CIENCIA DE DATOS APLICADA AL ESTUDIO LA FAUNA ÍCTICA EN LA ZONA DEL RÍO PARANÁ.

Cinthia A. Cuba L.[(1, 2, *)], Paola V. Britos[(3,*)] y Gladis G. Garrido[(1,2,*)]
[(1)] Proyecto Biología Pesquera – Instituto de Biología Subtropical (IBS). Rivadavia 2307, N3300LQH, Posadas - Misiones, Argentina.
[(2)] Facultad de Ciencias Exactas, Químicas y Naturales (FCEQyN). Universidad Nacional de Misiones (UNaM). Félix de Azara 1552, N3300LQH, Posadas - Misiones, Argentina.
[(3)] Universidad Nacional de Rio Negro (UNRN). Laboratorio de Informática Aplicada. Río Negro, Argentina.
[*]E-mail: cinthiaacuba@gmail.com, pbritos@unrn.edu.ar, gladysgarrido@gmail.com

RESUMEN

El constante crecimiento del volumen de información es transversal a todos los ámbitos de nuestra vida cotidiana y hacen necesaria la utilización de nuevos métodos y enfoques analíticos que nos permitan navegar entre el volumen de información que se genera día a día y nos brinde la posibilidad de extraer conocimiento útil u oculto en esa gran cantidad de datos. La ciencia de datos, minería de datos o descubrimiento de conocimiento, son términos que tienen una amplia trayectoria y una gran cantidad de estudios realizados. En contraparte, en el ámbito biológico, existe una gran cantidad de estudios sobre la comunidad y dinámica de la fauna íctica; pero no se ha localizado ningún estudio específico que unifique las dos áreas. El presente trabajo propone la aplicación de técnicas de Ciencia de Datos a un set de información que describe el monitoreo de la fauna íctica en un tramo del Río Paraná.

El objetivo es poder determinar y entender las relaciones entre variables biológicas y ambientales e intentar descubrir comportamientos no apreciables a simple vista.

El resultado de este trabajo será un aporte al Proyecto de Biología Pesquera Regional y a través de ellos, a las entidades que toman decisiones y gestionan los recursos para la conservación y preservación de la fauna íctica en el río Paraná.

Palabras clave: ciencia de datos, minería de datos, descubrimiento de conocimiento en bases de datos, ictiofauna de agua dulce.

CONTEXTO

Este trabajo se lleva a cabo dentro del Proyecto de Biología Pesquera en el Laboratorio del Instituto de Biología Subtropical (IBS) de la Facultad de Ciencias Exactas, Químicas y Naturales (FCEQYN), de la Universidad Nacional de Misiones (UNaM), en el marco de un plan de trabajo final de la Maestría en Tecnologías de Información de la FCEQYN - UNaM

1. INTRODUCCION

El propósito de este trabajo es el desarrollo y aplicación de técnicas de Ciencia de Datos que permitan el descubrimiento de patrones y/o regularidades sobre la dinámica de la fauna íctica en un tramo del río Paraná. De esta manera se espera entender las relaciones entre variables biológicas y ambientales y que éstas sirvan de soporte a la toma de decisiones sobre la conservación y estudio de comunidades ícticas.

Los sistemas acuáticos en particular, se caracterizan por presentar gran variabilidad espacial y temporal en sus componentes físicos y químicos [1,2], los cuales

determinan la dinámica de las diferentes comunidades. Para analizar la variación espacio-temporal de la comunidad de peces, evaluar el rendimiento pesquero en término de biomasa, abundancia, riqueza y su relación con variables ambientales (como temperatura del agua y del ambiente, turbidez del agua, oxigenación, entre otros) se tomará como caso de estudio, datos de monitoreos sobre la fauna íctica que el Proyecto de Biología Pesquera Regional de la UNaM ha realizado en un tramo del río Paraná que abarca casi 200 km y comprende cinco sitios, siendo éstos: Toma de Agua Eriday (TAE), Puerto Garapé (GPE), Puerto Nemesio Parma - Candelaria (NPM), Puerto Maní (PMI) y Arroyo Yabebiry (YBY). Este monitoreo se realiza desde el año 1993 a la actualidad.

Para el presente trabajo se acota la información al rango de fechas que fue establecida después de fijar la cota definitiva en el río, y esto va desde 2010 a 2020.

