Bosques aleatorios y cadenas de Márkov para clasificación de taxonomías sobre cadenas de ARN

AHUMADA SANTIAGO[1], ALDANA YEFERSSON[2], SÁNCHEZ JORGE[3], AND REINA JUAN[2]

[1] *Departamento de Matemáticas, 2022. Universidad Nacional de Colombia, Bogotá D.C.*
[2] *Facultad de Ingeniería, 2022. Universidad Nacional de Colombia, Bogotá D.C.*
[3] *Facultad de Medicina, 2022. Universidad Nacional de Colombia, Bogotá D.C.*

Compiled November 28, 2022

Resumen: La clasificación de cadenas de ARN es relevante en el mundo de la biomedicina, un ejemplo de esto es su uso en el desarrollo de las vacunas en contra de determinado virus. Sin embargo, catalogar este material genético se puede tornar bastante ineficaz de realizar por un profesional de la ciencia. En este artículo se abordará un método para solventar esto usando modelos de aprendizaje automatizado, por medio de un modelo matemático denominado Cadenas de Márkov, además de usar Bosques Aleatorios para ordenar estas en diferentes agrupaciones taxonómicas.

Palabras clave: ARN, codón, nucleótido, cadena de Márkov, norma matricial, autovalor, bosque aleatorio.

https://github.com/santiagoahl/RNA-taxonomy-prediction

CONTENIDO

1. INTRODUCCIÓN

El campo de la genética, desde su estructura y composición, se presta perfectamente para la utilización de herramientas de bases de datos, procesamiento de datos, y análisis de resultados, convirtiéndose estas en la piedra angular sobre la que se cimienta el conocimiento científico en esta materia.

Esto nos permite su comprensión y posterior investigación a través de funciones como las cadenas de Márkov, la cual se utiliza en este proyecto, con la idea de convertirse en un medio inicial para la organización y modelamiento de grandes cadenas de ARN de diferentes taxonomías, las cuales serian una tarea colosal sin la utilización de la tecnología de la información, que no la reduce, pero si nos da una herramienta a la altura del reto.

Por ende, este proyecto se convierte en una tarea multidisciplinar, que requiere la comprensión del modelo genético inicialmente propuesto por Watson y Crick, la comprensión de la teoría fundamental de la genética, así como el dominio de las normas estadísticas, la verificación de la información, sumado al dominio de las ciencias de la información, la programación, y la computación a gran escala. En este texto se plantea una herramienta para la clasificación de grandes cadenas taxonómicas, y que se espera, sea un aporte para la Universidad Nacional de Colombia y su departamento de Genética.

2. TAXONOMÍAS DE ARN

La taxonomía de acuerdo a la escuela cladísta (que dentro de las múltiples que existen actualmente es la aproximación más usada) es la ciencia que estudia la clasificación de cada una de las ramificaciones que se obtienen de un árbol filogenético, a estas clasificaciones se les denomina clados, además, un árbol filogenético se forma cuando a partir de un ente biológico (ani-

males, planta, hongos, bacterias, etc.) se rastrea su antepasado y se comienzan a desprender todos sus descendientes en un grafo con forma de árbol, independientemente de la escuela que la defina, la taxonomía pretende presentar un modelo ordenado y preciso de toda la variedad de organismos en unidades discretas, esto con el fin de que sea menos extenuante para los científicos trabajar con ellos.

Los sistemas de clasificación actuales están compuestos por taxones ubicados en sus respectivas categorías taxonómicas, las cuales, de acuerdo a heurísticas todavía discutidas, corresponderían a los mismos celadones según la escuela clasista, ya que son arbitrarias y poco científicas, además, generan confusión en la comunidad lo que es totalmente contradictorio con el propósito de formular un modelo clasificatorio como la taxonomía, con el fin de corregir este problema Judd y colaboradores (2002) coinciden en que:

- Cada taxón debe tener evidencia suficiente y fiable de formar un grupo monofilético, en palabras del autor "para convertir un cladón en taxón debe haber suficientes sinapomorfías que lo justifiquen", es decir, para clasificar un taxón este debe tener la suficiente cantidad de características físicas y genéticas que sus parientes cercanos para ser clasificado.

