Usar datos para desenterrar tragedias

¿Cómo se puede utilizar la ciencia de datos para predecir la existencia de fosas clandestinas en México?

Jorge Ruiz - PDH, Universidad Iberoamericana

05 de marzo de 2021

Estructura de la presentación

- Contexto del proyecto
- ¿Cómo funciona nuestro modelo de predicción de fosas clandestinas?
- Resultados
- Pasos a seguir

Contexto del proyecto

México está atravesando por una crisis de derechos humanos, como consecuencia del uso de las fuerzas armadas para llevar a cabo labores de seguridad pública.

- Sexenio de Felipe Calderón Hinojosa (2006 - 2012)
- Sexenio de Enrique Peña Nieto (2012 - 2018)
- Sexenio de Andrés Manuel López Obrador (2018 - presente)

Desapariciones generalizadas

> *"El Comité lamenta profundamente que se mantiene una situación de desapariciones generalizadas en gran parte del territorio del Estado parte y que imperen la impunidad y la revictimización"* *Comité CED de Naciones Unidas, 2018*

Según los datos oficiales, actualmente en México hay **84 mil 134 personas desaparecidas**.

Más del 95% de los casos son posteriores a 2007.

¿Por qué estamos haciendo esto?

Queremos responder una pregunta: **¿dónde buscar?**

En México, quienes buscan a las personas desaparecidas son sus mismos familiares.

Creemos que esta herramienta puede fortalecer los procesos de incidencia de colectivas de familiares en búsqueda en diferentes entidades de México, así como apoyar las labores de búsqueda de personas desaparecidas de autoridades federales y estatales.

¿Cómo funciona el modelo?

Partimos de la premisa de que las fosas que han sido observadas en el país son solo una fracción del universo de fosas clandestinas (porque además se encuentran fosas casi a diario).

¿Podemos conocer el universo entero de fosas en México?

No sabemos, pero esto es un paso más para acercarnos a él. Queremos, a grandes razgos, clasificar nuestros municipios en municipios donde puede haber fosas clandestinas y municipios donde no.

¿Cómo funciona el modelo?

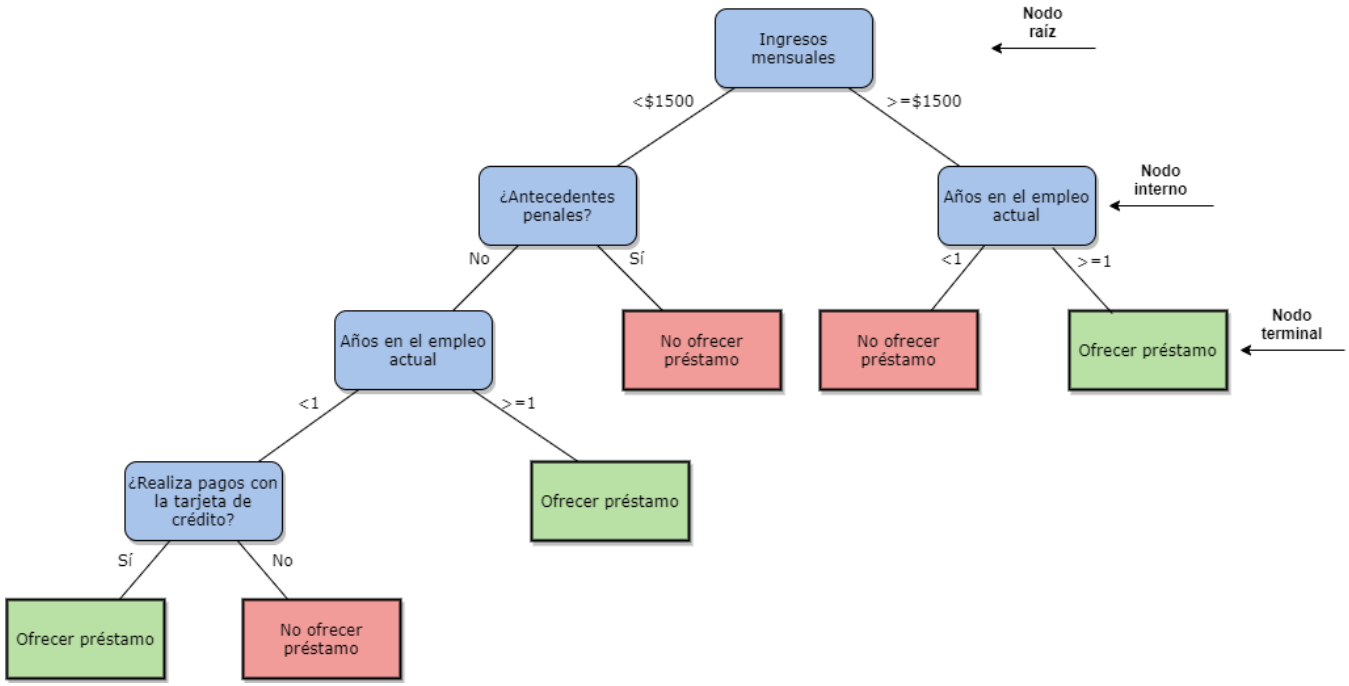
Utilizamos un modelo conocido en jerga de *machine learning* como *random forest*.

