# Análisis y estudio de Denuncias Registradas en Defensa al consumidor

Noemi Chuquichambi
Leg: 124644-6
nchuquichambi@gmail.com

Mireya Mamani
Leg: 127949-0
mireyamamani@gmail.com

## 1 .INTRODUCIÓN

La ciencia de los datos, es el proceso de adquirir conocimientos y conocimientos de los datos, es estructurada y no estructurada, utilizando métodos científicos, procesos, algoritmos y sistemas científicos para hacerlo en este campo cada vez más creciente.

El análisis aquí presentado tiene como intención explicar el comportamiento utilizando Machine Learning y la existencia de relaciones entre las variables de análisis.

Se realizó la búsqueda de un Data Set que contenga samples y features que permitan aplicar alguno de los modelos explicados a lo largo de la catedra.

Dimos con un data set de la Secretaria de modernización de la Presidencia de la Nación que cumplía con los requisitos y además consideramos que el análisis de las denuncias puede tener alguna utilidad práctica para algunos sectores.

## 2 .DATASETS

Secretaria de Modernización- Presidencia de la Nación

El archivo utilizado en el proyecto es una tabla donde se registran las denuncias registradas, fecha del mismo, el rubro, motivo, género y edad del reclamante.

https://datos.gob.ar/dataset/produccion-defensa-consumidor

## 3 .OBJETIVO

En base al tipo de reclamo y demás condiciones dadas predecir si los próximos reclamos serán estimados y tenidos en cuenta o de lo contrario serán desestimados.

## 4 .DATA SCIENCE WORKFLOW

El proceso de Ciencia de Datos consiste en los siguientes pasos:

**Objetivo:** En el caso de nuestro proyecto es el de poder predecir según diferentes variables si un reclamo será o no tenido en cuenta.

**Dataset:** Búsqueda según los tipos de features y cantidad de samples

**EDA:** Exploración de los datos contenidos en el dataset

**Modelización:** Aprendizaje supervisado

**Conclusiones**

## 5 .FEATURES

**Fecha**: Fecha del reclamo.

**Razón social**: Razón social de la empresa reclamada

**Ingreso por**: Mecanismo de ingreso del reclamo a la Dirección.

**Derivado a**: Derivación del reclamo una vez ingresado a la Dirección

**Categoría**: Categoría de la empresa reclamada, clasificada por la Dirección

**Rubro**: Rubro de la empresa reclamada, clasificado por la Dirección

**Subrubro**: Subrubro de la empresa reclamada, clasificado por la Dirección

**Modalidad**: Modalidad de realización del reclamo

**Motivo**: Clasificación asignada por la Dirección al motivo del reclamo ingresado

**Especificación motivo**: Detalle de la clasificación en el campo motivo

**Estado**: Estado del reclamo dentro de la Dirección. Sólo se indica si fue rechazado

**Motivo rechazo**: Motivo asignado por la Dirección en caso de ser rechazado el reclamo

**Región reclamante**: Región de la persona que realizó el reclamo

**Id provincia:** Número de identificación de la provincia de la persona que realizó el reclamo

**Provincia reclamante:** Provincia de la persona que realizó el reclamo

**Genero reclamante:** Género de la persona que realizó el reclamo.

**Edad reclamante:** Edad de la persona que realizó el reclamo

## 6 .EDA

Limpieza del data set y acondicionamiento de la tabla.

Verificamos que no haya filas con valores null. Resultado: Ninguna sample con valores nulos en variables relevantes.

Verificamos que no haya filas con valores nan. Resultado: Ninguna sample con valores nan en variables relevantes.

Análisis del tipo de columna. Para poder realizar el análisis por fecha tuvimos que modificar la columna “fecha” a formato de fecha ya que estaba configurada como string.

Verificación de valores numéricos fuera de lugar. Inspeccionamos viendo el mínimo y el máximo de cada columna observando si encontramos valores fuera de lugar. Esto sirve para ver, por ejemplo, si en las columnas numéricas tenemos valores negativos o extremadamente altos que nos darían un indicio de que hay algún problema con la data set.

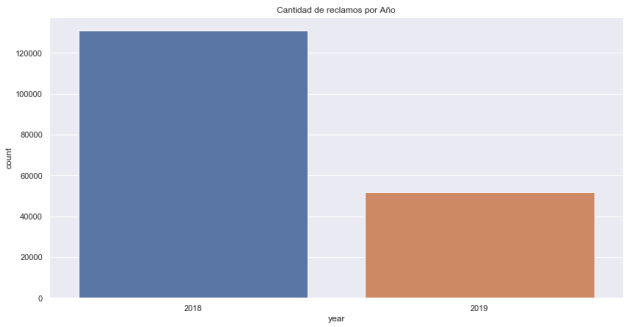
Se puede observar que existen valores de edad_reclamante igual a 0 o 118 o 119, valores que fueron desestimados.

