Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Evaluación comparativa de modelos predictivos en ciencia de datos con Physarum Polycephalum como referencia biológica

Tesis presentada para obtener el grado de:

Licenciatura en Ingeniería en Ciencias de la Computación

Presenta: Raúl Eduardo Serrano Gutiérrez

Director de Tesis: Dr. María Teresa Torrijos Muñoz

Codirector: M.I. Jaime Alejandro Romero Sierra

Puebla, Pue., abril de 2025

Agradecimientos

Quiero expresar mi más profundo agradecimiento primero a padres, quienes con su amor, paciencia y sacrificio han sido el pilar fundamental en mi vida. Gracias por enseñarme el valor del esfuerzo y la perseverancia, por apoyarme en cada desafío y por confiar en mí incluso cuando yo mismo dudaba. Sin su aliento constante y su ejemplo de dedicación, este logro no sería posible.

A mi familia por fomentar valores y aconsejarme en momentos críticos de mi vida, ya que sin esas muestras de apoyo e interés genuinos no hubiese considerado otros puntos de vista que fueron fundamentales en mi toma de decisiones.

A la Benemérita Universidad Autónoma de Puebla por formarme como ingeniero y brindarme las herramientas y conocimientos necesarios para desarrollar las habilidades y destrezas que se necesitan en el ámbito profesional.

A mi directora de tesis, la Dra. María Teresa Torrijos Muñoz por su confianza, guía y enseñanzas para la realización de este proyecto. He aprendido mucho. Gracias.

A mi director de tesis, el Maestro Jaime Alejandro Romero Sierra por sus enseñanzas, paciencia y gran apoyo para desarrollar esta tesis, he aprendido y explorado mucho. Gracias.

A los creadores de contenido y divulgadores científicos que despertaron en mi la curiosidad y me regresaron la emoción que sentí al ser admitido a la universidad, a Aldo Barta que lo conocí en su canal “El robot de Platón” por su contenido que busca generar curiosidad en su comunidad y hacerlos más críticos con lo que nos rodea, a Paulina Aguilar que en redes sociales es conocida como “Polilinker” sus videos cortos que hablan sobre su carrera y experiencia como ingeniera en Biotecnología de forma fácil y divertida, hicieron que me interesara en este grupo de moho y querer adaptarlo a esta carrera.

Resumen

El objetivo de este estudio es modelar el crecimiento y el proceso de toma de decisiones de Physarum polycephalum a través de técnicas y herramientas propias de la Ciencia de Datos, con el fin de comparar su eficiencia en la creación y optimización de redes frente a modelos tradicionales de optimización, para poder comprender el comportamiento del moho mucilaginoso Physarum polycephalum se usaran datos obtenidos ante cuatro estímulos químicos diferentes (Control, Ácido, Cafeína y Quinina). Para poder realizar este análisis, se recopilaron variables fundamentales como el Contacto, el Tiempo_de_cruce, la Arborización y el Área, provenientes de una investigación llevada a cabo en 2016. Se utilizaron métodos de agrupación como el K-Means y de clasificación supervisada como la Regresión Logística.

Los descubrimientos señalan que las variables Contacto y Tiempo_de_cruce son los factores más determinantes para segmentar y prever el tipo de tratamiento, demostrando así un rendimiento particularmente ventajoso para Quinina y Cafeína. Sin embargo, también surgieron dificultades para distinguir entre las variables Control y Ácido, lo que indica similitudes entre la reacción del moho bajo estas circunstancias y la posibilidad de agregar variables extra o modelos más avanzados para comprender mejor el comportamiento del moho.

Para terminar, la combinación de análisis no supervisado y supervisado demostró la capacidad de Physarum polycephalum para mostrar patrones reconocibles en su conducta. Este estudio no solo contribuye a la comprensión de organismos no neuronales, sino que también propone posibles usos en la inspiración biológica de sistemas de optimización y redes adaptativas.

Índice

Contenido

Introducción

Esta investigación del moho mucilaginoso conocido como Physarum polycephalum ha cobrado importancia en varios ámbitos, desde la biología básica hasta la investigación de ciencias aplicadas como biotecnología. Este ser unicelular, a pesar de que no tiene un sistema nervioso, demostró comportamientos complejos como la habituación, la optimización de rutas y patrones de crecimiento distintivos que se esperaría ver únicamente en seres vivos pluricelulares que cuenten con un sistema nervioso. Por esta razón, se ha propuesto la hipótesis de que el moho reacciona de forma única frente a diferentes estímulos químicos —Control (C), Ácido (CA), Cafeína (CC) y Quinina (Q)—, evidenciados en medidas como la cantidad de contactos, el tiempo requerido para cruzar un puente experimental y la zona o nivel de arborización durante su crecimiento.

El objetivo de este estudio es comprender el comportamiento del hongo mucilaginoso Physarum polycephalum ante los estímulos químicos de control, ácido, cafeína y quinina esto mediante técnicas de ciencia de datos. Se decidió implementar una táctica dual:

1. Clustering (K-Means), con el propósito de analizar la capacidad de agrupar naturalmente observaciones y establecer qué variables tienen un mayor efecto en la creación de patrones.

2. Clasificación (Regresión Logística), con el fin de evaluar la habilidad para anticipar el tratamiento en función de las variables adquiridas, confirmando la relevancia de Contacto, Tiempo_de_cruce, Arborización y Área.

Se anticipa que los descubrimientos enriquezcan tanto el entendimiento del aprendizaje no neuronal como la creación de soluciones bioinspiradas, ya que el moho proporciona tácticas de adaptación que podrían ser útiles en campos como la optimización de redes y la administración de recursos.

Para comprender el problema, este documento esta organizado en diversos capítulos usando la metodología para ciencia de datos propuesta por IBM comenzando desde la recopilación y limpieza de datos, el uso de K-Means y Regresión Logística, hasta el debate de los resultados y las conclusiones que resumen la contribución de la investigación.

Antecedentes

El moho mucilaginoso Physarum Polycephalum ha sido objeto de estudio desde hace varias décadas por su capacidad para mostrar comportamientos complejos a pesar de carecer de un sistema nervioso. Sus primeras observaciones científicas se remontan desde el año 1973 aunque se usaba para el estudio de las estructuras de las células procariotas, sin embargo, se pudo observar su habilidad para crecer y moverse de manera aparentemente “inteligente” sobre superficies húmedas. A lo largo del siglo XX, diversos investigadores documentaron su habilidad para navegar laberintos, encontrando rutas relativamente cortas para acceder a fuentes de alimento.

En los últimos años, estos hallazgos han tomado un giro más preciso y cuantitativo. Tero et al. (2010) demostraron la capacidad del moho para diseñar redes de transporte eficientes, al compararlas con sistemas ferroviarios y carreteras. Según sus experimentos, Physarum Polycephalum reorganiza su red protoplásmica para conectar diversos puntos de nutrientes de la forma más efectiva, minimizando la longitud total de las "vías" y reduciendo los costos de transporte, a pesar de tener diferencias con la red ferroviaria esta era muy similar y las únicas diferencias fueron geográficas que no se pudieron replicar como montañas y lagos. Este comportamiento, interpretado como un "algoritmo biológico" de optimización, dio origen a un campo de estudio que combina la Biología con la ciencia de datos y la computación bioinspirada.

También la investigación de Boisseau et al. (2016) expuso otro fenómeno de gran importancia: la habituación en organismos no neuronales. Su trabajo evidenció que el moho, al ser expuesto repetidamente a un estímulo químico como cafeína o quinina, reduce gradualmente su respuesta de aversión. Este descubrimiento es crucial, pues desvela que el aprendizaje no es exclusivo de organismos con cerebro y sugiere la existencia de mecanismos de memoria y adaptación en entidades unicelulares. Tal investigación ha llevado a la comunidad científica a plantear preguntas acerca de cómo se originó el aprendizaje a lo largo de la evolución y cómo este comportamiento puede modelarse matemáticamente.

Sumado a ello, se han ido generando bases de datos experimentales que recogen variables como el Contacto (número de veces que el moho interactúa con el puente), el Tiempo de cruce (segundos que tarda en atravesar una superficie con estímulo o sin él), el grado de Arborización y el Área cubierta a lo largo de su crecimiento. Dichas variables permiten una aproximación cuantitativa al estudio de la conducta de P. polycephalum. De hecho, con la llegada de técnicas de aprendizaje automático —tales como K-Means (para segmentación de patrones) y Regresión Logística (para clasificación supervisada)— se ha abierto la posibilidad de analizar estadísticamente los datos y hacer predicciones sobre el tratamiento aplicado o la respuesta del moho ante estímulos químicos.

A pesar de la creciente literatura, persisten interrogantes en torno a la eficacia comparativa de distintos modelos predictivos. como K-Means, Regresión Logística o incluso Árboles de Decisión para clasificar o agrupar comportamientos de Physarum polycephalum. Asimismo, se desconoce cómo interactúan variables como el Contacto, el Tiempo de cruce, la Arborización o el Área para discriminar efectivamente entre los tratamientos de Control (C), Ácido (CA), Cafeína (CC) y Quinina (Q). Estos puntos plantean la necesidad de profundizar en la analítica de datos del moho a fin de ampliar tanto el conocimiento biológico como las aplicaciones bioinspiradas.

Sumado a ello, se han ido generando bases de datos experimentales que recogen variables como el Contacto (número de veces que el moho interactúa con el puente), el Tiempo de cruce (segundos que tarda en atravesar una superficie con estímulo o sin él), el grado de Arborización y el Área cubierta a lo largo de su crecimiento. Dichas variables permiten una aproximación cuantitativa al estudio de la conducta de **P. polycephalum**.

Planteamiento del problema

A pesar de que se han realizado estudios sobre el comportamiento y la capacidad de aprendizaje de Physarum polycephalum para optimizar rutas, como resolver laberintos o demostrar su capacidad de aprendizaje sin contar con un sistema nervioso, no se dispone de un análisis cuantitativo que determine su comportamiento para ser utilizado para algoritmos de ciencia de datos. Para esta investigación se usará la base de datos obtenida en la investigación Boisseau et al. (2016) de las cuáles son las variables más influyentes en la diferenciación de sus respuestas bajo diferentes tratamientos químicos (Control, Ácido, Cafeína y Quinina). Asimismo, se desconoce hasta qué punto es factible clasificar o agrupar correctamente las observaciones experimentales de P. polycephalum registradas en grandes bases de datos utilizando enfoques de ciencia de datos.

Por una parte, los modelos de aprendizaje no supervisado, como K-Means, permiten segmentar los datos en clusters que podrían correlacionarse con cada tratamiento o manifestaciones conductuales. Sin embargo, aún no está claro si estos clusters coinciden o no con las categorías de tratamiento y cómo se podrían interpretar biológicamente. Por otra parte, los modelos de clasificación supervisada, como la Regresión Logística, ofrecen la posibilidad de predecir el tratamiento a partir de variables observadas (Contacto, Tiempo de cruce, Arborización, Área). Sin embargo, los hallazgos iniciales señalan que algunas clases, en particular el Tratamiento "Control" (C) y "Ácido" (CA), resultan complicadas de diferenciar.

Por tanto, la pregunta central que motiva este estudio es:

¿Qué tan eficientes y precisos son los distintos modelos predictivos K-Means y Regresión Logística para agrupar y clasificar el comportamiento de Physarum polycephalum bajo diferentes tratamientos químicos, considerando un conjunto de variables cuantitativas?

Para responder a esta interrogante, se requiere:

1. Una evaluación minuciosa de la base de datos obtenida del estudio de Boisseau et al. (2016), comprobando la distribución, la existencia de indicadores de desviación y las correlaciones entre las variables.
2. La aplicación de métodos de agrupación (K-Means) y de clasificación (Regresión Logística) facilita la comparación de la exactitud y solidez de cada técnica.

3. La valoración de la capacidad de interpretación y la relevancia de estos modelos para entender los procesos de habituación y respuesta adaptativa en el moho.

En consecuencia, el problema radica en la falta de un estudio comparativo que analice la eficacia y relevancia de diferentes modelos de ciencia de datos, aprovechando las múltiples dimensiones (Contacto, Tiempo de cruce, Arborización, Área) que caracterizan el comportamiento de Physarum polycephalum y que podrían ayudar a diferenciar los tratamientos químicos utilizados en los experimentos.

Justificación

La presente investigación se justifica por su doble aporte, tanto en el ámbito biológico como en el tecnológico:

Relevancia Teórica Biológica

1. Desde la perspectiva científica, comprender cómo Physarum polycephalum procesa estímulos repetidos y se habitúa (sin poseer un sistema nervioso) amplía la visión sobre el origen del aprendizaje en la evolución. Las conclusiones que resulten de este análisis podrían respaldar la hipótesis de que la habituación es un proceso altamente conservado y podría haber surgido antes de los sistemas neurales complejos.

2. Asimismo, la comparación entre distintos modelos de ciencia de datos permite un avance en la metodología de análisis de organismos no convencionales, ofreciendo un enfoque que pocas veces se ha visto aplicado en este tipo de estudios biológicos.

Relevancia en el ámbito de las Ciencias Computacionales

1. En el área de computación bioinspirada, la forma en que Physarum polycephalum optimiza redes y se adapta a estímulos podría trasladarse a soluciones de optimización de rutas y gestión de recursos. Contar con un modelo predictivo robusto para clasificar diferentes estados del moho facilita la detección temprana de comportamientos anómalos o la adaptación a nuevos estímulos.

2. Para los investigadores que buscan desarrollar algoritmos de aprendizaje basados en la conducta de este organismo, identificar con precisión las variables más discriminantes (por ejemplo, Contacto y Tiempo de cruce) es fundamental para reproducir o simular dichas conductas en entornos computacionales.

Viabilidad y Contribución

1. Se cuenta con datos experimentales de alta calidad (Boisseau et al., 2016), lo que brinda la viabilidad necesaria para aplicar herramientas estadísticas y modelos predictivos de manera rigorosa. Esta base de datos incluye suficientes observaciones y variables para garantizar un análisis representativo y confiable.

2. El proyecto contribuirá a establecer o refinar metodologías estándar para explorar, clasificar y analizar la conducta de organismos no neuronales. Además, servirá de punto de partida para investigaciones posteriores que busquen integrar otros enfoques, como técnicas de Deep Learning o algoritmos genéticos, en la predicción del comportamiento de Physarum polycephalum u otros organismos similares.

En síntesis, esta investigación no solo tiene relevancia académica por fortalecer el entendimiento de la habituación en organismos unicelulares, sino que también ofrece un potencial práctico a futuro en campos como la optimización de redes de transporte, la

planeación urbana y el desarrollo de modelos bioinspirados en el ámbito de las Ciencias Computacionales. El papel pionero que podría desempeñar Physarum polycephalum como plataforma experimental para el estudio del aprendizaje no neuronal y de la optimización en entornos complejos permite justificar la realización de este proyecto.

Objetivos General y Específicos

General

Modelar el crecimiento y el proceso de toma de decisiones de Physarum polycephalum a través de técnicas y herramientas propias de la Ciencia de Datos, con el fin de comparar su eficiencia en la creación y optimización de redes frente a modelos tradicionales de optimización. Este enfoque busca identificar las ventajas y desventajas de las redes adaptativas inspiradas en el comportamiento biológico del moho, en contraste con los algoritmos y modelos matemáticos ampliamente empleados en el campo de la optimización. Los hallazgos proporcionarán una base sólida para proponer mejoras en el diseño de infraestructuras y sistemas complejos (p. ej., transporte, comunicaciones, logística), al incorporar principios de adaptabilidad y autoorganización inspirados en la naturaleza.

Específicos

1. Analizar cuantitativamente el patrón de crecimiento y toma de decisiones del moho
    a. Describir y medir las principales variables que influyen en la forma en que Physarum polycephalum se expande y construye sus redes (p. ej., contacto, tiempo de cruce, arborización, área cubierta).
    b. Evaluar la repercusión de distintos tratamientos químicos (ácido, cafeína, quinina, control) en la configuración final de las redes biológicas.

2. Comparar la eficiencia y robustez de los algoritmos de ciencia de datos con el comportamiento emergente de P. polycephalum
    a. Aplicar enfoques como K-Means, Regresión Logística y Árboles de Decisión para identificar patrones y predecir el tratamiento o la configuración de red en el organismo.
    b. Determinar, mediante métricas de rendimiento (coeficiente de silueta, F1-score, exactitud, etc.), hasta qué punto las soluciones propuestas por los algoritmos se asemejan o superan las estrategias naturales del moho.

3. Diseñar e implementar una metodología de modelado bioinspirado
    a. Desarrollar un procedimiento sistemático para capturar el comportamiento de P. polycephalum en escenarios de optimización de redes, incorporando el proceso de habituación y las respuestas adaptativas del moho.
    b. Integrar herramientas de análisis estadístico y aprendizaje automático (visualización de datos, reducción de dimensionalidad, etc.) para explicar y reproducir la dinámica de crecimiento del moho.

4. Explorar la aplicabilidad de los principios de P. polycephalum en la mejora de redes tecnológicas

a. Identificar áreas (infraestructura de transporte, topologías de redes de comunicación, logística de distribución) donde la adaptabilidad y capacidad de autoorganización del moho puedan ofrecer beneficios significativos frente a los métodos tradicionales.
b. Proponer lineamientos o prototipos de prueba que transfieran el comportamiento observado en P. polycephalum a entornos computacionales y/o reales, validando si se obtiene un aumento en la eficiencia de las redes o sistemas complejos.

5. Generar nuevas perspectivas y recomendaciones para el diseño y la gestión de sistemas complejos
    a. Sintetizar los resultados en forma de lineamientos, marcos o guías de implementación que integren la visión bioinspirada del moho con estrategias de optimización consolidadas.
    b. Establecer oportunidades de investigación futura, como la combinación de enfoques bioinspirados con métodos de aprendizaje profundo u otras metodologías de vanguardia en la ciencia de datos.

Marco teórico

El aprendizaje y su relevancia en organismos no neuronales

El aprendizaje se define como un cambio en el comportamiento generado por la experiencia, y se ha asociado históricamente a la presencia de un sistema nervioso (Boisseau et al., 2016). No obstante, este concepto, sin embargo, ha sido objeto de revisión a partir de evidencias que señalan la existencia de formas primitivas de aprendizaje en organismos sencillos —incluyendo algunos unicelulares— antes de la aparición de sistemas nerviosos complejos.
Estudios recientes han demostrado que el aprendizaje no es exclusiva de animales con cerebro; por ejemplo, se ha observado que ciertas bacterias y protistas son capaces de adaptar su comportamiento o anticipar cambios en su entorno, lo cual podría considerarse una forma elemental de memoria (Ginsburg & Jablonka, 2010, citado en Boisseau et al., 2016). Desde esta perspectiva, el hallazgo de mecanismos de habituación en organismos unicelulares resulta particularmente interesante, pues abre la posibilidad de que el aprendizaje haya evolucionado de manera independiente y se presente en una amplia variedad de linajes en el árbol de la vida (Boisseau et al., 2016).

