

Casos prácticos de aplicación.



Caso práctico



[@casfatesvano \(CC BY-SA\)](#)

Eva ha logrado terminar su formación en Inteligencia Artificial y está buscando ofertas de empleo relacionadas con el desarrollo de modelos de Aprendizaje Automático.

Para mejorar sus Currículu y oportunidades de cara a posibles entrevistas de empleo ha decidido desarrollar varios modelos de aprendizaje automático con alguna de las herramientas online gratuitas que existen en Internet, para así poder utilizar como ejemplo en el caso de que en las entrevistas le pidan que demuestre de manera práctica lo que sabe.

En concreto va a utilizar la herramienta BigML que tiene una versión gratuita para entrenamientos con bases de datos pequeñas, pero que también tiene versiones de pago que están siendo utilizadas por empresas e instituciones para desarrollos complejos y reales.

A lo largo de esta unidad vas a ver varios casos prácticos de aplicación del aprendizaje automático a través de la herramienta **BigML**. En concreto vamos a hacer tres ejercicios guiados usando:

- ✓ Árbol de decisión.
- ✓ Clustering.
- ✓ Redes Neuronales.

La tarea final de esta unidad consistirá, por tanto, en que desarrolles por tu cuenta un cuarto ejercicio de aprendizaje automático.

**Materiales formativos de FP Online propiedad del Ministerio de
Educación y Formación Profesional.**

[Aviso Legal](#)

1.- Árbol de decisión.



Caso práctico

Eva ha encontrado una base de datos sobre vinos, con información sobre la composición de cada vino y un valor de calidad asignado a cada uno. Decide hacer su primer entrenamiento de aprendizaje automático supervisado con esta base de datos, definiendo como campo objetivo la calidad. Y como ni es una experta en vinos ni conoce de antemano que exista alguna correlación entre alguno de los datos y la calidad del vino, decide utilizar el algoritmo de Árbol de decisión.



[@Casfatesvano \(CC BY-SA\)](#)

Introducción a BigML

BigML es una plataforma online que permite plantear y resolver problemas de aprendizaje automático (supervisado y no supervisado) con los principales algoritmos que conocemos de esta técnica de Inteligencia Artificial. Incorpora herramientas útiles que facilitan analizar los datos de las bases de datos con las que se vaya a trabajar, así como hacer una evaluación de los modelos una vez finaliza su entrenamiento.

Para principiantes o para problemas con bases de datos no muy grandes tiene una versión gratuita, por lo que resulta especialmente interesante en nuestro caso como estudiantes, para hacer ejercicios prácticos y entender de manera práctica cómo funciona el aprendizaje automático.

Haz clic sobre la imagen o busca la web de BigML.com en tu navegador:

[Acceso a BigML](#)

The screenshot shows the BigML homepage. At the top, there's a navigation bar with links for PRODUCT, GETTING STARTED, PRICING, and SUPPORT. The main headline reads "Machine Learning made beautifully simple for everyone". Below it, a sub-headline says "Take your business to the next level with the leading Machine Learning platform". There are sign-up options for Amazon, GitHub, GitLab, and Google, followed by a "SIGN UP" button. A note below the sign-up form states: "By signing up, you agree to our terms of service and privacy policy. We'll occasionally send you BigML related emails." A circular callout on the right side contains the text "Clic en la imagen para acceder".

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

En el apartado "Pricing" puedes comprobar, como ya te hemos explicado, que tienes disponible una versión gratuita para siempre, limitada a bases de datos de hasta 16 MB:

The screenshot shows the "PRICING" section of the BigML website. The main heading is "Machine Learning for Everyone". A sub-section titled "FREE FOREVER!" offers "Perform unlimited tasks for datasets up to 16MB. No limited trials and no credit card required. SIGN UP now!". A circular callout on the right side contains the text "Bases de datos de hasta 16 MB". Below this, there are sections for "SUBSCRIPTION PLANS", "PRIVATE DEPLOYMENTS", "SUPPORT & ASSISTANCE", "PERSONALIZED TRAINING", and "BigML CERTIFICATIONS".

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

En el apartado "Getting Started" puedes encontrar diversos vídeos tutoriales que explican desde lo más básico hasta las herramientas más sofisticadas de esta plataforma. Es recomendable revisar estos vídeos a medida que vayas terminando esta lección, para afianzar conocimientos y descubrir más utilidades:

Tutoriales disponibles de BigML



The landing page for BigML Education Videos features a dark background with a green graduation cap icon at the top. Below it, the text "BigML Education Videos" is displayed. A paragraph explains that BigML offers a wide variety of basic Machine Learning resources that can be composed together to solve complex Machine Learning tasks. It mentions access via the BigML Dashboard, REST API, and libraries/tools, and includes introductory videos for the Dashboard, Sources, and Datasets. A call to action encourages users to provide feedback.

INTRODUCTION

June 2017 / 8:03 min

SOURCES

June 2017 / 16:55 min

DATASETS

June 2017 / 36:34 min

Take a brief tour of the BigML interface. Learn how to work with resources and navigate the BigML Dashboard.

Sources are the first step of any BigML workflow. Learn the basic features of BigML Sources, including file formats and upload options, or advanced parsing configuration.

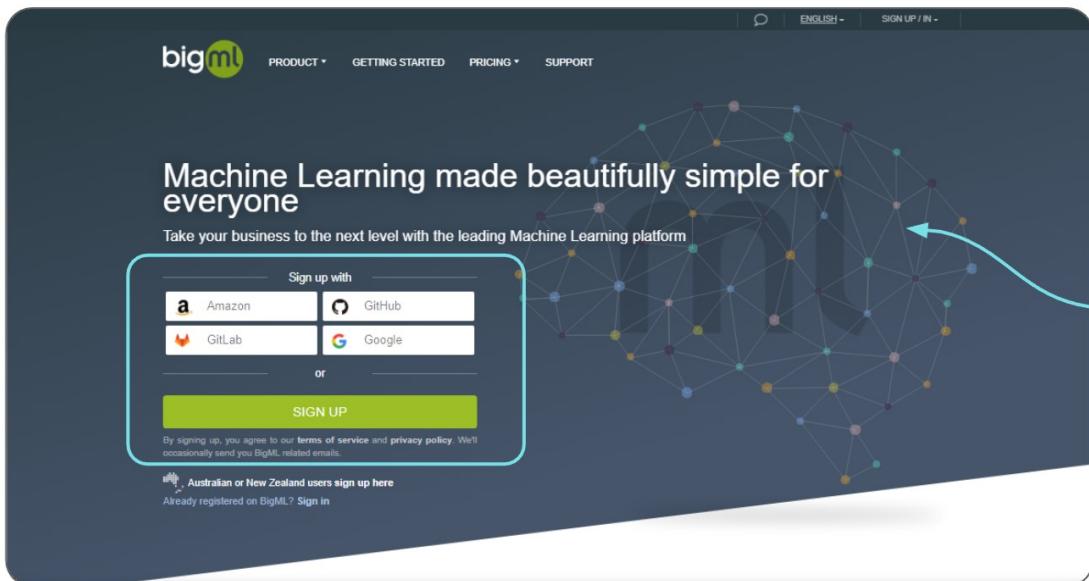
Datasets are the fundamental building block for your BigML workflows. Learn how to filter, sample, add new fields, or split a dataset into training and test datasets.

En el
menú
principal
*Getting
Started*

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Para poder utilizar la plataforma BigML **debes crearte una cuenta de usuario**, bien con alguna cuenta que ya poseas en las plataformas que se ven en la imagen, bien con un correo electrónico y contraseña.

Creación de Usuario en BigML



The sign-up page for BigML features a dark blue header with the BigML logo and navigation links for PRODUCT, GETTING STARTED, PRICING, and SUPPORT. The main area has a dark blue background with a network graph pattern. The text "Machine Learning made beautifully simple for everyone" is prominently displayed. Below it, a subtext reads "Take your business to the next level with the leading Machine Learning platform". A sign-up form is shown with fields for "Sign up with" (Amazon, GitHub, GitLab, Google) and a "SIGN UP" button. A note below the form states "By signing up, you agree to our terms of service and privacy policy. We'll occasionally send you BigML related emails." At the bottom, there are links for "Australian or New Zealand users sign up here" and "Already registered on BigML? Sign in". A yellow callout bubble on the right side points to the sign-up form with the text "Clic en la imagen para acceder".

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Lo primero que vas a ver cuando te hayas creado tu usuario es el apartado "Sources" (Recursos). Probablemente lo tengas vacío o con algunos archivos de prueba que te haya puesto BigML por defecto.

En esta sección puedes cargar tus propias bases de datos (en formato csv o xls, entre otros posibles tipos de archivos). Pero de momento, para ir más ágiles vamos a utilizar datasets disponibles y ya preparados para trabajar con ellos en BigML sin perder mucho tiempo. Pero no te olvides de esta sección, pues cuando quieras hacer tus propios desarrollos será aquí donde tengas que subir tus archivos de datos.

Recursos disponibles en BigML

The BigML dashboard interface includes a top navigation bar with links for PRODUCT, GETTING STARTED, PRICING, SUPPORT, and a user account (FRANBVG). Below this is a header for 'FRANBVG - My Dashboard' and 'BigML Intro Project'. The main content area is divided into sections: Sources, Datasets, Supervised, Unsupervised, Predictions, and Tasks. The 'Sources' section is currently active, showing a list of 10 datasets. The 'Supervised' and 'Unsupervised' tabs are also visible. At the bottom of the page is a footer with links for COMPANY (About, Blog, Brand Style Guide, Contact, Internships, Milestones, Openings, Team), PRODUCT (API, Documentation, Features, Labs, Organizations, Private Deployments, Releases, Tools, WhizzML), BUSINESS (Customers, Events, Newsletters, Partners, Pricing), TRAINING (Certifications, Education, ML Schools), and GALLERY (Datasets, Models, WhizzML). A search bar and a 'Show [10] sources' button are located at the bottom left. The footer also contains links for Terms of Service, Privacy Policy, and FAQ, along with the BigML logo.

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Busca, en la parte más baja de la pantalla (en el "footer" de la web), el apartado "Datasets"

Cómo obtener Datasets en BigML

The BigML dashboard interface includes a top navigation bar with links for PRODUCT, GETTING STARTED, PRICING, SUPPORT, and a user account (FRANBVG). Below this is a header for 'FRANBVG - My Dashboard' and 'BigML Intro Project'. The main content area is divided into sections: Sources, Datasets, Supervised, Unsupervised, Predictions, and Tasks. The 'Sources' section is currently active, showing a list of 10 datasets. The 'Supervised' and 'Unsupervised' tabs are also visible. At the bottom of the page is a footer with links for COMPANY (About, Blog, Brand Style Guide, Contact, Internships, Milestones, Openings, Team), PRODUCT (API, Documentation, Features, Labs, Organizations, Private Deployments, Releases, Tools, WhizzML), BUSINESS (Customers, Events, Newsletters, Partners, Pricing), TRAINING (Certifications, Education, ML Schools), and GALLERY (Datasets, Models, WhizzML). A search bar and a 'Show [10] sources' button are located at the bottom left. The footer also contains links for Terms of Service, Privacy Policy, and FAQ, along with the BigML logo. A large green circle with the text 'Abajo del todo, acceder a Datasets' has an arrow pointing to the 'Datasets' link in the footer.

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

En el apartado de Datasets vas a poder encontrar un buen número de bases de datos optimizadas para trabajar con ellas en BigML. Han sido aportadas por diferentes usuarios de la plataforma, por lo que... igual que en Yotube te puedes encontrar con contenidos geniales y bien hechos, o con trabajos de muy poca calidad o que tienen que ver poco con lo que buscas.

