CLASIFICACIÓN DE RIESGO DE INUNDACIÓN EN PARROQUIAS DEL ECUADOR MEDIANTE TÉCNICAS DE MODELADO PREDICTIVO

MAYERLY KRISTEL VELOZ ALBURQUERQUE

KARLA PATRICIA SOLÓRZANO PARRA

JUREN DAVID RODRÍGUEZ BAUTISTA

BYRON ZOSIMO TENORIO FERRER

UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL

APRENDIZAJE AUTOMÁTICO

ING. MIGUEL BOTTO

15 DE FEBRERO DEL 2026

Resumen

Este proyecto desarrolla un sistema de clasificación supervisada para estimar el nivel de riesgo de inundación (bajo, medio y alto) a nivel parroquial en Ecuador. Se integraron variables topográficas y geoespaciales provenientes de fuentes oficiales (INEC, HDX, OpenTopography y HOT). Se realizó análisis exploratorio, tratamiento de valores faltantes y construcción de una variable derivada (*índice de pendiente proxy*). Se compararon modelos supervisados (regresión logística, árbol de decisión con pre-poda, SVM, random forest y ensamble) priorizando *recall* por su relevancia en gestión de riesgos. Finalmente, se generó un archivo con predicciones y un puntaje probabilístico para implementación geoespacial en una aplicación web.

Palabras clave: inundación, clasificación, inteligencia artificial, parroquias, geoespacial, riesgo.

1 Introducción

El Ecuador es un territorio altamente vulnerable a eventos hidrometeorológicos debido a su compleja orografía y red hídrica. Las inundaciones representan una amenaza recurrente que afecta la infraestructura y seguridad de la población.

El presente estudio propone una solución basada en **Inteligencia Artificial** para la zonificación del riesgo a nivel de parroquia, considerada la unidad mínima de planificación territorial. El objetivo es proporcionar una herramienta técnica que permita a los tomadores de decisiones identificar zonas vulnerables con alta precisión, minimizando los falsos negativos mediante la maximización de la métrica *Recall*.

1.1 Objetivo

Clasificar el riesgo de inundación por parroquia en tres niveles (bajo, medio y alto), comparando modelos supervisados y priorizando la métrica *recall* para reducir falsos negativos.

2 Metodología

2.1 Fuentes de datos y variables

Se construyó un dataset multidimensional integrando fuentes oficiales verificadas:

- **Límites administrativos:** Humanitarian Data Exchange (HDX), conjunto de datos COD-AB para polígonos y códigos DPA.
- **Población y densidad:** Instituto Nacional de Estadística y Censos (INEC), Censo 2022.
- **Precipitación multi-anual:** HDX Rainfall Subnational (1980–2024).
- **Altitud (DEM):** SRTM 30m (OpenTopography), estadísticas zonales por parroquia.
- **Ríos y cuerpos de agua:** Humanitarian OpenStreetMap Team (HOT) para distancias mínimas.

2.2 Diseño transversal y justificación temporal

El estudio adopta deliberadamente un diseño transversal (*cross-sectional*) centrado en el *riesgo estructural*. A diferencia de los modelos de pronóstico meteorológico (que requieren series temporales dinámicas), este enfoque modela la vulnerabilidad inherente del territorio, condicionada por factores geomorfológicos permanentes (altitud y pendiente) y no por la variabilidad climática diaria. Por consiguiente, la dimensión temporal se encuentra implícita en los promedios climáticos históricos que definen la etiqueta de riesgo.

