Análisis de distribución, riesgo y proyecciones de reptiles y anfibios en México

**Autor: Erik Manuel Barajas Torres**

**Fecha de entrega: 13/05/2025**

**Diplomado en Ciencia de Datos – La dirección General de Cómputo y Tecnologías de la Universidad Autónoma del Estado de México**

**Emisión: 10**

# Introducción

México es uno de los países con mayor biodiversidad en el mundo, y dentro de su riqueza biológica destacan los reptiles y anfibios, conocidos colectivamente como herpetofauna. Estos grupos cumplen un papel fundamental en los ecosistemas: controlan poblaciones de insectos, participan en redes tróficas y actúan como indicadores ecológicos sensibles al deterioro ambiental.

Sin embargo, enfrentan diversas amenazas que han provocado la disminución de muchas de sus poblaciones. Entre las principales causas están la pérdida y fragmentación de hábitat, la contaminación del agua, la urbanización acelerada, el cambio climático y el comercio ilegal de especies. Estas problemáticas no solo afectan a nivel local, sino que también comprometen la resiliencia ecológica a gran escala.  
En este contexto, el presente estudio busca aplicar herramientas de ciencia de datos para analizar los patrones históricos y actuales de distribución de estas especies, identificar factores de riesgo, y generar modelos que permitan anticipar tendencias futuras. De este modo, se pretende contribuir con información útil para fortalecer estrategias de conservación.

# Resumen ejecutivo

Este trabajo se desarrolló a partir de una base de datos oficial de CONABIO, con más de 150 mil registros de reptiles y anfibios observados en distintas regiones de México. En una primera etapa, se llevó a cabo una limpieza de datos, filtrando años poco confiables y eliminando columnas con altos porcentajes de valores nulos o irrelevantes para el análisis. Posteriormente, se codificaron variables categóricas importantes como el ambiente, el estado de conservación (NOM-059 e IUCN), el endemismo, y el estado geográfico de registro.  
Se aplicaron técnicas de reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) y modelos de agrupamiento como K-Means y DBSCAN para detectar patrones de similitud entre especies. Además, se utilizó la herramienta Prophet para generar modelos de series de tiempo, con los cuales se proyectaron tendencias futuras para 40 especies relevantes.  
Finalmente, se entrenaron modelos supervisados de clasificación y regresión para analizar la capacidad predictiva de las variables codificadas, evaluando su rendimiento mediante métricas como RMSE, MAE y F1-score.

Los resultados permitieron identificar estados prioritarios en materia de conservación, especies con proyecciones preocupantes, y validaron hipótesis ecológicas clave, todo ello presentado también en un dashboard interactivo para consulta dinámica. Además, se recuperó información de la columna CITES, identificando especies reguladas por comercio internacional. Algunas de ellas coinciden con las especies en declive, fortaleciendo su carácter prioritario en conservación.

# Problema / Caso de negocio

En los últimos años, se ha observado una reducción sostenida en los registros de diversas especies de herpetofauna en México. Esta disminución puede deberse tanto a una real pérdida poblacional como a una baja en el esfuerzo de muestreo, lo cual genera incertidumbre sobre el verdadero estado de conservación de muchas especies.  
Ante este panorama, las autoridades ambientales, académicos y conservacionistas necesitan herramientas que no solo muestren los datos históricos, sino que también permitan detectar patrones, anticipar riesgos y focalizar acciones.  
El problema no se queda solo en un tema ecológico, es también de información. Existen datos, pero no están plenamente aprovechados mediante técnicas modernas de análisis que permitan tomar decisiones estratégicas y basadas en evidencia. Este proyecto busca investigar esas tendencias y que se puedan profundizar más sobre estos temas a futuro, incluso se pueda promover las buenas prácticas con información sustentada, de la cría en cautiverio.

# Propósito del estudio

El objetivo principal del estudio es aplicar técnicas de ciencia de datos al análisis de registros biológicos para:

• Detectar especies en riesgo a partir de su comportamiento histórico,

• Proyectar su evolución a corto y mediano plazo,

• Identificar regiones geográficas donde se concentran especies amenazadas,

# • Y generar información útil para orientar esfuerzos de conservación, educación ambiental y políticas públicas.

# Todo esto desde un enfoque técnico que combina análisis exploratorio, aprendizaje automático y visualización interactiva.

# Descripción del proyecto

## Hipótesis

Se plantearon las siguientes hipótesis de trabajo:

* **H1**: Las especies endémicas presentan una mayor tendencia a la disminución de registros que las no endémicas, lo cual puede relacionarse con su limitada distribución geográfica.
* **H2**: Las especies que habitan ambientes dulceacuícolas muestran mayor vulnerabilidad, dado el alto grado de impacto que sufren los cuerpos de agua en México.
* **H3**: Las especies clasificadas en la NOM-059 o en alguna categoría de riesgo según la IUCN tienden a tener comportamientos descendentes en su número de registros.