Este ejemplo lo vamos a hacer con un dataset que se llama "Red Wine Quality" (Calidad del vino rojo). Así que haz esta búsqueda (normalmente con la palabra "wine" ya sale):

Ejemplo de Dataset: red wine quality

The screenshot shows the BigML interface with a search bar at the top containing the word "wine". Below the search bar, there are three dataset cards. The first card, highlighted with a blue border, is titled "Red Wine quality" and is described as being related to red and white variants of Portuguese "Vinho Verde" wine. It has 1599 instances and 12 fields. The second card is titled "Portuguese Red Wine Quality" and has 1599 instances and 12 fields. The third card is titled "Sensory's dataset" and has 576 instances and 12 fields. Each card includes a histogram and some basic statistics.

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

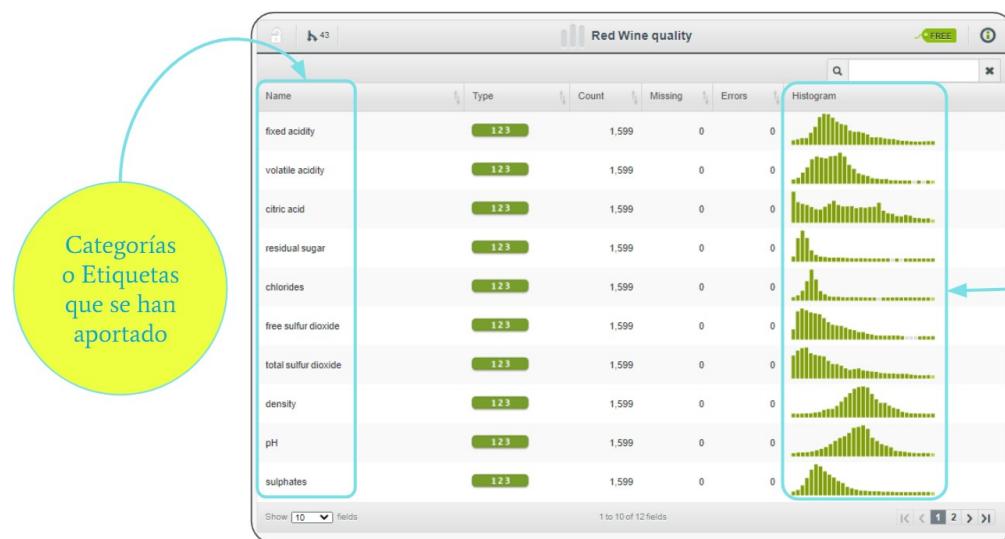
Hacer
búsqueda
con **wine**

Al hacer clic sobre el dataset accedemos a una previsualización de sus datos, en este caso vemos a la izquierda categorías o etiquetas relacionadas con las sustancias del vino: pH, sulfatos, acidez, ácido cítrico, azúcar residual, etc.

También podemos ver qué tipo de categoría es (en este caso la mayoría son numéricos), la cantidad de datos de cada categoría, si hay algún dato perdido o erróneo, y a la derecha del todo vemos un histograma con la distribución de dichos datos.

Como estamos utilizando un dataset "de los buenos" no hay ni datos perdidos ("missing") ni erróneos ("Errors"). Pero en otros casos sí puede que los haya...

Previsualización del Dataset en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Categorías
o Etiquetas
que se han
aportado

Revisando por
encima los
histogramas
podemos
detectar a
simple vista
algún posible
dato erróneo

Como el dataset tiene buena pinta (aunque no tiene muchas instancias... 1599 son pocas para lo que se suele trabajar en aprendizaje automático) vamos a "comprarla". Tranquilidad, que no hay que pagar nada:

Adquirir un Dataset en BigML

The screenshot shows the BigML interface with the 'Red Wine quality' dataset selected. A modal window titled 'Clone this dataset' is open, prompting the user to 'Clone the dataset Red Wine quality in' their account ('FRANBVG - My Dashboard') and project ('BigML Intro Project'). A message in the modal states: 'This dataset was already cloned into that project 1 time'. A large green button labeled 'Clone' is at the bottom right of the modal. The background shows histograms for each feature: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, and sulphates. A 'FREE' badge is visible in the top right corner of the main interface.

Darle aquí para
“comprar” el
data set

Fran Bartolomé - Elaboración propia (CC BY-SA)

Te va a llevar directamente al apartado "Datasets", en el que como mínimo debería aparecerte este que acabas de "comprar". A medida que vayas trabajando con otros datasets te irán apareciendo aquí. Podrás hacer diversos entrenamientos a partir del mismo dataset (no hace falta que "compres" el mismo dataset varias veces si quieras hacer varios entrenamientos).

Dataset ya adquirido y disponible en tu cuenta de BigML

The screenshot shows the BigML dashboard under the 'FRANBVG - My Dashboard' account. The 'Datasets' tab is selected. A list of datasets is shown, with the 'Red Wine quality' dataset highlighted. This dataset has 1599 instances and 12 fields. Below it, several other datasets are listed, including training and testing splits of the Red Wine quality dataset, and datasets for Diabetes Diagnosis and Heart Disease. The interface includes a toolbar with various icons for managing datasets.

En tu sección
de Datasets
puedes acceder
siempre que
quieras a
RedWine

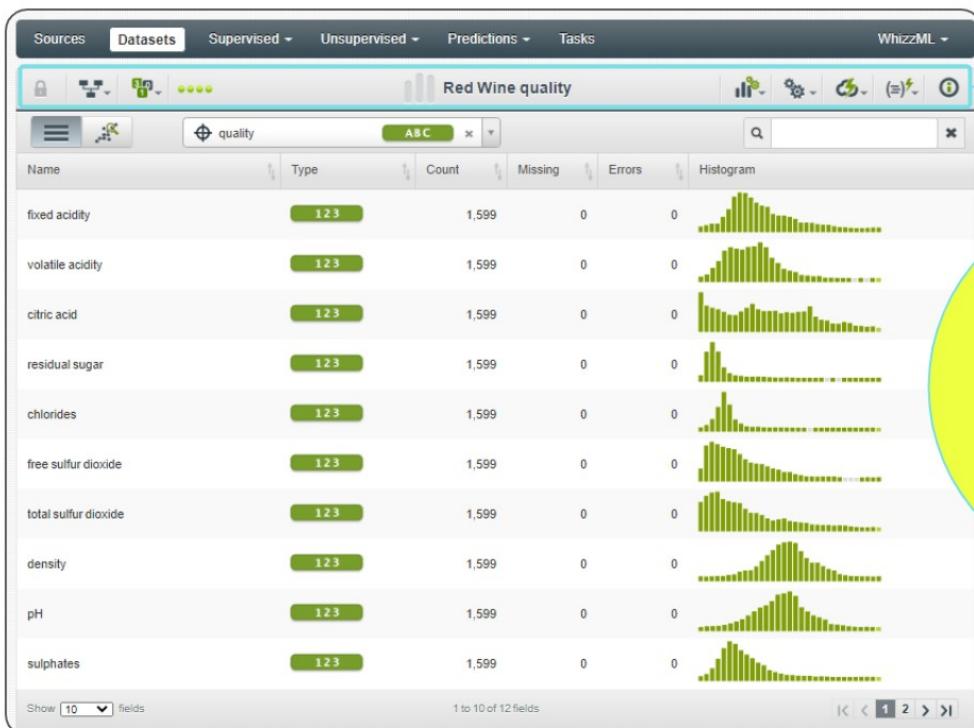
Fran Bartolomé - Elaboración propia (CC BY-SA)

Preparación de datos:

Antes de proceder al entrenamiento propiamente dicho, vamos a revisar un poco mejor nuestros datos. Recuerda que ya hemos explicado en las unidades anteriores la importancia que tiene la fase previa a los entrenamientos revisando los datos y asegurándonos que tenemos los justos y necesarios para que nuestra máquina ni trabaje en balde ni le falten partes importantes.

Vamos a revisar los datos de nuestro dataset "Red Wine quality".

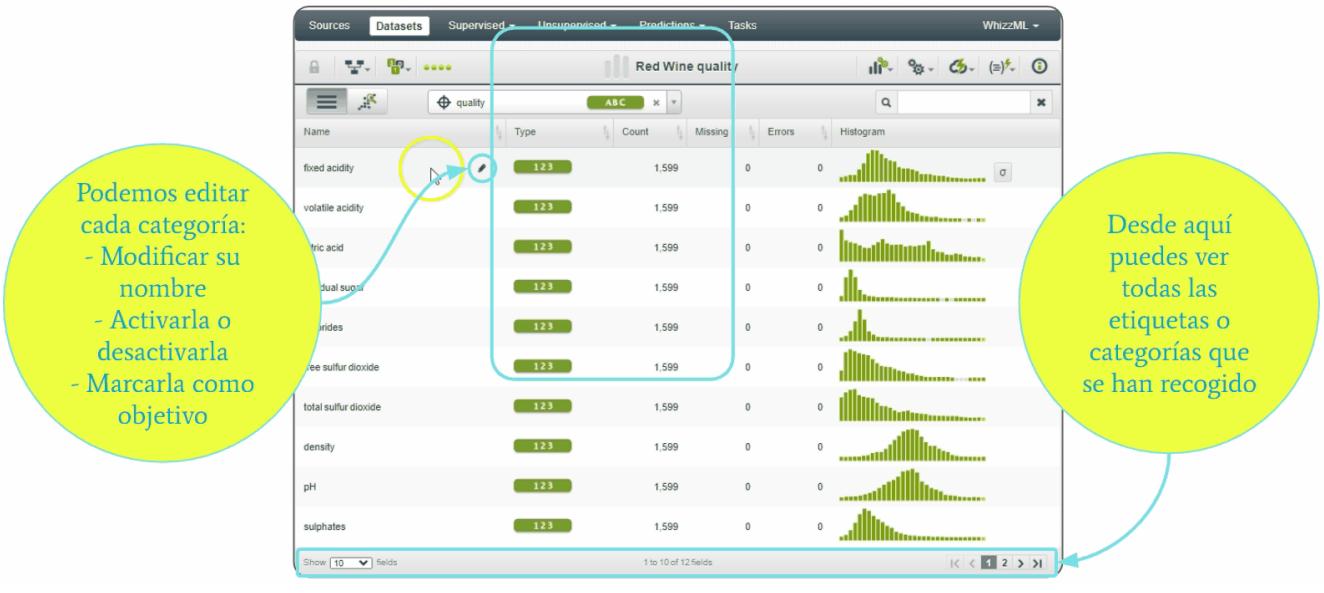
Aspecto del dataset en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Podemos editar el nombre de cada categoría (¿te animas a traducirlas?). En algunas ocasiones nos vamos a encontrar con categorías que no aportan nada al ejercicio de Inteligencia Artificial, y en esos casos es preferible desactivarlas para que no gasten recursos en el entrenamiento. También, para aprendizajes supervisados tendremos que definir qué categoría es nuestro objetivo. Estas dos acciones se editan junto al nombre:

Editar categorías de un dataset en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Definición del campo objetivo para aprendizaje supervisado



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Podemos hacer repaso de relación entre las diferentes categorías de nuestro dataset pasando a la vista Scatterplot, para confrontar dos de ellas y ver las gráficas que generan. En ejercicios complejos, nos va a permitir ya intuir categorías que tienen relación proporcional directa o inversa, o aquellas que aporten prácticamente la misma información y que por tanto pudiéramos prescindir de una de las dos, etc.

