

Sistemas de aprendizaje automático



Nombre: Victoria Jiménez Martín

Módulo: Sistemas de aprendizaje automático

Curso: Especialización de Inteligencia Artificial y Big Data

Índice

Apartado 1: Localiza el dataset "Diabetes completo (spanish)":	3
Utiliza el buscador de datasets que tiene la propia plataforma para ello.	3
Incorpora una captura de pantalla del dataset mencionado incorporado a tu apartado Datasets.	5
Apartado 2: Observación del dataset:	6
Incorpora una captura de pantalla del dataset donde se vean al menos 10 categorías con sus tipologías, errores, histogramas, etc.	6
Explica cómo es el dataset: Número de instancias y número de categorías.	8
Explica el tipo de categorías (numéricas, texto, items, categóricas...).	8
Analiza los histogramas de cada categoría y comenta aquellos en los que consideres que hay algún tipo de anomalía.	9
Apartado 3: Preparación del dataset para entrenamiento y test:	11
Incorpora una captura de pantalla del proceso en el que defines los porcentajes de datos reservados para entrenamiento y para test.	11
Apartado 4: Entrenamiento:	12
Incorpora una captura de pantalla que muestre el árbol de decisión del modelo ya entrenado.	12
Explica los principales resultados: Casos en los que haya resultado positivo o negativo con suficiente confiabilidad.	15
Incorpora capturas de pantalla de los diagramas de confiabilidad (confidence) y predicción (prediction).	17
Apartado 5: Evaluación:	19
Incorpora una captura de pantalla en la que se muestre la evaluación del modelo entrenado realizada con el dataset reservado en el apartado 3.	19
Explica el resultado de dicha evaluación, indicando el nivel de confianza obtenido (Accuracy) y el nivel de precisión (Precision).	20

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

Utiliza el buscador de datasets que tiene la propia plataforma para ello.

Nos vamos a la página de bigml, al apartado de Sources y buscamos el dataset de Diabetes Diagnosis:

The screenshot shows the BigML web interface. At the top, there's a navigation bar with the BigML logo, links for PRODUCT, GETTING STARTED, PRICING, and SUPPORT, and a user profile section for VJIMENEZMARTI... with a Dashboard button. Below this is a search bar with filters: PUBLIC, DATASETS, a search input containing 'Dia', POPULAR, ALL CATEGORIES, and a FREE tag. The main content area displays three dataset cards. The first card, 'Diabetes completo (spanish)' by miniadax, has a description in Spanish and shows 38.7 KB, 16 fields, and 768 instances. The second card, 'Arrhythmia' by czuriaga, describes distinguishing cardiac arrhythmia and shows 403.9 KB, 280 fields, and 452 instances. The third card, 'People killed by guns in USA' by czuriaga, describes gun deaths in the USA and shows 205 instances. A right sidebar shows the user's profile with options for 'Área personal', 'Perfil', 'Preferencias', and 'Cerrar sesión'.

Le damos al apartado de Free y clonamos el Dataset:

The screenshot shows the BigML web interface. The top navigation bar includes the BigML logo, links for PRODUCT, GETTING STARTED, PRICING, and SUPPORT, and a user profile section for VJIMENEZMARTIN058 with a Dashboard button. The main navigation bar has tabs for MIRIADAX, SCRIPTS, MODELS, and DATASETS. The central area displays the 'Diabetes completo (spanish)' dataset, which is marked as 'FREE'. A modal dialog titled 'Clone this dataset' is open, prompting the user to clone the dataset 'Diabetes completo (spanish)' into the 'VJIMENEZMARTIN058 - My Dashboard' and within the 'BigML Intro Project'. The dialog has 'Cancel' and 'Clone' buttons. On the right side, a sidebar menu lists 'Área personal', 'Perfil', 'Preferencias', and 'Cerrar sesión'.

bigml

PRODUCT GETTING STARTED PRICING SUPPORT

VJIMENEZMARTIN058 Dashboard

MIRIADAX SCRIPTS MODELS DATASETS

Diabetes completo (spanish) FREE

Clone this dataset

Clone the dataset **Diabetes completo (spanish)** in

VJIMENEZMARTIN058 - My Dashboard

within the project

BigML Intro Project

Cancel Clone

Área personal

Perfil

Preferencias

Cerrar sesión

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

Comprobamos que se añadió correctamente:

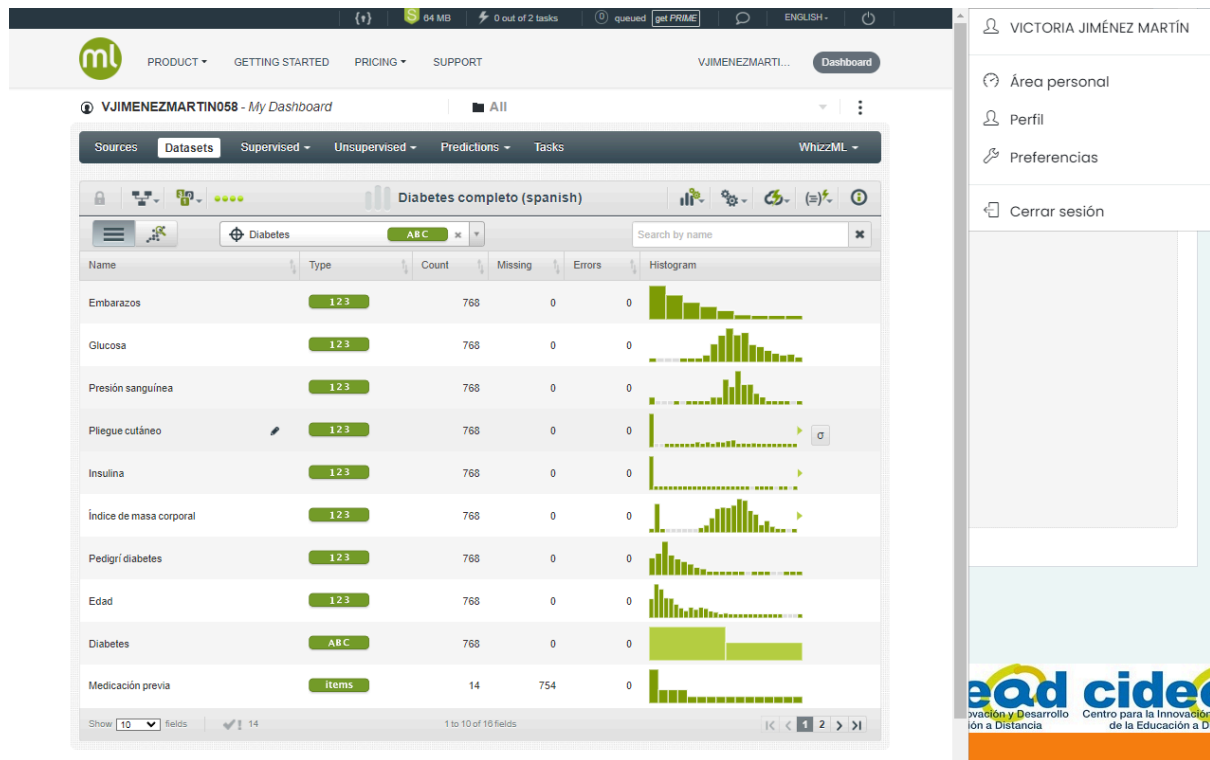
The screenshot displays the WhizzML dashboard for user VJIMENEZMARTIN058. The main content area shows the 'Diabetes completo (spanish)' dataset with a table of features and their histograms. The features listed are:

Name	Type	Count	Missing	Errors	Histogram
Embarazos	123	768	0	0	[Histogram]
Glucosa	123	768	0	0	[Histogram]
Presión sanguínea	123	768	0	0	[Histogram]
Pliegue cutáneo	123	768	0	0	[Histogram]
Insulina	123	768	0	0	[Histogram]
Índice de masa corporal	123	768	0	0	[Histogram]
Pedigrí diabetes	123	768	0	0	[Histogram]
Edad	123	768	0	0	[Histogram]
Diabetes	ABC	768	0	0	[Histogram]
Medicación previa	Items	14	754	0	[Histogram]

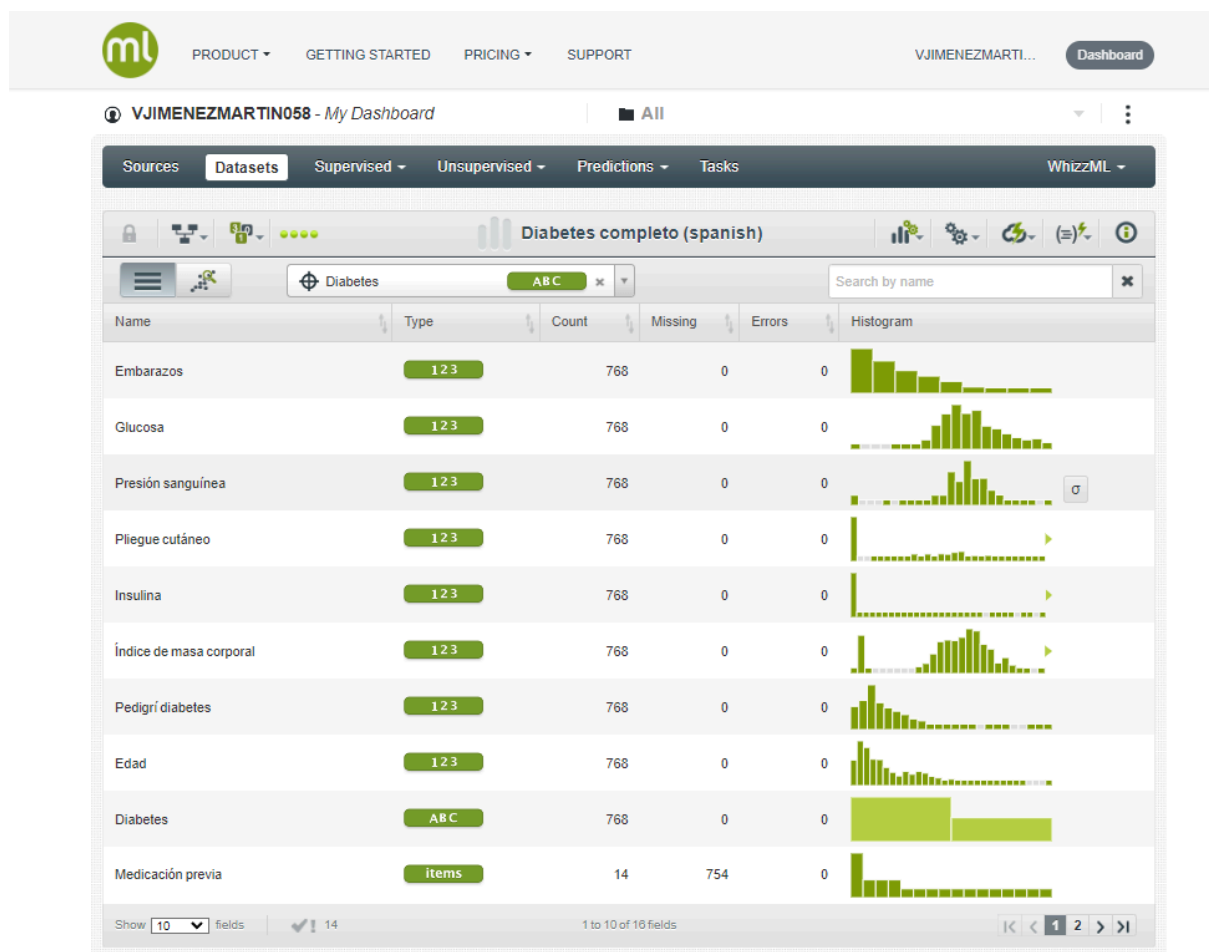
The sidebar on the right shows the user's profile information: VICTORIA JIMÉNEZ MARTÍN, with links to 'Área personal', 'Perfil', 'Preferencias', and 'Cerrar sesión'. The bottom of the sidebar features the 'ead cidec' logo and the text 'Innovación y Desarrollo de la Educación a Distancia' and 'Centro para la Innovación de la Educación a Distancia'.

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.



Realizamos otra captura para que se puedan ver correctamente:



Vemos que las categorías son: Pregnancies, Glucose, Blood pressure, Skinfold, Insulin, BMI, Diabetes pedigree, Age y Diabetes.

(Se adjunta otra captura con el resto de categorías, de la siguiente página)

VJIMENEZMARTIN058 - My Dashboard | All

Sources Datasets Supervised Unsupervised Predictions Tasks WhizzML

Diabetes completo (spanish)

Diabetes ABC x Search by name x

Name	Type	Count	Missing	Errors	Histogram
Observaciones	text	13	755	0	
Fecha de diagnóstico	YYYY-MM-DD	768	0	0	START: Not available END: Not available
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	MTWTFSS	766	2	0	

Show 10 fields 14 11 to 16 of 16 fields

El resto de categorías son: Observaciones, Fecha de diagnóstico, Fecha de diagnóstico.year, Fecha de diagnóstico.month, Fecha de diagnóstico.day-of-month, Fecha de diagnóstico.day-of-week.