## Flujo de trabajo

## Mapeo del sistema

El sistema se construyó a partir de registros públicos proporcionados por CONABIO, los cuales fueron tratados y analizados dentro del entorno en la nube Google Colab, que permitió trabajar con Python en un ambiente interactivo y compatible con visualización y modelado de datos.

Se utilizaron las siguientes herramientas y bibliotecas:

**Pandas**: para la carga, limpieza, filtrado, agrupamiento y transformación de los datos tabulares.

**Matplotlib y plotly**: para la visualización de series de tiempo, mapas de calor, análisis exploratorio y gráficas interactivas.

**Scikit-learn**: para aplicar modelos de:

**Reducción de dimensionalidad (PCA)**: permitió simplificar los datos codificados y facilitar análisis como clustering.

**Clustering con KMeans y DBSCAN:** se usaron para agrupar especies por características similares, donde DBSCAN ayudó a detectar grupos no lineales y especies "ruido".

**Modelos supervisados como RandomForestRegressor, RandomForestClassifier, LinearRegression y LogisticRegression:** para predecir tanto valores numéricos (regresión) como categorías (clasificación de especies en riesgo).

**Imbalanced-learn (SMOTE)**: se utilizó para resolver el problema del desbalanceo de clases en los modelos de clasificación. Es decir, ayudó a que el modelo aprendiera mejor a reconocer las especies en peligro, que eran la clase minoritaria.

**Prophet (Meta)**: para construir modelos de series de tiempo robustos que permitieran proyectar el comportamiento de especies con tendencia descendente o ascendente hasta el año 2030.

**Numpy y scipy:** para realizar operaciones matemáticas, estadísticas y de álgebra lineal.

**Seaborn y yellowbrick:** para facilitar análisis visual de modelos y gráficas de evaluación.

Todo el flujo fue documentado paso a paso, generando outputs intermedios que permitieron validar la consistencia del análisis y tomar decisiones informadas sobre qué técnicas aplicar en cada etapa.

## Definición de métricas

Durante el desarrollo del proyecto, se utilizaron diversas métricas para evaluar el rendimiento de los modelos aplicados, dependiendo del tipo de técnica empleada.

Para regresión:

**RMSE (Root Mean Squared Error)**: mide el error cuadrático medio entre los valores reales y los predichos. Es sensible a errores grandes y útil para evaluar precisión en modelos de predicción de registros.

**MAE (Mean Absolute Error)**: calcula el promedio del error absoluto. Es más robusto frente a outliers y complementa el RMSE para una visión más balanceada del desempeño.

Para clasificación supervisada:

**Accuracy**: proporción de predicciones correctas sobre el total.

**Recall (Sensibilidad)**: indica la capacidad del modelo para identificar correctamente las especies en riesgo (clase positiva).

**Precision**: mide cuántas de las especies predichas como en riesgo realmente lo están.

**F1-score**: combina precisión y recall en una sola métrica. Es especialmente útil cuando las clases están desbalanceadas, como en este caso.

Para clustering no supervisado:

**Silhouette Score**: mide qué tan bien definidas están las agrupaciones (clusters). Toma valores entre -1 y 1:

### Valores cercanos a 1 indican que los datos están bien agrupados.

### Valores cercanos a 0 indican que los datos están en el borde entre clusters.

### Valores negativos indican que los datos están mal asignados.

### Se utilizó para evaluar la calidad de los agrupamientos generados por KMeans y DBSCAN, ayudando a seleccionar el número óptimo de clusters y validar si las especies se agrupan de forma natural según sus características.

### Métodos y modelos

* **PCA**: reducción de dimensiones para facilitar clustering.
* **K-Means y DBSCAN**: para agrupar especies según similitud.
* **Prophet**: modelo de series de tiempo para proyecciones por especie.
* **Random Forest y regresión lineal**: para estimar el número de registros.
* **SMOTE**: técnica de sobremuestreo para balancear clases en modelos supervisados.

Evaluación de modelos

La evaluación de los modelos se llevó a cabo mediante una combinación de métricas cuantitativas, visualizaciones gráficas y validación cruzada. Para los modelos de regresión y clasificación, se graficaron comparaciones de predicción vs. valores reales, se calcularon métricas como RMSE, MAE, F1-score, accuracy y se aplicaron técnicas como SMOTE para mejorar el desempeño en clases desbalanceadas.

