

Máster Universitario en Datos
Cloud Y Gestión TI

Fundamentos de ingeniería de datos



BIGML

- PREDICTIONS -



25/01/2021

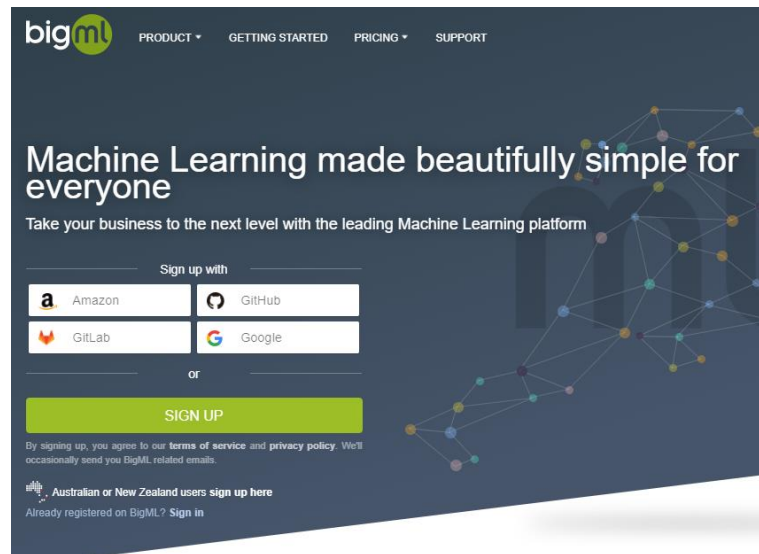
Grupo 5

Como ejercicio adicional para poder obtener la máxima nota hemos realizado un estudio de la herramienta BIGML para realizar predicciones.

Sobre esta herramienta cabe destacar que nos ofrece una buena capacidad de cómputo en la nube, tiene una amplia variedad de modelos y que es realmente intuitiva, es por ello una herramienta de aprendizaje especialmente interesante para iniciarse en el análisis de datos ya que en unos pocos pasos estaremos realizando predicciones sobre nuestro modelo.

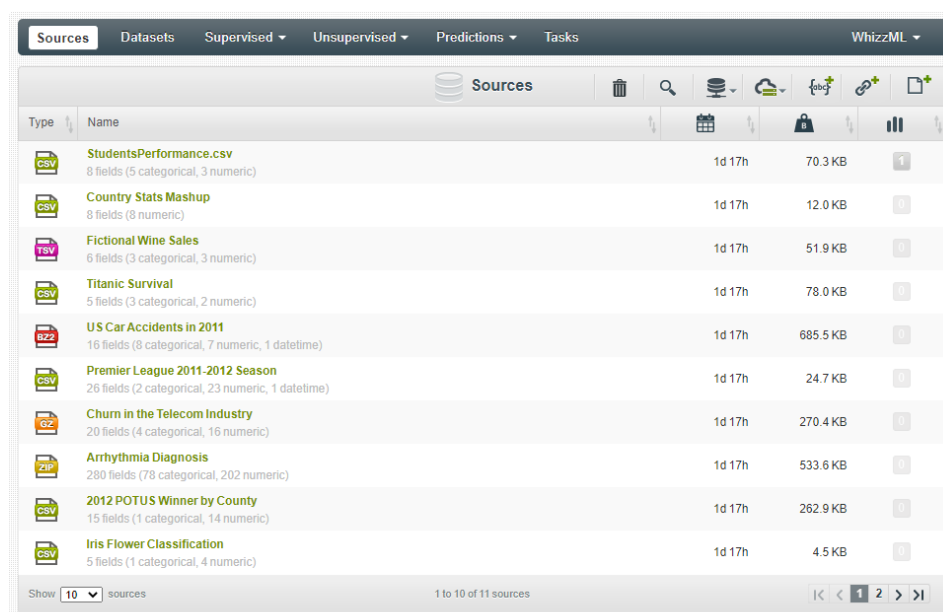


El primer paso en la utilización de BigML es el registro en la plataforma, para ello nos dirigimos a la página









Podemos registrarnos de cualquiera de las formas que ahí se indica.

