PBI

POC DE DATASETS

Análisis del Dataset de Cultivos y Presencia de Moscas y visualización en Power BI

Carmen Felipe Navarro

ÍNDICE

1. Introducción

En el mundo de la agricultura, el manejo adecuado de los cultivos y el control de plagas son fundamentales para asegurar una producción eficiente y sostenible. Este trabajo se centra en el análisis de un dataset obtenido del CIES (Centro de Investigación, Experimentación y Servicios del Champiñón) de Quintanar del Rey, que reúne información sobre cultivos de champiñones y la presencia de moscas en ellos. A través de este análisis, se busca identificar patrones y correlaciones que pueden ser útiles para mejorar la gestión agrícola.

El dataset incluye variables que describen las condiciones del cultivo, la ubicación, la estacionalidad y el impacto de las moscas en la producción. Todas las variables están en inglés, dado que el dataset está orientado a su publicación en una revista internacional.

La relevancia de este análisis radica en la importancia de los cultivos de champiñones, que son un recurso valioso en la agricultura, y en el impacto que las moscas pueden tener sobre su producción. Comprender cómo interactúan estos factores es crucial para desarrollar estrategias efectivas de manejo de plagas y mejorar la productividad en el sector.

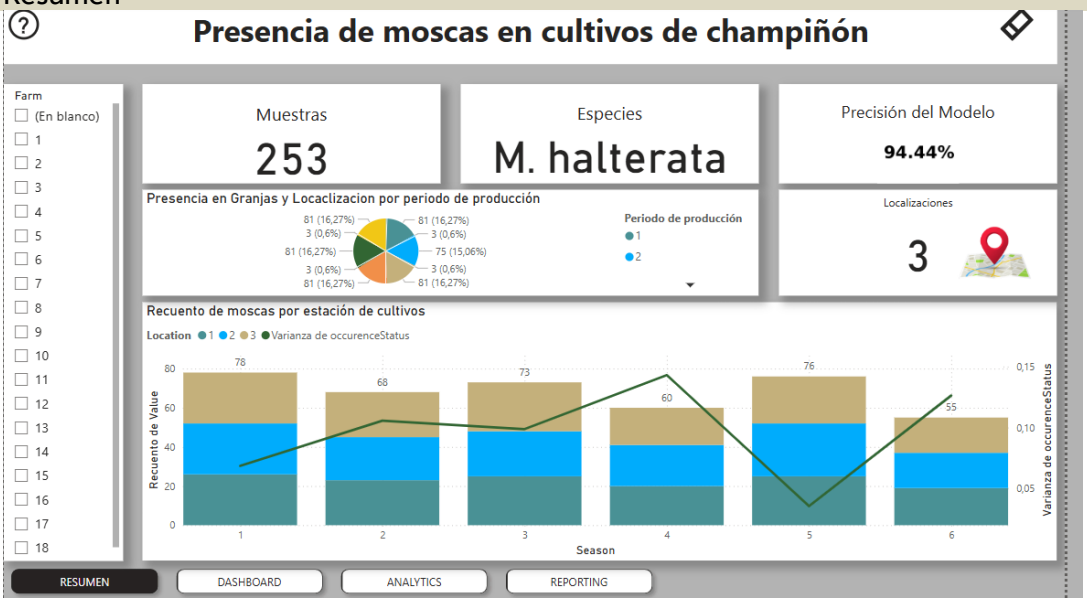
Variables

Las principales columnas del dataset son:

- **Species**: Tipo de mosca encontrada.
- **OccurenceStatus**: Indica si la mosca está presente (1) o ausente (0).
- **Value**: Nivel de impacto de las moscas en los cultivos.
- **Farm**: Identificación de la finca donde se tomaron las muestras.
- **Season**: Temporada en la que se registraron los datos.
- **Growin Stage**: Etapa de crecimiento del cultivo.
- **Location**: Ubicación específica dentro de la finca.

