Modelos de Machine Learning para la Clasificación e Identificación Taxonómica de Bacterias utilizando Secuencias Cortas del Gen rRNA 16S

Godinez Bravo Diego
Maestría en Cómputo Estadístico, Centro de Investigación en Matemáticas A. C., Parque de Investigación e Inovación Tecnológica, C.P. 66629, Monterrey, Nuevo León.

La taxonomía es una de las disciplinas científicas más relevantes en biología. En sus inicios, la caracterización y clasificación de especies se basaban exclusivamente en características fenotípicas. En la actualidad, la sistemática y la taxonomía de bacterias han evolucionado hacia el uso de información molecular. La introducción de métodos modernos de secuenciación ha optimizado y escalado el análisis de genomas completos, destacándose la secuenciación por shotgun y la secuenciación de amplicones. En este proyecto se aborda el trabajo taxonómico utilizando secuencias cortas obtenidas mediante la secuenciación de amplicones del RNA ribosómico (rRNA) de la subunidad pequeña 16S. El análisis de los datos genómicos incluye la extracción de características a través de la representación de frecuencias de subsecuencias, conocidas como k-mers. Además, se aplican técnicas de visualización de datos de alta dimensión, junto con modelos de Machine Learning, para explorar patrones subyacentes en los datos, identificar diferencias en la composición de nucleotidos entre secuencias y clasificar eficazmente las especies en distintos niveles jerárquicos. Este enfoque destaca el poder de la extracción de características en combinación con métodos computacionales modernos. Los resultados ofrecen una visión completa del desempeño de cada modelo, destacando su capacidad para diferenciar entre géneros bacterianos a partir de secuencias genómicas.

Keywords: ***Taxonomía, Secuenciación de Amplicones, k-mers, Machine Learning, Clasificación***

1 Introducción

La taxonomía ha experimentado un avance significativo a lo largo del tiempo, impulsada principalmente por los avances en biología molecular y genómica, que a su vez han sido potenciadas por las técnicas de secuenciación. En sus inicios, la clasificación se basaba principalmente en los rasgos fenotípicos, lo que condujo a clasificaciones incorrectas, ya que las diferencias observadas no siempre reflejan una divergencia evolutiva real. Esto es especialmente relevante en el dominio Bacteria, donde su gran diversidad y plasticidad fenotípica, entendida como la capacidad de un organismo para modificar su fenotipo en respuesta a diferentes condiciones ambientales, puede provocar que especies diferentes muestren rasgos fenotípicos completamente distintos, mientras que, a nivel genético, pertenezcan al mismo género.

Dentro del dominio Bacteria, se encuentra uno de los filos más grandes y fenotípicamente más diversos, llamado Proteobacteria. Este filo consta de más de 460 géneros y 1600 especies, lo que implica una notable diversidad morfológica y fisiológica. Las bacterias de este filo pueden presentar formas tan variadas como bastones, cocos o espirales, y tener características fisiológicas que van desde mesófilas hasta termófilas. Además, muestra una gran variedad de interacciones ecológicas, la mayoría de las especies conocidas son de vida libre. Sin embargo, algunas establecen asociaciones simbióticas con plantas, otras viven como endosimbiontes intracelulares de protozoos e invertebrados, mientras que otras son parásitos intracelulares obligados de humanos o mamíferos. Esta enorme diversidad, tanto en su morfología como en sus características metabólicas y ecológicas, subraya la importancia de emplear herramientas moleculares para lograr una clasificación precisa y reflejar adecuadamente las relaciones evolutivas dentro de este extenso dominio 1.

