CURSO DE INTELIGENCIA ARTIFICIAL MÁLAGA, 30 mayo a 20 julio 2021

Modelar sobre plataforma de ML
Ejercicio 1

Andrés González

Ejercicio 1. Mi primer modelo. Enunciado

El objetivo académico de este ejercicio es que des tus primeros pasos en BigML. Para ello vas a crear un primer modelo y ver su calidad. En este caso concreto vas a entrenar un modelo que indique si un crédito es de buena o mala calidad. **El conjunto de datos que vas a revisar divide los créditos en dos tipos: "good" y "bad", es decir, los que se devolvieron y los que no.**

Para ello vas a seguir los siguientes pasos:

- Harás una revisión básica de los datos que vas a usar.
- Subirás a BigML los datos que están en una dirección de internet.
- Ya en BigML crearás un conjunto de datos de entrenamiento (80%) y otros de test (20%).
- Después crearás un modelo de árbol de decisión con los datos de entrenamiento.
- Y finalmente evaluarás la calidad del modelo con los datos de test.

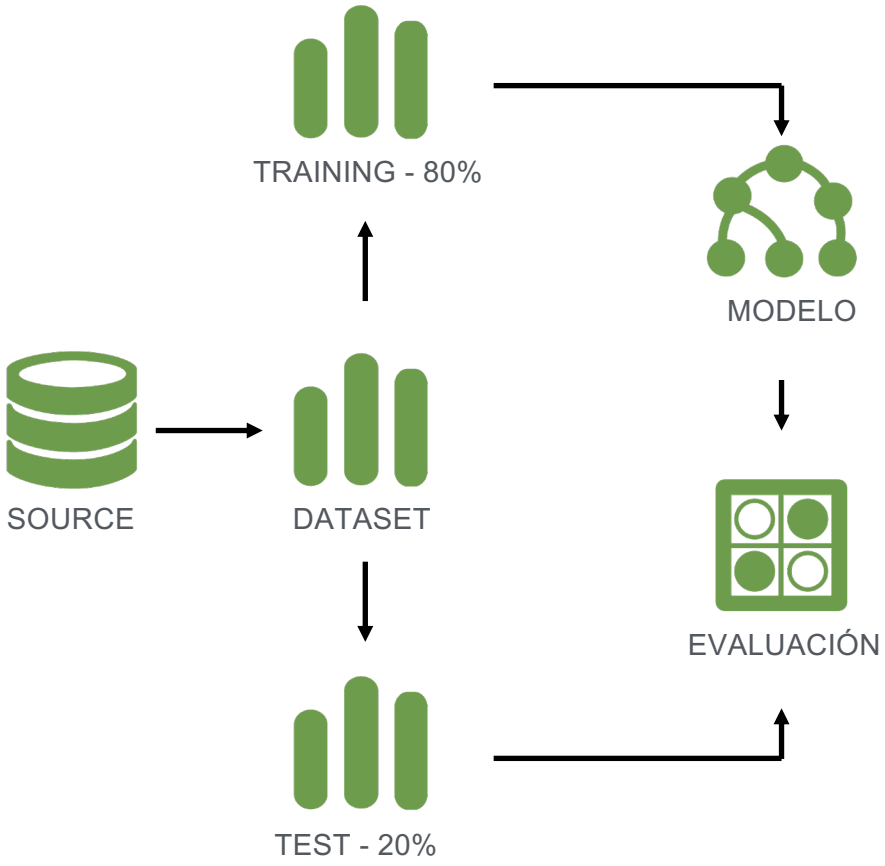
Después de crear y evaluar el modelo de árbol de decisión, vas a usar otro algoritmo, un *ensemble* de tipo *bagging*. Compararás los resultados de ambos algoritmos. Los pasos son:

- Crearás el modelo con el mismo conjunto de datos de entrenamiento de los apartados anteriores (80%).
- Evaluarás la calidad del modelo con el conjunto de datos de test.

¿Vamos?

Ejercicio 1. Mi primer modelo. Enunciado

- Este es el flujo que vas a seguir:



Ejercicio 1. Mi primer modelo. Enunciado

1. Revisión preliminar de datos

- Descarga el csv de entrenamiento y ábrelo con tu programa de hojas de cálculo favorito (copia y pega esta dirección en tu navegador):

https://cleverdata.io/csv/Loan-Data.csv

- **PREGUNTA 1.1**: Revisa las columnas, los tipos de datos y rellena la siguiente tabla:

	Respuesta
Número de filas, sin contar los nombres de las columnas (cabecera)	
Número de columnas	
¿Eliminarías alguna columna por poder introducir sesgos sexistas o racistas? Si es que sí, ¿cuál o cuáles?	

