Sistema para identificar bacterias MDR: Pseudomonas aeruginosa

Daniel Pardo Navarro

Resumen– A partir de los datos de antibiogramas del Hospital Clínic de Barcelona se diseña una metodología basada en algoritmos de clustering que permite detectar las habitaciones contaminadas e identificar la cepa de pseudomonas aeruginosa existente. Del total de 35 habitaciones identificadas como infectadas por la bacteria, 12 quedan excluidas del análisis al no haber muestras de pacientes, en otras 9 se detecta más de una cepa y en 14 una única cepa. En estas últimas 14 habitaciones se recomienda el antibiótico más adecuado para el tratamiento del paciente.

Palabras clave– Clustering, detección de cepas, infección hospitalaria, inteligencia artificial, pseudomonas aeruginosa, tratamiento antibiótico

Abstract– With the antibiogram data from the Hospital Clinic de Barcelona, a methodology based on clustering algorithms is designed to detect contaminated rooms and identify the existing strain of Pseudomonas aeruginosa. Of the total of 35 rooms identified as infected by the bacteria, 12 were excluded from the analysis as there were no patient samples, 9 had more than one strain detected and 14 had a single strain. In these last 14 rooms, the most suitable antibiotic is recommended for the treatment of the patient.

Keywords– Clustering, strain detection, hospital infection, artificial intelligence, pseudomonas aeruginosa, antibiotic treatment

1 INTRODUCCIÓN

En los últimos años la mayor parte de las inversiones en investigación de inteligencia artificial se centran en el sector sanitario. En medicina se puede distinguir la inteligencia artificial virtual, centrada en algoritmos para la toma de decisiones, y la física, con el desarrollo de robots que ayudan a realizar cirugías de forma más precisa o prótesis inteligentes para personas discapacitadas [1] [2]. Uno de los instrumentos más útiles son los datos de las históricas clínicas electrónicas que permiten a los algoritmos realizar diagnósticos, detección de enfermedades o ayudar en la toma decisiones médicas [3].

Este proyecto tiene como objeto de estudio las infecciones que se producen en el entorno hospitalario. En este campo existen diversos trabajos de inteligencia artificial que se centran en el control y la propagación de las infecciones. Por ejemplo, la predicción de sufrir una infección a partir variables estadísticamente significativas [4] o el uso de tecnología activa de Identificación por Radiofrecuencia (RFID) para crear redes de contactos de las personas que permite identificar supercontactores y rutas de transmisión para poder actuar de forma preventiva [5] [6].

Las infecciones nosocomiales (enfermedades adquiridas en el entorno hospitalario) son actualmente uno de los problemas médicos más importantes. Se estima que en la Unión Europea cada año se producen más de 4 millones de infecciones nosocomiales y cerca de 37.000 muertes se relacionan con ellas según el European Centre for Disease Prevention and Control (ECDC). Existen tres elementos clave para que se produzca una infección nosocomial: un microorganismo infeccioso que viva en el entorno hospitalario, una persona con predisposición de ser infectado y una cadena de transmisión que permite al microorganismo llegar a dicha persona vulnerable [7].

Dentro de estas enfermedades, los principales organismos internacionales como la Organización Mundial de la Salud (OMS) o el Consorcio Internacional para el Control de Infecciones Nosocomiales destacan a la bacteria pseudomonas aeruginosa como uno de los principales microorganismos causante de infecciones nosocomiales y se considera un problema de salud mundial [8].

Las pseudomonas son un conjunto de bacterias gramnegativas móviles que pueden habitar en tierra, agua, plantas, insectos y animales, aunque prefieren ambientes húmedos [9] y cuya frecuencia de infección oscila dependiendo del

• E-mail de contacto: daniel.pardo@e-campus.uab.cat
• Mención realizada: Computación
• Trabajo tutorizado por: Ramón Grau Sala (Arquitectura de Computadors i Sistemes Operatius)
• Curso 2020/2021

país, área geográfica, tipo de hospital y servicios, perfil de los pacientes y patrón de uso de los tratamientos con antibióticos [10].Gracias a la capacidad de mutación y adaptación de la bacteria en Europa es intrínsecamente resistente a muchos agentes antimicrobianos importantes y se ha observado que el 32,1 % de los casos aislados de pseudomonas aeruginosa son resistentes al menos a uno de los antimicrobianos utilizados [11]. En España, las infecciones hospitalarias por pseudomonas aeruginosa representan el 10,23 % solo superadas por la *escherichia coli*, con la diferencia de que el contagio de esta última es común en la comunidad [12]. Los principales mecanismos de transmisión de la bacteria es mediante el contacto directo de las manos entre el personal sanitario y los pacientes, el uso de dispositivos médicos invasivos o realizarse pruebas diagnósticas [13].

El diagnóstico de la infección por pseudomonas aeruginosa se debe realizar con un cultivo. La bacteria es resistente y sensible a antibióticos dependiendo de la cepa por lo que el tratamiento debe prescribirse en base a los resultados de su antibiograma. No obstante, los resultados no se obtienen de forma automática y el tiempo depende de la máquina y técnica utilizada, por lo que inicialmente no se puede saber la cepa que ha infectado al paciente. Dado que dentro del entorno hospitalario existen multitud cepas de pseudomonas aeruginosa cada una con resistencias particulares es importante tenerlas identificadas para poder suministrar un tratamiento adecuado [14].

En la sección 2 se exponen los objetivos del proyecto, en la sección 3 se analiza la base de datos, en la sección 4 se explica la metodología empleada, en la sección 5 se presentan los resultados obtenidos y en la sección 6 las principales conclusiones.

2 Objetivos

El proyecto tiene los siguientes **objetivos principales**:

1. Diferenciar e identificar las diferentes cepas de pseudomonas aeruginosa que hay en el hospital a partir de la multiresistencia a los medicamentos observada en los antibiogramas.
2. Determinar el lugar más probable en el que habita cada tipo de cepa de la bacteria de forma que se pueda identificar el lugar origen de la infección por pseudomonas de un paciente.
3. Conectar los resultados al sistema para que la información pueda ser visualizada por los responsables oportunos.

Objetivos secundarios:

1. Analizar la base de datos y hacer un tratamiento adecuado de los valores perdidos y outliers que puedan afectar al resultado final.
2. Implementar múltiples algoritmos de clustering y seleccionar el que obtenga una mejor clasificación.
3. Correlacionar las cepas detectadas con los pacientes y las ubicaciones del hospital.

3 Base de datos

Se utiliza la base de datos del Hospital Clínic de Barcelona con un total de 1236 muestras pseudoanonimizadas que contienen los resultados de antibiogramas de pseudomonas aeruginosa (los datos recogidos corresponden al año 2020). El conjunto de datos original utilizado en el estudio contiene 624 muestras. Posteriormente, se han continuado realizando antibiogramas hasta un total de 612 muestras adicionales que se tendrán en cuenta para el proceso de validación.

