ABORDAJE METODOLÓGICO DE MINERÍA DE DATOS APLICADA A INVESTIGACIONES AGROPECUARIAS SOBRE ARTRÓPODOS

Mariano R. Droz [1,2], Carlos E. Alvez [1], Pedro D. Benitez [2], Juan A. Ramos [2], Beatriz M. Diaz [3], Luz M. Zapata [2]

[1] Facultad de Ciencias de la Administración – Universidad Nacional de Entre Ríos
Av. Monseñor Tavella 1424 – (CP 3.200) Concordia, Entre Ríos, República Argentina

[2] Facultad de Ciencias de la Alimentación – Universidad Nacional de Entre Ríos
Av. Monseñor Tavella 1450 – (CP 3.200) Concordia, Entre Ríos, República Argentina

[3] Estación Experimental Agropecuaria Concordia – Instituto Nacional de Tecnología Agropecuaria
Ruta Provincial 22 y vías del Ferrocarril – (CP 3.200) Estación Yuquerí, Concordia, Entre Ríos, República Argentina

{mariano.droz, carlos.alvez, pedrodaniel.benitez, juan.ramos, luzmarina.zapata}@uner.edu.ar

diaz.beatriz@inta.gob.ar

RESUMEN

La Minería de Datos (*Data Mining*) trata de resolver problemas o comprender fenómenos o situaciones mediante el análisis de datos digitales. Minería de Datos y extracción o "descubrimiento" de conocimiento en bases de datos (*Knowledge Discovery in Databases, KDD*) se han empleado indistintamente, pero existen diferencias entre ambas, siendo la primera solamente el núcleo de un proceso KDD. No obstante, actualmente Ciencia de Datos (*Data Science*) es una expresión mucho más utilizada en el contexto de descubrimiento de conocimiento a partir de datos, por eso en este trabajo se aborda un proyecto de Minería de Datos con un enfoque de ciencia de datos aplicada. Llevar a cabo un proyecto de este tipo requiere de un abordaje metodológico, se ha optado por CRISP-DM (*CRoss Industry Standard Process for Data Mining*) ya que es considerado como el estándar de facto para proyectos de analítica, Minería de Datos y Ciencia de Datos. El objetivo consiste en ejecutar un proyecto para descubrir, determinar o relacionar la incidencia de los factores abióticos (temperatura y humedad relativa ambiente, nivel de luz, y temperatura y humedad del suelo) en el comportamiento de la artropofauna edáfica existente en sistemas productivos hortícolas de la región de Salto Grande.

Palabras clave: *Bases de Datos, Minería de Datos, Ciencia de Datos, CRISP-DM, Artrópodos.*

CONTEXTO

Este trabajo es desarrollado en el marco de las actividades de la Tesis denominada "Minería de Datos aplicada a estudios de biodiversidad de artrópodos de suelo" de la Maestría en Sistemas de Información (Facultad de Ciencias de la Administración. Universidad Nacional de Entre Ríos. Resolución "C.D." N° 144/20). A su vez, está vinculado al Proyecto PID Novel N° 8112 "Diseño y desarrollo de una trampa de caída por tiempo con sensores y datalogger para estudios de biodiversidad de artrópodos de suelo" de la Universidad Nacional de Entre Ríos (UNER).

Por otra parte, se enmarca en la línea de trabajo "Desarrollo de tecnologías de bajo impacto ambiental aplicadas a la horticultura", que lleva a cabo el Grupo Hortícola de la Estación Experimental Agropecuaria Concordia (EEA Concordia) del Instituto Nacional de Tecnología Agropecuaria (INTA).

1. INTRODUCCIÓN

De acuerdo con [1] la Minería de Datos (*Data Mining*, *DM*) trata de resolver problemas

mediante el análisis de datos existentes en bases de datos. A través del tiempo, varias expresiones se han utilizado para hacer referencia a la Minería de Datos, entre ellas: descubrimiento de conocimiento en bases de datos *(Knowledge Discovery -mining- in Databases, KDD)*, extracción de conocimiento *(knowledge extraction)*, análisis de patrones *(data/pattern analysis)*, arqueología de datos *(data archeology)* y recolección de información *(information harvesting)* [2, 3].

Si bien Minería de Datos y extracción o "descubrimiento" de conocimiento en bases de datos (KDD) se han empleado indistintamente, existen diferencias entre ambas. El término KDD se utiliza para referirse a un proceso que consta de una serie de fases, mientras que la Minería de Datos es sólo una de estas fases [4].

