Análisis comparativo entre algoritmos de aprendizaje de reglas para identificar indicadores que influyen en el bajo rendimiento industrial

Comparative Analysis between Rule Learning Algorithms to Identify Indicators that Influence Low Industrial Yield

Gil Rodríguez, Yohan; Socorro Llanes, Raisa; Rosete Suárez, Alejandro; Bravo Ilisástigui, Lisandra

Yohan Gil Rodríguez
yohan.gil@datazucar.cu
DATAZUCAR, GRUPO AZUCARERO AZCUBA, Cuba

Raisa Socorro Llanes
raisa@ceis.cujae.edu.cu
UNIVERSIDAD TECNOLÓGICA DE LA HABANA JOSÉ ANTONIO HECHEVERRÍA (CUJAE), Cuba

Alejandro Rosete Suárez
rosete@ceis.cujae.edu.cu
UNIVERSIDAD TECNOLÓGICA DE LA HABANA JOSÉ ANTONIO HECHEVERRÍA (CUJAE), Cuba

Lisandra Bravo Ilisástigui
lbravo@ceis.cujae.edu.cu
UNIVERSIDAD TECNOLÓGICA DE LA HABANA JOSÉ ANTONIO HECHEVERRÍA (CUJAE), Cuba

Revista Cubana de Transformación Digital
Unión de Informáticos de Cuba, Cuba
ISSN-e: 2708-3411
Periodicidad: Trimestral
vol. 3, núm. 1, 2022
rctd@uic.cu

Recepción: 05 Marzo 2022
Aprobación: 16 Mayo 2022

URL: http://portal.amelica.org/ameli/journal/389/3893118006/

Esta obra está bajo una Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional.

Resumen: La informatización de los procesos de la industria azucarera genera cuantio- sos datos. En la actualidad la aplicación de los programas de la Plataforma Agro-Industrial existente en el Grupo Azucarero Azcuba, ha garantizado la rapidez y calidad de las informaciones de zafra y los beneficios que de ello se derivan. La industria azucarera cubana requiere implementar herramientas y métodos científicos que permitan analizar y cuantificar con mayor precisión la influencia de las variables tecnológicas del proceso industrial en la eficiencia de la fabricación del azúcar de caña. Por eso, es necesario descubrir cuáles son las causas principales que están incidiendo en los bajos rendimientos industriales en el proceso de fabricación del azúcar de caña en Cuba a partir de los datos históricos de la zafra azucarera. Se utiliza la metodología CRISP-DM para el modelado del proceso de minería de datos. Se realiza como punto de partida para análisis posteriores más profundos una comparación entre algoritmos de
aprendizajes de reglas, donde se obtienen patrones que influyen en los bajos rendimientos industriales.

Palabras clave: Minería de datos, CRISP-DM, Rendimiento Industrial, Aprendizaje de Reglas.

Abstract: *The computerization of the processes of the sugar industry generates abundant data. At present, the application of the programs of the existing Agro-Indus- trial Platform in Azcuba has guaranteed the speed and quality of harvest in- formation and the benefits derived from it. The cuban sugar industry needs to implement scientific tools and methods that allow the influence of the techno- logical variables of the industrial process on the efficiency of cane sugar manu- facturing to be analyzed and quantified with greater precision. For this reason, it is necessary to discover what are the main causes that are influencing the low industrial yields in the cane sugar manufacturing process in cuba based on the historical data of the sugar harvest. The CRISP-DM methodology is used for modeling the data mining process. As a starting point for deeper analysis, a comparison between rule learning algorithms is made, where patterns that influence low industrial yields are obtained.*

Keywords: Data Mining, CRISP-DM, Industrial Yield, Rule Learning..

INTRODUCCIÓN

Cuba posee una rica tradición de más de cuatro siglos en la producción de azúcar de caña, la que ocupa el mayor uso de la tierra cultivable del país, por lo que constituye una de las fuentes principales de alimentación para el hombre, además del amplio uso que tienen los productos derivados a partir de procesos industriales de este cultivo (Concep- ción Cruz et al., 2015).

En la industria azucarera cubana existe una base de datos amplia que necesita ser utilizada

en forma eficaz para guiar el desarrollo productivo hacia escenarios más rentables. La utili- zación correcta de esta información ayudaría a la toma de 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 indus- trial. Se necesita prever el comportamiento de su proceso productivo con el fin de planificar y optimizar el uso de los recursos técnicos, humanos y financieros para mejorar aquellas varia-

bles tecnológicas que tienen mayor peso sobre el rendimiento industrial (Ribas García, Con- suegra del Rey, y Alfonso Alfonso, 2016).

La informatización de los procesos de la industria azucarera genera cuantiosos datos. En la actualidad la aplicación de los programas de la Plataforma Agro-Industrial existente en Azcuba, ha garantizado la rapidez y calidad de las informaciones de zafra y los beneficios que de ello se derivan. La plataforma está integrada por varios sistemas, entre ellos el sistema *IPlus* que es el sistema informativo de zafra del Grupo Azcuba que posibilita la conexión de los resultados ope- rativos de lo conducción del proceso agroindustrial. Se visualiza a diferentes niveles de dirección y contiene números módulos tanto para la empresa, provincia y nación, fundiendo como un todo único la información industrial de la zafra («Iplus – Datazucar», s. f.).