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

A través de la imagen proporcionada en el caso anterior, podemos ver que el Dataset consta de 768 instancias y 16 categorías.

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

La columna que definirá el tipo de las categorías es Type y vemos que la de Diabetes es de tipo Categórica, la categoría de Observaciones es de tipo texto, y el resto de categorías de Fecha de diagnóstico, Fecha de diagnóstico.year, Fecha de diagnóstico.month, Fecha de diagnóstico.day-of-month, Fecha de diagnóstico.day-of-week, tienen diferentes tipos de formato fecha y el resto de categorías son numéricas.

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

Si nos fijamos en los diferentes histogramas, podemos extraer las siguientes deducciones:

- En la categoría de **Embarazos** vemos que la mayoría de las pacientes han estado embarazadas pocas veces y a medida que aumenta el número de embarazos, se produce una disminución de los pacientes.
- En la categoría de **Glucosa**, muestra los niveles de glucosa en plasma 2 horas después de ingerir la glucosa y podemos ver que el mayor número de pacientes se da en los rangos más altos, que son de 100 a 130 de glucosa en sangre 2 horas después de ingerirla.
- En la categoría de **Presión sanguínea**, vemos que la presión arterial muestra un repunte para los datos de 70-75, y también que, tiene una concentración de valores en el rango medio lo que podríamos decir que son los valores normales.
- En la categoría de **Pliegue cutáneo**, podemos ver que la mayoría de los valores se encuentran en el extremo izquierdo, lo que indica que los pacientes tienen mediciones del pliegue cutáneo relativamente bajas.
- En la categoría de **Insulina**, podemos ver que muestra las concentraciones de insulina pasadas 2 horas después de la prueba de la glucosa y comprobamos como en el caso anterior que los valores se encuentran en el extremo izquierdo. Estos valores se podrían corresponder a pacientes que suelen tener niveles más altos de insulina en sangre.
- En la categoría de **índice de masa corporal**, la mayoría de los valores se concentran en el centro por lo que la mayoría de los pacientes se centran ahí y el rango de mayor índice es de 320-340.
- En la categoría de **Pedigri diabetes**, vemos que los valores mayores se centran en el extremo izquierdo y después sufren una disminución en la frecuencia a medida que aumenta el valor de la función, aunque podemos observar un repunte entre los valores de 200-300.
- En la categoría de **Edad**, podemos ver que hay una alta frecuencia de pacientes más jóvenes y una disminución progresiva en la cantidad de personas de más edad. La edad donde más valores hay, se centra entre los 20-22.
- En la categoría de **Diabetes**, muestra dos barras que podrían ser los casos positivos y negativos de diabetes. Los casos para los que tienen diabetes, muestra el mayor repunte, mientras que los que sí tienen diabetes presentan la barra menor.
- En la categoría de **Medicación previa**, vemos que solo ha tomado 14 valores diferentes y que representan los diferentes medicamentos. Este histograma

representa la frecuencia de cada categoría de los medicamentos para cada uno de los 754 casos que presenta el Dataset.

- En la categoría de **Observaciones**, podemos ver que ha tomado 13 valores diferentes apartados para comentarios como: patologías, problemas, resultados...
- En la categoría de **Fecha de diagnóstico_year**, muestra el año del diagnóstico que presenta la totalidad de los datos que corresponden al año 2016.
- En la categoría de **Fecha de diagnóstico_month**, muestra el mes del diagnóstico, que en este caso, los datos están recogidos desde Enero hasta diciembre.
- En la categoría de **Fecha de diagnóstico_day-of-month**, muestra el día del mes en el que se realizó el diagnóstico, que representa del 1 al 31.
- En la categoría de **Fecha de diagnóstico_day-of-week**., muestra el día de la semana en el que se realizó el diagnóstico. Representa desde el Lunes al Viernes.

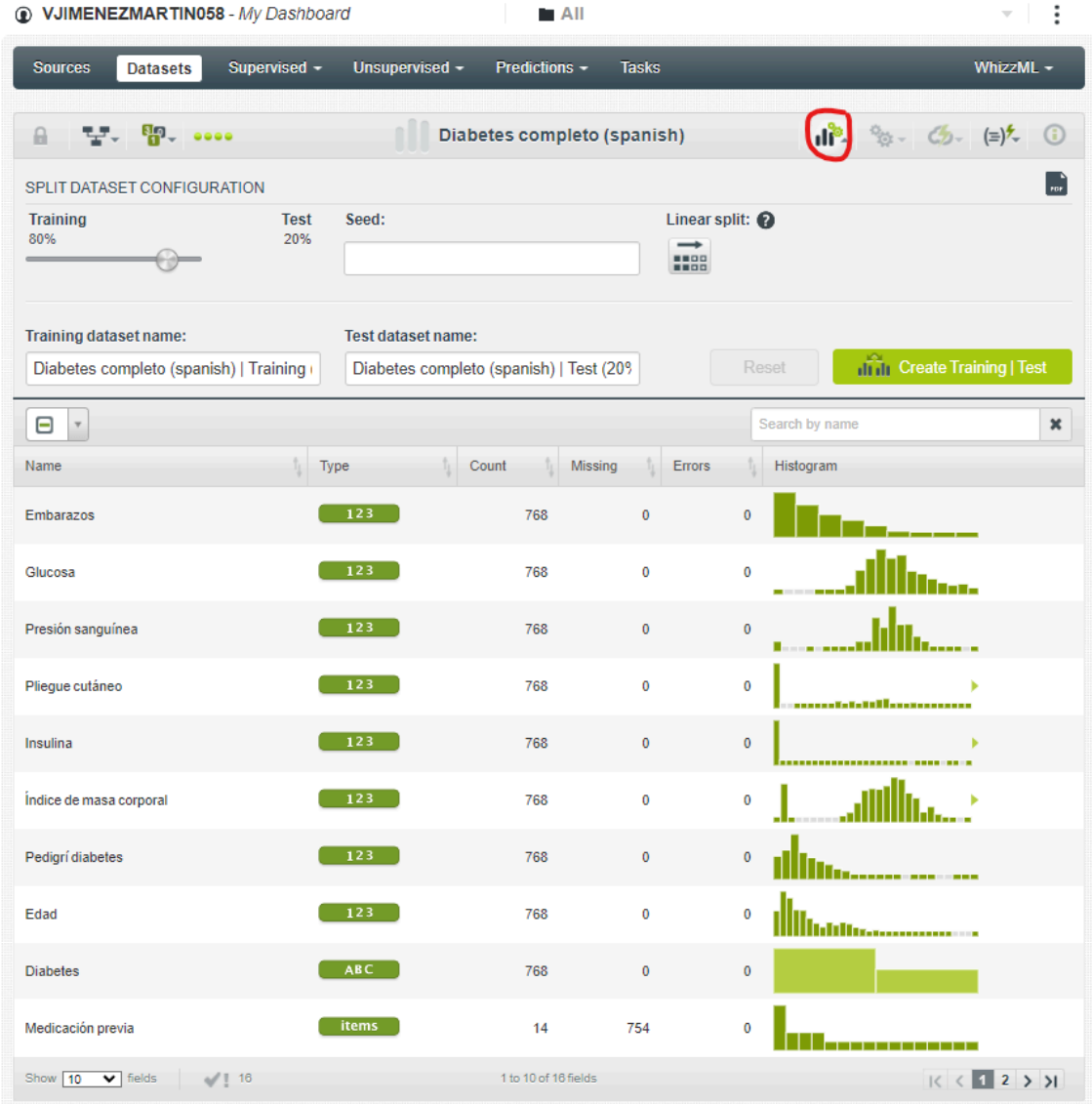
En base al análisis anterior, las categorías que podrían ser más propensas a error pueden ser, la presión sanguínea debido a que existe la presencia de un número considerable de medidas muy bajas de la presión arterial (cerca del extremo izquierdo) es atípica. Esto podría indicar que podrían haber datos erróneos o con errores de entrada debido a que la presión arterial en adultos, raramente es tan baja.

Y la otra que posiblemente pueda ser propensa a errores puede ser la categoría de Pliegue cutáneo debido a que el grosor del pliegue de la piel del tríceps muestra un gran número de valores en el extremo inferior, con lo que podemos intuir que pueden haber medidas incorrectas o imprecisiones en las mediciones.




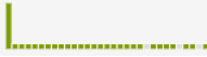






Apartado 3: Preparación del dataset para entrenamiento y test:

Incorpora una captura de pantalla del proceso en el que defines los porcentajes de datos reservados para entrenamiento y para test.

Definimos los porcentajes de entrenamiento en el apartado de Training | Test Split:



The screenshot shows the WhizzML interface for the 'Diabetes completo (spanish)' dataset. The 'SPLIT DATASET CONFIGURATION' section is active, showing a Training percentage of 80% and a Test percentage of 20%. The 'Create Training | Test' button is highlighted in green. Below the configuration, a table lists the dataset fields with their types, counts, missing values, and errors.

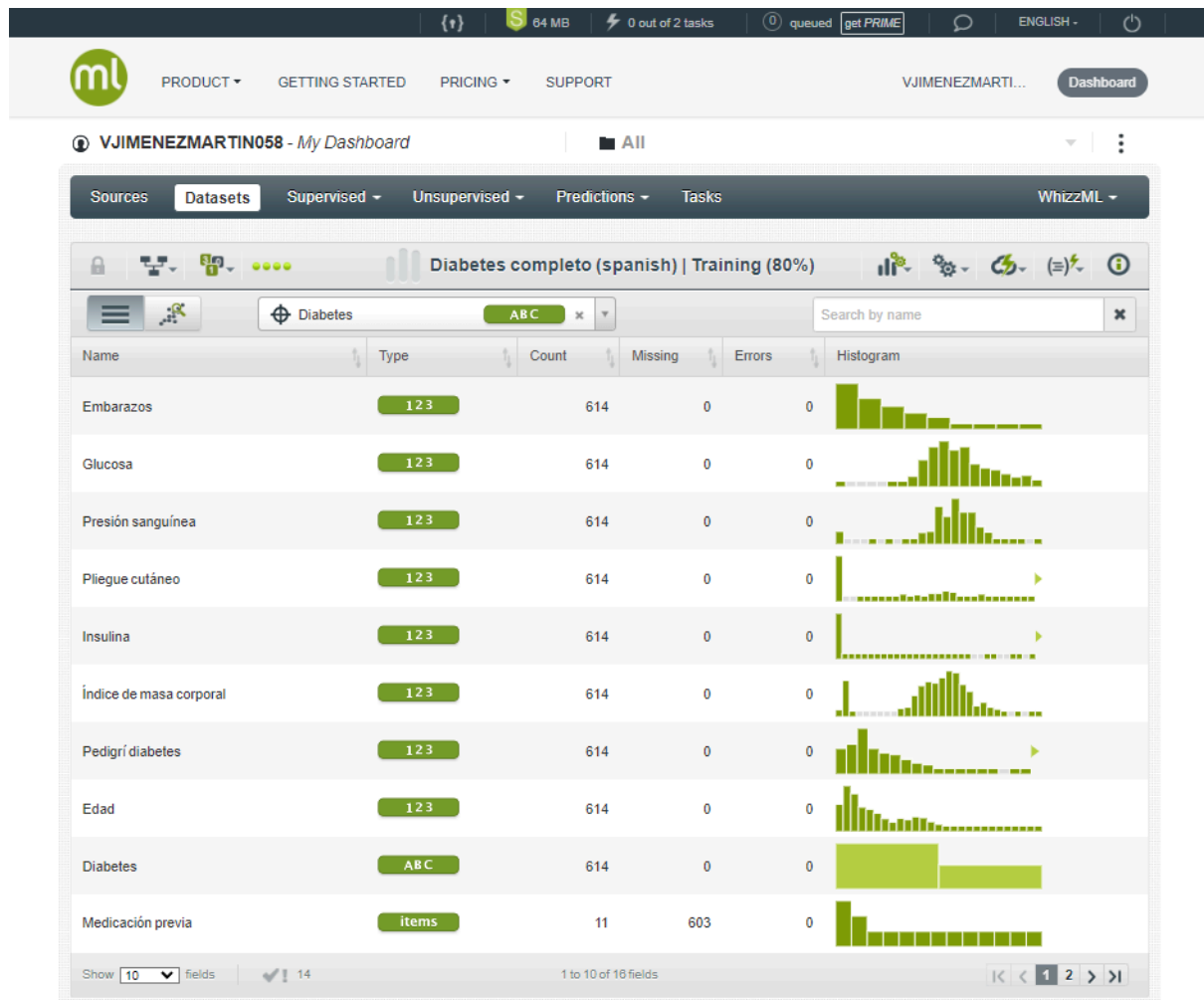
Name	Type	Count	Missing	Errors	Histogram
Embarazos	1 2 3	768	0	0	
Glucosa	1 2 3	768	0	0	
Presión sanguínea	1 2 3	768	0	0	
Pliegue cutáneo	1 2 3	768	0	0	
Insulina	1 2 3	768	0	0	
Índice de masa corporal	1 2 3	768	0	0	
Pedigrí diabetes	1 2 3	768	0	0	
Edad	1 2 3	768	0	0	
Diabetes	A B C	768	0	0	
Medicación previa	items	14	754	0	

Le damos a Creating Training | Test.


Apartado 4: Entrenamiento:

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

Una vez realizado el Test, comprobamos que han disminuido los datos totales recabados:



Para crear el árbol de decisión nos iremos al apartado de Model:



PRODUCT ▾GETTING STARTEDPRICING ▾SUPPORT

VJIMENEZMARTI...Dashboard

VJIMENEZMARTIN058 - My DashboardAll

SourcesDatasetsSupervised ▾Unsupervised ▾Predictions ▾TasksWhizzML ▾

Diabetes completo (spanish) | Training (80%)

MODEL CONFIGURATION

Objective field:
Diabetes ABC







Automatic optimization

Advanced configuration

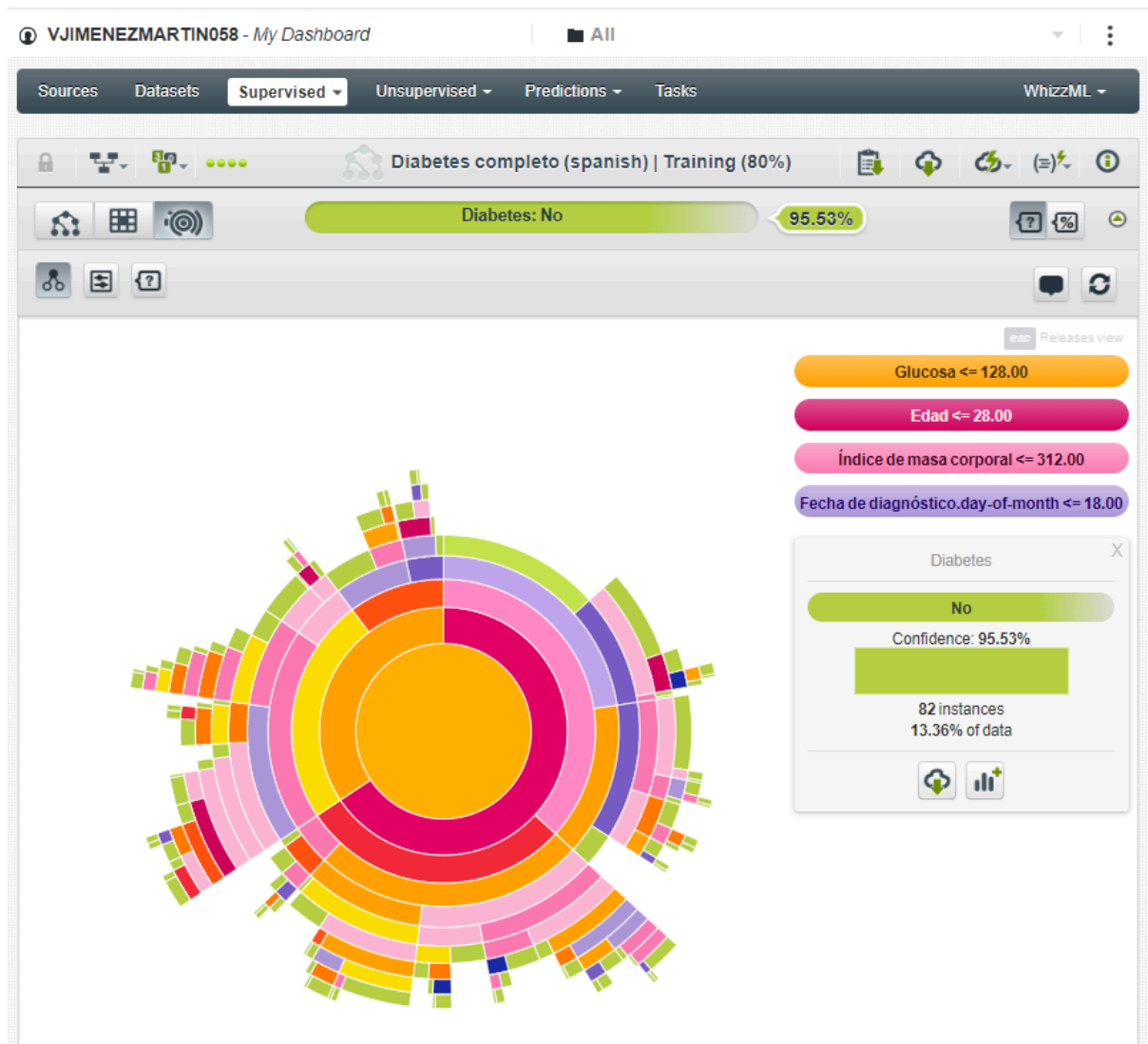
Model name:
Diabetes completo (spanish) | Training (80%)

ResetCreate model

Search by name

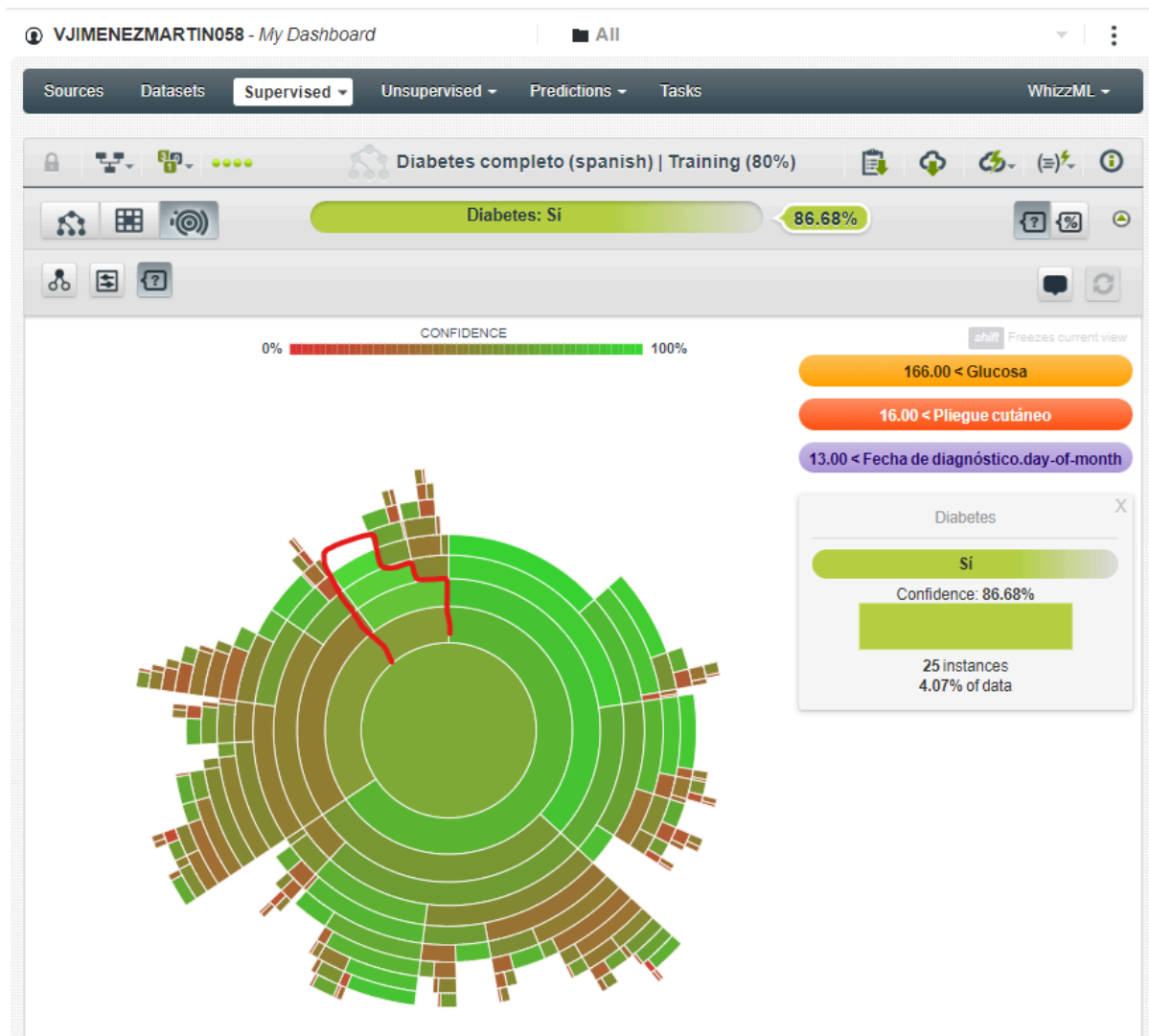
Name	Type	Count	Missing	Errors	Histogram
Embarazos	123	614	0	0	
Glucosa	123	614	0	0	
Presión sanguínea	123	614	0	0	
Pliegue cutáneo	123	614	0	0	
Insulina	123	614	0	0	
Índice de masa corporal	123	614	0	0	

Y le damos a Create Model y nos crea el siguiente árbol de precisión:



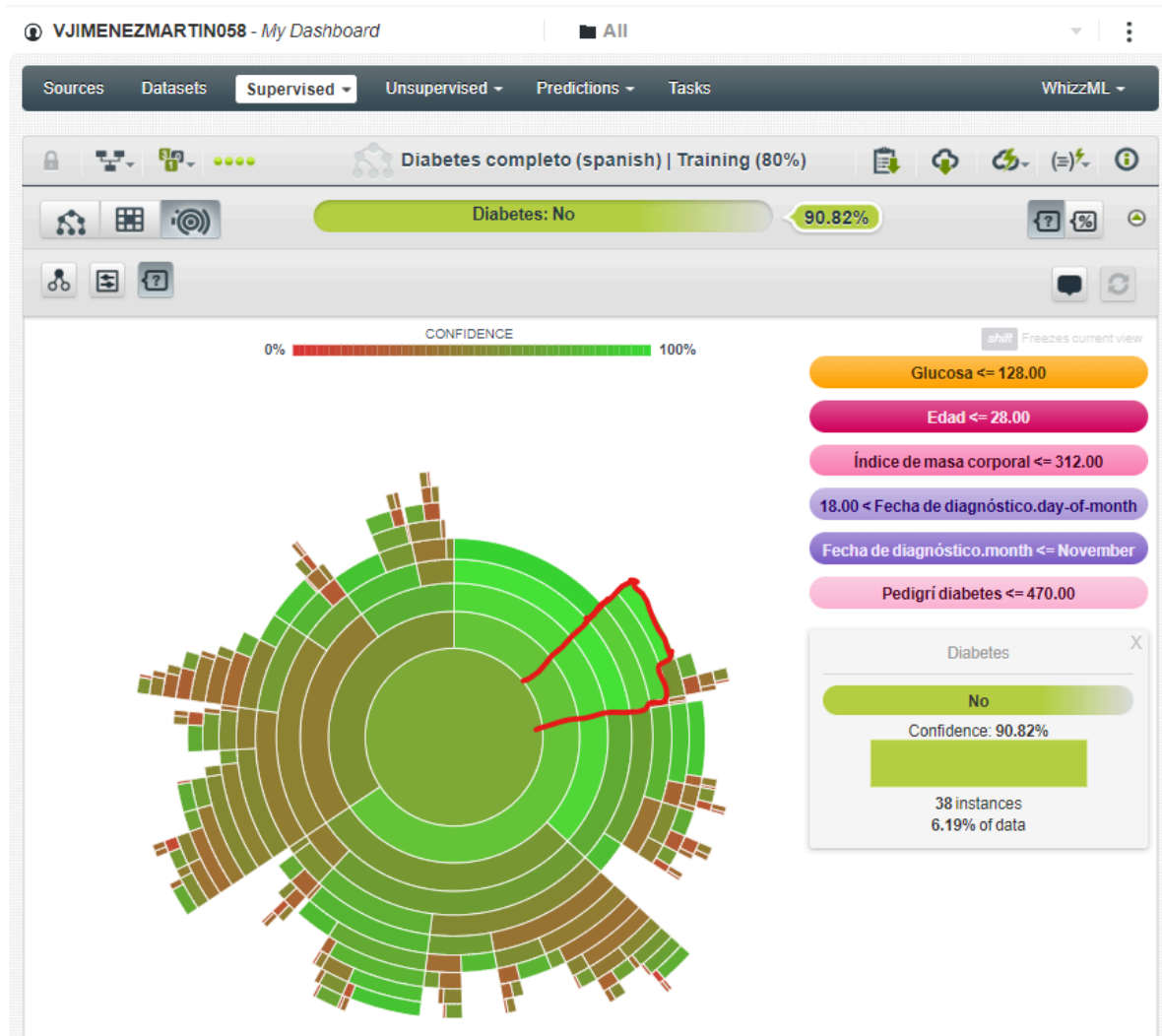
Explica los principales resultados: Casos en los que haya resultado positivo o negativo con suficiente confiabilidad.

Para los casos de positivos en diabetes, según el árbol de precisión por confiabilidad, para poder saber qué datos son más fiables, nos fijamos en el color verde claro que sería lo más confiable. Nos quedamos con los datos de la siguiente parte del diagrama:



El resultado sería: la glucosa por encima de 166 y el pliegue cutáneo y los datos se tomaron posterior al día 13 de algún mes. Estos datos presentan una confiabilidad del 86.68%.