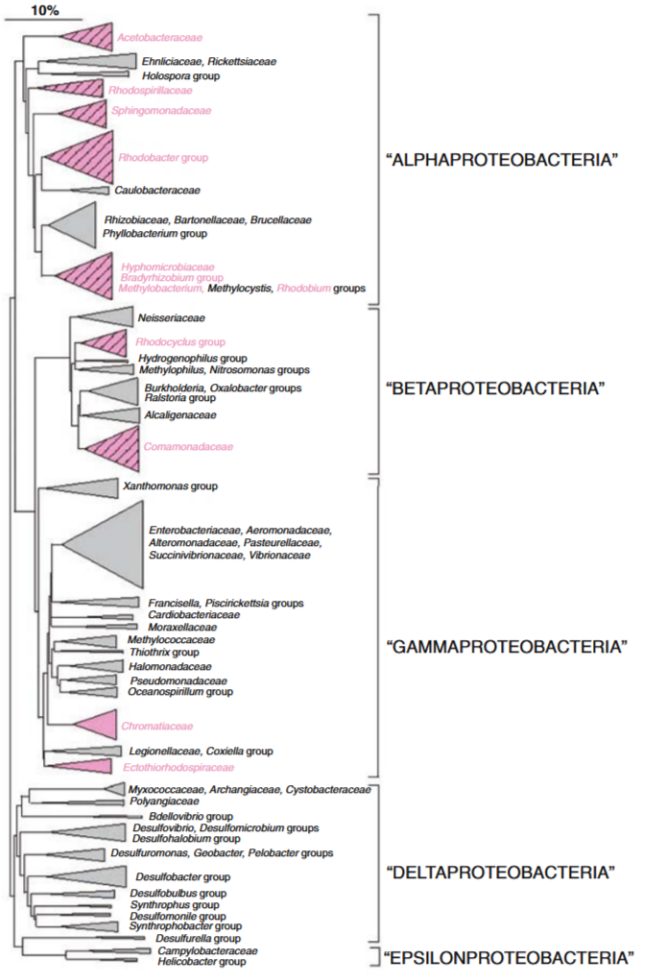

FIGURE 1. Árbol filogenético construido a partir de secuencias 16S rRNA, representando la filogenia entre géneros del filo Proteobacteria (M. Dworkin *et al.* (2006))

Una de las moléculas más robustas en el estudio de la filogenia y taxonomía de procariontes es la molécula de RNA ribosomal (rRNA) de la subunidad pequeña, conocida como 16S. Esta molécula, presente en todos los procariontes, resulta especialmente útil debido a su resistencia a la transferencia horizontal de genes (HGT), lo que facilita una representación más precisa de las relaciones evolutivas entre diferentes especies. La estructura del rRNA 16S se compone de nueve regiones (V1-V9), algunas de las cuales son altamente conservadas, mientras que otras presentan mayor variabilidad. Esta combinación la convierte en una excelente herramienta para la clasificación filogenética, permitiendo identificar relaciones evolutivas incluso en organismos con una notable diversidad fenotípica, fisiológica o ecológica.

La implementación de metodologías de secuenciación de nueva generación (NGS) ha impulsado la generación de datos metagenómicos, brindando nuevas oportunidades para el estudio de la filogenia y la clasificación de organismos al proporcionar una mayor cantidad de información para analizar las relaciones evolutivas entre especies. En particular, la secuenciación de amplicones, que se centra en regiones específicas del genoma, como el rRNA 16S en procariontes, ha permitido una identificación más precisa y eficiente de especies en comunidades microbianas complejas. Esta disponibilidad de datos ha favorecido el uso de enfoques más avanzados, como los modelos de clasificación basados en Machine Learning (ML), que permiten procesar grandes volúmenes de datos y realizar clasificaciones de manera precisa. Los modelos de ML son capaces de identificar patrones complejos y hacer predicciones que serían difíciles de alcanzar con enfoques tradicionales, mejorando así la precisión de las clasificaciones y ofreciendo nuevas perspectivas en el campo de la taxonomía.