Revisión de datos previa al entrenamiento en BigML



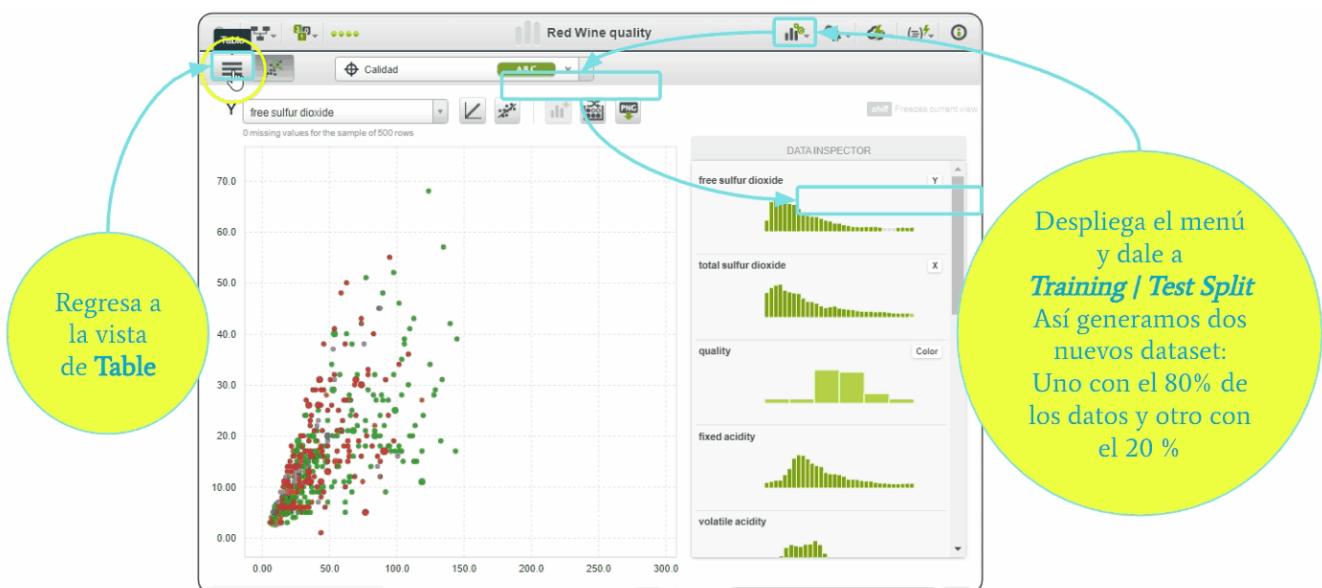
Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Entrenamiento:

¡Por fin! Ya ha llegado el momento. Vamos a iniciar el entrenamiento, pero como probablemente la mayoría de los que trabajemos con este dataset no tendremos ni idea de vinos ni de sus propiedades, vamos a necesitar "guardarnos" una parte de los datos para luego poder hacer comprobaciones que demuestren si el modelo resultante es eficaz o no.

Así que vamos a partir en dos partes el dataset. Nos vamos a quedar con el 80% de los datos para hacer el entrenamiento, y vamos a guardar el 20% de los mismos para luego hacer comprobaciones.

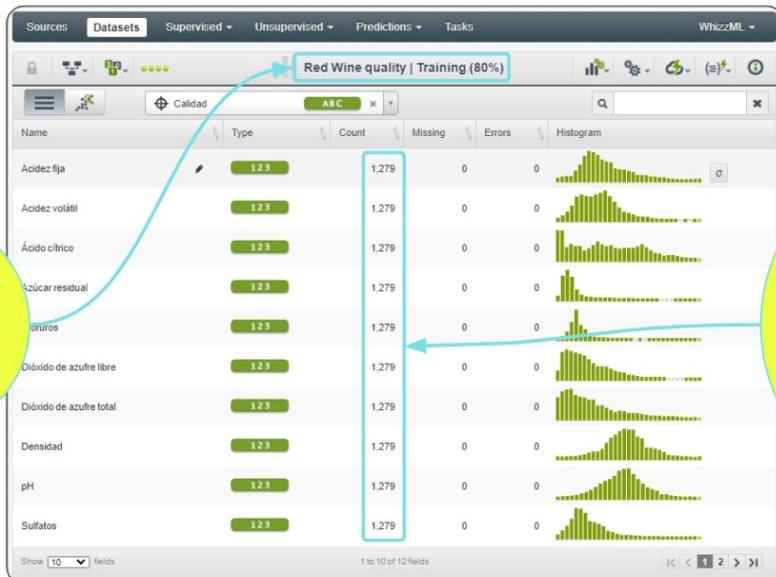
Partición del dataset en 80/20 para entrenar y luego comprobar



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Te llevará directamente a la vista del dataset con el 80% de los datos.

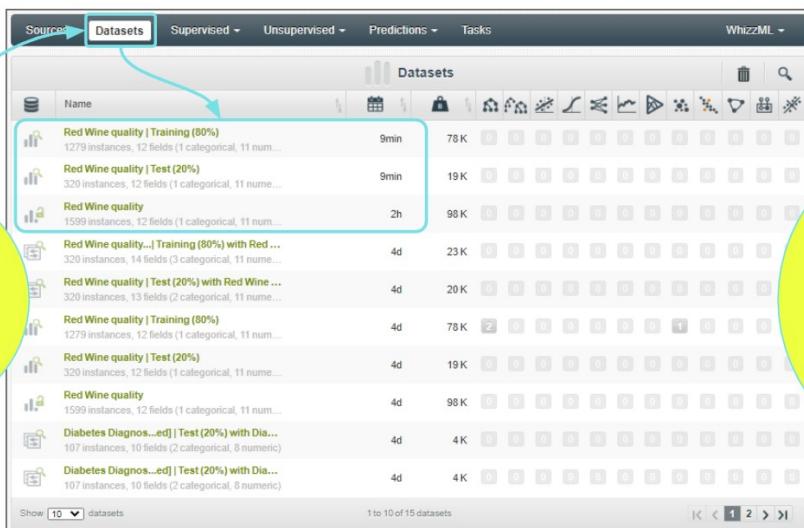
Nuevo dataset con el 80% de los datos iniciales



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Puedes comprobar que se han generado dos nuevos datasets mirando en tu apartado de datasets. Recuerda que estos datasets se quedarán aquí para poder usarlos tantas veces quieras:

Comprobación en tu lista de datasets

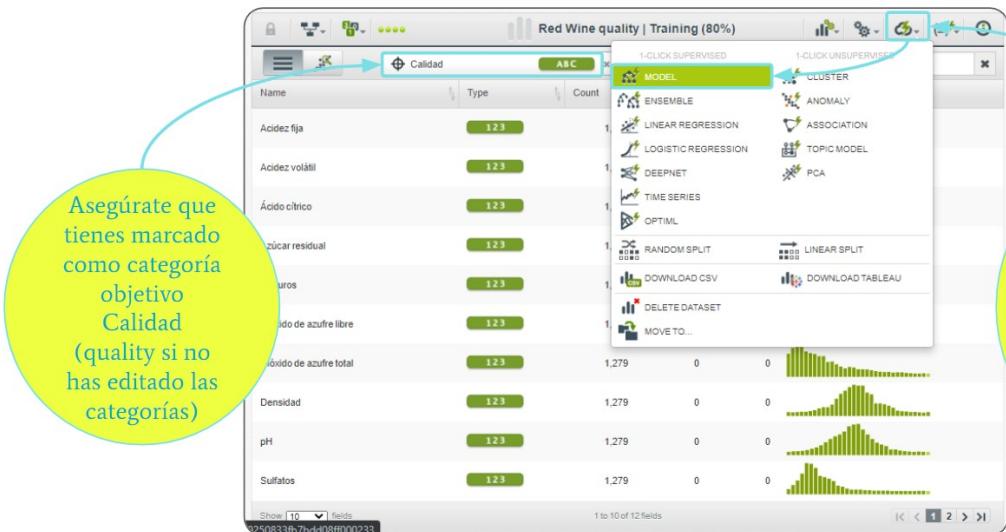


Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Volvemos a entrar en el dataset con el 80% de los datos, y sin más preámbulos ¡vamos a entrenar! En este ejemplo vamos a utilizar el algoritmo de Árbol de Decisión, que si recuerdas lo tratado en unidades anteriores es como el comodín de los algoritmos... En este caso en el que queremos predecir la calidad del vino, y no vemos con claridad que se pudiera utilizar otro algoritmo más específico, el árbol de decisión es la mejor opción.

En BigML llaman "model" al algoritmo de árbol de decisión:

Entrenamiento con algoritmo de Árbol de Decisión (Model) en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Modelo entrenado:

¡Y ya estaría! ¿Ha tardado mucho? Con tan pocos datos (1279, después de haber hecho la partición del 80%), el modelo se entrena en pocos segundos.

Lo que vemos ahora es la representación gráfica del modelo. Evidentemente, con forma de árbol de decisión.

¡Ojo! Tu modelo no tiene por qué parecerse al de la siguiente imagen. Al haber partir la muestra de datos entre 80/20 los datos que haya tomado en tu caso y en el de este ejemplo no tienen por qué ser los mismos, por lo que a partir de ahora las imágenes que se muestren probablemente no coincidan con lo que veas en tu ejercicio.

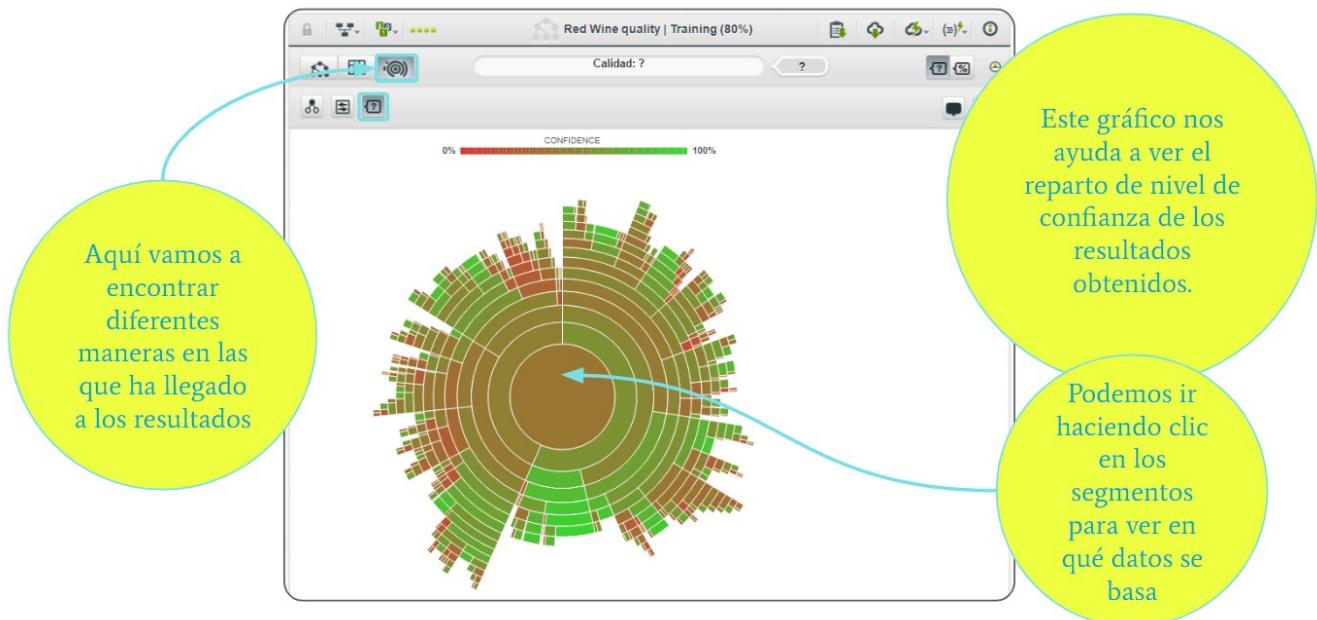
Modelo ya entrenado con árbol de decisión en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Podemos hacer un recorrido por las diferentes maneras de presentar los datos, para entender mejor qué hemos obtenido. Por ejemplo, podemos ver el conjunto de resultados o de caminos que ha seguido, relacionándolos con la confiabilidad de que el resultado sea fiable:

Vista Confianza del árbol de decisión en BigIML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

O también, para este tipo de algoritmo, podemos ver qué valores de calidad, de los posibles, ha encontrado para cada caso. Por ejemplo, en este caso hay una predominancia del resultado azul (calidad=5), seguido del resultado naranja (calidad = 6), etc.