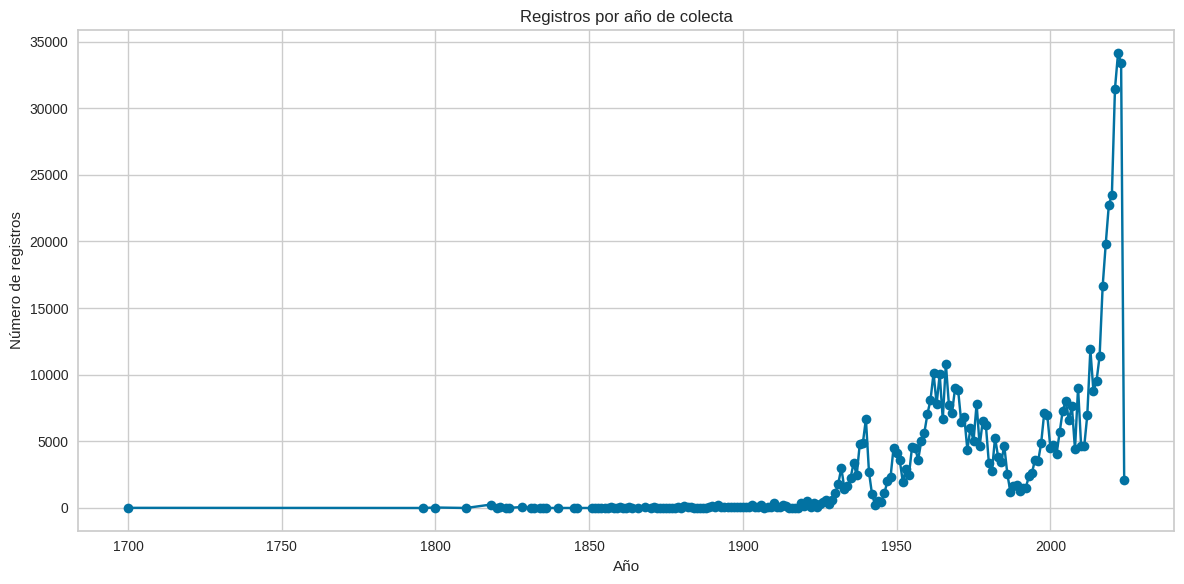
En el caso de clustering no supervisado, se utilizó el Silhouette Score como métrica principal para evaluar la cohesión y separación de los grupos formados, ayudando a determinar el número óptimo de clusters (k en KMeans) y a validar la segmentación lograda por DBSCAN.

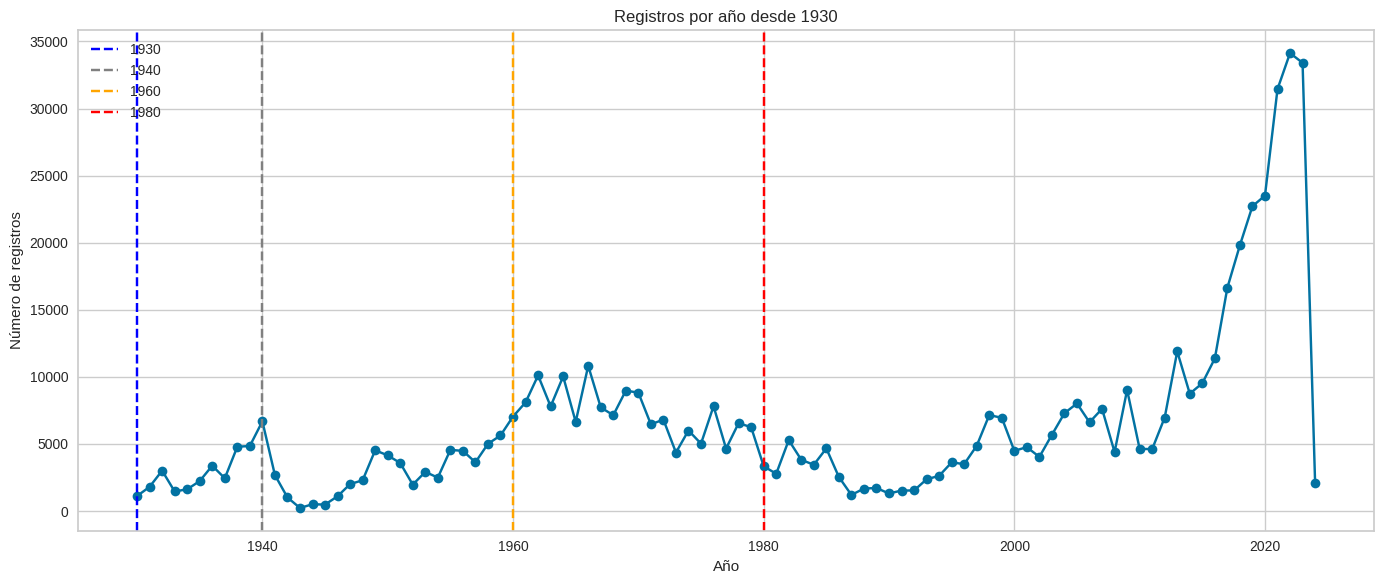
Para las series de tiempo, el modelo Prophet fue evaluado visualmente a través de sus gráficas de ajuste y proyección, observando la tendencia histórica y la confianza de las predicciones hasta el año 2030. Estas gráficas permitieron identificar especies con descenso sostenido y alta incertidumbre futura.

En conjunto, estas evaluaciones aseguraron que los modelos utilizados no solo funcionaran matemáticamente, sino que también aportaran valor interpretativo al análisis ecológico del proyecto.

## Análisis

El análisis comenzó con una exploración temporal de los registros, donde se observó un crecimiento acumulado en el número de registros desde 1950, con un pico en años recientes debido a un mayor esfuerzo de muestreo y digitalización de datos. Sin embargo, al analizar el comportamiento específico por especie, se identificaron 40 especies con tendencias extremas: 30 con clara disminución sostenida en el tiempo y 10 con aumentos significativos.





