Modelo predictivo de resistencia antibiótica en bacterias bucales mediante análisis fenotípico y taxonómico.

Ana Luisa Sotelo Ariza 8993444.

Jorge Ivan Barrera Salgado 8993445.

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos.

Dirigido por

Fabian Tobar Tosse.

Pontificia Universidad Javeriana Cali

Facultad de Ingeniería y Ciencias.

Maestría en Ciencia de datos.

Santiago de Cali

DICIEMBRE 1 DE 2025.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1 INTRODUCCIÓN

La resistencia antibiótica se ha consolidado como una de las amenazas más críticas para la salud pública global, al favorecer la aparición de infecciones difíciles de tratar, prolongar los tiempos de hospitalización y aumentar la morbilidad y mortalidad asociadas a enfermedades infecciosas. La Organización Mundial de la Salud (OMS) advierte que la proliferación de bacterias multirresistentes compromete seriamente la eficacia de los tratamientos disponibles y demanda el desarrollo de estrategias analíticas capaces de anticipar comportamientos clínicos emergentes [1]. En este escenario, la detección temprana de patrones de resistencia dentro de microbiomas complejos constituye un componente esencial para la vigilancia epidemiológica y para la toma de decisiones informadas en el ámbito clínico.

Históricamente, la determinación de resistencia se ha basado en cultivos microbiológicos y pruebas fenotípicas, métodos que, si bien son precisos, resultan limitados frente a la vasta diversidad bacteriana y requieren tiempos de procesamiento prolongados. El avance de las tecnologías bioinformáticas ha impulsado enfoques que integran grandes volúmenes de datos taxonómicos, genómicos y fenotípicos. En particular, los métodos de aprendizaje supervisado permiten identificar patrones no evidentes a simple vista y modelar relaciones entre características bacterianas y perfiles de resistencia, lo que abre nuevas posibilidades para la predicción temprana y la caracterización computacional de riesgos microbiológicos.

En este proyecto se integró un conjunto de datos proveniente del repositorio BV-BRC, compuesto por información taxonómica (familia, género y especie), metadatos genómicos y perfiles fenotípicos de susceptibilidad a antibióticos. A partir de un proceso de depuración, estandarización y unificación de estas fuentes, se construyó un dataset consolidado que permitió realizar análisis exploratorios enfocados en la distribución taxonómica de las bacterias bucales y su correspondencia con los niveles de resistencia observados. Posteriormente, se implementó un modelo de regresión logística binaria empleando codificación categórica, validación mediante división train–test y métricas de desempeño como exactitud, precisión, recall y F1-score. El modelo alcanzó un accuracy del 75.2 % y un F1-score de 0.49, resultados que reflejan tanto el desbalance entre clases como la complejidad inherente a la predicción de resistencia antibiótica. No obstante, el análisis de los coeficientes permitió identificar señales fenotípicas y taxonómicas asociadas a niveles diferenciados de resistencia.

Los resultados demuestran la viabilidad de este enfoque computacional para caracterizar tendencias de resistencia antibiótica en bacterias bucales y constituyen una base metodológica para el desarrollo de modelos predictivos más robustos, con aplicaciones potenciales en vigilancia epidemiológica y análisis microbiológico basado en datos.

2 DEFINICIÓN DEL PROBLEMA

2.1 PLANTEAMIENTO DEL PROBLEMA

La resistencia antibiótica constituye una de las mayores amenazas para la salud pública mundial, debido al incremento de bacterias capaces de evadir los tratamientos antimicrobianos convencionales y a su impacto directo en la morbimortalidad global [1]. El ecosistema bucal, caracterizado por una elevada diversidad microbiana, incluye especies que pueden actuar como reservorios de resistencia y participar en infecciones oportunistas cuando los mecanismos defensivos del huésped se ven comprometidos [2]. En este contexto, comprender cómo las características biológicas de estas bacterias se relacionan con sus perfiles de resistencia resulta fundamental para apoyar la vigilancia epidemiológica y optimizar la toma de decisiones clínicas.

Aunque las pruebas fenotípicas de susceptibilidad continúan siendo el método estándar para determinar resistencia antibiótica, su implementación requiere infraestructura especializada, condiciones de laboratorio estrictas y puede implicar tiempos prolongados para la obtención de resultados [3]. Paralelamente, el crecimiento de plataformas bioinformáticas como BV-BRC ha facilitado el acceso a grandes volúmenes de información taxonómica, fenotípica y de muestras clínicas [4]. Sin embargo, estas fuentes presentan desafíos importantes para su utilización directa, tales como duplicados, inconsistencias, ausencia de estandarización y datos faltantes, lo que demanda procesos rigurosos de depuración e integración antes de su uso analítico [5].

Diversos estudios han mostrado que la taxonomía bacteriana puede asociarse con patrones diferenciales de resistencia, evidenciando que componentes como la familia, el género o la especie pueden servir como marcadores útiles para anticipar el comportamiento fenotípico frente a antibióticos [6]. No obstante, aún es limitada la disponibilidad de modelos computacionales que integren simultáneamente características taxonómicas y fenotípicas para predecir resistencia antibiótica en bacterias bucales. Esta brecha dificulta el aprovechamiento del potencial analítico de los datos disponibles y limita el desarrollo de herramientas predictivas que contribuyan a la identificación temprana de riesgos microbiológicos.

Ante este escenario, surge la necesidad de diseñar un enfoque computacional que permita integrar y depurar datos provenientes de múltiples fuentes, analizar la relación entre fenotipo

y taxonomía, y evaluar la capacidad de modelos supervisados para predecir resistencia antibiótica. El problema central radica en determinar si es posible estimar la resistencia de aislamientos bacterianos bucales a partir de características fenotípicas y taxonómicas, y establecer cuál es el desempeño alcanzable por un modelo predictivo bajo las condiciones reales del dataset consolidado.

2.2 FORMULACIÓN DEL PROBLEMA

¿Es posible predecir la resistencia antibiótica en bacterias bucales mediante la integración y análisis de características fenotípicas y taxonómicas provenientes de bases de datos bioinformáticas como BV-BRC, utilizando técnicas de aprendizaje supervisado aplicadas sobre un dataset previamente depurado y estandarizado?

2.2.1 PREGUNTAS DE SISTEMATIZACIÓN

- ¿Cómo integrar, limpiar y estandarizar las diferentes fuentes de datos provenientes de BV-BRC para construir un dataset coherente y apto para análisis predictivo?

- ¿Qué patrones descriptivos se observan en las variables clínicas, taxonómicas y fenotípicas de las bacterias bucales incluidas en el dataset consolidado?

- ¿Qué relación existe entre los niveles taxonómicos (familia, género y especie) y los fenotipos de resistencia antibiótica registrados en el conjunto de datos?

- ¿Qué desempeño logra un modelo de regresión logística al predecir la resistencia antibiótica de bacterias bucales a partir de características fenotípicas y taxonómicas?

3 OBJETIVOS DEL PROYECTO

3.1 OBJETIVO GENERAL

Implementar un modelo predictivo supervisado para estimar la resistencia antibiótica en bacterias bucales mediante la integración, depuración y análisis de datos fenotípicos y taxonómicos provenientes de BV-BRC, evaluando su desempeño a través de métricas de clasificación.

3.2 OBJETIVOS ESPECÍFICOS

- Integrar, limpiar y normalizar las diferentes fuentes de información provenientes de BV-BRC y de muestras clínicas, con el fin de construir un conjunto de datos estructurado, coherente y apto para análisis estadístico y modelado predictivo.

- Realizar el análisis exploratorio de las variables clínicas, taxonómicas y fenotípicas para identificar patrones relevantes en la distribución de bacterias bucales y en los niveles de resistencia antibiótica.

- Examinar la relación entre la clasificación taxonómica de las bacterias (familia, género y especie) y los fenotipos de resistencia antibiótica, con el propósito de identificar asociaciones significativas entre ambos componentes.

- Implementar un modelo predictivo supervisado, basado en regresión logística, para estimar la probabilidad de resistencia antibiótica en bacterias bucales y evaluar su desempeño mediante métricas de validación.

4 MARCO TEÓRICO Y ANTECEDENTES

4.1 MARCO TEÓRICO

En los últimos años, el estudio de las comunidades microbianas y su relación con la salud humana ha avanzado significativamente gracias al acceso a bases de datos bioinformáticas y al desarrollo de algoritmos de aprendizaje automático. Aunque el análisis genómico ha sido ampliamente utilizado en investigaciones moleculares, la disponibilidad de datos taxonómicos y fenotípicos estructurados permite abordar problemas clínicos clave sin depender de técnicas de secuenciación complejas. En este contexto, el machine learning se ha consolidado como una herramienta poderosa para analizar grandes volúmenes de datos biológicos, identificar patrones relevantes y apoyar la predicción de fenómenos como la resistencia antibiótica [12][13].

Este marco teórico presenta los conceptos fundamentales que sustentan este proyecto, incluyendo la resistencia antimicrobiana, la taxonomía bacteriana, los fenotipos de susceptibilidad antibiótica, la integración de datos bioinformáticos y el uso de algoritmos supervisados como la regresión logística. Cada uno de estos elementos aporta perspectivas clave para comprender y modelar el comportamiento de bacterias bucales frente a agentes antimicrobianos.

Taxonomía bacteriana

La taxonomía bacteriana permite organizar y clasificar a los microorganismos en niveles jerárquicos como familia, género y especie. Esta estructura facilita la identificación y comunicación precisa de las características biológicas de cada organismo [7][10]. Estudios recientes han mostrado que ciertos grupos taxonómicos presentan tendencias específicas de susceptibilidad o resistencia antibiótica, lo que los convierte en marcadores útiles para análisis predictivos [6][8][9]. Bases como NCBI Taxonomy proporcionan la nomenclatura estandarizada empleada en investigaciones biomédicas [7].

Resistencia antibiótica

La resistencia antibiótica es la capacidad de un microorganismo para sobrevivir ante la presencia de un fármaco diseñado para inhibirlo o eliminarlo. La Organización Mundial de la Salud ha catalogado este fenómeno como una de las principales amenazas para la salud global [1]. En el entorno bucal, la diversidad microbiana genera un ecosistema dinámico en el que pueden coexistir bacterias con perfiles diferenciales de resistencia [2]. Comprender estas dinámicas resulta fundamental para la vigilancia epidemiológica y para orientar decisiones clínicas.

Fenotipos de susceptibilidad antibiótica

Las pruebas de susceptibilidad fenotípica representan el estándar para clasificar a un aislamiento en categorías como susceptible, intermedio o resistente. Estas metodologías se encuentran estandarizadas bajo guías internacionales como CLSI, lo que permite comparar resultados entre estudios y establecer puntos de corte uniformes [3]. En aplicaciones computacionales, es común transformar estas categorías en un esquema binario (resistente/no resistente) para facilitar el modelado estadístico y algorítmico.

Integración de datos bioinformáticos

Plataformas como BV-BRC consolidan datos provenientes de múltiples fuentes clínicas y experimentales, incluyendo taxonomía, metadatos de aislamiento y fenotipos de resistencia [4]. Sin embargo, integrar estos datos requiere procesos de limpieza, normalización, estandarización y eliminación de duplicados debido a su heterogeneidad y las inconsistencias inherentes a su recolección. Investigaciones recientes destacan la necesidad de metodologías robustas para integrar datos heterogéneos y permitir un análisis confiable en estudios de resistencia antimicrobiana [5].

Algoritmos de Machine Learning aplicados a microbiología

El aprendizaje supervisado se ha convertido en una herramienta clave para analizar datos biológicos y generar predicciones en sistemas complejos. Modelos como árboles de decisión, Random Forest, SVM o regresión logística permiten identificar relaciones entre variables taxonómicas y fenotípicas, y predecir comportamientos clínicos relevantes [12][13]. En el ámbito microbiológico, estos algoritmos se han empleado para clasificar bacterias, predecir resistencia y detectar patrones difíciles de identificar mediante análisis estadísticos tradicionales.

Métricas de evaluación del modelo

El desempeño del modelo de clasificación se evaluó utilizando métricas derivadas de la matriz de confusión, ampliamente aceptadas en problemas de clasificación binaria en el área biomédica [14].

A partir de los valores de verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN), se definieron las siguientes métricas:

Exactitud (Accuracy):

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}$$

Precisión (Precision):

$$\text{Precisión} = \frac{VP}{VP + FP}$$

Recall (Sensibilidad):

$$\text{Recall} = \frac{VP}{VP + FN}$$

F1-score:

$$\text{F1-score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}}$$

Estas métricas permiten evaluar el rendimiento del modelo desde distintas perspectivas, siendo especialmente relevantes en el análisis de resistencia antimicrobiana, donde los errores de clasificación pueden tener consecuencias clínicas significativas [1].

Regresión logística binaria

La regresión logística es un algoritmo de aprendizaje supervisado ampliamente empleado en problemas de clasificación binaria, especialmente en el ámbito biomédico, debido a su simplicidad, interpretabilidad y fundamento estadístico [15]. Este modelo estima la probabilidad de pertenencia a una clase mediante la función logística o sigmoide.

Matemáticamente, la regresión logística se define como:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

donde:

- $P(y = 1 \mid \mathbf{x})$es la probabilidad de que una observación pertenezca a la clase positiva,
- $\beta_0$corresponde al término independiente,
- $\beta_i$representan los coeficientes asociados a las variables predictoras x_i.

En este trabajo, se implementó un modelo de regresión logística binaria, utilizando codificación categórica para las variables taxonómicas y fenotípicas. La validación del modelo se realizó mediante una división *train–test*, estrategia común para evaluar la capacidad de generalización del modelo sobre datos no utilizados durante el entrenamiento [12].

Posibles Modelos en Base al Marco Teórico y Referencias Aplicadas en el Proyecto

1. **Modelo de clasificación**: Se implementó un modelo supervisado de clasificación para identificar patógenos y categorizar muestras de metagenomas según su resistencia antibiótica. *Random Forest* y *XGBoost* demostraron ser los más eficaces.

2. **Modelo de regresión:** Se utilizó un enfoque de regresión logística como modelo de clasificación binaria para analizar la relación entre características fenotípicas y la respuesta a antibióticos, permitiendo estimar la probabilidad de pertenencia a una clase (susceptible o resistente) y apoyar la predicción de la efectividad terapéutica.
3. **Modelo de detección de anomalías**: Aunque no fue el enfoque central, se valoró la utilidad de modelos como **Isolation Forest** para detectar patrones genómicos inusuales que podrían estar vinculados a nuevas formas de resistencia.

4.2 ANTECEDENTES

El estudio del ecosistema microbiano de la cavidad bucal ha permitido identificar una amplia diversidad de especies bacterianas, cuya clasificación taxonómica ha sido fortalecida por bases de referencia como NCBI Taxonomy y los avances recientes en la estandarización de nomenclatura [7]–[11]. La caracterización de estos microorganismos ha mostrado que distintas especies presentan comportamientos diferenciados frente a agentes antimicrobianos, lo que ha motivado investigaciones orientadas a analizar la relación entre taxonomía bacteriana y perfiles de susceptibilidad.

En paralelo, la estandarización de las pruebas de susceptibilidad fenotípica ha sido documentada ampliamente por entidades internacionales, permitiendo que los resultados obtenidos en estudios clínicos sean comparables y utilizables en análisis posteriores [3]. Estos fenotipos se han convertido en una fuente clave para investigaciones que buscan evaluar patrones de comportamiento bacteriano sin depender de datos genómicos complejos.

La consolidación de plataformas bioinformáticas como BV-BRC ha facilitado la disponibilidad de información taxonómica y fenotípica proveniente de diversos laboratorios e instituciones de investigación [4]. No obstante, estudios recientes enfatizan las dificultades inherentes a la integración de datos heterogéneos, especialmente cuando se combinan fuentes con distintos formatos, niveles de precisión o estándares de curación [5]. Estas limitaciones han motivado la creación de metodologías de depuración y estandarización orientadas a mejorar la calidad de los datos utilizados en modelos analíticos.

Desde el campo del aprendizaje supervisado, se han desarrollado modelos predictivos aplicados a microbiología para tareas de clasificación, predicción de resistencia y análisis de patrones. Modelos como la regresión logística, los métodos basados en árboles y otros algoritmos clásicos han demostrado utilidad para generar predicciones interpretables a partir de variables categóricas y clínicas [12][13]. Aunque gran parte de la literatura se centra en datos moleculares o genómicos, los estudios basados exclusivamente en taxonomía y fenotipo siguen siendo menos frecuentes, lo que evidencia una oportunidad de aportar nuevo conocimiento mediante enfoques más accesibles y reproducibles.

5 DISEÑO Y ARQUITECTURA DEL PROYECTO

5.1 DISEÑO DE LA ARQUITECTURA

El presente proyecto se estructuró bajo un enfoque de Ciencia de Datos aplicada a la bioinformática, con el objetivo de desarrollar un modelo de aprendizaje automático capaz de predecir la resistencia antibiótica en bacterias del microbioma bucal a partir de la integración y análisis de características taxonómicas y fenotípicas obtenidas de bases de datos especializadas como BV-BRC.