Se conoce la influencia que tienen algunas variables tecnológicas en el rendimiento in- dustrial, ya sea por conocimiento empírico o por investigaciones científicas, como la de Ribas García et al., 2016, que en su investigación sólo analiza valores anuales de 39 variables tecno- lógicas en tres años de zafra.

En la actualidad se necesita conocer, partiendo del comportamiento histórico del proceso productivo, relaciones interesantes entre las variables tecnológicas que tienen mayor peso en el bajo rendimiento industrial. Donde a partir del análisis de la información histórica diaria de más de 500 indicadores en 10 años de zafra, se identificarán reglas sólidas, desconocidas o como confirmación de las relaciones utilizadas actualmente.

El objetivo del presente trabajo es realizar un análisis comparativo de diferentes algorit- mos de aprendizaje de reglas que permitan realizar un preprocesamiento de los datos, para identificar los indicadores que influyen en la clasificación del rendimiento industrial en los históricos de diez años de la zafra azucarera cubana (2010-2017).

METODOLOGÍA

Materiales y Métodos

Metodología

El aumento del volumen y 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 histó- rica es útil para explicar el pasado, entender el presente y predecir la información futura. La mayoría de las decisiones de empresas, organizaciones e instituciones se basan también en información sobre experiencias pasadas extraídas de fuentes muy diversas. Además, ya que los datos pueden proceder de fuentes diversas y pertenecer a diferentes dominios, parece clara la inminente necesidad de analizar los mismos para la obtención de información útil para la organización.

En muchas situaciones, el método tradicional de convertir los datos en conocimiento consiste en un análisis e interpretación realizada de forma manual. Esta forma de actuar es lenta, cara y altamente subjetiva. De hecho, el análisis manual es impracticable en dominios donde el volumen de los datos crece exponencialmente: la enorme abundancia de datos desborda la capacidad humana de comprenderlos sin la ayuda de herramientas potentes. Consecuen- temente, muchas decisiones importantes se realizan, no sobre la base de la gran cantidad de datos disponibles, sino siguiendo la propia intuición del usuario al no disponer de las herra- mientas necesarias. Éste es el principal cometido de la minería de datos: resolver problemas analizando los datos presentes en las bases de datos (Hernández Orallo, Ramárez Quintana, & Ferri Ramírez, 2004).

La minería de datos es un sistema de información basado en computación que explora grandes repositorios de datos para generar información y descubrir conocimiento. La palabra se origina en la minería tradicional; sin embargo, el objetivo es buscar conocimiento que per- mita descubrir elementos como: patrones interesantes, relación entre los datos, definir reglas, predecir valores desconocidos, agrupar objetos homogéneos y otros aspectos que son difíciles de descubrir en sistemas de información tradicionales (Peña-Ayala, 2014).

La minería de datos trata, en términos generales, de resolver problemas mediante el aná- lisis de datos presentes en bases de datos reales. Hoy en día, está calificado como ciencia y tecnología para explorar datos para descubrir patrones desconocidos ya presentes. Se distin- gue la minería de datos como sinónimo del proceso de descubrimiento de conocimientos en bases de datos (KDD, del inglés *Knowledge Discovery in Databases*), mientras que otros ven a la minería de datos como el paso principal de KDD.

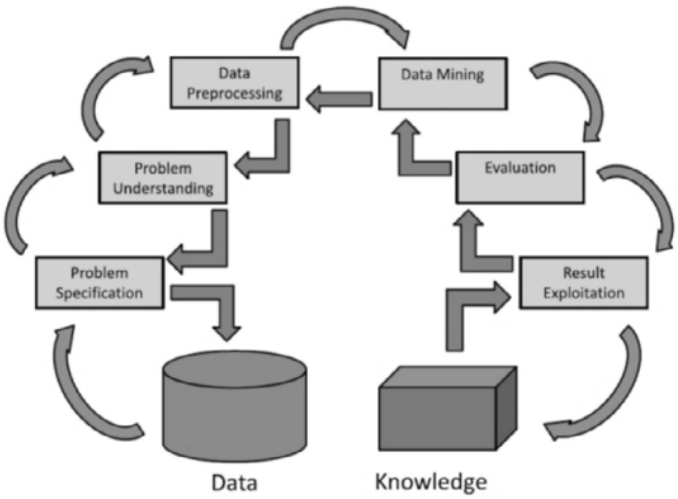

Figura 1.
Proceso de descubrimiento de conocimientos en bases de datos.

FIGURA 1.
Proceso de descubrimiento de conocimientos en bases de datos.
sf

La Figura 1 resume el proceso KDD y revela las seis etapas mencionadas previamente. Cabe mencionar que todas las etapas están interconectadas, mostrando que el proceso KDD es en realidad un esquema auto organizado donde cada etapa condiciona las etapas restantes y el camino inverso también está permitido (García, Luengo, y Herrera, 2015).

