FUNDACIÓN YPF

# INGENIAS+ | DATA SCIENCE

# DEFORESTACIÓN

## *Proyecto de análisis de datos de la deforestación en Argentina (2001-2020)*


***Integrantes:***
**Ulloa Melina**
**Cruz Nicole**

FUNDACIÓN YPF

# CONTENIDO

FUNDACIÓN YPF

# INTRODUCCIÓN

## ¿POR QUÉ ES UN PROBLEMA IMPORTANTE EN ARGENTINA?

*La deforestación es un problema crítico que afecta tanto al medio ambiente como a las comunidades. En regiones como la Amazonía, la tala de árboles contribuye al cambio climático, pone en riesgo especies y altera los ciclos del agua. Además, impacta a las comunidades locales que dependen de los bosques para su supervivencia. A través de este análisis, buscamos entender mejor las causas y consecuencias de la deforestación, y explorar soluciones para reducir sus efectos.*

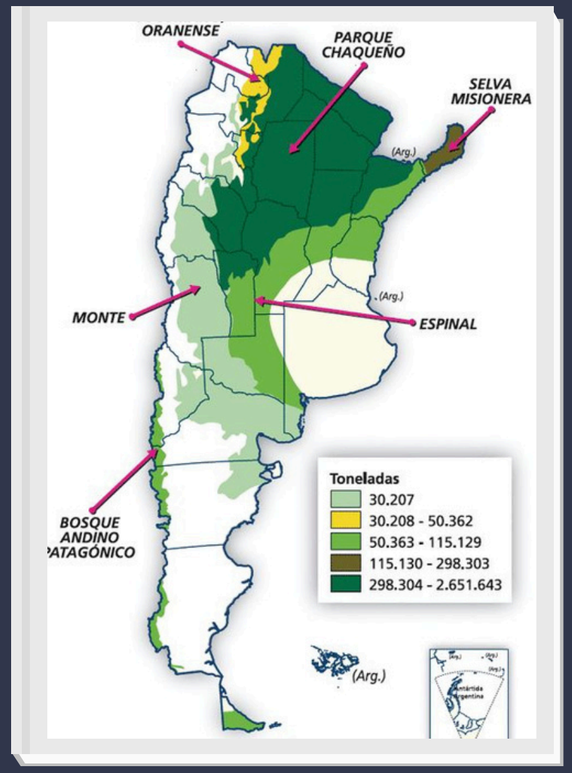
FUNDACIÓN YPF

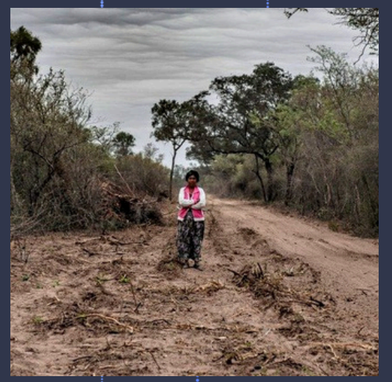
# 1. OBJETIVO DEL PROYECTO

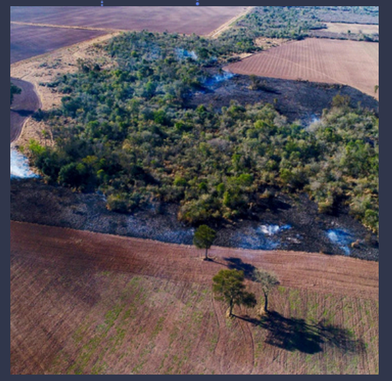
- Analizar datos reales sobre la deforestación.
- Visualizar el impacto mediante herramientas de Data Science.
- Generar conciencia ambiental.
- Detectar provincias con mayor impacto de deforestación.
- Predecir la cantidad de hectáreas deforestadas por año y región.
- Agrupar zonas según su comportamiento ambiental.
- Generar visualizaciones claras para apoyar políticas públicas.
- Establecer bases para futuros estudios con más variables.
- aportar información útil para la toma de decisiones ambientales.