Ejercicio 1. Mi primer modelo. Enunciado

2. Sube los datos a la plataforma

- Entra en la web de BigML con tu usuario y contraseña.
- Sube a BigML el source desde el enlace web.
- Revisa si ha detectado correctamente todos los tipos de campo (los campos con números como 123 y los categóricos como ABC. Si alguno no lo ha detectado correctamente, modifícalo editando el Source.

3. Crea los conjuntos de datos de entrenamiento y de test

- Primero crea un conjunto con el 100% de los datos con un 1-CLIC DATASET.
- Verifica si el número de instancias es el mismo que has visto tú en tu hoja de cálculo. Si no es así, modifícalo.
- Ahora vas a crear los dos conjuntos de datos, uno para el entrenamiento con el 80% de las instancias y otro para medir la calidad de modelo, con el 20% restante. Hazlo con la opción **RANDOM SPLIT**.

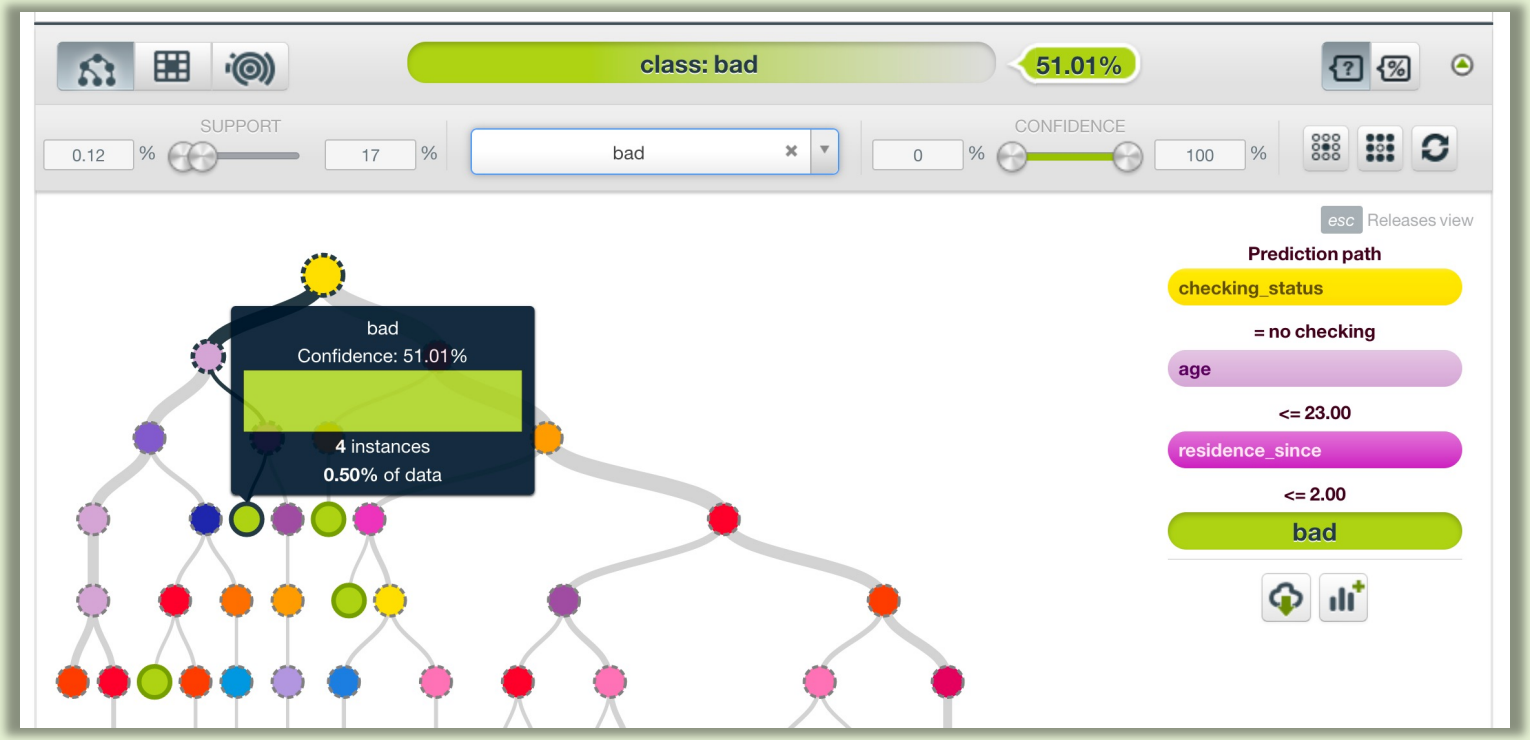
Ejercicio 1. Mi primer modelo. Enunciado