- Cada taxón debería tener caracteres morfológicos obvios que permitan identificarlo, en el caso de las aves existe una clara diferenciación en la forma de sus picos y este permite que al encontrar una nueva especie de ave esta pueda ser inicialmente clasificada por la forma de su pico al ser comparado con la de sus parientes cercanos.

- De acuerdo a Davis y Heywood (1963) dentro de los diferentes taxones deben existir subcategorías que permitan diferenciar con mayor precisión pequeños cambios generados en las especies, en palabras de Davis y Heywood “Debemos ser capaces de ubicar a los taxones en sub taxones de categoría más alta de forma que podamos encontrarlos de nuevo".

- Finalmente, el último criterio es la estabilidad de la nomenclatura, si un grupo es nombrado en el pasado, este debe mantenerse con el mismo nombre.

Podemos observar que la manera en la que se clasifican los seres vivos de los distintos reinos no presenta una claridad absoluta para la comunidad científica y los investigadores dedicados a la taxonomía en la naturaleza, esto genera un problema de información ambigua y difusa en los textos, esto ocurre, por ejemplo, con relativa frecuencia en el reino animal que presenta la cifra poco apreciable de entre 1,5 y 2 millones de especies, esto se traduce en una incertidumbre de medio millón de especies sin clasificar, en cifras conservadoras se cree que existen aún 8,7 millones de especies vivas sin clasificar.

Debido a la gran cantidad de información que irá surgiendo conforme se investiga y se sigan descubriendo más y más especies es relevante proponer el uso de nuevas herramientas más actuales para clasificar a los seres vivos, una de ellas es la inteligencia artificial, la cual permite el procesamiento de inmensas cantidades de información mediante tecnología capaz de realizar múltiples cantidades de operaciones matemáticas casi instantáneamente, en la naturaleza es bien conocido que, desde el descubrimiento del ADN, las especies son diferenciables con relativa facilidad, ya que analizando detenidamente la estructura proteica del ADN podemos saber con exactitud a qué familia pertenece determinada especie, pero esto conlleva a un problema y es que.

Por ejemplo, el ADN de un ser humano tiene en su interior aproximadamente 700 megabytes de información, alrededor de 180 GB, lo cual hace extremadamente difícil todo su procesamiento, ahora extrapolar esta situación a la cantidad de especies que irán surgiendo con el tiempo representa una cifra de información casi ridícula de mencionar.

No debe ser apartado el hecho de que existen seres más simples en la medida en que su estructura genética es reducida, estos son los virus, seres microscópicos de los que no se puede decir con certeza que están vivas, pero que interactúan con seres vivos de manera común, los virus han estado presentes en la tierra desde mucho antes que los dinosaurios, estos son microorganismos compuestos de material genético protegido por un envoltorio proteico, que causa diversas enfermedades introduciéndose como parásito en una célula para reproducirse en ella (Oxford Languages) y causar graves problemas en seres multicelulares como los humanos.

Su clasificación (taxonomía) es vital a la hora de elaborar antídotos, ya que estos actúan dependiendo de la estructura del determinado virus, en la actualidad se tienen cerca de 320.000 virus diferentes que afecten a los mamíferos, muchos de los cuales son beneficiosos y otros inofensivos para los humanos, sin embargo, hay otros extremadamente contagiosos y letales como son la fiebre hemorrágica de Marburgo, ébola, virus Hanta, gripe aviar, fiebre de Lassa, SIDA, etc. El caso de los virus es especial, ya que es, a diferencia de los animales, el reino de los virus produce nuevas especies a ritmos acelerados. Se estima que en la tierra existen 10 quintillones de microbios virales, si bien la mayoría no son letales para los seres humanos, no hay tolerancia que soporte la cantidad elevada de posibles virus letales para los seres humanos, el más reciente de ellos que provocó la pandemia de 2019 COVID es un ejemplo de que la humanidad no estaba del todo preparada para un virus con la capacidad de contagiar a todo el planeta y menos de que produjera varios millones de muertes directas e indirectas, sin embargo, la humanidad no se quedó atrás en el desarrollo de una vacuna apropiada mediante el método CRISPR.