Una vez logados veremos la vista principal de bigML, el denominado dashboard, además veremos que bigML pone a nuestra disposición varios datasets sobre los que ir trabajando. En nuestro caso, como queremos trabajar con nuestro propio dataset, nos dirigimos a la parte superior derecha y hacemos click en el folio en blanco, desde ahí podremos subir a la plataforma nuestro dataset.



Una vez registrado el dataset podemos realizar las transformaciones de variables y atributos pertinentes en el caso de que nuestro dataset contenga inconsistencia o por si la plataforma no hubiera clasificado correctamente alguno de los atributos de nuestro dataset (esto es relativamente frecuente con datasets pequeños o con columnas con muchos valores nulos).

Name	Type	Count	Missing	Errors	Histogram
gender	ABC	1,000	0	0	
race/ethnicity	ABC	1,000	0	0	
parental level of education	ABC	1,000	0	0	
lunch	ABC	1,000	0	0	
test preparation course	ABC	1,000	0	0	
math score	123	1,000	0	0	

En nuestro caso, evalúa el dataset correctamente por lo que no tenemos que hacer ningún tipo de tratamiento adicional.

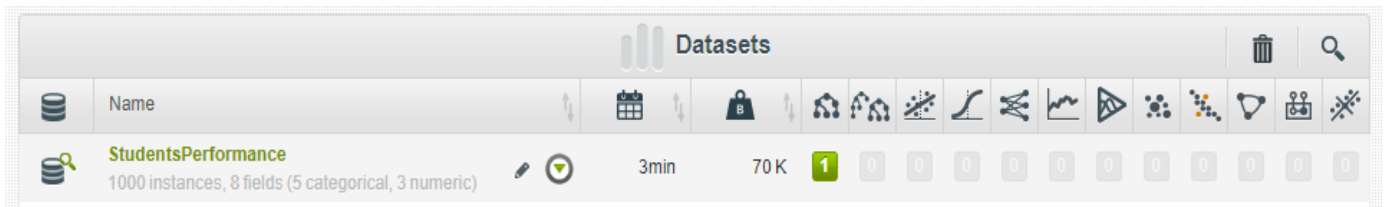
Con el dataset listo el siguiente paso consistiría en el entrenamiento de algún modelo, y es aquí donde se ve la capacidad computacional y el alto nivel de abstracción que nos proporciona la plataforma, donde sin escribir ni una sola línea de código podemos generar una cantidad de modelos diferentes muy elevada.

En nuestro caso, como queremos centrarnos en la parte de predicción el modelo que vamos a generar es un árbol de decisión relativamente sencillo.

Para realizar un árbol de decisión normalmente tendríamos que realizar una discretización de las variables numéricas, pero como veremos a continuación esto tampoco hace falta realizarlo porque la plataforma lo hace por nosotros, veremos que en nuestro dataset hay 3 atributos numéricos y que cuando introduzcamos el modelado

