



Implementación de un modelo para predecir la resistencia a carbapenémicos en *Klebsiella pneumoniae* mediante un algoritmo de *Machine Learning*

Paredes Escobar, Michelle Marcela

Departamento de Ciencias de la Vida y de la Agricultura

Carrera de Ingeniería en Biotecnología

Trabajo de titulación, previo a la obtención del título de Ingeniera en Biotecnología

Grijalva Silva, Rodrigo Marcelo M. D. PhD.

05 de marzo del 2021

Certificación Urkund



Document Information

Analyzed document	Tesis_Michelle_Paredes_Revisada 032021.docx (D97220643)
Submitted	3/4/2021 3:14:00 PM
Submitted by	
Submitter email	rmgrijalva@espe.edu.ec
Similarity	3%
Analysis address	rmgrijalva.espe@analysis.orkund.com

Sources included in the report

W	URL: https://doi.org/10.1093/bib/bbx083 Fetched: 3/4/2021 3:46:00 PM	 1
SA	Cubiella_Victorero-Tamara.pdf Document Cubiella_Victorero-Tamara.pdf (D54375916)	 6
SA	veraalejandra_prog4232.pdf Document veraalejandra_prog4232.pdf (D77760118)	 3
SA	Trabajo de Titulación_Estefanía Mariño_FOSFOMICINA v8.docx Document Trabajo de Titulación_Estefanía Mariño_FOSFOMICINA v8.docx (D58316987)	 2
W	URL: https://zagan.unizar.es/record/47451/files/TAZ-TFM-2015-161.pdf Fetched: 1/18/2020 6:43:21 PM	 1



Firmado electrónicamente por:
**RODRIGO MARCELO
GRIJALVA SILVA**



Departamento de Ciencias de la Vida y de la Agricultura

Carrera de Ingeniería en Biotecnología

Certificación

Certifico que el trabajo de titulación, denominado “**Implementación de un modelo para predecir la resistencia a carbapenémicos en *Klebsiella pneumoniae* mediante un algoritmo de *Machine Learning***” fue realizado por la señorita **Paredes Escobar, Michelle Marcela** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto, cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 05 de marzo del 2021

Firma:



Firmado electrónicamente por:
**RODRIGO MARCELO
GRIJALVA SILVA**

.....
Dr. Grijalva Silva, Rodrigo Marcelo M.D., PhD.

C.C.: 1706590641



Departamento de Ciencias de la Vida y de la Agricultura

Carrera de Ingeniería en Biotecnología

Responsabilidad de Autoría

Yo, **Paredes Escobar, Michelle Marcela**, con C.C. 1722869300, declaro que el contenido, ideas y criterios del trabajo de titulación: **“Implementación de un modelo para predecir la resistencia a carbapenémicos en *Klebsiella pneumoniae* mediante un algoritmo de *Machine Learning*”** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 05 de marzo del 2021

Firma:



Paredes Escobar, Michelle Marcela

C.C. 1722869300



Departamento de Ciencias de la Vida y de la Agricultura

Carrera de Ingeniería en Biotecnología

Autorización de publicación

Yo, **Paredes Escobar, Michelle Marcela**, con C.C. 1722869300, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Implementación de un modelo para predecir la resistencia a carbapenémicos en *Klebsiella pneumoniae* mediante un algoritmo de *Machine Learning*"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad

Sangolquí, 05 de marzo del 2021

Firma:



Paredes Escobar, Michelle Marcela

C.C. 1722869300

Dedicatoria

Dedicado a la memoria de mi abuela Olga Alicia Barberán Solórzano y mi amigo Geovanny

Antonio Meza Maldonado.

Agradecimientos

A mis padres y hermanos por su apoyo incondicional en el transcurso de mi carrera universitaria.

Al Dr. Marcelo Grijalva por abrirme las puertas de su laboratorio y permitirme desarrollar mis habilidades en investigación contribuyendo a mi formación profesional.

A la Dra. Sonia Cárdenas por su constante apoyo, guía y motivación en la realización de este proyecto.

A la M. Sc. Silvana Granda, quien fue como una segunda madre, por las incontables enseñanzas, consejos, reflexiones, ánimos, risas y demás apoyo brindado durante estos años.

A los docentes Alma Koch, Claudia Segovia, Ligia Ayala y Rodrigo Ávalos por ver en mí diversas cualidades y otorgarme oportunidades en el ámbito investigativo.

Al Ing. Leonardo Tayupanta por ser la inspiración y motivación en este proyecto.

A mis compañeras de laboratorio Lizeth, Tatiana, Kirsty, Judith, Andrea, Tannya y Abigail por su apoyo ante cualquier inquietud que presentase. Sus risas y compañerismo en este último par de años formaron parte de mi construcción personal y profesional.

A todos los que conforman la Asociación de Estudiantes de Biotecnología AEIB-ESPE por permitirme explorar una nueva faceta en mi vida.

A mis amigos Andrés, Anita, Iván, Santiago, Ricardo, Martín, Michelle, Melanie y Dani. Fueron varias horas las que pasábamos juntos a diario demostrando que somos una familia.

A mis mejores amigos Daniela, David, Juan y Paola. La vida es más sencilla si siempre estamos juntitos.

Índice de Contenido

Certificación Urkund	2
Certificación	3
Responsabilidad de Autoría.....	4
Autorización de publicación	5
Dedicatoria.....	6
Agradecimientos.....	7
Índice de Contenido	8
Índice de Tablas	12
Índice de Figuras	14
Abreviaturas.....	16
Resumen	18
Abstract.....	19
Capítulo I: Introducción.....	20
Antecedentes	20
Justificación.....	22
Objetivos de trabajo de titulación.....	23
Objetivo general.....	23
Objetivo específico.....	23
Capítulo II: Revisión Bibliográfica.....	25

Generalidades	25
Taxonomía	25
Morfología	26
Epidemiología	27
Factores de virulencia	27
Lipopolisacáridos.....	28
Polisacáridos de cápsula.....	29
Fimbrias.....	29
Proteínas de membrana externa.....	30
Sideróforos.....	30
Fuentes de nitrógeno	31
Resistencia los antimicrobianos	31
Resistencia a los carbapenémicos	31
Clase A.....	32
Clase B.....	32
Clase D.....	33
Estudio de asociación de genoma completo (GWAS).....	34
Machine learning (ML)	37
Capítulo III: Metodología.....	40
Esquema de desarrollo.....	40

	10
Hardware	41
Software.....	41
Set de datos	41
Preprocesamiento de las secuencias de genoma completo.....	42
Conteo de unitigs	43
Filtrado de unitig asociados al fenotipo	43
Entrenamiento y evaluación del modelo	44
Capítulo IV: Resultados	47
Set de datos	47
Preprocesamiento de las secuencias de genoma completo.....	48
Análisis de resultados de aprendizaje del modelo de predicción de susceptibilidad a imipenem	50
Análisis del experimento 1	50
Análisis del experimento 2	51
Análisis del experimento 3	52
Análisis del experimento 4	53
Análisis general de los experimentos	54
Análisis de resultados de aprendizaje del modelo de predicción de susceptibilidad a meropenem	55
Análisis del experimento 1	55
Análisis del experimento 2	56

	11
Análisis del experimento 3	57
Análisis del experimento 4	58
Análisis general de los experimentos	59
Capítulo V: Discusión	61
Capítulo VI: Conclusiones y Recomendaciones	68
Conclusiones	68
Recomendaciones	69
Bibliografía	70

Índice de Tablas

Tabla 1. <i>Clasificación taxonómica de K. pneumoniae.</i>	25
Tabla 2. <i>Parámetros de búsqueda de genoma de K. pneumoniae en PATRIC</i>	42
Tabla 3. <i>Ejemplo de la matriz presencia/ausencia de unitigs en el aprendizaje del modelo</i>	43
Tabla 4. <i>Parámetros de entrenamiento</i>	45
Tabla 5. <i>Diseño experimental</i>	45
Tabla 6. <i>Resumen de fenotipos de susceptibilidad del set de datos</i>	47
Tabla 7. <i>Resumen del contenido de la matriz presencia/ausencia de unitigs</i>	49
Tabla 8. <i>Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento</i> <i>1</i>	50
Tabla 9. <i>Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento</i> <i>2</i>	51
Tabla 10. <i>Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento</i> <i>3</i>	52
Tabla 11. <i>Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento</i> <i>4</i>	53
Tabla 12. <i>Desempeño del modelo de predicción de susceptibilidad a meropenem en el</i> <i>experimento 1</i>	55
Tabla 13. <i>Desempeño del modelo de predicción de susceptibilidad a meropenem en el</i> <i>experimento 2</i>	56

Tabla 14. <i>Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 3</i>	57
Tabla 15. <i>Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 4</i>	58

Índice de Figuras

Figura 1. <i>Línea de tiempo de identificación de genes de resistencia que expresan carbapenemasas.....</i>	21
Figura 2. <i>Estructura y morfología de K. pneumoniae.....</i>	26
Figura 3. <i>Factores de virulencia.....</i>	28
Figura 4. <i>Estructura del lipopolisacárido.....</i>	29
Figura 5. <i>Construcción del gráfico De Bruijn compactado.....</i>	36
Figura 6. <i>Estructura de un árbol de decisión.....</i>	39
Figura 7. <i>Generación de in árbol de decisión.....</i>	39
Figura 8. <i>Esquema de desarrollo.....</i>	40
Figura 9. <i>Porcentaje de aislados de K. pneumoniae susceptibles reportados en el set de datos...47</i>	
Figura 10. <i>Gráfico Q-Q del análisis de GWAS entre el los unitigs de K. pneumoniae y el fenotipo de susceptibilidad a imipenem.....</i>	48
Figura 11. <i>Gráfico Q-Q del análisis de GWAS entre los unitigs de K. pneumoniae y el fenotipo de susceptibilidad a meropenem.....</i>	49
Figura 12. <i>Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 1.....</i>	50
Figura 13. <i>Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 2.....</i>	51
Figura 14. <i>Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 3.....</i>	52

Figura 15. <i>Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 4</i>	53
Figura 16. <i>Curva ROC del modelo de predicción de susceptibilidad a imipenem</i>	54
Figura 17. <i>Tiempo de entrenamiento de los modelos de predicción de susceptibilidad a imipenem</i>	55
Figura 18. <i>Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 1</i>	56
Figura 19. <i>Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 2</i>	57
Figura 20. <i>Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 3</i>	58
Figura 21. <i>Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 4</i>	59
Figura 22. <i>Curva ROC del modelo de predicción de susceptibilidad a meropenem</i>	60
Figura 23. <i>Tiempo de entrenamiento de los modelos de predicción de susceptibilidad a meropenem</i>	60

Abreviaturas

AI	Inteligencia artificial (siglas en inglés)
ML	Machine learning (aprendizaje automático)
NGS	Secuenciación de siguiente generación (siglas en inglés)
GWAS	Estudio de asociación de genoma completo (siglas en inglés)
XGBoost	Extreme gradient boosting
KPC	<i>Klebsiella pneumoniae</i> carbapenemasas
IMP	Imipenemasa
VIM	Verona metalo- β -lactamasa codificada por integrones
OXA	Oxacilinasas
NMD	New Delhi metalo- β -lactamasa
CRE	Enterobacterias resistentes a carbapenémicos
LPS	Lipolisacáridos (siglas en inglés)
CPS	Polisacáridos de cápsula (siglas en inglés)
BLEE	β -lactamasas de espectro extendido
CHDL	β -lactamasas de clase D hidrolizantes de carbapenem
SNP	Polimorfismo de nucleótido simple (siglas en inglés)
cDBG	Gráfico De Bruijn compactado (siglas en inglés)
LMM	Modelos lineales mixtos (siglas en inglés)

LD	Desequilibrio de ligamiento (siglas en inglés)
NB	Naive Bayes (siglas en inglés)
DT	Árboles de decisión (siglas en inglés)
RF	Bosques aleatorios (siglas en inglés)
SVM	Máquinas de vectores de soporte (siglas en inglés)
ANN	Redes neuronales artificiales (siglas en inglés)
NCBI	Centro Nacional de Información Biotecnológica (siglas en inglés)
PATRIC	Centro de Integración de Recursos de Pathosystems (siglas en inglés)
MB	Megabits
GATB	Genome Assembly & Analysis Tool Box (siglas en inglés)
MD	Profundidad máxima (siglas en inglés)
CV	Validación cruzada (siglas en inglés)
VP	Verdaderos positivos
FP	Falsos positivos
VN	Verdaderos negativos
FN	Falsos negativos
ROC	Característica del funcionamiento del receptor (siglas en inglés)
AUC	Área bajo la curva (siglas en inglés)
Q-Q	Cuantil-Cuantil

Resumen

Klebsiella pneumoniae es un patógeno oportunista asociado al ambiente intrahospitalario y que provoca infecciones graves en pacientes en riesgo. El tratamiento antibiótico puede incluir carbapenémicos en episodios infecciosos causados por *K. pneumoniae* resistente a otros antibióticos. No obstante, *K. pneumoniae* ha adquirido resistencia a múltiples antibióticos, convirtiéndose en un problema de salud pública. El diagnóstico tardío y la prescripción no específica han aumentado las tasas de mortalidad. Las herramientas de Inteligencia Artificial (AI) se han convertido en un apoyo a las técnicas tradicionales para el diagnóstico y prescripción de un tratamiento. En este proyecto se implementó un modelo basado en *Machine Learning* (ML) para predecir la resistencia de *K. pneumoniae* a imipenem y meropenem. La extensa cantidad de datos de secuenciación de nueva generación (NGS) permitió formar un set de datos para el entrenamiento del modelo de ML. El preprocesamiento basado en estudios de asociación de genoma completo (GWAS) corrigió los problemas asociados a la estructura poblacional bacteriana y la dimensionalidad. Los modelos se entrenaron mediante un algoritmo *Extreme Gradient Boosting* (XGBoost) optimizando tiempo y recursos computacionales. Los resultados de aprendizaje demostraron la capacidad de los modelos basados en ML para predecir fenotipos de susceptibilidad antimicrobiana. No obstante, las métricas de evaluación se podrían mejorar aumentando la cantidad de aislados.

Palabras clave:

- **KLEBSIELLA PNEUMONIAE**
- **CARBAPENÉMICOS**
- **MACHINE LEARNING (ML)**
- **EXTREME GRADIENT BOOSTING (XGBOOST)**

Abstract

Klebsiella pneumoniae is an opportunistic pathogen associated with the hospital environment and able to cause serious infections in vulnerable patient groups. The treatment, in infections caused by resistant isolates includes the administration of carbapenems. However, *K. pneumoniae* has acquired resistance to multiple antibiotics, becoming a public health problem. Untimely diagnosis and non-targeted prescription have increased mortality rates. Artificial intelligence (AI) tools have become a support for traditional techniques for diagnosing and prescribing a treatment. In this project, a model based on Machine Learning (ML) was implemented to predict the resistance of *K. pneumoniae* to imipenem and meropenem. The large amount of New Generation Sequencing (NGS) data made it possible to create a data set for training the ML model. The preprocessing based in genome-wide association studies (GWAS) corrected problems associated with bacterial population structure and dimensionality. The models were trained using an Extreme Gradient Boosting (XGBoost) algorithm, optimizing time and computational resources. The learning outcomes demonstrated the ability of ML-based models to predict antimicrobial susceptibility phenotypes. However, the evaluation metrics could be improved by increasing the number of isolates.