2. Análisis en Power BI

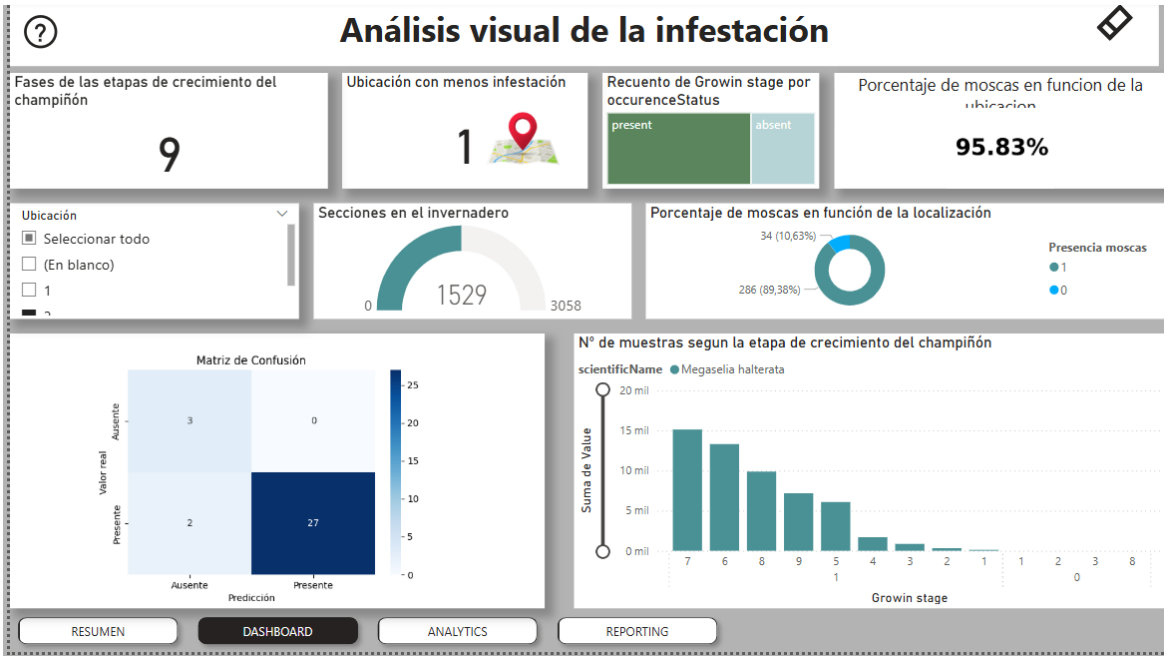
Resumen



El resumen muestra una visión general de la presencia de moscas en cultivos de champiñón. Se analizaron 253 muestras, detectando la especie M. halterata, con un modelo de predicción del 94.44% de precisión. Además, las infestaciones se distribuyen en 3 ubicaciones y varían según el periodo de producción.

Los gráficos reflejan cómo cambian las infestaciones por estación de cultivo y ubicación. Se observa que la cantidad de moscas fluctúa a lo largo del tiempo, con algunos picos y caídas. También hay un filtro por granjas que permite analizar datos específicos.

