

IBM

Data Science

Experience Local



Manual para Workshop
Febrero 2018

Índice

Introducción

Workshop 1. Comenzar con DSX

1. 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. DSX Local Workshop

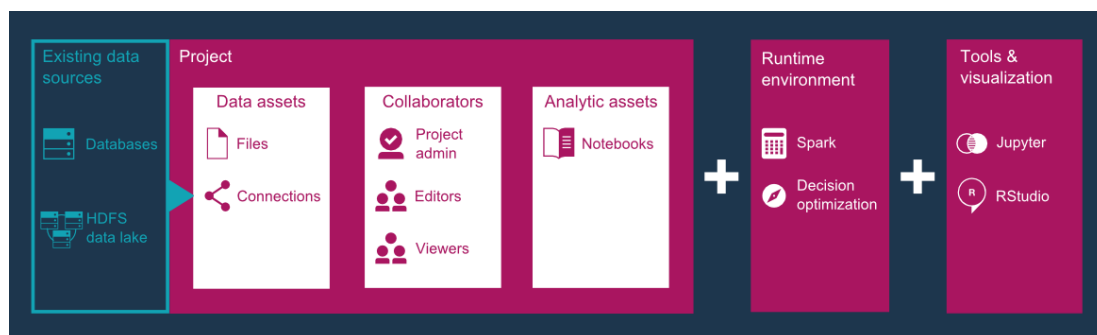
1. DSX Local Workshop
2. Decision Optimization en DSX Local
3. Deep Learning

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, pero incluye 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 variantes: IBM Data Science Experience en Local y en Cloud. En este workshop nos centraremos en IBM Data Science Experience en Local.

IBM Data Science Experience (DSX) Local es una solución empresarial lista para usar en tus instalaciones dirigida no solo a científicos de datos sino que también está dirigida a ingenieros de datos. Ofrece un conjunto de herramientas de data science, como los notebooks RStudio, Spark, Jupyter y Zeppelin, que están integradas con tecnologías propietarias de IBM.



La interfaz intuitiva de DSX proporciona un espacio de proyectos de colaboración para equipos e individuos para optimizar su tiempo y rendimiento. Los proyectos pueden contener notebooks, assets de datos y colaboradores.

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**.

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.

Entra en: <https://9.172.229.220/>

Registrare:

IBM Data Science Experience Local

Data Science Experience

[Sign in](#) [Sign up](#)

Complete fields to request access from your organization's administrator.

First Name
Candela ✓

Last Name
Retolaza ✓

Username
candela_retolaza ✓

Email
candela_retolaza@hotmail.com ✓

Password
••••••••

Confirm password
••••••••

[Sign Up](#)

El administrador/administradores tiene que aceptar la petición y podrás comenzar a utilizar DSX.

Sample notebooks

NOTEBOOK - JUPYTER	NOTEBOOK - JUPYTER	NOTEBOOK - JUPYTER	NOTEBOOK - JUPYTER
Analyze open data sets with...	Train and predict with Scala machine...	Using the Progress Callbacks with...	How to make targeted offers to customers?
AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM-DD
DATE: Aug 29, 2016	DATE: Mar 07, 2017	DATE: Mar 07, 2017	DATE: Mar 07, 2017
TOPIC: Society	TOPIC: Transportation	TOPIC: Science & Technology	TOPIC: Economy & Business
SOURCE: External	SOURCE: External	SOURCE: External	SOURCE: Self-Contained

Recently updated projects View all

NAME	PROJECT TYPE	ROLE	COLLABORATORS	LAST UPDATED
DSX_Workshop	Standard	Admin		2018-02-08
FastStartTELCO	Standard	Admin		2018-02-02
DD Workshop	Standard	Admin		2018-01-31
DSX_Local_WorkshopJacques	Standard	Viewer	..	2018-01-30
DSX_Local_WorkshopCandela_Carmen	Standard	Admin		2018-01-30

Helpful links

- Docs: Find the information you need
- Discussion forum: Start a discussion or ask a question
- Blog: Read and follow our latest news
- Got ideas?: Have feedback on Data Science Experience?

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

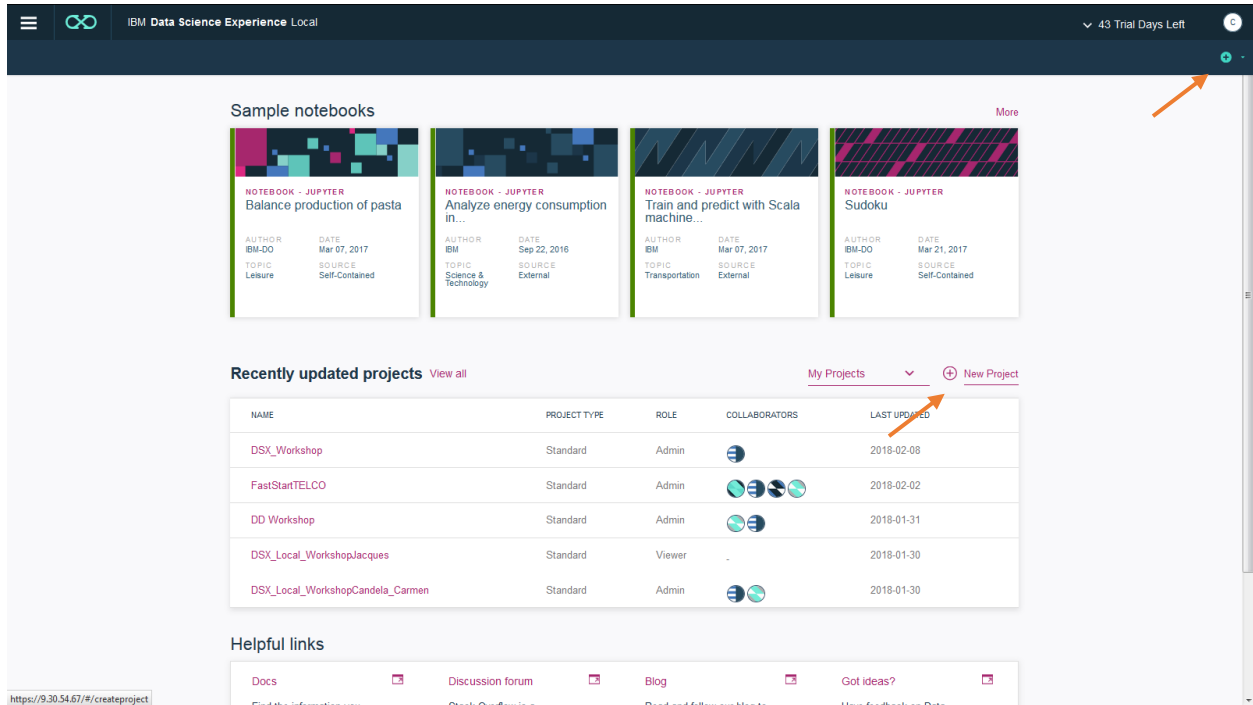
1. Configura un proyecto para organizar sus recursos.
2. Agrega colaboradores a tu proyecto.
3. Agrega datos a tu proyecto.
4. Comience a analizar datos. Por ejemplo, puede crear notebooks, y hacer un deployment para crear aplicaciones o servicios web.

¿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.

Workshop 2.

Crea y organiza los recursos en un proyecto

1. Crear un proyecto



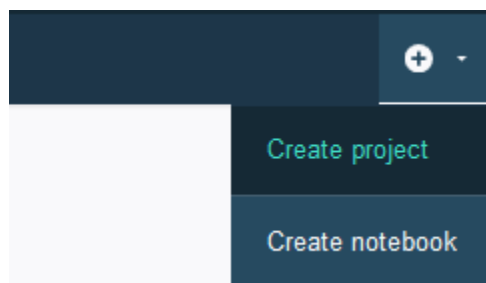
The screenshot shows the IBM Data Science Experience Local dashboard. In the top right corner, there is a dark blue header bar with a '43 Trial Days Left' indicator and a circular icon containing a plus sign. An orange arrow points to this icon. Below the header, the main content area displays 'Sample notebooks' and 'Recently updated projects'. The 'Recently updated projects' section includes a table with columns: NAME, PROJECT TYPE, ROLE, COLLABORATORS, and LAST UPDATED. A 'New Project' button is visible next to the 'My Projects' dropdown menu.

NAME	PROJECT TYPE	ROLE	COLLABORATORS	LAST UPDATED
DSX_Workshop	Standard	Admin		2018-02-08
FastStartTELCO	Standard	Admin		2018-02-02
DD Workshop	Standard	Admin		2018-01-31
DSX_Local_WorkshopJacques	Standard	Viewer		2018-01-30
DSX_Local_WorkshopCandela_Carmen	Standard	Admin		2018-01-30

Podemos hacerlo de varias maneras, por ejemplo, arriba a la derecha, pinchando en



el icono, creamos un nuevo proyecto



Sample notebooks

NOTEBOOK - JUPYTER
Balance production of pasta

AUTHOR: IBM-DO, DATE: Mar 07, 2017, TOPIC: Leisure, SOURCE: Self-Contained

NOTEBOOK - JUPYTER
Analyze energy consumption in...

AUTHOR: IBM, DATE: Sep 22, 2016, TOPIC: Science & Technology, SOURCE: External

NOTEBOOK - JUPYTER
Train and predict with Scala machine...

AUTHOR: IBM, DATE: Mar 07, 2017, TOPIC: Transportation, SOURCE: External

NOTEBOOK - JUPYTER
Sudoku

AUTHOR: IBM-DO, DATE: Mar 21, 2017, TOPIC: Leisure, SOURCE: Self-Contained

Recently updated projects [View all](#) My Projects [New Project](#)

NAME	PROJECT TYPE	ROLE	COLLABORATORS	LAST UPDATED
DSX_Workshop	Standard	Admin		2018-02-08
FastStartTELCO	Standard	Admin		2018-02-02
DD Workshop	Standard	Admin		2018-01-31
DSX_Local_WorkshopJacques	Standard	Viewer		2018-01-30
DSX_Local_WorkshopCandela_Carmen	Standard	Admin		2018-01-30

Helpful links

[Docs](#) [Discussion forum](#) [Blog](#) [Got ideas?](#)

<https://9.30.54.67/#/createProject>

Por ejemplo, llamamos al nuevo proyecto: Workshop_CR con tus iniciales, pues estando en el mismo clúster el nombre de los proyectos deber ser único. Añadimos una descripción opcional: por ejemplo, Workshop DSX. Vemos que existen tres maneras de crear un proyecto: En blanco, Desde un fichero o desde GitHub. Vamos a comenzar por crear uno en blanco.

