Aplicaciones de la Minería de datos en la agroindustria azucarera de caña

Osvaldo Gozá-León
Facultad de Ingeniería Química, Universidad Tecnológica de La Habana, José Antonio Echeverría, Cujae. La Habana, Cuba.
ogoza@quimica.cujae.edu.cu

RESUMEN

El aumento continuo de la disponibilidad de datos de toda naturaleza, unido al desarrollo vertiginoso de las tecnologías de la información y las comunicaciones, hace imprescindible el desarrollo de técnicas de Minería de datos que permitan procesar y analizar grandes volúmenes de datos y extraer de ellos información útil. En el campo de la ingeniería, en general, se colecta y almacena gran cantidad de datos, a través de diferentes sensores, en los que la Minería de datos juega un rol importante para la creación de modelos y patrones. Este artículo tiene como objetivo presentar una revisión de las aplicaciones que ha tenido la Minería de datos en la agroindustria azucarera de caña durante los últimos 25 años, tanto la agricultura como la industria y describir las técnicas aplicadas y sus posibilidades en la solución de problemas. La revisión se basó, fundamentalmente, en motores de búsqueda en internet, en combinación con directorios especializados en la temática azucarera. De los trabajos consultados se observa una aplicación creciente de las técnicas de Minería de datos en la agroindustria azucarera de caña, tanto en el sector agrícola como en el industrial, especialmente en los últimos 5 años.
Palabras clave: Minería de datos, azúcar, caña.

ABSTRACT

The continuous increase in the availability of data of all kinds, together with the rapid development of information and communication technologies, makes it essential to develop data mining techniques that process and analyze large volumes of data and extract useful information from them. In the engineering field in general, a large amount of data is collected and stored through different sensors, where data mining plays an important role in the creation of models and patterns. This article aims to present a review of the applications that data mining has had in the sugarcane agroindustry during the last 25 years, including both agriculture and industry, describing which techniques have been applied and their possibilities in the problem solving. The review was based primarily on internet search engines, in combination with sugar specialized directories. From the studies consulted, a growing application of data mining techniques is observed in the sugarcane agroindustry, both in the agricultural and industrial sectors, especially in the last 5 years.
Key words: data mining, sugar, cane.

INTRODUCCIÓN

En las últimas décadas, los avances tecnológicos han posibilitado reunir y almacenar grandes cantidades de datos en las industrias de procesos, que han estado acompañados por el desarrollo de programas de computación capaces de analizar estos datos y extraer información útil para múltiples aplicaciones. En un principio, esas grandes cantidades de datos, rara vez se utilizaban para análisis detallados, sino para comprobaciones técnicas de rutina y cumplimiento de registros de procesos. Posteriormente, la importancia de extraer información de los datos almacenados ha

asumido un papel preponderante en la industria de procesos. Al mismo tiempo, el desarrollo técnico y el análisis de bases de datos se ha desarrollado a un ritmo vertiginoso, lo que ha hecho que las investigaciones sobre Minería de datos y análisis en la industria de procesos sean muy populares. Al analizar los patrones de los datos del proceso y las relaciones entre las variables, se puede extraer información útil, a partir de la cual se pueden desarrollar modelos estadísticos para diversas aplicaciones, como el monitoreo de procesos, el diagnóstico de fallas, el control avanzado, entre otros. En resumen, el objetivo principal de la Minería de datos y el análisis de datos es extraer información útil y procesarla para mejorar la comprensión y la toma de decisiones del proceso (1).

La industria azucarera de caña tiene siglos de existencia y ha experimentado muchos cambios tecnológicos en el cultivo y procesamiento de la caña de azúcar, a los que se han ido incorporando progresivamente avances logrados en las nuevas tecnologías de la información y las comunicaciones; no obstante, aún es operada heurísticamente. Aunque dicha industria no se encuentra entre las que más aplicaciones ha tenido de las técnicas de Minería de datos, se observa un uso creciente de estas en los últimos años, orientadas a mejorar el entendimiento de los procesos y la adquisición de conocimientos efectivos que permitan resolver problemas que, con frecuencia, son difíciles o imposibles de solucionar por las vías tradicionales y permite, con ello, revelar potencialidades para incrementar la eficiencia.

Este artículo tiene como objetivo presentar una revisión de las aplicaciones que ha tenido la Minería de datos en la agroindustria azucarera de caña, durante los últimos 25 años, tanto la agricultura como la industria, y describe las técnicas aplicadas y sus posibilidades en la solución de problemas.

MATERIALES Y MÉTODOS

Minería de datos

El término Minería de datos se utiliza, mayoritariamente, para referirse al proceso genérico correspondiente a las técnicas y herramientas de investigación usadas para extraer información útil de una base de datos. Dentro de estas técnicas se pueden considerar todos los modelos matemáticos y técnicas basadas en aplicaciones de software, para el análisis inteligente de los datos y búsqueda de patrones o tendencias, aplicados de forma iterativa e interactiva.

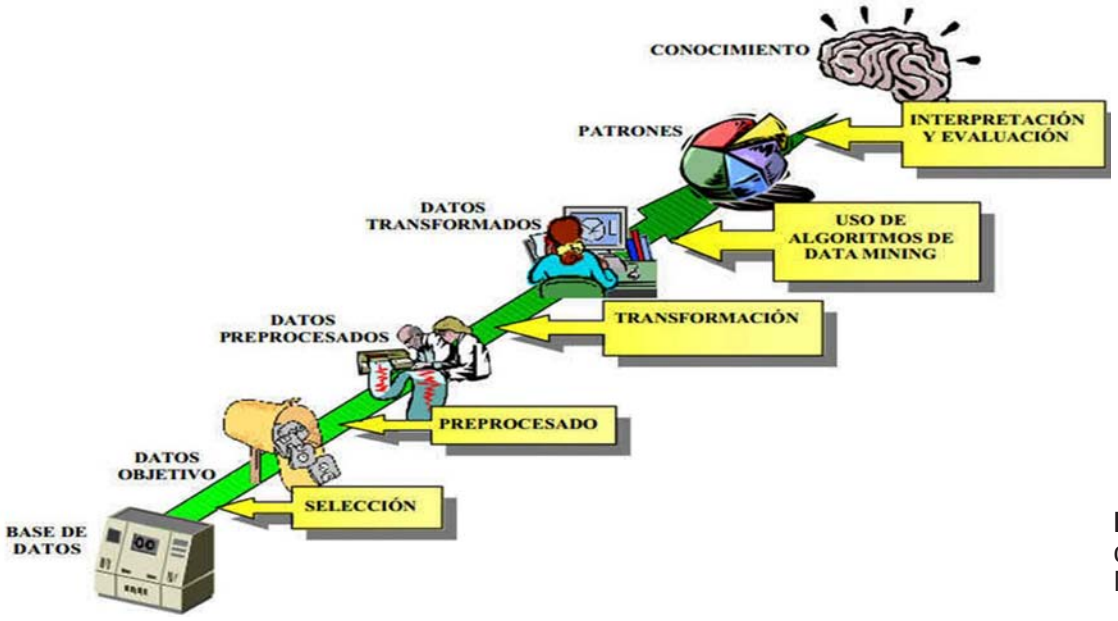

Figura 1. Fases típicas de un proceso de Minería de datos.

Las fases en el proceso global de Minería de datos no están claramente diferenciadas, lo que hace que sea un proceso iterativo e interactivo con el usuario experto. Las interacciones entre las decisiones tomadas en diferentes fases, así como los parámetros de los métodos utilizados y la forma de representar el problema, suelen ser extremadamente complejos. Típicamente el proceso se estructura en las fases que se ilustran en la figura 1 (2).

Los modelos utilizados en Minería de datos pueden ser descriptivos o predictivos. Los modelos descriptivos exploran las propiedades de los datos que se examinan e identifican patrones que explican, resumen o caracterizan dichos datos. Estos modelos permiten acometer tareas, tales como: asociación y agrupamiento. En las tareas de asociación se identifican relaciones no explícitas entre atributos nominales, para reconocer como la ocurrencia de un suceso puede inducir o generar la aparición de otros. En las tareas de agrupamiento se obtienen grupos o *clusters,* a partir de los datos, de manera que los objetos de un mismo grupo son muy similares entre sí y muy distintos a los de otros grupos.

Los modelos predictivos estiman o predicen valores futuros de la variable objetivo del análisis, parten de datos de entrada que se consideran influyentes en su comportamiento. Estos modelos permiten acometer tareas tales, como: clasificación, regresión y predicción. En las tareas de clasificación se examinan los datos y, en función de ellos, se asigna a la variable objetivo (nominal) uno de sus posibles valores. En las tareas de regresión y predicción se examinan los datos y, de acuerdo con ellos, se asigna a la variable objetivo (numérica) uno de sus posibles valores.

La preparación de los datos es el paso inicial en el desarrollo del modelo, el objetivo de este paso es adquirir una visión de los datos del proceso y, con ello, entonces seleccionar los más apropiados para la modelación. La tarea principal de este paso consiste en extraer la información de las bases de datos históricas, examinar la estructura del conjunto de datos y seleccionar las muestras y variables (1). Para asegurar la eficiencia de este paso se deben analizar las características de los datos del proceso, tales como: el alejamiento de la distribución normal y el nivel de correlación entre las variables. Otro aspecto importante es la selección de las variables y las muestras, que está estrechamente relacionado con el paso de desarrollo del modelo (3). Dicha selección depende del tipo de modelo que se quiere desarrollar y constituye la tarea principal a acometer con ese modelo.

En el paso inicial, al conformar el conjunto de datos, es necesario realizar el pre-procesamiento para mejorar su calidad y, pudieran ser necesarias algunas transformaciones apropiadas de los datos, para que el modelado sea más eficiente (4).

Al tener preparado el conjunto de datos de entrenamiento, es posible seleccionar un algoritmo de Minería de datos, apropiado para la construcción del modelo. De acuerdo con el análisis detallado de las características de los datos, la complejidad del modelo a obtener puede ser valorada (3). Seleccionada la estructura del modelo, sus parámetros pueden ser determinados con la implementación de un algoritmo de Minería de datos, con el conjunto de datos de entrenamiento. Finalmente, para que el modelo pueda ser utilizado necesita ser validado, y requiere para ello, también, de un conjunto de datos de validación.

En el proceso de Minería de datos, normalmente, la preparación de los datos es la actividad que mayor cantidad de tiempo y esfuerzo requiere, pues los resultados dependen, en gran medida, de la calidad de los datos. Especialmente, la poca disponibilidad de datos y de calidad dudosa, constituyen un gran obstáculo para el desarrollo de proyectos de inteligencia artificial en países en vías de desarrollo (5).

La Minería de datos utiliza una amplia variedad de modelos; entre los más comunes aplicados en la industria azucarera, valorados en esta revisión, se encuentran: Análisis de componentes principales (ACP), como modelo descriptivo y Redes neuronales artificiales (RNA); Máquinas de soporte vectorial (MSV); Bosque aleatorio (BA) y Árboles de decisión (AD), como modelos predictivos.

Análisis de componentes principales

El ACP es una de las técnicas estadísticas multivariantes más difundidas en el análisis de datos (6). Sus principales objetivos son: extraer la información más importante de un conjunto de datos multivariantes; comprimir un conjunto de datos multivariantes y mantener solo la información que se considere importante (reducir la dimensionalidad de los datos); simplificar la descripción de un conjunto de datos y analizar la estructura de las observaciones y de las variables (7).

La idea central del ACP es reducir la dimensionalidad de un conjunto de datos correspondientes a un gran número de variables y retener, tanto como sea posible, la variación de los datos originales. Esto se logra transformando las variables originales en un nuevo conjunto de variables, combinación lineal de las primarias, que se denominan componentes principales, los cuales no están correlacionados entre sí y son ordenados, de forma tal, que el primer componente retiene la mayor parte de la variación presente en las variables originales (8).

Redes neuronales artificiales

Las RNA son modelos que intentan simular la estructura y los aspectos funcionales de las Redes neuronales biológicas.

Aunque existen muchos tipos de Redes neuronales, estas poseen varias características comunes. Las Redes neuronales están compuestas por numerosos elementos de procesamiento, llamados nodos que, en su conjunto, forman una red. La selección de la arquitectura de la red depende de la tarea que se vaya a realizar y comprende: especificar las características de los nodos, la topología de la red y el algoritmo de entrenamiento. Típicamente las neuronas son agrupadas en diferentes capas, como: capa de entrada, capas de salida y capas ocultas. El uso de las capas ocultas les confiere a las Redes neuronales la habilidad de describir sistemas no lineales. Teóricamente, las RNA son capaces de aproximar cualquier función lineal o no lineal y aprender de los datos, que las convierten en una poderosa herramienta para la modelación, en tareas de clasificación y regresión (1).

Máquinas de soporte vectorial

Las MSV pueden ser utilizadas para problemas de clasificación y de regresión. La idea principal consiste en construir un hiperplano o conjunto de hiperplanos en un espacio de alta dimensionalidad, basado en diferentes tipos de datos, tanto lineales como no lineales, pueden ser separados (1).

En contraste con las RNA, que se desarrollaron de forma heurística con aplicaciones y experimentación y precedieron la teoría, las MSV se desarrollaron primero sobre la base de una teoría sólida y, después, fueron implementadas en la práctica. Una ventaja significativa de las MSV es que, mientras las RNA pueden tener múltiples mínimos locales, la solución de una MSV es global y única.

Árboles de decisión

Un AD, como su nombre lo indica, es una herramienta que utiliza un gráfico en forma de árbol, para describir las relaciones entre las distintas variables y la toma de decisiones. Los AD son comúnmente usados en investigación de operaciones, particularmente en el análisis de decisiones, para ayudar a identificar la estrategia más probable y lograr el objetivo. Recientemente han sido, además, introducidos en las industrias de procesos asociados a tareas de clasificación, encontrándose entre sus aplicaciones más comunes el monitoreo de procesos, el diagnostico de fallas y la predicción de la calidad (1).

Lo que hace especiales a los árboles de decisión entre los modelos de Minería de datos es su claridad en la representación de la información. El conocimiento aprendido en el entrenamiento es directamente formulado con una estructura jerárquica que lo representa de manera tal que es fácilmente comprendido, incluso, por los que no son expertos.

Bosques aleatorios

El BA es un conjunto de árboles de decisión, cada uno de los cuales realiza una predicción, de forma independiente. Puede ser usado para tareas, tanto de clasificación como de regresión. Los BA ajustan árboles de decisión separados a un número predefinido de conjuntos de datos de arranque. Para cada árbol, los datos se dividen de forma recursiva en unidades más homogéneas, que comúnmente se denominan nodos, con el fin de mejorar la previsibilidad de la variable de respuesta. Los puntos de división se basan en valores de variables predictoras, las que se consideran variables explicativas importantes. El valor predicho de una respuesta categórica (clasificación) es la moda de las clases de todos los árboles de decisión ajustados individuales y el valor predicho de una respuesta continua (regresión) es la respuesta media ajustada de todos los árboles individuales que resultaron de cada muestra procesada (1).

A diferencia de las Redes neuronales que pueden procesar muchos tipos diferentes de datos, los BA solo pueden trabajar con datos en formato tabular; sin embargo, son más fáciles de entrenar que las Redes neuronales.

RESULTADOS Y DISCUSIÓN

Aplicaciones en la agroindustria azucarera

La revisión se basó, fundamentalmente, en búsquedas electrónicas con el motor *Google Scholar*, con palabras claves, como: azúcar y caña, junto con los nombres de las técnicas de Minería de datos más usadas, en combinación con directorios especializados en la temática azucarera; abarcó el periodo de 1997 a 2021.

En la figura 2 se muestra, de forma resumida, la proporción en que se han usado las técnicas aplicadas en su conjunto; se puede observar que las técnicas RNA son las de más amplio uso y las técnicas de AD las de menor uso.

Las RNA han sido aplicadas mayoritariamente en el sector industrial, vinculadas a la solución de problemas, sobre todo, en la cristalización (9 - 18) y, en menor medida, en la clarificación (19 - 21) y en la molienda de la caña (22, 23).

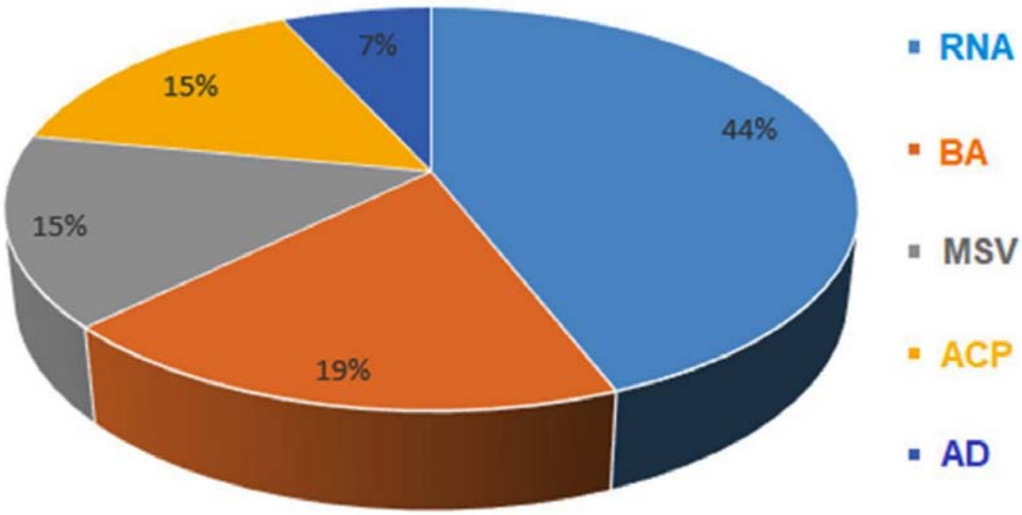

Figura 2. Resumen de aplicaciones de la Minería de datos en la agroindustria azucarera de caña.

Las aplicaciones de las RNA en la cristalización están asociadas al monitoreo y control automático, y son capaces de lidiar con la dinámica compleja y el alto carácter no lineal de este proceso. En una de las primeras aplicaciones se realizó una modelación híbrida en la que se combinó el balance

de masa de los cristales de sacarosa con una Red neuronal que modeló la velocidad de crecimiento de los cristales (9), que incidió también en su fase de crecimiento, en otras aplicaciones se logró mejorar la estrategia de control en comparación con el control standard PID, mediante el control predictivo basado en Redes neuronales (14, 16). Con el uso de Redes neuronales, ha sido posible desarrollar un sistema para la medición automática, en tiempo real, de la distribución del tamaño de los cristales, que presenta ventajas significativas en comparación con el método de muestreo tradicional, basado en mediciones individuales de cristales sencillos (12). El uso de las Redes neuronales ha permitido realizar una modelación preliminar de la pureza de referencia de las mieles finales cubanas, y tener en cuenta los contenidos de glucosa, fructosa, cenizas, polisacáridos, oligosacáridos y la relación impurezas/agua en las mieles, en contraste con los modelos tradicionales que solo tienen en cuenta los contenidos de glucosa, fructosa y cenizas y presentan una correlación pobre (15). También el uso de la Redes neuronales ha facilitado el monitoreo en línea, de la pureza del licor madre en masas cocidas de tercera (13), así como la estimación de la sobresaturación en la etapa de cristalización, en función de la concentración, la temperatura y la pureza de la solución (17). Más recientemente, con las Redes neuronales profundas, se ha implementado un sistema preciso de clasificación de imágenes de los cristales de azúcar como base para el control automático de la cristalización (18).

Un sistema de clasificación basado en procesamiento de imágenes y Redes neuronales, que permite realizar la identificación automática de los tipos de azúcar, a partir de las características de color y textura extraídas de las imágenes de los cristales, supera en rapidez y precisión al procedimiento tradicional que se hace manualmente, ya sea visualmente o con instrumentos en el laboratorio (24).

Las aplicaciones de las RNA en la clarificación están asociadas al control automático del pH. Su uso combinado con la programación dinámica ha mejorado el control en tiempo real y en línea del pH del jugo clarificado (19, 20). Recientemente, fue concebido un sistema en el que también se utiliza la lógica difusa (21).

En el caso de la molienda de la caña, un modelo neuronal permitió optimizar los parámetros en el tándem de molinos y encontrar las distancias entre las mazas con las que se obtiene la máxima extracción de jugo (22). Con este mismo objetivo se utilizó una Red neuronal en combinación con un controlador inteligente que supera el clásico controlador PID (23).

Se desarrolló un método analítico rápido y de bajo costo, basado en Redes neuronales, que utiliza un sistema de adquisición de imágenes y permite determinar el porciento de materias extrañas en la caña, como materia prima de la fábrica (25).

Los sistemas agrícolas tienen interacciones complejas con el medioambiente y la tierra, que pudieran ser mejor entendidas con la aplicación de la computación. Las interacciones son tan complejas que es imposible cuantificar sus efectos acumulativos sin la aplicación de las últimas herramientas computacionales. La modelación de estos sistemas tiene un futuro promisorio y puede abrir nuevas fronteras que pueden ser de ayuda en las transiciones agroecológicas, dado el vínculo existente entre las variables ambientales y varios procesos fisiológicos, pues es posible predecir la respuesta de las cosechas, a partir de un conjunto de condiciones ambientales. Asimismo, la aplicación de diferentes modelos a diferentes escalas pudiera ayudar a comprender los mecanismos desde el punto de vista cualitativo y cuantitativo (26).

En el sector agrícola, las RNA también han tenido diversas aplicaciones. Con Redes neuronales profundas, a partir de imágenes de las hojas de la caña de azúcar, es posible identificar si existen enfermedades en la planta (27). Una Red neuronal permite predecir el rendimiento en caña de cultivos, basado en un conjunto amplio de parámetros que diferencian las distintas variedades de caña, como son: la estructura, el color y tamaño de las hojas, que ha sido de ayuda para identificar el tipo de caña a ser cultivada en determinada región, en busca de los mayores beneficios (28). Un modelo neuronal permitió predecir la cosecha de la caña de azúcar (toneladas de caña), a partir de datos

históricos de variables climatológicas de los cultivos agrícolas de caña de azúcar (temperaturas máximas y mínimas, oscilación térmica, precipitaciones, heliofanía, humedad relativa, evaporación) y hectáreas de los cultivos sembrados, con resultados muy aproximados a los obtenidos por el método tradicional de aforo (29). Un sistema implementado en una máquina cosechadora, con la ayuda de una Red neuronal integra los datos de múltiples sensores para medir el flujo másico de caña en tiempo real y facilitar, de esta forma, la automatización y el monitoreo de la cosecha (30). En contraste con la vía tradicional para estimar el azúcar recuperable, a partir de la pureza del jugo, una técnica simple permite realizar esta estimación mediante un modelo neuronal, en función de propiedades bioeléctricas de la caña (31). Como una vía alternativa más rápida y de menor costo, en comparación con el método tradicional, se presenta un modelo neuronal que permite estimar, con alta precisión, el valor de Pol del jugo de la caña, a partir de los grados Brix y el peso de la fibra húmeda (32). La aplicación de un modelo neuronal fue de utilidad en la identificación de los factores que influyen en las emisiones de CO2 provenientes de la tierra, inducidas por el manejo de las cosechas en áreas agrícolas cañeras de Brasil, con variables, como: la humedad y temperatura de la tierra, las precipitaciones, el pH y el carbono orgánico (33).

Las Redes neuronales también se han aplicado para la estimación de emisiones en la industria; se aplicó un modelo neuronal para el monitoreo de las emisiones de CO2 en los hornos de la industria azucarera, se tuvieron en cuenta las cantidades de los combustibles quemados (bagazo, madera y petróleo) y los factores de emisión, de acuerdo con las pautas establecidas para su cálculo por el Grupo intergubernamental de expertos sobre el cambio climático (34).

Los BA, en la revisión realizada, han encontrado aplicación solo en el sector agrícola, útiles por su capacidad predictiva para manejar muchas combinaciones diferentes de variables climáticas y de las cosechas. Por medio de los Bosques aleatorios ha sido posible explicar, tanto con modelos de clasificación como de regresión, la variación anual del rendimiento en caña, en regiones de Australia (35); así como también predecir dicho rendimiento, a partir de bases de datos de varios centrales, en Brasil (36) y de datos de sensores remotos en áreas cañeras de China (37). Los BA junto con los MSV permitieron clasificar la variedad de la caña y el ciclo de cosecha con imágenes captadas por sensores remotos (38). Basado en la capacidad de clasificación de los Bosques aleatorios, se desarrolló un método para el mapeo de la caña de azúcar, al inicio de la temporada a partir de imágenes de alta resolución espacio-temporal, que superan limitaciones de investigaciones anteriores relacionadas con las condiciones del tiempo y el periodo de crecimiento (39). Con el uso de BA se implementó un método que brinda la posibilidad de diferenciar la planta de la caña de azúcar de las malas hierbas a partir de imágenes espectrales de las hojas, que facilitan la aplicación localizada de herbicidas solo donde se necesita, como un precepto de la agricultura de precisión (40). Los BA presentan resultados similares a las RNA en la clasificación de plantaciones cañeras en Brasil con imágenes satelitales (41). Los BA permiten, con resultados similares a los modelos de regresión lineal múltiple, mejorar la predicción de la biomasa de la caña al integrar los datos de la cosecha, tradicionalmente utilizados con variables biométricas, tales como: el número de tallos y la altura de la planta (42).

La estimación del contenido de azúcar en la caña, es una información muy valiosa para las fábricas, que tradicionalmente han utilizado los promedios históricos o las curvas de maduración como una vía adicional; con Bosques aleatorios, se ha desarrollado un modelo que permite realizar esta estimación de forma satisfactoria, a partir de información sobre la cosecha (43). Un modelo de clasificación basado en BA permite predecir si el porciento de azúcar recuperable de la caña supera determinado valor deseable de referencia, a partir de datos sobre diversas variables que incluyen factores climáticos y características de la tierra y el cultivo, este constituye una herramienta útil para la determinación de estrategias apropiadas que logren mayores producciones (44).

Las MSV han encontrado mayor aplicación en el sector agrícola (36, 38, 45 - 48) que en el industrial (49 - 51).

Las aplicaciones industriales de las MSV se han basado en su capacidad predictiva. Las MSV han superado a las RNA en los resultados predictivos del estado de cristalización del azúcar, en función de la distribución del tamaño de los cristales, a partir del Brix y temperatura del sirope, el vacío, la presión y temperatura del vapor y el flujo de licor alimentado (49). En el control automático de la cristalización han tenido una aplicación efectiva las MSV al estimar, con gran precisión, la pureza y la sobresaturación del licor madre y considerar como variables de entrada: el vacío, la temperatura, Brix y nivel de la masa cocida, la presión y temperatura del vapor y el flujo de alimentación (51). A partir de mediciones de espectroscopia infrarroja, un modelo basado en MSV permite monitorear de forma más efectiva la calidad, mediante la determinación del Brix y el Pol en jugo mezclado, meladuras, masas cocidas y mieles finales (50).

En la agricultura, dos aplicaciones ya presentadas hacen un uso combinado de las MSV con los BA (38, 36). Al igual que los BA, las MSV han sido también utilizadas para predecir el rendimiento en caña (48, 47) y en la clasificación y mapeo de campos de caña (45, 46).

El ACP ha tenido la mayoría de sus aplicaciones en la agricultura (52 - 58). Ha sido de utilidad para investigar una asociación entre la enfermedad de la roya y los nutrientes potasio, calcio, magnesio y silicio (52). Facilitó la identificación de los indicadores agronómicos y fisiológicos que pueden ayudar a determinar la eficiencia en el uso del nitrógeno de las variedades de caña (57). Su uso combinado con BA posibilitó obtener un modelo que permite predecir la precipitación media superior, en una región cañera de Australia (54). El ACP proporciona una herramienta gráfica adecuada para el análisis visual de datos provenientes de experimentos, en los cuales un grupo de genotipos de caña de azúcar son evaluados en diferentes ambientes, cuyos resultados permiten identificar cultivos de alto rendimiento (53). Asimismo, ha sido una técnica efectiva en la selección de diferentes variedades de caña, con alto contenido de biomasa (55). Mediante regresión de mínimos cuadrados parciales, cuyo fundamento lo constituye el ACP, se ha desarrollado un sistema de medición rápido y preciso de los contenidos de sacarosa y componentes iónicos en el jugo de la caña con datos espectroscópicos (56). También mediante regresión de mínimos cuadrados parciales y con datos espectroscópicos, se implementó un método de monitoreo de la calidad de la caña que determina el Brix, el Pol, la fibra y el total de azúcar recuperable (58).

El método tradicional ICUMSA de determinación de color en la industria ha sido mejorado mediante el ACP, pues permite realizar una distinción más fina entre los colorantes (59). Un modelo basado en regresión de mínimos cuadrados parciales facilita determinar con rapidez y precisión las concentraciones de sacarosa, glucosa y fructosa en mieles finales, a partir de datos de cromatografía líquida de alta resolución (60).

Los AD, en la revisión realizada, han encontrado aplicación solo en el sector agrícola, basadas en su capacidad clasificatoria. Han sido de utilidad en la discriminación de variedades de caña de azúcar en diferentes tipos de suelos, a partir de imágenes satelitales (61, 62). Se ha demostrado su potencial para explicar la variación del rendimiento en caña al considerar un gran número de variables que incluyen factores relacionados con el clima, la química del suelo y las estrategias de corte (63). También han sido de ayuda en la determinación de los factores que propician altos rendimientos en caña, a partir de la medición de un conjunto amplio de variables climáticas (64).

CONCLUSIONES

Se consultaron 59 trabajos del período comprendido entre 1997 y 2021, en ellos se observó una utilización creciente de las técnicas de Minería de datos en la agroindustria azucarera de caña, tanto

en el sector agrícola como en el industrial, especialmente en los últimos 5 años, con 42 trabajos, que representan el 71 % del total revisado.

Las técnicas aplicadas son, por su orden: las Redes neuronales artificiales (26), Bosques aleatorios (11), Máquinas de soporte vectorial (9), Análisis de componentes principales (9) y Árboles de decisión (4).

En los trabajos revisados se observó que las técnicas RNA, MSV y ACP han sido aplicadas, tanto en el sector agrícola como el industrial, mientras que BA y AD han sido aplicadas solo en el sector agrícola.

En el sector industrial**,** la mayoría de las aplicaciones están relacionadas, por su orden, con los procesos de cristalización, clarificación y molienda; mientras que en el sector agrícola, la mayoría de las aplicaciones están relacionadas con el rendimiento en caña de los cultivos y los parámetros de calidad de la caña.

La aplicación de estas técnicas ha sido de mucha utilidad en la solución de variados problemas que, con frecuencia, son difíciles o imposibles de resolver por las vías tradicionales y han permitido revelar potencialidades para mejorar los procesos e incrementar la eficiencia.

REFERENCIAS BIBLIOGRÁFICAS

1. Zhiqiang Ge, Zhihuan Song, Steven X. Ding and Biao Huang. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. IEEE Access. 5: 20590–20616, 2017.
2. Martínez, F.J. Optimización mediante técnicas de Minería de datos del ciclo de recocido de una línea de galvanizado [Tesis doctoral]. Universidad de La Rioja; 2003.
3. Bishop, C. M. Pattern Recognition and Machine Learning. New York, NY, USA: Springer-Verlag, 2006.
4. Shu Xu, Bo Lu, Michael Baldea, Thomas F. Edgar, Willy Wojsznis, Terrence Blevins et al. Data cleaning in the process industries. Rev. Chem. Eng. 31(5): 453–490, 2015.
5. Kshetri, Nir. Artificial intelligence in developing countries. IEEE IT Professional, 22(4): 63-68, 2020.
6. Polanco J.M. El papel del análisis por componentes principales en la evaluación de redes de control de la calidad del aire. Comunicaciones en Estadística. 9(2): 271–294, Agosto 2016.
7. Abdi, H. & Williams, L. J. Principal Component Analysis. Wiley Interdisc. Rev. Comp. Stat. 2(4): 433–459, 2010.
8. Gupta V., Mittal M., Chand P. And Kumar P. Principal Component Analysis & Factor Analysis as an Enhanced Tool of Pattern Recognition. International Journal of Electrical and Electronic Engineering & Telecommunications. Special Issue, 1(2): 73-78, July 2015.
9. P. Lauret, H. Boyer, J.C. Gatina. Hybrid modelling of a sugar boiling process. Control Engineering Practice. 8: 299-310, 2000.
10. SD Peacock. An introduction to neural networks and their application in the sugar industry. Proc S Afr Sug Technol Ass. 72: 184–191, 1998.
11. M. Benne, B. Grondin-Perez, J.-D. Lan-Sun-Luk, J.-P Chabriat. Neural Networks Models of Evaporators and Crystallisation Processes in Sugar Cane Industry. Proceedings of the XXIII ISSCT Congress, New Delhi, India, 22-26 February, 1:173-181, 1999.
12. Aubrey Z Mhlongo and Michael J Alport. Application of artificial neural network techniques for measuring grain sizes during sugar crystallisation. Proc S Afr Sug Technol Ass. 76: 460–468, 2002.
13. C. Bonnecaze, B. Grondin-Perez, M. Benne And J.P. Chabriat. Neural networks model of mother liquor purity of C massecuite. Proc. Int. Soc. Sugar Cane Technol. 24: 140-142, 2001.

14. S. Beyou, B. Grondin-Perez, M. Benne, C. Damour, and J.-P. Chabriat. Control Improvement of a C Sugar Cane Crystallization Using an Auto-Tuning PID Controller Based on Linearization of a Neural Network. Control and Information Engineering. 3(6): 1646–1651, 2009.
15. Gozá León O., Santana R., Kazemian H., Hormaza J., Montero-Sánchez Y. Aplicación de las Redes neuronales en la modelación de la pureza de referencia de mieles finales cubanas. Revista ATAC. 3: 9–12, 2010.
16. Paz LA, Georgieva P. and Feyo S.. Model Predictive Control Strategies for Batch Sugar Crystallization Process. Advanced Model Predictive Control. InTech, 2011, 225-246.
17. Morales H., di Sciascio F., Amicarelli A. Estimation of Supersaturation in the Crystallization Process of the Sugar Industry. 2018 Argentine Conference on Automatic Control (AADECA), Buenos Aires, Argentina, 7-9 Nov 2018.
18. Zhang J., Meng Y., Wu J., Qui J., Wang H., Yao T. et al. Monitoring sugar crystallization with deep neural networks. Journal of Food Engineering. 280, September 2020, 109965.
19. Xiaofeng Lin, Shengyong Lei and Huixia Liu. Neural Network Modeling and HDP for Neutralized pH Value Control in the Clarifying Process of Sugar Cane Juice. Proceedings of the World Congress on Engineering and Computer Science WCECS 2008 (San Francisco, USA), October 22-24, 2008.
20. Lin X., Yang J., Liu H., Song S., Song C. An improved method of DHP for Optimal Control in the Clarifying Process of Sugar Cane Juice. Proceedings of International Joint Conference on Neural Networks (Atlanta, Georgia, USA), June 14-19, 2009.
21. Kumar S., Singh G., Singh A., Kumar P. Neuro-Fuzzy Modeling of PH Neutralization Process in Sugar Mill. International Conference on Advances in Computing, Communication & Materials (ICACCM), August 21-22, 2020.
22. D Oktarini, A S Mohruni, S Sharif, M Yanis. Optimum Milling Parameters of Sugarcane Juice Production Using Artificial Neural Networks (ANN). Journal of Physics: Conf. Series 1167 (2019) 012016.
23. V. Saravana, R. Bharani. Design and Analysis of an Intelligent PID Controller for Sugar Industry Process. 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). 11-12 December, 2019.
24. Rizky A., Susanto A., Litasari. Comparison between Colour Models in Automatic Identification of Cane Sugar. 2013 IEEE International Conference on Computational Intelligence and Cybernetics, Indonesia, 3–4 Dic. 2013.
25. Nascimento W., Janoni L., Regina E., Verbi F. Sugarcane Stalk Content Prediction in the Presence of a Solid Impurity Using an Artificial Intelligence Method Focused on Sugar Manufacturing. Food Analytical Methods. 13: 140–144, 2020.
26. Ahmed M., Ahmad S. Systems Modeling. Springer Singapore, 2020.
27. Sammy V. Militante, Bobby D. Gerardo, Ruji P. Medina. Sugarcane Disease Recognition using Deep Learning. 2019 IEEE Eurasia Conference on IOT, Communication and Engineering, 575-578, 2019.
28. Rajesh S Budihal, S Krishna Anand. Constructing an appropriate neural network for maximising sugarcane yield in a particular region. Australian Journal of Wireless Technologies, Mobility & Security. 1(1), 2019.
29. Mendoza-Haro I., Marquetti-Nodarse H. Redes Neuronales Artificiales: factores que determinan la cosecha de caña en la industria azucarera. Revista Ciencia UNEMI. 12(29): 36-50, Enero-Abril 2019.
30. Lima J.d.J.A.d., Maldaner L.F., Molin J.P. Sensor Fusion with NARX Neural Network to Predict the Mass Flow in a Sugarcane Harvester. Sensors, 21, 4530, 2021.

31. Sucipto S., Arwani M., Hendrawan Y., Widaningtyas S., Al Riza D.F., Yuliatun S. et al. Bioelectrical measurement for sugar recovery of sugarcane prediction using artificial neural network. Proceeding of EECSI 2018: 652-656, Malang-Indonesia, 16-18 Oct 2018.
32. Coelho AP, Bettiol JV, Dalri AB, Fischer JA, de Faria RT & Palaretti LF. Application of artificial neural networks in the prediction of sugarcane juice Pol. Revista Brasileira de Engenharia Agrícola e Ambiental. 23(1): 9-15, 2019.
33. Farhate CVV, Souza ZMd, Oliveira SRdM, Tavares RLM, Carvalho JLN. Use of data mining techniques to classify soil CO2 emission induced by crop management in sugarcane field. PLoS ONE 13(3): e0193537, 2018.
34. Saleh C., Chairdino RA, Nizam M, Baba Md Deros, Rachman N. Prediction of CO2 Emissions Using An Artificial Neural Network The Case of the Sugar Industry. Advanced Science Letters. 21: 3079–3083, 2015.
35. Everingham Y., Sexton J., Skocaj D., Inman-Bamber G. Accurate prediction of sugarcane yield using a random forest algorithm. Agron. Sustain. Dev. 36: 27, 2016.
36. Hammer R.G., Sentelhas P.C. & Mariano J.C.Q. Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. Sugar Tech 22: 216–225, 2020.
37. Jing-Xian Xu, Jun Ma, Ya-Nan Tang, Wei-Xiong Wu, Jin-Hua Shao, Wan-Ben Wu 1 et al. Estimation of Sugarcane Yield Using a Machine Learning Approach Based on UAV-LiDAR Data. Remote Sensing. 12: 2823, 2020.
38. Everingham Y., Lowe K.H., Donald D.A., Coomans D.H., Markley J. Advanced satellite imagery to classify sugarcane crop characteristics. Agron. Sustain. Dev. 27: 111–117, 2007.
39. Jiang H, Li D, Jing W, Xu J, Huang J, Yang J et al. Early Season Mapping of Sugarcane by Applying Machine Learning Algorithms to Sentinel-1A/2 Time Series Data: A Case Study in Zhanjiang City, China. Remote Sensing. 11: 861, 2019.
40. de Souza MF, do Amaral LR, de Medeiros SR, Neris MA, Ferreira C. Spectral differentiation of sugarcane from weeds. Biosystems Engineering. 190: 41-46, 2020.
41. dos Santos FH, dos Santos JP and Falcão AC. Evaluating methods to classify sugarcane planting using convolutional neural network and random forest algorithms. International Journal of Development Research. 10(12): 42807-42811, December 2020.
42. Grespan M., Martins F.M., de Medeiros S.R., Rios L. Biometric characteristics and canopy reflectance association for early-stage sugarcane biomass prediction. Sci. Agric. 76(4): 274-280, July/August 2019.
43. Pires M., Ferreira F., Antunes L.H. Rodrigues. From spreadsheets to sugar content modeling: A data mining approach. Computers and Electronics in Agriculture, 132: 14-20, 2017.
44. Nadernejad, F., Din mohammad, I., Rasouli, M. A Data-driven Model for Predicting the Yield of Recoverable Sugar from Sugarcane. Journal of Agricultural Machinery, 2021; (): -. doi: 10.22067/jam.2021.69805.1034
45. Virnodkar S.S., Pachghare V.K., Patil V.C., Jha S.K., Virnodkar S.S. Application of Machine Learning on Remote Sensing Data for Sugarcane Crop Classification: A Review. ICT Analysis and Applications. 93: 539-555, 2020.
46. Moreira EFA, Barbosa MHP, Peternelli LA. Can statistical learning models make early selection among sugarcane families easier and still efficient?. Crop Science. 61:456–465, 2020.
47. AWM Gaffar and IS Sitanggang. Spatial model for predicting sugarcane crop productivity using support vector regression. 2019 IOP Conf. Ser.: Earth Environ. Sci. 335 012009, 2019.
48. Medar R. A., Rajpurohit V. S. and A.M. Sugarcane Crop Yield Forecasting Model Using Supervised Machine Learning. I.J. Intelligent Systems and Applications, 8:11-20, 2019.

49. Meng Y., Yu X., He H., Tang Z., Wang X., Chen J. Knowledge-based modeling for predicting cane sugar crystallization state. International Journal On Smart Sensing and Intelligent Systems. 7(3), September 2014.
50. Ramírez-Morales I., Rivero D., Fernández-Blanco E., Pazos A. Optimization of NIR calibration models for multiple processes in the sugar industry. Chemometrics and Intelligent Laboratory Systems. 159:45-57, 2016.
51. Meng Y., Lan Q., Qin J., Yu S., Pang H., Zheng K. Data-driven soft sensor modeling based on twin support vector regression for cane sugar crystallization. Journal of Food Engineering. 241:159-165, January 2019.
52. P Cadet, SA McFarlane and JH Meyer. Association between nutrients and rust in sugarcane in Kwazulu-Natal. Proc S Afr Sug Technol Ass 7: 223-229, 2003.
53. J.L. Queme, H. Orozco and M. Melgar. GGE biplot analysis used to evaluate cane yield of sugarcane (Saccharum spp.) cultivars across sites and crop cycles. Proc. Int. Soc. Sugar Cane Technol., Vol. 27, 2010.
54. McKinna, L., & Everingham, Y. Seasonal climate prediction for the Australian sugar industry using data mining techniques. Knowledge-Oriented Applications in Data Mining, 109, 2011.
55. Santchurn D., Ramdoyal K., Houssen MG., Labuschagne M. From sugar industry to cane industry: investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. Euphytica. 185:543–558, 2012.
56. E Taira, M Ueno, K Watanabe, Y Kawamitsu and K Yoshimoto. Non-destructive measurement system for process control using combined spectroscopic data. Proc. Int. Soc. Sugar Cane Technol., 29:656-659, 2016.
57. Yang Y, Gao S, Jiang Y, Lin Z, Luo J, Li M et al. The Physiological and Agronomic Responses to Nitrogen Dosage in Different Sugarcane Varieties. Front. Plant Sci. 10:406, 2019.
58. Corrêdo, L.d.P.; Maldaner, L.F.; Bazame, H.C.; Molin, J.P. Evaluation of Minimum Preparation Sampling Strategies for Sugarcane Quality Prediction by vis-NIR Spectroscopy. Sensors, 21, 2195, 2021.
59. Mersad, A., Lewandowski, R., Heyd, B., & Decloux, M. Colorants in the sugar industry. Int. Sugar Jnl, 105(1254):269-281, 2003.
60. R Mueangmontri, P Chapanya, R Nootas, C Ngasan and U Pliansinchai. Evaluation of a near-infrared spectrophotometer for determining molasses quality. Proc. Int. Soc. Sugar Cane Technol., 29:651-655, 2016.
61. Goltz, E., Arcoverde, G. F. B., de Aguiar, D. A., Rudorff, B. F. T., & Maeda, E. E. Data mining by decision tree for object oriented classification of the sugar cane cut kinds. In 2009 IEEE International Geoscience and Remote Sensing Symposium (Vol. 5, pp. V-405). IEEE, July 2009.
62. Kai P.M., da Costa R.M., de Oliveira D.M., Fernandes D.S.A., Felix J. Discrimination of Sugarcane Varieties by Remote Sensing: A Review of Literature. 2020 IEEE 44th Annual Computers,Software and Applications Conference (COMPSAC), 13-17 July 2020.
63. Rodrigues P., Ferreira F., Antunes L.H. Identification of patterns for increasing production with decision trees in sugarcane mill data. Sci. Agric. 76(4): 281-289, July/August 2019.
64. R. Revathy, P. Murali and S. Balamurali. Hadoop Big Data Mining An Effective Mapreduce Tool For Classifying Sugarcane Yield Data. Plant Archives Volume, 20(2):4245-4250, 2020.