La humanidad se dio cuenta entonces que es necesaria una aceleración en la producción de antídotos efectivos, ya que no será el último virus que afecte a la humanidad, uno de los pasos más importantes a la hora de diseñar un antídoto es reconocer a que especie de virus pertenece para saber de qué manera se podría comportar ante distintos antídotos, y con la gran cantidad de virus que restan por descubrir es conveniente crear una herramienta que los pueda clasificar para eso se propone en este documento el uso de herramientas computacionales elevadas y algoritmos matemáticos como las cadenas de Márkov para clasificar diferentes virus de acuerdo a un conjunto de datos de su ADN.

3. OBJETIVOS

Se plantea como objetivo principal la clasificación de distintas familias de virus de acuerdo a las distintas combinaciones de codones presentes en familias ya estudiadas, así como la predicción

Además, Se plantean como objetivos secundarios la implementación de las cadenas de Márkov para analizar conjuntos extensos de datos, el uso de distintas estrategias de simplificación de estructuras de datos con forma de grafos y finalmente analizar con detenimiento la eficacia con la cual el modelo predic-

tivo propuesto determina la familia de virus a la cual pertenecen diversas familias ya estudiadas.

4. PLANTEAMIENTO DEL PROBLEMA

Inicialmente, se tiene un conjunto de datos que consiste en la información genética de la taxonomía de unos virus, con estos se calcularán las respectivas Cadenas de Márkov, este es visto en la teoría de la probabilidad como un tipo especial de procesos estocástico discreto en el que la probabilidad de que ocurra un evento depende única y exclusivamente del evento anterior obviando los anteriores, este procedimiento para cada secuencia dispuesta de ARN

Con las cadenas ya generadas, se les es calculada métricas matriciales, a partir de las normas matriciales y los autovalores podemos obtener un nuevo conjunto de datos, a este nuevo conjunto de datos se les debe asignar un clasificador, que en este caso es un clasificador Random Forest, este es un método de aprendizaje automático capaz de realizar tareas tanto de clasificación como de regresión.

Escogimos este método, pues, resulta ser de bastante utilidad por tratarse de un método de aprendizaje supervisado por conjuntos donde un grupo de modelos débiles se combinan para formar uno poderoso, con esta estrategia evitar sesgos generados por el mismo modelo o generados de manera intrínseca por el tipo de datos que se están tratando, en secciones siguientes se tratará con más detalle el modelo que se utilizó y los cálculos realizados.

5. CONJUNTO DE DATOS

Se importaron cadenas de ARN, cada una de ellas perteneciente a 1 entre 19 diferentes taxonomías, desde una base de datos suministrada por el docente Juan David García Arteaga, Ph. D., la cual puede ser encontrada en este enlace.

Esta fue suministrada en un formato de dos columnas, la primera corresponde al nombre del virus con una taxonomía retroviral específica:

- Orthomyxoviridae
- Rhabdoviridae
- Arteriviridae
- Coronaviridae
- Reoviridae
- Caliciviridae
- Phenuiviridae
- Hantaviridae
- Picornaviridae
- Betaflexiviridae
- Astroviridae
- Closteroviridae
- Flaviviridae
- Potyviridae
- Retroviridae
- Togaviridae
- Paramyxoviridae
- Hepeviridae
- Pneumoviridae

La segundo columna constituye la cadena de ARN verificada, correspondiente a la taxonomía de la primera columna, siendo constituida por alrededor de 2000 a 4000 caracteres, cada uno representando una base nitrogenada de la cadena de ARN del virus.