Create Project

Blank From File From Github

Name*

Workshop

This name is valid 92

Description

Type your description here

☐ Library Project

Cancel Create

Gestión de proyectos:

Los proyectos son la base de operaciones para la colaboración. Puede usar proyectos para recopilar activos, como cuadernos, modelos, fuentes de datos y conjuntos de datos remotos y locales, en un solo lugar. Los proyectos son una excelente manera de trabajar con un equipo: compartir un conjunto de activos y luego construir nuevos modelos y análisis juntos. Puede colaborar en DSX Local o en un repositorio externo de GitHub.

DSX Local administra sus recursos y proporciona opciones automáticas de escalado y recuperación. La consola de administración (IBM Data Platform Manager) lo ayuda a administrar y supervisar su hardware, usuarios y servicios. La seguridad está integrada en DSX desde el cifrado de datos en reposo y en movimiento hasta la administración de usuarios.

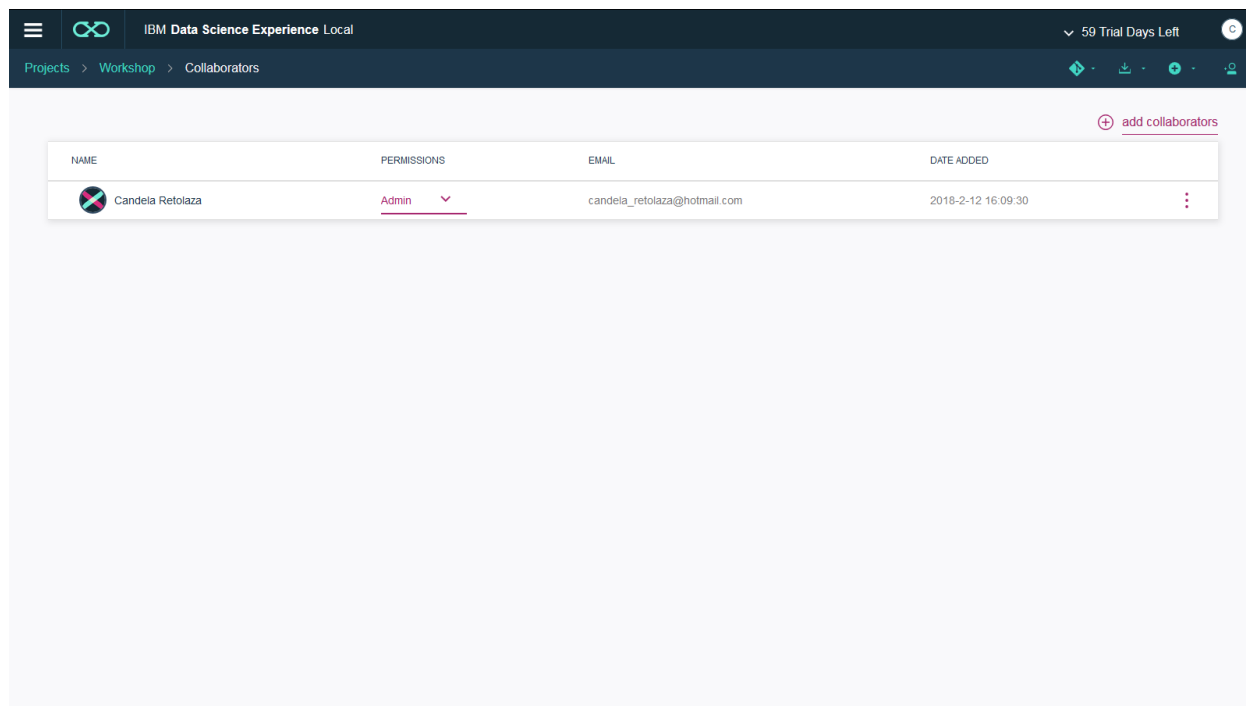
Si tienes permisos de administrador en un proyecto, tienes control total sobre él. Si tiene permisos de Editor, puedes agregar activos y colaboradores a un proyecto.

The screenshot shows the IBM Data Science Experience Local interface. At the top, there's a header with the IBM logo and 'IBM Data Science Experience Local'. Below the header, there's a navigation bar with 'Projects' and 'Worshop'. The main content area displays the 'Worshop' project details. It includes a description 'Description not available.', the date created 'Fri Feb 09 2018', and a summary of resources: 0 Assets, 5 Environments, 0 Jobs, 0 Data Sources, and 1 Collaborator. Below this, there's a section for 'Collaborators' showing 'Candela Retolaza admin' with a 'View all (1)' link. To the right, there's a 'Recent Assets' section with a table header: NAME, ASSET TYPE, and LAST MODIFIED. The table is currently empty, with a message 'you have no recently modified assets'.

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.

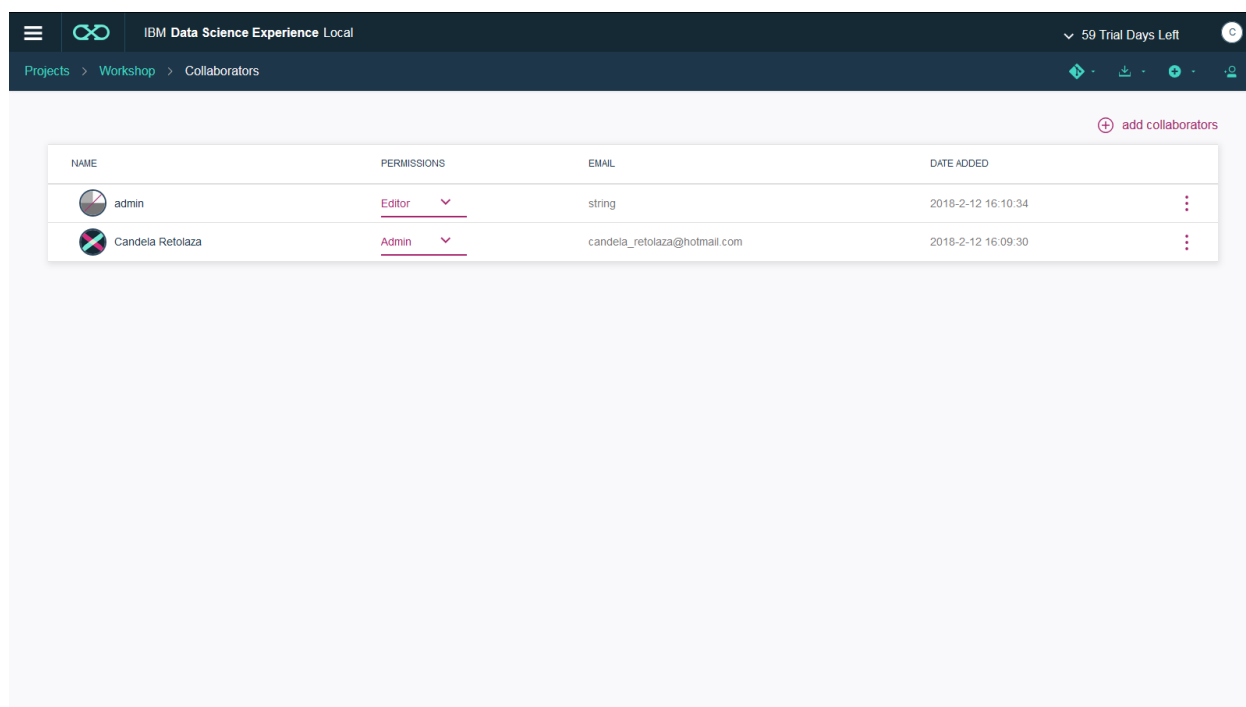
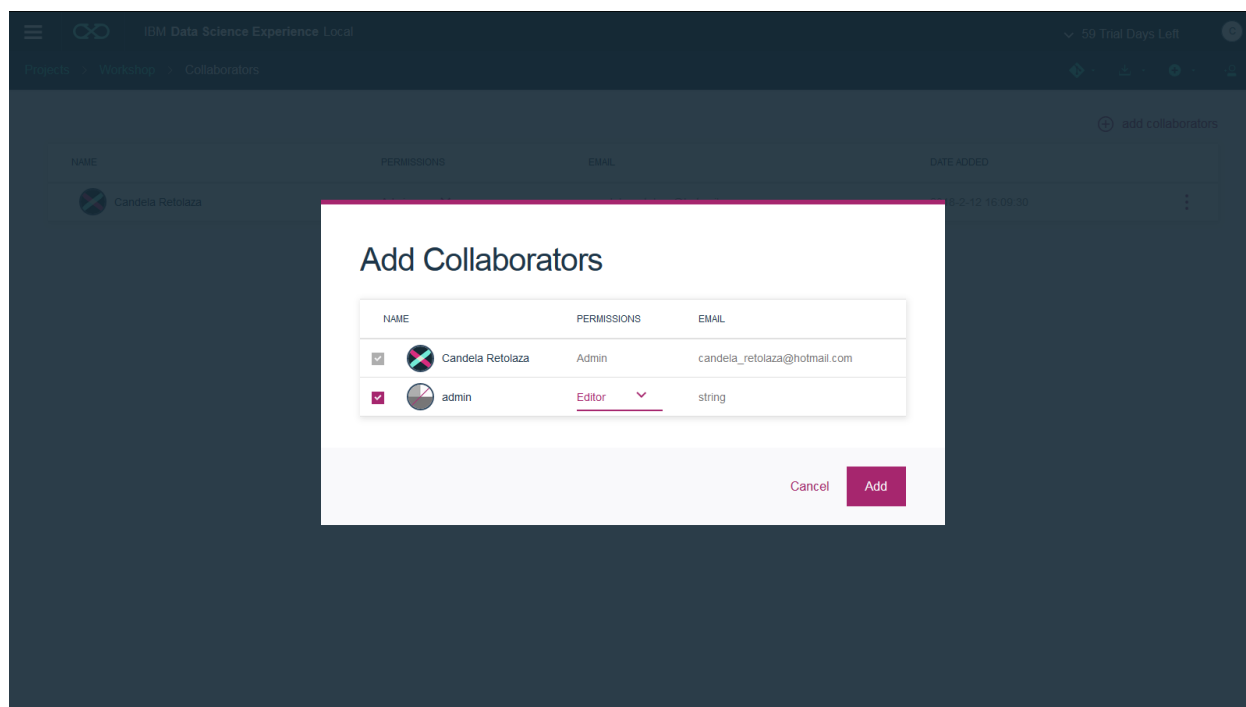
2. Añade colaboradores

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.