Para cada muestra se recoge la información en 34 atributos. Los dos primeros corresponden a un identificador único y al tipo de forma de adquisición de la muestra. Los 32 atributos restantes hacen referencia a un tipo de antibiótico utilizado para el antibiograma de la muestra. No se analizan todos los antibióticos en cada muestra por lo que hay algo más del 72 % de valores nulos. En el resultado del antibiograma si la cepa de pseudomonas es sensible se identifica con un 1 y si es resistente con un 2. De los 15 posibles tipos de muestra diferentes cinco de ellas representan sobre el 93 % del total: 5 (29 %), 6 (27,2 %), 3 (19,2 %), 1 (12,5 %) y 4 (5,1 %).

Por lo que respecta a la distribución de los valores, teniendo en cuenta todas las muestras y los 32 posibles antibióticos para el conjunto original se tiene un total de 14.380 valores nulos donde no se ha analizado el antibiótico (72,02 % del total), 4.741 de casos sensibles (23,74 % del total) y 847 casos resistentes (4,24 % del total). En el conjunto adicional se obtiene una distribución similar con 14.259 valores nulos (72,81 % del total), 4.458 de casos sensibles (22,76 % del total) y 867 casos resistentes (4,43 % del total).

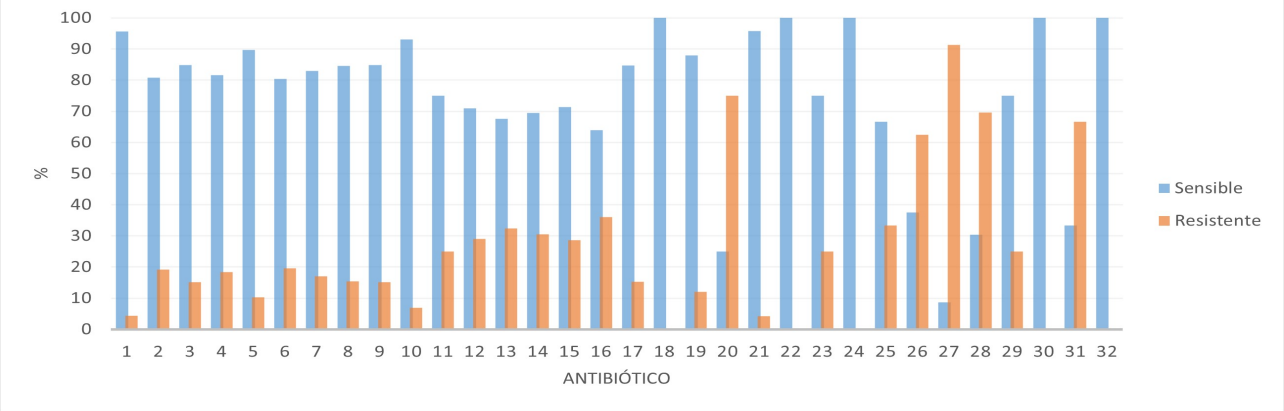

Fig. 1: Porcentaje de resistencia observada en cada antibiótico

La Figura 1 muestra para cada uno de los antibióticos el porcentaje de casos sensibles y casos resistentes teniendo en cuenta el total de las muestras. Destacan algunos casos extremos de resistencia como el antibiótico 20, 26, 27 y 28 pero cuentan con un número pequeño de muestras resistentes (73) que representan solo el 8,6 % del total de casos resistentes. En este sentido, la mayor parte de los casos resistentes se engloban en los antibióticos2 (14,05 %), 5 (7,56 %), 6 (14,4 %), 7 (12,52 %), 8 (11,33 %), 9 (11,1 %) y 10 (5,08 %) que en conjunto suman 644 casos resistentes y representan un 76,04 % del total. No obstantes, en ninguno de estos antibióticos los casos resistentes llegan a alcanzar el 20 %.

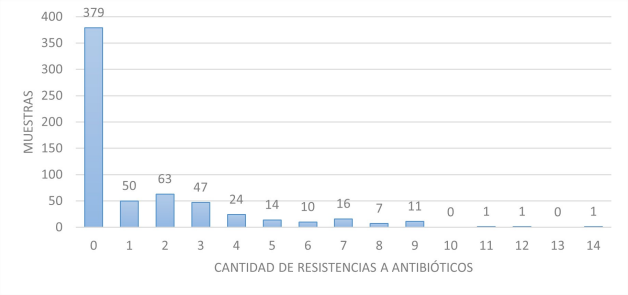

Fig. 2: Recuento de la cantidad de antibióticos resistentes por cada muestra

En la Figura 2 se puede observar la cantidad de resistencias a antibióticos en las muestras analizadas. En este caso, destaca significativamente que 379 muestras (60,7 % del total) no presentaban ninguna resistencia y la bacteria era sensible a los antibióticos. 50 muestras (8 %) presentan 1 resistencia, 63 muestras (10 %) 2 resistencias o 47 muestras (7,5 %) 3 resistencias. También destacan algunos valores extremos donde se puede observar cómo hay una muestra que presenta 11, 12 o hasta 14 resistencias en el antibiograma. El conjunto de datos adicional presenta una distribución equivalente que se pude consultar en el apéndice A.1)

Para complementar estas datos se tiene la información relativa a los pacientes de la base de datos anterior con la fecha de entrada y salida de una determinada habitación. En total hay 345 habitaciones diferentes y 253 pacientes únicos que cambian de habitación durante el ingreso. De media, un paciente pasa por 3,21 habitaciones antes de que se le detecte un positivo por pseudomonas. La Figura 3 muestra la cantidad de pacientes que pasan por un determinado número de habitaciones. Como caso extremo se observa un paciente que pasa por hasta 15 habitaciones diferentes. Además, el 58,08 % de los pacientes que han pasado por más de una habitación una de las habitaciones ha sido urgencias.

En el proceso de análisis numérico se deberán considerar las dos bases de datos disponible y contrastar los resultados de las muestras con los pacientes, los días de ingreso y las habitaciones para validar la coherencia. Para este paso se tienen en cuenta las habitaciones en las que ha estado un paciente los 30 días anteriores a la detección de la infección por pseudomonas. Además, siempre se dejan 5 días de margen entre que se detecta la infección en un paciente y se empieza a contabilizar ya que es el tiempo mínimo de incubación de la bacteria.

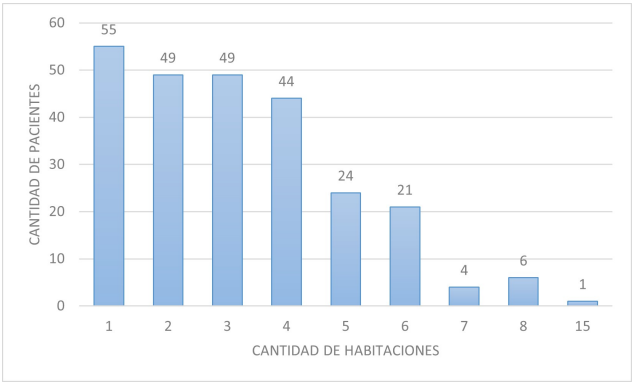

Fig. 3: Cantidad habitaciones por las que pasa un paciente

4 METODOLOGÍA

La Figura 4 muestra las principales etapas que se debe seguir en el proceso de clustering para generar conocimiento a partir de datos. Siempre se puede volver a la etapa anterior en caso de detectar algún error. En este apartado se exponen las fases de procesamiento y tratamiento de la base de datos (4.1), los algoritmos de clustering utilizados (4.2), las medidas de distancia (4.3) y el análisis numérico (4.5).

