PROYECTO FINAL - ID BOOTCAMPS DATA SCIENCE

# Juanito 2.0

Stefani Pereira
Benjamin Kutz

# INTRODUCCIÓN

# PRIMERA PRESENTACIÓN

- PRESENTACIÓN DE JUANITO (PROBLEMA DE MERCADO PÚBLICO LICITACIONES)
- DÓNDE ESTÁN LOS DATOS (EXPLICACIÓN OBTENCIÓN DATA)
- CÓMO RESOLVEMOS EL PROBLEMA (ARMAMOS UN MODELO QUE HACE…)
- OBTENEMOS LOS SIGUIENTES RESULTADOS

# PRIMERA PRESENTACIÓN

- PROBLEMA / DESAFÍO (MUCHA DIAPO EXPLICANDO EL PROBLEMA)
  - Marco/Escenario - Mercado Público
  - Problema - Identificar oportunidades para ofrecer productos
  - Acercamiento analítico - Data disponible (EDA)
- METODOLOGÍA EMPLEADA
  - Estrategia - Selección de variables categóricas y tratamiento variable descripción con técnicas de NLP
  - Preparación Data - transformación, desbalance clases, alta dimensionalidad
- MODELAMIENTO
  - Estrategia - iteración prueba/error modelos seleccionados
- EVALUACIÓN
  - Estrategia - evaluación resultado métricas y estabilidad del modelo

QUIÉN ES

# JUANITO

¡Hoy será un gran día, nada puede malir sal!
client metrica
ChileCompra
MercadoPúblico

Sé muy bien lo que hago, me sale casi automático.
client metrica

Son solo 999 descripciones al día lo que tengo que leer, eso no es nada para mí.
client metrica
MercadoPublico.cl
ChileCompra
Búsqueda de Licitaciones

# DESAFÍO A RESOLVER

¿Cómo puedo identificar oportunidades de negocio en el mercado público?

¿Se puede dar una solución con Machine Learning?

# ¡POR SUPUESTO!

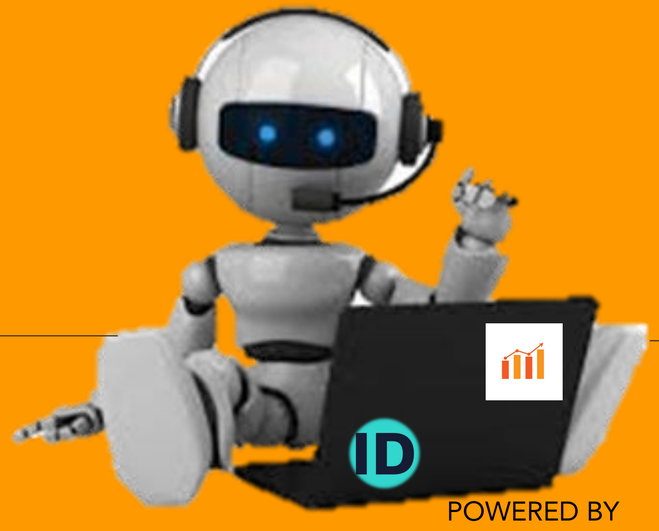
Data scientist

MODELO DE APRENDIZAJE SUPERVISADO

# JUANITO 2.0

| PRODUCTOS DE LA EMPRESA "Xs" |
| --- |
| LECHE |
| ARROZ |
| LECHUGA |
| POLLO |

| LICITACIONES PASADAS | |
| --- | --- |
| Productos Licitados | Resultado Empresa "Xs" |
| Leche | SI |
| Patata | NO |
| Leche | NO |
| Arroz | SI |




| PREDICCIÓN MODELO | |
| --- | --- |
| Productos Licitados | Resultado Predicción |
| Arroz | SI |
| Patata | NO |
| Leche | SI |
| Pollo | SI |

# ESQUEMA SENCILLO PROCESO, MARCO DE MERCADO PUBLICO Y ACTORES, LICITACION BENEFICIAROS CLIENTMETRICA JUANITO

TODO

# ¿A quién va dirigido el proyecto?

client metrica

# QUIENES SON LOS BENEFICIARIOS

- Clientmetrica
  - empresa chilena de tecnológica que provee productos y servicios de analytics a empresas medianas y pequeñas.

- Cliente Beneficiario (anónimo)
  - Empresa de materiales y productos de seguridad industrial
  - Participa en Mercado Público
  - Mix amplio de productos
  - Prospección manual de productos

13.000
SKU's

# MERCADO PÚBLICO

- ¿QUÉ ES mercadopublico.cl ?
  - Plataforma transaccional: gestiona las compras del estado.
  - Licitaciones Públicas: busca transparentar la compra de bienes y servicios del estado.