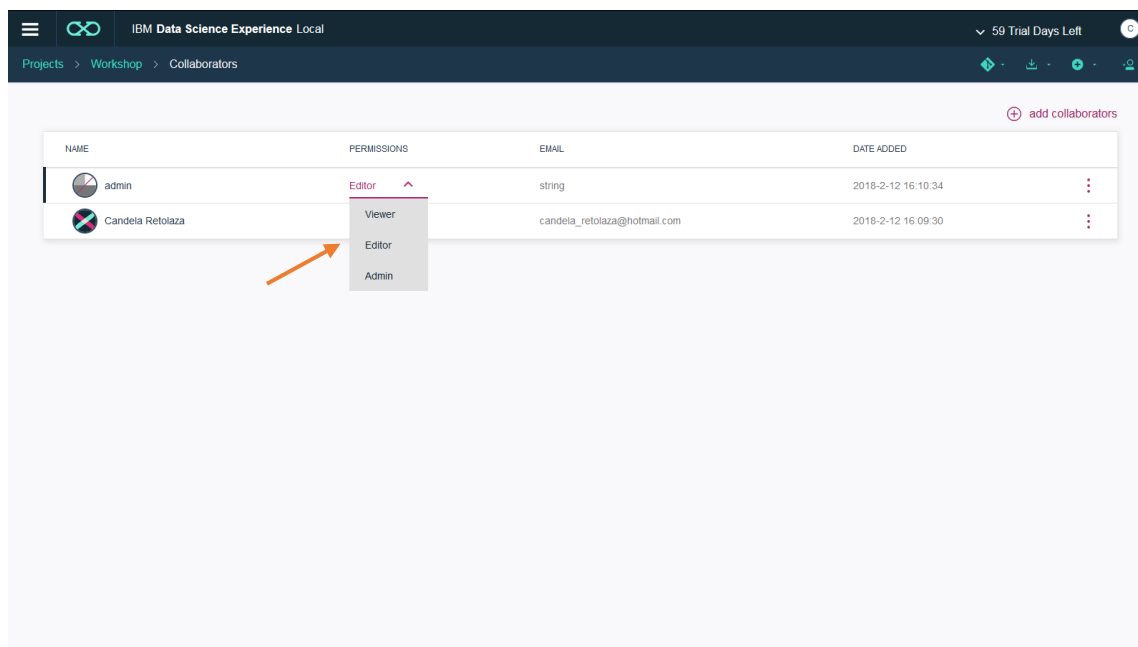


Para poder hacer esto, solo necesito que la persona a la que quiero añadir tenga acceso al mismo cluster.

Elegimos a quienes creemos que estén en nuestro proyecto, seleccionamos el tipo de acceso que queremos dar a esa persona y lo añadimos. Y ya estará en nuestro proyecto.



Una vez añadido, 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, puede agregar recursos como conexiones, datos de fichero plano, datos de bases de datos, Notebooks, flujos de aprendizaje automático, modelos.

Puede crear un cuaderno Jupyter en R, Python o Scala, o un cuaderno Zeppelin con una combinación de R, Python o Scala. También puede crear modelos automáticos o semiautomáticos y flows de SPSS Modeler. Puede comenzar desde cero, importar un bloc de notas existente o usar uno de los ejemplos de la comunidad.

Cuando abra un notebook nuevo o existente, la barra de acción está disponible. Desde la barra de acciones, puede ver y usar fuentes de datos conectadas, ver los comentarios y versiones, y compartir su proyecto o notebook con otras personas.

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

Workshop 3.

Parte predictiva

1. 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. Podemos entrar directamente desde la pantalla principal, o desde el proyecto, pinchando en crear un nuevo Flow de SPSS Modeler.

IBM Data Science Experience Local

Community

Published Assets

Projects

Model Management

Tools

Jupyter

RStudio

Zeppelin

H2O Flow

SPSS Modeler ^{RC}

Environments

Sample notebooks

More

NOTEBOOK - JUPYTER
Model a Golomb ruler using DO
AUTHOR: IBM-DO, DATE: Dec 12, 2016, TOPIC: Science & Technology, SOURCE: Self-Contained

NOTEBOOK - JUPYTER
Using the Progress Callbacks with...
AUTHOR: IBM, DATE: Mar 07, 2017, TOPIC: Science & Technology, SOURCE: External

NOTEBOOK - JUPYTER
Work with a remote Spark Sample One
AUTHOR: IBM, DATE: Jul 07, 2017, TOPIC: Science & Technology, SOURCE: Self-Contained

NOTEBOOK - JUPYTER
Work with a remote Spark Sample...
AUTHOR: IBM, DATE: Mar 07, 2017, TOPIC: Science & Technology, SOURCE: External

Recently updated projects

View all

My Projects New Project

NAME	PROJECT TYPE	ROLE	COLLABORATORS	LAST UPDATED
Intro_DeepLearning_Candela	Standard	Admin		2018-02-13
DSX_Local_Workshop_CR	Standard	Admin		2018-02-12
Workshop	Standard	Admin		2018-02-12
dsx-samples	-	Viewer	-	2018-02-12

Helpful links

Docs
Find the information you need. Watch videos of key tasks.

Discussion forum
Stack Overflow is a community of 6.9 million programmers just like you, helping each other. Join the conversation on Data Science Experience

Blog
Read and follow our blog to keep up with the latest updates about Data Science Experience.

Got ideas?
Have feedback on Data Science Experience? Submit your ideas in our product forum or vote on ideas submitted by others.

IBM Data Science Experience Local

Projects > Workshop > All

All Notebooks RStudio Models SPSS Modeler Streams Scripts Data Sets Other Files Published Assets

Notebooks

view all (0) add notebook

NAME	STATUS	ENVIRONMENT	TOOL	LAST MODIFIED
you have no notebooks				

RStudio

view all (0) open RStudio

NAME	TYPE	LAST MODIFIED
you have no rstudio files		

Models

view all (0) add model

NAME	TYPE	STATUS	LAST MODIFIED
you have no model files			

SPSS Modeler Streams

view all (1) add stream

NAME	TYPE	LAST MODIFIED	ACTIONS
Drug Study Example	SPSS	02-14-2018	

Scripts

view all (0) add script

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:

IBM Data Science Experience Local

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 CART algorithms to predict the effect of advertising promotions on the sale of various items. Input data of sales before and after a past promotion are used to train the model to predict the effectiveness of advertising.

Cancel Create

Projects Workshop Drug Study Example

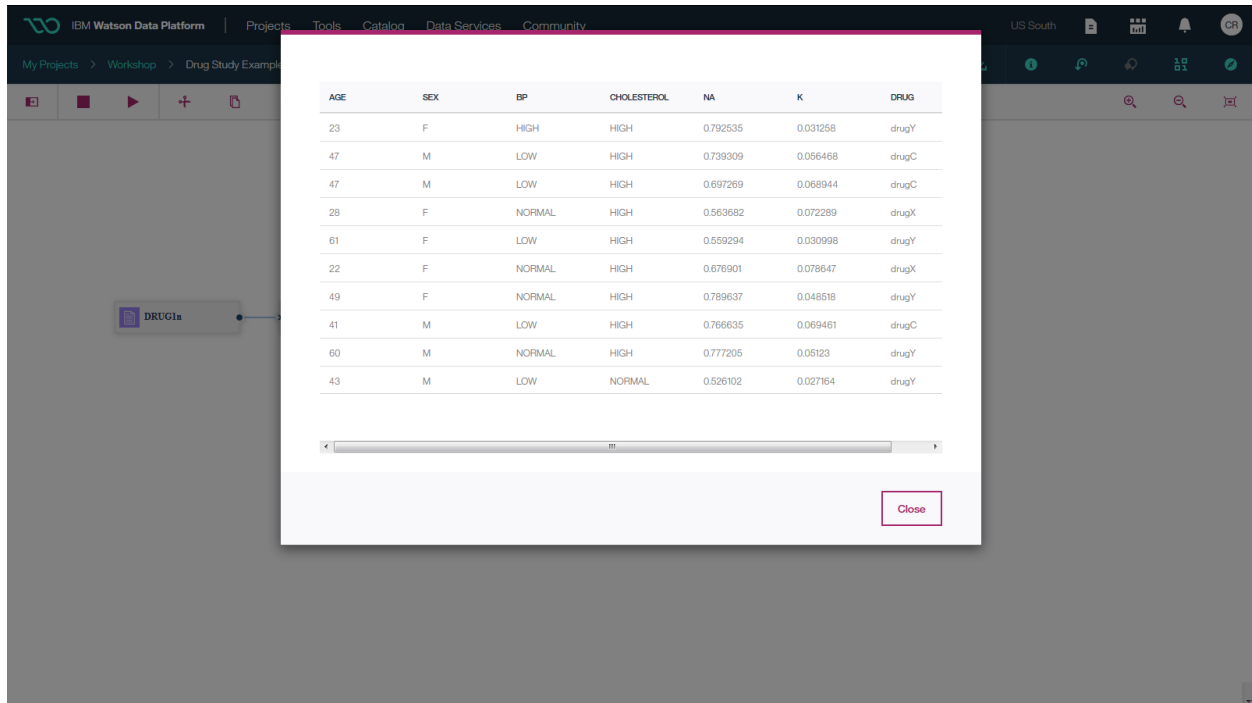
Palette

Drug Na_Bu_K Discard Fields Define Types Drug Drug Drug Drug

Abre la paleta para ver todos los nodos que se pueden utilizar.