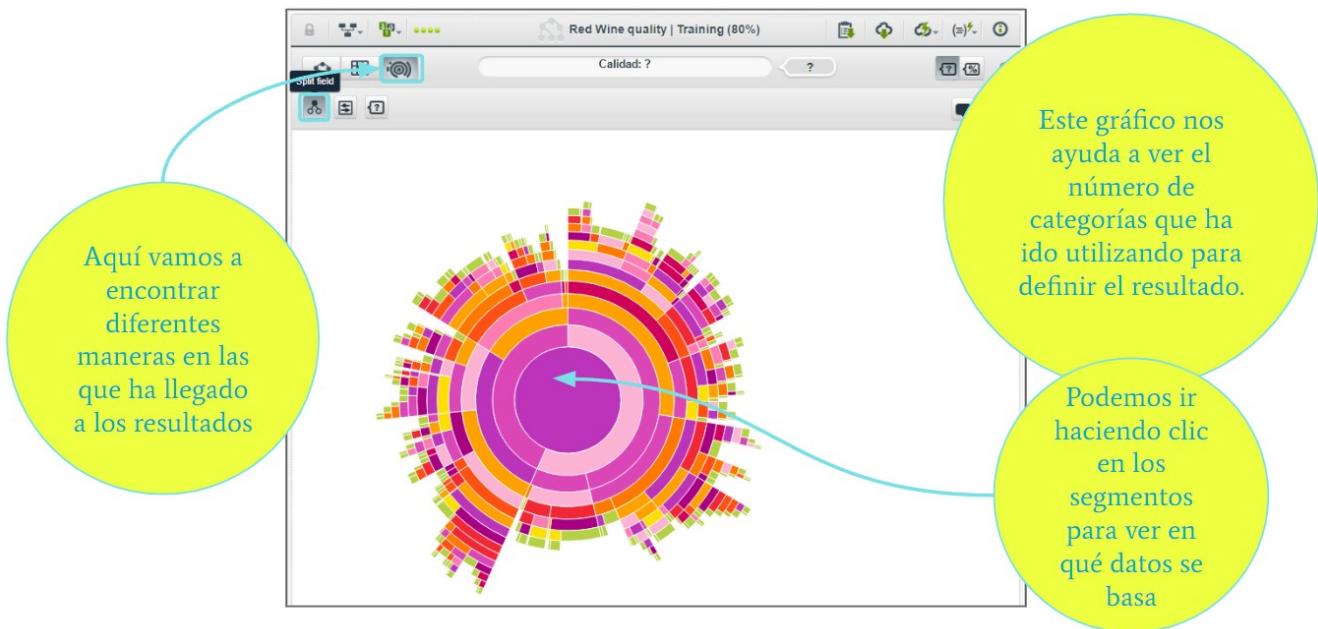
Vista Predicción del árbol de decisión en BigIML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

O el peso que han tenido las diferentes categorías:

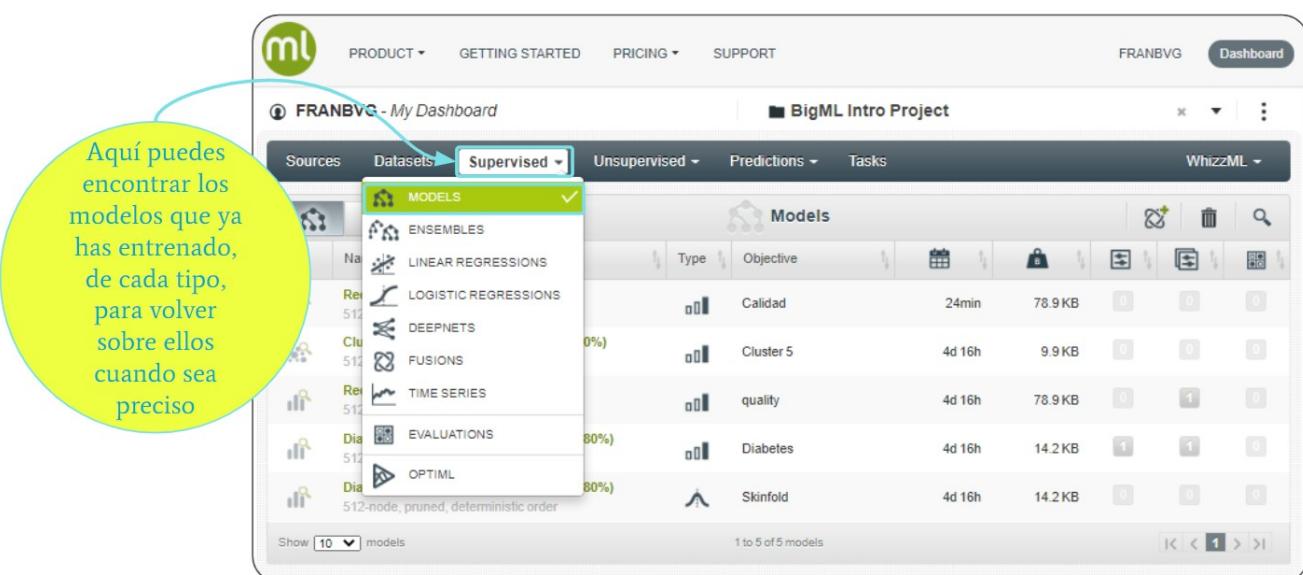
Vista Split Field del árbol de decisión en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Este resultado, el modelo entrenado, a partir de ahora se quedará guardado en tu apartado de modelos supervisados:

Comprobación en tu lista de modelos supervisados



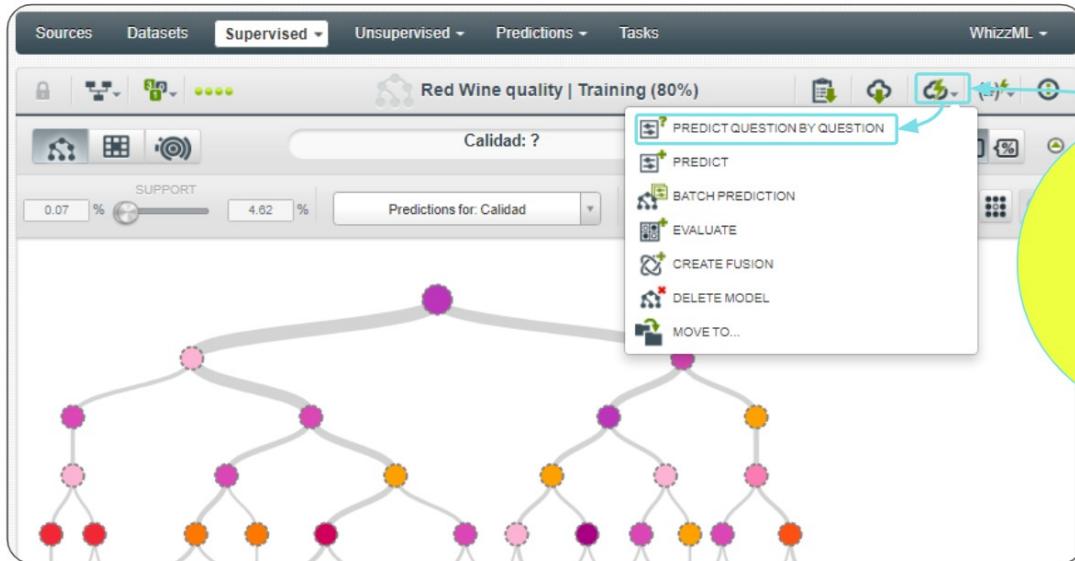
Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Uso del modelo:

A partir de ahora podemos utilizar este modelo entrenado para predecir la calidad de un vino según su composición. En BigML vamos a encontrar tres formas de introducir los datos de nuevos casos:

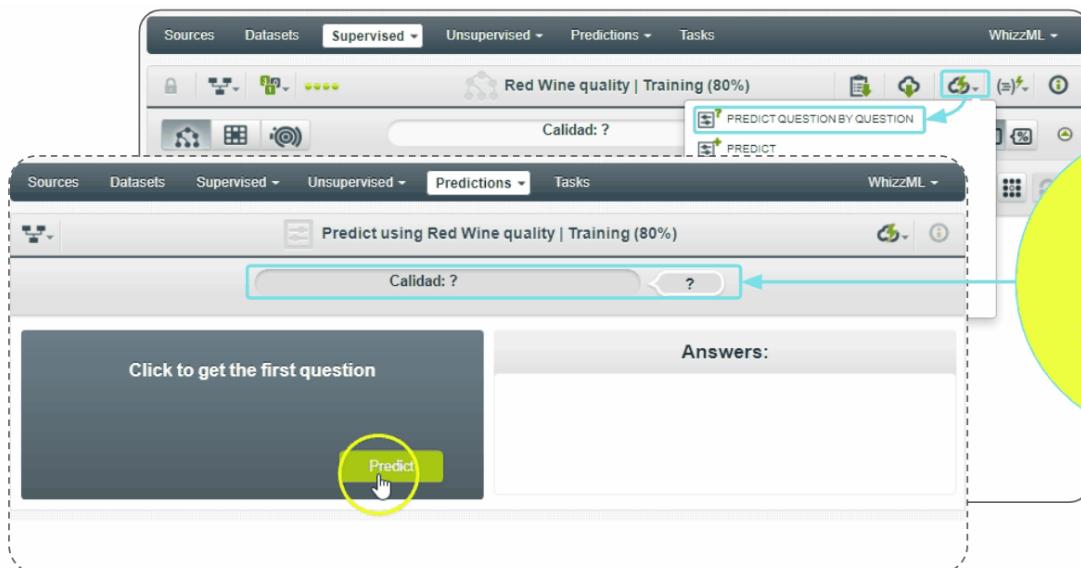
- ✓ Dato a dato.
- ✓ Por instancias.
- ✓ Base de datos con varias instancias a la vez.

Nueva predicción con modelo entrenado, dato a dato.



Puedes probar el algoritmo de predicción de calidad introduciendo dato a dato

Introducción de datos para nueva predicción



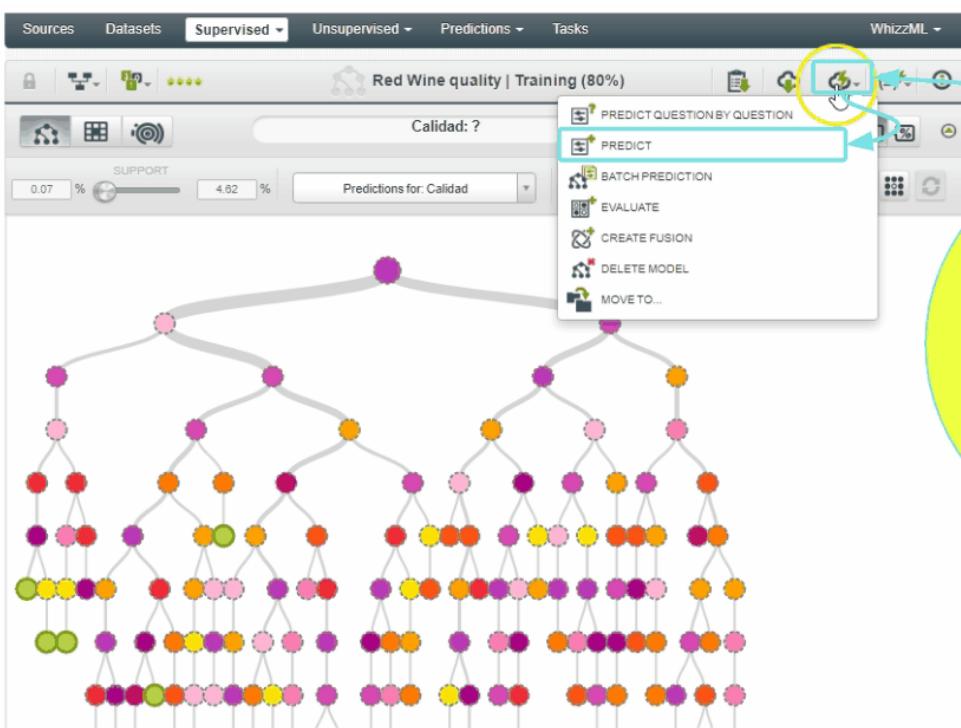
Según los parámetros que indiques para diferentes categorías, te indica la calidad del vino y el % de confiabilidad

Nueva predicción con modelo entrenado, por instancia.