A través de la segmentación por grupos biológicos, se observó que los anfibios presentan una tasa de disminución mayor en comparación con los reptiles, lo cual se relaciona con su alta sensibilidad a cambios en humedad, temperatura y contaminación, factores intensamente presentes en los ecosistemas acuáticos donde habitan.

También se realizó un análisis geográfico utilizando las columnas de estadomapa, lo que permitió identificar los estados con mayor concentración de especies en riesgo. Entre ellos destacaron Chiapas, Oaxaca, Veracruz y Guerrero, estados que además presentan alta biodiversidad y múltiples ecosistemas. Se construyeron indicadores como el número de especies NOM-059, IUCN, y aquellas con proyección negativa según Prophet, lo cual permitió crear un índice de prioridad territorial.

Mediante técnicas de clustering (KMeans y DBSCAN), se agruparon especies con características similares, revelando patrones de distribución no evidentes en el análisis tradicional. Por ejemplo, algunas especies endémicas y de agua dulce formaron grupos con tendencias similares de riesgo.

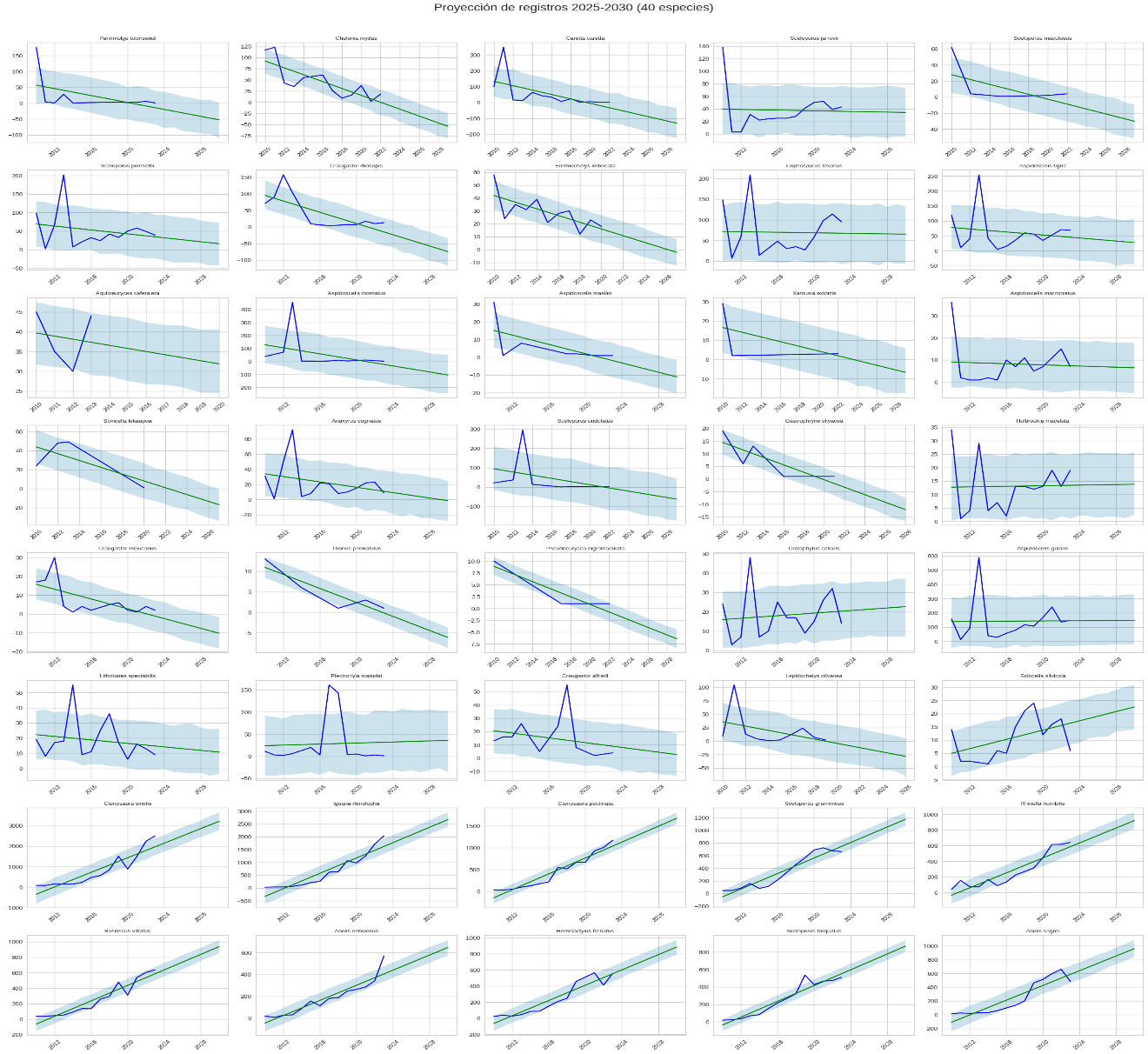
Adicionalmente, se exploró el estatus CITES de ciertas especies, observándose coincidencias entre aquellas en peligro de extinción y su presencia en los apéndices I y II, lo cual refuerza su vulnerabilidad frente al comercio internacional.