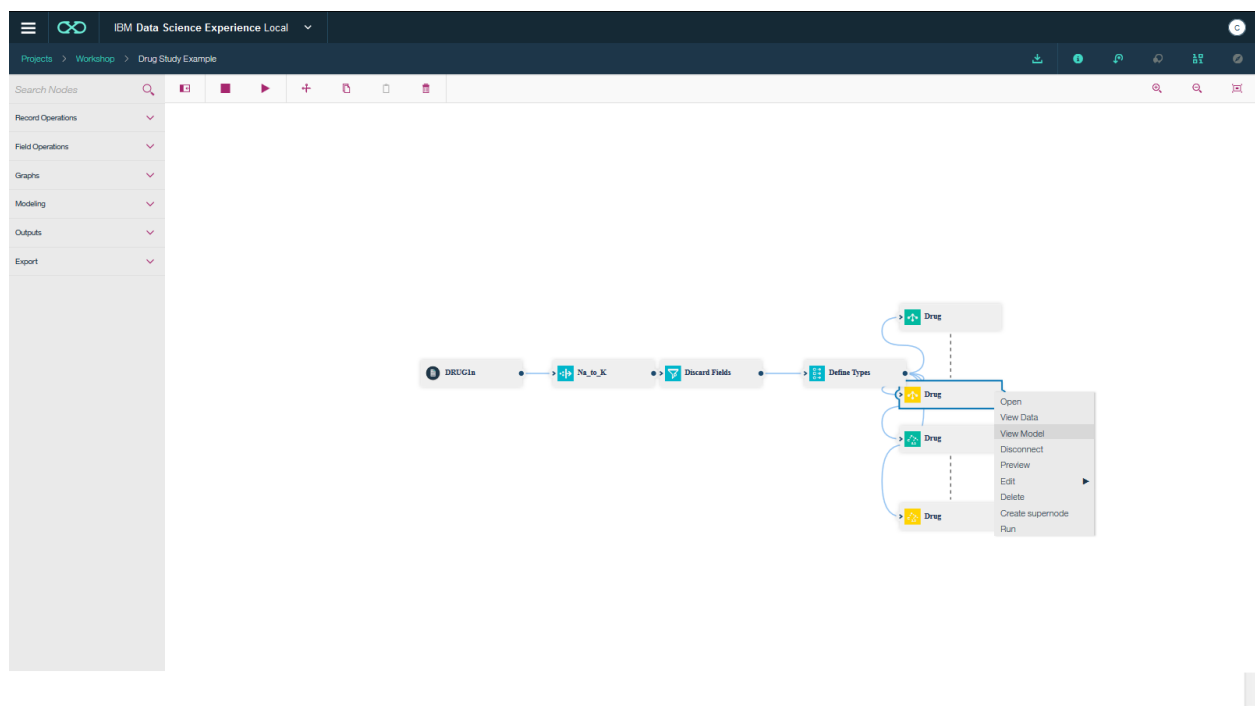
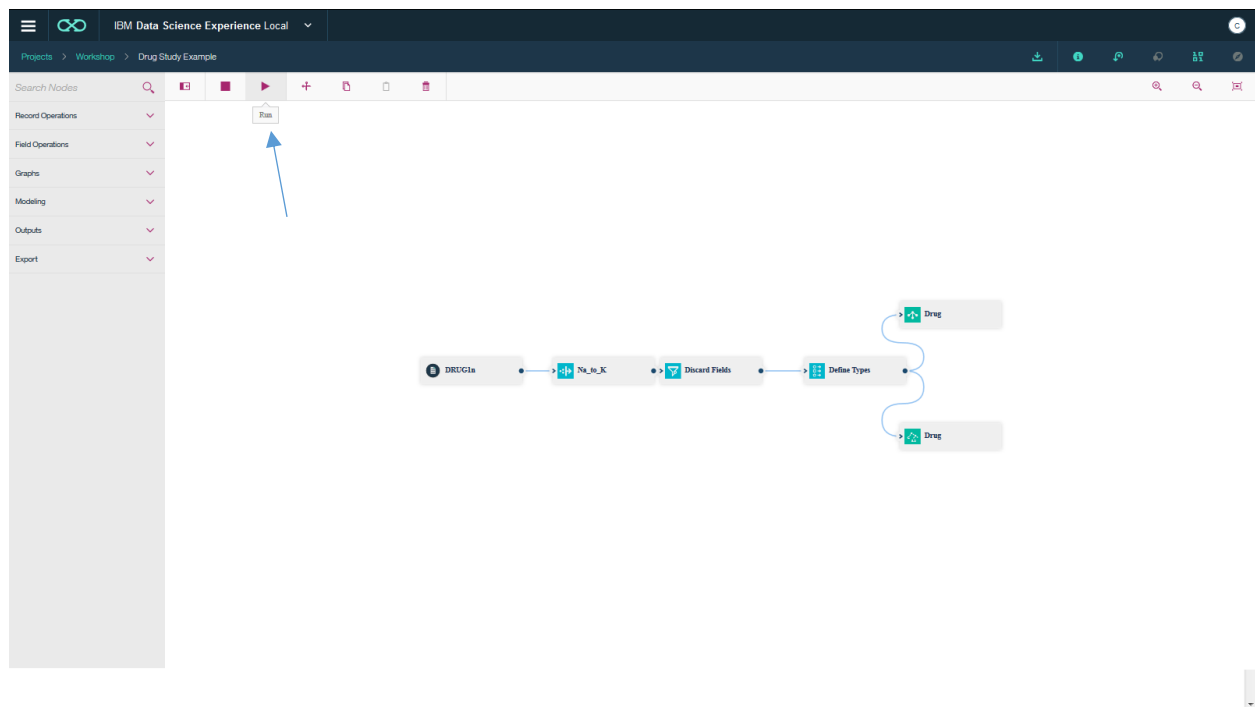
En este ejemplo, imagina que es un investigador médico que está recopilando datos para un estudio. Has recopilado información sobre un conjunto de pacientes, de los

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.

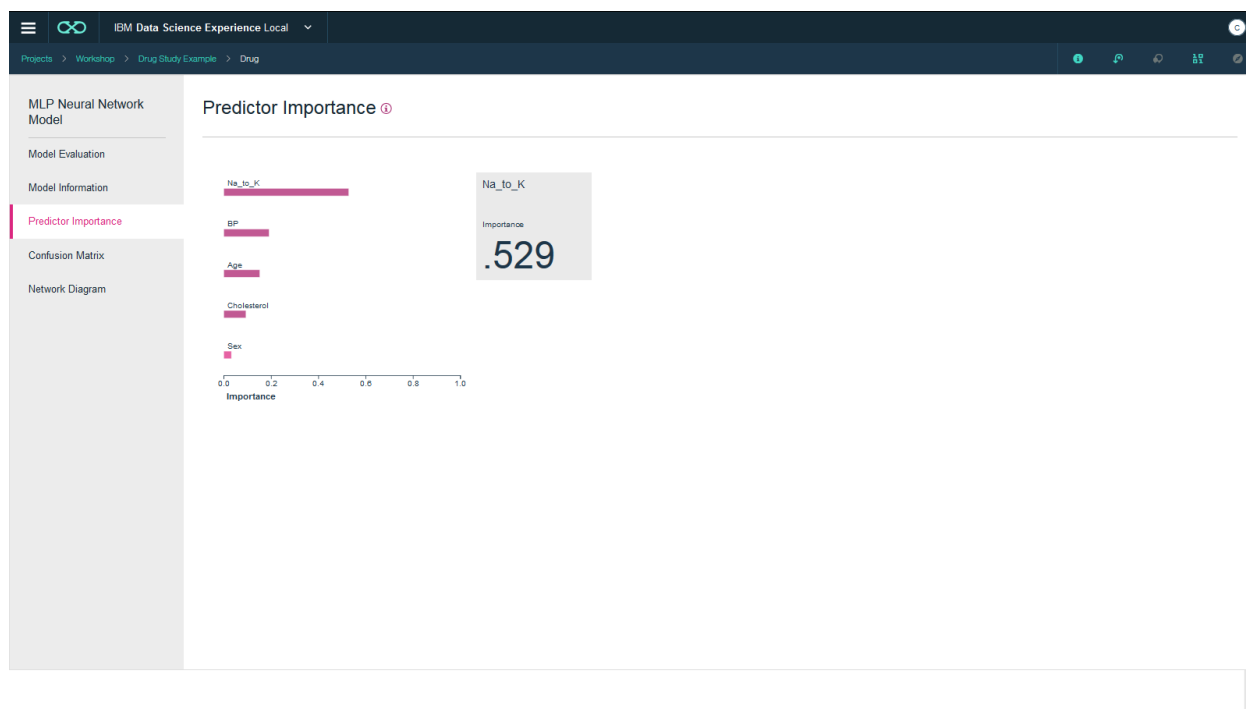


AGE	SEX	BP	CHOLESTEROL	NA	K	DRUG
23	F	HIGH	HIGH	0.792535	0.031258	drugY
47	M	LOW	HIGH	0.739309	0.056468	drugC
47	M	LOW	HIGH	0.697269	0.068944	drugC
28	F	NORMAL	HIGH	0.563682	0.072289	drugX
61	F	LOW	HIGH	0.559294	0.030998	drugY
22	F	NORMAL	HIGH	0.676901	0.078647	drugX
49	F	NORMAL	HIGH	0.789637	0.048518	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. Se propone que el lector entienda el modelo, y lo ejecute:



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



IBM Data Science Experience Local

Projects > Workshop > Drug Study Example > Drug

MLP Neural Network Model

Model Evaluation

Model Information

Predictor Importance

Confusion Matrix

Network Diagram

Confusion Matrix ⓘ

Observed	Predicted					Percent Correct
	drugA	drugB	drugC	drugX	drugY	
drugA	23	0	0	0	0	100.0%
drugB	0	16	0	0	0	100.0%
drugC	0	0	16	0	0	100.0%
drugX	0	0	0	54	0	100.0%
drugY	0	0	0	0	91	100.0%
Percent Correct	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Si damos a **view data** podemos visualizar los datos y hacer auditoria de mis datos, ver la calidad, ver los estadísticos, ver gráficos, modificarlos, descargarlos, seleccionar datos desde el gráfico y crear nodos directamente desde el propio gráfico, etc.



[illegible]

IBM Data Science Experience Local

Projects

Workshop

Drug Study Example

Audit

Quality

Statistics

Pearson Correlations

Age

Statistics

Count	200
Mean	44,315
Min	15
Max	74
Range	59
Variance	273,714
Standard Deviation	16,544
Standard Error of Mean	1.17
Median	45
Mode	47

Pearson Correlations

Na_to_K	-0.063	Weak
SNC-Drug	-0.095	Weak

SNC-Drug

Statistics

Count	200
Mean	0.986
Min	0.531
Max	1
Range	0.469
Variance	0.002
Standard Deviation	0.046
Standard Error of Mean	0.003
Median	0.998
Mode	1

*Multiple modes exist. The smallest value is shown.

Pearson Correlations

Age	-0.095	Weak
Na_to_K	0.134	Weak

Na_to_K

Statistics

Count	200
Mean	16,084
Min	6,269
Max	38,247
Range	31,978
Variance	52,195
Standard Deviation	7,224
Standard Error of Mean	0.811
Median	13,938
Mode	10,977

*Multiple modes exist. The smallest value is shown.