Este modelo tiene muchas ventajas:

- Es MUY sencillo.
- No es muy sensible a *missing values* (valores perdidos/datos incompletos).
- Nos ayuda a evitar de mejor manera el *overfitting* que otros modelos.
- Se puede utilizar tanto para regresiones como clasificaciones, **nosotros queremos lo segundo**.

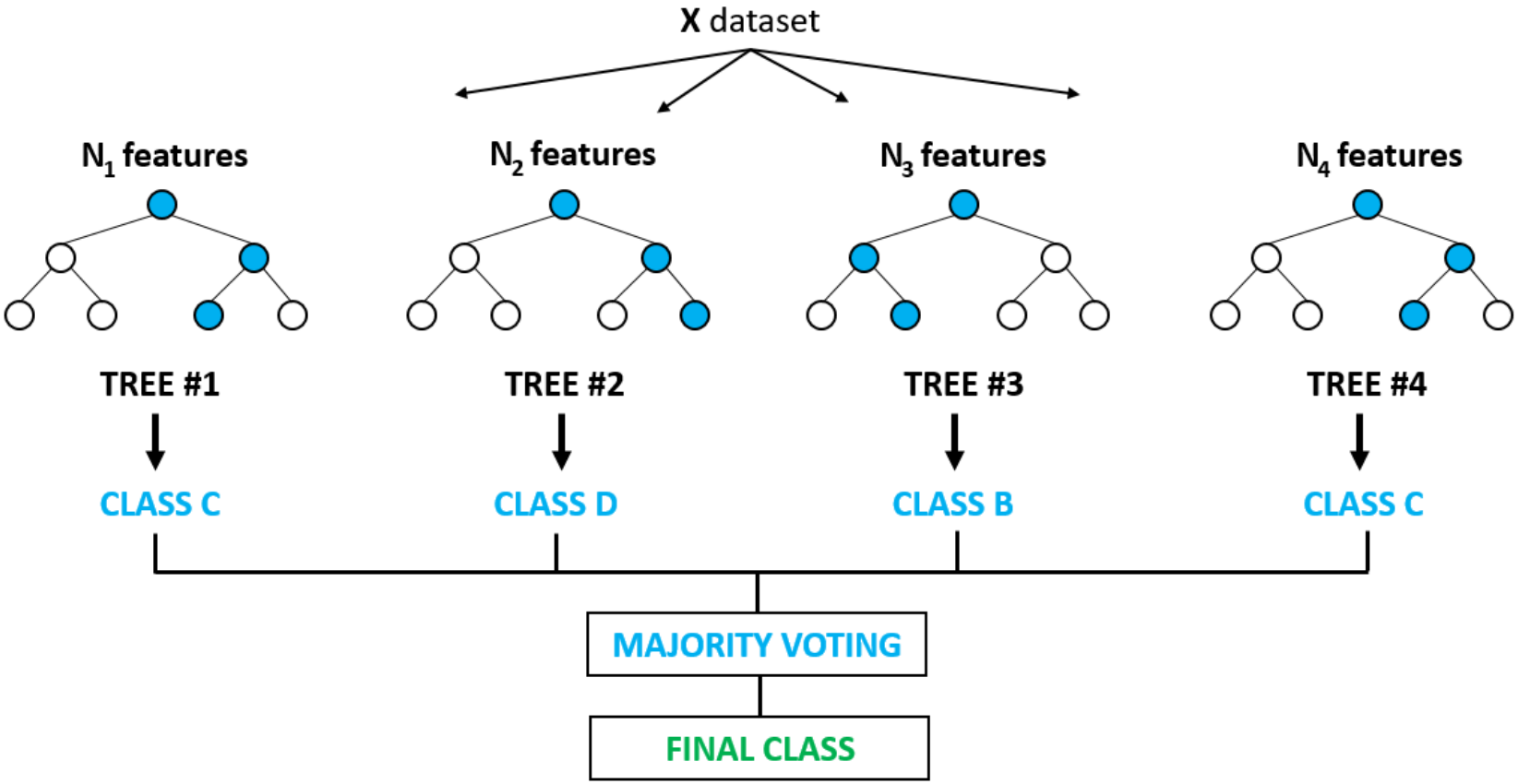
¿Cómo funciona el modelo?

Un paso atrás... ¿Cómo funciona un *random forest*? Suponemos que han visto todos alguna vez una de estas cosas



¿Cómo funciona el modelo?