Key words:

- **KLEBSIELLA PNEUMONIAE**
- **CARBAPENÉMICOS**
- **MACHINE LEARNING (ML)**
- **EXTREME GRADIENT BOOSTING (XGBOOST)**

Capítulo I: Introducción

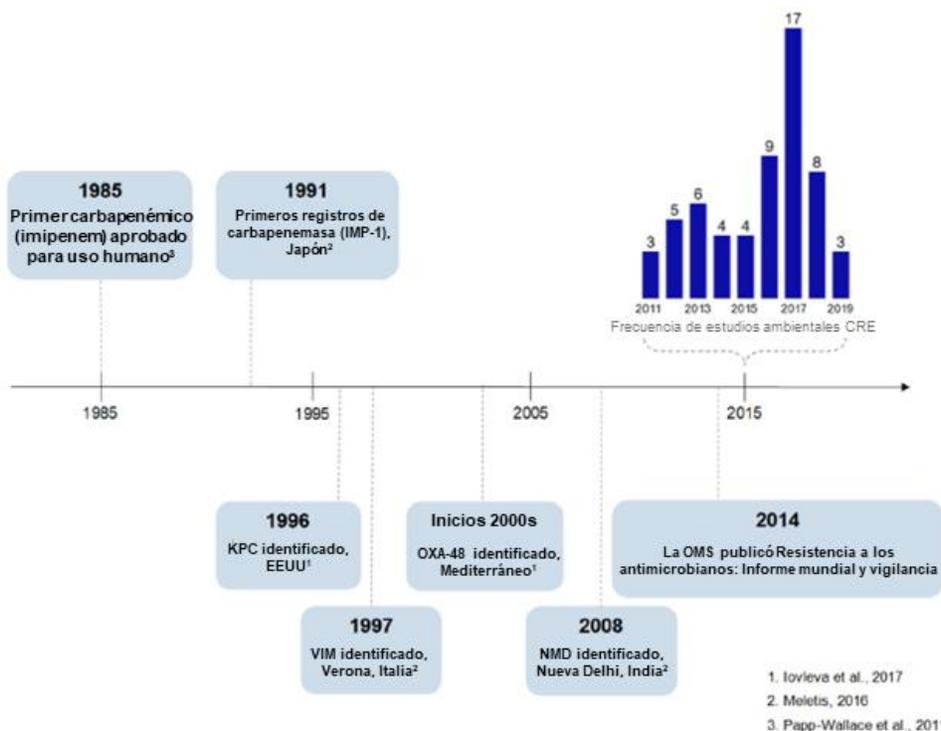
Antecedentes

Klebsiella pneumoniae es un microorganismo de distribución ubicua que se encuentra presente de manera asintomática en el tracto gastrointestinal en la tercera parte de la población mundial. Sin embargo, es un patógeno oportunista asociado al ambiente intrahospitalario y responsable de infecciones graves en grupos vulnerables (Opoku-Temeng et al., 2019; Prado-Vivar et al., 2019). Posee resistencia intrínseca a la amoxicilina y tetraciclina. No obstante, en el transcurso del tiempo, ha desarrollado resistencia a múltiples antibióticos, adquirida por mutaciones o por elementos genéticos móviles (Figura 1) (Meletis, 2016). Fue descrita por primera vez por Edwin Klebs en 1875 mientras estudiaba las vías respiratorias de pacientes que murieron por neumonía. En 1882, Carl Friedlander detalló formalmente la especie considerada responsable de esta patología (Bengoechea & Sa Pessoa, 2019; Long et al., 2017).

Klebsiella pneumoniae carbapenemasas (KPC) son las enzimas más comunes en todo el mundo capaces de hidrolizar penicilina, cefalosporinas, monobactámicos y carbapenémicos. En 1996, en Carolina del Norte – EE. UU., se reportó el primer caso de *K. pneumoniae* productora de KPC codificada por el gen blaKPC-1, posteriormente se identificaron 22 isotipos logrando propagarse rápidamente por el mundo llegando a América del Sur, Europa y Asia. Las infecciones provocadas por microorganismos productores de KPC poseen elevadas tasas de mortalidad, entre el 32 y 72%, y están asociadas comúnmente con diagnóstico tardío y administración de antibióticos inapropiados (Mills & Lee, 2019; Plazak et al., 2018).

Figura 1

Línea de tiempo de identificación de genes de resistencia que expresan carbapenemasas



Nota. El gráfico muestra la línea de tiempo del primer uso de los antibióticos carbapenémicos y posterior identificación de genes de resistencia que expresan carbapenemasas: IMP, KPC, VIM, OXA y NDM. Modificado de Mills & Lee (2019).

Las primeras metalo- β -lactamasas se reportaron en aislados de *Pseudomonas aeruginosa*: imipenasa (IMP) en Japón en 1988 y Verona metalo- β -lactamasa codificada por integrones (VIM) en Verona – Italia en 1997. Posteriormente se reportaron casos endémicos de *K. pneumonia* productora de IMP y VIM en Grecia, Taiwan y Japón (Escandón-Vargas et al., 2017; Poirel et al., 2004).

Las oxacilinasas son enzimas con alta afinidad para hidrolizar imipenem. En 2001, se reportó el primer caso de *K. pneumoniae* productora de OXA-48 aislado de un hombre de 54

años con infección en el tracto urinario y quemaduras en la piel en un hospital de Estambul, Turquía. Desde el primer reporte las cepas productoras de OXA-48 se han distribuido ampliamente en Oriente Medio, África del Norte, Europa, Estados Unidos y Japón (Escandón-Vargas et al., 2017; Poirel et al., 2004).

El primer caso de New Delhi metalo- β -lactamasa (NMD) se describió en 2008 en aislados de *E. coli* y *K. pneumoniae* de un paciente sueco hospitalizado por una infección en el tracto urinario en Nueva Delhi, India. Actualmente NMD se ha extendido alrededor del mundo principalmente en las familias Enterobacteriaceae, Vibrionaceae y en bacterias Gram-negativas no fermentadoras (Escandón-Vargas et al., 2017; Yong et al., 2009).

K. pneumoniae resistentes a carbapenémicos se han convertido en amenaza para la salud pública con altas tasas de mortalidad. La falta de métodos de diagnóstico rápidos y de plataformas de detección precisa para la administración temprana de antibióticos apropiados ha contribuido a esta problemática (Plazak et al., 2018).

Justificación

Las bacterias resistentes a antibióticos son una creciente amenaza para la salud pública que podría llegar a ser una de las principales causas de muerte en el mundo junto con otras patologías como el cáncer, diabetes y enfermedades cardiovasculares. Las Enterobacterias Resistentes a Carbapenémicos, CRE por sus siglas en inglés, son la familia de bacterias resistentes que se encuentran dentro del grupo de prioridad crítica de la Organización Mundial de la Salud debido a su alta tasa de morbilidad y mortalidad a nivel mundial (Suay-García & Pérez-Gracia, 2006). En el ambiente hospitalario, se han identificado diversas carbapenemasas: KPC, IMP, VIM, NDM y OXA-48 cuyas prevalencias varían según la región. NDM posee una prevalencia del 50% en India, OXA-48 es predominante en el Mediterráneo y KPC es endémica

en Grecia, Italia e Israel con incidencia del 30% y en Colombia, Argentina y Brasil es asociada con brotes nosocomiales. En el Ecuador, el Programa Nacional de Vigilancia de la Resistencia a los Antimicrobianos reportó una resistencia al imipenem y al meropenem en aproximadamente 20% de aislados de *K. pneumoniae* en muestras sanguíneas (Soria-Segarra et al., 2020).

Las elevadas tasas de mortalidad se encuentran asociadas a la prescripción y administración inapropiada de antibióticos, al diagnóstico tardío y a las limitadas opciones terapéuticas. El diagnóstico temprano y preciso es necesario para combatir la problemática mundial. Los métodos convencionales de cultivo que implican la siembra en múltiples medios llegando a tomar hasta 36 horas para emitir un diagnóstico. Las herramientas de diagnóstico moleculares, pese a ser altamente precisas, necesitan de un conocimiento *a priori* de los genes responsables de la resistencia. Los modelos de inteligencia artificial basados en algoritmos de *Machine Learning* se muestran como una alternativa de diagnóstico rápida y precisa capaz de predecir fenotipos resistentes a partir de secuencias del genoma bacteriano brindando apoyo a las herramientas de diagnóstico tradicional y mejorando la selección de terapias (Nguyen et al., 2018, 2019).

Objetivos de trabajo de titulación

Objetivo general

- Implementar un modelo para predecir la resistencia a carbapenémicos en *Klebsiella pneumoniae* mediante un algoritmo de *Machine Learning*.

Objetivo específico

- Obtener un set de datos con las secuencias del genoma de *Klebsiella pneumoniae* y el fenotipo de resistencia.

- Realizar el preprocesamiento del set de datos con las secuencias del genoma *Klebsiella pneumoniae*.
- Elaborar un modelo *de Machine Learning* basado en *Extreme Gradient Boosting* (XGBoost).
- Analizar la sensibilidad del modelo para predecir la resistencia a carbapenémicos de *Klebsiella pneumoniae*.

Capítulo II: Revisión Bibliográfica

Generalidades

Klebsiella pneumoniae es un bacillo entérico Gram-negativo de la familia Enterobacteriaceae. Es un organismo aerobio facultativo, fermentador de lactosa, productor de gas, encapsulado, no móvil. Se encuentra ampliamente distribuido en agua, suelo, plantas y animales. En el ser humano forma parte de la microbiota normal de la mucosa oral e intestinal. Provoca patologías graves como bacteriemia, infecciones urinarias, infecciones de tejidos blandos, sepsis y neumonía en grupos vulnerables como neonatos, ancianos y personas inmunodeprimidas. Coloniza material quirúrgico y sanitario provocando una diseminación nosocomial (Wang et al., 2019). Los antibióticos más utilizados para su tratamiento son los β -lactámicos así como los carbapenémicos. Sin embargo, la hidrólisis del anillo lactámico por la enzima β -lactamasa confiere resistencia a los mismos (Opoku-Temeng et al., 2019).

Taxonomía

De acuerdo con el sitio web Interagency Taxonomic Information System (ITIS) la clasificación taxonómica para *K. pneumoniae* corresponde a la presentada en la Tabla 1.

Tabla 1

Clasificación taxonómica de K. pneumoniae.

Reino	Bacteria
Subreino	Negibacteria
Phylo	Proteobacteria
Clase	Gammaproteobacteria
Orden	Enterobacteriales

Familia	Enterobacteriaceae
Genero	<i>Klebsiella</i>
Especie	<i>Klebsiella pneumoniae</i>
Subespecie	<i>Klebsiella pneumoniae pneumoniae</i>

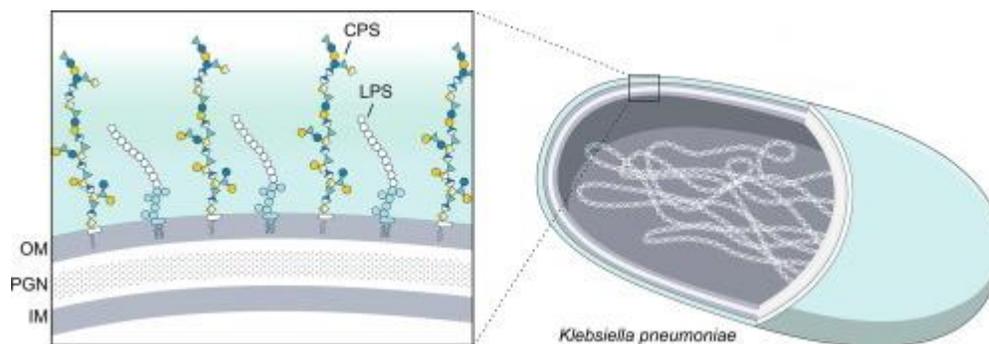
Nota. Adaptado de ITIS (2019).

Morfología

K. pneumoniae es un bacilo Gram-negativa encapsulado, con un tamaño de alrededor de 0,6-0,8 μm de longitud y 0,3-1,0 μm de diámetro. Son inmóviles por la falta de flagelo. Su citoplasma se encuentra confinado por la membrana interna, peptidoglicano, espacio periplasmático y membrana externa en la cual se hallan anclados lipopolisacáridos, polisacáridos de cápsula y varios tipos de pili o fimbrias (Figura 2). En medio selectivo MacConkey se observan colonias grandes de color rosado y consistencia mucoide producto de los polisacáridos de membrana (Murray et al., 2018).

Figura 2

Estructura y morfología de K. pneumoniae.



Nota. Membrana interna (IM), Peptidoglicano (PGN), Membrana externa (OM),

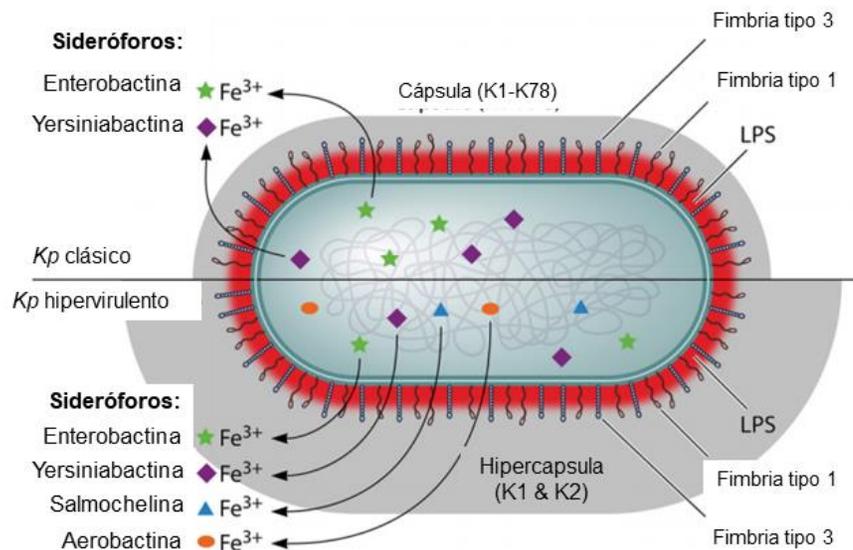
Lipopolisacáridos (LPS), Polisacáridos de capsula (CPS). Adaptado de Opoku-Temeng et al. (2019)

Epidemiología

K. pneumoniae es un organismo oportunista de distribución ubicua. En el humano se encuentra, sin ocasionar patologías, en las mucosas nasofaríngeas e intestinal con un porcentaje de detección del 1-6% y 5-38% respectivamente; no se presenta en la piel. Dentro del ambiente intrahospitalario se localiza en batas, guantes, equipos y material quirúrgico. Su diseminación se asocia principalmente al uso de catéter e intervenciones quirúrgicas (Reyes et al., 2019). Es uno de los 8 patógenos hospitalarios más relevantes, siendo responsable de infecciones nosocomiales entre el 3-8% de los casos. Es el segundo agente causal de infecciones provocadas por Gram-negativos tras *Escherichia coli*. Las diferentes variantes antigénicas en la cápsula confieren diversas características de virulencia y promueven la interacción con el huésped, aumentando la probabilidad de supervivencia (Vargas & Toro, 2010).

