TÉCNICAS DE MINERÍA DE DATOS PARA LA REDUCCIÓN DE COSTOS EN LA AUTOMATIZACIÓN DEL PROCESO DE CULTIVO DE TILAPIA.

DATA MINING TECHNIQUES FOR THE REDUCTION OF COSTS IN THE AUTOMATION OF THE TILAPIA CULTIVATION PROCESSDATA MINING TECHNIQUES FOR THE REDUCTION OF COSTS IN THE AUTOMATION OF THE TILAPIA CULTIVATION PROCESS.

Manuel Alejandro Coronado Arjona[1]
Miguel Ángel Perera Collí[2]
Víctor Manuel Bianchi Rosado[3]
Mariano de Jesús Matú Sansores[4]
Miguel Ángel Cohuo Ávila[5]

RESUMEN

La inversión económica en tecnología dificulta la automatización de los procesos de las personas dedicadas a actividades agrícolas, ganaderas o pesqueras, debido principalmente al precio elevado de algunos sensores utilizados. En las instalaciones del Centro de Bachillerato Tecnológico Agropecuario No. 14 se realizó la automatización del proceso de cultivo de tilapia con el objetivo de evaluar la ganancia en el peso y tamaño de esta especie usando un sistema mecanizado frente al sistema tradicional. Las variables físicas y químicas que influyen en el desarrollo biológico de estos peces son diversas; sin embargo, el oxígeno disuelto (OD) es el factor que mayor relevancia tiene en el proceso de cultivo. Desafortunadamente, el costo del sensor requerido para este proceso es caro. Por esta razón se utilizaron técnicas de minería de datos con el fin de determinar un modelo que permita predecir los niveles de oxígeno disuelto a partir de los valores obtenidos por los sensores de turbidez, temperatura y potencial de hidrógeno (pH). Las técnicas utilizadas para este estudio fueron los algoritmos de perceptrón multicapa, M5P y regresión lineal. El modelo matemático que se obtenga permitirá pronosticar el grado del OD sin la necesidad de contar con el sensor correspondiente. El impacto radica en que el modelo podría ser extrapolado a otros proyectos de automatización similares reduciendo con ello sus costos de inversión.

Palabras clave: Arduino, Automatización, Calidad del agua, Minería de datos

[1] Doctor en Sistemas Computacionales por la Universidad del Sur Campus Mérida. Maestría en Tecnologías de la Información egresado de la Universidad Latino. manuel.coronado@ittizimin.edu.mx
[2] Maestro en Pedagogía por la Universidad del Sur Campus Mérida. Ingeniero en Electrónica egresado del Instituto Tecnológico de Mérida. miguel.perera@ittizimin.edu.mx
[3] Maestro en Tecnologías de la Información egresado de la Universidad Interamericana para el Desarrollo Campus Tizimín. Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Mérida. victor.bianchi@ittizimin.edu.mx
[4] Maestro en Matemáticas por la Escuela Normal Superior de Yucatán, Licenciado en Ciencias de la Computación por la Universidad Autónoma de Yucatán campus Tizimín. mariano.matu@ittizimin.edu.mx
[5] Doctor en Sistemas Computacionales por la Universidad del Sur Campus Mérida. Maestría en Tecnologías de la Información egresado de la Universidad Latino. macohuo@itescam.edu.mx

Fecha de recepción: 15 de enero, 2019.

Fecha de aceptación: 01 de abril, 2019.

ABSTRACT.

The economic investment in technology hinders the automation of the processes of people engaged in agricultural, livestock or fishing activities, mainly due to the high price of some sensors used. In the facilities of the Centro de Bachillerato Tecnológico Agropecuario No. 14, the automation of the tilapia cultivation process was carried out in order to evaluate the gain in the weight and size of this species using a mechanized system versus the traditional system. The physical and chemical variables that influence the biological development of these fish are diverse; however, dissolved oxygen (DO) is the most important factor in the cultivation process. Unfortunately, the cost of the sensor required for this process is expensive. For this reason, data mining techniques were used in order to determine a model that allows predicting the levels of dissolved oxygen from the values obtained by turbidity, temperature and hydrogen potential (pH) sensors. The techniques used for this study were multilayer perceptron algorithms, M5P and linear regression. The mathematical model obtained will allow predicting the degree of the DO without the need to have the corresponding sensor. The impact is that the model could be extrapolated to other similar automation projects, thereby reducing its investment costs.

keywords: Arduino, Automation, Water quality, Data mining.