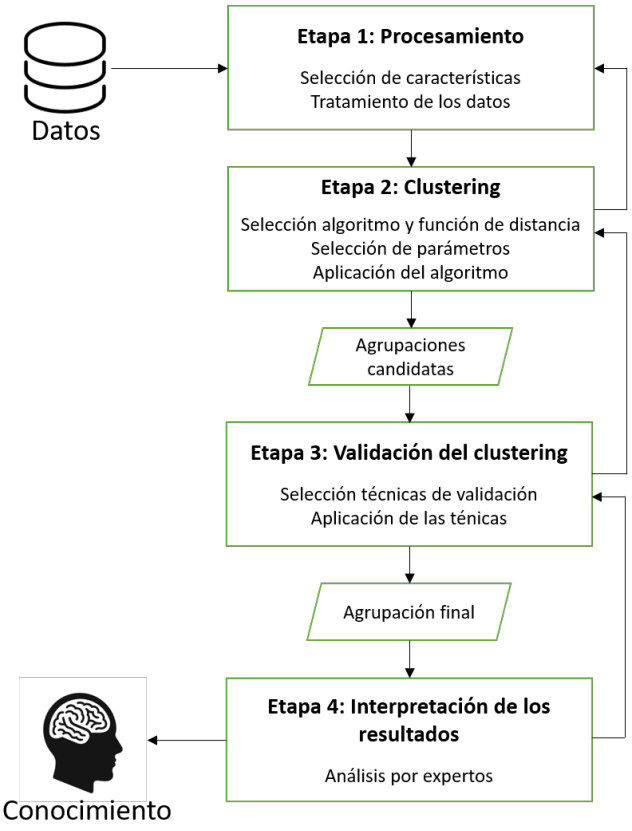

Fig. 4: Etapas del proceso de clustering

4.1. Procesamiento

En el apartado 3 se ha determinado que hay una gran cantidad de valores nulos. Los algoritmos de clustering no permiten que haya datos vacíos por lo que se implementan diferentes técnicas para el tratamiento de los datos. Las operaciones en secuencia que se realizan sobre la base de datos

son las siguientes:

⇒ Eliminar atributos vacíos

El primer paso consiste en eliminar los atributos que no tengan ningún valor. En el caso de que algún antibiótico no se utilice para ninguna muestra de los antibiogramas debe eliminarse directamente de los datos.

⇒ Eliminar atributos con un % de valores desconocidos (parametrizable)

Se eliminan los atributos con un número de valores desconocidos determinado. En este caso, se indica un % de valores nulos a partir del cual el atributo debe ser eliminado si lo supera en valores desconocidos.

⇒ Eliminar instancias con un % de valores desconocidos (parametrizable)

Se aplica la misma lógica que en el caso anterior, pero para las muestras.

⇒ Agrupar por tipo de muestra (opcional)

Si la opción está seleccionada las muestras se agrupan en subdatasets dependiendo del tipo de muestra.

⇒ Imputar un valor a los valores desconocidos

Para imputar un valor a los valores nulos se implementan los siguientes criterios: media, mediana, moda, Bayesian ridge, Decision tree regressor, Extra Trees regressor, KNeighbors regressor. Para los últimos cuatro modelos se utiliza librería IterativeImputer que modela cada característica con valores perdidos en función de las otras características y usa esa estimación para atribuir un valor. Lo hace de manera iterativa por turnos: en cada paso, una columna de características se designa como salida Y y los otros atributos se tratan como entradas X. Se ajusta un regresor que se usa para predecir los valores perdidos de Y. Esto se hace para cada característica de forma iterativa y se repite según el valor de iteraciones que se indique.

⇒ Redondear valores (opcional)

Los valores imputados pueden ser decimales dependiendo de la opción seleccionada. Por ello, se implementa la opción redondear los valores imputados para que la base de datos sea completamente binaria con valores a 1 ó 2 o seguir con el resultado obtenido en este proceso.

⇒ Eliminar atributos con todos los valores iguales

Si después del proceso de imputar valores un atributo contiene el mismo valor en todas las instancias no aporta información por lo que debe eliminarse.

⇒ Aplicar PCA (opcional)

Cabe tener en cuenta que aplicar el PCA sobre bases de datos binarias tampoco es recomendable ya que reduce la influencia de la dirección de varianza máxima y enfatiza las direcciones ortogonales de varianza más baja. Esta opción puede funcionar si al imputar los datos se obtienen valores decimales y se sigue la ejecución. En caso de que se quiera utilizar PCA opcionalmente se pueden hacer configuraciones adicionales. Para determinar el número de componentes de la PCA se puede indicar el porcentaje de capacidad explicativa mínimo que se quiere mantener.

4.2. Algoritmos

La finalidad del análisis de la base de datos es identificar las diferentes cepas existentes y clasificar cada muestra en una cepa. El conjunto de datos no ésta etiquetado y se debe buscar automáticamente la estructura interna de los datos realizando agrupamientos. La estrategia de machine learning más adecuada para realizar esta tarea es recurrir a técnicas de aprendizaje no supervisado utilizando algoritmos de clustering.

El clustering es una técnica que permite agrupar grandes cantidades de datos multidimensionales en conjuntos diferenciados o clústeres que comparten características similares. Los datos en clústeres distintos no deben guardar ningún tipo de relación para que el análisis funcione de forma efectiva. La Tabla 1 muestra un resumen de los parámetros utilizados en las diferentes configuraciones de los algoritmos. Los algortimos de clustering implementados para el análisis de los datos son los siguientes:

•K-means

La principal aplicación del algoritmo es realizar agrupamientos o agrupamientos por similitud. El objetivo del algoritmo es minimizar la suma de distancias entre los puntos de un clúster y su centroide. Generalmente se utiliza la distancia euclidiana como medida de distancia, pero en el caso de los datos binarios se puede considerar el uso de la distancia de Hamming.

Inicialmente el algoritmo recibe una matriz de M puntos (muestras) con N dimensiones, un valor K de clúster introducidos por el usuario y toma K puntos arbitrarios de la muestra de datos como centroide de clúster inicial. Es un algoritmo iterativo en el que se destacan dos pasos principales. Un primer paso donde cada punto de la muestra de datos se asigna al clúster con el centroide más cercano. Una vez se han asignado todos los puntos a un clúster, el segundo paso consiste en recalcular el valor del centroide y repetir el algoritmo hasta la convergencia donde el valor del centroide no varía o la variación no es significativa [15].

