Tarea para SAA04

Título de la tarea: Definición de propiedades de una red neuronal para ejercicio de clasificación.

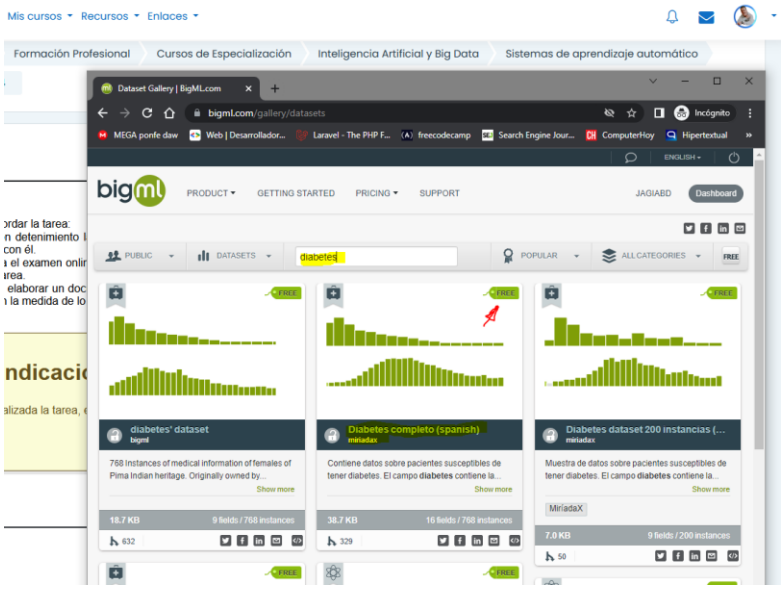
Ciclo formativo y módulo: Curso especialización en Inteligencia Artificial y Big Data - Casos prácticos de aplicación.

Curso académico: 2022-2023

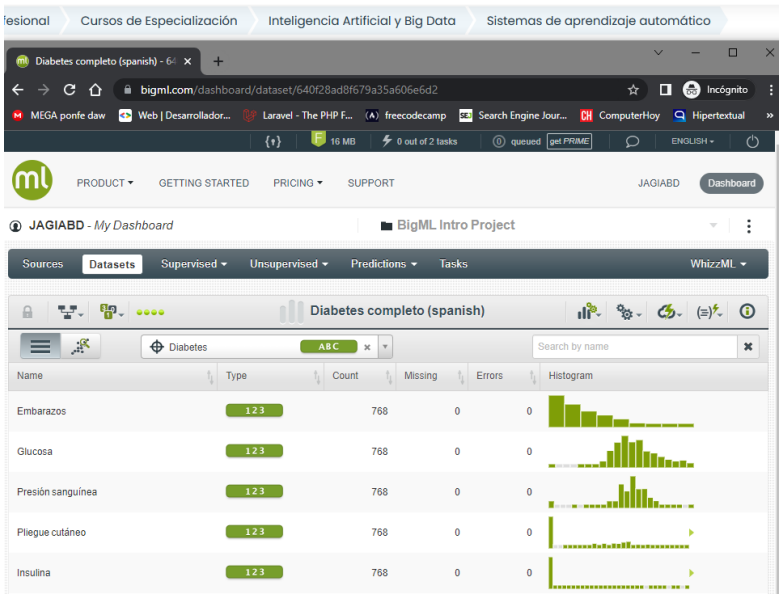
¿Qué te pedimos que hagas?

Te pedimos que utilices la plataforma BigML, tal como se ha explicado en el caso práctico del primer capítulo de esta unidad. Se da por sentado que sabes acceder, y que tienes usuario para trabajar en ella. Tendrás que entregar un documento con la siguiente información:

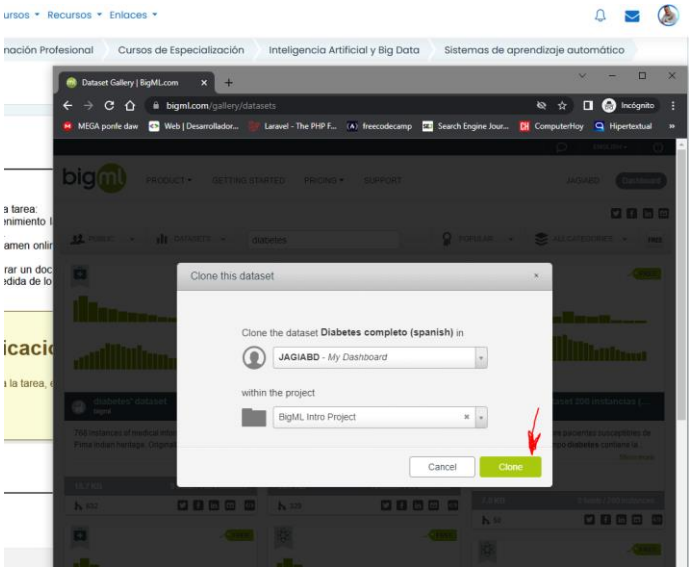
- **Apartado 1: Localiza el dataset "Diabetes completo (spanish)":**
    - Utiliza el buscador de datasets que tiene la propia plataforma para ello.



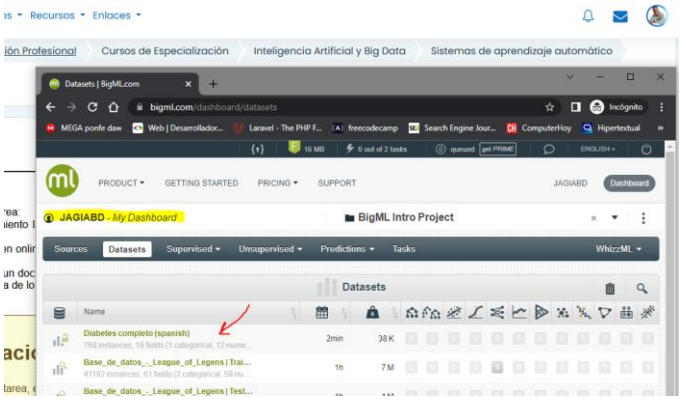
Introduciendo una palabra clave podemos encontrar mejor el dataset.