Factores de virulencia

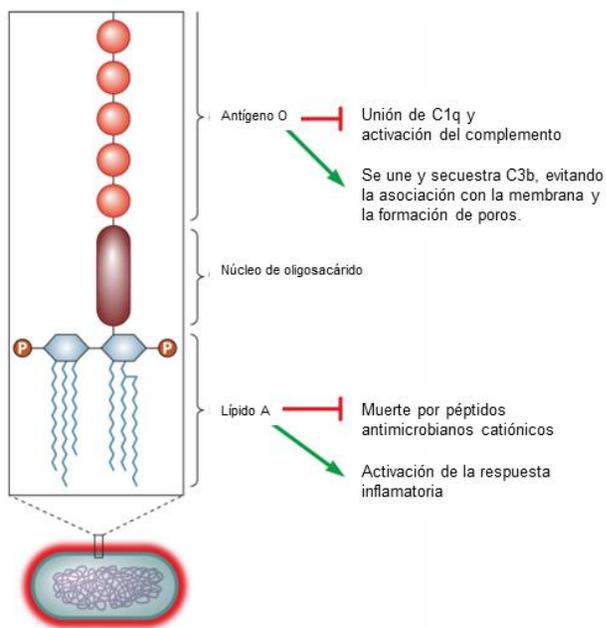
K. pneumoniae posee factores de virulencia para evadir el sistema inmunitario del huésped. Los componentes que favorecen la virulencia son: lipopolisacáridos (LPS), polisacáridos de cápsula (CPS), fimbrias, proteínas de membrana externa, sideróforos y fuentes de nitrógeno (Figura 3) (Opoku-Temeng et al., 2019).

Figura 3*Factores de virulencia*

Nota. Modificado de Paczosa & Meccas (2016).

Lipopolisacáridos

Los lipopolisacáridos (LPS) están formados por tres componentes: un lípido A anclado a la membrana, un núcleo de oligosacárido y el antígeno O (Figura 4). El lípido A se une al receptor TLR4 activando el sistema inmune y puede sufrir modificaciones estructurales que dificultan la respuesta inmunitaria (Opoku-Temeng et al., 2019). El núcleo de oligosacárido puede ser de dos clases: tipo I y tipo II, que se diferencian estructuralmente por los sustituyentes GlcN (Li et al., 2014). El antígeno O está formado por polímeros de oligosacáridos. Se han identificado nueve tipos que varían en la composición de monómeros de azúcar. LPS son capaces de evitar el sistema de complemento por la longitud del antígeno. Los fenotipos lisos o de longitud completa son más eficientes en evadir el sistema de complemento que los fenotipos truncados (Bengoechea & Sa Pessoa, 2019).

Figura 4*Estructura del lipopolisacárido*

Nota. Modificado de Paczosa & Meccas (2016).

Polisacáridos de cápsula

Los polisacáridos de cápsula (CPS) están compuesto por monómeros de tres a seis azúcares que forman una capa gruesa que protege de la opsonización y la fagocitosis. El grosor del polisacárido bloquea la unión e internalización por macrófagos, neutrófilos, células dendríticas y epiteliales (Bengoechea & Sa Pessoa, 2019). Se han identificado 79 serotipos o antígenos K, 77 tipos reconocidos por pruebas serológicas y 2 caracterizados por genotipado molecular y tipificación de fagos (Pan et al., 2015).

Fimbrias

Las fimbrias son apéndices filamentosos que se extienden en la superficie bacteriana para facilitar la adhesión a las células del hospedero y sobre superficies inertes. Se han descrito

cuatro tipos para *Klebsiella pneumoniae*: fimbria tipo 1, fimbria tipo 3, fimbria Kpc y adhesina KPF-28 (Li et al., 2014). Las fimbrias tipo 1 son estructuras delgadas que se adhieren a los residuos de manosa presentes en las células del huésped y son esenciales para la colonización del tracto urinario (Struve et al., 2008). Las fimbrias tipo 3 son apéndices de 2–4 nm de ancho y 0.5–2 µm de largo. No están asociadas a la virulencia, pero contribuyen, junto con las fimbrias Kpc, a la formación de biopelículas en dispositivos médicos. La adhesina KPF-28 favorece la colonización de *K. pneumoniae* en el intestino de los mamíferos (Li et al., 2014).

Proteínas de membrana externa

Las principales proteínas de membrana son OmpA, porinas y bombas de eflujo. Las proteínas OmpA disminuyen la respuesta inflamatoria y aumentan la resistencia a la fagocitosis. Las porinas pueden ser de cuatro tipos: OmpK35, OmpK36, KpnO y OmpK26. La ausencia de OmpK35, KpnO y OmpK26 favorece la resistencia a cefalosporinas y carbapenémicos, mientras que la pérdida de OmpK36 previene la fagocitosis. La bomba de eflujo AcrAB expulsa antibióticos y moléculas antimicrobianas propias del huésped, aumentando la resistencia a agentes antimicrobianos (Bengoechea & Sa Pessoa, 2019; Li et al., 2014; Tsai et al., 2011).

Sideróforos

Los sideróforos son moléculas quelantes que captan el hierro necesario para el crecimiento. La aerobactina secretada posee mayor afinidad al hierro que las proteínas del huésped. Enterobactina, yersiniabactina y salmochelina permiten evadir el sistema inmune del hospedero. Los sideróforos son un mecanismo de supervivencia en entornos escasos de hierro y aumentan la patogenicidad (Effah et al., 2020; Li et al., 2014; Russo et al., 2011).

Fuentes de nitrógeno

K. pneumoniae obtiene nitrógeno para su crecimiento de dos formas, sintetiza ureasa para hidrolizar la urea y compete por la alantoína presente dentro del huésped (Li et al., 2014).

Resistencia los antimicrobianos

K. pneumoniae posee un rango amplio de resistencia a antibióticos debido a elementos genéticos móviles como los plásmidos. El mecanismo más importante es la síntesis de β -lactamasas capaces de hidrolizar el anillo de los β -lactámicos manteniendo intacto el peptidoglicano de la pared celular. En la actualidad se conoce que *K. pneumoniae* es capaz de sintetizar dos enzimas: β -lactamasas de espectro extendido (BLEE) y carbapenemasas confiriendo resistencia a la mayoría de los antibióticos β -lactámicos y volviéndose un peligro para la salud pública (Opoku-Temeng et al., 2019; Reyes et al., 2019). Otros mecanismos alternativos como la expresión de bombas de eflujo y cambios en la síntesis de porinas no actúan directamente sobre la molécula, pero contribuyen a impedir la entrada del antibiótico a la célula (Suay-García & Pérez-Gracia, 2006).

Resistencia a los carbapenémicos

Los carbapenémicos son antibióticos de la familia de los β -lactámicos que se usan con frecuencia para tratar infecciones provocados por Enterobacteriaceae. Con la aparición de cepas de *K. pneumoniae* BLEE, los carbapenémicos, como el meropenem e imipenem, se volvieron los antibióticos de primera línea para el tratamiento. Sin embargo, la aparición de cepas de *K. pneumoniae* productoras de carbapenemasas representan una amenaza para la salud pública y un desafío terapéutico. De acuerdo con la clasificación molecular de Ambler, las

carbapenemasas se dividen en tres grupos: serina- β -lactamasas de clase A, metalo- β lactamasas de clase B y serina- β -lactamasas de clase D (Escandón-Vargas et al., 2017; Reyes et al., 2019; Russo & Marr, 2019).

Clase A

Las serina- β -lactamasas de clase A poseen un residuo de serina en su sitio activo y son capaces de hidrolizar gran variedad de β -lactámicos. Pueden estar presentes en el cromosoma bacteriano o en elementos genéticos móviles (Escandón-Vargas et al., 2017; Nordmann et al., 2011).

Klebsiella pneumoniae carbapenemasa (KPC)

Las enzimas KPC hidrolizan penicilinas, cefalosporinas, monobactamas y carbapenémicos. Sin embargo, se inhiben parcialmente por el ácido clavulánico y el tazobactam. En su mayoría son codificadas por plásmidos que propagan el gen blaKPC. Se han identificado 23 variantes que divergen en la sustitución de aminoácidos (Escandón-Vargas et al., 2017).

Clase B

Las metalo- β lactamasas de clase B requieren iones de zinc para su actividad enzimática y por lo tanto son inhibidos por agentes quelantes como el EDTA. Se dividen en tres subclases: B1, B2 y B3. La subclase B1 necesita dos iones de zinc, la subclase 2 se une a un ion de zinc y presenta inhibición a un segundo ion metálico, y la subclase B3 se liga a dos iones, pero hidroliza las cefalosporinas. Dentro de la subclase 1, se encuentran enzimas de gran relevancia clínica como: imipenemasa (IMP), metalo- β -lactamasa codificada por Verona (VIM) y metalo- β -lactamasa de Nueva Delhi (NDM) (Escandón-Vargas et al., 2017; Nordmann et al., 2011; Russo & Marr, 2019).

Verona metalo- β -lactamasa codificada por integrones (VIM)

Hidrolizan penicilinas, cefalosporinas y carbapenémicos, con excepción de aztreonam. Los genes blaVIM se encuentran en los integrones de clase 1 y se han identificado en plásmidos con diferentes tipos de replicación. Se han identificado 51 variantes y en *K. pneumoniae* predomina VIM-1 (Escandón-Vargas et al., 2017; Nordmann et al., 2011; Samuelsen et al., 2011; Tamma et al., 2016).

New Delhi metalo- β -lactamasa (NMD)

Es capaz de hidrolizar gran cantidad de β -lactámicos, excepto las monobactamas. Los genes blaNMD se encuentran en plásmidos los cuales pueden portar diferentes genes de resistencia. Se han descrito 16 variantes (Escandón-Vargas et al., 2017; Nordmann et al., 2011; Wu et al., 2019).

Imipenemasa (IMP)

Al igual que VIM, IMP es capaz de hidrolizar penicilinas, cefalosporinas y carbapenémicos. Es codificada por el gen blaIMP que se encuentra como cassette de genes dentro de integrones de clase 1. Se han identificado 58 variantes (Escandón-Vargas et al., 2017; Nordmann et al., 2011).

Clase D

Al igual que la clase A, las β -lactamasas de clase D u oxacilinasas necesitan un residuo de serina para formar su sitio activo. Es el grupo más grande con 500 variantes debido a su elevada capacidad de mutación y posee amplios perfiles de hidrolización, no es inhibida por EDTA o inhibidores de β -lactamasa. Solo el subgrupo de β -lactamasas de clase D hidrolizantes de

carbapenem (CHDL) es capaz de actuar sobre los carbapenémicos (Escandón-Vargas et al., 2017; Nordmann et al., 2011; Suay-García & Pérez-Gracia, 2006).

Subgrupo OXA-48

Las oxacilinasas 48 son enzimas de amplio espectro capaces de hidrolizar penicilinas, carbapenémicos y cefalosporinas. El gen blaOXA es transmitido por un plásmido de 62,5 kb. Se han identificado seis variantes que difieren en la secuencia de aminoácidos. Los aislados productores de OXA-48 son difíciles de identificar, por lo que representan una amenaza para la salud pública (Escandón-Vargas et al., 2017; Nordmann et al., 2011; Suay-García & Pérez-Gracia, 2006).

Estudio de asociación de genoma completo (GWAS)

Los estudios de asociación de genoma completo (GWAS por sus siglas en inglés) son un método para asociar estadísticamente variantes genéticas con un fenotipo. Estos estudios en la última década han despuntado en bacteriología. Se estudian los fenotipos relevantes en salud pública como: virulencia, resistencia a los antimicrobianos, concentración mínima inhibitoria, susceptibilidad del huésped a reinfección, entre otros. Dentro de las poblaciones bacterianas, las principales formas de variación incluyen polimorfismos de nucleótido simple (SNP), presencia/ausencia de genes y *k-mers* de longitud fija o variable (Falush, 2016; Read & Massey, 2014; San et al., 2020).

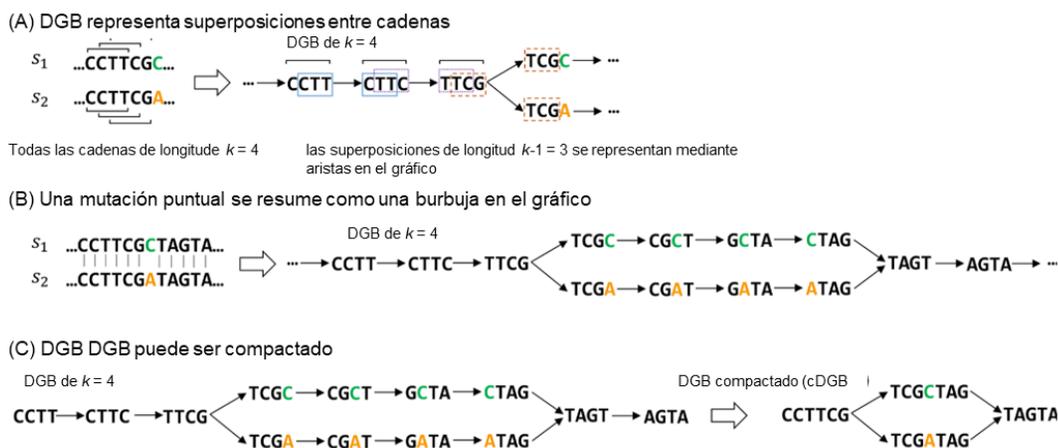
Los primeros GWAS realizados en humanos en 2005 usaron SNPs para identificar las variaciones génicas en función a un genoma de referencia. Sin embargo, este enfoque no es apropiado en bacterias debido a la dificultad para la elección de un genoma de referencia por la alta plasticidad genómica y su amplio genoma accesorio. Por otro lado, los métodos basados en genes no abarcan regiones no codificantes incluidos los determinantes genéticos encargados de

la regulación transcripcional y traduccional (Jaillard et al., 2018, 2017; San et al., 2020). Como alternativa a estas metodologías, se propone utilizar subsecuencias de ADN de longitud k denominados k -mers para describir eventos genéticos como: mutaciones, diferencia en el contenido de genes, indels y recombinaciones, sin requerir de un genoma de referencia otorgando flexibilidad al momento de estudiar un genoma completo. Los k -mers largos otorgan mayor especificidad y los cortos mayor sensibilidad. No obstante, este enfoque pierde interpretabilidad debido a que: (1) existe un amplio rango de redundancia entre k -mers y (2) la cantidad de subsecuencias aumenta en k por lo que se generan millones de datos (Jaillard et al., 2018, 2017; J. A. Lees et al., 2016).

Los gráficos De Bruijn compactados (cDBG) cierran el margen de diferencia entre los enfoques basados en SNPs, genes y k -mers. Elimina la redundancia al colocar todos los k -mer que pertenecen a una misma secuencia en una sola cadena de nucleótidos más larga, posteriormente se produce una bifurcación con los k -mers variables formando una burbuja y finalmente se asocia nuevamente a una secuencia larga. Cada nodo presente en la burbuja se denomina *unitig*. El análisis de *unitigs* permite realizar un estudio sin pérdida de datos y se adapta a la variabilidad de los genomas (Jaillard et al., 2018, 2017).

Figura 5

Construcción del gráfico De Bruijn compactado



Nota. El gráfico muestra la construcción del gráfico De Bruijn compactado a partir de un conjunto de secuencias S_1 y S_2 que difieren en un nucleótido. (A) Ambas secuencias se dividen en k -mers de longitud $k=4$ y se observa que dos nodos continuos poseen nucleótidos idénticos. Al presentarse el SNP, se forma una bifurcación. (B) La bifurcación forma una burbuja y cada rama representa un alelo. (C) Las secuencias semejantes se compactan en los extremos de la burbuja. En la figura se pueden observar 4 nodos (*unitigs*) que representan la misma variación. Modificado de Jaillard et al. (2018).