INTRODUCCIÓN

Se llevó a cabo la automatización del proceso de cultivo de tilapia con tecnología Arduino, a fin de evaluar la ganancia de peso y tamaño de estos peces en comparación con el método tradicionalista. En la implementación se utilizaron los sensores de: oxígeno disuelto (OD), pH, temperatura y turbidez.

El costo de la automatización con sensores Arduino puede ser elevado dependiendo del nivel de especialización y/o de su grado en los diferentes proyectos. Cualquier productor desea obtener el mayor beneficio con la menor inversión posible.

Ya que el nivel de OD en los estanques es un parámetro que influye mucho en el desarrollo de estos peces y debido al costo elevado del respectivo sensor, se propone describir un modelo predictivo que ayude a obtener el nivel de este parámetro a partir de los tres restantes. De encontrarse el modelo, el costo de automatización se vería reducido ya que sólo se utilizarían los sensores de pH, temperatura y turbidez en sustitución del de oxígeno disuelto.

Se generaron 378 registros correspondientes a igual número de lecturas llevadas a cabo en un estanque de geomembrana con 200 alevines durante un periodo de 2 meses y medio. Los registros fueron sometidos a tres técnicas de predicción utilizando el software Weka a fin de obtener un modelo que permita predecir con un bajo margen de error los niveles de oxígeno disuelto.

JUSTIFICACIÓN

Las granjas acuáticas ya tienen cierta antigüedad en nuestro país, sin embargo, la dificultad para dar seguimiento y controlar la calidad del agua, las enfermedades y parásitos de los peces hacen que quienes intentan proyectos de este tipo abandonen al poco tiempo por no contar con las herramientas tecnológicas necesarias para lograr una producción eficiente que permita incluso la exportación.

La automatización sólo es viable si al evaluar los beneficios económicos y sociales de las mejoras que se podrían obtener al automatizar, éstas son mayores a los costos de operación y mantenimiento del sistema (Mendiburu, 2003). Por lo tanto, un modelo matemático que ayude a predecir el nivel de oxígeno disuelto (OD) a partir de sensores para otros propósitos y a un menor precio en el mercado podría representar un beneficio económico a los productores y, al mismo tiempo, permitiría la generación de empleos, producción de alimentos de calidad, entre otros.

MARCO TEÓRICO

Calidad del agua en el proceso de cultivo de tilapia

Bautista y Ruiz (2011) mencionan que la acuicultura o acuacultura es el empleo de diferentes sistemas y técnicas en el proceso de cultivo de organismos que cumplen su ciclo de vida total o parcial en el agua, pudiendo ser estos, animales o vegetales. Los cultivos son generalmente destinados al consumo humano, esparcimiento, conservación y repoblamiento de ambientes naturales, en este último caso, para especies nativas. El cultivo de peces es una alternativa que los productores han incorporado a sus sistemas productivos, con el objeto de diversificar su producción.

La Secretaría de Agricultura y Desarrollo Rural (SEDER, 2014) en Jalisco, menciona que parámetros físicos, químicos y biológicos son determinantes en la calidad del agua de los estanques utilizados en el proceso de acuacultura; por tal motivo, deben ser mantenidos dentro de los rangos aceptables para un buen desarrollo de la población en cultivo ya que de otro modo, podría tener bajo crecimiento, proliferación de patógenos con brotes de enfermedad, eventuales mortalidades y baja calidad del producto final.