Si bien, desde el punto de vista ictiológico existen diversos estudios relacionados a la biología reproductiva y rendimientos pesqueros [3], no se han expuesto patrones o características que expliquen el porqué de esta dinámica. Para dar respuesta a lo expuesto, este trabajo propone la aplicación de técnicas de Ciencias de Datos que permitan descubrir patrones y/o regularidades sobre la dinámica de la fauna íctica en el tramo del río Paraná antes mencionado.

2. LINEAS DE INVESTIGACION Y DESARROLLO

Toda la información recolectada del 1993 a la actualidad se encuentra almacenada en una base de datos relacional y puesto que una de las actividades del proyecto es brindar información a través de distintos informes, el análisis de ese gran volumen de datos dio pie a la creación de un almacén de datos, que junto a un sistema OLAP, forman parte de las nuevas herramientas de trabajo del proyecto. Si bien este set tecnológico aporta al rápido análisis de información desde distintas perspectivas, aún queda pendiente explicar e interpretar la dinámica de la comunidad íctica. Se espera responder a esta problemática aplicando técnicas de Ciencia de Datos que permitan descubrir patrones sobre la dinámica de la fauna íctica.

Finalmente, plasmar los hallazgos de manera simple, resumida y comprensible para el usuario.

3. RESULTADOS OBTENIDOS / ESPERADOS

Para comenzar con el proceso de extracción de conocimiento se han estudiado y comparado diferentes metodologías, como los expuestos en [4], [5] y [6] sobre Explotación de Información y Minería de Datos; luego las metodologías emergentes para Ciencia de Datos [7] o la adaptación de alguna existente [8] y [9].

Se decidió realizar un análisis comparativo entre las metodologías ASUM y Educción de Requisitos para Proyectos de Explotación de Información. Ambas se basan en la metodología CRISP-DM y la finalidad del análisis entre ambas es tomar sus mejores cualidades. ASUM, planteada como un Método Unificado de Soluciones Analíticas se presenta como un Manual de Usuario en línea que permite navegar entre cada fase o proceso planteado e indica las tareas a realizar en cada una de ellas (orientada a su herramienta comercial SPSS). Educción de Requisitos realizada como una tesis doctoral, tiene como puntos fuertes una serie de plantillas entregables que son fáciles y comprensibles de presentar al cliente; además de mostrar ejemplos concretos sobre la aplicación de la misma.

 RedUNCI - UNdeC. ISBN: 978-987-24611-3-3

Del análisis de metodologías se obtienen las siguientes cinco fases: 1- Definición del Proyecto. 2- Educción de Procesos de Negocio. 3- Educción de Datos de Procesos de Negocio. 4- Conceptualización del Negocio. 5- Especificación de Procesos de Explotación de Información.

De la aplicación de la fase 1 se obtuvieron cinco objetivos generales. A continuación se describen las tareas realizadas para resolver la premisa "Determinar e identificar características sobre Estructura de Peces y su relación con variables ambientales".

Para realizar las tareas prácticas se utiliza la herramienta RapidMiner Studio bajo una licencia "*Educational Edition*" en su versión 9.8.

En primer lugar, se tomó una muestra correspondiente a 1 (un) año de monitoreo con el fin de realizar las tareas de pre-procesamiento, los cuales incluyen limpieza de los datos, tratamiento de valores nulos, faltantes y calidad en general de los datos. Con la muestra tomada se trabajaron 25 variables tanto cualitativas como cuantitativas, de las cuales 8 representan datos ambientales y 17 datos bilógicos de los peces. En total la muestra se representa por 5380 registros.

Para comprender qué ocurre con los datos y cómo se relacionan entre sí, es necesario analizar una porción representativa del problema; para ello se ejecutó el proceso "Descubrimiento de Grupos" [10] con el algoritmo *K-means*.

Una vez que la herramienta formó los grupos, se utilizó una matriz de correlación para identificar qué tan fuertemente correlacionadas se encontraban las variables y con qué atributos. Además se analizaron los pesos de cada variable para tener una noción sobre cuáles son las más representativas. Como resultado se pudo observar que las variables con mayor peso al momento de definir grupos fueron: peso y largo del pez (como principales atributos biológicos ícticos); temperatura, oxígeno y transparencia del agua (como principales atributos ambientales). Por otra parte, los pares de variables mayormente correlacionadas fueron altura-peso, altura-largo, largo-peso y temp_ambiente-temp_agua.