4. Crea el modelo de árbol de decisión

- Haz una pequeña preparación de los datos antes de entrenarlo. En el conjunto de datos que vas a usar para entrenar el modelo (el que tiene 80% de las instancias), pon como “not preferred” el campo “personal_status”.
- Si es necesario, pon “class” como campo objetivo.
- Entrena un modelo de árbol único ”1-CLICK MODEL” con el conjunto de datos que tiene el 80% de las instancias.
- Vamos a revisar el árbol de decisiones.
- Selecciona con el ratón un patrón que tenga como clase BAD en la hoja final (verde) y que tenga pocos nodos.
- Asegúrate de que está completo el ”Prediction Path” de la derecha, hasta el final BAD.

Ejercicio 1. Mi primer modelo. Enunciado

- **PREGUNTA 1.2**: Identifica un patrón “bad” con más del 80% de confianza y otro con menos del 25%. ¿Cuántas instancias tiene cada uno?



Ejercicio 1. Mi primer modelo. Enunciado

5. Evalúa el modelo con el conjunto de test (20%)

- En BigML evalúa el modelo con el conjunto de test (20%). Usa la opción de la nube con el rayo.

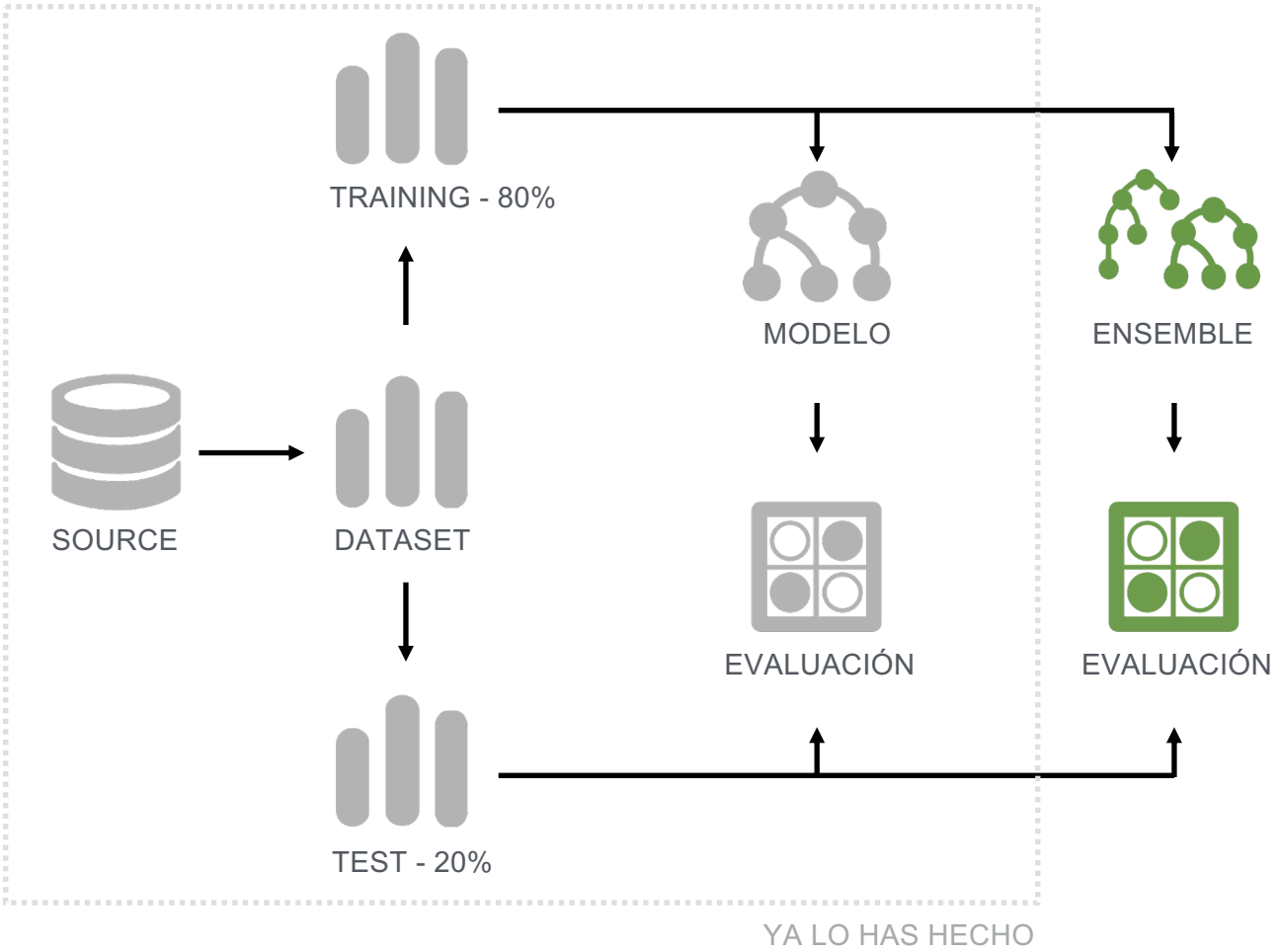
- **PREGUNTA 1.3**: ¿Qué valores de calidad ofrece el Dashboard para cada clase? Rellena los valores de la tabla.