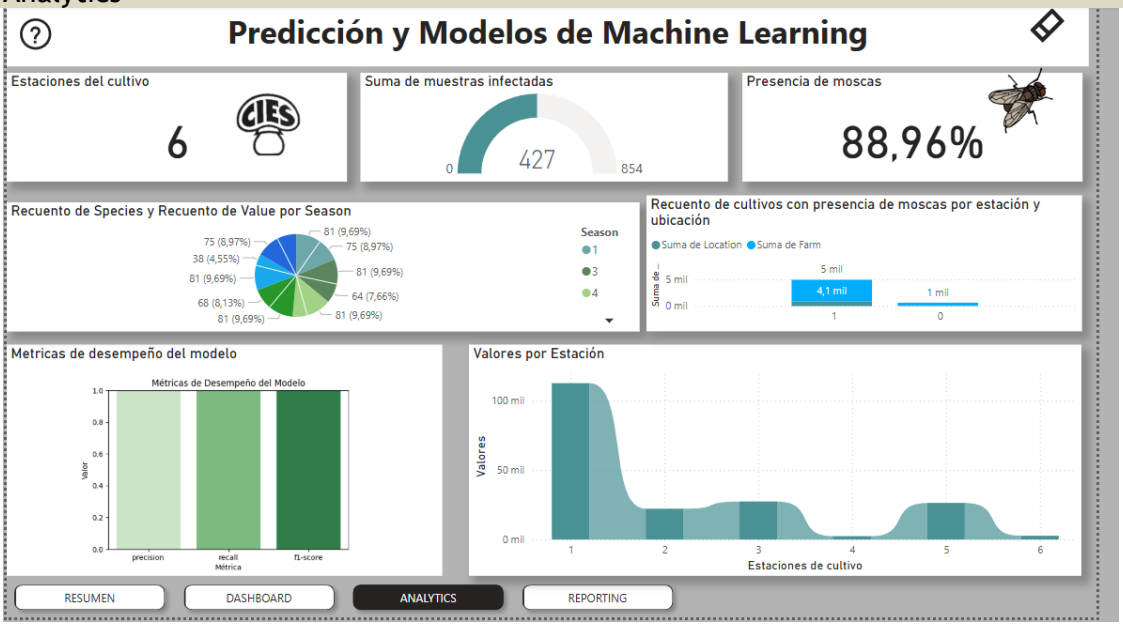
Dashboard



El dashboard "Análisis visual de la infestación" presenta un análisis detallado sobre la presencia de moscas en cultivos de champiñón, destacando métricas clave y visualizaciones interactivas. Se identifican 9 fases de crecimiento, con una ubicación específica que muestra la menor infestación. Además, se observa que el 95.83% de las infestaciones están influenciadas por la ubicación.

Entre las visualizaciones más relevantes, el gráfico de barras muestra la relación entre las etapas de crecimiento y la presencia de infestación, mientras que el gráfico de anillo indica que el 10.63% de las muestras presentan moscas. La matriz de confusión evalúa el rendimiento del modelo predictivo, evidenciando una buena precisión con solo 2 falsos negativos y ningún falso positivo.

Analytics



La parte de analytics se enfoca en la predicción y modelos de Machine Learning para detectar la presencia de moscas en cultivos de champiñón. Se identificaron 6 estaciones de cultivo, con 427 muestras infectadas de un total de 854. La presencia de moscas es alta, alcanzando un 88.96%.

El modelo de predicción tiene buenas métricas de desempeño, con valores altos en precisión, recall y F1-score, lo que indica que funciona bien para detectar infestaciones. Además, se muestra la distribución de especies por estación y el recuento de infestaciones por ubicación, con una diferencia notable entre estaciones, por lo que ayuda a entender el comportamiento de la infestación y la efectividad del modelo.

Reporting

Este reporting ofrece un análisis detallado de los datos de infestación, con 253 muestras de la especie Megaselia halterata, organizadas en tablas con información sobre ubicación, etapa de crecimiento y presencia de infestación.
Se complementa con tablas y gráficos que muestran la desviación de presencia por etapa de crecimiento, ayudando a identificar los momentos más críticos. También incluye una matriz sobre la presencia de moscas por granja y valor, destacando las áreas con mayor impacto.

Además, hay un enlace para descargar los datos en Excel, permitiendo un análisis más profundo y facilitando la toma de decisiones basada en información detallada.

3. ANÁLISIS CON PYTHON

Analizar este archivo con Python es súper útil porque nos ayuda a entender mejor la presencia de moscas Phoridae en distintas granjas, temporadas y etapas de crecimiento. Básicamente, los datos muestran si una especie está presente o no, en qué lugar y en qué cantidad. Si organizamos bien esta información, podemos encontrar patrones y sacar conclusiones interesantes.

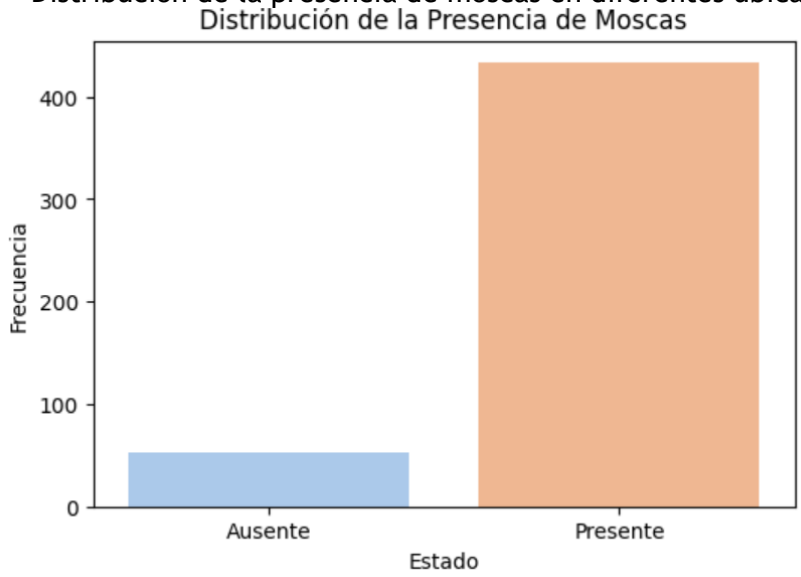
Lo primero que hay que hacer es limpiar los datos, porque en muchos archivos CSV suele haber errores, datos faltantes o formatos raros. Con Python, especialmente usando Pandas, se puede estructurar todo bien y asegurarse de que la información sea confiable antes de analizarla.