A continuación, se describen algunas de las características de los parámetros considerados en este proyecto, junto con los sensores utilizados:

Turbidez del agua: La turbidez se debe a la presencia de partículas suspendidas en cantidades variables. Su efecto impacta en el crecimiento de peces y de otros organismos naturales que forman parte de su alimentación debido a que la luz penetra a una corta distancia y la fotosíntesis se reduce. En la figura 1a se aprecia el sensor Arduino utilizado para detectar la calidad del agua midiendo los niveles de turbidez empleando la luz para detectar partículas suspendidas en el líquido.

pH: Su valor caracteriza la acidez y alcalinidad de las aguas afectando de diversas formas a los vegetales y animales que viven en ellas. Según la FAO (s.f.), los valores de pH varían de 0 a 14, un pH 7 indica que el agua es neutra. Los valores inferiores a 7 indican acidez y los superiores, alcalinidad. De acuerdo a SEDER (2014), los animales crecen mejor en aguas alcalinas que en aguas ácidas. Las aguas ácidas irritan las branquias de los peces. El sensor correspondiente se encuentra en la figura 1b.

Oxígeno disuelto: Según SEDER (2014), en la calidad del agua es el parámetro más importante. Si éste falta se afecta el crecimiento y la conversión alimenticia de los organismos. Se trata del elemento más importante en el agua para los organismos acuáticos en la realización de los procesos oxidativos que coadyuvan a la obtención de energía a partir del alimento. La concentración de oxígeno disuelto en el agua es medida, usualmente en

partes por millón (ppm) o en miligramos por litro (mg/l). En la figura 1C se puede apreciar el dispositivo lector de esta variable.

Temperatura: La temperatura rige algunos parámetros físicos, químicos y biológicos, tales como la evaporación y la solubilidad de los gases. Dentro de los biológicos están los procesos metabólicos como la respiración, nutrición, actividad de las bacterias en la descomposición de la materia orgánica, entre otros; de ahí la necesidad de conocer y evaluar los cambios de temperatura del agua (SEDER, 2014). En la figura 1d se muestra el sensor correspondiente.

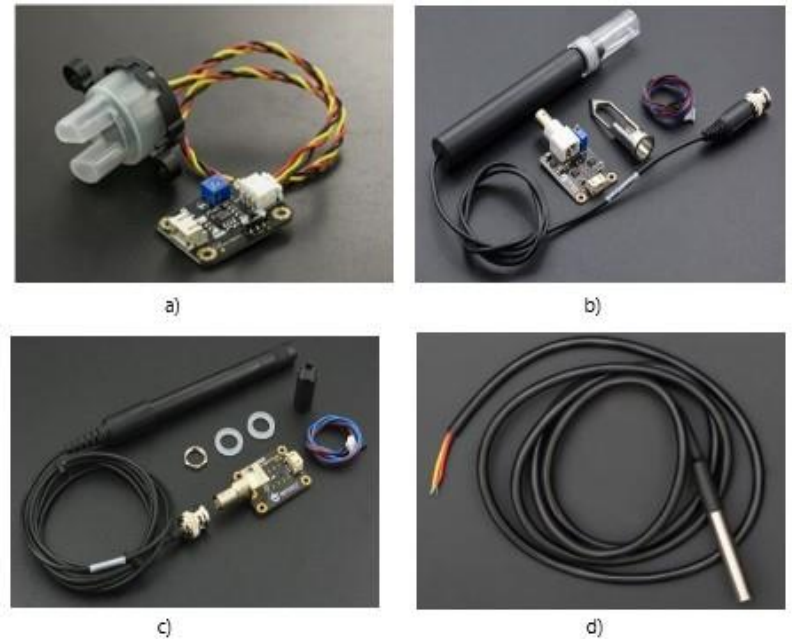

Figura 1. Sensores Arduino utilizados para determinar la calidad del agua en el proyecto (Cortesía dfrobot.com)

Minería de datos

Pérez y Satín (2007) definen a la minería de datos como el proceso de descubrir nuevas y significativas relaciones, patrones y tendencias al momento de examinar grandes bases de datos. Sus técnicas persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en esas extensas bases de datos.