Son diversas las metodologías que han sido propuestos para el desarrollo de proyectos de Minería de Datos, según Montequín et al., s. f., las metodologías SEMMA (*Sample, Explore, Modify, Model, Assess*) y CRISP-DM (*Cross-Industry Standard Process for Data Mining*) com- parten la misma esencia, estructurando el proyecto de Minería de Dato en fases que se en- cuentran interrelacionadas entre sí, convirtiendo el proceso de Minería de Dato en un proceso iterativo e interactivo. La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una

perspecti- va más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de Minería de Datos donde la metodología SEMMA co- mienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema téc- nico. Desde ese punto de vista más global se puede considerar que la metodología CRISP-DM está más cercana al concepto real de proyecto, pudiendo ser integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas.

Un gran número de técnicas para minería de datos son bien conocidas y se utilizan en mu- chas aplicaciones. Según Wirth y Hipp (2000) los principales métodos de minería de datos se dividen teniendo en cuenta el método utilizado para la obtención del conocimiento en predic- tivos y descriptivos. A continuación, daremos una breve descripción del método Aprendizaje de Reglas, incluidas las referencias de algunos representantes y algoritmos concretos y consi- deraciones importantes desde el punto de vista de los datos de preprocesamiento.

Los métodos basados en reglas son útiles y muy conocidos en el ámbito del aprendizaje automático debido a que son capaces de crear modelos interpretables. La principal caracte- rística del método es utilizar reglas basadas en lenguaje natural, teniendo en cuenta el grado de complejidad. Las reglas pueden o no asociarse a cada categoría, de manera que validen, invaliden o incluyan la categoría en los resultados, siempre que cumplan con los requisitos de las reglas. También sirven para el reordenamiento de los resultados expresado en un lenguaje básico computacional. El reordenamiento viene dado por el algoritmo de aprendizaje (Pérez, s. f.).

El aprendizaje de reglas también llamado algoritmos de reglas de separación y conquista o de cobertura. Todos los métodos comparten la operación principal. Buscan una regla que expliquen parte de los datos, separe estos ejemplos y conquiste recursivamente el resto. Hay muchas formas de hacer esto, y también muchas formas de interpretar las reglas producidas y utilizarlas en el mecanismo de inferencia. Desde el punto de vista de preprocesamiento de datos, en general, requieren datos nominales o discretizados (aunque esta tarea suele estar implícita en el algoritmo) y disponer de un selector de atributos interesantes de los datos. Sin embargo, ejemplos ruidosos y valores atípicos pueden perjudicar el rendimiento del modelo final. Buenos ejemplos de estos modelos son los algoritmos AQ, CN2, RIPPER, PART y FU- RIA (Wirth y Hipp, 2000).

El aprendizaje de reglas inductivas es uno de los campos más tradicionales en el aprendizaje automático. Sin embargo, al reflexionar sobre su larga historia, se puede ar- gumentar que, si bien los métodos modernos son algo más escalables que los algoritmos tradicionales de aprendizaje de reglas, no se ha logrado ningún avance importante. De hecho, el algoritmo de aprendizaje de reglas RIPPER sigue siendo muy difícil de superar en términos de precisión y simplicidad de los conjuntos de reglas aprendidos (Beck y Fürnkranz, 2021).

Herramientas

Knime Analytics Platform es una plataforma de análisis, informes e integración de datos de código abierto desarrollada y respaldada por Knime.com AG. Mediante el uso de una interfaz gráfica, *Knime* permite a los usuarios crear flujos de datos, ejecutar pasos de análisis seleccio- nados y revisar los resultados, modelos y vistas interactivas.

Escrito en Java y construido sobre *Eclipse, Knime Analytics Platform* aprovecha la capaci- dad de extensión del módulo de *Eclipse* mediante el uso de complementos y conectores. Los complementos disponibles admiten la integración, con métodos para minería de texto, mine- ría de imágenes y análisis de series de tiempo.

Knime también integra varios otros proyectos de código abierto, incluidos los algoritmos de aprendizaje automático de *Weka, R* y *JF*reeChart. Admite envoltorios para llamar a otro código y proporciona nodos para que los usuarios puedan ejecutar *Java, Python, Perl* y otros fragmentos de código. La plataforma de análisis de *Knime* aprovecha la capacidad del comple- mento *Eclipse*; como resultado, existen más de 1,000 módulos

que admiten conectores para todos los formatos de archivo y bases de datos principales, así como una amplia gama de tipos

de datos, funciones estadísticas y algoritmos avanzados de aprendizaje automático y predicti- vo (Equipo Técnico de *Krypton Solid*, 2021).

Para la realización del proceso, y luego de un estudio exploratorio, se decidió emplear la metodología CRISP-DM y la herramienta de análisis de datos *Knime*.

RESULTADOS Y DISCUSIÓN

Comprensión del negocio

Actualmente, ante la gran cantidad de datos que son recogidos y almacenados en la base de datos del *IPlus*, las herramientas tradicionales de gestión de datos y las herramientas estadís- ticas no son adecuadas para extraer conocimiento útil, comprensible y previamente descono- cido, por lo que resulta necesario la aplicación de técnicas de minería de datos a los históricos de la zafra azucarera.