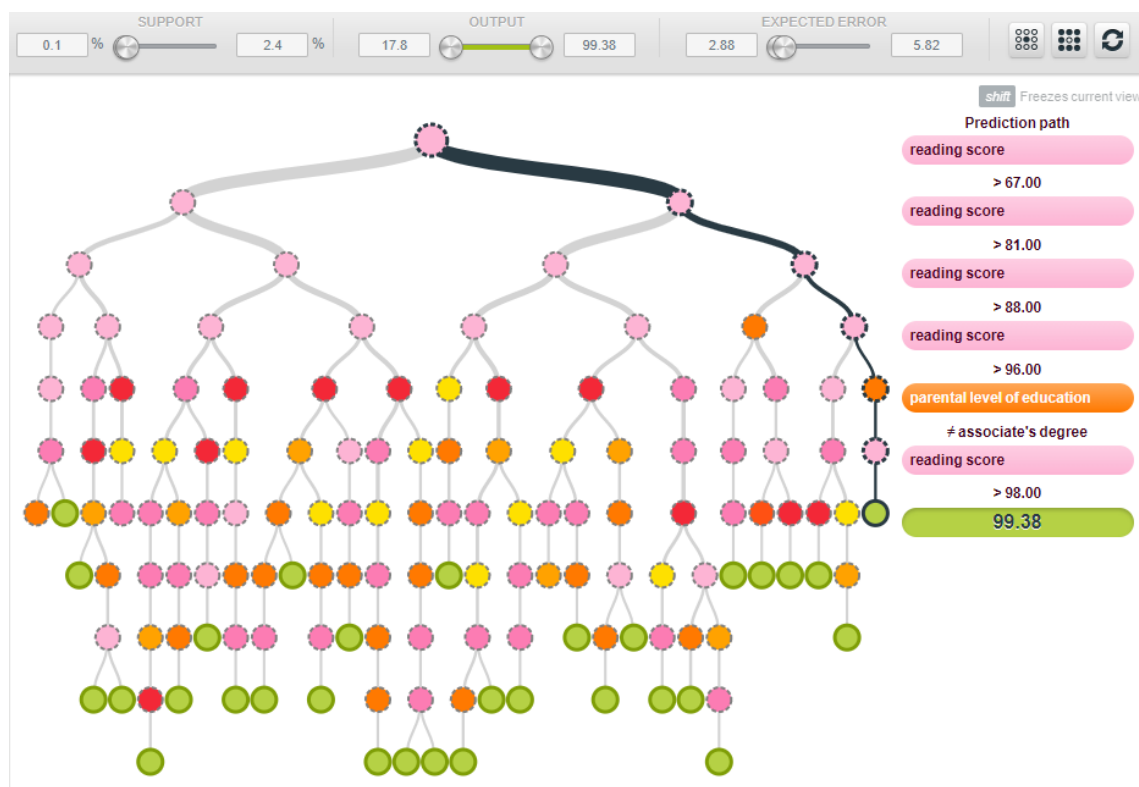
por árboles de decisión la plataforma nos agrupará las notas en varios conjuntos.

Para ello nos dirigimos una vez más a la parte superior derecha y seleccionaremos la opción de modelar por arboles de decisión:



Aquí podemos ajustar todos los parámetros de nuestro modelo para intentar que haga las mejores predicciones posibles.

Una vez generado el modelo, podemos ver el árbol generado



Con este árbol ya podemos predecir el comportamiento que tendrán nuevas instancias, para ello, nos dirigimos a la pestaña de predictions en la parte superior (símbolo de la nube).

The screenshot shows the 'Predictions' tab in the WhizzML interface. At the top, a navigation bar includes 'Sources', 'Datasets', 'Supervised', 'Unsupervised', 'Predictions', and 'Tasks'. The main header area displays 'Predict using StudentsPerformance' and a 'writing score: 71.40' with a '7.15' prediction value. Below this, a 'Missing strategy' section shows two icons. The main area contains eight input fields, each with a slider, a percentage, and a checked status icon. The inputs are: 'reading score' (0-120, 94.47%, 69), 'math score' (0-125, 1.98%, 66), 'parental level of education' (1.14%, dropdown: 'associate's degree'), 'test preparation course' (0.93%, dropdown: 'completed'), 'gender' (0.87%, dropdown: 'female'), 'race/ethnicity' (0.52%, dropdown: 'group A'), 'lunch' (0.10%, dropdown: 'free/reduced'), and 'All input fields:'. At the bottom, a 'New prediction name' field contains 'StudentsPerformance', and a green 'Save' button is visible.

Field	Value	Percentage	Status
reading score	69	94.47%	✓
math score	66	1.98%	✓
parental level of education	associate's degree	1.14%	✓
test preparation course	completed	0.93%	✓
gender	female	0.87%	✓
race/ethnicity	group A	0.52%	✓
lunch	free/reduced	0.10%	✓
All input fields:			✓

En esta vista de predicciones lo único que tenemos que hacer es especificar los parámetros del individuo que queremos clasificar, en la parte superior nos aparecerá la predicción del writing score de nuestro modelo (que es la variable que estamos clasificando).

Cabe destacar que además de esta forma de predecir, también podemos hacerlo en batch, es decir, pasar como parámetro un fichero csv con todas las instancias que queramos clasificar y el sistema nos añadirá una columna con la predicción realizada.