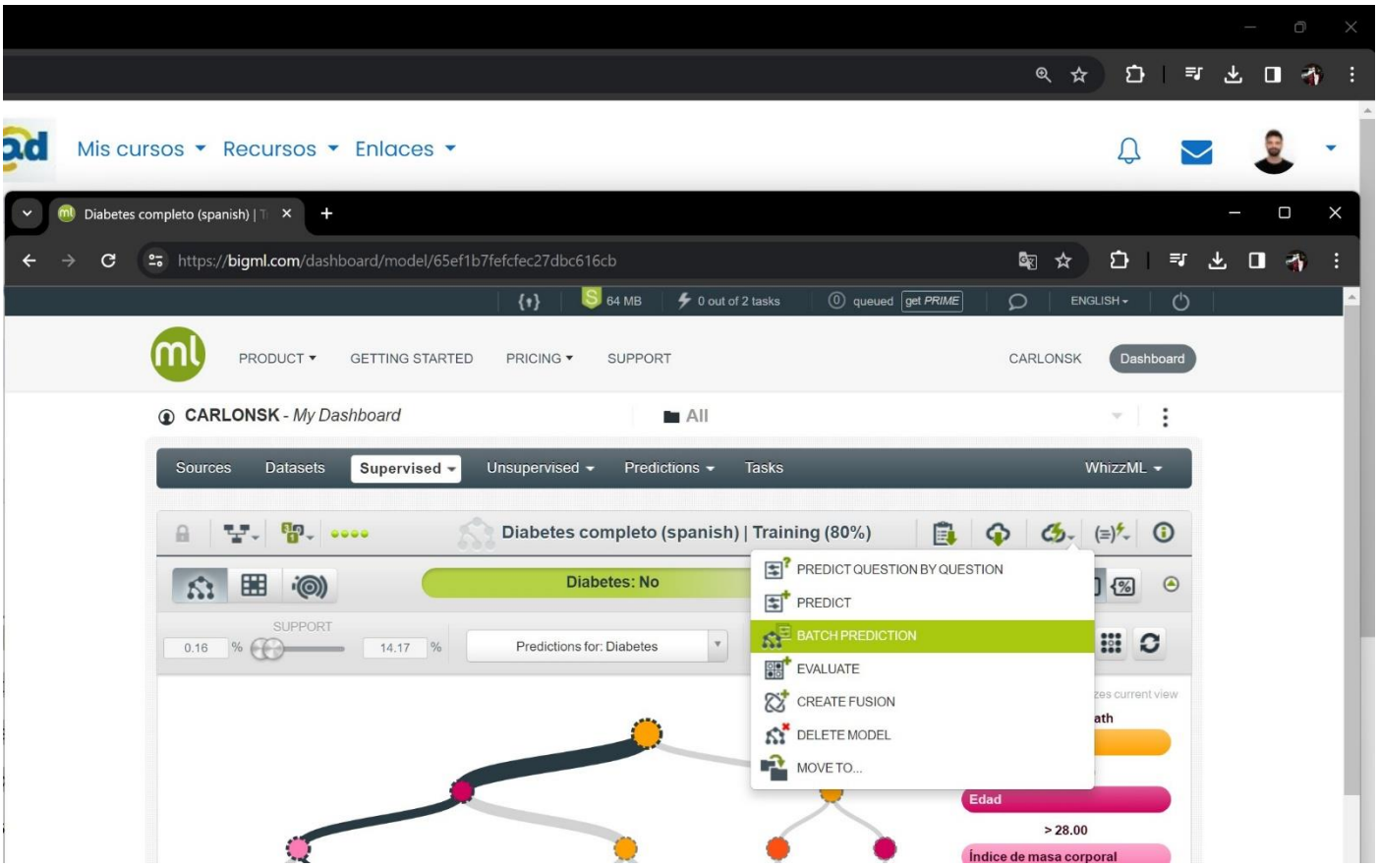
TAREA SISTEMAS DE APRENDIZAJE AUTOMATICO 05