Pearson Correlations

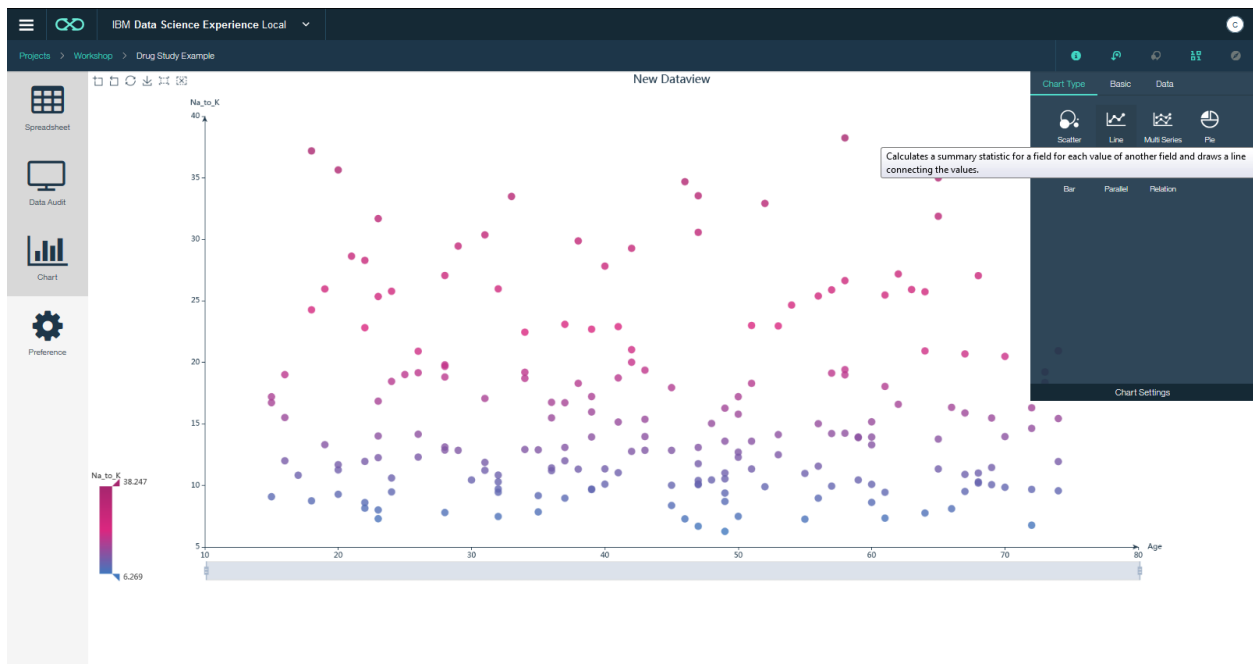
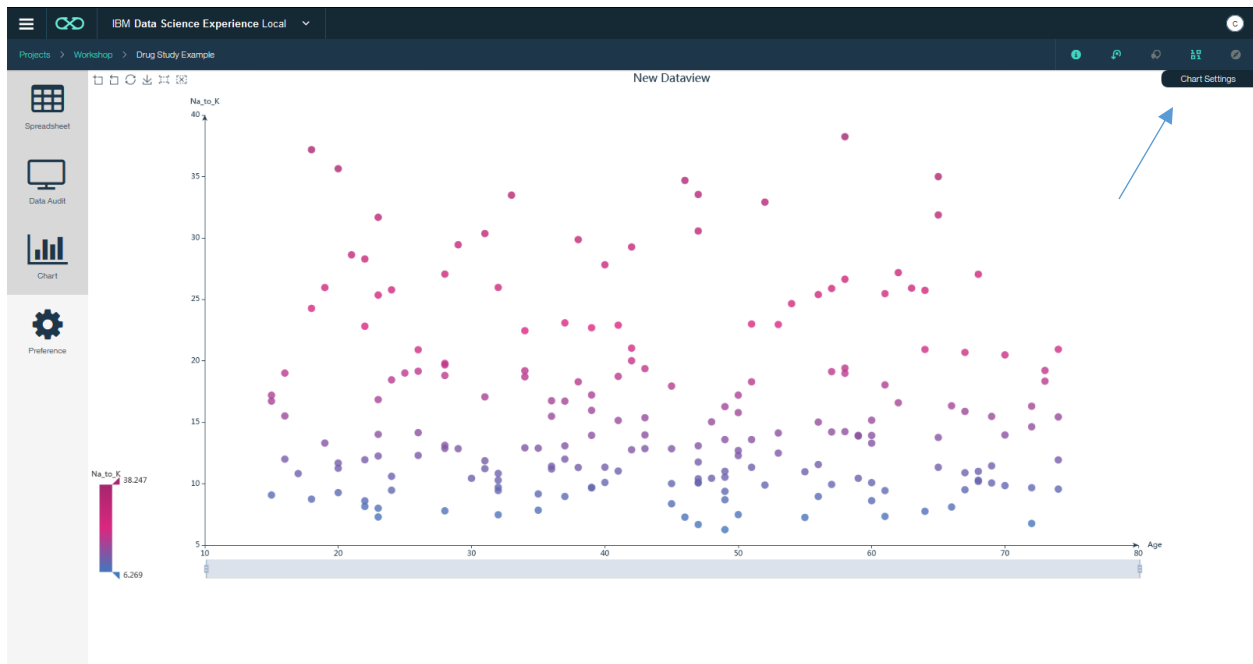
Age	-0.063	Weak
SNC-Drug	0.134	Weak

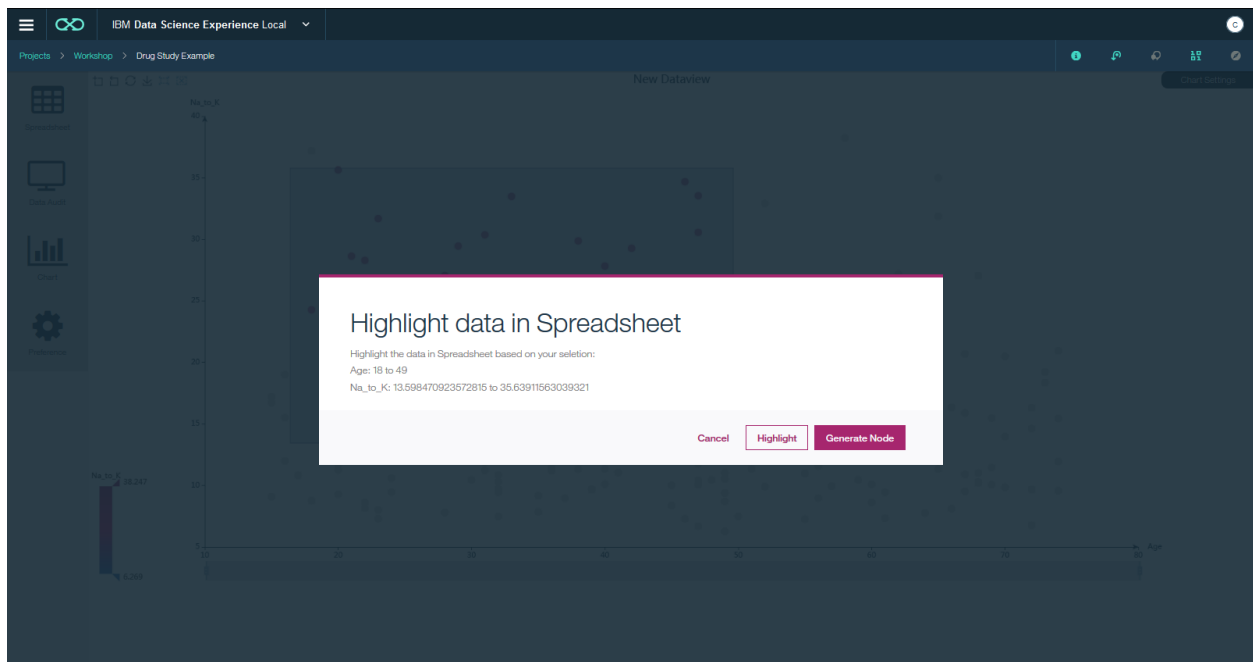
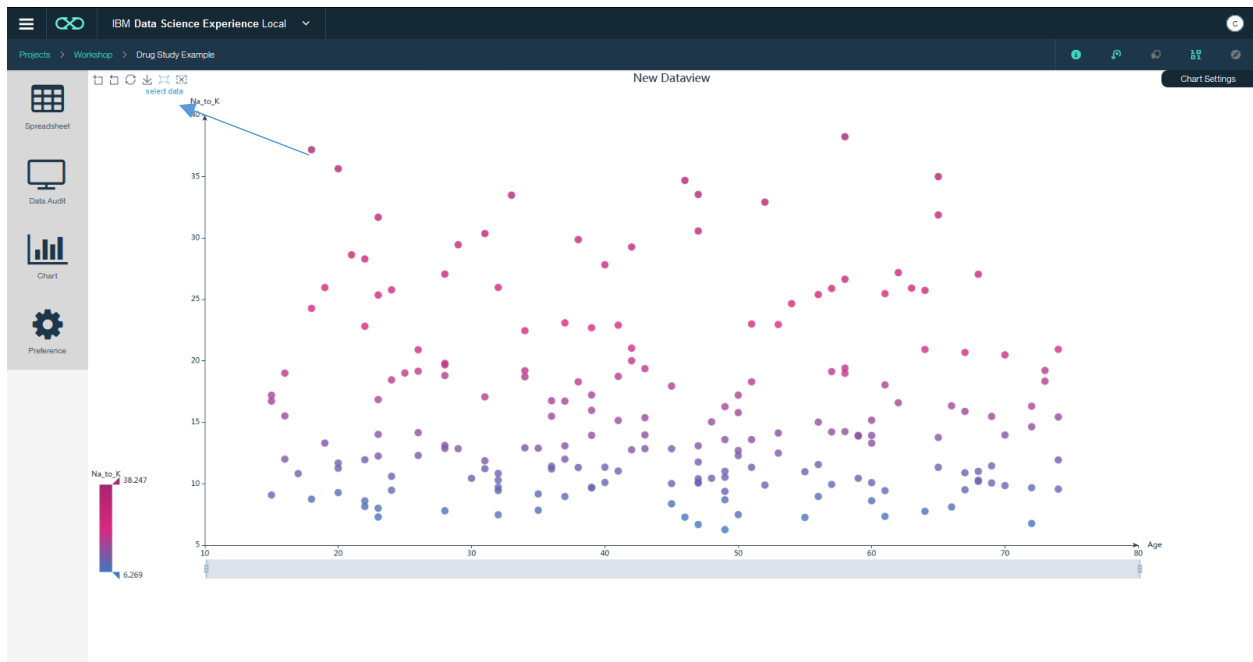
Spreadsheet

Data Audit

Chart

Preference





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

2. DSX Local Workshop

En este workshop, aprenderás a desarrollar e implementar aplicaciones en DSX Local. El workshop se ha dividido en varias partes independientes para aquellos que están interesados en una determinada herramienta de desarrollo o una determinada tarea de implementación.

Vamos a seguir el siguiente repositorio:

<https://github.com/CandelaRetolaza/Data-Science-Experience-en-LOCAL>

3. 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 se propone seguir el siguiente Workshop:

https://github.com/jc900/FastStart_DDLabs

4. 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-0bfb-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=18379e532a9953d4e97f2a75eee37a8ece9ee4745676e1a647493fbfdd7b16fb

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>

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/arvizga7/Self-Driving-Car-in-DSX>

Workshop EXTRA.

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

ícono  arriba a la derecha, o clickando en **New data asset**.

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

My Projects > Workshop

Overview **Assets** Bookmarks Deployments Collaborators Settings

What assets are you looking for?

Data assets [New data asset](#)

0 assets selected.

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				

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

My Projects > Workshop

Overview **Assets** Bookmarks Deployments Collaborators Settings

What assets are you looking for?

Data assets

0 assets selected.

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

Drop files here or browse for files to upload.

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 IBM Watson Data Platform interface. The main navigation bar includes 'Projects', 'Tools', 'Catalog', 'Data Services', and 'Community'. The 'Assets' tab is active, displaying a search bar and several sections: 'Data assets', 'Notebooks', 'Streams flows', and 'Models'. Under 'Data assets', a table lists 'datos_banca.csv' as a Data Asset. An orange arrow points to the file name. The right sidebar shows the 'Files' tab with 'datos_banca.csv' selected.