Physarum polycephalum: un modelo de estudio para la bioinspiración

El moho mucilaginoso *Physarum polycephalum* es un organismo unicelular multinucleado que pertenece al grupo de los mixomicetos. Presenta características de gran interés para la comunidad científica debido a su comportamiento emergente y altamente adaptable, ya que al ser una masa de protoplasma este se desliza en búsqueda de alimento mediante tubos llamados plasmodios, que son los encargados de explorar los alrededores en búsqueda de alimento, el cual sin contar con un sistema nervioso o un cerebro, es capaz de analizar, recordar y clasificar el alimento encontrado para así tomar decisiones a base de lo recolectado, también exhibe navegación compleja, resolución de laberintos y optimización

de rutas de manera aparentemente "inteligente" (Tero et al., 2010). Tradicionalmente, *P. polycephalum* habita en ambientes húmedos y sombreados, donde se alimenta de materia orgánica en descomposición.

Capacidad de aprendizaje en *Physarum polycephalum*

Aunque *P. polycephalum* carece de neuronas, se ha observado que puede memorizar ciertas condiciones ambientales y responder de forma diferenciada a estímulos repetidos. Boisseau et al. (2016) demostraron experimentalmente la aparición de habituación en *P. polycephalum* utilizando sustancias químicas como quinina y cafeína, típicamente repelentes. El experimento reveló que, tras exposiciones sucesivas, el moho reducía gradualmente su aversión al estímulo; no obstante, al retirar la sustancia durante un intervalo de tiempo y volverla a presentar, la respuesta aversiva reaparecía.
Este fenómeno cumple con los criterios de habituación:

1. **Disminución de la respuesta** tras presentaciones repetidas de un estímulo inofensivo.
2. **Recuperación espontánea** de la respuesta cuando el estímulo deja de presentarse y se reintroduce más tarde.
3. **Especificidad del estímulo**, ya que la habituación al estímulo A (por ejemplo, quinina) no se transfiere automáticamente a un estímulo B (cafeína), descartando así fatiga sensorial o motora (Boisseau et al., 2016).

Optimización y diseño de redes en *Physarum polycephalum*

Más allá de su capacidad de habituación, *Physarum polycephalum* ha inspirado diversas investigaciones por su habilidad para diseñar redes de transporte eficientes. Tero et al. (2010) describen cómo este organismo puede optimizar su protoplasma a fin de conectar fuentes de nutrientes reduciendo costes de transporte y maximizando la distribución de recursos. En sus experimentos, *P. polycephalum* fue capaz de formar redes jerárquicas, comparables en eficiencia a las redes ferroviarias o de carreteras diseñadas por el ser humano.
El éxito de *P. polycephalum* radica en la actualización continua de su estructura tubular a través de quimiotaxis y retroalimentación local, lo cual favorece rutas cortas y una alta redundancia de caminos. A partir de este comportamiento, Tero et al. (2010) proponen reglas de actualización matemáticas y computacionales que han derivado en algoritmos bioinspirados para problemas de optimización de grafos.

De la biología a la computación bioinspirada

Los resultados de Tero et al. (2010) han motivado la exploración de modelos computacionales que simulan la lógica de crecimiento de *P. polycephalum* para aplicaciones que van desde el diseño de rutas hasta la planificación de infraestructuras. Por un lado, se pueden comparar estos algoritmos bioinspirados con enfoques computacionales establecidos (por ejemplo, algoritmos genéticos o heurísticas clásicas para optimización de redes) para evaluar la eficiencia y robustez en distintos escenarios.
Por otro lado, las investigaciones de Boisseau et al. (2016) aportan la dimensión del aprendizaje en organismos no neuronales, de forma que las redes basadas en *P. polycephalum* no solo pueden optimizar rutas, sino también ajustar su comportamiento ante estímulos

repetidos. Esto abre la puerta a sistemas inteligentes capaces de habituarse a patrones de tráfico, costos energéticos o requerimientos de tiempo, incorporando mecanismos similares a la habituación.

Perspectiva integral de las redes bioinspiradas en *Physarum polycephalum*

En conjunto, las aportaciones de Boisseau et al. (2016) y Tero et al. (2010) sugieren que *Physarum polycephalum* es un modelo de estudio idóneo para comprender cómo un organismo aparentemente "simple" puede exhibir conductas adaptativas sofisticadas. Mientras que el primer trabajo se centra en la habituación y la capacidad de "aprender" sin sinapsis ni neuronas, el segundo enfatiza los procesos de optimización de redes y la posibilidad de trasladar ese comportamiento a sistemas artificiales.

En este estudio, se emplearán los siguientes algoritmos para representar enfoques tradicionales en la creación y optimización de redes:

1. K-Means:
   Un algoritmo de agrupamiento no supervisado que particiona un conjunto de datos en k grupos basándose en la similitud de sus características. En el contexto de redes, k-means se utiliza para identificar clústeres de nodos o puntos clave que puedan optimizar la conectividad general del sistema.

2. Regresión Logística: Una técnica de clasificación supervisada que modela la probabilidad de que un evento ocurra en función de uno o más predictores. En este proyecto, la regresión logística se utilizará para predecir patrones de conectividad en redes basándose en características observadas, permitiendo analizar la probabilidad de formación de enlaces entre puntos clave

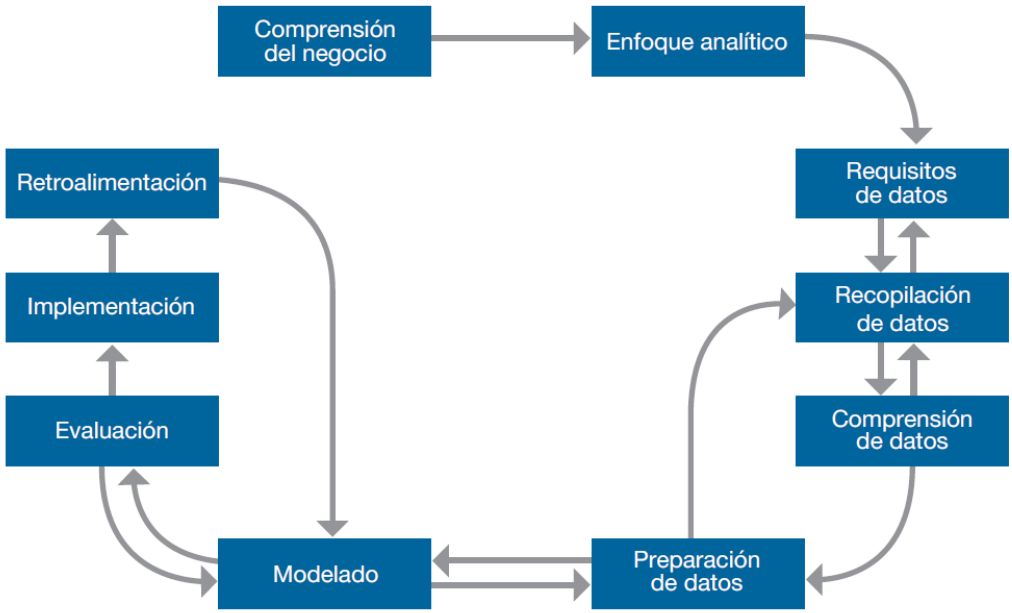
Estos algoritmos proporcionan una base sólida para comparar el comportamiento emergente de *Physarum polycephalum* con técnicas matemáticas y computacionales consolidadas. Al hacerlo, se espera identificar fortalezas y limitaciones de las redes bioinspiradas frente a los enfoques tradicionales.

En los siguientes apartados se describirán los métodos empleados para evaluar y contrastar estos diferentes enfoques de optimización y aprendizaje, estableciendo así una base para el desarrollo de soluciones basadas en la naturaleza de *P. polycephalum*.

Metodología

El desarrollo del proyecto se basa en la Metodología Fundamental para la Ciencia de Datos propuesta por John B. Rollis científico de datos .

Como se muestra en la imagen, esta metodología considera 10 etapas que se describen a continuación.



Etapa 1: Comprensión del negocio

Definición del problema de investigación

El proyecto se centra en analizar cómo Physarum polycephalum responde a diferentes tratamientos químicos (por ejemplo, ácido, cafeína, quinina, control) realizados en una investigación de Boisseau et al. (2016) y modelar su comportamiento para comparar su eficiencia en la creación de redes con modelos de optimización tradicionales. El problema radica en entender qué variables (Contacto, Tiempo_de_cruce, Arborización, Área, entre otras) influyen más en la formación de estas redes y en la adaptación del moho frente a estímulos repetidos.

Objetivos principales

- Identificar patrones de habituación o respuesta conductual frente a estímulos químicos.
- Construir modelos predictivos (regresión logística) y descriptivos (agrupamiento K-Means) basados en las variables recogidas de P. polycephalum.
- Evaluar cuáles variables son más influyentes en la respuesta del moho, determinando su relevancia para la clasificación o para la optimización de redes.

Participación del dominio

- Para validar hipótesis sobre el comportamiento del moho y la pertinencia de cada variable, resulta esencial contar con la retroalimentación de expertos en biología, quienes brindarán perspectiva sobre la habituación y el ciclo de vida de Physarum polycephalum, aunque por el momento solo se usarán los datos obtenidos de las

aportaciones de Boisseau et al. (2016), se espera que en futuras investigaciones se pueda obtener el moho para experimentar con diferentes parámetros.
- Asimismo, se busca que más científicos de datos contribuyan a definir el enfoque de modelado, también a evaluar la calidad de los datos y a proponer mejoras o ajustes en la metodología empleada para obtener modelos más eficientes.

Etapa 2: Enfoque analítico

En esta etapa, se traduce el problema de investigación sobre el comportamiento de Physarum polycephalum al contexto de técnicas analíticas propias de la ciencia de datos, con el fin de abordar los objetivos desde una perspectiva estadística y computacional.

Objetivo analítico

- Determinación del tratamiento al que fue sometido el moho (Control, Ácido, Cafeína, Quinina) y análisis de su respuesta a lo largo del tiempo, en función de variables como el Contacto, el Tiempo de cruce, la Arborización y el Área.
- Este objetivo se plantea como un problema de clasificación, donde la variable objetivo es el tipo de tratamiento, y las variables predictoras son las mediciones tomadas del moho.

Enfoque metodológico

Para cubrir los distintos ángulos del problema, se proponen las siguientes estrategias analíticas:

Algoritmos de agrupamiento (K-Means)

- Identificar patrones de agrupamiento en las respuestas del moho sin usar etiquetas predeterminadas (aprendizaje no supervisado).
- Mediante K-Means, se busca descubrir clusters en los datos que pudieran relacionarse con ciertos tratamientos o comportamientos emergentes del moho. Esta aproximación permite observar si el moho tiende a "auto agruparse" en patrones que correspondan o no a los tratamientos químicos aplicados.

Modelos predictivos (Regresión Logística)

- Propósito: Construir modelos supervisados que predigan el tipo de tratamiento (variable categórica: C, CA, CC, Q) a partir de los valores de las variables observadas (Contacto, Tiempo_de_cruce, Arborización, Área, etc.).
- Aplicación:
    1. Regresión Logística: Modelo estadístico que estima la probabilidad de cada clase (tratamiento) al analizar el efecto de cada variable explicativa. Se evaluarán métricas como la exactitud, precisión y F1-score para comparar el desempeño.

Comparación de enfoques tradicionales con los bioinspirados

- El moho Physarum polycephalum a menudo se ha comparado con algoritmos de optimización (p. ej., Dijkstra para rutas más cortas, algoritmos genéticos para diseño de redes) por su habilidad de autoorganizarse.
- Propósito: Confrontar los métodos tradicionales (por ejemplo, se podría tomar un algoritmo de optimización estándar para comprobar rutas eficientes) con la propuesta bioinspirada (comportamiento real del moho) y los modelos estadísticos (K-Means, Regresión Logística). Se observará si el moho "supera" o iguala los resultados de los algoritmos clásicos en la formación de redes o si existen escenarios donde la estrategia biológica resulta más o menos eficiente.

Justificación del enfoque

- El uso de K-Means permite descubrir patrones de comportamiento sin imponer etiquetas, explorando la posibilidad de que el moho genere ciertos "perfiles de respuesta" de forma natural.
- La Regresión Logística abordan directamente el objetivo de clasificación (qué tratamiento fue aplicado) y permiten medir el impacto relativo de cada variable (Contacto, Tiempo_de_cruce, etc.).
- Al conjuntar aprendizaje no supervisado (descubrimiento de patrones) y aprendizaje supervisado (predicción y clasificación), se obtendrá una visión integral del comportamiento de Physarum polycephalum.

Resultados esperados

- Identificar grupos internos (clusters) que reflejen diferentes niveles de habituación o eficiencia en la formación de redes, los cuales puedan correlacionarse con tratamientos específicos.
- Determinar con qué precisión se pueden clasificar las muestras según el tratamiento, observando si las variables escogidas son suficiente para distinguir, por ejemplo, entre el moho expuesto a cafeína (CC) y el control (C).
- Contribuir a la comparación de metodologías, mostrando en qué medida los enfoques tradicionales de optimización difieren o coinciden con el diseño de redes del moho, y cómo los modelos estadísticos capturan dicha dinámica biológica.

Etapa 3: Requisitos de datos

Para esta etapa se busca identificar y documentar los datos necesarios para cumplir con los objetivos y el enfoque analítico definido en el proyecto. Describiremos la fuente de datos, el formato y los atributos clave del conjunto obtenido.

Descripción rápida de la estructura

Columna	Descripción
Treatment	Tipo de tratamiento (C, CA, CC, Q)
Day	Día del experimento (1 a 9)
Bridge	Composición del puente (A, Q, CAF)

Contact	Número de contactos del moho con el puente
Crossing_time	Tiempo de cruce (en segundos)
Arborisation	Grado de ramificación (valor entre 0 y 1 aprox.)
area	Área cubierta por el moho

Fuente de datos

El conjunto de datos proviene del artículo:
Boisseau, R. P., Vogel, D., & Dussutour, A. (2016). Habituation in non-neural organisms: Evidence from slime moulds. Proceedings of the Royal Society B: Biological Sciences, 283(1829), 20160446.
https://doi.org/10.1098/rspb.2016.0446

Descripción de las variables incluidas:

- Treatment: Tipo de tratamiento aplicado (C: Control, CA: Ácido, CC: Cafeína, Q: Quinina).
- Day: Día del experimento (valores entre 1 y 9).
- Bridge: Composición del puente experimental (A: Ácido, Q: Quinina, CAF: Cafeína).
- Contact: Número de contactos del moho con el puente.
- Crossing_time: Tiempo de cruce del puente (en segundos)
- Arborisation: Grado de ramificación del moho (valor entre 0 y 1).area: Área cubierta por el moho durante el experimento (unidades cuadradas).

Descripción estadística

Los datos utilizados provienen de un estudio realizado en el año 2016, en el cual se experimentó con cuatro sustancias (Control, Ácido, Quinina y Cafeína). Mientras que la quinina y el ácido ya habían sido evaluados previamente en investigaciones de quimiotaxis con Physarum polycephalum, la cafeína no contaba con antecedentes claros. Por este motivo, se procedió a ensayar diferentes concentraciones (1, 2, 3 y 4 mM), descubriéndose que concentraciones superiores a 2 mM resultaban perjudiciales, ocasionando extrusiones de citoplasma e impidiendo al moho cruzar el puente experimental.
Esta variabilidad en la concentración de cafeína refleja la dificultad inicial para determinar la dosis adecuada, evidenciando que el diseño experimental involucró un proceso de ajuste "al tanteo" hasta establecer los rangos seguros para el moho.

Descripción técnica de los Datos:

1. Validez de Datos:
    a. Los valores deben estan dentro de los rangos esperados:
        i. Contact: Valores positivos sin inconsistencias extremas.
        ii. Crossing_time: Tiempo en segundos mayor a 0.
        iii. Arborisation: Escala numérica entre 0 y 1.
    b. Las categorías de Treatment y Bridge deben estan correctamente etiquetadas.
2. Formato:

   a. Las variables numéricas deben estar en un formato compatible con el análisis estadístico.
   b. Las variables categóricas (Treatment, Bridge) deben estar codificadas en formato adecuado.
3. Integridad de Datos:
   a. Verificar si hay valores faltantes en alguna columna y tratar los datos según sea necesario (relleno, eliminación o imputación).
   b. Detectar y manejar valores atípicos mediante análisis exploratorio.

Etapa 4: Recopilación de datos

En esta etapa, el objetivo principal es **identificar y reunir** todos los recursos de datos relevantes para el estudio de *Physarum polycephalum* bajo diferentes estímulos químicos. La información proviene, esencialmente, de la investigación realizada en 2016, donde se emplearon cuatro tratamientos (Control, Ácido, Quinina y Cafeína) y se midieron variables clave como el Contacto, el Tiempo_de_cruce y el Área cubierta por el moho. El enfoque aquí se centra en **asegurar** que los datos estén **completos** y **representen** adecuadamente la variabilidad del comportamiento del moho, considerando posibles nuevas fuentes o registros adicionales si se detectan lagunas en la información. Además, se evalúa la pertinencia de obtener nuevas mediciones de concentraciones de cafeína o de probar distintas intensidades en los tratamientos, con el fin de optimizar la cobertura de datos y enriquecer el posterior análisis exploratorio y modelado.

Descripción de las Columnas del DataFrame

A continuación, se detalla el significadoy contenido de cada columna:

1. **Tratamiento (Treatment)**
   Indica el tipo de tratamiento experimental aplicado al moho mucilaginoso (Physarum polycephalum) para estudiar su habituación a diferentes estímulos químicos. Los valores son:
   - C: Control. Grupo no expuesto a estímulos específicos, utilizado como referencia.
   - CA: Ácido. Exposición a una solución ácida.
   - CC: Cafeína. Exposición a cafeína.
   - Q: Quinina. Exposición a quinina, una sustancia amarga.
2. **Día (Day)**
   Representa el día específico dentro del ciclo experimental, utilizado para realizar un seguimiento temporal de las respuestas del moho.

3. **Puente (Bridge)**
   Describe la composición del puente utilizado para que el moho se desplace.

   Los valores son:
   - A: Ácido. El puente estaba tratado con una solución ácida.
   - Q: Quinina. El puente contenía quinina.
   - CAF: Cafeína. El puente estaba impregnado con cafeína.

4. **Contacto (Contact)**
   Mide si el moho estableció contacto con el puente, lo que indica su disposición a explorar o evitar determinados estímulos.

5. **Tiempo_de_cruce (Crossing_time)**
   Tiempo que tarda el moho en cruzar el puente experimental. Es un indicador de su comportamiento locomotor frente a estímulos químicos.

6. **Arborización (Arborisation)**
   Se refiere al grado de expansión o ramificación del moho en el entorno experimental.

7. **Área (area)**
   Tamaño del espacio cubierto por el moho durante el experimento. Es una medida de su crecimiento o expansión bajo diferentes condiciones.

Análisis Exploratorio de los Datos

Para comprender mejor la estructura y la distribución de las variables registradas, se realizó un análisis descriptivo mediante el método df.describe():

	DÍA	CONTACTO	TIEMPO_DE_CRUCE	ARBORIZACIÓN	ÁREA
COUNT	3662.0	3662.0	3662.0	3662.0	3662.0
MEAN	5.0	80.13	101.72	0.56	91.51
STD	2.59	80.05	73.30	0.19	26.94
MIN	1.0	5.0	20.0	0.09	8.50
25%	3.0	25.0	60.0	0.43	73.10
50%	5.0	60.0	80.0	0.59	92.02
75%	7.0	105.0	115.0	0.71	110.50
MAX	9.0	961.0	945.0	0.95	167.88

Observaciones Clave

1. Día (Day)

- **Media**: 5.0 días, indicando que el experimento se centró alrededor de este punto.
- **Rango**: De 1 a 9 días, con una desviación estándar de 2.59 días.
- **Percentiles**:
  - 25% de los experimentos concluyeron antes del día 3.
  - 50% (mediana) se realizó en 5 días.
  - 75% antes del día 7.

