Curso de Especialización de
Inteligencia Artificial y Big Data (IABD)

Sistemas de Aprendizaje Automático

UD05. Evaluación de un modelo a través de la matriz de confusión.
Tarea Online.

JUAN ANTONIO GARCÍA MUELAS

INDICE

Pag

1. Apartado 1: Realiza una predicción por lote en BigML 2
2. Apartado 2: Calcula la matriz de confusión 4
3. Apartado 3: Aplica la técnica de aprendizaje no supervisado de Detección de Anomalías 5

Tarea para SAA05

Título de la tarea: Evaluación de un modelo a través de la matriz de confusión.

Ciclo formativo y módulo: Curso especialización en Inteligencia Artificial y Big Data - Sistemas de Aprendizaje Automático

Curso académico: 2022-2023

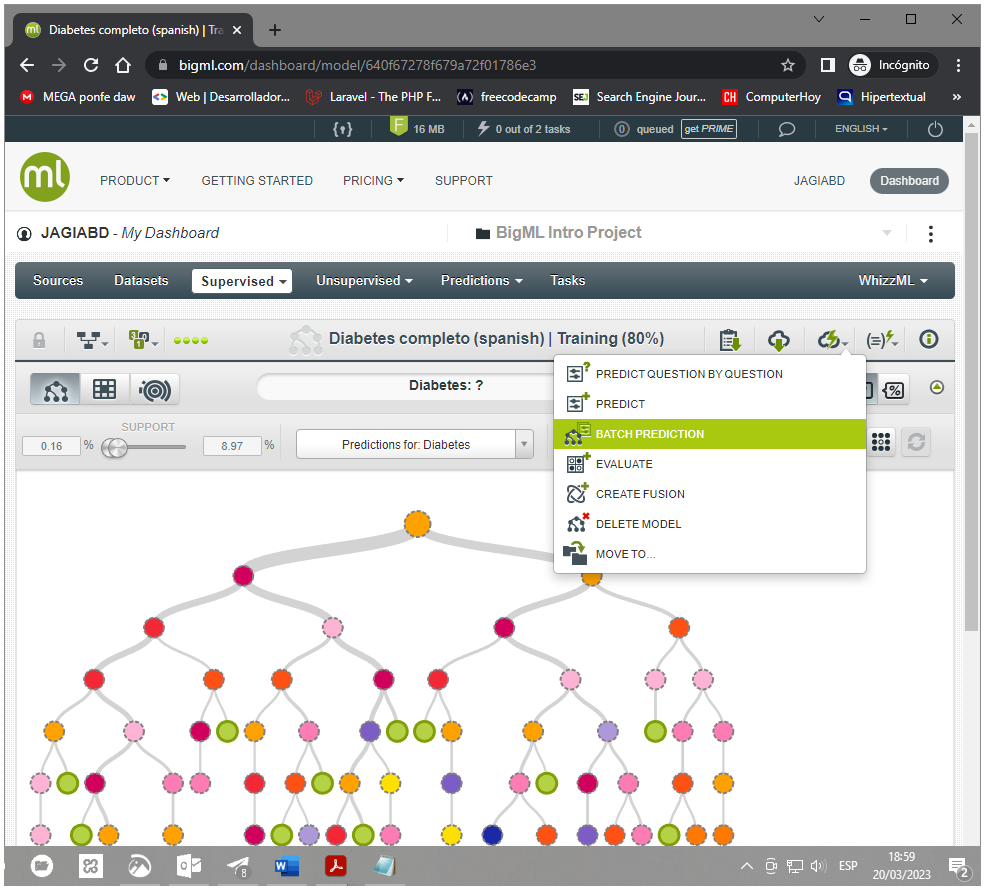
¿Qué te pedimos que hagas?

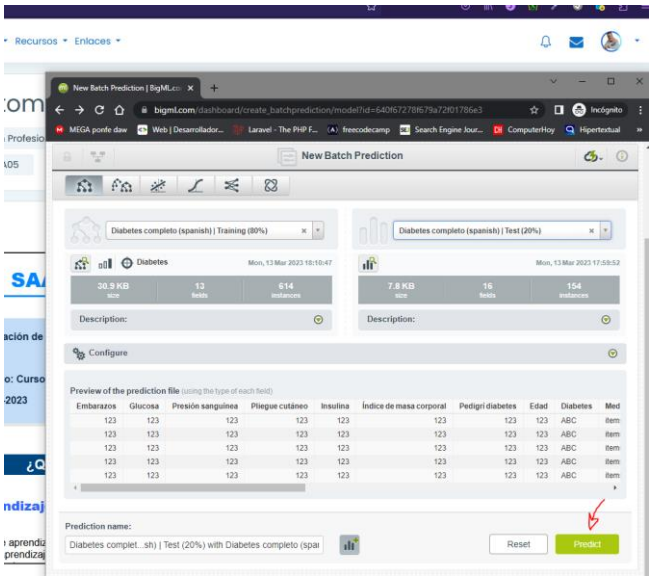
Utiliza uno de los modelos que ya has entrenado en BigML en las unidades anteriores, y evalúa los resultados de las predicciones sobre los datos de test utilizando la matriz de confusión.

- ✓ **Apartado 1: Realiza una predicción por lote en BigML**
  - Elige un dataset para clasificación binaria de los que vienen por defecto en BigML o carga uno que te parezca interesante.
  - Separa los datos en 80% para el entrenamiento y 20% para test.
  - Entrena un modelo de árbol de decisión.

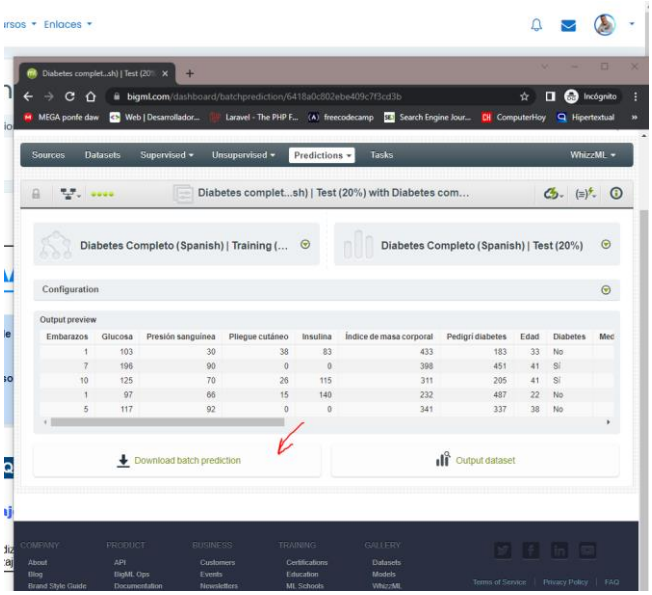
    Dado que en enunciado nos permite usar un **dataset** de los entrenados en unidades anteriores, retomo el **dataset** de la tarea anterior, al que le hemos separado los datos y entrenado con un árbol de decisión.
  - Realiza la predicción por lotes, seleccionando el conjunto de datos de test. Descarga el archivo csv resultante.

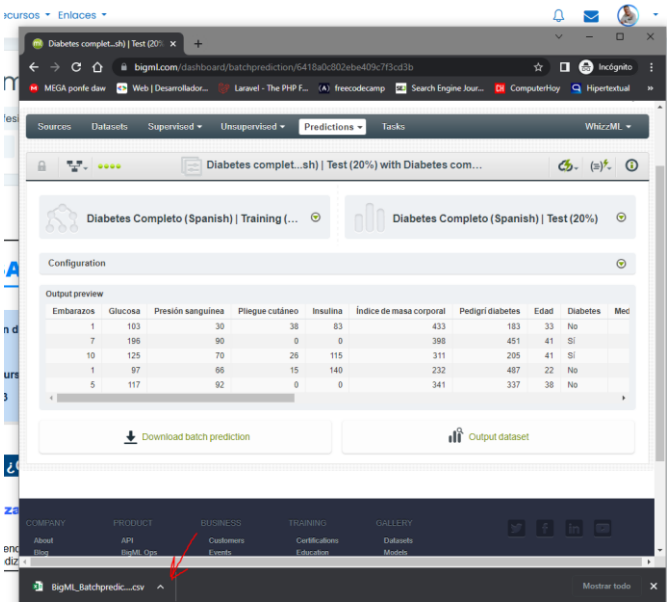
    Realizamos la predicción por lotes.





Tras generarla, descargo los datos en un archivo **`.csv`**

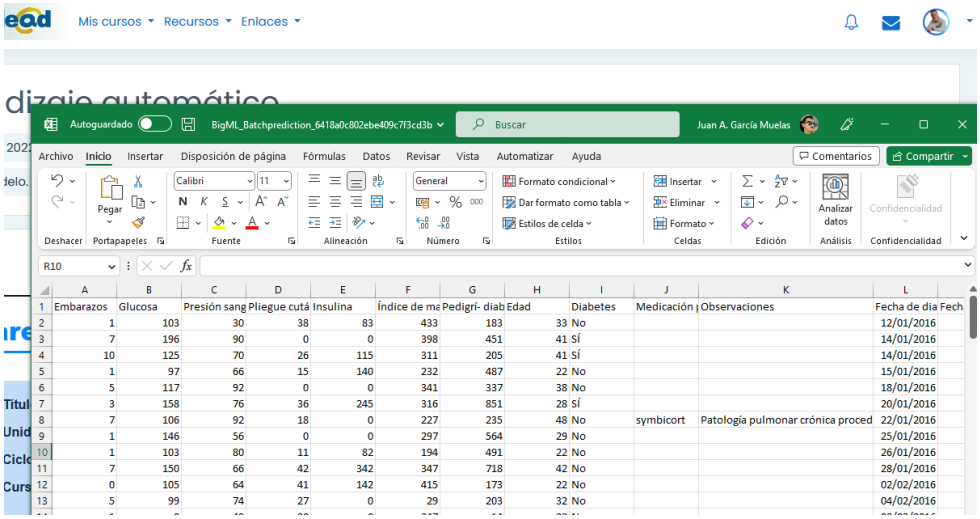




✓ Apartado 2: Calcula la matriz de confusión.

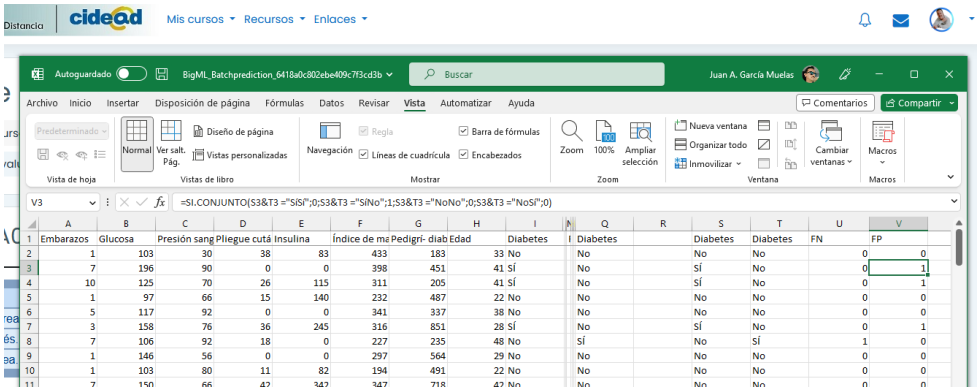
➢ Abre el archivo csv en una hoja de cálculo y aplica las fórmulas necesarias para obtener: errores totales, falsos negativos y falsos positivos.

Abro el archivo descargado y le doy formato



	Embarazos	Glucosa	Presión sang	Pliegue cutá	Insulina	Índice de ma	Pedigrí- diab	Edad	Diabetes	Medicación	Observaciones	Fecha de dia	Fech
2	1	103	30	38	83	433	183	33	No			12/01/2016	
3	7	196	90	0	0	398	451	41	SÍ			14/01/2016	
4	10	125	70	26	115	311	205	41	SÍ			14/01/2016	
5	1	97	66	15	140	232	487	22	No			15/01/2016	
6	5	117	92	0	0	341	337	38	No			18/01/2016	
7	3	158	76	36	245	316	851	28	SÍ			20/01/2016	
8	7	106	92	18	0	227	235	48	No	symbicort	Patología pulmonar crónica proced	22/01/2016	
9	1	146	56	0	0	297	564	29	No			25/01/2016	
10	1	103	80	11	82	194	491	22	No			26/01/2016	
11	7	150	66	42	342	347	718	42	No			28/01/2016	
12	0	105	64	41	142	415	173	22	No			02/02/2016	
13	5	99	74	27	0	29	203	32	No			04/02/2016	

Creo nuevas columnas para calcular datos como Falsos Negativos (**FN**) y Falsos Positivos (**FP**), mediante condicionales adaptados a la respuesta.



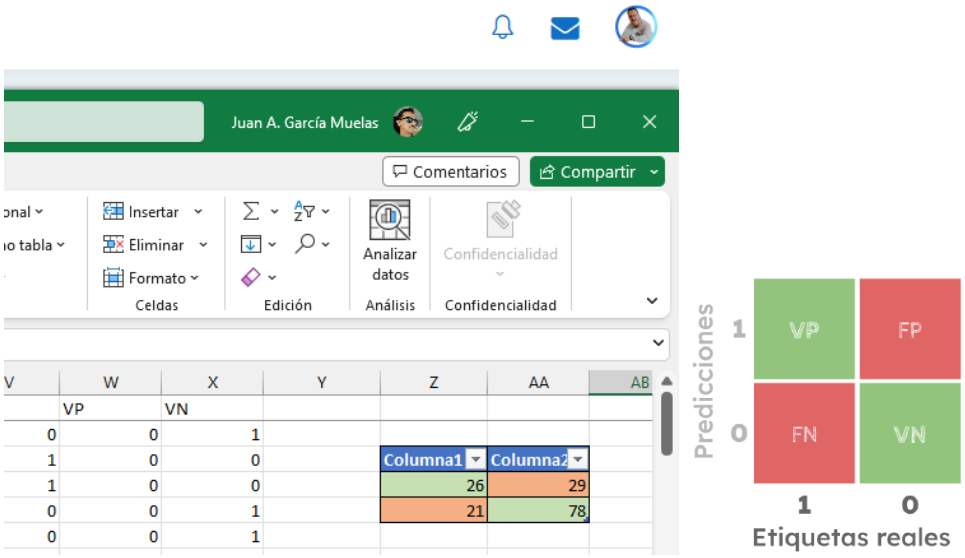
	Embarazos	Glucosa	Presión sang	Pliegue cutá	Insulina	Índice de ma	Pedigrí- diab	Edad	Diabetes	Diabetes	Diabetes	Diabetes	FN	FP
2	1	103	30	38	83	433	183	33	No	No	No	No	0	0
3	7	196	90	0	0	398	451	41	SÍ	No	SÍ	No	0	1
4	10	125	70	26	115	311	205	41	SÍ	No	SÍ	No	0	1
5	1	97	66	15	140	232	487	22	No	No	No	No	0	0
6	5	117	92	0	0	341	337	38	No	No	No	No	0	0
7	3	158	76	36	245	316	851	28	SÍ	No	SÍ	No	0	1
8	7	106	92	18	0	227	235	48	No	SÍ	No	SÍ	1	0
9	1	146	56	0	0	297	564	29	No	No	No	No	0	0
10	1	103	80	11	82	194	491	22	No	No	No	No	0	0

Q	R	S	T	U	V
Diabetes		Diabetes	Diabetes	FN	FP
No		No	No	0	0
No		No	No	0	0
				21	29

FN = 21, FP= 29 ET= 50

➢ Construye la matriz de confusión, rellenando los valores correspondientes.

Añado una pequeña tabla donde incluir los datos extraídos.



➢ Analiza los resultados. ¿Es fiable el modelo?

Ha acertado 104 de 154, por lo que obtiene una **`Exactitud`** del 0.675.

El **`Recall`** (26/(26+21)) es de 0.553

La **`Precisión`** obtenida (26/(26+29)) es de 0.473

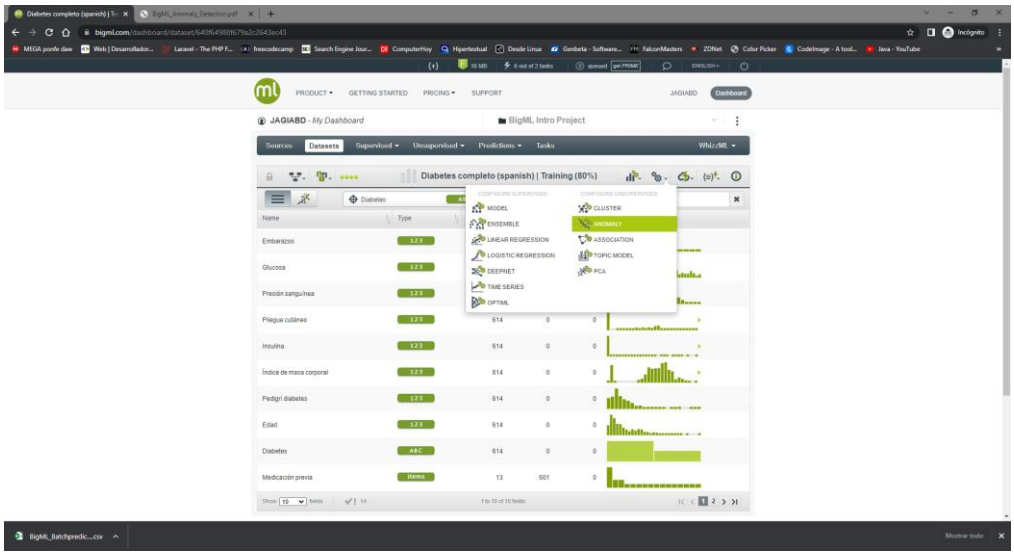
Visto que hay un dato sensiblemente mayor, nos fiaremos del **`F1 score`**:

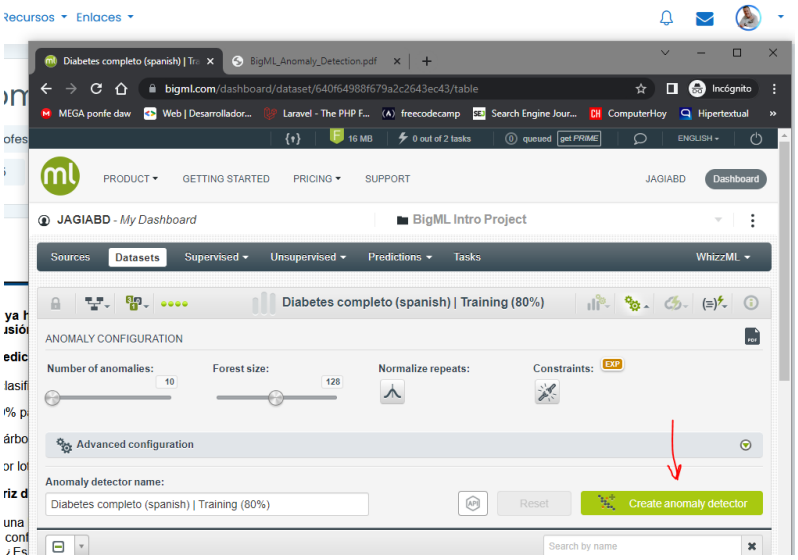
F1= (2*0.473*0.553)/(0.473+0.553)=0.523/0.996 =0.525, observando un score bastante bajo, por lo que aunque fiable, es mejorable.

✓ **Apartado 3: Aplica la técnica de aprendizaje no supervisado de Detección de Anomalías.**

➢ Aplica el modelo de detección de anomalías en BigML dentro de las funciones rápidas de algoritmos no supervisados.

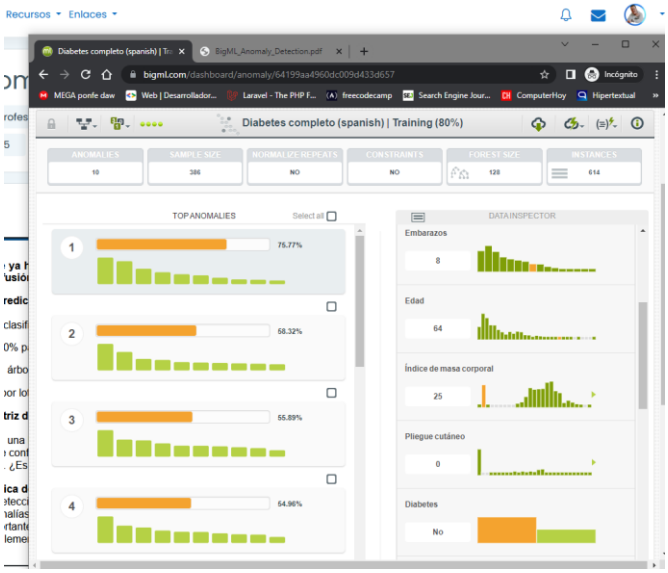
Accedo desde el menú:

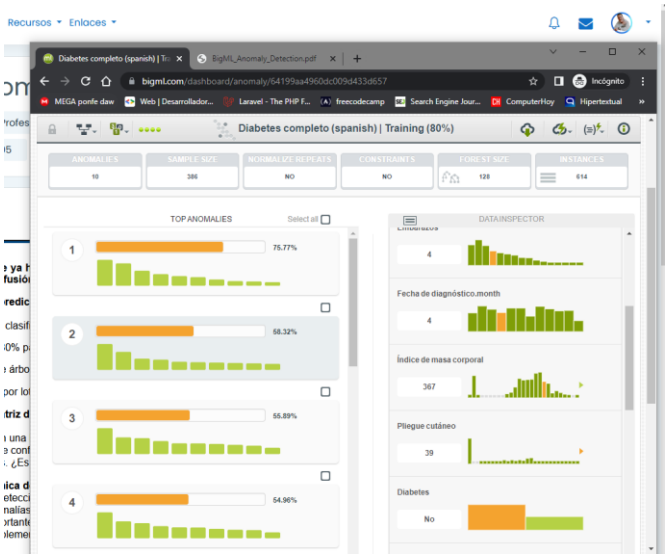


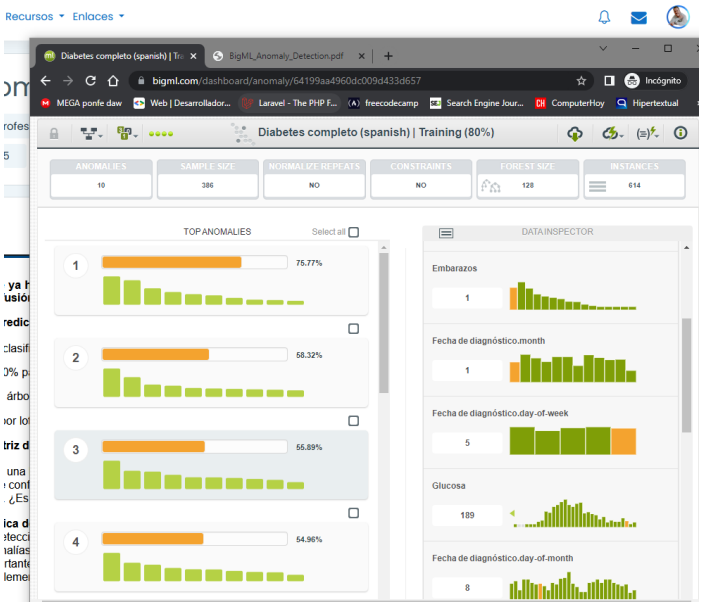


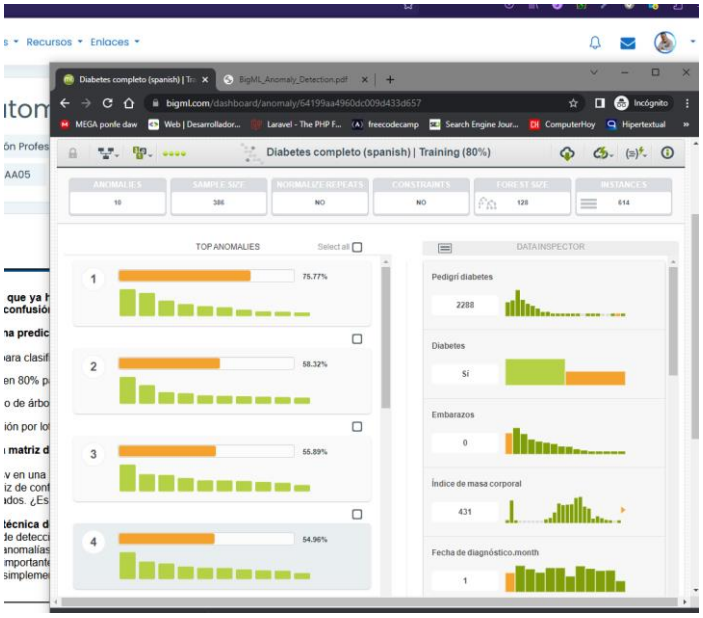
- Analiza las top 5 anomalías de tu problema y decide si merece la pena analizarlas a parte.

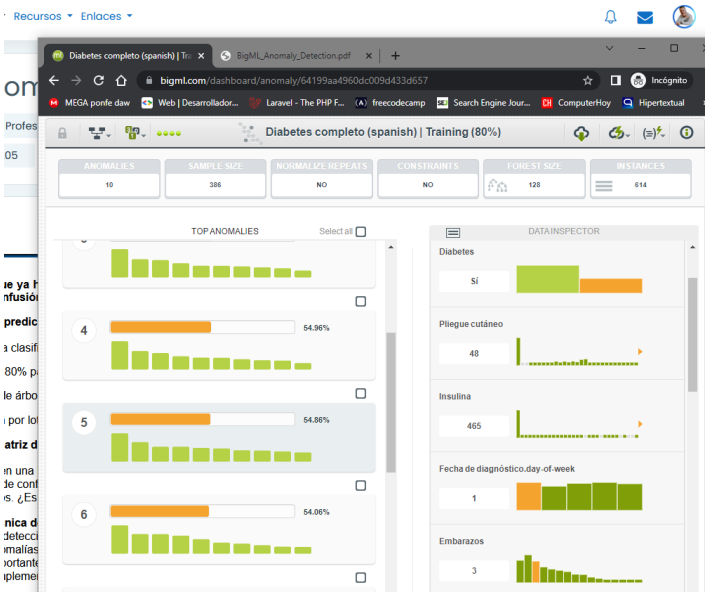
Abro las cinco vistas para poder observar los resultados






Recursos
Enlaces
Diabetes completo (spanish) | Training (80%)
ANOMALIES
SAMPLE SIZE
NORMALIZE REPEATS
CONSTRAINTS
FOREST SIZE
INSTANCES
TOP ANOMALIES
DATAINSPECTOR
Embarazos
Fecha de diagnóstico.month
Fecha de diagnóstico.day-of-week
Glucosa
Fecha de diagnóstico.day-of-month


Recursos
Enlaces
Diabetes completo (spanish) | Training (80%)
TOP ANOMALIES
DATAINSPECTOR
Pedigrí diabetes
Diabetes
Embarazos
Índice de masa corporal
Fecha de diagnóstico.month