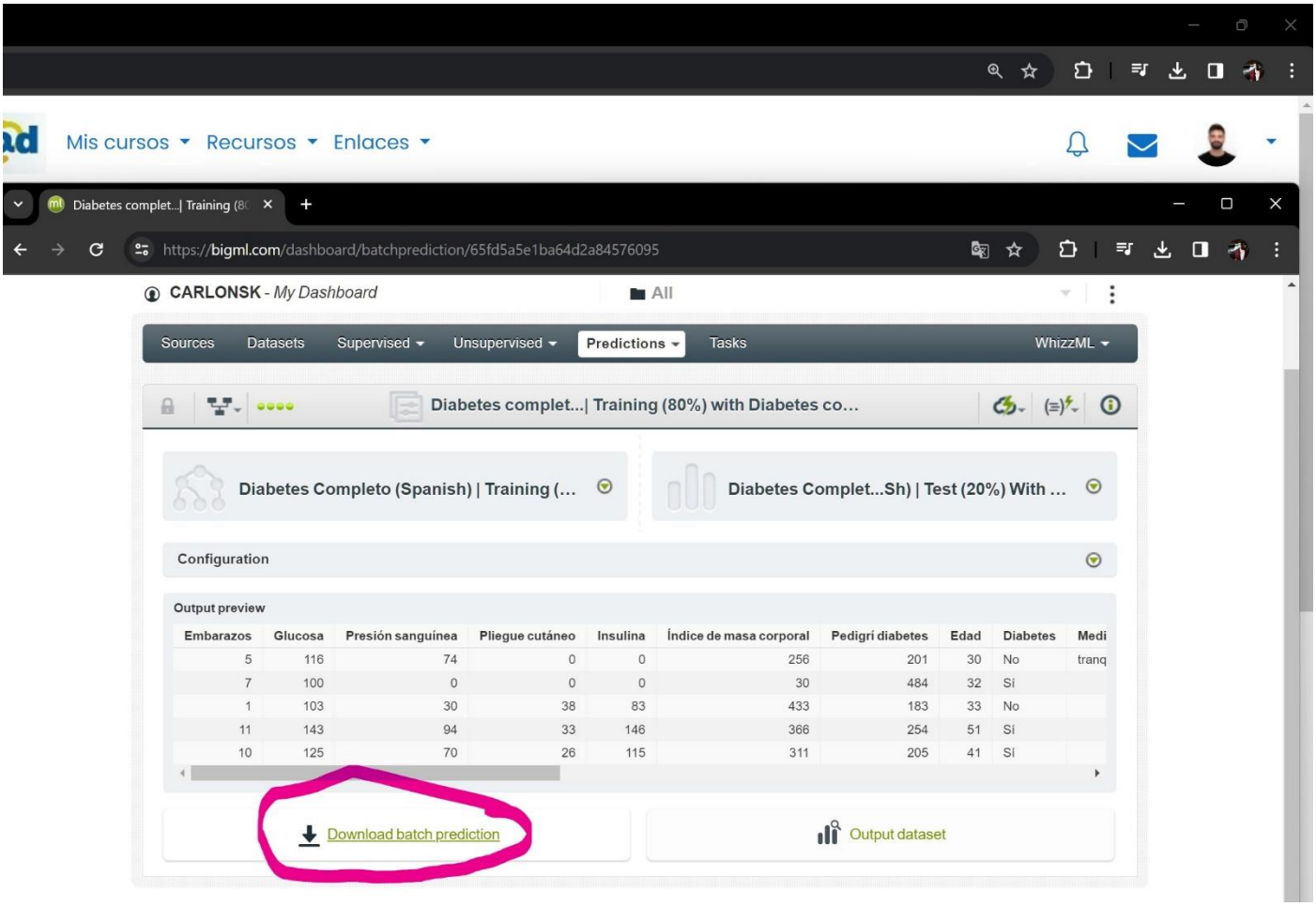
Utiliza uno de los modelos que ya has entrenado en BigML en las unidades anteriores, y evalúa los resultados de las predicciones sobre los datos de test utilizando la matriz de confusión.

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.*
- *Realiza la predicción por lotes, seleccionando el conjunto de datos de test. Descarga el archivo csv resultante.*

Escojo para el ejercicio el dataset de diabetes, ya entrenado en la tarea anterior. Realizo la predicción por lotes con los datos de test y descargo el archivo.