En 9 días de observación, el moho fue expuesto gradualmente a los tratamientos. La distribución relativamente uniforme sugiere que no hubo concentraciones excesivas en una sola fase temporal.

2. Contacto (Contact)

- **Media**: 80.13 contactos.

- **Rango**: De 5 a 961 contactos.
- **Desviación Estándar**: 80.05, indicando una **alta dispersión**.
- **Percentiles**:
    - 25% tienen menos de 25 contactos.
    - 50% alcanzan 60 contactos.
    - 75% superan los 105 contactos.

El número de contactos puede variar drásticamente, sugiriendo que en algunos casos el moho exploró intensamente (más de 900 contactos), mientras que en otros apenas interaccionó con el puente (5 contactos). Esta variabilidad se ha asociado con la repulsión o atracción frente a sustancias como la quinina o la cafeína en distintas concentraciones.

3. Tiempo_de_cruce (Crossing_time)

- **Media**: 101.72 segundos.
- **Rango**: De 20 a 945 segundos.
- **Desviación Estándar**: 73.30, reflejando un **alto grado de dispersión**.
- **Percentiles**:
    - 25% de los experimentos, el moho cruzó en menos de 60 segundos.
    - El 50% (mediana) fue de 80 segundos.
    - El 75% por debajo de 115 segundos.

Existen valores elevados (945 segundos) que indican casos donde el moho tardó considerablemente en atravesar el puente, posiblemente por aversión a altos niveles de cafeína en las primeras etapas del experimento o a la quinina, reconocida como un repelente potente.

4. Arborización (Arborisation)

- **Media**: 0.56, indicando un **nivel moderado** de expansión.
- **Rango**: De 0.09 a 0.95, con una desviación estándar de 0.19.
- **Percentiles**:
    - 25% < 0.43.
    - 50% (mediana) = 0.59.
    - 75% > 0.71.

La arborización mide la ramificación del moho. Un valor cercano a 0 indica poca ramificación, mientras que valores cercanos a 1 señalan una alta dispersión celular. El intervalo amplio (0.09–0.95) apunta a que el moho presenta conductas muy distintas en su crecimiento, seguramente relacionadas con la intensidad del estímulo químico.

5. Área (area)

- **Media**: 91.51 unidades cuadradas.
- **Rango**: De 8.5 a 167.88 unidades cuadradas.
- **Desviación Estándar**: 26.94.
- **Percentiles**:
    - 25% < 73.1.
    - 50% (mediana) = 92.02.
    - 75% > 110.5.

La mayoría de los experimentos reportaron una cobertura significativa del área. Algunos casos excepcionales superaron 150 unidades, mostrando una respuesta de crecimiento

extraordinaria, ya sea por condiciones favorables (tratamiento menos agresivo) o por estrategias adaptativas frente al estímulo.

Etapa 5: Comprensión de datos

Tras la recolección inicial, es fundamental comprender la calidad y la estructura de la información antes de pasar al modelado. Para ello, se recurre a estadísticas descriptivas y a visualizaciones que permitan identificar la distribución de las variables, la presencia de valores atípicos y los primeros indicios de relaciones entre datos. En el caso de Physarum polycephalum, esta etapa resulta especialmente valiosa para detectar cómo influyen los diferentes tratamientos (en particular las concentraciones de cafeína) sobre el Contacto, el Tiempo de cruce y otras métricas de comportamiento.
A continuación, se presentan algunas observaciones generales derivadas del análisis preliminar de los datos:

Observaciones Generales

1. **Variabilidad Alta**
   El Contacto y el Tiempo_de_cruce muestran un rango muy amplio, evidenciando respuestas conductuales heterogéneas del moho ante los diferentes tratamientos.
2. **Concentraciones de Cafeína**
   Se detectó que concentraciones superiores a 2 mM provocan efectos adversos, dificultando el cruce y prolongando notablemente los tiempos de cruce.
3. **Datos Atípicos**
   Existen outliers (por ejemplo, valores de Contacto > 900), que se gestionaron con técnicas de rango intercuartílico (IQR) para evitar que distorsionen de forma excesiva el análisis y el posterior modelado.
4. **Distribución Temporal**
   El experimento abarca 9 días, lo que permite capturar una visión longitudinal de la habituación y la respuesta conductual del moho frente a los estímulos, contribuyendo a un análisis más robusto de su comportamiento adaptativo.

Con estos primeros **insights**, se sientan las bases para un tratamiento de datos adecuado en etapas posteriores, donde se afinarán las estrategias de modelado y la comparación con enfoques de optimización tradicionales.

A continuación, se mostrará una serie de códigos parciales y graficas donde se hace una exploración mas detallada de los datos usando herramientas de Python como *pandas*, *matplotlib* y *seaborn* para visualizar la distribución de un conjunto de observaciones clasificadas por el tipo de tratamiento aplicado a *Physarum polycephalum* en esta investigación realizada en el 2016.

Códigos y su funcionamiento:

1. **Conteo de frecuencias y configuración de color**
   Se determina cuántas observaciones tiene cada tipo de tratamiento (counts), cuál es el más frecuente (max_treatment), el menos frecuente (min_treatment) y cuántos datos hay en total.

```
# Gráfica de tratamiento
plt.figure(figsize=(8, 6))
counts = df['Tratamiento'].value_counts()
max_treatment = counts.idxmax()
min_treatment = counts.idxmin()
total = len(df)

colors = ['gray' if x not in [max_treatment, min_treatment]
          else 'steelblue' if x == max_treatment
          else 'orange'
          for x in df['Tratamiento'].unique()]
```

2. **Creación de la gráfica**
   Se dibuja un **countplot** mostrando la cantidad de datos para cada valor de Tratamiento.
   hue='Tratamiento' permite aplicar la paleta de colores según el valor de la columna.

```
ax = sns.countplot(x='Tratamiento', hue='Tratamiento', data=df,
palette=colors, dodge=False)
```

3. **Ajustes estéticos**
   Se eliminan ejes y bordes (spines) para un diseño minimalista.
   Se desactiva el eje **Y** (ax.yaxis.set_visible(False)) y se quita la leyenda por defecto.

```
# Ajustes de estilo
ax.set_xlabel('')
ax.set_ylabel('')
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
ax.spines['left'].set_visible(False)
ax.spines['bottom'].set_visible(False)
```

4. **Etiquetado de porcentajes**
   Se calcula el porcentaje correspondiente a cada barra (p.get_height() / total * 100) y se lo coloca en su interior, centrado y en blanco para resaltar.

```
# Añadir porcentaje en cada barra
for p in ax.patches:
    porcentaje = f'{(p.get_height() / total * 100):.1f}%'
    ax.annotate(f'{porcentaje}',
                (p.get_x() + p.get_width() / 2., p.get_height() / 2),
                ha='center',       va='center',       fontsize=10,
color='white', fontweight='bold')
```

5. **Título y recuadro explicativo**
   Se añade un título a la gráfica y un **recuadro** que describe cada sigla de tratamiento.

```
plt.title('Tipo de tratamiento', fontsize=14)
```

```
# Cuadro explicativo
explicacion = """C: Control
CA: Ácido
CC: Cafeína
Q: Quinina"""

plt.text(-0.4, 1100, explicacion, fontsize=12, color='black',
         bbox=dict(facecolor='white',            edgecolor='black',
boxstyle='round,pad=0.5'))
```

Como resultado nos muestra la siguiente grafica.

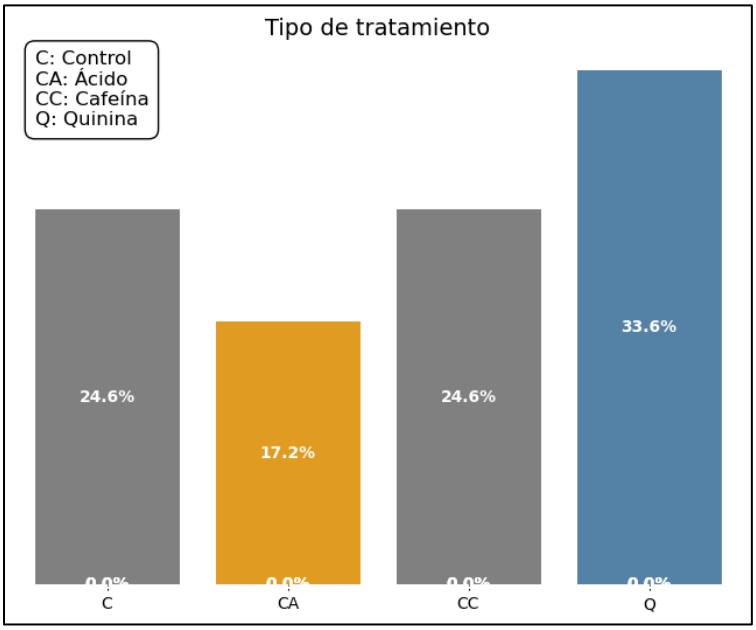


Figura 1

El resultado de la compilación de este código nos generó la Figura 1 dando así la siguiente interpretación de la gráfica:

- Cada barra refleja el número de observaciones pertenecientes a un tipo de tratamiento.
- El porcentaje sobre la barra indica el porcentaje de dichas observaciones respecto al total del dataset.
- Las diferencias de altura permiten ver, por ejemplo, si la Quinina (Q) se aplicó en una proporción mayor que el Control (C) o si el Ácido (CA) es el menos frecuente.
- El uso de colores para destacar el tratamiento más común (steelblue) y el menos frecuente (orange) ayuda a identificar las categorías más y menos representadas de un vistazo.

De esta forma, la gráfica proporciona un resumen rápido de cómo se distribuye el conjunto de datos según el tratamiento aplicado, lo que resulta útil para detectar desbalances en la muestra (por ejemplo, si uno de los tratamientos tuvo muchas menos réplicas que los demás) y para contextualizar los análisis estadísticos y de modelado posteriores.

Después se busca obtener la frecuencia de registros para cada día con **df['Día'].value_counts().sort_index()**, lo que brinda un conteo ordenado del día 1 al 9.

Luego, esa información se convierte a un DataFrame **(df_dias)** que contiene dos columnas: “Día” y “Frecuencia”, facilitando la creación del gráfico.

A continuación en la Figura 2 se usa **sns.barplot** para trazar un gráfico de barras donde el eje X representa los días y el eje Y la cantidad de observaciones. Se añaden las etiquetas numéricas sobre cada barra con **plt.text** para indicar la frecuencia exacta, y así proporcionar una lectura directa de los valores en cada jornada.

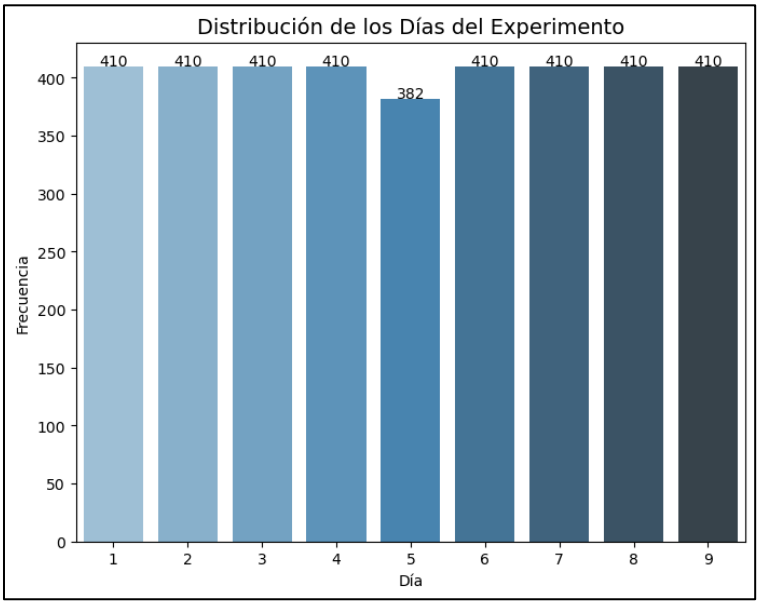


Figura 2.

La figura resultante muestra que en la mayoría de los días se registran alrededor de 410 observaciones, salvo el día 5 con 382. Este hallazgo sugiere una distribución bastante uniforme del muestreo a lo largo del tiempo, lo cual favorece la comparación temporal y reduce la posibilidad de sesgos por un registro desigual de datos.
Al igual que en la gráfica anterior (que mostraba la distribución de los tratamientos), este código emplea sns.countplot para representar la frecuencia de cada categoría en la variable “Puente”. Se determinan el puente más frecuente y el menos frecuente para asignarles colores distintivos (teal y gold, respectivamente), mientras que los restantes se representan en lightcoral. Al eliminar ejes y leyendas y añadir el porcentaje dentro de cada barra, la gráfica se centra en la comparación visual entre las clases de puente.

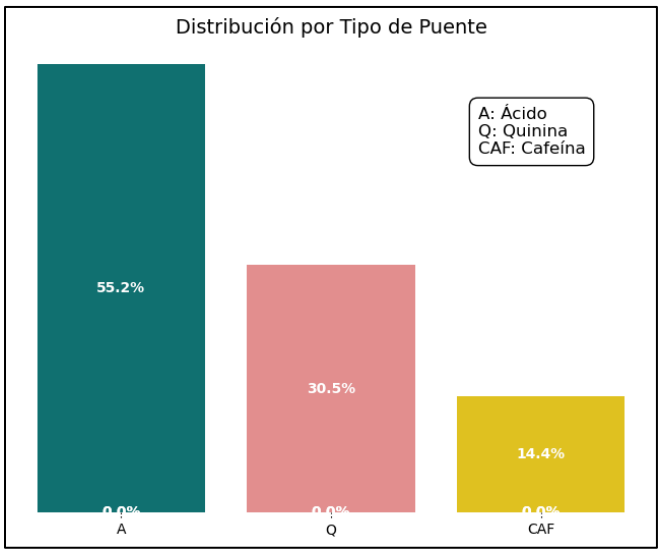


Figura 3.

La Figura 3 resultante del código pone de manifiesto que "A" (Ácido) ocupa el porcentaje más alto (55.2%), seguido de "Q" (Quinina) con 30.5% y, en menor medida, "CAF" (Cafeína) con 14.4%. Esta diferencia de proporciones sugiere que, durante el experimento, el puente ácido se utilizó con mucha más frecuencia que los demás, lo cual podría influir en los resultados de conducta de *Physarum polycephalum* frente a distintos estímulos químicos. Asimismo, es importante considerar esta disparidad en las frecuencias para evitar sesgos en los análisis posteriores, especialmente si se comparan respuestas del moho en condiciones de distintos puentes.

A continuación, se crearán 2 graficas con el siguiente código:

```
# Gráfica de contactos
fig, axes = plt.subplots(1, 2, figsize=(16, 6))

sns.histplot(df['Contacto'], bins=10, kde=True, ax=axes[0],
color='skyblue')
axes[0].set_title('Frecuencia de Contactos', fontsize=14)
axes[0].set_xlabel('Número de Contactos')
axes[0].set_ylabel('Frecuencia')

sns.boxplot(x='Contacto', data=df)
axes[1].set_title('Distribución del Número de Contactos', fontsize=14)
axes[1].set_xlabel('Número de Contactos')
axes[1].set_ylabel('')

# Quitar bordes laterales en ambos gráficos
sns.despine(left=True, bottom=True, ax=axes[0])
sns.despine(left=True, bottom=True, ax=axes[1])

# Mostrar los gráficos
plt.tight_layout()
plt.show()
```

En este caso, el script genera dos gráficas colocadas en paralelo (Figura 4):

- A la izquierda, un histograma (sns.histplot) que muestra la distribución del número de contactos, superponiendo además una curva de densidad (kde=True).
- A la derecha, un boxplot (sns.boxplot) de la misma variable, resaltando los valores centrales y los posibles outliers.

Finalmente, se aplican ciertos ajustes estéticos (sns.despine) para eliminar ejes innecesarios y se dispone todo en un espacio ordenado con plt.tight_layout().

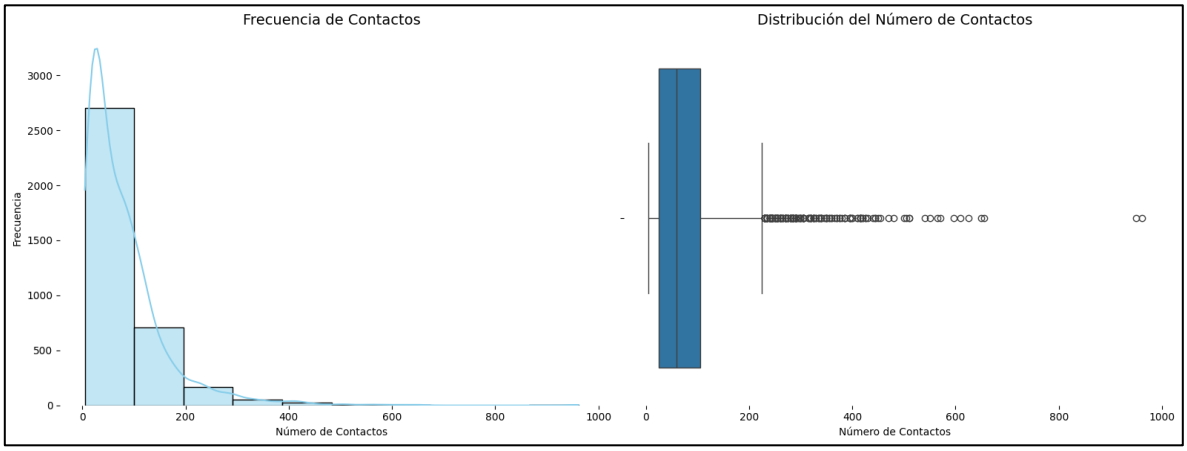


Figura 4.

El histograma revela que la mayoría de los valores de Contacto se concentran cerca de la franja entre 0 y 100, con una caída rápida a medida que aumenta el número de contactos. El boxplot confirma este comportamiento, pues se observa un gran rango de valores y la presencia de datos atípicos (outliers) a la derecha, algunos de los cuales superan 400 e incluso se acercan a 1000. Esto indica que, aunque la mayoría de observaciones de *Physarum polycephalum* se agrupan en niveles de contacto relativamente bajos, un conjunto pequeño de pruebas registró interacciones muy elevadas, lo cual podría reflejar condiciones experimentales específicas o respuestas extremas del moho.

Tras identificar la presencia de valores atípicos en la variable “Contacto”, se aplicó el **método de rango intercuartílico (IQR)** para filtrar aquellos puntos que se encontraban muy alejados de la distribución central. Con ello, se calcula Q1(cuartil 25), Q3(cuartil 75), y el IQR como la diferencia entre ambos. Posteriormente, se definen límites inferior y superior para mantener únicamente los datos que caen en un rango razonable, reduciendo la influencia de valores extremos.

```
Q1 = df['Contacto'].quantile(0.25)
Q3 = df['Contacto'].quantile(0.75)
IQR = Q3 - Q1

limite_inferior = Q1 - 1.5 * IQR
limite_superior = Q3 + 1.3 * IQR

df_filtrado = df[(df['Contacto'] >= limite_inferior) & (df['Contacto'] <=
limite_superior)]
```

Al recompilar el mismo código de visualización (histograma y boxplot) con el DataFrame filtrado (df_filtrado), se obtiene una representación más ajustada de la mayoría de los registros. De esta manera, el histograma muestra una concentración más clara de las observaciones en rangos bajos de "Contacto", y el boxplot reduce la cantidad de puntos atípicos, facilitando la interpretación de la distribución real del moho sin que los outliers distorsionen su análisis.

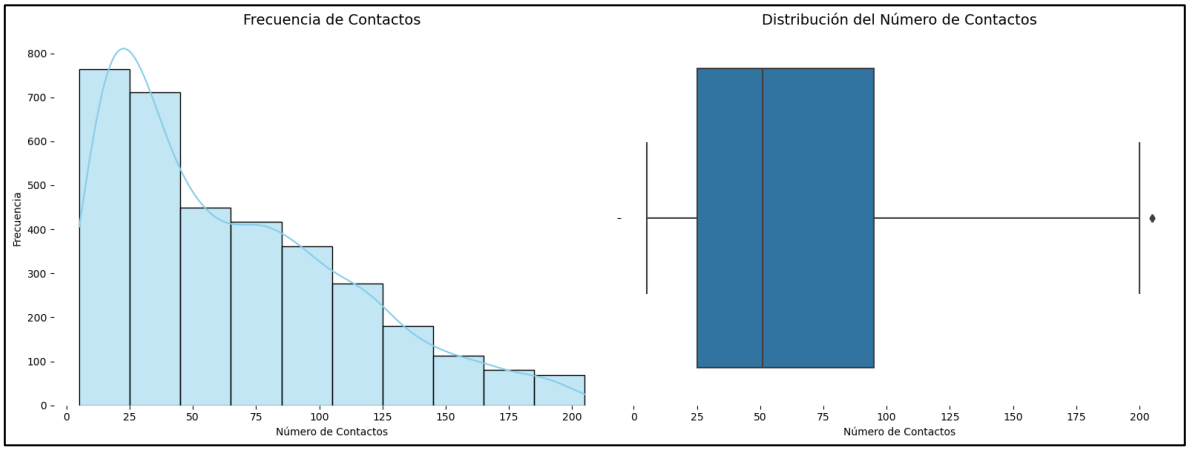


Figura 5.

Así, del análisis filtrado se puede concluir que alrededor del 75% de los datos presentan valores de contacto inferiores a 100, lo que confirma la alta concentración de observaciones en la franja baja de la escala. Con la eliminación de outliers, la distribución refleja de forma más fidedigna cómo, en la mayoría de los casos, el moho interactúa en un rango moderado de contactos, mientras que las respuestas extremas (valores muy elevados) se consideran situaciones puntuales que podrían responder a condiciones experimentales muy específicas.

Teniendo en cuenta los datos anteriores, se mostrarán 2 graficas sobre el área y la arborización ya que buscamos encontrar patrones de crecimiento del moho.

Distribución de la Arborización

La gráfica de histogramas (Figura 6) (con la curva de densidad kde) muestra cómo el valor de *Arborización* se concentra en un rango intermedio que gira en torno a 0.5–0.6, reflejando la tendencia del moho a alcanzar un nivel moderado de ramificación. Sin embargo, se aprecian valores tanto más bajos (cercanos a 0.1–0.2) como más altos (0.8–0.9), lo que indica que, si bien en promedio el organismo tiene una expansión media, existen casos de ramificación limitada y otros de ramificación muy amplia.

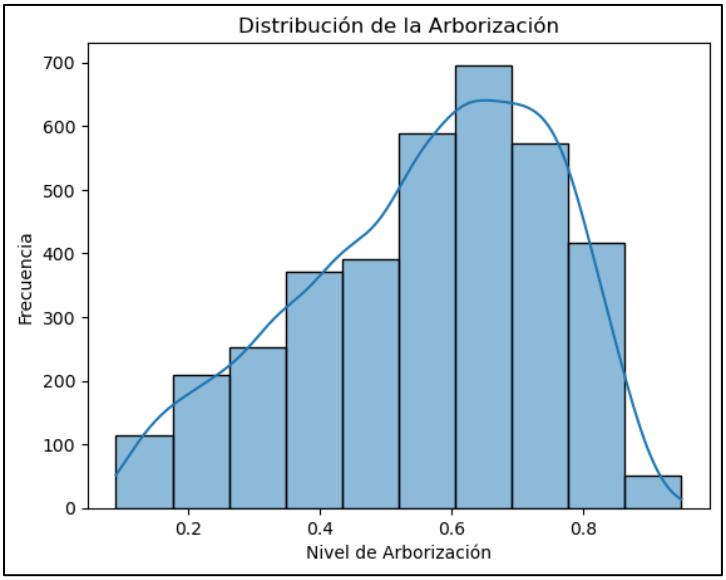


Figura 6

Distribución del Área Cubierta

La gráfica ilustra la variabilidad en la extensión cubierta por el moho (Figura 7). El grueso de los datos (caja) se ubica aproximadamente entre 70 y 110, con una mediana cercana a 90–95. Esto significa que la mayor parte de las observaciones se concentran en ese rango, sugiriendo que el moho cubre comúnmente un área media. No obstante, se observan puntos atípicos por debajo de 40 y por encima de 160, reflejando ensayos en los que el organismo apenas se expandió, así como casos excepcionales con una cobertura muy elevada.

Ambas gráficas proporcionan una visión general de cómo se comportan las variables de interés para el moho, ayudando a plantear la segmentación que se llevará a cabo con el algoritmo K-Means y a considerar posibles valores extremos antes de la fase de modelado.

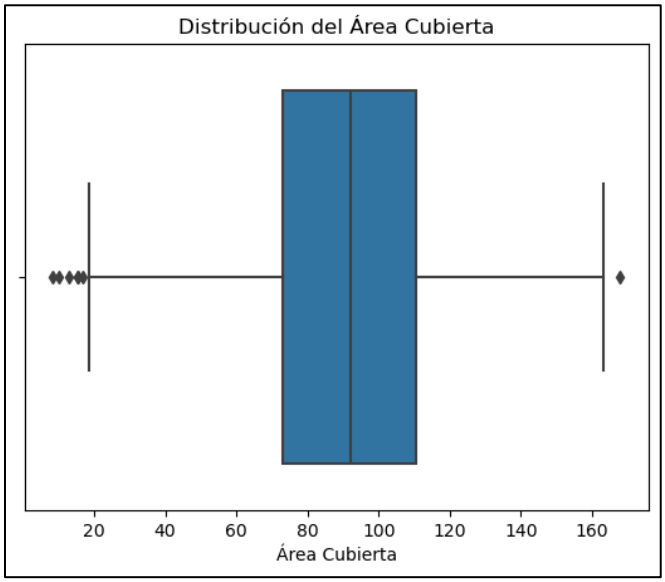


Figura 7.

Ambas gráficas proporcionan una visión general de cómo se comportan las variables de interés para el moho, ayudando a plantear la segmentación que se llevará a cabo con el algoritmo K-Means y a considerar posibles valores extremos antes de la fase de modelado.

Etapa 6: Preparación de datos

En esta etapa, el objetivo principal es **ajustar** y **estructurar** la información necesaria para que el **algoritmo K-Means** pueda **segmentar** eficazmente el comportamiento de *Physarum polycephalum*. Tras un análisis preliminar de todas las variables (Contacto, Tiempo_de_cruce, Arborización, Área), se determinó que **Contacto** y **Tiempo_de_cruce** ofrecen la mejor capacidad de discriminación, tal y como lo indican las puntuaciones de **inercia** y **coeficiente de silueta**. Con el fin de reducir el impacto de escalas y valores atípicos, se aplican técnicas de escalado (por ejemplo, StandardScaler) y filtrados (rango intercuartílico, IQR). Finalmente, se ejecuta **K-Means** para encontrar cuatro clusters, fundamentados en la hipótesis de que cada uno podría corresponder a un **patrón conductual** del moho —relacionado, de manera indirecta, con los diferentes tratamientos—. Esta preparación asegura una **base sólida** para agrupar las observaciones de modo que la formación de clusters refleje de forma fidedigna las dinámicas de interacción y tiempo de cruce.

Para iniciar, se seleccionan las variables numéricas *Contacto*, *Tiempo_de_cruce*, *Arborización* y *Área*, y se **escalan** mediante StandardScaler para uniformar su rango. A continuación, se aplica **KMeans** con 4 clusters, asumiendo la hipótesis inicial de que podrían agruparse en torno a los cuatro tratamientos. Finalmente, se agrega una columna 'Cluster' a df con las asignaciones de cada observación y se crea un countplot para observar cuántas muestras de cada tipo de tratamiento caen en cada grupo.

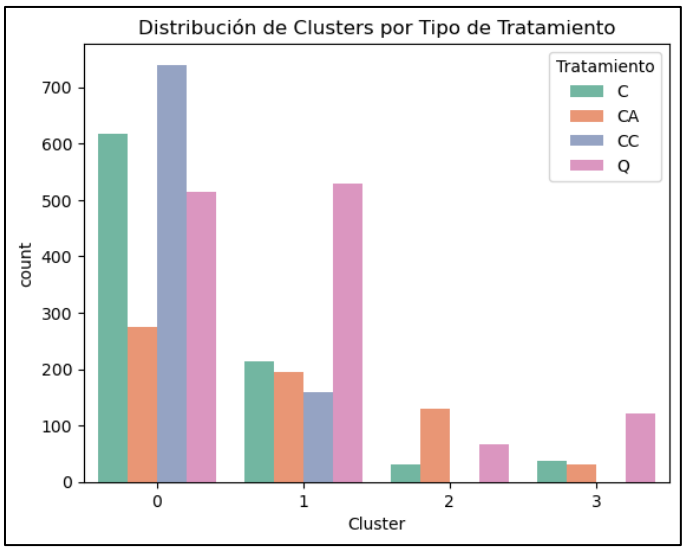


Figura 8.

La gráfica resultante muestra la **distribución de observaciones** en cada uno de los 4 clusters, diferenciando según el tratamiento (C, CA, CC, Q) a través de diferentes colores. Si, por ejemplo, el Cluster 0 alberga mayormente observaciones de *CC* (Cafeína) y *Q* (Quinina), se podría inferir que sus variables numéricas (Contacto, Tiempo_de_cruce, Arborización, Área)

presentan patrones similares para esos dos tratamientos. En cambio, si un cluster está dominado por el Control (C), indica que sus valores en las variables de comportamiento del moho se agrupan de manera distinta. Estos resultados aportan **indicios** acerca de qué tratamientos comparten rasgos conductuales y facilitan la interpretación de posibles **diferencias** o **similitudes** en el comportamiento de *Physarum polycephalum*.

A continuación se generaran una serie de graficas con las variables (Contacto, Tiempo_de_cruce, Arborización y Área) dentro de cada cluster generado por K-Means. Al compararlos de esta manera, se identifica qué rasgos caracterizan a cada grupo y se evalúa hasta qué punto los clusters reflejan comportamientos diferenciados de *Physarum polycephalum*. Dado que K-Means se basa en distancias en un espacio de variables escaladas, es crucial comprobar si los conjuntos resultantes efectivamente mantienen patrones únicos en la variable analizada.

Interpretación General de las Gráficas

- Contacto: El boxplot de Contacto revela diferencias notorias entre los cuatro clusters. Por ejemplo, un cluster podría mostrar valores medianos bajos y pocos outliers, mientras que otro presenta una mediana más elevada con numerosos valores atípicos. Esto sugiere que algunos grupos de moho interactúan intensamente con el entorno (alto Contacto), mientras que otros registran interacciones mucho más modestas.

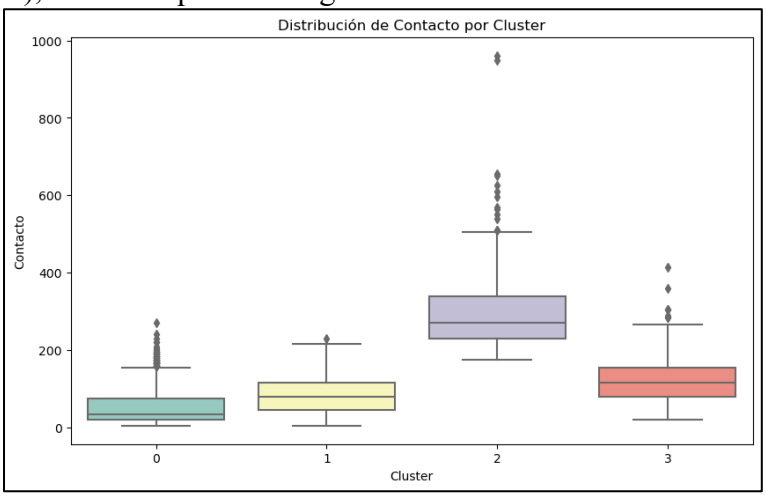


Figura 9.

- Tiempo_de_cruce: Examinar esta variable por cluster permite distinguir qué grupos tardan más en atravesar el puente. Si un cluster exhibe tiempos de cruce significativamente mayores, indicaría una respuesta de aversión o cautela del moho ante el estímulo correspondiente.

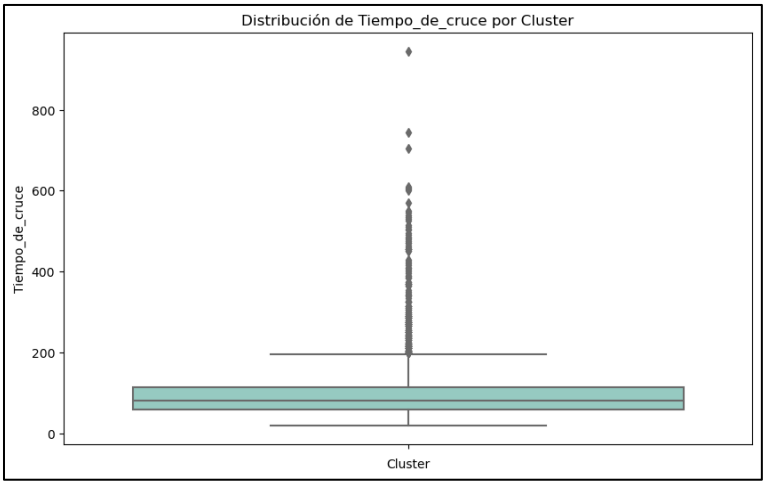


Figura 10

- Arborización: Esta métrica refleja la ramificación del moho. Un cluster con valores altos indica un moho más expansivo, mientras que valores bajos sugieren un patrón de crecimiento restringido. Las diferencias en el boxplot sugieren variaciones marcadas en la estrategia de expansión celular en cada grupo.

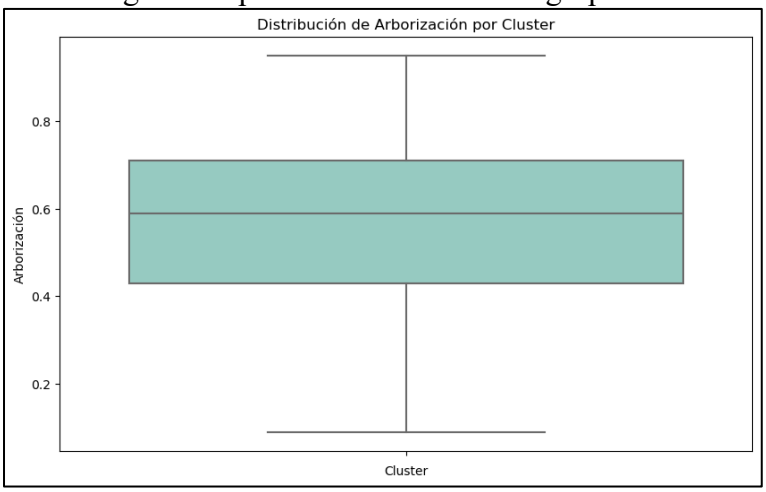


Figura 11.

- Área: El área cubierta muestra si el moho se extiende ampliamente o permanece más concentrado. Un cluster con mayor mediana y menor dispersión indicaría un comportamiento de crecimiento estable, mientras que valores muy dispersos podrían denotar reacciones erráticas o adaptaciones específicas a los tratamientos.

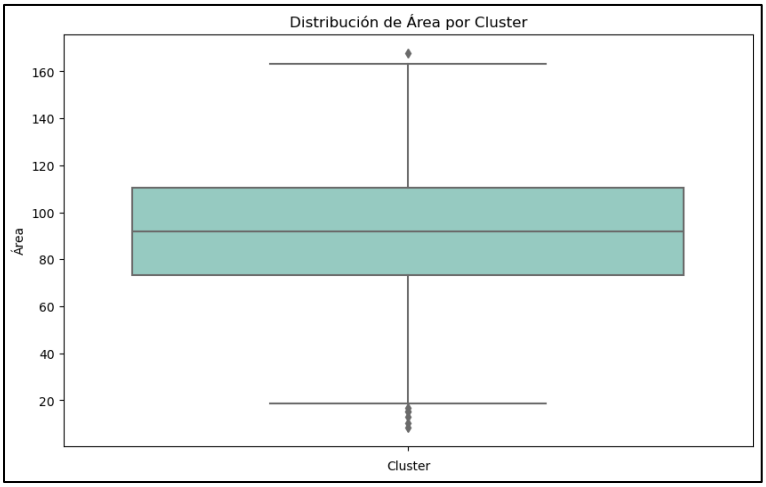


Figura 12.

En conjunto, estos boxplots permiten verificar la cohesión interna de cada cluster y comprender por qué K-Means clasificó las observaciones de cierta manera, facilitando la identificación de grupos con rasgos distintivos en el comportamiento de *Physarum polycephalum*.