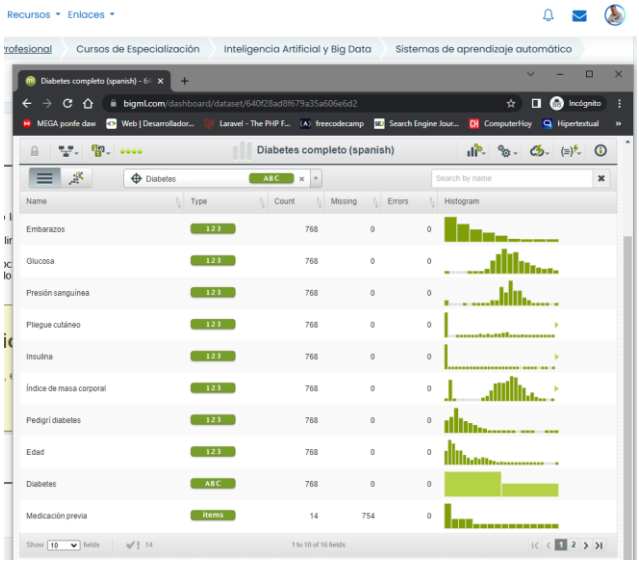
- Incorpora una captura de pantalla del dataset mencionado incorporado a tu apartado *Datasets*.



Tras clonarlo, podemos verlo ya entre nuestros datasets



- **Apartado 2: Observación del dataset:**
  - Incorpora una captura de pantalla del dataset donde se vean al menos 10 categorías con sus tipologías, errores, histogramas, etc.

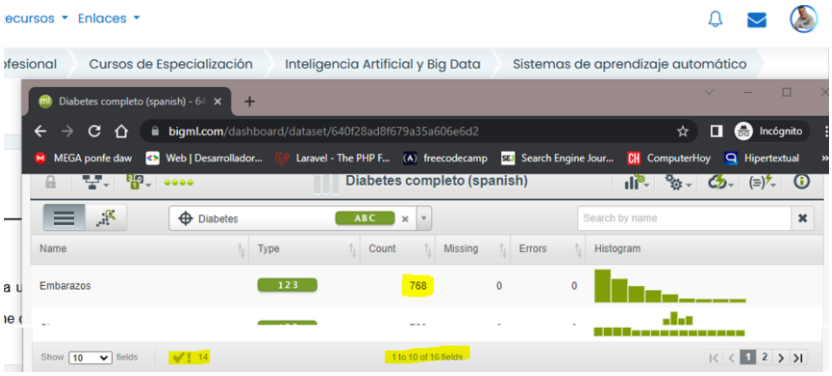


- Explica cómo es el dataset: Número de instancias y número de categorías.

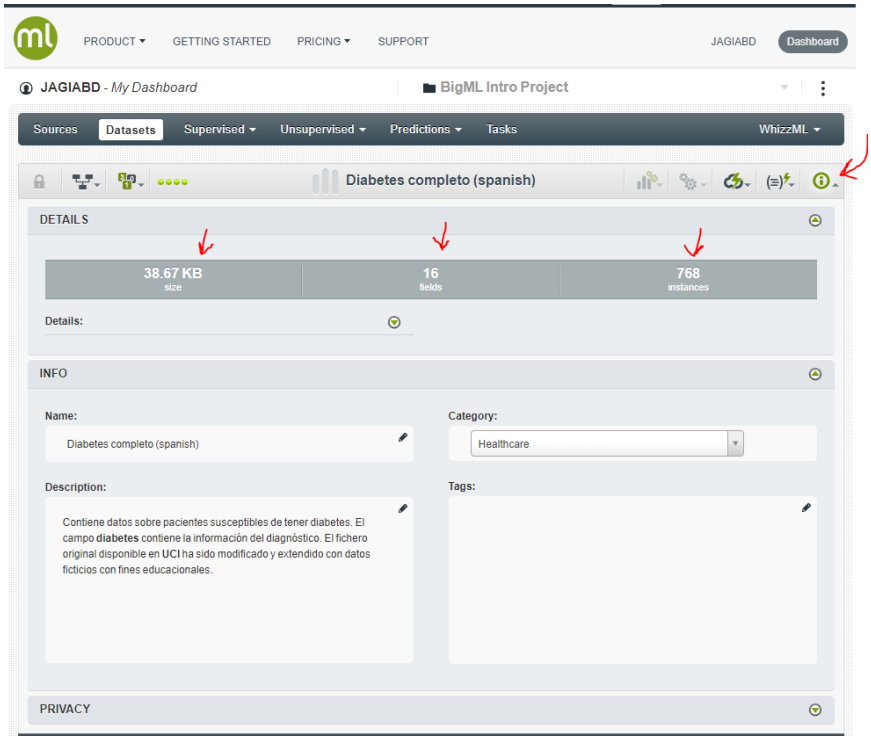
Como ya nos avanzaba desde el apartado introductorio en el "Caso Práctico", es un **`dataset`** de poco tamaño.

Ya en un primer vistazo, se pueden observar en concreto un total de 768 instancias desde la columna **`Count`**, y 16 categorías, que se ven contabilizadas en la parte inferior del **`dataset`**.

A la izquierda de este dato, he resaltado también que de esas 16, **BigML** ha detectado que sólo 14 tienen valores significativos, aligerando así los datos con los que trabajar.



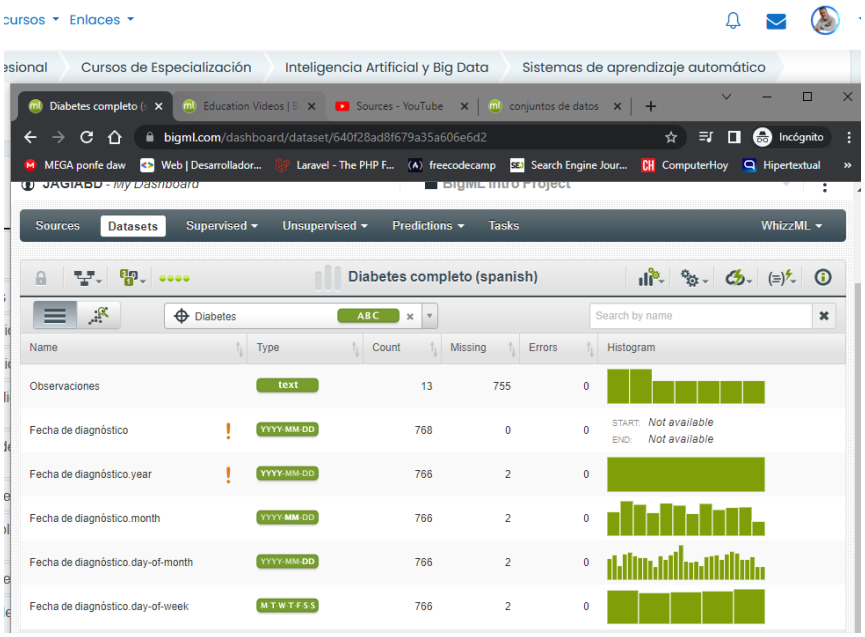
Toda esta información también la tenemos haciendo clic sobre el icono de información, así como su fecha de creación, la categoría del sector o la descripción:



- Explica el tipo de categorías (numéricas, texto, items, categóricas...).