Teniendo este panorama general, se procedió con el análisis de cada partición (verificando si la correlación de los datos en general es correcta.). Por cada grupo conformado se identifica el atributo cluster - con el que fue catalogado en el paso anterior - y en base a este atributo se aplica el algoritmo de inducción para obtener "Reglas de pertenencia a cada grupo". Esta tarea se ejecutó con el algortimo *Decision Tree*.

Como resultado se obtuvo que las variables que definieron el agrupamiento fueron tres: peso, largo y temperatura del agua, en el orden indicado.

Actualmente la investigación para este trabajo se encuentra en la tercera fase de la metodología, donde se realizan tareas asociadas a la Educción de Datos de Procesos de Negocio. Estas tareas se subdividen en dos grandes actividades, por un lado el relevamiento de los datos del negocio y por el otro, el análisis de los repositorios de datos. En esta etapa se está analizando la incorporación de información a través de la transformación de datos existentes, por ejemplo: a partir de la fecha de pesca, transformar este dato a estaciones del año y con este nuevo valor, generar otras salidas que sirvan al cumplimiento de los objetivos del trabajo.

4. FORMACION DE RECURSOS HUMANOS

El presente se lleva a cabo como Trabajo Final de la Maestría en Tecnologías de Información de la FCEQyN –UNaM y se conforma por tres integrantes: el maestrando, Lic. Cinthia Cuba de la FCEQyN – UnaM; la Dra. Paola Britos, perteneciente al Laboratorio de Informática Aplicada de la UNRN; y la Mgter. Gladys G. Garrido, Docente de categoría III (Sistema Nacional de Incentivo a la Investigación) perteneciente al Departamento de Biología de la FCEQYN – UNAM.

5. BIBLIOGRAFIA

[1] Fontoura, Nelson Ferreira, Ceni, Gianfranco, Braun, Aloisio Sirangelo, & Marques, Camilla da Silva. (2018). "*Defining the reproductive period of freshwater fish species using the Gonadosomatic Index: a proposed protocol applied to ten species of the Patos Lagoon basin*". Neotropical Ichthyology, 16 (2), e170006. Epub July 16, 2018.

[2] Stebniki, Samanta; González, Iván; D'Anatro, Alejandro y Teixeira de Mello, Franco. (2016) "*Relaciones entre variables ambientales y la comunidad de peces en el río Uruguay bajo"* Uruguay. Aqua-LAC - Vol. 8 - Nº 1 - Mar. 2016. pp. 62- 67.

[3] Torruco, Daniel, González-Solis, Alicia y Torruco-González, Ángel D. (2017). "*Diversidad y distribución de peces y su relación con variables ambientales, en el sur del Golfo de México*." Revista de Biología Tropical, vol. 66, no. 1, 2018, p. 438.

[4] Pollo Cattaneo, M. Florencia (2017). "*Modelo de proceso para elicitación de requerimientos en Proyectos de Explotación de Información*" (Tesis doctoral). Universidad Nacional de La Plata. Argentina.

[5] M. T. Rodríguez Montequín, J. V. Álvarez Cabal, J. M. Mesa Fernández and A. González Valdés. (2005) *"Metodologías para la realización de proyectos de Data Mining"*, in VII Congreso Internacional de Ingeniería de Proyectos, pp. 257-265.

[6] J. M. Moine, S. Gordillo and A. S. Haedo, (2011) *"Análisis comparativo de metodologías para la gestión de proyectos de minería de datos"*, in XVII Congreso Argentino de Ciencias de la Computación (CACIC), pp. 931-938.

[7] Angée Santiago - Lozano, Silvia, Montoya-Munera, Edwin - Ospina Arango, Juan - Tabares, Marta. (2018). *"Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects"*: 13th International Conference, KMO 2018, Žilina, Slovakia, August 6–10, 2018, Proceedings. 10.1007/978-3-319-95204-8_51

[8] IBM – (2015) "*Analytics Solutions Unified Method (ASUM)*", Publicado en: http://gforge.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/deliveryprocesses/ASUM-DM_8A5C87D5.html <Último acceso: Febrero 2021>

[9] Fois, Giuliana, Agüero, Gustavo, Britos, Paola. (2020). "*Comparación de metodologías agiles para Ciencia de Datos*" – Publicación InGenio 2020

[10] Britos, Paola. (2008). *"Procesos de explotación de información basados en sistemas inteligentes"*. (Tesis doctoral). Universidad Nacional de La Plata. Argentina.