Descripción e Interpretación del Análisis con PCA

A continuación, se usara el método de Análisis de Componentes Principales (PCA) para proyectar las variables numéricas escaladas (Contacto, Tiempo_de_cruce, Arborización, Área) en un espacio bidimensional:

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
pca_result = pca.fit_transform(variables_numericas_scaled)

df['PCA1'] = pca_result[:, 0]
df['PCA2'] = pca_result[:, 1]

plt.figure(figsize=(10, 8))
sns.scatterplot(x='PCA1',    y='PCA2',    hue='Cluster',    palette='Set1',
data=df, style='Tratamiento')
plt.title('Visualización de Clusters usando PCA')
plt.show()
```

Explicación general del código:

1. **Cálculo de PCA**
   Mediante PCA(n_components=2), se extraen dos ejes principales (*PCA1* y *PCA2*) que capturan gran parte de la variabilidad presente en los datos originales. Cada punto (observación) es proyectado en este espacio nuevo, permitiendo visualizar la dispersión y relaciones entre los registros de forma más clara.
2. **Incorporación de PCA1 y PCA2 al DataFrame**
   Al asignar df['PCA1'] = pca_result[:, 0] y df['PCA2'] = pca_result[:, 1], se añaden columnas con las coordenadas de cada observación en estas dos primeras componentes.

3. **Visualización con Seaborn**
   La función sns.scatterplot traza los datos según *PCA1* y *PCA2*, coloreando cada punto de acuerdo con su cluster (0, 1, 2 o 3) y variando la forma según el Tratamiento (C, CA, CC, Q). Esto posibilita analizar de un vistazo si ciertos tratamientos tienden a agruparse en uno u otro cluster o si se reparten de manera más heterogénea.

Interpretación de la Gráfica de Clusters usando PCA

1. **Ejes *PCA1* y *PCA2***
   - *PCA1* y *PCA2* son los ejes que **condensan** la mayor parte de la varianza de los datos.
   - Cada punto en el plano (PCA1, PCA2) representa una observación del dataset con base en sus características originales (*Contacto*, *Tiempo_de_cruce*, *Arborización*, *Área*).

2. **Clusters (0, 1, 2, 3)**
   - K-Means ha asignado un color distinto a cada cluster.
   - **Cluster 0 (Rojo)** aparece con valores más bajos en PCA1 y PCA2.
   - **Cluster 1 (Azul)** se distribuye horizontalmente, ocupando el núcleo de la gráfica.
   - **Cluster 2 (Verde)** se concentra hacia el lado derecho (valores mayores de PCA1).
   - **Cluster 3 (Morado)** se sitúa mayormente en la parte superior (valores elevados de PCA2).

3. **Tratamiento (Forma de los Puntos)**
   - **C (círculo)**: Control, sin estímulos específicos. Se ve principalmente en los clusters 1 y 3.
   - **CA (estrella)**: Ácido, con más incidencia en clusters 0 y 1.
   - **CC (cuadrado)**: Cafeína, frecuentemente ubicado en el cluster 2.
   - **Q (cruz)**: Quinina, concentrada sobre todo en los clusters 2 y 3.

4. **Observaciones**
   - El **Tratamiento Q (cruz)** sobresale en clusters 2 (verde) y 3 (morado), lo que sugiere rasgos similares en sus variables para estos dos grupos.
   - **CC (cuadrado)** comparte notable presencia en el cluster 2, apuntando a posibles afinidades en la forma de crecimiento o respuesta del moho en presencia de cafeína y quinina.
   - **CA (estrella)** se dispersa entre 0 y 1, indicando tal vez una variabilidad media en su comportamiento.
   - **C (círculo)** se observa en clusters 1 y 3, denotando que incluso el grupo control manifiesta dos patrones de respuesta distintos.

5. **Interpretación General**
   - **Cluster 0** podría agrupar casos con **características más extremas** o comportamientos atípicos en valores de PCA1 y PCA2.

- **Cluster 1** funciona como un cluster **central**, con mayor heterogeneidad de tratamientos.
- **Cluster 2** se asocia con **Cafeína y Quinina**, reforzando la idea de que esas variables comparten rasgos conductuales en el moho.
- **Cluster 3** congrega observaciones con **valores altos en PCA2**, probablemente relacionadas con mayor variabilidad en variables como *Arborización* o *Área*.

El diagrama final (Figura 13) muestra **agrupaciones claras** en el espacio de PCA, pero con cierta mezcla de tratamientos. Esto sugiere que, aunque existen rasgos distintivos en cada cluster, los tratamientos pueden exhibir **patrones conductuales similares**, resultando en su dispersión entre diversos grupos. Esta representación facilita la comprensión de cómo los cuatro clusters se relacionan con la respuesta de *Physarum polycephalum* frente a diferentes estímulos químicos.

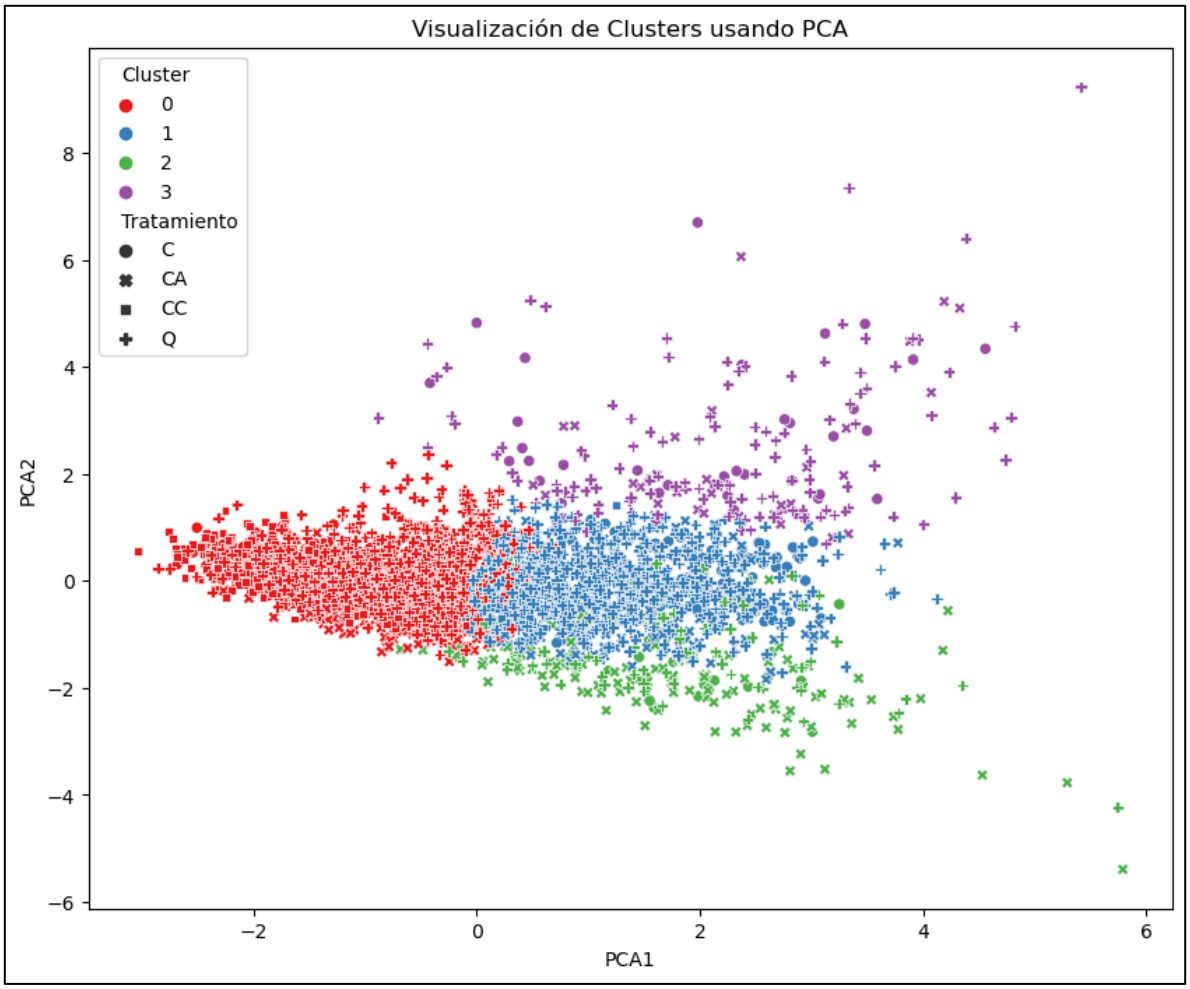


Figura 13

Para determinar qué combinación de variables es más adecuada en el modelado con K-Means, se realiza un muestreo (resample) de 6000 registros y luego se prueban todas las combinaciones posibles de las características [Contacto, Tiempo_de_cruce, Arborización, Área], desde subconjuntos de tamaño 2 hasta incluir las 4 variables. En cada caso:

1. Se escalan las variables elegidas (StandardScaler) para uniformar sus rangos.
2. Se ejecuta K-Means con n_clusters=4, asumiendo la hipótesis de que el moho podría agruparse en torno a las cuatro categorías de tratamiento.
3. Se miden la inercia (inertia_) y el coeficiente de silueta (silhouette_score) para evaluar la calidad de la agrupación en términos de cohesión interna y separación entre clusters.

El DataFrame resultados_sample_df almacena los valores de inercia y silueta para cada combinación, y luego se ordena (sort_values) según la silueta de mayor a menor (la silueta máxima indica la mejor partición en clusters). Esto se mostrara en la Figura 14.

[27]:

	Combinación	Inercia	Silhouette
0	(Contacto, Tiempo_de_cruce)	3603.703848	0.484283
6	(Contacto, Tiempo_de_cruce, Arborización)	6776.650025	0.405420
3	(Tiempo_de_cruce, Arborización)	3276.446835	0.383350
1	(Contacto, Arborización)	3348.290349	0.379860
5	(Arborización, Área)	2954.193043	0.376198
4	(Tiempo_de_cruce, Área)	3823.584008	0.357742
10	(Contacto, Tiempo_de_cruce, Arborización, Área)	10683.557418	0.354046
2	(Contacto, Área)	3652.103840	0.350150
7	(Contacto, Tiempo_de_cruce, Área)	7424.592976	0.343084
9	(Tiempo_de_cruce, Arborización, Área)	6460.068868	0.300000

Figura 14

La combinación de **Contacto y Tiempo_de_cruce** reporta la **mejor** puntuación de silueta, indicando que estas dos variables juntas ofrecen la segmentación más clara del comportamiento de *Physarum polycephalum*. Esto sugiere que la intensidad de contacto con el puente y la velocidad para atravesarlo son factores diferenciadores más potentes que arborización o área, al menos para la agrupación en cuatro clusters. Con esta evidencia, se justifica focalizar el análisis en Contacto y Tiempo_de_cruce si se busca un particionado más nítido de observaciones.

Se selecciona la combinación de características que obtuvo la mejor puntuación de silueta (en este caso, ('Contacto', 'Tiempo_de_cruce')) y se extraen dichas columnas de la muestra df_sample. Luego, sns.pairplot genera una matriz de dispersión entre ambas variables, presentando histogramas o curvas de densidad en la diagonal (diag_kind='kde'). Esto permite visualizar cómo se distribuyen los datos en cada dimensión y si existe alguna relación aparente (por ejemplo, un agrupamiento natural o un patrón creciente/decreciente).

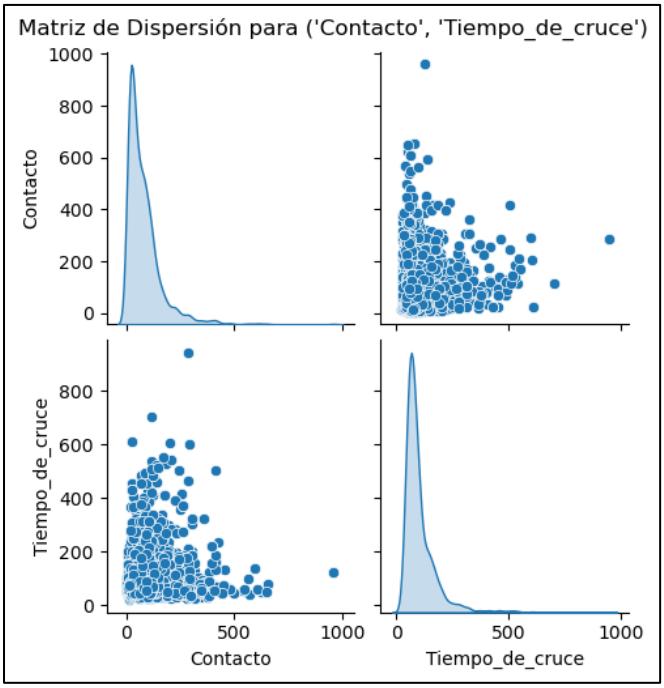


Figura 15.

El gráfico muestra en el eje X y Y a las dos variables de interés: **Contacto** y **Tiempo_de_cruce**. Se aprecia que la mayoría de los puntos se concentran en valores bajos de Contacto y Tiempo de cruce, con algunos casos dispersos a valores muy altos en ambas variables. En el eje diagonal de la matriz de dispersión se observan las distribuciones univariadas para cada variable (Contacto, Tiempo_de_cruce), notando que ambas presentan una fuerte concentración de valores bajos y una progresiva disminución de frecuencia a medida que se incrementan las mediciones. Esto sugiere distribuciones asimétricas (con colas largas hacia la derecha), en las que un número reducido de observaciones se distingue por valores muy altos.

En los gráficos de dispersión cruzados, la mayoría de los puntos se concentra por debajo de aproximadamente 200 en Contacto y 150 en Tiempo_de_cruce, delineando una gran nube de datos en la zona baja de ambas variables. A su vez, se aprecian algunos casos que se alejan de dicho conglomerado, indicando instancias donde el moho registró un número de contactos excepcionalmente elevado o tiempos de cruce prolongados.

Esta separación en “nubes” sugiere la posible existencia de subgrupos (clusters) naturales: uno donde el moho tiene un comportamiento más “típico” (bajo contacto y tiempo de cruce), y otros en que, ya sea por motivaciones experimentales (diferente tratamiento químico) o por condiciones particulares, presenta valores mucho más altos. Así, el alto contraste en ambas variables (Contacto y Tiempo_de_cruce) explica por qué resultan tan eficaces al distinguir comportamientos en un análisis de clustering: capturan, de forma sintética, la heterogeneidad en la interacción del moho con el entorno y en la rapidez con que cruza el puente experimental.

En esta sección, presentamos el **proceso de clustering** empleado para agrupar las observaciones en torno a las variables *Contacto* y *Tiempo_de_cruce*. Estas dos características

se identificaron como las más representativas para distinguir el comportamiento del moho *Physarum polycephalum*, según un análisis previo de puntuaciones de inercia y coeficiente de silueta. El código abarca desde la selección y el escalado de variables hasta la ejecución de K-Means y la visualización de los clusters, permitiendo observar qué patrones afloran en la respuesta del moho ante distintos tratamientos.

```
import seaborn as sns
import matplotlib.pyplot as plt

mejor_combinacion = resultados_sample_sorted.iloc[0]['Combinación']
data_mejor = df_sample[list(mejor_combinacion)]
scaler = StandardScaler()
data_mejor_scaled = scaler.fit_transform(data_mejor)

kmeans_mejor = KMeans(n_clusters=4, random_state=42)
df_sample['Cluster_Mejor'] = kmeans_mejor.fit_predict(data_mejor_scaled)

plt.figure(figsize=(10, 8))
sns.scatterplot(x=mejor_combinacion[0],            y=mejor_combinacion[1],
hue='Cluster_Mejor',            data=df_sample,            palette='Set2',
style='Cluster_Mejor')
plt.title(f'Clusters basados en {mejor_combinacion}')
plt.xlabel(mejor_combinacion[0])
plt.ylabel(mejor_combinacion[1])
plt.legend(title='Cluster')
plt.show()
```

Generando así la siguiente grafica

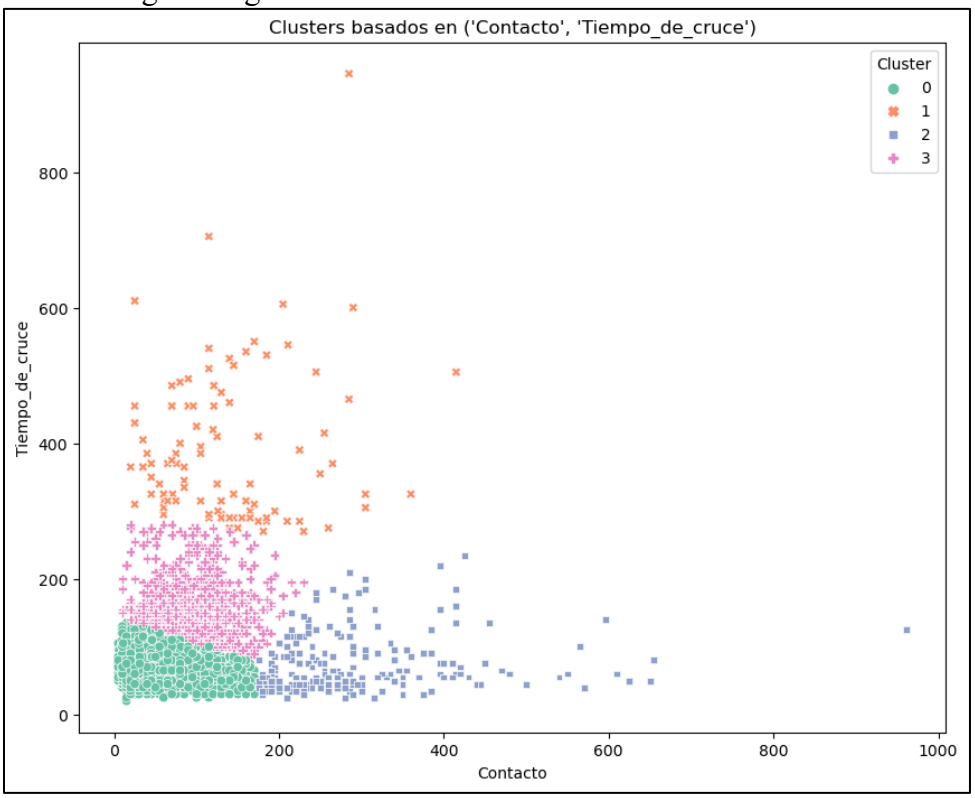


Figura 16.

Interpretación de los Clusters en (Contacto, Tiempo_de_cruce)

1. **Cluster 0 (círculos verdes)**
    - **Región**: Bajo Contacto (0–100) y Tiempo de cruce menor a 100.
    - **Posible Tratamiento**: Se asocia con **Control (C)**, pues exhibe un comportamiento estable y poco afectado por estímulos.
2. **Cluster 1 (cruces naranjas)**
    - **Región**: Contacto y tiempo de cruce **elevados**, algunos puntos superan 500.
    - **Posible Tratamiento**: Podría relacionarse con **Ácido (CA)**, al reflejar un moho que evita firmemente el estímulo (tiempos de cruce prolongados).
3. **Cluster 2 (cuadrados azules)**
    - **Región**: Contacto de 100 a más de 500, con Tiempos de cruce muy variables (100–400).
    - **Posible Tratamiento**: Se sugiere **Cafeína (CC)**, ya que el moho presenta una alta interacción, pero con rapidez de cruce diversa.

4. **Cluster 3 (cruces rosadas)**
    - **Región**: Contacto y tiempo de cruce **moderados** (entre 100–300).
    - **Posible Tratamiento**: Se vincula con **Quinina (Q)**, mostrando niveles de interacción y velocidades de cruce intermedias.

Los resultados confirman que **Contacto** y **Tiempo_de_cruce** son **factores determinantes** para la segmentación del comportamiento del moho, diferenciando cuatro posibles patrones o clusters. No obstante, para la *Etapa 7: Modelado*, se buscará **contrastar** la eficacia de estos hallazgos a través de una **Regresión Logística**, profundizando en la capacidad de predecir el tipo de tratamiento (C, CA, CC, Q) a partir de estas variables clave. De esta manera, se integrará un enfoque de **clasificación supervisada**, complementando los resultados obtenidos con K-Means y permitiendo una visión más robusta de la dinámica conductual de *Physarum polycephalum*.

Etapa 7: Modelado

Una vez que se han identificado los clusters y se ha constatado la relevancia de **Contacto** y **Tiempo_de_cruce** en el comportamiento de *Physarum polycephalum*, se procederá a **profundizar** en los hallazgos a través de un **modelo supervisado**: la **Regresión Logística**. Este enfoque permitirá **predecir** el tipo de tratamiento al que fue sometido el moho, basándose en los patrones detectados de interacción y velocidad de cruce. Al complementar el análisis no supervisado (K-Means) con el **enfoque supervisado** de la regresión, se espera confirmar la **validez** de estas variables para discriminar los tratamientos y obtener conclusiones más sólidas sobre las estrategias de habituación y adaptación del moho en diferentes condiciones experimentales.

