

# Informe Final — Clasificación de Riesgo de Inundación

**Alumna: Miriam Velazque**

**Materia: Aprendizaje Automático**

## 1. Introducción y Objetivo

El proyecto busca desarrollar un sistema de clasificación binaria para zonas afectadas por inundaciones en Argentina, combinando variables geoespaciales y ambientales. El objetivo es predecir si una zona se inunda parcialmente (1) o totalmente (2), para apoyar decisiones estratégicas como evacuación, asignación de recursos y planificación territorial.

## 2. Dataset y Fuentes

- **Fuente principal:** DU-Tech y RENABAP (datos.gob.ar).
- **Capas utilizadas:**
  - du\_techo\_zonas\_inundables.csv: contiene la variable objetivo se\_inunda\_.
  - du\_techo\_cuerpos\_de\_agua.geojson, du\_techo\_cursos\_de\_agua.geojson, du\_techo\_microbasurales.geojson: capas geoespaciales ambientales.
- **Cantidad de registros:** 2746 zonas.
- **Formato:** CSV y GeoJSON, procesados con pandas y geopandas.

## 3. Exploración y Preprocesamiento

### 3.1 Limpieza semántica

- Corrección de errores tipográficos (provincia, departamen).
- Normalización de texto y codificación categórica (provincia, departamento).

### 3.2 Codificación de variables

- se\_inunda\_: transformada en se\_inunda\_binaria (1 = parcial, 2 = total).
- con\_que\_fr: agrupada en frecuencia\_codificada (Alta, Media, Baja, Otro).

### 3.3 Transformaciones espaciales

- Conversión de geometrías WKT a objetos MultiPolygon.
- Cálculo de centroides y reproyección a EPSG:5347.
- Cálculo de distancias:
  - distancia\_cuerpo\_agua: distancia al cuerpo de agua más cercano.
  - distancia\_curso\_agua: distancia al curso de agua más cercano.
- Conteo de microbasurales en un radio de 500m (microbasurales\_cercanos).

## 4. Preparación del Dataset Final

- Verificación de tipos y nulos.
- Codificación one-hot de frecuencia\_codificada.
- Exportación como dataset\_modelo\_inundaciones.csv.

## 5. Modelado Supervisado

### 5.1 Variables utilizadas

- Predictoras: distancia\_cuerpo\_agua, distancia\_curso\_agua, microbasurales\_cercanos, frecuencia\_num.
- Objetivo: se\_inunda\_binaria.

### 5.2 Modelos entrenados

Se entrenaron y evaluaron cuatro modelos clásicos de clasificación binaria:

Modelo	Accuracy	Recall clase 2	F1 clase 2	ROC AUC
Árbol de Decisión	0.560	0.393	0.472	—
Regresión Logística	0.587	0.607	0.595	—
Random Forest	0.580	0.549	0.567	0.612
KNN (k=7)	0.573	0.567	0.570	—

## 6. Justificación de Modelos

### Árbol de Decisión

Explorado por su interpretabilidad. Visualiza reglas claras, pero bajo recall en clase 2 (0.393) lo descarta como modelo final.

### Regresión Logística

Modelo lineal, útil para interpretar coeficientes. Buen recall en clase 2 (0.607), pero menor ROC AUC que Random Forest.

### Random Forest

Modelo final elegido. Maneja no linealidades, evita overfitting y ofrece mejor discriminación entre clases (ROC AUC = 0.612).

### KNN (k=7)

Explorado con distintos valores de  $k$ . El mejor rendimiento se obtuvo con  $k = 7$ , con métricas equilibradas y aceptables.

## 7. Visualizaciones Clave

- Histograma de frecuencia de inundación (frecuencia\_codificada).
- Boxplot de distancia a cuerpos de agua según tipo de inundación.
- Matriz de confusión de cada modelo.

- **Curva ROC de Random Forest** (ya generada).

## **8. Aplicaciones del Sistema**

- Clasificación automática de zonas no etiquetadas.
- Actualización dinámica ante cambios ambientales (ej. aparición de microbasurales).
- Soporte a decisiones de evacuación, asignación de recursos y planificación urbana.

## **9. Conclusión**

El sistema desarrollado es reproducible, técnicamente sólido y aplicable a escenarios reales. Aunque las métricas son moderadas, el enfoque permite escalar y mejorar en futuras versiones. La elección de Random Forest como modelo final se justifica por su rendimiento y robustez.