La fuerte estructura poblacional bacteriana es un factor de confusión dentro de los GWAS y provoca un aumento en la tasa de falsos positivos (error tipo I). El control de la estructura poblacional es un paso crítico para evitar asociaciones falsas producto del desequilibrio de ligamiento. Los modelos lineales mixtos (LMM) han demostrado ser capaces de corregir los factores de confusión por medio de una matriz K de parentesco que establece similitudes genotípicas por pares de individuos reduciendo el error tipo I (Kaler & Purcell, 2019; J. Lees, 2017; J. A. Lees et al., 2019; Lippert et al., 2011; Listgarten et al., 2012; Price et al., 2010)

Machine learning (ML)

Machine Learning (ML) es una técnica de la Inteligencia Artificial que mediante la implementación de algoritmos, probabilidad y estadística, permite que un sistema computacional identifique patrones en datos masivos y aprenda por sí mismo sobre un conjunto de datos creando un modelo capaz de predecir. Los algoritmos de *Machine Learning* son entrenados por un conjunto de valores de entrada que permiten recopilar observaciones de un sin número de variables encontrando combinaciones que predicen un resultado de forma confiable. El algoritmo entrenado es evaluado con una nueva base de datos que no pertenecen al conjunto inicial de entrenamiento para conocer la precisión de los resultados de salida (Baştanlar & Özuysal, 2013; Macesic et al., 2017; Obermeyer & Emanuel, 2016; Sandoval Serrano, 2018).

ML se subdivide en dos clases: el aprendizaje supervisado y el aprendizaje no supervisado. El primero requiere una base de datos de entrada que incluya la variable de interés (variable de salida). El algoritmo entrena al modelo para que sea capaz de predecir la variable de salida en función a los datos de entrada a partir de las características observadas durante el entrenamiento. Por el contrario, el aprendizaje no supervisado no requiere de un resultado previamente observado para ser entrenado. Los algoritmos hacen uso de bases de datos con diferentes variables y encuentran combinaciones o agrupaciones dentro de los datos que permitan obtener un resultado (Baştanlar & Özuysal, 2013; Deo, 2015; Macesic et al., 2017).

Las fases de desarrollo a seguir en ML son: Entrenamiento y Prueba. En el entrenamiento se separa un segmento de datos para entrenar el algoritmo y se entrega toda la información para que encuentre los patrones necesarios y posteriormente se realicen las predicciones. La otra parte del segmento de los datos se usa para realizar las pruebas. Se

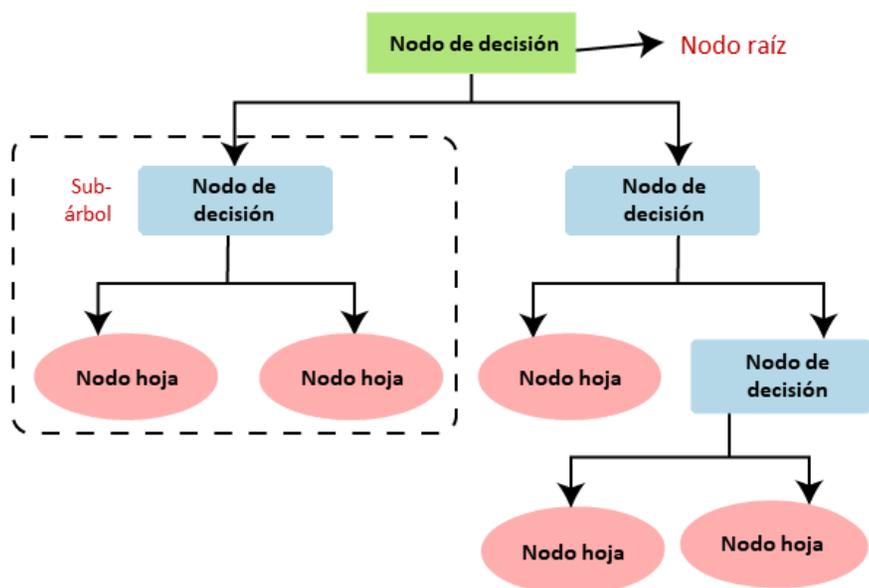
realizan las preguntas al algoritmo y se evalúan las respuestas para verificar el aprendizaje, si los resultados están entre el 80% y 90% de respuestas correctas, se recomienda utilizar el algoritmo (Sandoval Serrano, 2018).

Las técnicas de Inteligencia Artificial y los modelos de *Machine Learning* han demostrado su elevado rendimiento para el control de resistencia a los antimicrobianos gracias a la elevada recopilación de datos clínicos de la última década. Los modelos más utilizados en este campo incluyen: Naive Bayes (NB), árboles de decisión (DT), bosques aleatorios (RF), máquinas de soporte vectorial (SVM) y redes neuronales artificiales (ANN). Los modelos basados en árboles de decisión se entrenan en menor tiempo, mayor velocidad y facilidad para interpretar en comparación con los otros modelos (Lv et al., 2020).

Los árboles de decisión (DT) son modelos utilizados para resolver problemas de clasificación o de regresión. Gráficamente mantienen la estructura de un árbol (figura 6). El nodo raíz se forma seleccionando una característica, las ramas dividen la raíz en función a una regla establecida y los nodos hoja representan el resultado. Este proceso se realiza de forma recursiva hasta obtener el mejor árbol (figura 7). Entrenar los modelos basados en árboles de decisión con el algoritmo XGBoost permite aumentar la velocidad de entrenamiento y mejorar el rendimiento del modelo (Chen & Guestrin, 2016).

Figura 6

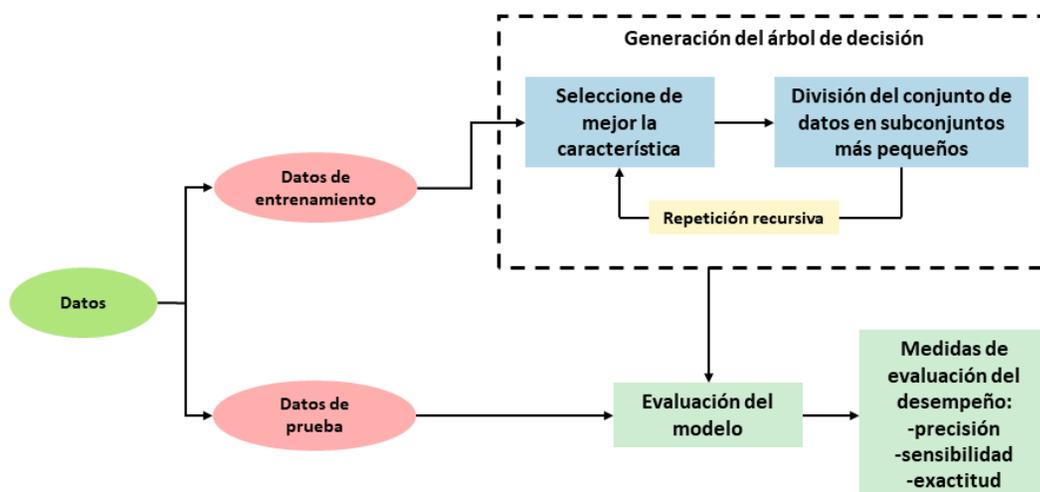
Estructura de un árbol de decisión



Nota. Modificado de *Árbol de decisión en Machine Learning (Parte 1)* - sitiobigdata.com (n.d.).

Figura 7

Generación de in árbol de decisión



Nota. Modificado de *Árbol de decisión en Machine Learning (Parte 1)* - sitiobigdata.com (n.d.).

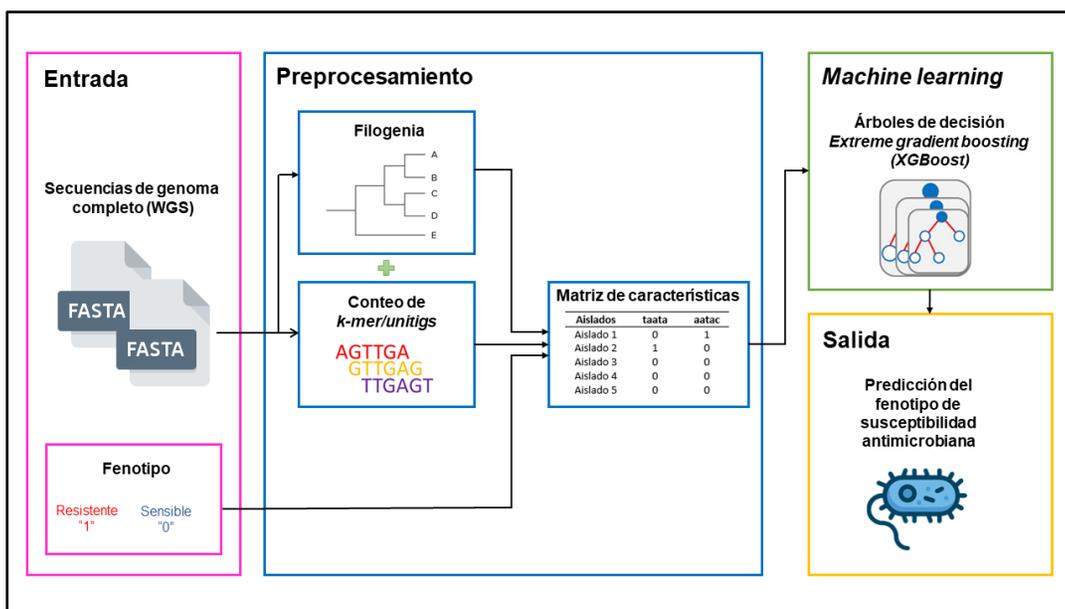
Capítulo III: Metodología

Esquema de desarrollo

El proyecto de investigación consiste en un modelo capaz de predecir la resistencia a dos carbapenémicos en *Klebsiella pneumoniae* mediante un algoritmo de *Machine Learning*. En la figura 8 se muestra el esquema de desarrollo del proyecto. En el bloque de entrada se toman secuencias de genoma completo (*contigs*) de *K. pneumoniae* y su fenotipo de susceptibilidad antimicrobiana. Con estos datos, se realiza un conteo de *unitigs* y se construye la filogenia con el fin de aplicar un GWAS para filtrar los *unitigs* asociadas al fenotipo y formar la matriz de características. El modelo de *Machine Learning* supervisado basado en árboles de decisión toma la matriz como entrada para aprender las características responsables del fenotipo y finalmente en la salida se evalúa la sensibilidad del modelo con un set de datos de evaluación.

Figura 8

Esquema de desarrollo



Hardware

Todos los procedimientos fueron realizados en una computadora personal marca HP modelo Probook 450 con memoria RAM de 16 GB y procesador Intel[®] Core™ i5 CPU 2.50 GHz con 4 núcleos.

Software

El presente proyecto de investigación se realizó utilizando el sistema operativo Unix distribución Ubuntu 16.04 LTS. Los programas cuyos binarios fueron ejecutados desde la terminal de Linux fueron: Prokka v1.14.6 (Seemann, 2014), Roary v3.13.0 (Page et al., 2015) y RAxML v8.2.12 (Stamatakis, 2014) para la anotación de secuencias, obtención de pangenoma y construcción de la filogenia respectivamente. De igual forma, se utilizó Python v3.7 para ejecutar unitig-counter (Jaillard et al., 2018) para el conteo de *unitigs*, y Pyseer (J. A. Lees et al., 2018) para el filtrado de estos. Finalmente, el modelo de *Machine Learning* se entrenó y evaluó con el paquete Scikit-learn en Python (Pedregosa et al., 2011). Los análisis estadísticos se realizaron en R Studio v1.2.5033 (RStudio Team, 2020) utilizando el paquete Tidyverse (Wickham et al., 2019)

Set de datos

Se recolectó un total de 416 secuencias de genoma completo de *K. pneumoniae* del Centro Nacional de Información Biotecnológica (NCBI) GenBank (Clark et al., 2016) a través del sitio web Pathosystems Resource Integration Center (PATRIC) (Antonopoulos et al., 2018) con los parámetros de búsqueda indicados en la Tabla 2. Las secuencias se descargaron en formato “.fna” y cada una posee una longitud aproximada de 5 millones de nucleótidos equivalente a 5 megabits (MB). Cada secuencia se etiquetó con un código único otorgado por PATRIC.

Se recopilaron los fenotipos de susceptibilidad para imipenem y meropenem, clasificados por PATRIC de acuerdo con los criterios del CLSI para MIC, en documento “.csv”. Los fenotipos se clasificaron en una matriz binaria donde los aislados resistentes se etiquetaron con “1” y los sensibles “0”. No se consideraron los fenotipos intermedios.

Tabla 2

Parámetros de búsqueda de genoma de K. pneumoniae en PATRIC

Parámetro	Selección
Organism	<i>Klebsiella pneumoniae</i>
Genome status	WGS
Antimicrobial Resistance	Susceptible, resistance
Host Name	Homo sapiens
Genome Quality	Good

Nota. En la tabla se muestran los parámetros de búsqueda utilizados en PATRIC para secuencias de genoma completo de *K. pneumoniae* cuyo fenotipo de resistencia haya sido reportado.

Preprocesamiento de las secuencias de genoma completo

Los algoritmos de *Machine Learning* utilizan una tabla de características para realizar tareas de regresión o clasificación (Baştanlar & Özuysal, 2013). Al tratarse de un problema de clasificaciones de fenotipos resistentes y sensibles, se utilizó *unitigs* como característica de análisis para construir la matriz de aprendizaje de dimensiones $m \times n$, donde cada fila m pertenece a un aislado, las columnas n a un único *unitig* y cuyos elementos corresponden al número de veces que ocurre cada *unitig* en un determinado aislado (Tabla 3). Se utilizó el

enfoque de los GWAS para identificar los *unitigs* asociados estadísticamente a los fenotipos de susceptibilidad antimicrobiana.

Tabla 3

Ejemplo de la matriz presencia/ausencia de unitigs en el aprendizaje del modelo

Aislados	taata	aatac	acctg	cctgc
Aislado 1	0	1	1	0
Aislado 2	1	0	1	0
Aislado 3	0	0	0	0
Aislado 4	0	0	0	1
Aislado 5	0	0	1	1

Nota. Las secuencias de *unitigs* se encuentran en las columnas y las etiquetas de los aislados en las filas. El conteo de aparición de cada *unitig* conforma los elementos de la matriz.

Conteo de unitigs

Se utilizó el paquete independiente unitig-counter (Jaillard et al., 2018) que utiliza la biblioteca Genome Assembly & Analysis Tool Box (GATB) (Drezen et al., 2014) para crear una matriz presencia/ausencia de *unitigs* en todos los genomas. Se empleó todas las secuencias en formato “.fna” y se definió una longitud de *k-mer* de 31 nucleótidos para el análisis.

Filtrado de unitig asociados al fenotipo

Para filtrar los *unitigs* asociados al fenotipo se realizó un GWAS por medio del paquete de Python, Pyseer que utiliza una matriz de parentesco entre pares de aislados para evaluar

estadísticamente las asociaciones entre una variante genética y el fenotipo, además estima el umbral de significancia a partir de la corrección de Bonferroni (J. A. Lees et al., 2018).

La matriz de parentesco se obtuvo a partir del script provisto por Pyseer que toma como entrada un árbol filogenético de genoma central. Para la construcción de la filogenia, las secuencias de genoma completo se anotaron en Prokka v1.14.6 (Seemann, 2014) y los archivos GFF3 se usaron para crear un pangenoma en Roary v3.13.0 (Page et al., 2015) con un porcentaje mínimo de identidad BLAST del 90% y alineación de los genes centrales con MAFFT (Katoh & Standley, 2013). El árbol filogenético se realizó en RAxML v8.2.12 (Stamatakis, 2014) aplicando un modelo GTR + gamma y 1000 bootstraps.

La estimación del valor-p se realizó con Pyseer utilizando un modelo lineal mixto y se determinó el umbral significativo para el filtrado de *unitigs* asociados con el fenotipo de resistencia.