También es clave sacar estadísticas básicas, como cuántas veces aparece cada especie, en qué lugares se encuentran más seguido o cómo cambia su presencia según la temporada. Esto nos ayuda a identificar tendencias y entender mejor qué factores pueden estar influyendo en su aparición.
Otro punto importante es la visualización. Con Python se pueden hacer gráficos que muestran de forma clara la distribución de los datos, lo que hace mucho más fácil detectar patrones sin tener que revisar números uno por uno.

En resumen, usar Python para analizar este archivo nos ahorra tiempo, organiza la información y nos permite descubrir cosas que, de otra forma, serían difíciles de ver. Es una herramienta súper potente para entender qué está pasando con estas especies y cómo afectan su entorno.

3.1 Estadísticas descriptivas

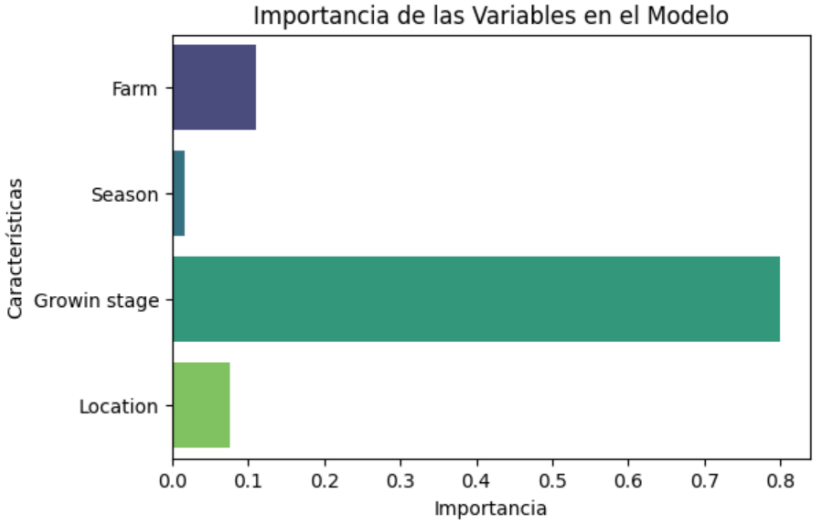
- Distribución de la presencia de moscas en diferentes ubicaciones y temporadas.



Este gráfico de barras muestra cuántas veces se registraron cultivos con moscas ("Presente") y sin moscas ("Ausente") en el dataset.
Si hay muchos más casos de "Ausente" que "Presente", el dataset está desequilibrado, lo que puede afectar al modelo. Si los valores están balanceados, el modelo tendrá datos más equitativos para aprender.

3.2 Probabilidades y tendencias

- Probabilidad de ocurrencia de moscas según temporada, ubicación y etapa de crecimiento.



Muestra qué variables son más importantes para el modelo a la hora de predecir la presencia de moscas.

- Variables con valores altos → Son las más influyentes en la predicción.
- Variables con valores bajos → Apenas afectan la decisión del modelo.
- Si la variable "Ubicación" tiene alta importancia → La zona influye en la aparición de moscas.
- Si la variable "Etapa de crecimiento" tiene alta importancia → Ciertas etapas favorecen la presencia de moscas.

3.3 Modelo de Machine Learning

3.3.1 Árbol de decisión

Modelo utilizado

Se utilizó un Árbol de Decisión para predecir la presencia de moscas en distintas ubicaciones y condiciones. Este modelo permite analizar cómo distintas variables influyen en la aparición de las moscas y facilita la interpretación de los resultados mediante reglas de decisión.
Además, se implementó un segundo modelo basado en regresión para estimar el impacto en términos de cantidad de moscas presentes.

Procesamiento de datos

Antes de entrenar los modelos, se realizaron los siguientes pasos:

- Se convirtió la variable occurenceStatus a valores numéricos (0 = Absent, 1 = Present).
- Se convirtió la variable Value a tipo numérico y se eliminaron los valores nulos.
- Se seleccionaron las variables predictoras: Farm, Season, Growin stage y Location.
- Se dividieron los datos en conjunto de entrenamiento (80%) y conjunto de prueba (20%).