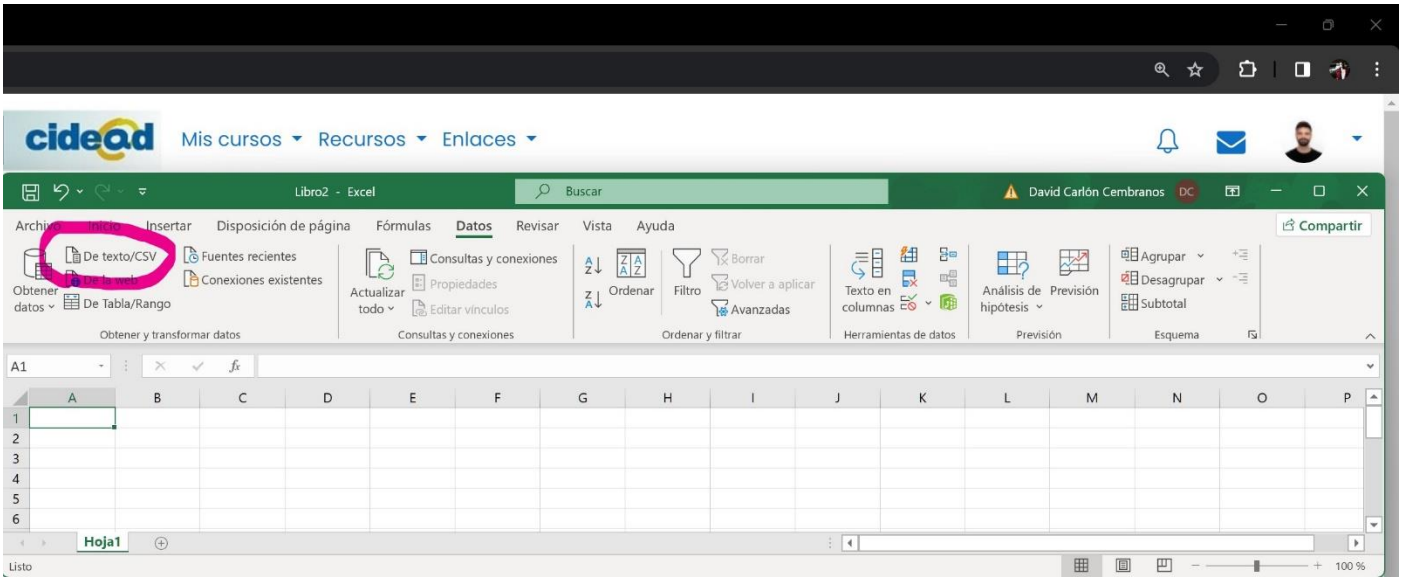




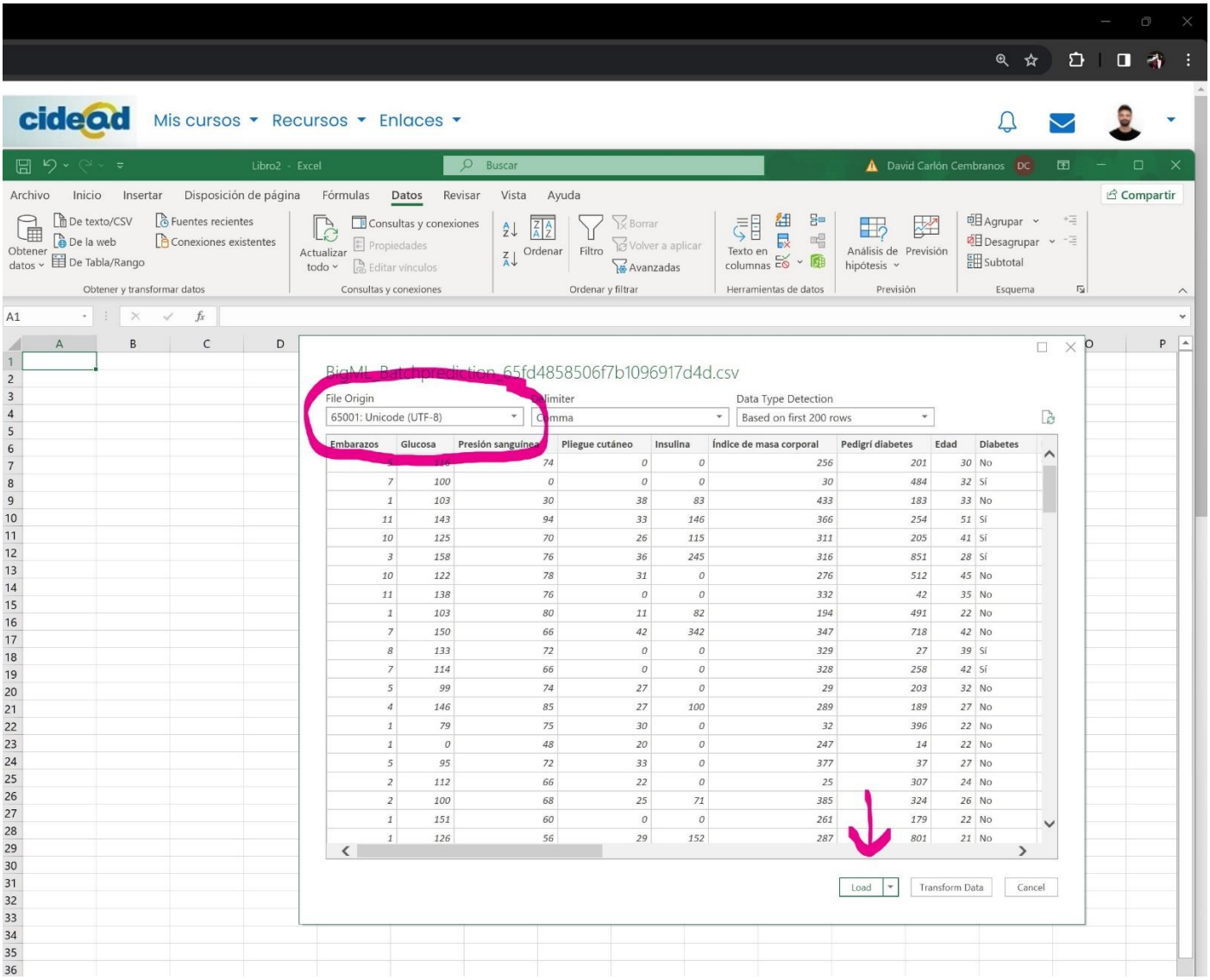
Apartado 2: Calcula la matriz de confusión.

- *Abre el archivo csv en una hoja de cálculo y aplica las fórmulas necesarias para obtener: errores totales, falsos negativos y falsos positivos.*
- *Construye la matriz de confusión, rellenando los valores correspondientes.*