Las tareas más comunes de la minería de datos, según Riquelme, Ruiz y Gilbert (2006) son:

a. Clasificación: clasifica un dato dentro de una de las clases categóricas predefinidas.
b. Regresión: el propósito de este modelo es hacer corresponder un dato con un valor real de una variable.
c. Clustering: se refiere a la agrupación de registros, observaciones o casos en clases de objetos similares. Un clúster es una colección de registros que son similares entre sí, y distintos a los registros de otro clúster.
d. Generación de reglas: aquí se extraen o generan reglas de los datos. Estas reglas hacen referencia al descubrimiento de relaciones de asociación y dependencias funcionales entre los diferentes atributos.
e. Análisis de secuencias: se modelan patrones secuenciales, como análisis de series temporales, secuencias de genes, etc. El objetivo es modelar los estados del proceso, o extraer e informar de la desviación y tendencias en el tiempo.

Métricas matemáticas para la evaluación de la calidad de los modelos

De acuerdo con Hornick, Marcadé y Venkayala (2007) en lo que respecta a los modelos de regresión, su calidad es obtenida calculando los errores acumulativos al comparar valores predichos con valores conocidos. Generalmente, cuanto menor sea el error acumulativo, mejor será el rendimiento del modelo.

Hernández, Ramírez y Ferri (2004), mencionan que, si la tarea es agrupamiento, las medidas de evaluación dependen de la técnica utilizada, aunque normalmente, suelen ser función de la cohesión de cada grupo y de la separación entre grupos. Es posible formalizar estas dos características utilizando la distancia media al centro del grupo de los miembros de un grupo y la distancia media entre grupos, respectivamente.

En el software de minería de datos, Weka, las mediciones de error utilizadas son:

a. Coeficiente de Kappa: El índice Kappa se usa para evaluar la concordancia de métodos cuyo resultado es categórico, con dos o más clases. Este índice representa la proporción de acuerdos observados respecto del máximo acuerdo posible más allá del azar (Borrás *et al*). En caso de concordancia perfecta el valor de kappa es 1; si la concordancia observada es igual a la esperada kappa vale 0; y en el caso de que el acuerdo observado sea inferior al esperado el índice kappa es menor que 0.
b. Coeficiente de correlación: De acuerdo a Vinuesa (2016), la correlación es una medida de la relación (covariación) lineal entre dos variables cuantitativas continuas (x, y). La manera más sencilla de saber si dos variables están correlacionadas es determinar si co-varían (varían conjuntamente). La principal diferencia entre regresión y correlación es que con esta última se determina el grado de asociación entre variables y si dicha relación es o no significativa. Por otro lado, la regresión trata de definir la función que mejor explica la relación entre las variables. El rango de valores se encuentra entre 1 y -1 (ver figura 2).

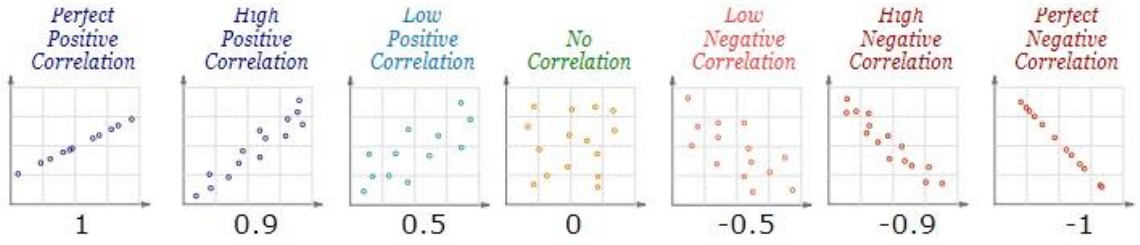

Figura 2. Rango de valores para la interpretación del coeficiente de correlación

c. Error Absoluto Medio: es la cantidad utilizada para medir qué tan cerca están los pronósticos o predicciones respecto a los resultados eventuales. La media del error absoluto es como la varianza, pero en lugar de cuadrar la diferencia, se utiliza su valor absoluto.
d. Error Cuadrático Medio: es una medida de las diferencias entre valores (valores de muestra y población) predichos por un modelo o un estimador y los valores realmente observados. Representa la muestra desviación estándar de las diferencias entre los valores predichos y los valores observados.
e. Error relativo: es una medida que indica cuánto se desvía un resultado del valor real.
f. Error absoluto: Es una medida en porcentaje comparado con el valor real.