Bien, pues un *random forest* no es más que la combinación de muchos árboles de decisión



Regresando a las fosas en México

Datos generales:

- **2 mil 458 municipios.**
- Hasta ahora tenemos datos de fosas 2009 a 2018. Las observaciones vienen de prensa escrita y de fiscalías estatales.
- 56 variables predictoras: sociodemográficos, características físicas y topográficas, infraestructura, y algunas cosas de violencia, pero tratamos en realidad de evitarlas.
- Tenemos en total entonces estimaciones realizadas con 20 modelos: **10** años * **2** fuentes.

¿Cómo funciona el modelo?

El primer paso, clasificamos los municipios en tres categorías conforme a la variable dependiente (de forma manual):

1 = municipios que tuvieron observaciones de fosas por alguna de las fuentes entre 2009 y 2018.

0 = municipios donde -dado un análisis de contexto- consideramos poco posible que existan fosas clandestinas.

-1 = municipios donde no conocemos la respuesta. **Estos son los que nos interesan**.

¿Cómo funciona el modelo?

El segundo paso es imputar nuestros valores perdidos o nuestros datos incompletos.

Utilizamos un método de imputación conocido como *MissForest* (Stekhoven y Bühlmann, 2012).

El *MissForest* es muy útil cuando tenemos bases de datos que están mezcladas con variables continuas y variables categóricas.

¿Cómo funciona el modelo?

Para el tercer paso, entrenamos los modelos con 2/3 de los municipios y después evaluamos su rendimiento con el resto de la información.

Entrenamos = Estimamos

El modelo *"aprende"* las características de los municipios con fosas y los que nosotros clasificamos como "sin" fosas.

Dadas esas características, los árboles votan.

OJO: vamos a identificar sólo fosas con las características de las fosas que *ya han sido observadas antes*.

¿Y los sesgos?

WARNING: usar datos NO ES IGUAL (!=) a ser objetivos

Los procesos de generación de datos importan mucho.

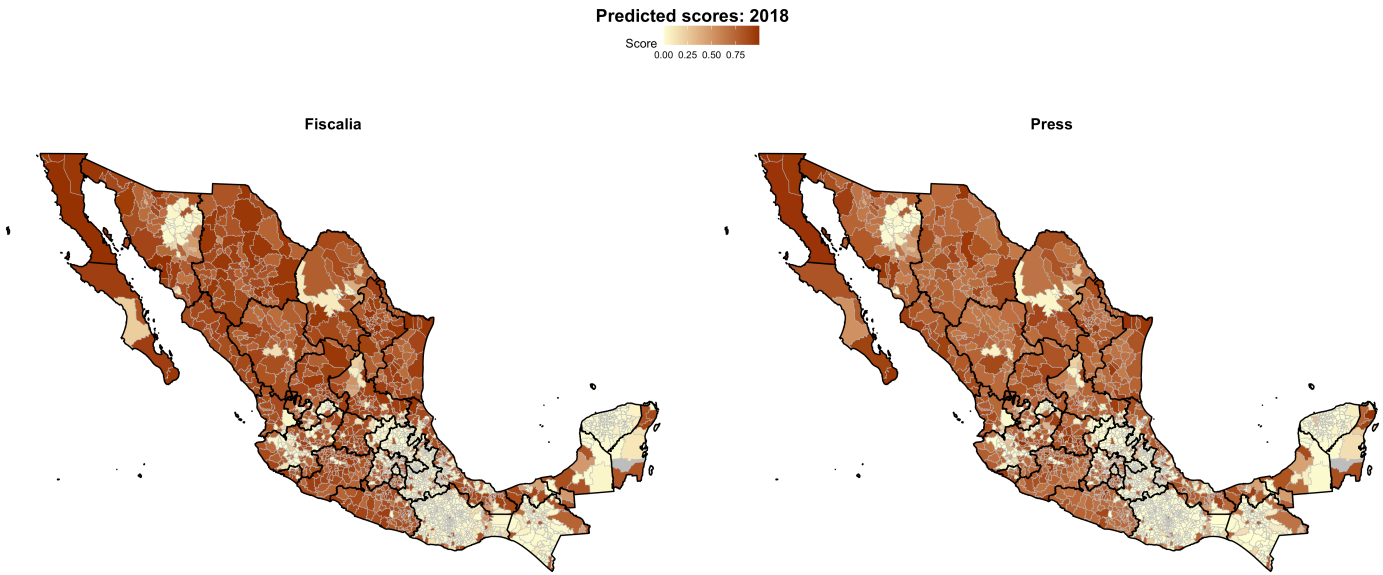
¿Cómo se ven los resultados?

Por eso usamos dos fuentes: fiscalías y prensa.

Cada una tiene distintos sesgos potenciales, pero en lo general coinciden en las estimaciones.