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 table displays 12 columns: AGE, ED, EMPLOY, ADDRESS, INCOME, DESTING, CREDIT, OTHECIBT, DEFAULT, PREDDEF1, PREDDEF2, and PREDDEF3. The first row of data is highlighted.

AGE	ED	EMPLOY	ADDRESS	INCOME	DESTING	CREDIT	OTHECIBT	DEFAULT	PREDDEF1	PREDDEF2	PREDDEF3
41.000000	3.000000	17.000000	12.000000	176.000000	9.300000	11.359392	5.008608	1.000000	0.806394	0.788640	0.213043
27.000000	1.000000	10.000000	6.000000	31.000000	17.300000	1.362202	4.000796	0.000000	0.198297	0.128445	0.436903
40.000000	1.000000	15.000000	14.000000	55.000000	5.500000	0.856075	2.168925	0.000000	0.010036	0.002987	0.141023
41.000000	1.000000	15.000000	14.000000	120.000000	2.900000	2.658720	0.821280	0.000000	0.022138	0.010273	0.104422
24.000000	2.000000	2.000000	0.000000	26.000000	17.300000	1.767436	3.056564	1.000000	0.781588	0.737885	0.436903
41.000000	2.000000	5.000000	5.000000	25.000000	10.200000	0.392700	2.157300	0.000000	0.216709	0.328195	0.233577
39.000000	1.000000	20.000000	9.000000	67.000000	30.600000	3.833874	16.668126	0.000000	0.185960	0.179256	0.817092
43.000000	1.000000	12.000000	11.000000	38.000000	3.600000	0.128592	1.238408	0.000000	0.014709	0.010573	0.113358
24.000000	1.000000	3.000000	4.000000	19.000000	24.400000	1.358348	3.277652	1.000000	0.748041	0.619443	0.663904
36.000000	1.000000	0.000000	13.000000	25.000000	19.700000	2.777700	2.147300	0.000000	0.815057	0.797227	0.515530
27.000000	1.000000	0.000000	1.000000	16.000000	1.700000	0.182512	0.089488	0.000000	0.350309	0.610511	0.090548
25.000000	1.000000	4.000000	0.000000	23.000000	5.200000	0.252356	0.843644	0.000000	0.239054	0.219024	0.136307
52.000000	1.000000	24.000000	14.000000	64.000000	10.000000	3.929600	2.470400	0.000000	0.009790	0.006285	0.226898
37.000000	1.000000	9.000000	29.000000	16.300000	1.719901	3.010399	0.364494	0.000000	0.340468	0.404835	0.404835
48.000000	1.000000	22.000000	15.000000	100.000000	9.100000	3.703700	5.396300	0.000000	0.011874	0.007714	0.208663

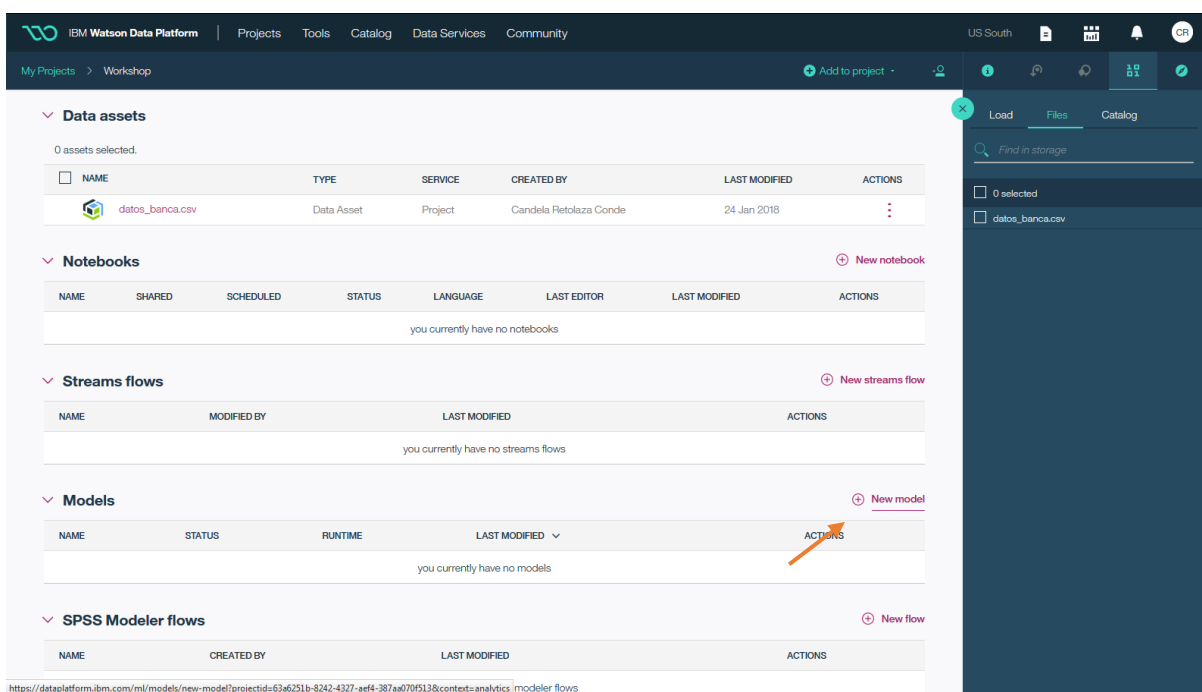
El fichero que acabamos de subir a nuestro proyecto contiene información de demográfica de clientes de un banco ficticio, con datos como: edad, nivel educativo, años en el trabajo actual, años en la misma vivienda, salario... en el que, además, tenemos información sobre los créditos que tiene cada cliente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



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.

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](#).

Cancel 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

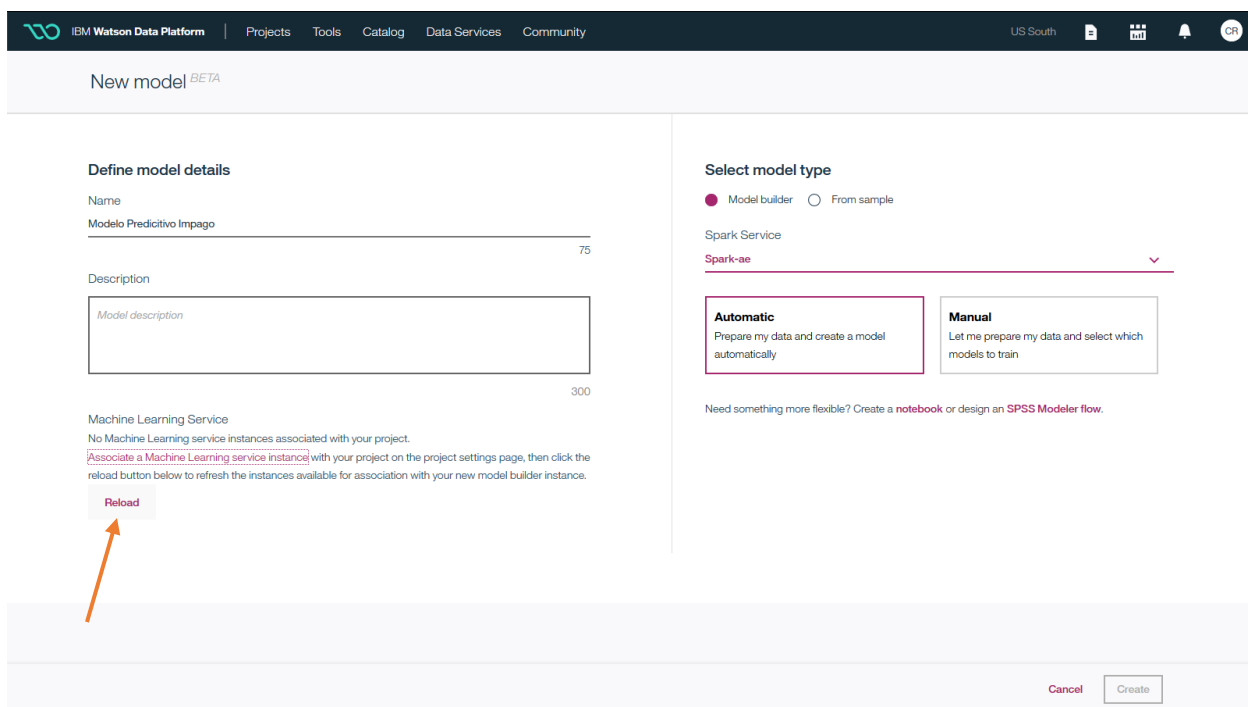
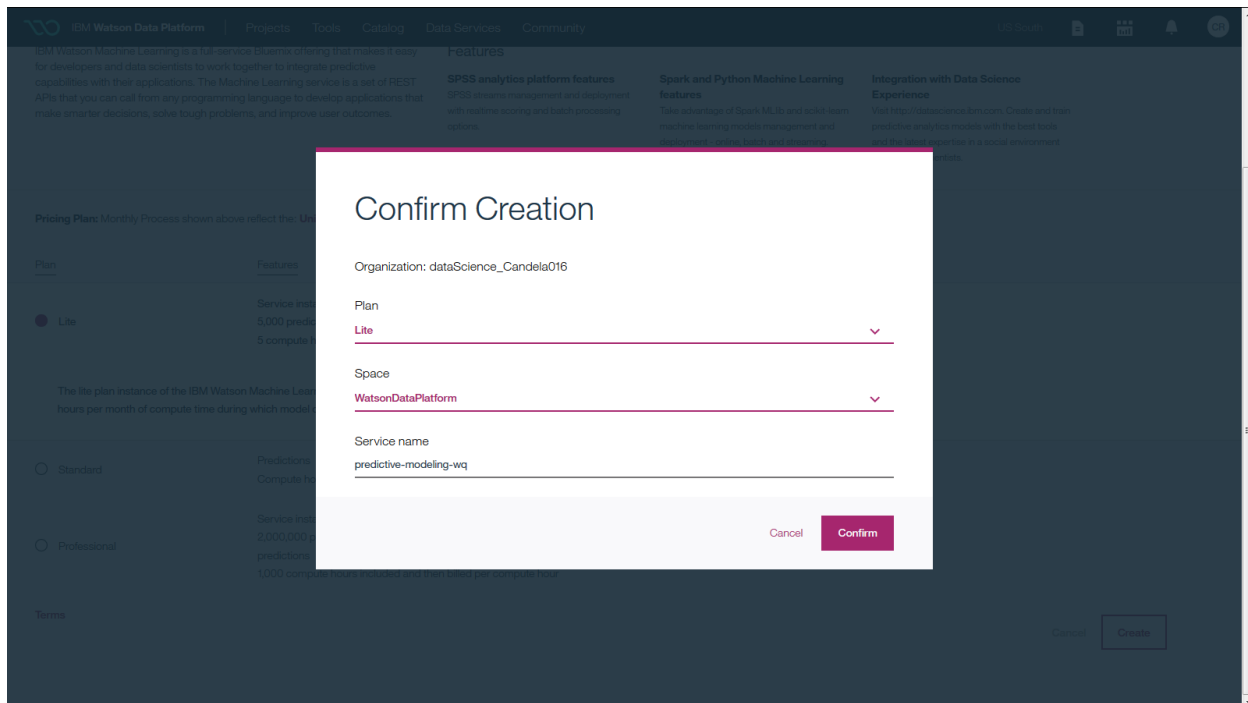
SPSS analytics platform features
SPSS streams management and deployment with realtime scoring and batch processing options.