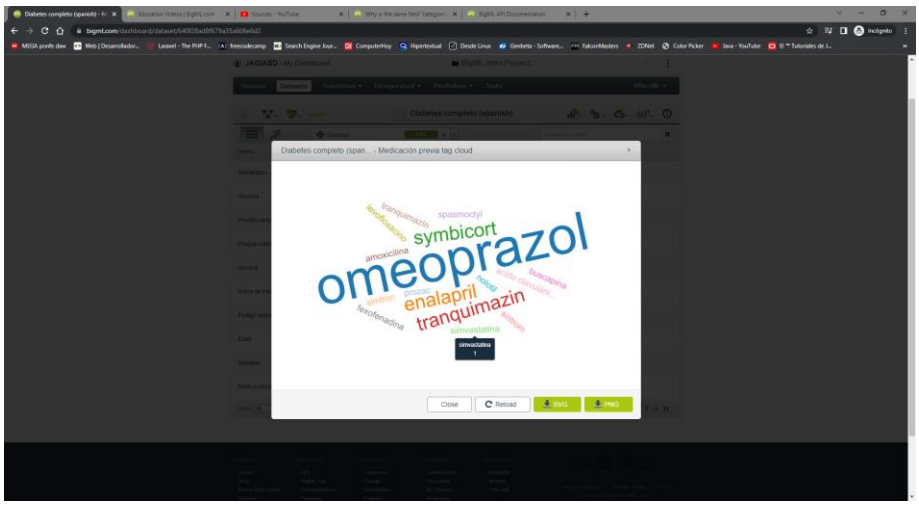
En la columna **Type**, se muestra de forma muy visual el tipo de dato que utiliza cada categoría. La mayoría, como se observaba en la anterior captura del listado, son de carácter numérico.

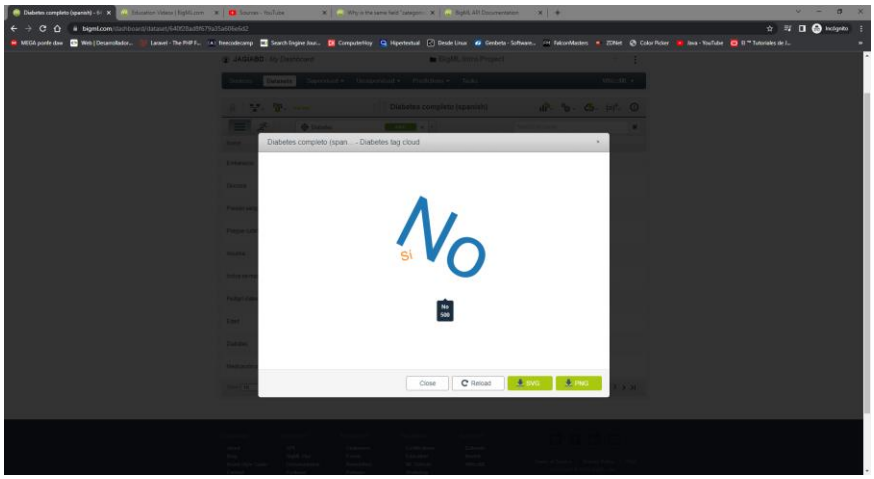
También se presentan varias con distintos tipos de formato de fecha como se puede ver aquí:



El tipo de formato de fecha para cada una de las categorías queda reflejado en la misma descripción (ej: **`YYYY-MM-DD`**).

Por último, tenemos categorías de tipo **`text`** (texto) como ***Observaciones***, **`items`** (elementos con muchos valores categóricos diferentes en los que el propio **BigML** intentará detectar el mejor separador no alfanumérico para ellos) como ***Medicación previa***, y **`categorical`**, un tipo de dato de texto que representa un conjunto de datos para una categoría (***Diabetes***). Estos dos últimos, podemos verlos aquí de forma gráfica para una mejor idea de su significado:

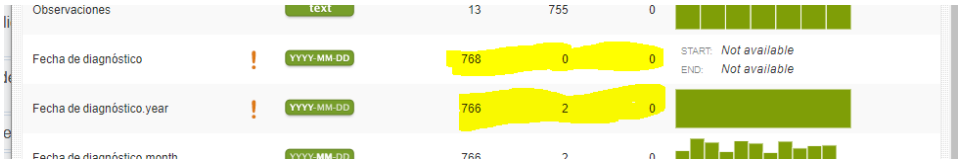




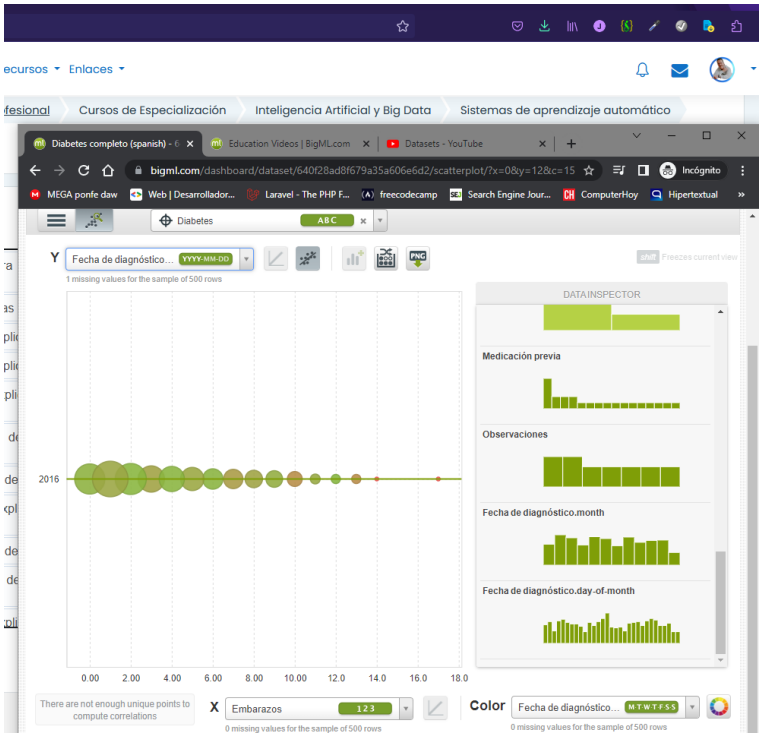
- Analiza los histogramas de cada categoría y comenta aquellos en los que consideres que hay algún tipo de anomalía.