En este contexto, se explora el uso de datos metagenómicos obtenidos a partir de un proceso de secuenciación por amplicones, una método novedoso que permite obtener secuencias de las regiones hipervariables del rRNA 16S. El análisis se enfoca específicamente en las regiones V3-V4, utilizando el enfoque de extracción de características basadas en k-mers, el cual divide las secuencias en subsecuencias de longitud fija k. Esta representación de las secuencias, expresada como vectores de frecuencias, se aplica a modelos de ML para clasificar y visualizar patrones taxonómicos a diferentes niveles jerárquicos. Para facilitar la interpretación de los resultados, se integran técnicas de reducción de dimensionalidad, como t-distributed Stochastic Neighbor Embedding (t-SNE), que permiten la exploración de datos de alta dimensionalidad. De esta manera, se combina la potencia de las herramientas computacionales con técnicas genómicas para mejorar el proceso de clasificación de organismos, mostrando cómo el enfoque de k-mers no solo ofrece una representación eficiente de las secuencias genómicas, sino que también potencia el rendimiento de los modelos de ML, capturando información clave en la composición de nucleótidos. Los resultados obtenidos validan la capacidad de los modelos de ML para diferenciar entre géneros de bacterias, demostrando la efectividad de este enfoque en la clasificación taxonómica basada en datos metagenómicos, específicamente en secuencias cortas de rRNA 16S.

2 Métodos

2.1 Conjunto de Datos

Los datos metagenómicos utilizados en este estudio provienen del trabajo publicado por Fiannaca *et al.* (2018). Se tratan de datos simulados de un proceso de secuenciación de amplicones, basados en el gen rRNA 16S correspondiente a la subunidad ribosomal pequeña. Para la simulación, se seleccionaron únicamente las regiones hipervariables V3-V4 (aproximadamente 469 pares de bases), ya que estas regiones, menos conservadas, han sido ampliamente reportadas como discriminantes para la identificación y clasificación de especies.

El conjunto de datos consta de 28,000 secuencias cortas, las cuales se encuentran en formato FASTA, un formato comúnmente utilizado en bioinformática para almacenar secuencias de aminoácidos o nucleótidos. Los datos están disponibles públicamente en el siguiente URL: `http://tblab.pa.icar.cnr.it/public/BMC-CIBB_suppl/datasets/`.

2.2 Representación Mediante k-mers y Análisis de Frecuencias

Un enfoque ampliamente utilizado en bioinformática para el procesamiento de datos genéticos altamente complejos es la fragmentación de secuencias genómicas en subsecuencias consecutivas de longitud k, conocidas como k-mers. Este método permite el análisis de grandes secuencias genómicas con el objetivo de revelar información sobre la variación entre diferentes genomas 2.

Aunque este enfoque elimina la información de codificación de la secuencia, conserva detalles clave sobre la proporción de nucleótidos, lo cual es crucial en genómica. Aspectos como la proporción de GC se mantienen, lo que tiene importantes implicaciones biológicas relacionadas con la estabilidad del DNA, la expresión génica, y otros procesos celulares.

Para representar esta información de manera efectiva, se genera un vector de frecuencias que captura la frecuencia de cada subsecuencia en un espacio total de k-mers, de forma análoga al modelo de Bag-of-Words (BoW) utilizado en Procesamiento de Lenguaje Natural (NLP) 2. En el caso de BoW, cada palabra en un corpus se representa por su frecuencia en un conjunto determinado, sin tener en cuenta el orden de aparición. De manera similar, en el análisis de k-mers, cada subsecuencia k-mer es contada a lo largo de una secuencia genómica, creando un vector que refleja la frecuencia de aparición de cada k-mer en esa secuencia.