6. MANEJO DE DATOS

El manejo del conjunto de datos en donde convertimos cada lista de nucleótidos, agrupados de manera individual:

- A,G,C,A,A,T,C,A,C,A,A,T,C,T,G

En una lista de codones:

- AGC, AAT, CAC, AAT, CTG

En donde se estructuraron las cadenas genéticas agrupadas a través del siguiente algoritmo:

Algorithm 1. Algoritmo para particionar secuencias de ARN

1: **procedure** COMPUTECODONS(*chain*) ▷ Convierte una secuencia de nucleotidos en una secuencia de codones.
2: $l \leftarrow len(chain)$ ▷ Longitud de la cadena
3: $codons \leftarrow []$ ▷ Secuencia inicial vacía de codones
4: **for** $i = 0$ to $l - 4$, $i \leftarrow i + 3$ **do** ▷ Hacemos saltos en la secuencia cada 3 nucleotidos
5: $codons \leftarrow codons.insert(chain[i : i + 3])$ ▷ Insertamos un nuevo codón a la cadena
6: **return** *codons* ▷ Secuencia de codones

Utilizando las reglas de la genética básica, y su cumplimiento junto con las relaciones de codones correctas, en donde después de aplicar el algoritmo, la información queda organizada de forma tal, que posteriormente cada uno de estos codones será utilizada como nodo de las cadenas de Márkov al ejercutarse la aplicación.

7. CADENAS DE MÁRKOV

Tras hacer la partición de cada cadena de ARN en codones. Nuestro plan es ser capaces de desmantelar la estructura interna de las cadenas, de forma que podamos discernir la distribución de aparición y escritura de cada codón y cómo la aparición de un codón influye en que otro aparezca o no.

En modelos de procesamiento de lenguaje natural, tanto para clasificación como para generación de texto, las cadenas de Márkov son muy populares. Lo anterior es dado que una cadena de Márkov mide la probabilidad de transición de cada estado a otro y modela dichas probabilidades con grafos dirigidos ponderados.

Para ser precisos, hemos trabajado con una estructura de dato equivalente a un grafo dirigido ponderado, es decir, con la matriz de transición de estados asociada al grafo. Se denota cada nodo por un número entero $0, 1, \ldots, n$ y se define la matriz

$$T = [t_{i,j}]_{i,j \in \{0,\ldots,n\}}$$

Donde

$$t_{i,j} := \mathbb{P}[j|i]$$

Es decir, $t_{i,j}$ mide la probabilidad de que después de observar el estado i, se observe el estado j.

La figura 1 muestra un ejemplo de una cadena de Márkov asociada a una cadena de ARN.