Para el caso de negativos en diabetes, nos iremos a otra parte del diagrama siguiendo los mismos pasos con respecto al color, y los datos serían los siguientes:



Para estos resultado vemos que: la glucosa será menor o igual a 128, la edad es menor o igual a 28, el índice de masa corporal será menor o igual a 312, la fecha del diagnóstico es algún día superior al 18 de Noviembre y la pedigrí diabetes será igual o menor que 470. Estos datos tienen una confiabilidad del 90,82%

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

Diagrama de confiabilidad:

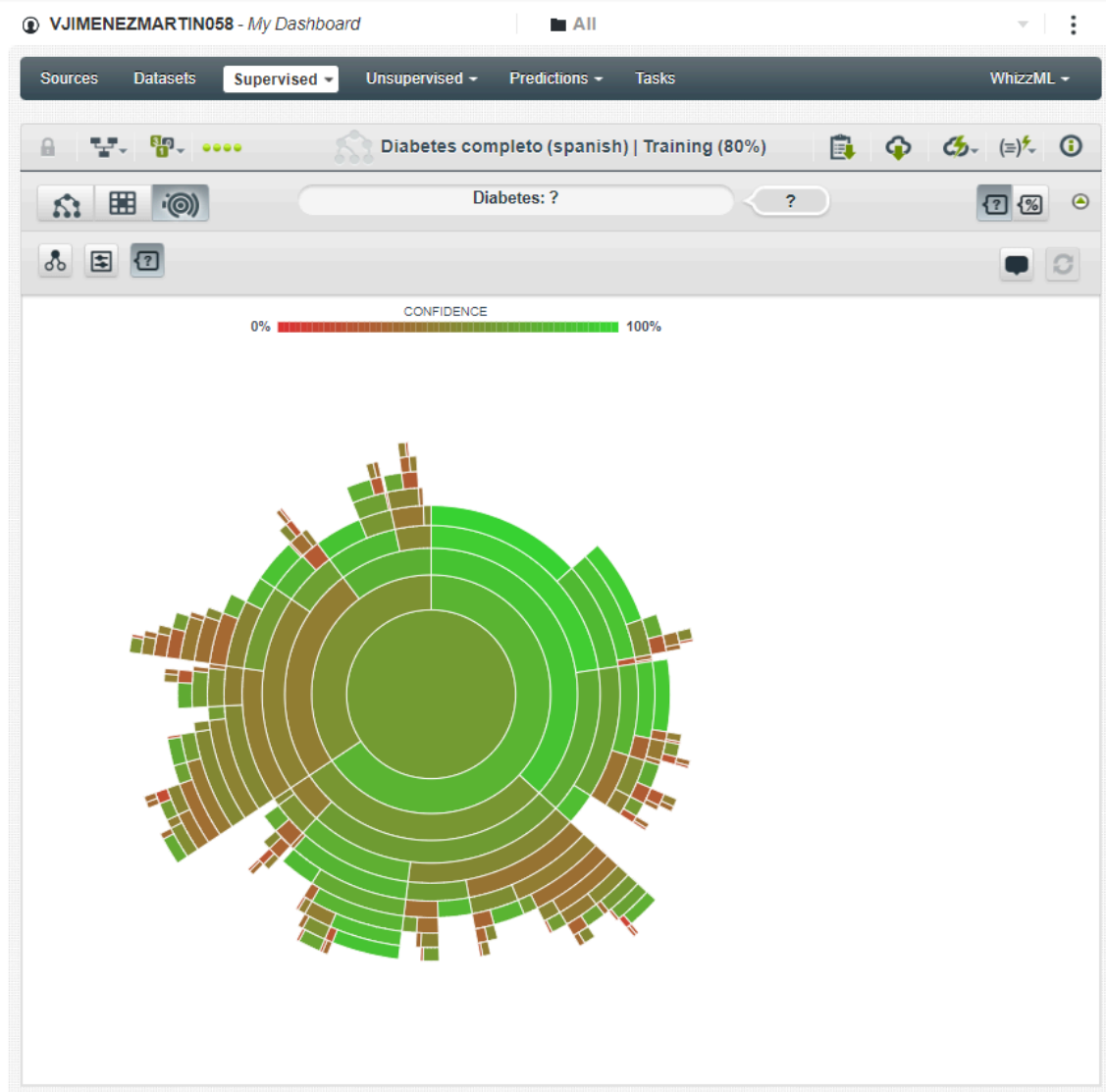
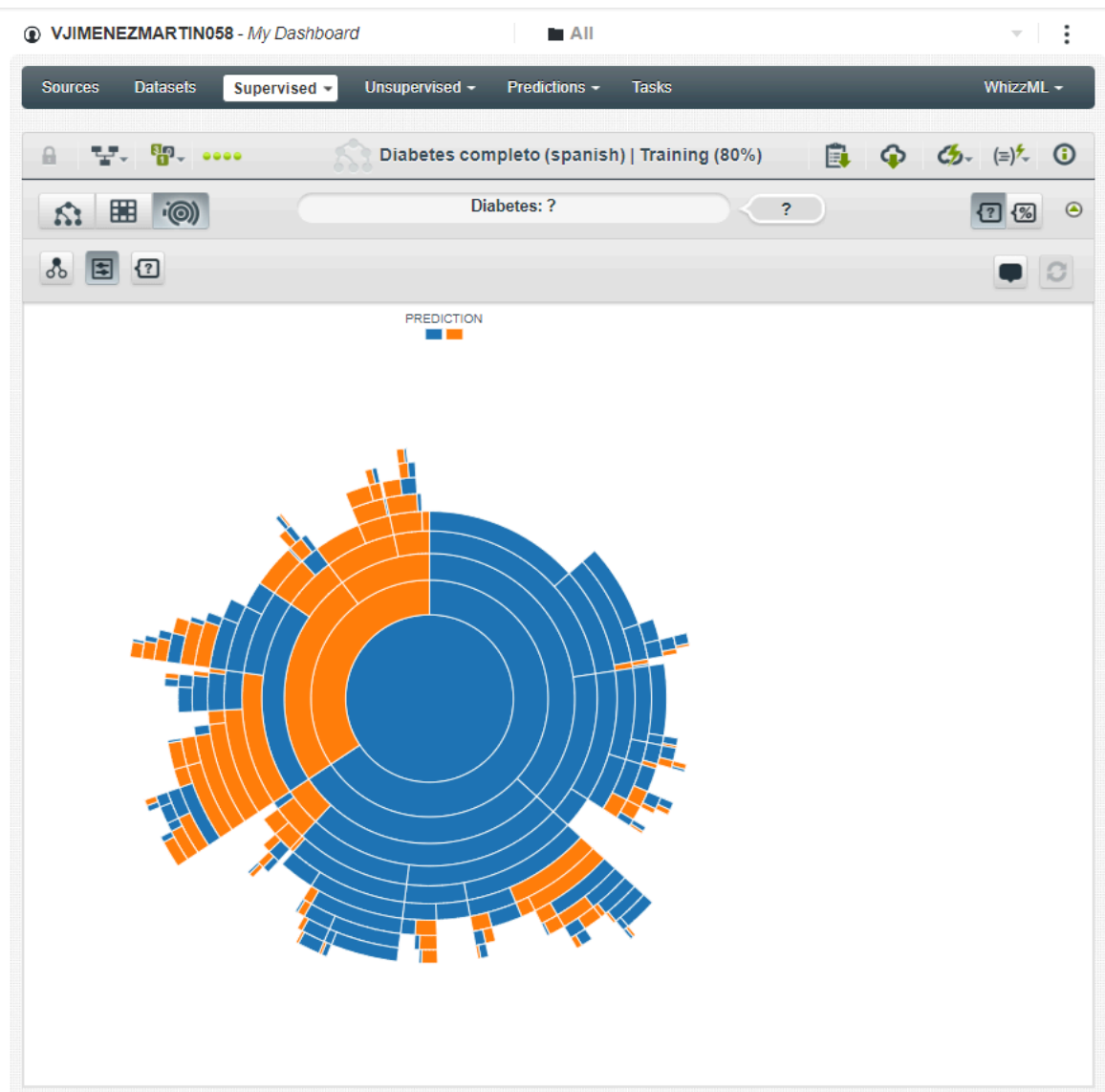
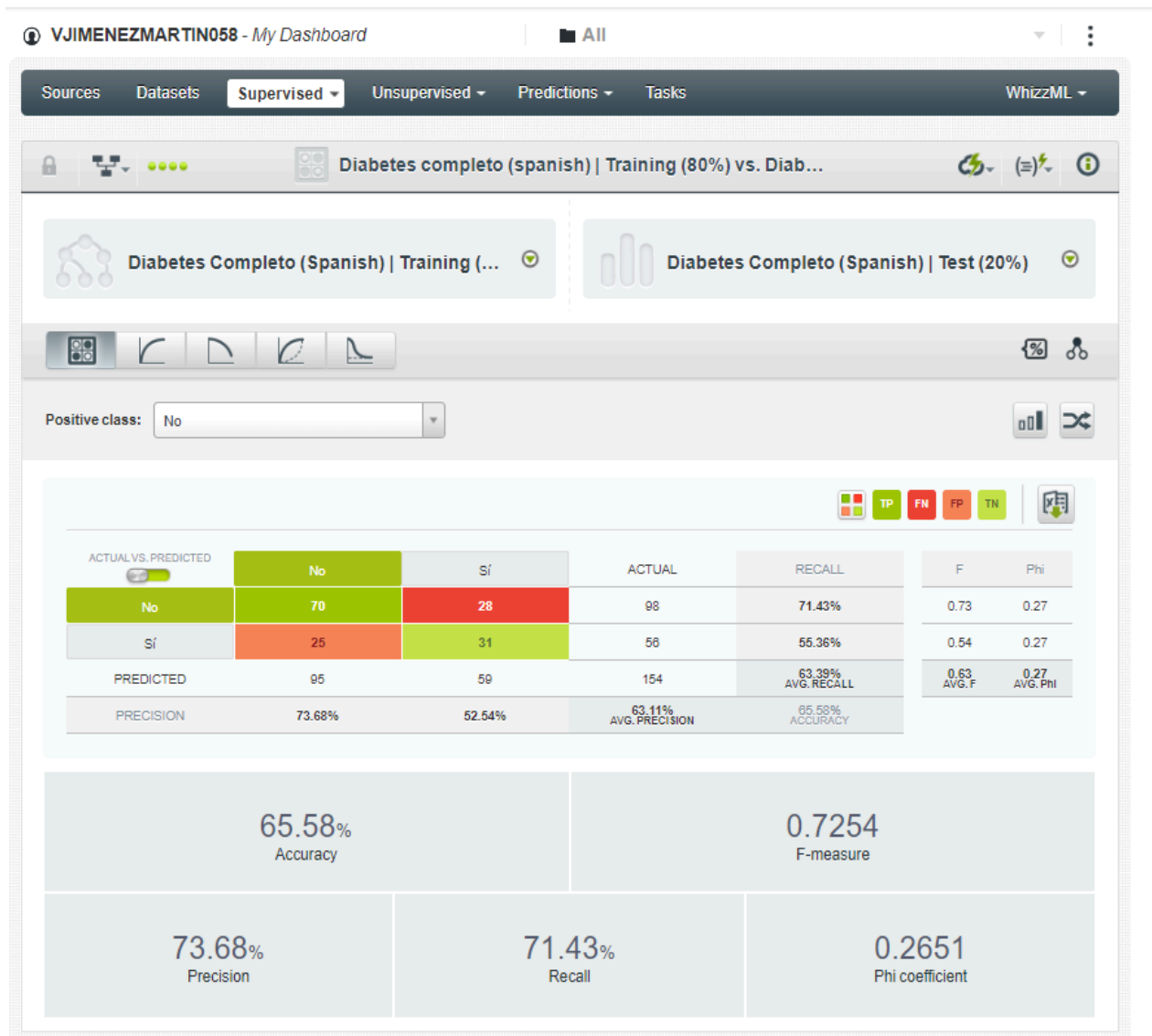


Diagrama de predicción:



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*).

La evaluación que se puede realizar de forma general es la siguiente:

- Verdaderos positivos: 31 resultados (predijo correctamente cuando había diabetes).
- Falsos positivos: 28 resultados (predijo incorrectamente cuando decía que había diabetes pero no había diabetes).
- Verdaderos negativos: 70 resultados (predijo correctamente que no había diabetes).
- Falsos negativos: 25 resultados (predijo incorrectamente cuando decía que no había diabetes pero sí había diabetes).

Más específicamente, para saber el nivel de confianza y el nivel de precisión, nos muestra los siguientes resultados:

- Precisión (proporción de predicciones positivas que son correctas): para la clase de "No" (no diabetes) es del 73,68% y para la clase de "Si" (si diabetes) es del 52,54%. Por tanto, el promedio de este modelo es 63,11%.
- Accuracy o Confiabilidad (indica que tan bien el modelo ha predicho ambas clases, tanto positivas como negativas): la exactitud global del modelo es de 65.58%