

TAREA SISTEMAS DE APRENDIZAJE AUTOMATICO 04

A Eva le han pedido que entrene un modelo de Inteligencia Artificial que ayude a los médicos a predecir posibles casos de diabetes en pacientes.

Para ello le han facilitado una base de datos con el historial de las personas que han pasado por el hospital de su localidad y se les han hecho análisis de sangre para decidir después si tenían o no Diabetes.

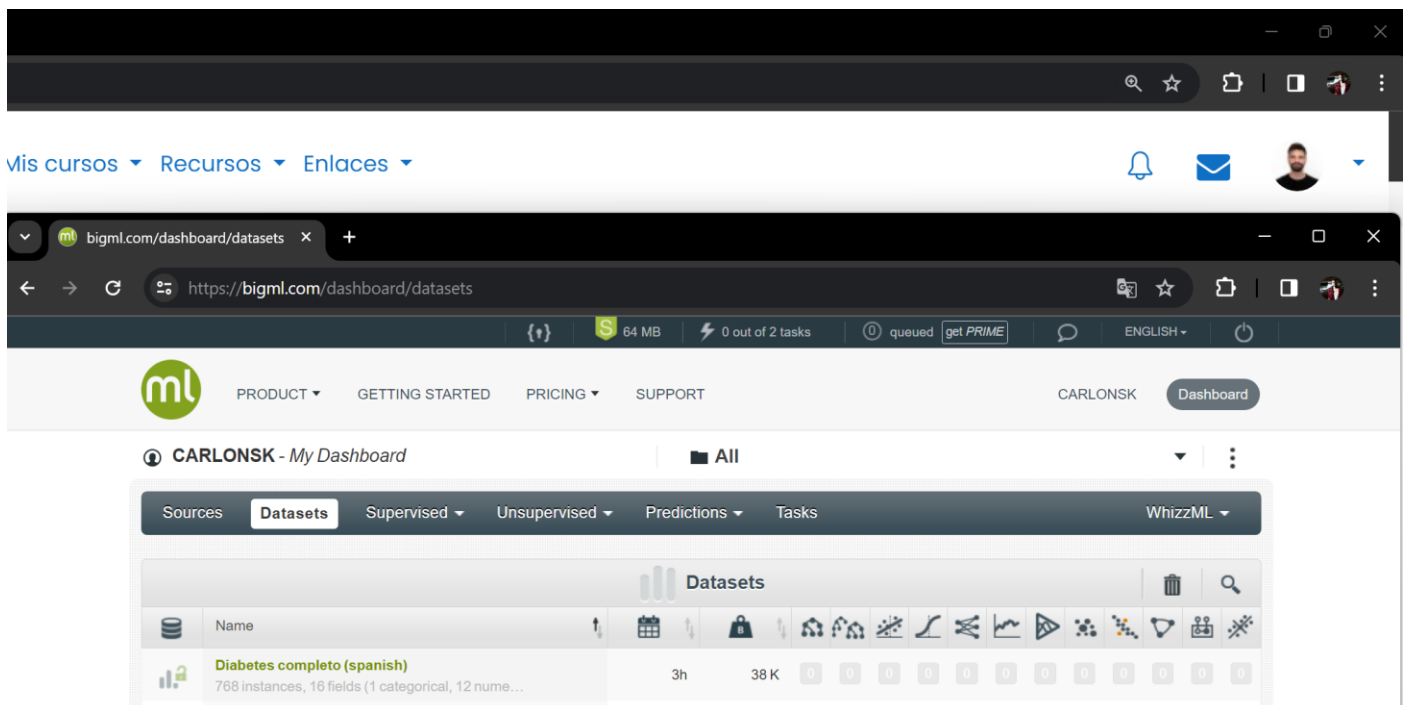
La base de datos se la han dado en un archivo .csv, y no es especialmente grande, por lo que Eva va a hacer una primera aproximación con la herramienta BigML, que le va a permitir hacer un primer prototipo y detectar posibles aspectos a revisar en la propia base de datos.

Va a utilizar un algoritmo de Árbol de Decisión, que es el que generalmente da mejores resultados en este tipo de primeras aproximaciones con bases de datos pequeñas.