El algoritmo es muy dependiente de una buena inicialización de los centroides de los clústeres iniciales para obtener agrupaciones aceptables ya que puede converger en mínimos locales. Por este motivo se han elaborado diferentes técnicas que optimizan la selección de los K centroides iniciales [16].

Otra limitación del algoritmo es que el rendimiento depende de la exactitud del valor K que además debe ser introducido por el usuario, por lo que se debe utilizar algún método para encontrar el valor más adecuado para la distribución de datos concreta. En este proyecto se utiliza el método de *elbow* para realizar esta tarea. Por lo que respecta a la distribución, k-means es adecuado para distribuciones isotrópicas de los datos, pero falla para formas alargadas o con superposición de grupos. En este sentido, tampoco sabe diferenciar entre densidades y solo se basa en distancia a los centros para asignar un punto a un clúster, por lo que puntos extremos pueden quedar mal clasificados si tiene el centro de otro clúster a menor distancia.

También es importante tener en consideración que el algoritmo es muy sensible al ruido y los valores *outliers* de forma que debe hacerse un tratamiento de los datos previamente a la ejecución. En este contexto, se entiende que generan ruido aquellos atributos que no son útiles para hacer una predicción y provocan que el resultado empeore o que crean una relación de casualidad en lugar de causalidad con las que llegar a conclusiones inexactas y equivocas [17].

•Spectral clustering

Parte de la idea de representar los datos de las muestras ori-

TABLA 1: OPCIONES DE CONFIGURACIÓN ALGORITMOS

	K-means	**Spectral clustering**	**Mean-Shift**	**GMM**	**POPC**
Distancia	Euclidiana Hamming Minkowski	Euclidiana	None	Mahalanobis	Euclidiana
Número de clústeres	Elbow	Elbow	None	BIC AIC Silhouette	None
Inicialización	K-means++	K-means++	None	None	K-means++
Tolerancia	0.001	0.001	None	None	0.0001
Iteraciones	200	200	200	100	300

ginales mediante un grafo de similitud. En primer lugar, mediante métodos de álgebra lineal se debe realizar el cálculo de los valores propios para construir la matriz de similitud de los datos, siendo las matrices de vectores propios laplacianas una de las principales herramientas. A partir de esta nueva matriz se puede reducir la dimensionalidad de los datos seleccionando las características más importantes y realizar un proceso de clusterización usando un algoritmo estándar como k-means sobre esta nueva representación de los datos. Dado que las muestras a clasificar contienen datos binarios en este trabajo se construye la matriz calculando un grafo de *k-nearest neighbor* [18].

Alguno de los puntos fuertes de este algoritmo es que se puede utilizar para grandes conjuntos de datos y no realiza una suposición sólida sobre la forma de los clústeres, a diferencia del algoritmo de k-means que tiene una tendencia a realizar grupos esféricos y con poca varianza. Tampoco existe el problema de quedarse atascado en mínimos locales o ser dependiente de la inicialización. Con spectral clustering se pueden resolver problemas muy generales como espirales entrelazadas [18].

La principal limitación del algoritmo es que antes de poder realizar la construcción del grafo de similitud se debe definir una función de similitud adecuada. Encontrar esta función es una tarea compleja ya que va a depender de la distribución de los datos. Para este proyecto se utiliza la función de similitud *nearest neighbors* para construir el grafo de similitud ya que las muestras de una misma cepa deben estar próximas entre ellas. Además, una vez construida la matriz de similitud se realiza una clusterización con el algoritmo de k-means, por lo que comparte la debilidad de tener que introducir el valor K adecuado [18].

•Gaussian Mixture Model

GMM se trata de un modelo estadístico en el que se asume que todos los puntos del conjunto de los datos multidimensionales se generan a partir de una mezcla de un número finito de distribuciones gaussianas multivariadas. El modelo debe conocer el número de componentes de los datos ya que se tendrán tantas gaussianas como componentes. Para realizar esta estimación en este trabajo se utilizan tres criterios: el Criterio de información bayesiano (BIC), Criterio de información de Akaike (AIC) o el método Silhouette.

Para muestras de datos no etiquetadas ajustar los parámetros de las distribuciones es complicado y se utiliza junto al algoritmo Expectation–maximization [19].

Se trata un algoritmo iterativo en dos pasos donde en la etapa de Expectación genera distribuciones con parámetros razonables en función de los datos y calcula probabilidades de cada uno. Para ello usa la media actual y la suposición de la desviación estándar para calcular las probabilidades. En la fase de Maximización actualiza la media y la desviación estándar para maximizar las probabilidades. El algoritmo calcula dónde ha de estar centrada la gaussiana y su orientación comprobando la cercanía de cada muestra realizando un descenso coordinado hasta la convergencia [20].

Se trata de un método de soft clustering ya que cada punto tiene una probabilidad de pertenecer a los diferentes grupos con lo que se distingue de k-means que realiza hard clustering donde cada punto pertenece a un único grupo. Por otro lado, k-means funciona con distribuciones esféricas, pero falla cuando la forma de la distribución es alargada, hay superposición de clases o no tienen en cuenta las diferentes densidades de la distribución. Por su parte GMM asume distribuciones normales y realiza agrupaciones suaves para cada clase con probabilidades, por lo que es bueno para estimaciones de densidad. Ajusta una serie de distribuciones normales al conjunto de datos mediante la estimación de un parámetro. No obstante, esta característica de asumir distribuciones normales es una limitación del algoritmo que va a depender del tipo de datos que se deben agrupar [20].

•Mean-Shift

Considera el espacio de características como una densidad de probabilidades. Tiene un funcionamiento básico similar al k-means en el hecho de que se basa en centroides que se actualizan en función de los puntos asignados al clúster determinado. Todas las muestras evolucionan y se mueven hacia la zona de máxima probabilidad que tienen alrededor en función de sus vecinos. Por lo tanto, para determinar la dirección se tiene en cuenta una vecindad que debe ser delimitada en cuanto a su tamaño. El algoritmo recibe como parámetro la dimensión de la ventana deslizante que se debe analizar alrededor de cada punto y establece de forma automática el número de clústeres [21].

Cuanto mayor sea el tamaño de la ventana mayor será la simplificación que se realiza. A partir de la ventana el

algoritmo calcula la media local de los puntos que contiene y mueve la ventana hacia la dirección en el que se encuentra el valor de la media más alto. De esta forma, si hay zonas densas se acercará a esas agrupaciones para aumentar esa densidad hasta establecerse como clústeres. El algoritmo no asume ninguna forma predefinida en el conjunto de datos y puede usarse en espacios de características arbitrarios. Por otro lado, mientras que en k-means se debe indicar el número de clústeres, el algoritmo de mean-shift los encuentra de forma automática. No obstante, el tamaño de la ventana que se debe introducir es decisivo a la hora de que el algoritmo realice una correcta clasificación ya que puede fusionar o separar clústeres si no es el adecuado. No existe un tamaño de ventana correcto por definición y depende de cada conjunto de datos, por lo que para este trabajo se utiliza un algoritmo de estimación lo cual añade coste computacional [21].