# 2.HIPÓTESIS

La deforestación en Argentina entre 2001 y 2020 no se distribuye de forma homogénea en el territorio nacional, sino que se concentra en ciertos grupos de provincias con patrones específicos de intensidad y evolución temporal, los cuales pueden ser identificados y caracterizados mediante técnicas de machine learning.





- No todas las provincias sufren deforestación igual.
- Algunas tienen **más deforestación, otras menos, y otras más reciente**.
- Esas diferencias se pueden agrupar: por ejemplo, provincias del norte que pierden bosques todo el tiempo (como **Chaco** o **Santiago del Estero**), y otras que tienen un impacto bajo.
- Con herramientas de machine learning, como **Random Forest** o **KMeans**, podemos encontrar esos grupos automáticamente y hacer predicciones de **cuánto se deforesta cada año y dónde**.
- Eso ayuda a tomar decisiones: por ejemplo, **saber dónde se necesita más control ambiental.**

# 3.METODOLOGÍA

## *DataSet Utilizados:*

1. ***Argentina_Deforestacion.csv:***
   - **Contenía registros por año, provincia y región.**
   - **Incluía 2.381 filas.**
   - **Columnas clave: year, region, parent_region, deforestation_hectares.**

2. ***Forest_ConservationAreas_Funding_inSouthAmerica_PAN2024.csv***
   - ***Aportó 10.328 filas (antes de limpieza).***
   - **Sumó nuevas regiones y años, no contemplados en el primer archivo.**
   - **Trajo información territorial más detallada.**

*Después de limpiar y unir ambos datasets, obtuvimos un total de 11.417 observaciones.*
*Esto fue clave para:*
- *Aumentar la cobertura territorial y temporal*
- *Mejorar la calidad del análisis y las predicciones.*
- *Modelos más precisos y generalizables → Más datos y más variedad hacen que los modelos de Machine Learning tengan mejor capacidad de detectar patrones reales y no se queden "ciegos" ante zonas sin datos.*

| year | country | country_iso2 | region | region_trase_id | parent_region | parent_region_trase_ | deforestation_hecta |
|---|---|---|---|---|---|---|---|
| 2001 | ARGENTINA | AR | 12 DE OCTUBRE | AR-22036 | CHACO | AR-22 | 2281.40134 |
| 2001 | ARGENTINA | AR | 1° DE MAYO | AR-22126 | CHACO | AR-22 | 199.1756951 |
| 2001 | ARGENTINA | AR | 25 DE MAYO | AR-22168 | CHACO | AR-22 | 315.4334006 |
| 2001 | ARGENTINA | AR | 2 DE ABRIL | AR-22039 | CHACO | AR-22 | 293.3001777 |
| 2001 | ARGENTINA | AR | 9 DE JULIO | AR-22105 | CHACO | AR-22 | 671.2301017 |
| 2001 | ARGENTINA | AR | 9 DE JULIO | AR-82077 | SANTA FE | AR-82 | 2751.792784 |
| 2001 | ARGENTINA | AR | AGUIRRE | AR-86007 | SANTIAGO DEL EST | AR-86 | 3080.422561 |
| 2001 | ARGENTINA | AR | ALBERDI | AR-86014 | SANTIAGO DEL EST | AR-86 | 16151.43562 |
| 2001 | ARGENTINA | AR | ALMIRANTE BROWN | AR-22007 | CHACO | AR-22 | 4754.504445 |
| 2001 | ARGENTINA | AR | AMBATO | AR-10007 | CATAMARCA | AR-10 | 52.07355947 |
| 2001 | ARGENTINA | AR | ANCASTI | AR-10014 | CATAMARCA | AR-10 | 29.25274934 |