## 7 .ANÁLISIS EXPLORATORIO DE DATOS

### 7.1 Reclamos por año

Como nos interesa saber la cantidad de reclamos por año procedemos a realizar las siguientes acciones:

- Creamos la columna 'year' con el año indicado en la columna 'fecha'
- Contamos frecuencia por año y graficamos los resultados




Los datos de nuestro data set van de marzo de 2018 a junio de 2019. Podemos observar con este análisis que en el 2018 se registran mayor cantidad de datos

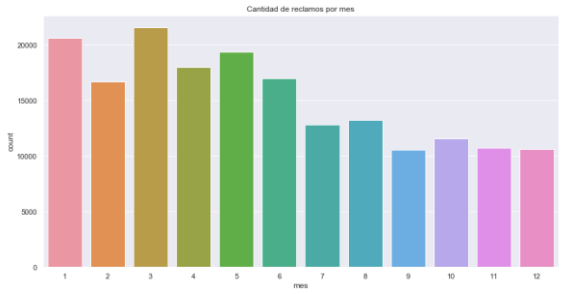
En el caso de 2019 la cantidad se ve reducida por tener datos hasta junio 2019.

### 7.2 Reclamos por mes

Como nos interesa saber la cantidad de reclamos por mes procedemos a realizar las siguientes acciones:

1- Creamos la columna 'mes' con el mes indicado en la columna 'fecha'

2- Contamos la frecuencia por mes y graficamos los resultados




Podemos observar en este análisis un pico de reclamos en Marzo. Las razones pueden ser las siguientes y una baja de reclamos en los meses de agosto, septiembre, octubre y diciembre.

**Marzo** – regreso del receso de Vacaciones de verano.

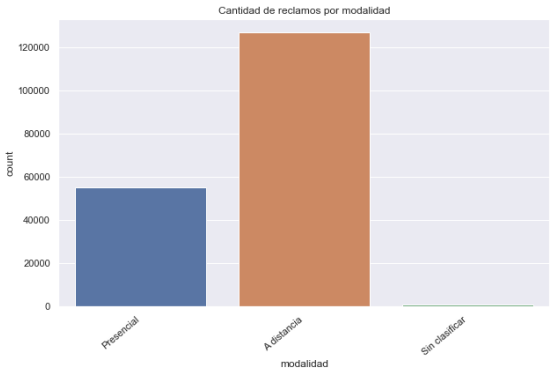
**Septiembre-Diciembre** – Inicio el receso de Vacaciones de Verano.

## 7.3 Reclamos por modalidad

Como nos interesa saber la cantidad de reclamos por modalidad, es decir por qué medio se hacen efectivos lo reclamos.

Se puede decir que hay una mayor cantidad de reclamos que se hacen a distancia, debido a la facilidad de iniciar el reclamo a través de la web este medio es el preferido por los denunciantes.

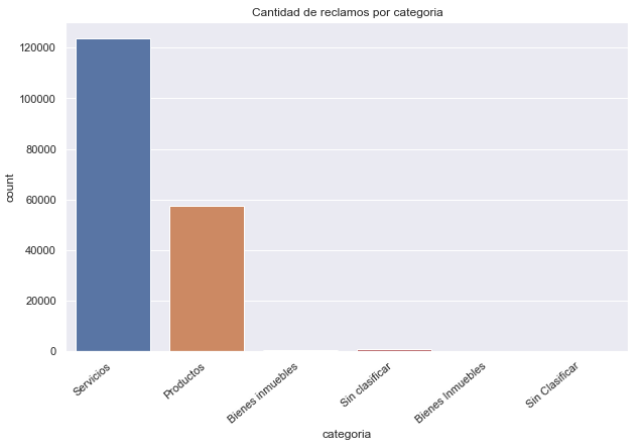
https://www.argentina.gob.ar/iniciar-un-reclamo-ante-defensa-del-consumidor




## 7.4 Reclamos por categoría

Como nos interesa saber la cantidad de reclamos por categoría se realizar las siguientes acciones:

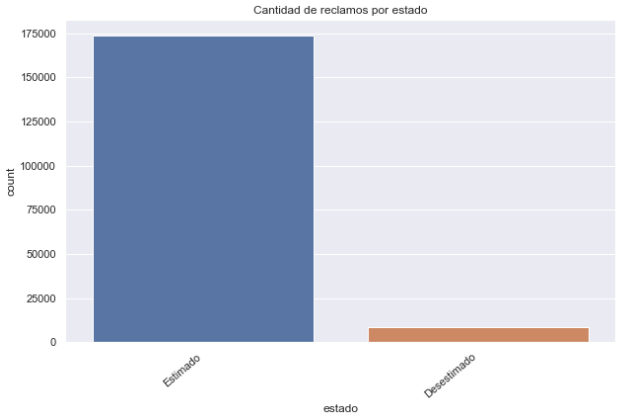
1- Contamos la frecuencia por categoría y graficamos los resultados.




Se puede observar que la mayor proporción de los reclamos son por servicios, seguido de productos.