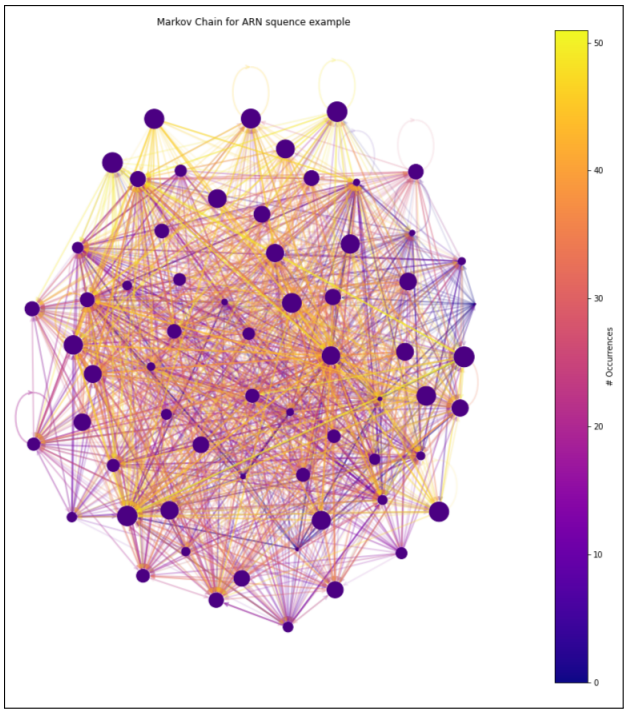

Fig. 1. Visualización de una cadena de Márkov asociada a un ejemplar de cadena de ARN. Cada nodo es un codón. Se omiten etiquetar los nodos por estética.

Hemos observado que la mayoría de valores de las matrices eran 0 o muy cercanos a 0. Por tanto, hemos normalizado las entradas de las matrices tomando la normalización máximo mínimo, la cual mapea linealmente

$$[min, max] \to [0, 1]$$

Donde tomamos $min = 0$ y $max = 2.32 \times 10^{-2}$. La figura 2 muestra un ejemplar de matriz por cada taxonomía.

8. NUEVO CONJUNTO DATOS

Las cadenas de Márkov nos proveen bastante información sobre la estructura interna de las secuencias de ARN. Es por ello que optamos por construir un nuevo conjunto de datos que describe de forma numérica a las cadenas de ARN con base en las matrices de transición de estado asociada a cada cadena.

En total se calcularon 18 métricas por cada cadena de Márkov: Las 8 primeras corresponden a normas matriciales (Para más información sobre las normas matriciales usadas, ver [2]) y las 10 restantes corresponden a las 10 primeras normas (Norma compleja) de autovalores de la matriz por ordenadas de forma ascendente. La figura 4 nos muestra la distribución del nuevo conjunto de datos.

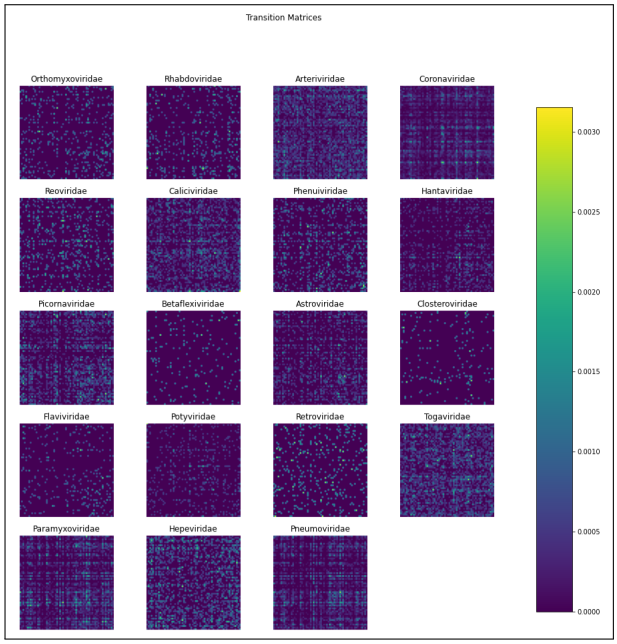

Fig. 2. Distribución de valores de matrices. Se tomó un ejemplo por cada taxonomía.

	A	B	C	D	F	J	T
1	Column1	Frobenius Norm	Nuclear Norm	Infty Norm	Neg L1 Norm	eig 1	Taxonomy
2	0	4,14902E+16	2,35852E+16	5,60399E+15	1,24533E+16	1,67757E+15	Orthomyxoviridae
3	1	4,06888E+16	2,38189E+15	4,52261E+15	2,51256E+15	3,21591E+11	Orthomyxoviridae
4	2	4,6728E+14	2,77136E+15	4,4905E+15	0	0	Orthomyxoviridae
5	3	4,69124E+15	2,78054E+16	3,71747E+15	0	0	Orthomyxoviridae
6	4	7,6685E+14	4,8257E+16	4,62428E+16	0	0	Orthomyxoviridae
7	5	6,22718E+15	3,7147E+15	5,66667E+15	0	0	Orthomyxoviridae
8	6	4,18288E+15	2,37168E+16	4,35323E+15	0	0	Orthomyxoviridae
9	7	4,24102E+16	2,36756E+16	6,36704E+15	0	0	Orthomyxoviridae
10	8	4,73095E+16	2,77414E+15	5,73913E+15	0	0	Orthomyxoviridae
11	9	4,83997E+16	2,90538E+14	5,41045E+16	0	0	Orthomyxoviridae

