

# **Predicción Taxonómica de Especies**

## **Usando Frecuencias de Codones**

**Alumno:** Esconjaureguy leonel

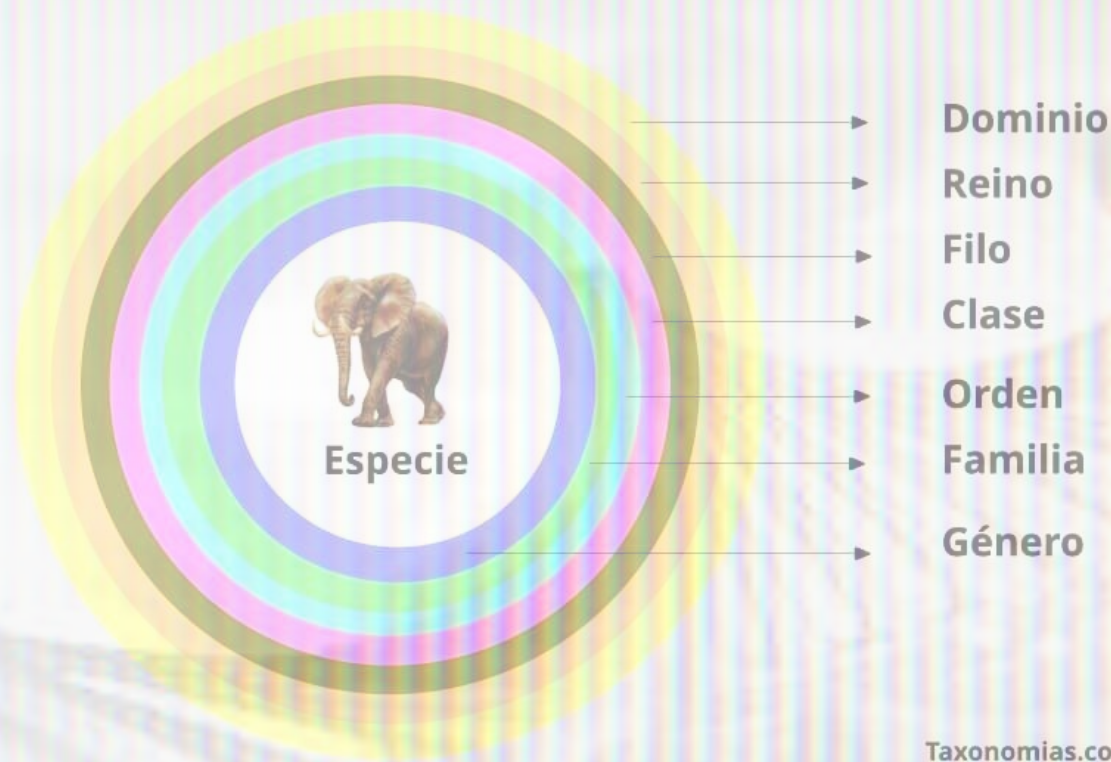
**profesor:** Germán Rodriguez

**Tutor:** Silvia Vilela

**Comisión:** #61190

# Introducción:

En la actualidad, los investigadores lidian con un océano de datos genéticos. Clasificar especies manualmente puede tomar semanas, mientras que la biodiversidad enfrenta amenazas crecientes. Con el poder del Machine Learning, podemos transformar este desafío en una oportunidad: predecir con rapidez y precisión a qué reino pertenece una especie a partir de patrones ocultos en sus secuencias de codones.



# Introducción:

La capacidad de predecir el reino taxonómico tiene aplicaciones transformadoras en biotecnología, agricultura y medicina, impulsando innovaciones en áreas clave:



## Desarrollo de nuevos medicamentos:

Identificar especies microbianas que produzcan compuestos bioactivos para tratamientos más efectivos.

Detectar patógenos y acelerar el desarrollo de **terapias personalizadas**.