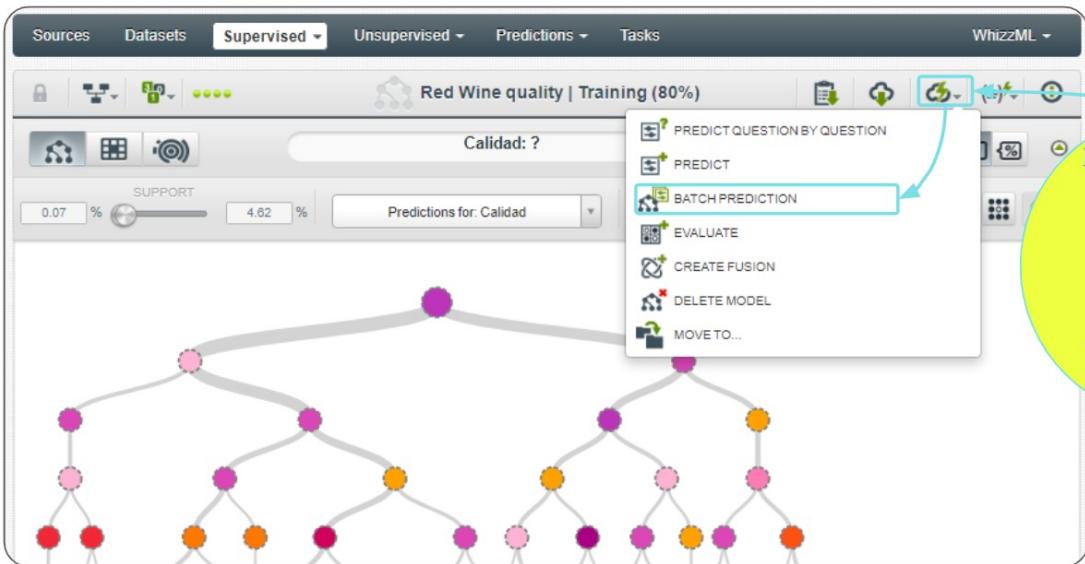
Puedes probar el algoritmo de predicción de calidad definiendo todos los parámetros a la vez

Nueva predicción con modelo entrenado, por instancia.



Puedes probar el algoritmo de predicción de calidad definiendo todos los parámetros a la vez

Nueva predicción con modelo entrenado, con base de datos



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Y es aquí donde vamos a utilizar ese 20% de datos que nos hemos guardado antes. Vamos a probar el modelo entrenado por el 80% de datos reales, con este otro 20% de casos que también son reales. Y como en realidad el dato real de calidad lo conocemos, vamos a poder compararlo con el dato pronosticado por el modelo:

Nueva predicción con modelo entrenado, con base de datos

Aquí ponemos el dataset que se ha creado antes, con el 20% de los datos

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Cómo descargar csv de los resultados del entrenamiento con árbol de decisión en BigML

Nos ofrece una previsualización de los resultados, según categorías, y con la Calidad en la última columna

Podemos descargar un archivo .csv o generar un nuevo dataset para trabajar con él en bigML

The screenshot shows the bigML interface with the following details:

- Top Bar:** Sources, Datasets, Supervised, Unsupervised, Predictions, Tasks, WhizzML.
- Title:** Red Wine quality | Test (20%) with Red Wine quality | T...
- Dataset Preview:** Red Wine Quality | Training (80%) and Red Wine Quality | Test (20%).
- Configuration:** Shows the output preview of the dataset.
- Output preview Table:**

Acidez volátil	Ácido cítrico	Azúcar residual	Cloruros	Dióxido de azufre libre	Dióxido de azufre total	Densidad	pH	Sulfatos	Alcohol	Calidad
0.66	0.0	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
0.5	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.8	10.5	5
0.41	0.24	1.8	0.08	4.0	11.0	0.9962	3.28	0.59	9.5	5
0.43	0.21	1.6	0.106	10.0	37.0	0.9966	3.17	0.91	9.5	5
0.645	0.0	5.5	0.086	5.0	18.0	0.9986	3.4	0.55	9.6	6
- Buttons:** Download batch prediction, Output dataset.

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

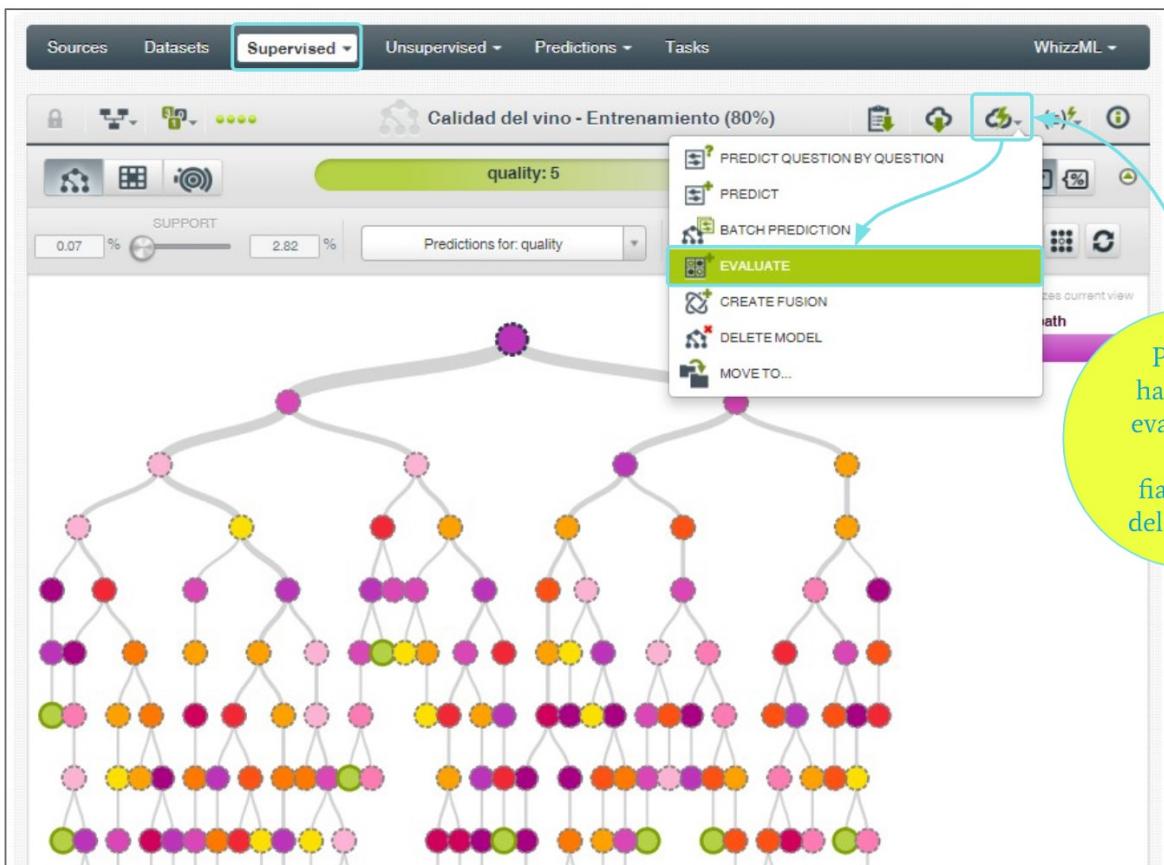
Aquí podemos previsualizar 5 instancias en las que al final podremos ver el dato real de calidad y a la derecha del todo dato predicho por el modelo. Para comparar bien todos los datos podemos descargarnos el archivo en formato csv.

En cualquier caso, BigML también nos permite hacer un análisis del modelo entrenado.

Analizando fiabilidad del modelo entrenado:

Asegúrate de estar dentro del modelo supervisado, y dale a **Evaluate**

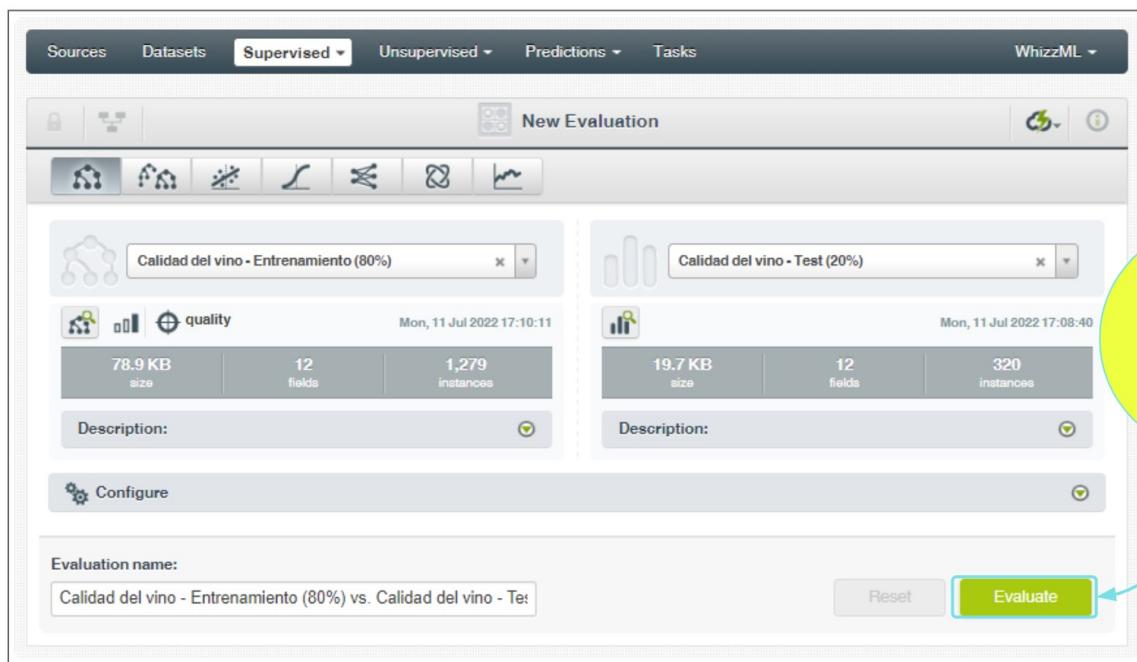
Obtener evaluación del modelo entrenado con árbol de decisión en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Te va a pedir que confirmes que quieres evaluar este modelo con la muestra del 20% de datos originales con el campo "calidad" conocido.

Iniciar análisis del modelo entrenado en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Y así podemos ver datos de confiabilidad, eficacia, etc.

Datos de evaluación del modelo

The screenshot shows the WhizzML interface with the 'Supervised' tab selected. It displays two datasets: 'Calidad Del Vino - Entrenamiento (80%)' and 'Calidad Del Vino - Test (20%)'. A green callout bubble points to the evaluation metrics table, which includes Accuracy (95.00%), F-measure (0.1111), Precision (8.33%), Recall (16.67%), and Phi coefficient (0.094).

Accuracy	F-measure
95.00%	0.1111

Precision	Recall	Phi coefficient
8.33%	16.67%	0.094

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Te de información sobre el grado de exactitud del modelo, su precisión y otros coeficientes

Datos ampliados de la evaluación del modelo

The screenshot shows the WhizzML interface with the 'Supervised' tab selected. It displays two datasets: 'Calidad Del Vino - Entrenamiento (80%)' and 'Calidad Del Vino - Test (20%)'. A green callout bubble points to the detailed metrics table, which compares MODEL, MODE, and DIFFERENCE for Accuracy, F-measure, Precision, Recall, and Phi coefficient. A button labeled 'Hide metrics comparing with mode-based decision' is visible.