Lo bueno de hacerlo en la plataforma de BigML es que también va a poder enriquecer el informe que entregará a los responsables del hospital un pequeño análisis del caso, pudiendo aportar buenas recomendaciones para el tratamiento previo de los datos que se pretendan utilizar en el desarrollo definitivo.

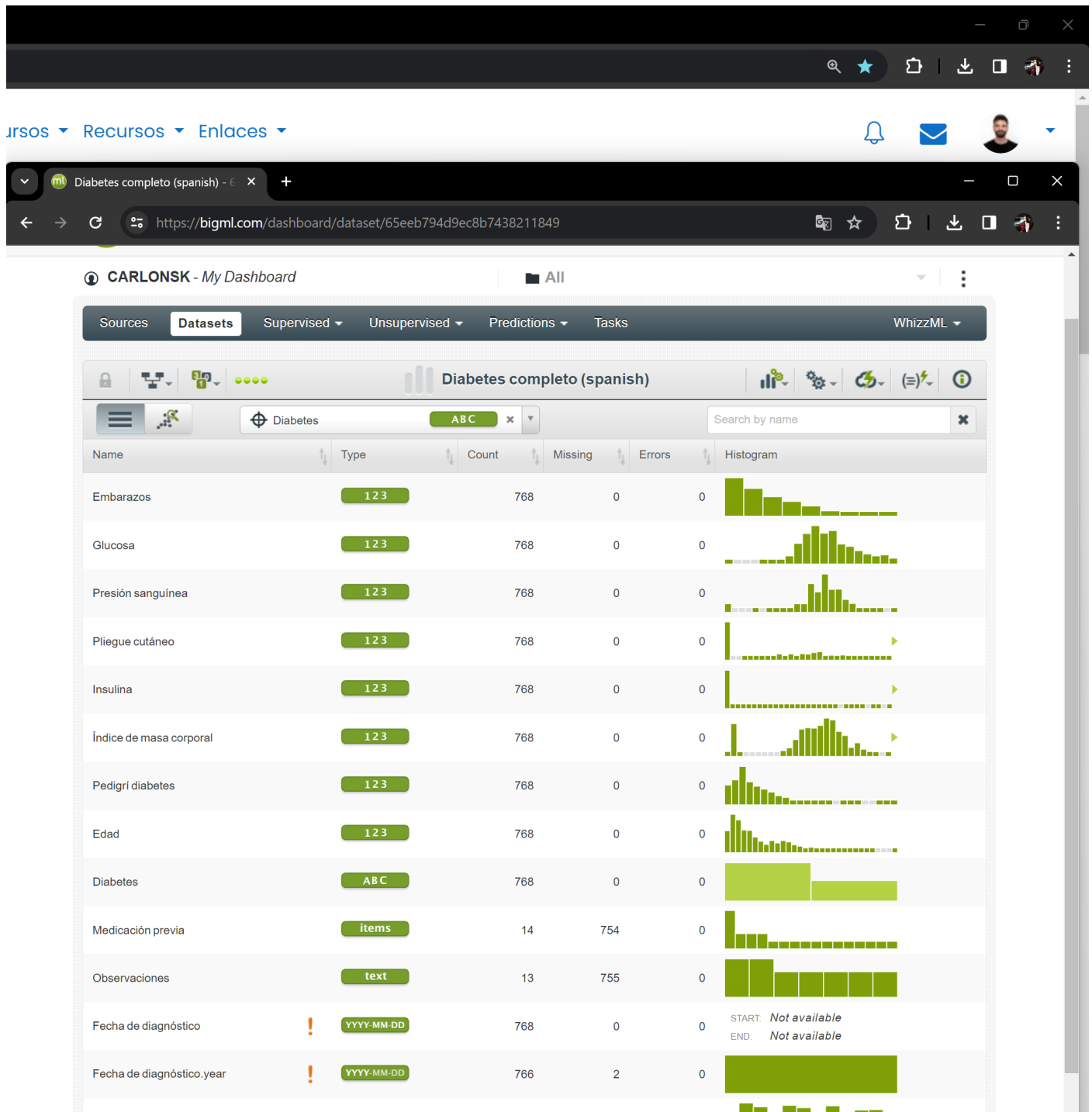
Apartado 1: Localiza el dataset "Diabetes completo (spanish)":

- Utiliza el buscador de datasets que tiene la propia plataforma para ello.
- Incorpora una captura de pantalla del dataset mencionado incorporado a tu apartado 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.
- Explica el tipo de categorías (numéricas, texto, items, categóricas...).
- Analiza los histogramas de cada categoría y comenta aquellos en los que consideres que hay algún tipo de anomalía.