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://data.science.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
The lite plan instance of the IBM Watson Machine Learning service provides you with a maximum of 5 deployed models, 5,000 predictions per month, and 5 hours per month of compute time during which model can be trained, evaluated, and deployed to be available to accept prediction events.		
<input type="radio"/> Standard	Predictions Compute hours Service instance	\$0.5 USD/1,000 predictions \$0.45 USD/hour €1,000 USD/instance



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

New model BETA

Define model details

Name
Modelo Predictivo Impago 75

Description
Model description 300

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

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

Select data asset

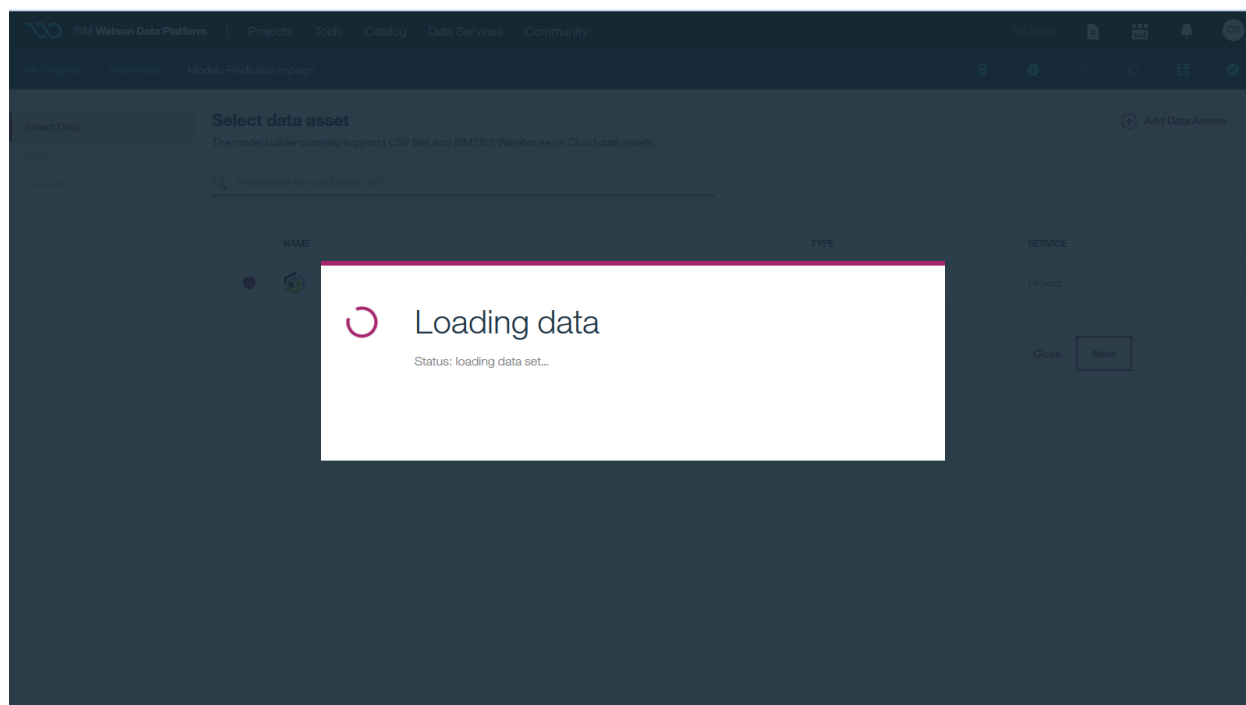
The model builder currently supports CSV files and IBM Db2 Warehouse on Cloud data assets.

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.

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

My Projects > Workshop > Modelo Predictivo Impago

Select Data

Train

Evaluate

Select a technique

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

Feature columns
All (default)

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: 60 Test: 20 Holdout: 20

Close Previous Next

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.

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

My Projects > Workshop > Modelo Predictivo Impago

Select Data

Train

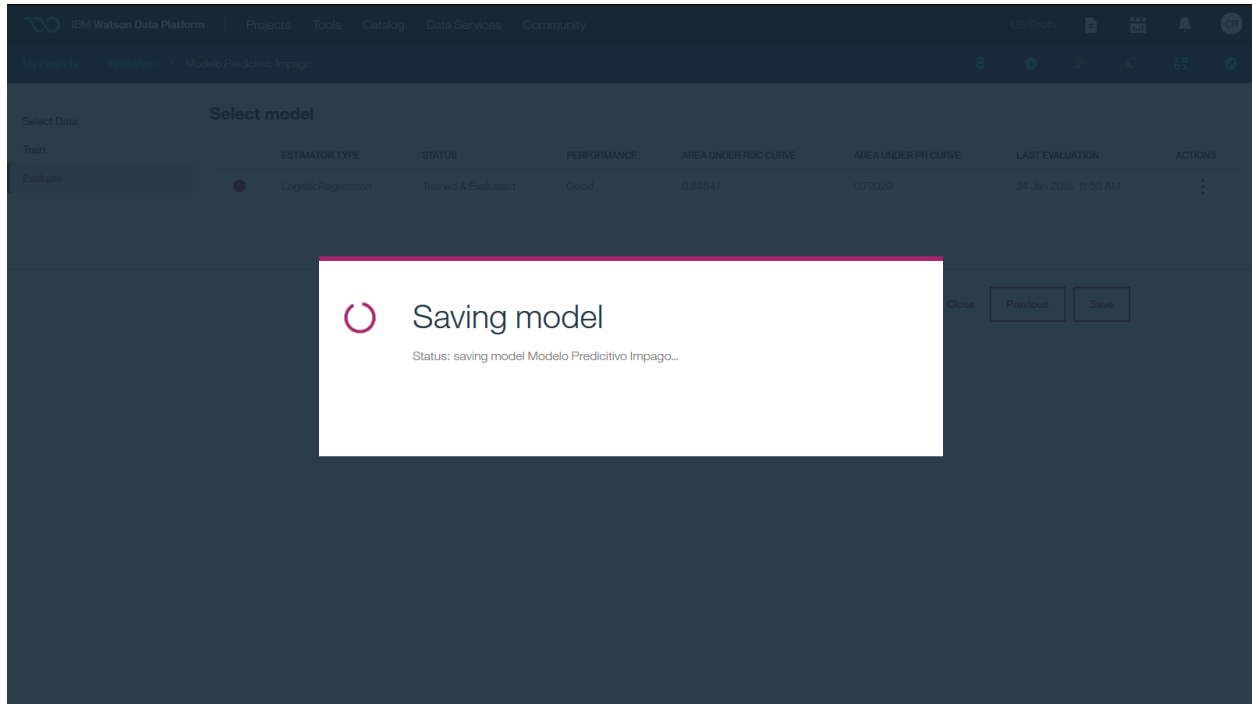
Evaluate

Select model

ESTIMATOR TYPE	STATUS	PERFORMANCE	AREA UNDER ROC CURVE	AREA UNDER PR CURVE	LAST EVALUATION	ACTIONS
LogisticRegression	Trained & Evaluated	Good	0.84647	0.72029	24 Jan 2018, 11:53 AM	

Close Previous Save

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.



3. Despliega el modelo en Watson Machine Learning

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

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(31,0)
ed	decimal(31,0)
employ	decimal(31,0)
address	decimal(31,0)

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](#)

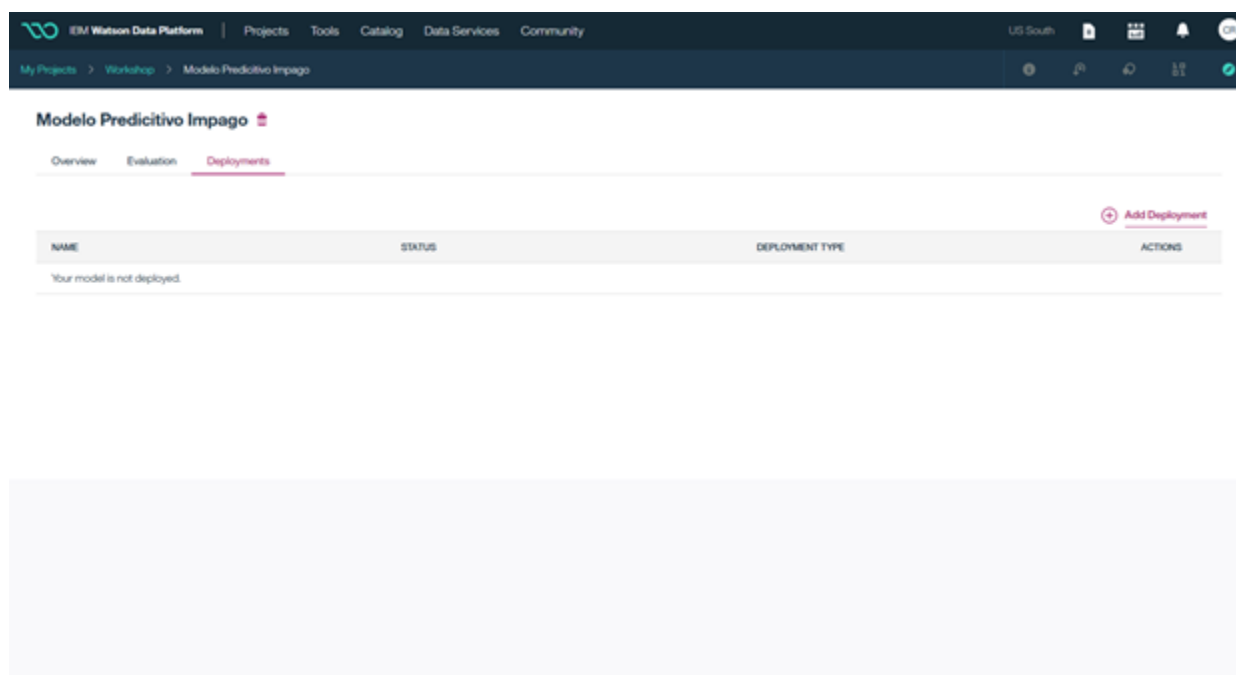
Versions

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.



