

Sistemas de aprendizaje automático



Nombre: Victoria Jiménez Martín

Módulo: Sistemas de aprendizaje automático

Curso: Especialización de Inteligencia Artificial y Big Data

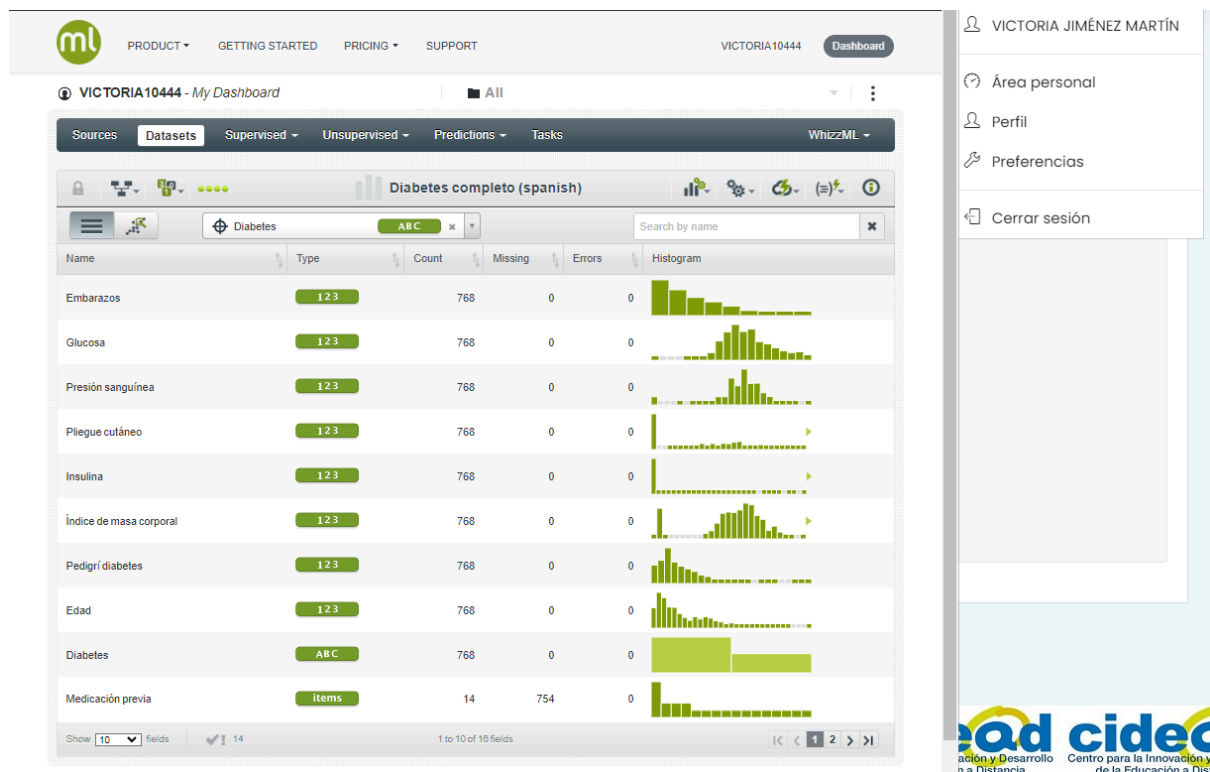
Índice

Apartado 1: Realiza una predicción por lote en BigML	3
Elige un dataset para clasificación binaria de los que vienen por defecto en BigML o carga uno que te parezca interesante.	3
Separa los datos en 80% para el entrenamiento y 20% para test.	3
Entrena un modelo de árbol de decisión.	4
Realiza la predicción por lotes, seleccionando el conjunto de datos de test. Descarga el archivo csv resultante.	6
Apartado 2: Calcula la matriz de confusión.	8
Abre el archivo csv en una hoja de cálculo y aplica las fórmulas necesarias para obtener: errores totales, falsos negativos y falsos positivos.	8
Construye la matriz de confusión, rellenando los valores correspondientes.	11
Analiza los resultados. ¿Es fiable el modelo?	11
Apartado 3: Aplica la técnica de aprendizaje no supervisado de Detección de Anomalías.	12
Aplica el modelo de detección de anomalías en BigML dentro de las funciones rápidas de algoritmos no supervisados.	12
Analiza las top 5 anomalías de tu problema y decide si merece la pena analizarlas a parte.	13
Si crees que son importantes, crea un dataset con ellas para analizarlas	14

Apartado 1: Realiza una predicción por lote en BigML

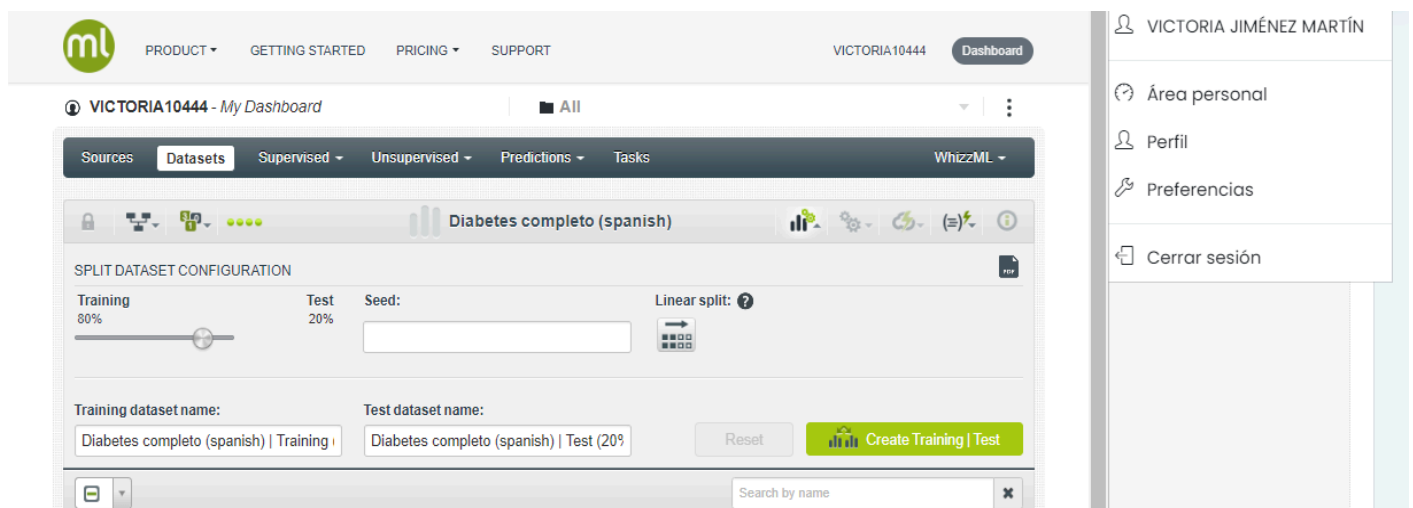
Elige un dataset para clasificación binaria de los que vienen por defecto en BigML o carga uno que te parezca interesante.