- *Analiza los resultados. ¿Es fiable el modelo?*

Abro el archivo csv con Excel, a través de la opción de datos de un .csv.



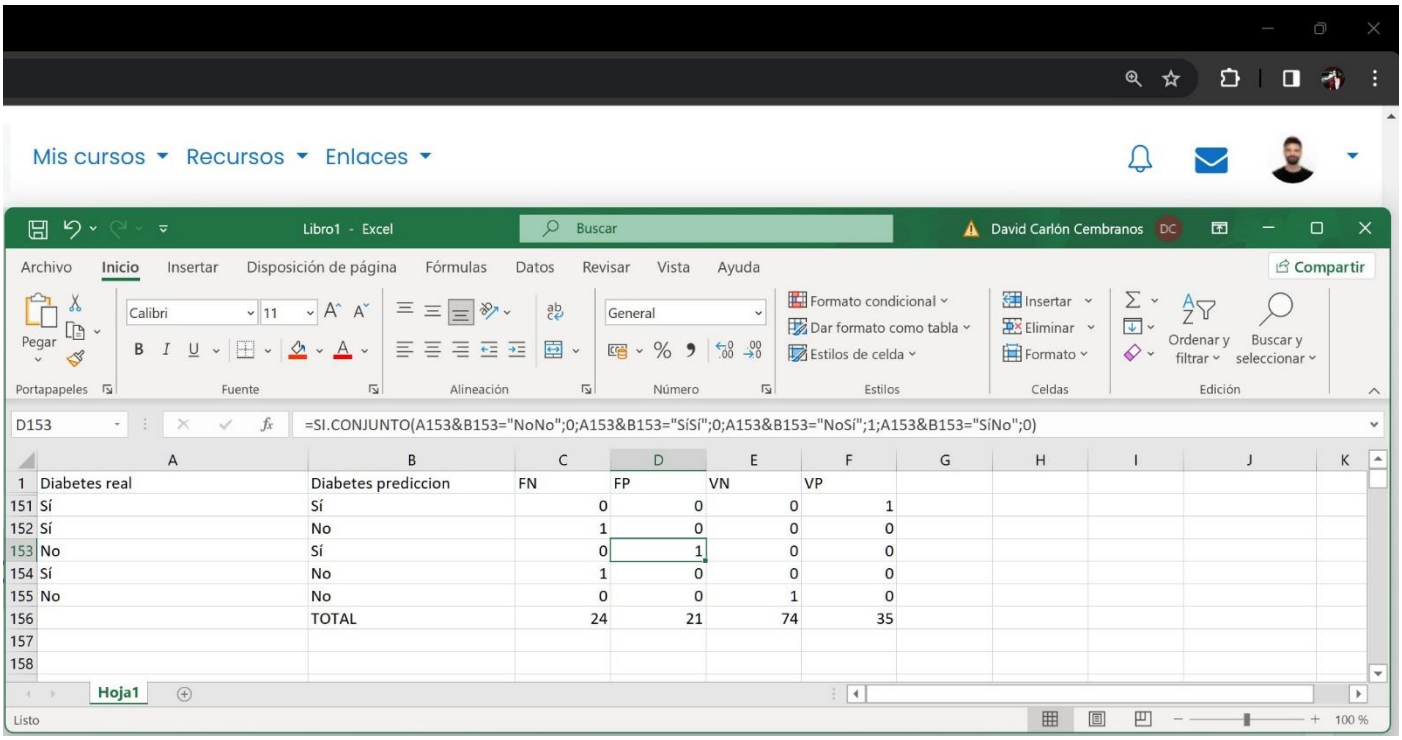
Cargo los datos en formato Unicode (UTF-8) para que los datos tengan, aparezcan bien distribuidos en filas y columnas, y bien escritos.