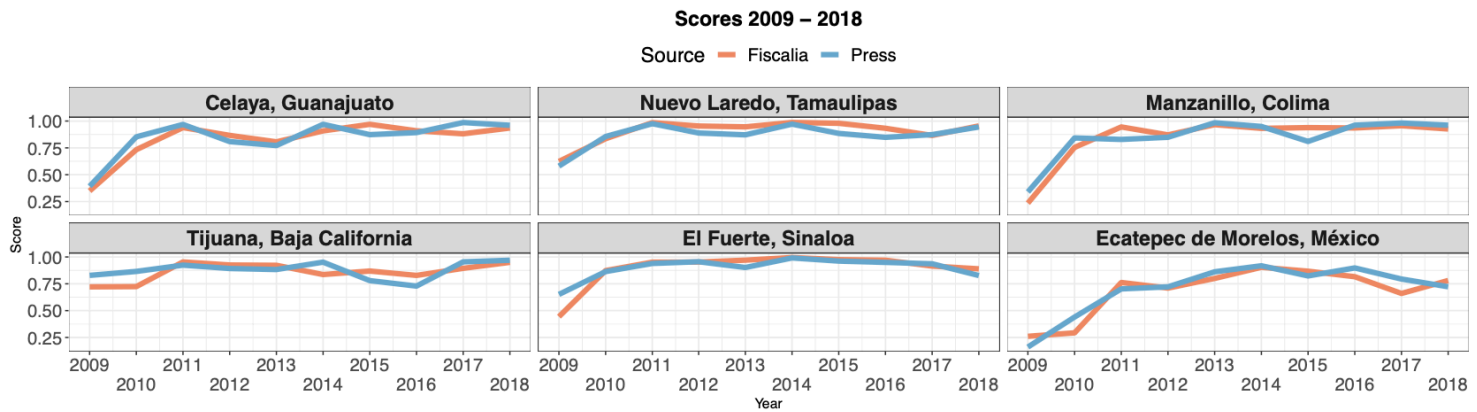
¿Cómo se ven los resultados?

Podemos ver también que la distribución geográfica de probabilidades es similar para ambas.



¿Y son consistentes en el tiempo?

De lo que hemos visto hasta ahora, en realidad son bastante consistentes.



¿Y son consistentes con la realidad?

Hallan 13 bolsas con restos humanos en Celaya

Suman 26 bultos similares encontrados en los últimos dos días en tres municipios

NOTICIAS

Localizan fosa clandestina con cadáveres en Tijuana, Baja California

La fosa contaba con una escalera, y estaba cubierta entre maleza, lonas, y tierra

24 DE NOVIEMBRE DE 2019 | 09:30 PM CST

La señora Angélica Ramírez agradeció el apoyo de las diferentes brigadas y colectivos que están en busca de personas desaparecidas. (Omar Martínez /Cuartoscuro.Com)

¿Cómo se han utilizado nuestros resultados?

Buscamos que sirva para la implementación de programas contemplados en la *Ley General en Materia de Desaparición Forzada y Desaparición Cometida por Particulares*:

- La ley contempla la creación de un Sistema Nacional de Búsqueda
- Asistencia para la implementación del Programa Nacional de Exhumaciones

Nuestros resultados han sido utilizados por colectivas de familiares con personas desaparecidas en estados como Chihuahua, Nuevo León y Guanajuato.

Se han preparado reportes para la Fiscalía General del Estado de Jalisco y para la Fiscalía General del Estado de Veracruz, así como al Equipo Argentino de Antropología Forense (EAAF).

Siguientes pasos

- Obtener más datos de fosas y más datos para nuestros predictores.
- Trabajar en los *ceros* del modelo.
- Mejorar la desagregación geográfica de las predicciones del modelo.

Materiales para consultar

- Data Cívica, PDH Ibero, HRDAG: *Predecir la existencia de fosas clandestinas en municipios mexicanos: Una primera aproximación estadística*: https://datacivica.org/assets/pdf/Fosas_web.pdf
- SCIENCE. *Mapping Mexico´s Hidden graves*: https://www.sciencemag.org/news/2017/06/mapping-mexico-s-hidden-graves
- La Lista. *Ciencia de datos para trazar un mapa de la crueldad a la mexicana (I)*: https://la-lista.com/seguridad/2021/01/18/ciencia-de-datos-para-trazar-un-mapa-de-la-crueldad-a-la-mexicana-i
- PDH Ibero, Data Cívica (et.al). *Informe sobre la situación de fosas clandestinas en Guanajuato*: https://fosas-guanajuato.datacivica.org/#intro

¡Muchas gracias por su atención!

Presentación disponible en:
https://github.com/JorgeRuRe/presentacion-fosas-iberoSD.git

- Jorge Ruiz: @jorgerure / jorge.ruiz@ibero.mx