| year | parent_region | deforestation_hectar |
|---|---|---|
| 2001 | CHACO | 2281.40134 |
| 2001 | CHACO | 199.1756951 |
| 2001 | CHACO | 315.4334006 |
| 2001 | CHACO | 293.3001777 |
| 2001 | CHACO | 671.2301017 |
| 2001 | SANTA FE | 2751.792784 |
| 2001 | SANTIAGO DEL EST | 3080.422561 |
| 2001 | SANTIAGO DEL EST | 16151.43562 |
| 2001 | CHACO | 4754.504445 |
| 2001 | CATAMARCA | 52.07355947 |
| 2001 | CATAMARCA | 29.25274934 |
| 2001 | SALTA | 8833.862742 |
| 2001 | SANTIAGO DEL EST | 865.0087292 |

# 4. LIMPIEZA Y PREPARACIÓN DE DATOS.

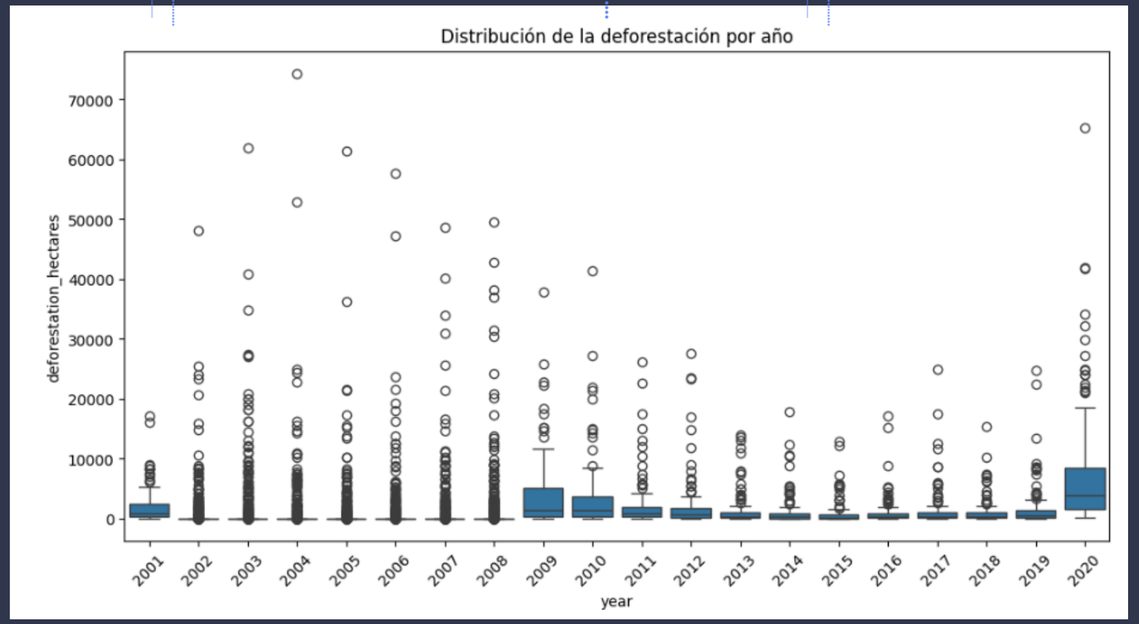
- Eliminación de columnas irrelevantes: Se descartaron IDs y códigos sin valor analítico.
  - ***drop() de Pandas***

- Revisión de nulos y duplicados: No se detectaron valores nulos ni duplicados.
  *-df.isnull().sum() para nulos*
  *-df.duplicated().sum() para duplicados*

- Renombrado de columnas: Se estandarizaron nombres al español y con mejor formato.
  ***-Renombrado con df.rename(columns={...})***

- Conversión de tipos de datos(*LabelEncoder*) : Nos aseguramos que los datos tengan el tipo correcto, **Año (int), Deforestación (float), Región (string)**
  **-df.astype({'columna': tipo})**