	bad	good
Accuracy		
Precision		
Recall		
PHI Coefficient		

- **PREGUNTA 2.4**: ¿Qué tipo de créditos (“good” o “bad”) detecta mejor el modelo de árbol único? Señala dos indicadores de la tabla anterior para razonar tu respuesta.

Ejercicio 1. Mi primer modelo. Enunciado

- Ahora vas a entrenar los mismos datos con un *ensemble*. Los vas a evaluar con el mismo conjunto de datos del 20%. Este es el flujo:



Ejercicio 1. Mi primer modelo. Enunciado

6. Entrena un modelo con un conjunto de 10 árboles (*ensemble*)

En los apartados anteriores has usado un modelo de árbol de decisión.

Ahora vas a entrenar un nuevo modelo con un ***ensemble***, es decir, varios árboles combinados.

- Usa el mismo conjunto de datos del 80% y crea un 1-CLIC ENSEMBLE.
- Recuerda que el 1-CLIC ENSEMBLE entrena el modelo con 10 árboles, en lugar de 1.
- Cada uno de los 10 árboles usa una muestra aleatoria de los datos de entrenamiento (no usa el conjunto de datos completo).

7. Evalúa el conjunto de árboles con el conjunto de test

- La calidad de un *ensemble* en general es mejor que la de un único árbol. Vamos a comprobarlo.
- Evalúa el modelo *ensemble* con el conjunto de datos del 20%. Usa otra vez la opción de la nube con el rayo.

Ejercicio 1. Mi primer modelo. Enunciado

- **PREGUNTA 1.5**: ¿Qué valores de calidad ofrece el Dashboard del *ensemble*? Rellena los valores de la tabla.

	1-CLIC MODEL		1-CLIC ENSEMBLE	
	bad	good	bad	good
Accuracy				
Precision				
Recall				
PHI Coefficient				

- **PREGUNTA 1.6**: ¿Qué modelo da mejores resultados? Razona tu respuesta señalando en qué indicador **global** (tal y como dijimos en la clase) del apartado anterior te has fijado.

  NOTA: no necesariamente sale mejor el *ensemble* que el modelo de árbol único.

Ejercicio 1. Mi primer modelo. Enunciado

- **PREGUNTA 1.7**: ¿Cuáles son las 3 variables que tienen más peso en ambos modelos y en qué orden?

	1-CLIC MODEL		1-CLIC ENSEMBLE	
	Nombre	%	Nombre	%
Primer campo				
Segundo campo				
Tercer campo				

- **PREGUNTA 1.8**: Si los datos de entrenamiento son exactamente los mismos (has usado el 80% tanto para el árbol como para el *ensemble*), ¿por qué salen valores diferentes en las variables con más peso?

Ejercicio 1. Mi primer modelo. Enunciado

- **PREGUNTA 1.9**: Entrega los siguientes enlaces a tus recursos:
  IMPORTANTE: los enlaces **NO** son la URL del navegador. Repasa la clase del martes para saber cómo los tienes que conseguir.

  - Dataset:
  - Dataset 80%:
  - Dataset 20%:
  - Modelo Árbol 80%:
  - Modelo *Ensemble* 80%:
  - Evaluación Árbol:
  - Evaluación *Ensemble*:

Material de apoyo:

- Video (5 min.): Clasificación y regresión: árboles de decisión
- Vídeo (9 min.): Predicción y evaluación
- Artículo: Machine Learning: predicciones basadas en datos con BigML

Vamos.