Escogeremos el mismo Dataset que en la unidad anterior que es el de la diabetes:



Separa los datos en 80% para el entrenamiento y 20% para test.

Entrenamos el Dataset:



Entrena un modelo de árbol de decisión.

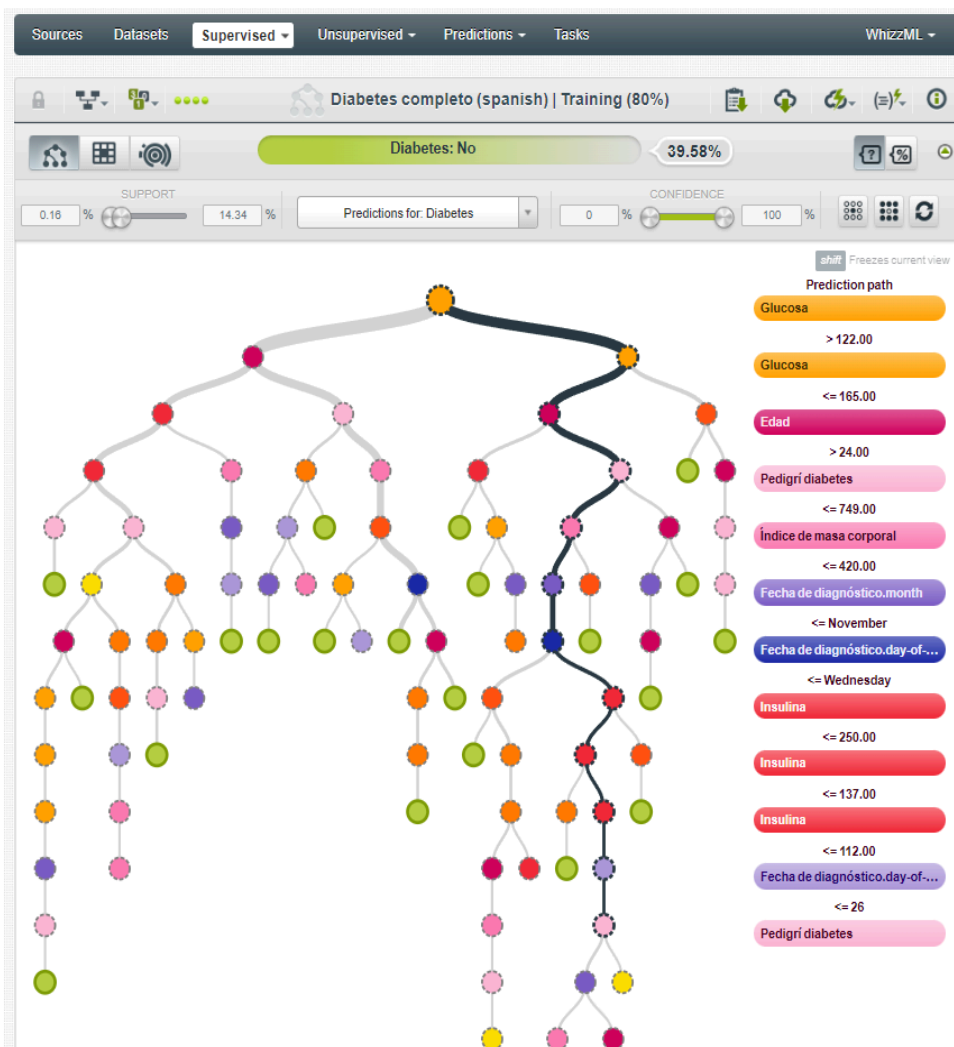
Nos iremos al apartado de Model:

The screenshot shows the WhizzML dashboard for user VICTORIA10444. The main panel displays the 'Diabetes completo (spanish)' dataset in 'Training (80%)' mode. A modal menu is open, showing options for 'CONFIGURE SUPERVISED' (MODEL, ENSEMBLE, LINEAR REGRESSION, LOGISTIC REGRESSION, DEEPNET, TIME SERIES, OPTIML) and 'CONFIGURE UNSUPERVISED' (CLUSTER, ANOMALY, ASSOCIATION, TOPIC MODEL, PCA). The 'MODEL' option is highlighted. The background shows a table of features: Embarazos, Glucosa, Presión sanguínea, Pliegue cutáneo, Insulina, Índice de masa corporal, Pedigrí diabetes, and Edad, each with a 'Type' column and a '123' button. To the right, a sidebar shows the user's profile and navigation options: Área personal, Perfil, Preferencias, and Cerrar sesión.

Y crearemos el model:

The screenshot shows the 'MODEL CONFIGURATION' screen for the 'Diabetes completo (spanish)' dataset. The 'Objective field' is set to 'Diabetes' and 'Automatic optimization' is enabled. The 'Model name' is 'Diabetes completo (spanish) | Training (80%)'. The 'Create model' button is visible. The background shows the same dataset table as the previous screenshot. To the right, the sidebar shows the user's profile and navigation options: Área personal, Perfil, Preferencias, and Cerrar sesión.