- Codificación de variables categóricas: parent_region → parent_region_encoded con ***LabelEncoder*.**




***El boxplot muestra valores atípicos en la deforestación por año. Detectarlos fue clave para evitar sesgos y asegurar un análisis más preciso.***

# 5. PREPARACIÓN PARA EL ANÁLISIS / CLUSTERING

◆ ***División del dataset:***

- 80% de los datos para entrenamiento
- 20% para prueba

🧠 ***"Esta división se usa para evaluar modelos supervisados como Random Forest, asegurando una evaluación objetiva."***

◆ **Agrupación de datos por región:**

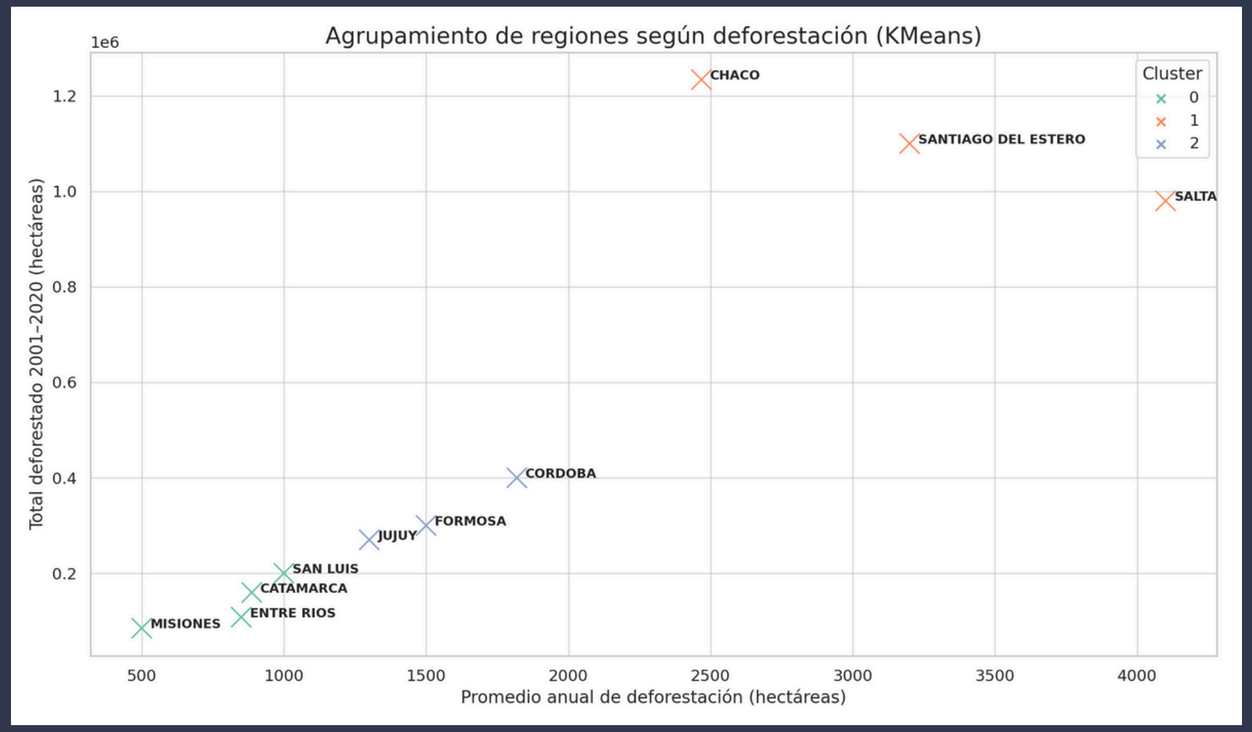
- Se calculó el promedio de deforestación
- Se obtuvo el total acumulado
- Se contó la cantidad de años registrados

🧠 ***"Estas variables resumen el comportamiento de cada región y permiten compararlas para aplicar clustering (agrupamiento)."***

# 6. AGRUPAMIENTO CON KMEANS

*-Machine learning no supervisado*

Se aplicó el algoritmo KMeans (k=3) para agrupar provincias con comportamientos similares de deforestación.
Las regiones se agrupan naturalmente según cómo se comporta la deforestación.