Entrenamiento y evaluación del modelo

Se generó un modelo de predicción basado en árboles de decisión usando el algoritmo XGBoost (Chen & Guestrin, 2016) por medio de la biblioteca Scikit-learn en Python (Pedregosa et al., 2011). En la Tabla 4 se muestran los parámetros utilizados para el entrenamiento del modelo. Para determinar los valores óptimos de profundidad máxima del árbol y validación cruzada, se planteó un diseño experimental 2^2 estableciendo cuatro experimentos (Tabla 5).

Tabla 4*Parámetros de entrenamiento*

Parámetro	Valor
Alfa	1e-4
Submuestreo por columna	0.6
Submuestreo por fila	0.6
Tasa de aprendizaje	0.01
Objetivo	<i>binary:hinge</i>

Tabla 5*Diseño experimental*

		Validación cruzada	
		5	10
Profundidad máxima	3	Experimento 1	Experimento 2
	4	Experimento 3	Experimento 4

Nota. Se estableció un diseño experimental 2² para determinar la combinación óptima entre dos parámetros de entrenamiento: profundidad máxima (MD) y validación cruzada (CV). El experimento 1 probó CV de 3 y MD de 5, el experimento 2 CV de 3 y MD de 10, experimento 3 CV 4 y MD de 5, finalmente el experimento 4 CV de 4 y MD de 10.

El set de datos se dividió en tres conjuntos no superpuestos: 80% para entrenamiento, 20% pruebas y validación. La evaluación del rendimiento del modelo se realizó con la medición

de los siguientes parámetros: tasa de verdaderos positivos (VP), tasa de falsos positivos (FP), tasa de verdaderos negativos (VN), tasa de falsos negativos (FN), curva de característica del funcionamiento del receptor (curva ROC) y el área bajo la curva ROC (AUC). Adicionalmente se estimó:

Exactitud (A)

$$A (\%) = \frac{VP + VN}{VP + VN + FP + FN} \cdot 100$$

Precisión (P)

$$P (\%) = \frac{VP}{VP + FP} \cdot 100$$

Sensibilidad (R)

$$R (\%) = \frac{VP}{VP + FN} \cdot 100$$

Valor F (F1)

$$F1 (\%) = 2 \cdot \frac{P \cdot R}{P + R} \cdot 100$$

Capítulo IV: Resultados

Set de datos

El set de datos se formó con un total de 416 aislados de *K. pneumoniae* y sus fenotipos de resistencia a dos antibióticos carbapenémicos. Para imipenem, se reportaron 237 aislados resistentes y 179 sensibles. Mientras que, para meropenem, se obtuvieron 227 aislados resistentes y 189 sensibles (Tabla 6). Se observa una proporción mayor de aislados resistentes para ambos antibióticos (Figura 9).

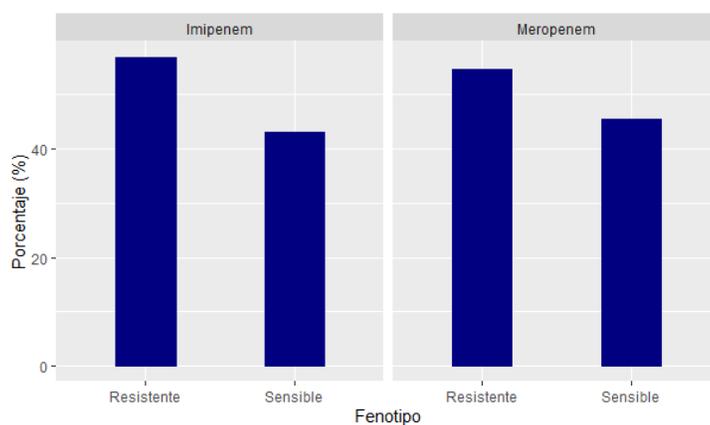
Tabla 6

Resumen de fenotipos de susceptibilidad del set de datos

Fenotipo	Imipenem	Meropenem
Resistente	237 (56.97%)	227 (54.57%)
Sensible	179 (43.03%)	189 (45.43%)
Total	416 (100%)	416 (100%)

Figura 9

Porcentaje de aislados de K. pneumoniae susceptibles reportados en el set de datos



Preprocesamiento de las secuencias de genoma completo

El análisis de las 416 secuencias de genoma completo de *Klebsiella pneumoniae* en unitigs-counter dio como resultado un total de 48 551 563 *k-mers* cuyas secuencias redundantes se compactaron en 1 600 166 *unitigs*.

El GWAS evaluó la asociación entre los *unitigs* y el fenotipo de susceptibilidad a cada uno de los antibióticos utilizando un LMM para identificar los valores-p inflados y evitar un aumento en el error tipo I. Los resultados se presentaron en un gráfico cuantil-cuantil (Q-Q) entre los $-\log_{10}$ (valor-p) observados y esperados. En las Figuras 10 y 11 se muestra el gráfico Q-Q para la asociación de *unitigs* a imipenem y meropenem respectivamente. En ambos casos, los *unitigs* fuertemente asociados al fenotipo se encuentran desviados hacia la esquina superior derecha.

Figura 10

Gráfico Q-Q del análisis de GWAS entre el los *unitigs* de *K. pneumoniae* y el fenotipo de susceptibilidad a imipenem

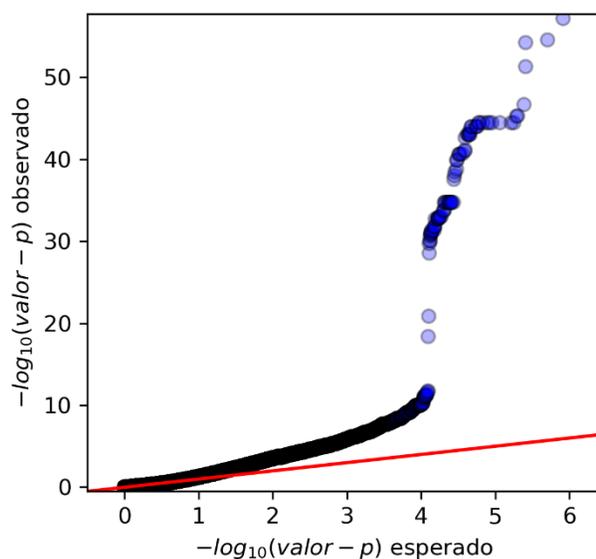
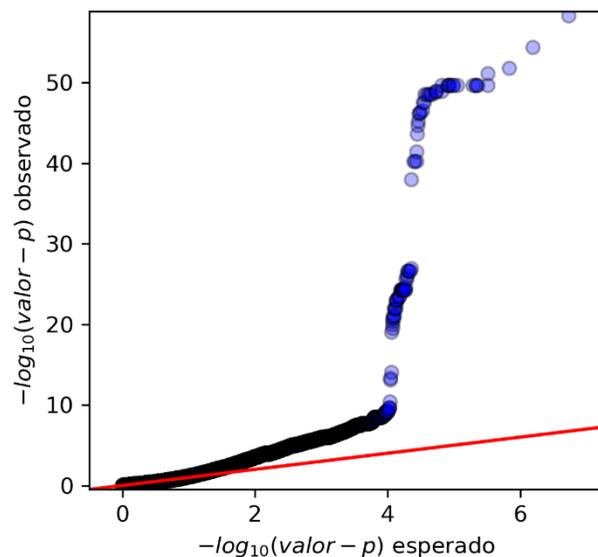


Figura 11

Gráfico Q-Q del análisis de GWAS entre los *unitigs* de *K. pneumoniae* y el fenotipo de susceptibilidad a meropenem



La corrección de Bonferroni realizada en Pyseer estimó un umbral de significancia de $6,78 \times 10^{-08}$ para imipenem y meropenem. Se filtraron los *unitigs* que superaban el umbral dando como resultado 417 *unitigs* fuertemente asociados a imipenem y 312 a meropenem. En la Tabla 7 se muestra un resumen del contenido de la matriz de presencia/ausencia para el aprendizaje del modelo.

Tabla 7

Resumen del contenido de la matriz presencia/ausencia de *unitigs*

	Imipenem	Meropenem
Características (dimensiones)	417	312
Aislados	416	416

Análisis de resultados de aprendizaje del modelo de predicción de susceptibilidad a imipenem

Análisis del experimento 1

En el experimento 1 se evaluó el rendimiento del modelo aplicado una profundidad máxima de aprendizaje de 3 y una validación cruzada de 5. Se obtuvo una exactitud de 78.57% y se entrenó en un periodo de 2 minutos y 17 segundos. El desempeño general del experimento 1 se presentan en la Tabla 8 y en la Figura 12.

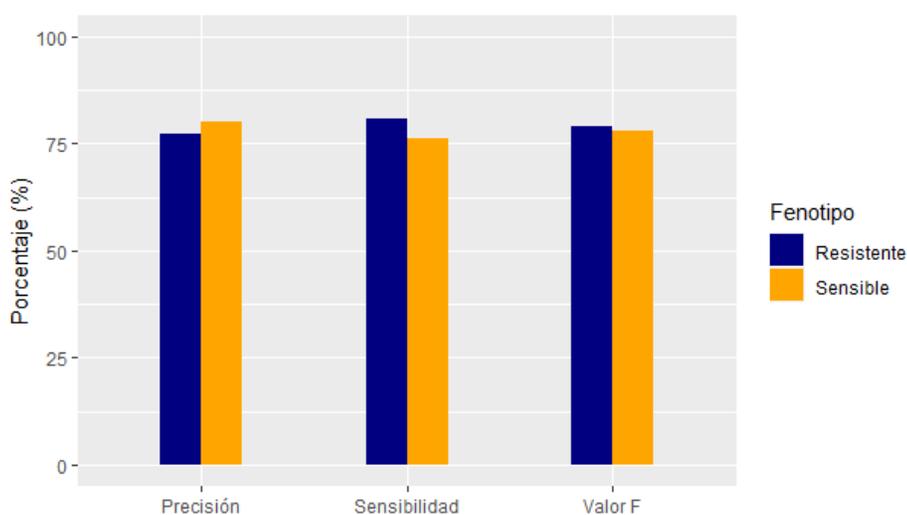
Tabla 8

Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 1

Fenotipo	Precisión (%)	Sensibilidad (%)	Valor F (%)
Resistente	77.27	80.95	79.07
Sensible	80.00	76.19	78.15

Figura 12

Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 1



Análisis del experimento 2

En el experimento 2 se midió el rendimiento del modelo manteniendo la profundidad máxima de aprendizaje de 3 y aumentando la validación cruzada a 10. La exactitud del modelo empeoró en relación con el experimento 1 obteniendo una exactitud de 76.19% en un periodo de entrenamiento de 4 minutos y 58 segundos. El desempeño del experimento 2 se presenta en la Tabla 9 y Figura 13.

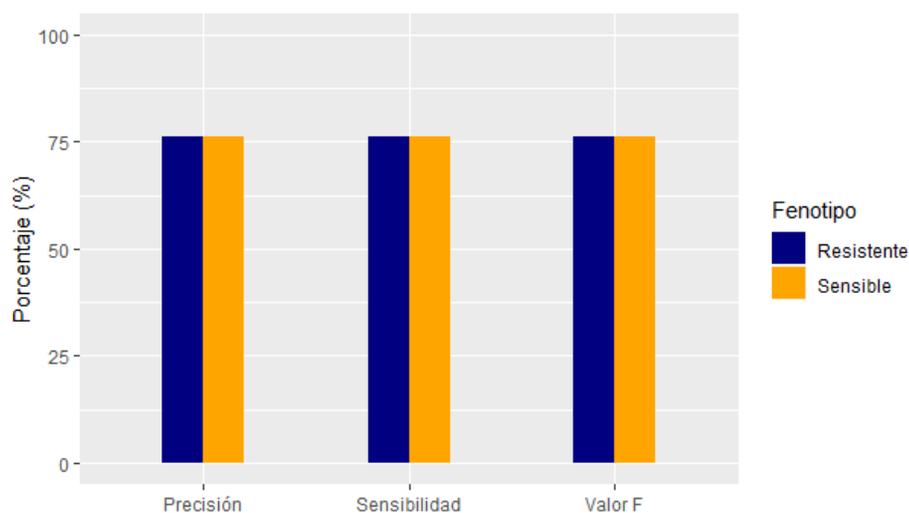
Tabla 9

Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 2

Fenotipo	Precisión (%)	Sensibilidad (%)	Valor F (%)
Resistente	76.19	76.19	76.19
Sensible	76.19	76.19	76.19

Figura 13

Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 2



Análisis del experimento 3

En el experimento 3 se evaluó el rendimiento del modelo a una profundidad máxima de aprendizaje de 4 y una validación cruzada a 5. La exactitud del modelo mejoró en relación con los dos experimentos anteriores. La exactitud fue de 80.95% en un periodo de 2 minutos y 17 segundos. El desempeño del modelo en el experimento 3 se muestra en la Tabla 10 y Figura 14.

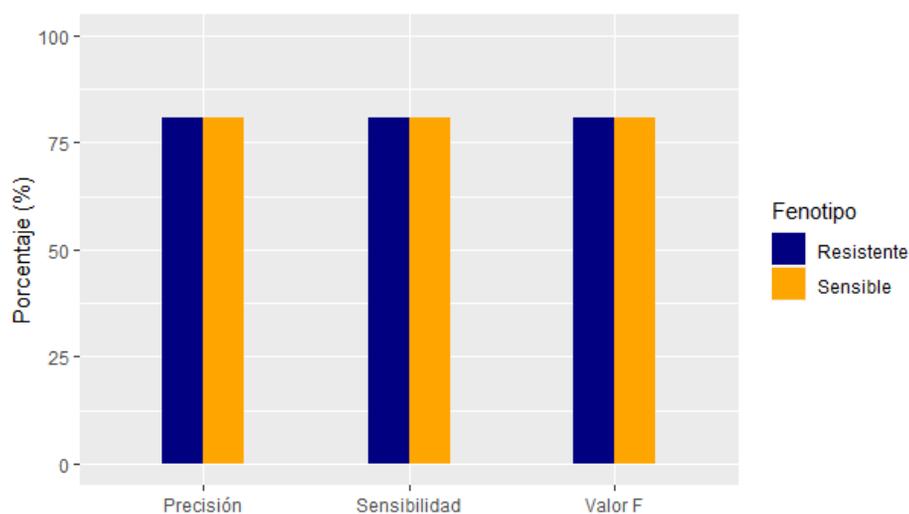
Tabla 10

Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 3

Fenotipo	Precisión (%)	Sensibilidad (%)	Valor F (%)
Resistente	80.95	80.95	80.95
Sensible	80.95	80.95	80.95

Figura 14

Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 3



Análisis del experimento 4

En el experimento 4 se evaluó el rendimiento del modelo manteniendo la profundidad máxima de aprendizaje de 4 y aumentando la validación cruzada a 10. La exactitud del modelo no mejoró en relación el experimento 3. La exactitud fue de 78.57% en un periodo de 4 minutos y 57 segundos. El desempeño del modelo en el experimento 4 se muestra en la Tabla 11 y Figura 15.