Y vemos que crea el árbol de decisión:



VICTORIA JIMÉNEZ MARTÍN

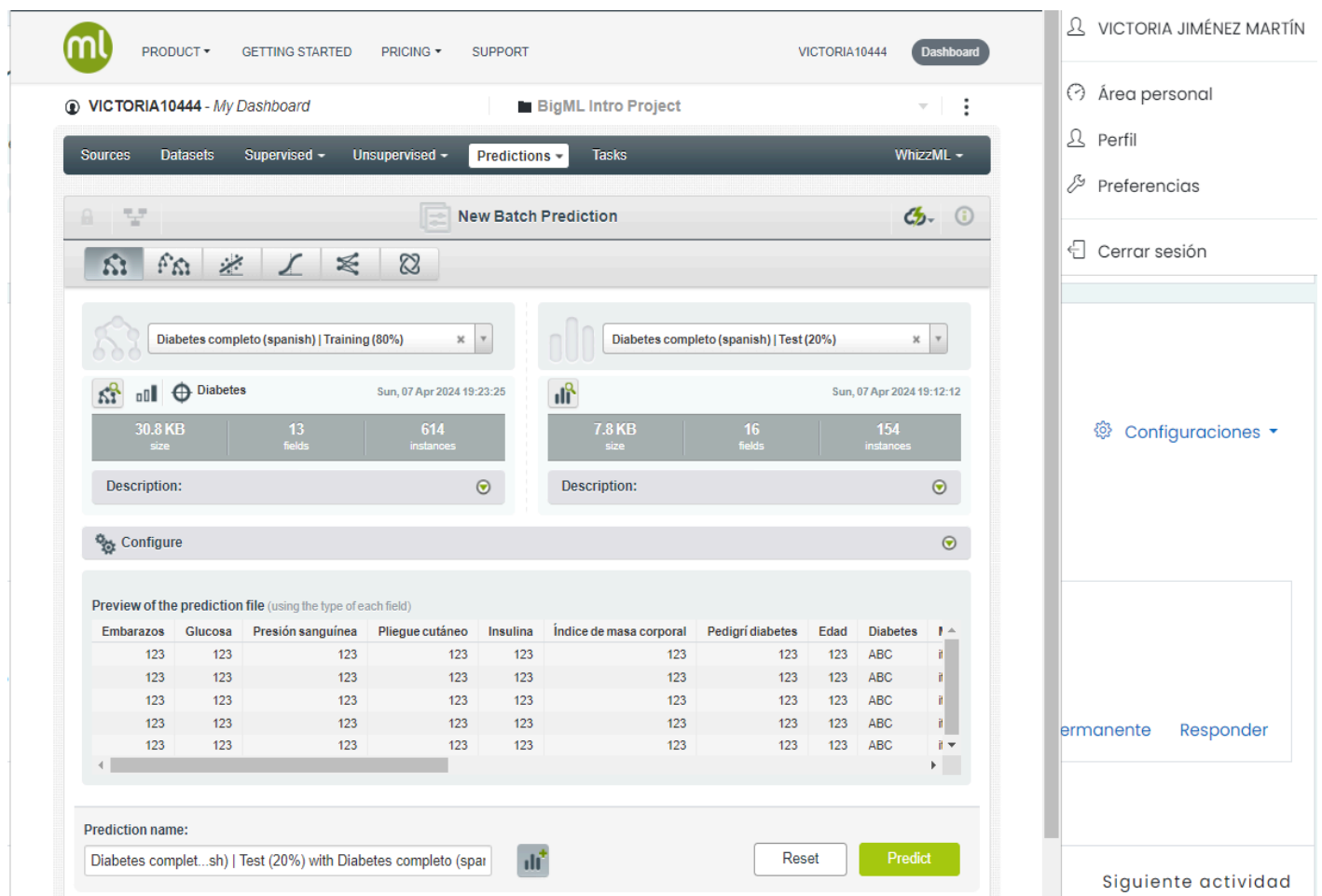
Área personal

Perfil


Preferencias

Cerrar sesión

Realizamos la predicción por lotes en Batch prediction:



Y descargamos el dataset como .csv:

PRODUCT ▾ GETTING STARTED PRICING ▾ SUPPORTVICTORIA10444 Dashboard

VICTORIA10444 - My Dashboard BigML Intro Project

Sources Datasets Supervised ▾ Unsupervised ▾ Predictions ▾ Tasks WhizzML ▾

Diabetes complet...sh | Test (20%) with Diabetes com...

Diabetes Completo (Spanish) | Training (...)

Diabetes Completo (Spanish) | Test (20%)

Configuration

Output preview

Embarazos	Glucosa	Presión sanguínea	Pliegue cutáneo	Insulina	Índice de masa corporal	Pedigrí diabetes	Edad	Diabetes
6	148	72	35	0	336	627	50	Sí
10	168	74	0	0	38	537	34	Sí
7	100	0	0	0	30	484	32	Sí
1	115	70	30	96	346	529	32	Sí
3	126	88	41	235	393	704	27	No

Download batch prediction

Output dataset

VICTORIA JIMÉNEZ MARTÍN

Área personal

Perfil

Preferencias

Cerrar sesión

Configuraciones ▾

ermanente Responder

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.

Importamos los datos en una hoja de cálculo:

Diabetes										VICTORIA JIMÉNEZ MARTÍN
Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones ...										Área personal
100% 123 Predet... 10 B I A										Perfil
A1	Embarazos									Preferencias
	A	B	C	D	E	F	G	H	I	Cerrar sesión
1	Embarazos	Glucosa	Presión sanguínea	Plegue cutáneo	Insulina	Índice de masa corporal	Pedigrí diabetes	Edad	Diabetes	
2	6	148	72	35	0	336	627	50	Sí	
3	10	168	74	0	0	38	537	34	Sí	
4	7	100	0	0	0	30	484	32	Sí	
5	1	115	70	30	96	346	529	32	Sí	
6	3	126	88	41	235	393	704	27	No	
7	8	99	84	0	0	354	388	50	No	
8	11	143	94	33	146	366	254	51	Sí	
9	1	97	66	15	140	232	487	22	No	
10	5	109	75	26	0	36	546	60	No	
11	3	88	58	11	54	248	267	22	No	
12	7	103	66	32	0	391	344	31	Sí	
13	1	101	50	15	36	242	526	26	No	
14	8	176	90	34	300	337	467	58	Sí	
15	2	84	0	0	0	0	304	21	No	
16	2	109	92	0	0	427	845	54	No	
17	13	126	90	0	0	434	583	42	Sí	
18	15	136	70	32	110	371	153	43	Sí	
19	7	81	78	40	48	467	261	42	No	
20	1	151	60	0	0	261	179	22	No	
21	5	124	74	0	0	34	22	38	Sí	
22	5	78	48	0	0	337	654	25	No	
23	0	113	76	0	0	333	278	23	Sí	
24	3	120	70	30	135	429	452	30	No	
25	1	117	88	24	145	345	403	40	Sí	
26	2	96	68	13	49	211	647	26	No	
27	0	93	60	25	92	287	532	22	No	
28	1	136	74	50	204	374	399	24	No	