## Finalmente, se llevaron a cabo pruebas de hipótesis relacionadas con endemismo, ambiente y estatus de conservación. Estas pruebas demostraron que las especies endémicas y acuáticas, así como aquellas presentes en listas de riesgo, tienden a mostrar una pendiente más negativa en su número de registros a lo largo del tiempo, validando las hipótesis planteadas al inicio del proyecto.

# Resultados

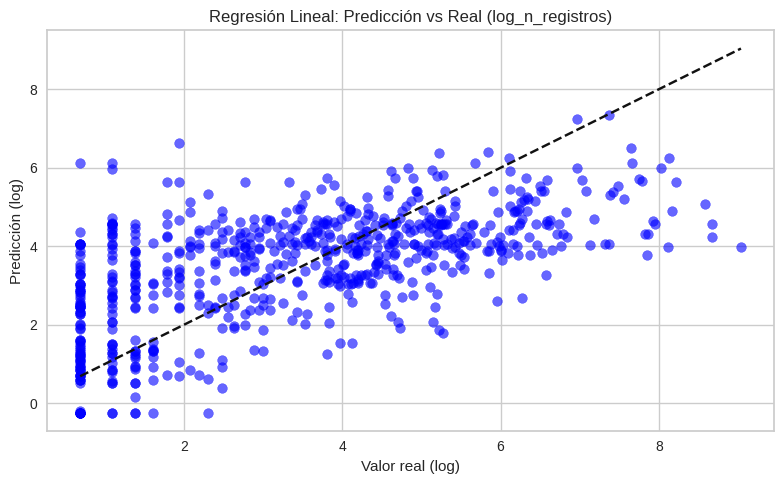
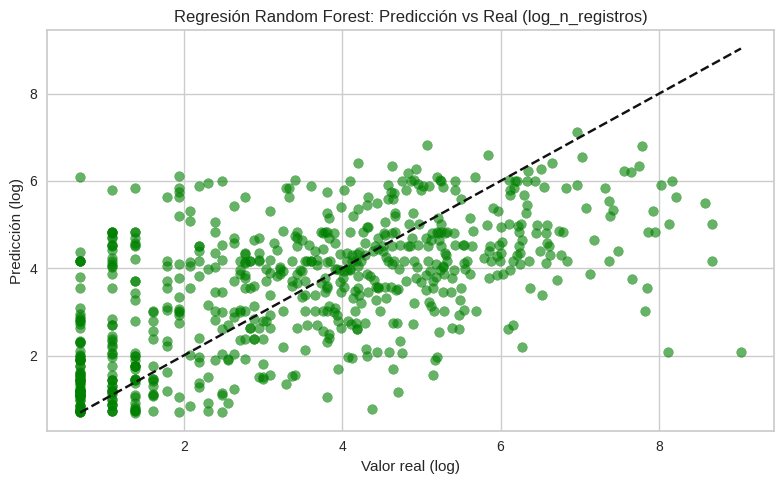
A lo largo del análisis se obtuvieron hallazgos clave que permiten responder al propósito del proyecto:

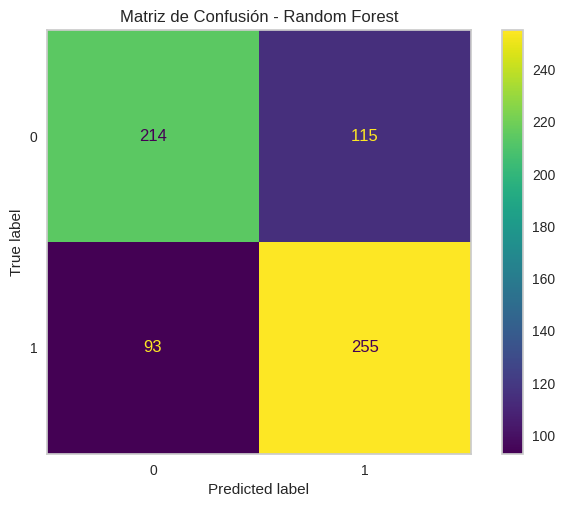
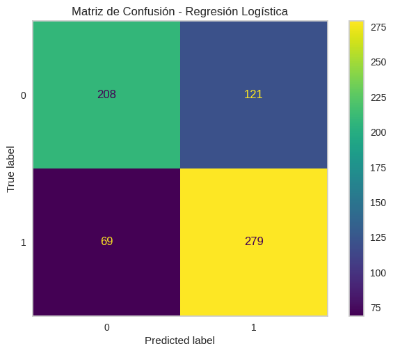
Se identificaron 40 especies con tendencias extremas en su número de registros entre 2010 y 2023. De ellas, 30 presentan una disminución sostenida y 10 un incremento notable. Estas especies fueron seleccionadas como base para las proyecciones a futuro mediante modelos de series de tiempo (Prophet).