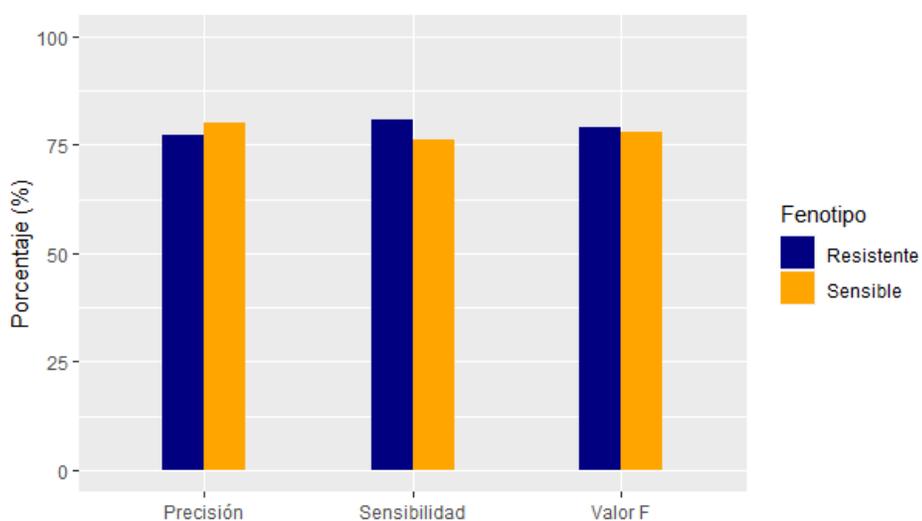
Tabla 11

Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 4

Fenotipo	Precisión (%)	Sensibilidad (%)	Valor F (%)
Resistente	77.27	80.95	79.07
Sensible	80.00	76.19	78.05

Figura 15

Desempeño del modelo de predicción de susceptibilidad a imipenem en el experimento 4



Análisis general de los experimentos

El desempeño del modelo para clasificar fenotipos resistentes y sensibles a imipenem se evaluó por medio de una curva ROC y el AUC. En la Figura 16 se compara el rendimiento de clasificación de los cuatro experimentos. Se observó que el modelo entrenado en el experimento 3 posee un AUC de 0.81 por lo que es capaz de clasificar entre las clases. El tiempo del experimento 3 fue de 2 minutos y 17 segundos (Figura 17)

Figura 16

Curva ROC del modelo de predicción de susceptibilidad a imipenem

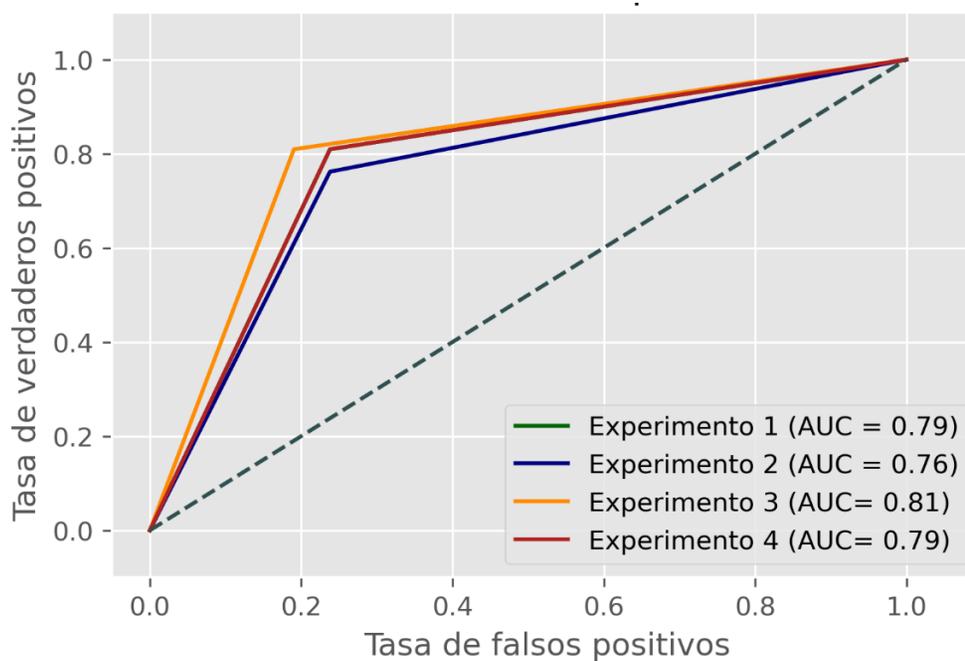
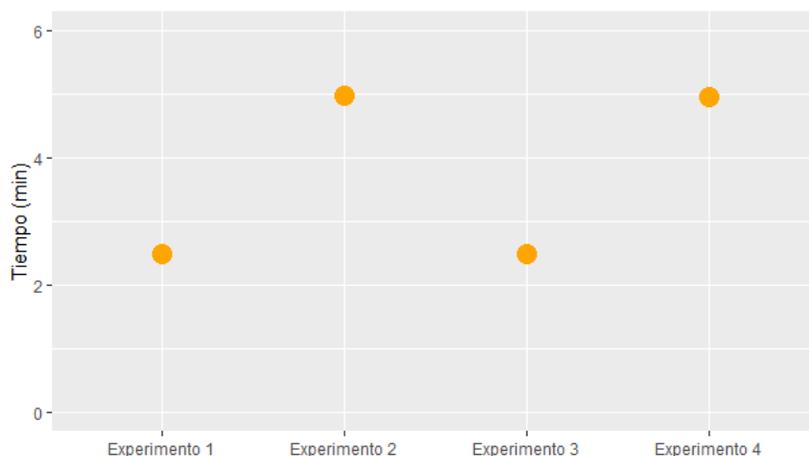


Figura 17

Tiempo de entrenamiento de los modelos de predicción de susceptibilidad a imipenem



Análisis de resultados de aprendizaje del modelo de predicción de susceptibilidad a

meropenem

Análisis del experimento 1

En el experimento 1 se evaluó el rendimiento del modelo aplicado una profundidad máxima de aprendizaje de 3 y una validación cruzada de 5. Se obtuvo una exactitud de 78.57% y se entrenó en un periodo de 2 minutos y 18 segundos. El desempeño general del experimento 1 se presentan en la Tabla 12 y en la Figura 18.

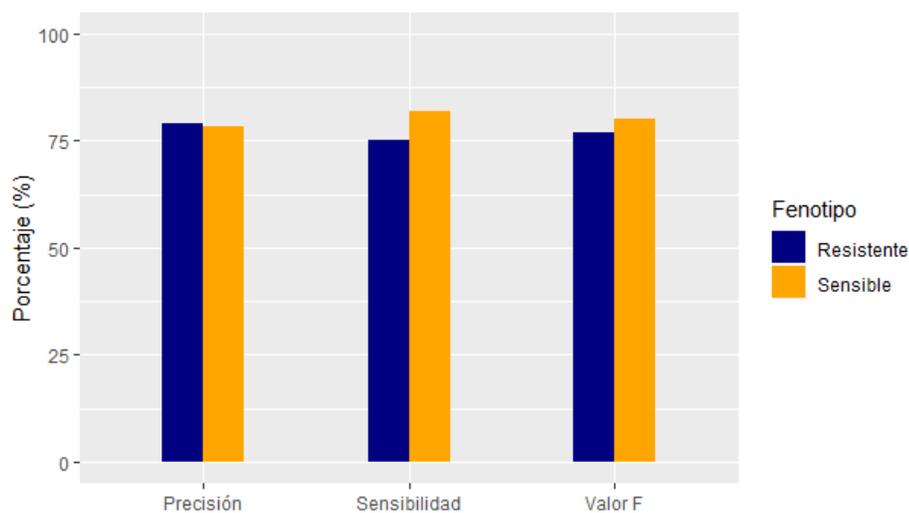
Tabla 12

Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 1

Fenotipo	Precisión (%)	Sensibilidad (%)	Valor F (%)
Resistente	78.95	75.00	76.92
Sensible	78.26	81.81	80.00

Figura 18

Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 1



Análisis del experimento 2

En el experimento 2 se midió el rendimiento del modelo manteniendo la profundidad máxima de aprendizaje de 3 y aumentando la validación cruzada a 10. La exactitud del modelo no mejoró en relación con el experimento 1 obteniendo una exactitud de 78.57% en un periodo de entrenamiento de 4 minutos y 26 segundos. El desempeño del experimento 2 se presenta en la Tabla 13 y Figura 19.

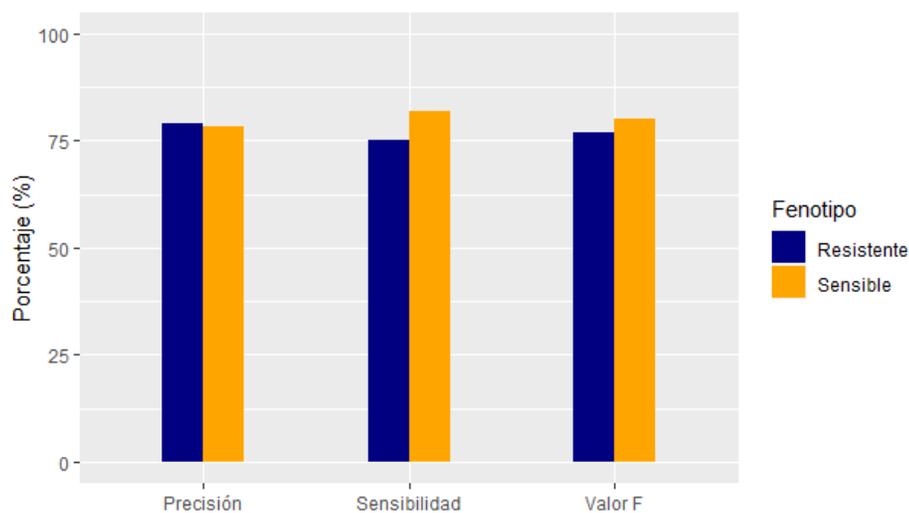
Tabla 13

Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 2

Fenotipo	Precisión (%)	Sensibilidad (%)	Valor F (%)
Resistente	78.94	75.00	76.92
Sensible	78.26	81.81	80.00

Figura 19

Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 2



Análisis del experimento 3

En el experimento 3 se evaluó el rendimiento del modelo a una profundidad máxima de aprendizaje de 4 y una validación cruzada a 5. La exactitud del modelo empeoró en relación con los dos experimentos anteriores. La exactitud fue de 76.19% en un periodo de 2 minutos y 21 segundos. El desempeño del modelo en el experimento 3 se muestra en la Tabla 14 y Figura 20.

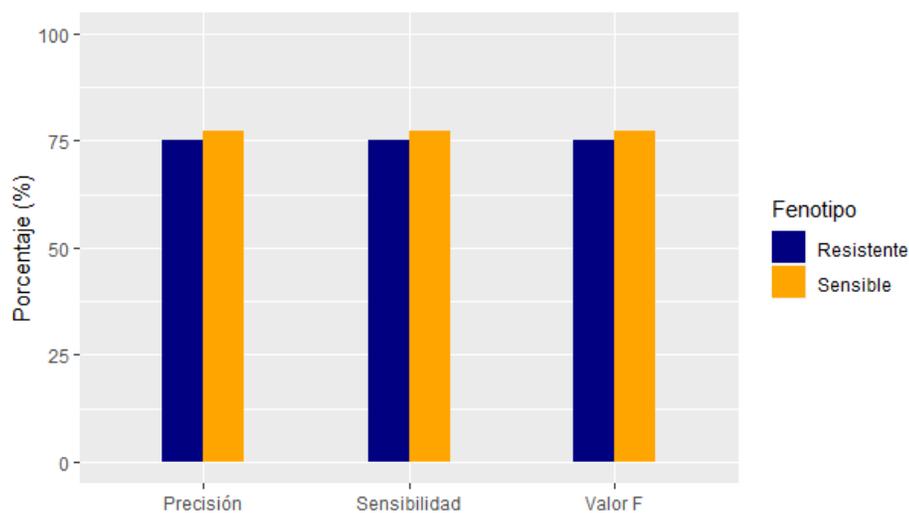
Tabla 14

Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 3

Fenotipo	Precisión (%)	Sensibilidad (%)	Valor F (%)
Resistente	77.27	77.27	77.27
Sensible	75	75	75

Figura 20

Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 3



Análisis del experimento 4

En el experimento 4 se evaluó el rendimiento del modelo manteniendo la profundidad máxima de aprendizaje de 4 y aumentando la validación cruzada a 10. La exactitud del modelo mejoró en relación con los tres experimentos anteriores. La exactitud fue de 80.95% en un periodo de 4 minutos y 26 segundos. El desempeño del modelo en el experimento 4 se muestra en la Tabla 15 y Figura 21.

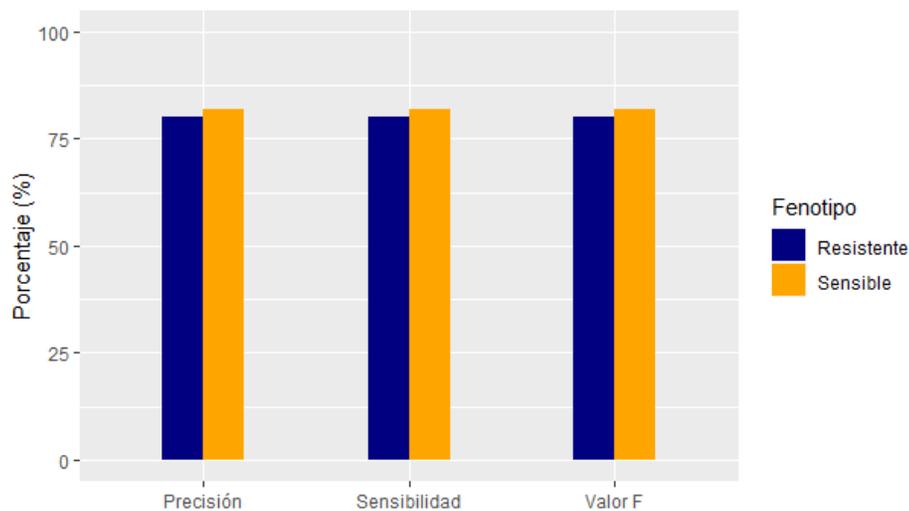
Tabla 15

Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 4

Fenotipo	Precisión (%)	Sensibilidad (%)	Valor F (%)
Resistente	80	80	80
Sensible	81.82	81.82	81.82

Figura 21

Desempeño del modelo de predicción de susceptibilidad a meropenem en el experimento 4

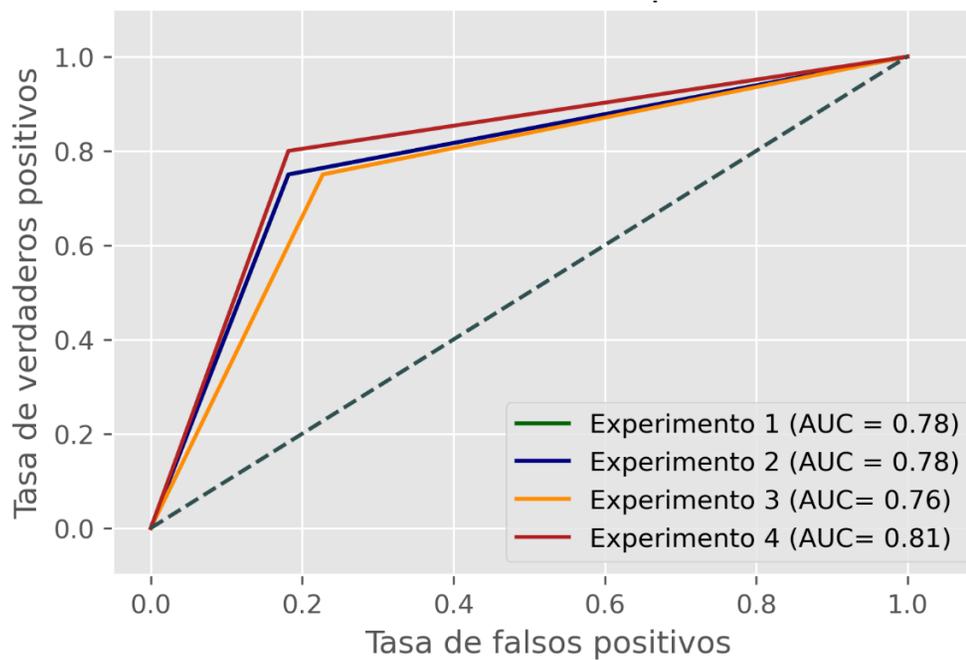


Análisis general de los experimentos

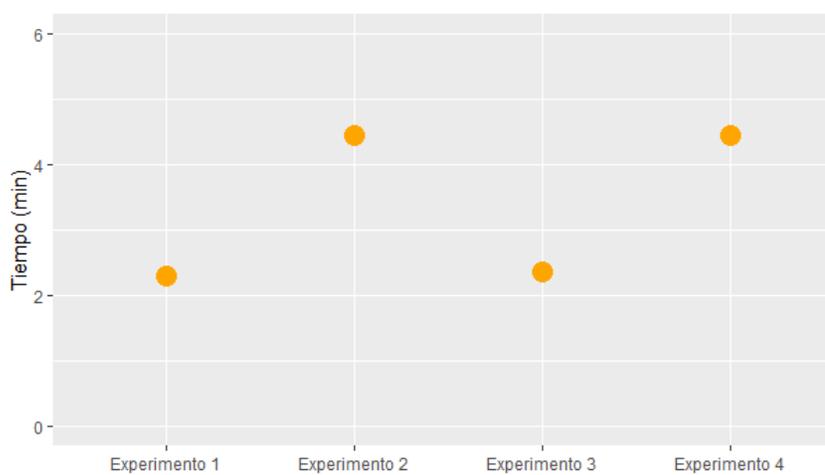
El desempeño del modelo para clasificar fenotipos resistentes y sensibles a meropenem se evaluó por medio de una curva ROC y el AUC. En la Figura 22 se compara el rendimiento de clasificación de los cuatro experimentos. Se observó que el modelo entrenado en el experimento 4 posee un AUC de 0.81 por lo que es capaz de clasificar entre las clases. El tiempo del experimento 3 fue de 2 minutos y 17 segundos (Figura 23).