Configuraciones

Permanente Responder

Ahora tendremos que hallar, los valores siguientes:

- VP: Verdaderos positivos: =CONTAR.SI.CONJUNTO(I:I; "Sí"; Q:Q; "Sí")

	P	Q	R	S	T	U
1	Fecha de diagnóstico	Diabetes	VP	27	154	
2		1 Sí	FN	25		
3		4 Sí	FP	28		
4		5 Sí	VN	74		
5		3 No	Errores totales	53		
6		3 No				
7		3 No				
8		4 No				
9		5 No				
10		1 Sí				
11		3 No				

- FN: Falsos negativos: =CONTAR.SI.CONJUNTO(I:I; "Sí"; Q:Q; "No")

	P	Q	R	S	T	U
1	Fecha de diagnóstico	Diabetes	VP	27	154	
2		1 Sí	FN	25		
3		4 Sí	FP	28		
4		5 Sí	VN	74		
5		3 No	Errores totales	53		
6		3 No				
7		3 No				
8		4 No				
9		5 No				
10		1 Sí				

- FP: Falsos positivos: =CONTAR.SI.CONJUNTO(I:I; "No"; Q:Q; "Sí")

	P	Q	R	S	T	U
1	Fecha de diagnóstico	Diabetes	VP	27	154	
2		1 Sí	FN	25		
3		4 Sí	FP	28		
4		5 Sí	VN	74		
5		3 No	Errores totales	53		
6		3 No				
7		3 No				
8		4 No				
9		5 No				

- VN: Verdaderos negativos: =CONTAR.SI.CONJUNTO(I:I; "No"; Q:Q; "No")

Diabetes

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones ...

100% 123 Predet... - 10 + B I A :

S4 fx =CONTAR.SI.CONJUNTO(I:I; "No"; Q:Q; "No")

	P	Q	R	S	T	U
1	Fecha de diagn	Diabetes	VP	27	154	
2	1	Sí	FN	25		
3	4	Sí	FP	28		
4	5	Sí	VN	74		
5	3	No	Errores totales	53		
6	3	No				
7	3	No				
8	4	No				
9	5	No				
10	1	Sí				
11	3	No				
12	2	Sí				

VICTORIA JIMÉNEZ MARTÍN

Área personal

Perfil

Preferencias

Cerrar sesión

Siguiente actividad

- Total errores: =S2+S3 (FN + FP)

Diabetes

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones ...

100% 123 Predet... - 10 + B I A :

S5 fx =S2+S3

	P	Q	R	S	T	U
1	Fecha de diagn	Diabetes	VP	27	154	
2	1	Sí	FN	25		
3	4	Sí	FP	28		
4	5	Sí	VN	74		
5	3	No	Errores totales	53		
6	3	No				
7	3	No				
8	4	No				
9	5	No				
10	1	Sí				

VICTORIA JIMÉNEZ MARTÍN

Área personal

Perfil

Preferencias

Cerrar sesión

Construye la matriz de confusión, rellenando los valores correspondientes.

A continuación creamos la matriz de confusión en el excel:

Diabetes									
Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones ...									
100% 123 Predet... 10 B I A									
X9		Q	R	S	T	U	V	W	X
1	diagnó	Diabetes	VP	27					
2	1	Sí	FN	25					
3	4	Sí	FP	28					
4	5	Sí	VN	74					
5	3	No	Errores totales	53					
6	3	No							
7	3	No							
8	4	No							
9	5	No							
10	1	Sí		Predicción Sí	Predicción No				
11	3	No	Realidad Sí	27	25				
12	2	Sí	Realidad No	28	74				
13	2	Sí							
14	4	Sí							
15	3	No							
16	4	Sí							
17	5	Sí							
18	5	No							
19	1	No							
20	4	No							
21	2	Sí							
22	2	Sí							
23	5	No							
24	5	No							
25	5	Sí							
26	1	No							

Analiza los resultados. ¿Es fiable el modelo?

$$\text{Exactitud} = \frac{VP + VN}{VP + FN + FP + VN} = 0,6558 \times 100 = 65,58\%$$

$$\text{Precisión} = \frac{VP}{VP + FP} = 0,490 \times 100 = 49,00\%$$

$$\text{Sensibilidad} = \frac{VP}{VP + FN} = 0,5192 \times 100 = 51,92\%$$

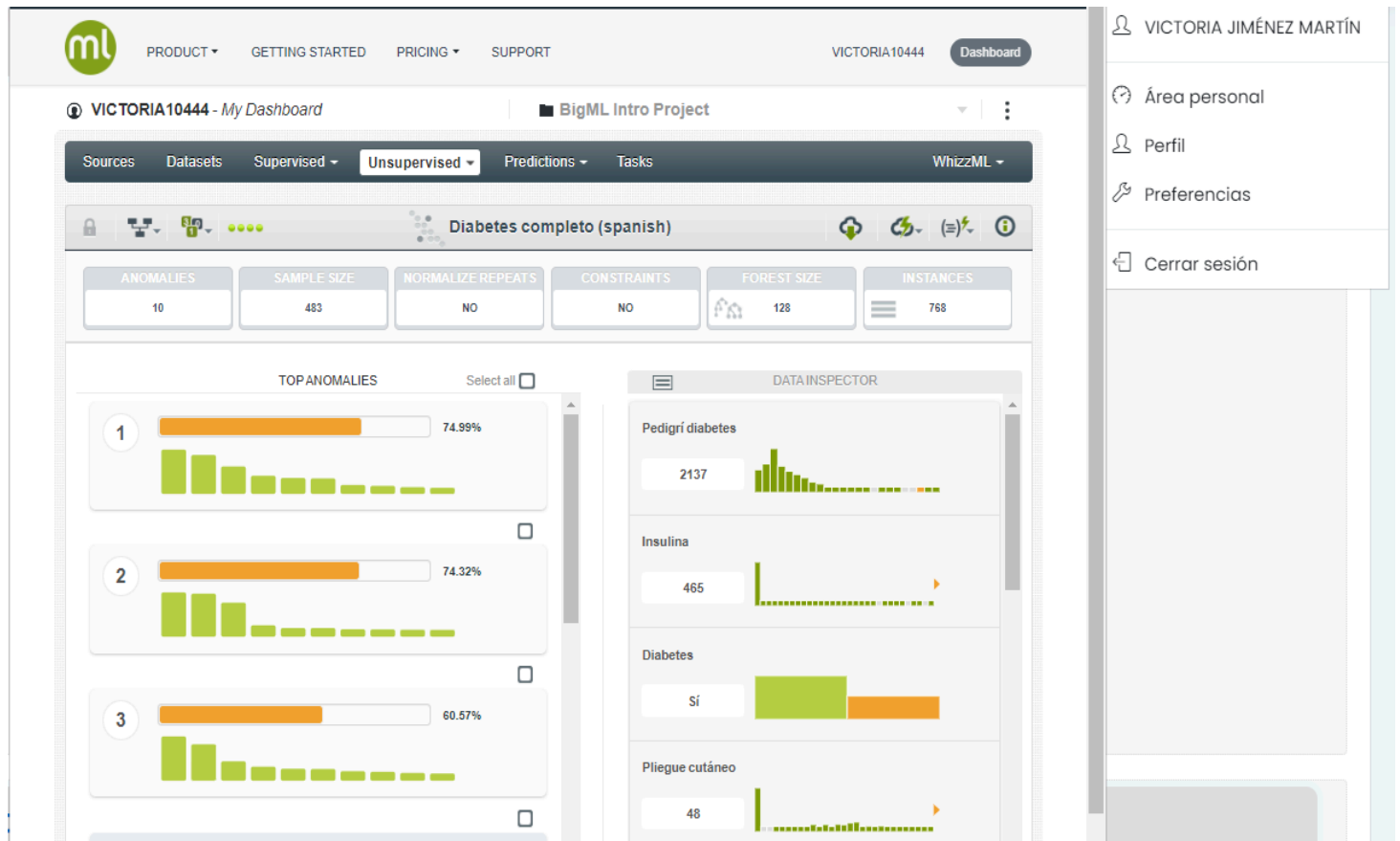
$$\text{Especificidad} = 2x \frac{\text{Precisión} \times \text{sensibilidad}}{\text{Precisión} + \text{sensibilidad}} = 0,5042 \times 100 = 50,42\%$$

En base a los siguientes datos, podemos decir que el modelo no es muy fiable para predecir la diabetes ya que tanto la precisión como la sensibilidad son solo un poco mejores que el azar, pero los resultados son demasiado bajos, como para considerarlos fiables.

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

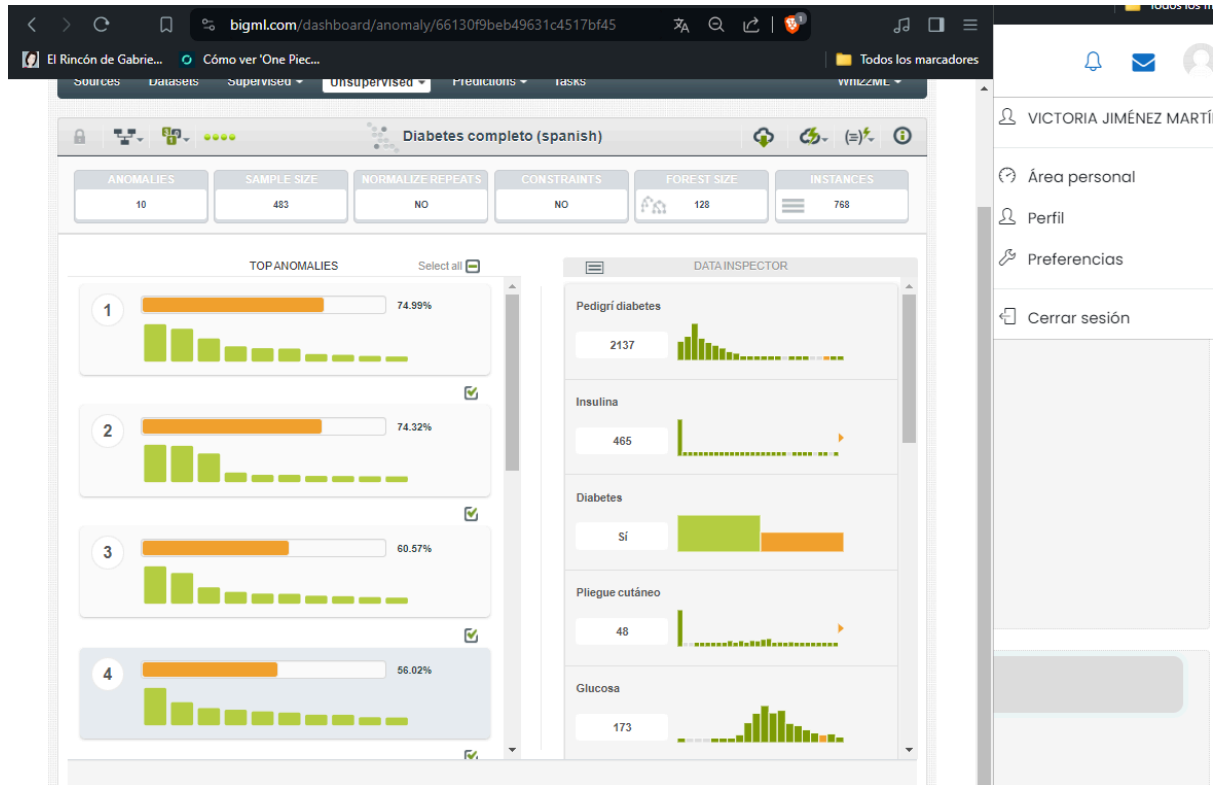
Aplica el modelo de detección de anomalías en BigML dentro de las funciones rápidas de algoritmos no supervisados.

Aplicamos el modelo de detección de anomalías y nos muestra lo siguiente:



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

Seleccionamos las anomalías que consideremos:

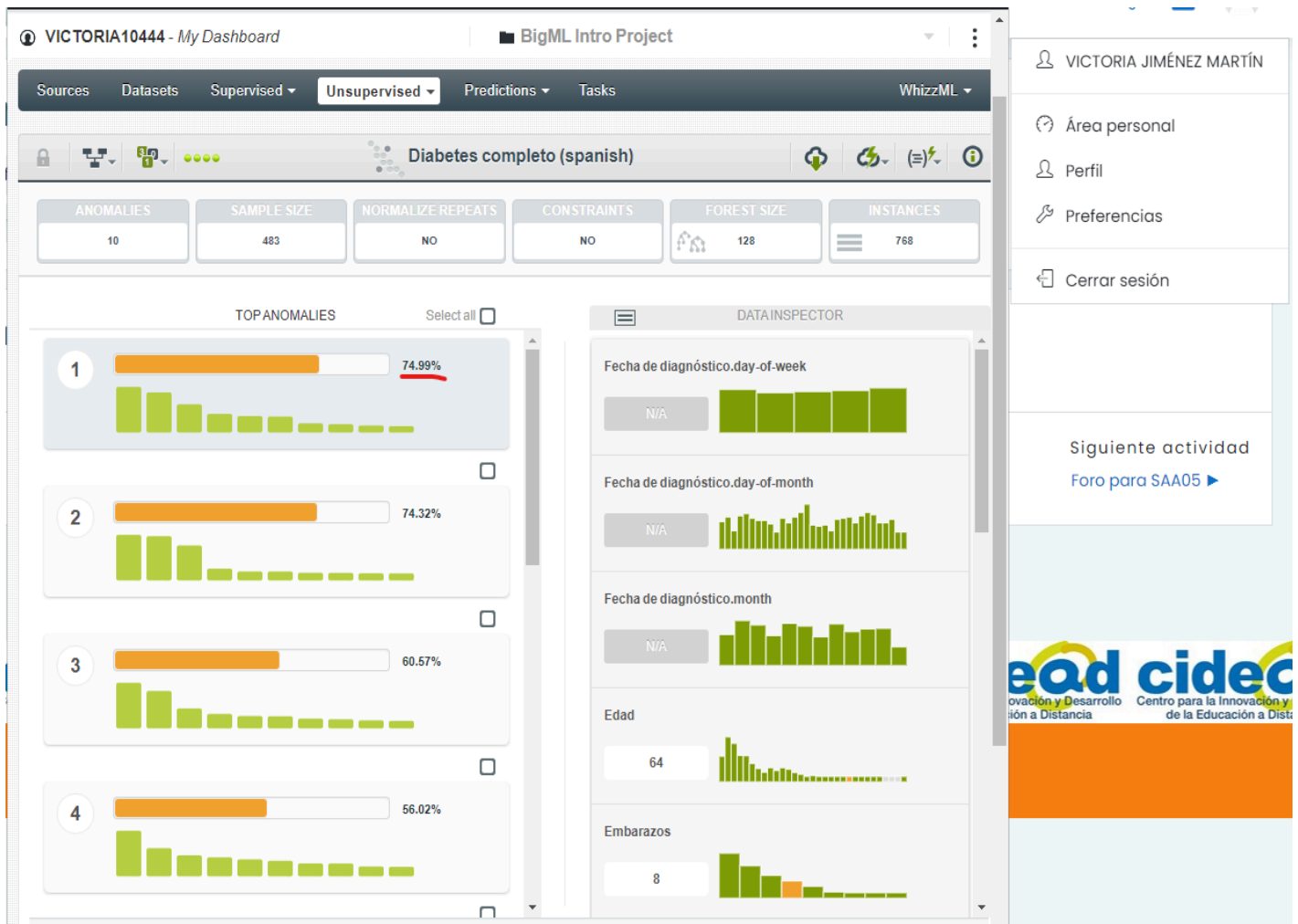


Merece la pena ya que tienen un alto porcentaje de anomalías, sobre todo las 3 primeras.

