

Machine Learning orientado a Limpieza de datos

Algo sobre mí



Dicen que la curiosidad mató al gato...pero murió
sabiendo

Eli Carreño

- ⌘ Ingeniera Civil Industrial, Mención en Modelamiento matemático.
- ⌘ Diplomado en Dirección de Proyectos.
- ⌘ ~~Unicørniø~~ Data Scientist hace 3 años (?).
- ⌘ Diplomado en Inteligencia Artificial.
- ⌘ Lead Data Science en Women Who Code.
- ⌘ Data Scientist en Walmart.

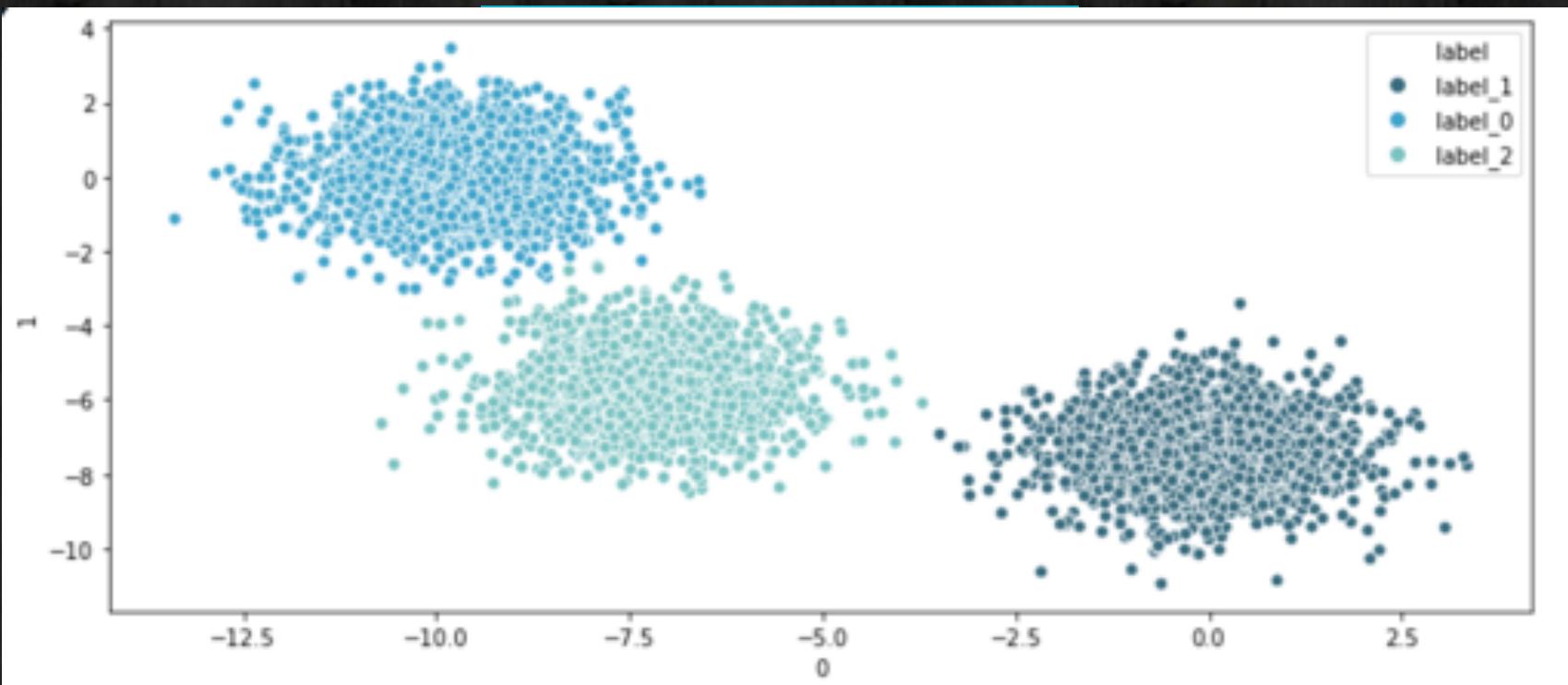


BASADO EN
HECHOS REALES

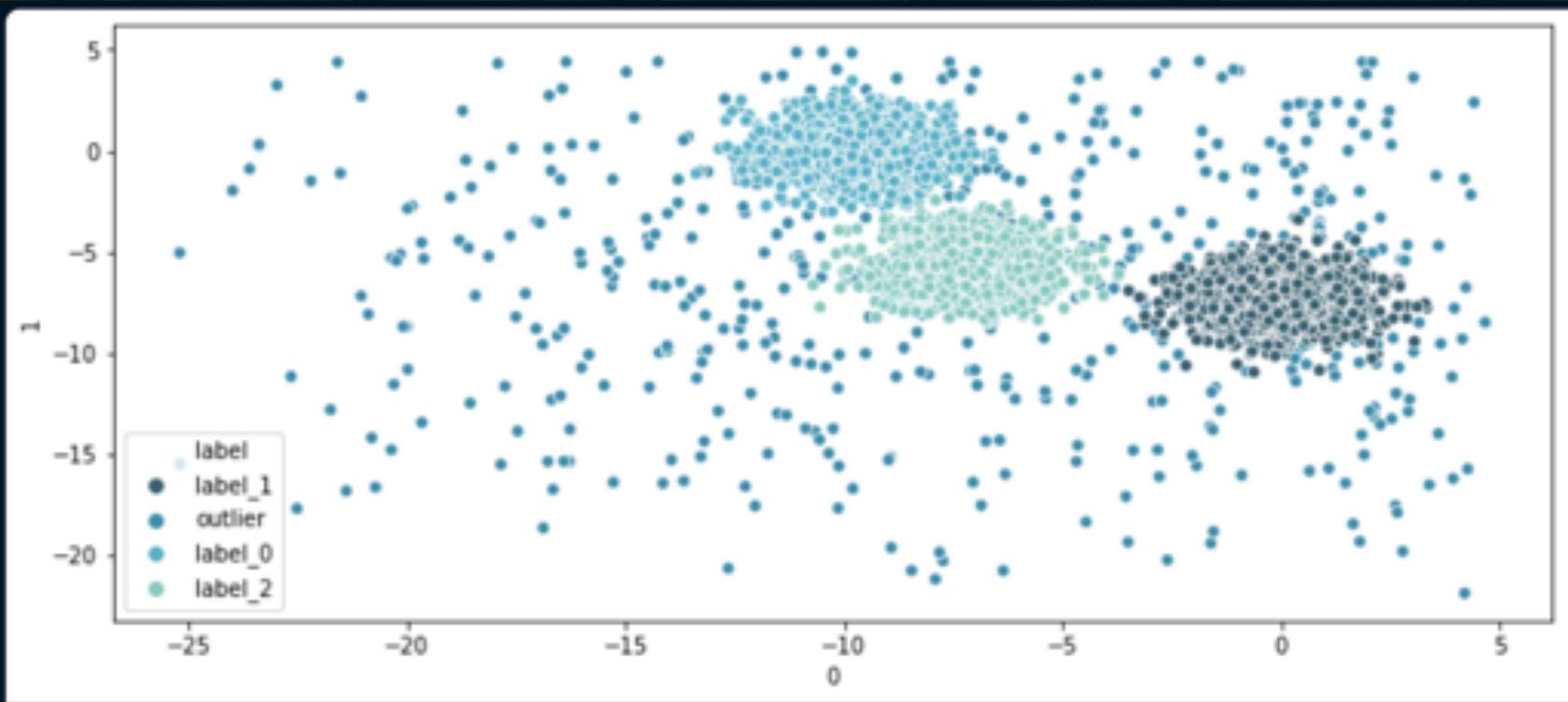
1. El problema

Limpieza de datos....