La mayor parte de los histogramas nos reflejan los datos numéricos guardados en cada categoría.

Tenemos como veíamos en las capturas anteriores campos de texto en los que **BigML** nos muestra los datos mediante esos tokens, y finalmente, dos campos de fecha que han sido descartados:

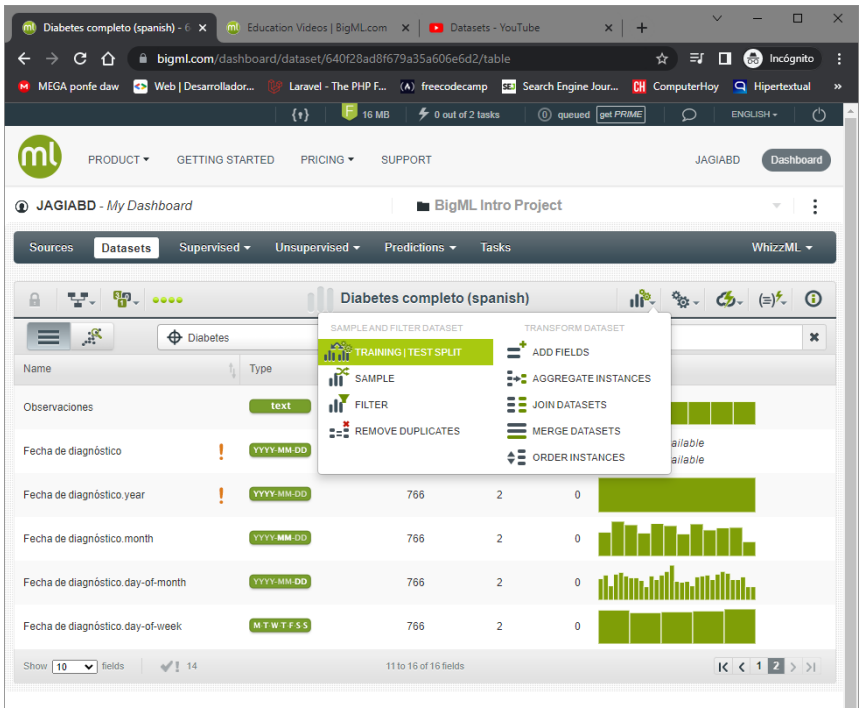


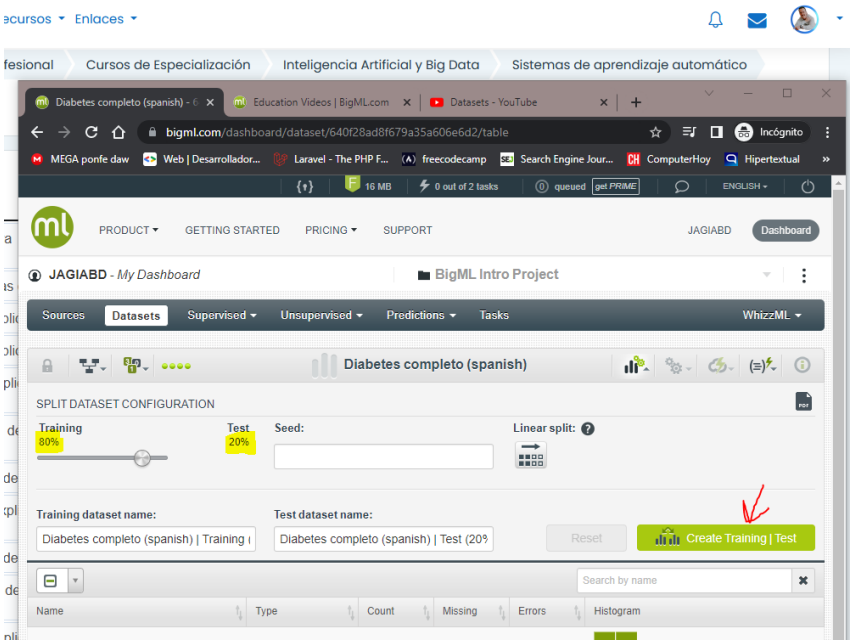
En el primero, vemos que no se han obtenido datos, pues las fechas han sido guardadas en otros formatos posteriormente, y la categoría que nos fecha por año, nos arroja que las 766 instancias recogidas, tienen el valor 2016.



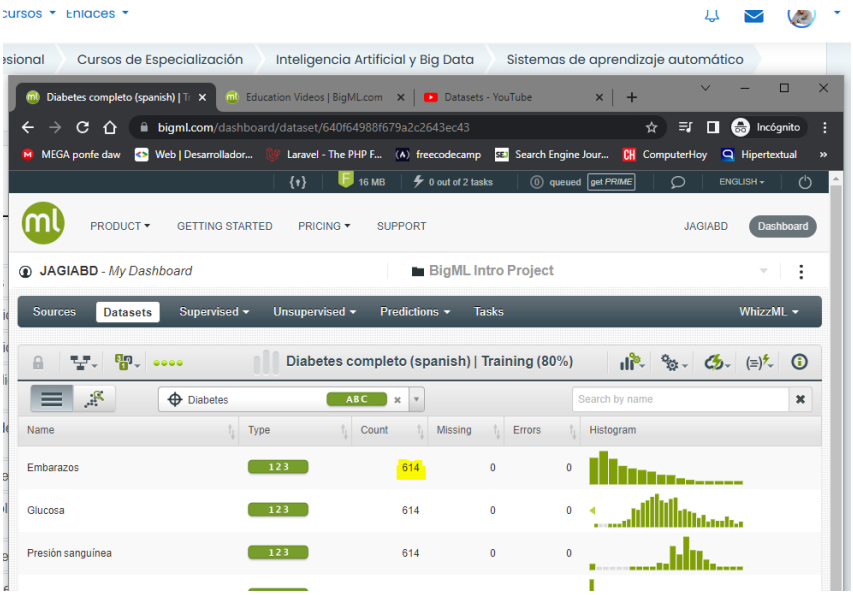
- **Apartado 3: Preparación del dataset para entrenamiento y test:**
    - Incorpora una captura de pantalla del proceso en el que defines los porcentajes de datos reservados para entrenamiento y para test.

Procedemos desde el menú.

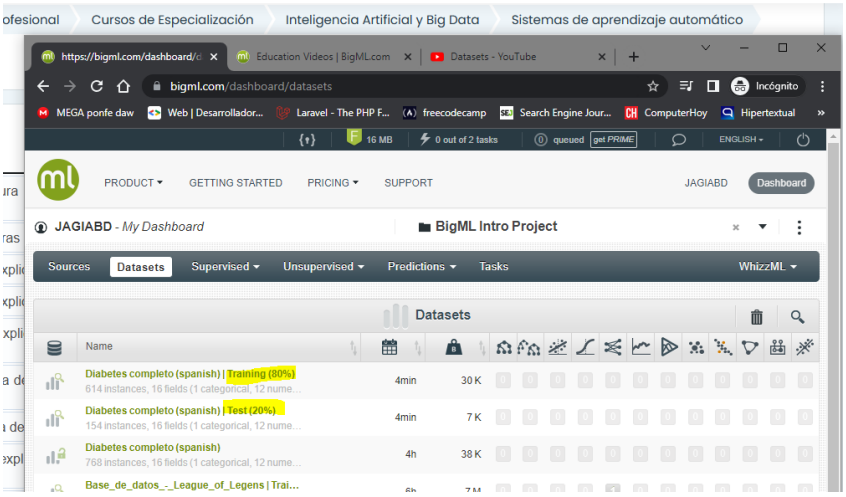




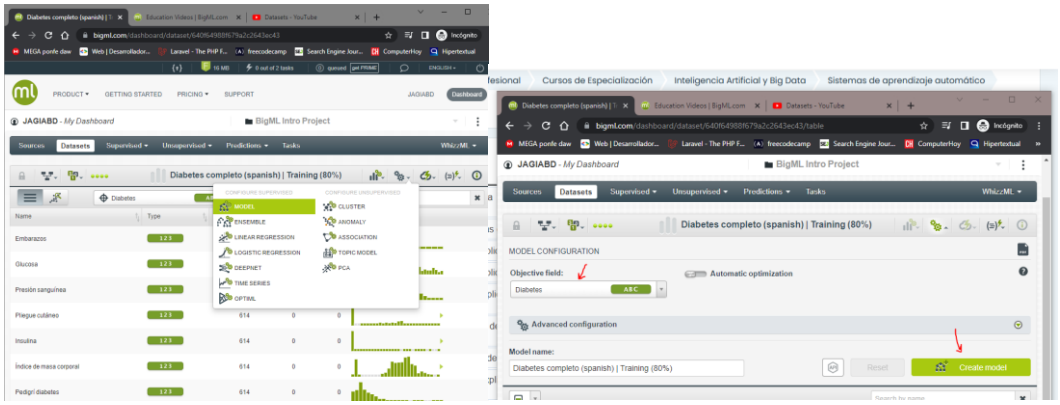
Tras seleccionar los porcentajes hacemos clic sobre el botón de crear y nos lleva al nuevo **`dataset`** con el 80% de los datos (como se puede además observar en el número de instancias)



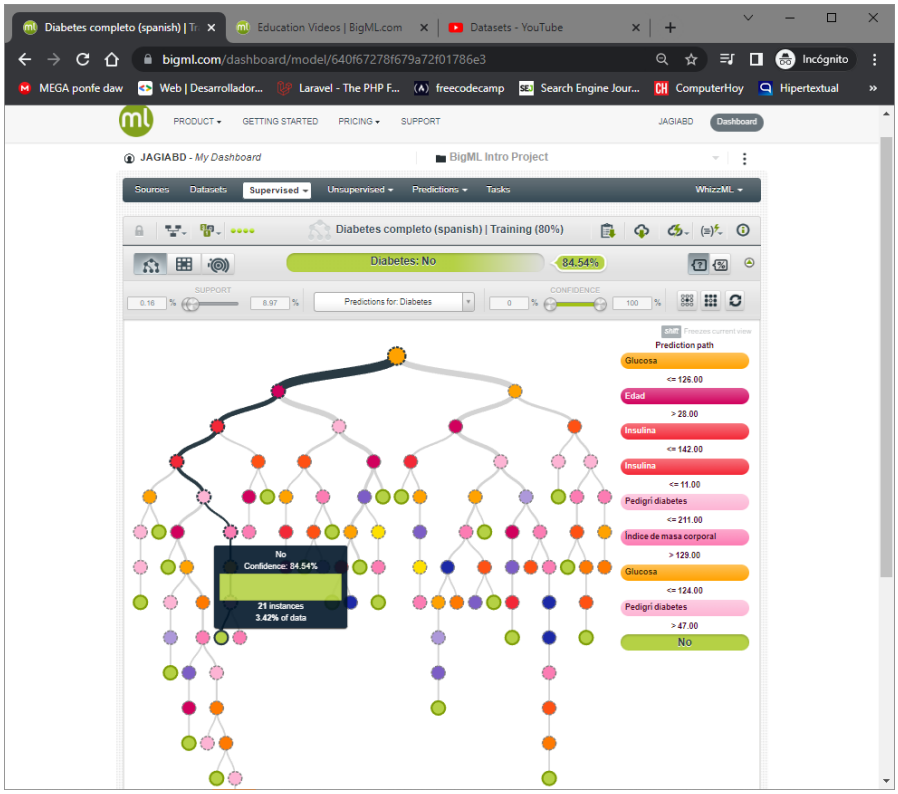
También los podemos ver incorporados a nuestra lista de **`Datasets.`**