Las proyecciones con Prophet mostraron que varias especies tienden a continuar su descenso en los próximos años, lo cual representa un potencial escenario de crisis ecológica si no se implementan medidas de conservación. Estas predicciones, validadas visual y analíticamente, refuerzan la utilidad del enfoque prospectivo en conservación.

Los modelos supervisados (clasificación con Random Forest y Logistic Regression) obtuvieron F1-scores aceptables, especialmente tras la aplicación de SMOTE, que permitió balancear las clases y mejorar la detección de especies en categorías de riesgo como NOM-059 e IUCN. Esto valida el uso de machine learning como herramienta de apoyo para predicción de riesgo.

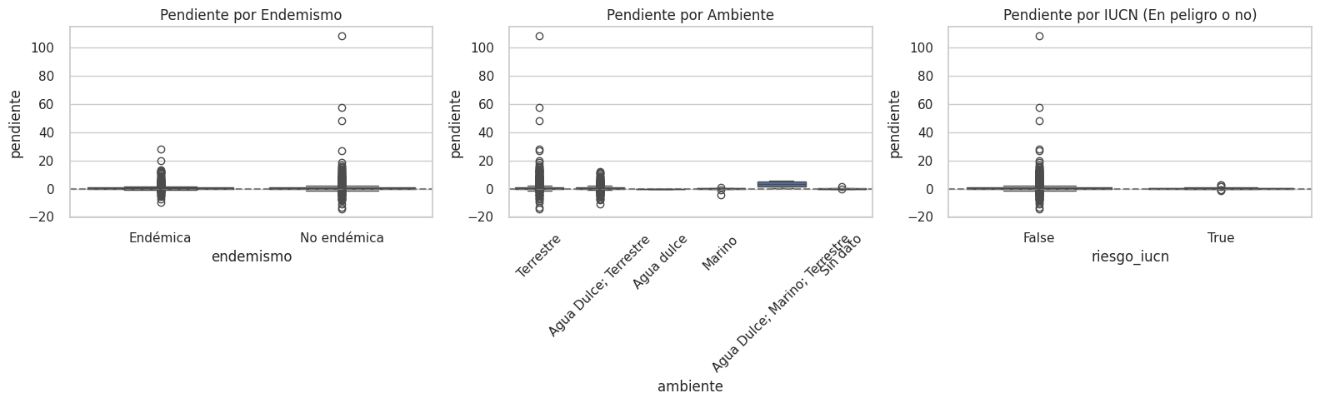
 

Se construyó un índice de prioridad territorial, cruzando el número de especies amenazadas (según NOM-059, IUCN y proyecciones) por estado y municipio. Este análisis permitió ubicar a Chiapas, Oaxaca, Veracruz y Guerrero como zonas críticas de atención ecológica.

Las pruebas de hipótesis reforzaron varios patrones esperados:

* **H1 –** Endémicas vs. No endémicas: Aunque las especies endémicas mostraron una mayor variabilidad y pendientes más negativas, la diferencia no fue estadísticamente significativa (t = -1.38, p = 0.1689). Esto sugiere una tendencia, pero no suficiente evidencia concluyente.
* **H2 –** Agua Dulce vs. otros ambientes: Se encontró una diferencia significativa (t = -2.30, p = 0.0213), lo que confirma que las especies acuáticas están disminuyendo más que las terrestres o marinas. Esto valida la hipótesis H2 con evidencia estadística sólida.
* **H3 –** Especies en NOM-059/IUCN en peligro vs. otras: En el caso de NOM-059, el valor fue casi significativo (t = -1.84, p = 0.0656), y para IUCN no fue concluyente (t = -1.09, p = 0.2756). Aun así, ambas muestran una tendencia hacia pendientes más negativas, lo que apoya parcialmente la hipótesis.



Además, el análisis geográfico y ecológico permitió contextualizar los resultados:  
Estados como Chiapas, Oaxaca y Veracruz concentran especies en riesgo debido a factores como; alta biodiversidad y endemismo regional, cambio climático, que altera los hábitats sensibles de anfibios, deforestación y cambio de uso de suelo, contaminación de cuerpos de agua, afectando a especies acuáticas y en algunos casos, el comercio ilegal de especies o su captura no regulada.

Aunque la columna cites fue excluida del análisis principal por su alto porcentaje de datos nulos (86%), se realizó un filtrado final para identificar las especies que sí cuentan con clasificación en alguno de los Apéndices de CITES. Estas clasificaciones regulan el comercio internacional de especies amenazadas.

Se identificaron especies incluidas en los Apéndices I, II y III, principalmente tortugas y serpientes. Algunas de ellas también coinciden con las especies proyectadas como en declive o incluidas en NOM-059/IUCN, lo que refuerza su estatus de vulnerabilidad.

En resumen, los modelos analíticos aplicados, las proyecciones de tendencias y las evaluaciones espaciales permitieron no solo identificar especies en riesgo, sino también priorizar zonas y validar hipótesis ecológicas, fortaleciendo la interpretación y utilidad del estudio en el contexto de conservación.

# Implementación (opcional)

Se desarrolló un dashboard interactivo utilizando las bibliotecas Panel y Plotly, implementado en el entorno de Google Colab. Este dashboard permite explorar de manera dinámica los resultados obtenidos a lo largo del proyecto.

El panel incluye:

Gráficas interactivas de series de tiempo por especie (generadas con Prophet), con opción de seleccionar especies específicas.

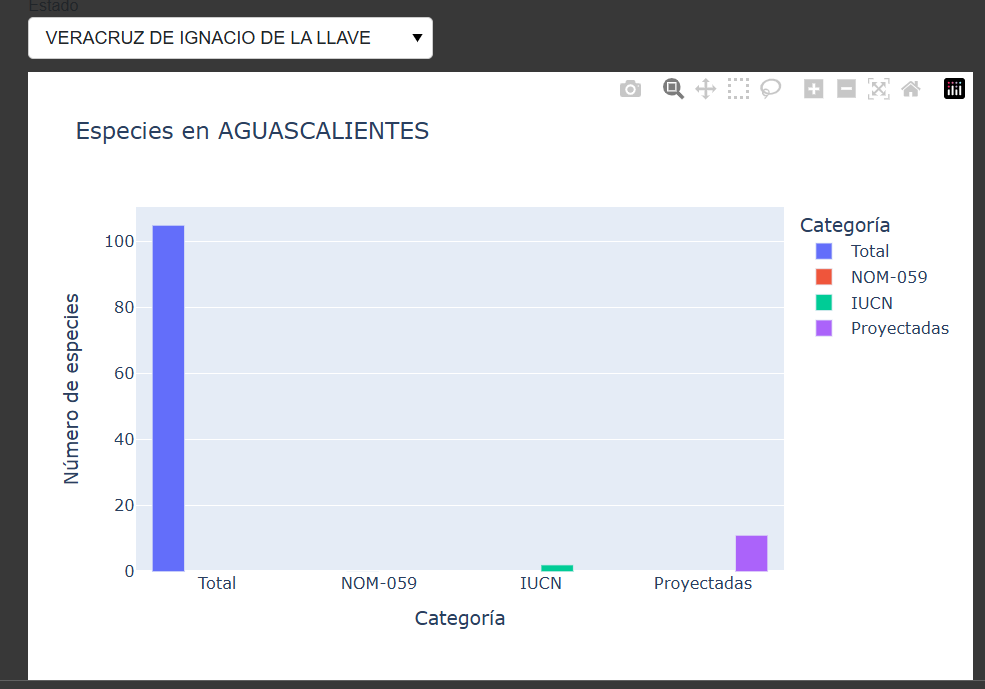
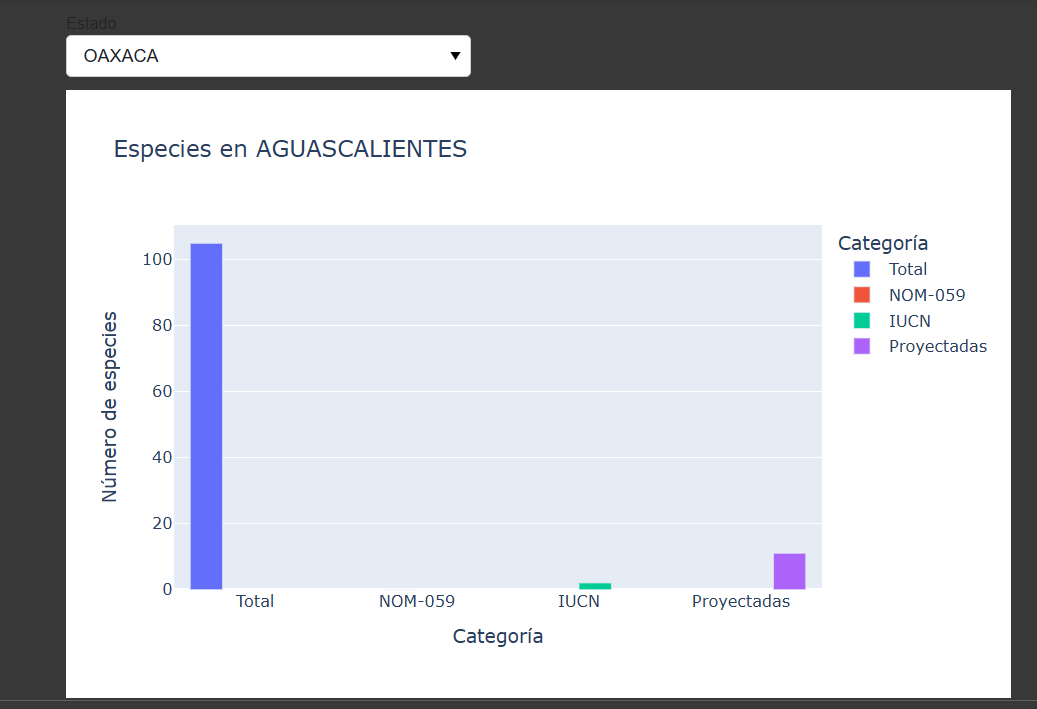
Visualizaciones por estado o municipio, que muestran el número de especies en riesgo según distintas fuentes (NOM-059, IUCN o predicciones).