Para alcanzar este propósito, se diseñó una arquitectura modular e integrada que garantiza la trazabilidad completa desde la adquisición, depuración y estandarización de las fuentes de datos hasta la construcción de modelos supervisados interpretables, orientados a identificar patrones de resistencia asociados a la clasificación biológica de los microorganismos. El sistema se concibió como un flujo end-to-end, compuesto por las siguientes fases principales:

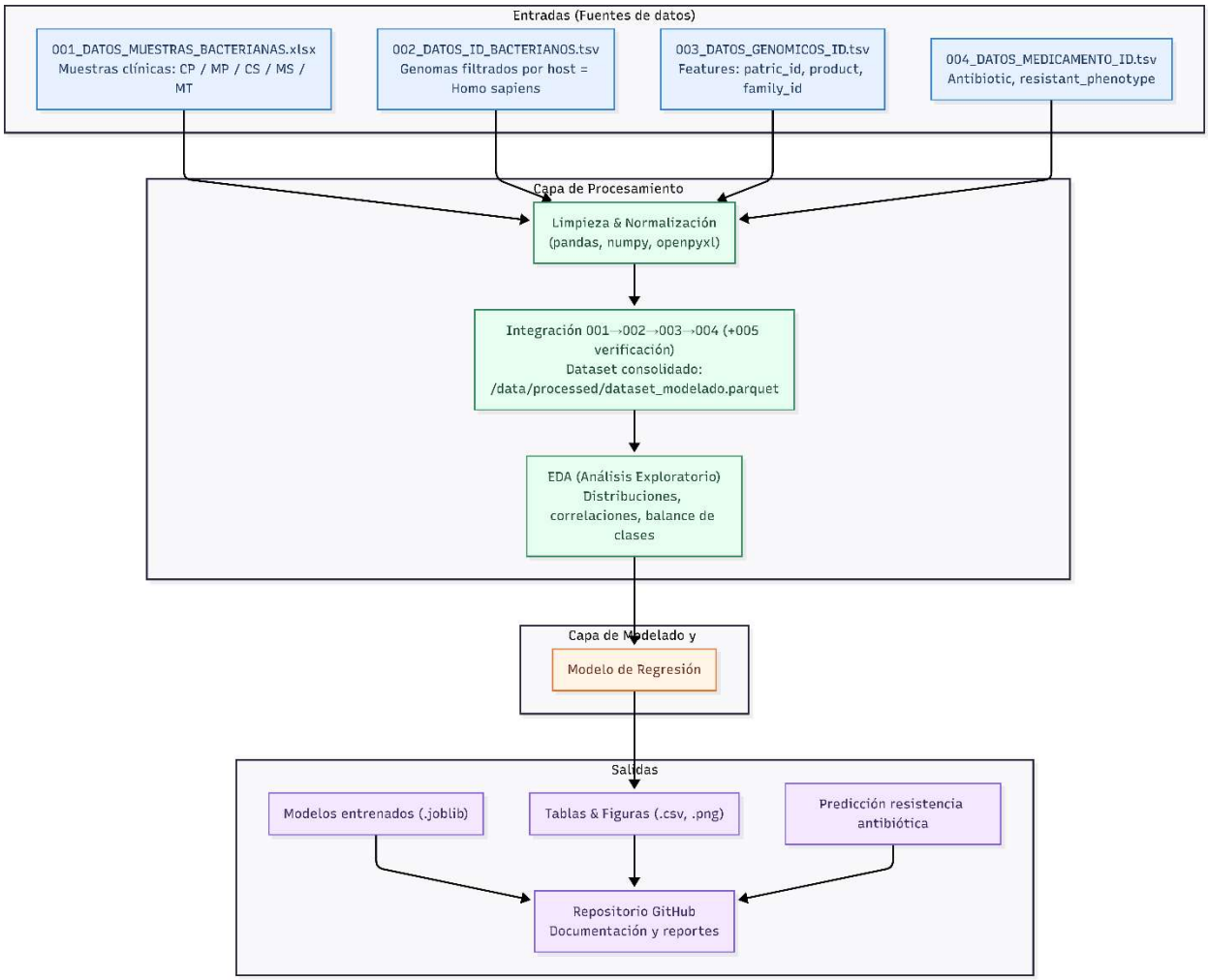


Figura 1: Diagrama de Arquitectura.

5.1.1 ENTRADAS (FUENTES DE DATOS)

Incluye la recopilación de muestras clínicas (placa dental, saliva y tejido tumoral) y la descarga de datos genómicos desde la base *BV-BRC*, empleando su interfaz de línea de comandos `(p3-scripts)`.
En esta etapa se generaron los archivos 002 a 004, que contienen información biológica, taxonómica y fenotípica de las bacterias analizadas.

5.1.2 CAPA DE PROCESAMIENTO

A través de scripts en Python, se normalizan y unifican los archivos obtenidos, estandarizando nombres de variables, limpiando valores inconsistentes y vinculando las diferentes fuentes mediante el identificador único genome_id. Durante este proceso se eliminan duplicados, se homogenizan formatos y se asegura la integridad entre los datos taxonómicos y los fenotípicos.

El resultado es un conjunto de datos consolidado que constituye la base para la construcción del dataset de entrenamiento utilizado en el modelo predictivo.
Asimismo, se realiza un análisis exploratorio en el cual se aplican técnicas descriptivas y de minería de datos para identificar patrones de resistencia, distribuciones de especies y relaciones entre las distintas categorías taxonómicas y los fenotipos antibióticos.

5.1.3 CAPA DE MODELADO Y EVALUACIÓN

En esta capa se implementa un modelo supervisado de regresión logística orientado a predecir la resistencia antibiótica a partir de variables taxonómicas y fenotípicas previamente depuradas. Para ello, se realiza la codificación de las variables categóricas mediante técnicas como One-Hot Encoding, seguida de la partición del dataset en conjuntos de entrenamiento y prueba.

La evaluación del modelo se lleva a cabo mediante métricas de clasificación, incluyendo accuracy, recall y F1-score, lo que permite determinar su capacidad para discriminar entre aislamientos resistentes y no resistentes. Adicionalmente, se analizan los coeficientes del modelo con el fin de identificar la contribución de cada especie, género o familia a la probabilidad de resistencia, aportando interpretabilidad al proceso predictivo.

5.1.4 SALIDAS

Finalmente, los resultados del proceso se organizan en carpetas estructuradas dentro del repositorio del proyecto, incluyendo:

- Dataset final consolidado.

- Métricas de desempeño del modelo.
- Gráficas generadas durante el análisis exploratorio.
- Coeficientes e interpretaciones del modelo.
- Reportes técnicos necesarios para la documentación del flujo analítico.

Esta estructura garantiza la trazabilidad, reproducibilidad y accesibilidad del proyecto

5.2 CONFIGURACIÓN DEL ENTORNO

5.2.1 HERRAMIENTAS UTILIZADAS

Para la implementación del proyecto se emplearon herramientas de código abierto que permiten garantizar la **reproducibilidad**, **eficiencia** y **trazabilidad** de todo el proyecto. A continuación, se describen las principales:

- **Anaconda:** Utilizada para la gestión del entorno de desarrollo, instalación de dependencias y control de versiones de librerías. Su uso permitió mantener un entorno aislado y reproducible durante todo el ciclo del proyecto.
- **Python 3.10:** Lenguaje base para la ejecución de los procesos de limpieza, análisis exploratorio y modelado de datos, gracias a su versatilidad y amplia disponibilidad de librerías científicas.
- **Git y GitHub:** utilizados para el control de versiones del código, documentación y datos. Se trabajó mediante ramas (main, dev, experiment) y commits descriptivos que aseguran la
- **Mermaid:** Herramienta de diagramación empleada para la representación visual de la arquitectura del sistema, flujos de datos y relaciones entre procesos. Su integración en el repositorio facilita la documentación estructurada y la comprensión de la arquitectura del proyecto.
- **BV-BRC Command Line Interface (p3-scripts):** La línea de comandos de BV-BRC (p3-scripts) fue una de las herramientas principales utilizadas para la adquisición de datos en este proyecto. Esta interfaz permite realizar consultas automatizadas al Bacterial and Viral Bioinformatics Resource Center (BV-BRC) y extraer información estructurada en formato .tsv, incluyendo metadatos de aislamiento, taxonomía (familia, género, especie) y fenotipos de resistencia antibiótica.

  Su uso facilitó el proceso de descarga y estandarización de datos, permitiendo filtrar registros, seleccionar campos específicos y recuperar únicamente la información

relevante para el análisis fenotípico y taxonómico. Entre sus principales ventajas se destacan:

- Automatización reproducible: las consultas pueden ejecutarse mediante scripts, asegurando consistencia en la extracción de datos.
- Acceso estructurado a múltiples tablas: comandos como
  - p3-all-genomes,
  - p3-get-genome-metadata,
  - p3-get-genome-drugs
- Permiten obtener información específica de manera rápida y organizada.
- Infraestructura bioinformática integrada: BV-BRC mantiene datos curados provenientes de diversas instituciones, lo cual garantiza calidad y actualización constante.
- Compatibilidad con flujos de análisis: las salidas generadas en .tsv se integran fácilmente con librerías de Python utilizadas en el preprocesamiento y análisis exploratorio.
- El proyecto utilizó exclusivamente los recursos de BV-BRC relacionados con taxonomía, metadatos del aislamiento y perfiles fenotípicos de susceptibilidad antibiótica, sin incluir información genómica detallada como secuencias, contigs, anotaciones de genes o familias de proteínas, dado que estos niveles de información no formaban parte del alcance definido.

¿Cómo se estructura?

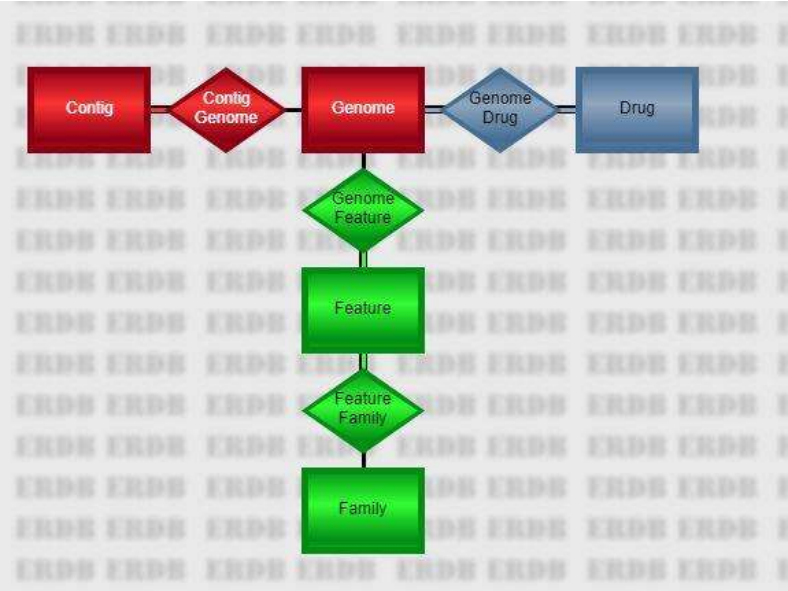


Figura 2: Modelo base BV-BRC.

- **Genome (Genoma)**

Un **genoma** corresponde al conjunto completo de secuencias de ADN y sus respectivas anotaciones, que representan la mejor estimación del material genético de un organismo.

El BV-BRC provee múltiples herramientas para acceder a la información genómica mediante su *CLI*, entre las que se destacan:

- *p3-all-genomes*: lista todos los genomas disponibles o un subconjunto filtrado.
- *p3-get-genome-data*: obtiene metadatos de genomas específicos.
- *p3-get-genome-features*: accede a las características (genes, regiones o elementos funcionales) asociadas a cada genoma.
- *p3-get-genome-contigs*: recupera las secuencias de ADN correspondientes a cada genoma.
- *p3-get-genome-drugs*: extrae información relacionada con la resistencia antimicrobiana vinculada a los genomas.

En los archivos exportados, las columnas asociadas a esta entidad utilizan el prefijo **genome.**

Ejemplo: *genome.genome_name* → nombre del genoma bacteriano.

- **Contig**

Un **contig** es una secuencia continua de ADN dentro de un genoma. Puede representar un **cromosoma completo**, un **plásmido** o un **fragmento parcial de ADN** ensamblado.
Los contigs se obtienen a partir de los identificadores de genoma mediante el comando:

- *p3-get-genome-contigs*

Las columnas asociadas a esta entidad utilizan el prefijo **contig.**

Ejemplo: *contig.length* → longitud del contig en pares de bases.

- **Drug (Fármaco o agente antimicrobiano)**

La entidad **Drug** almacena la información relacionada con los **fármacos antimicrobianos** utilizados en tratamientos y pruebas de resistencia. Constituye la base de los datos de resistencia antimicrobiana en BV-BRC y permite vincular cada fármaco con los genomas en los que se ha identificado resistencia o sensibilidad, los comandos utilizados para obtener la información:

- *p3-all-drugs*: lista todos los fármacos registrados.
- *p3-get-drug-genomes*: obtiene datos de resistencia a dichos fármacos en distintos genomas.

Las columnas correspondientes emplean el prefijo **drug.**

Ejemplo: *drug.molecular_formula* → fórmula molecular del compuesto.

- **Feature (Característica o elemento genómico)**

Una **feature** representa una región funcional de interés dentro del genoma, que puede corresponder a:

- Un **gen** (zona que codifica una proteína),
- Un **sitio de ARN**,
- Una **secuencia CRISPR**, o
- Una **región reguladora**.

Cada feature pertenece exclusivamente a un único genoma, aunque puede estar distribuida en varios contigs, algunos de los comandos que se pueden usar son:

- *p3-get-genome-features*: obtiene las características de uno o varios genomas.
- *p3-get-family-features*: lista las features asociadas a familias de proteínas.
- *p3-get-feature-data*: recupera información detallada de una feature específica.

El identificador único de cada feature es *patric_id*, y las columnas de salida emplean el prefijo **feature**.

Ejemplo: *feature.location* → ubicación de la feature dentro del genoma.

- **Family (Familia de proteínas)**

Una **familia** agrupa proteínas o features que se consideran **homólogas isofuncionales**, es decir, que comparten similitud estructural y funcional. Este nivel de información permite estudiar relaciones evolutivas y patrones de funcionalidad conservados entre diferentes especies bacterianas, algunos comandos para obtener la información son:

- *p3-get-family-features*: obtiene las features pertenecientes a una o varias familias.
- *p3-get-family-data*: recupera la información general de dichas familias.

Las columnas correspondientes a esta entidad utilizan el prefijo **family.**

Ejemplo: *family.product* → producto o función asociada a la familia proteica.

5.2.2 LIBRERÍAS DE PYTHON

El entorno incluye librerías de uso general y especializado para el análisis, modelado y visualización de datos, en la Tabla1 se describen las principales librerías utilizadas y su función dentro del flujo de trabajo.

LIBRERÍA	FUNCIÓN PRINCIPAL
PANDAS	Manipulación y limpieza de datos tabulares (.csv, .tsv, .xlsx).
NUMPY	Cálculo numérico, manejo de matrices y arrays multidimensionales.
SCIPY	Aplicación de pruebas estadísticas y métricas matemáticas.
SCIKIT-LEARN	Implementación de modelos supervisados (Regresión Logística, *Random Forest, XGBoost*) y no supervisados (*K-Means, PCA*).
MATPLOTLIB / SEABORN	Visualización de datos y resultados (gráficos, matrices de confusión, distribuciones).
OPENPYXL	Lectura de archivos Excel (.xlsx) y conversión a estructuras pandas.
JOBLIB	Serialización y almacenamiento de modelos entrenados.
TQDM	Seguimiento del progreso de tareas en bucles (entrenamiento).
PYARROW / FASTPARQUET	Almacenamiento eficiente en formato *Parquet* para datasets procesados.

Tabla 1: Descripción librerías Python.

5.3 GESTIÓN DEL CONTROL DE VERSIONES

Con el fin de garantizar una adecuada gestión del código, la trazabilidad de los cambios y la colaboración sincrónica del equipo de trabajo se procedió a la creación de un repositorio en la plataforma GitHub bajo el nombre:
PROYECTO_GRADO_MAESTRIA_DATOS_2025
https://github.com/anaariza1119/PROYECTO_GRADO_MAESTRIA_DATOS_2025
Este repositorio constituye el espacio centralizado donde se almacenan y versionan los datos, scripts, notebooks, experimentos y documentación relacionados con el proyecto titulado:
"Modelo predictivo de resistencia antibiótica en bacterias bucales mediante análisis fenotípico y taxonómico.".