Entrenamiento del modelo

- Se entrenó un Árbol de Decisión de Clasificación con una profundidad máxima de 4 para predecir la presencia de moscas.
- Se entrenó un Árbol de Decisión de Regresión con la misma profundidad para predecir el impacto de su presencia en términos de cantidad.

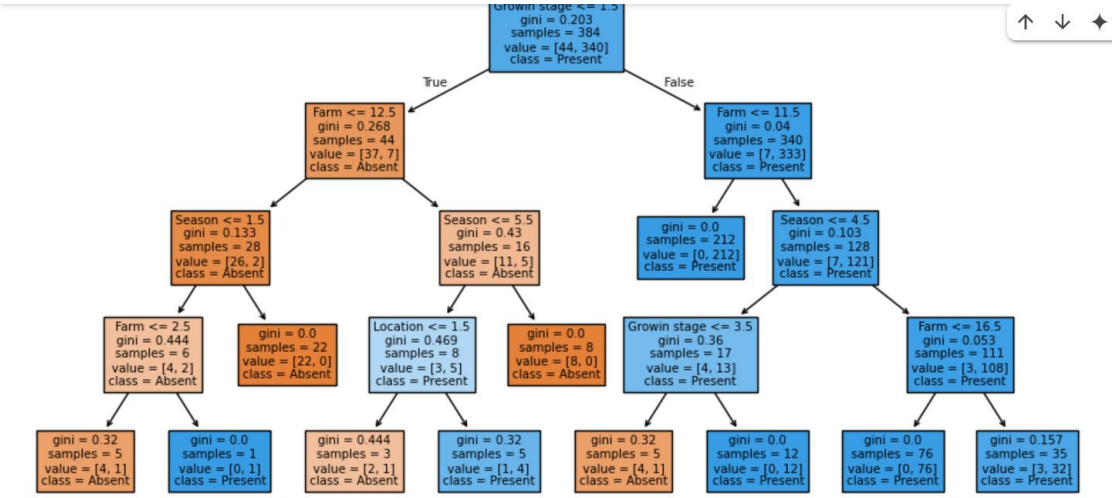
Evaluación del modelo

- Precisión del modelo de clasificación: Se obtuvo una precisión del 99% en la predicción de presencia de moscas.
- Error cuadrático medio (MSE) del modelo de regresión: Se obtuvo un error cuántico medio (predicción de impacto) del 2703.19 en la predicción del número de moscas.

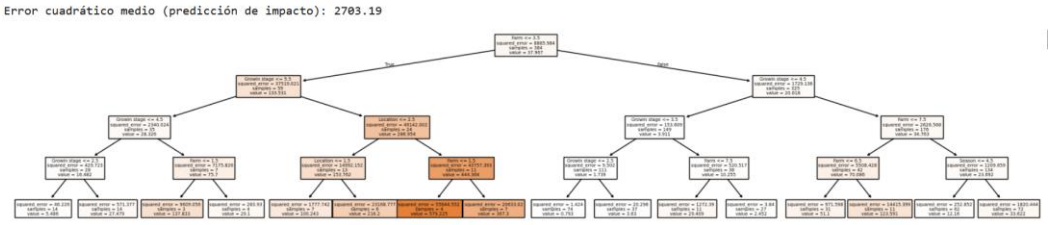
Visualización y resultados

A continuación, se muestran los árboles de decisión generados, los cuales indican los criterios utilizados para clasificar la presencia de moscas y estimar su impacto. En la gráfica se pueden observar los nodos de decisión y los factores más relevantes en la predicción.