- **Apartado 4: Entrenamiento:**



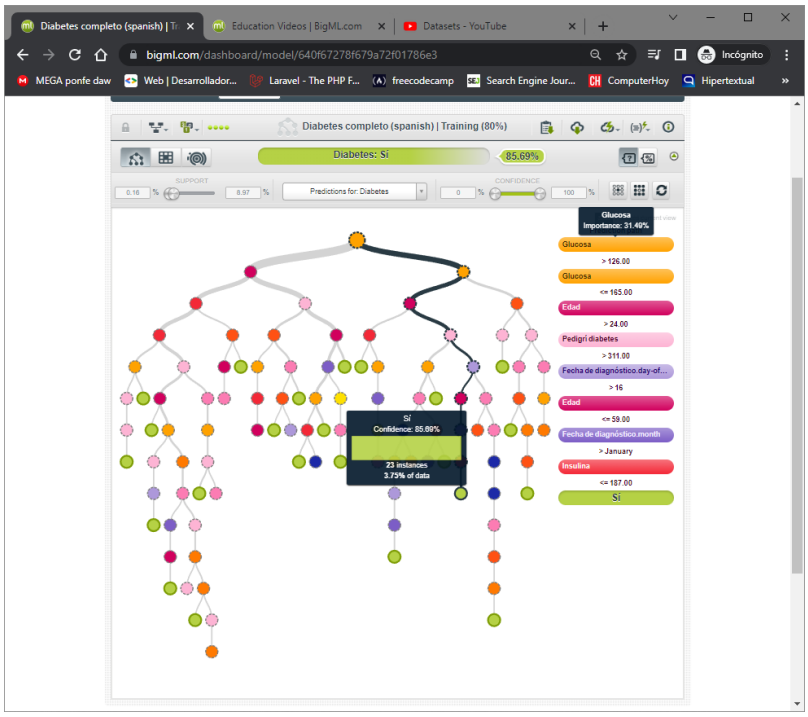
- Incorpora una captura de pantalla que muestre el árbol de decisión del modelo ya entrenado.

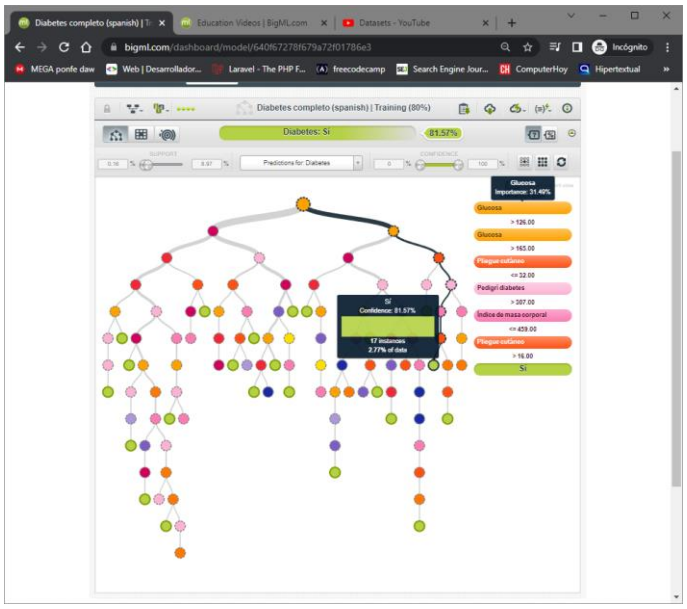


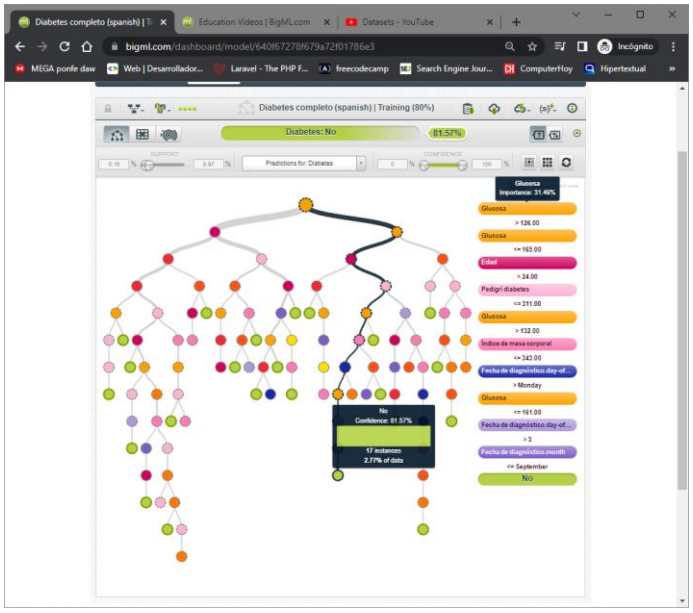
Podemos ver ya el diagrama de árbol y los resultados.

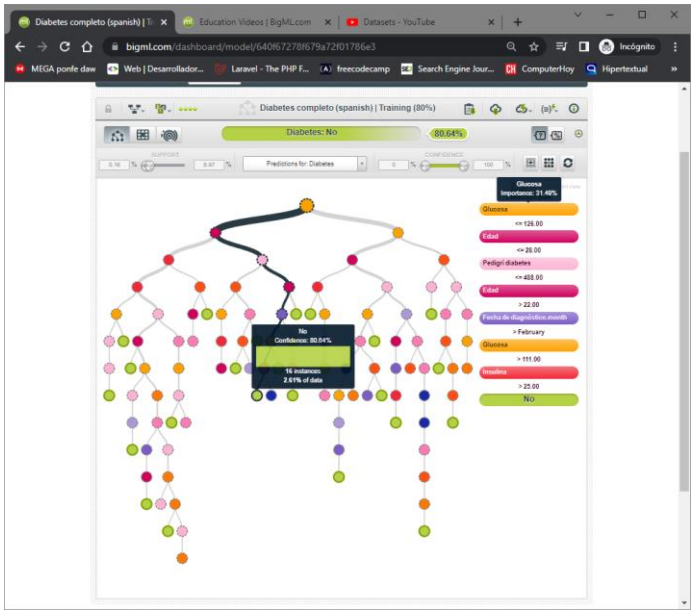
- Explica los principales resultados: Casos en los que haya resultado positivo o negativo con suficiente confiabilidad.

Los puntos verdes nos dan una pista de hacia donde fijar nuestra mirada para el análisis.