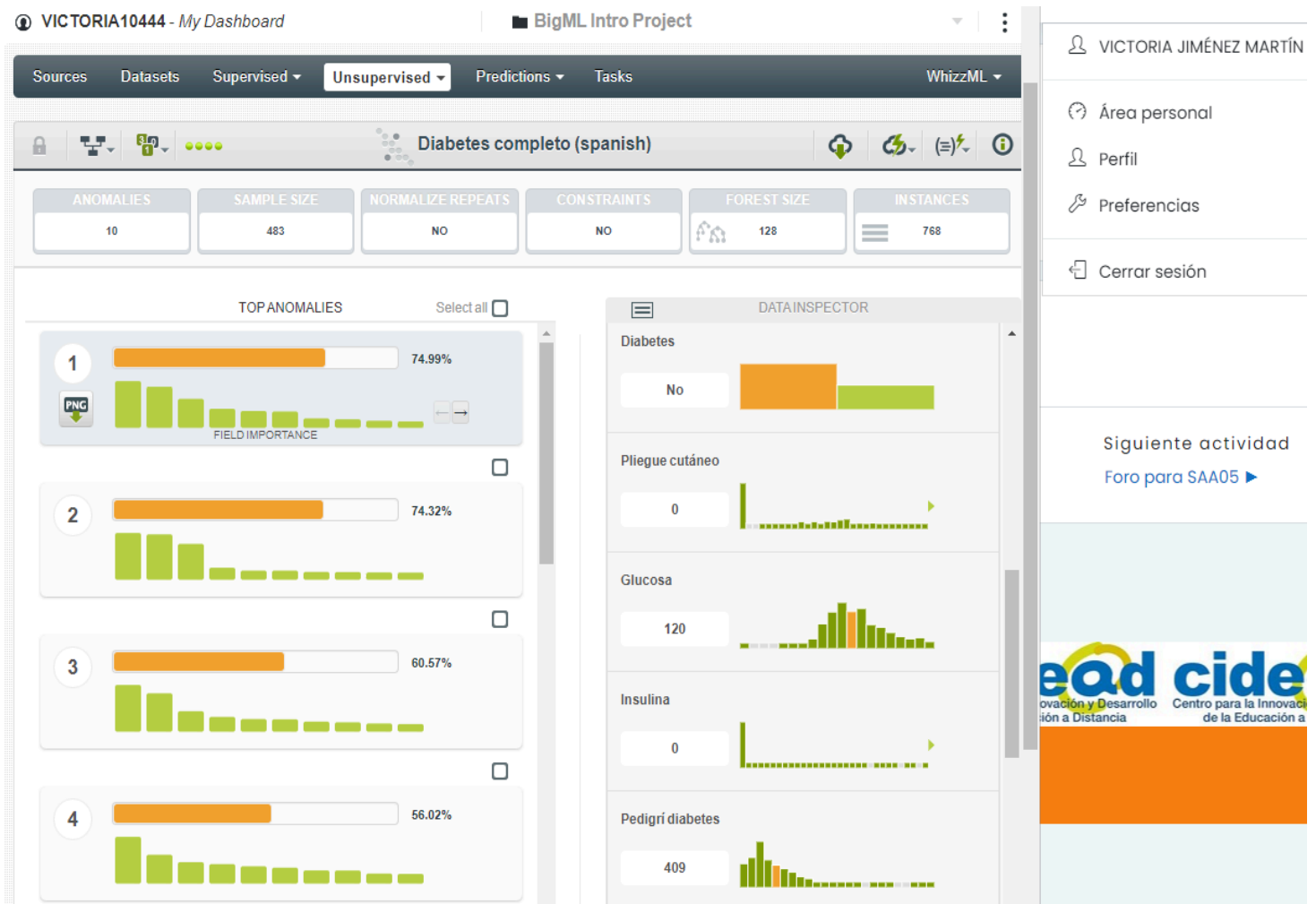
Si crees que son importantes, crea un dataset con ellas para analizarlas

Nos vamos a basar en el primer registro para analizar y mostrar que el dataset, es propenso a errores:

En el primer registro en el que se un mayor porcentaje de anomalías es el de 74,99%.



Si observamos el panel de la derecha que es el histograma del paciente, no se presentan anomalías muy notables., pero si nos vamos a datos más concretos como el de la pedigrí diabetes y los niveles de glucosa, observamos lo siguiente:



Vemos que presenta que no tiene diabetes pero el nivel de glucosa son altos (120) y según la página de <https://medlineplus.gov/spanish/ency/article/000305.htm>, la diabetes de tipo 1 se podría llegar a diagnosticar con una glucemia de 126, por lo que podría estar al límite de padecer diabetes en un futuro o puede que tenga antecedentes en la familia de diabetes y sea penpenso a tenerla.

Pruebas y exámenes

La diabetes se diagnostica con los siguientes exámenes de sangre:

- Nivel de **glucemia en ayunas**. La diabetes se diagnostica si este es de 126 mg/dL (7 mmol/L) o superior en dos ocasiones diferentes.
- Nivel de glucemia aleatoria (sin ayunar) -- Usted puede tener diabetes si este es de 200 mg/dL (11.1 mmol/L) o superior y tiene síntomas como aumento de la sed, de la orina y fatiga, (Esto se debe confirmar con un examen en ayunas).
- **Prueba de tolerancia a la glucosa oral**. La diabetes se diagnostica si el nivel de glucosa es de 200 mg/dL (11.1 mmol/L) o superior 2 horas después de tomar una bebida azucarada especial.
- **Examen de hemoglobina A1C (A1C)**. La diabetes se diagnostica si el resultado del examen es 6.5% o superior.

Y por el apartado de función de pedigrí de diabetes, vemos que también presenta unos altos para no tener diabetes y que según el siguiente estudio <https://ri.conicet.gov.ar/handle/11336/61863>, la pedigrí diabetes se presenta para datos que si se tiene diabetes:

Análisis multidimensional de una base de datos de mujeres PIMA

Título: Multidimensional analysis from a database of PIMA women

Tarres, Maria Cristina; **Moscoloni, Nora Ana Maria** ; **Navone, Hugo Daniel** ; **D'Ottavio, Adriana Leticia**

Fecha de publicación: 11/2016

Editorial: Universidad de Sonora. División de Ciencias Biológicas y de la Salud

Revista: Biotecnia

ISSN: 1665-1456

Idioma: Español

Tipo de recurso: Artículo publicado

Clasificación temática:

Tópicos Sociales

Resumen

Se presenta una tipología multidimensional de mujeres estadounidenses Pima con datos de la Pima Indians Diabetes Database, analizada mediante Componentes Principales y posterior clasificación. Fueron construidos tres clusters individuales con: (1) 64% de positividad para diabetes, glucemia acorde con tolerancia alterada a la glucosa, hiperinsulinemia, obesidad según índice de masa corporal, grosor del pliegue de la piel del tríceps superior al promedio general y presión diastólica cercana a prehipertensión; (2) 50% de positividad para diabetes, glucemia cercana al límite inferior de tolerancia alterada a la glucosa, índice de masa corporal correspondiente a obesidad y grosor del pliegue de la piel del tríceps, presión diastólica, edad y número de embarazos mayores a la media general; (3) 16% de positividad para diabetes, índice de masa corporal indicando sobrepeso y número de embarazos, función de **pedigrí** de diabetes, edad, glucemia, insulinemia y grosor del pliegue de la piel del tríceps menores al promedio general. Resulta destacable el mayor grado de antecedentes familiares en el grupo de riesgo y la relevancia del sobrepeso así como la configuración espacial de los individuos revelando posibles fases en el desarrollo diabético de estas mujeres. Finalmente, se subraya la utilidad de la clusterización en problemas biológicos.