El dataset cuenta con un total de **768** instancias y **16** categorías. El tipo de categorías se muestra en la columna Type. A continuación, comentare los tipos de categorías:

- Numérico, datos numéricos como enteros, decimales y porcentajes. Y se representan con en la columna Type con 123.
- Categórico, en el que tenemos Diabetes, que es el dato objetivo del modelo y sirve para clasificar en categorías como colores, países y en este caso SI/NO.
- Texto, en este caso Observaciones y Medicación previa, que son como su nombre indica texto.
- Fecha, la fecha en diferentes formatos.

Medicación previa	!	items	14	754	0	
Observaciones		text	13	755	0	
Fecha de diagnóstico	!	YYYY-MM-DD	768	0	0	START: <i>Not available</i> END: <i>Not available</i>
Fecha de diagnóstico.year	!	YYYY-MM-DD	766	2	0	
Fecha de diagnóstico.month	!	YYYY-MM-DD	766	2	0	
Fecha de diagnóstico.day-of-month	!	YYYY-MM-DD	766	2	0	
Fecha de diagnóstico.day-of-week	!	M T W T F S S	766	2	0	

Como se aprecia en la imagen, el campo de fecha de diagnostico no se han obtenido datos, así que BigML lo desactiva. La categoría de año son todos 2016 así que también decide que es irrelevante. Después bajo mi criterio decido desactivar las otras tres fechas, que, aunque están bien los histogramas llego a la conclusión de que son irrelevantes a la hora de entrenar el modelo, para que este sea más liviano. Todos los demás histogramas creo que están correctos.

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.

