Identificación de patrones que influyen en los bajos rendimientos industriales, en la fabricación de azúcar

Identification of patterns that influence low industrial yield in sugar manufacture

Yohan Gil Rodríguez
yohangilrod@gmail.com • https://orcid.org/0000-0002-8239-4124
EMPRESA AZCUBA

Raisa Socorro Llanes
raisa@ceis.cujae.edu.cu • https://orcid.org/0000-0002-2627-1912
UNIVERSIDAD TECNOLÓGICA DE LA HABANA, CUJAE

Lérida Hernández Nodarse
lerida.hernandez@datazucar.cu
AZCUBA

Recibido: 2024-01-11 • Aceptado: 2024-03-30

RESUMEN

La informatización de los procesos de la industria azucarera genera cuantiosos datos que son almacenados de forma creciente al paso de los años. En la actualidad, la aplicación de los programas de la plataforma agroindustrial existente en la Organización Superior de Dirección para la Agroindustria Azucarera (AZCUBA), ha garantizado la rapidez y calidad de las informaciones de zafra. La industria azucarera cubana requiere implementar herramientas y métodos científicos que permitan identificar patrones y comportamientos ocultos en sus datos históricos. En este artículo se expone el empleo de técnicas de extracción de conocimientos, a partir de datos para la identificación de las causas que están incidiendo en los bajos rendimientos industriales. Entre los materiales empleados están las bases de datos de diez años de zafra (2010-2019), que presentan cada una más de 4 millones de registros transaccionales y una media de 578 indicadores por año. La metodología seleccionada para establecer un marco de trabajo del ciclo de vida del proceso de minería de datos fue CRISP-DM. La herramienta seleccionada para aplicar las técnicas de minería de datos fue la plataforma de análisis de datos KNIME. Se realizó un análisis predictivo de los datos, en el cual se emplean los métodos simbólicos. Se comparan las métricas de siete algoritmos de aprendizaje automático: CONJUNCTIVERULE, DECISIONTABLE, RIDOR, FURIA, PART, JRIP, J48, para la selección de características, y se determinó la selección del algoritmo J48 para la clasificación. Se obtienen y validan los atributos que influyen en los bajos rendimientos industriales. Se logra crear las bases para un análisis más profundo de las medidas organizativas y de control necesarias, con el objetivo de perder azúcar en el proceso industrial.

Se recomienda realizar un análisis prescriptivo de los datos, para predecir escenarios logísticos.

Palabras clave: análisis predictivo, aprendizaje de reglas, árboles de decisión, minería de datos, rendimiento industrial azucarero

ABSTRACT

The computerization of the processes of the sugar industry generates large amounts of data that are increasingly stored over the years. At present, the application of the programs of the Agro-Industrial Platform existing in AZCUBA, has guaranteed the speed and quality of the harvest information. The Cuban sugar industry requires the implementation of scientific tools and methods that allow the identification of hidden patterns and behaviors in its historical data. This paper exposes the use of knowledge extraction techniques from data to identify the causes that are influencing low industrial yields. Among the materials used are the databases of ten years of harvest (2010-2019) that each present more than 4 million transactional records and an average of 578 indicators per year. The selected methodology to establish a data mining process life cycle framework was CRISP-DM. The tool selected to apply the data mining techniques was the KNIME data analysis platform. A predictive analysis of the data was carried out, in which the symbolic methods of the family of prediction methods are used. Metrics of seven machine learning algorithms are compared: CONJUNCTIVERULE, DECISIONTABLE, RIDOR, FURIA, PART, JRIP, J48 for feature selection and J48 algorithm selection for classification was determined. The attributes that influence low industrial yields are obtained and validated. It is possible to create the bases for a deeper analysis of the organizational and control measures necessary to stop losing sugar in the industrial process. It is recommended to carry out a prescriptive analysis of the data to predict future logistics scenarios.

Keywords: Predictive Analysis, Rules Learning, Decision Trees, Data Mining, Sugar Industrial Yield

INTRODUCCIÓN

Cuba posee una rica tradición de más de cuatro siglos en la producción de azúcar de caña. Esta industria hoy está afectada por carencias de materia prima, ineficiencia productiva, altos precios del petróleo y afectaciones climatológicas (como la sequía), todo lo cual provoca que se produzcan bajos rendimientos (Cruz et al., 2015).