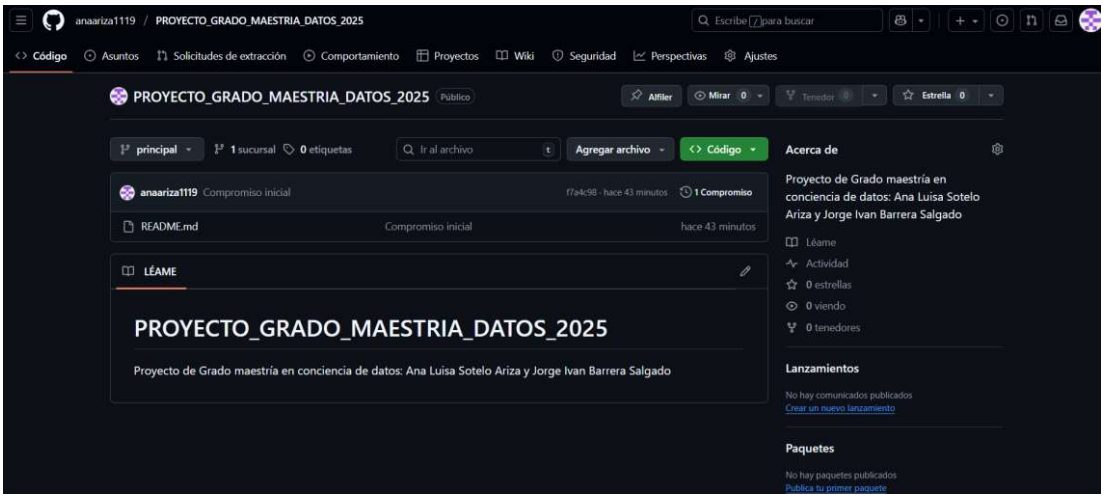


Figura 3: Creación repositorio.

5.3.1 ORGANIZACIÓN

La organización del repositorio responde a las buenas prácticas de ingeniería de software y ciencia de datos, lo cual permite una adecuada separación entre datos, código, reportes y resultados experimentales.

- **data/:** Contiene los conjuntos de datos en sus diferentes estados: crudos **(raw)**, procesados y fuentes externas suministradas por el director de proyecto (external).
- **notebooks/:** Almacena los cuadernos de análisis exploratorio-utilizados durante las fases iniciales de experimentación.
- **src/:** Concentra el código fuente del proyecto, organizado en subcarpetas para el preprocesamiento de datos, la implementación de modelos, la evaluación de resultados, la generación de visualizaciones y utilidades de apoyo.

- **experiments/:** Incluye los resultados de los experimentos realizados, así como los registros de ejecución, métricas de validación y modelos entrenados guardados.
- **reports/:** Concentra la documentación formal, incluyendo figuras, gráficos y el informe final de la investigación.
- **tests/:** Reservado para la ejecución de pruebas de código que respalden la calidad y consistencia del desarrollo.
- **Archivos base:** En la raíz del repositorio se encuentran documentos fundamentales como README.md, .gitignore, requirements.txt y environment.yml, que describen el proyecto, definen exclusiones de control de versiones y especifican las dependencias necesarias para la reproducción de los experimentos.

6 OBTENCIÓN Y PREPARACIÓN DE LOS DATOS

6.1 OBJETIVO DEL CAPITULO

Integrar, limpiar y normalizar las diferentes fuentes de información provenientes de BV-BRC y de muestras clínicas, con el fin de construir un conjunto de datos estructurado, coherente y apto para análisis estadístico y modelado predictivo.

6.1.1 DESCRIPCIÓN

La fase de obtención y preparación de los datos constituye el punto de partida del proceso analítico, pues permite construir un conjunto de datos estructurado, confiable y trazable que servirá como base para los modelos de aprendizaje automático.

En este proyecto, todas las fuentes de información fueron obtenidas directamente del repositorio *Bacterial and Viral Bioinformatics Resource Center (BV-BRC)* mediante su interfaz de línea de comandos (*Command Line Interface – CLI)*, conocida como `p3-scripts`.

Cada consulta permitió recuperar información funcional y fenotípica de bacterias aisladas de muestras humanas del microbioma bucal (*host_name = Homo sapiens*). A continuación, se describen las cuatro fuentes de datos utilizadas, el proceso de limpieza, normalización e integración, y la estructura final del *dataset* consolidado.

6.2 OBTENCIÓN DE DATOS

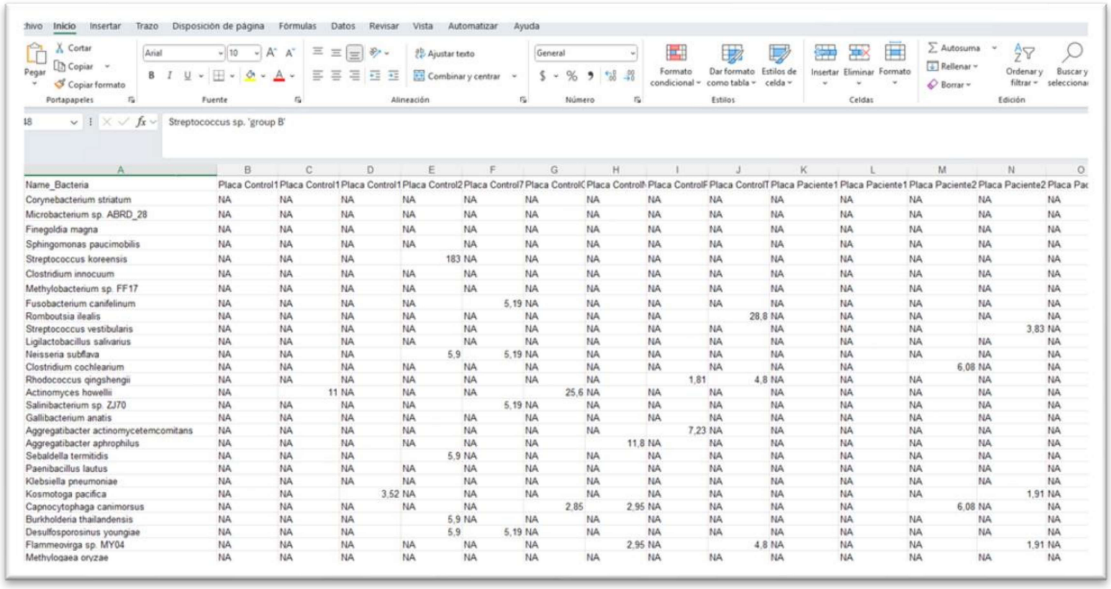
6.2.1 ARCHIVO: 001_DATOS_MUESTRAS_BACTERIANAS.xlsx

6.2.1.1 DESCRIPCIÓN

Este archivo constituye la base inicial del estudio y contiene los nombres de las bacterias identificadas en las muestras biológicas humanas, clasificadas por tipo de muestra.
Las muestras incluyen **placa dental**, **saliva** y **tejido tumoral oral**, diferenciadas entre grupos **control** y **paciente**

Este archivo se encuentra alojado en el repositorio:

https://github.com/anaariza1119/PROYECTO_GRADO_MAESTRIA_DATOS_2025/tree/main/data/external



Name_Bacteria	Placa Control1	Placa Control1	Placa Control1	Placa Control2	Placa Control7	Placa Control(	Placa Control(	Placa Control(	Placa Control(	Placa Control(	Placa Paciente1	Placa Paciente1	Placa Paciente2	Placa Paciente2	Placa Pac
Corynebacterium striatum	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Microbacterium sp. ABRD_28	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Finegoldia magna	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Sphingomonas paucimobilis	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Streptococcus koreensis	NA	NA	NA		183	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Clostridium innocuum	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Methylobacterium sp. FF17	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fusobacterium canifelinum	NA	NA	NA	NA		5,19	NA	NA	NA	NA	NA	NA	NA	NA	NA
Romboutsia ilealis	NA	NA	NA	NA	NA	NA	NA	NA	NA		28,8	NA	NA	NA	NA
Streptococcus vestibularis	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		3,83	NA
Ligilactobacillus salivarius	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Neisseria subflava	NA	NA	NA		5,9	5,19	NA	NA	NA	NA	NA	NA	NA	NA	NA
Clostridium cochlearium	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		6,08	NA	NA
Rhodococcus qingshengii	NA	NA	NA	NA	NA	NA	NA		1,81		4,8	NA	NA	NA	NA
Actinomyces howellii	NA		11	NA	NA		25,6	NA	NA	NA	NA	NA	NA	NA	NA
Salinibacterium sp. ZJ70	NA	NA	NA	NA		5,19	NA	NA	NA	NA	NA	NA	NA	NA	NA
Gallibacterium anatis	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Aggregatibacter actinomycetemcomitans	NA	NA	NA	NA	NA	NA	NA		7,23	NA	NA	NA	NA	NA	NA
Aggregatibacter aphrophilus	NA	NA	NA	NA	NA	NA		11,8	NA	NA	NA	NA	NA	NA	NA
Sebaldella termitidis	NA	NA	NA		5,9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Paenibacillus lautus	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Klebsiella pneumoniae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Kosmotoga pacifica	NA	NA		3,52	NA	NA	NA	NA	NA	NA	NA	NA	NA		1,91	NA
Capnocytophaga canimorsus	NA	NA	NA	NA	NA		2,85		2,95	NA	NA	NA	NA	6,08	NA	NA
Burkholderia thailandensis	NA	NA	NA		5,9	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA
Desulfosporosinus youngiae	NA	NA	NA		5,9	5,19	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Flammeovirga sp. MY04	NA	NA	NA	NA	NA	NA		2,95	NA		4,8	NA	NA	NA		1,91	NA
Methylogaea oryzae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Figura 4: Archivo 001_DATOS_MUESTRAS_BACTERIANAS.

6.2.1.2 ORIGEN

Datos clínicos obtenidos de pacientes humanos. Este archivo fue proporcionado como insumo inicial para realizar las búsquedas correspondientes en BV-BRC.

6.2.1.3 COLUMNAS

COLUMNA	DESCRIPCIÓN
NAME_BACTERIAS	Nombre científico de la bacteria aislada.
CP	*Control Placa*: muestra de placa dental del grupo control.
MP	*Muestra Paciente Placa*: muestra de placa dental de pacientes.
CS	*Control Saliva*: muestra de saliva del grupo control.
MS	*Muestra Paciente Saliva*: muestra de saliva de pacientes.
MT	*Muestra Tumoral*: tejido tumoral oral de pacientes.

Tabla 2: Variables archivo 001_DATOS_MUESTRAS_BACTERIANAS.

6.2.1.4 PROPÓSITO

Establecer la lista base de especies bacterianas presentes en diferentes tipos de muestras para su posterior búsqueda e identificación genómica en BV-BRC.

6.2.2 ARCHIVO: 002_DATOS_ID_BACTERIANOS.tsv

6.2.2.1 DESCRIPCIÓN

Este archivo está conformado por los ID de cada bacteria, con sus nombres completos, se filtra por medio del parámetro ***'hostname': 'Humanos/Homo sapiens'.***

Este archivo se encuentra alojado en el repositorio:

https://github.com/anaariza1119/PROYECTO_GRADO_MAESTRIA_DATOS_2025/tree/main/data/raw

6.2.2.2 ORIGEN

El origen del archivo corresponde a la tabla **`p3-all-genomes`** de la base de datos BV-BRC, en donde se encuentran los ID de las bacterias, los comandos usados para su extracción son:

```
p3-all-genomes --eq genome_name,Corynebacterium striatum |
p3-get-genome-data  --eq host_name,"Homo sapiens"
 --attr genome_name --attr host_name >> 002_DATOS_ID_BACTERIANOS.tsv
```

Figura 5: Comando 002_DATOS_ID_BACTERIANOS.

002_DATOS_ID_BACTERIANOS.tsv

	genome.genome_id	genome.genome_name	genome.host_name
2	82380.59	Microbacterium oxydans strain VHA	Homo sapiens
3	82380.60	Microbacterium oxydans strain VHA	Homo sapiens
4	82380.61	Microbacterium oxydans strain VHA	Homo sapiens
5	82380.63	Microbacterium oxydans VHA	Homo sapiens
6	2029109.3	Microbacterium sp. UMB0228	Human, Homo sapiens
7	2916212.3	Microbacterium sp. p3-SID336	Homo sapiens
8	2916213.3	Microbacterium sp. p3-SID337	Homo sapiens
9	2916214.4	Microbacterium sp. p3-SID338	Homo sapiens

Figura 6: Datos 002_DATOS_ID_BACTERIANOS.

6.2.2.3 COLUMNAS

COLUMNA	DESCRIPCIÓN
GENOME_ID	ID del genoma.
GENOME_NAME	Nombre de la bacteria
HOST_NAME	Especie huésped de donde se aisló la bacteria.

Tabla 3: Variables archivo 002_DATOS_ID_BACTERIANOS.

6.2.2.4 PROPÓSITO

El propósito de este *dataset* es Identificar los ID de las bacterias que se encuentran en las muestras bucales.

6.2.3 ARCHIVO: 003_DATOS_FAMILIAS.tsv

6.2.3.1 DESCRIPCIÓN

Este archivo contiene las propiedades taxonómicas de cada bacteria como la familia, especie y género.

Este archivo se encuentra alojado en el repositorio:

https://github.com/anaariza1119/PROYECTO_GRADO_MAESTRIA_DATOS_2025/tree/main/data/raw

6.2.3.2 ORIGEN

El origen del archivo corresponde a la tabla **`p3-get-genome-data`** de la base de datos BV-BRC, contiene las familias de bacterianas filtradas por el genoma id, los comandos usados para la extracción son los siguientes:

```
p3-echo -t genome_id    43770.574   |p3-get-genome-data --attr genome_id --attr genome_name --attr
genus --attr species --attr taxon_id --attr taxon_lineage_names --attr taxon_lineage_ranks --attr
family --eq host_name,"Homo sapiens" >>  003_DATOS_FAMILIAS.tsv
```

Figura 7: Comando 003_DATOS_GENOMICOS_ID.

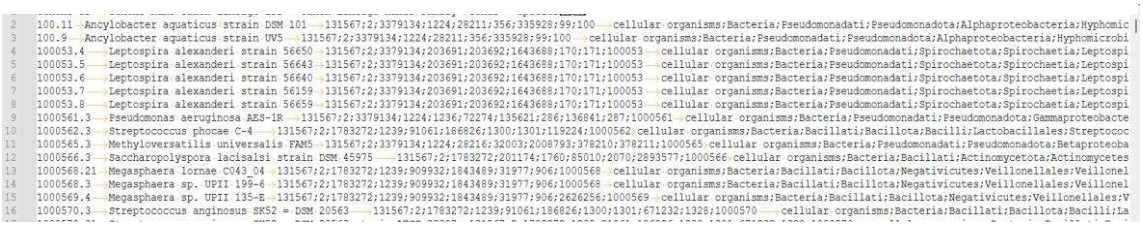


Figura 8: Datos 003_DATOS_GENOMICOS_ID.

6.2.3.3 COLUMNAS

COLUMNA	DESCRIPCIÓN
GENOME ID	Identificador único del genoma en BV-BRC.
GENOME NAME	Nombre del genoma o cepa.
TAXON LINEAGE IDS	Identificadores numéricos de cada nivel taxonómico.
TAXON LINEAGE NAMES	Nombres jerárquicos de la clasificación taxonómica.
FAMILY	Familia taxonómica del organismo.
GENUS	Género al que pertenece el organismo.
SPECIES	Especie del organismo.

Tabla 4: Variables archivo 003_DATOS_GENOMICOS_ID.

6.2.3.4 PROPÓSITO

El propósito de este *dataset* es relacionar las familias bacterianas y agruparlas.

6.2.4 ARCHIVO: 004_DATOS_RESISTENCIA.tsv

6.2.4.1 DESCRIPCION

Este archivo contiene los antibióticos y las resistencias de las bacterias.