Comprensión de los datos

Existe multitud de registros y atributos para procesar en una aplicación de minería de da- tos. Las bases de datos presentan las siguientes dimensiones:

- • Cantidad de Registros: la cantidad de registros con la información contable es como promedio de más de 4 millones. La base de datos que menor cantidad de registros posee es la del año 2011 con 2 369 119 registros y la que más posee es la del 2019 con 6 652 282 registros (Figura 2).

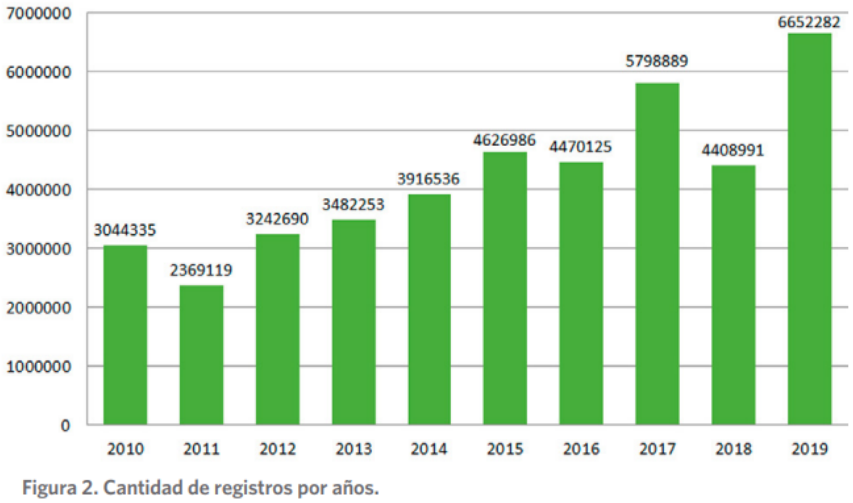

Figura 2. Cantidad de registros por años.

FIGURA 2.
Cantidad de registros por años.
sf

Cantidad de Indicadores: la cantidad de indicadores gestionados por el sistema es de 3605 como promedio, pero sólo 578 como promedio son almacenados en los históricos en cada base de datos. La base de datos que menor cantidad de indicadores posee es la del año 2010 con 518 indicadores y la que más posee es la del 2019 con 676 indicadores (Figura 3).

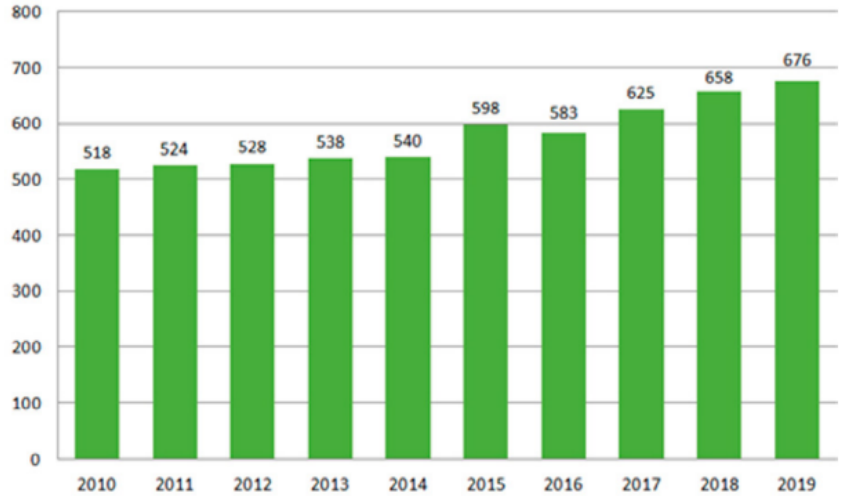

Figura 3. Cantidad de indicadores por años.

FIGURA 3
Cantidad de indicadores por años.
sf

Estos orígenes de datos almacenan los valores numéricos de los indicadores que describen el proceso industrial azucarero diariamente.

Se realiza una exploración inicial de los orígenes de datos disponibles, revelando informa- ción interesante acerca del comportamiento de los indicadores de las zafras azucareras en el país.

Se realiza una exploración de 578 indicadores como promedio en cada base de datos, que arroja los siguientes resultados:

- • Indicadores Capturados: indicadores que se capturan de forma manual en el sistema,
  valores de entrada con que el sistema IPlus realiza los cálculos del resto de los indica- dores. La cantidad de indicadores capturados por el sistema es de 236 como promedio, pero no se gestionan la misma cantidad en cada base de datos. La cantidad de indicado- res capturados aumenta con el paso de los años, esto incluye en el análisis y es necesario realizar un estudio más profundo para determinar las características de los mismos. La base de datos que menor cantidad de indicadores capturados posee es la del año 2011 con 209 indicadores y la que más posee es la del 2019 con 275 registros. La cantidad de registros transaccionales para los indicadores capturados representan como promedio el 39,68 % del total de registros.
  • Indicadores Calculados: indicadores que se calculan a partir de los datos de entrada y
  de las más de 10 000 fórmulas que maneja el sistema IPlus. La cantidad de indicadores calculados por el sistema es de 342 como promedio. La cantidad de indicadores calcu- lados aumenta con el paso de los años. La base de datos que menor cantidad de indica- dores capturados posee es la del año 2010 con 306 indicadores y la que más posee es la del 2019 con 401 registros. La cantidad de registros transaccionales para los indicadores calculados representan como promedio el 60,32 % del total de registros.