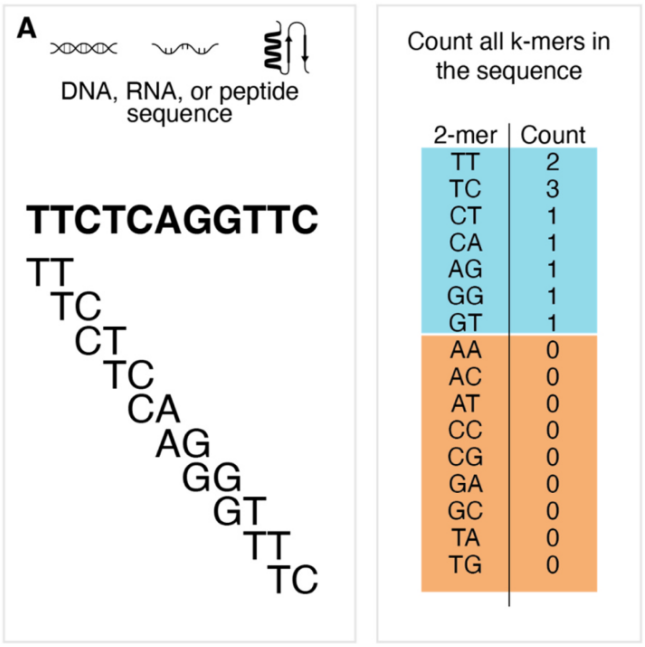

FIGURE 2. Posibles k-mers con $k = 2$. Para una secuencia especifica, el conteo de k-mers se realiza considerando el espacio total de k-mers, generando un vector de caracteristicas que captura la frecuencia de cada k-mer (C. Moeckel *et al.* (2020)).

La selección del valor k depende exclusivamente de la aplicación específica y del conjunto de datos, lo cual es fundamental para capturar patrones significativos. Para valores pequeños de k, puede ser insuficiente para representar o distinguir secuencias largas, ya que los k-mers posibles pueden aparecer repetidamente entre genomas y revelar patrones similares en diversas especies. Fiannaca *et al.* (2018) reportaron valores óptimos de $k = \{5, 6, 7\}$. En este trabajo, se eligió un valor de $k = 6$, con el objetivo de no solo preservar la información sobre la composición de nucleótidos, sino también de considerar los codones y trasladar la interpretación de las frecuencias de k-mers hacia el conteo de aminoácidos.

De esta manera, el vector de frecuencias generado actúa como una representación compacta de la secuencia genómica, permitiendo que patrones de composición de nucleótidos sean extraídos y utilizados por modelos de ML para tareas de clasificación, similar a cómo los modelos BoW se utilizan para clasificar y analizar texto.

2.3 Modelos de Machine Learning para la Visualización de Datos de Alta Dimensionalidad y Clasificación de Especies

La aplicación de modelos de ML en el análisis de datos metagenómicos ha transformado la forma en que se analizan las secuencias genéticas. Estos métodos no solo permiten manejar la alta dimensionalidad de los datos, identificando y visualizando patrones complejos, sino también realizar la clasificación de especies a partir de secuencias de DNA, RNA o péptidos. Al combinar técnicas de reducción de dimensionalidad con algoritmos de clasificación, se optimiza el procesamiento de datos metagenómicos, lo que contribuye significativamente al campo de la taxonomía y proporciona un enfoque innovador para la identificación precisa de especies.

2.3.1 Técnicas de Reducción de Dimensionalidad para la Exploración de Datos Genómicos

El análisis de secuencias genómicas genera datos de alta dimensionalidad debido a la gran diversidad de información contenida en los vectores de frecuencia generados a partir de los k-mers. Al representar una secuencia mediante k-mers, cada uno de ellos corresponde a una dimensión en un espacio vectorial de alta dimensión 2. Para ilustrar el problema de la dimensionalidad, consideremos los cuatro posibles nucleótidos: Adenina (A), Timina (T), Citosina (C) y Guanina (G). Para $k = 1$, existen $4^1 = 4$ posibles k-mers (A, T, G, C). Para $k = 2$, el número de posibles k-mers aumenta a $4^2 = 16$ (AA, AT, AG, etc.). De esta manera, la dimensionalidad crece de forma exponencial a medida que aumenta el valor de k. Técnicas como Análisis de Componentes Principales (PCA) o t-distributed Stochastic Neighbor Embedding (t-SNE) permiten proyectar estos datos en espacios de menor dimensión, preservando las relaciones intrínsecas entre los datos, con el objetivo de facilitar la visualización de patrones o agrupamientos en los datos.