METODOLOGÍA

Etapa 1. Automatización mediante Arduino: se adquirieron e implementaron los sensores para determinar la calidad del agua en un estanque de geomembrana. En la figura 3 se puede observar el correspondiente diagrama de circuitos.

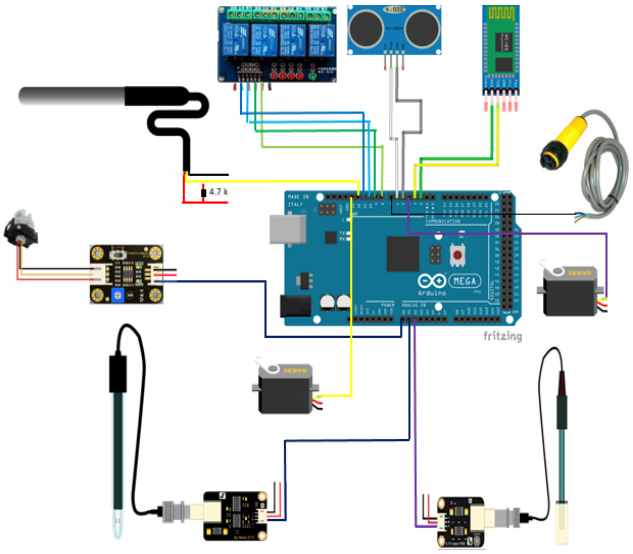

Figura 3. Diagrama de circuitos utilizado para el monitoreo de variables y automatización del proceso de cultivo de tilapia

Etapa 2. Instalación de un servidor de base de datos. Fue instalado el software servidor phpmyadmin en un dispositivo Raspberry pi. El servidor phpmyadmin permite la creación y administración de bases de datos vía web con la intención de que la información de las lecturas realizadas por los sensores pueda estar disponible en cualquier momento y lugar a través de una conexión a Internet.

Etapa 3. Diseño lógico y físico de la base de datos. Se procedió a realizar el diseño lógico (ver figura 4) de la base de datos y su implementación física en el servidor de bases de datos a través del lenguaje de consulta SQL. La base de datos consiste en 10 tablas.

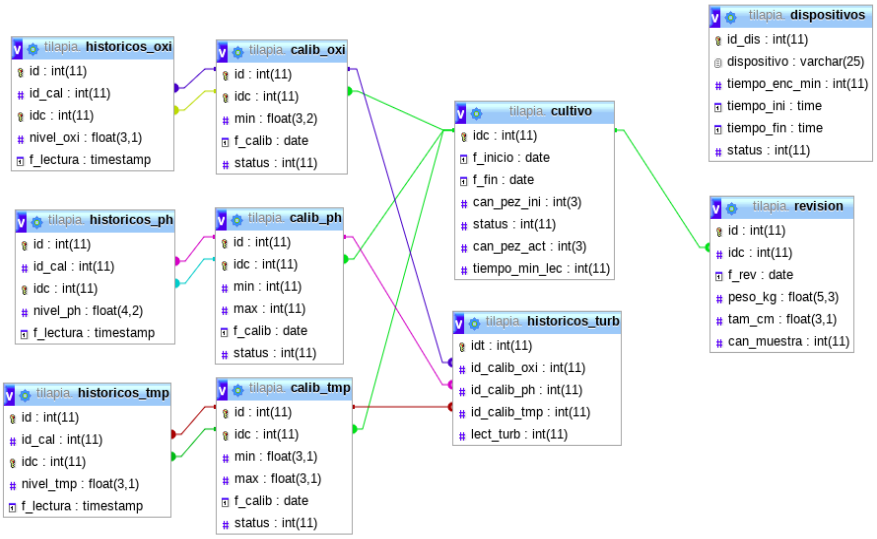

Figura 4. Diseño lógico de la base de datos del proyecto

Etapa 4. Selección del conjunto de datos. A través de una sentencia SQL se obtuvieron los valores correspondientes a las lecturas realizadas por los sensores de pH, temperatura, oxigenación y turbidez. A continuación, se describe la sentencia SQL utilizada:

```
select historicos_oxi.nivel_oxi, historicos_ph.nivel_ph, historicos_tmp.nivel_tmp, historicos_turb.lect_turb from historicos_oxi INNER JOIN historicos_ph on historicos_oxi.id=historicos_ph.id INNER join historicos_tmp on historicos_oxi.id=historicos_tmp.id INNER JOIN historicos_turb on historicos_oxi.id=historicos_turb.idt
```