Así, la Minería de Datos es el núcleo de un proceso KDD, ya que implica aplicar algoritmos para explorar los datos, desarrollar modelos y descubrir patrones previamente desconocidos. Estos modelos resultantes son utilizados para analizar, comprender o predecir fenómenos a partir de los datos [5].

Sin embargo, actualmente Ciencia de Datos (*Data Science*) es una expresión mucho más utilizada que Minería de Datos en el contexto de descubrimiento de conocimiento a partir de datos [6].

Ahora bien, ¿qué es la Ciencia de Datos?, según [6] existen dos perspectivas: a) ciencia *de* los datos; y b) aplicar métodos científicos *a* los datos. Para estos autores estos dos enfoques podrían llamarse: a) ciencia de datos teórica; y b) ciencia de datos aplicada.

Desde el punto de vista de la primera perspectiva, se hace un abordaje académico, se estudian los datos en todas sus manifestaciones, junto con los métodos y algoritmos para manipular, analizar, visualizar y enriquecer los datos. Esta perspectiva es metodológicamente cercana a la informática y a la estadística, combinando trabajo teórico, algorítmico y empírico [6].

En tanto, desde la segunda perspectiva, la ciencia de datos abarca tanto el enfoque académico como el industrial, extrayendo valor de los datos utilizando métodos científicos, como las pruebas de hipótesis estadísticas o el aprendizaje automático. Aquí el énfasis está en resolver problemas específicos de un dominio en función de los datos. Éstos se utilizan para construir modelos, diseñar artefactos y, en general, aumentar la comprensión de un tema [6].

A partir de lo anterior, en este trabajo se hace referencia a un proyecto de Minería de Datos como un proceso de extracción de conocimiento a partir de datos, lo que también se considera ciencia de datos aplicada.

Un proyecto de Minería de Datos de este tipo es recomendable que sea abordado siguiendo una metodología. Así, CRISP-DM [7] (*CRoss Industry Standard Process for Data Mining*), presentada en el año 1999 por las empresas NCR, SPSS y Daimler Chrysler, es actualmente considerado el estándar de facto para proyectos de analítica, Minería de Datos y Ciencia de Datos [6].

CRISP-DM elabora y amplía las fases de la propuesta original de KDD en seis fases: Comprensión o entendimiento del negocio (*Business understanding*), Comprensión de los datos (*Data understanding*), Preparación de los datos (*Data preparation*), Modelización (*Modelling*), Evaluación (*Evaluation*) y Despliegue (*Deployment*). Cada fase se descompone en un conjunto de tareas genéricas (o generales) de segundo nivel. A partir del tercer nivel de abstracción, se realiza un "mapeo" de las tareas genéricas definidas en el modelo a situaciones específicas. En el cuarto nivel, se encuentran las instancias de proceso, donde se describen las acciones, decisiones y resultados de un proyecto particular de Minería de Datos [6, 8].

En cuanto al campo de uso de la Minería de Datos, es imposible efectuar una enumeración pormenorizada de todas las aplicaciones donde juega un papel crítico, ya que donde hay datos, hay aplicaciones [2]. Más allá de esto, específicamente este trabajo apunta a llevar a cabo un proyecto de Minería de Datos para extraer conocimiento con el fin de descubrir o explicar patrones de comportamiento de

artrópodos de suelo a partir de ciertos factores abióticos.

Cabe señalar, que en los ecosistemas terrestres la fauna edáfica juega un papel clave en la provisión de funciones (ciclado de nutrientes, descomposición de la materia orgánica, mantención de hábitats para organismos benéficos, etc.) y servicios ecosistémicos muy valorados en la actualidad como componentes de la biodiversidad funcional [9, 10].

Una de las formas de realizar los estudios de la biodiversidad de los artrópodos epigeos del suelo es mediante la utilización de trampas de caída, llamadas habitualmente "pitfall" [11, 12]. El avance tecnológico ha permitido la evolución de las trampas pitfall, surgiendo así las trampas de caída por tiempo (*time-sorting pitfall trap*) y las trampas de caída por tiempo con sensores [11–14].

El presente trabajo da continuidad a la cooperación científico-tecnológica entre dos grupos de investigación de la Universidad Nacional de Entre Ríos y la Estación Experimental Agropecuaria Concordia del Instituto Nacional de Tecnología Agropecuaria. Los datos recolectados en ocho ensayos de campo realizados durante el año 2021, utilizando una trampa automatizada de caída por tiempo con sensores y datalogger, diseñada y construida para estudios de biodiversidad de artrópodos de suelo, serán integrados a una base de datos con el propósito de llevar a cabo un proyecto de Minería de Datos guidado metodológicamente por CRISP-DM.