Embarazos	Glucosa	Presión sanguínea	Pliegue cutáneo	Insulina	Índice de masa corporal	Pedigrí diabetes	Edad	Diabetes
7	100	0	0	0	30	484	32	Sí
1	103	30	38	83	433	183	33	No
11	143	94	33	146	366	254	51	Sí
10	125	70	26	115	311	205	41	Sí
3	158	76	36	245	316	851	28	Sí
10	122	78	31	0	276	512	45	No
11	138	76	0	0	332	42	35	No
1	103	80	11	82	194	491	22	No
7	150	66	42	342	347	718	42	No
8	133	72	0	0	329	27	39	Sí
7	114	66	0	0	328	258	42	Sí
5	99	74	27	0	29	203	32	No
4	146	85	27	100	289	189	27	No
1	79	75	30	0	32	396	22	No
1	0	48	20	0	247	14	22	No
5	95	72	33	0	377	37	27	No
2	112	66	22	0	25	307	24	No
2	100	68	25	71	385	324	26	No
1	151	60	0	0	261	179	22	No
1	126	56	29	152	287	801	21	No

Con los datos en Excel, distribuidos en filas y columnas, me quedo, solo con las dos columnas que me interesan para crear la matriz de confusión (Diabetes y su predicción). Con la función SI.CONJUNTO establezco los parámetros, para calcular los datos de la matriz:

- Falsos Negativos (FN): **24**
- Falsos Positivos (FP): **21**
- Verdaderos Negativos (VN): **74**
- Verdaderos Positivos (VP): **35**
- Casos Totales (CT): **154**



	A	B	C	D	E	F
1	Diabetes real	Diabetes prediccion	FN	FP	VN	VP
151	Sí	Sí	0	0	0	1
152	Sí	No	1	0	0	0
153	No	Sí	0	1	0	0
154	Sí	No	1	0	0	0
155	No	No	0	0	1	0
156		TOTAL	24	21	74	35
157						
158						

Creo la Matriz de confusión:

		PREDICCION 0	PREDICCION 1
REAL	1	24	35
REAL	0	74	21

Errores Totales = FN + FP = 24 + 22 = **46**

Acuracy = VN + VP / CT = 109 / 155 = **0,70**

Recall = VP / (VP +FN) = 35 / (35 + 24) = **0,59**

Precisión = VP / (VP + FP) = 35 / (35 + 21) = **0,63**

En este caso, como se destaca en la teoría del tema, la métrica más apropiada cuando tenemos un conjunto de datos no balanceado. Como es el caso que tenemos mas datos de negativos que de positivos vamos a calcular también el **F1score**:

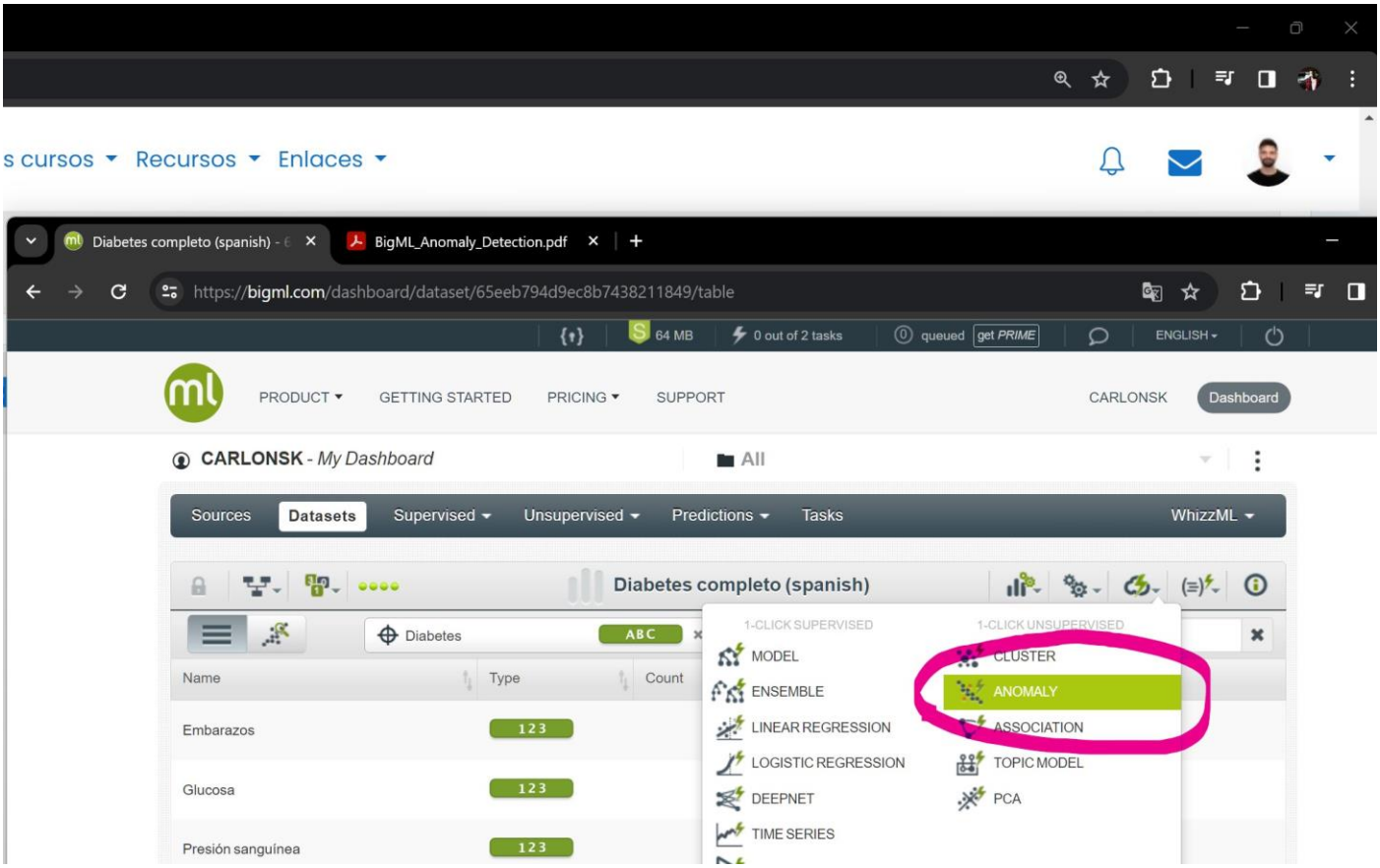
F1 Score = (2 * Precision * Recall) / (Precision + Recall) = (2 * 0,63 * 0,59) / (0,63 + 0,59) = **0,61**