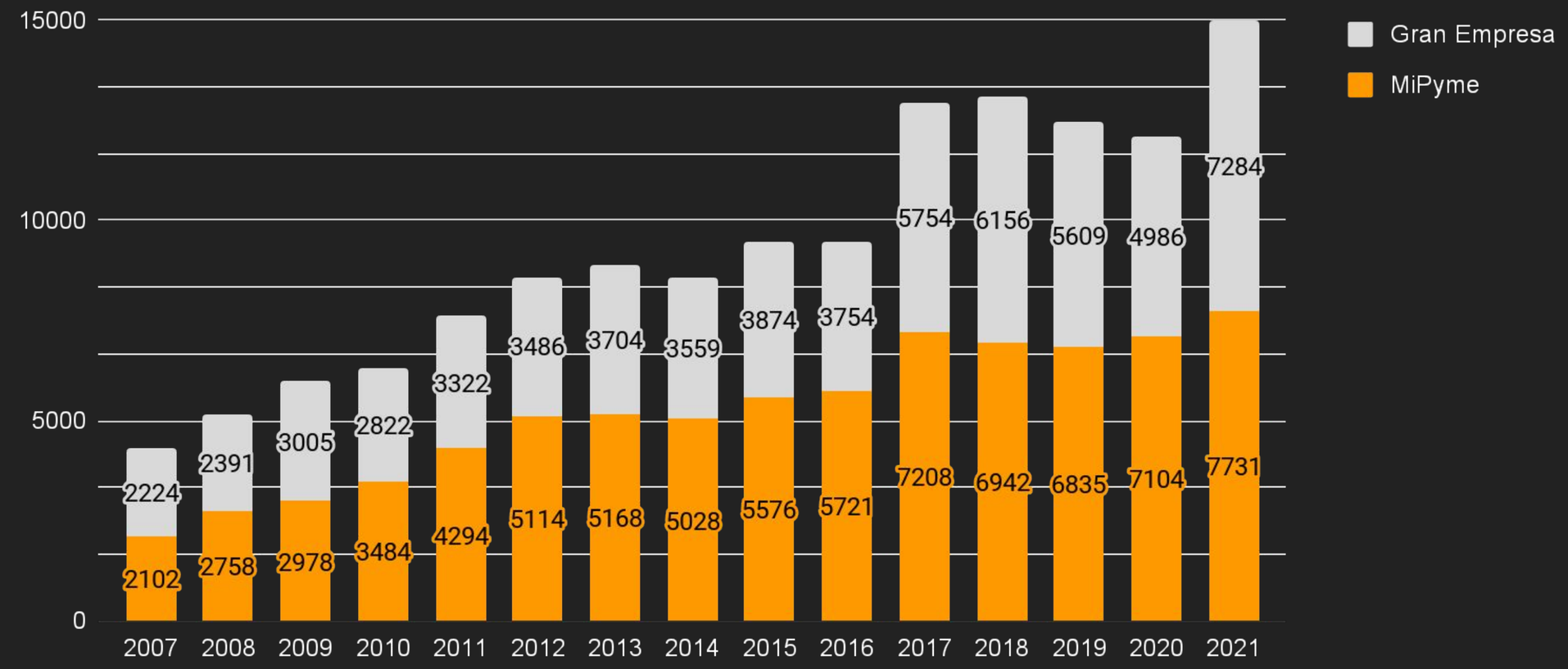
$15 MILLONES USD

GASTO TOTAL 2021

1,6 mm TOTAL TRANSACCIONES

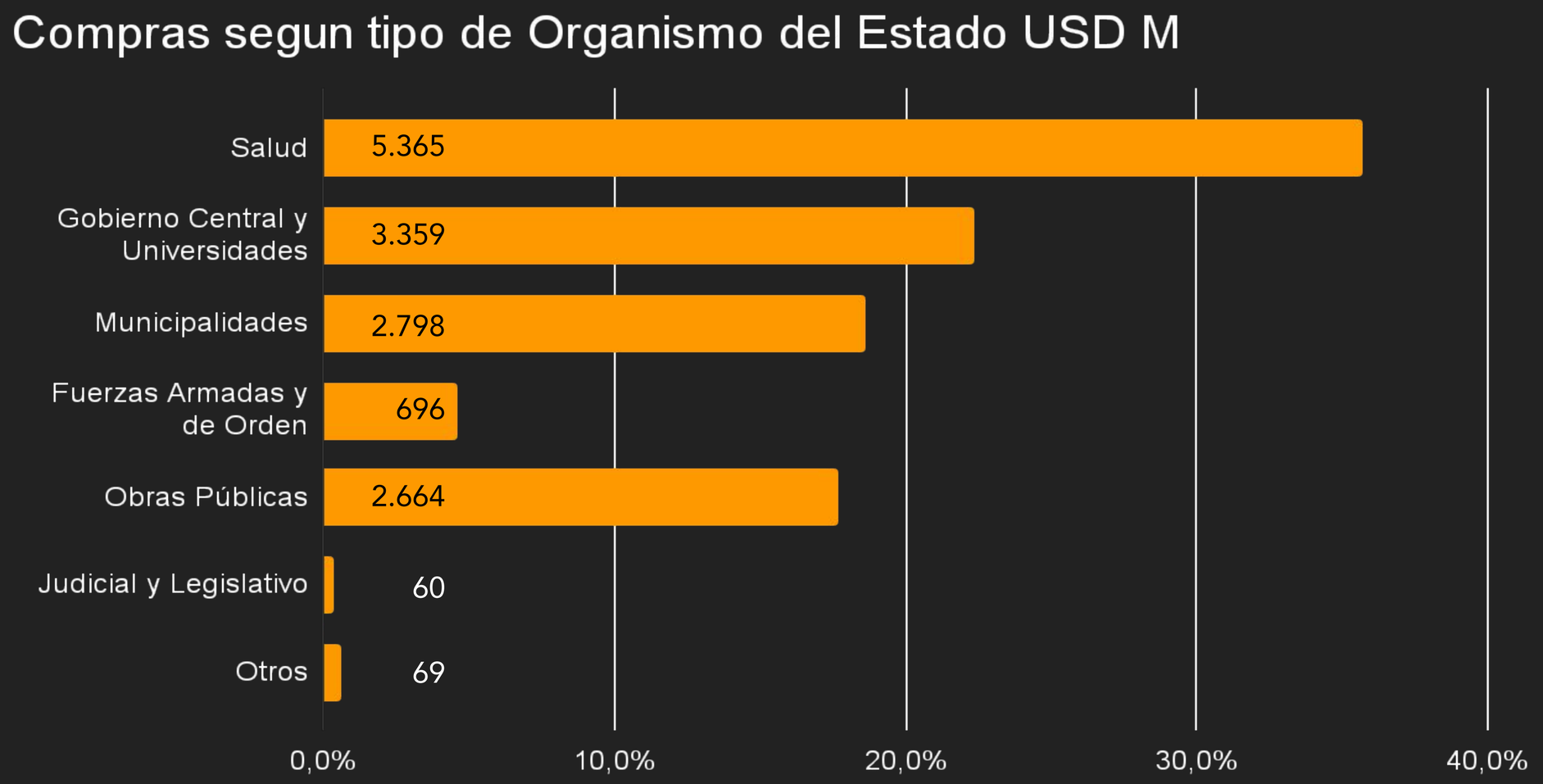
# MERCADO PÚBLICO - TAMAÑO MERCADO

- EVOLUCIÓN DE GASTO (USD Miles) TRANSADO POR AÑO Y TAMAÑO EMPRESAS

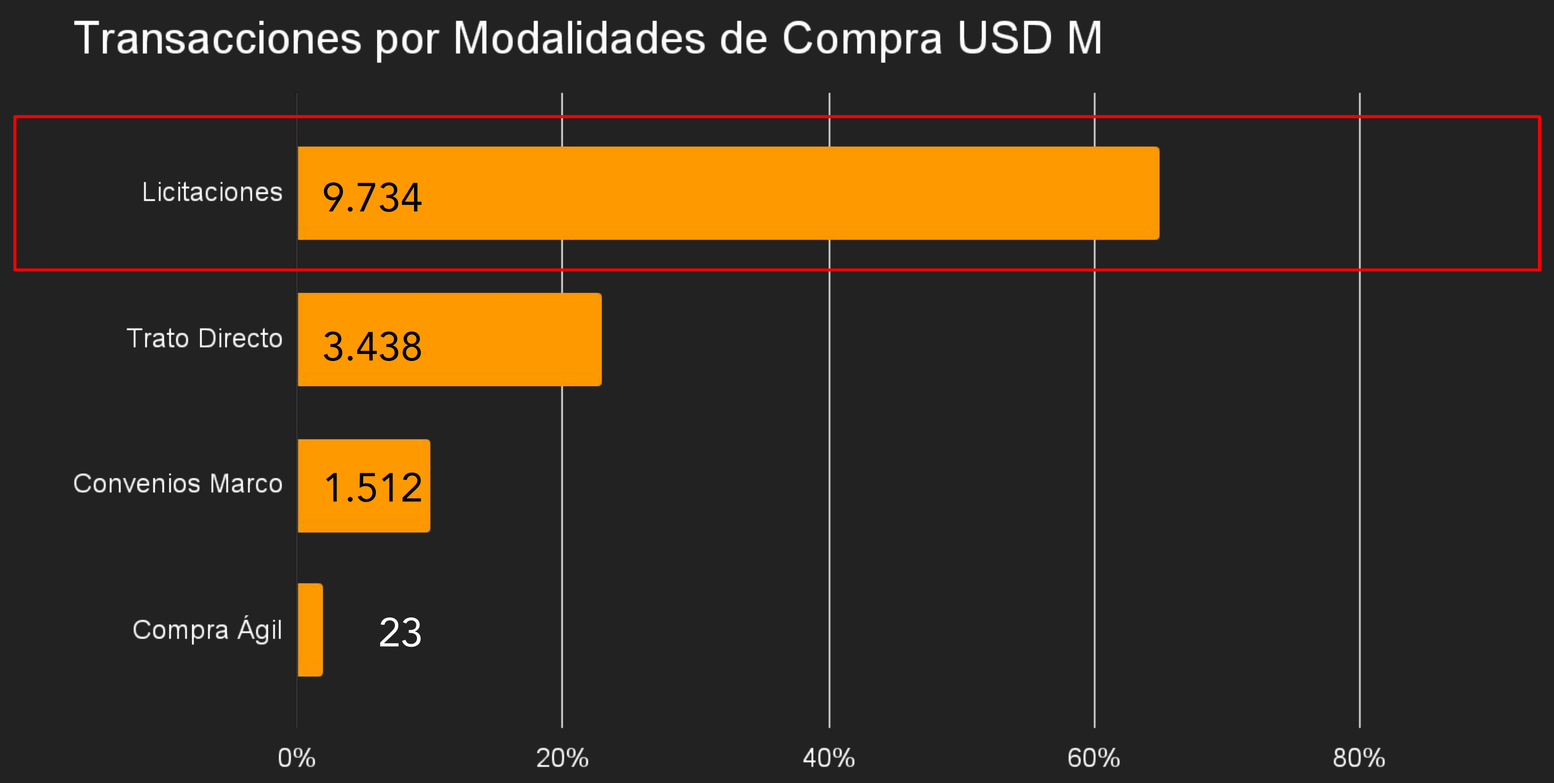


# MERCADO PÚBLICO - ¿QUIÉNES COMPRAN?

- 850 ORGANISMOS DEL ESTADO (15.000 compradores)



# MERCADO PÚBLICO - MECANISMOS COMPRAS



# ¿Cómo resolvimos el problema?

# ESTRATEGIA PARA ABORDAR EL DESAFÍO

- Definiciones
  - Brainstoming
  - Planificación de proyecto y definición de objetivos
  - Selección de metodología y herramientas