De esta manera, el aporte de esta propuesta consiste en aplicar Minería de Datos en una base de datos específicamente diseñada para descubrir, determinar o relacionar la incidencia de los factores abióticos (temperatura y humedad relativa ambiente, nivel de luz, y temperatura y humedad del suelo) en el comportamiento de la artropofauna edáfica existente en sistemas productivos hortícolas de la región de Salto Grande.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

El presente trabajo se enmarca en el análisis y aplicación de Minería de Datos con fines científicos para estudios de biodiversidad de artrópodos de suelo.

Contempla el diseño, implementación y carga de datos en una base de datos específica para este propósito. A su vez, comprende aplicar la metodología CRISP-DM para guiar el proceso de extracción de conocimiento sobre patrones de comportamiento de artrópodos de suelo, a partir de datos genuinos obtenidos en ensayos de campo realizados con una trampa "pitfall" convencional y otra automatizada por tiempo con sensores de temperatura y humedad relativa ambiente, nivel de luz y temperatura y humedad del suelo.

Asimismo, esta propuesta se encuadra en líneas de investigaciones prioritarias de dos facultades de la Universidad Nacional de Entre Ríos. Por un lado, en la línea de investigación *"Bases de Datos"* de la Facultad de Ciencias de la Administración, según las Resoluciones "C.D." Nº 160/11 y 203/11. Por otro lado, también se inscribe en la línea de investigación *"Bioinformática"* correspondiente a la Facultad de Ciencias de la Alimentación, de acuerdo a la Resolución "C.D." N° 313/19.

3. RESULTADOS ALCANZADOS Y ESPERADOS

El objetivo general del Proyecto PID Novel N° 8112 "Diseño y desarrollo de una trampa de caída por tiempo con sensores y datalogger para estudios de biodiversidad de artrópodos de suelo" fue diseñar y desarrollar una trampa de caída por tiempo automatizada para capturar artrópodos de suelo y medir factores abióticos con el fin de evaluar tecnologías de bajo impacto ambiental aplicadas a la horticultura.

Dicho objetivo general y los objetivos específicos vinculados al diseño, construcción, comprobación de su funcionamiento a través de ensayos de campo y evaluación de la capacidad de captura del dispositivo desarrollado, ya fueron alcanzados.

A partir de lo anterior, y del trabajo de acondicionamiento y clasificación de los individuos que componen las muestras, realizado por profesionales del Grupo Hortícola de la EEA Concordia del INTA, se disponen de datos propios que serán la materia prima principal sobre la cual se podrá llevar a cabo este proyecto de Minería de Datos.

El proyecto precitado será guiado por la metodología CRISP-DM y tiene como objetivo general descubrir conocimiento válido, útil y novedoso en una base de datos diseñada para estudios de biodiversidad de artrópodos de suelo.

En función a lo anterior, entre los objetivos específicos se encuentra diseñar e implementar en un motor de libre distribución una base de datos para unificar los datos obtenidos de orígenes y formatos diferentes (archivos del datalogger, planillas de cálculo, etc.). A su vez, el proyecto apunta a la construcción de modelos descriptivos a partir de los datos, con el propósito de extraer conocimiento sobre el comportamiento de la artropofauna edáfica estudiada.

Para dar cumplimiento a ello, en primer lugar se efectuará una búsqueda, revisión, clasificación y análisis de libros, trabajos y documentación científica y tecnológica relacionados con la temática. El foco de atención principal estará puesto en el proceso de extracción de conocimiento en bases de datos y en las tareas, los modelos y las técnicas de Minería de Datos.

Asimismo, se tiene previsto llevar a cabo y documentar las salidas de las tareas que prevé CRISP-DM en cada una de las seis fases, a saber [6, 7, 15]:

- Comprensión del problema o negocio: Determinar los objetivos del negocio, evaluar la situación, determinar los objetivos de la Minería de Datos y crear un plan para el proyecto.
- Entendimiento de los datos: Recolectar los datos iniciales, describir los datos, explorar los datos y verificar la calidad de los datos.
- Preparación de los datos: Seleccionar, limpiar, construir, integrar y dar forma a los datos.
- Modelado: Seleccionar la técnica de modelado, diseñar las pruebas del modelo, construir y evaluar el modelo.
- Evaluación: Evaluar los resultados, revisar el proceso y determinar las próximas etapas.
- Implementación: Planificar la implementación, planificar el monitoreo y el mantenimiento, crear un reporte final y revisar el proyecto.