La consulta SQL arrojó un total de 378 registros. Posteriormente, se seleccionaron 370 filas para generar con ellas los modelos y, las 8 filas restantes, servirían para comprobar la validez de los mismos. Estos dos grupos de datos fueron dispuestos en dos archivos diferentes en formato ARFF (ver figura 5) para ser cargados al software de minería de datos, Weka. Esta aplicación provee de una colección de técnicas para realizar análisis de datos y modelado predictivo, también consta de herramientas para la visualización de los datos.

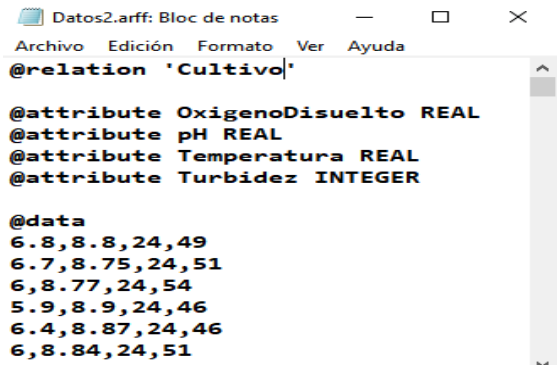

Figura 5. Archivo con los valores de las lecturas provenientes de los sensores de OD, pH, temperatura y turbidez

Etapa 5. Selección de las técnicas de minería de datos. En esta etapa se decidió que la tarea a realizar es la predicción ya que lo que se pretende es obtener el valor del oxígeno disuelto a partir de los sensores de pH, turbidez y temperatura. Por lo tanto, se optó por someter los datos del archivo ARFF a tres técnicas de predicción, éstas son:

a. Regresión: Técnica estadística que estudia la relación entre variables cuantitativas, puede ser simple o múltiple (Guisande *et al*, 2006).
b. Perceptrón multicapa (MPL): es una red neuronal con al menos tres capas, una de entrada, una de salida y una o más intermedias. De acuerdo a Flórez y Fernández (2008), constituye el paradigma más utilizado para la resolución de problemas tanto de clasificación como de regresión. Según Robertson (2017), este modelo de red opera con valores que son pasados a la capa de entrada, procesados por las capas ocultas (intermedias) y devueltos a través de la capa de salida.
c. Algoritmo M5P: es un método de aprendizaje mediante árboles de decisión en donde cada hoja tiene asociada una clase que permite calcular el valor estimado de la instancia mediante una regresión lineal (Vizcaino, 2008).

Etapa 6. Construcción y evaluación del modelo con el software Weka. Del total de 378 registros, se seleccionaron los primeros 370 para generar los modelos utilizando cada una de las técnicas anteriormente mencionadas. Las restantes 8 filas de datos permitieron determinar la eficiencia de los prototipos obtenidos. Desafortunadamente, los valores predichos distaban demasiado de los valores reales ubicados en el segundo archivo ARFF. En el software Weka, para corroborar la validez con datos distintos a los de entrenamiento de los algoritmos, se utiliza la opción “Supplied test set”.

RESULTADOS

En la tabla 1 se especifican los valores de las métricas seleccionadas utilizadas para comparar las distintas técnicas de minería de datos en la obtención de un modelo matemático que ayude a predecir los niveles de oxígeno disuelto.

Tabla 1. Comparación de resultados por técnica

Algoritmo	Coeficiente de correlación	Error en la media absoluta	Error absoluto relativo	Error cuadrático relativo
Regresión Lineal	0.4167	0.3852	92.6%	90.906%
Perceptrón Multicapa	0.5198	0.3568	85.7856%	86.8862%
M5P	0.0228	9.9483	2392.1627%	2146.898%

El algoritmo que presenta un mayor coeficiente de correlación positiva con los menores errores es el perceptrón multicapa. Por otro lado, el M5P dista bastante de las otras dos técnicas con 0.0228 de correlación junto con 2392.1627% y 2146.898% en sus errores. La baja correlación positiva hace que sea imposible determinar el OD a partir de las otras variables, es decir, no se aprecie que cavarían de manera conjunta.