La política agroindustrial vigente en el país reconoce la necesidad de continuar incrementando la eficiencia agrícola e industrial del sector, así como asegurar el cumplimiento de los programas de producción de caña de azúcar, la modernización del equipamiento y la mejoría del aprovechamiento de la capacidad de molida. No obstante, en la evaluación realizada de 12 años de zafra, se exponen los bajos niveles de eficiencia de indicadores industriales y el

decrecimiento promedio anual de 3,5 % del rendimiento industrial, de 5,3 % de la producción azucarera y la alta volatilidad del rendimiento industrial (Cala, Pacheco, & Sánchez, 2020).

El aumento del volumen y la variedad de información que se encuentra informatizada en bases de datos digitales y otras fuentes, ha crecido espectacularmente en las últimas décadas. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. Aparte de su función de «memoria de la organización», la información histórica es útil para explicar el pasado, entender el presente y predecir la información futura. Este es el principal cometido de la minería de datos: resolver problemas analizando los datos que se encuentran en las bases de datos (Hernández, Ramárez, & Ferri, 2004). El objetivo es crear valor y este radica, en gran medida, en la capacidad analítica. La cuestión es ser capaces de preguntar a los datos de tal manera que la información recopilada pueda dar las respuestas necesarias para entender lo que sucede, por qué sucede y hasta lo que pudiera suceder. Estas variantes son en realidad lo que conocemos como análisis descriptivo, predictivo y prescriptivo (Prometeus GS-Editor Team, 2019).

En la industria azucarera cubana existe una base de datos amplia, que necesita ser utilizada de forma eficaz para guiar el desarrollo productivo hacia escenarios más rentables. El empleo correcto de esta información ayudaría a tomar decisiones con bases objetivas. El sector azucarero cubano requiere implementar métodos que permitan cuantificar con mayor precisión la influencia de las variables tecnológicas del proceso, sobre el rendimiento industrial. Se necesita prever el comportamiento de su proceso productivo, para planificar y optimizar el uso de los recursos técnicos, humanos y financieros, y así mejorar aquellas variables tecnológicas que tienen mayor peso sobre el rendimiento industrial (Ribas, Consuegra, & Alfonso, 2016).

Por este motivo, el objetivo de nuestro estudio fue descubrir los patrones que permitan identificar las causas de los bajos rendimientos industriales en el proceso de fabricación del azúcar de caña, a partir de la aplicación de técnicas predictivas de minería de datos a los datos históricos de la industria azucarera cubana (2010-2019).

METODOLOGÍA

Entre los materiales empleados están las bases de datos de los históricos de zafra, las cuales contienen gran cantidad de registros transaccionales. La cantidad de registros es como promedio por año de más de 4 millones, según se detalla en la figura 1.

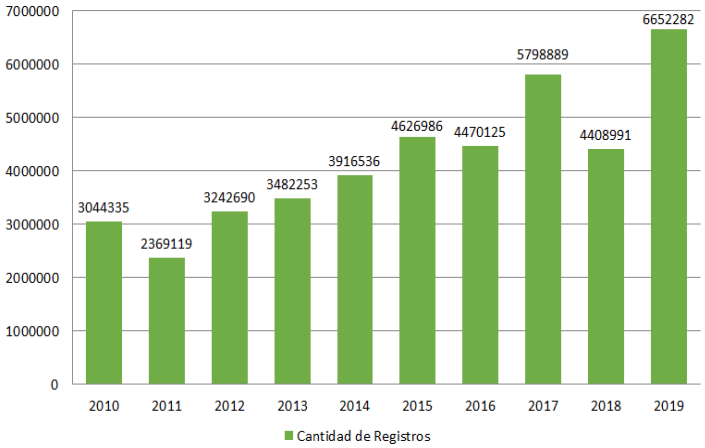

Fig. 1 Cantidad de registros por años.

La cantidad de registros transaccionales capturados representan el 39,68 %, mientras que los calculados son 60,32 %. El número medio de indicadores que se gestionan en cada central es de 3 605 como promedio, pero solo quedan registrados en los históricos a nivel nacional una media de 578, como se detalla en la figura 2.

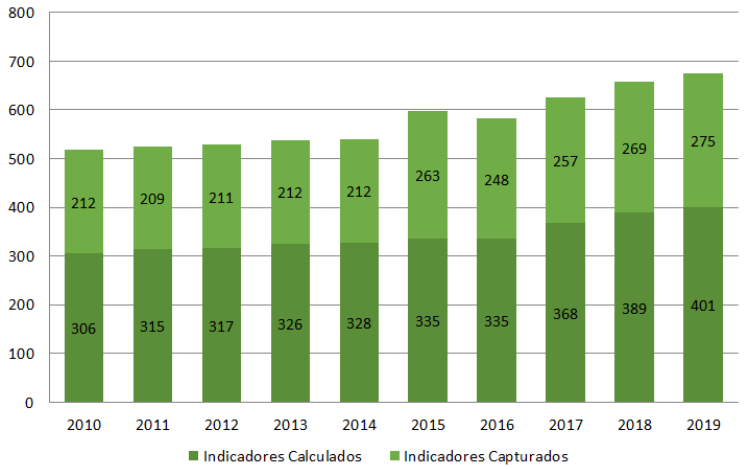

Fig. 2 Cantidad de indicadores por años.

La metodología seleccionada para establecer un marco de trabajo del ciclo de vida del proceso de minería de datos fue CRISP-DM (Cross Industry Standard Process for Data Mining), ya que mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. La herramienta seleccionada para aplicar las técnicas de minería de datos fue KNIME, plataforma de análisis que proporciona una interfaz gráfica fácil de usar y permite la integración, el procesamiento, el análisis y la exploración de datos.

El proceso de descubrimiento de conocimiento en base de datos (KDD) se desarrolló sobre un conjunto inicial de datos formado por seis atributos y 42 012 206 de instancias. Luego del proceso de selección, limpieza y transformación de datos, se obtuvo una vista minable, compuesta por 59 387 instancias y dos conjuntos de atributos. En este proceso se excluyeron del análisis los indicadores: RPC, rendimiento guía, rendimiento a reportar, norma potencial, kilogramos de azúcar, caña/t azúcar, azúcar propia física, azúcar física hecha, aprovechamiento rpc %, % de PH en norma, entre otros atributos que tienen una influencia directa conocida sobre el rendimiento. Para entrenar y probar los modelos en los experimentos se emplearon conjuntos de datos distintos, a fin de no sobrestimar su precisión. Se realizó una partición de los 59 387 registros donde se utilizan 41 570 (70 %) de los datos para el conjunto de entrenamiento y 17 817 (30 %) para el conjunto de prueba.

Se efectuó un análisis predictivo de los datos, en el cual se emplearon los métodos simbólicos, de la familia de los métodos de predicción, ya que cuentan con el conocimiento y producen modelos más interpretables para los humanos (García, Luengo, & Herrera, 2015). Se utilizaron siete algoritmos de aprendizaje automático para entrenar el modelo; de ellos, seis son algoritmos de aprendizaje de reglas, útiles y muy conocidos en el ámbito del aprendizaje automático, ya que son capaces de crear modelos interpretables (Mesa, 2019), así como un algoritmo de árbol de decisión, estrechamente relacionado con los algoritmos de aprendizaje (García et al., 2015). Se planificaron dos tareas en el modelado, para cumplir con el objetivo planteado, según se detalla en la tabla 1.

Tabla 1. Tareas planificadas para el modelado

N°	Tarea	Algoritmo	Objetivos
1	Selección de atributos	Selección hacia adelante (J48, CONJUNCTIVERULE, PART, JRIP, DECISIONTABLE, RIDOR, FURIA)	Seleccionar los atributos que más influyen en el bajo rendimiento industrial. Seleccionar el algoritmo que mejor resultado reporta para aplicar la clasificación
2	Clasificación	Árbol de decisión (C4.5)	Seleccionar los patrones que influyen en el bajo rendimiento industrial.

La selección de atributos permite automatizar la búsqueda de los subconjuntos de atributos más apropiados, para explicar un atributo objetivo (Lemarie, 2021). Se empleó el método de Selección hacia adelante y se implementó un ciclo por cada algoritmo. La comparación entre las métricas de la matriz de confusión de estos algoritmos determinó la selección del algoritmo para realizar la clasificación. Se llevó a cabo un análisis con la frecuencia de aparición de los atributos en cada uno de los algoritmos, lo que permitió conocer aquellos que predominan. Se calculó la curva ROC, se determinó el área bajo la curva como indicador de la calidad del modelo y se evaluaron los atributos por la escala de Swets. Según Mauricio, Rodríguez, & Muñoz (2022), el desempeño del modelo se clasifica en Fallido ($0,5 \leq AUC < 0,6$), Pobre ($0,6 \leq AUC < 0,7$), Razonable ($0,7 \leq AUC < 0,8$), Bueno ($0,8 \leq AUC < 0,9$) y Excelente ($0,9 \leq AUC \leq 1,0$). Se obtienen los atributos cuya área bajo la curva es superior a 0,5.

Con la tarea de clasificación, gracias a un conjunto de ejemplos ya clasificados, se pretende construir un modelo que permita clasificar nuevos casos (García, 2017). Se realizaron dos experimentos por cada conjunto de datos:

1. Se utilizaron todos los atributos de cada conjunto de datos.

2. Se filtraron los atributos de cada conjunto de datos, de acuerdo con los atributos seleccionados en la tarea de selección de características. Se realizó un análisis con la frecuencia de aparición de los atributos en cada experimento, lo que permitió conocer aquellos atributos más representativos.

Se validaron los patrones que influyen en los bajos rendimientos industriales, que se obtuvieron en la investigación y se realizó una validación cruzada por cada experimento, analizando el comportamiento de los patrones propuestos en 10 iteraciones con el total del conjunto de datos.

RESULTADOS Y DISCUSIÓN

De los experimentos realizados en la selección de características para el conjunto de datos 1, el algoritmo RIDDOR presentó los mejores resultados en las métricas Precision (0,866) y Specificity (0,858) y el CONJUNCTIVERULE, la mejor Sensitivity (0,920), mientras que el J48 presentó la mejor F-measure (0,823), Acurracy (0,799), Cohen's kappa (0,590) y el menor Error (0,252). El algoritmo que mayor cantidad de iteraciones realizó fue el J48 con 99 iteraciones; el de menor cantidad fue CONJUNCTIVERULE con 59 iteraciones. El algoritmo que mayor cantidad de atributos seleccionó fue el CONJUNCTIVERULE con 44 atributos, mientras que los que menor cantidad seleccionaron fueron el DECISIONTABLE y PART con 30 atributos. Los atributos que más predominan, según su frecuencia, son dos que están presentes en el 100 % de los algoritmos, así como ocho atributos presentes en seis (85,71 %) algoritmos. La mayor cantidad de atributos seleccionados están presentes en cuatro algoritmos (57,14 %)

y la menor cantidad se encuentra en un algoritmo que representa 14,29 %. La curva ROC para este experimento se realizó con los 59 387 registros y 63 atributos. Se obtuvieron 49 atributos, cuya área bajo la curva es superior a 0,5.

Para el conjunto de datos 2, el algoritmo J48 presentó los mejores resultados en todas las métricas: Precision (0,997), Sensitivity (0,997), Specificity (0,996), F-measure (0,997), Acurracy (0,996), Cohen's kappa (0,993) y el menor Error (0,004). El algoritmo que mayor cantidad de iteraciones realizó fue el J48 con 105 iteraciones, mientras que los que menor cantidad realizó fue PART con 59 iteraciones. El algoritmo que mayor cantidad de atributos seleccionó fue el JRIP con 56 atributos, mientras que los que menor cantidad seleccionó fue el DECISIONTABLE con 40 atributos. Los atributos que más predominan según su frecuencia son dos que están en el 100 % de los algoritmos, así como tres en seis (85,71%) algoritmos. La mayor cantidad de atributos seleccionados se encuentran en tres algoritmos (42,86 %); la menor cantidad está presente en siete algoritmos que representan el 100 %. La curva ROC para este experimento se realizó con los 59 387 registros y 100 atributos. Se obtuvieron 73 atributos, cuya área bajo la curva es superior a 0,5.

Un análisis integral de ambos resultados permite recomendar los atributos que se deben tener en cuenta para la confección del modelo. El análisis estadístico entre estos algoritmos demostró que el J48 mostró mejores métricas para realizar la clasificación.

De los experimentos realizados en la clasificación, para el conjunto de datos 1 el experimento 1 obtuvo los mejores resultados: 539 patrones, de los cuales 301 (55,84 %) representan el bajo rendimiento y 238 (44,16 %), el no bajo rendimiento. De los 17 817 registros del conjunto de prueba, 13 594 son correctos y 4 223 incorrectos. El experimento muestra una exactitud de 76,3 % y un error en la predicción de 23,7 %.

Para el conjunto de datos 2, el experimento 2 presentó los mejores resultados: 56 patrones, de los cuales 28 (50,00 %) representan el bajo rendimiento y 28 (50,00 %), el no bajo rendimiento. De los 17 817 registros del conjunto de prueba, 17 752 son correctos y 65 incorrectos. El experimento presenta una exactitud de 99,6 % y un error en la predicción de 0,365 %.

Finalmente, se analizaron los patrones obtenidos y de ellos se identificaron 32 atributos más representativos, que influyen en los bajos rendimientos industriales, reflejo o alerta de posibles situaciones existentes. Entre ellos se destacan:

- Metros cúbicos de agua gastados: su consumo puede estar relacionado con paradas de cierta duración, lo que obliga a limpiar con agua los equipos, reponer agua en el enfriadero por contaminaciones de azúcar, mal manejo de los retornos, problemas de contaminación de las aguas de rechazo, arrastres de azúcar a los condensadores de los tachos o de los evaporadores.

- Temperatura agua imbibición: la temperatura debe garantizar la mayor extracción de sacarosa del bagazo, sin que se funda la cera contenida en este; si no se extrae menos azúcar de la caña, sube la pérdida en bagazo y baja el rendimiento.

- Agua imbibición t: a más agua más lavado del bagazo, más sacarosa que pasa en los jugos a la fábrica y menos pérdida en bagazo.

- Molienda de diferentes cepas de caña (T Fríos, % Fríos, T caña madurez tardía, % retoños quedados, T retoños quedados, % primaveras quedadas, T Primaveras quedadas, T Socas, % Socas, % Total quedadas, T Retoños, T caña 12 meses o más, % Retoños): define la influencia de una cepa en el rendimiento, es su desfase, es

decir, su molienda en momentos no apropiados comienza a perder rendimiento y a formar sustancias indeseables que afectan el buen desarrollo del proceso y la calidad del azúcar.

- Porcentaje de caña en programación: valores bajos indica desorden en la cosecha, es decir, que no se escogen las cañas más maduras para la molienda.
- Caña atrasada t: determina que la caña se muele con alto nivel de deterioro y que se pierde una cantidad considerable de rendimiento.
- Caña con madurador: el madurador se aplica a las cañas con las que se inicia la zafra, acelerando su maduración o en otros momentos de la zafra.
- Caña tiro directo t: se supone que la caña de tiro directo se muela con mayor frescura que la caña que se almacena sobre carros, pero en Cuba, últimamente, la caña de tiro directo permanece largas horas cortada en los camiones y sufre atrasos después de cortada.
- Masa cocida A Brix %: el bajo Brix puede ser síntoma de que la materia prima procesada sea mala y los materiales sean viscosos, lo que obliga a dar bajo Brix para que el azúcar salga con calidad.
- Jugo clarificado Brix %: cuando es bajo hay que evaporar más para obtener meladura concentrada, solo la sacarosa que se puede descomponer al evaporar los jugos.
- Meladura Brix %: un bajo Brix obliga a cocinar durante más tiempo en los tachos, se pierde sacarosa por esa causa.
- Miel B extraída Brix %: el Brix de la miel depende del Brix de la masa cocida B y de la aplicación de agua en las centrífugas para obtener más calidad en el azúcar. En este último caso, se sube la pureza de la miel B y baja el rendimiento.
- Derrames t Pol: afecta directamente al rendimiento porque es producto azucarado que se pierde. Se contabilizan como pérdidas indeterminadas.
- Pérdida miel final % Pol caña: es la pérdida más importante en el proceso azucarero y expresa qué porcentaje de la Pol que trajo la caña se perdió con la miel final.
- Pérdida bagazo % Pol caña (R379): es la pérdida de Pol de la que entró con la caña, que se va en el bagazo.
- Jugo última extracción Pol % (i63d): a menor pureza o contenido de Pol de ese jugo, menor pérdida en bagazo y más rendimiento.
- Pérdida en molienda: es una expresión de la Pol que se pierde en el bagazo por cada parte de fibra que trae la caña.
- Ácido clorhídrico kg: se usa para limpiar la superficie calórica principalmente de los evaporadores. Cuando se producen paradas frecuentes hay que limpiarlos con mayor reiteración.
- Floculante kg: se utiliza en la etapa de clarificación, con el fin de disminuir el tiempo de residencia del jugo en los clarificadores. Se puede gastar más floculante por moler materia prima de mala calidad.

- Cal kg: se utiliza para precipitar las impurezas y obtener sacarosa purificada. El consumo puede estar asociado a inestabilidad en la molida, mala calidad de la caña.
- pH jugo filtros: un bajo pH indica que la cachaza estuvo retenida mucho tiempo en el clarificador o en el cachazón, puede indicar descomposición de la sacarosa.
- pH agua enfriadero: es un elemento de control para evaluar fugas de azúcar hacia el enfriadero, no cuantifica la pérdida pero alerta e influye en la búsqueda y detección de la causa.
- Cachaza agotada t: mientras más cachaza se produce, más azúcar se pierde en ella. El exceso de materias extrañas puede incrementar la cantidad de cachaza que se produce.
- Porcentaje extracción primer molino: influye directamente en la eficiencia del tándem. Es elemento que se debe considerar, porque al aumentar las pérdidas en bagazo baja el rendimiento.
- Porcentaje extracción último molino: determina la pérdida de azúcar o Pol, que al no pasar a los jugos se queda en el bagazo y se quema en los hornos.
- Miel B extraída t: es una pérdida considerable que afecta el rendimiento.
- Pureza meladura: refleja la calidad de la caña molida, además puede denotar un buen trabajo en fábrica.
- Color ICUMSA azúcar alta calidad a granel: al lavar más, suben las purezas de las mieles incluyendo la miel final, que influye directamente en el rendimiento.
- Azúcar alta calidad en sacos humedad %: cuando se procesan productos viscosos, provenientes de cañas atrasadas y/o deterioradas este parámetro puede alterarse.
- Jugo desmenuzadora cenizas %: las cenizas son sustancias melasigénicas, es decir, que producen miel y por tanto contribuyen a que el rendimiento sea menor.
- Caída pureza masa B: la pureza de la masa cocida B depende de la calidad de la caña, a mayor pureza en la miel B más pérdidas y menos rendimiento.
- Horas remoción sin extracciones: representa el tiempo que transcurre desde que el azúcar se extrae en los molinos hasta que sale producida como tal. Los materiales en el proceso pueden deteriorarse en el tiempo que están almacenados, depende mucho de la molida.

Al validar los patrones que influyen en los bajos rendimientos industriales, que se han obtenido como resultado de esta investigación, se realizó una validación cruzada por cada experimento, analizando el comportamiento de los patrones propuestos en 10 iteraciones con el total del conjunto de datos. Se validaron los conjuntos de prueba por cada experimento, de los cuales el error promedio fue de 3,226 %.

Similares a esta investigación se han realizado estudios que utilizan la selección de características para la reducción de la dimensionalidad (Casillas, Cordón, Del Jesús, & Herrera, 2001; Li & Wu, 2008). En Huang et al. (2020), los autores proponen un algoritmo de selección de características basado en reglas de asociación y trabajos que utilizan los árboles de decisión para la reducción de la dimensionalidad (Akhiat, Asnaoui, Chahhou, & Zinedine, 2020; Akhiat, Manzali, Chahhou, & Zinedine, 2021; Topouzelis & Psyllos, 2012; Zhou, Si, & Fujita, 2017), como los seleccionados en este trabajo. También se identificaron otros que utilizan árboles de decisión para resolver

problemas similares de clasificación (Peloia, Bocca, & Rodrigues, 2019; Sasikanth, Krishnam Raju, Naveen Kumar, & Kurumalla, 2019; Siqueira, Rodrigues, Bocca, & Oliveira, 2017; Vieira Ribeiro, Antunes Rodrigues, & Pires Gravina de Oliveira, 2017).

Al igual que los hallazgos de nuestro estudio, en trabajos consultados, como Ribas et al. (2016), los autores realizan un análisis de los factores que más inciden sobre el rendimiento industrial. Es de resaltar que las variables escogidas por los autores para realizar este análisis, también son identificadas en primer lugar por la fase de preparación de los datos, así como en la fase de modelado en las tareas de selección de atributos y clasificación de esta investigación. Se identifican algunos estudios que demuestran la influencia de las variables identificadas en el rendimiento industrial. En un estudio de González, Castellanos, & Puertas (2010), se señala que el esquema de imbibición, la cantidad y temperatura del agua utilizada, así como la composición de la caña utilizada influyen en las perdidas de azúcar en el bagazo. Guerra (2019) plantea que el manejo adecuado en tiempo y en la forma adecuada de la caña, y el trabajo oportuno de los retoños garantizan mantener un buen rendimiento. Según Matute, Bedoya, & Feo (2012), la concentración del floculante en la clarificación del jugo de caña de azúcar es una parte esencial en el proceso de fabricación del azúcar refinado, debido a que afecta el rendimiento y la calidad del producto final.