•POPC
Powered Outer Probabilistic Clustering se basa en el cálculo de probabilidades descontadas de diferentes características que pertenecen a diferentes grupos. Para el cálculo de las características se tienen en cuenta si está activa o no, por lo que es ideal para conjunto de datos binarios. El algoritmo se inicia utilizando el algoritmo k-means con un número *k* de clústeres igual a la mitad del conjunto de las muestras. Se utiliza un sistema de back-propagation para reducir el número de clústeres y utiliza probabilidades externas potenciadas para reorganizar las muestras a partir una función de evaluación. El proceso finaliza cuando la reorganización de las muestras no aumenta la función de evaluación. Este algoritmo converge en un número óptimo de clústeres y con agrupaciones de mayor calidad que el algoritmo clásico de k-means [22].

4.3. Medidas de distancia

Al utilizar un algoritmo de clustering es importante el concepto de distancia y las diferentes posibles medidas aplicables ya que afectan al resultado de la clasificación dependiendo del conjunto de datos. Las medidas de distancias permiten medir la similitud de muestras normalizadas. Las distancias utilizadas en este trabajo son las siguientes:

•Euclidiana
Es la distancia entre dos puntos en un espacio euclidiano. Se trata de una medida adecuada cuando los datos son de baja dimensión y es importante medir la magnitud de los vectores.

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (1)$$

•Hamming
Se utiliza para medir el número de valores diferentes entre dos cadenas de la misma longitud. Generalmente se usa para comparar dos vectores binarios.

1 0 1 0 0 0 1
0 0 1 0 1 0 1 → d(H) = 2

•Hamming (implementación)
Se realiza una implementación de la distancia de Hamming que redondea el valor de la resta a 0 o 1 para valores decimales. Si un punto es [1,0,0] y el otro [0.6, 0.4, 0.6] la distancia de Hamming será [0.4, 0.4, 0.6] que acabará redondeada a [0,0,1] por lo que la distancia resultante es 1. De esta forma el valor de la distancia siempre será un valor entero.

•Manhattan
Medida geométrica donde distancia entre dos puntos es la suma de las diferencias absolutas de sus coordenadas. Es una medida adecuada para conjuntos de datos con atributos discretos o binarios.

$$d_{MH}(p, q) = \sum_{i=1}^{n} |p_i - p_i| \qquad (2)$$

donde $\mathbf{p} = (p_1, p_2, ..., p_n)$ y $(\mathbf{q} = q_1, q_2, ..., q_n)$ son vectores.

•Minkowski
Es una generalización tanto de la distancia euclidiana como de la distancia de Manhattan. Se utiliza en el espacio vectorial normalizado y con el parámetro p se manipula la métrica para que se parezca a otras:

- p = 1 → Distancia de Manhattan
- p = 2 → Distancia euclidiana
- p = 3 → Distancia de Chebyshev

$$d_{MK}(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \qquad (3)$$

•Mahalanobis
Es la medida de distancia del algoritmo GMM y permite determinar la similitud entre dos variables aleatorias multidimensionales. Se utiliza en problemas en los que se busca conocer la distancia y la correlación entre las variables, superando las limitaciones de la distancia euclidiana.

$$d_{MN}(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \sum{}^{-1} (\vec{x} - \vec{y})} \qquad (4)$$

4.4. Validación interna

Al no disponer de ground truth se deben utilizar métricas de validación que permitan medir el rendimiento a partir del propio modelo. Esta validación se basa en el concepto de cohesión (la distancia que hay entre los puntos de un mismo clúster) y separación (la distancia entre dos clústeres diferentes). Las métricas utilizadas son:

•Coeficiente Silhouette
Esta métrica se calcula a partid de dos componentes: la distancia media entre una muestra y cada uno de los puntos restantes de la misma clase y la distancia media entre la muestra y todos los puntos del clúster más cercano. Las puntuaciones oscilan entre -1 y 1 donde una puntuación de 1 indica grupos densos y bien separados mientras que valores cercanos a 0 superposición de clases. Generalmente se obtiene resultados más altos para grupos convexos que para otro tipo de formas [23].

•Índice Calinski-Harabasz
El índice mide la media de la dispersión entre clústeres

cercanos y la dispersión interna de cada clúster donde la dispersión se define como la suma de las distancias al cuadrado. Obtiene una puntuación alta para grupos densos y bien separados. Hay que tener en cuenta que los clústeres con forma convexa obtienen una puntuación más alta que otro tipo de formas [24].

•Índice Davies-Bouldin
El índice calcula la similitud promedio entre clústeres comparando las distancias del clúster con su tamaño. La mejor puntuación posibles es 0 que indica la mejor separación entre clústeres. Los grupos convexos tienden a obtener un peor resultado con esta métrica [25].

4.5. Análisis numérico

Fase 1. Para cada clusterización realizada se correlacionan los pacientes a los que corresponden las muestras de antibiogramas, la habitación en la que ha estado y el tipo de cepa que se le ha asignado en cada caso. A continuación, se analizan los datos de los pacientes infectados respecto del total de pacientes para evidenciar si una habitación está contaminada por al menos una cepa de pseudomonas. Se comprueba para cada caso el número de positivos y negativos que hay en la habitación y el número de positivos y negativos de la planta.

Para evaluar la relación entre una habitación y los casos positivos por pseudomonas se utiliza el estadístico chi cuadrado. Se trata una prueba de estadística descriptiva que permite extraer información de una muestra para determinar la existencia o no de independencia entre dos variables.

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - n_{ij})^2}{n_{ij}} \tag{5}$$

Como resultado de aplicar el estadístico, Un nivel de significancia α de 0.01 indica que hay un 1 % de concluir que existe una relación entre variables y sea trate de un error. Si un valor $p \leq \alpha$ se rechaza la hipótesis nula H_0 y se concluye que las variables son dependientes ya que existe una relación estadísticamente significativa entre las variables. Por el contrario, si un valor $p > \alpha$ no se puede concluir que las variables estén relacionadas. En nuestro estudio si el resultado de aplicar el test es estadísticamente significativo implica que la habitación está contaminada.

Fase 2. Por otro lado, para verificar el sistema se utiliza cada muestra del conjunto de datos adicional seleccionado las habitaciones en las que ha estado hospitalizado el paciente de la muestra y se comprueba en las diferentes configuraciones si aparece la habitación clasificada en alguno de los clústeres. En caso de coincidencia se recuperan los clústeres en los que aparece asociada la habitación. Si la distancia de Hamming de la muestra es 0 a uno y solo uno de estos clústeres implica que la muestra se puede asociar correctamente a dicho clúster y la cepa se puede identificar. En el caso de que haya dos clústeres con distancia 0 implica que las fronteras no son claras y no se puede determinar la cepa a la que pertenece la muestra con certeza. Si la distancia es diferente de 0 la muestra puede ser totalmente nueva nunca antes vista o que se encuentre en otra habitación.