Por consiguiente, la contribución del trabajo radica en que es una continuidad de una labor ya iniciada, la cual contempló desde la construcción del dispositivo, la realización de los ensayos de campo para obtener datos genuinos y, finalmente, la realización de un proyecto de Minería de Datos, sobre una base de datos surgida del trabajo en colaboración de investigadores de dos instituciones públicas de la República Argentina.

4. FORMACIÓN DE RECURSOS HUMANOS

Este trabajo forma parte de las actividades planificadas para una Tesis de la Maestría en Sistemas de Información de la Facultad de Ciencias de la Administración de la Universidad Nacional de Entre Ríos.

El equipo de trabajo está compuesto por el tesista, Director y Co-Directora de la Tesis precitada y por el Director e investigadores del Proyecto PID Novel N° 8112 de la Facultad de Ciencias de la Alimentación de la Universidad Nacional de Entre Ríos. Asimismo, forma parte la investigadora responsable del Grupo Hortícola de la EEA Concordia del INTA.

A su vez, este trabajo es un ámbito idóneo para que estudiantes de carreras de grado afines de la Universidad Nacional de Entre Ríos, puedan llevar a cabo sus proyectos de Trabajo Final o participen como Becarios de Iniciación en la Investigación.

5. BIBLIOGRAFÍA.

1. Witten H. Ian, Frank Eibe, H.A.M.: Data Mining: Practical Machine Learning Tolls and Techniques. Elseiver Inc. (2011)
2. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. (2011)
3. Vasilakos, A. V., Wan, J., Chen, F., Rong, X., Zhang, D., Deng, P.: Data Mining for the Internet of Things: Literature Review and Challenges. Int. J. Distrib. Sens. Networks. 11, 431047 (2015). https://doi.org/10.1155/2015/431047
4. Orallo, J.H., Quintana, M.J.R., Ramírez, C.F.: Introducción a la Minería de Datos. Pearson Educación (2004)
5. Maimon, O., Rokach, L.: The Data Mining and Knowledge Discovery Handbook.
6. Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M.J., Flach, P.: CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Trans. Knowl. Data Eng. 33, 3048-3061 (2021). https://doi.org/10.1109/TKDE.2019.2962680
7. Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., Wirth, R.: Crisp-Dm 1.0. Cris. Consort. 76 (2000)
8. Moine J.: Metodologías para el descubrimiento de conocimiento en bases de datos : un estudio comparativo. XVII Congr. Argentino Ciencias La Comput. XVII CACIC, 15, 18-22 (2013)
9. ELN-FAB: Functional agrobiodiversity: Nature serving Europe's farmers. ECNC-European Centre for Nature Conservation, Tilburg, the Netherlands (2012)
10. Ruiz-Lupión, D., Pascual, J., Melguizo-Ruiz, N., Verdeny-Vilalta, O., Moya-Laraño, J., Activity, S.A.: New litter trap devices outperform pitfall traps for studying arthropod activity. Insects. 10, 1-15 (2019). https://doi.org/10.3390/insects10050147
11. Droz, M.R., Diaz, B.M., Ramos, J.A., Benitez, P.D., Zapata, L.M.: Trampa pitfall con embudo y data logger para estudios de biodiversidad de artrópodos de suelo. An. CAI 2020. Congr. ARGENTINO AGROINFORMATICA (JAIIO). ISSN 2525-0949. 354-367 (2020)
12. Droz, M.R., Ramos, J.A., Benitez, P.D., Zapata, L.M., Diaz, B.M.: Diseño y desarrollo de un sistema embebido para una trampa pitfall con data logger. XXVI Congr. Argentino Ciencias la Comput. (CACIC 2020). 571-580 (2021)
13. Buchholz, S.: Design of a time-sorting pitfall trap for surface-active arthropods. Entomol. Exp. Appl. 133, 100-103 (2009). https://doi.org/10.1111/j.1570-7458.2009.00902.x
14. McMunn, M.S.: A time-sorting pitfall trap and temperature datalogger for the sampling of surface-active arthropods. HardwareX. 1, 38-45 (2017). https://doi.org/10.1016/j.ohx.2017.02.001
15. Espinosa Zúñiga, J.J.: Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. Ing. Investig. y Tecnol. 21, 1-13 (2020). https://doi.org/10.22201/fi.25940732e.2020.21n1.008