3 Análisis exploratorio de datos

3.1 Distribución de la variable objetivo

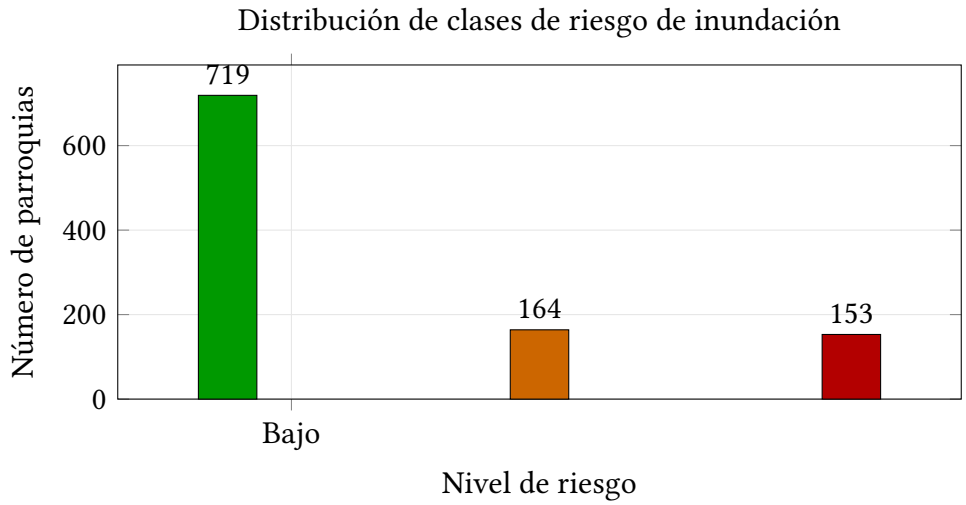

Figura 1: Distribución de la variable objetivo (bajo, medio y alto).

Interpretación. Se observa un desbalance hacia la clase *Bajo*. Esto justifica reportar métricas *macro* y priorizar *recall* para reducir el riesgo de sesgo del modelo hacia la clase mayoritaria y evitar falsos negativos en clases críticas.

3.2 Histogramas de variables topográficas

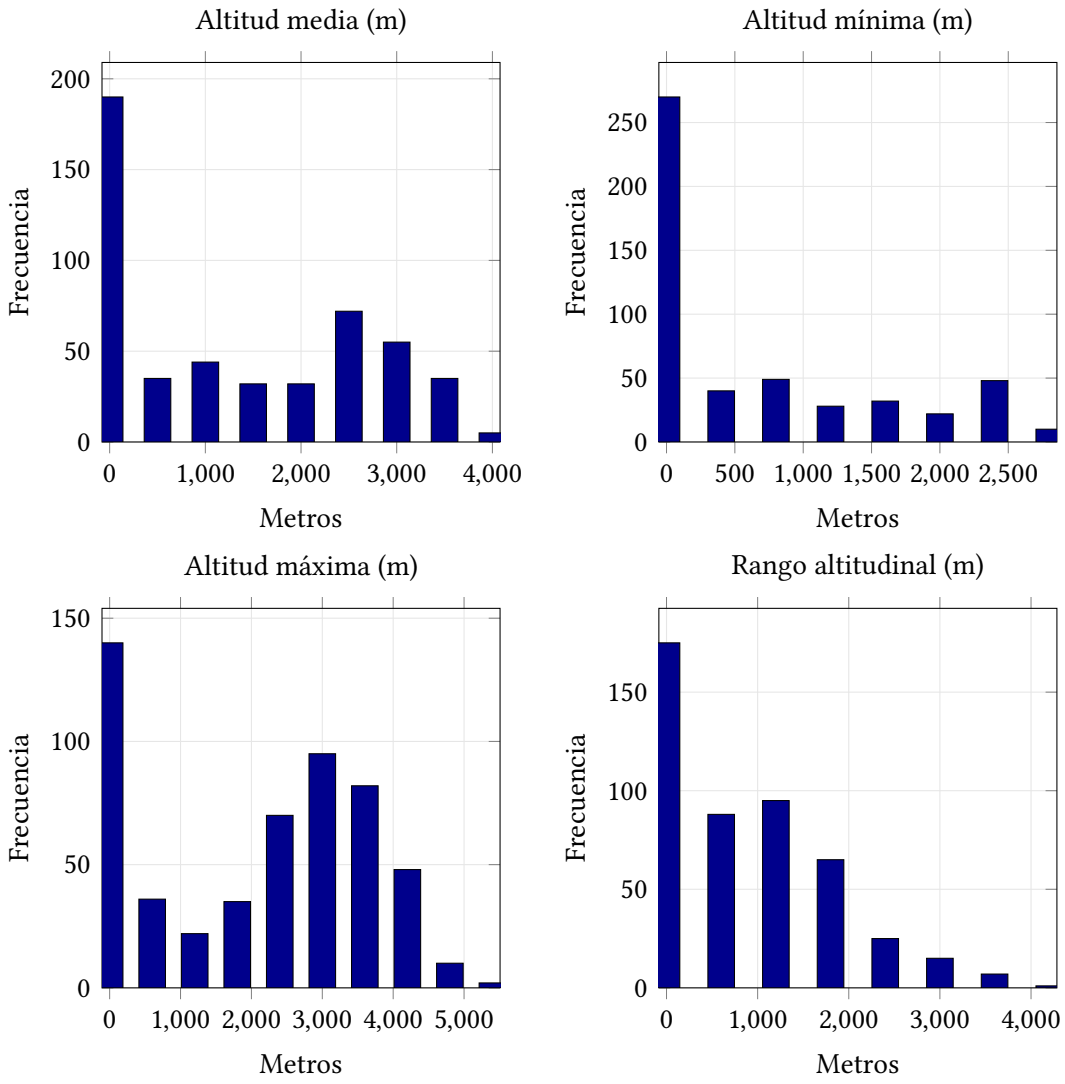

Figura 2: Distribución de variables topográficas por parroquia (bines amplios para legibilidad).