◆ **Cluster 1 – Alto impacto sostenido**

- Promedio: ~4.435 hectáreas/año
- Total: ~2.3 millones de hectáreas
- Provincias típicas: Chaco, Santiago del Estero, Salta

Son zonas con deforestación fuerte y constante durante los 20 años.

◆ **Cluster 2 – Alto impacto pero con menor historial**

- Promedio: ~4.710 hectáreas/año
- Total: ~1.15 millones de hectáreas
- Menos años registrados

Zonas que empezaron a ser afectadas más recientemente o que se monitorean desde hace menos tiempo.

◆ **Cluster 0 – Bajo impacto**

- Promedio: ~899 hectáreas/año
- Total: ~108.000 hectáreas
- Más años de datos

Zonas con deforestación baja o controlada.

FUNDACIÓN YPF

# 7. SILHOUETTE SCORE

Para validar la calidad del agrupamiento
Obtuvimos uno de 0.40, lo que muestra una separación moderada entre los grupos
Confirma que aunque hay cierta suposición, las regiones se agrupan con características similares de deforestación. Permitiendo identificar zonas con diferentes niveles de impacto ambiental.

```
from sklearn.metrics import silhouette_score

# Usamos los datos escalados para el cálculo del Silhouette Score
score = silhouette_score(X_scaled, kmeans.labels_)

# Mostramos el resultado
print("Silhouette Score:", score)
```

```
Silhouette Score: 0.4072886032906074
```

# 8. MODELO PREDICTIVO CON RANDOM FOREST

***Se entrenó usando variables predictoras: año (year) y región codificada (parent_region_encoded).***

- ***División de datos:***

*80% para entrenar*
*20% para evaluar*

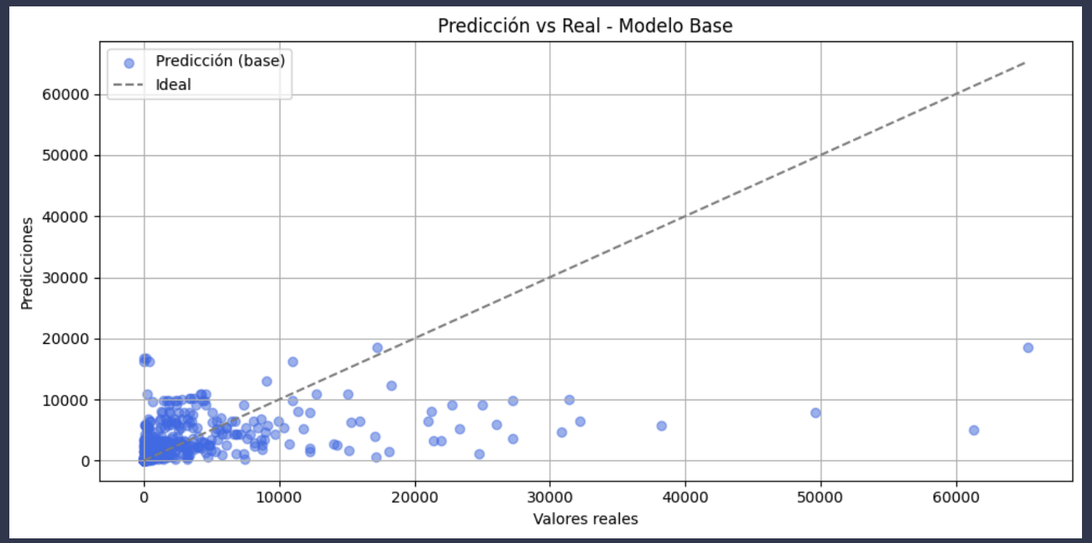
- ***Métricas de evaluación:***

*MAE (Error Absoluto Medio): 658,06 hectáreas (error promedio en predicción)*
*MSE (Error Cuadrático Medio): 8.625.511,13 (sensibilidad a valores extremos)*
*$R^2$ (Coeficiente de Determinación): 0,29 (el modelo explica el 29% de la variabilidad)*