- Herramientas
  - Miro → Mapa mental, brainstorming, diseño de flujos
  - Trello → definición de tareas, coordinación y seguimiento de actividades
  - Github → Control de versionamiento
  - Lenguaje → Python
  - VSCode → Ide para progamación del código
  - Librerias → Pandas, NumPy, Sklearn, NLTK, XGBoost, Matplotlib, Regex

COMO RESOLVER EL PROBLEMA

# METODOLOGÍA DS

# CICLO DATA SCIENCE

# CICLO DATA SCIENCE

- Herramientas y Técnicas
  - APIs, CSV, SQL
  - Lenguaje Python
  - Limpieza data
  - Creación Variables
  - NLP
  - Bag of Words
  - Reducción de dimensionalidad (PCA)
  - Balance Muestreo (SMOTE)
  - GridSearch
  - Algoritmos de ML (Reg. Logística+XGBoost)

# DEFINICIÓN DEL PROBLEMA

- PROBLEMA
  - Dificultad para detectar licitaciones (1000 diarias) en las cuales se puedan ofrecer alguno de los productos (13mil SKU's) del cliente.
- DATOS
  - Historia de participación del cliente en licitaciones
  - Casos (licitaciones) marcados por el cliente si aplican o no participar
  - Catálogo de productos del cliente
- ESTRUCTURA
  - Supervisado (contamos con label de los casos)
  - Clasificación (licitación aplica o no)

*Buscador no permite filtrar por producto*

Filtrado

*Dificultad para leer campo descripción (texto libre)*

Reconocimiento

*Diariamente se publican cerca de 1000 licitaciones*

Alto volumen

# OBJETIVO

- Crear un modelo de machine learning que clasifique, dentro de un listado de licitaciones públicas (de Chile), aquellas donde una empresa o persona pueda (o no) participar con su catálogo de productos y/o servicios.

- Objetivos Específicos
  - Identificar fuentes de datos de licitaciones públicas de Chile
  - Identificar variables
  - Construir algoritmo de clasificación
  - Grado de confianza mínimo del modelo

# FUENTES DE DATOS

- Mercado Público
  - historia de todas las licitaciones publicadas
  - detalle de cada licitación con su resultado
  - detalla quienes participaron en la licitación → identificando quién ganó
  - detalle de los productos y servicios en cada licitación

- Clientmetrica
  - BBDD con marcas de productos en licitaciones que aplican o no

# MERCADO PÚBLICO

a) Portal transaccional web
[https://www.mercadopublico.cl/Home](https://www.mercadopublico.cl/Home)

b) Descarga de archivo planos
[https://datos-abiertos.chilecompra.cl/descargas/ordenes-y-licitaciones](https://datos-abiertos.chilecompra.cl/descargas/ordenes-y-licitaciones)

c) Mediante API API API API API API
[https://desarrolladores.mercadopublico.cl/docs/services/](https://desarrolladores.mercadopublico.cl/docs/services/)

**LICITACIONES_APINUEVA_TENDER (dbo)**

| |
|---|
| ocid |
| Periodo |
| lid |
| Lic_date |
| Tag |
| Lic_Name |
| Parties_id |
| Id_id |
| Id_legalname |
| Id_scheme |
| StreetAddress |
| Region |
| Country |
| Contactname |
| Email |
| Telephone |
| Roles |
| Tender_id |
| Tender_title |
| Tender_desc |
| Tender_status |
| Tender_value_am |
| Tender_value_cur |
| Tender_proc_name |
| Tender_proc_name_id |
| Tender_proc_id |
| Tender_procmet |
| Tender_procmetdet |
| Startdate |
| Enddate |
| Duration |
| Lic_Url |

**LICITACIONES_APIVIEJA_ITEMS (dbo)**

| |
|---|
| Item_Key |
| Tender_ID |
| Correlativo |
| Periodo |
| Cod_Producto |
| Cod_Categoria |
| Categoria |
| Nombre_Producto |
| Descripcion |
| Unidad_Medida |
| Cantidad |
| Adjudicacion |
| Rubros1 |
| Rubros2 |
| Rubros3 |

# CLIENTMETRICA

BBDD SQL

a) Data histórica de Licitaciones

b) Tabla con licitaciones en las cuales el cliente puede participar marcadas con SI o NO

*Dataset de entrenamiento se construye a partir de historia previa (participación del cliente) y marcas "manuales" del cliente.*

*IMPORTANTE: marcas manuales de licitaciones donde NO debe participar.*

# DATOS

Variables Independientes

Target!!

| | id_licitacion | id_producto | Rubro1 | Rubro2 | Rubro3 | nombre_producto | descripcion | label |
|---|---|---|---|---|---|---|---|---|
| 0 | 2792-88-LE21 | 2792-88-LE21-1 | Equipamiento y suministros médicos | Productos quirúrgicos | Equipo quirúrgico, accesorios y productos rela... | Equipo quirúrgico de fragmatome retinal vítreo... | CÁMARA RETINAL NO MIDRIÁTICA - ESPECIFICACIONE... | 0 |
| 1 | 1704-52-LP21 | 1704-52-LP21-1 | Servicios de transporte, almacenaje y correo | Transporte de correo y carga | Transporte de carga por carretera | Servicios de transporte local en camión | SERVICIO DE MUDANZA DE ENSERES PROPIOS Y EN US... | 0 |
| 2 | 1003-20-L121 | 1003-20-L121-1 | Productos derivados de minerales, plantas y an... | Chatarra y desechos no metálicos | Desechos no metálicos | Desperdicios o restos de madera | LEÑA SECA CERTIFICADA | 0 |
| 3 | 903556-3-LP21 | 903556-3-LP21-1 | Tecnologías de la información, telecomunicacio... | Software | Software de administración de redes | Software de administración | NaN | 0 |
| 4 | 4236-21-L121 | 4236-21-L121-1 | Equipos, accesorios y suministros de oficina | Suministros de oficina | Instrumentos de escritura | Marcadores | MARCADOR BLANCO PARA METALES | 0 |

Todas Categóricas!! → nombre producto (sí también es categórica)

la más importante … que contiene el nombre verdadero del producto
→ Descripción (texto libre 😭)