MODEL	MODE	DIFFERENCE
95.00%	98.13%	▼-3.13%

MODEL	MODE	DIFFERENCE
0.1111	0	▲0.1111

MODEL	MODE	DIFFERENCE
8.33%	0.00%	▲8.33%

MODEL	MODE	DIFFERENCE
16.67%	0.00%	▲16.67%

MODEL	MODE	DIFFERENCE
0.094	0	▲0.094

Te de información sobre el grado de exactitud del modelo, su precisión y otros coeficientes

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))



Autoevaluación

¿Por qué dividimos el dataset en dos partes (80% y 20%)?

- Porque no son necesarios tantos datos, así gastamos menos tiempo y recursos de computación.
- Para entrenar el modelo solo con la parte del 80% de los datos y guardar el otro 20% de datos reales para luego poder hacer comprobaciones y medir la confiabilidad del modelo entrenado.
- Porque así funciona mejor el algoritmo de Árbol de Decisión.
- Para hacer el entrenamiento del modelo en dos partes y no superar los 16 MB disponibles en la versión gratuita de BigML

[Mostrar retroalimentación](#)

Solución

1. Incorrecto
2. Correcto
3. Incorrecto
4. Incorrecto

2.- Clustering.



Caso práctico



[@casfatesvano \(CC BY-SA\)](#)

Eva ha pensado seguir trabajando con la base de datos de vinos que ha utilizado para hacer su primer entrenamiento con algoritmo de Árbol de Decisión.

Ahora que ya tiene este modelo entrenado con campo objetivo supervisado quiere probar a ver qué encuentra si plantea, para la misma base de datos un entrenamiento no supervisado. En concreto a través de la técnica de clustering.

Cuando tenga ambos entrenamientos finalizados, podrá comparar uno con otro, y quizás, quién sabe, descubrir relaciones

entre los distintos tipos de componentes del vino. O quizás cuál es la mejor combinación para obtener la mejor calidad, y viceversa.

Empezamos este ejercicio asumiendo que ya sabes acceder a BigML, que tienes una cuenta en dicha herramienta y que sabes llegar al apartado de datasets. Si alguno de estos pasos no los recuerdas bien, revisa los primeros pasos descritos en el punto anterior.

Selección del dataset

Vamos a aprovechar el dataset "Red Wine quality". En concreto vamos a buscar el dataset del 80% de los datos, pues igual que en el ejemplo anterior, vamos a hacer una predicción al final del ejercicio y lo adecuado es utilizar datos que no hayan formado parte del entrenamiento.

Asegúrate de entrar en tu sección de datasets, y selecciona el del 80%:

Elección del dataset

The screenshot shows the BigML interface with the 'Datasets' tab selected. A yellow circle on the left contains the text 'Regresa al Dataset del 80% de los datos'. An arrow points from this circle to the 'Datasets' tab in the top navigation bar. The main table lists 17 datasets, with the first one highlighted in blue.

Name	Time	Size	Actions														
Red Wine quality Test (20%) with Red Wine ...	5min	20 K															
Red Wine quality Test (20%) with Red Wine ...	10min	20 K															
Red Wine quality Training (80%)	1h	78 K															
Red Wine quality	1h	19 K															
Red Wine quality	3h	98 K															
Red Wine quality... Training (80%) with Red ...	4d	23 K															
Red Wine quality Test (20%) with Red Wine ...	4d	20 K															
Red Wine quality Training (80%)	4d	78 K															
Red Wine quality Test (20%)	4d	19 K															
Red Wine quality	4d	98 K															
Red Wine quality	1599 instances, 12 fields (1 categorical, 11 num...																
Red Wine quality	320 instances, 12 fields (1 categorical, 11 num...																
Red Wine quality	320 instances, 13 fields (2 categorical, 11 num...																
Red Wine quality	1279 instances, 12 fields (1 categorical, 11 num...																
Red Wine quality	320 instances, 12 fields (1 categorical, 11 num...																
Red Wine quality	1599 instances, 12 fields (1 categorical, 11 num...																

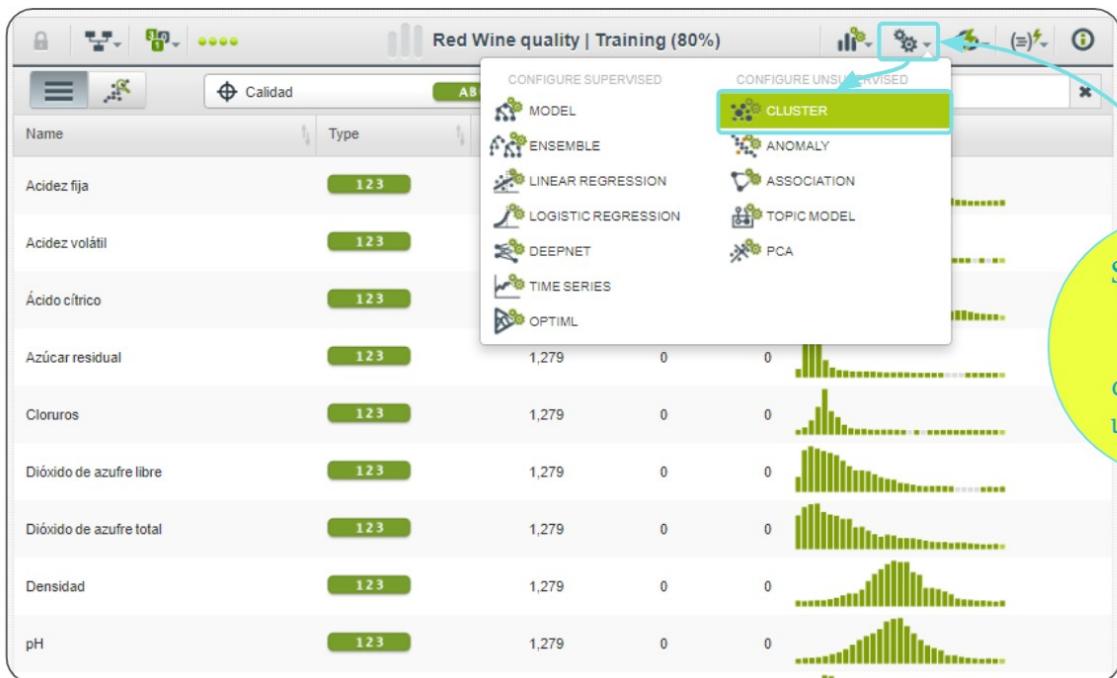
Show 10 datasets 1 to 10 of 17 datasets

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Entrenamiento

En esta ocasión vamos a lanzar el entrenamiento desde el menú que no tiene el rayo. No se comentó en el capítulo anterior por no abrumar con datos. Pero ahora que ya tienes algo de familiaridad con BigML, te contamos que a la hora de lanzar un entrenamiento se puede hacer desde el menú con el rayo (y lo hará inmediatamente con una serie de valores preestablecidos para los parámetros que se pueden definir), o con el menú sin el rayo (que antes de iniciar el entrenamiento nos dará opción a configurar algunos parámetros del algoritmo).

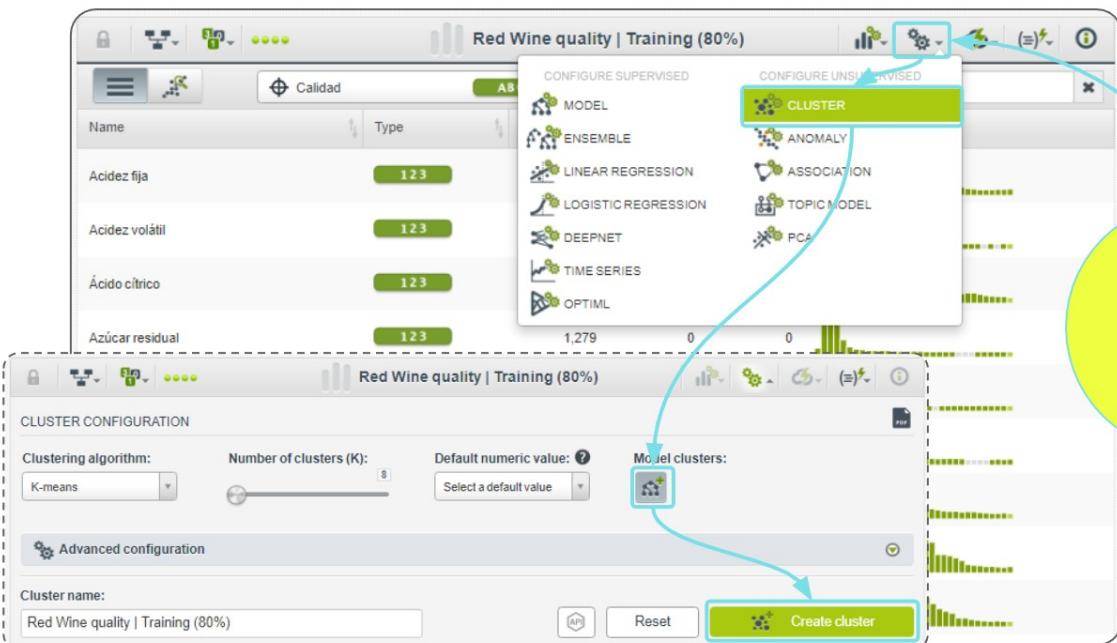
Entrenamiento con clustering



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Podremos decidir si optamos por la variante K-Menas o G-Means (es decir, si somos nosotros quienes definimos el número de clusters a crear o si el propio algoritmo define ese número). En este ejemplo vamos a utilizar K-Means con K=8.

Parámetros del entrenamiento con clustering



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Modelo entrenado:

En este caso BigML nos proporciona una representación gráfica de círculos, que representan a cada cluster, de mayor o menor diámetro en función del número de instancias que formen

parte del mismo.

Hay que mover el ratón y hacer clic sobre dichos círculos para ver información sobre los mismos. En esta imagen se ha hecho clic sobre el cluster número 7, y en la columna derecha podemos ver los valores del centroide. Recuerda que el centroide es un punto o valor que representa el valor medio, predominante o más "cercano" a los verdaderos valores de la muestra que ha servido para hacer el entrenamiento. Eso con cada categoría. De manera que en este ejemplo vemos que los valores del centroide del cluster 7 son 7,97 para la acidez fija, 0,57 para la acidez volátil, etc. También vemos que hay 266 instancias que se han considerado pertenecientes a este cluster.

