UADY
UNIVERSIDAD AUTÓNOMA DE YUCATÁN

Proyecto de investigación
Gestión e inferencia de información utilizando Machine Learning para una biorrefinería que produce hidrógeno a partir de aguas residuales.

Responsable: Dr. Jorge Ricardo Gómez Montalvo

Alumno: Ing. Luis Arturo Rodríguez Filigrana

Colaboradores:

- Dr. Francisco Moo Mena, Universidad Autónoma de Yucatán.
- Dr. Jesús Ixbalank Torres Zúñiga, Universidad de Guanajuato.
- MC. Luis Ramiro Basto Díaz, Universidad Autónoma de Yucatán.
- Dr. Víctor H. Domínguez Menéndez, Universidad Autónoma de Yucatán.
- MC. Juan Francisco Garcilazo Ortiz, Universidad Autónoma de Yucatán.

1. Introducción

La gran cantidad de datos, que se genera día a día, está creciendo a un ritmo sin precedentes como resultado de la evolución de las tecnologías web, las redes sociales y los dispositivos móviles, que utilizan el servicio de internet [1]. Un ejemplo es Twitter, que genera más de 50 millones de tweets por día, produciendo más de 6 TB de datos diarios [2]. A los grandes volúmenes de datos, que se generan diariamente, se les conoce como Big Data [3].

Big Data es un término para conjuntos de datos masivos que tienen una estructura grande, variada y compleja, con las dificultades de almacenamiento, análisis y visualización para futuros procesos o resultados [3]. Estos datos se generan a partir de transacciones en línea, correos electrónicos, videos, audios, imágenes, sensores, etc., [4]. Debido a la gran cantidad de información, las bases de datos y los repositorios, que almacenan esta información, también están creciendo exponencialmente. De hecho, se ha sugerido que los datos crecen a la misma velocidad que los recursos computacionales, que, según la ley de Moore, se duplican cada 18 meses [5]. Uno de los mayores desafíos en la ciencia de la computación, en general, es la generación de sistemas que sean capaces de procesar conjuntos de datos en crecimiento [5].

La ciencia computacional de extraer información útil de grandes volúmenes de datos o bases de datos se conoce como minería de datos. Es una disciplina, que se encuentra en la intersección de estadísticas, aprendizaje automático, administración de datos y bases de datos, reconocimiento de patrones, inteligencia artificial y otras áreas [6]. Así pues, El objetivo de la minería de datos es descubrir información nueva y útil en bases de datos, y repositorios, empleando diversos algoritmos de minería de datos, tales como, Support Vector Machines (SVM), Classification And Regression Trees (CART), A priori, entre otros, para encontrar patrones de información para un análisis más claro en grandes volúmenes de datos [7].

Al igual que la minería de datos, el Aprendizaje Automático (Machine Learning) juega un papel importante como un componente fundamental del análisis de datos, y es uno de los principales impulsores de la revolución del Big Data [8]. La razón de esto se debe a su capacidad de aprender de los datos y proporcionar información basada en ellos, así como decisiones y predicciones [9]. Así pues, uno de los principales objetivos del aprendizaje automático, es descubrir conocimiento a partir de la gran cantidad de información almacenada en las bases de datos y repositorios con el fin de tomar decisiones inteligentes.

Hoy día, diversas áreas del conocimiento (e.g., biología, inteligencia artificial, bioquímica, etc.) han generado y almacenado enormes volúmenes de datos, que describen sus operaciones, productos y procesos. Tal cantidad de datos complica la inferencia de información y conocimiento, ya que las diversas áreas del conocimiento no cuentan con los algoritmos necesarios de minería de datos (e.g., SVM, CART, A priori, etc.) y aprendizaje automático (e.g., Supervised Learning (SL), Unsupervised Learning (UL) y Reinforcement Learning (RL)) necesarios para procesar toda la información. El campo de la minería de datos y aprendizaje automático, abordan la cuestión de extraer patrones interesantes, asociaciones, reglas, cambios y anomalías de los datos para mejorar el proceso de toma de decisiones en las diversas áreas del conocimiento [10].

En este proyecto de investigación se propone diseñar y desarrollar una plataforma web para recibir y almacenar información, que utiliza métodos y algoritmos de aprendizaje automático y

minería de datos para procesar y analizar grandes cantidades de datos que son generados por una biorrefinería que produce hidrógeno a partir de aguas residuales.

2. Contexto y Problemática

La figura 1 muestra la biorrefinería propuesta para tratar residuos agroindustriales. Primeramente, se propone un digestor anaerobio para tratar los residuos agroindustriales y generar principalmente ácidos grasos volátiles (AGVs). Adicionalmente, la digestión anaerobia generará dióxido de carbono (CO_2) como un subproducto. El efluente por tratar serán vinazas provenientes de la industria tequilera, conocidas por tener una alta carga orgánica (40 – 50 DQO/l), y alta generación de AGV en la etapa acidogénica cuando son tratados por la digestión anaerobia (DA). Los ácidos grasos volátiles generados por la digestión anaerobia se utilizarán como sustrato en las celdas electroquímicas microbianas, las cuales producirán hidrógeno. Por otro lado, el CO_2 generado como subproducto se inyectará en el PBR de producción de microalgas para producir biomasa revalorizable. En esta etapa se estudiarán dos consorcios de microalgas nativas. El primero fue aislado desde la laguna de Yuriria, Estado de Guanajuato, y contiene mayoritariamente microalgas pertenecientes al género de *Chlorella sp*. El segundo fue aislado desde una planta de tratamiento de aguas residuales urbanas y aún no ha sido caracterizado molecularmente. Ambos consorcios serán cultivados a distintas proporciones CO_2/aire con el fin de seleccionar el consorcio que presente la mayor capacidad de crecimiento a elevadas concentraciones de CO_2. La biomasa microalgal obtenida será caracterizada bioquímicamente mediante técnicas de determinación de carbohidratos, lípidos, proteínas y actividad antioxidante total (DPPH, carotenoides totales y fenoles totales) para su posterior revalorización [11], además, será caracterizada mediante técnicas de biología molecular.

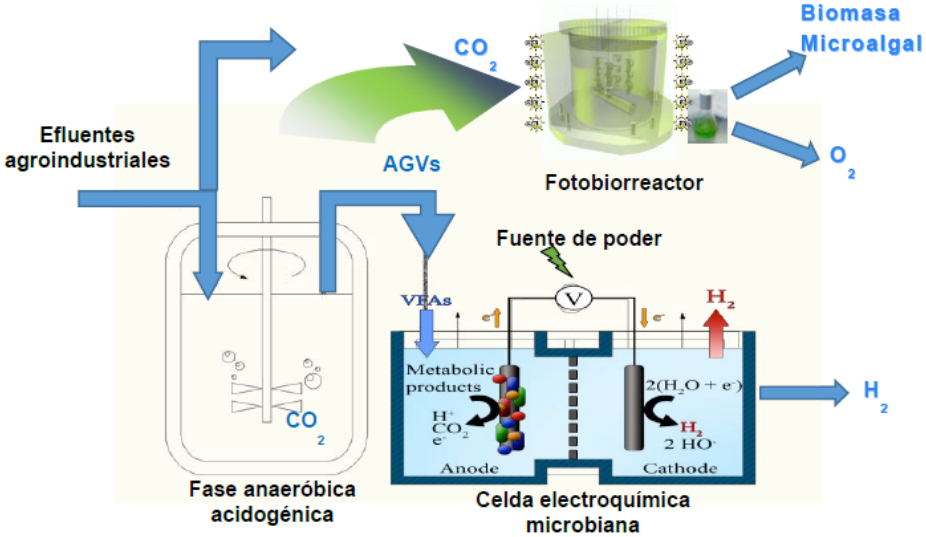

Figura 1.- Biorrefinería propuesta

En la misma figura 1 se puede observar la Fase anaeróbica acidogénica que se encarga de tratar los residuos agroindustriales y generar, principalmente, Ácidos Grasos Volátiles (AGVs). La fase anaeróbica acidogénica utiliza las siguientes entradas: Tasa de dilución (Dil), Demanda Química de Oxígeno (DQO) y Ácidos Grasos Volátiles (AGVs). De igual manera, la fase anaeróbica acidogénica tiene como variables del estado de interés las siguientes: biomasa, DQO, AGVs y CO_2.

Los AGVs generados por la fase anaeróbica acidogénica, se utilizan como dato de entrada en las celdas electroquímicas microbianas, las cuales tienen como variables de interés al hidrógeno, biomasa de bacterias anodofílicas, biomasa de bacterias metanogénicas, biomasa de bacterias hidrogenotrópicas, mediador de oxidación, corriente (en amperes), flujo de hidrógeno (QH_2) y AGVs.

Por otro lado, el CO_2 generado por la fase anaeróbica acidogénica se inyecta, como dato de entrada, en el fotobiorreactor de producción de microalgas. Otros datos de entrada que utiliza el fotobiorreactor son: flujo de entrada (Q_in), concentración de nutriente.Las variablesd ee estado de interés del fotobiorreactor son la biomasa, el nutriente, la cuota de nutriente y el carbono inorgánico. Este proyecto ocurre en el contexto de esta biorrefinería, la cual utiliza sensores para la medición y/o estimación de variables de estado de la fase anaeróbica acidogénica, el fotobiorreactor y la celda electroquímica microbiana.

En la figura 2, se puede observar que la biorrefinería cuenta con 30 sensores aproximadamente, que toman muestras cada 10 segundos. Las muestras pasan por un filtro que elimina el ruido de las mediciones y se envían esos datos cada 10 minutos (2a). Estos datos son capturados de forma manual por un biotecnólogo que trabaja con la biorrefinería (2b). La información capturada de forma manual (2c), puede ocasionar que algún dato contenga valores erróneos y, afecte el proceso de toma de decisiones (2d). La toma de decisiones se lleva a cabo cuando se tiene información recopilada, de la biorrefinería, de 3 o 5 meses dependiendo de los criterios establecidos por los biotecnólogos (2e). Una vez que se tiene una decisión, un biotecnólogo se encarga de aplicar las correcciones, si las hay, a la biorrefinería (2g).

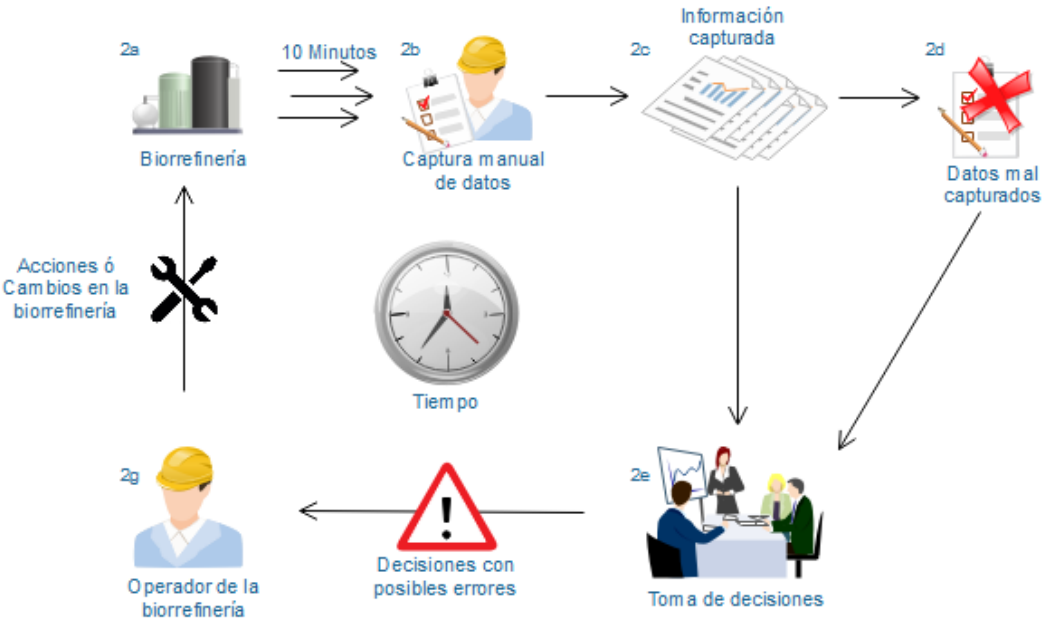

Figura 2.- Proceso de toma de decisiones y sus posibles consecuencias

En las siguientes subsecciones se explican las problemáticas de los procesos planteados anteriormente.

2.1. Recepción y almacenamiento de datos

Como ya se mencionó, la biorrefinería genera aproximadamente 550 datos cada 10 minutos, y los captura en un formato establecido por la misma. La inferencia de conocimiento a partir de estos datos resulta ser complicado, ya que no se cuenta con un formato o estándar de datos establecido por la biorrefinería y, sin los cálculos computacionales necesarios (e.g., Minería de datos, Machine Learning, etc), el usuario final no puede realizar la inferencia de conocimiento adecuadamente. En este proyecto se propone especificar un formato para los datos generados por la biorrefinería, los cuales serán almacenados, por un módulo de almacenamiento, en una base de datos.

Generalmente, los biotecnólogos capturan de forma manual los resultados generados por la biorrefinería. Los resultados pueden tener información mal capturada, esto es un problema al momento de tomar una decisión para realizar un cambio en la biorrefinería, ya que, puede afectar al rendimiento de la biorrefinería y causar errores. Debido a esto, se propone un módulo de recepción de datos que se encargue de recopilar toda la información generada por la biorrefinería. De este modo se pretende agilizar el proceso de captura de información y evitar errores de captura.

2.2. Toma de decisiones

En el proceso de toma de decisiones, los biotecnólogos realizan cálculos estadísticos básicos utilizando la información generada por la biorrefinería para verificar que la biorrefinería esté funcionando correctamente o no. Un problema es cuando hay información con errores de captura, esto causa que se tomen decisiones erróneas y se efectúen cambios en los procesos de la biorrefinería que pueden causar errores en su funcionamiento. Otro problema es el tiempo que se tarda en tomar una decisión ya que, al momento de aplicar una acción, la biorrefinería ya está en otro estado de tiempo. Es decir, ya generó nueva información y la toma de decisiones se realizó con resultados de un estado de tiempo anterior y, de igual manera, puede afectar al funcionamiento de la biorrefinería y causar errores.

2.3. Visualización de resultados

Cuando se realiza un cambio en la biorrefinería debido a una decisión tomada, los biotecnólogos esperan visualizar si esos cambios realizaron algún efecto en los procesos de la biorrefinería, pero, esos resultados solo se ven cada 10 minutos cuando la biorrefinería produce información. Para tener un mejor control del funcionamiento de la biorrefinería, es necesario conocer los estados de la biorrefinería en casi tiempo real.

3. Objetivo General

- Desarrollar una plataforma web que permita recibir, almacenar, inferir y visualizar información, para la toma de decisiones, cercano a tiempo real, utilizando algoritmos de minería de datos y de aprendizaje automático, para una biorrefinería que produce hidrógeno a partir de aguas residuales, a partir de los datos que aquella genera.

4. Objetivos Específicos

- Desarrollar un sistema de adquisición y procesamiento de datos recaudados directamente de la biorrefinería para su almacenamiento en un repositorio y posteriormente realizar un procesamiento de estos para inferencia de información.

- Desarrollar un sistema que permita obtener, almacenar y procesar los datos de la biorrefinería en un repositorio para generar información estadística que pueda ser visualizada a través de una interfaz amigable en ambiente Web y que permita conocer el estado del proceso.

- Desarrollar una Interfaz Humano-Máquina (IHM), en un ambiente Web, para la visualización de la información almacenada recaudada de la biorrefinería, con el propósito de que cualquier usuario pueda acceder a esta información para la inferencia de datos

- Establecer un estándar de almacenamiento de información extraída de la biorrefinería para su uso genérico para que cualquier otra biorrefinería pueda tener acceso a este sistema y poder brindarle las mismas herramientas para el procesamiento de la información.

- Realizar una selección de algoritmos de Machine Learning para realizar pruebas de desempeño de la plataforma y verificar que el rendimiento sea óptimo en el momento que se este realizando todas las actividades de la biorrefinería.

En la figura 3 se muestra una propuesta de la plataforma web que es resultado del cumplimiento de los objetivos específicos antes mencionados.

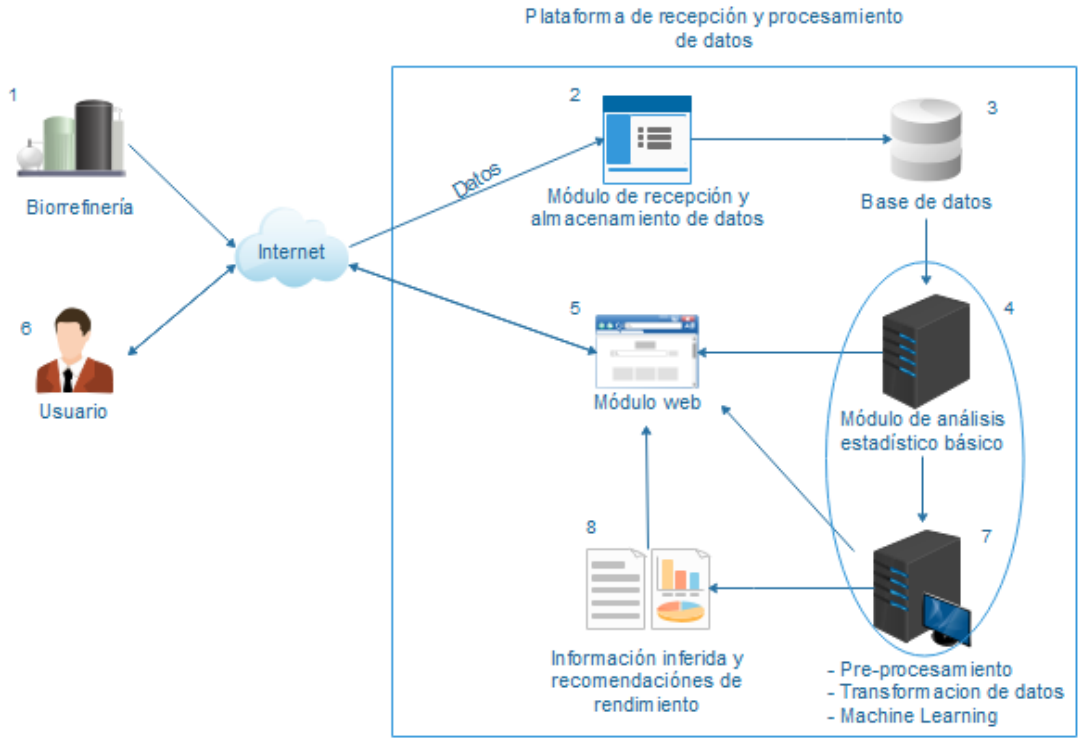

Figura 3.- Plataforma web de recepción y procesamiento de datos producidos por una biorrefinería

5. Metodología

La mteodología propuesta para alcanzar los objetivos de este proyecto son:

- Investigar especificaciones de formatos de datos para almacenar información en una base de datos, para determinar el formato adecuado para la información generada por la biorrefinería.

- Realizar análisis sobre funcionamiento de bases de datos relacionales y no relacionales, con el fin de verificar cuál de éstas permite una mejor gestión de la información generada por la biorrefinería.

- Desarrollar un módulo de recepción y almacenamiento de información que se encargue de recibir toda la información generada por la biorrefinería y que, a su vez, la almacene utilizando un formato establecido en un sistema de base de datos (relacional o no relacional).

- Desarrollar un módulo web que permita a los biotecnólogos realizar una selección de datos específicos entre toda la información almacenada en la base de datos y, de igual manera, que permita visualizar los resultados estadísticos del procesamiento de la información.

- Desarrollar un módulo que se encargue del preprocesamiento de la información seleccionada anteriormente. El preprocesamiento de los datos consiste en tres pasos [12]:

    - **Formatear**: Consiste en definir un formato para los datos con los cuales se va a trabajar.

    - **Limpiar**: Es posible que haya instancias de datos incompletas y que no contengan los datos necesarios, estas instancias pueden ser eliminadas dependiendo si, son útiles para el usuario final o no.

    - **Muestrear**: De los datos seleccionados (e.g, 1000 datos), seleccionar solo una muestra de ellos (e.g, 300 datos) para explorarlos mucho más rápido y reducir tiempos de ejecución, memoria y requisitos de computación.

- Desarrollar un sistema que se encargue de la transformación de los datos, que consiste en tres pasos [12]:

    - **Escalar**: Los datos pre-procesados pueden contener atributos con una combinación de escalas y, los procesos de Machine Learning manejan escalas entre 0 y 1 y, debido a esto, hay que considerar un tipo de escala que sea compatible.

    - **Descomponer**: Algunas instancias de los datos se pueden dividir en partes constituyentes que pueden ser más útiles cuando se aplican métodos de Machine Learning. Un ejemplo es la fecha que tiene los componentes de día y hora que, a su vez, podrían dividirse en más.

    - **Agregar**: Puede haber instancias de datos que tengan el mismo significado cuando se está capturando información. Debido a esto, estas instancias se pueden combinar en una sola.

- Desarrollar un tercer proceso que se encargue de realizar inferencia estadística aplicando algoritmos de Machine Learning correspondientes para determinar si la biorrefinería continuará funcionando correctamente o se detecte un posible fallo.

- Realizar pruebas a la plataforma de recepción y almacenamiento para evaluación de resultados. Las pruebas se realizarán en coordinación con los biotecnólogos encargados de la biorrefinería.

6. Calendario de actividades

Actividades	Periodo Otoño 2018 (Julio - Diciembre)					
	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Investigar formatos de datos	x	x				
Análisis de bases de datos		x	x			
Desarrollo de módulo de recepción y almacenamiento			x	x	x	
Desarrollo de módulo web				x	x	x
Estancia en la Universidad de Guanajuato (Responsable y Estudiante)						x

Actividades	Periodo Primavera 2019 (Enero – Julio)						
	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio
Desarrollo de módulo de preprocesamiento	x	x	x				
Desarrollo de sistema de transformación de datos			x	x			
Desarrollo de módulo de Machine Learning				x	x	x	
Realizar pruebas para evaluación de resultados						x	x

7. Trabajos previos relacionados

En esta sección se presenta un breve análisis de trabajos previos que tienen cierta relación con el tipo de proyecto que se realiza en esta tesis.

En [12], se hace una investigación exhaustiva en todos los aspectos de la diabetes ha llevado a la generación de grandes cantidades de datos. El objetivo del presente estudio es realizar una revisión sistemática de las aplicaciones de aprendizaje automático, técnicas de minería de datos y herramientas en el campo de la investigación de la diabetes con respecto a predicción y diagnóstico, complicaciones diabéticas, antecedentes genéticos y medioambiente, y atención y gestión de la salud.

En [13], se presenta un método de minería de datos web en la adopción de comparación y análisis de paginas web utilizando bases de datos gratuitas del área de química. El bloque de datos de la pagina web se recupera después de comparar y analizar, y luego los datos se extraen para actualizarse en minería de información técnica. El artículo elabora las arquitecturas y composiciones del sistema y el proceso del sistema probado por las bases de datos de propiedades físicas de la química.

En [14], se menciona que las campañas de detección de alto rendimiento (High-throughput screening) en compañías farmacéuticas han acumulado una gran cantidad de datos para varios millones de compuestos en un par de cientos de ensayos. A pesar de la conciencia generalizada de que la gran cantidad de información esta oculta dentro de la gran cantidad de datos, se ha informado poco acerca de un método de minería de datos sistemático que pueda extraer de manera confiable conocimientos relevantes de interés para químicos y biólogos. El artículo menciona el desarrollo de un enfoque de minería de datos basado en un algoritmo llamado identificación de patrones basada en ontología (OPI) y lo aplicaron a una base de datos HTS interna.