Stochastic Neighbor Embedding (SNE) se enfoca en transformar la distancia Euclidiana entre pares de datos en el espacio de alta dimensión en probabilidades condicionales que representen similaridades. Para cualquier par de datos x_i y x_j, la similaridad se expresa como la probabilidad condicional $p_{j|i}$, que representa la probabilidad de que x_i seleccione a x_j como su vecino, basándose en una distribución Gaussiana centrada en x_i, con varianza proporcional a la distancia entre x_i y x_j.

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)}$$

Donde $\|x_i - x_j\|$ es la distancia euclidiana entre los puntos x_i y x_j, y σ_i es un parámetro de escala que depende de la densidad local alrededor del punto x_i.

De manera análoga, en el espacio de baja dimensión, es posible calcular la probabilidad condicional entre dos puntos y_i y y_j denotada como q_{ji}.

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|y_i - y_k\|^2\right)}$$

El objetivo de SNE es minimizar la diferencia entre las probabilidades en el espacio de alta dimensión y sus correspondientes en el espacio de baja dimensión. Se busca hacer las similaridades p_{ij} lo más semejantes posibles a las similaridades q_{ij}, lo cual se mide mediante la divergencia de Kullback-Leibler.

$$C = \sum_i KL(P_i\|Q_i) = \sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}$$

A diferencia de SNE, t-SNE utiliza la distribución t-Student en lugar de la distribución Gaussiana para calcular las similaridades entre dos puntos en el espacio de baja dimensión. Además, emplea una versión simétrica de la función de costo.

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Una de las principales ventajas de t-SNE sobre otros métodos como PCA es su capacidad para capturar relaciones no lineales entre los datos. A diferencia de PCA, que solo puede capturar relaciones lineales, t-SNE es capaz de manejar relaciones no lineales, lo que lo convierte en una herramienta especialmente útil para tareas de visualización de datos de alta dimensión. Sin embargo, una de las principales desventajas de t-SNE es su costo computacional elevado, especialmente cuando se trabaja con grandes volúmenes de datos. Además del ajuste de la perplejidad, un parámetro que influye en cómo se calculan las probabilidades de los puntos vecinos. La perplejidad afecta directamente la escala de la distribución de distancias y puede tener un gran impacto en los resultados. Un valor demasiado bajo puede hacer que t-SNE enfoque demasiado en las relaciones locales, mientras que un valor demasiado alto puede hacer que se pierdan las relaciones locales importantes y se enfoque más en la estructura global. Ajustar la perplejidad de manera adecuada es crucial para obtener buenos resultados en la visualización.

2.3.2 Modelos de Aprendizaje Supervisado para la Clasificación de Especies

En el campo de la taxonomía, la identificación y clasificación de organismos plantea un desafío debido al gran número de categorías existentes en cada nivel taxonómico. Por ejemplo, dentro del filo Proteobacteria se han descrito más de 450 géneros y más de 1600 especies. Trasladando este problema al contexto de ML, se convierte en un problema de clasificación multiclase. Para abordar este tipo de problemas, existen diversos modelos de clasificación como Árboles de decisiones, Bosques Aleatorios y Métodos de Ensamble. Además, muchos algoritmos diseñados para clasificación binaria, como la Regresión Logística y las Máquinas de Soporte Vectorial (SVM), pueden extenderse para aplicarse en problemas multiclase. La elección del modelo adecuado depende directamente de las características específicas del problema y del tipo de datos involucrados.

Para garantizar una comparación confiable del desempeño de los modelos y su capacidad de generalización, se emplean metodologías diseñadas para evaluar de manera robusta los algoritmos de ML. Un enfoque ampliamente utilizado es la Validación Cruzada (CV), que permite obtener estimaciones precisas del rendimiento del modelo, reduciendo el riesgo de que este se ajuste excesivamente a los datos de entrenamiento (sobreajuste), conduciendo a un desempeño deficiente en datos no vistos previamente. Entre los distintos métodos de CV, la Validación Cruzada k-fold es uno de los más utilizados. En este procedimiento, el conjunto de datos se divide en k subconjuntos (folds). En cada iteración, $k-1$ de estos subconjuntos se utilizan para entrenar el modelo, mientras que el subconjunto restante se emplea para validarlo. Este proceso se repite k veces, alternando los subconjuntos de validación. Finalmente, el rendimiento se calcula como el promedio de una métrica específica (e.g., accuracy, recall, AUC) obtenida a lo largo de las iteraciones. Aunque este método puede ser computacionalmente costoso, es especialmente útil cuando se trabaja con conjuntos de datos pequeños, ya que permite aprovechar al máximo la información disponible para entrenar y evaluar los modelos.

3 Resultados

3.1 Visualización de Patrones Taxonómicos Mediante t-SNE a partir de Vectores de Frecuencias de k-mers

t-SNE permite proyectar los datos en un espacio de menor dimensión para facilitar su visualización. En este caso, se utilizaron únicamente las dos primeras dimensiones generadas por el algoritmo. Las visualizaciones generadas con t-SNE permiten identificar patrones y agrupamientos en los datos, facilitando la distinción entre los grupos taxonómicos en distintos niveles jerárquicos.

Cada punto en la gráfica representa una observación, correspondiente a un vector de características derivado de una secuencia específica. Los puntos están codificados por colores y marcadores para reflejar su grupo taxonómico. En los resultados, se decidió omitir la visualización del nivel taxonómico inferior (i.e., género), ya que el elevado número de categorías en este nivel dificulta la identificación de agrupamientos claros. Sin embargo, al analizar los niveles taxonómicos de Clase y Orden, se observa que las observaciones tienden a agruparse y distribuirse según el grupo taxonómico al que pertenecen 3. En otras palabras, las observaciones que se encuentran más próximas en las visualizaciones suelen compartir el mismo nivel jerárquico superior. No obstante, es notable que la clase Betaproteobacteria se dispersa a lo largo del gráfico sin mostrar un patrón o agrupamiento claro 4.

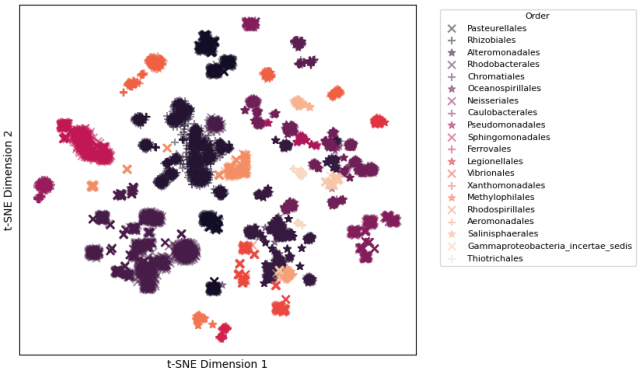

FIGURE 3. Reducción de dimensionalidad mediante t-SNE. Las observaciones están diferenciadas de acuerdo al nivel taxonómico de Orden.

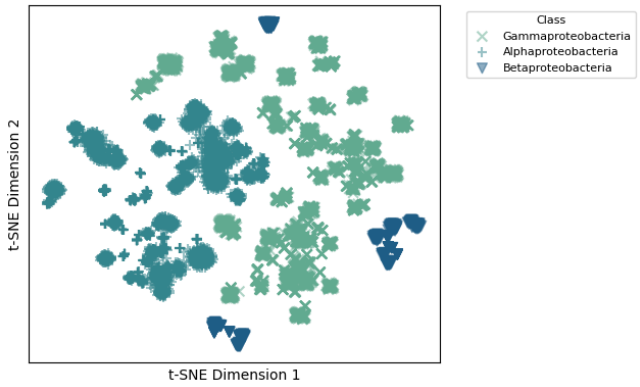

FIGURE 4. Reducción de dimensionalidad mediante t-SNE. Las observaciones están diferenciadas de acuerdo al nivel taxonómico de Clase.

Aunque las visualizaciones no evidencian agrupaciones definidas para todos los niveles jerárquicos, es importante considerar que solo se están considerando las dos primeras dimensiones generadas por t-SNE. Esto implica que parte de la información crucial para distinguir los niveles taxonómicos puede no estar reflejada en estas proyecciones. A pesar de esta limitación, las visualizaciones destacan la utilidad del enfoque basado en frecuencias de k-mers, demostrando que estas características contienen información suficiente para discriminar efectivamente entre grupos taxonómicos. Es importante mencionar que los modelos de ML utilizan como entrada el conjunto completo de vectores de características, lo que les permite aprovechar toda la información disponible en las frecuencias de k-mers para realizar clasificaciones precisas.

3.2 Evaluación del Desempeño de Modelos de Machine Learning para la Clasificación de Géneros de Bacterias

Para la evaluación del desempeño de los modelos de clasificación, se utilizó el método de Validación Cruzada k-Fold con 10 particiones. Este enfoque permitió obtener una estimación confiable de la capacidad de generalización de los modelos. A través de este proceso, se evaluó el rendimiento de los modelos en términos del *accuracy*, definida como la proporción de instancias correctamente clasificadas sobre el total de instancias en los datos.

$$Accuracy = \frac{Numero\ de\ Predicciones\ Correctas}{Numero\ Total\ de\ Predicciones}$$

Los resultados presentados en la tabla I destacan el buen desempeño general de los modelos de clasificación evaluados. Los modelos de Máquinas de Soporte Vectorial (SVM) y Bagging lograron la mayor precisión media (0.88), seguidos de cerca por los Bosques Aleatorios y el Perceptrón Multicapa (MLP), ambos con una precisión media de 0.87. Por su parte, el modelo basado en Extreme Gradient Boosting (XGBoost) mostró un desempeño competitivo, lo que reafirma la eficacia de los algoritmos de Gradient Boosting. En contraste, el modelo de Árboles de Decisión tuvo un rendimiento notablemente inferior, con una precisión media de aproximadamente 0.67.

Considerando un balance entre precisión y tiempo de ejecución, destacan los modelos SVM y Bosques Aleatorios como opciones particularmente eficientes para este problema. Aunque XGBoost alcanzó resultados competitivos en términos de precisión, su tiempo de ejecución promedio es considerablemente alto, con aproximadamente 250.47 segundos por fold, evaluado bajo un esquema de validación cruzada con $K = 10$.

TABLE I. Rendimiento Comparativo de Modelos de Clasificación Evaluados mediante Validación Cruzada K-Fold.

Modelo	Precisión Media	Tiempo de Ejecución
SVM	0.88	28.48
SVM (poly)	0.88	26.35
Decision Tree	0.67	1.75
MLP	0.87	94.74
Bagging	0.88	536.41
Random Forest	0.87	5.27
XGBoost	0.85	250.47

Los resultados se muestran mediante boxplots para facilitar su interpretación y comparar la distribución de las precisiones obtenidas por cada modelo a lo largo de los 10 folds del método CV K-fold. En los diagramas, se observa que los modelos SVM, Bagging y Bosques Aleatorios presentan distribuciones estrechas y centradas en valores altos, reflejando un rendimiento consistente y robusto. De manera similar, los modelos SVM (kernel polinomial), y MLP muestran distribuciones comparables, aunque con una ligera mayor dispersión. Por otro lado, XGBoost, a pesar de ser competitivo en términos de precisión, se centra a lo largo de un valor menor de precisión. Finalmente, el modelo basado en Árboles de Decisión exhibe una distribución más amplia y desplazada hacia valores bajos, confirmando su menor rendimiento 5.

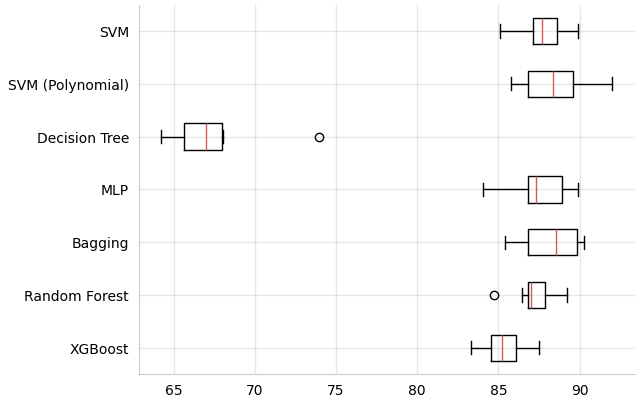

FIGURE 5. Distribución de los valores de la precisión obtenidos de los modelos de clasificación evaluados mediante CV K-fold con $K = 10$.