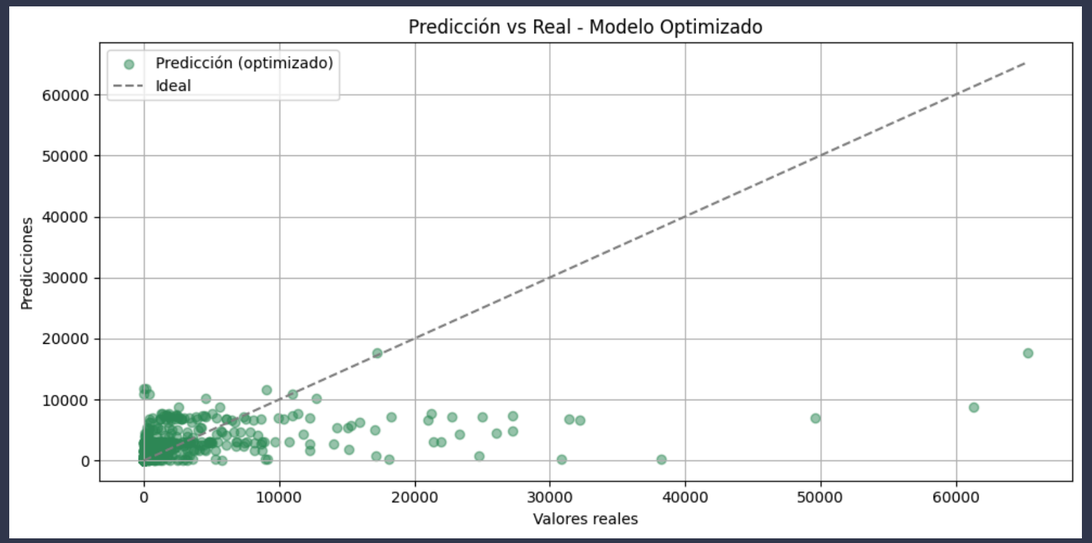
🧠 ***Este desempeño moderado sirve como base para mejorar el modelo con más datos o variables.***

# 9.PREDICCION VS DISPERSIÓN REAL

## *Modelo base*

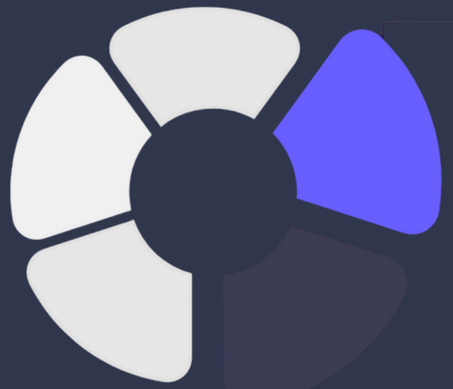



## *Modelo optimizado*



# 10. CONCLUSIONES

- *Las provincias presentan patrones ambientales diferentes*
- *Es posible agruparlas por su comportamiento frente a la deforestación*
- *Algunas requieren control y monitoreo constante*
- *Otras necesitan educación ambiental o asistencia territorial*



# 11. PROPUESTA O SOLUCIÓN

- *Campañas de concientización en zonas críticas*
- *Uso de imágenes satelitales para monitoreo en tiempo real*
- *Educación ambiental en escuelas, adaptada a cada región*
- *Políticas públicas diferenciadas por tipo de clúster*

# MUCHAS GRACIAS

INGENIAS+

FUNDACIÓN YPF