Este archivo se encuentra alojado en el repositorio:

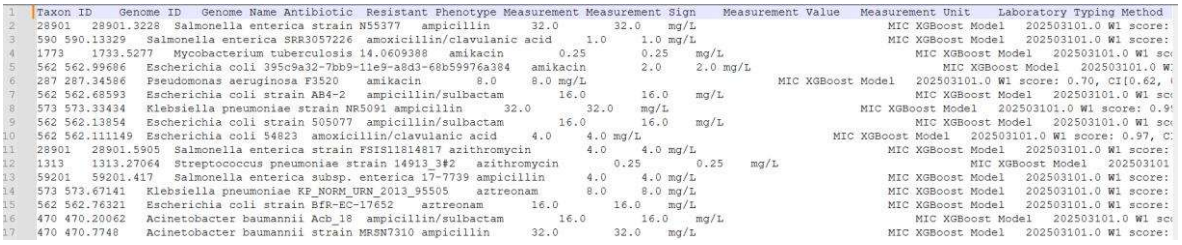
https://github.com/anaariza1119/PROYECTO_GRADO_MAESTRIA_DATOS_2025/tree/main/data/raw

6.2.4.2 ORIGEN

El origen de la base corresponde a la tabla **`p3-get-genome-drugs`** de la base BV_BRC**,** el comando usado para su extracción es el siguiente:

```
p3-echo -t genome_id 1318.969 | p3-get-genome-drugs --attr genome_id --
attr antibiotic --attr resistant_phenotype >> resistencia.tsv
```

Figura 9: Comando 004_DATOS_MEDICAMENTO_ID



Taxon ID	Genome ID	Genome Name	Antibiotic	Resistant Phenotype	Measurement	Measurement Sign	Measurement Value	Measurement Unit	Laboratory Typing Method
28901	28901.3228	Salmonella enterica strain N55377	ampicillin		32.0		32.0	mg/L	MIC XGBoost Model 202503101.0 W1 score:
590	590.13329	Salmonella enterica SRR3057226	amoxicillin/clavulanic acid		1.0		1.0	mg/L	MIC XGBoost Model 202503101.0 W1 score:
1773	1733.5277	Mycobacterium tuberculosis 14.0609388	amikacin		0.25		0.25	mg/L	MIC XGBoost Model 202503101.0 W1 sc
562	562.99686	Escherichia coli 395c9a32-7bb9-11e9-a8d3-68b59976a384	amikacin		2.0		2.0	mg/L	MIC XGBoost Model 202503101.0 W
287	287.34586	Pseudomonas aeruginosa F3520	amikacin		8.0		8.0	mg/L	MIC XGBoost Model 202503101.0 W1 score: 0.70, CI[0.62,
562	562.68593	Escherichia coli strain AB4-2	ampicillin/sulbactam		16.0		16.0	mg/L	MIC XGBoost Model 202503101.0 W1 sc
573	573.33434	Klebsiella pneumoniae strain NR5091	ampicillin		32.0		32.0	mg/L	MIC XGBoost Model 202503101.0 W1 score: 0.9
562	562.13854	Escherichia coli strain 505077	ampicillin/sulbactam		16.0		16.0	mg/L	MIC XGBoost Model 202503101.0 W1 sc
562	562.111149	Escherichia coli 54823	amoxicillin/clavulanic acid		4.0		4.0	mg/L	MIC XGBoost Model 202503101.0 W1 score: 0.97, C
28901	28901.5905	Salmonella enterica strain FSIS11814817	azithromycin		4.0		4.0	mg/L	MIC XGBoost Model 202503101.0 W1 score:
1313	1313.27064	Streptococcus pneumoniae strain 14913_3#2	azithromycin		0.25		0.25	mg/L	MIC XGBoost Model 202503101
59201	59201.417	Salmonella enterica subsp. enterica 17-7739	ampicillin		4.0		4.0	mg/L	MIC XGBoost Model 202503101.0 W1 score:
573	573.67141	Klebsiella pneumoniae KP_NORM_URN_2013_95505	aztreonam		8.0		8.0	mg/L	MIC XGBoost Model 202503101.0 W1 score:
562	562.76321	Escherichia coli strain BfR-EC-17652	aztreonam		16.0		16.0	mg/L	MIC XGBoost Model 202503101.0 W1 score:
470	470.20062	Acinetobacter baumannii Acb_18	ampicillin/sulbactam		16.0		16.0	mg/L	MIC XGBoost Model 202503101.0 W1 sc
470	470.7748	Acinetobacter baumannii strain MRSN7310	ampicillin		32.0		32.0	mg/L	MIC XGBoost Model 202503101.0 W1 score:

Figura 10: Datos 004_DATOS_RESISTENCIA.

6.2.4.3 COLUMNAS

COLUMNA	DESCRIPCIÓN
TAXON ID	Identificador taxonómico del organismo.
GENOME ID	Identificador único del genoma en BV-BRC.
GENOME NAME	Nombre del genoma o cepa.
ANTIBIOTIC	Antibiótico evaluado.
RESISTANT PHENOTYPE	Resultado de resistencia o susceptibilidad.
MEASUREMENT	Tipo de medición realizada.
MEASUREMENT SIGN	Signo que acompaña la medición (>, <, =).
MEASUREMENT VALUE	Valor numérico de la medición.
MEASUREMENT UNIT	Unidad del valor medido (µg/mL, mm).
LABORATORY TYPING METHOD	Método de laboratorio usado.
LABORATORY TYPING METHOD VERSION	Versión del método de laboratorio.
LABORATORY TYPING PLATFORM	Plataforma o equipo empleado.
VENDOR	Proveedor del kit o método.
TESTING STANDARD	Norma usada (CLSI, EUCAST, etc.).
TESTING STANDARD YEAR	Año de la norma aplicada.
COMPUTATIONAL METHOD	Método computacional de análisis.
COMPUTATIONAL METHOD VERSION	Versión del método computacional.
COMPUTATIONAL METHOD PERFORMANCE	Desempeño del método computacional.
EVIDENCE	Tipo de evidencia (experimental o computacional).
SOURCE	Fuente del dato o base de origen.

Tabla 5: Variables archivo 004_DATOS_RESISTENCIA.

6.2.4.4 PROPÓSITO

El propósito de este *dataset* es poder identificar los antibióticos que son efectivos contra cada bacteria para encontrar tendencias en sus características.

6.3 PREPARACIÓN Y LIMPIEZA DE DATOS

La limpieza y preparación de los datos se realizaron utilizando Python (versión 3.10) y las librerías `pandas, numpy y openpyxl`, dentro de un entorno Anaconda.
Las principales tareas ejecutadas fueron:

6.3.1 NORMALIZACIÓN DE COLUMNAS

Se estandarizan los nombres de columnas y el contenido textual para garantizar la compatibilidad entre las distintas fuentes de datos.
Se aplicó la función `normalizar_columnas(df)` para:

- Eliminar espacios en blanco en los encabezados.
- Convertir todos los nombres a minúsculas.
- Reemplazar espacios por guiones bajos (_).
- Eliminar caracteres especiales con expresiones regulares ([^a-z0-9_]).

```python
def normalizar_columnas(df: pd.DataFrame) → pd.DataFrame:
    df = df.copy()
    df.columns = (
        df.columns
          .str.strip()
          .str.lower()
          .str.replace(" ", "_")
          .str.replace("[^a-z0-9_]", "", regex=True)
    )
    return df
```

Figura 11: Código normalización columnas.

Posteriormente se ejecutó la función limpiar_textos_y_nulos(df) para normalizar el contenido de cada celda:

- Remover espacios iniciales/finales.
- Convertir a minúsculas.
- Reemplazar valores vacíos, “NA”, “N/A” o guiones por NaN.
- Sustituir todos los valores nulos (NaN) por 0, tanto en variables numéricas como de texto.
- Eliminar valores duplicados.
- En el caso particular de la fuente 001_DATOS_MUESTRAS_BACTERIANAS.xlsx, los valores negativos fueron convertidos a cero.
  Esta decisión se fundamenta en que las variables presentes en este conjunto de datos representan magnitudes biológicas o experimentales (por ejemplo, conteos

de bacterias, concentraciones o identificadores numéricos), las cuales no pueden asumir valores negativos en un contexto real.

- Por tanto, cualquier valor negativo identificado se considera un error de registro, digitación o transformación durante la captura de los datos. Sustituirlos por cero garantiza la coherencia biológica de la información y evita distorsiones en los análisis estadísticos posteriores, ya que el valor cero refleja de forma más adecuada la ausencia o no detección de un valor medible, en lugar de representar una magnitud inválida.

```python
def limpiar_textos_y_nulos(df: pd.DataFrame) -> pd.DataFrame:
    """
    Limpieza general:
      - Strings: strip/lower; vacíos comunes → NaN.
      - Elimina filas totalmente vacías y duplicadas.S
      - Numéricas: NaN → 0; negativos → 0.
      - Texto: NaN → "0" y, si el texto representa un número negativo, → "0".
    """
    df = df.copy()
    vacios = {"", " ", "-", "na", "n/a", "NA", "N/A"}

    # ── 1) Limpieza de texto base ──
    for col in df.columns:
        if df[col].dtype == "object":
            df[col] = df[col].apply(lambda x: x.strip().lower() if isinstance(x, str) else x)
            df[col] = df[col].apply(lambda x: np.nan if isinstance(x, str) and x in vacios else x)

    # ── 2) Quitar filas vacías y duplicadas ──
    df = df.dropna(how="all").drop_duplicates()

    # ── 3) Numéricas reales: NaN → 0; negativos → 0 ──
    num_cols = df.select_dtypes(include=[np.number]).columns
    if len(num_cols):
        df[num_cols] = df[num_cols].fillna(0)
        df[num_cols] = df[num_cols].applymap(lambda x: 0 if x < 0 else x)

    # ── 4) Objetos que contienen números negativos en texto ──
    # patrón: acepta guion normal o unicode (-) y decimales con punto o coma
    neg_pattern = re.compile(r"^\s*[--]\s*\d+(?:[.,]\d+)?\s*$")

    obj_cols = df.select_dtypes(include=["object"]).columns
    for col in obj_cols:
        # a) reemplazar NaN por "0"
        df[col] = df[col].fillna("0")

        # b) detectar strings que representan números negativos
        mask_neg_text = df[col].astype(str).apply(lambda s: bool(neg_pattern.match(s)))

        # c) para esas filas, poner "0"
        if mask_neg_text.any():
            df.loc[mask_neg_text, col] = "0"

        # d) adicional: intentar coercer a número para detectar negativos ocultos tipo " -3,5 "
        tmp = (
            df[col]
            .astype(str)
            .str.replace("-", "-", regex=False)   # guion unicode → normal
            .str.replace(",", ".", regex=False)   # coma decimal → punto
        )
        tmp_num = pd.to_numeric(tmp, errors="coerce")
        mask_neg_num = tmp_num < 0
        if mask_neg_num.any():
            df.loc[mask_neg_num, col] = "0"

    return df
```

Figura 12: Código normalización celdas.

En la siguiente figura se observa el resultado de la normalización de columnas y celdas aplicada a las cinco fuentes de datos (001, 002, 003, 004).

En el encabezado de cada *DataFrame* se evidencia la eliminación de espacios y caracteres especiales, la conversión a minúsculas y la sustitución de espacios por guiones bajos, esta estandarización garantiza que todas las tablas mantengan una estructura coherente y compatible para los procesos de integración posteriores.
Además, durante la limpieza del contenido se reemplazaron los valores nulos y los campos vacíos por **0**, lo que evita errores en los cruces o análisis numéricos.

```
Head 001:
            name_bacteria  10cp  15cp  ... 3mt  5mt     nombre_normalizado
0  [clostridium] hylemonae  1.78  1.83  ...   0    0  Clostridium hylemonae
1   [clostridium] innocuum     0     0  ...   0  111   Clostridium innocuum
2    abiotrophia defectiva     0  25.7  ...   0    0  Abiotrophia defectiva

[3 rows x 48 columns]

Head 002 (muestra en memoria):
  genomegenome_id                  genomegenome_name genomehost_name
0        82380.59  microbacterium oxydans strain vha    homo sapiens
1        82380.60  microbacterium oxydans strain vha    homo sapiens
2        82380.61  microbacterium oxydans strain vha    homo sapiens

Head 003 (muestra):
    genome_id  ... featurefamily_id
0  43770.1307  ...              0.0
1  43770.1307  ...              0.0
2  43770.1307  ...              0.0

[3 rows x 4 columns]

Head 004 (muestra):
    genome_id  ... genome_drugresistant_phenotype
0  43770.1401  ...                              0
1  43770.1401  ...                              0
2  43770.1401  ...                              0

[3 rows x 4 columns]

Head 005 (muestra):
  genomegenome_id  ...      genomehost_name
0        1318.969  ...  human, homo sapiens
1        1328.227  ...  human, homo sapiens
2        1334.234  ...  human, homo sapiens

[3 rows x 3 columns]
```

Figura 13: Resultado estabilización columnas y celdas.

También se aplica la función `limpiar_textos_y_nulos()`, para realizar la depuración de duplicados la cual permitió reducir significativamente el tamaño de las fuentes de datos, eliminando registros repetidos que podían generar inconsistencias en los cruces.

Por ejemplo, el archivo **002_DATOS_ID_BACTERIANOS.tsv** pasó de **1.302.275 registros iniciales** a **543.716 registros únicos**, lo que representa una reducción del **58%**, según los logs de ejecución.

```
Backup → C:\Users\anaso\OneDrive\Documentos\PROYECTO_GRADO_MAESTRIA_DATOS_2025\data\interim\001_normalizado.csv (275, 48)
Backup → C:\Users\anaso\OneDrive\Documentos\PROYECTO_GRADO_MAESTRIA_DATOS_2025\data\interim\002_normalizado.csv | filas in=1302275 → out=543716 | dups removidos=758559
Backup → C:\Users\anaso\OneDrive\Documentos\PROYECTO_GRADO_MAESTRIA_DATOS_2025\data\interim\003_normalizado.csv | filas in=5111836 → out=5111836 | dups removidos=0
Backup → C:\Users\anaso\OneDrive\Documentos\PROYECTO_GRADO_MAESTRIA_DATOS_2025\data\interim\004_normalizado.csv | filas in=23750 → out=22599 | dups removidos=1151
Backup → C:\Users\anaso\OneDrive\Documentos\PROYECTO_GRADO_MAESTRIA_DATOS_2025\data\interim\005_normalizado.csv | filas in=818056 → out=394911 | dups removidos=423145
```

Figura 14: Log eliminar duplicados.

6.3.2 ESTANDARIZACIÓN DE NOMBRES CIENTÍFICOS

La estandarización de los nombres científicos constituyó una de las etapas más importantes dentro del proceso de preparación y limpieza de los datos, ya que de ella dependía la correcta vinculación entre los registros bacterianos del archivo 001 y los identificadores genómicos del archivo 002.

En las bases de datos biológicas, es común encontrar variaciones en la escritura de los nombres taxonómicos debido a diferencias en las fuentes, errores de digitación o convenciones propias de cada laboratorio. Ejemplos frecuentes incluyen el uso de abreviaturas *(sp., spp., strain),* códigos de colección *(ATCC, DSM, NCTC),* tildes, mayúsculas o corchetes para indicar reclasificaciones. Estas inconsistencias pueden provocar que un mismo organismo sea tratado como diferentes entidades al realizar cruces o análisis comparativos, generando falsos negativos y pérdida de información biológicamente relevante [27],[28],[29].

Para evitarlo, se desarrolló la función `limpiar_nombre_cientifico()`, la cual normaliza todos los nombres bacterianos al formato "Género especie", siguiendo las reglas del sistema binomial de nomenclatura establecido por *Linnaeus* [30]. Este formato asegura que:

- El género se escriba con inicial mayúscula.
- La especie se escriba en minúscula.
- Se eliminen tildes, corchetes, paréntesis y cualquier caracter especial.
- Se supriman abreviaturas taxonómicas no válidas como *sp., spp., strain, ATCC, DSM,* entre otras.

```python
def limpiar_nombre_cientifico(nombre: str) → str | None:
    """
    Estandariza a 'Género especie':
      - Quita tildes, corchetes, paréntesis, comillas y puntos.
      - Elimina 'sp', 'spp' y metadatos de cepa (strain, ATCC, DSM, etc.).
      - NO fragmenta tokens con '_' o '-' (se descartan completos si no son alfabéticos).
      - Evita tomar como 'especie' términos de sitio/aislado comunes (eye, skin, soil, etc.).
    """
    import re, unicodedata

    if not isinstance(nombre, str) or not nombre.strip():
        return None

    # 1) Normalización básica
    n = ''.join(c for c in unicodedata.normalize('NFD', nombre) if unicodedata.category(c) ≠ 'Mn')
    n = re.sub(r'[\[\]\(\)\'\"{}\.]', '', n)

    # 2) Cortar metadatos de cepa/colecciones y lo que siga
    n =
re.sub(r'\b(strain|type|isolate|atcc|dsm|nctc|ccug|jcm|kctc|nbrc|cbs|subsp|subspecies|umb|mmrc|cw|we
)\b.*',
                '', n, flags=re.IGNORECASE)

    # 3) Quitar sp/spp como palabras completas
    n = re.sub(r'\bsp\.?\b|\bspp\.?\b', '', n, flags=re.IGNORECASE)

    # 4) Espacios prolijos (¡sin reemplazar '_' ni '-'!)
    n = re.sub(r'\s+', ' ', n).strip()

    # 5) Tokenizar por ESPACIO y quedarse solo con tokens 100% alfabéticos
    tokens_orig = n.split()
    tokens_alpha = [t for t in tokens_orig if re.fullmatch(r'[A-Za-z]+', t)]

    if not tokens_alpha:
        return None

    genero = tokens_alpha[0].capitalize()

    # Lista corta de términos no-especie comunes (puedes ampliarla si ves otros)
    STOP_ESPECIE = {
        "eye","skin","soil","water","aquifer","sediment","lake","river","seawater","marine",
        "stool","feces","fecal","urine","blood","saliva","oral","nasal","vaginal","gut",

    return genero
```

Figura 15: Código limpiar nombre científico.