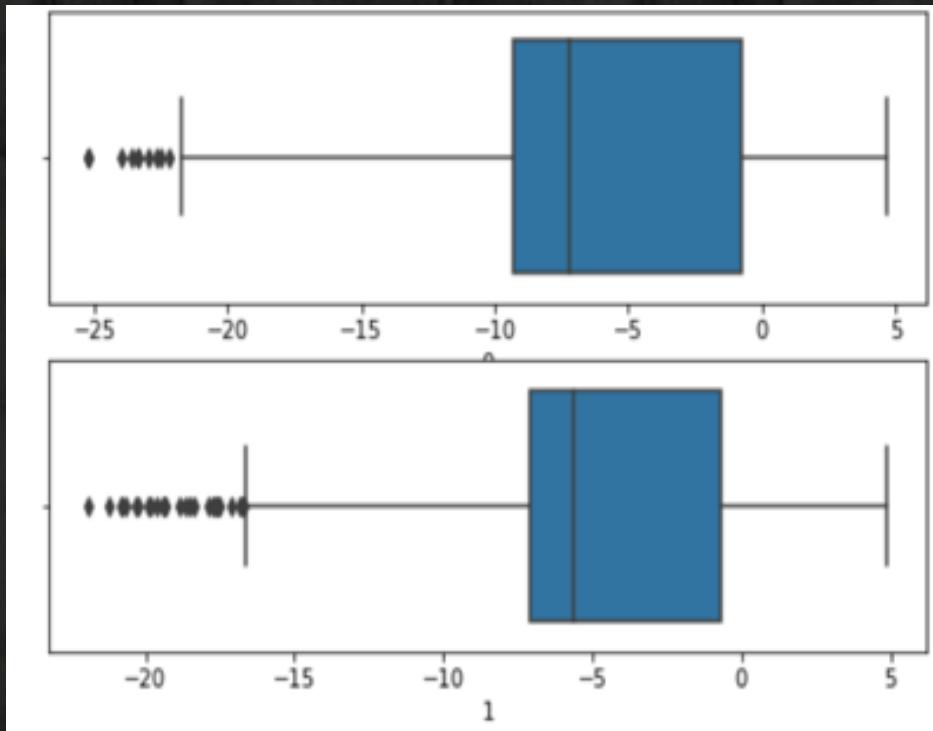
Considerando el siguiente dataset



Mejor consideremos este

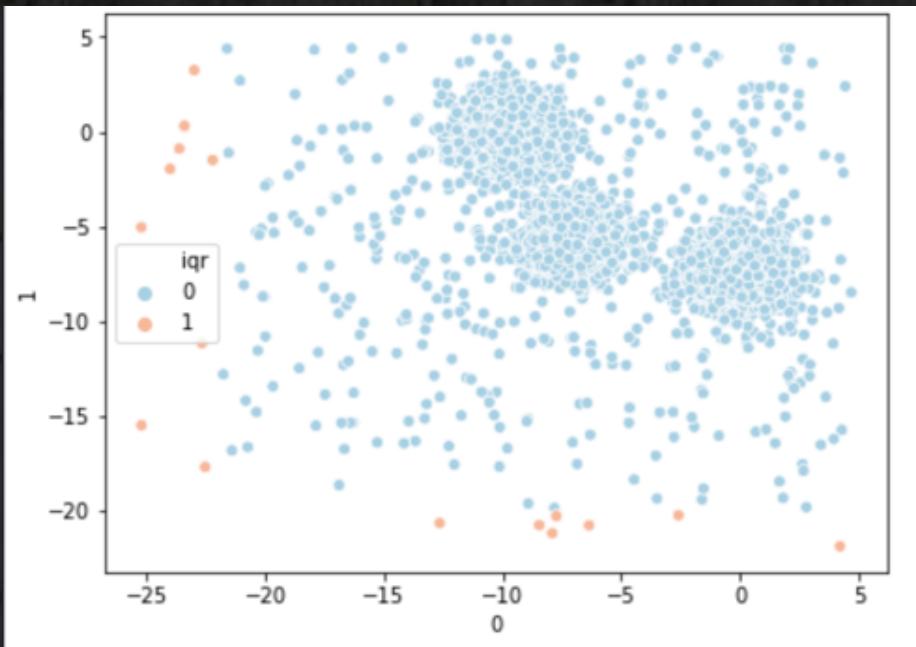


El dataset del terror.



- Sensores viejos y descalibrados.
- Correa de transporte inestable.
- Valores corregidos por reglas de negocio perdidas

IQR = FAIL



*¿Cómo borrar esos
puntos que sé que son
atípicos?*



DBSCAN



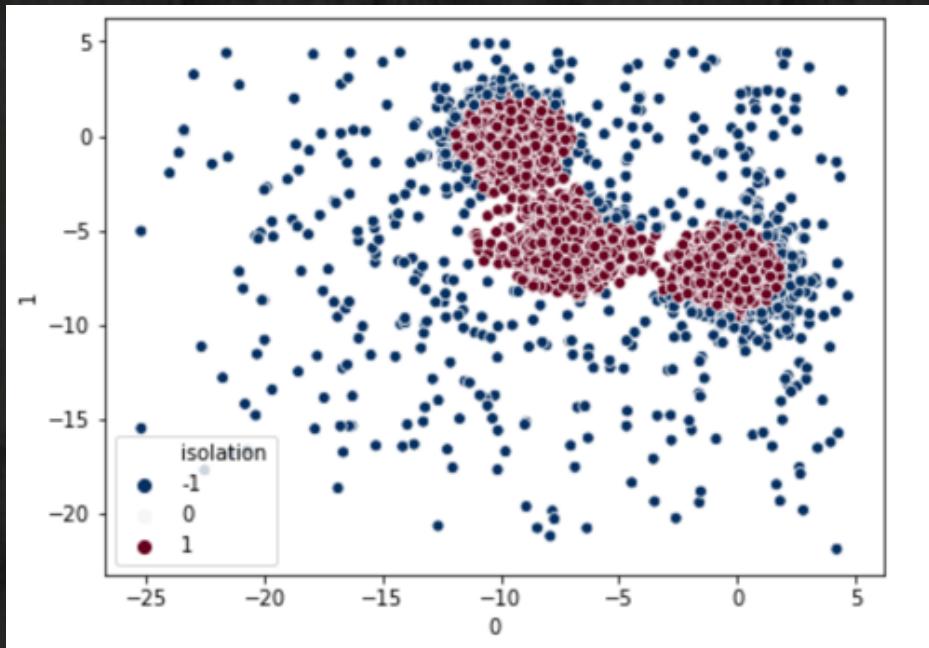
DBSCAN



Isolation Forest

Isolation Forest para reconocimiento de datos atípicos

- Algoritmo específico para anomalías
“Divide y vencerás”
- Liviano de ejecutar, finaliza en tiempos prudentes
- Funciona mejor que la idea del DBSCAN o incluso un HDBSCAN.



2. Agregando Complejidad

Izzi pizza limpiar un dataset con 2 variables

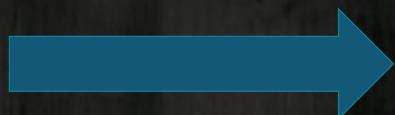
Elevando la apuesta

Prueba 1

5.000 registros

2 variables

2 categorías



Prueba 2

100.000 registros

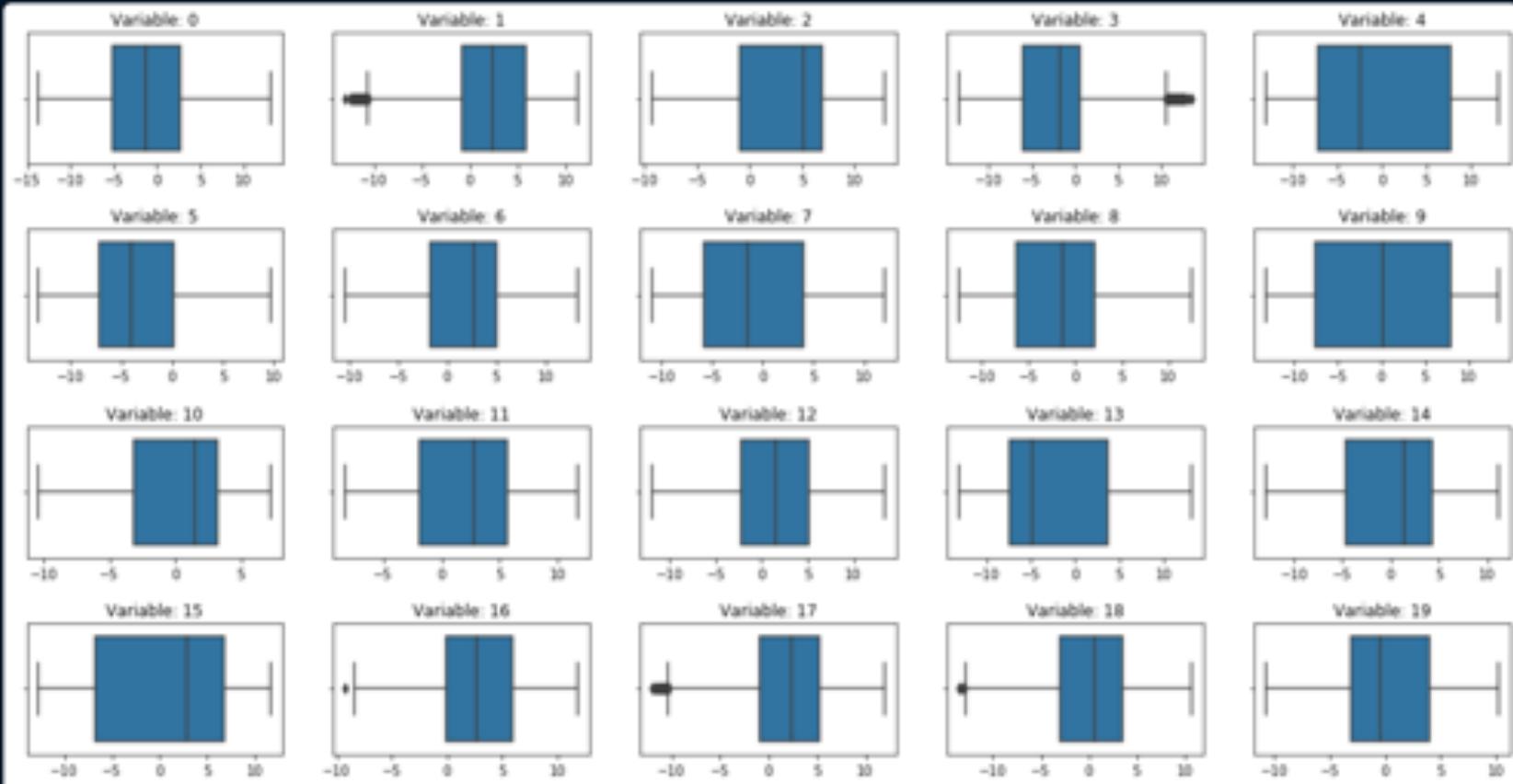
20 variables

10 categorías

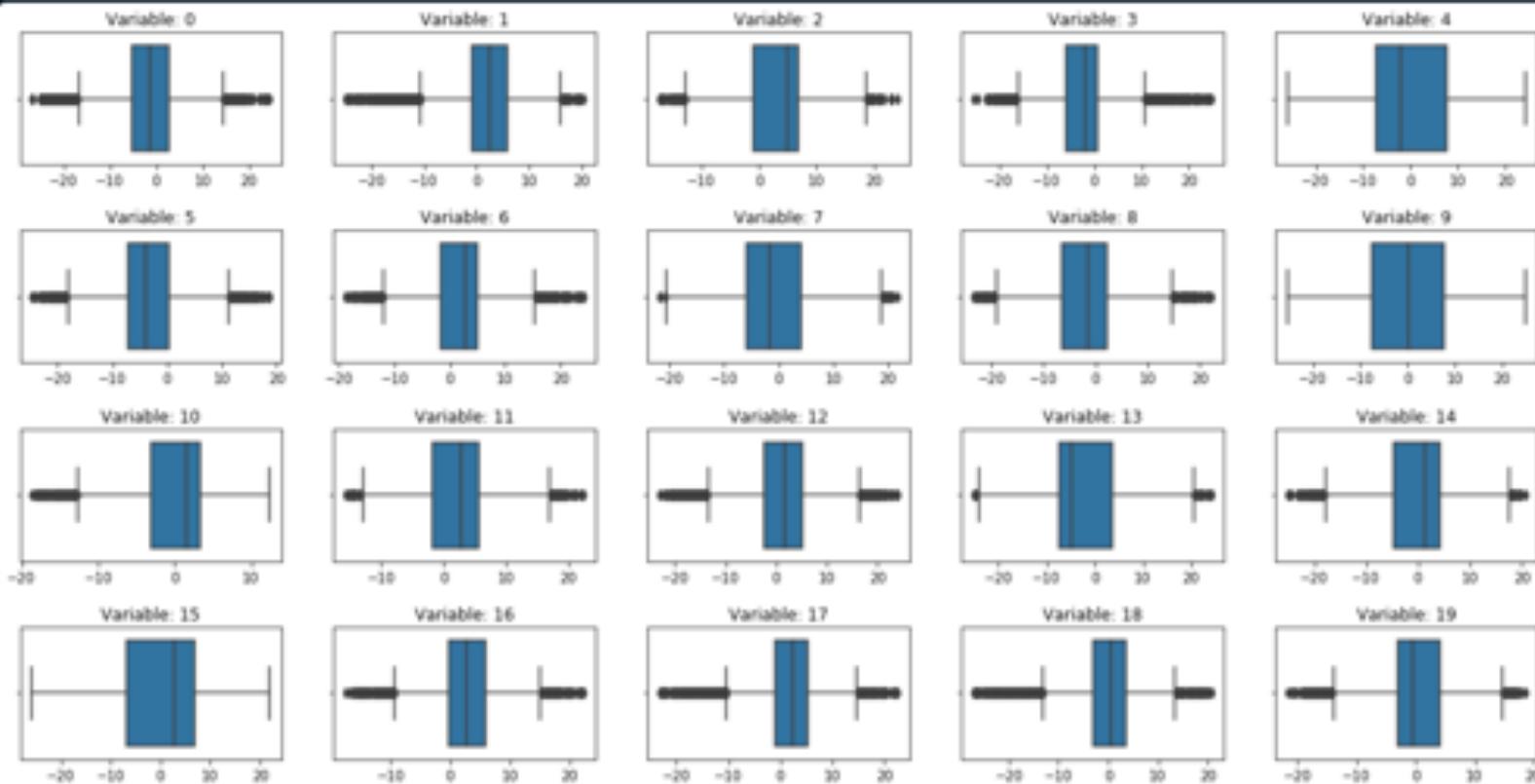
To-do list

- Obtener la data
- Agregar datos sucios
- Isolation Forest sobre datos puros
- Revisar influencia de modelos que reducen dimensiones(PCA, UMAP)

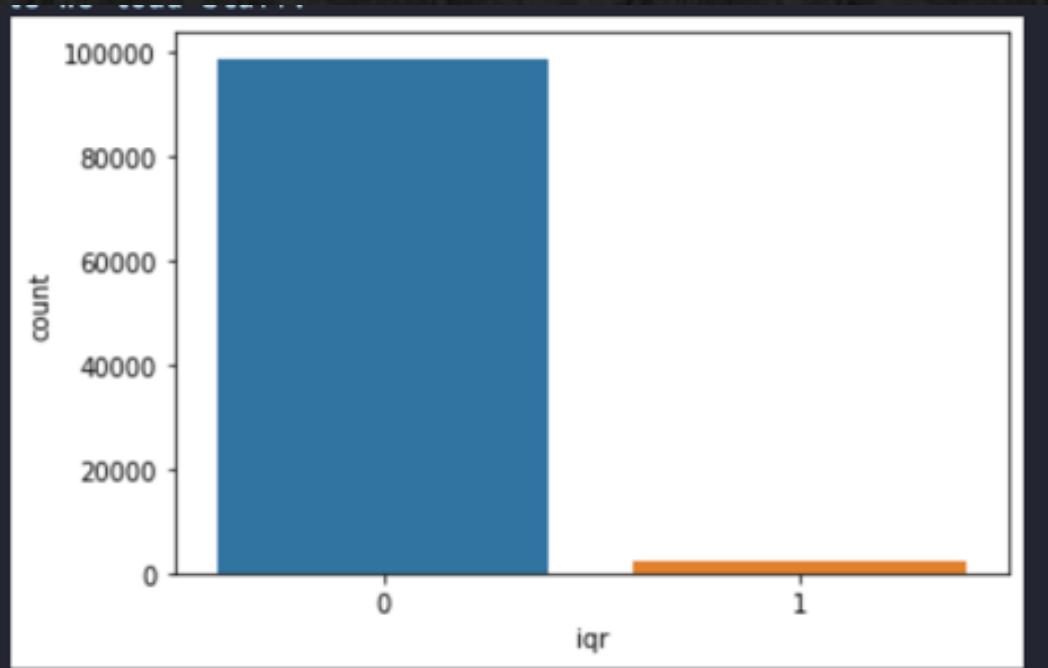
Datos originales



Data arruinada



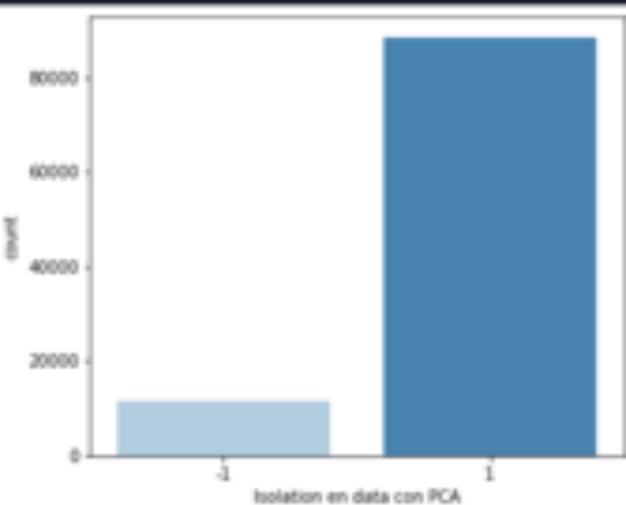
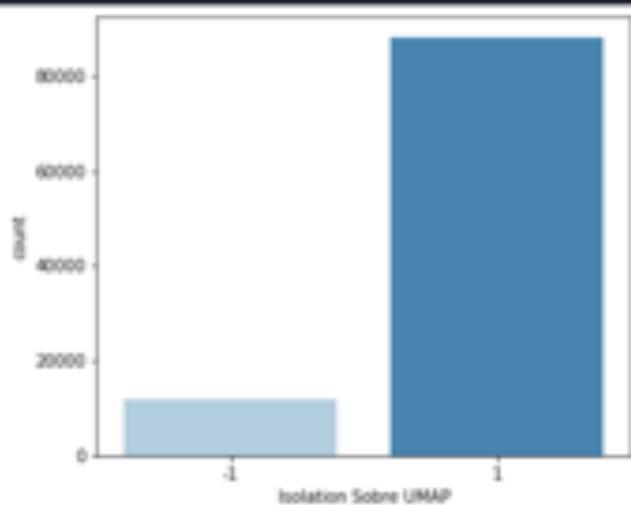
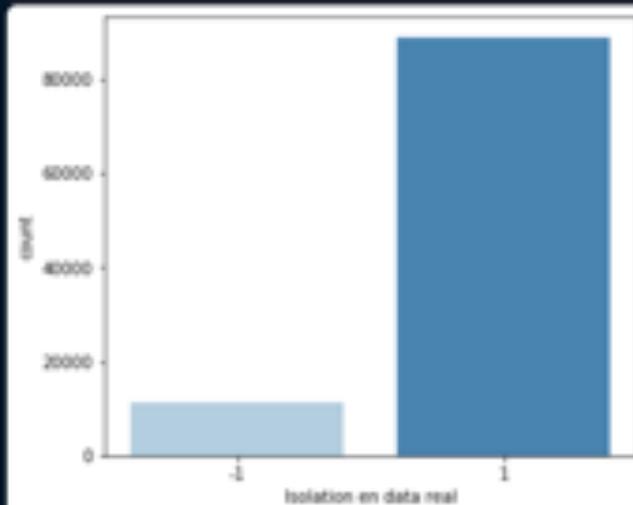
Just in case



IQR de nuevo Fail

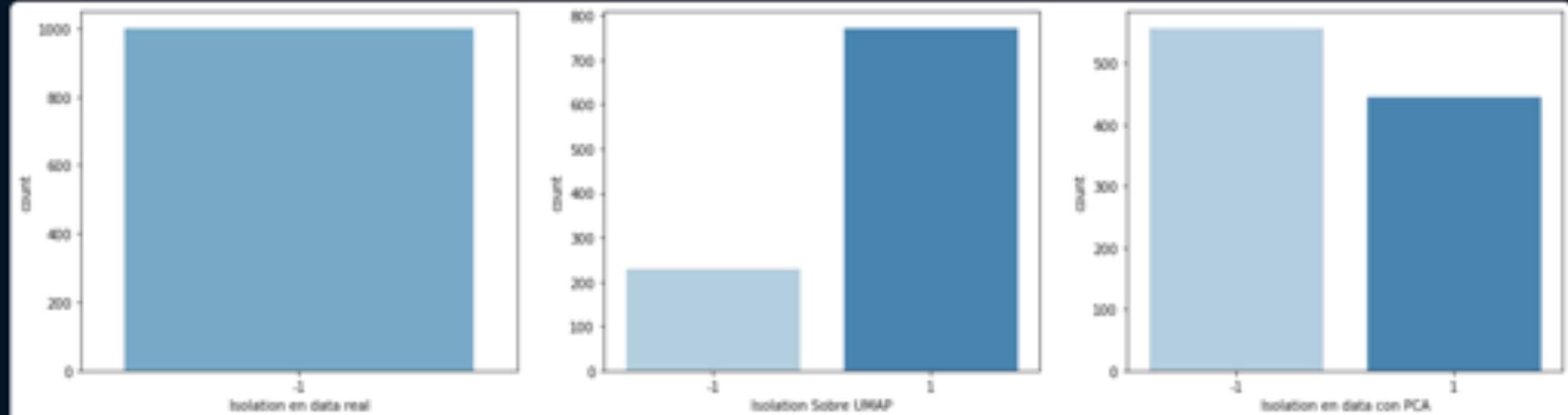
Resultados Isolation Forest

- ✓ Sigue siendo liviano de ejecutar
- ✓ Limpia más datos que IQR.
- ✓ A priori es insensible a reducción de dimensiones.



Resultados Isolation Forest

- ✓ El resultado de mejor calidad se obtiene aplicando Isolation Forest sobre los datos puros.



3. Que el mundo arda.

Lo mismo, pero ahora con un dataset de grandes ligas

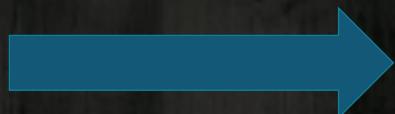
Elevando la apuesta

Prueba 2

100.000 registros

20 variables

10 categorías



Prueba 3

25 Millones de
registros

20 variables

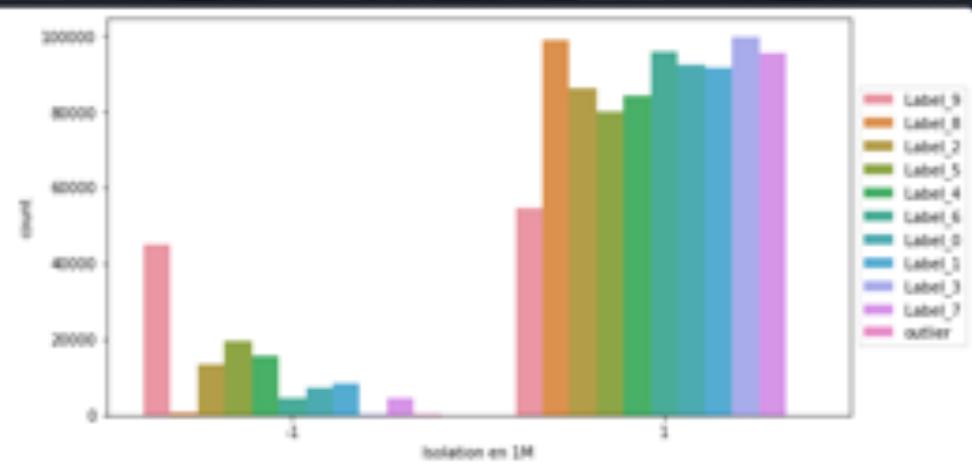
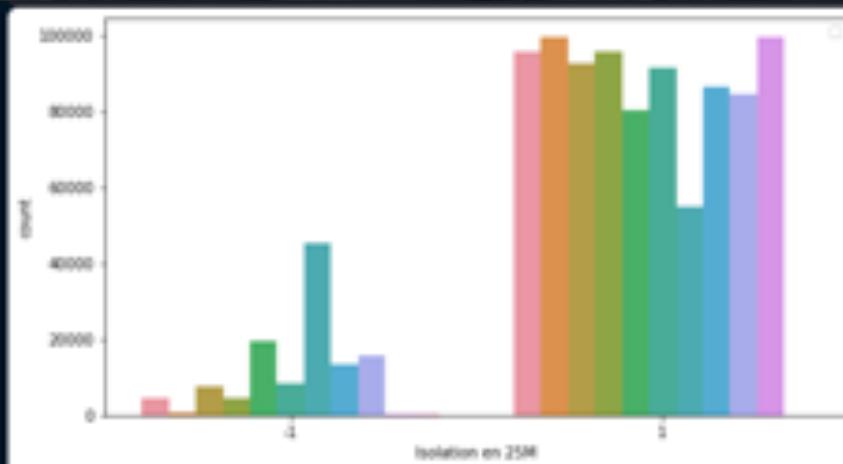
10 categorías

To-do list

- Obtener la data
- Agregar datos sucios
- Isolation Forest a Morir
- Samplear una muestra (1 millón de registros) y con eso probar 2 alternativas:
 - Entrenar Isolation Forest y luego extrapolarlo.
 - Usar un Random Forest para la extrapolación.

Isolation Forest en datasets grandes

- ✓ Curiosamente, sigue ejecutándose en poco tiempo.
- ✓ Pierde precisión para detector atípicos



Random Forest en acción.

88%

Isolation Forest entrenado con 1Mi y usado para predecir los 24Mi restantes.

93%

Isolation Forest con 1Mi de registros. Se usa ese dataset para entrenar un RF y con ese predecir los 24 Mi restantes

4. Conclusiones.

Finally the end.

*Machine learning no es solo el
entregable final de Data Science*

Consideraciones Finales

- ✿ Todo el material de esta presentación está en Github.
 - ✿ <https://github.com/aoiymk/BigDataWeek>
- ✿ Búscame en  : Eli Carreño



Muchas gracias por la atención