○



## Innovación en biotecnología y expresión heteróloga de genes:

- Seleccionar especies con atributos únicos para la **producción de proteínas recombinantes** o biomateriales avanzados.
- Acelerar la creación de **productos biotecnológicos** eficientes y sostenibles

○

## Mejora de la calidad y rendimiento de los cultivos:

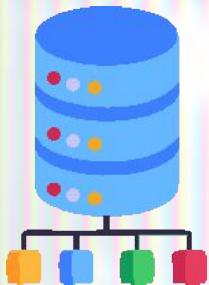
- Identificar genes clave en plantas y microorganismos para potenciar cultivos más resistentes y productivos.
- Facilitar la ingeniería genética para enfrentar desafíos como la sequía o plagas.

# Objetivo:



- Desarrollar un modelo de clasificación supervisado que aprende patrones distintivos en las frecuencias de codones para predecir con precisión el reino biológico de una especie (virus, bacterias, eucariotas).

# Dataset:



El conjunto de datos examina las frecuencias de uso de codones en el ADN codificante de muestras de organismos de diferentes taxones.

## Variables:

### categorías:

- Kingdom
- SpeciesName

### Dimensiones:

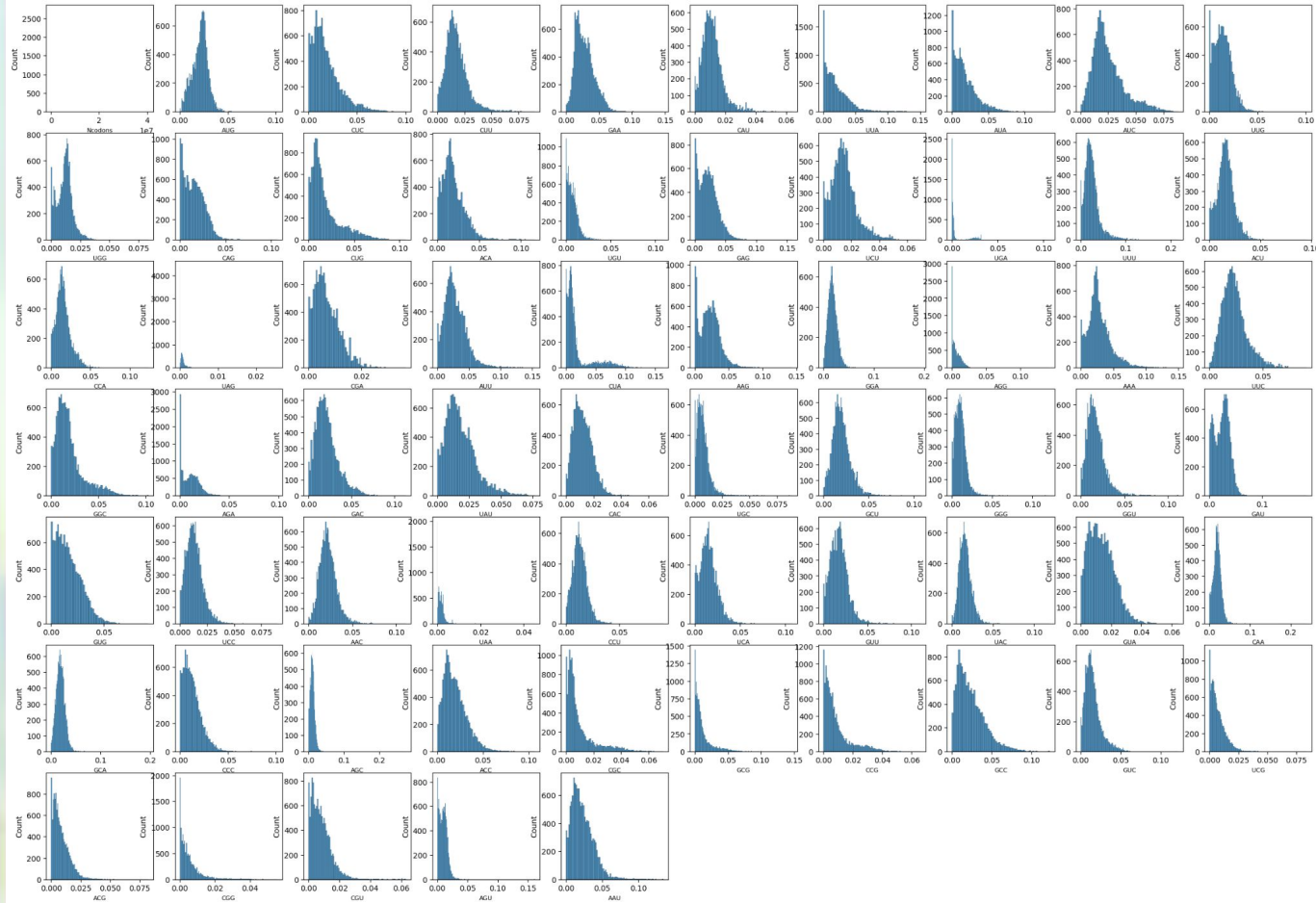
- columnas: 69
- filas: 13028

### categorías:

- Codon Frequencies:
- Ncodons
- DNAtype

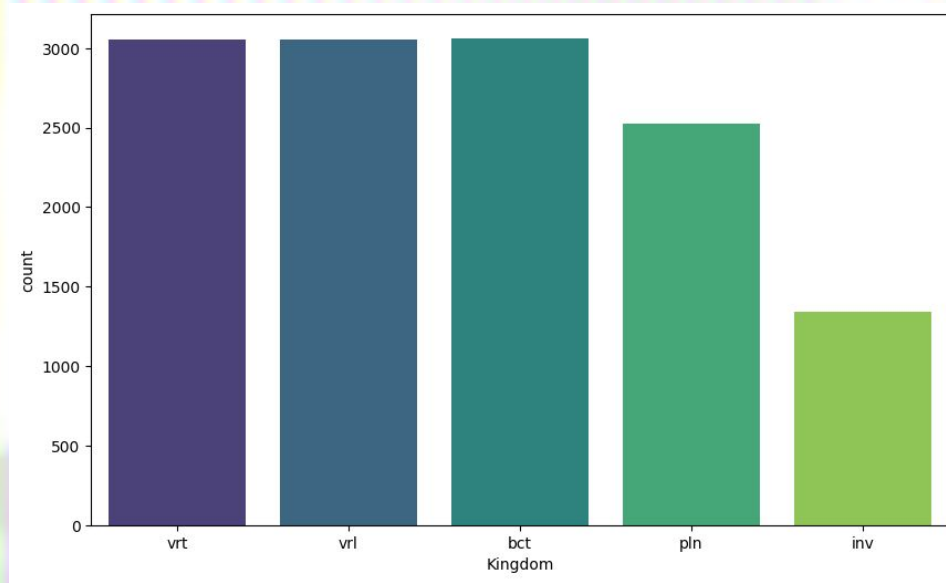


# Análisis univariado:



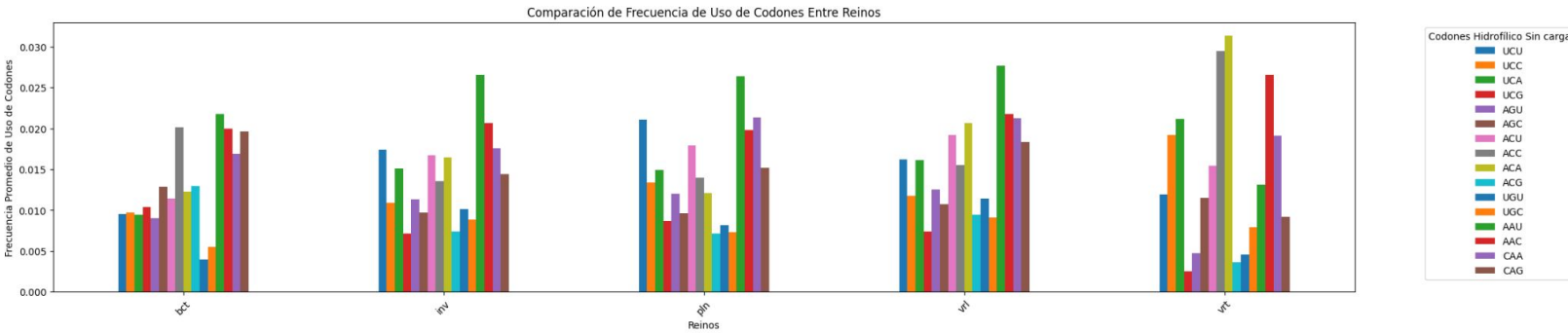
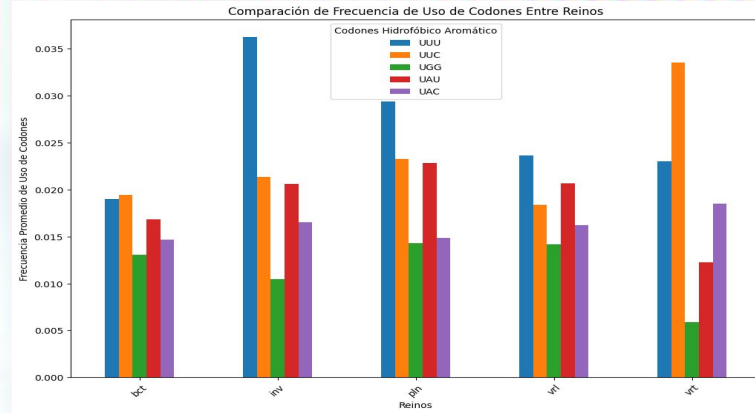
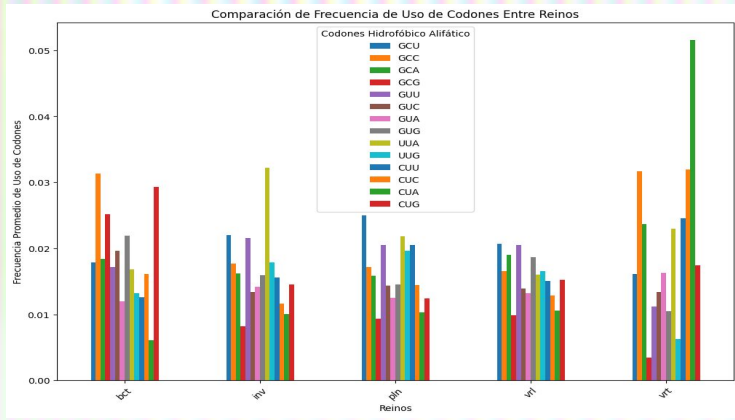
El gráfico muestra que ciertos codones se utilizan con más frecuencia que otros. Esto sugiere que podrían ser más eficientes para crear proteínas, lo que podría ayudar a los organismos a adaptarse mejor a su entorno

# Análisis univariado:



La variable que queremos predecir es el reino al que pertenece el gen en estudio. Conocer esta información es crucial, ya que nos orienta sobre los pasos a seguir en nuestra investigación y la implementación de herramientas adecuadas

# Análisis bivariado:



Los análisis univariado y bivariado ayudan a identificar patrones y relaciones entre la frecuencia de codones y el reino biológico, lo cual es una buena base para inferir su potencial como predictor.



# Selección de modelos:

categorias	accuracy	Presicion	recall	F1_score
arbol de desición	0.92	0.92	0.92	0.92
knn	0.61	0.63	0.61	0.60
random forest	0.97	0.97	0.97	0.97
regresión logistica	0.74	0.77	0.74	0.73

Random Forest es el modelo que tiene el mejor rendimiento general, con un accuracy, precision, recall y F1-score de 0.94. Esto indica que Random Forest es el modelo más robusto y equilibrado en cuanto a las métricas clave de rendimiento. La alta precisión y recall indican que este modelo es efectivo tanto para identificar correctamente las clases positivas como para minimizar los falsos negativos.