El proceso permitió que todas las observaciones siguieran un estándar taxonómico uniforme, indispensable para la correcta identificación y trazabilidad entre las distintas fuentes de información [31].

6.3.2.1 EJEMPLOS DEL PROCESO DE ESTANDARIZACIÓN

NOMBRE ORIGINAL	NOMBRE ESTANDARIZADO	DESCRIPCIÓN DEL CAMBIO APLICADO
[Clostridium] innocuum	**Clostridium innocuum**	Eliminación de corchetes y mayúscula inicial en género.
Clostridium sp.	**Clostridium**	Eliminación de abreviatura *sp.* (especie no determinada).
Bacillus subtilis strain 168	**Bacillus subtilis**	Remoción de “strain 168”.
Actinomyces sp.	**Actinomyces**	Supresión de punto y espacio adicional.
Lactobacillus spp.	**Lactobacillus**	Eliminación de *spp.* (múltiples especies).
Escherichia coli ATCC 25922	**Escherichia coli**	Limpieza de código de colección.
Microcella sp.	**Microcella**	Abreviatura *sp.* eliminada.

Tabla 6: Ejemplos del proceso de estandarización.

Tras la ejecución del proceso, se verificó la ausencia de sufijos como *sp.* o *spp.* en los campos normalizados, confirmando que todos los registros bacterianos quedaron correctamente estandarizados. Este control se realizó mediante expresiones regulares, obteniéndose 0 ocurrencias de dichos términos, lo que evidencia la eficacia del proceso de limpieza.

```
Head 001:
            name_bacteria  10cp  15cp  ... 3mt  5mt     nombre_normalizado
0  [clostridium] hylemonae  1.78  1.83  ...   0    0  Clostridium hylemonae
1   [clostridium] innocuum     0     0  ...   0  111   Clostridium innocuum
2    abiotrophia defectiva     0  25.7  ...   0    0  Abiotrophia defectiva
```

Figura 16: Resultado estandarizar nombre científico.

6.4 INTEGRACIÓN DE DATOS

Desarrollar un conjunto de datos de entrenamiento compuesto por bacterias bucales y sus respectivos marcadores genómicos de patogenicidad y resistencia.

Las principales tareas ejecutadas fueron:

6.4.1 CONVERSIÓN DE TIPO DE DATOS

La función `_ensure_str()` convierte columnas específicas a tipo *string* para evitar errores de comparación durante las uniones (*merge*).

```
#──────────────────────────────────────────────────
# Paso 1 - Asegurar tipo str
#──────────────────────────────────────────────────
def _ensure_str(df, cols):
    for c in cols:
        if c in df.columns:
            df[c] = df[c].astype(str)
    return df
```

Figura 17: Comando conversión de tipo de datos

6.4.2 GRÁFICOS DE COINCIDENCIAS

La función `pie_match_vs_nomatch()` calcula el número y porcentaje de coincidencias y no coincidencias entre dos conjuntos de datos y genera un gráfico tipo torta que representa esa proporción.

```python
def pie_match_vs_nomatch(matched_count, base_count, title, out_path=None):
    matched = int(matched_count)
    base = int(base_count) if base_count else 0
    no_match = max(base - matched, 0)
    pct_m = round((matched / base * 100), 2) if base else 0
    pct_nm = round((no_match / base * 100), 2) if base else 0

    sizes = [matched, no_match]
    labels = ['Cruzan', 'No cruzan']
    colors = [PALETTE[0], PALETTE[1]]
    explode = (0.05, 0)

    plt.figure(figsize=(6, 6))
    wedges, texts, autotexts = plt.pie(
        sizes,
        explode=explode,
        labels=None,
        colors=colors,
        startangle=90,
        shadow=True,
        autopct='%1.1f%%',
        pctdistance=0.75,
        wedgeprops={'linewidth': 1, 'edgecolor': 'white'}
    )
    plt.title(title, fontsize=14, fontweight='bold')
    plt.axis('equal')
    plt.legend(
        wedges,
        [f"{lab} ({val/base*100:.1f}%)" for lab, val in zip(labels, sizes) if base > 0],
        loc='upper left',
        bbox_to_anchor=(0.02, 0.98),
        frameon=True,
        fancybox=True,
        shadow=False,
        fontsize=10
    )
    plt.tight_layout()
    if out_path:
        os.makedirs(os.path.dirname(out_path), exist_ok=True)
        plt.savefig(out_path, dpi=300, bbox_inches='tight')
        plt.close()
    else:
        plt.show()

    return {
        "matched": matched,
        "nomatch": no_match,
        "base": base,
        "pct_matched": pct_m,
        "pct_nomatch": pct_nm
    }
```

Figura 18:Comando gráficos de coincidencias

6.4.1.1 EXPORTACIÓN DE REPORTES DE UNIÓN

La función `export_join_reports()` realiza las uniones entre tablas, identifica los registros coincidentes y no coincidentes, guarda los resultados en archivos .csv y genera el gráfico correspondiente.

```python
#————————————————————————————————————
# Paso 3 - Unión y reportes
#————————————————————————————————————
def export_join_reports(left_df, right_df, key, left_name, right_name, out_dir, base_is='left'):
    os.makedirs(out_dir, exist_ok=True)

    inner = left_df.merge(right_df, how='inner', on=key, suffixes=(f"_{left_name}",
f"_{right_name}"))
    left_only = left_df[~left_df[key].isin(inner[key])]
    right_only = right_df[~right_df[key].isin(inner[key])]

    if base_is == 'left':
        base_total = left_df[key].nunique(dropna=True)
    elif base_is == 'inner':
        base_total = inner[key].nunique(dropna=True)
    else:
        base_total = left_df[key].nunique(dropna=True)

    matched_total = inner[key].nunique(dropna=True)

    inner_path = os.path.join(out_dir, f"inner_{left_name}_vs_{right_name}.csv")
    left_only_path = os.path.join(out_dir, f"no_{right_name}_en_{left_name}.csv")
    right_only_path = os.path.join(out_dir, f"no_{left_name}_en_{right_name}.csv")

    inner.to_csv(inner_path, index=False, encoding='utf-8')
    left_only.to_csv(left_only_path, index=False, encoding='utf-8')
    right_only.to_csv(right_only_path, index=False, encoding='utf-8')

    pie_path = os.path.join(out_dir, f"pie_{left_name}_vs_{right_name}.png")
    title = f"Relación entre {left_name} y {right_name}"
    pie_stats = pie_match_vs_nomatch(matched_total, base_total, title, pie_path)

    resumen_row = {
        "union": f"{left_name}_vs_{right_name}",
        "clave": key,
        "base": base_is,
        "base_total": base_total,
        "matched": pie_stats["matched"],
        "no_matched": pie_stats["nomatch"],
        "pct_matched": pie_stats["pct_matched"],
        "pct_no_matched": pie_stats["pct_nomatch"],
        "inner_csv": inner_path,
        "left_only_csv": left_only_path,
        "right_only_csv": right_only_path,
        "pie_png": pie_path
    }
    return inner, left_only, right_only, resumen_row
```

Figura 19: Comando exportación de reportes de unión

6.4.1.2 INTEGRACIONES PROGRESIVAS

Se realizan tres integraciones consecutivas: (a) entre 001 y 002 por el campo *nombre_normalizado*, (b) entre la unión previa y 003 por *genome_id*, y (c) entre la unión resultante y 004 por el mismo campo.

```
#-----------------------------------------------
# Paso 5 - Integraciones
#-----------------------------------------------
df001_unicos = _ensure_str(df001_unicos, ['nombre_normalizado'])
df002 = _ensure_str(df002, ['nombre_normalizado'])

inner_001_002, no_002_en_001, no_001_en_002, row_001_002 = export_join_reports(
    left_df=df001_unicos,
    right_df=df002,
    key='nombre_normalizado',
    left_name='001',
    right_name='002',
    out_dir=os.path.join(ruta_out, "001_vs_002"),
    base_is='left'
)

df001_002 = inner_001_002.rename(columns={
    "genomegenome_id": "genome_id",
    "genomegenome_name": "genome_name"
})
df001_002 = _ensure_str(df001_002, ['genome_id'])
df003 = _ensure_str(df003, ['genome_id'])

inner_001002_003, no_003_en_001002, no_001002_en_003, row_001002_003 = export_join_reports(
    left_df=df001_002[['nombre_normalizado','genome_id','genome_name']].drop_duplicates(),
    right_df=df003[['genome_id']].drop_duplicates(),
    key='genome_id',
    left_name='001_002',
    right_name='003',
    out_dir=os.path.join(ruta_out, "001_002_vs_003"),
    base_is='left'
)

df004 = _ensure_str(df004, ['genome_id'])
inner_0010023_004, no_004_en_0010023, no_0010023_en_004, row_0010023_004 = export_join_reports(
    left_df=inner_001002_003[['genome_id']].drop_duplicates(),
    right_df=df004[['genome_id']].drop_duplicates(),
    key='genome_id',
    left_name='001_002_003',
    right_name='004',
    out_dir=os.path.join(ruta_out, "001_002_003_vs_004"),
    base_is='left'
)
```

Figura 20:Comando integraciones progresivas.

6.4.1.3 RESUMEN GENERAL DE INTEGRACIONES

Los resultados de cada integración se resumen en un archivo *(resumen_integracion.csv)* que contiene los totales de registros base, coincidencias, porcentajes y rutas de salida generadas.

```
#-----------------------------------------------
# Paso 6 - Resumen general
#-----------------------------------------------
resumen = pd.DataFrame([row_001_002, row_001002_003, row_0010023_004])
resumen_path = os.path.join(ruta_out, "resumen_integracion.csv")
resumen.to_csv(resumen_path, index=False, encoding='utf-8')
print("Resumen guardado en:", resumen_path)
```

Figura 21:Comando resumen general de integraciones.

6.4.1.4 RESUMEN IMPRESO DE RESULTADOS

Se imprime en consola el resumen final de las tres integraciones con sus respectivos totales, porcentajes y nombres de archivos exportados.

```
#-----------------------------------------------
# Paso 8 - Mostrar resumen
#-----------------------------------------------
print("\n========== RESUMEN DE INTEGRACIÓN ==========\n")
for fila in [row_001_002, row_001002_003, row_0010023_004]:
    print(f" Unión: {fila['union']}")
    print(f"  - Base de referencia: {fila['base']}")
    print(f"  - Total base: {fila['base_total']:,}")
    print(f"  - Coinciden: {fila['matched']:,} ({fila['pct_matched']}%)")
    print(f"  - No coinciden: {fila['no_matched']:,} ({fila['pct_no_matched']}%)")
    print(f"  - CSV coincidencias: {os.path.basename(fila['inner_csv'])}")
    print(f"  - CSV no coinciden (der): {os.path.basename(fila['right_only_csv'])}")
    print(f"  - Gráfico: {os.path.basename(fila['pie_png'])}")
    print("-" * 60)
```

Figura 22:Comando resumen impreso de resultados.

6.4.1.5 RESULTADOS DE LA INTEGRACIÓN

El análisis de integración de las bases de datos **001, 002, 003 y 004** permitió evaluar el grado de coincidencia entre las distintas fuentes de información. En la primera comparación (**001_vs_002**) se alcanzó una concordancia del **78,85%**, lo que evidencia una relación inicial consistente entre ambas. Con la incorporación de la base **003**, la coincidencia aumentó a **90,62%**, indicando una mejora en la homogeneidad y compatibilidad de los registros. Sin embargo, al añadir la base **004**, el nivel de concordancia descendió a **82,99%**, posiblemente debido a la falta de estudios clínicos, en la base genómica inicial. En conjunto, estos resultados confirman una integración globalmente estable y confiable, para el entrenamiento del modelo de datos, destacando la preservación de **23** bacterias, **109.831** *Genome_id* y más de **4** millones de registros asociados a las resistencias antibióticas.

Estas figuras se encuentran alojadas en el repositorio:

https://github.com/anaariza1119/PROYECTO_GRADO_MAESTRIA_DATOS_2025/tree/main/reports/figures

Gráficos por cada cruce:

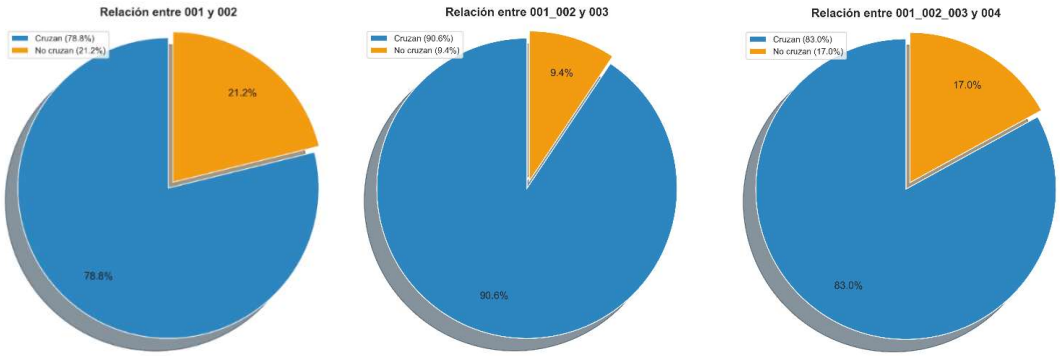


Figura 23:Graficos pie_001_vs_002_vs_003_vs_004.

RELACION	TOTAL BASE	COINCIDEN	% COINCIDEN	NO COINCIDEN	% NO COINCIDEN
001_vs_002	260	205	78.85%	55	21.15%
001_002_vs_003	146,034	132,339	90.62%	13,695	9.38%
001_002_003_vs_004	132,339	109,831	82.99%	22,508	17.01%

Tabla 7: Resultado integración Objetivo 1.

Durante la integración se realizaron tres uniones progresivas:

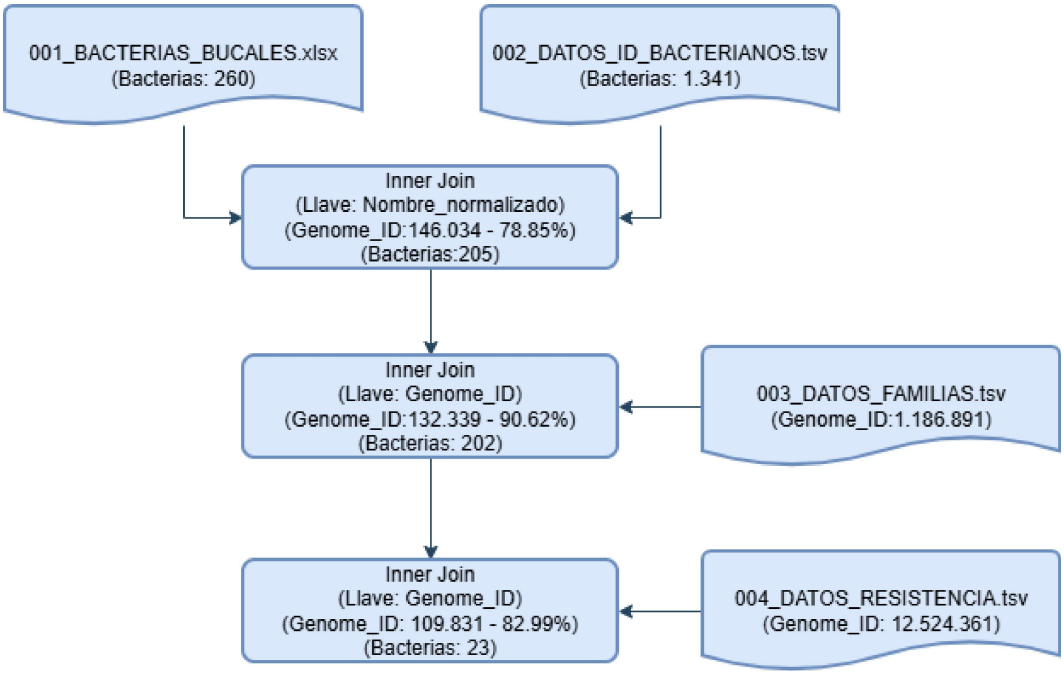


Figura 24:Diagrama relación de resultados

1. Entre los archivos 001 y 002, el 78.85 % de las especies (205 de 260) presentaron coincidencias por el campo nombre_normalizado.

2. En la segunda integración, al incorporar el archivo 003, se mantuvieron 132,339 registros coincidentes sobre un total de 146,034, equivalentes al 90.62 %.

3. En la última unión, con el archivo 004, se obtuvieron 109,831 coincidencias de 132,339 registros, lo que corresponde al 82.99 %.

El número de especies únicas se redujo de 260 iniciales a 23 después de las integraciones, como resultado de la eliminación de registros duplicados o sin correspondencia entre tablas.