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

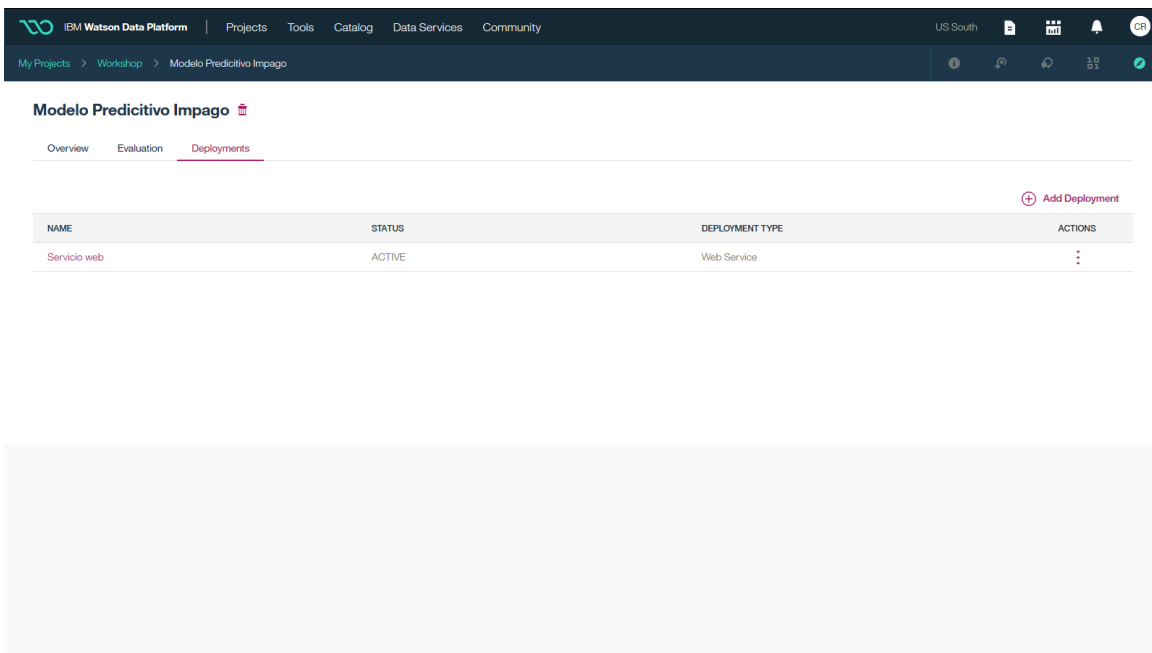
Web Service Batch Prediction Real-time Streaming Predictions

Name
Servicio web

Description
Modelo de predicción de impagos

209

Cancel Save



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

My Projects > Workshop > Modelo Predictivo Impago

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**.

IBM Watson Data Platform | Projects Tools Catalog Data Services Community

US South

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 [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
Modelo Predictivo Impago	trained	spark-2.0	24 Jan 2018	

SPSS Modeler flows [New flow](#)

NAME	CREATED BY	LAST MODIFIED	ACTIONS

<https://dataplatform.ibm.com/ml/models/new-model?projectId=63a6251b-8242-4327-ae14-387a80705138&context=data> spss modeler flows

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

IBM Watson Data Platform | Projects Tools Catalog Data Services Community

US South

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

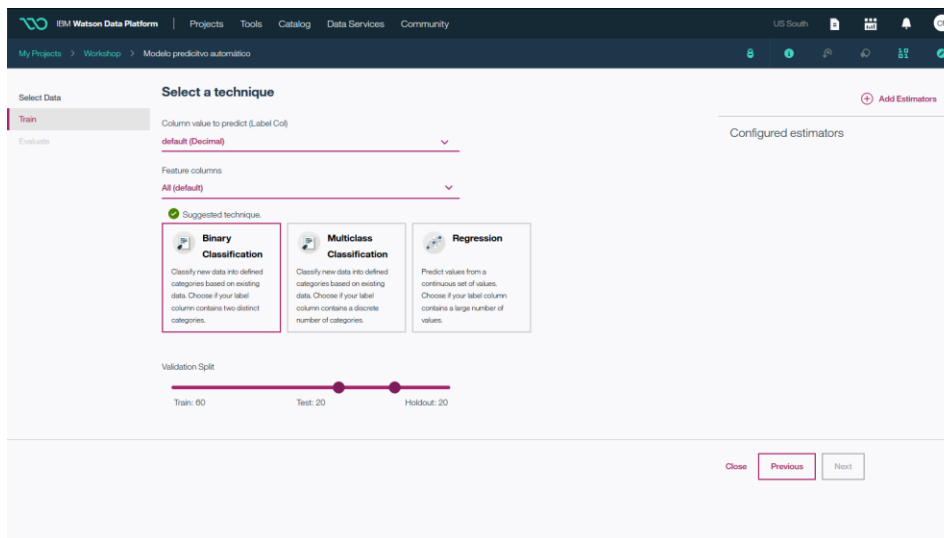
Manual
Let me prepare my data and select which models to train

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

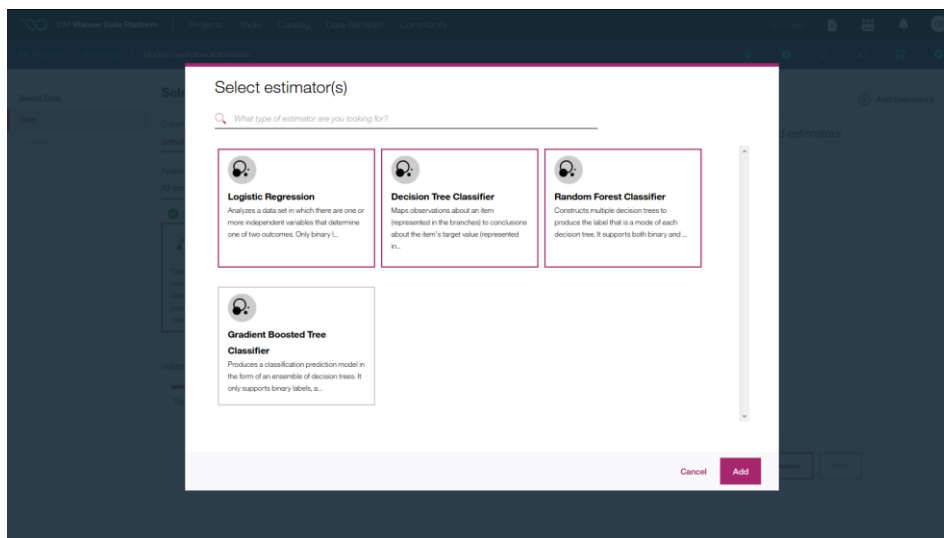
Cancel Create

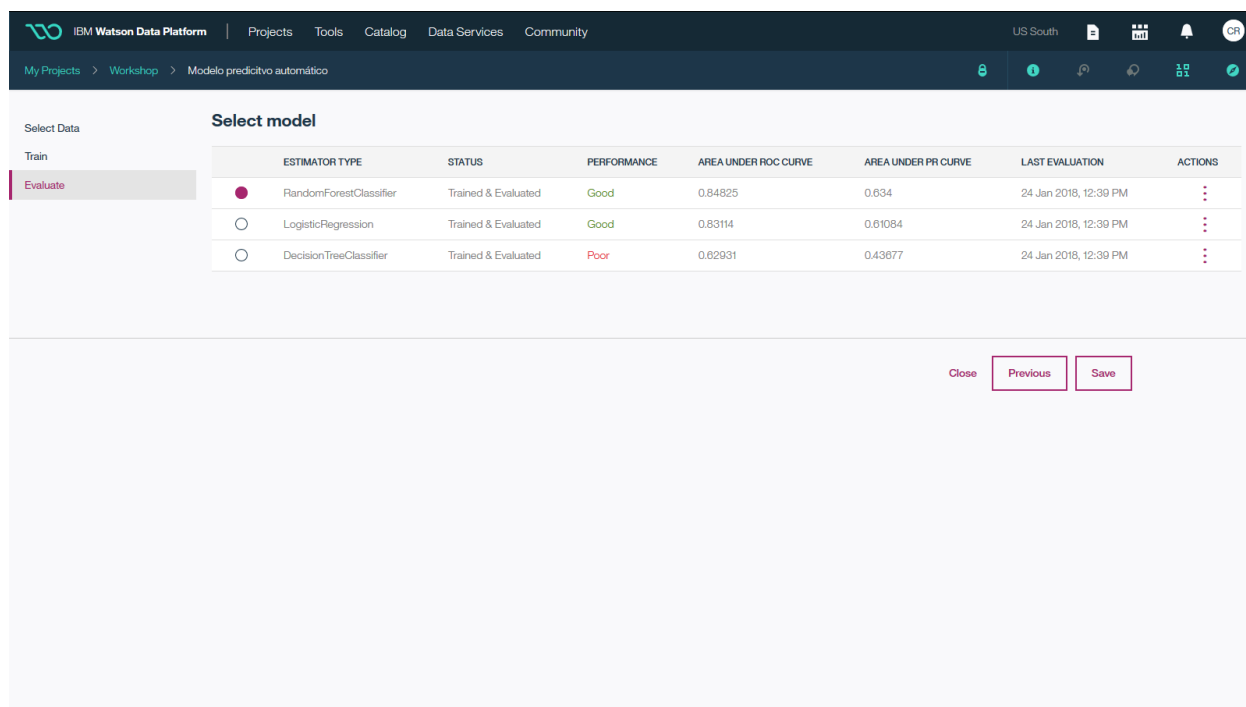
Igual que antes, seleccionamos el fichero de datos de banca.

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



Podemos seleccionar uno o varios estimadores. Añadimos y ejecutamos.





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 user is logged in as 'US South'. The breadcrumb trail shows 'My Projects > Workshop > Modelo predictivo automático'. The left sidebar has 'Select Data' and 'Train' tabs, with 'Evaluate' selected. The main area is titled 'Select model' and displays a table of three estimators.

ESTIMATOR TYPE	STATUS	PERFORMANCE	AREA UNDER ROC CURVE	AREA UNDER PR CURVE	LAST EVALUATION	ACTIONS
<input checked="" type="radio"/> RandomForestClassifier	Trained & Evaluated	Good	0.84825	0.634	24 Jan 2018, 12:39 PM	⋮
<input type="radio"/> LogisticRegression	Trained & Evaluated	Good	0.83114	0.61084	24 Jan 2018, 12:39 PM	⋮
<input type="radio"/> DecisionTreeClassifier	Trained & Evaluated	Poor	0.62931	0.43677	24 Jan 2018, 12:39 PM	⋮

At the bottom right of the dialog, there are three buttons: 'Close', 'Previous', and 'Save'.

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

5. 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.