Por tanto, los modelos pueden considerarse aceptados. Desde el punto de vista analítico ayudará a apoyar la toma de decisiones con bases objetivas, brindándoles un análisis matemático a partir de un gran conjunto de datos sobre las causas que más inciden en el bajo rendimiento industrial.

Se realizó un proceso de minería de datos con métodos predictivos de los datos históricos de la zafra azucarera, que permitió obtener las causas que influyen en los bajos rendimientos y, a su vez, posibilitó a partir de los modelos obtenidos vaticinar lo que va a ocurrir con antelación.

Es importante señalar que para futuros trabajos sería muy oportuno aplicar la analítica prescriptiva a los datos históricos de la zafra azucarera. Además, se recomienda realizar estos análisis a otros aspectos del proceso de fabricación del azúcar de caña, como la calidad de la materia prima, el rendimiento del campo, variedades y maduración de la caña, la productividad de los equipos de corte, alza y tiro, el tiempo de inicio y fin de la zafra.

CONCLUSIONES

- Enfocados en el objetivo del negocio, fueron planteadas dos tareas de minería, relacionadas con la selección de características y la clasificación. Se desarrollaron los experimentos necesarios en cada caso y como resultado se obtuvo que resultan factibles para la identificación de los atributos que más influyen en el bajo rendimiento industrial.
- Con el propósito de validar los atributos y patrones seleccionados, se realizó un análisis comparativo de estos, considerándose los modelos como aceptados.
- Con la identificación de estos atributos se lograron crear las bases para un análisis más profundo de las medidas organizativas y de control necesarias, para dejar de perder azúcar en el proceso industrial.
- La realización de análisis prescriptivo a los datos de los históricos de la zafra azucarera permitirá ir más allá de las predicciones y recomendar el mejor plan de acción en el futuro.

REFERENCIAS

Akhiat, Y., Asnaoui, Y., Chahhou, M., & Zinedine, A. (2020). A new graph feature selection approach. 2020 6th IEEE Congress on Information Science and Technology (CiSt), 156-161. Agadi Essaouira, Morocco: IEEE. https://doi.org/10.1109/CiSt49399.2021.9357067

Akhiat, Y., Manzali, Y., Chahhou, M., & Zinedine, A. (2021). A New Noisy Random Forest Based Method for Feature Selection. Cybernetics and Information Technologies, 21(2), 10-28. https://doi.org/10.2478/cait-2021-0016

Cala Jústiz, Y., Pacheco Feria, U., & Sánchez Jiménez, M. (2020). Análisis de indicadores de eficiencia productiva y perspectivas de la industria azucarera en Santiago de Cuba. Anuario Facultad de Ciencias Económicas y Empresariales, 91-106.

Casillas, J., Cordón, O., Del Jesus, M. J., & Herrera, F. (2001). Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems. Information Sciences, 136(1), 135-157. https://doi.org/10.1016/S0020-0255(01)00147-5

Concepción Cruz, E., Caraballoso Torrecilla, V., Nápoles Alberto, R. G., Morales Fundora, L., Cruz Coca, O., & Viñas Quintero, Y. (2015). Problemas asociados al rendimiento agrícola de la caña de azúcar en la Cooperativa Potrerillo, provincia Sancti Spíritus. Centro Azúcar, 42(2), 83-92.

García Fernández, J. (2017). Modelos híbridos de aprendizaje basados en instancias y reglas para Clasificación Monotónica (Tesis de Doctorado, Jaén: Universidad de Jaén). Jaén: Universidad de Jaén. Recuperado de http://ruja.ujaen.es/jspui/handle/10953/864

García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

González Pérez, F., Castellanos Álvarez, J. A., & Puertas Fernández, J. F. (2010). Método para determinar la cantidad de agua de imbibición a utilizar en la industria de azúcar de caña. Ingeniería Mecánica, 13(1), 41-48.

Guerra González, J. D. (2019). La estructuración de las cepas y los cultivares de caña de azúcar en la Cooperativa de Producción Agropecuaria 10 de Octubre. (Thesis, Universidad de Matanzas. Facultad de Ciencias Agropecuarias). Universidad de Matanzas. Facultad de Ciencias Agropecuarias. Recuperado de http://rein.umcc.cu/handle/123456789/829

Hernández Orallo, J., Ramárez Quintana, M. J., & Ferri Ramírez, C. (2004). Introducción a la Minería de Datos. España: Pearson Educacion. S.A.