Como resultado del proceso de integración y comparación entre las distintas bases de datos, se generaron los siguientes archivos, los cuales recopilan la información obtenida en cada cruce.
Estos archivos se encuentran disponibles en el repositorio del proyecto, en la ruta:

https://github.com/anaariza1119/PROYECTO_GRADO_MAESTRIA_DATOS_2025/tree/main/data/processed

Los archivos generados son los siguientes:

- **Inner_001_vs_002.csv:** Contiene los registros coincidentes entre las bases 001 y 002.
- **Inner_001_002_vs_003.csv:** Consolida las coincidencias resultantes de la unión entre las bases 001, 002 y 003.
- **Inner_001_002_003_vs_004.csv:** Integra la información de las cuatro bases de datos, mostrando las coincidencias globales.
- **No_001_en_002.csv:** Agrupa los registros presentes en la base 001 que no se encontraron en la base 002.
- **No_001_002_en_003.csv:** Identifica los registros de las bases 001 y 002 que no aparecen en la base 003.
- **No_001_002_003_en_004.csv:** Muestra los registros existentes en las tres primeras bases que no están presentes en la base 004.

7 ANÁLISIS EXPLORATORIO DE MUESTRAS CLÍNICAS.

7.1 OBJETIVO DEL CAPITULO

Realizar el análisis exploratorio de las variables clínicas, taxonómicas y fenotípicas para identificar patrones relevantes en la distribución de bacterias bucales y en los niveles de resistencia antibiótica.

7.1.1 DESCRIPCIÓN

En esta fase mediante el uso de herramientas de análisis en Python (principalmente pandas, seaborn y matplotlib), se generaron resúmenes cuantitativos y gráficos comparativos que describen la distribución de abundancias relativas, la variabilidad intragrupal e intergrupal y la identificación de bacterias potencialmente asociadas con muestras tumorales.

Este proceso incluye la normalización de proporciones, la visualización de las 30 bacterias más representativas mediante mapas de calor, la determinación de los 10 taxones más abundantes por tipo de muestra y la evaluación de la distribución global mediante diagramas de caja (boxplots).

Asimismo, se efectuó una comparación cruzada entre los perfiles tumorales y los de otros grupos, con el fin de detectar bacterias comunes y aquellas exclusivas de los tejidos tumorales, aportando así una primera aproximación exploratoria al comportamiento diferencial del microbioma clínico.

En conjunto, este análisis constituye una fase descriptiva y comparativa preliminar, fundamental para orientar los siguientes capítulos enfocados en la validación estructural y el estudio taxonómico y de resistencia antimicrobiana.

7.1.2 AGRUPACIÓN DE MUESTRAS.

El código desarrollado realiza un análisis de las muestras clínicas mediante los grupos de muestras (placa y saliva de controles y pacientes, además de tejido tumoral) a partir de los nombres de las columnas. Posteriormente, se calcula la abundancia total y relativa de cada bacteria en los distintos grupos, generando un resumen comparativo.

```
# ──────────────────────────────────────────────────────────
# Paso 2: Identificar grupos de muestras
# ──────────────────────────────────────────────────────────
grupos = {
    'CP': [c for c in df.columns if 'control' in c and 'placa' in c],
    'MP': [c for c in df.columns if 'placa_paciente' in c],
    'CS': [c for c in df.columns if 'saliva_control' in c],
    'MS': [c for c in df.columns if 'saliva_paciente' in c],
    'MT': [c for c in df.columns if 'turmor_paciente' in c]
}

for g, cols in grupos.items():
    print(f"{g}: {len(cols)} columnas detectadas")
```

Figura 25:Comando Identificar Grupos.

```
CP: 9 columnas detectadas
MP: 10 columnas detectadas
CS: 7 columnas detectadas
MS: 10 columnas detectadas
MT: 10 columnas detectadas
```

Figura 26:Resultado Identificar Grupos

7.1.3 ABUNDANCIA POR GRUPO.

En este paso se elabora un resumen de abundancia por grupo, sumando los valores de cada tipo de muestra para obtener la distribución total de bacterias. Luego, se aplica una normalización proporcional que convierte los valores absolutos en abundancias relativas, permitiendo comparar la composición microbiana entre los distintos grupos clínicos.

```
# ──────────────────────────────────────────────
# Paso 3: Resumen de abundancia por grupo
# ──────────────────────────────────────────────
resumen = {}
for grupo, cols in grupos.items():
    resumen[grupo] = df[cols].sum(axis=1)

df_resumen = pd.DataFrame(resumen)
df_resumen["nombre_normalizado"] = df["nombre_normalizado"]

# Normalizar proporciones
df_resumen_prop = df_resumen.set_index("nombre_normalizado")
df_resumen_prop = df_resumen_prop.div(df_resumen_prop.sum(axis=0), axis=1)

plt.figure(figsize=(10, 12))
sns.heatmap(df_resumen_prop.head(30), cmap="YlGnBu", cbar_kws={'label':
'Proporción relativa'})
plt.title("Abundancia relativa de las 30 bacterias principales",
fontsize=14, fontweight='bold')
plt.xlabel("Tipo de muestra", fontsize=12)
plt.ylabel("Bacterias", fontsize=12)
plt.tight_layout()
plt.show()
```

Figura 27:Comando Abundancia por Grupos.

El mapa de calor muestra la distribución de abundancia relativa de las bacterias detectadas en los diferentes tipos de muestras (CP, MP, CS, MS y MT). Se observan perfiles microbianos diferenciados entre grupos, con una mayor representación de géneros como *Streptococcus*, Prevotella y *Capnocytophaga* en las muestras de pacientes, especialmente en tejido tumoral, lo que sugiere una posible asociación entre estas bacterias y las condiciones patológicas analizadas.

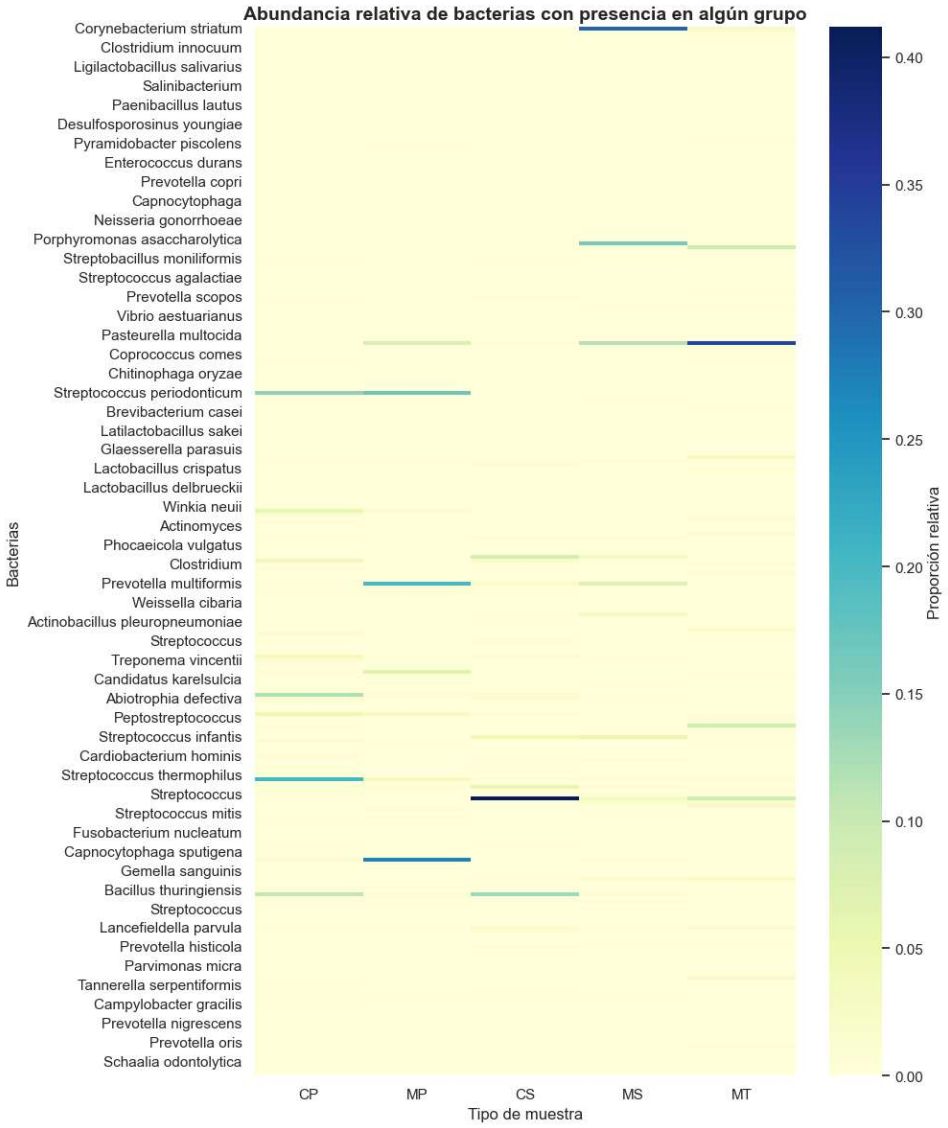


Figura 28:Grafico Abundancia 30 bacterias.

7.1.4 GRÁFICOS DE BARRAS DE ABUNDANCIA

El código identifica las bacterias más abundantes en los diferentes tejidos y las compara con otros grupos, determinando cuáles son comunes y cuáles son exclusivas o dominantes en cada muestra, con el fin de resaltar posibles asociaciones microbianas específicas del tejido.

```python
# ──────────────────────────────────────────────────────────
# Paso 5: Top bacterias por tipo de muestra (barras multicolor sin error
bars)
# ──────────────────────────────────────────────────────────
top_n = 10
for grupo in grupos.keys():
    top = df_resumen.nlargest(top_n, grupo)[["nombre_normalizado", grupo]]

    plt.figure(figsize=(8, 5))
    sns.barplot(
        data=top,
        x=grupo,
        y="nombre_normalizado",
        palette=PALETTE[:len(top)],  # cada barra con color distinto
        errorbar=None  # ← evita mostrar la línea negra
    )

    plt.title(f"Top {top_n} bacterias más abundantes en {grupo}",
fontsize=13, fontweight='bold')
    plt.xlabel("Nivel de abundancia", fontsize=11)
    plt.ylabel("Bacteria", fontsize=11)
    plt.grid(axis='x', linestyle='--', alpha=0.5)
    plt.tight_layout()
    plt.show()
```

Figura 29:Código de gráficos de barras.

Gráfico de barras top 10 de bacterias con mayor abundancia en las muestras de tejido tumoral (MT):

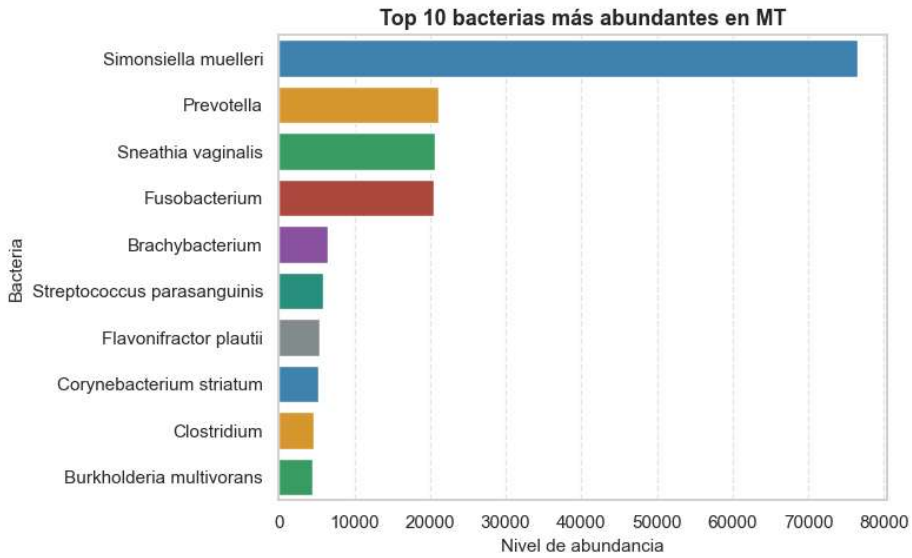


Figura 30:Grafica bacterias más abundantes (MT).

Se observa que *Simonsiella muelleri* muestra una dominancia marcada, superando ampliamente a las demás especies. En niveles intermedios destacan *Prevotella*, *Sneathia vaginalis* y *Fusobacterium*, todas ellas previamente asociadas con procesos inflamatorios y disbiosis en tejidos orales. Las demás bacterias, aunque menos abundantes, contribuyen al perfil microbiano característico del tejido tumoral.

Gráfico de barras diez bacterias más abundantes en las muestras de saliva de pacientes (MS).

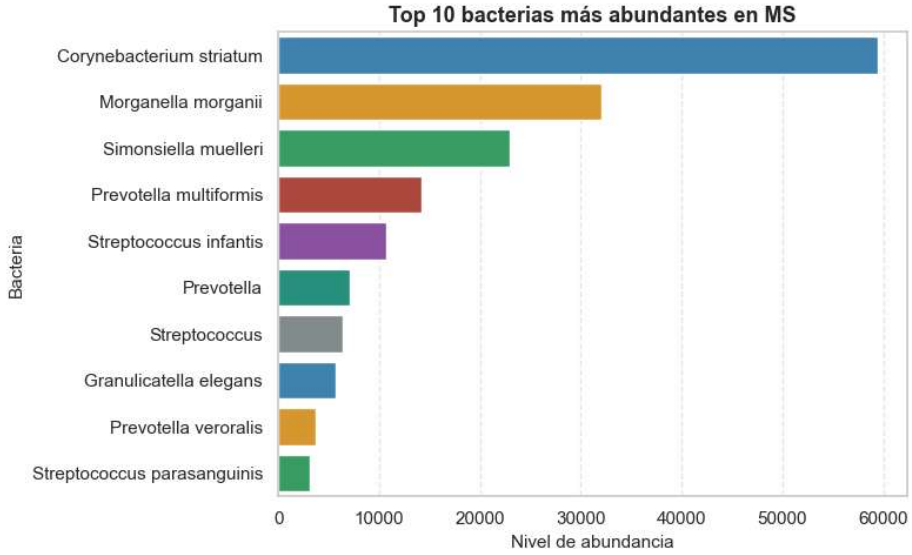
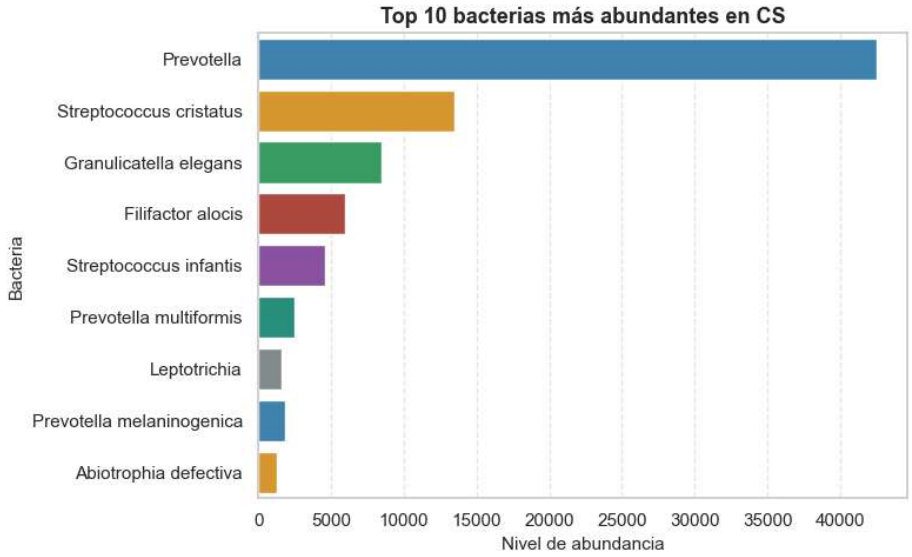


Figura 31:Grafica bacterias más abundantes (MS).

Se observa un claro predominio de *Corynebacterium striatum* y *Morganella morganii*, seguidas por *Simonsiella muelleri* y *Prevotella multiformis*. Estas especies destacan por su alta prevalencia en ambientes orales alterados y su posible implicación en procesos infecciosos o inflamatorios. La presencia de géneros como *Streptococcus* y *Prevotella* evidencia una comunidad microbiana mixta, con coexistencia de bacterias comensales y oportunistas características de las condiciones patológicas analizadas.

Gráfico de barras diez bacterias más abundantes en las muestras de saliva de pacientes (MS).



El gráfico muestra las bacterias más abundantes en saliva control (CS), destacando el predominio de *Prevotella*, seguida de *Streptococcus cristatus* y *Granulicatella elegans*, lo que refleja un perfil microbiano típico de microbiota oral saludable.

7.1.5 COMPARACIÓN DE MUESTRAS

El código compara las bacterias más abundantes en tejido tumoral (MT) con las de otros grupos. Selecciona las 15 especies más representativas de cada conjunto, identifica las bacterias compartidas y las exclusivas del tumor, y luego imprime ambos listados para resaltar posibles diferencias microbianas asociadas al tejido tumoral.

```
# ──────────────────────────────────────────────────
# Paso 6: Comparación con tumores (MT)
# ──────────────────────────────────────────────────
top_mt = set(df_resumen.nlargest(15, 'MT')["nombre_normalizado"])
otros_totales = df_resumen.drop(columns=['nombre_normalizado',
'MT']).sum(axis=1)
top_otros = set(df.loc[otros_totales.nlargest(15).index,
"nombre_normalizado"])

comunes = top_mt.intersection(top_otros)
solo_mt = top_mt.difference(top_otros)

print("\n═══ Bacterias presentes tanto en tumores como en otros grupos ═══")
for b in comunes:
    print(f" - {b}")
print("\n═══ Bacterias exclusivas o dominantes en tumores (MT) ═══")
for b in solo_mt:
    print(f" - {b}")
```

Figura 32:Código comparación con tumores.

Bacterias presentes tanto en tumores como en otros grupos

- Prevotella
- Corynebacterium striatum
- Simonsiella muelleri

Bacterias exclusivas o dominantes en tumores (MT)

- Streptococcus parasanguinis
- Prevotella veroralis
- Sneathia vaginalis
- Leptotrichia trevisanii
- Brachybacterium
- Clostridium
- Burkholderia multivorans
- Lancefieldella párvula
- Fusobacterium
- Brachybacterium huguangmaarense
- Flavonifractor plautii

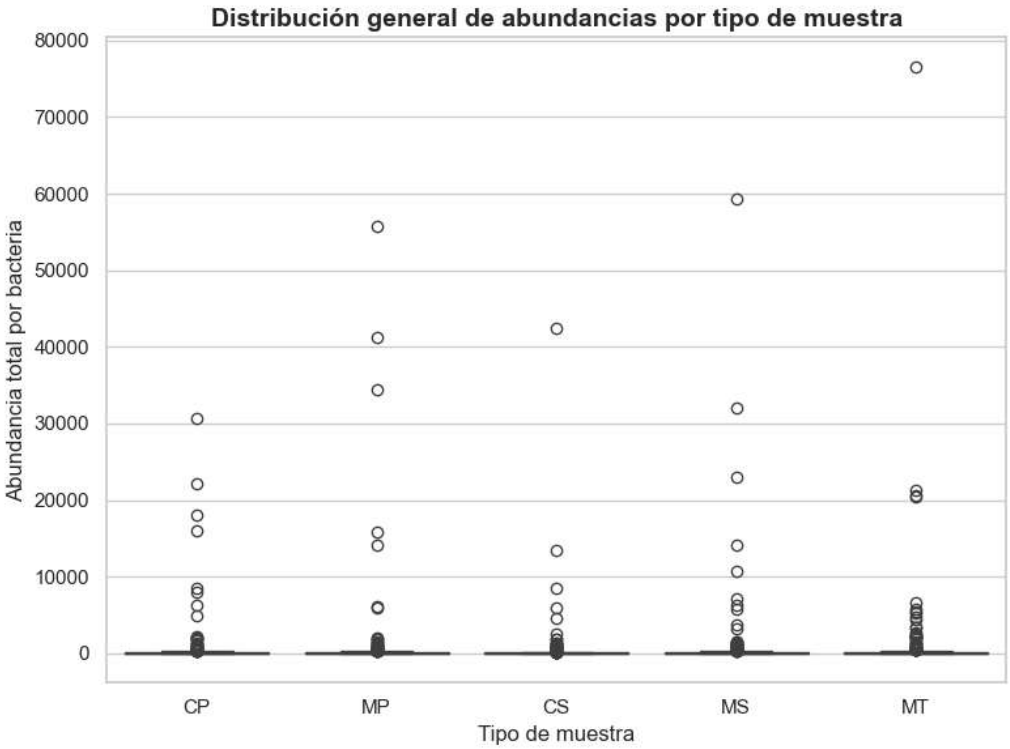


Figura 33:Grafico distribución general por tipo.

El gráfico presenta la distribución de la abundancia total por bacteria en diferentes tipos de muestras (CP, MP, CS, MS, MT).
En el eje vertical se muestra la abundancia total, mientras que en el eje horizontal se encuentran los diferentes tipos de muestras. Los puntos representan las mediciones individuales de abundancia bacteriana para cada tipo de muestra, y la dispersión de estos puntos indica la variabilidad dentro de cada grupo.

Se observa una concentración de los puntos cercanos a la base del gráfico, lo que sugiere que la mayoría de las muestras tienen abundancias bacterianas bajas.

Sin embargo, también se identifican algunos puntos dispersos a alturas mucho mayores, lo que indica que existen muestras con abundancias significativamente altas. Este patrón refleja la heterogeneidad de la abundancia bacteriana entre los tipos de muestra analizados.

8 ANÁLISIS EXPLORATORIO TAXONÓMICO Y DE RESISTENCIA

8.1 OBJETIVO DEL CAPÍTULO

Examinar la relación entre la clasificación taxonómica de las bacterias (familia, género y especie) y los fenotipos de resistencia antibiótica, con el propósito de identificar asociaciones significativas entre ambos componentes.

8.1.1 DESCRIPCIÓN

El análisis se centró en evaluar la relación entre las características fenotípicas de las bacterias del microbioma bucal específicamente los niveles taxonómicos de familia, género y especie y sus perfiles de resistencia antibiótica. Para ello, se integraron las bases de datos fenotípicas y de resistencia mediante la llave *genome_id*, lo que permitió vincular cada organismo con su correspondiente fenotipo de resistencia. Posteriormente, se desarrolló un análisis bivariado que examinó la asociación entre cada nivel taxonómico y la variable *resistant_phenotype*, empleando tanto exploración visual mediante gráficos de distribución como validación estadística utilizando pruebas de Chi-cuadrado. Este enfoque permitió identificar patrones fenotípicos asociados a la resistencia antibiótica y evaluar su potencial como marcadores predictivos, estableciendo la base analítica necesaria para la posterior implementación de modelos supervisados.

Este análisis permitió identificar tendencias diferenciales de resistencia en cada nivel taxonómico, justificando la construcción de un modelo supervisado capaz de evaluar su valor predictivo

Dimensiones del dataset integrado:

- Fenotipos: 1,186,797 registros
- Resistencias: 12,519,617 registros

Distribución del nivel de resistencia:

- BAJO: 10,540,706
- ALTO: 1,960,305
- MEDIO: 18,606

8.1.2 ANÁLISIS BIVARIADO

Para profundizar en la relación entre la estructura taxonómica y el fenotipo antibiótico, se realizó un análisis bivariado entre cada nivel de clasificación (familia, género y especie) y la variable de resistencia. Este análisis permitió identificar patrones diferenciales de comportamiento dentro de cada agrupación taxonómica y evaluar la presencia de tendencias consistentes entre categorías.

8.1.2.1 FAMILIA VS RESISTENCIAS

El análisis entre las familias taxonómicas y el fenotipo antibiótico permitió examinar si existen grupos biológicos con patrones de resistencia diferenciados. Este nivel proporciona una visión general del comportamiento de grandes agrupaciones microbianas y permite identificar familias con tendencias marcadamente resistentes o susceptibles.

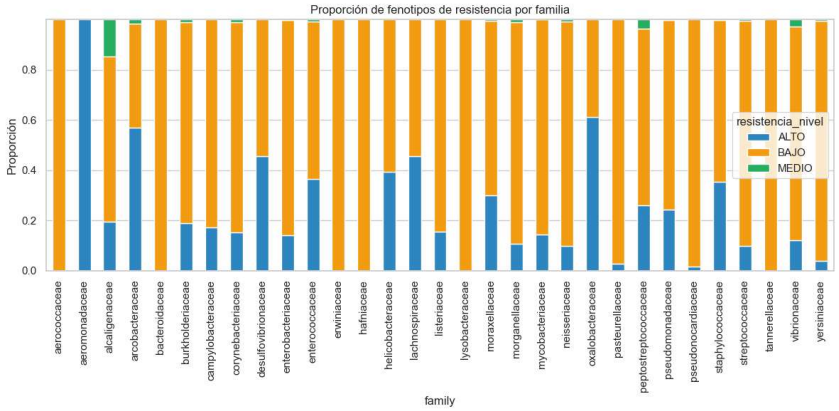


Figura 34:Grafica Familia Vs Resistencias.

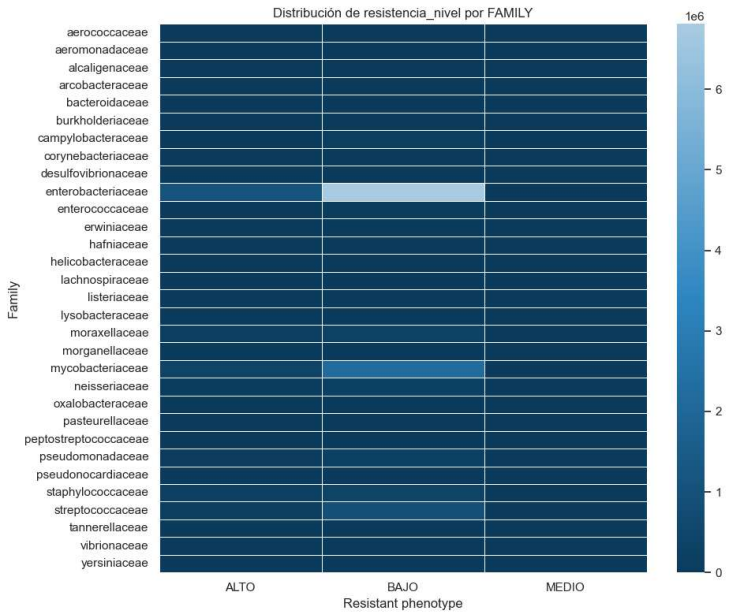


Figura 35:Mapa calor Familia Vs Resistencias.

Se analizaron las proporciones de los niveles de resistencia (ALTO, MEDIO, BAJO) dentro de cada familia. Aunque la mayoría presenta predominio del nivel BAJO, algunas muestran proporciones significativamente mayores de resistencia ALTA, indicando patrones fenotípicos no homogéneos. Estas diferencias sugieren que la variabilidad en resistencia puede estar fuertemente influenciada por características compartidas a nivel de familia.

8.1.2.2 GENERO VS RESISTENCIAS

La exploración de la relación entre los géneros bacterianos y la resistencia antibiótica buscó determinar si las diferencias observadas a nivel de familia se mantienen o se intensifican cuando se analizan agrupaciones más específicas. Este nivel intermedio ofrece una mejor resolución para detectar variabilidad fenotípica.

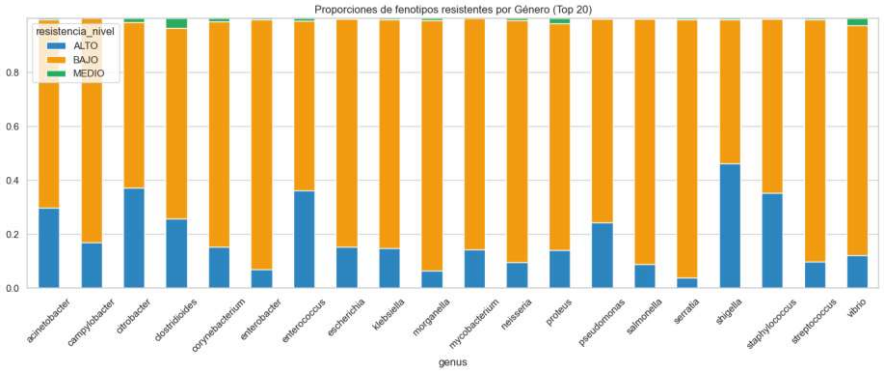


Figura 36:Grafica Genero Vs Resistencias.

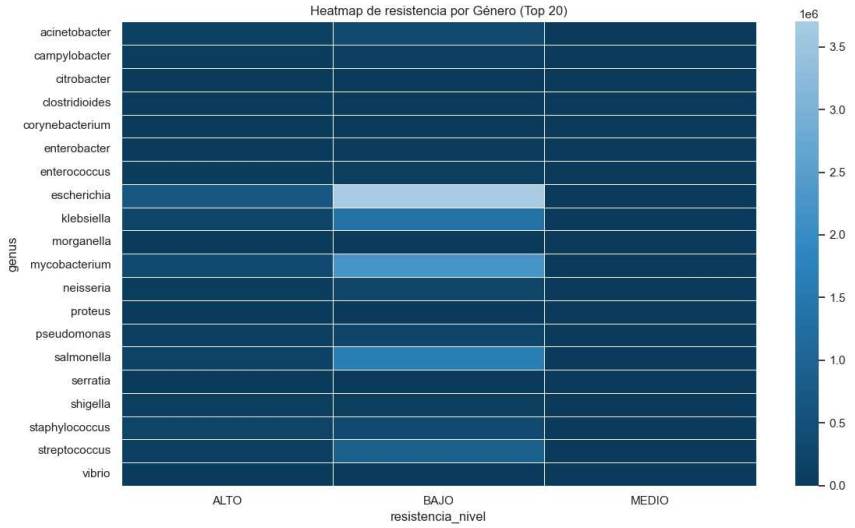


Figura 37:Mapa Calor Genero Vs Resistencias.

El análisis a nivel de género reflejó un patrón similar: predominio del nivel BAJO, pero con géneros que concentran valores notablemente más altos de resistencia ALTA. Estas

diferencias permiten identificar agrupaciones taxonómicas con mayor riesgo relativo de resistencia, lo cual coincide con la evidencia estadística obtenida posteriormente

8.1.2.3 ESPECIE VS RESISTENCIAS

El análisis a nivel de especie permitió evaluar el grado más fino de diferenciación taxonómica, donde se espera encontrar mayor especificidad en los patrones de resistencia. Este nivel ofrece la perspectiva más detallada sobre cómo se distribuye el fenotipo antibiótico dentro del microbioma bucal.

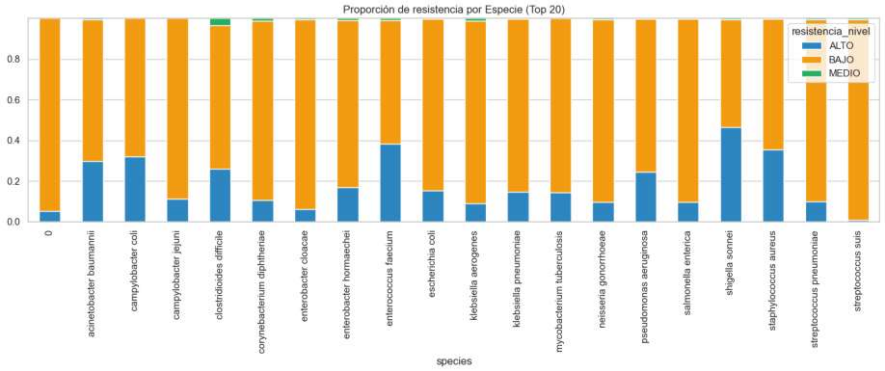


Figura 38:Grafica Especie Vs Resistencias.

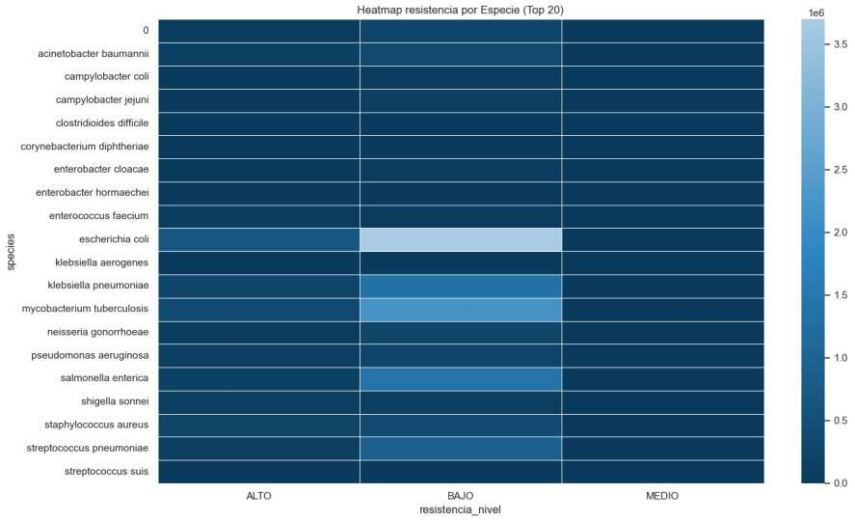


Figura 39: Mapa de calor Especie Vs Resistencias.

A nivel de especie, la heterogeneidad se incrementa, mostrando proporciones diferenciadas y en algunos casos dominancia marcada de resistencia ALTA. Este comportamiento refuerza la importancia de evaluar la resistencia en escalas taxonómicas específicas, dado que la variabilidad interna entre especies de un mismo género puede ser considerable.

8.1.3 VALIDACIÓN ESTADÍSTICA

Con el fin de confirmar si las diferencias observadas en el análisis bivariado son estadísticamente significativas, se aplicó una prueba de Chi-cuadrado para evaluar la asociación entre las variables taxonómicas y el fenotipo antibiótico. Esta validación permite determinar si las variaciones en la resistencia se deben al azar o si reflejan patrones estructurales propios de cada nivel taxonómico.

La prueba de Chi-cuadrado confirmó una asociación estadísticamente significativa entre las categorías taxonómicas y el nivel de resistencia antibiótica. En todos los casos, los valores de Chi² fueron extremadamente altos y el p-value = 0.0, lo que indica que la distribución de los niveles de resistencia no es aleatoria dentro de los grupos taxonómicos.

- **Familia vs resistencia:** χ^2 = 374,913 (df = 60)
  → Diferencias marcadas entre familias y patrones consistentes de resistencia asociada.
- **Género vs resistencia:** χ^2 = 505,491 (df = 38)
  → Variabilidad significativa en los perfiles de resistencia entre géneros.
- **Especie vs resistencia:** χ^2 = 515,327 (df = 38)
  → Mayor heterogeneidad y especificidad predictiva a nivel de especie.

```
FAMILY vs resistencia_nivel
Chi2: 374913.0173021943  p-value: 0.0  df: 60
GENUS vs resistant_phenotype
Chi2: 505491.4232295078  p-value: 0.0  df: 38
SPECIES vs resistant_phenotype
Chi2: 515327.334991516  p-value: 0.0  df: 38
```

Figura 40:Resultado Chi-cuadrado

Estos resultados corroboran que los niveles taxonómicos poseen información estructuralmente relevante y no aleatoria respecto al fenotipo antibiótico, posicionándolos como variables con potencial explicativo para la modelación supervisada.

9 CONSTRUCCIÓN DEL MODELO

9.1 OBJETIVO DEL CAPÍTULO

Implementar un modelo predictivo supervisado, basado en regresión logística, para estimar la probabilidad de resistencia antibiótica en bacterias bucales y evaluar su desempeño mediante métricas de validación.

9.1.1 DESCRIPCIÓN

Los modelos se desarrollaron utilizando una formulación binaria de la variable resistencia_nivel, donde la clase 1 representa los aislamientos con resistencia alta, mientras que la clase 0 agrupa los niveles bajo y medio. Este enfoque permite replicar un escenario clínico en el que el objetivo principal es distinguir los casos que representan riesgo terapéutico significativo.

Para preparar los datos, las variables categóricas de alta cardinalidad (*family, genus, species y antibiotic*) fueron transformadas mediante *One-Hot Encoding*, generando una matriz dispersa altamente dimensional. La división del conjunto de datos se realizó utilizando un 80 % para entrenamiento y un 20 % para prueba, con muestreo estratificado para preservar la proporción real entre clases, un aspecto fundamental dada la marcada descompensación observada entre ellas.

La Regresión Logística se entrenó utilizando el *solver saga*, dada su eficiencia en contextos de gran escala y su comportamiento estable frente a matrices dispersas. Este modelo constituye el principal referente del capítulo debido a su interpretabilidad y a su capacidad para generar estimaciones probabilísticas robustas. En paralelo, se incorporó un modelo *Naive Bayes (BernoulliNB)* como línea base probabilística, seleccionado por su alta eficiencia computacional y su desempeño aceptable en estructuras dispersas generadas por *One-Hot Encoding*. La comparación entre ambos algoritmos permitió evaluar cómo responden dos enfoques clásicos uno lineal y otro probabilístico ante el reto de un *dataset* masivo, desbalanceado y de alta dimensionalidad.

9.1.2 MÉTRICAS GENERALES DE DESEMPEÑO

Con el fin de evaluar el desempeño global de los modelos, se analizaron métricas agregadas que permiten una comparación inicial entre enfoques, considerando el fuerte desbalance presente en el conjunto de datos.

Modelo	Accuracy	F1-score
Regresión Logística	0.752	0.498
Naive Bayes (Bernoulli)	0.822	0.341

Tabla 8: Métricas generales de desempeño.

La Regresión Logística alcanzó una exactitud global de 0.752 y un F1-score de 0.498, lo que indica un desempeño equilibrado bajo condiciones de desbalance extremo. En contraste, el modelo Naive Bayes obtuvo un accuracy superior0.822, sin embargo, este valor está fuertemente influenciado por la predominancia de la clase no resistente, por lo que resulta insuficiente como métrica única de comparación.
En consecuencia, aunque Naive Bayes presenta una mayor exactitud global, la Regresión Logística ofrece un mejor balance entre precisión y sensibilidad, lo cual resulta más adecuado para el objetivo del estudio.

9.1.3 MÉTRICAS POR CLASE:

Dado que la identificación de la resistencia antibiótica constituye el foco principal del análisis, se evaluó el desempeño de los modelos a nivel de clase.

Clase	Precision	Recall	F1-score
No resistente	0.95	0.75	0.84
Resistente	0.37	0.79	0.50

Tabla 9: Métricas por clase – Regresión Logística.

Clase	Precision	Recall	F1-score
No resistente	0.88	0.92	0.90
Resistente	0.41	0.29	0.34

Tabla 10: Métricas por clase – Naive Bayes.

La **Regresión Logística** mostró un desempeño consistente en la clase no resistente y destacó especialmente en la clase resistente, donde alcanzó un **recall de 0.79**, lo que indica una alta capacidad para identificar correctamente los aislamientos resistentes.
Por el contrario, el modelo **Naive Bayes** presentó una recuperación limitada de la clase resistente **recall de 0.29**, lo que implica una mayor proporción de casos resistentes no detectados. Desde una perspectiva clínica y epidemiológica, este comportamiento representa una limitación significativa.

9.1.4 MATRIZ DE CONFUSIÓN

El análisis de la matriz de confusión permite examinar de forma detallada los aciertos y errores de clasificación de cada modelo.

	Predicción: No resistente	Predicción: Resistente
Real: No resistente	1,691,969	573,992
Real: Resistente	90,060	330,063

Tabla 11: Matriz de confusión – Regresión Logística.

	Predicción: No resistente	Predicción: Resistente
Real: No resistente	2,084,954	181,007
Real: Resistente	296,314	123,809

Tabla 12: Matriz de confusión – Naive Bayes.

En la Regresión Logística, el número de falsos negativos 90,060 es considerablemente menor, lo que refleja una estrategia orientada a priorizar la detección de resistencia, aun a costa de un mayor número de falsos positivos. Este enfoque resulta coherente con los objetivos del estudio, donde la omisión de casos resistentes tiene mayores implicaciones clínicas.

En contraste, el modelo Naive Bayes presentó un volumen elevado de falsos negativos en la clase resistente, lo que confirma su menor capacidad para identificar patrones complejos asociados a la resistencia antibiótica.

9.1.5 IMPORTANCIA DE LAS VARIABLES

La Regresión Logística permite interpretar cada coeficiente como la contribución relativa de las categorías correspondientes a la probabilidad estimada de resistencia. Las variables con los coeficientes más altos representan mayor asociación con resistencia, mientras que los coeficientes negativos indican contextos donde la resistencia es menos probable.

Principales variables asociadas positivamente con resistencia

- *species_citrobacter sp. fdaargos_156*
- *species_corynebacterium striatum*
- *species_streptococcus agalactiae*

- *species_serratia rubidaea*
- *antibiotic_sulfonamides*
- *species_proteus mirabilis*
- *antibiotic_doxycycline*

Estos predictores señalan especies y antibióticos fuertemente vinculados con la probabilidad de observar resistencia elevada en los datos analizados.

Variables asociadas a menor probabilidad de resistencia

- *antibiotic_fosfomycin*
- *antibiotic_tigecyklin*
- *species_burkholderia cepacia*
- *antibiotic_ceftazidime/clavulanic acid*

Estas categorías aparecen consistentemente asociadas con niveles más bajos de resistencia, proporcionando información relevante para el análisis comparativo de efectos.

En el caso de *Naive Bayes*, la interpretación de los parámetros no posee la misma capacidad explicativa debido a la naturaleza del modelo. Aunque permite calcular probabilidades condicionadas, la suposición de independencia entre características limita la inferencia sobre la interacción real entre las variables. Esto refuerza la ventaja interpretativa de la Regresión Logística como herramienta analítica dentro del estudio.

10 CONCLUSIONES

El presente proyecto tuvo como propósito analizar e identificar patrones de resistencia antibiótica en bacterias bucales mediante la integración de características fenotípicas y taxonómicas y el uso de técnicas de aprendizaje supervisado. Para ello, se desarrolló un flujo metodológico que incluyó la depuración y estandarización de cuatro fuentes de datos provenientes del repositorio BV-BRC, la consolidación de un *dataset* unificado, el análisis exploratorio de variables clínicas y taxonómicas, la caracterización de la relación entre clasificación biológica y resistencia, y finalmente la implementación y evaluación de un modelo de regresión logística binaria. Este enfoque permitió combinar estadística descriptiva, minería de datos y modelado predictivo para comprender mejor la dinámica de resistencia en microbiomas bucales.

Se integró, limpió y normalizó el conjunto de datos procedente de cuatro fuentes heterogéneas, lo cual permitió construir un *dataset* estandarizado, coherente y libre de duplicados. Durante el proceso de integración se evaluó la relación entre las bases de datos mediante métricas de coincidencia. La unión entre las muestras clínicas y los identificadores genómicos mostró un 78.85 % de coincidencia, con una pérdida del 21.15 % debido a registros que no lograban vincularse. En la integración posterior con la información taxonómica, la coincidencia aumentó al 90.62 %, mientras que la unión final con los datos fenotípicos de resistencia alcanzó un 82.99 % de correspondencia. Estos resultados evidenciaron la calidad del enlace entre las fuentes y permitieron obtener un *dataset* final completamente funcional para análisis estadístico y modelado, en el que quedaron correctamente alineadas las variables taxonómicas, fenotípicas y los metadatos clínicos.

Se realizó el análisis exploratorio de las variables clínicas y taxonómicas, lo que permitió identificar patrones relevantes en la distribución de tipo de muestra, edad, sexo y zona anatómica, así como caracterizar la abundancia de especies presentes en el microbioma bucal. Los resultados mostraron variabilidad en los fenotipos de resistencia entre grupos y pusieron en evidencia que un subconjunto de especies concentró una proporción importante de los aislamientos observados, lo que sugiere la existencia de perfiles clínicos y microbiológicos recurrentes asociados a determinados contextos de resistencia.

Se examinó la relación entre clasificación taxonómica y niveles de resistencia antibiótica, considerando tres niveles jerárquicos principales: familia, género y especie. Este análisis permitió identificar asociaciones claras entre determinados taxones y la prevalencia de fenotipos resistentes. En particular, especies como *Citrobacter sp., Corynebacterium striatum, Streptococcus agalactiae y Proteus mirabilis* mostraron una mayor presencia dentro del grupo de aislamientos resistentes frente a los susceptibles, mientras que otras especies presentaron patrones fenotípicos más favorables. Estas evidencias confirmaron que la taxonomía constituye un marcador informativo relevante para apoyar la predicción de resistencia en ausencia de datos genómicos detallados.

Se implementó y evaluó un modelo de regresión logística binaria empleando codificación categórica y validación mediante división *train–test*. El modelo alcanzó un desempeño moderado, con un *accuracy* del 75.2 % y un *F1-score* de 0.49 para la clase resistente, lo que evidencia una capacidad aceptable de discriminación en un contexto de desbalance entre categorías. El análisis de los coeficientes del modelo mostró asociaciones significativas entre determinadas combinaciones fenotípicas y taxonómicas y una mayor probabilidad de resistencia, confirmando el valor predictivo de estas variables en el contexto de las bacterias bucales.

En conjunto, los resultados del proyecto demuestran la viabilidad de integrar datos fenotípicos y taxonómicos para caracterizar y predecir resistencia antibiótica en bacterias bucales mediante métodos computacionales. El estudio aporta una metodología reproducible que combina procesamiento de datos, análisis descriptivo y modelado supervisado, constituyendo una base sólida para sistemas de vigilancia microbiológica basados en datos.

11 TRABAJOS FUTUROS

El desarrollo de este proyecto permitió establecer una metodología integral para la integración, análisis y modelado predictivo de datos fenotípicos y taxonómicos asociados a resistencia antibiótica en bacterias bucales. No obstante, los resultados alcanzados y las limitaciones identificadas abren múltiples líneas de investigación que permitirían fortalecer y ampliar el alcance del estudio.

En primer lugar, se recomienda explorar modelos de aprendizaje supervisado más avanzados que superen las restricciones inherentes a la regresión logística en contextos de alta dimensionalidad y relaciones no lineales. Técnicas *como Random Forest, XGBoost, Support Vector Machines (SVM)* o redes neuronales profundas podrían mejorar el rendimiento predictivo, manejar de forma más eficiente el desbalance entre clases y capturar patrones complejos entre la taxonomía bacteriana y la resistencia fenotípica.

En segundo lugar, resulta pertinente ampliar el conjunto de antibióticos y fenotipos analizados, incorporando nuevos agentes antimicrobianos y aprovechando datos cuantitativos como valores de concentración inhibitoria mínima (MIC), cuando estén disponibles. El uso de información más granular permitiría construir modelos más sensibles y robustos que reflejen con mayor precisión la variabilidad real de los perfiles de resistencia.

Una línea de investigación especialmente relevante consiste en integrar información genética o metagenómica a nivel de genes de resistencia, factores de virulencia o variantes específicas. La combinación de datos fenotípicos, taxonómicos y genómicos podría potenciar la capacidad de los modelos para identificar marcadores predictivos con mayor precisión, permitiendo una caracterización más completa de los mecanismos asociados a la resistencia antibiótica.

Asimismo, se recomienda aplicar técnicas de interpretabilidad de modelos, como *SHAP values o Permutation Importance,* con el fin de complementar el análisis de coeficientes y permitir una comprensión más profunda de las variables que influyen en la predicción de resistencia. Esto fortalecería la transparencia del modelo y facilitaría su uso en entornos clínicos y epidemiológicos.

Otra línea de trabajo consiste en expandir el tamaño y diversidad del *dataset* clínico, integrando muestras provenientes de diferentes regiones geográficas, contextos clínicos y poblaciones. Esto permitiría evaluar la estabilidad del modelo, mejorar su capacidad de generalización y reducir posibles sesgos asociados a la distribución original de los datos.

Finalmente, una proyección de alto impacto consiste en desarrollar un sistema automatizado de vigilancia microbiológica basado en predicciones, capaz de procesar continuamente nuevos datos fenotípicos y taxonómicos y generar alertas tempranas sobre patrones emergentes de resistencia antibiótica. Un sistema de este tipo podría integrarse en flujos clínicos o herramientas de

investigación, aportando valor en la toma de decisiones terapéuticas y en la vigilancia epidemiológica.

En conjunto, estos trabajos futuros constituyen un marco sólido para avanzar hacia modelos predictivos más precisos, explicables y aplicables en el estudio de la resistencia antibiótica, potenciando el uso de herramientas computacionales en beneficio del diagnóstico, la vigilancia y la salud pública.

12 REFERENCIAS BIBLIOGRÁFICAS

[1] World Health Organization, *Antimicrobial Resistance: Global Report on Surveillance*, WHO Press, 2014.

[2] A. R. Marsh and D. A. Zaura, "The oral microbiome: structure, function, and its role in resistance," *Journal of Oral Microbiology*, vol. 12, no. 1, pp. 1–12, 2020.

[3] Clinical and Laboratory Standards Institute (CLSI), *Performance Standards for Antimicrobial Susceptibility Testing*, 32nd ed., 2022.

[4] BV-BRC Data Portal, "Bacterial and Viral Bioinformatics Resource Center," 2024. [Online]. Available: https://www.bv-brc.org

[5] H. Li et al., "Challenges in integrating heterogeneous biological datasets for antimicrobial resistance analysis," *Bioinformatics*, vol. 38, no. 4, pp. 1021–1029, 2022.

[6] C. F. Gonzalez, A. Wilson, and P. Harrison, "Taxonomic signatures associated with antimicrobial resistance profiles in clinical isolates," *Microbial Genomics*, vol. 7, no. 11, pp. 1–12, 2021.

[7] S. Federhen, "The NCBI Taxonomy database," *Nucleic Acids Research*, vol. 40, no. D1, pp. D136–D143, 2012.

[8] C. Wittouck et al., "A genome-based species taxonomy of the Lactobacillus genus complex," *Nature Microbiology*, vol. 5, pp. 251–259, 2020.

[9] P. Parks et al., "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life," *Nature Biotechnology*, vol. 36, pp. 996–1004, 2018.

[10] C. Linnaeus, *Systema Naturae*, 10th ed., Stockholm: Laurentius Salvius, 1758.

[11] M. K. Tindall, "The use of taxonomic names in microbiology and the role of nomenclature," *International Journal of Systematic and Evolutionary Microbiology*, vol. 61, pp. 2775–2780, 2011.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[13] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed. Packt Publishing, 2020.

[14] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[15] D. W. Hosmer, S. Lemeshow y R. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.