4 Discusión

Los resultados obtenidos destacan el desempeño de los modelos de clasificación aplicados a datos metagenómicos derivados de secuencias cortas del gen rRNA 16S, utilizando el enfoque de k-mers para la representación de las secuencias. Este método demostró ser eficaz para capturar patrones distintivos en los datos, los cuales fueron aprovechados de manera eficiente por los algoritmos evaluados. Además, los modelos lograron discriminar géneros de bacterias con un nivel significativo de precisión, resaltando su capacidad para discriminar instancias en este contexto específico.

En particular, los modelos SVM, Bosques Aleatorios y MLP destacaron entre los modelos evaluados, mostrando métricas de precisión favorables con tiempos de ejecución razonables. Por otro lado, si bien Bagging y XGBoost mostraron resultados competitivos en términos de precisión, su elevado costo computacional podría limitar su uso en estudios a gran escala. En contraste, el rendimiento reducido de los Árboles de Decisión evidencia que enfoques más simples son insuficientes para abordar la complejidad de los datos metagenómicos de alta dimensionalidad.

El desempeño de los modelos evaluados puede mejorarse mediante técnicas de ajuste de hiperparámetros, que permitirían identificar las configuraciones óptimas para cada algoritmo. Esta estrategia no solo incrementaría la precisión de los modelos, sino que también contribuiría a reducir la variabilidad observada en algunos casos, al tiempo que optimizaría su eficiencia computacional, como es el caso de XGBoosting.

5 Conclusiones

El enfoque de representar secuencias mediante vectores de frecuencias basados en k-mers, en combinación con modelos de clasificación de ML, es una estrategia eficiente para la identificación taxonómica de especies a partir de secuencias derivadas del gen rRNA 16S. Esto se fortalece por los métodos de reducción de dimensionalidad aplicados, donde únicamente dos dimensionaes son suficientes para distinguir claramente diferentes niveles jerárquicos entre las observaciones. Esto evidencia la capacidad del método para preservar de manera efectiva la información contenida en las secuencias.

Además, estos resultados destacan cómo las regiones hipervariables V3-V4 son suficientes para capturar y analizar relaciones filogenéticas en procariontes, reafirmando su utilidad en estudios metagenómicos. Estos hallazgos son particularmente relevantes en campos como la biología molecular, para el análisis de las regiones V1-V9, así como en microbiología y ecología microbiana, donde la identificación y clasificación taxonómica precisa son fundamentales para comprender la composición y dinámica de comunidades bacterianas complejas.

1. A. Fiannaca, et al., Deep learning models for bacteria taxonomic classification of metagenomic data, BMC Bioinformatics 18 (2017) S19, 10.1186/s12859-017-1971-6
2. C. Moeckel, et al., A survey of k-mer methods and applications in bioinformatics, Briefings in Bioinformatics 21 (2020) 594, 10.1093/bib/bbz073
3. L. van der Maaten and G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579
4. M. Dworkin, et al., eds., The Prokaryotes: A Handbook on the Biology of Bacteria, Third Edition, Volume 5: Proteobacteria: Alpha and Beta Subclasses (Springer, New York, NY, 2006), pp. 1–1234, 10.1007/0-387-30745-1, URL https://doi.org/10.1007/0-387-30745-1.
5. M. Wattenberg, F. Viégas, and I. Johnson, How to Use t-SNE Effectively, Distill (2016), 10.23915/distill.00002
6. S. learn Contributors, Cross-validation: evaluating estimator performance, https://scikit-learn.org/1.5/modules/cross_validation.html (2024), Accessed: 2024-11-26.