```
plt.figure(figsize=(10, 8))
corr_matrix = df[['Contacto', 'Tiempo_de_cruce', 'Arborización',
'Área']].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Mapa de Correlación entre Variables Numéricas')
plt.show()
```

Este bloque de instrucciones calcula la **matriz de correlación** (corr()) de las variables numéricas —Contacto, Tiempo_de_cruce, Arborización y Área— y genera un **mapa de calor** (heatmap) mediante Seaborn. Al excluir la variable categórica "Tratamiento", se busca comprobar si existe **correlación significativa** entre las variables cuantitativas que podría ocasionar multicolinealidad en la **Regresión Logística** generando así la figura 17.

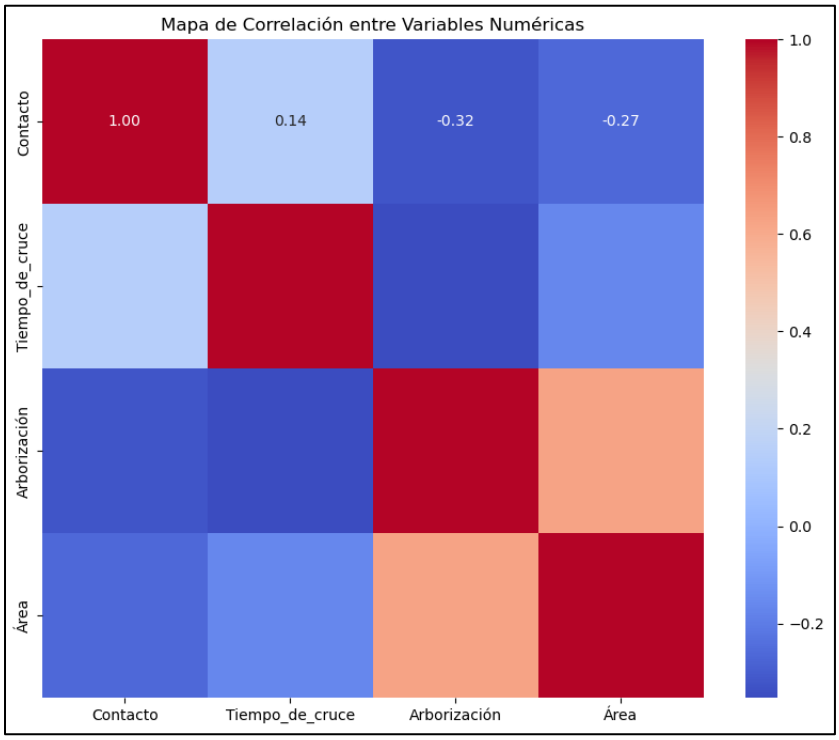


Figura 17

Interpretación del Mapa de Correlación

- Cada celda refleja la **correlación de Pearson** entre dos variables (por ejemplo, Contacto vs. Tiempo_de_cruce). Un valor cercano a **+1** indica correlación positiva alta; cerca de **-1**, correlación negativa alta.
- En la figura se observa que las correlaciones entre variables son relativamente **moderadas**, sin superar ±0.7, lo que sugiere que no habría un problema crítico de multicolinealidad.
- Hay ligeras correlaciones negativas (por ejemplo, Contacto con Arborización) implican que a mayor Contacto, el grado de Arborización tiende a ser menor (o viceversa), pero sin ser un factor determinante.

En conjunto, la matriz indica que **Contacto** y **Tiempo_de_cruce** —variables previamente identificadas como influyentes en el comportamiento del moho— no presentan una correlación tan alta como para imposibilitar su uso conjunto en la Regresión Logística, respaldando la decisión de incluirlas como **factores clave** en el próximo modelado.

Relación con la Variable Objetivo

Como estamos interesados en el tipo de tratamiento, que es una variable categórica, podemos analizar cómo las variables numéricas se relacionan con esta variable objetivo.

Coeficientes de Correlación de Pearson o ANOVA: Aunque la variable objetivo es categórica, podemos evaluar si hay diferencias significativas en las variables numéricas para cada categoría.

```
import scipy.stats as stats

for col in ['Contacto', 'Tiempo_de_cruce', 'Arborización', 'Área']:
    f_val, p_val = stats.f_oneway(*[df[df['Tratamiento'] == cat][col] for
cat in df['Tratamiento'].unique()])
    print(f"ANOVA para {col}: F-valor = {f_val:.2f}, p-valor =
{p_val:.4f}")
```

```
ANOVA para Contacto: F-valor = 245.42, p-valor = 0.0000
ANOVA para Tiempo_de_cruce: F-valor = 151.67, p-valor = 0.0000
ANOVA para Arborización: F-valor = 173.11, p-valor = 0.0000
ANOVA para Área: F-valor = 72.17, p-valor = 0.0000
```

Figura 18

Este segmento de código realiza un **Análisis de Varianza (ANOVA)** para verificar si existen **diferencias estadísticamente significativas** en cada variable numérica (*Contacto, Tiempo_de_cruce, Arborización, Área*) al comparar los diferentes tratamientos (C, CA, CC, Q). La función stats.f_oneway permite calcular el **F-valor** y el **p-valor** para cada variable, basándose en la distribución de sus valores dentro de cada categoría de tratamiento.

Interpretación de los Resultados del ANOVA

El análisis de varianza (ANOVA) realizado para evaluar la relación entre las variables numéricas y la variable categórica **Tratamiento** arrojó los siguientes resultados:

Resultados:

1. **Contacto**:
   - **F-valor**: 245.42
   - **p-valor**: 0.0000
   - Interpretación: Existe una diferencia estadísticamente significativa entre los tratamientos en términos de la variable **Contacto**. Esto indica que los valores de contacto varían significativamente según el tratamiento aplicado.

2. **Tiempo_de_cruce**:
   - **F-valor**: 151.67
   - **p-valor**: 0.0000
   - Interpretación: También muestra diferencias significativas entre los tratamientos, lo que sugiere que el tiempo que tarda el moho en cruzar el puente está influenciado por el tratamiento.

3. **Arborización**:
   - **F-valor**: 173.11
   - **p-valor**: 0.0000
   - Interpretación: La arborización del moho tiene diferencias importantes entre tratamientos, lo que podría reflejar un efecto directo del estímulo químico.

4. **Área**:
    - **F-valor**: 72.17
    - **p-valor**: 0.0000
    - Interpretación: Aunque el F-valor es menor en comparación con las otras variables, el área también muestra diferencias significativas entre los tratamientos.

Conclusión:

- **Todas las variables tienen diferencias estadísticamente significativas** entre los tratamientos (p-valor < 0.05).
- Las variables **Contacto** y **Tiempo_de_cruce** tienen los valores de F más altos, lo que sugiere que podrían ser las más discriminativas para el modelo.
- La selección de variables para la regresión logística puede incluir estas cuatro variables, pero podríamos priorizar **Contacto** y **Tiempo_de_cruce** como las más importantes.

En esta fase, se crea un **modelo de Regresión Logística** para cada variable por separado (*Contacto*, *Tiempo_de_cruce*, *Arborización* y *Área*) y se valida su capacidad para **predecir** el tipo de tratamiento aplicado al moho (C, CA, CC, Q). Para ello:

```
from sklearn.linear_model import LogisticRegression
from    sklearn.metrics    import    accuracy_score,    confusion_matrix,
classification_report
import seaborn as sns
import matplotlib.pyplot as plt

# Variables para modelar de forma individual
variables = ['Contacto', 'Tiempo_de_cruce', 'Arborización', 'Área']

for var in variables:
    # Seleccionar una sola variable
    X_train_var = X_train[[var]].values
    X_test_var = X_test[[var]].values

    # Entrenar el modelo multinomial (sin especificar multi_class)
    model     =     LogisticRegression(random_state=42,     max_iter=1000,
solver='lbfgs')
    model.fit(X_train_var, y_train)

    # Predecir en el conjunto de prueba
    y_pred = model.predict(X_test_var)
    y_pred_proba = model.predict_proba(X_test_var)

    # Graficar la probabilidad de predicción para cada clase
    plt.figure(figsize=(10, 6))
    for i, class_name in enumerate(le.classes_):
        plt.scatter(X_test_var,  y_pred_proba[:,  i],  label=f'Probabilidad
Clase {class_name}', alpha=0.5)
    plt.title(f'Regresión Logística Multinomial para {var}')
    plt.xlabel(var)
    plt.ylabel('Probabilidad')
    plt.legend()
    plt.show()
```

```
    # Generar matriz de confusión gráficamente
    cm = confusion_matrix(y_test, y_pred)
    plt.figure(figsize=(8, 6))
    sns.heatmap(cm,        annot=True,        fmt="d",        cmap="Blues",
xticklabels=le.classes_, yticklabels=le.classes_)
    plt.title(f'Matriz de Confusión para {var}')
    plt.xlabel('Predicción')
    plt.ylabel('Verdadero')
    plt.show()

    # Mostrar métricas de evaluación
    report         =         classification_report(y_test,         y_pred,
target_names=le.classes_, zero_division=0)
    print(f"Métricas para {var}:")
    print(report)
```

1. Se **codifica** la variable objetivo (Tratamiento) con LabelEncoder, transformándola en valores numéricos.
2. Se **dividen** los datos en entrenamiento y prueba (train_test_split) para evaluar la capacidad de generalización de cada modelo.
3. Se **entrena** un modelo de regresión logística usando únicamente una columna (X_train_var) y se **predice** sobre la muestra de prueba.
4. Se generan **gráficas** de probabilidad (y_pred_proba) para cada clase y se elabora una **matriz de confusión** que muestra cómo se distribuyen los aciertos y los errores de clasificación.
5. Se **miden** las métricas de precisión, recall y F1 (classification_report) para cuantificar el rendimiento en cada una de las categorías de tratamiento.

A continuación, se mostraran las gráficas y métricas obtenidas con el análisis de cada variable.

Contacto

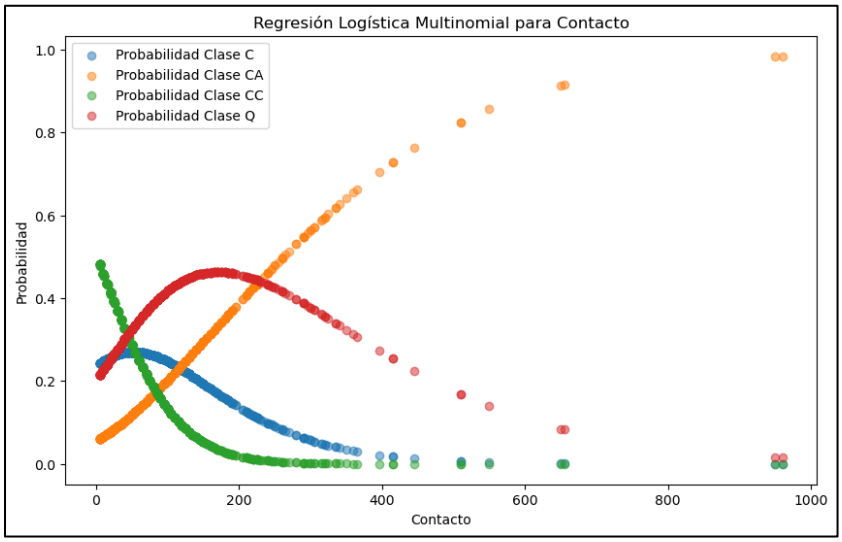


Figura 19

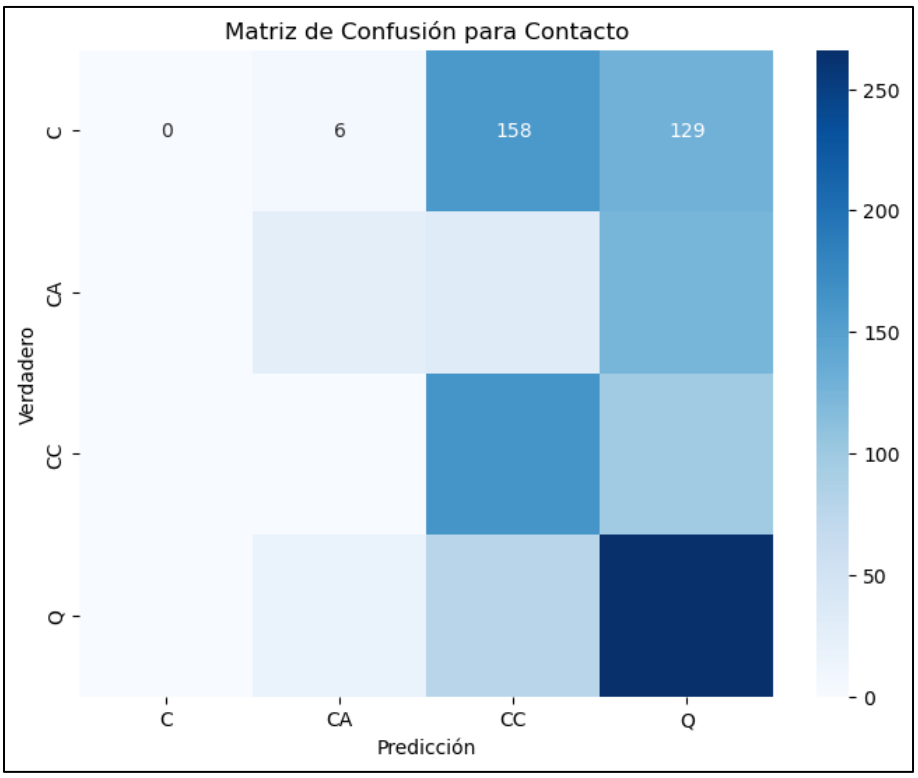


Figura 20.

Métricas para Contacto:

	precision	recall	f1-score	support
C	0.00	0.00	0.00	293
CA	0.52	0.14	0.22	183
CC	0.38	0.62	0.47	261
Q	0.43	0.73	0.54	362
accuracy			0.41	1099
macro avg	0.33	0.38	0.31	1099
weighted avg	0.32	0.41	0.33	1099

Figura 21.

Tiempo_de_cruce

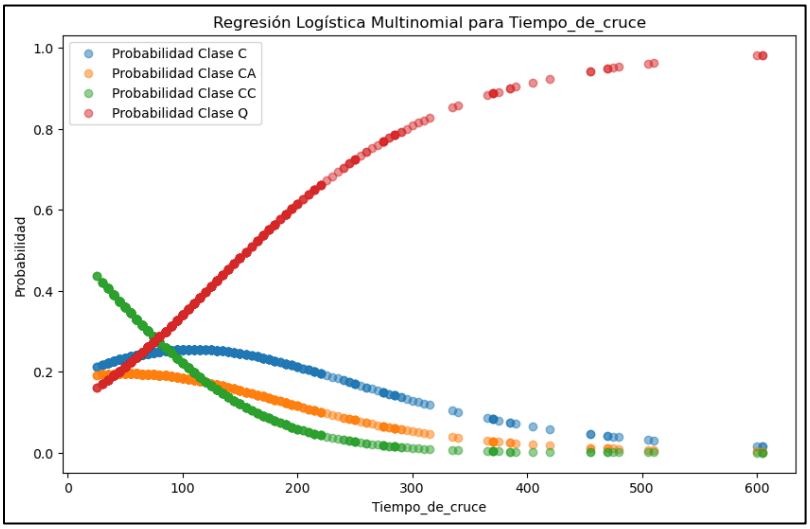


Figura 22.

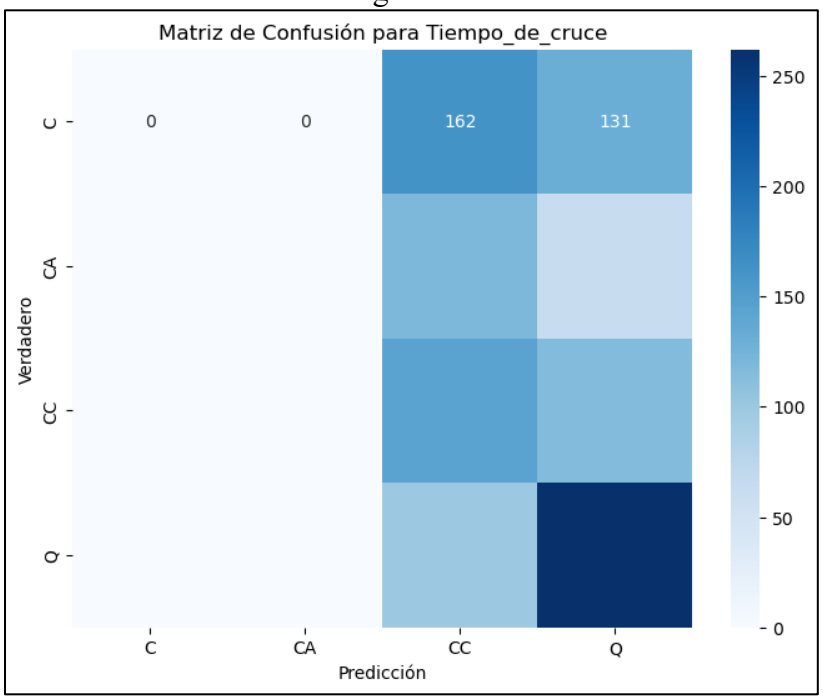


Figura 23

Métricas para Tiempo_de_cruce:

	precision	recall	f1-score	support
C	0.00	0.00	0.00	293
CA	0.00	0.00	0.00	183
CC	0.28	0.56	0.37	261
Q	0.46	0.72	0.56	362
accuracy			0.37	1099
macro avg	0.18	0.32	0.23	1099
weighted avg	0.22	0.37	0.27	1099

Figura 24

Arborización

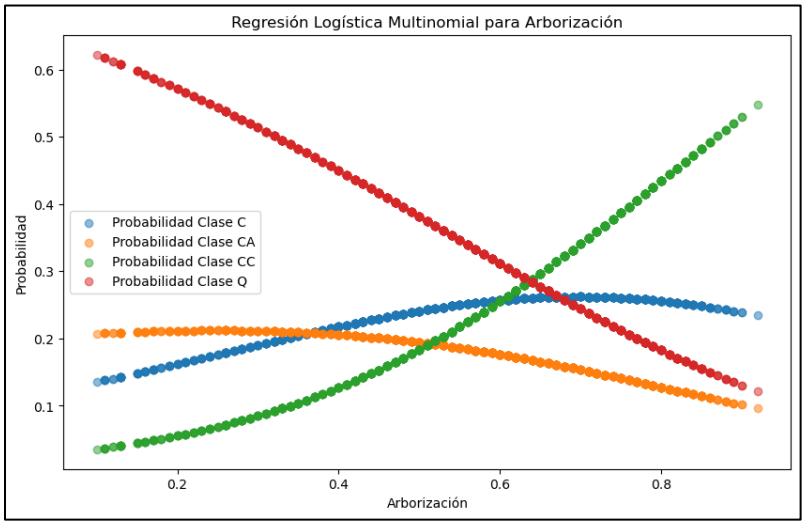


Figura 25

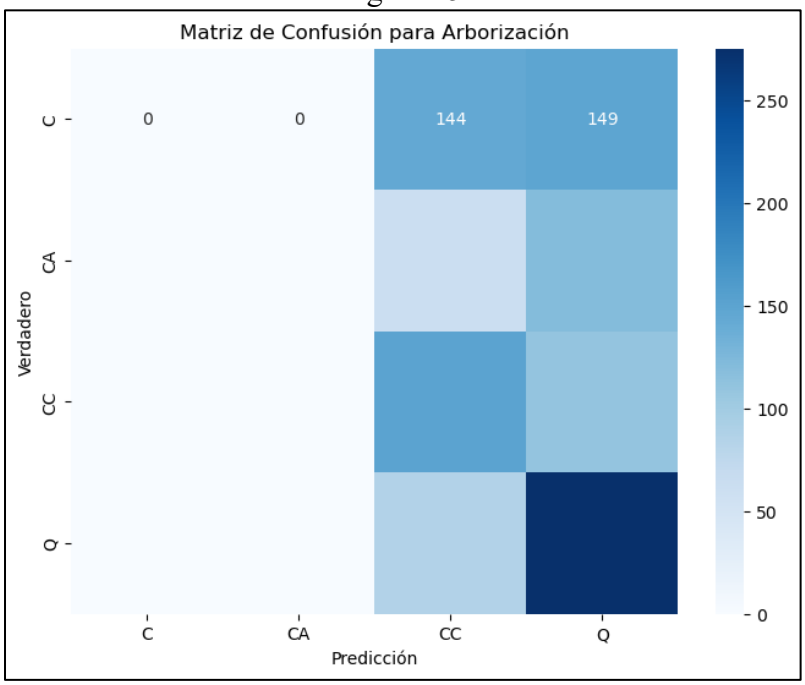


Figura 26

Métricas para Arborización:

	precision	recall	f1-score	support
C	0.00	0.00	0.00	293
CA	0.00	0.00	0.00	183
CC	0.34	0.58	0.43	261
Q	0.42	0.76	0.54	362
accuracy			0.39	1099
macro avg	0.19	0.33	0.24	1099
weighted avg	0.22	0.39	0.28	1099

Figura 27

Área

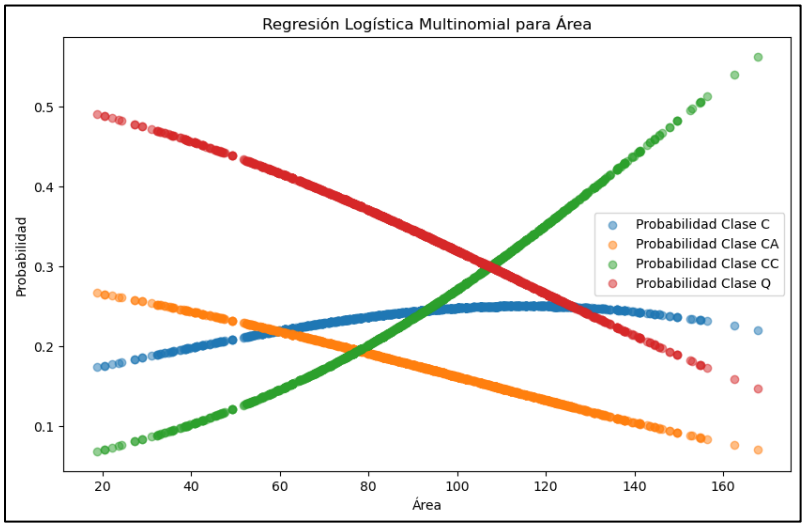


Figura 28

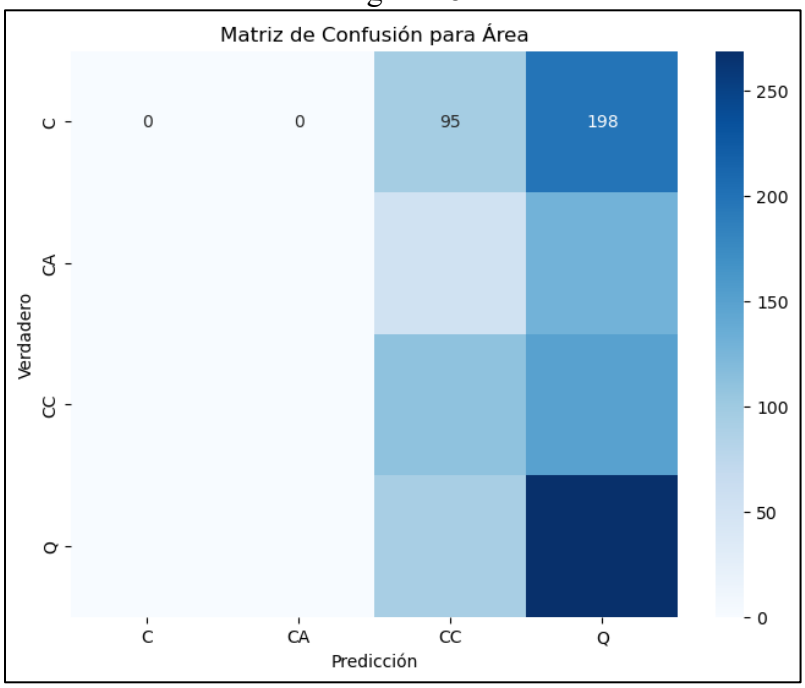


Figura 29

Métricas para Área:

	precision	recall	f1-score	support
C	0.00	0.00	0.00	293
CA	0.00	0.00	0.00	183
CC	0.32	0.43	0.36	261
Q	0.36	0.74	0.49	362
accuracy			0.35	1099
macro avg	0.17	0.29	0.21	1099
weighted avg	0.19	0.35	0.25	1099

Figura 30

Conclusiones del Análisis de Regresión Logística por Variable

Se han entrenado modelos de **regresión logística** individuales para predecir el tipo de tratamiento (**C**, **CA**, **CC**, **Q**) usando las variables **Contacto**, **Tiempo_de_cruce**, **Arborización** y **Área** de forma independiente. A continuación, se presentan las conclusiones basadas en las métricas obtenidas.

1. Contacto
Métricas Clave:

- **Accuracy**: 41%
- **F1-score más alto**:
  - **Clase Q** (0.54), con una precisión de 43% y un recall de 73%.
  - **Clase CC** (0.47), con una precisión de 38% y un recall de 62%.

Observaciones:

- La variable **Contacto** tiene una capacidad moderada para identificar las clases **Q** y **CC**.
- Las clases **C** (Control) y **CA** (Ácido) no se predicen correctamente (precisión y recall = 0).

2. Tiempo_de_cruce
Métricas Clave:

- **Accuracy**: 37%
- **F1-score más alto**:
  - **Clase Q** (0.56), con una precisión de 46% y un recall de 72%.
  - **Clase CC** (0.37), con una precisión de 28% y un recall de 56%.

Observaciones:

- La variable **Tiempo_de_cruce** muestra un comportamiento similar al de Contacto:
  - Se predicen mejor las clases **Q** y **CC**.
  - Las clases **C** y **CA** no tienen predicciones satisfactorias.

3. Arborización
Métricas Clave:

- **Accuracy**: 39%
- **F1-score más alto**:
  - **Clase Q** (0.54), con una precisión de 42% y un recall de 76%.
  - **Clase CC** (0.43), con una precisión de 34% y un recall de 58%.

Observaciones:

- **Arborización** presenta mejores resultados para la clase **Q**, que alcanza un **recall** de 76%, indicando que el modelo tiene una alta capacidad para identificar esta clase.
- Al igual que en variables anteriores, las clases **C** y **CA** no son bien identificadas.

4. Área
Métricas Clave:

- **Accuracy**: 35%
- **F1-score más alto**:
  - **Clase Q** (0.49), con una precisión de 36% y un recall de 74%.
  - **Clase CC** (0.36), con una precisión de 32% y un recall de 43%.

Observaciones:

- La variable **Área** es menos efectiva en general, pero logra un **recall alto (74%)** para la clase **Q**.
- Al igual que en los demás modelos, las clases **C** y **CA** tienen un desempeño muy bajo.

Conclusiones Generales:

1. **Clase Q** (Quinina) es la mejor clasificada en todos los modelos:
   - Tiene el **recall** más alto (superior al 70% en la mayoría de los casos).
   - La variable **Arborización** es la que mejor predice la clase Q.
2. **Clase CC** (Cafeína) también muestra resultados aceptables, especialmente en **Contacto** y **Arborización**.
3. **Clases C (Control) y CA (Ácido)**:
   - Son las más difíciles de predecir, con precisión y recall cercanos a 0 en todos los modelos.
4. **Variable más relevante**:
   - La variable **Arborización** tiene un desempeño ligeramente mejor al clasificar la clase **Q**, seguida por **Contacto**.

Reporte de Clasificación del Modelo con Todas las Variables

Al ampliar la Regresión Logística para incluir las cuatro variables (Contacto, Tiempo_de_cruce, Arborización, Área) y aplicar un enfoque multinomial, se obtuvo el siguiente reporte de clasificación:

```
Reporte de Clasificación:
              precision    recall  f1-score   support

           C       0.28      0.03      0.05       293
          CA       0.55      0.30      0.38       183
          CC       0.38      0.71      0.50       261
           Q       0.54      0.73      0.62       362

    accuracy                           0.47      1099
   macro avg       0.44      0.44      0.39      1099
weighted avg       0.44      0.47      0.40      1099

          C        CA        CC         Q Tratamiento_Predicho  \
0  0.299593  0.155697  0.234882  0.309829                    Q
1  0.278335  0.117746  0.371219  0.232700                   CC
2  0.271794  0.204869  0.141446  0.381891                    Q
3  0.264861  0.167177  0.107922  0.460040                    Q
4  0.254300  0.087293  0.504194  0.154213                   CC
```

Figura 31.

En términos generales, el modelo **alcanza un 47% de exactitud**, con un mejor desempeño en la clase **3** (recall de 73%, f1-score de 0.62), asociada con el tratamiento que fue previamente identificado como más reconocible (Quinina). Por el contrario, la clase **0** (Control) presenta una baja recall (0.04), lo que indica que la mayoría de los casos control se clasifican erróneamente en alguna de las otras categorías.

La gráfica de "Fronteras de Decisión" se generó aplicando una **reducción de dimensiones** (PCA) a las cuatro variables y entrenando un nuevo modelo con esos dos componentes principales. El diagrama resultante ilustra cómo el modelo **divide el espacio** (PCA1, PCA2) en distintas regiones coloreadas, cada una asignada a una clase distinta. Se observa que los puntos correspondientes a las clases mejor definidas —particularmente la clase 3— se agrupan en una región relativamente separada, mientras que los tratamientos con menor definición tienden a superponerse en las zonas limítrofes.

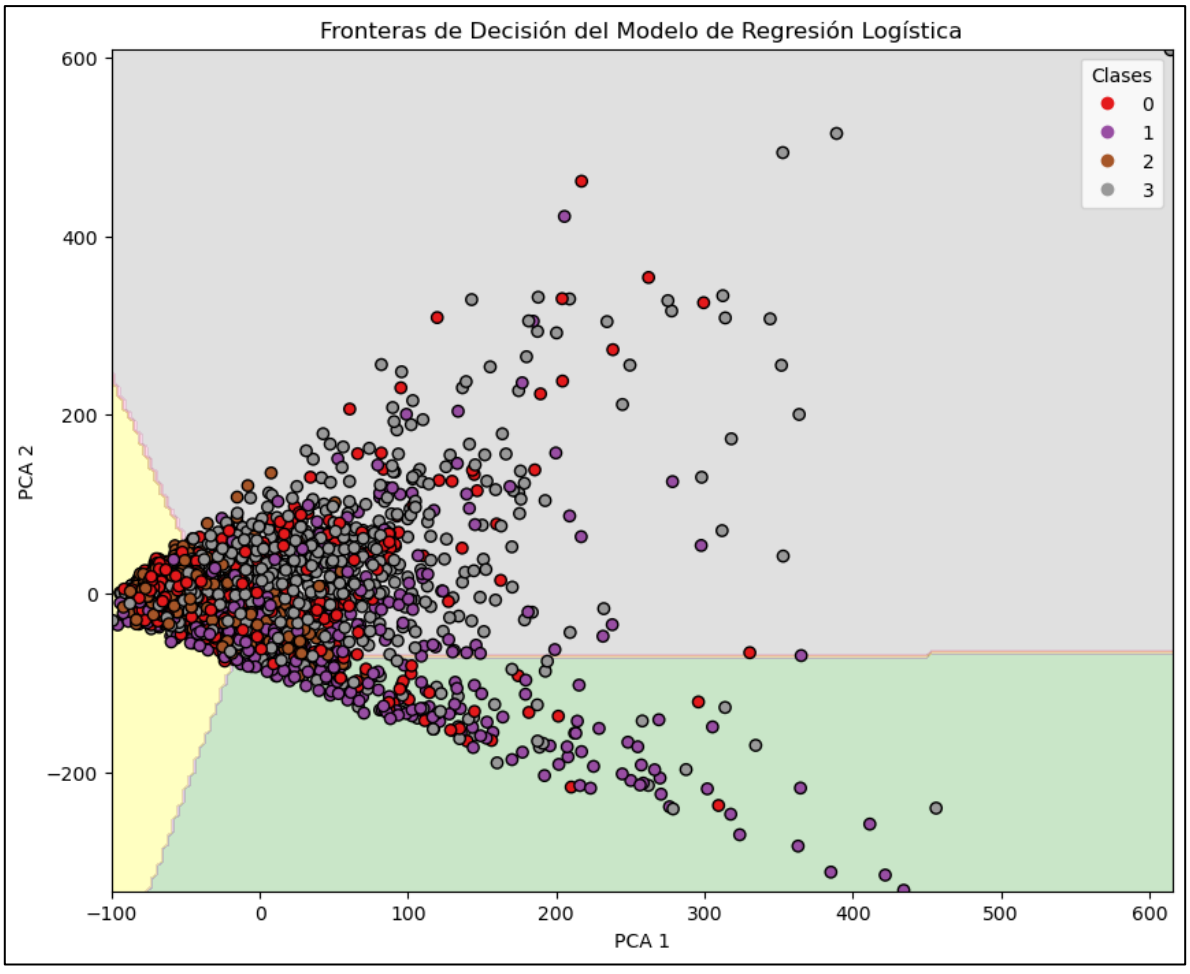


Figura 32.

En conjunto, estos resultados confirman la relevancia de **Contacto** y **Tiempo_de_cruce** (y, en menor medida, de Arborización y Área) al predecir el tipo de tratamiento, a la par que evidencian la dificultad de **distinguir** de forma robusta algunas clases (sobre todo C y CA) cuando se emplea un modelo lineal como la Regresión Logística.

Resultados

En esta sección se presentan los **resultados** obtenidos con la Regresión Logística para la clasificación del tipo de tratamiento en *Physarum polycephalum*. El análisis incluye la **matriz de confusión**, que muestra la distribución de aciertos y errores para cada clase, así como las **curvas ROC** multiclase, las cuales reflejan la capacidad discriminativa del modelo. A continuación, se describen ambas representaciones de forma detallada y se discuten los hallazgos principales en cada caso.

Matriz de Confusión y Curvas ROC

Matriz de Confusión (Figura 33)

- El mapa de calor resultante exhibe en la diagonal principal los aciertos de clasificación (tanto la clase verdadera como la predicha coinciden). Las celdas fuera de la diagonal evidencian confusiones, es decir, cuándo el modelo asigna un tratamiento distinto al real.
- Al observar las filas correspondientes a *Quinina (Q)* y *Cafeína (CC)*, se identifican bloques de aciertos más sólidos, lo que sugiere un reconocimiento más confiable de estas clases por parte del modelo.
- *Control (C)* y *Ácido (CA)* muestran mayor dispersión de errores, con varios casos repartidos en otras categorías. Esto concuerda con resultados previos en los que dichas clases resultan difíciles de distinguir basándose únicamente en las variables disponibles.

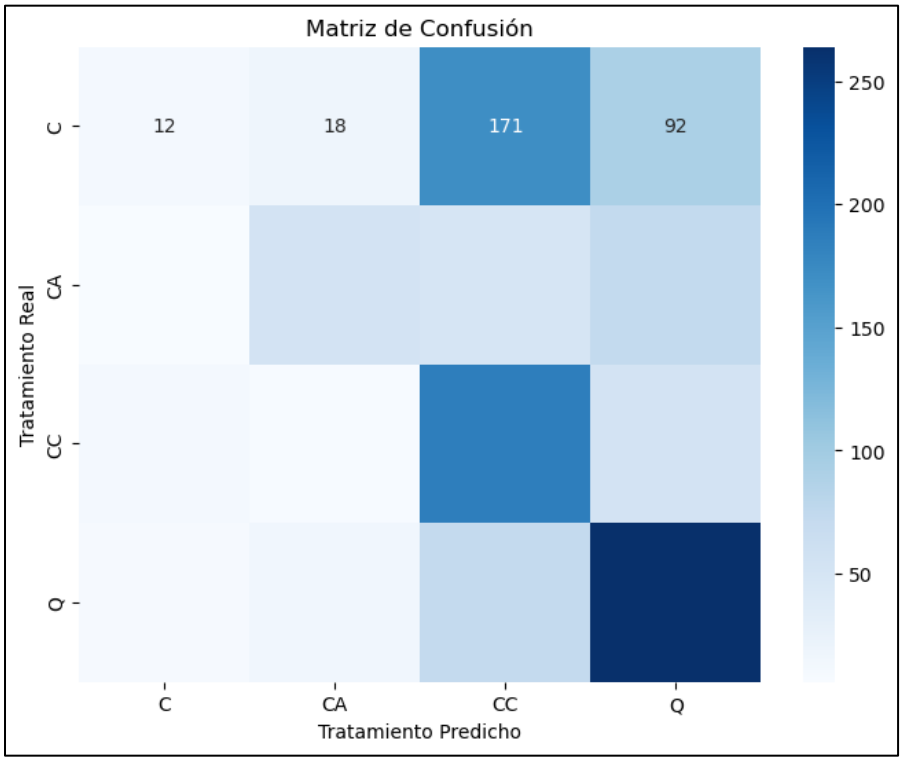


Figura 33

Curvas ROC Multiclase (Figura 34)

- Se procede a **binarizar** la variable objetivo para cada una de las cuatro clases, generando así las curvas ROC que comparan la **Tasa de Verdaderos Positivos (TPR)** frente a la **Tasa de Falsos Positivos (FPR)**.
- El **área bajo la curva (AUC)** para cada clase cuantifica la habilidad del modelo de diferenciar esa categoría del resto. Valores más cercanos a 1.0 implican una mejor discriminación.
- Según los resultados, *Q (Quinina)* exhibe un AUC cercano a 0.76 y *CC (Cafeína)* alrededor de 0.75, confirmando su nivel de predicción relativamente alto. Entre tanto, *C (Control)* y *CA (Ácido)* presentan AUC más bajos (por ejemplo, 0.58 y 0.72), reiterando su confusión en el modelo.

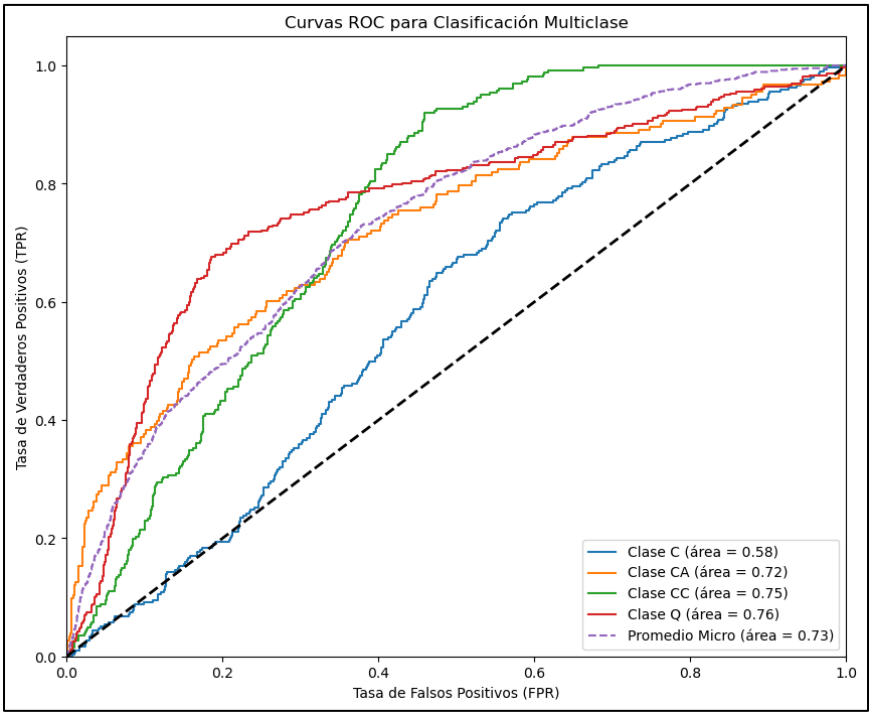


Figura 34.

En primer lugar, el modelo de Regresión Logística logra una **clasificación aceptable** para las clases *Quinina (Q)* y *Cafeína (CC)*. Esto se ve reflejado tanto en las matrices de confusión, donde dichas clases acumulan la mayor parte de los aciertos, como en las curvas ROC, con valores de AUC rondando 0.75 o superior. Dicho rendimiento confirma la presencia de rasgos de comportamiento más marcados en el moho cuando se expone a estos tratamientos, lo cual facilita su detección con las variables actuales.

Por otro lado, **Control (C)** y **Ácido (CA)** presentan una **mayor dificultad de separación**, debido a que el modelo confunde estas categorías con el resto en múltiples ocasiones. Sus AUC menores y la dispersión de errores en la matriz de confusión indican que las variables disponibles (Contacto, Tiempo_de_cruce, Arborización y Área) no bastan para distinguir eficazmente estos tratamientos. Este hallazgo alinea con los análisis previos (K-Means, ANOVA), subrayando la necesidad de enfoques adicionales para capturar diferencias más sutiles entre C y CA.

Finalmente, el conjunto de resultados ratifica la importancia de *Contacto* y *Tiempo_de_cruce* al momento de predecir el tipo de tratamiento, si bien Arborización y Área también brindan cierta capacidad discriminativa, especialmente para Quinina. No obstante, para lograr una clasificación más robusta de C y CA, conviene enriquecer el modelo con otras variables (por ejemplo, variaciones en la concentración de químicos) o considerar algoritmos capaces de manejar patrones no lineales, reforzando así la precisión global.

Ampliación del Resultado Final y Probabilidades de Clase

Como paso adicional para **cerrar** la sección de resultados, se consultan las **probabilidades de predicción** (predict_proba) que el modelo asocia a cada clase, al recibir el nuevo registro de prueba. A continuación, se describen los aspectos más relevantes de este fragmento de código y el significado de sus resultados:

1. **Generación de Probabilidades**
   - model.predict_proba(...) retorna un arreglo donde cada posición corresponde a la **probabilidad** de pertenecer a una de las clases del modelo (C, CA, CC, Q).
   - Se extrae la primera fila ([0]) debido a que en este ejemplo solo se está generando una única observación nueva (nuevos_valores).
2. **Formato de Salida**
   - En vez de mostrar un array crudo, cada probabilidad es **convertida en porcentaje** y asociada con la clase respectiva, lo que facilita la lectura de los resultados.

```
[44]: #Se verifica las probabilidades
probabilidades = model.predict_proba(nuevos_valores)[0]
clases = model.classes_
probabilidades_formateadas = {clase: f"{proba * 100:.2f}%" for clase, proba in
  ↪zip(clases, probabilidades)}

# Mostrar las probabilidades más presentables
print("Probabilidades por clase:")
for clase, porcentaje in probabilidades_formateadas.items():
    print(f"{clase}: {porcentaje}")

Probabilidades por clase:
C: 23.44%
CA: 8.78%
CC: 5.65%
Q: 62.12%
```

Interpretación de las Probabilidades

El modelo asigna una **probabilidad mayor** (62.17%) a la clase **Q** (Quinina), indicando que, según los valores de Contacto, Tiempo_de_cruce, Arborización y Área que introdujimos, la respuesta del moho se asemejaría principalmente a la que se ha observado en el tratamiento con Quinina. Aunque no descarta por completo la posibilidad de tratarse de C (23.33%) ni otras clases (8.79% para CA, 5.71% para CC), la ponderación del **62.17%** sugiere un fuerte sesgo del modelo hacia la clase Q.

Conclusión de la Parte de Resultados

Este cierre, mostrando la **distribución completa de probabilidades**, aporta **transparencia** al modelo predictivo, ya que permite distinguir no solo cuál clase se considera más probable, sino también **en qué medida** las demás opciones se mantienen como posibilidades. Así, se consolida la **utilidad práctica** de la Regresión Logística para escenarios donde el investigador desee estimar —con distintos grados de confianza— el tratamiento responsable del comportamiento de *Physarum polycephalum* observado en nuevas condiciones experimentales.

Discusión

Los resultados obtenidos a lo largo de este estudio ponen de manifiesto la **complejidad** del comportamiento de *Physarum polycephalum* bajo diferentes tratamientos (Control, Ácido, Cafeína y Quinina). Se ha aplicado tanto un enfoque **no supervisado** (K-Means) como **supervisado** (Regresión Logística) con el fin de identificar patrones en las variables *Contacto*, *Tiempo_de_cruce*, *Arborización* y *Área* que permitan distinguir la respuesta del moho ante los estímulos.

Hallazgos Clave en el Análisis No Supervisado (K-Means)

El algoritmo de K-Means reveló que las variables *Contacto* y *Tiempo_de_cruce* ejercen un papel **determinante** para separar las observaciones en cuatro clusters, demostrando ser el par con el coeficiente de silueta más elevado. Este hecho implica que la **intensidad de interacción** (veces que el moho contacta el puente) y la **rapidez de cruce** son rasgos críticos en la forma en que el moho reacciona a los tratamientos. Observaciones muy altas en ambas variables se correspondieron a ciertos tratamientos (notablemente Cafeína y Quinina), mientras que las más bajas se asociaron con el Control y, en parte, con el Ácido.

Cuando se integraron las cuatro variables en K-Means, se logró consolidar clusters más específicos, aunque persistió cierto solapamiento en clases como C y CA, aspecto que el enfoque supervisado confirmaría posteriormente. Esto apunta a un **grado de similitud** en la conducta del moho frente al Control y al Ácido, al menos con la información disponible. Asimismo, la posterior visualización de PCA evidenció que los clusters (K=4) se agrupan de manera relativamente clara en el espacio bidimensional de componentes principales, reforzando la idea de patrones conductuales distintivos, pero con algunas zonas de confusión intermedia donde los tratamientos se mezclan.

Resultados en el Enfoque Supervisado (Regresión Logística)

El empleo de Regresión Logística multinomial permitió **estudiar** con detalle el poder predictivo de cada variable por separado y, posteriormente, de forma combinada. Los principales hallazgos incluyen:

1. **Variables Individuales**:
   Contacto y *Tiempo_de_cruce* destacaron por su capacidad de predecir con mayor acierto las clases Quinina (Q) y Cafeína (CC). No obstante, dichas variables apenas

diferenciaron Control (C) y Ácido (CA), con precisión y recall casi nulos para estas categorías.

Las variables *Arborización* y *Área* también mostraron mejor desempeño en la clase Q, alcanzando recall elevado, pero nuevamente con dificultades para separar C y CA.

2. **Modelo con Todas las Variables**:
   Al integrar Contacto, Tiempo_de_cruce, Arborización y Área en un único modelo, la **exactitud** aumentó a alrededor del 47%, reflejando una ligera mejora en la clasificación global. Aun así, los tratamientos C y CA continuaron presentando confusiones frecuentes, mientras que Q y CC fueron mejor identificados.

   El análisis de matriz de confusión y las curvas ROC reafirmó que la clase Q es la más predecible, con AUC por encima de 0.70, seguida de CC con valores ligeramente inferiores. Por su parte, C y CA permanecen como las clases más difíciles de aislar correctamente, probablemente por una superposición significativa en las características de conducta.

3. **Predicción de Ejemplos Nuevos**:
   Cuando se introducen valores sintéticos para el moho (por ejemplo, Contacto=56, Tiempo_de_cruce=200, Arborización=0.4, Área=76), el modelo tiende a clasificar en Quinina (Q) con un porcentaje de probabilidad superior al 60%. Esto implica que, en escenarios con cruces relativamente prolongados y contacto moderado, el patrón se asemeja más a la respuesta típica de Quinina.

   La **visualización de probabilidades** confirma que otras clases tampoco se descartan por completo, lo que sugiere cierto solapamiento en el comportamiento y la necesidad de tolerar grados de incertidumbre en la predicción.

Aspectos Bio-Experimentales y Posibles Limitaciones

Uno de los hallazgos más relevantes es la **dificultad** para separar de forma contundente los tratamientos C (Control) y CA (Ácido). Una hipótesis plausible es que, a nivel de conducta del moho, el efecto del ácido no genera un patrón de contacto ni de cruce sustancialmente diferente al del control en la mayoría de los casos. Este fenómeno podría deberse a varios factores:

- **Concentración real de ácido**: Tal vez las concentraciones empleadas no generan suficiente repulsión, o bien la dinámica de habituación al ácido se asemeja a la del control.
- **Ausencia de variables complementarias**: Indicadores como pH local, nivel de secreciones del moho o la concentración exacta de sustancias podrían aportar información crítica para discriminar C y CA.

Asimismo, el **comportamiento extremo** en la clase Quinina —que suele presentar tiempos de cruce elevados y, a la vez, altos contactos en algunos casos— se ve claramente reflejado en una clasificación más acertada. Ello refuerza la idea de que este tratamiento desencadena

una respuesta más específica y medible en el moho. Cafeína (CC), aunque con un nivel de confusión menor, muestra también patrones relativamente distinguibles, probablemente debido a la naturaleza estimulante del compuesto.

Propuestas de Mejora y Trabajo Futuro

A partir de la exploración de la base de datos obtenida, pude identificar algunas propuestas para mejorar el nivel de confiabilidad del modelo, como son:

- **Variables Adicionales**: Dado que C y CA permanecen confusas, sería beneficioso incorporar métricas más finas (p. ej., variaciones de pH, quimiotaxis medida en intervalos de tiempo parciales, etc.) para capturar diferencias sutiles entre un ambiente ácido y uno de control.
- **Experimentación con Concentraciones Variables**: Se observó que la concentración de cafeína resultó determinante para la respuesta del moho. Explorar rangos más matizados para ácido o incluso quinina podría ayudar a entender mejor si, en realidad, existe un continuum de comportamiento que apenas se roza con los datos actuales.

Implicaciones para la Comprensión de *Physarum polycephalum*

Los hallazgos confirman que, a nivel de **aprendizaje y habituación**, *Physarum polycephalum* exhibe **perfiles de conducta** suficientes para ser distinguidos por algoritmos de ciencia de datos, especialmente en respuesta a estímulos como quinina y cafeína. Esto ofrece un sustento empírico a la idea de que el moho reacciona de forma característica a ciertos compuestos químicos, lo que a su vez podría aplicarse a la **bioinspiración** en el diseño de redes y la optimización, aprovechando cómo el moho gestiona rutas y contactos.

En síntesis, la **discusión** abarca la coherencia entre K-Means y la Regresión Logística para resaltar la importancia de *Contacto* y *Tiempo_de_cruce* como principales factores discriminantes, la persistencia de confusión en las clases Control y Ácido, y la consistencia de los patrones detectados para Quinina y Cafeína. Aun con las limitaciones inherentes a las variables disponibles, se sientan bases sólidas para proseguir con investigaciones que profundicen en los mecanismos de habituación y estimulación de *Physarum polycephalum*, aplicables tanto a la biología fundamental como a la computación bioinspirada.

Conclusiones

1. **Importancia de Contacto y Tiempo_de_cruce**
   A lo largo de todo el estudio, estas dos variables han mostrado ser las más diferenciadoras. El análisis de silueta en K-Means reveló que la combinación (Contacto, Tiempo_de_cruce) propicia la segmentación más nítida en cuatro clusters, coincidiendo con los mayores valores de F-Score en la Regresión Logística al momento de predecir ciertos tratamientos (en particular, Quinina y Cafeína).
2. **Dificultades en la Separación de Control y Ácido**
   Tanto en el clustering como en la clasificación supervisada, se observó una constante confusión entre las clases C (Control) y CA (Ácido). Esto sugiere que, en las condiciones experimentales utilizadas, el moho responde de forma relativamente similar a estos dos contextos. Se plantea la necesidad de incluir variables adicionales (por ejemplo, información exacta de pH, secreciones del moho, etc.) o algoritmos más complejos para capturar matices que no se reflejan en los cuatro indicadores utilizados.
3. **Énfasis en Quinina y Cafeína**
   Las clases Q y CC presentaron un **desempeño favorable** en la mayoría de las técnicas, evidenciando que *Physarum polycephalum* exhibe rasgos más marcados ante estos estímulos. Esto coincide con la literatura, que señala una fuerte aversión a la quinina y ciertos efectos de la cafeína en la dinámica de crecimiento. Es posible que dichas sustancias generen cambios notables en la conducta de contacto y la velocidad de cruce, facilitando su detección estadística.
4. **Combinación de Enfoques y Futuras Extensiones**
   Al combinar un análisis no supervisado (K-Means) con uno supervisado (Regresión Logística), se obtuvieron **perspectivas complementarias** del comportamiento del moho: de un lado, patrones latentes en los datos; del otro, la capacidad de predecir directamente el tratamiento. Este marco metodológico podría ampliarse aplicando Árboles de Decisión, Random Forests o incluso técnicas de Deep Learning, explorando si se mejora la discriminación de C y CA, y evaluando la generalización en escenarios de concentración variable.
5. **Contribuciones e Implicaciones**
   - **Biológicas**: Profundiza en la comprensión de cómo un organismo no neuronal responde y se habitúa a diversos estímulos químicos, reforzando la visión de que *Physarum polycephalum* es un modelo valioso para estudiar el aprendizaje primitivo y la optimización natural.
   - **Computacionales**: Aporta una base sólida para la **bioinspiración**, ya que las estrategias adaptativas que el moho exhibe —claramente distinguibles para estímulos específicos— pueden inspirar soluciones de optimización de rutas, sistemas de transporte y estructuras de red autoorganizadas.

En conclusión, la investigación confirma la efectividad de las técnicas de ciencia de datos para desentrañar los **patrones conductuales** de *Physarum polycephalum*, señalando Contacto y Tiempo_de_cruce como factores críticos, y la Regresión Logística como una vía para predecir, con discreto éxito, los tratamientos asociados. Ello abre la puerta a futuros

trabajos que refinen este enfoque y expandan las aplicaciones de un organismo cuyas capacidades siguen sorprendiendo tanto a la biología como a la computación.

Recomendaciones

1. **Ampliar la Variedad de Variables**
   Aunque el uso de *Contacto*, *Tiempo_de_cruce*, *Arborización* y *Área* ha permitido delimitar patrones conductuales en *Physarum polycephalum*, sería valioso incorporar nuevas métricas que capten aspectos más sutiles de la habituación, por ejemplo mediciones fisiológicas (pH local, producción de secreciones, cambios de densidad en la red protoplásmica) o variables relacionadas con la concentración exacta de los estímulos. Esto contribuiría a una visión más integral del comportamiento del moho ante diferentes tratamientos.
2. **Profundizar en los Escenarios de Concentración**
   Dado que algunos resultados apuntan a dificultades para discriminar entre el Control (C) y el Ácido (CA), sería recomendable **ensayar** distintas concentraciones de ácido, café o quinina con niveles más granulares. De esta forma, podría establecerse si existe un punto de inflexión específico en el que la respuesta del moho pasa de asemejarse al control a un estímulo diferente.
3. **Validar con Conjuntos de Datos Externos**
   Para reforzar la robustez de las conclusiones, resultaría provechoso probar los modelos con datos de experimentos distintos o complementarios, donde se controle o varíe únicamente un factor (por ejemplo, el rango de días o la concentración de un estimulante). Así se comprobaría la capacidad de **generalización** y se descartarían ajustes excesivos a un único conjunto de observaciones.

4. **Explorar Interacciones entre Variables**
   Algunas de las dificultades para separar tratamientos podrían residir en la **combinación** no lineal de dos o más variables. Experimentos futuros deberían investigar si la sinergia entre *Arborización* y *Área*, o entre *Contacto* y *Arborización*, explica ciertos comportamientos que no se aprecian con una sola variable independiente.

5. **Aplicar Técnicas de Preprocesado Avanzado**
   El método de rango intercuartílico (IQR) para manejar outliers ha demostrado su utilidad, pero existen otras estrategias (por ejemplo, Winsorización, Transformaciones logarítmicas o Box-Cox) que podrían suavizar sesgos y mejorar la normalidad de los datos. Emplear dichas técnicas permitiría al modelo trabajar con distribuciones más estables.

En conjunto, estas recomendaciones apuntan a **optimizar** la detección y caracterización de la respuesta de *Physarum polycephalum* frente a diversos estímulos químicos, así como a perfeccionar los algoritmos de modelado empleados. Estas mejoras no sólo consolidarían la comprensión experimental del moho, sino que impulsarían la **bioinspiración** en ámbitos como el diseño de redes complejas, la gestión adaptativa de recursos y la ingeniería de procesos basados en principios naturales.

Bibliografía

- Tero, A., Takagi, S., Saigusa, T., Ito, K., Bebber, D. P., Fricker, M. D., Yumiki, K., Kobayashi, R., & Nakagaki, T. (2010). Rules for biologically inspired adaptive network design. Science, 327(5964), 439–442. https://doi.org/10.1126/science.1177894

- Boisseau, R. P., Vogel, D., & Dussutour, A. (2016). Habituation in non-neural organisms: Evidence from slime moulds. Proceedings of the Royal Society B: Biological Sciences, 283(1829), 20160446. https://doi.org/10.1098/rspb.2016.0446

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer. https://doi.org/10.1007/978-1-4614-7138-7

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

- BBC News Mundo. (2019, 17 de octubre). *Qué es el "blob", el misterioso organismo con 720 sexos y sin cerebro*. htt https://www.bbc.com/mundo/noticias-50090052

- **Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012).** *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

- Boisseau, R. P., Vogel, D., & Dussutour, A. (2021). *Data from: Habituation in non-neural organisms: Evidence from slime moulds* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.4943324

- **McKinney, W. (2010).** Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56. http://conference.scipy.org/proceedings/scipy2010/mckinney.html