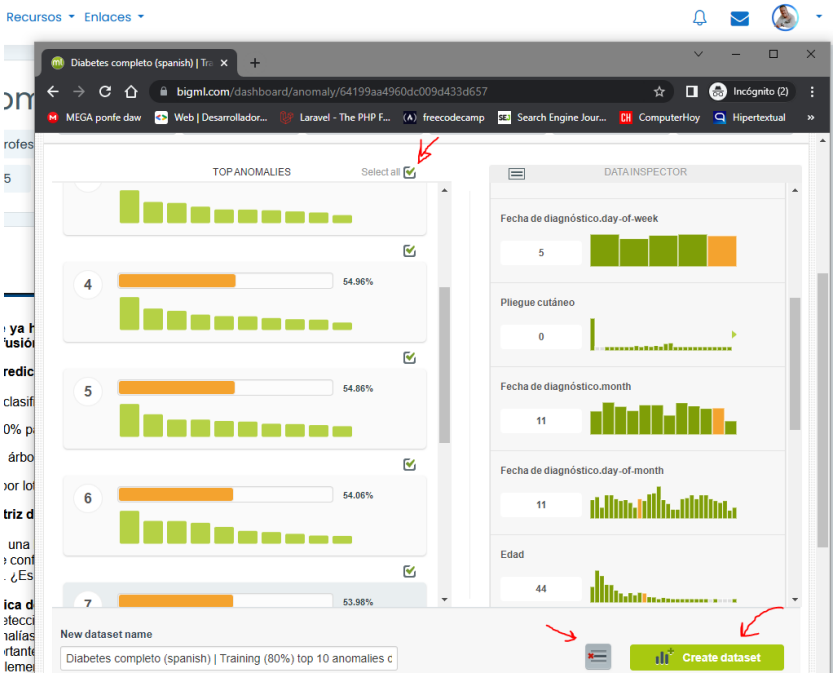
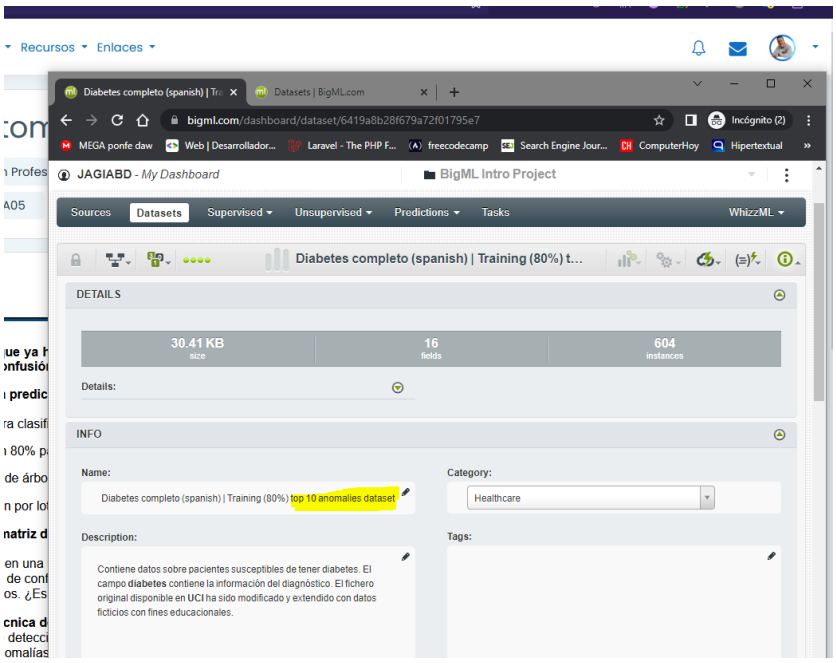

Recursos
Enlaces
Diabetes completo (spanish) | Training (80%)
TOP ANOMALIES
DATAINSPECTOR
Diabetes
Pliegue cutáneo
Insulina
Fecha de diagnóstico.day-of-week
Embarazos

- Si crees que son importantes, crea un **dataset** con ellas para analizarlas

Guiándome por la documentación voy a prestar una mayor atención a las dos primeras, que han obtenido un **score** de entorno al 60% o superior, haciéndolas viables para su análisis.

Sin embargo, tras revisar los datos concretos que arrojan y viendo el detalle de las anomalías detectadas y su impacto, creo que no son lo suficientemente relevantes (en la mayoría de los casos están por debajo del 5%), por lo que avanzaré al siguiente punto de la tarea.

- Si crees que son simplemente errores de medida, crea un **dataset** sin ellas.



Selecciono las anomalías, hago clic para eliminarlas y creo el **dataset**, pudiendo ya verlo entre todos los creados.