Modelo de Clasificación

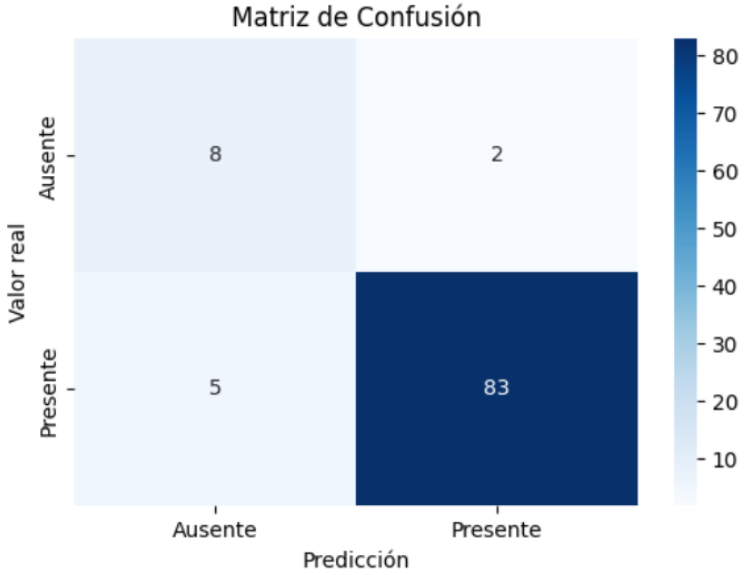


Modelo de Regresión



3.3.2 Matriz de confusión

La matriz de confusión se generó para evaluar el rendimiento del modelo de clasificación en la predicción de la presencia de moscas. A continuación, se muestra la matriz de confusión obtenida, que compara las predicciones realizadas por el modelo con las etiquetas reales.



A partir de esta matriz, se pueden calcular las métricas de rendimiento que se detallan a continuación.

Precisión (%) y Métricas Adicionales

El modelo alcanzó una **precisión** del 92.86%. Las métricas adicionales que permiten un análisis más detallado del rendimiento del modelo son las siguientes:

Precisión (Accuracy)

$$\text{Precisión} = \frac{VP + VN}{VP + VN + FP + FN} = \frac{83 + 8}{83 + 8 + 5 + 2} = 92.86\%$$

Recall (Sensibilidad)

$$\text{Recall} = \frac{VP}{VP + FN} = \frac{83}{83 + 5} = 0.94\,(94\%)$$

Precision (Precisión)

$$\text{Precision} = \frac{VP}{VP + FP} = \frac{83}{83 + 2} = 0.98\,(98\%)$$

F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.98 \times 0.94}{0.98 + 0.94} = 0.96\,(96\%)$$

Resultados y conclusiones

```
Precisión del modelo: 92.86%

Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.62      0.80      0.70        10
           1       0.98      0.94      0.96        88

    accuracy                           0.93        98
   macro avg       0.80      0.87      0.83        98
weighted avg       0.94      0.93      0.93        98
```

Este reporte muestra que el modelo es extremadamente preciso al predecir la clase positiva (presencia de moscas), con un recall y un F1-score muy altos para la clase "1" (presencia de moscas). Sin embargo, para la clase "0" (ausencia de moscas), el modelo presenta una precisión más baja (0.62), lo que sugiere que tiene dificultades para detectar correctamente los casos en los que las moscas están ausentes, aunque el recall para esa clase es más alto (0.80), lo que significa que no está dejando pasar demasiados casos negativos.

4. CONCLUSIONES

En este trabajo, analicé cómo las moscas afectan el cultivo de champiñón y qué factores influyen en su presencia. Usando Power BI y modelos de Machine Learning, pude identificar patrones importantes, como que la ubicación y la etapa de crecimiento del cultivo tienen un gran impacto en la cantidad de infestaciones.

Los datos muestran que el 95.83% de las infestaciones están relacionadas con la ubicación, y que hay ciertas etapas del crecimiento donde las moscas aparecen con más frecuencia. Además, el modelo predictivo que utilicé tuvo una precisión de más del 90%, lo que indica que es bastante efectivo para detectar infestaciones.

Para complementar el análisis, incluí tablas y gráficos que facilitan la interpretación de la información. También agregué un enlace para descargar los datos en Excel, por si alguien quiere analizarlos con más detalle.

5. BIBIOGRAFIA

Navarro Lozano, M. J. (2002). *Biología y control del ácaro miceliófago Brennandania lambi (Krczal) (Acari: Pygmephoridae) en los cultivos de champiñón de Castilla-La Mancha* (Tesis doctoral, Universidad de Castilla-La Mancha). Dialnet. https://dialnet.unirioja.es/servlet/tesis?codigo=71945

Navarro, M. J., Escudero, A., Ferragut, F., López-Lorrio, A., & Gea, F. J. (2001). *Evolución de las poblaciones de los dípteros Megaselia halterata y Lycoriella auripila (Diptera: Phoridae y Sciaridae) en el cultivo de champiñón de Castilla-La Mancha.* Boletín de Sanidad Vegetal. Plagas, 27, 373-381. Recuperado de https://www.mapa.gob.es/ministerio/pags/biblioteca/plagas/bsvp-27-03-373-381.pdf