CONCLUSIONES

Niveles bajos en el coeficiente de relación y porcentajes elevados en los errores absoluto y relativo hacen difícil que técnicas de minería de datos puedan ser utilizadas para predecir cantidad de oxígeno disuelto en los estanques destinados al cultivo de tilapia a partir de las lecturas efectuadas por los sensores de turbidez, temperatura y pH.

Los datos resumidos en la tabla 1 son con base en las lecturas efectuadas en los primeros 80 días del proceso de cultivo, iniciando en la tercera semana de octubre de 2018 y con un clima alternándose entre caluroso y frio en el oriente del estado de Yucatán. Dado que el proceso de cultivo dura aproximadamente 6 meses, se pretende seguir guardando datos obtenidos de los sensores para determinar, si es posible o no, la generación del modelo matemático en etapas futuras bajo ambientes calurosos y con los peces ya de mayor tamaño.

BIBLIOGRAFÍA

Bautista, J. y Ruiz, J. (2011). Calidad del agua para el cultivo de tilapia en estanques de geomembrana. Recuperado el 5 de Diciembre de 2018, de http://fuente.uan.edu.mx/publicaciones/03-08/2.pdf

Borrás, J., Delegido, J. Pezzola, A., Pereira, M., Morassi, G. y Camps, G. (2017). Clasificación de usos del suelo a partir de imágenes Sentinel-2. Recuperado el 6 de Diciembre de 2018, de https://riunet.upv.es/bitstream/handle/10251/83604/7133-28392-1-PB.pdf?sequence=1

FAO (s.f.). Mejora de la calidad del agua en los estanques. Recuperado el 5 de Diciembre de 2018, de http://www.fao.org/fishery/static/FAO_Training/FAO_Training/General/x6709s/x6709s02.htm

Flórez, R. y Fernández, J. (2008). Las redes neuronales artificiales. Fundamentos teóricos y aplicaciones prácticas. España: Netbiblo

Guisande, C., Barreiro, A., Maneiro, I., Riveiro, I., Vergara, A. y Vaamonde, A. (2006). Tratamiento de datos. España: Ediciones Díaz de Santos

Hernández, J., Ramírez, M. y Ferri, C. (2004). Introducción a la minería de datos. España: Pearson

Hornick, M., Marcadé, E. y Venkayala, S. (2007). Java Data Mining. Strategy, Standard, and practice. A practical guide for architecture, design and implementation. E.E.U.U.: Morgan Kaufmann publications

Mendiburu, H. (2003). Automatización medioambiental. Recuperado el 5 de Diciembre de 2018, de http://www.liceus.com/cgi-bin/ac/pu/AutomatizacionMedioambiental.pdf

Pérez, C. y Satín, D. (2007). Minería de datos. Técnicas y herramientas. España: Paraninfo

Riquelme, J., Ruiz, R. y Gilbert, K. (2006). Minería de datos. Conceptos y tendencias. Recuperado el 6 de Diciembre de 2018, de https://www.redalyc.org/html/925/92502902/

Robertson, D. (2017). An Evaluation of Fast Multi-Layer Perceptron Training Techniques for Games. Recuperado el 6 de Diciembre de 2018, de https://rke.abertay.ac.uk/ws/portalfiles/portal/14242051/Robertson_AnEvaluationOfFastMultiLayerPerceptron_Published_2017.pdf

SEDER (2014). Calidad del agua en la acuacultura. Recuperado el 5 de Diciembre de 2018, de https://seder.jalisco.gob.mx/fomento-acuicola-y-pesquero-e-inocuidad/519

Vinuesa, P. (2016). Correlación: teoría y práctica. Recuperado el 6 de Diciembre de 2018, de http://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema8_correlacion.pdf

Viscanino, P. (2008). Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de weka (waikato environment for knowledge analysis). Recuperado el 5 de Diciembre de 2018, de http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf