Proyecto de Regresión - Salvemos a las Abejas
Paulina Adams Arteche

Introducción

La desaparición masiva de abejas ha generado preocupación mundial, dado su rol clave en la polinización y el equilibrio ecológico. Diversas amenazas como plagas, pesticidas, enfermedades y otros factores han sido relacionadas con esta pérdida. En este proyecto se utilizará la plataforma KNIME para analizar un dataset relacionado con colonias de abejas en Estados Unidos y construir un modelo predictivo que estime cuántas colonias se pierden dependiendo de distintas amenazas.

Descripción del problema

Pregunta: ¿Cuántas colonias de abejas se pierden en función de amenazas como varroa mites, pesticidas, enfermedades u otras causas?

Objetivo: Predecir el número de colonias perdidas (`lost_colonies`) usando un modelo de regresión basado en las variables causales.

Tipo de aprendizaje: Aprendizaje supervisado (regresión).

Variable objetivo: `lost_colonies`

Preparación de los datos

1. Lectura de datos

Nodo: `CSV Reader` Carga el archivo con los datos brutos desde un .csv.

2. Exploración preliminar

Nodo: `Statistics` Permite revisar valores máximos, mínimos, medias, nulos y distribución de datos.

3. Selección de columnas relevantes

Nodo: `Column Filter` Se eliminaron columnas no necesarias como `state`, `year` y `quarter` para enfocarse solo en las variables numéricas que pueden influir directamente en la pérdida de colonias.

Tratamiento de valores faltantes

Nodo: `Missing Value`

- Columnas numéricas: se usó la mediana para imputar los valores faltantes.

5. Verificación de la limpieza

Nodo: `Table View` permite visualizar si ya no existen valores nulos y revisar la forma final de la tabla.

6. Codificación (opcional)

Nodo: `One to Many` Transforma variables categóricas (si las hubiera) en variables binarias. En este caso, el uso fue opcional pero se incluyó por si se trabajaba con algún campo no numérico.

Selección y entrenamiento del modelo

7. División de datos

Nodo: `Table Partitioner`

- Divide los datos en entrenamiento (70%) y prueba (30%).
- Se aseguró aleatoriedad con semilla fija para reproducibilidad.

8. Entrenamiento del modelo

Nodo: `Linear Regression Learner`

- Variable objetivo: `lost_colonies`
- Variables predictoras:
  - `varroa_mites`
  - `pesticides`
  - `diseases`
  - `other_pests_and_parasites`
  - `unknown`
  - `other`

9. Predicción sobre datos de prueba

Nodo: `Regression Predictor` Aplica el modelo entrenado a la partición de prueba para generar predicciones de `lost_colonies`.

Análisis de resultados

10. Evaluación del modelo

Nodo: `Numeric Scorer` Proporciona métricas clave de regresión:

Métrica	Valor
R^2 (coeficiente de determinación)	**0.964**
Root Mean Squared Error (RMSE)	10,187.7
Error medio (mean error)	-676.46

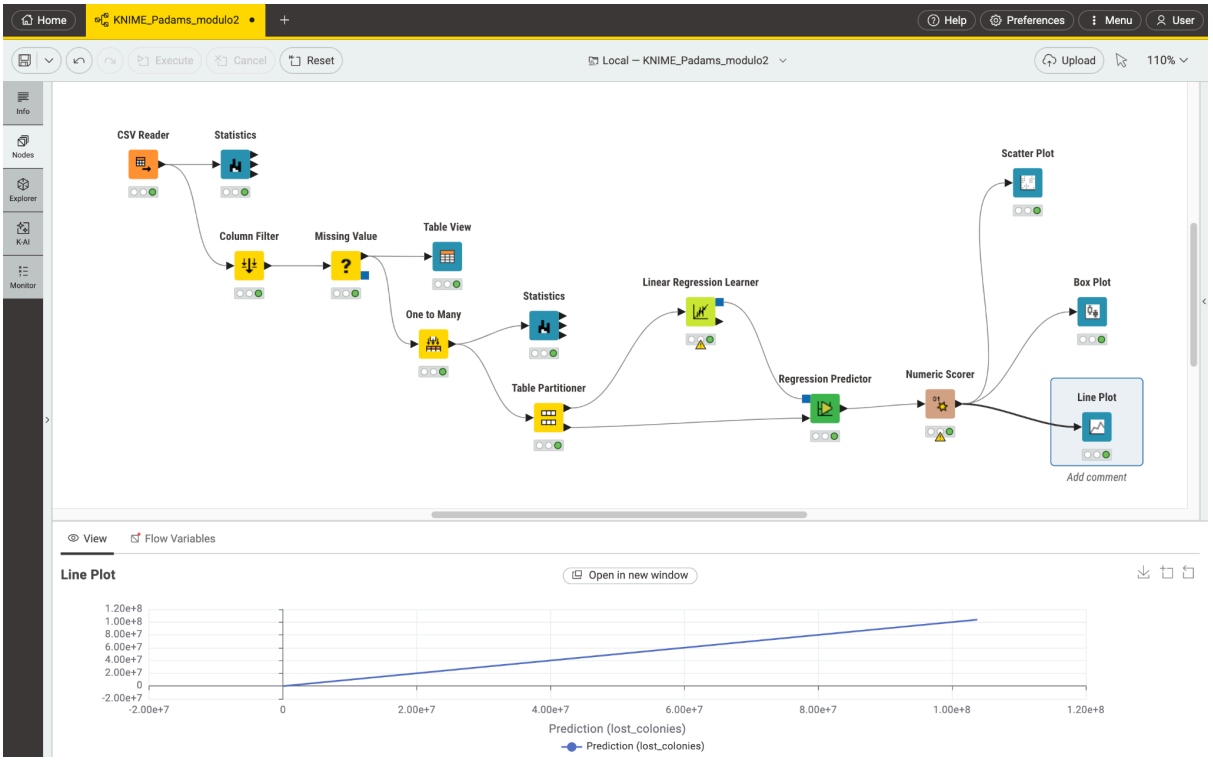
Interpretación: El modelo tiene un poder explicativo del 96.4%, lo que indica que captura muy bien la variabilidad en la pérdida de colonias. Sin embargo, el RMSE elevado sugiere que existen outliers o escalas grandes que podrían mejorarse con normalización o transformaciones.

Visualización de datos

Se recomienda complementar con los siguientes nodos para visualización:

- `Scatter Plot`: comparar predicción vs real
- `Box Plot`: analizar distribución de cada variable
- `Line Plot`: ver evolución temporal si se agregan columnas de tiempo

(Estos no fueron incluidos en el flujo original, pero pueden agregarse para presentaciones o reportes).



Conclusión y reflexión final

Este proyecto demostró que herramientas de machine learning pueden ayudarnos a entender fenómenos ecológicos complejos. KNIME permitió estructurar un flujo sin necesidad de programación, permitiendo limpiar datos, entrenar un modelo de regresión y evaluar su rendimiento.

El modelo resultante tiene un buen ajuste y podría emplearse para prever pérdidas de colonias en diferentes escenarios o regiones. El siguiente paso podría ser aplicar modelos más robustos como Random Forest o ajustar hiperparámetros para reducir errores.

Este tipo de análisis puede informar políticas públicas, estrategias de conservación y educación ambiental.