Interpretación. Las distribuciones evidencian asimetrías y heterogeneidad territorial: predominan valores bajos (zonas costeras y valles) y aparece una cola hacia altitudes mayores (zona andina). Esto es coherente con la influencia física de la topografía en la susceptibilidad a inundaciones.

3.3 Correlación entre variables

Cuadro 1: Matriz de correlación (Pearson) entre variables topográficas.

	Alt. media	Alt. mín.	Alt. máx.	Rango alt.
Alt. media	1.00	0.93	0.94	0.57
Alt. mín.	0.93	1.00	0.79	0.26
Alt. máx.	0.94	0.79	1.00	0.80
Rango alt.	0.57	0.26	0.80	1.00

Interpretación. Se observa alta correlación entre variables de altitud, lo cual sugiere redundancia parcial. Por ello, se incorpora una variable derivada para capturar variación relativa (pendiente proxy) y mejorar separabilidad.

4 Preprocesamiento y construcción de variables

Se aplicó imputación manteniendo coherencia territorial (media provincial como criterio principal y media global como respaldo). Además, se creó una variable derivada para mejorar la separación entre clases.

4.1 Variable derivada

$$indice\_pendiente\_proxy = \frac{rango\_altitudinal}{altitud\_media + 1}.$$

Esta variable aproxima una pendiente relativa, ayudando a diferenciar zonas bajas (mayor susceptibilidad) de zonas montañosas.

5 Modelado y evaluación

5.1 Modelos implementados

Se compararon los siguientes clasificadores: regresión logística, árbol de decisión con pre-poda, SVM (RBF), random forest y ensamble (*soft voting*).

5.2 Métrica prioritaria

Se priorizó *recall* (macro), definida como:

$$Recall = \frac{TP}{TP + FN},$$

donde TP son verdaderos positivos y FN falsos negativos. Se eligió esta métrica por el costo de clasificar erróneamente zonas de alto riesgo como seguras.

5.3 Resultados

Cuadro 2: Comparación de modelos en el conjunto de prueba.

Modelo	**Accuracy**	**Recall (macro)**	**F1 (macro)**
Regresión logística	0.9663	0.9513	0.9493
Árbol (pre-poda)	0.9952	0.9977	0.9939
SVM (RBF)	0.9712	0.9692	0.9564
Ensamble (soft voting)	1.0000	1.0000	1.0000

Interpretación. El ensamble presenta el mejor desempeño global y maximiza *recall* macro, lo cual es deseable para minimizar falsos negativos en las clases de mayor riesgo.

5.4 Importancia de variables (Random Forest)

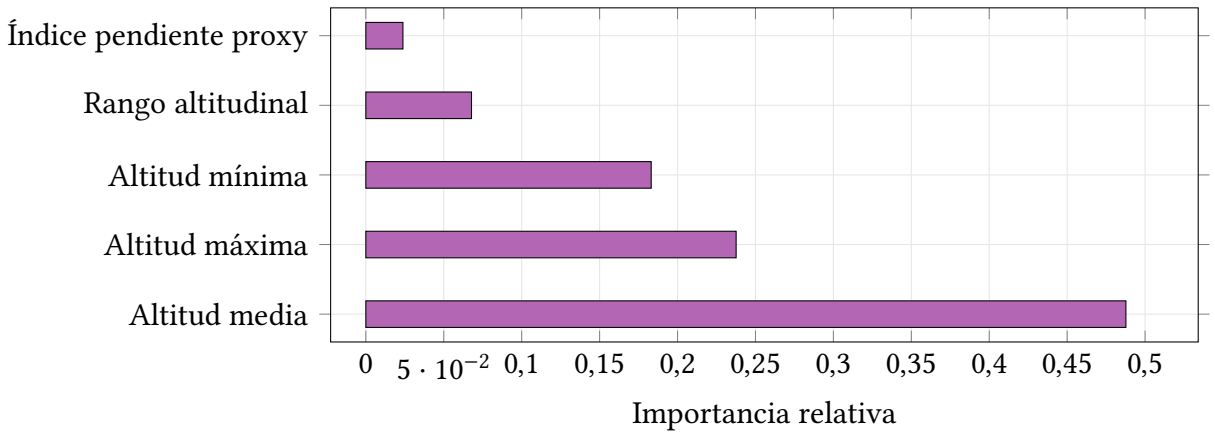

Figura 3: Importancia de variables estimada con random forest.