Se decide utilizar solo los indicadores capturados y no así los calculados, ya que los calcu- lados se conoce las relaciones o fórmulas que los generan, es necesario determinar la influen- cia de estos con el rendimiento.

Se realiza la verificación de la calidad de los datos durante el proceso de descripción y ex- ploración. Algunos de los problemas detectados son los siguientes:

- • Los indicadores aumentan con el paso de los años, lo que implica que los primeros alma-
  cenes de datos cuentan con menos indicadores que los últimos, esto no constituye una incoherencia en la codificación ya que se adaptan a las necesidades a medida del paso del tiempo, estos indicadores varían por añadidura o eliminación de un año al otro.
  • Existen valores de indicadores (*Valor_Diario*) en cero, pero no es un dato perdido, este
  puede interpretarse como valor real, ya que cuando no se ha molido caña en el día se arrastran los valores hasta fecha, el rendimiento y los indicadores capturados serán cero.

Preparación de los datos

Se recolectó información de las bases de datos de los históricos de la zafra azucarera, pro- porcionada por la Dirección de informática, comunicaciones y análisis del Grupo Azcuba. Se
realiza la selección de los atributos o características de interés para la investigación actual, en la base de datos donde se guarda toda la información referente los valores de los indicadores analizados en la zafra azucarera. Los atributos de Indust_Indicador_Diario son de gran utili- dad:

- • El atributo ID_Indicador es principal, ya que el mismo identifica el indicador analizado
  por medio de una relación con Iplus_Indicador.
  • El atributo ID_Entidad es principal, ya que el mismo identifica la entidad o central azu- carero analizado por medio de una relación con Iplus_Entidad.
  • El atributo Fecha_Carga_Datos es necesario ya que permite la ubicación de los valores en el tiempo, es necesario en el proceso de preparación de los datos a partir de este atri- buto construir nuevos como Mes y Año, para facilitar la ubicación de los valores en ran- gos de tiempo.
  • El atributo Valor_Diario es principal, ya que es en este atributo donde se guardan los
  valores necesarios para la investigación.
  • Los atributos como Valor_Semana y Valor_HF se pueden excluir, ya que son acumula- dos de Valor_Diario.
  • Se realiza una transformación al conjunto de datos originales con vistas a obtener en una sola fila los valores de los indicadores.
  » Datos de Indicadores: se guarda toda la información que describe a los indicadores.
  Los atributos de *Iplus_Indicador* pueden ser de gran utilidad.
  » Datos de Categoría: se guarda toda la información referente las entidades. Los atribu- tos de *Iplus_Categoria* pueden ser de gran utilidad para análisis posteriores.
  » Datos de Entidad: se guarda toda la información referente las entidades. Los atributos de *Iplus_Entidades* pueden ser de gran utilidad para análisis posteriores.

La construcción de nuevos datos se realiza por medio de sentencias SQL propias del ori- gen de datos. Entre las características agregadas se hallan:

- • A partir del atributo Id_Indicador y su valor contenido en el atributo Valor_Diario se
  genera un nuevo atributo con valores numéricos por cada Id_Indicador diferente. Este nuevo atributo se nombrará de acuerdo al valor del atributo Descripción por medio de la relación existente entre Indust_Indicador_Diario - Iplus_Indicador. Estos nuevos atri- butos se generan a partir del proceso transponer las filas en columnas.
  » Se realiza para todos los indicadores, generando atributos con la forma i10, i11, i325a,
  i42a, etc. Este proceso se realiza en cada origen de datos con el objetivo, así mismo,
  para realizar esta acción es necesario utilización del Nodo Pivoting – KNIME. Al transponer las filas en columnas se transforma los registros transaccionales en regis- tros minables.
  » A partir del valor Rendimiento a reportar, se genera el atributo Evaluación del Ren-
  dimiento (Eval_BajoRend) que puede tomar los siguientes valores ordinarios:

Bajo: para R295 < 10, asignándole el valor 1 NoBajo: para R295 >= 10, asignándole el valor 0

Para realizar esta acción es necesario la utilización de un Nodo *Math Formula – KNI- ME*, el cual por medio de la función if(x,y,z), permitirá asignar un valor numérico a los rangos y condiciones definidos anteriormente. Después es necesario aplicar un Nodo *Cell Replacer – KNIME* para sustituir los valores numéricos (0, 1) por los valores ordi- narios (Bajo, No Bajo) respectivamente.

Se dispone de varios orígenes de datos, correspondiente a uno por cada año de zafra azu- carera. Se utiliza el método básico de adición para integrar dos o más conjuntos de datos con atributos similares, pero con registros diferentes. Aplicando la herramienta *Knime* se diseña un flujo de trabajo donde se utiliza un Nodo *Concatenate - KNIME*, para integrar los diferen- tes orígenes de datos previamente obtenidos por medio de consulta SQL. Una vez realizado

este proceso, se realizan las acciones para agregar nuevos datos y se obtiene las vistas mina- bles por cada conjunto de datos, que son salvadas para su posterior uso en un fichero .CSV, como se ilustra en la Figura 4.