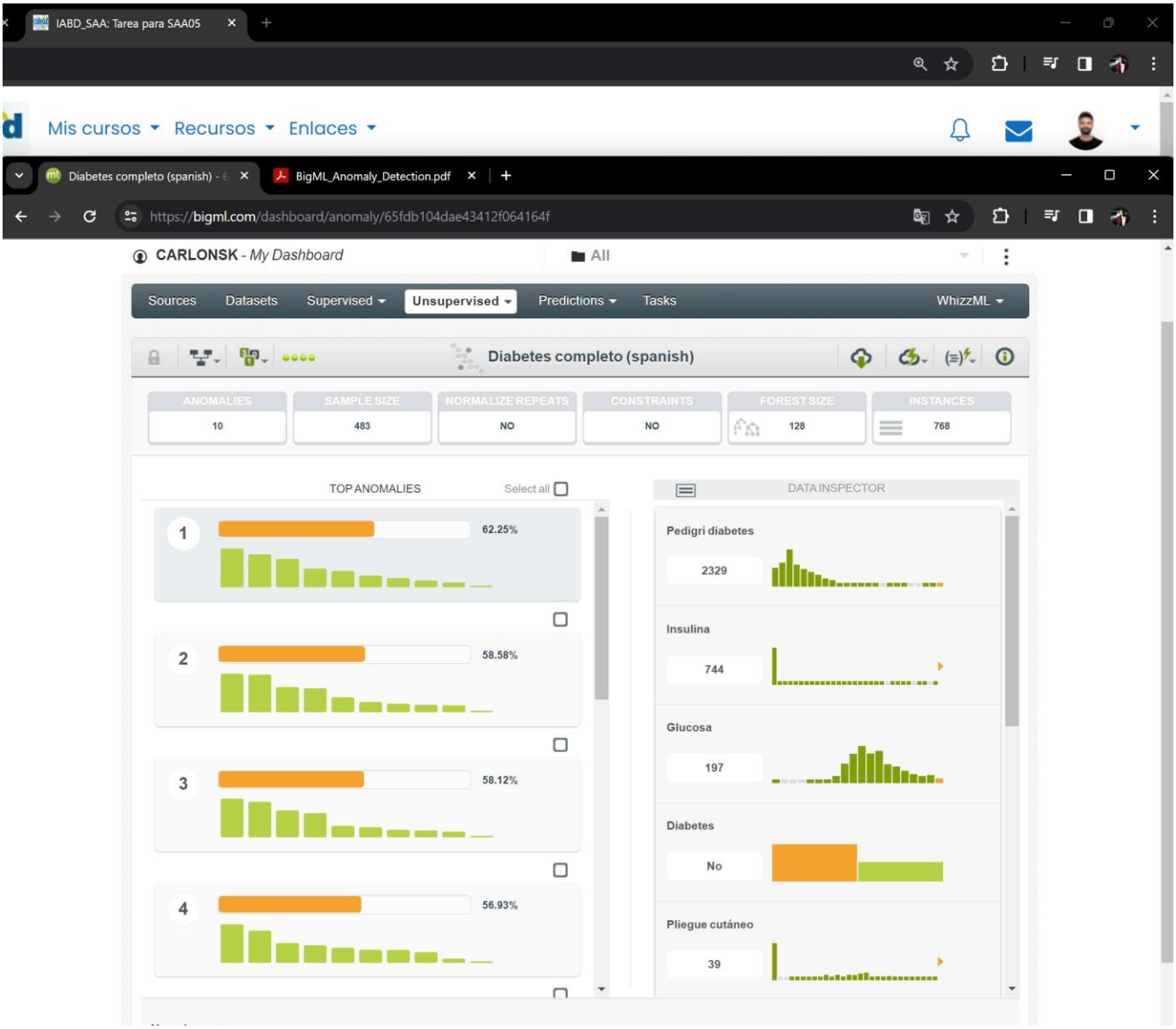
Bajo mi punto de vista el modelo un es demasiado fiable, teniendo en cuenta además que es de un ámbito médico, creo que para que pudiera usarse tendría que ser mucho más preciso.

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.*
- *Analiza las top 5 anomalías de tu problema y decide si merece la pena analizarlas a parte.*
- *Si crees que son importantes, crea un dataset con ellas para analizarlas*
- *Si crees que son simplemente errores de medida, crea un dataset sin ellas.*

Aplico el modelo al dataset completo. Y observo el top 5 anomalias.





Como dice la documentación, solo deberíamos tener en cuenta a partir de 60%, por lo tanto no es muy preocupante, dado mi poco conocimiento sobre el tema, decido que pueden ser errores de medida.

Por consecuencia los borro y genero el dataset.