Diabetes completo (spanish) | Training (80%)
Diabetes: Sí
81.57%
Glucosa
Importance: 31.40%
Sí
Confidence: 81.57%
17 instances
2.77% of data

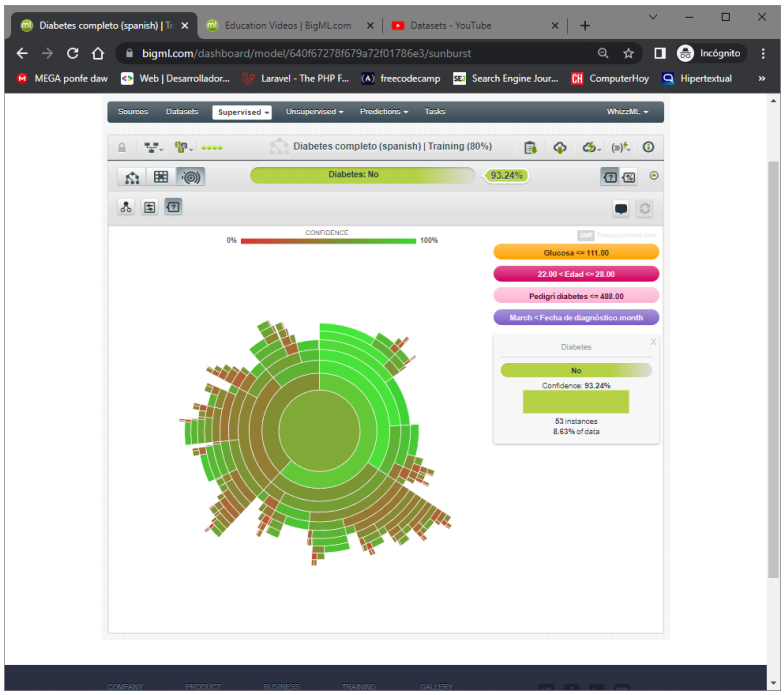

Diabetes completo (spanish) | Training (80%)
Diabetes: No
81.57%
Glucosa
Importance: 31.40%
No
Confidence: 81.57%
17 instances
2.77% of data

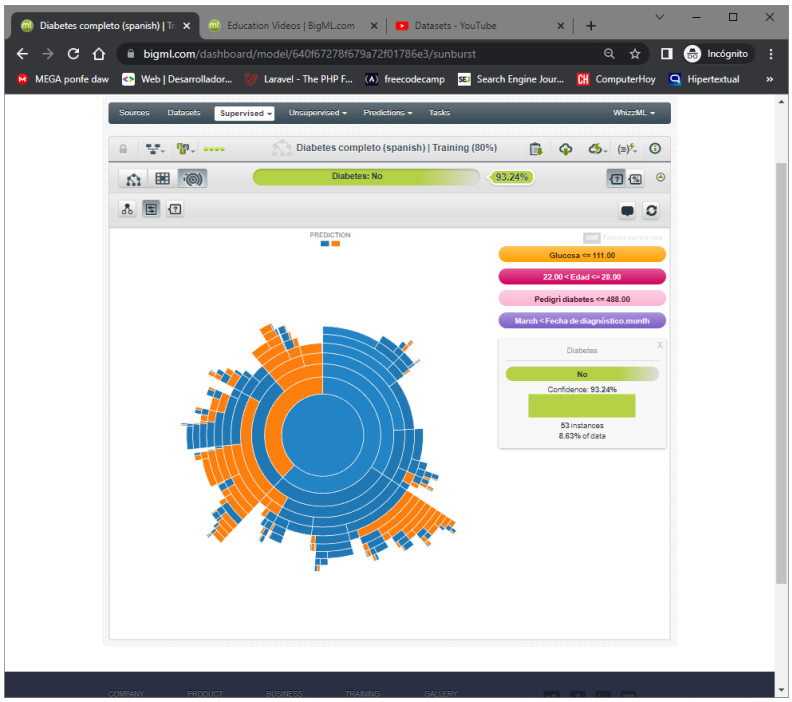

Diabetes completo (spanish) | Training (80%)
Diabetes: No
80.64%
Glucosa
Importance: 31.40%
No
Confidence: 80.64%
16 instances
2.61% of data

Revisados, podemos deducir que el riesgo de diagnóstico de diabetes positivo aumenta:

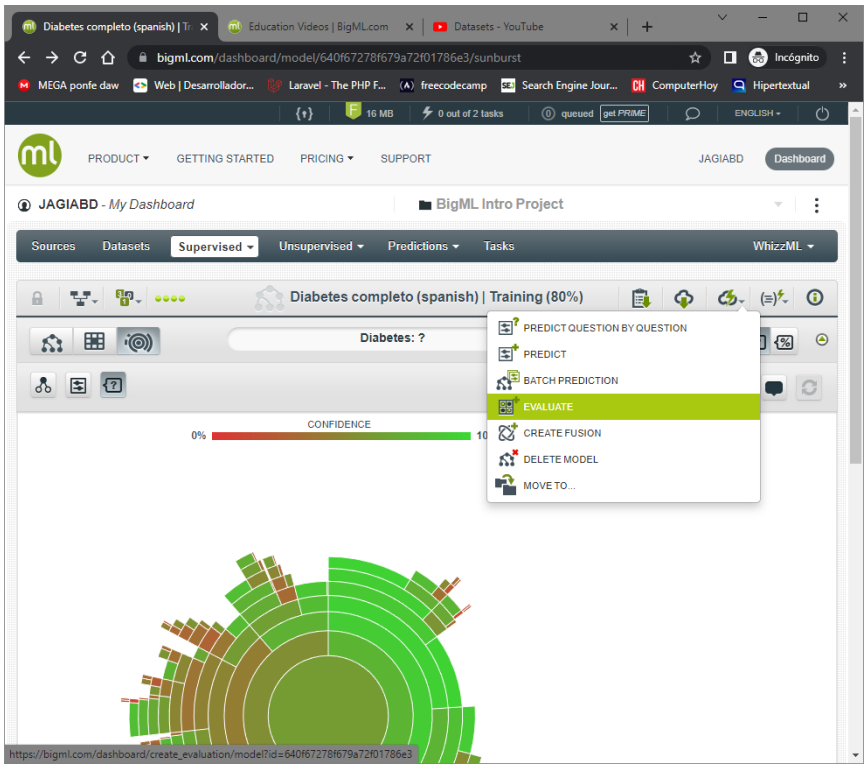
- Cuanto mayor es el porcentaje de antecedentes familiares o pedigrí.
- Entre los 28 y los 59 años. Con índices de glucosa superiores a 126.
- Con pliegues cutáneos de entre 16 y 32mm.
- IMC de obesidad mayores de 32
- Cuestiones como la presión sanguínea o los niveles de insulina no aportan por su datos una relevancia en el diagnóstico para este **`dataset`**