Fig. 3. Nuevo conjunto de datos

9. MODELO DE CLASIFICACIÓN: RANDOM FOREST

Los nuevos datos creados a partir de métricas asociadas a la matriz de transición de estado nos proveen bastante información sobre las secuencias de ARN. Al ser datos puramente numéricos. Hemos optado por implementar un modelo de clasificación multi clase utilizando **bosques aleatorios**.

El método de bosques aleatorios (o en inglés, Random Forest) Es un método de ensamble el cual construye una cantidad finita de árboles de decisión (En este contexto se denominan predictores o estimadores) y genera un consenso entre las predicciones de cada uno de ellos para una muestra dada. El componente aleatorio de los bosques aleatorios se debe a que cada predictor elige características al azar y con base en ellas genera su predicción. La ventaja de esto es que los árboles serán entonces débilmente correlacionados y con base en la teoría de aprendizaje automatizado y el conocimiento de múltiples expertos, estas predicciones serán más precisas.

El bosque aleatorio entrenado se construyó con ayuda de la librería `scikit-learn` con las siguientes características:

- Una cantidad de 100 estimadores.
- Una profundidad máxima de 30 ramas por cada árbol.
- El conjunto de entrenamiento de cada árbol se escogió con la técnica **bootstrap**.

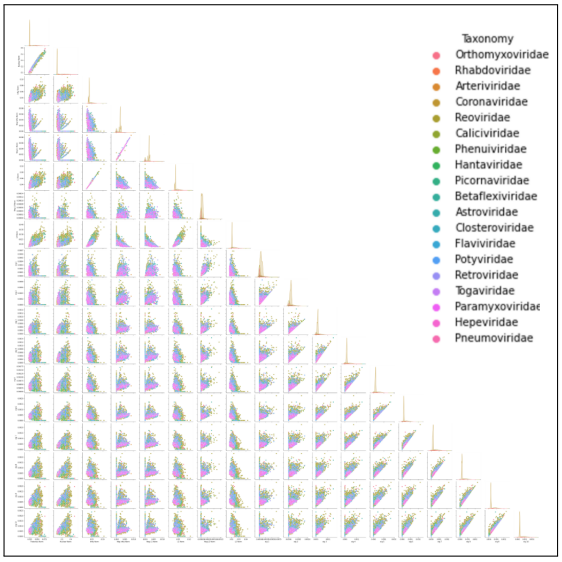

Fig. 4. Distribución de las variables del nuevo conjunto de datos. Las etiquetas corresponden a la normas matriciales y normas de autovalores mencionados

- El criterio de optimización es la medida **Gini**.

10. RESULTADOS

El modelo obtuvo un **95.16%** de score en el conjunto de testeo (Véase figura 5).

```
Score en conjunto de entrenamiento:  100 %
Score en conjunto de testeo:  95.16 %
```

Fig. 5. Score del modelo.

La figura 6 nos muestra las curvas de aprendizaje conseguidas (La primera de ellas, usando validación cruzada). Mientras la figura 7 nos muestra la matriz de confusión del modelo.