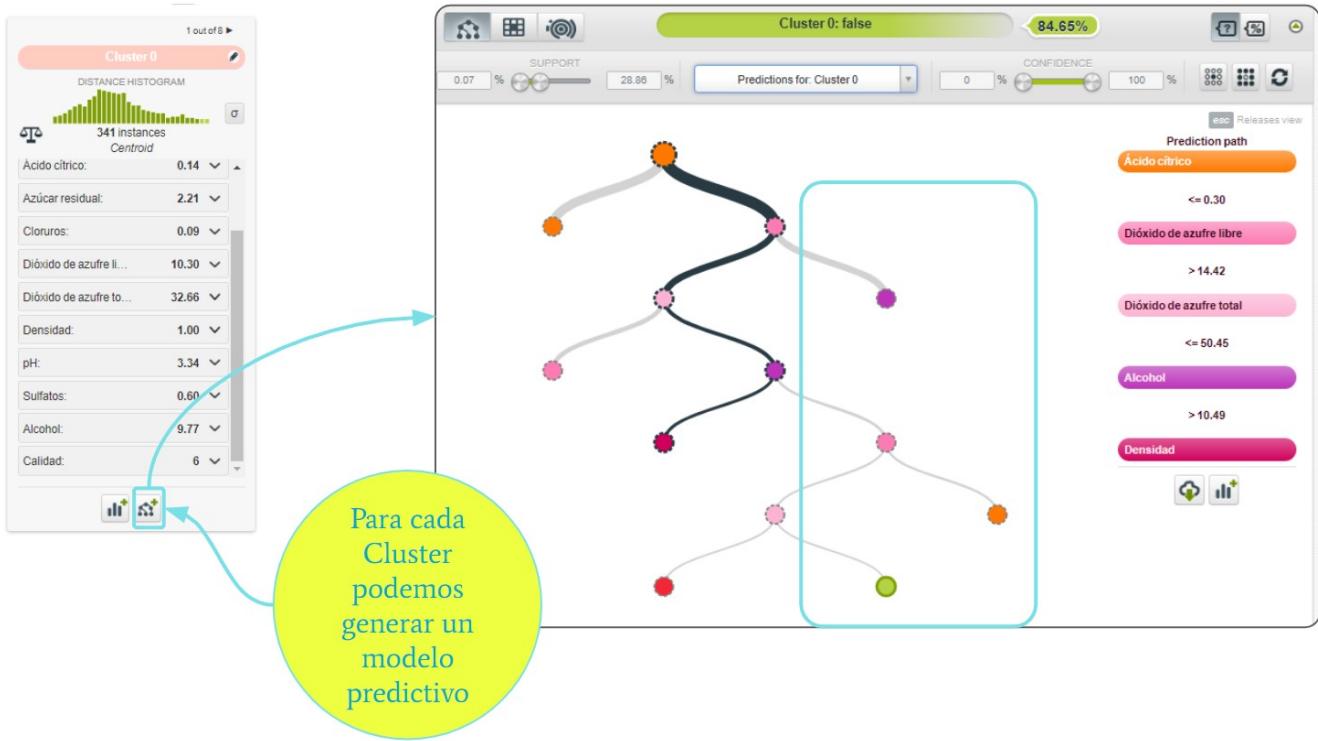
Modelo entrenado por clustering



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Si nos movemos abajo del todo de la columna del cluster 7, veremos la opción de generar un modelo predictivo para este cluster en concreto. Esto nos permite conocer en profundidad qué parámetros y valores han sido más determinantes para generar dichos clústeres. En la medida en que conozcamos el mundo de los vinos y sus composiciones, este análisis cluster por cluster y con sus modelos generativos, nos va a permitir identificar propiedades, asociaciones, dependencias e independencias, etc.

Modelo predictivo para uno de los clusters



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Predicción con el modelo entrenado:

Igual que en el ejercicio del capítulo anterior, dentro de nuestro modelo entrenado buscamos el menú de la nube con el rayo y le damos a "Batch Centroid" para iniciar una predicción a partir de la base de datos del 20% de datos reservados. En este caso el resultado que esperamos es que nos diga a qué cluster pertenece cada instancia de vino.

Predicción con el modelo para base de datos

Sources Datasets Supervised Unsupervised Predictions Tasks WhizzML

New Batch Centroid

Red Wine quality | Training (80%)
Thu, 11 Feb 2021 12:14:39
78.9 KB size 12 fields 1,279 instances 8 clusters
Description:

Red Wine quality | Test (20%)
Thu, 11 Feb 2021 10:34:27
19.7 KB size 12 fields 320 instances
Description:

Configure

Preview of the prediction file (using the type of each field)

Acidez fija	Acidez volátil	Ácido cítrico	Azúcar residual	Cloruros	Dióxido de azufre libre	Dióxido de azufre total	Densidad	pH	Sulfatos	Alcohol
123	123	123	123	123	123	123	123	123	123	123
123	123	123	123	123	123	123	123	123	123	123
123	123	123	123	123	123	123	123	123	123	123
123	123	123	123	123	123	123	123	123	123	123
123	123	123	123	123	123	123	123	123	123	123

Prediction name:
Red Wine quality | Test (20%) with Red Wine quality | Training (80%)

Reset Centroid

Especificamos un dataset que no contenga muestras usadas en el entrenamiento

Y, efectivamente, el modelo entrenado nos da como respuesta el listado de las instancias del 20% definiendo en qué cluster estarían:

Resultado de la predicción

The screenshot shows the WhizzML interface with the following details:

- Top Bar:** Sources, Datasets, Supervised, Unsupervised, Predictions, Tasks, WhizzML.
- Toolbar:** Lock, Filter, Refresh, Red Wine quality | Test (20%) with Red Wine quality | T...
- Left Panel:** Red Wine Quality | Training (80%)
- Right Panel:** Red Wine Quality | Test (20%)
- Output Preview Table:**

plátíl	Ácido cítrico	Azúcar residual	Cloruros	Dióxido de azufre libre	Dióxido de azufre total	Densidad	pH	Sulfatos	Alcohol	Calidad	cluster
0.66	0.0	1.8	0.075		13.0	40.0	0.9978	3.51	0.56	9.4	5 Cluster 0
0.5	0.36	6.1	0.071		17.0	102.0	0.9978	3.35	0.8	10.5	5 Cluster 2
0.41	0.24	1.8	0.08		4.0	11.0	0.9962	3.28	0.59	9.5	5 Cluster 0
0.43	0.21	1.6	0.106		10.0	37.0	0.9966	3.17	0.91	9.5	5 Cluster 0
0.645	0.0	5.5	0.086		5.0	18.0	0.9986	3.4	0.55	9.6	6 Cluster 0
- Buttons:** Download batch centroid, Output dataset.

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Igual que en el ejercicio del punto anterior, podemos descargarnos el archivo de resultados de la predicción en formato csv para revisar dichos resultados de manera más precisa.

En el análisis del resultado del modelo entrenado, podríamos, por ejemplo, revisar cada cluster para localizar aquel que para el parámetro "Calidad" tiene un valor más elevado. Y una vez localizado ese cluster, podríamos revisar los valores del resto de parámetros, para intentar localizar qué tipo de vinos conocidos cumplen esas condiciones, o yendo ya a posibilidades más técnicas, podríamos intentar influir en las condiciones de cultivo de las vides para que las plantas y sus frutos se aproximen lo más posible a esos parámetros. Evidentemente en este punto tendría que intervenir un técnico agrícola que conozca y entienda bien cómo se puede influir en estos parámetros a nivel cultivo...



Autoevaluación

¿Sería posible trabajar en este ejercicio con el algoritmo G-Means en lugar de con el K-Means?

Sí

Sólo si utilizamos cualquier otro algoritmo distinto de clustering.



No

[Mostrar retroalimentación](#)

Solución

1. Correcto
2. Incorrecto
3. Incorrecto

3.- Redes neuronales.

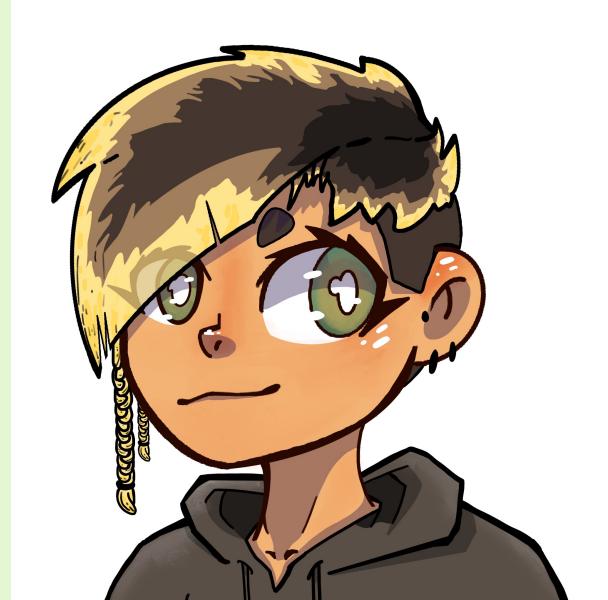


Caso práctico

El último ejercicio que quiere practicar Eva en la plataforma BigML para tener un buen portfolio de ejercicios prácticos de Aprendizaje Automático va a ser con el algoritmo de redes neuronales.

Como es una apasionada de los videojuegos ha pensado que sería buena idea analizar a través de Deep Learning cuál es la mejor manera de jugar a League of Legens. Y de paso descubrir si hay algún tipo de ventaja según la estrategia que desarrolle cada equipo. Espera, incluso, poder determinar si el juego es imparcial o si beneficia más a alguno de los participantes.

Como es lógico, lo primero de todo es hacerse con una base de datos que recoja todos los datos posibles de multitud de partidas, y a partir de ahí preparar un dataset con el que trabajar en BigML.



[@Casfatesvano \(CC BY-SA\)](#)

Vamos a aprovechar este caso práctico para ver cómo podemos incorporar en BigML cualquier base de datos que tengamos u obtengamos por nuestra cuenta. Así que lo primero de todo que tienes que hacer es descargar en tu ordenador el archivo csv que tiene la base de datos con la que vamos a trabajar.

Descárgate el archivo "[Base de datos - League of Legens.csv](#) (csv - 9,29 MB)"

Una vez tengas el archivo en tu ordenador inicia sesión en tu cuenta de BigML (recuerda que tienes descritos en el capítulo 1 de esta unidad el paso a paso para acceder a BigML y entrar en tu usuario).

En el apartado **Sources** abre el menú para cargar un archivo local (**Upload a local file**). Se abrirá una ventana para localizar dicho archivo.

En este caso el archivo ocupa 9,29 MB. Recuerda que la cuenta gratuita no permite trabajar con archivos de más de 16 MB, por lo que cuando vayas a trabajar con otras bases de datos ten en cuenta que si ocupan más no te va a dejar dar el paso siguiente.

[Cargar un nuevo recurso en BigML](#)

The screenshot shows the BigML interface with the 'Sources' tab selected. The toolbar at the top includes options like 'Datasets', 'Supervised', 'Unsupervised', 'Predictions', 'Tasks', and 'WhizzML'. A yellow circle highlights the 'UPLOAD A LOCAL FILE' button in the top right corner of the toolbar. The main area displays a list of sources, including various CSV, XLS, and TSV files, along with their details like type, name, size, and last modified. A yellow circle also highlights the first item in the list: 'Base de datos - League of Legends.csv'.

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Ahora que ya tenemos el archivo csv cargado en el apartado Sources, tenemos que generar el dataset correspondiente. Digamos que BigML necesita "traducir" el archivo que hemos subido nosotros al formato que utiliza internamente. Así que entra dentro del recurso:

Abrir recurso propio en BigML

The screenshot shows the BigML interface with the 'Sources' tab selected. A yellow circle highlights the 'Base de datos - League of Legends.csv' entry in the list, indicating it's selected. The toolbar at the top includes options like 'Datasets', 'Supervised', 'Unsupervised', 'Predictions', 'Tasks', and 'WhizzML'. The main area displays a list of sources, including various CSV, XLS, and TSV files, along with their details like type, name, size, and last modified.

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Una vez dentro del recurso concreto, dale al menú para crear el dataset. Se despliega un apartado en el que puedes decidir el nombre que le vas a dar al dataset, y el tamaño que va

a ocupar.

Imagínate que en el tamaño te salieran más de 16MB. Entonces tendrías que renunciar a parte de las instancias para reducir su tamaño y poder seguir adelante.

En este sentido también volvemos a recordar que si te tomas la molestia de "trabajar" y revisar la base de datos antes de subirla a BigML, puedes aligerarla eliminando campos que no sean útiles, o instancias que contengan errores o a las que les falten muchos datos.

Generar un dataset a partir de un recurso en BigML

The screenshot shows the BigML interface with the following elements:

- Sources** tab selected.
- Datasets** tab.
- Supervised**, **Unsupervised**, **Predictions**, and **Tasks** dropdowns.
- WhizzML** dropdown.
- Dataset configuration** section:
 - Dataset name:** `Base de datos - League of Legends`.
 - Size:** `8.86 MB`.
 - Create dataset** button.
- Data preview** table:

Name	Type	Instance 1	Instance 2	Instance 3
gameId	123	3326086514	3229566029	33273
creationTime	123	1504279457970	1497848803862	1504
gameDuration	123	1949	1851	1493
seasonId	123	9	9	9
winner	123	1	1	1
- Search by name** input field.
- A green callout bubble with the text: **Accede al recurso que acabas de cargar, haciendo clic sobre él**.