# Mejorar el modelo:

## **Cantidad de variables**

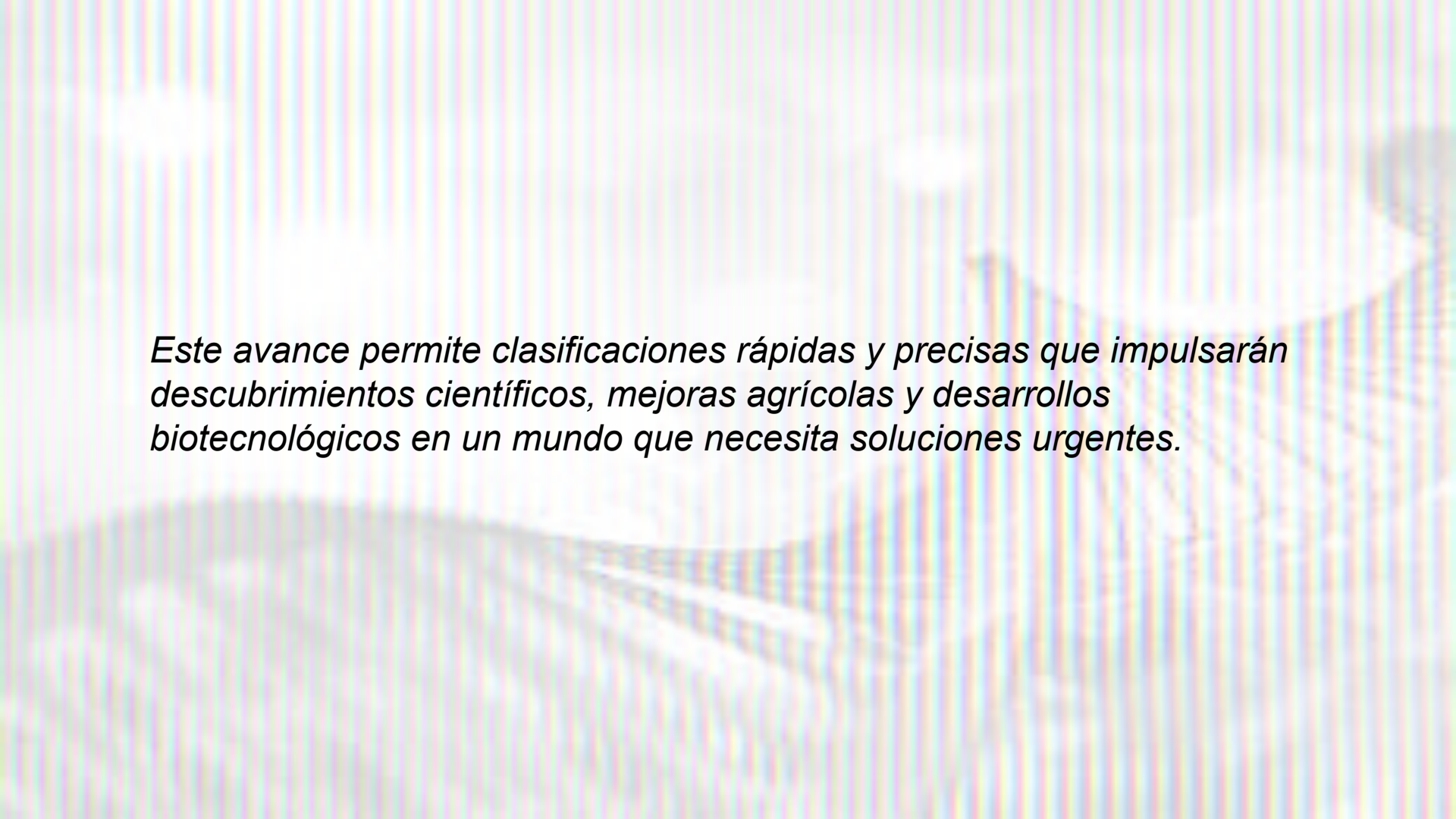
Los mejores valores métricos se logran cuando `max_features` es de 21, alcanzando una precisión del 96%, una sensibilidad del 96% y una exactitud del 96%

## **Cantidad de ramas**

Los mejores valores métricos se logran cuando el número de ramas es de 24, alcanzando una precisión del 97%, una sensibilidad del 97% y una exactitud del 97%

## **Cantidad de árboles**

Los mejores valores métricos se logran cuando el `n_estimators` es de 75, alcanzando una precisión del 96%, una sensibilidad del 96% y una exactitud del 96%



*Este avance permite clasificaciones rápidas y precisas que impulsarán descubrimientos científicos, mejoras agrícolas y desarrollos biotecnológicos en un mundo que necesita soluciones urgentes.*