Al realizar este análisis para cada configuración se contabilizan los errores entendidos como aquellas muestras de una habitación que se pueden asociar a más de un clúster y como acierto las que pertenecen a un único clúster.

5 Resultados

5.1. Análisis general de habitaciones contaminadas

La Tabla 2 muestra la distribución de los casos positivos en las diferentes zonas del hospital y la cantidad de habitaciones significativas al aplicar el test de chi cuadrado sobre los datos generales del hospital. Entre ellas destaca la zona de *Planta* con el menor porcentaje de habitaciones significativas (3,68 %) y, en el caso opuesto, *Otras* con más de un 28 % de habitaciones significativas.

A partir de estos resultados no se pueden identificar las cepas ya que son un conglomerado. Al analizar las cepas de cada habitación por separado no se puede identificar una cepa de forma clara y los resultados obtenidos pueden no reflejar la realidad. Además, un paciente está de media en más de 3 habitaciones diferentes durante la estancia en el hospital y en una habitación puede haber más de una cepa. Por lo tanto, es necesario continuar con el análisis de los resultados de los algoritmos de clustering para tratar de identificar las cepas en función de la habitación y determinar así la resistencia y el antibiótico de tratamiento más adecuado.

TABLA 2: Cantidad de positivos por zona

	Pacientes positivos	**Habitaciones con positivos**	**Habitaciones significativas**
Urgencias	265	50	3 (6,00 %)
Planta	287	190	7 (3,68 %)
UCI	123	46	3 (6,52 %)
Intermedio	60	31	3 (9,68 %)
Otras	78	28	8 (28,57 %)

5.2. Clusterización

Las combinaciones de opciones posibles de ejecutar los algoritmos de clustering producen un total de 1624 configuraciones diferentes. La mitad de estas configuraciones agrupan los datos en subconjuntos en función del tipo de muestra mientras que la otra mitad no lo tenía en cuenta. Agrupar las muestras produce una segregación del total de las 624 muestras en 15 conjuntos, algunos muy reducidos o con una única muestra. Los resultados obtenidos por los algoritmos de clustering para estas configuraciones es inestable ya que el número de muestras para los conjuntos es demasiado pequeño. Las 812 configuraciones que realizan este proceso de clustering se eliminan antes de realizar un análisis más profundo de los resultados, aunque puede ser una práctica relevante en el caso de tener un mayor número de muestras totales.

5.3. Validación interna

Se calculan los índices de validación interna mencionados en la sección 4.4 sobre las 812 configuraciones restantes

y se hace una selección de las mejores configuraciones de la siguiente manera: en el coeficiente Silhouette las configuraciones con una puntuación mayor a 0.45 (319 casos), en el índice Calinski-Harabasz las configuraciones por debajo del 0.3 % respecto del peor resultado (255 casos) y en el índice Davies-Bouldin las configuraciones que estén por encima del 0,5 % respecto del mejor resultado (210 casos). A continuación se cruzan los 3 índices y se obtiene un total de 92 muestras (11,33 % del total de 812) como las mejores opciones posibles atendiendo a las métricas empleadas. La Tabla 3 un resumen de las características de las configuraciones donde destaca el algoritmo de k-means con 85 casos (92,39 % del total). Por lo tanto, estos resultados indican que las muestras se pueden clasificar en grupos esféricos y con poca varianza ya que es una características del algoritmo k-means. Además, el hecho de que aparezcan configuraciones significativas de GMM puede deberse a que haya una ligera superposición de clases y fronteras difusas.

TABLA 3: CONFIGURACIONES SELECCIONADAS

		K-means	**GMM**
Total		85	7
Tratamiento	Media	10	2
	Bayesian ridge	13	1
	Decision tree	11	-
	Extra tree	9	3
	KNN	42	1
Distancia	Euclidiana	67	Mahalanobis
	Manhattan	8	
	Minkowski	11	

Teniendo en cuenta estas 92 configuraciones al aplicar el test de chi cuadrado se detectan 11 habitaciones significativas adicionales por lo que el número aumenta hasta 35 respecto de las 24 identificadas en la Tabla 2. Por lo tanto, el test de chi cuadrado falla si se utiliza sobre los datos iniciales ya que no tiene en cuenta un desglose por cepa entre las habitaciones.

Se utiliza el conjunto de datos de muestras adicional que se adquirieron posteriormente al inicio del proyecto para validar las configuraciones anteriores. En este conjunto de datos no hay pacientes que han estado en 12 de las habitaciones significativas (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) por lo que se pueden evaluar 23 de las 35 habitaciones significativas detectadas después de utilizar los algoritmos de clusterización.

5.4. Verificación del sistema

El resultado de la distancia de Hamming de las muestras del conjunto adicional a las configuraciones seleccionadas desprende un valor mínimo de errores de 8 y máximo 11. Por lo que respecta a los aciertos el mínimo es 12 y el máximo 16. Además, en el caso de los aciertos, cabe tener en cuenta la distancia a la que se encuentra la muestra de algún elemento del clúster, siendo la mejor puntuación 0.

No obstante, algunas habitaciones tienen muestras muy dispares y más de una cepa asociada. En estos casos, la habitación se representa en más de un clúster lo que pude producir que esté contabilizada tanto en los aciertos como en los errores a la vez dependiendo de cada muestras. Por lo tanto, se deben excluir de los aciertos todas las habitaciones que también aparecen como error para asegurarnos de que todas las habitaciones restantes en los aciertos permitan identificar de manera unívoca una muestra con una cepa. Una vez filtradas las habitaciones, el valor mínimo de aciertos es 6 y el máximo 10.

TABLA 4: HABITACIONES CON MÁS DE UNA CEPA

	Habitaciones
Urgencias	13
Planta	-
UCI	14, 15, 16
Intermedio	17
Otras	18, 19, 20, 21

Este hecho también se debe tener en cuenta a la hora de contabilizar los errores ya que las habitaciones que contienen diferentes cepas producen un error al no poder clasificar de forma clara una muestra. En la Tabla 4 se muestran hasta 9 habitaciones identificadas con más de una cepa por todas las configuraciones, por lo que no se puede determinar la cepa concreta de pseudomonas de la infección de un paciente a partir de estas habitaciones.