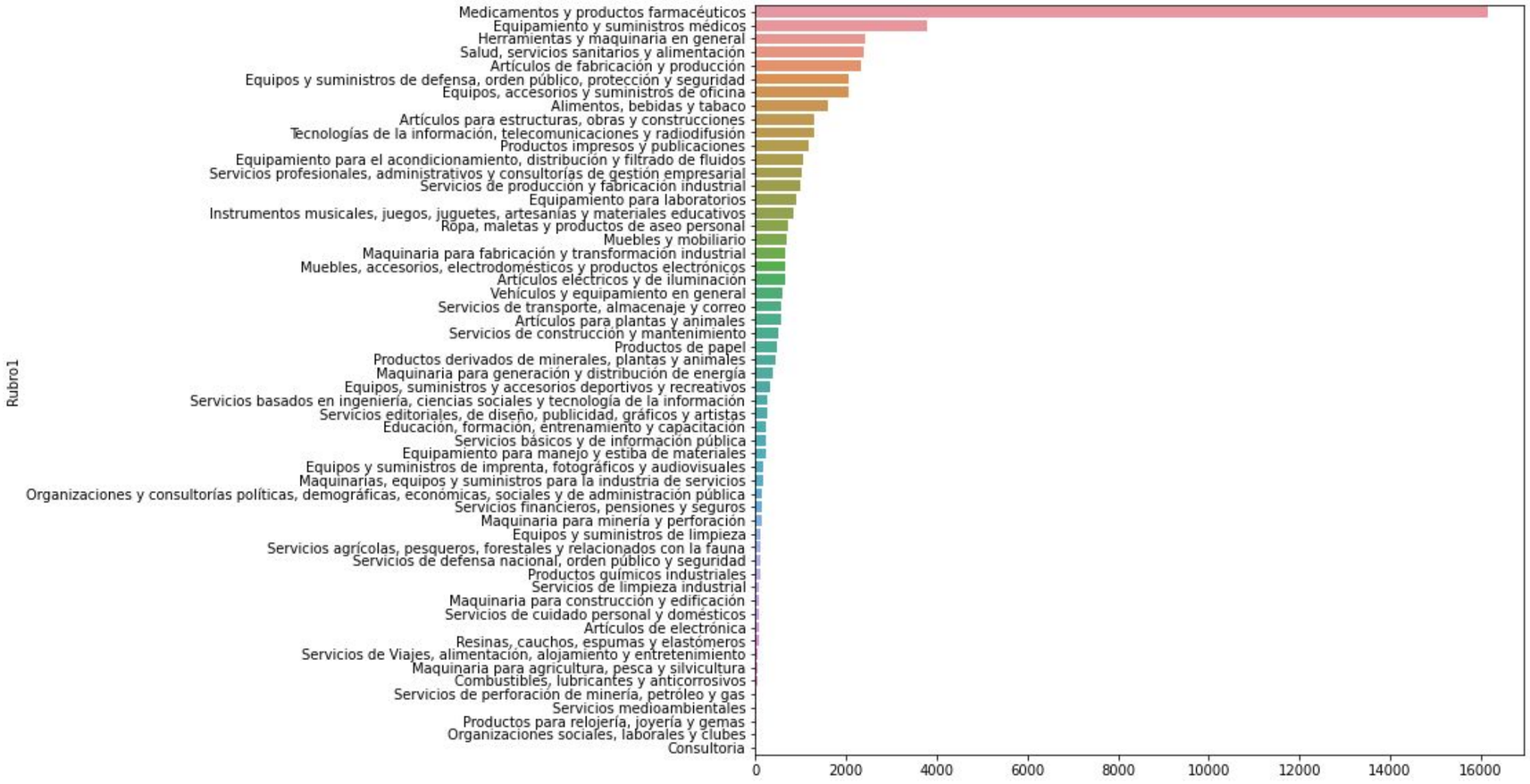
# ESTRATEGIA - EDA

- Rubros
- Mirada “completa” Mercado Público
- Foco en ventas de cliente
- NLP sobre campo descriptivo
- NLP catálogo productos cliente
- Balance de label

# EDA - Mercado Público

```
Total categorías Rubros 1: 56
Total categorías Rubros 2: 307
Total categorías Rubros 3: 1284
Total categorías Productos: 4740
```

- El Rubro de Salud es la categoría con mayor cantidad de registros (30%)
- Muestra (y su distribución) es representativa del universo de licitaciones.



# EDA - Mercado Público

## Campo descripción

- Aún cuando el Rubro de Salud es el de mayor peso, existen palabras "genéricas" con mayor frecuencia, como servicio.

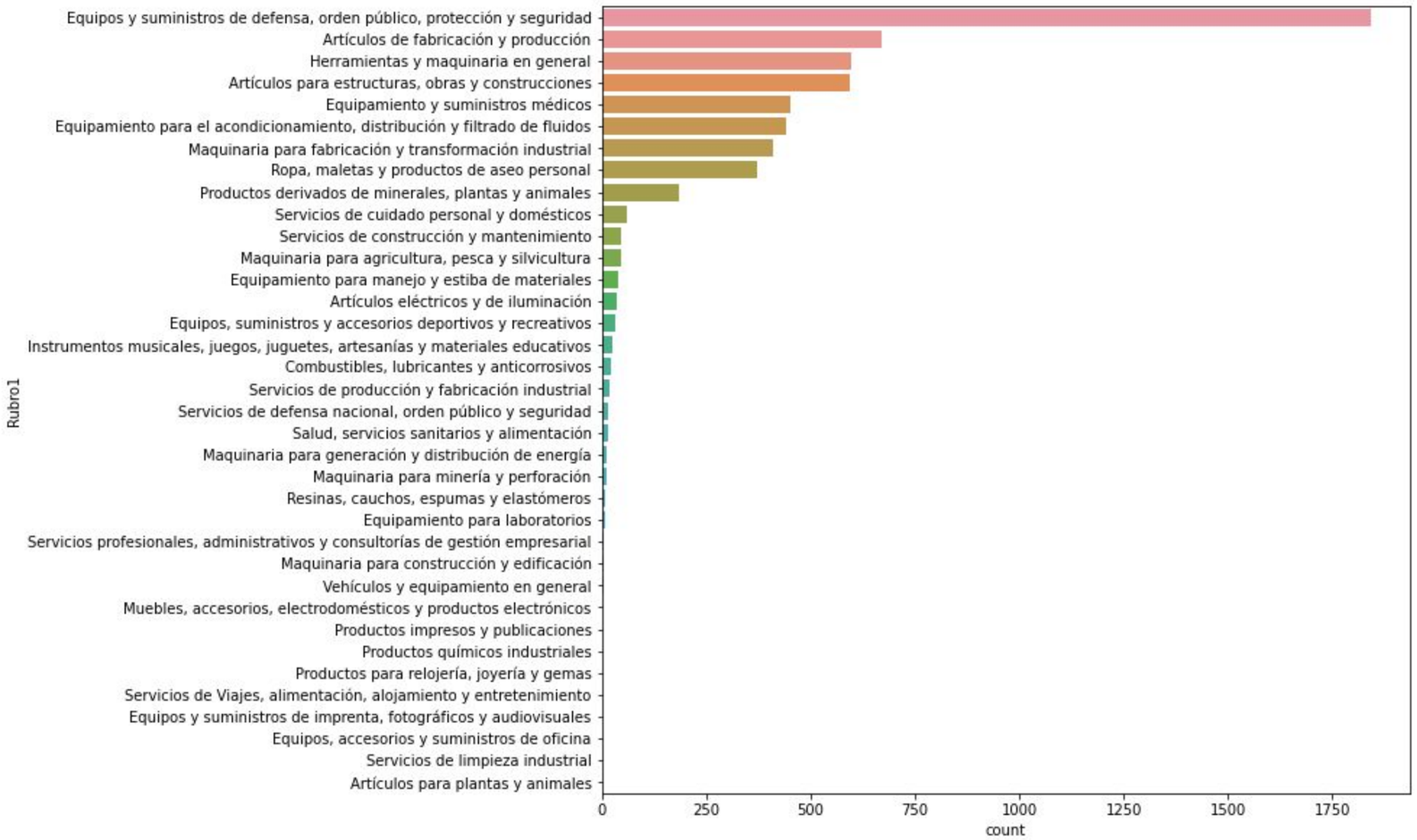
- Al transformar las palabras a su raíz (stemming) se mantienen las principales palabras pero surgen algunas palabras nuevas como quirurg, medic.

servici
accesori
clorhidrat
equip
medic
sodic
papel
domest
cabl
cint
soport
sistem
quirurg
kits
segur
dental
laboratori
herramient
unidad
filtr
electr
product
aliment
conjunt
pintur
vehicul
analiz
agu
congel
sulfat
manten
public
nader
acetat
montaj
aceit
mes
acer
disc
metal
motor
general
control
bol
can
piez
llav
soldadur
tub
electron
protector
material
suministr
tej
maquin
alambr
sets
pacient
proteccion
libr
cuchill
comercial
semill
almacig
disposit
contenedor
tornill
hidroclorur
limpiez
maquinari
puert
digital
aparat
personal
plastic
barr
calent
instrument
valvul
transport
air
caden
tuberi
planch
bomb
caj
tarjet
jueg
impresor
punt
cuerd
aguj
acid
aceit
sill
automovil
baterí
comput
uniform
aliment
producto
filtr
acetat
nader
public
manten
sulfat
analiz
agu
congel
conjunt
pintur
vehicul
unidad
montaj
digital
aparat

# EDA - Cliente

## Categorias

- Los registros de licitaciones se condicen con el sector Ventas centrado en productos industriales de seguridad y construcción.