Ahora seleccionaremos los 5 primeros registros y crearemos el dataset:

VICTORIA10444 - My Dashboard | **BigML Intro Project**

Sources | Datasets | Supervised | **Unsupervised** | Predictions | Tasks | WhizzML

Diabetes completo (spanish)

ANOMALIES: 10 | SAMPLE SIZE: 483 | NORMALIZE REPEATS: NO | CONSTRAINTS: NO | FOREST SIZE: 128 | INSTANCES: 768

TOP ANOMALIES | Select all

- 1 | 74.99%
- 2 | 74.32%
- 3 | 60.57%
- 4 | 56.02%

DATAINSPECTOR

Pedigrí diabetes: 629

Edad: 24

Presión sanguínea: 65

Medicación previa: N/A

Observaciones: N/A

New dataset name: Diabetes completo (spanish) top 10 anomalies dataset

Create dataset


Victoria Jiménez Martín

- Área personal
- Perfil
- Preferencias
- Cerrar sesión

Siguiente actividad: [Foro para SAA05](#)

ead cidec
Innovación y Desarrollo | Centro para la Innovación y de la Educación a Distancia

Vemos que nos ha creado el siguiente dataset:


PRODUCT ▾
GETTING STARTED
PRICING ▾
SUPPORT
VICTORIA10444
Dashboard

VICTORIA10444 - My Dashboard
BigML Intro Project










Sources
Datasets
Supervised ▾
Unsupervised ▾
Predictions ▾
Tasks
WhizzML ▾

Diabetes completo (spanish) top 5 anomalies ...

Diabetes

ABC

Search by name

Name	Type	Count	Missing	Errors	Histogram
Embarazos	123	5	0	0	
Glucosa	123	5	0	0	
Presión sanguínea	123	5	0	0	
Pliegue cutáneo	123	5	0	0	
Insulina	123	5	0	0	
Índice de masa corporal	123	5	0	0	
Pedigrí diabetes	123	5	0	0	
Edad	123	5	0	0	
Diabetes	ABC	5	0	0	
Medicación previa	items	0	5	0	

Procedemos a realizar el entrenamiento y vemos que del dataset creado y posteriormente entrenado, encuentra dos anomalías:

[illegible]

Clicamos sobre este, y nos iremos a Anomaly Score:

VICTORIA10444 - My Dashboard | BigML Intro Project

Sources Datasets Supervised Unsupervised Predictions Tasks WhizzML

Diabetes completo (spanish) top 5 anomalies ...

ANOMALIES: 10 SAMPLE SIZE: 2 NORMALIZE REPEATS: NO

ANOMALY SCORE

- ANOMALY SCORE
- BATCH ANOMALY SCORE
- DELETE ANOMALY DETECTOR
- MOVE TO...

TOP ANOMALIES

1 0.03%

2 0.02%

3 0.02%

Embarazos: 2

Fecha de diagnóstico.day-of-week: N/A

Presión sanguínea: 65

VICTORIA JIMÉNEZ MARTÍN

- Área personal
- Perfil
- Preferencias
- Cerrar sesión

Siguiete actividad

[Foro para SAA05](#)

Pero vamos el porcentaje es mínimo:

Sources Datasets Supervised Unsupervised Predictions Tasks WhizzML

Diabetes completo (spanish) top 5 anomalies ...

Score: 0.02%

All input fields: ☒

Embarazos: 14.00%

Fecha de diagnóstico.day-of-week: 11.72%

Presión sanguínea: 10.94%

Índice de masa corporal: 10.10%

Pedigri diabetes: 9.38%

Fecha de diagnóstico.day-of-month: 8.56%

Insulina: 7.03%

Edad: 7.03%

Fecha de diagnóstico.month: 7.03%

Glucosa: 6.25%

Pliegue cutáneo: 5.47%

Diabetes: 2.34%

New anomaly score name

Diabetes completo (spanish) top 5 anomalies dataset | Training (80%)

Score

VICTORIA JIMÉNEZ MARTÍN

- Área personal
- Perfil
- Preferencias
- Cerrar sesión

Siguiete actividad

[Foro para SAA05](#)

Ahora vamos a sacar el Batch Score a ver si encuentra más anomalías:

VICTORIA10444 - My Dashboard

BigML Intro Project

SourcesDatasetsSupervisedUnsupervisedPredictionsTasksWhizzML

New Batch Anomaly Score

Diabetes completo (spanish) top 5 anomalies data...

Mon, 08 Apr 2024 19:46:23

0.2 KBsize15 fields4 instances10 top anomalies

Description:

Configure

Preview of the prediction file (using the type of each field)

Embarazos	Glucosa	Presión sanguínea	Pliegue cutáneo	Insulina	Índice de masa corporal	Pedigrí diabetes	Edad	Diabetes	Med
123	123	123	123	123	123	123	123	ABC	item:
123	123	123	123	123	123	123	123	ABC	item:
123	123	123	123	123	123	123	123	ABC	item:
123	123	123	123	123	123	123	123	ABC	item:
123	123	123	123	123	123	123	123	ABC	item:

Prediction name:

Batch Anomalyset | Test (20%) with Diabetes completo (spanish)

Reset

Score

Diabetes completo (spanish) top 5 anomalies data...

Mon, 08 Apr 2024 19:14:24

0.0 KBsize16 fields1 instances

Description:

Configure

VICTORIA JIMÉNEZ MARTÍN

Área personal

Perfil

Preferencias

Cerrar sesión

Siguiente actividad

Foro para SAA05

Y vemos que la puntuación obtenida es la siguiente:

ml

PRODUCTGETTING STARTEDPRICINGSUPPORT

VICTORIA10444Dashboard

VICTORIA10444 - My Dashboard

BigML Intro Project

SourcesDatasetsSupervisedUnsupervisedPredictionsTasksWhizzML

Batch Anomalyset | Test (20%) with Diabetes complet...

Diabetes Completo (Spanish) Top 5 Ano...

Diabetes Completo (Spanish) Top 5 Ano...

Output preview

stico	Fecha de diagnóstico.year	Fecha de diagnóstico.month	Fecha de diagnóstico.day-of-month	Fecha de diagnóstico.day-of-week	score
					2.5E-4

Download batch score

Output dataset

VICTORIA JIMÉNEZ MARTÍN

Área personal

Perfil

Preferencias

Cerrar sesión

Siguiente actividad

Foro para SAA05