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

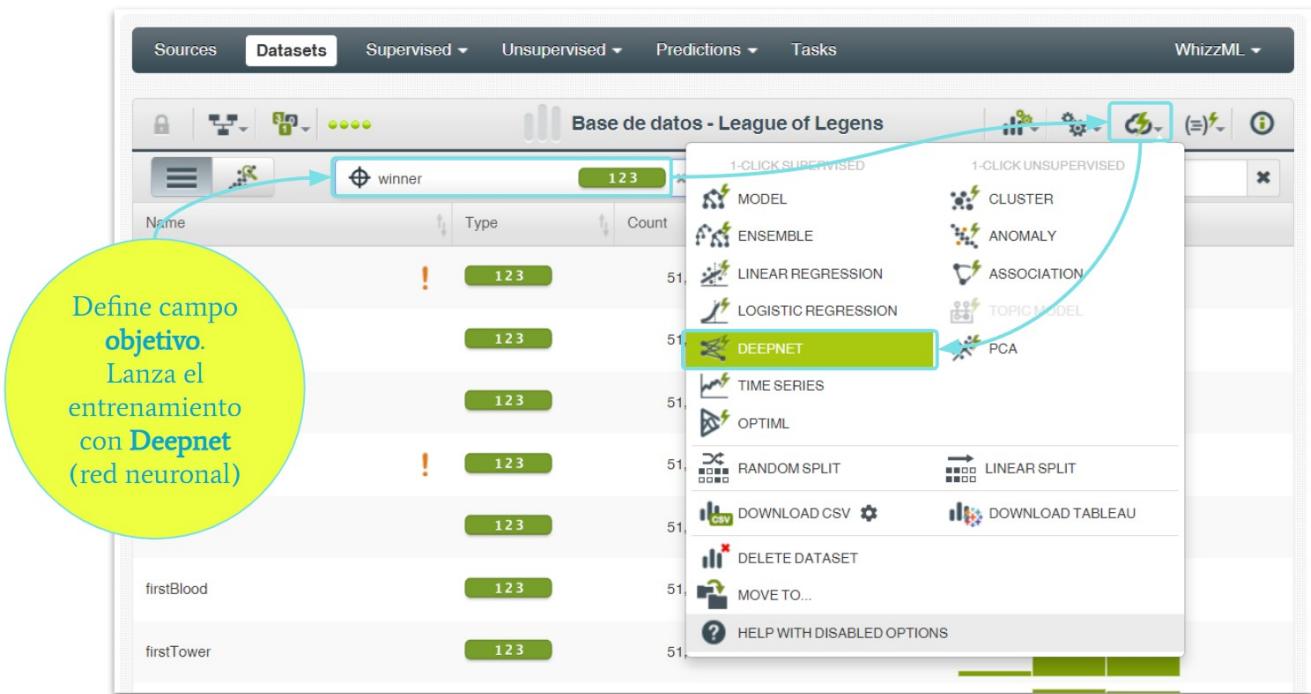
Ahora que ya hemos generado el dataset, hay que darle un buen repaso.

Revisa todas las categorías. Fíjate que directamente te aparecen algunas con un signo de admiración. Eso significa que el propio BigML ha detectado que no es un valor representativo, por ejemplo el número o nombre usuario, o el número de partida. En cualquier caso, haz la revisión, y activa o desactiva las categorías que consideres. Pues todo lo que sea "aligerar" los datos con los que trabaje la red neuronal, será tiempo y gasto de cómputo que nos ahorraremos.

También tienes que definir cuál es tu categoría objetivo. Por defecto te va a venir marcada una que no es lo que pretendemos. Así que asegúrate de marcar "winner" como el objetivo del problema.

Y después de todo esto, ya sí, dale a entrenar el modelo. En esta ocasión será con **DEEPNET**, que es como en BigML llaman al algoritmo de red neuronal profunda.

Iniciar entrenamiento con Red Neuronal en BigML



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

¿Te has dado cuenta que en esta ocasión ha tardado algo más en terminar el entrenamiento? Eso es porque tenemos un número bastante más elevado de instancias y categorías que en los ejemplos que realizamos en los capítulos 1 y 2.

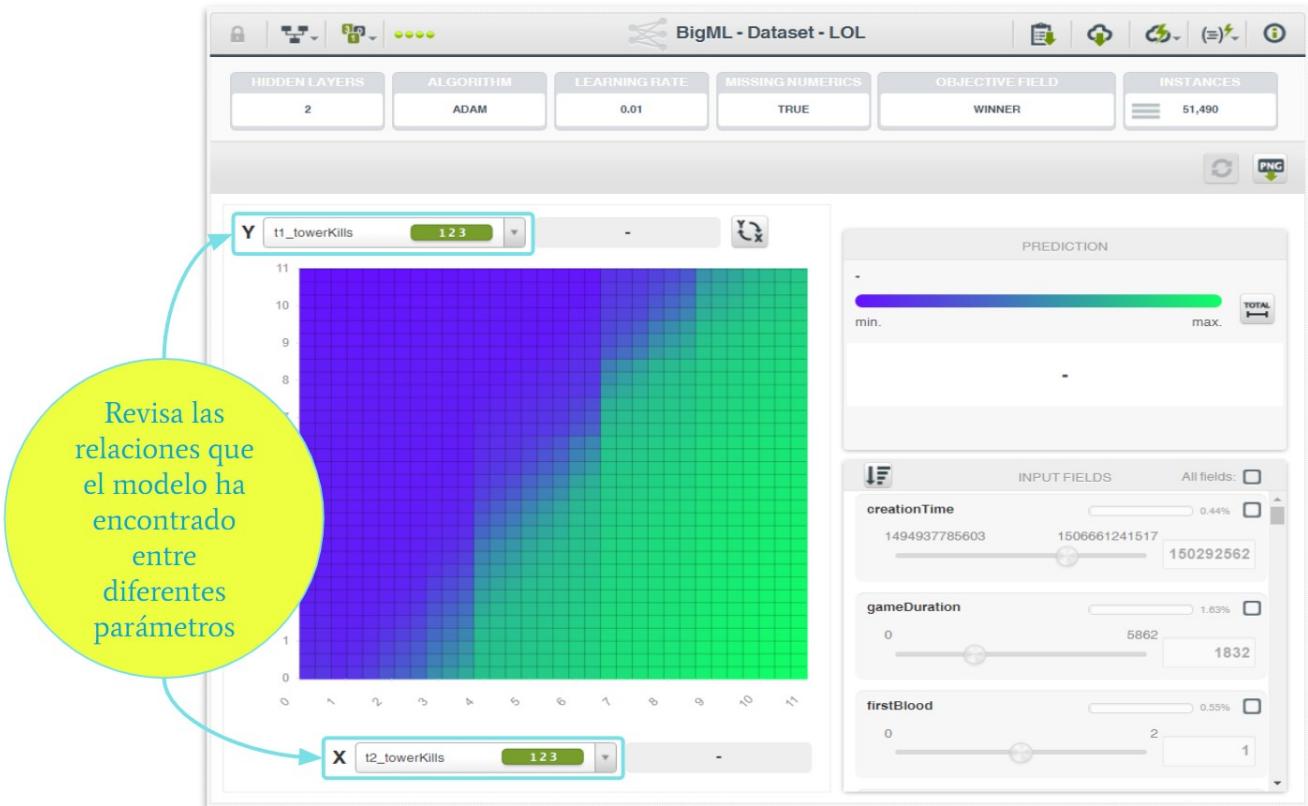
En este ejemplo, como nuestro campo objetivo era "winner" y esa categoría solo tiene como posibles resultados 1 ó 2 (equipo 1 o equipo 2), la presentación gráfica que nos ofrece BigML es con estas gráficas en las que el color azul se corresponde con que gana el equipo 1 y el color verde se corresponde con que gana el equipo 2.

En la siguiente imagen vemos el resultado (probabilidad de que gane uno u otro equipo) en función del número de torres destruidas. Cuantas más torres destruye el equipo 1 (en el eje y, vertical), más probabilidades tiene de ganar (más color azul). Cuantas más torres destruye el equipo 2 (en el eje x, horizontal), más probabilidades tiene de ganar (más color verde). Pero... ¿notas algo llamativo?

La frontera entre los tonos azules y verdes no inicia en el valor (0,0) (en la esquina inferior izquierda)... Por lo que parece que tiene más posibilidades de ganar el equipo 1. Fíjate en la posición x=2, y=1 (es decir, cuando el equipo 2 destruye dos torres y el equipo 1 destruye solo una torre). Segundo nuestro modelo tiene más probabilidades de ganar el equipo 1 que el 2, aún habiendo destruido menos torres.

Y esto, ¿a qué se debe? ¡esta es la pregunta que hay que hacerse! Hay unas cuantas posibilidades: Los desarrolladores del juego no han hecho bien su trabajo y hay un equipo con más probabilidades de ganar que el otro, o a lo mejor las torres no son determinantes para ganar el juego, o a lo mejor el juego está perfecto y lo que pasa es que tenemos un problema con nuestros datos: que nuestra base de datos esté sesgada, o que tenga deficiencias en los datos...

Análisis de resultados del modelo entrenado



Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Puedes ir cambiando las categorías representadas en el eje X y el eje Y, para ir analizando qué más información te ofrece este modelo.

También es interesante que pruebes a hacer alguna predicción con el modelo ya entrenado. Puedes modificar los valores de cada categoría (y también desactivar los que consideres no relevantes).

En la imagen a continuación te mostramos un ejemplo, en el que el modelo pronostica que con esos parámetros tiene más probabilidades de ganar el equipo 1 que el 2: recuerda que la categoría **winner** en nuestra base de datos solo tomaba valores 1 o 2. De manera que cualquier valor entre 1 y 2 debemos leerlo mirando cuánto más cerca o lejos está del 1 y del 2. En el caso "winner: 1.12" está mucho más cerca de 1 que de 2. El empate estaría en el caso "winner: 1.5".

Realizar predicción con modelo entrenado

Revisa las relaciones que el modelo ha encontrado entre diferentes parámetros

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))

Igual que en los ejercicios anteriores podríamos hacer dentro de BigML una evaluación de la eficacia del modelo. Pero... en este caso ¡nos faltan datos!... No hemos tenido la precaución de dividir nuestro dataset en dos partes (80% para entrenar el modelo, y 20% para hacer ahora la evaluación). Así que... habría que volver al apartado datasets, hacer la partición, volver a entrenar un modelo Deepnet con el dataset del 80% y entonces sí tendríamos datos realistas para poder evaluar dicho modelo.

¿Crees que eres capaz de hacerlo por tu cuenta? ¡Seguro que sí!

Evaluar modelo

Haz una evaluación de la confiabilidad del modelo

Fran Bartolomé - Elaboración propia ([CC BY-SA](#))



Para saber más

Para poder sacarle partido a este ejercicio es mucho mejor si conoces la mecánica del juego League of Legens. En realidad, para hacer cualquier modelo de aprendizaje automático o profundo es necesario conocer bien el mundillo del que proceden los datos.

Es relativamente sencillo encontrar información y tutoriales rápidos sobre qué es y cómo se juega a League of Legens. Pero, ojo, no te distraigas mucho... que una cosa es conocer el juego y otra distinta enviciarse a jugar con él 😊.



Autoevaluación

Indica si la siguiente afirmación es verdadera o falsa:

La revisión y procesamiento de los datos antes de generar el dataset en BigML es importante porque nos permite "aligerar" de categorías inútiles o instancias erróneas la base de datos.

- Verdadero Falso

Verdadero

Sí, la gestión de los datos ya hemos visto en numerosas ocasiones que es fundamental para optimizar el proceso de entrenamiento de cualquier modelo de inteligencia artificial