En [15], se menciona que la Base de Datos de Biocatálisis / Biodegradación de la Universidad de Minnesota (UM-BBD) proporciona información sobre catabolismo microbiano y biotransformaciones relacionadas, principalmente para contaminantes ambientales. En los últimos dos años, ha aumentado su aliento para incluir más ejemplos de metabolismo microbiano de metales y metaloides, y expandió los tipos de información que incluye para contener biotransformaciones microbianas e interacciones vinculantes con muchos elementos químicos. También ha aumentado las formas en que se puede acceder a estos datos (Minería). La extracción de los datos de la UM-BBD proporciona una visión única de cómo el mundo microbiano recicla los grupos funcionales orgánicos.

En [16], se explica que los métodos de minería de datos del campo de Programación Lógica Inductiva (ILP) tienen ventajas potenciales para los datos químicos estructurales. En este artículo presentan Warmr, el primer algoritmo de minería de datos ILP que se aplicará a los datos quimio informáticos. Se ilustra el valor de Warmr al aplicarlo a una base de datos bien estudiada de compuestos químicos probados por carcinogenicidad en roedores. La extracción de datos se utilizo para encontrar todas las subestructuras frecuentes en la base de datos, y el conocimiento de estas subestructuras frecuentes muestra un valor agregado a la base de datos. Un uso de las subestructuras frecuentes fue convertirlas en reglas de predicción probabilísticas que relacionan la descripción del compuesto con la carcinogénesis. Se descubrió que estas reglas son precisas den los datos de prueba y proporcionan una idea de la relación entre la estructura y la actividad en la carcinogénesis.

En [17], se presenta una introducción de alto nivel a la minería de datos relacionada con la vigilancia de los datos de la salud. La minería de datos se compara con las estadísticas tradicionales, se identifican algunas ventajas de los sistemas de datos automatizados y se describen algunas estrategias y algoritmos de minería de datos. Un ejemplo concreto ilustra los pasos involucrados en el proceso de minería de datos, y se describen tres aplicaciones exitosas de minería de datos en el ámbito de la atención médica.

8. Referencias

1) Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS quarterly*, 1165-1188.
2) Mathioudakis, M., & Koudas, N. (2010, June). Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 1155-1158). ACM.
3) Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.
4) C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012
5) G. Bell and J. N. Gray (1997), The revolution yet to happen, in Beyond Calculation (P. J. Denning and R. M. Metcalfe, eds), Springer, pp. 5–32.
6) Hand, D. J. (2007). Princi9ples of data mining. *Drug safety*, *30*(7), 621-622.
7) Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.
8) M. James, C. Michael, B. Brad, and B. Jacques, Big Data: The Next Frontier for Innovation, Competition, and Productivity. New York, NY: McKinsey Global Institute, 2011.
9) M. Rouse. (2011). Machine Learning Definition. [Online]. Available: http://whatis.techtarget.com/definition/machine-learning
10) Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.
11) Maadane, A., Merghoub, N., Ainane, T., El Arroussi, H., Benhima, R., Amzazi, S., ... & Wahby, I. (2015). Antioxidant activity of some Moroccan marine microalgae: Pufa profiles, carotenoids and phenolic content. *Journal of biotechnology*, *215*, 13-19.Jason Brownlee. (2013). *How to Prepare Data For Machine Learning.* [Online]. https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/
12) Jason Brownlee. (2016). *Supervised and Unsupervised Machine Learning Algorithms.* [Online]. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
13) Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal.
14) Shan, L., Zhao, Y., & Zhang, J. (2007, August). Developing the System of Web-Data Mining from Chemical Database Based on Internet. In Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on (Vol. 4, pp. 26-30). IEEE.
15) Yan, S. F., King, F. J., He, Y., Caldwell, J. S., & Zhou, Y. (2006). Learning from the data: mining of large high-throughput screening databases. Journal of chemical information and modeling, 46(6), 2381-2395.
16) Ellis, L. B., Hou, B. K., Kang, W., & Wackett, L. P. (2003). The University of Minnesota biocatalysis/biodegradation database: post-genomic data mining. Nucleic Acids Research, 31(1), 262-265
17) King, R. D., Srinivasan, A., & Dehaspe, L. (2001). Warmr: a data mining tool for chemical data. Journal of Computer-Aided Molecular Design, 15(2), 173-181.
18) Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. Infection Control & Hospital Epidemiology, 25(8), 690-695.