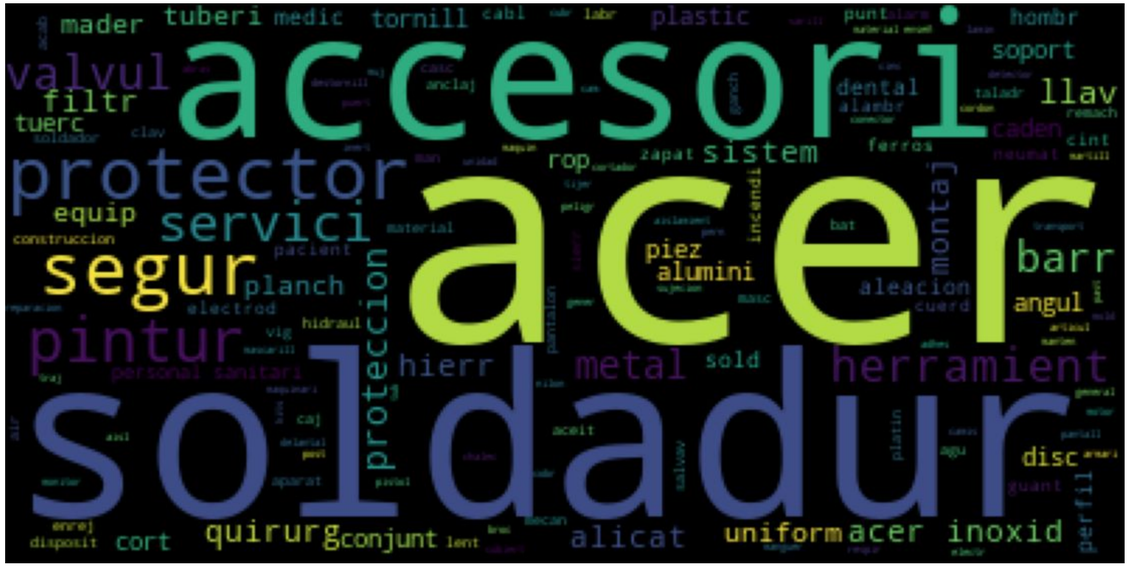


# EDA - Cliente

Campo descripción “limpio”

- Los productos que más se repiten en licitaciones en las cuales ha participado el cliente tienen relación con soldaduras, pintura, barra, acero, válvula.

- Las palabras tienen relación con el catálogo de productos del cliente



# EDA - Cliente

Catálogo de Productos (SKU's) "limpio"

- Los productos que más se repiten en el catálogo de productos tienen relación con herramientas, materiales y vestimenta de seguridad.

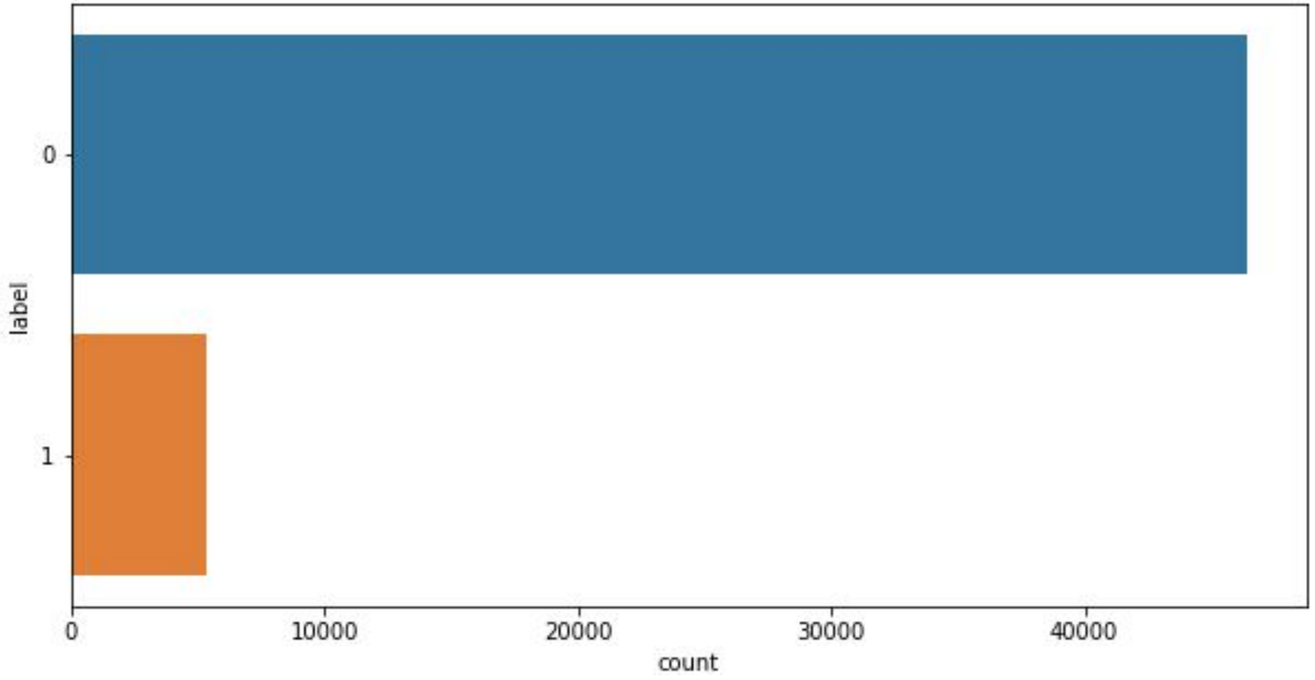
- La frecuencia tiene relación a variaciones del producto (ej. medidas, tallas)

- El sr. Parker es relevante en la ideación de productos industriales de seguridad 😂

# EDA - Balance label

- Existe un desbalance de la variable target respecto a sus marcas en las licitaciones
  - 89,6% → NO
  - 10,4% → SI

- Es de esperar dada la realidad de cómo se distribuyen las licitaciones (mayor peso área salud)



# LIMPIEZA DE DATOS

- Cambio de nombre de columnas.
- Reemplazo de index por ID de productos licitados.
- Eliminación de NaNs.
- Retiro de registros duplicados.

# ESTRATEGIA - CONSTRUCCIÓN VARIABLES

- Categóricas → Dummies y selección de aquellas más relevantes
- Texto libre → Bolsa de Palabras normalizada

# CREACIÓN DE VARIABLES - CATEGÓRICAS

- Dummies
  - total valores únicos entre Rubro 1, 2, 3, nom_producto = 6366

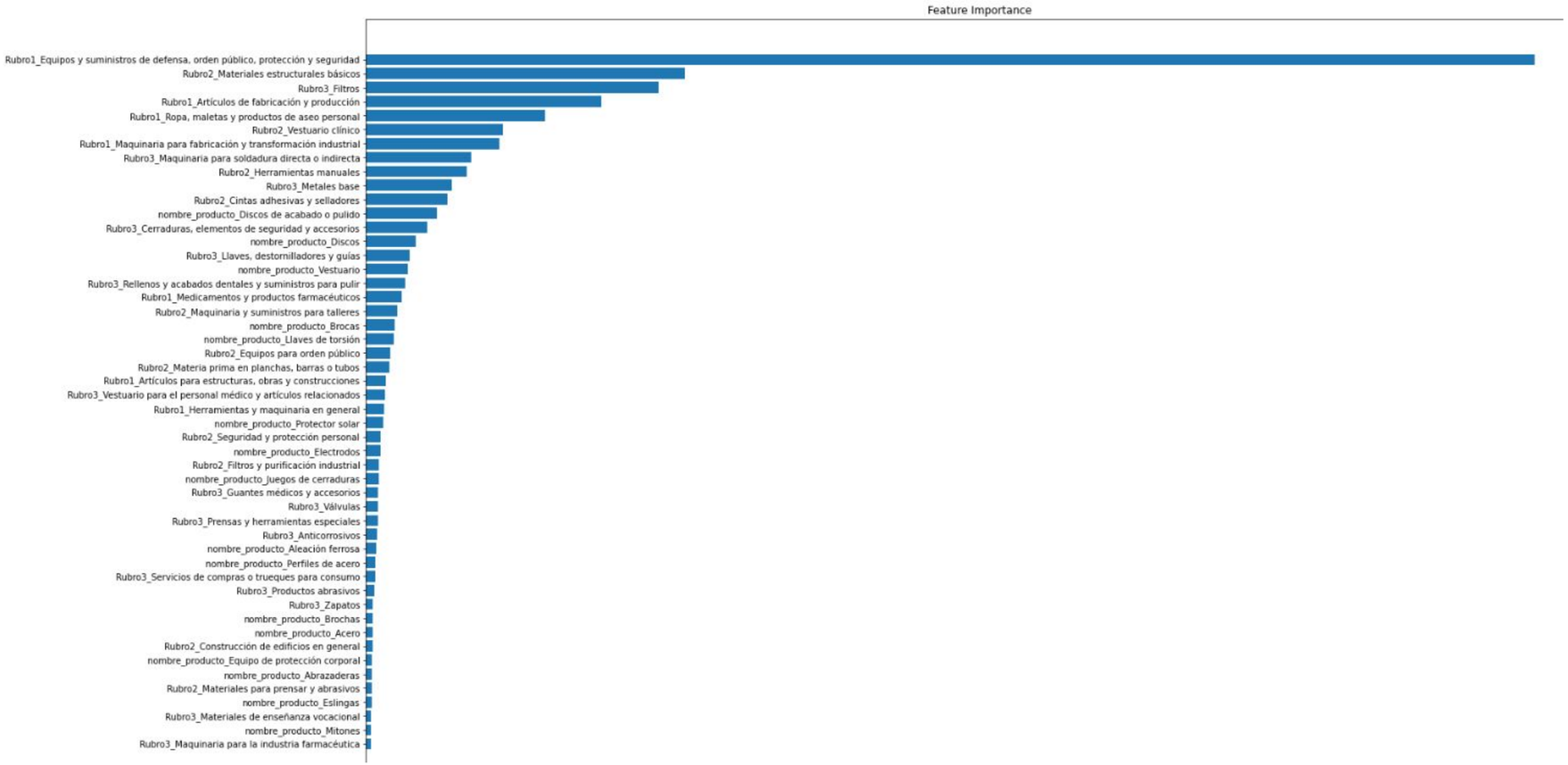
| | label | Rubro1_Artículos de electrónica | Rubro1_Artículos de fabricación y producción | Rubro1_Artículos eléctricos y de iluminación | Rubro1_Artículos para estructuras, obras y construcciones | Rubro1_Artículos para plantas y animales | Rubro1_Combustibles, lubricantes y anticorrosivos | Rubro1_Consultoria | Rubro1_Educación, formación, entrenamiento y capacitación | Rubro1_Equipamiento para el acondicionamiento, distribución y filtrado de fluidos | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| id_producto | | | | | | | | | | | |
| 2792-88-LE21-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 1704-52-LP21-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 1003-20-L121-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 4236-21-L121-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 4236-21-L121-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1058141-35-LQ21-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 1058141-35-LQ21-4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 1173455-12-LE21-41 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 1057461-64-LE21-17 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 4309-381-LQ21-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |

54484 rows × 6366 columns

# CREACIÓN DE VARIABLES - CATEGÓRICAS

- Catboost para selección de dummies
- Criterio de selección:

  importancia >= 0.000000000001
- Resultado: reducción a 113 variables categóricas (*Rubro 1 tiene mayor peso*)



# CREACIÓN DE VARIABLES - DESCRIPCIÓN

- Procesamiento de texto (Regex+NLP)

  *se crean funciones para facilitar transformación*

  - mínusculas
  - sacar tíldes
  - caracteres especiales
  - sacar stopwords

```
CÁMARA RETINAL NO MIDRIÁTICA - ESPECIFICACIONE...
SERVICIO DE MUDANZA DE ENSERES PROPIOS Y EN US...
LEÑA SECA CERTIFICADA
MARCADOR BLANCO PARA METALES
PLOMADA
...
ALTEPLASA FA 50 MG CAJA X 2 FA [216P008]
TENECTEPLASA FA 10.000 UI (50 MG) [216P185]
manguera transparente para nivel (mts)
CINTA ADHESIVA 18 MM X 30 MT
10-100-214-208-01 Olanzapina tableta liofiliza...
```

Regex + NLP

servici
accesori
clorhidrat
equip
domest
papel
cabl
soport
cint
medic
sodic
quirurg
kits
sistem
segur
dental
laboratori
herramient
electr
filtr
product
aliment
unidad
conjunt
analiz
agu
manten
sulfat
congel
maquin
alambr
protector
control
general
motor
piez
soldadur
software
electron
disposit
comercial
almacig
semill
proteccion
cuchill
libr
pacient
sets
metal
tornill
hidroclorur
personal
plastic
transport
instrument
valvul
digital
aparat
calent
barr
tub
llav
suministr
material
mader
public
acid
tuberi
planch
bomb
caj
tarjet
jueg
impresor
vehicul
automovil
montaj
aceit
acetat
limpiez
maquinari
air
caden
comprimido
recubierto
caja comprimido
similar
chile certificado
bases tecnica
publico conveniente
practicas fabricacion
cumplimiento buena
fabricado extranjero
fabricado chile
fabricado paise
tipo
bases
segun bases
extranjero certificado
certificado acredite
acero inoxidable
servicio
caja
fecha vencimiento
eficacia certificado
plastico
desechable
adquisicion
sistema
acredite fabricado
agencias especificaciones
agencias agencia
vigilancia sanitaria
material
certificado
grande
aluminio
liberacion prolongada
aluminio anodizado
aluminio
aluminio anodizado
aluminio anodizado
aluminio anodizado

# CREACIÓN DE VARIABLES - DESCRIPCIÓN

*luego de transformar el texto descriptivo se convierte a un array de frecuencia de palabras y se normaliza*

- Bag of Words + TF-IDF
  - 19.535 palabras/variables

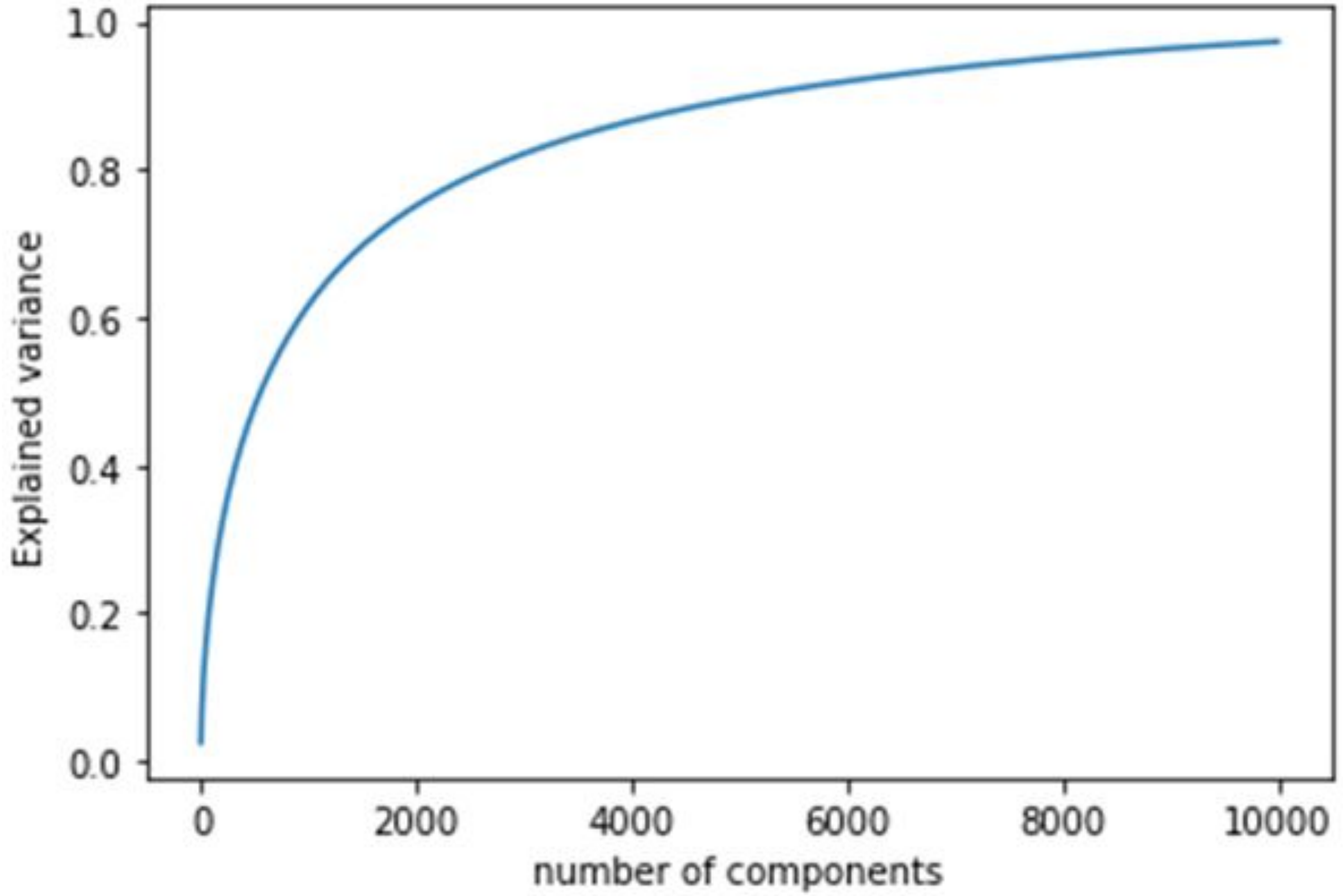
| | aac | aacc | aaguj | aalt | aasht | aatic | aavv | abacav | abadiaisbn | abaj | ... | zurd | zurn | zuñig | zymar | ñielol | ñilhu | ñip | ñiqu | ñubl | ñuño |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