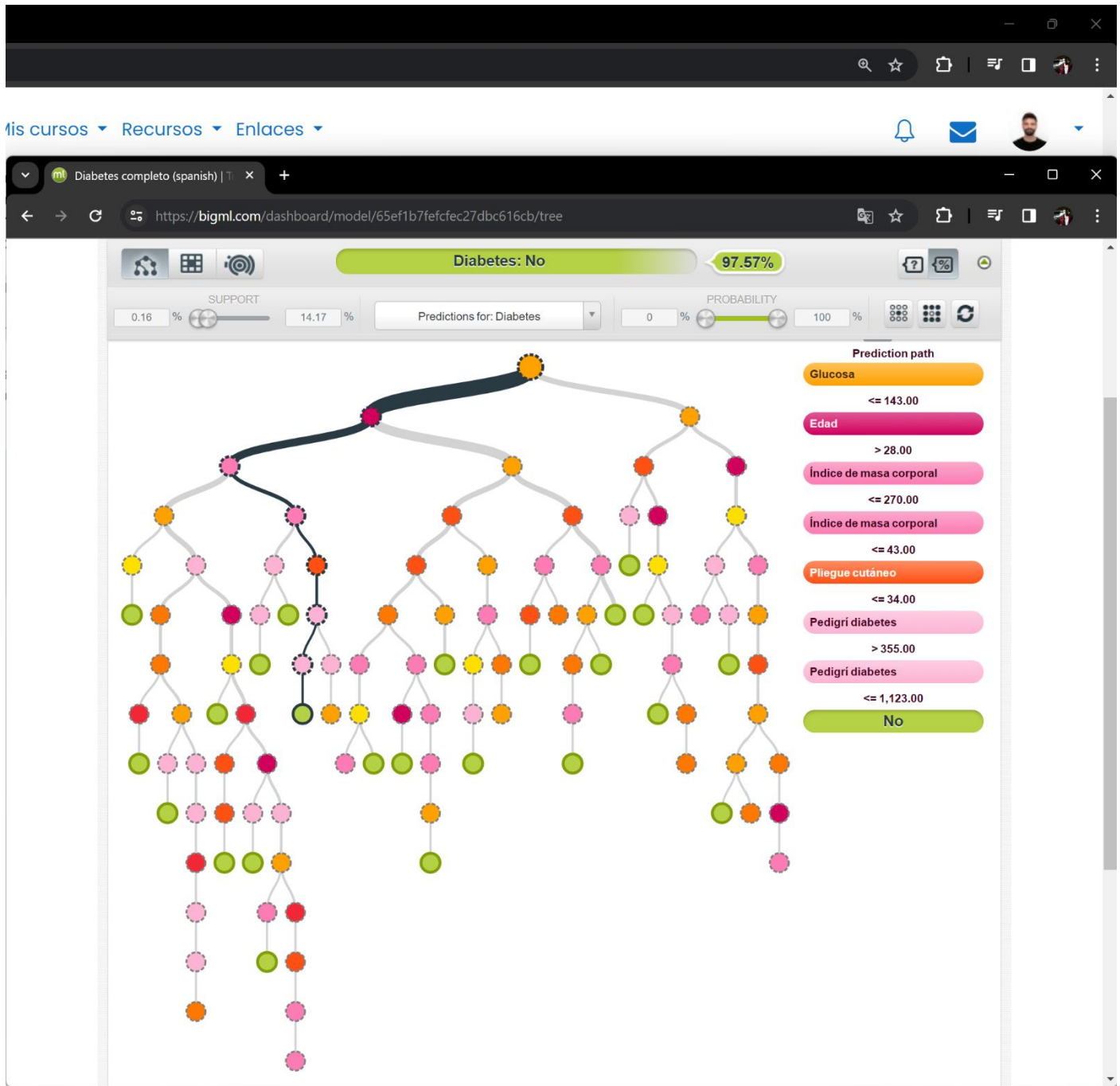
The screenshot shows the BigML web interface. The browser address bar displays <https://bigml.com/dashboard/dataset/65eeb794d9ec8b7438211849/table>. The user is logged in as CARLONSK. The dashboard title is "CARLONSK - My Dashboard". The main section is titled "Diabetes completo (spanish)". Under "SPLIT DATASET CONFIGURATION", the "Training" slider is set to 80% and the "Test" slider is set to 20%. The "Seed" field is empty. The "Linear split" button is visible. Below the configuration, the "Training dataset name" is "Diabetes completo (spanish) | Training" and the "Test dataset name" is "Diabetes completo (spanish) | Test (20%)". A "Create Training | Test" button is present. A table below shows the dataset structure:

Name	Type	Count	Missing	Errors	Histogram
Embarazos	1 2 3	768	0	0	
Glucosa	1 2 3	768	0	0	

Preparamos el dataset eligiendo el 80% de los datos para el entrenamiento, y reservamos el 20% para después evaluar el modelo.

Apartado 4: Entrenamiento:

- Incorpora una captura de pantalla que muestre el árbol de decisión del modelo ya entrenado.
- Explica los principales resultados: Casos en los que haya resultado positivo o negativo con suficiente confiabilidad.
- Incorpora capturas de pantalla de los diagramas de confiabilidad (confidence) y predicción (prediction).



Basándonos en los datos revisados, podemos ver que varios factores influyen en el riesgo de diagnóstico positivo de diabetes:

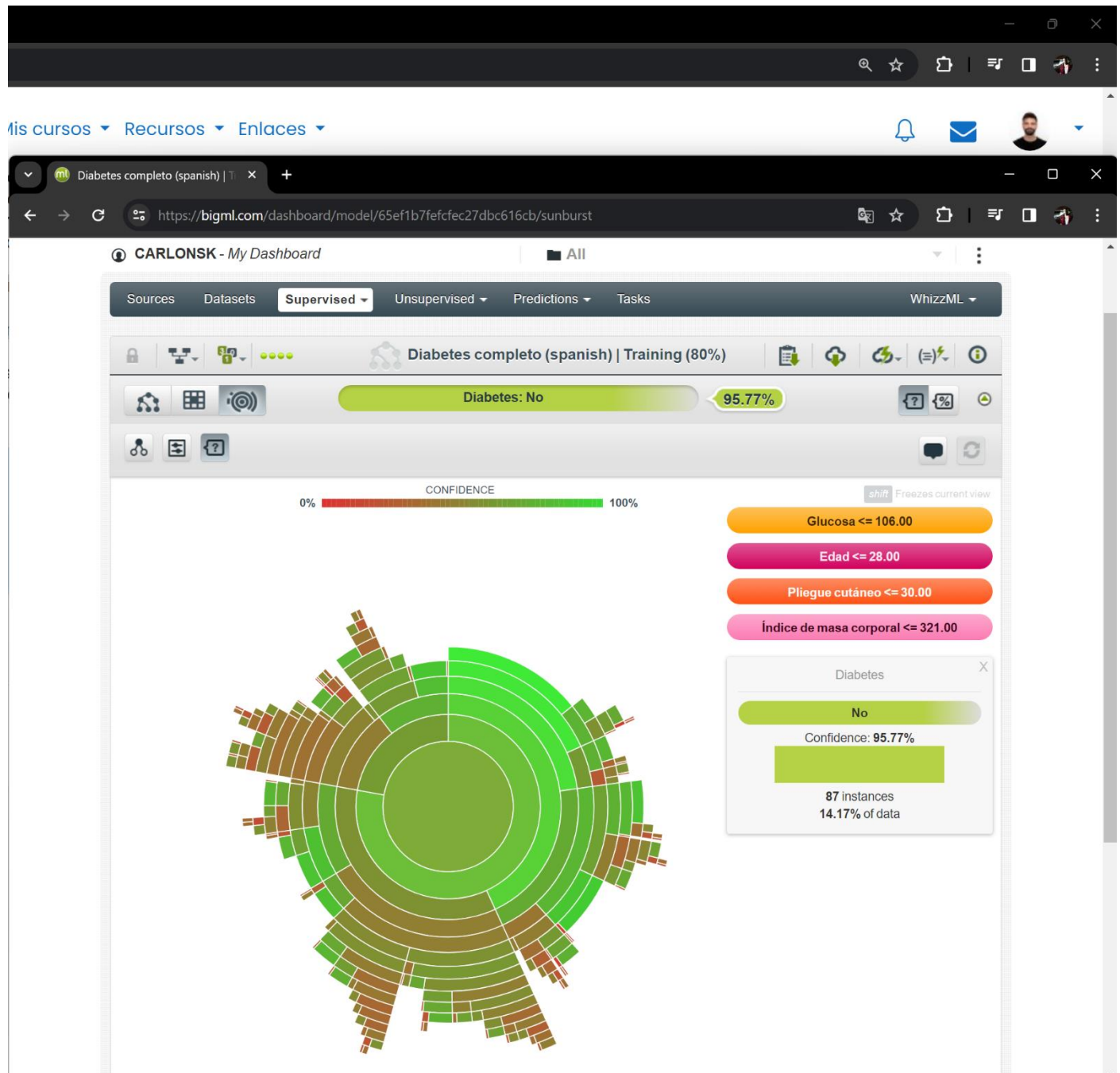
Pedigrí diabetes: Cuanto mayor sea el porcentaje de antecedentes familiares o pedigrí relacionados con la diabetes, mayor será el riesgo.

Edad: El riesgo aumenta entre los 29 y los 59 años, especialmente si los índices de glucosa superan los 126.