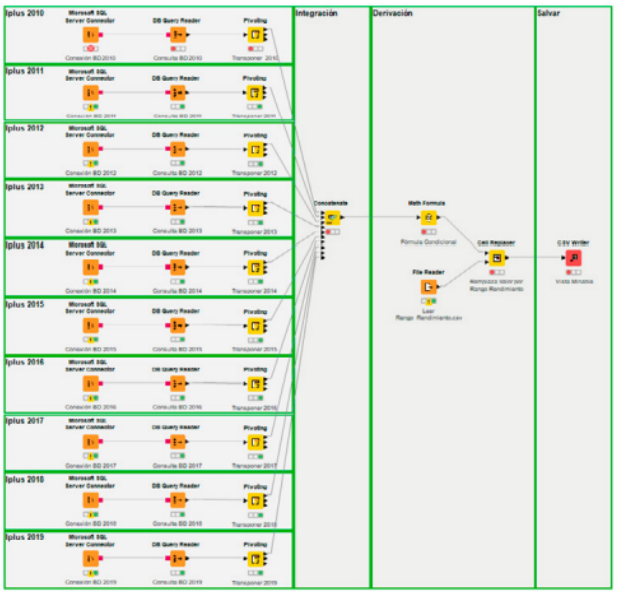

Figura 4.
Flujo de Trabajo para integrar diferentes orígenes de datos.

FIGURA 4
Flujo de Trabajo para integrar diferentes orígenes de datos.
sf

Modelado y Evaluación

Se realiza una partición los datos donde se utiliza el 90 % de los datos para el conjunto de en- trenamiento y el 10 % para el conjunto de prueba. Aplicando *TAKE FROM TOP*, donde, se coloca las filas superiores en la primera tabla de salida (conjunto de entrenamiento) y el resto en la segunda tabla (conjunto de prueba). Para estimar la precisión del modelo es necesario comparar los casos etiquetados en el conjunto de prueba con el resultado de aplicar el modelo, para obtener un porcentaje de clasificación. Si la precisión del clasificador es aceptable, podre- mos utilizar el modelo para clasificar nuevos casos (de los que desconocemos realmente su

clase). Se describen los siguientes diseños de comprobación. De un total de 59 387 registros, se utilizan 53 448 de los datos para el conjunto de entrenamiento y 5 939 para el conjunto de prueba.

Se diseña un modelo de entrenamiento para clasificación en *Knime* (Figura 5). Este flujo

de trabajo demuestra cómo se aplican seis algoritmos de aprendizajes de reglas para identifi- car los indicadores que influyen en el rendimiento a partir de los datos históricos de la zafra azucarera.

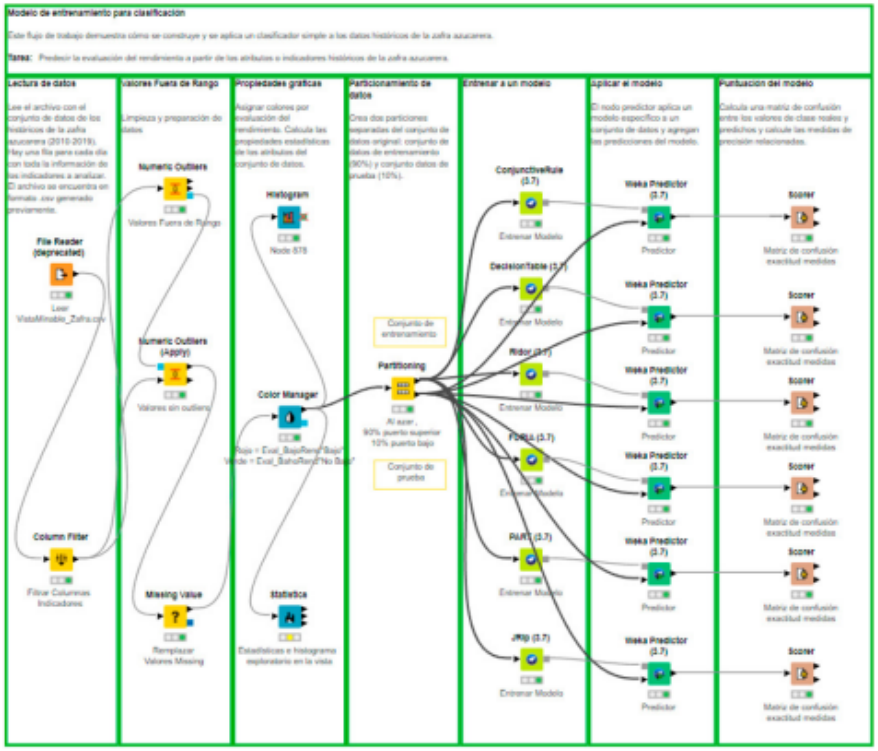

Figura 5. Modelo de entrenamiento para clasificación

FIGURA 5
Modelo de entrenamiento para clasificación
sf

Se obtiene los siguientes modelos para cada uno de los algoritmos aplicados:

- • **Conjunctive Rule**: implementa una sola regla de aprendizaje conjuntiva a partir de la comparación de un conjunto de datos validados (Núñez, Velandia, Hernández, Melén- dez, y Vargas, 2013). Una regla consiste en varios antecedentes "Y" juntos y el valor de la clase para la clasificación. Lo siguiente que realiza el algoritmo es la distribución de las clases disponibles a el término medio parea un valor numérico. Si la instancia experi- mental es poco notoria para esta regla, entonces es prevista utilizando una distribución predeterminada de la clase en los datos cubiertos por la regla de aprendizaje (Ortega y

  Suárez, 2010). La distribución de la clase es la siguiente: Cubierta por la regla (Bajo 0.80, No Bajo 0.19), No cubierta por la regla (Bajo 0.41, No Bajo 0.58).