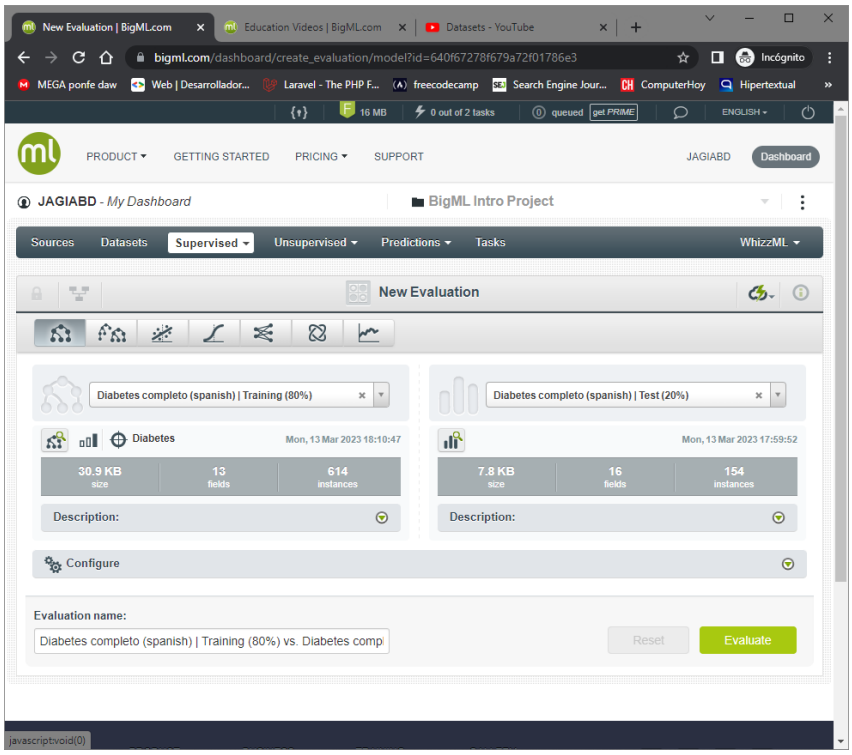
- Incorpora capturas de pantalla de los diagramas de confiabilidad (**`confidence`**) y predicción (**`prediction`**).



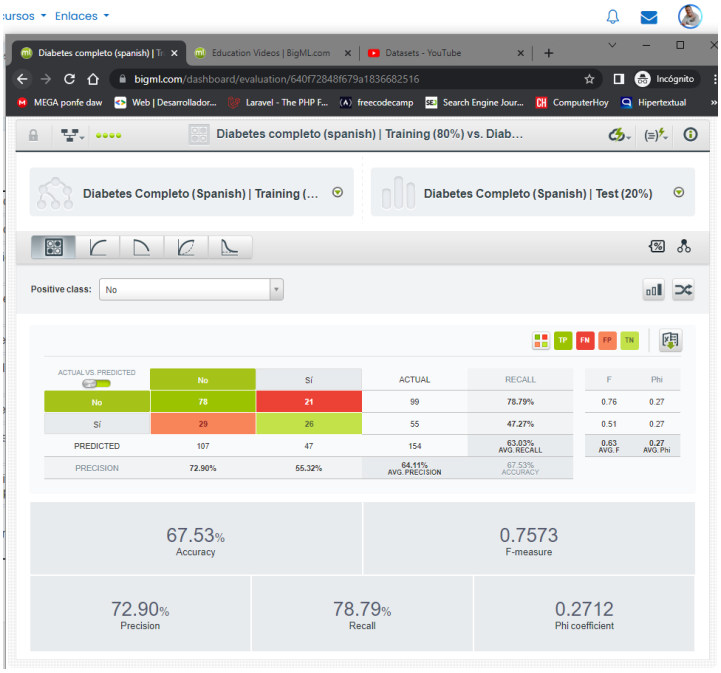


- **Apartado 5: Evaluación:**
    - Incorpora una captura de pantalla en la que se muestre la evaluación del modelo entrenado, realizada con el **`dataset`** reservado en el apartado 3.





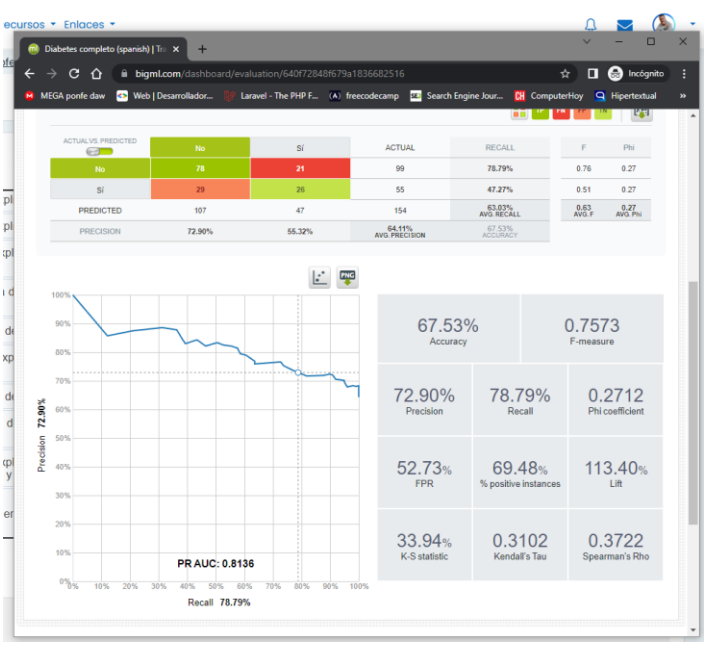
Tras hacer clic sobre el botón **`Evaluate`**, podemos observar los resultados.



- Explica el resultado de dicha evaluación, indicando el nivel de confianza obtenido (**`Accuracy`**) y el nivel de precisión (**`Precision`**).

  El hecho de tener un **`dataset`** muy pequeño puede llevarnos a obtener conclusiones engañosas respecto a los datos obtenidos y su grado de "positividad".

  Aun así, nos entrega unos valores de precisión del 72.90% y una fiabilidad o nivel de confianza de 67.53%.



  Otros valores reseñables durante su análisis son el 52.73% de falsos positivos o el 69.48% de instancias positivas.