Figura 22

Curva ROC del modelo de predicción de susceptibilidad a meropenem

**Figura 23**

Tiempo de entrenamiento de los modelos de predicción de susceptibilidad a meropenem



Capítulo V: Discusión

El set de datos utilizado para el estudio contiene 416 secuencias de genoma completo de *K. pneumoniae* obtenidas por medio de PATRIC en archivos extensión “.fna” que poseen secuencias divididas en *contigs*. Utilizar *contigs* proporciona ventajas frente al uso de *reads* porque permiten estudiar cualquier variación en el genoma y evita el uso de secuencias con errores de secuenciación o que se encuentran mal anotadas (Jaillard et al., 2018).

La variable objetivo utilizada para el aprendizaje automático supervisado es el fenotipo de susceptibilidad antimicrobiana para dos carbapenémicos, imipenem y meropenem. Ambos son antibióticos de primera línea utilizados prioritariamente para infecciones provocadas por *K. pneumoniae*. Imipenem, en combinación con cilastatin, es ampliamente usado debido a su versatilidad para el tratamiento de infecciones. Mientras que meropenem es empleado debido a su escasa actividad frente a bacterias Gram-positivas y alta en contra de Gram-negativas (Elshamy & Aboshanab, 2020).

El etiquetado de fenotipos sensibles y resistentes es abordado como un problema de clasificación entre dos clases (Van Camp, Haslam, & Porollo, 2020). Por lo tanto, no se consideraron los fenotipos intermedios como una tercera clase. Se etiquetaron los fenotipos de forma binaria de tal forma que los sensibles se representan como “0” y los resistentes “1”.

La proporción entre las clases que conforman la variable objetivo debe ser proporcional para evitar el sesgo al momento de entrenar el modelo. Un conjunto de datos desequilibrado provoca un aumento en la tasa de falsos negativos y por lo tanto el rendimiento de clasificación disminuye (Thabtah, Hammoud, & Kamalov, 2019).

En la investigación realizada por Nguyen et al. (2019) para predecir la resistencia a azitromicina en *Neisseria gonorrhoeae* se utilizó un set de datos conformado por 392 aislados de

los cuales el 45.40% fueron sensibles y 54.60% resistentes obteniendo una exactitud de predicción del 94%.

El estudio realizado por Santerre, Davis, Xia, & Stevens (2016) propone predecir la resistencia a carbapenémicos en *Acinetobacter baumannii* con un set de datos compuesto por 110 aislados sensibles y 122 resistentes alcanzando una exactitud del 92%.

Para este estudio la proporción de aislados sensibles y resistentes, tanto para imipenem y meropenem, no muestran una diferencia significativa y por lo tanto se descartó la existencia de sesgo en la variable objetivo.

La elección de la característica de entrenamiento influye en la precisión de las predicciones (Mahé & Tournoud, 2018). Estudios realizados en *Staphylococcus aureus* y *Mycobacterium tuberculosis* para predecir resistencia a antimicrobianos lograron predicciones precisas a partir de un catálogo de mutaciones y genes responsables de la resistencia (Sequencing et al., 2014; Yang et al., 2018). Sin embargo, las variaciones genéticas asociadas a la resistencia no han sido reportadas para todas las especies bacterianas por lo que este enfoque es limitado.

Los *k-mer* como característica de entrenamiento es un enfoque novedoso que no requiere conocimiento *a priori* de los genes relacionados a la resistencia. Además, el uso de *k-mers* otorga flexibilidad al estudio debido a su capacidad de identificar variantes genéticas como: presencia/ausencia de genes, SNPs e *indels* (Steinkey, Moat, Gannon, Zovoilis, & Laing, 2020).

La longitud de *k-mers* definida por *k* varía entre 10 a 100 nucleótidos. Determinar la longitud para los estudios de predicción depende de la calidad del ensamblaje y la complejidad de los genomas de entrada. Los *k-mers* largos proporcionan una mayor especificidad, pero a su

vez pierden capacidad para detectar asociaciones entre el fenotipo y genotipo. Por otro lado, los *k-mers* cortos otorgan mayor sensibilidad, pero aumentan la redundancia (Jaillard et al., 2017).

El estudio realizado por Nguyen et al. (2019) evalúa el rendimiento del modelo de predicción para longitud de *k-mer* entre 5 y 10. Mientras mayor fue la longitud *k*, la exactitud del modelo aumentó al igual que el tiempo de entrenamiento.

Santerre et al., (2016) asegura que a mayor longitud de *k* las métricas de la curva ROC y el área bajo la curva ROC (AUC) mejoran. Caso contrario, cuando se utilizó un valor de *k* de 21 y 41 no se detectaron las pequeñas variaciones presentes en el genoma.

En las investigaciones realizadas por Mahé & Tournoud (2018) y Jaillard et al. (2018) se utilizó el valor por defecto $k = 31$ demostrando que los modelos de predicción mantienen el equilibrio entre las variaciones presentes en las secuencias y la eficiencia computacional. Por lo tanto, para la investigación se mantuvo esta variable en su valor por defecto.

El análisis de *k-mers* genera millones de subsecuencias de ADN en proporción 4^k cuya obtención y análisis consume recursos computacionales y tiempo de ejecución. Jaillard et al. (2017) proponen utilizar un gráfico De Bruijn compactado y unir las secuencias redundantes para formar *unitigs* de tal forma que se disminuyen los millones de subsecuencias obtenidas. En este estudio se obtuvieron 48 551 563 *k-mers* que se compactaron en 1 600 166 *unitigs* y posteriormente se formó la matriz presencia/ausencia de *unitigs*.

Pese a que las secuencias redundantes fueron comprimidas al formar *unitigs*, la proporción de datos sigue siendo extensa y un gran número de variantes pueden no estar relacionadas con el fenotipo. Para obtener las variantes verdaderamente asociadas se utilizaron los principios de un estudio de asociación de genoma completo (GWAS) el cual utiliza métodos

estadísticos para determinar si una variante genética es verdaderamente responsable del fenotipo (J. A. Lees, Tien Mai, Galardini, Wheeler, & Corander, 2019).

La fuerte estructura poblacional bacteriana es una limitación técnica que se aborda en los GWAS. La recombinación del cromosoma bacteriano puede ocurrir en múltiples ocasiones en la vida de la célula o simplemente puede que nunca ocurra. La falta de recombinación y la transmisión clonal del cromosoma llevan a un fuerte desequilibrio de ligamiento (LD). De esta forma todos los sitios del genoma se encuentran correlacionados y cuando se introduce una nueva mutación responsable de un fenotipo, la asociación es confusa y se atribuye erróneamente a todas las mutaciones del genoma (J. Lees, 2017).

El estudio realizado por (Chen & Shapiro, 2015) propone combinar pruebas de selección positiva y GWAS para controlar la estructura poblacional. Sin embargo, la selección positiva por sí sola no es suficiente para asociar el genotipo con el fenotipo de interés.

Los métodos basados en regresión han demostrado ser capaces de corregir las falsas asociaciones originadas por la fuerte estructura poblacional bacteriana. Los modelos lineales mixtos (LMM) determinan el fenotipo en función del genotipo más la similitud genotípica entre dos individuos (Price, Zaitlen, Reich, & Patterson, 2010).

La asociación entre las variantes genéticas y el fenotipo se evalúan por medio de una prueba de hipótesis donde el *valor-p* indica qué tan probable es que una variante se encuentre asociada a un fenotipo. Las variantes deben superar un umbral para ser consideradas como asociadas. La corrección de Bonferroni permite determinar un *valor-p* umbral que debe alcanzar valores a partir de 1×10^{-08} para aceptar la hipótesis nula y limitar los errores tipo I (Zhang, 2016).

J. A. Lees et al., 2016 utilizaron un GWAS para determinar variantes responsables de la resistencia y encontrar nuevos factores asociados a la capacidad de invasión del huésped en *Streptococcus pyogenes*. Por medio de un LMM junto con la corrección de Bonferroni, se corrigió la estructura poblacional y determinó el umbral a un nivel de significancia de 0.05. Los resultados demostraron la capacidad de LMM para mapear las variantes asociadas al fenotipo.

En nuestro estudio se utilizó un LMM con el fin de corregir la estructura poblacional y determinar el valor-p de cada una de las variantes. Los gráficos Q-Q muestra una desviación de un grupo de variantes hacia la hipótesis nula indicando que dichas variables se encuentran fuertemente asociadas al fenotipo (Figuras 10 y 11). La corrección de Bonferroni determinó un umbral de significancia de $6,78 \times 10^{-08}$ ideal para disminuir las falsas asociaciones.

Tras la aplicación de GWAS para el filtrado de variantes asociadas, la matriz presencia/ausencia utilizada como entrada para el entrenamiento del modelo de predicción a imipenem posee una dimensión de 417 características y la matriz para la predicción a meropenem está formada por 312 características. De esta forma la dimensión de características se aproximó al tamaño de la muestra.

Las dimensiones de la matriz de entrada se deben controlar debido a que medida que el número de características aumenta, el tamaño de la muestra también debe aumentar de forma proporcional. Caso contrario, un gran número de características de entrada influye en el rendimiento del modelo volviéndolo más complejo, el tiempo de entrenamiento es mayor y existen tendencias al sobreajuste (Reddy et al., 2020).

Los modelos de *Machine Learning* más utilizado para predecir la resistencia a los antimicrobianos son los modelos basados en árboles de decisión debido a que son capaces de manejar set de datos heterogéneos (numéricos y categóricos), pueden ejecutarse en set de

datos con valores perdidos y son fáciles de interpretar (Lv, Deng, & Zhang, 2020). Los árboles de decisión potenciados con el algoritmo XGBoost muestran un entrenamiento escalable, portátil y preciso pese a que los recursos computacionales sean escasos y el tamaño de la muestra reducido (T. Chen & Guestrin, 2016).

En la investigación realizada por Steinkey et al. (2020) se demostró que los árboles de decisión potencializados con XGBoost se entrenan más rápido, usan menos memoria y son más precisos que los métodos de aprendizaje profundo.

Debido a que el proyecto de investigación estuvo enfocado a ser realizado bajo capacidades computacionales en máquinas de escritorio de libre acceso, se utilizó el algoritmo XGBoost para el entrenamiento del modelo.

La selección de parámetros se realizó de acuerdo con lo descrito por Nguyen et al. (2018) que plantea que la profundidad máxima óptima para el entrenamiento del modelo de predicción de resistencia en *K. pneumoniae* debe ser 3 o 4. Valores más elevados pueden originar el sobreajuste del modelo y valores menores con llevan a predicciones espurias.

Martínez, Mora, Lériada, Álvarez, & Soguero (2019) utilizaron una validación cruzada de 5 en su modelo de predicción de resistencia a antimicrobianos en múltiples patógenos presentes en unidades de cuidados intensivos. Por otro lado, Van Camp et al. (2020) utilizaron una validación cruzada de 10. Mientras mayor sea el número de validaciones cruzadas realizadas por el modelo, mayor será el tiempo de entrenamiento, pero se evitará el sobreajuste.

La presente investigación estableció un diseño experimental 2^2 para probar la mejor combinación entre la profundidad máxima y la validación cruzada. En el modelo de predicción de susceptibilidad a imipenem se estableció como mejor combinación la propuesta en el experimento 3 donde se utilizó una profundidad máxima de 4 y una validación cruzada de 5. Se

alcanzó una exactitud del 81%, precisión de 80.95%, sensibilidad del 80.95% y una capacidad de clasificación del 81%.

En caso de meropenem, el mejor resultado se obtuvo en el experimento 4 en el cual se estableció la profundidad máxima de 4 y una validación cruzada de 10. La exactitud fue del 80,95%, la precisión del 81%, la sensibilidad del 81% y la capacidad de clasificación del 80,95%.

De acuerdo con lo descrito por Shi et al. (2019) las métricas pueden mejorar aumentando el tamaño de la muestra y reduciendo las dimensiones de la matriz de entrada. El estudio realizado por Nguyen et al. (2018) se llevó acabo con 1497 aislados y los *k-mer* se alinearon con genes responsables de la resistencia. De esta forma se obtuvo una matriz de entrada robusta y con dimensiones reducidas. Se logró una exactitud de predicción de susceptibilidad a imipenem y meropenem del 94% y 93% respectivamente.

Capítulo VI: Conclusiones y Recomendaciones

Conclusiones

El set de datos se conformó de forma proporcional para evitar el sesgo durante el entrenamiento hacia alguna de las dos clases. Para imipenem se recopiló 237 aislados resistentes y 179 sensibles. Para meropenem se obtuvieron 227 aislados resistentes y 189 sensibles.

Se utilizó un GWAS para el preprocesamiento de las secuencias y obtener una matriz de características con dimensiones reducidas para evitar el sobre ajuste en el entrenamiento. Pyseer se muestra como una herramienta eficiente para controlar la estructura poblacional por medio de un modelo lineal mixto y la corrección de múltiples pruebas o corrección de Bonferroni.

Los mejores parámetros de entrenamiento para el modelo de predicción de susceptibilidad a imipenem fueron: valor alfa $1e-4$, submuestreo por columna y fila de 0.6, tasa de aprendizaje de 0.01, objetivo de aprendizaje *binary:hinge*, profundidad máxima de 4 y una validación cruzada de 5. El modelo mostró una exactitud del 81% para clasificar fenotipos sensibles y resistentes.

En el caso del modelo de predicción de susceptibilidad de meropenem, los mejores parámetros fueron: valor alfa $1e-4$, submuestreo por columna y fila de 0.6, tasa de aprendizaje de 0.01, objetivo de aprendizaje *binary:hinge*, profundidad máxima de 4 y validación cruzada de 10. El modelo mostró una exactitud del 80,95% para clasificar fenotipos sensibles y resistentes.