  • **Decision Table:** clase para construir y usar un clasificador mayoritario de tabla de deci-

  sión simple (Ian H. y Frank, 2011). Las tablas de decisión son una de las formas más sim- ples de representación del conocimiento dentro del ámbito de la clasificación. Su forma más básica de utilización es el almacenamiento de las ocurrencias de los atributos más relevantes sobre cada una de las clases. Estas tablas están formadas por un esquema y un cuerpo. El primero se refiere a los valores de los atributos que mejor representen a las clases; mientras que el cuerpo está representado por varios arreglos que contienen con- teos de correspondencia entre cada valor de atributo y cada clase. Es válido notar que la precisión de este clasificador depende en gran medida del proceso de selección de atri- butos que se realiza en su primera etapa. Es por esto que generalmente se utiliza como función de evaluación para la selección de atributos a la precisión de la propia tabla de decisión utilizando el proceso de validación cruzada. En grandes conjuntos de datos realizar la selección mediante este procedimiento es bastante lento; sin embargo se han ideado técnicas para mejorar la

complejidad computacional del proceso de validación cruzada (Rivas Méndez, 2014). A partir del modelo entrenado se obtiene 8119 reglas para 53448 instancias de entrenamiento.

• **Ridor:** basado en el algoritmo *Ripple Down Rule*. Genera una regla por defecto (predeterminada) y luego toma un conjunto de reglas que predicen clases para la regla predeterminada con el mínimo de error. Entonces genera el mejor conjunto de reglas hasta lograr disminuir el error. Realiza una expansión de excepciones en forma de árbol. Las excepciones son un conjunto de reglas que predicen clases distintas a las predeterminadas (Núñez, Velandia, Hernández, Meléndez, y Vargas, 2013). Las re- glas que son producidas en este algoritmo tienen como propiedad que la mayoría de los ejemplos son cubiertos por las reglas de mejor evaluación y las de menor poder discriminatorio representan las excepciones. Por lo que una persona puede solo mi- rar los primeros niveles e ignorar toda la estructura más profunda, eso es lo especial de las reglas con excepciones (Rivas Méndez, 2014). A partir del modelo entrenado se obtiene 91 reglas.

• **Furia:** este algoritmo constituye una ampliación o extensión de *RIPPER*, un algoritmo de aprendizaje de reglas del estado del arte, preservando sus conocidas ventajas, como por ejemplo, los conjuntos de reglas simples y comprensibles. En particular, *FURIA* apren- de reglas difusas en lugar de reglas convencionales, y conjuntos de reglas no ordenadas en lugar de listas de reglas. Además, para tratar con ejemplos no cubiertos hace uso de una regla eficiente de estiramiento. Los resultados experimentales presentados mues- tran que *FURIA* supera significativamente al algoritmo *RIPPER* original, así como a otros clasificadores como por ejemplo el C4.5, en términos de precisión de clasificación (Coto Palacio, Jiménez Martínez, y Nowé, 2020). La principal diferencia entre una regla difusa y una regla convencional es que la regla difusa suele abarcar más, por lo que juega con ventaja respecto a la regla convencional. Por ejemplo, en las reglas convencionales se producen una serie de modelos con agudos, lo que atrae a transiciones abruptas da- das entre distintas clases, siendo una peculiaridad no intuitiva y, además, puede cuestio- narse. Se esperaría que una clase dada por una regla se disminuyera desde el núcleo a la frontera (de lleno a cero) de manera gradual (Pérez, s. f.). A partir del modelo entrenado se obtiene 45 reglas.

• **Part:** genera una lista de reglas de decisión en orden de jerarquía. En esencia construye una regla, elimina las instancias que cubre y continúa creando reglas recursi- vamente para las instancias finales hasta que no queda ninguna instancia (Núñez, Velandia, Hernández, Meléndez, y Vargas, 2013). Este algoritmo, además, se conside- ra un estándar en la industria como algoritmo de clasificación. Se trata de una im- plementación mejorada de reglas del algoritmo C4.5. Se considera un algoritmo muy mejorado en cuanto a la precisión en materia de predicción (Pérez, s. f.). A partir del modelo entrenado se obtiene 948 reglas. Las estadísticas de precisión del modelo son las siguientes:

- • **JRIP:** basado en el algoritmo *RIPPER* (Poda Incremental Repetida para la Reducción del Error). Utiliza varias comparaciones al mismo tiempo, construye un conjunto de reglas por separado y luego establece comparaciones entre ellas (Núñez, Velandia, Hernández, Meléndez, y Vargas, 2013). Es un algoritmo de aprendizaje que se basa en distintas re- glas que utiliza para crear un conjunto de reglas que se encarga de identificar las clases posibles, mientras minimiza la cantidad de errores. El error se define por el número de ejemplos de formación mal clasificados por las reglas. El algoritmo asume que los datos con los cuáles se ha entrenado previamente, son similares de alguna manera a los datos no vistos sobre los que realizara los cálculos para obtener las distintas reglas. (Pérez, s. f.). A partir del modelo entrenado se obtiene 64 reglas.