Huang, C., Huang, X., Fang, Y., Xu, J., Qu, Y., Zhai, P., … Li, J. (2020). Sample imbalance disease classification model based on association rule feature selection. Pattern Recognition Letters, 133, 280-286. https://doi.org/10.1016/j.patrec.2020.03.016

Lemarie, F. (2021). Capítulo 1. Técnicas de Análisis de Datos en WEKA. Recuperado de https://www.academia.edu/61030769/Cap%C3%ADtulo_1_T%C3%A9cnicas_de_An%C3%A1lisis_de_Datos_en_WEKA_CAP%C3%8DTULO_1_T%C3%89CNICAS_DE_AN%C3%81LISIS_DE_DATOS_EN_WEKA

Li, Y., & Wu, Z.-F. (2008). Fuzzy feature selection based on min–max learning rule and extension matrix. Pattern Recognition, 41(1), 217-226. https://doi.org/10.1016/j.patcog.2007.06.007

Matute, L., Bedoya, C., & Feo, J. (2012). Determinación de la concentración óptima de floculante a usar en la clarificación de jugos de caña en un central azucarero. Revista de la Facultad de Agronomía, 38(3). Recuperado de http://saber.ucv.ve/ojs/index.php/rev_agro/article/view/5903

Mauricio Munar, A., Rodríguez Carlosama, A., & Muñoz España, J. L. (2022). Potenciales áreas cultivables de pasifloras en una región tropical considerando escenarios de cambio climático. Revista de Investigación Agraria y Ambiental (RIAA), 3(1). Recuperado de http://portal.amelica.org/ameli/journal/130/1302674008/html/

Mesa Pérez, F. (2019). Estudio y análisis del funcionamiento de técnicas de minería de datos en conjuntos de datos relacionados con la Biología (Tesis de Grado, Universidad de Jaén). Universidad de Jaén, España. Recuperado de http://tauja.ujaen.es/jspui/handle/10953.1/10372

Peloia, P. R., Bocca, F. F., & Rodrigues, L. H. A. (2019). Identification of patterns for increasing production with decision trees in sugarcane mill data. Scientia Agricola, 76(4), 281-289. https://doi.org/10.1590/1678-992x-2017-0239

Prometeus GS-Editor Team. (2019, febrero 21). Análisis de datos predictivo, descriptivo y prescriptivo ¿En qué consisten? Recuperado 26 de febrero de 2023, de Prometeus Global Solutions website: https://prometeusgs.com/analisis-de-datos-diferencias/

Ribas García, M., Consuegra del Rey, R., & Alfonso Alfonso, M. (2016). Análisis de los factores que más inciden sobre el rendimiento industrial azucarero. Centro Azúcar, 43(1), 51-61.

Sasikanth, T., Krishnam Raju, M., Naveen Kumar, E., & Kurumalla, S. (2019). Prediction of crop yield using data mining techniques. IJESRT, 8(3), 6.

Siqueira, T., Rodrigues, L. H., Bocca, F., & Oliveira, M. (2017, octubre 21). Decision trees for knowledge discovery on the yield decline of sugarcane ratoons. https://doi.org/10.19146/pibic-2017-78279

Topouzelis, K., & Psyllos, A. (2012). Oil spill feature selection and classification using decision tree forest on SAR image data. ISPRS Journal of Photogrammetry and Remote Sensing, 68, 135-143. https://doi.org/10.1016/j.isprsjprs.2012.01.005

Vieira Ribeiro, N., Antunes Rodrigues, L. H., & Pires Gravina de Oliveira, M. (2017). Development of predictive models using Data Mining techniques to detect borer infestation (Diatraea saccharalis) in sugarcane culture | Galoá Proceedings. Presentado en XXV Congresso de Iniciação Científica da Unicamp, Brasil. Brasil. Recuperado de https://proceedings.science/unicamp-pibic/pibic-2017/papers/development-of-predictive-models-using-data-mining-techniques-to-detect-borer-infestation--diatraea-saccharalis--in-suga#

Zhou, L., Si, Y.-W., & Fujita, H. (2017). Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method. Knowledge-Based Systems, 128, 93-101. https://doi.org/10.1016/j.knosys.2017.05.003