5 rows × 19535 columns

# CREACIÓN DE VARIABLES - DESCRIPCIÓN

*se decide reducir la dimensionalidad de la bolsa de palabras*

- Reducción de dimensiones utilizando PCA
  - 1.000 componentes → 0,6
  - 2.000 componentes → 0,75
  - 5.000 componentes → 0,85
  - 7.000 componentes → 0,9
  - 10.000 componentes → 0,98



# BALANCE DE MUESTRAS

- SMOTE
  - Muestra: “0” 90% | “1” 10%
  - Estrategia de oversampling → 30%

# RESULTADO DATASET ENTRENAMIENTO

- 54.484 registros x 10.113 variables independientes

| | Rubro1_Equipos y suministros de defensa, orden público, protección y seguridad | Rubro2_Materiales estructurales básicos | Rubro1_Artículos de fabricación y producción | Rubro3_Filtros | Rubro1_Ropa, maletas y productos de aseo personal | Rubro2_Vestuario clínico | Rubro1_Maquinaria para fabricación y transformación industrial | Rubro3_Maquinaria para soldadura directa o indirecta | Rubro2_Herramientas manuales | Rubro2_Cintas adhesivas y selladores | ... | PC9990 | PC9991 | PC9992 | PC9993 | PC9994 | PC9995 | PC9996 | PC9997 | PC9998 | PC9999 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0.000475 | -0.000919 | -0.000795 | 0.001668 | -0.000468 | 0.001385 | -0.000188 | -0.002308 | 0.000297 | -9.016536e-05 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0.000123 | 0.000021 | -0.000059 | 0.000030 | 0.000181 | 0.000030 | -0.000114 | 0.000006 | 0.000028 | 1.183444e-07 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0.004396 | -0.000007 | -0.001900 | 0.002460 | 0.000395 | -0.000117 | -0.001567 | 0.000693 | 0.001213 | -1.094438e-03 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0.000003 | -0.000030 | -0.000008 | -0.000104 | 0.000057 | 0.000067 | 0.000033 | 0.000006 | 0.000095 | 7.119809e-05 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | -0.001304 | -0.000111 | -0.001878 | -0.000201 | 0.000342 | 0.000025 | -0.001541 | -0.000360 | -0.001103 | -4.511249e-04 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 42268 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | -0.000103 | 0.000095 | 0.000421 | -0.000187 | -0.000714 | -0.000092 | -0.000280 | -0.000716 | -0.000345 | 5.312532e-04 |
| 42269 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | -0.000088 | 0.000400 | 0.000068 | -0.000095 | 0.000527 | 0.000331 | -0.000336 | 0.000179 | -0.000003 | -8.556863e-04 |
| 42270 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0.001715 | -0.003220 | -0.000318 | 0.002456 | 0.000548 | -0.000667 | -0.003014 | 0.001004 | -0.000756 | -1.332532e-03 |
| 42271 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0.000534 | -0.001018 | 0.000708 | 0.000026 | -0.000501 | 0.000690 | -0.001193 | 0.000073 | 0.002796 | -9.104459e-04 |
| 42272 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | -0.000004 | 0.000148 | 0.000160 | 0.000030 | -0.000023 | 0.000070 | 0.000008 | -0.000238 | -0.000070 | -2.913763e-04 |

42273 rows × 10113 columns

IMPACTO Y RESULTADO

# MODELADO

# ENTRENAMIENTO MODELOS

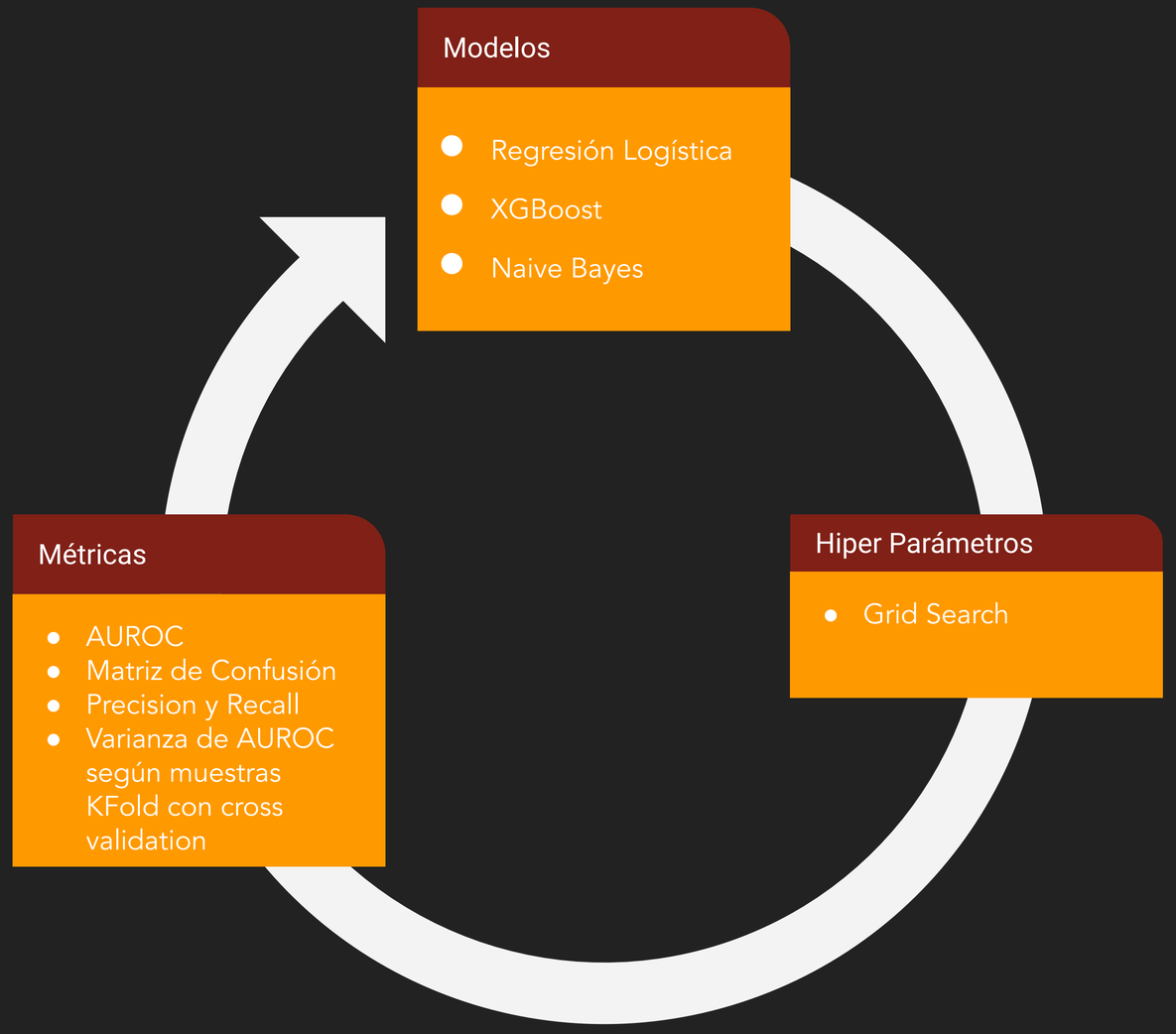
- ESTRATEGIA
  - Modelos seleccionados - Regresión Logística, Naive Bayes, XGBoost
  - Métricas - AUROC, Matriz de Confusión, Precisión, Sensibilidad
  - Hiper Parámetros - Gridsearch
  - Representatividad variable y en cada muestreo - CV (KFold)

# ENTRENAMIENTO MODELOS

- VARIABLE A PREDECIR

# ENTRENAMIENTO MODELOS

- ESTRATEGIA



# ENTRENAMIENTO MODELOS

- MÉTRICAS

- Acuracy
- AUROC

# ¿Cuál fue el resultado?

# EVALUACIÓN MODELOS

- ESTRATEGIA EMPLEADA
  - Evaluación resultado métricas en Train y Test - Matriz de Confusión, AUROC, Precisión, Recall (sensibilidad)
  - Estabilidad del modelo - varianza media de cada Fold utilizado en el Cross Validation
  - Análisis de los errores FN, FP
- JUSTIFICACIÓN SELECCIÓN DEL MODELO
  - Modelos seleccionados

# Iteración 1

- Sólo variables categóricas (sin campo descripción)

| Modelo | ROC | Accuracy | Precision 0 | Precision 1 | Recall 0 | Recall 1 | F1-score 0 | F1-score 1 |
|---|---|---|---|---|---|---|---|---|
| Regresión Logística | 0,87 | 0,89 | 0,89 | 0.87 | 1 | 0.10 | 0.94 | 0,17 |

# Iteración 2

- Creación variable campo descripción como bolsa de palabras reducido con PCA + 1 variable categórica (Rubro1)

| Modelo | ROC | Accuracy | Precision 0 | Precision 1 | Recall 0 | Recall 1 | F1-score 0 | F1-score 1 |
|---|---|---|---|---|---|---|---|---|
| Regresión Logística | 0,97 | 0,95 | 0,96 | 0.84 | 0.98 | 0.67 | 0.97 | 0,74 |
| XGBoost | 0,96 | 0,95 | 0,95 | 0.87 | 0.99 | 0.57 | 0.97 | 0,69 |
| Naive Bayes | 0,83 | 0,70 | 1 | 0.26 | 0.67 | 0.99 | 0.80 | 0,41 |

# Iteración 3

- Idem iteración 2 + creación de variables categóricas seleccionadas con Catboost

| Modelo | ROC | Accuracy | Precision 0 | Precision 1 | Recall 0 | Recall 1 | F1-score 0 | F1-score 1 | STD AUROC CV= 3 |
|---|---|---|---|---|---|---|---|---|---|
| Regresión Logística | 0,988 | 0,95 | 0,95 | 0.93 | 0.98 | 0.83 | 0.97 | 0,88 | 0.0011 |
| XGBoost | 0,96 | 0,90 | 0,90 | 0.92 | 0.98 | 0.62 | 0.94 | 0,74 | 0.0048 |

# Iteración 4

- Idem iteración 3 pero se corrige desbalance de la muestra utilizando SMOTE

| Modelo | ROC | Accuracy | Precision 0 | Precision 1 | Recall 0 | Recall 1 | F1-score 0 | F1-score 1 | STD AUROC CV= 3 |
|---|---|---|---|---|---|---|---|---|---|
| Regresión Logística | 0,98 | 0,96 | 0.98 | 0.82 | 0.98 | 0.86 | 0.98 | 0.84 | 0.0021 |
| XGBoost | 0.98 | 0.97 | 0.98 | 0.86 | 0.98 | 0.82 | 0.98 | 0.84 | 0.0051 |

# ANÁLISIS DE ERRORES

- FALSOS POSITIVOS para categorías específicas

| label_y | y_pred | id_licitacion | Rubro1 | Rubro2 | Rubro3 | nombre_producto | descripcion |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 4236-25-L121 | Artículos de fabricación y producción | Pinturas, diluyentes y accesorios para pintado | Pinturas | Pinturas acrílicas | PINTURA ACRILICA ALTO TRAFICO AMARILLA |
| 0 | 1 | 4236-26-L121 | Artículos de fabricación y producción | Pinturas, diluyentes y accesorios para pintado | Pinturas | Pinturas acrílicas | PINTURA ACRILICA ALTO TRAFICO AMARILLA |

# ANÁLISIS DE ERRORES

- FALSOS NEGATIVOS para rubro con más licitaciones

| label_y | y_pred | id_licitacion | Rubro1 | Rubro2 | Rubro3 | nombre_producto | descripcion |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1013609-24-LE21 | Ropa, maletas y productos de aseo personal | Ropa | Uniformes | Delantal de doctor | delantal blanco talla M |

RESULTADOS

# CONCLUSIONES

# MODELOS

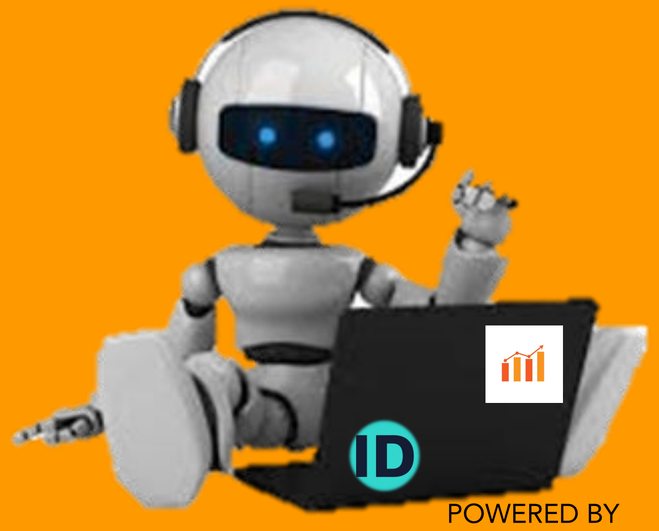
- Modelo Seleccionado(Iteración 4)

| Modelo | ROC | Recall 1 |
|---|---|---|
| Regresión Logística | 0,98 | 0.86 |

| HIPERPARÁMETROS | |
|---|---|
| solver | ‘lbfgs’ |
| c | 1.00 |
| max_iter | 500 |
| multi_class | ‘OVR’ |
| Penalty | 12 |

| params | split0_test_roc_auc | split1_test_roc_auc | split2_test_roc_auc | mean_test_roc_auc | std_test_roc_auc | rank_test_roc_auc |
|---|---|---|---|---|---|---|
| {'C': 1.0, 'max_iter': 500, 'multi_class': 'ov... | 0.981006 | 0.984968 | 0.985702 | 0.983892 | 0.002062 | 1 |

# ¿Cuáles son entonces los beneficios esperados de Juanito 2.0?



# VENTAJAS Y DESVENTAJAS DE JUANITO 2.0

## VENTAJAS

- Clasifica correctamente
- Rapidez
- Pocos errores

## DESVENTAJAS

- Falsas predicciones
- No es genérico
- Utiliza muchas variables
- Tiempo de entrenamiento

BÚSQUEDA
AUTOMATIZADA

- TIEMPO
- ERRORES
+ AUTOMÁTICO

---

= - COSTOS
+ OPORTUNIDADES

# CONCLUSIONES

- Es posible automatizar el proceso de búsqueda de oportunidades de negocio en licitaciones públicas mediante el uso de técnicas de machine learning.
- El campo de descripción de la licitación (texto libre) es el más relevante al momento de clasificar de forma correcta una licitación como oportunidad de negocio.
- Juanito resuelve la problemática de la búsqueda manual de un ser humano al reducir los tiempos y costos de búsqueda así como aumentar las oportunidades de negocios.
- Juanito es capaz de detectar oportunidades donde el ser humano no es capaz de encontrar (debido a restricciones propias del ser humano).

# QUÉ APRENDIMOS

- Como trabajar en equipo desarrollando un proyecto de data science
- Aplicar las distintas técnicas y herramientas que aprendimos durante el bootcamp
- Dependiendo de la problemática hay modelos que generan mejores resultados que otros
- Aprendimos a evaluar los errores de los modelos
- Pusimos en práctica y conocimos la utilidad de distintas herramientas como Github, VScode, Trello, Mural/Miro, distintas librerias en Python

# TRABAJO FUTURO

- Buscar nuevas variables para hacer más genérica la clasificación del modelo.
- Crear un diccionario para “normalizar” las palabras.
- Incluir el catálogo de productos de la empresa.
- Modelo de recomendación para sugerencia de productos y precio a licitar.
- Modelo de tasa de éxito para poder ganar la licitación.

PROYECTO FINAL - ID BOOTCAMPS DATA SCIENCE

# Juanito 2.0

Benjamin Kutz
Stefani Pereira

“Ley de Alzheimer de la programación: si lees un código que escribiste hace más de dos semanas, es como si lo vieras por primera vez”.

*Via Dan Hurvitz*