Filtros dinámicos para analizar únicamente especies endémicas, acuáticas, o clasificadas en categorías de amenaza.

Mapas y tablas comparativas con indicadores calculados por región, como el índice de prioridad ecológica.

Además, la implementación en Google Colab permite portabilidad y acceso sin instalación, lo que vuelve al dashboard ideal para presentaciones académicas, análisis participativos o difusión en instituciones ambientales.

"Dashboard interactivo desarrollado en Panel para consulta por especie y estado."



# Conclusiones

Este proyecto demuestra cómo el uso de ciencia de datos puede ser una herramienta poderosa para entender y anticipar problemas ambientales, en este caso enfocados en reptiles y anfibios de México. Al analizar miles de registros históricos, logramos identificar especies con tendencias preocupantes, patrones regionales de riesgo y posibles escenarios futuros si no se actúa a tiempo.

Uno de los hallazgos más importantes es que las especies que habitan en ambientes de agua dulce son las que muestran una mayor tendencia a disminuir. Esto puede explicarse por varios factores: los cuerpos de agua dulce suelen estar más expuestos a la contaminación, la urbanización, el desvío de ríos o la sequía por cambio climático. Además, estos ecosistemas suelen ser más pequeños y fragmentados, por lo que cualquier alteración impacta rápidamente a las especies que dependen de ellos.

También se encontró que los anfibios presentan una mayor tasa de disminución que los reptiles, lo cual coincide con estudios internacionales. Los anfibios, al tener piel permeable y depender del agua para reproducirse, son particularmente sensibles a los cambios en temperatura, humedad y calidad del agua.

Gracias al uso de modelos como Prophet, pudimos proyectar que muchas de las especies ya en declive podrían seguir disminuyendo hacia 2030, lo cual sirve como una señal de alerta temprana. Por otro lado, los modelos supervisados nos permitieron identificar con bastante precisión cuáles especies ya están clasificadas en riesgo por instituciones como la NOM-059 o la IUCN, validando los patrones observados en los datos.

# Además, al analizar la distribución geográfica, se detectaron estados como Chiapas, Oaxaca, Veracruz y Guerrero como regiones con alta concentración de especies en riesgo. Esto puede deberse tanto a su gran biodiversidad como a los problemas ambientales que enfrentan, como la deforestación, el cambio climático, la fragmentación del hábitat y el desarrollo urbano no regulado.

La inclusión en CITES refuerza la necesidad de monitoreo y control sobre estas especies, no solo por su estado poblacional, sino también por los riesgos asociados al tráfico ilegal de fauna.

Del total de especies clasificadas en los apéndices de la Convención CITES, cinco también coinciden con las especies que presentan una tendencia de disminución poblacional entre 2010 y 2023. Todas ellas pertenecen al Apéndice I, lo que indica un nivel máximo de protección debido a su alto riesgo de extinción por comercio internacional.

Entre las especies identificadas se encuentran tortugas marinas emblemáticas como Caretta caretta (tortuga caguama), Chelonia mydas (tortuga verde), Eretmochelys imbricata (tortuga carey) y Lepidochelys olivacea (tortuga golfina), todas reconocidas tanto por su vulnerabilidad ecológica como por su importancia cultural y económica.

Lo relevante es que además de estar en CITES, todas están también incluidas en la NOM-059 como "En peligro de extinción", y en la lista IUCN con clasificaciones de "Vulnerable", "En peligro" o "En peligro crítico". Esta triple coincidencia valida que existe una urgencia real de atención y conservación, reforzando que su declive observado no es un dato aislado, sino una confirmación multifuente del riesgo que enfrentan.

Concluyendo, el análisis permitió:

Detectar especies prioritarias para la conservación.

Visualizar tendencias de riesgo hacia el futuro.

Construir herramientas interactivas para comunicar resultados.

Proponer zonas geográficas clave donde se deben enfocar los esfuerzos.

Este trabajo no solo sirve como ejercicio académico, sino como una propuesta real de cómo los datos pueden ayudar a tomar decisiones más informadas, y por qué es urgente actuar antes de que las proyecciones se conviertan en realidades irreversibles.

# Referencias (formato APA)

CONABIO. (2023). Sistema Nacional de Información sobre Biodiversidad. https://www.conabio.gob.mx  
Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45.  
Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.  
McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

IUCN. (2023). *The IUCN Red List of Threatened Species*. <https://www.iucnredlist.org/>

Semarnat. (2019). *NOM-059-SEMARNAT-2010: Especies nativas de México de flora y fauna silvestres – categorías de riesgo y especificaciones para su inclusión, exclusión o cambio*. Diario Oficial de la Federación.

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences, 114*(30), E6089–E6096. https://doi.org/10.1073/pnas.1704949114