En el resto de habitaciones significativas no hay una única configuración que consiga identificar las cepas de cada habitación, por lo que la solución será un conjunto de configuraciones que funcionan mejor o peor en base a la habitación concreta. Para seleccionar las configuraciones finales se establecen dos criterios esenciales: que permita (a distancia 0). Con estos criterios hay 4 configuraciones (990, 1121, 1487 y 1527) que permiten identificar las cepas de hasta 9 habitaciones de forma correcta aunque coinciden en gran parte de las habitaciones. Para complementar este grupo se buscan las configuraciones que permitan identificar cepas de las habitaciones que faltan a distancia 0 el mayor número de muestras posibles. Al grupo anterior se suman 3 configuraciones nuevas (1093, 1331 y 1558) y para evitar duplicar información de las habitaciones el grupo de configuraciones finales lo componen 4 configuraciones (1093, 1121, 1331 y 1558). La Tabla 5 muestra un resumen de las habitaciones y la configuración que es capaz de identificar la cepa de pseudomonas de forma unívoca y la Tabla 6 las características generales de las configuraciones finales. Cabe destacar que la habitación *25* es el único caso en el que no se detectan dos cepas diferentes, pero la cepa se identifica con un error diferente a 0 lo cual se debe a que las muestras analizadas nunca se hayan visto antes o que la cepa provenga de otra lugar. Como dato destacable solo las configuraciones que utilizan PCA pueden identificar correctamente las cepas de la habitación *26*.

De las 24 habitaciones significativas de estar contaminadas por pseudomonas vistas en la Tabla 3, en la Tabla 4 se pueden ver 9 habitaciones en las que existe más de una cepa y no se puede discernir la cepa que causa la infección mientras que en la Tabla 5 se muestran 15 habitaciones y

TABLA 5: HABITACIONES Y CONFIGURACIONES

	Habitación	Configuración			
Urgencias	22	1121	1558	1093	1331
	23	1121		1093	
	24		1558		
	25		*1558**		
	26				1331
Planta	27	1121	1558	1093	
	28	1121			
	29	1121			1331
	30	1121	1558		
	31	1121	1558	1093	
UCI	32			1093	
Intermedio	33		1558	1093	1331
Otras	34	1121	1558	1093	1331
	35	1121	1558	1093	

* *Habitación con error diferente a 0*

la configuración que permite identificar la cepa de forma unívoca.

TABLA 6: CONFIGURACIONES FINALES

Configuracion	Algoritmo	Clústeres	Tratamiento	Distancia	PCA
1093	K-means	9	Bayesian ridge	Manhattan	Sí
1121	K-means	5	KNN	Euclidiana	NO
1331	K-means	6	KNN	Euclidiana	NO
1558	K-means	5	Media	Euclidiana	NO

Se validan los resultados obtenidos comparando para cada habitación las resistencias y los antibióticos más adecuados de las muestras de pacientes nuevos respecto del conjunto de datos original. Además, se comparan la coherencia en los grupos creados de las 4 configuraciones seleccionadas para validar que un conjunto de habitaciones agrupadas como una misma cepa se mantiene en el resto. La Tabla 7 muestra los resultados obtenidos donde se detectan 5 cepas diferentes y los antibióticos recomendados en base a la resistencia de la cepa ordenados de mejor opción a peor hasta un error máximo inferior al 5 %. Los resultados obtenidos no coincide en todos los casos con las resistencias analizadas en la Figura 1 de la base de datos lo que indica que no eran visibles o evidentes y ha sido necesario realizar la clusterización.

6 CONCLUSIONES

Se ha desarrollada una metodología para la detección de cepas de pseudomonas aeruginosa que se puede volver a procesar en una futura evolución de la bacteria.

El estadístico chi cuadrado antes de tener en cuenta los datos de la clusterización produce una subestimación de la habitaciones significativas de estar contaminadas por una cepa de pseudomonas.

TABLA 7: CEPAS POR HABITACIÓN Y ANTIBIÓTICOS

Cepa	Habitación	Antibióticos
A	22	1, 10, 2, 5, 9
	23	
	24	
	27	
	30	
	31	
	33	
	34	
	35	
B	25	10, 1, 5
C	26	1, 10, 5, 2
	29	
D	32	10, 5, 1
E	28	8, 7, 9

Se identifican un total de 35 habitaciones significativas de estar infectadas por pseudomonas de las cuales 12 han quedado excluidas del análisis al no existir muestras, en 9 se identifica más de una cepa de pseudomonas y en 14 se puede detectar la cepa de forma unívoca.

Para las 14 habitaciones en las que se identifica la cepa de pseudomonas se realiza una recomendación de los antibióticos más adecuados para el tratamiento.

No se puede determinar de una manera precisa el tratamiento más adecuado en habitaciones con más de 2 cepas de pseudomonas.

Los resultados obtenidos permiten desarrollar políticas para la asignación de habitaciones dependiendo de las características del paciente.

Relacionado con el punto anterior, un trabajo futuro del proyecto es poder predecir los pacientes que son susceptibles de ser infectados por pseudomonas aeruginosa en base a sus características y comorbilidades. Otra línea de investigación sería realizar el estudio sobre los 60, 90 o 365 días anteriores de que se detecte un paciente como infectado y comparar la evolución.

AGRADECIMIENTOS

A todas las personas que me han ayudado de diferentes formas a que este proyecto haya sido una realidad. Mención especial Gemma Sanjuan por parte del Hospital Clínic de Barcelona y al tutor del proyecto Ramón Grau por todo el tiempo y esfuerzo prestado.

REFERENCIAS

[1] Amisha, P. Malik, M. Pathania, and V. K. Rathaur, "Overview of artificial intelligence in medicine," *Journal of Family Medicine and Primary Care*, vol. 8, no. 7, pp. 2328-2331, 2019.

[2] J. A. Tenreiro Machado, "An Evolutionary Perspective of Virus Propagation", *Mathematics*, vol. 8, no. 5, pp. 779-799, 2020.

[3] K. Shailaja, B. Seetharamulu, and M.A. Jabbar, "Machine Learning in Healthcare: A Review", *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018.

[4] Y.J. Chang, M.L. Yeh, Y.C. Li et al., "Predicting Hospital-Acquired Infections by Scoring System with Simple Parameters", *PLoS ONE*, vol. 6, no. 8, 2011.

[5] P. Vanhems, A. Barrat, C. Cattuto et al., "Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors", *PLoS ONE*, vol. 8, no. 9, 2013.

[6] A. Haque, et al., "Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance", *Proceedings of the 2nd Machine Learning for Healthcare*, vol. 68, pp. 75-87, 2017.

[7] C.E. Jaramillo, "Modelización y simulación de la transmisión por contacto de una infección nosocomial en el servicio de urgencias hospitalarias", Ph.D dissertation, Departamento de Arquitectura de Computadores y Sistemas Operativos, Universitat Autònoma de Barcelona, Barcelona, june 2017.

[8] V. D. Rosenthal, et al., "International Nosocomial Infection Control Consortium (INICC) report, data summary of 45 countries for 2012-2017: Device-associated module", *American Journal of Infection Control*, vol. 48, no. 4, pp. 423–432, 01 april 2020.

[9] J. E. Bennett, R. Dolin, and M. J. Blaser, "Capítulo 221: Pseudomonas aeruginosa y otro tipo de pseudomonas", en *Mandell, Douglas y Bennett, enfermedades infecciosas: principios y práctica*. Barcelona: Elsevier España, 2015.