11. CONCLUSIONES

A lo largo del presente proyecto hemos logrado implementar con relativo éxito un modelo de aprendizaje supervisado que clasifica secuencias de ARN en taxonomías. Algunas conclusiones finales son:

- Se observó una mejoría en el desempeño de los modelos de clasificación al normalizar las entradas de las matrices y posteriormente el nuevo conjunto de datos. Se utilizaron escaladores `MinMaxScaler()` y `StdScaler()`.
- El tratamiento de los datos es una de las etapas claves para conseguir el éxito en el desarrollo de modelos de aprendizaje. En particular, la ingeniería de características y la visualización de datos jugó un papel importante para optimizar el desempeño de nuestro modelo.
- Actualmente, el modelo de bosque aleatorio es uno de los más efectivos en el campo de aprendizaje automático y es popular entre los métodos de ensamble.

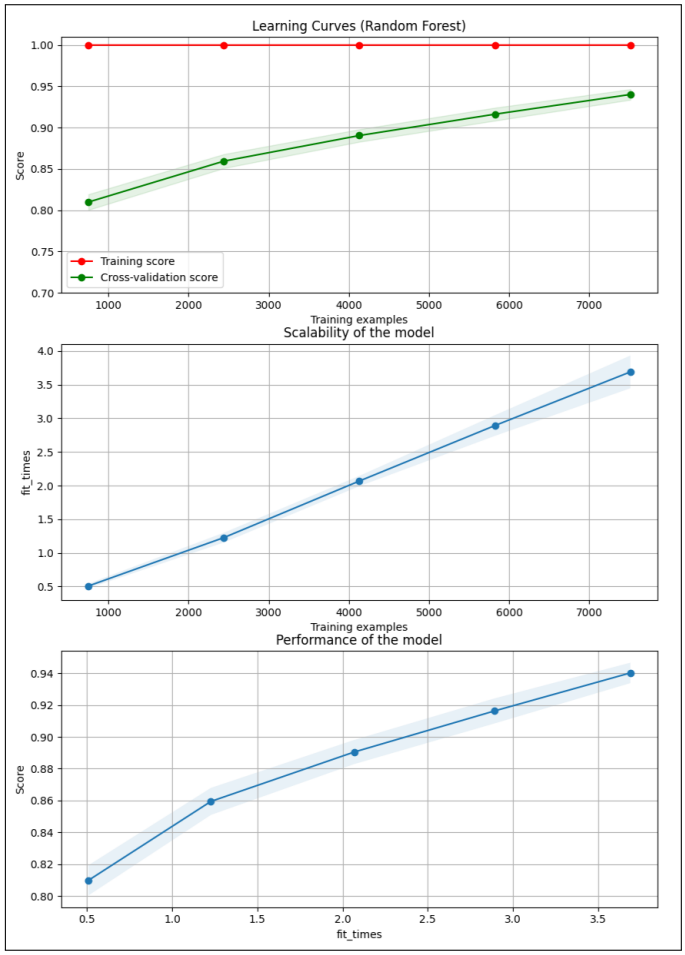

Fig. 6. Curvas de aprendizaje del modelo

A. Posibles mejoras

- Los bosques aleatorios se desempeñan mejor si se les dan bastantes datos, por lo que consideramos que el modelo mejoraría si de 5000 ejemplos podemos conseguir un estimado de 30.000 ejemplos.
- Respecto a los datos, calculamos una métrica adicional para medir de manera equitativa qué tan bien predicha fue cada clase. Esta métrica se conoce como el **f1 score**. Nuestro modelo obtuvo 58.5% en esta métrica. Investigando encontramos que se debe a que los datos suministrados estaban muy desbalanceados. En efecto, el 46% de los datos perteneces a una sola clase. Por tanto, consideramos que se deberían suministrar nuevos datos que balanceen esta desproporción.
- Encontramos que algunas normas matriciales estaban fuertemente correlacionadas (Anteriormente eran las 11 posibles que ofrece `NumPy`, por lo que eliminamos 3) por tanto, consideramos que a futuro se debería hacer una limpieza más exhaustiva de estos datos para reducir variables correlacionadas. Una opción que se baraja es el algoritmo de extracción de componentes principales `PCA`.
- En el nuevo conjunto de datos consideramos que podemos agregar distintos tipos de variables nuevas. Por ejemplo: valores singulares, determinantes, cofactores, etc.