Ambos modelos fueron entrenados con el algoritmo XGBoost con recursos computacionales estándar alcanzando una exactitud, precisión, sensibilidad y valor F superiores al 80% y tiempos de entrenamiento menores a los 5 minutos.

Recomendaciones

- Aumentar la cantidad de aislados que formarán el set de datos para obtener un análisis más profundo y mejorar los resultados del modelo. El sitio web FTP PATRIC permite descargar múltiples secuencias de genoma y sus referencias de anotación en diversos formatos. El enlace del sitio web: <ftp://ftp.patricbrc.org/>
- Construir la filogenia es un servidor o RAxML a través del portal web CIPRES Science optimiza tiempo en su elaboración y análisis. El sitio web de CIPRES Science: <http://www.phylo.org/>
- Probar diferentes longitudes de *k-mers* para la obtención y conteo de *unitigs* de tal forma que se determine la longitud óptima para el análisis.
- El conteo de *unitigs* se puede realizar en unitig-caller. Los archivos de salida son más fáciles de interpretar y también se pueden utilizar como entrada en Pyseer. Repositorio de unitig-caller: <https://github.com/johnlees/unitig-caller>

Bibliografía

- Antonopoulos, D. A., Assaf, R., Aziz, R. K., Brettin, T., Bun, C., Conrad, N., Davis, J. J., Dietrich, E. M., Disz, T., Gerdes, S., Kenyon, R. W., Machi, D., Mao, C., Murphy-Olson, D. E., Nordberg, E. K., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., ... Yoo, H. (2018). PATRIC as a unique resource for studying antimicrobial resistance. *Briefings in Bioinformatics*, 20(4), 1094–1102. <https://doi.org/10.1093/bib/bbx083>
- Árbol de decisión en Machine Learning (Parte 1) - sitiobigdata.com. (n.d.). Retrieved February 21, 2021, from <https://sitiobigdata.com/2019/12/14/arbOL-de-decision-en-machine-learning-parte-1/#>
- Baştanlar, Y., & Özuysal, M. (2013). Introduction to Machine Learning. In *Methods in Molecular Biology* (Vol. 1107, pp. 105–128). <https://doi.org/10.1007/978-1-62703-748-8>
- Bengoechea, J. A., & Sa Pessoa, J. (2019). Klebsiella pneumoniae infection biology: Living to counteract host defences. *FEMS Microbiology Reviews*, 43(2), 123–144. <https://doi.org/10.1093/femsre/fuy043>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 42(8), 785–794. <https://doi.org/https://doi.org/10.1145/2939672.2939785>
- Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, 44(D1), D67–D72. <https://doi.org/10.1093/nar/gkv1276>
- Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, 132(20), 1920–1930. <https://doi.org/10.1161/CIRCULATIONAHA.115.001593>
- Drezen, E., Rizk, G., Chikhi, R., Deltel, C., Lemaitre, C., Peterlongo, P., & Lavenier, D. (2014).

- GATB: Genome Assembly & Analysis Tool Box. *Bioinformatics (Oxford, England)*, 30(20), 2959–2961. <https://doi.org/10.1093/bioinformatics/btu406>
- Effah, C. Y., Sun, T., Liu, S., & Wu, Y. (2020). *Klebsiella pneumoniae*: An increasing threat to public health. *Annals of Clinical Microbiology and Antimicrobials*, 19(1), 1–9. <https://doi.org/10.1186/s12941-019-0343-8>
- Escandón-Vargas, K., Reyes, S., Gutiérrez, S., & Villegas, M. V. (2017). The epidemiology of carbapenemases in Latin America and the Caribbean. *Expert Review of Anti-Infective Therapy*, 15(3), 277–297. <https://doi.org/10.1080/14787210.2017.1268918>
- Falush, D. (2016). Bacterial genomics: Microbial GWAS coming of age. *Nature Microbiology*, 1(5). <https://doi.org/10.1038/nmicrobiol.2016.59>
- ITIS. (2019). *SingleRpt @ www.itis.gov*. https://www.itis.gov/servlet/SingleRpt/SingleRpt?search_topic=TSN&search_value=32272&print_version=PRT&source=to_print#null
- Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., & Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genetics*, 14(11), 1–28. <https://doi.org/10.1371/journal.pgen.1007758>
- Jaillard, M., Tournoud, M., Lima, L., Lacroix, V., Veyrieras, J. B., & Jacob, L. (2017). Representing Genetic Determinants in Bacterial GWAS with Compacted De Bruijn Graphs. *BioRxiv*, 1–24. <https://doi.org/10.1101/113563>
- Kaler, A. S., & Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. *BMC Genomics*, 20(1), 1–8. <https://doi.org/10.1186/s12864-019-5992->

- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Lees, J. (2017). *The background of bacterial GWAS*. *October*, 1–30.
- Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., & Corander, J. (2018). pyseer: A comprehensive tool for microbial pangenome-wide association studies. *BioRxiv*, 15–17. <https://doi.org/10.1101/266312>
- Lees, J. A., Tien Mai, T., Galardini, M., Wheeler, N. E., & Corander, J. (2019). Improved inference and prediction of bacterial genotype-phenotype associations using pangenome-spanning regressions. *BioRxiv*. <https://doi.org/10.1101/852426>
- Lees, J. A., Vehkala, M., Välimäki, N., Harris, S. R., Chewapreecha, C., Croucher, N. J., Marttinen, P., Davies, M. R., Steer, A. C., Tong, S. Y. C., Honkela, A., Parkhill, J., Bentley, S. D., & Corander, J. (2016). Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications*, *7*. <https://doi.org/10.1038/ncomms12797>
- Li, B., Zhao, Y., Liu, C., Chen, Z., & Zhou, D. (2014). *Molecular pathogenesis of Klebsiella pneumoniae*. *9*, 1071–1081.
- Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835. <https://doi.org/10.1038/nmeth.1681>
- Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., & Heckerman, D. (2012).

- Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6), 525–526. <https://doi.org/10.1038/nmeth.2037>
- Long, W., Linson, S. E., Saavedra, M., Cantu, C., Davis, J., Brettin, T., & Olsena, R. (2017). *Whole-Genome Sequencing of Human Clinical Klebsiella pneumoniae Isolates Reveals Misidentification and Misunderstandings of Klebsiella pneumoniae, Klebsiella variicola, and Klebsiella quasipneumoniae*. 2(4), 1–15.
- Lv, J., Deng, S., & Zhang, L. (2020). A review of artificial intelligence applications for antimicrobial resistance. *Biosafety and Health*. <https://doi.org/10.1016/j.bsheal.2020.08.003>
- Macesic, N., Polubriaginof, F., & Tatonetti, N. P. (2017). Machine learning: Novel bioinformatics approaches for combating antimicrobial resistance. *Current Opinion in Infectious Diseases*, 30(6), 511–517. <https://doi.org/10.1097/QCO.0000000000000406>
- Meletis, G. (2016). Carbapenem resistance: overview of the problem and future perspectives. *Therapeutic Advances in Infectious Disease*, 3(1), 15–21. <https://doi.org/10.1177/2049936115621709>
- Mills, M. C., & Lee, J. (2019). The threat of carbapenem-resistant bacteria in the environment: Evidence of widespread contamination of reservoirs at a global scale. *Environmental Pollution*, 255, 113143. <https://doi.org/10.1016/j.envpol.2019.113143>
- Murray, P., Rosenthal, K., & Pfaller, M. (2018). *Medical Microbiology* (7th ed.). McGraw-Hill Education.
- Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R., Shukla, M., Stevens, R. L., Xia, F., Yoo, H., & Davis, J. J. (2018). Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Scientific Reports*, 8(1), 1–11.

<https://doi.org/10.1038/s41598-017-18972-w>

Nguyen, M., Wesley Long, S., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., Tyson, G. H.,

Zhao, S., & Davisa, J. J. (2019). Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. *Journal of Clinical Microbiology*, 57(2), 1–15. <https://doi.org/10.1128/JCM.01260-18>

Nordmann, P., Naas, T., & Poirel, L. (2011). Global spread of carbapenemase producing Enterobacteriaceae. *Emerging Infectious Diseases*, 17(10), 1791–1798.

<https://doi.org/10.3201/eid1710.110655>

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375(13), 1212–1216.

<https://doi.org/10.1056/NEJMp1609300>

Opoku-Temeng, C., Kobayashi, S. D., & DeLeo, F. R. (2019). Klebsiella pneumoniae capsule polysaccharide as a target for therapeutics and vaccines. *Computational and Structural Biotechnology Journal*, 17, 1360–1366. <https://doi.org/10.1016/j.csbj.2019.09.011>

Paczosa, M. K., & Mecsas, J. (2016). Klebsiella pneumoniae: Going on the Offense with a Strong DefensePaczosa, M. K., & Mecsas, J. (2016). Klebsiella pneumoniae: Going on the Offense

with a Strong Defense. *Microbiology and Molecular Biology Reviews*, 80(3), 629–661.

<https://doi.org/10.1128/mm>. *Microbiology and Molecular Biology Reviews*, 80(3), 629–661. <https://doi.org/10.1128/mnbr.00078-15>

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22), 3691–3693.

<https://doi.org/10.1093/bioinformatics/btv421>

Pan, Y., Lin, T., Chen, C., Chen, Y., & Hsieh, P. (2015). Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of *Klebsiella* spp. *Nature Publishing Group, October*, 1–10. <https://doi.org/10.1038/srep15573>

Pedregosa, F., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Dubourg, V., Pedregosa, F., Gramfort, A., Michel, V., Thirion, B., Pedregosa, F., & Weiss, R. (2011). Scikit-learn : Machine Learning in Python To cite this version : Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12.

Plazak, M. E., Tamma, P. D., & Heil, E. L. (2018). The antibiotic arms race: current and emerging therapy for *Klebsiella pneumoniae* carbapenemase (KPC) - producing bacteria. *Expert Opinion on Pharmacotherapy*, 19(18), 2019–2031. <https://doi.org/10.1080/14656566.2018.1538354>

Poirel, L., Héritier, C., Tolun, V., & Nordmann, P. (2004). Emergence of Oxacillinase-Mediated Resistance to Imipenem in *Klebs. pneu.* *Antimicrobial Agents and Chemotherapy*, 48(1), 15–22. <https://doi.org/10.1128/AAC.48.1.15>

Prado-Vivar, M. B., Ortiz, L., Reyes, J., Villacis, E., Fornasini, M., Baldeon, M. E., & Cardenas, P. A. (2019). Molecular typing of a large nosocomial outbreak of KPC-producing bacteria in the biggest tertiary-care hospital of Quito, Ecuador. *Journal of Global Antimicrobial Resistance*, 19, 328–332. <https://doi.org/10.1016/j.jgar.2019.05.014>

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7), 459–463. <https://doi.org/10.1038/nrg2813>

- Read, T. D., & Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: A new direction for bacteriology. *Genome Medicine*, 6(11), 1–11. <https://doi.org/10.1186/s13073-014-0109-z>
- Reyes, J., Aguilar, A. C., & Caicedo, A. (2019). Carbapenem-resistant *Klebsiella pneumoniae*: Microbiology key points for clinical practice. *International Journal of General Medicine*, 12, 437–446. <https://doi.org/10.2147/IJGM.S214305>
- Russo, T. A., & Marr, C. M. (2019). *crossmA* structural, epidemiological & genetic overview of *Klebsiella pneumoniae* carbapenemases (KPCs). 32(3), 1–42.
- Russo, T. A., Shon, A. S., Beanan, J. M., Olson, R., Macdonald, U., Pomakov, A. O., & Visitacion, M. P. (2011). Hypervirulent *K. pneumoniae* Secretes More and More Active Iron-Acquisition Molecules than “Classical” *K. pneumoniae* Thereby Enhancing its Virulence. 6(10). <https://doi.org/10.1371/journal.pone.0026734>
- Samuelsen, Toleman, M. A., Hasseltvedt, V., Fursted, K., Leegaard, T. M., Walsh, T. R., Sundsfjord, A., & Giske, C. G. (2011). Molecular characterization of VIM-producing *Klebsiella pneumoniae* from Scandinavia reveals genetic relatedness with international clonal complexes encoding transferable multidrug resistance. *Clinical Microbiology and Infection*, 17(12), 1811–1816. <https://doi.org/10.1111/j.1469-0691.2011.03532.x>
- San, J. E., Baichoo, S., Kanzi, A., Moosa, Y., Lessells, R., Fonseca, V., Mogaka, J., Power, R., & de Oliveira, T. (2020). Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Frontiers in Microbiology*, 10(January). <https://doi.org/10.3389/fmicb.2019.03119>
- Sandoval Serrano, L. (2018). *Algoritmos de aprendizaje automático para análisis y predicción de*

datos. 36–40.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>

Soria-Segarra, C., Soria-Segarra, C., Catagua-González, A., & Gutiérrez-Fernández, J. (2020). Carbapenemase producing Enterobacteriaceae in intensive care units in Ecuador: Results from a multicenter study. *Journal of Infection and Public Health*, *13*(1), 80–88. <https://doi.org/10.1016/j.jiph.2019.06.013>

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>

Struve, C., Bojer, M., & Krogfelt, K. A. (2008). *Characterization of Klebsiella pneumoniae Type 1 Fimbriae by Detection of Phase Variation during Colonization and Infection and Impact on Virulence*. *76*(9), 4055–4065. <https://doi.org/10.1128/IAI.00494-08>

Suay-García, B., & Pérez-Gracia, M. T. (2006). Present and Future of Carbapenem-Resistant Enterobacteriaceae (CRE) Infections. *AIP Conference Proceedings*, *827*, 248–261. <https://doi.org/10.1063/1.2195216>

Tamma, P. D., Suwantararat, N., Rudin, S. D., Logan, L. K., Simner, P. J., Rojas, L. J., Mojica, M. F., Carroll, K. C., & Bonomo, R. A. (2016). First report of a verona integron-encoded metallo- β -lactamase-producing *Klebsiella pneumoniae* infection in a child in the United States. *Journal of the Pediatric Infectious Diseases Society*, *5*(3), e24–e27. <https://doi.org/10.1093/jpids/piw025>

Tsai, Y., Fung, C., Lin, J., Chen, J., Chang, F., Chen, T., & Siu, L. K. (2011). *Klebsiella pneumoniae*

Outer Membrane Porins OmpK35 and OmpK36 Play Roles in both Antimicrobial Resistance and Virulence *J. Clin. Microbiol.* 55(4), 1485–1493. <https://doi.org/10.1128/AAC.01275-10>

Vargas, J. A. L., & Toro, L. M. E. (2010). *K. pneumoniae*: ¿The new “superbacteria”?

Pathogenicity, epidemiology and resistance mechanisms. *Antonie van Leeuwenhoek*, 23(2), 157–165.

Wang, B., Zhang, P., Li, Y., & Wang, Y. (2019). *Klebsiella pneumoniae*-induced multiple invasive abscesses. *Medicine*, 98(39), e17362. <https://doi.org/10.1097/md.00000000000017362>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

Wu, W., Feng, Y., Tang, G., Qiao, F., McNally, A., & Zonga, Z. (2019). *crossm* NDM Metallo-β-Lactamases and Their Bacterial Producers in. *Clinical Microbiology Reviews*, 32(2), e00115-18.

Yong, D., Toleman, M. A., Giske, C. G., Cho, H. S., Sundman, K., Lee, K., & Walsh, T. R. (2009).

Characterization of a new metallo-β-lactamase gene, bla NDM-1, and a novel erythromycin esterase gene carried on a unique genetic structure in *Klebsiella pneumoniae* sequence type 14 from India. *Antimicrobial Agents and Chemotherapy*, 53(12), 5046–5054.

<https://doi.org/10.1128/AAC.00774-09>