[10] A. Hernández, G. Yagüe, V. Garcia, et al., "Infecciones nosocomiales por Pseudomonas aeruginosa multirresistente incluido carbapenémicos: factores predictivos y pronósticos. Estudio prospectivo 2016-2017", *Rev Esp Quimioter*. vol. 31, no 2, pp. 123-30, april 2018.

[11] European Centre for Disease Prevention and Control, "Surveillance of antimicrobial resistance in Europe 2018", ECDC, November 2019.

[12] EPINE, "Prevalencia de infecciones", ESTUDIO EPINE-EPPS no. 30, 2019

[13] M. C. Fariñas and L. Martínez-Martínez, "Infecciones causadas por bacterias gramnegativas multirresistentes: enterobacterias, Pseudomonas aeruginosa, Acinetobacter baumannii y otros bacilos gramnegativos no fermentadores," *Enfermedades Infecciosas y Microbiología Clínica*, vol. 31, no. 6, pp. 402–409, 2013.

[14] L.M.Bush and M.T. Vazquez-Pertejo, "Infecciones por Pseudomonas y patógenos relacionados", Accedido: abril 2021. [Online]. Disponible:
`https://www.msdmanuals.com/es-es/professional/enfermedades-infecciosas/bacilos-gramnegativos/infecciones-por-pseudomonas-y-pat\%C3\%B3genos-relacionados`

[15] J. Macqueen, "Some methods for classification and analysis of multivariate observations", en *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967

[16] D. Arthur, and S. Vassilvitskii, "k-means++: the advantages of careful seeding" *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, page 1027-1035. Philadelphia, PA, USA, Society for Industrial and Applied Mathematics, 2007

[17] M. Kaushik and B.Mathur "Comparative Study of K-Means and Hierarchical Clustering Techniques", *International Journal of Software & Hardware Research in Engineering*, vol. 2, no. 6, pp. 93-98, 2014

[18] Von Luxburg, U. "A tutorial on spectral clustering", *Stat Comput* vol. 17, no. 4, pp. 395–416, 2007

[19] S. Misra, H. Li, and J. He, "Robust geomechanical characterization by analyzing the performance of shallow-learning regression methods using unsupervised clustering methods", *Machine Learning for Subsurface Characterization*, pp. 129–155, 2020.

[20] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977

[21] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[22] P. Taraba, "Powered Outer Probabilistic Clustering", *Proceedings of the World Congress on Engineering and Computer Science*, San Francisco, USA, October 25-27, 2017

[23] P.J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Journal of Computational and Applied Mathematics* vol. 20, pp. 53–65, 1987

[24] T. Caliński and J. Harabasz, "A Dendrite Method for Cluster Analysis", *Communications in Statistics-theory and Methods* vol. 3, pp. 1-27, 1974

[25] D.Davies and D. Bouldin, "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. PAMI-1, no. 2, pp. 224-227, 1979

APÉNDICE

A.1. Base de datos

Figuras adicionales del conjunto de de antibiogramas adicionales.

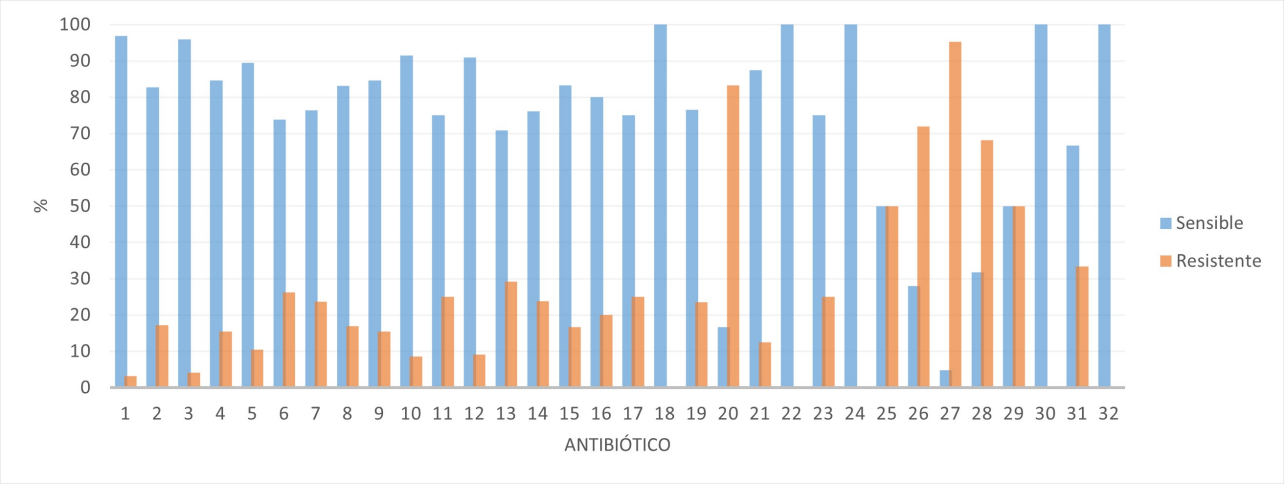

Fig. 5: Porcentaje de resistencia observada en cada antibiótico

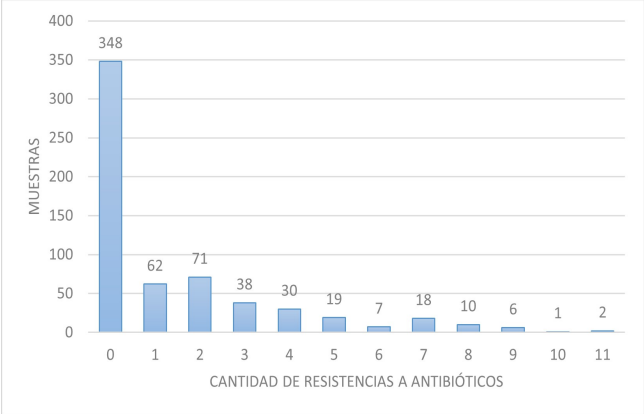

Fig. 6: Recuento de la cantidad de antibióticos resistentes por cada muestra

A.2. Medicamentos

TABLA 8: MEDICAMENTOS UTILIZADOS EN LOS ANTIBIOGRAMAS

Número de referencia	**Antibiótico**
1	Amikacina
2	Ciprofloxacina
3	Ceftolozano/tazobactam
4	Ceftazidima/avibactam
5	Gentamicina
6	Imipenem
7	Meropenem
8	Pipera/tazobactam
9	Ceftazidima
10	Tobramicina
11	Colistina
12	Ertapenem
13	Cefotaxima
14	Cotrimoxazol
15	Clindamicina
16	Eritromicina
17	Levofloxacina
18	Linezolid
19	Oxacilina
20	Penicilina
21	Rifampicina
22	Vancomicina
23	Aztreonam
24	Cefepime
25	Tetraciclina
26	Amoxicilina/clavulánico ácido
27	Ampicilina
28	Cefuroxima
29	Fosfomicina
30	Teicoplanina
31	Nitrofurantoina
32	Cefiderocol