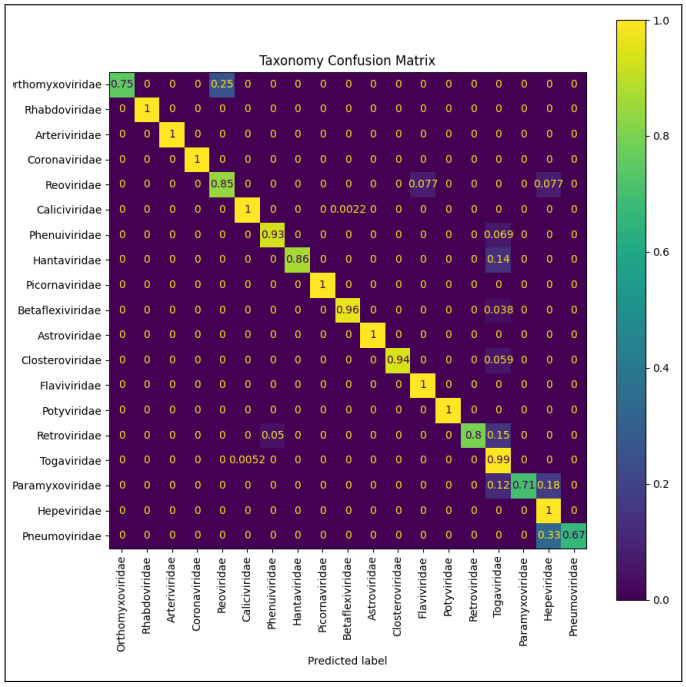

Fig. 7. Matriz de confusión del modelo.

12. REFERENCIAS

1 Scikit-learn: machine learning in Python — scikit-learn 1.1.3 documentation. (s. f.). scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. https://scikit-learn.org/stable/

2 numpy.linalg.norm. NumPy v1.23 Manual. (s. f.). NumPy. https://numpy.org/doc/stable/reference/generated/numpy.linalg.norm.html

3 Abid Ali Awan 2021, Deep learning model to predict mRNA degradation - Towards AI, Medium, Towards AI, viewed 20 October 2022, <https://pub.towardsai.net/deep-learning-model-to-predict-mrna-degradation-1533a7f32ad4>.

4 Kuhn, JH 2021, 'Virus Taxonomy', Encyclopedia of Virology, pp. 28–37, viewed 23 October 2022, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7157452/::text=Virus20%taxonomy%20is%20a%20virology,naming%20(nomenclature)%20for%20taxa.>.

5 Wren, JD, Hildebrand, WH, Chandrasekaran, S Melcher, U 2005, 'Markov model recognition and classification of DNA/protein sequences within large text databases', Bioinformatics, vol. 21, no. 21, pp. 4046–4053, viewed 25 October 2022,<https://academic.oup.com/bioinformatics/article/21/21/4046/2269262574699>.

6 National Center for Biotechnology Information 2022, Nih.gov, viewed 28 October 2022, <https://www.ncbi.nlm.nih.gov/>.

7 Brandis, G Hughes, D 2016, 'The Selective Advantage of Synonymous Codon Usage Bias in Salmonella', in M Ibba (ed.), PLOS Genetics, vol. 12, no. 3, p. e1005926, viewed 28 October 2022, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4786093/>.

8 Zimmer, C 2013, An Infinity of Viruses, Science, National Geographic, viewed 28 October 2022, <https://www.nationalgeographic.com/science/article/an-infinity-of-viruses>.

9 del 2021, El ARNm, el mensajero del genoma - El Blog de Genotipia, Genotipia, viewed 31 October 2022, <https://genotipia.com/arnm/>.