' tab is also present.

Podemos copiar celdas o trozos para reutilizar, etc. Además, podemos compartir o notebooks para que las personas que no tienen cuentas DSX puedan verlos.

Si deseas enseñar a otras personas su notebook pero no quieres que puedan ejecutarlo, puedes darles una URL con una vista de solo lectura.

Si deseas publicar tu notebook para que otras personas puedan copiarlo y ejecutarlo, puedes publicarlo en Github o como gist.

¡asegúrate de ocultar cualquier código, como credenciales, que no quieras que otros vean!

En el apartado siguiente veremos cómo visualizar los resultados.

2. Visualizaciones

Usa visualizaciones en sus notebooks para presentar datos visualmente para ayudar a identificar patrones, obtener información y tomar decisiones.

Muchas de las bibliotecas de visualización de código abierto más comunes, como **matplotlib**, están preinstaladas en DSX. Todo lo que tienes que hacer es importarlos.

Para ver la lista de bibliotecas instaladas, ejecuta el comando apropiado desde una celda de notebook:

Python: !pip list --isolated

R: installed.packages()

Para importar una biblioteca instalada en tu notebook, ejecute el comando apropiado desde una celda de tu notebook con el nombre de la biblioteca:

Python: import library_name

R: library(library_name)

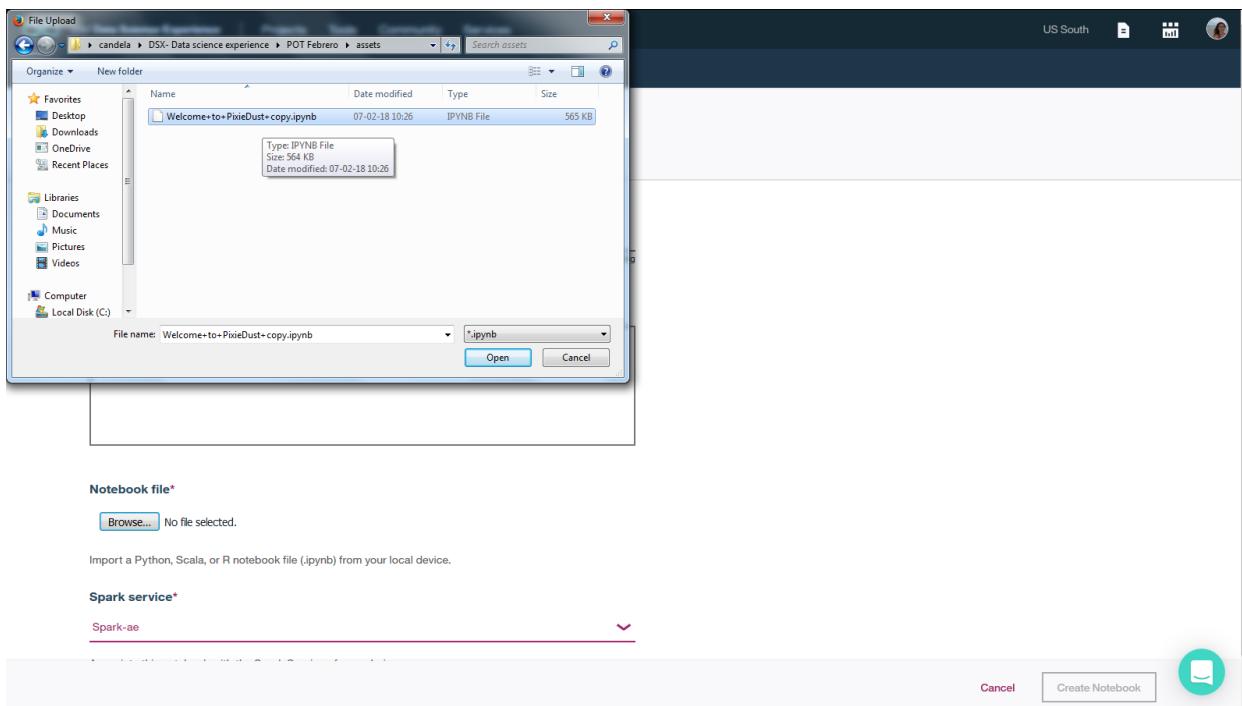
Puede instalar fácilmente otras bibliotecas y paquetes de visualización. Consulte en la ayuda de DSX: *Install custom or third-party libraries and packages*.

Además, puedes usar estas bibliotecas y herramientas de visualización de IBM:

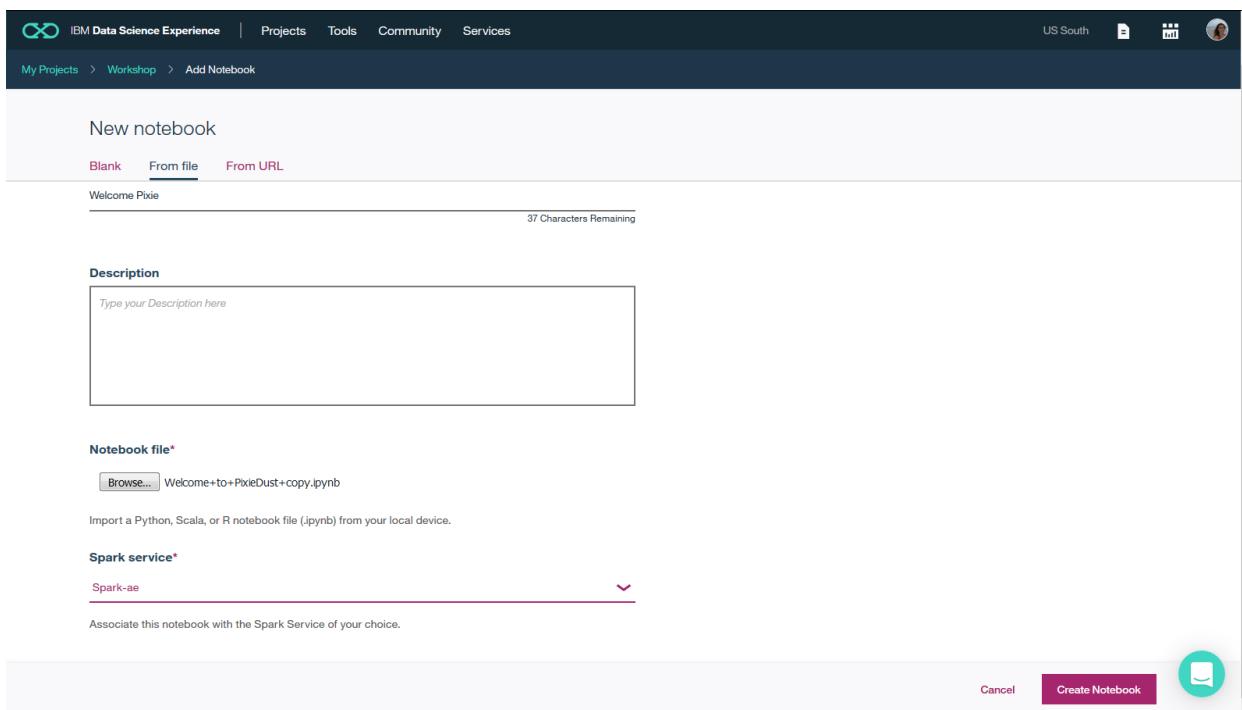
- PixieDust: cree gráficos con un comando de una sola palabra y luego explore con una interfaz de usuario integrada en lugar de código. Ejecute el código de Scala dentro de los cuadernos de Python.
- Brunel: crea gráficos interactivos con código simple. Prueba en un cuaderno.
- Modelos SPSS: cree tablas y gráficos interactivos para ayudarlo a evaluar y mejorar un modelo de análisis predictivo creado con algoritmos de aprendizaje automático SPSS.

Puedes usar las siguientes bibliotecas de visualización en Notebooks de Scala: PixieDust, Brunel for Scala y Lightning for Scala.

Vamos a subir un Notebook llamado 'Welcome Pixie Dust' para comenzar a desenvolvernos. Creamos un nuevo notebook, pero vamos a subirlo desde un fichero. Tenemos el siguiente fichero en nuestro escritorio: *Welcome+to+PixieDust+copy.ipynb* ponemos un nombre y lo subimos a DSX:



Solo podemos seleccionar un servicio de Spark y creamos el notebook:



Se propone al lector seguir los pasos del notebook para comenzar a familiarizarse con los notebooks y con las visualizaciones de PixelDust.

Otra forma de conseguir el mismo notebook:

Buscamos en la comunidad el notebook 'Welcome to PixieDust', lo abrimos, lo copiamos en un proyecto y lo editamos, para poder modificar o ejecutar celda a celda.

New notebook: Welcome to PixieDust

Project
Workshop

Add the notebook to an existing project.

Spark service*
Spark-ae

Associate this notebook with the IBM Analytics for Apache Spark Service of your choice.

Cancel Create Notebook

Una vez tenemos el notebook en nuestro proyecto, lo abrimos y le damos a editar. Podemos ejecutar celda a celda y seguir las instrucciones del notebook.

Welcome to PixieDust

This notebook features an introduction to [PixieDust](#), the Python library that makes data visualization easy.

This notebook runs on Python 2.7 and 3.5, with Spark 2.0.

Table of Contents

- [Get started](#)
- [Load text data from remote sources](#)
- [Mix Scala and Python on the same notebook](#)
- [Add Spark packages and run inside your notebook](#)
- [Stash your data](#)
- [Contribute](#)

Get started

This introduction is pretty straightforward, but it wouldn't hurt to load up the [PixieDust documentation](#) so it's handy.

New to notebooks? Don't worry. Here's all you need to know to run this introduction:

1. Make sure this notebook is in Edit mode
2. To run code cells, put your cursor in the cell and press **Shift + Enter**
3. The cell number will change to **[1]** to indicate that it is currently executing. (When starting with notebooks, it's best to run cells in order, one at a time.)

```
In [1]: # To confirm you have the latest version of PixieDust on your system, run this cell
!pip install --user --upgrade pixiedust
```

Requirement already up-to-date: pixiedust in /usr/local/src/bluemix_jupyter_bundle.v79/notebook/lib/extras

Collecting markdown (from pixiedust)

 Downloading Markdown-2.6.11-py2.py3-none-any.whl (78kB)

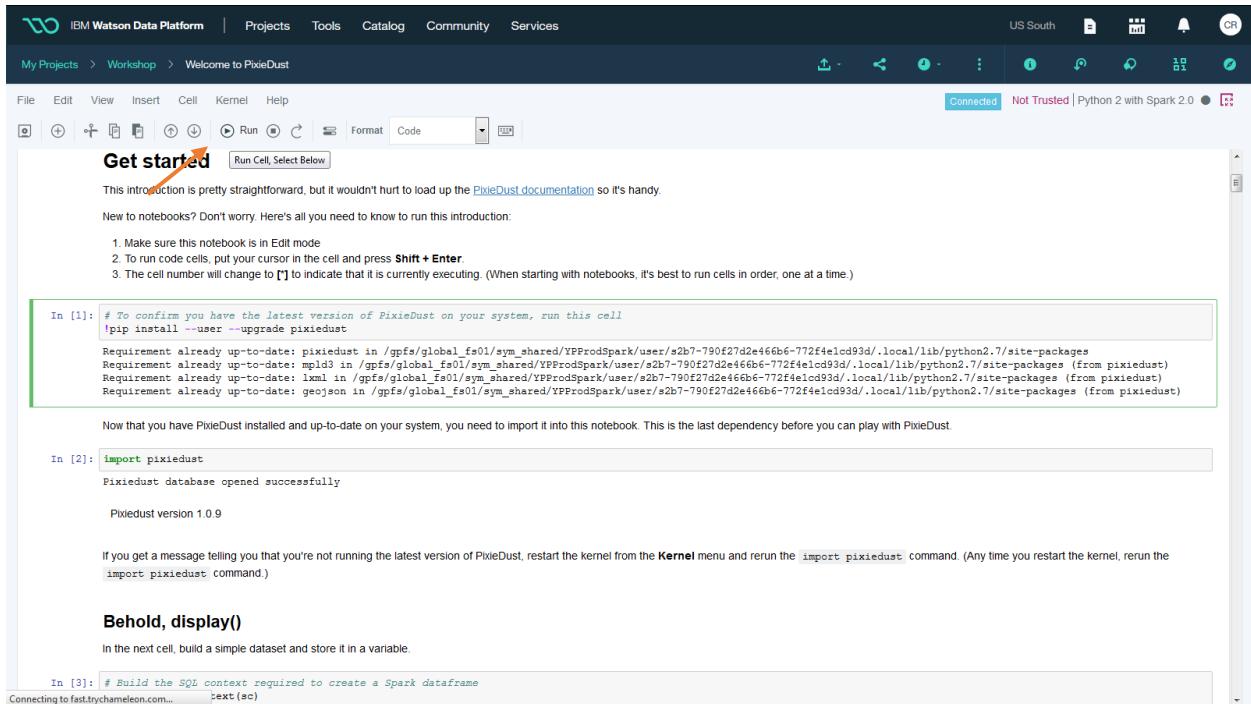
 100% |████████████████████████████████| 81kB 1.2MB/s eta 0:00:01

Collecting lxml (from pixiedust)

 Downloading lxml-4.1.1-cp27mu-manylinux1_x86_64.whl (5.6MB)

 100% |████████████████████████████████| 5.6MB 165kB/s eta 0:00:01

Requirement already up-to-date: astunparse in /usr/local/src/bluemix_jupyter_bundle.v79/notebook/lib/python2.7/site-packages (from pixiedust)



The screenshot shows the IBM Watson Data Platform Workshop interface. At the top, there are navigation links for Projects, Tools, Catalog, Community, and Services, along with a location indicator for US South and a user profile icon. The main area is a Jupyter notebook titled 'Welcome to PixieDust'. A 'Get started' button is highlighted with a red arrow. The notebook contains several code cells:

- In [1]:** A command to install the latest version of PixieDust: `# To confirm you have the latest version of PixieDust on your system, run this cell
!pip install --user --upgrade pixiedust`. The output shows requirements for pixeld3, lxml, and georjson.
- In [2]:** An import statement: `import pixiedust`. The output: 'Pixiedust database opened successfully' and 'Pixiedust version 1.0.9'.
- In [3]:** A command to build the SQL context: `# Build the SQL context required to create a Spark dataframe
text(sc)`. The output: 'Connecting to fast.trychameleon.com...'.

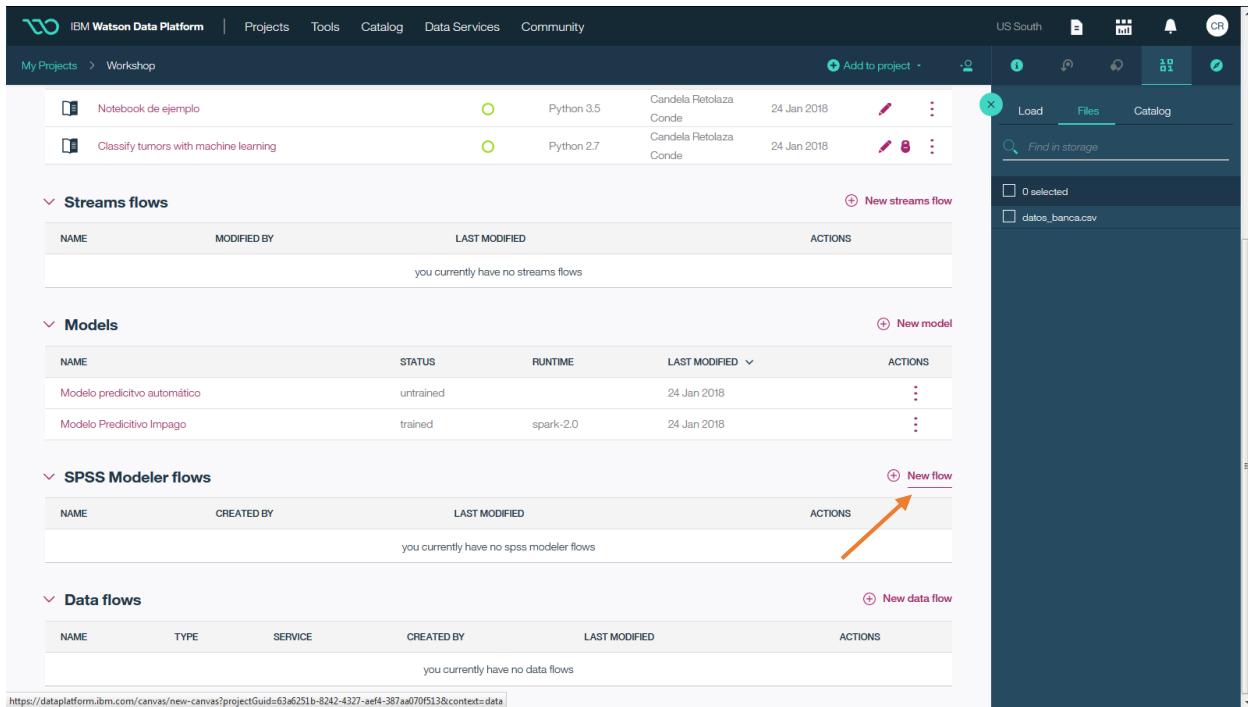
3. Algoritmos de analítica predictiva de SPSS

Una de las funcionalidades más interesantes que proporciona IBM como valor añadido a DSX es ésta.

SPSS Modeler es una herramienta muy estable y muy potente para realizar minería de datos. Es una herramienta que permite al equipo de científicos de datos realizar todo el proceso de minería de datos siguiendo CRISP-DM, es decir, en SPSS podemos acceder a los datos (ya estén en cualquier base de datos, o ficheros planos) podemos limpiar y modificar los datos, después tenemos más de 50 modelos (árboles de regresión, clústeres, redes neuronales, regresiones, etc.) para analizarlos y posteriormente podemos exportar esos datos, hacer gráficos, o ponerlo en producción.

DSX está adquiriendo cada vez más funcionalidades de SPSS Modeler para añadirla a sus funcionalidades propias. Gracias a esto, el usuario es capaz de hacer minería de datos de una manera más sencilla, y así enriquecer los proyectos.

Ahora vamos a hacer analítica con los flujos de SPSS.



The screenshot shows the IBM Watson Data Platform interface with the 'Workshop' tab selected. The main area displays two projects: 'Notebook de ejemplo' and 'Classify tumors with machine learning'. Below these, there are sections for 'Streams flows', 'Models', and 'SPSS Modeler flows'. The 'SPSS Modeler flows' section is the focus, showing a table with columns: NAME, CREATED BY, LAST MODIFIED, and ACTIONS. A red arrow points to the 'New flow' button in this section. On the right side, there is a sidebar with tabs for 'Load', 'Files', and 'Catalog', and a search bar for 'Find in storage'.

Podemos crear un flujo nuevo o si somos usuarios de SPSS Modeler, podemos importar modelos que ya tengamos hechos o empezar con un ejemplo. Vamos a comenzar con un ejemplo para familiarizarnos y posteriormente crearemos una ruta con los datos que hemos cargado antes.

IBM Watson Data Platform | Projects Tools Catalog Data Services Community US South CR

SPSS Modeler BETA

New From file From example

Select one of the samples below to get started with an existing stream that suits the kind of modeling you want to do. When you create the stream it will be added to your project, allowing you to modify and save your changes.

SPSS MODELER
Drug Study Example
Use neural network and C5.0 algorithms to build classification models that allow you to predict the correct type of drug for a patient based on various health metrics.

SPSS MODELER
Sales Promotion Study
Use neural network and C5.0 algorithms to predict the effect of advertising promotions on the sale of various items. Input data of sales before and after a post promotion are used to train the model to predict the effectiveness of advertising.

Cancel Creating...

My Projects > Workshop > Drug Study Example

DRUG1a → Drug

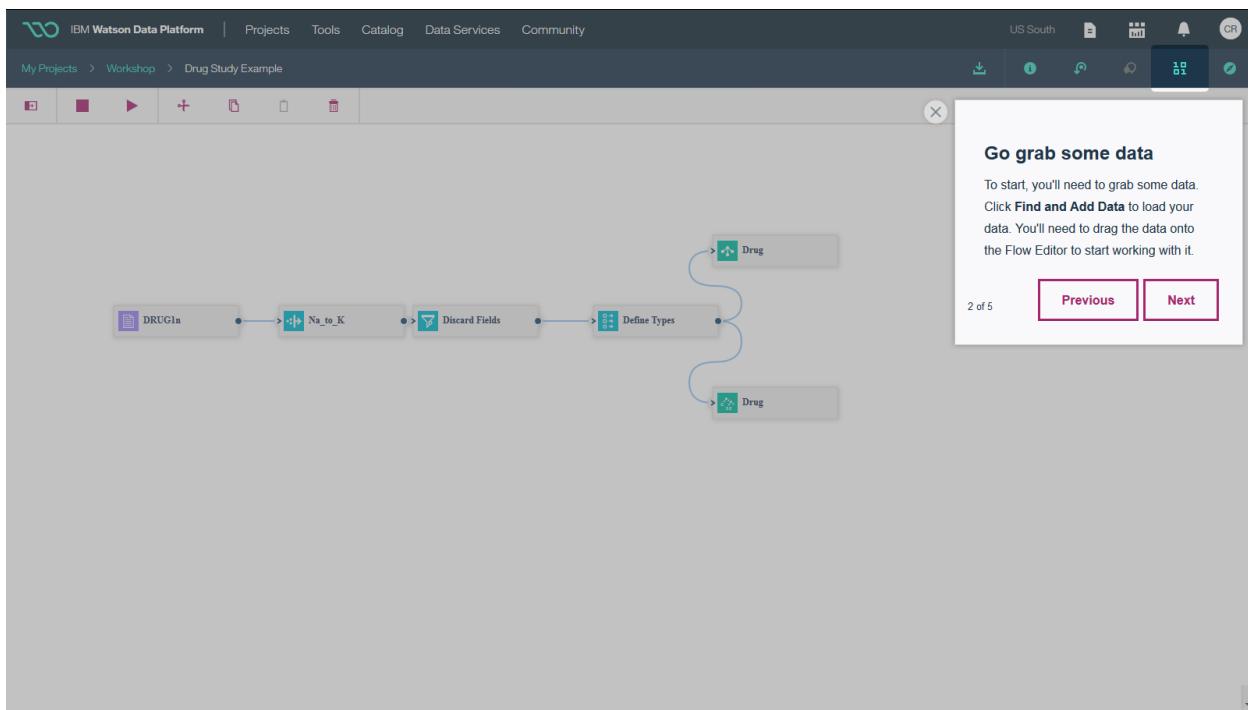
Na_to_K

Welcome to flows

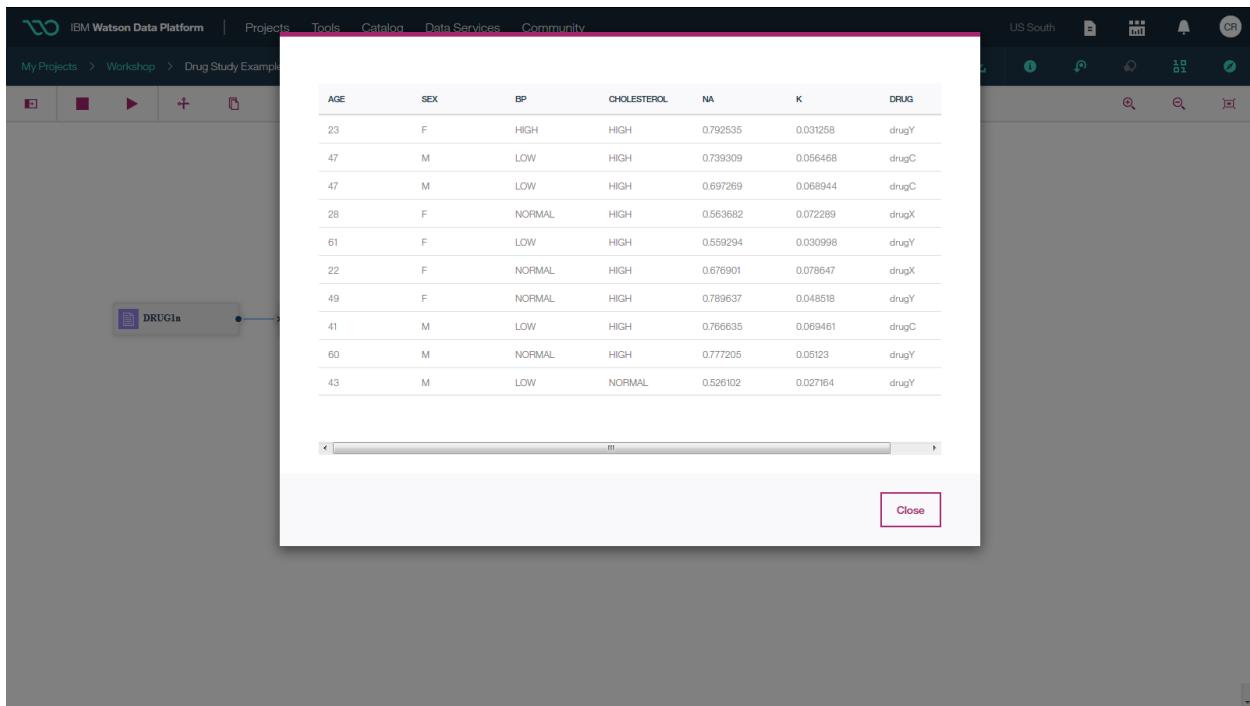
Use flows to transform data and create predictive models. This tour leads you through some of the opening steps to get you started.

Start Tour

1 of 5

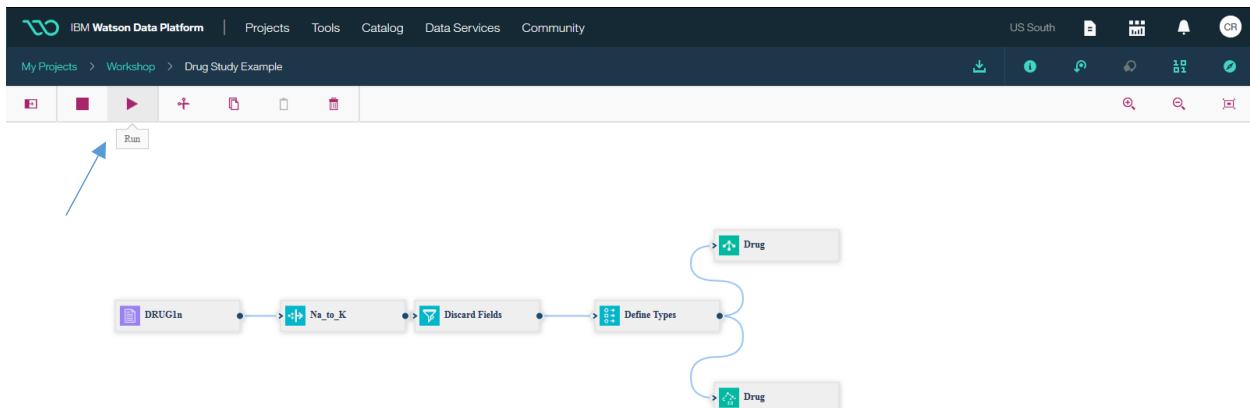


En este ejemplo, imagina que es un investigador médico que está recopilando datos para un estudio. Has recopilado información sobre un conjunto de pacientes, de los cuales todos sufrieron la misma enfermedad. Durante el curso del tratamiento, cada paciente respondió a un medicamento de un total de cinco. Parte de su trabajo consiste en utilizar minería de datos para averiguar qué medicamento es el adecuado para un futuro paciente con la misma enfermedad.



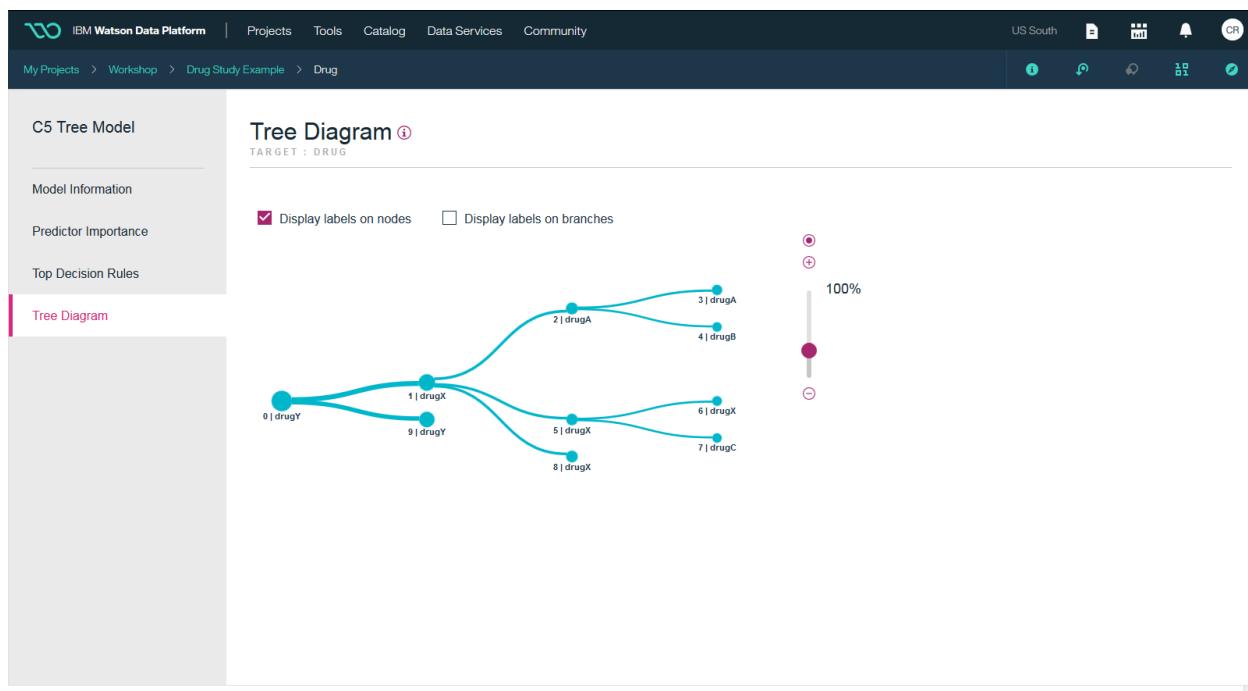
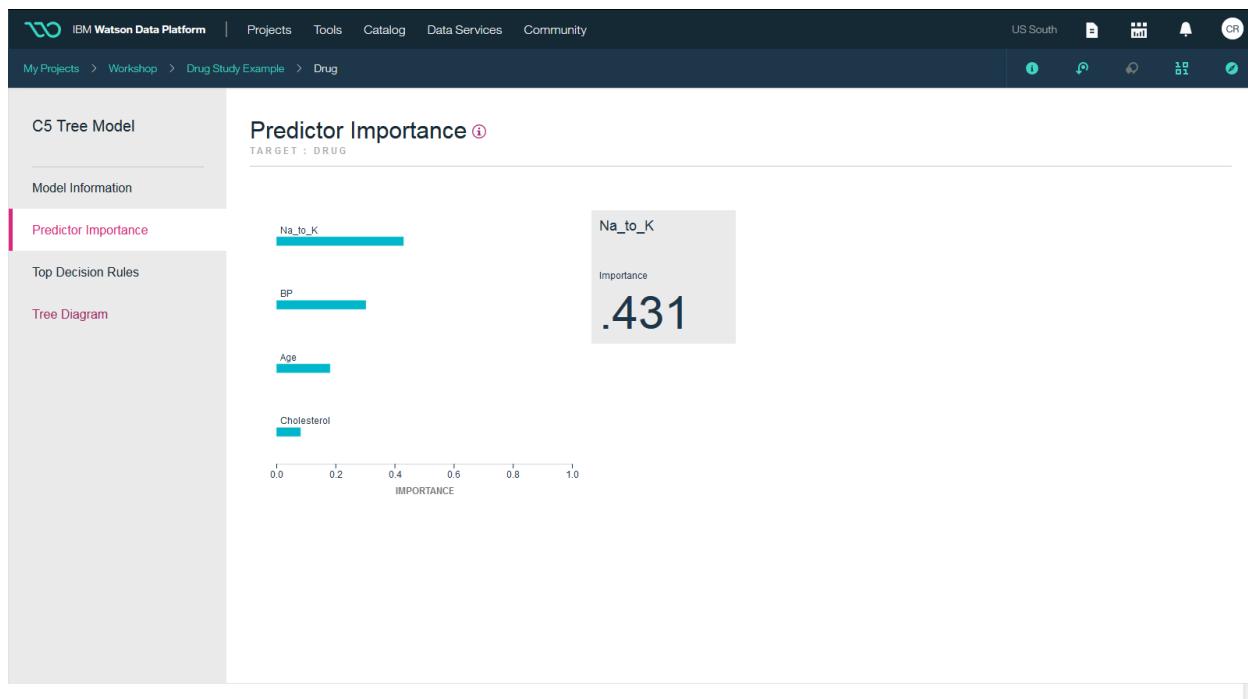
AGE	SEX	BP	CHOLESTEROL	Na	K	DRUG
23	F	HIGH	HIGH	0.792535	0.031258	drugY
47	M	LOW	HIGH	0.739309	0.056468	drugC
47	M	LOW	HIGH	0.697269	0.068944	drugC
28	F	NORMAL	HIGH	0.563682	0.072289	drugX
61	F	LOW	HIGH	0.559294	0.030998	drugY
22	F	NORMAL	HIGH	0.676901	0.078647	drugX
49	F	NORMAL	HIGH	0.789637	0.048518	drugY
41	M	LOW	HIGH	0.766635	0.069461	drugC
60	M	NORMAL	HIGH	0.777205	0.05123	drugY
43	M	LOW	NORMAL	0.526102	0.027164	drugY

Debido a que el sodio y el potasio tienen una alta correlación, se puede observar que es así haciendo un gráfico, se crea una nueva variable para que pueda ser utilizada en el modelo.

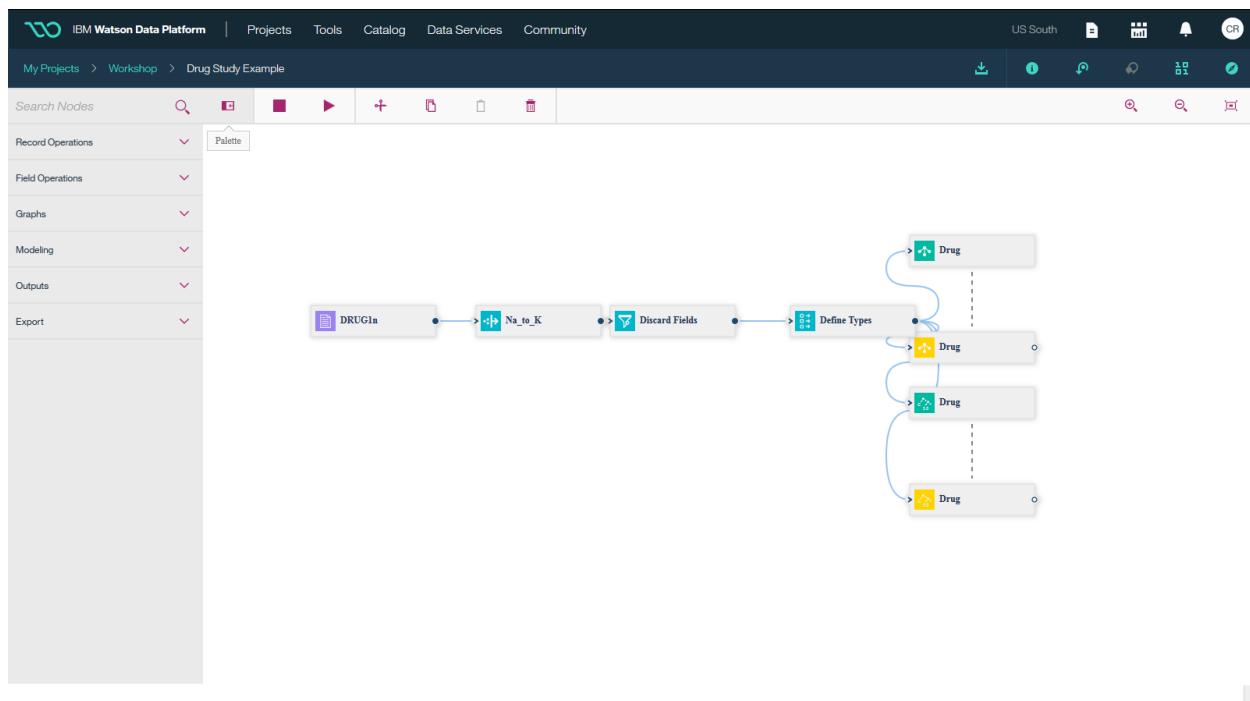


Para observar los resultados: Botón derecho: **view model**

Model Information	
Predictor Importance	Target Field: Drug
Top Decision Rules	Model Type: Multi-Class Decision Tree
Tree Diagram	Algorithm Name: C5
	Number of Features: 4
	Tree Depth: 4
	Number of Nodes: 10



Se sugiere al lector que entienda el modelo y que lo modifique si es necesario. Podemos copiar el modelo en nuestro proyecto.



A continuación, vamos a hacer uno nuevo.

New From file From example

Name*

Modelo de predicción de impago

Description

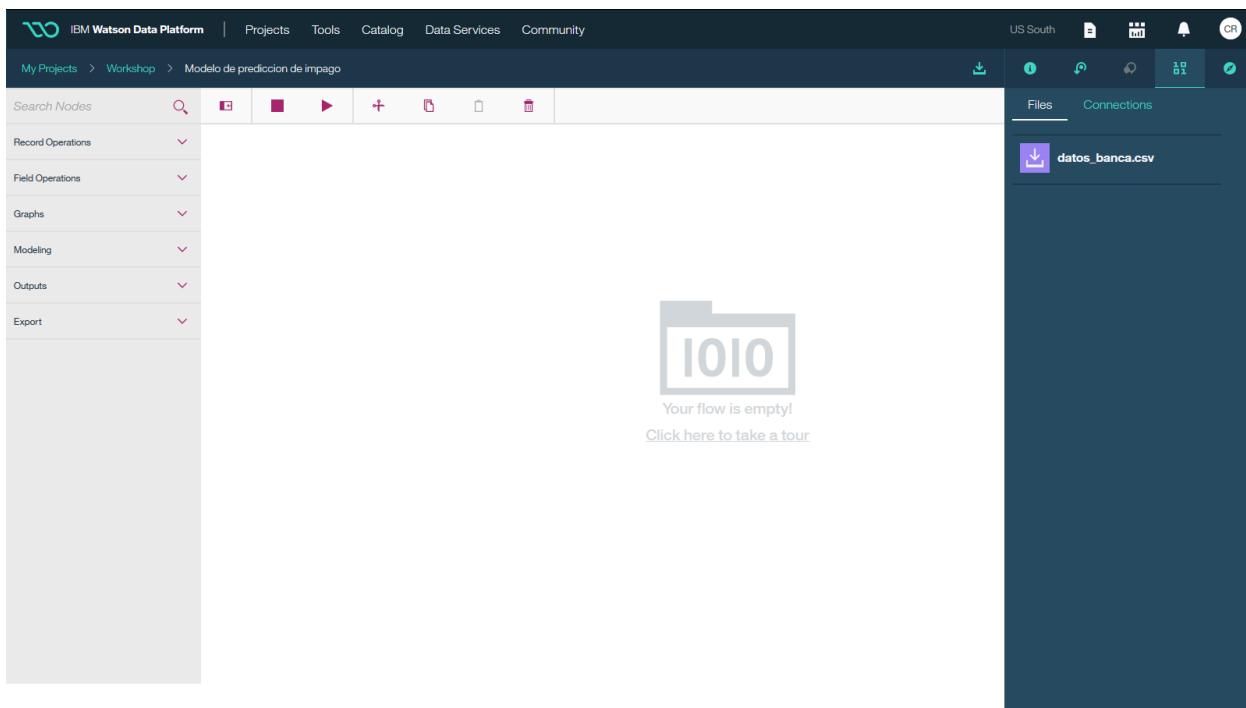
Type description here.

Runtime

IBM SPSS Modeler

Cancel Creating...

seleccionamos el runtime de spss modeler y creamos:



Como veis tenemos una paleta a la izquierda en la que tenemos los nodos, y un lienzo en blanco en el que hacer nuestra ruta o Flow. Se recomienda al lector que explore los nodos.

Por ejemplo, la pestaña de la paleta Operaciones con registro contiene nodos que puede utilizar para realizar operaciones en los registros de datos como, por ejemplo, seleccionar, fusionar y añadir.

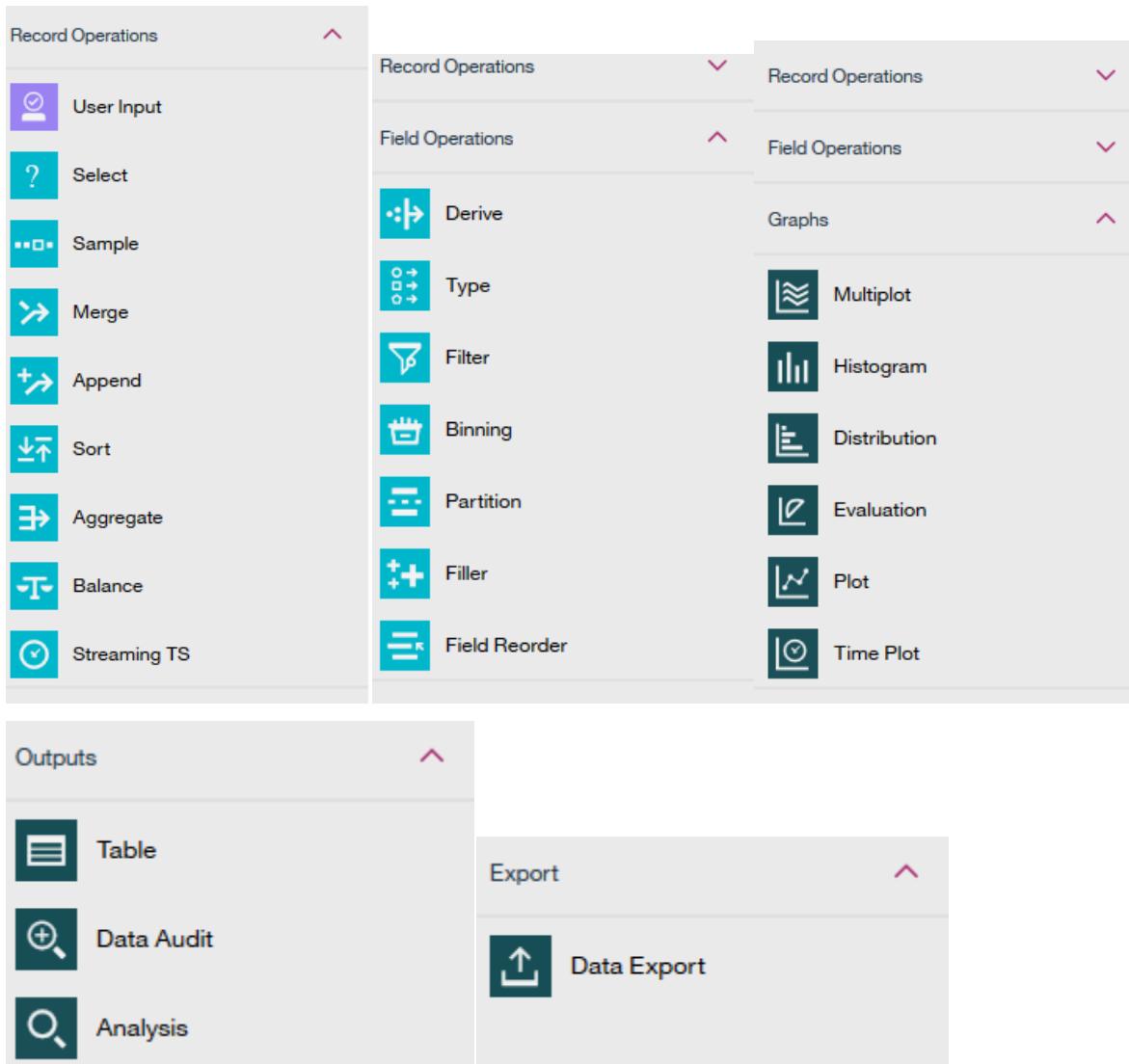
Los nodos Operaciones con campos realizan operaciones en campos de datos como, por ejemplo, filtrar, derivar campos nuevos y determinar el nivel de medición para campos dados.

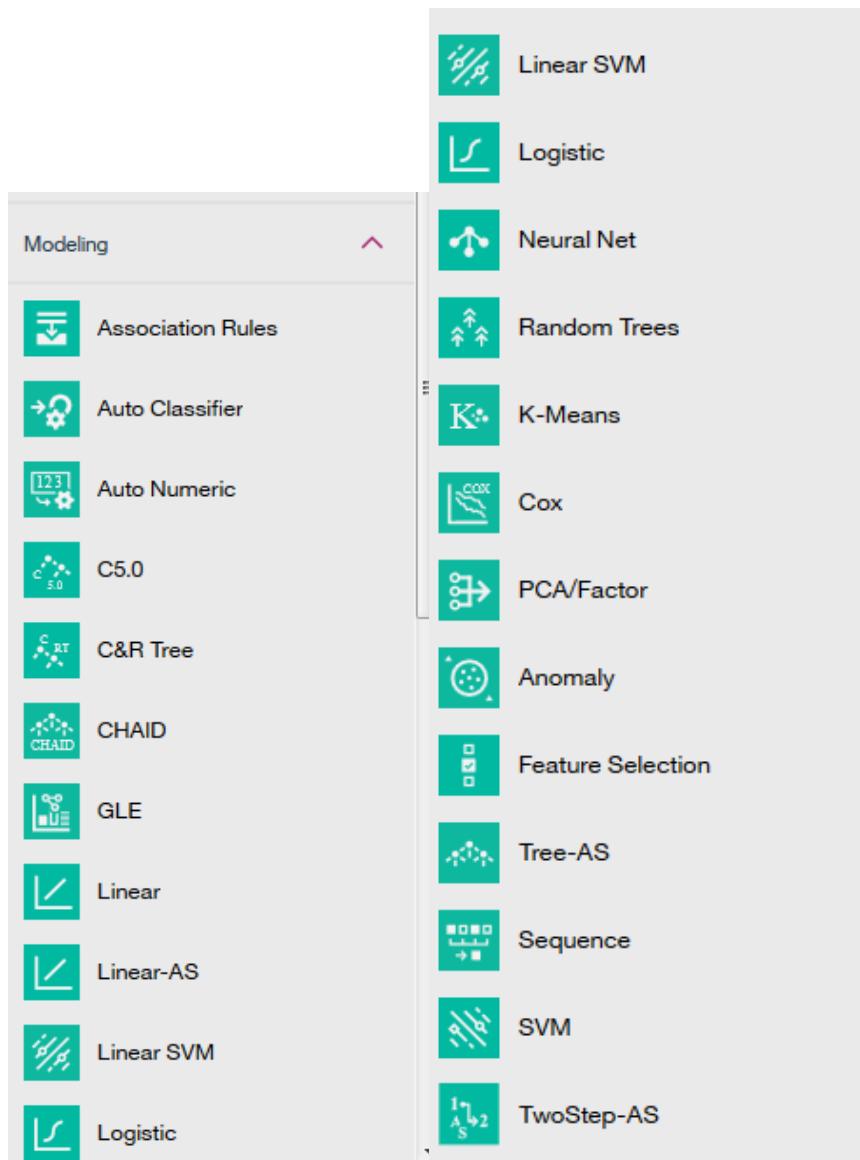
Los nodos Gráficos muestran gráficamente los datos antes y después del modelado. Entre ellos se incluyen gráficos, histogramas, nodos de malla y diagramas de evaluación. Los nodos Modelado utilizan los algoritmos de modelado disponibles en SPSS Modeler como, por ejemplo, redes neuronales, árboles de decisión, algoritmos de agrupación en clúster y secuenciación de datos.

Los nodos Salida generan diferentes salidas para resultados de datos, gráficos y modelos que se pueden visualizar en DSX.

En la documentación completa de DSX se encuentra una descripción completa de cada nodo.

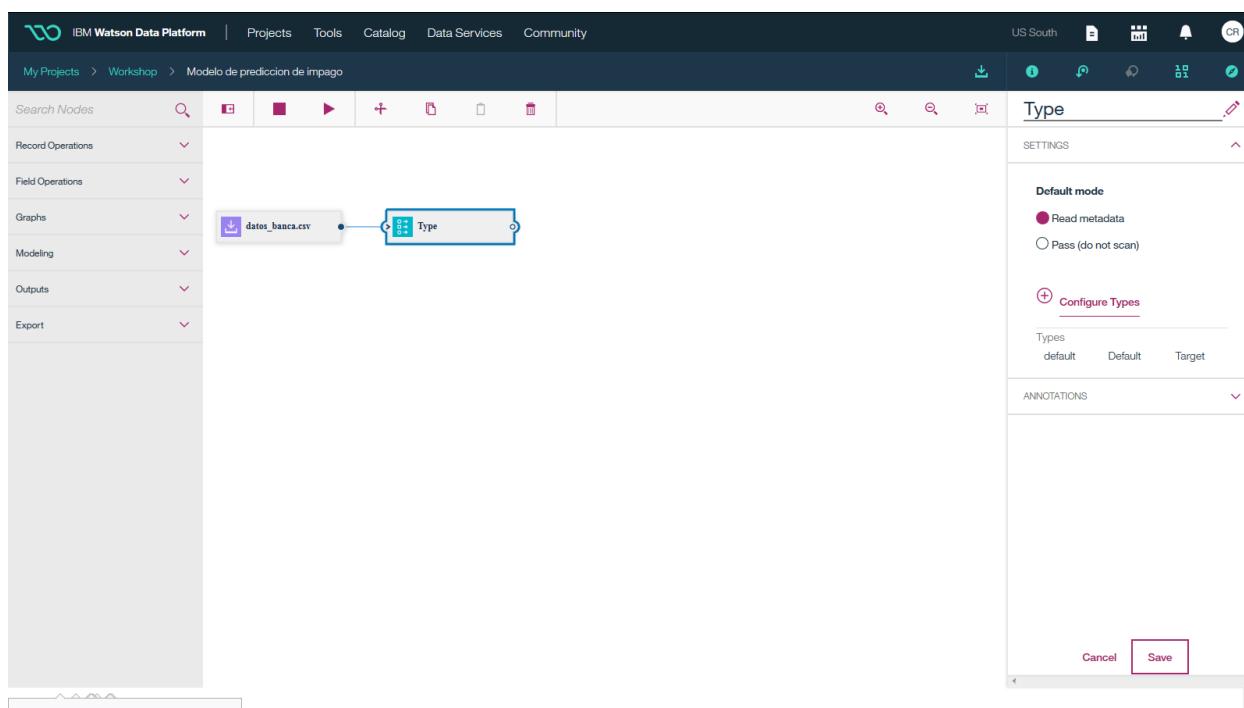
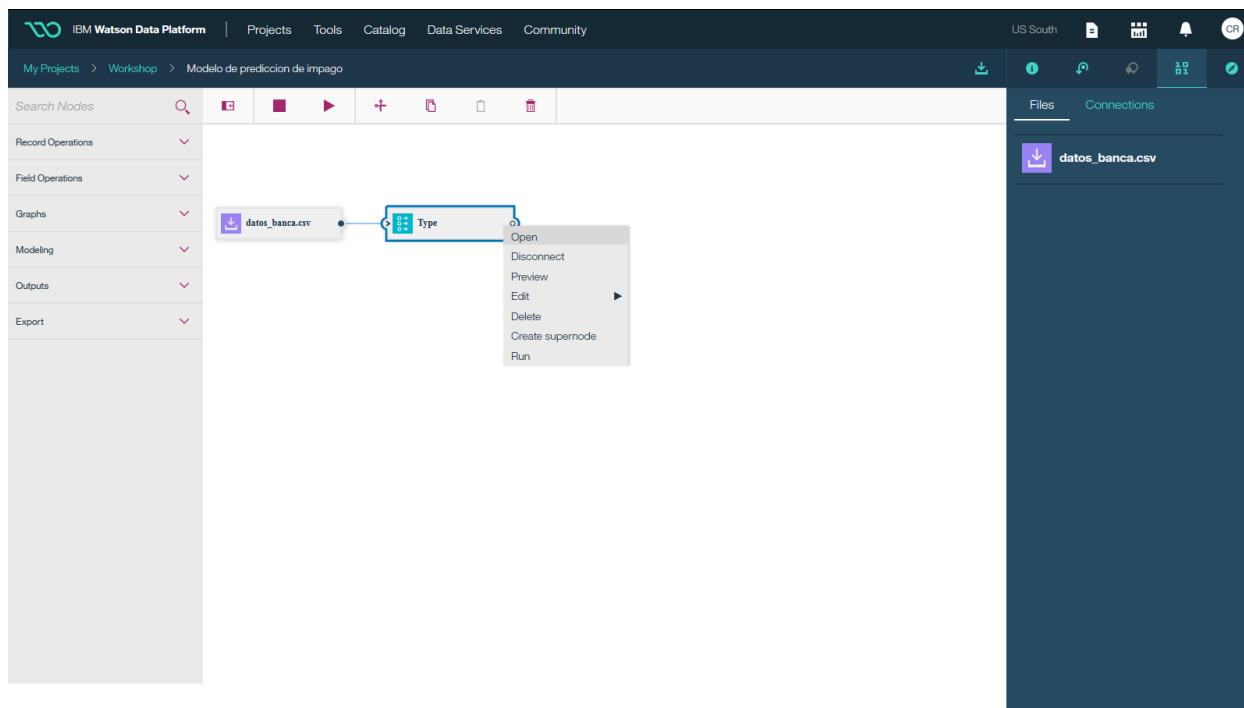
<https://dataplatform.ibm.com/docs/content/analyze-data/ml-canvas-spss.html?audience=wdp&context=analytics>

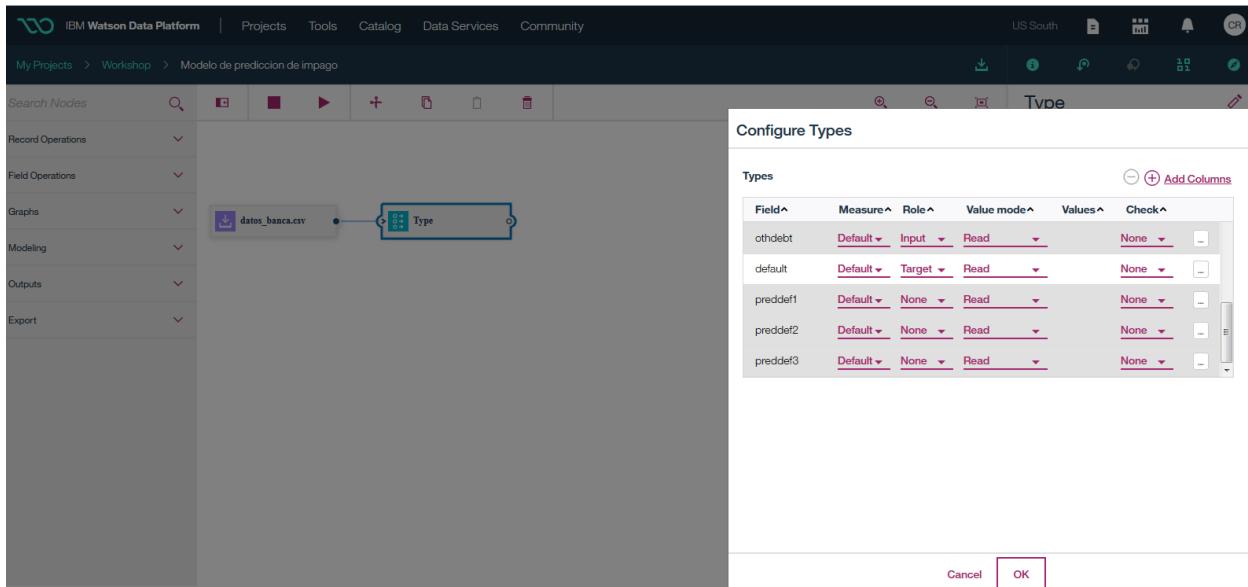
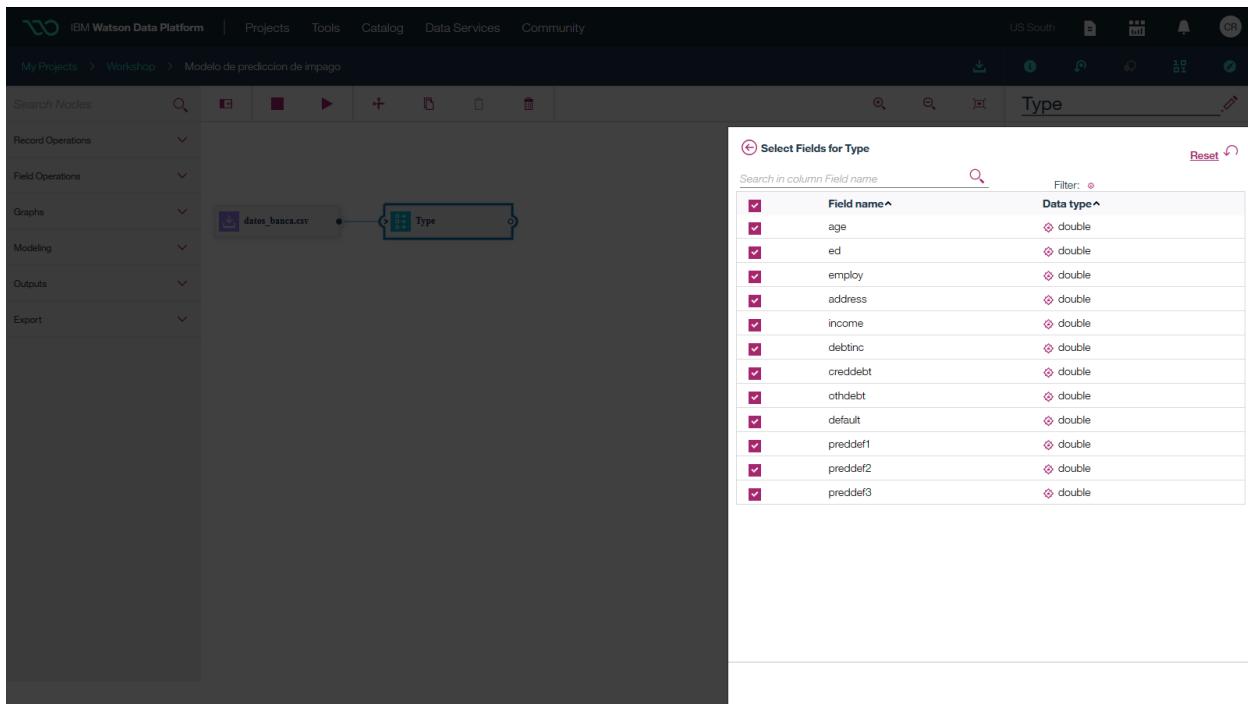




Para realizar el modelo, seleccionamos el fichero de entrada y vemos que incluso podemos añadir uno nuevo o conectarnos a una base de datos que tengamos. Arrastra el fichero al lienzo.

Y añade un nodo tipo, para instanciar los valores, y para seleccionar Default como **target**.





Como es un fichero preparado no se necesita hacer una limpieza de los datos, pero todo esto es posible hacerlo también aquí, con los nodos de operaciones con los datos, si no lo hemos hecho en un paso previo con la parte de **Data Refinery**. Además, podemos añadir varias fuentes de entrada y cruzar las tablas, elegir lo que queramos, etc.

Se invita al lector que acabe el modelo y consiga predecir las variables que generan impago.

Se deja al lector otro fichero, llamado baskets.csv para practicar ya sea con Modelos automáticos, con Notebooks, o con flows de SPSS. El objetivo será el siguiente:

Este ejemplo está relacionado con datos ficticios que describen el contenido de cestas de supermercado (es decir, una colección de artículos comprados a la vez) junto con los datos personales del comprador, que pueden obtenerse a través de las tarjetas de fidelidad. El objetivo es descubrir grupos de clientes que compren productos parecidos calificables desde el punto de vista demográfico, como por edad, ingresos, etc.

Este ejemplo muestra dos fases de la minería de datos:

- Modelado de reglas de asociación y una visualización de malla que muestra enlaces entre los artículos comprados.
- Perfilado de reglas de inducción C5.0 de los compradores de grupos identificados de productos.

Nota: Esta aplicación no utiliza directamente el modelado predictivo y, por tanto, no hay ninguna medición de la precisión de los modelos resultantes ni entrenamiento asociado/distinción de comprobaciones en el proceso de minería de datos.

4. RStudio

R es un popular paquete de análisis estadístico y aprendizaje automático que permite la gestión de datos e incluye pruebas, modelos, análisis y gráficos, y permite la gestión de datos. RStudio, incluido en IBM Data Science Experience, proporciona un IDE para trabajar con R.

Una sesión de RStudio creada en Data Science Experience incluye 2 GB de almacenamiento y 5 GB de memoria disponible para su uso.

También podemos hacer que nuestro análisis en R sea accesible para los no programadores a través de Shiny. Shiny es una excelente opción para implementar el análisis de minería de datos a los usuarios de negocio.

Shiny es un marco de aplicaciones web para R que permite convertir el análisis de R en aplicaciones web interactivas. No se requieren conocimientos de HTML, CSS o JavaScript.

The top screenshot shows the IBM Watson Data Platform interface. The top navigation bar includes 'Projects', 'Tools' (which is the active tab), 'Catalog', 'Data Services', and 'Community'. Below the navigation is a 'Recently updated' section with three items: 'Workshop', 'SPSS Modeler', and 'Data Refinery'. To the right is a table for 'Collaborators' with columns for 'ROLE', 'COLLABORATORS', 'DATE CREATED', and 'LAST UPDATED'. A 'New project' button is in the top right. The 'Get started' section contains a button to 'Create Catalog'. The bottom section, 'New in the community', displays four cards: 'Predicting Flight Cancellations Using...', 'Customer demographics and sales', 'Learn basics about notebooks and Apache Spark', and 'Continuous Learning on Watson Data Platform'.

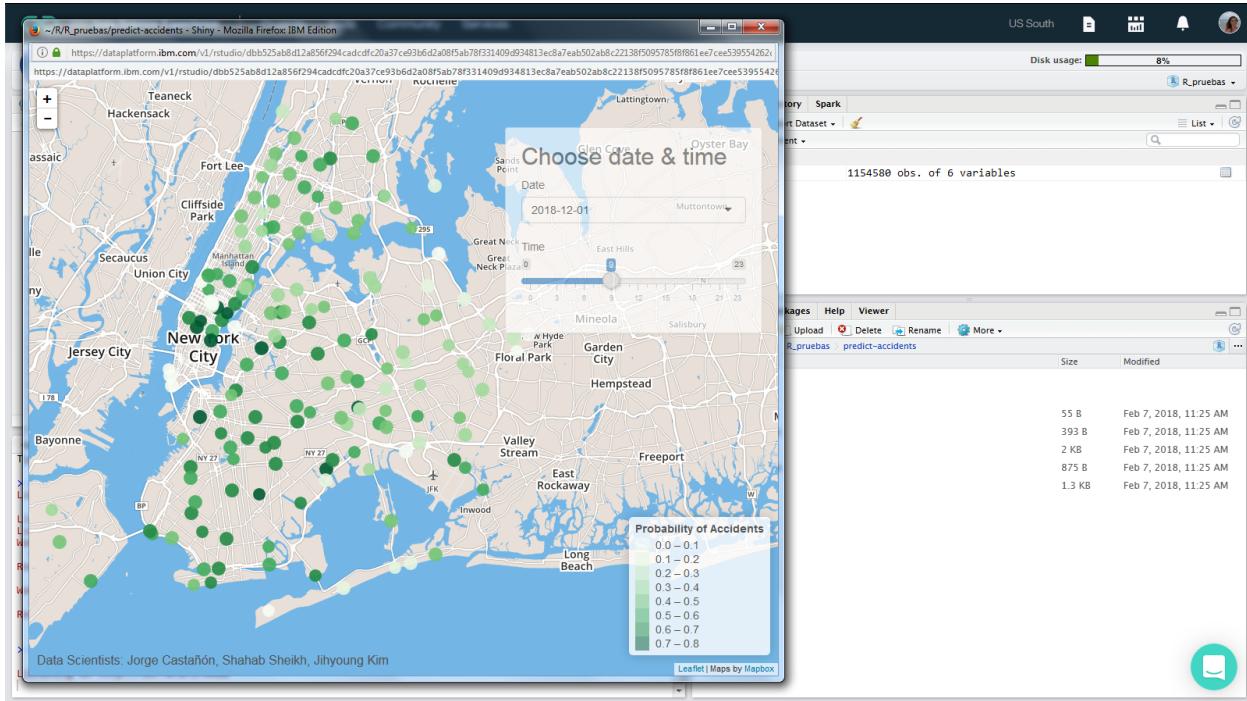
The bottom screenshot shows the RStudio interface within the IBM Watson Data Platform. The top navigation bar is identical. The main area has two panes: 'Console' on the left and 'Files' on the right. The 'Console' pane shows the R startup message: 'R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch" Copyright (C) 2016 The R Foundation for Statistical Computing Platform: x86_64-redhat-linux-gnu (64-bit)'. The 'Files' pane shows a directory structure with files like 'Rprofile', 'config.yml', 'item-sparka-demos', 'lost+found', and 'R'.

Proponemos al lector seguir los pasos de los siguientes proyectos (Predictions of car accident in NYC based on weather data, analyzing flight delays, y Blocpower) que se encuentran en el siguiente repositorio:

<https://github.com/IBMDatascience/dsx-shiny-apps>

Cuando accedas al repositorio de Github, primero lee el **Readme.md** y sigue los pasos.

El primero de los proyectos te muestra con la aplicación de Shiny los resultados de un modelo predictivo de accidentes de coche en New York. Es un modelo entrenado con datos históricos de accidentes e información meteorológica.



5. Librerías de Deep Learning

Deep learning es una rama de machine learning que utiliza grandes cantidades de datos para enseñar a las máquinas cómo hacer tareas o cosas que antes sólo eran capaces de hacer los seres humanos.

Buenos ejemplos de Deep learning son la percepción, el reconocimiento de lo que hay en una imagen, lo que las personas dicen cuando hablan, o ayudar a los robots a explorar el mundo e interactuar con él. El Deep learning está emergiendo como una herramienta central para resolver problemas de percepción en los últimos años. Son los modelos que están detrás de la visión artificial y el reconocimiento de voz. Cada vez más personas descubren que el Deep learning es una herramienta muy potente para resolver múltiples problemas.

Muchas empresas de hoy en día han convertido el Deep learning en una parte central de su conjunto de herramientas de aprendizaje automático. Por ejemplo, Facebook, Google y Uber están utilizando el Deep Learning en sus productos. En IBM estamos colaborando con los líderes en el mercado para impulsar la investigación y liderar en ese espacio.

Para comenzar con Deep Learning en Python con Data Science Experience:

Existe una comunidad cada vez mayor de investigadores, ingenieros y científicos de datos que comparten un conjunto común y muy potente de herramientas, y la mayoría de ellas son de código abierto.

Una de las cosas buenas del Deep learning es que es realmente una familia de técnicas que se adapta a todo tipo de datos y todo tipo de problemas, todos utilizan una infraestructura común y un lenguaje común para describir items.

Lo que se aconseja al lector es comenzar con modelos muy simples y posteriormente comenzar con los que son más complejos y grandes. Es sencillo comenzar con tu propio ordenador ya que con IBM Data Science Experience tienes todo lo que necesitas para comenzar a experimentar con las tecnologías de Deep Learning.

Las bibliotecas y tutoriales más populares de Deep Learning en Python son:

Theano: una de las bibliotecas de Deep Learning más conocidas.

→ TUTORIAL: https://dataplatform.ibm.com/analytics/notebooks/b4f6f269-6cd6-4adc-b63d-d19e5b0e90a0/view?access_token=647ed3ebaf725ffd9d4cf77fbc41066e093e15f764d5c810620a43044e362780

Tensorflow: es una biblioteca de bajo nivel que está menos madura que Theano. Sin embargo, es compatible con Google y ofrece computación distribuida lista para usar.

→Tutorial: https://dataplatform.ibm.com/analytics/notebooks/91440c8b-0fb-471e-b04e-235e4d9f510d/view?access_token=fb4380415a903111e26cec3bd95d8ba91a04746185c866fecde9d36643fa5585

Keras: Esta es nuestra biblioteca favorita de Python para Deep Learning y es el mejor lugar para comenzar si eres principiante.

→ Tutorial https://dataplatform.ibm.com/analytics/notebooks/d96fa67b-14f1-4db7-8b60-1af3c13699c3/view?access_token=c31fd96333af39811a78fe7773e421a50c7e20a450badb653bf4e0db39dc8f3f

Lasagne:

→ Tutorial

https://dataplatform.ibm.com/analytics/notebooks/c1bda39b-3fcd-4dae-a109-e71d11113633/view?access_token=18379e532a9953d4e97f2a75eee37a8ece9ee4745676e1a647493fbfd7b16fb

MXNet- Es otra biblioteca de alto nivel similar a Keras. Ofrece enlaces para múltiples idiomas y soporte para computación distribuida.

→ Tutorial https://dataplatform.ibm.com/analytics/notebooks/39e93a50-cfc1-4097-b671-5261ba56e166/view?access_token=b7bd65f58805daf1f39465395dbb239c2f03d2cdeb611d8f413c81c7b1b06791

Más información sobre Deep Learning en IBM Data Science Experience:

<https://medium.com/ibm-data-science-experience/deep-learning-with-data-science-experience-8478cc0f81ac>

Proponemos comenzar con el siguiente notebook, crea un proyecto o cópiale el notebook en algún proyecto que ya tengas creado, y sigue los pasos:

https://dataplatform.ibm.com/analytics/notebooks/d96fa67b-14f1-4db7-8b60-1af3c13699c3/view?access_token=c31fd96333af39811a78fe7773e421a50c7e20a450badb653bf4e0db39dc8f3f

Una vez el lector tenga soltura con Data Science Experience y comprenda qué es el Deep learning, se recomienda este interesante proyecto: **Self Driving Car tutorials with Data Science Experience.** Que se encuentra en el repositorio siguiente. Proponemos leer con atención el archivo README.md y seguir los pasos que se indican en él.

<https://github.com/aruizga7/Self-Driving-Car-in-DSX>

Workshop 5.

Recursos para trabajar en Data Science Experience en Local

Para todos aquellos que quieran aprender a utilizar Data Science Experience en Local:

https://github.com/elenalowery/DSX_Local_Workshop

ANEXO: Decision Optimization en DSX Local

Se usa con frecuencia el término optimización para referirse a hacer algo mejor. Aunque la optimización a menudo mejora las cosas, significa mucho más que eso: la optimización significa encontrar la solución más adecuada para una situación definida con precisión. Esta sofisticada tecnología, también llamada Analítica Prescriptiva, consiste en explorar una amplia gama de escenarios posibles antes de sugerir la mejor manera de responder a una situación presente o futura.

Generalmente se basa en problemas de negocios, como planificación compleja, programación, fijación de precios, inventario o administración de recursos. La analítica prescriptiva consiste en una multitud de problemas operacionales que están más allá de las capacidades del cerebro humano o del software de oficina estándar.

Para cualquier problema, se comienza a resolverlo con el modelo de optimización, que es la formulación matemática del problema que puede ser interpretada y resuelta por un motor de optimización. El modelo de optimización especifica las relaciones entre los objetivos, límites y elecciones que están involucradas en las decisiones. Pero son los datos de entrada los que hacen que estas relaciones sean concretas. Un modelo de optimización para la planificación de la producción, por ejemplo, puede tener la misma

forma si está produciendo tres productos o mil. El modelo de optimización más los datos de entrada crea una instancia de un problema de optimización.

Los motores de optimización (o solucionadores) aplican algoritmos matemáticos para encontrar una solución, un conjunto de decisiones que alcanza los mejores valores de los objetivos y respeta los límites impuestos. El motor de optimización implementa algoritmos especializados que se han desarrollado y ajustado para resolver de manera eficiente una gran variedad de problemas diferentes. Decision Optimization utiliza los motores de optimización IBM CPLEX que han demostrado ser especialmente útiles para las aplicaciones del mundo real.

Decision Optimization permite crear diferentes escenarios pues proporciona una plataforma configurable para dar soporte a los responsables de tomar las decisiones con analítica para resolver sus retos de planificación y programación. Reduce el esfuerzo, el tiempo y el riesgo asociado a la creación de soluciones personalizadas que mejoran los resultados de negocio.

Para comenzar y aprender a utilizarlo en DSX en Local:

https://github.com/jc900/FastStart_DDLabs