## 7.5 Reclamos por estado

Como nos interesa saber la cantidad de reclamos por estado procedemos a realizar las siguientes acciones:

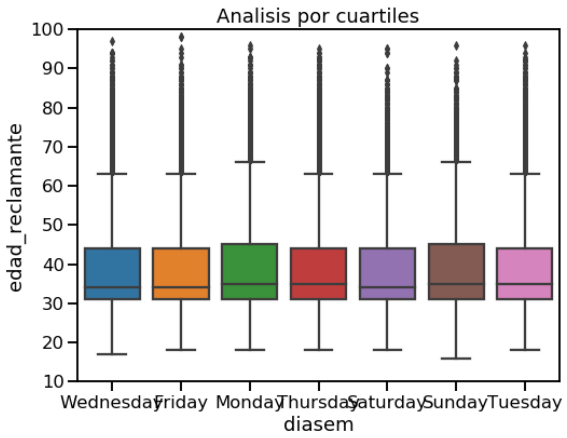



Contamos la frecuencia por estado del reclamo y graficamos los resultados.

Vamos a considerar los reclamos que se estiman y continúan curso.

### 7.6 Edad de los reclamantes por día de la semana

Como nos interesa saber la cantidad de reclamo por día de la semana contamos la frecuencia por edad del reclamante y graficamos los resultados.




## 8 .DUMMIES

Realizamos dummies a partir de la columna 'Estado' y obtenemos dos columnas a partir de esta obteniendo 'estimado' y 'desestimado' para crear nuevos valores en base a esta información esto se puede analizar gracias a que esta variable es categórica.

| Edad | Altura | Sexo |
|---|---|---|
| 18 | 1.70 | Masculino |
| 24 | 1.60 | Femenino |
| 30 | 1.90 | Femenino |
| 28 | 1.5 | Masculino |

→

| Edad | Altura | Sexo | Masculino | Femenino |
|---|---|---|---|---|
| 18 | 1.70 | Masculino | 1 | 0 |
| 24 | 1.60 | Femenino | 0 | 1 |
| 30 | 1.90 | Femenino | 0 | 1 |
| 28 | 1.5 | Masculino | 1 | 0 |

Se realizó un proceso de DUMMIES de manera manual.

## 9 .APRENDIZAJE SUPERVISADO

Decidimos utilizar el aprendizaje supervisado para poder analizar si había una relación entre las condiciones del reclamo y la cantidad de reclamos estimados.

Determinamos el feature que en nuestro modelo usaremos como función dependiente (y). Para eso extrajimos la columna 'Estimados' para predecirlas en una sola categoría. Luego borramos las columnas originales ya que no las usaremos más adelante.


- We assume Y variable depends on X.
- **What we do not know is the true function/rule y = f(x).**

## 10 MODELOS

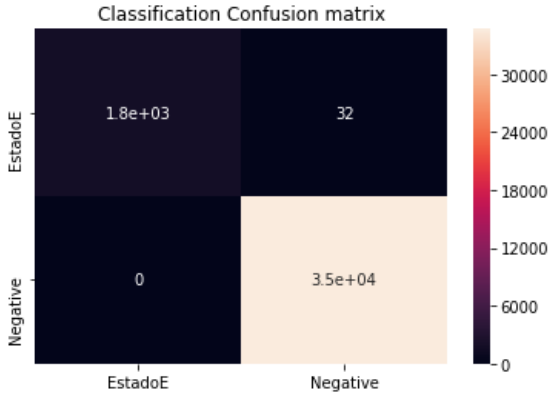
### 10.1 Knn

Se inicia la etapa de predicción intentando modelizar con KNN. En principio se obtiene un accuracy de 0.99 ingresando como parámetro la cantidad de vecinos = 23.

Con este valor, el accuracy sube hasta 0.99. Es en esta instancia cuando se decide recurrir a los metadatos implícitos en las features generando nueva información para alimentar al modelo (Ver sección Features).

#### 10.1.1 Matriz de confusión

Nos permite visualizar los True Positive y True Negative de los resultados del modelo KNN.



### 10.2 SVM

Una vez terminado el KNN se opta por evaluar el clasificador de SVM. Análisis del que no se pudo obtener datos significativos de variación de Accuracy concretos.

## 11 RESULTADOS

Usando el modelo KNN el máximo valor que logramos alcanzar de Accuracy es 0.99 es el valor más alto.

## 12 CONCLUSIONES

Cabe la posibilidad de que el modelo se haya "sobreajustado" a los datos de entrenamiento, por lo tanto se debería analizar una modificación de los datos de training.

Se debería replantear una cantidad menos restrictiva de features para el análisis SVM ya que solo se terminan eligiendo tres.

La aplicación práctica de este análisis tiene aplicación sobre servicios de atención al cliente, análisis de perfiles de consumidores insatisfechos.

El análisis hecho corresponde a una primera instancia del estudio del caso, con fines netamente académicos. Incorporando los datos externos mencionados y evaluando otros modelos pueden obtenerse herramientas con fines predictivos que generen impacto, con la **posibilidad real de salvar vidas**. La aplicación práctica tiene incidencia sobre los servicios de salud, seguridad, control vehicular.