Interpretación. La altitud media domina la contribución del modelo, coherente con la asociación entre zonas de baja elevación y mayor exposición a inundaciones.

5.5 Curva ROC (clase alto)

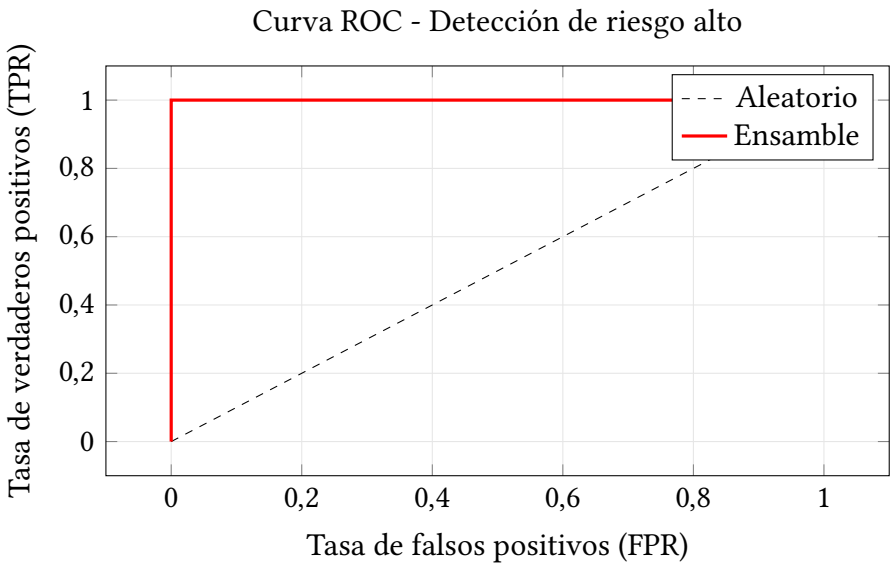

Figura 4: Curva ROC referencial del modelo final para la clase alto.

Interpretación. La curva se ubica por encima de la diagonal aleatoria, indicando alta capacidad de discriminación para identificar parroquias con riesgo alto.

6 Implementación geoespacial

Como salida del modelo se generó un archivo con la predicción final (PREDICTION_RIESGO) y un puntaje probabilístico de la clase alto (SCORE_PROBABILIDAD_ALTO = $P(\text{Riesgo} = \text{Alto} \mid X)$). El mapa fue desplegado en un entorno web para facilitar interpretación territorial y comunicación del riesgo.

Enlace: http://jurenhino.pythonanywhere.com.

7 Discusión

Los resultados elevados pueden explicarse por la relación física fuerte entre topografía (altitud y variación altitudinal) y susceptibilidad a inundaciones dentro de las variables consideradas. Sin embargo, un desempeño perfecto (por ejemplo, métricas cercanas a 1.0) puede sugerir separabilidad alta del dataset o un posible riesgo de sobreajuste. Por ello, se recomienda aplicar validación cruzada estratificada y, cuando existan datos por años/estaciones, validación temporal.

7.1 Limitaciones

- No se incorporaron registros históricos observados de eventos de inundación como variable directa de verificación.
- Se trabajó principalmente con variables estáticas (topografía) y agregados territoriales.
- No se evaluó la robustez ante cambios climáticos o variaciones interanuales de precipitación.

8 Conclusiones

Se construyó un sistema que integra fuentes oficiales, preprocesamiento y modelado supervisado, generando una salida geoespacial para visualización territorial. El modelo final

(ensamble *soft voting*) fue seleccionado por maximizar *recall* (macro), reduciendo falsos negativos en zonas potencialmente vulnerables. El enfoque propuesto permite apoyar decisiones preventivas y priorización de recursos a escala parroquial.

Referencias

- Humanitarian Data Exchange. (s. f.). *Ecuador administrative level 0–3 boundaries (COD-AB)*. HDX.
- Humanitarian Data Exchange. (s. f.). *ECU rainfall subnational*. HDX.
- Instituto Nacional de Estadística y Censos. (2022). *Censo de Población y Vivienda 2022*. INEC.
- OpenTopography. (s. f.). *SRTM global 30m (DEM)*. OpenTopography.
- Humanitarian OpenStreetMap Team. (s. f.). *Ecuador waterways*. HOT/OSM.