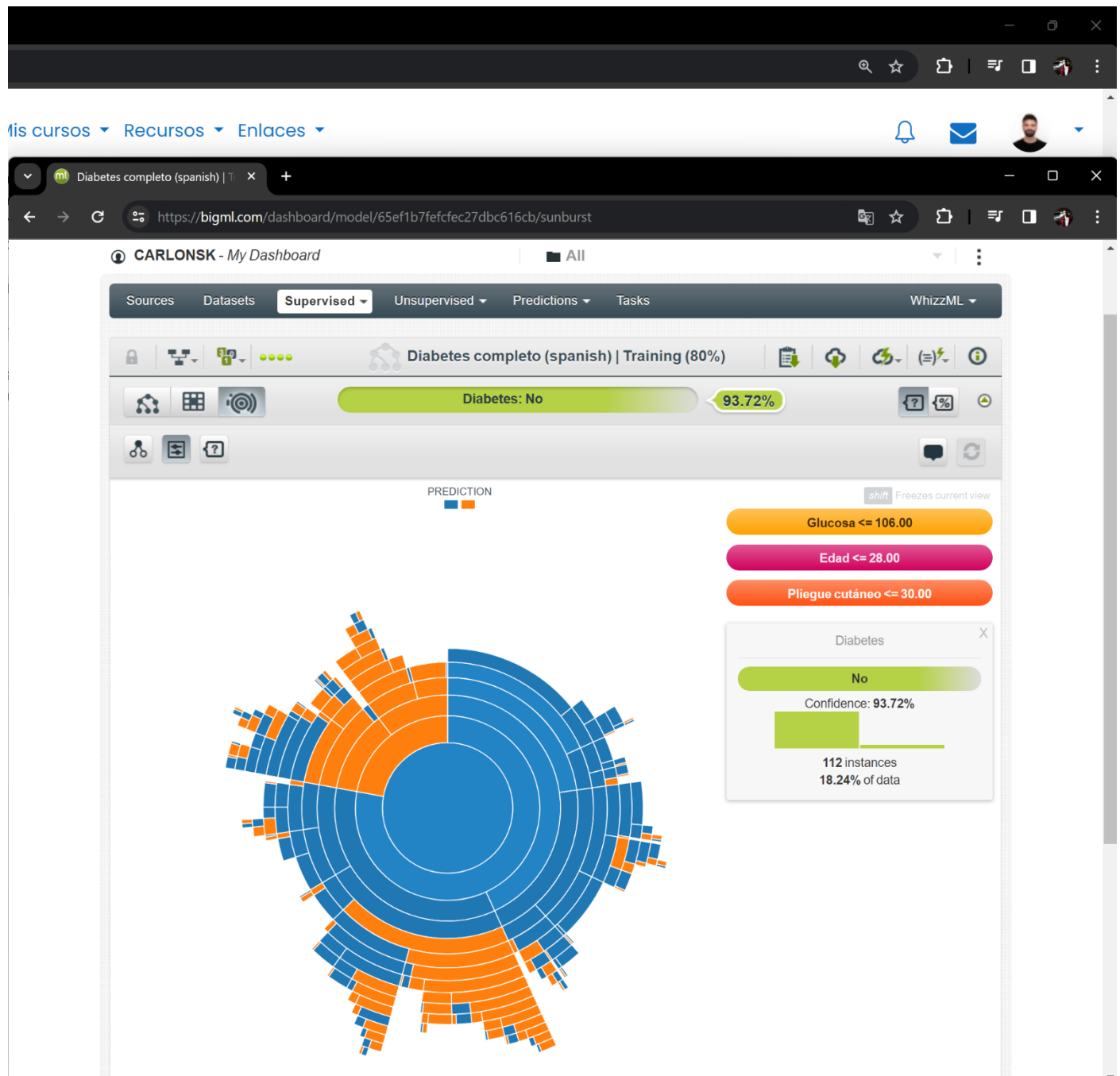
Pliegues cutáneos: Un rango de pliegues cutáneos entre 16 y 34 mm también está asociado con un mayor riesgo.

IMC de obesidad: Las personas con un índice de masa corporal (IMC) superior a 32 enfrentan un riesgo más alto.

Confiabilidad

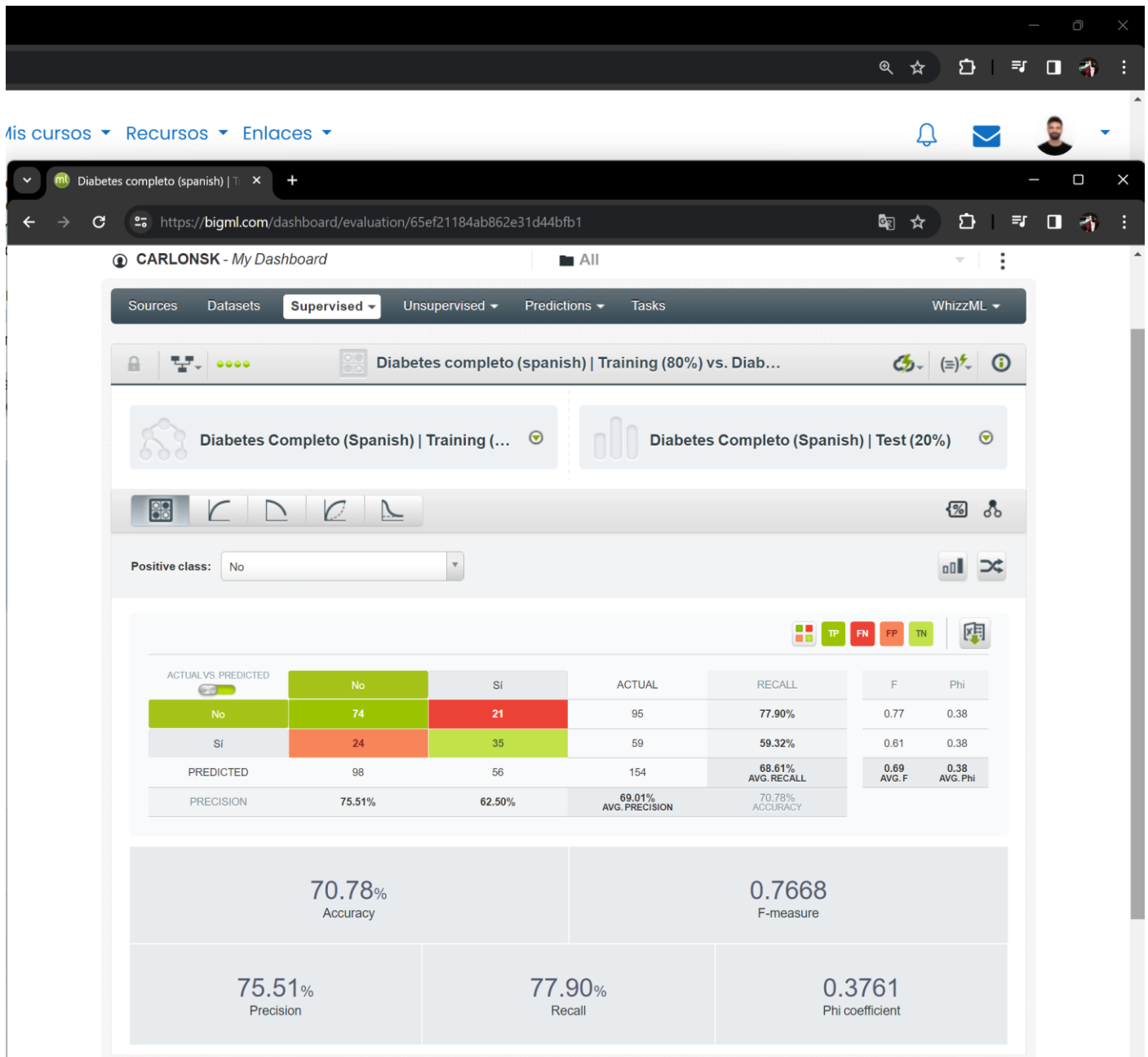


Precision



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



El **Accuracy** en este caso es la **proporción de pacientes correctamente clasificados por el modelo**, es decir, la cantidad de veces que el modelo ha acertado al predecir si un paciente tiene o no diabetes. En este caso, el **Accuracy es del 70.78%**, lo que significa que el modelo ha acertado en 2 de cada 3 casos.

La **Precisión** indica la **proporción de pacientes que el modelo ha predicho como diabéticos que realmente lo son**. En este caso, la **Precisión es del 75.51%**, lo que significa que 7 de cada 10 pacientes que el modelo ha predicho como diabéticos realmente lo son.

Para obtener resultados más confiables, se recomienda, utilizar un dataset más grande que permitirá una evaluación más precisa del rendimiento del modelo.