  Las estadísticas de precisión de por cada modelo son las siguientes:

Estadísticas de precisión	Conjunctive Rule		Decision Table		Ridor		FURIA		PART		JRIP	
	Bajo	No Bajo	Bajo	No Bajo	Bajo	No Bajo	Bajo	No Bajo	Bajo	No Bajo	Bajo	No Bajo
True Positives	1792	751	5061	8	4744	278	4179	537	4451	276	4194	504
False Positives	113	3283	856	14	586	331	327	896	588	624	360	881
True Negatives	751	1792	8	5061	278	4744	537	4179	276	4451	504	4194
False Negatives	3283	113	14	856	331	586	896	327	624	588	881	360
Recall	0.35	0.87	1	0.01	0.93	0.32	0.82	0.62	0.88	0.32	0.83	0.58
Precision	0.94	0.19	0.86	0.36	0.89	0.46	0.93	0.37	0.88	0.31	0.92	0.36
Sensitivity	0.35	0.87	1	0.01	0.93	0.32	0.82	0.62	0.88	0.32	0.83	0.58
Specificity	0.87	0.35	0.01	1	0.32	0.93	0.62	0.82	0.32	0.88	0.58	0.83
F-meansure	0.51	0.31	0.92	0.02	0.91	0.38	0.87	0.47	0.88	0.31	0.87	0.45
Acurrency	0.43		0.85		0.85		0.79		0.8		0.79	
Cohen's kappa	0.09		0.01		0.29		0.35		0.19		0.33	

SN
st
sf

La métrica **Recall** mide qué tan bueno es el modelo para detectar eventos positivos (Wid- mann, 2019), se obtiene que el algoritmo que es capaz de identificar mejor para clasificar el bajo rendimiento es *DECISIONTABLE* con (1.0), seguido del *RIDOR* con (0.93).

La métrica **Precision** mide qué tan bueno es el modelo para asignar eventos positivos a la

clase positiva (Widmann, 2019), se obtiene que el algoritmo que más precisión presenta para el entrenamiento realizado para clasificar el bajo rendimiento es el *CONJUNCTIVERULE* con (0.94), seguido del *JRIP* con (0.92).

La métrica **Sensitivity** mide qué tan apto es el modelo para detectar eventos en la clase positiva (Widmann, 2019), se obtiene que el algoritmo que más sensibilidad presenta para el entrenamiento realizado para clasificar el bajo rendimiento es el CONJUNCTIVERULE con (0.94), seguido del JRIP con (0.92).

La métrica **Specificity** mide cuán exacta es la asignación a la clase positiva (Widmann, 2019), se obtiene que el algoritmo que más especificidad presenta para el entrenamiento rea- lizado para clasificar el bajo rendimiento es el *CONJUNCTIVERULE* con (0.87), seguido del *FURIA* con (0.62).

La métrica **F-meansure**, es la media armónica de recuperación y precisión (Wid- mann, 2019), se obtiene que el algoritmo que presenta mejor precisión y recuperación para clasificar el bajo rendimiento es *DECISIONTABLE* con (0.92), seguido del *RIDOR* con (0.91).

El **Coeficiente Kappa de Cohen** (κ), un estadístico de concordancia entre dos investiga- dores que corrige el azar (Gordillo & Rodríguez, 2009), se obtiene que el algoritmo que más confiabilidad presenta para el entrenamiento realizado es *FURIA* con (0.35), seguido del *JRIP* con (0.33).

La métrica **Accurancy**, mide el porcentaje de casos que el modelo ha acertado (Martínez Heras, 2020), se obtiene que el algoritmo que presenta mejor precisión y recuperación para clasificar el bajo rendimiento es *DECISIONTABLE* y *RIDOR* con (0.85), seguido del PART con (0.8).

Como resultado de la revisión se obtienen las siguientes reglas que inciden en el rendi- miento:

- • Según el algoritmo ConjunctiveRule se obtiene la siguiente regla con su influencia en el bajo rendimiento:

(i368 <= 12.4485) and (i63d > 1.065) and (i64 <= 14.995) and (i113 <= 14.555) and (i10124 <= 1.34) and (i42a <= 1327.4805) => Eval_BajoRend = Bajo Donde:

» Agua imbibición Total t(i42a)
» Jugo Última Extracción Tándem A Pol %(i63d)

» Jugo Clarificado Brix %(i64)
» Jugo filtros Brix %(i113)
» Rendimiento Guía(i368)
» % caña madurez media(i10124)

- • •Según el algoritmo JRIP se obtiene entre otras, la siguiente regla con su influencia en el rendimiento:
  (i368 >= 11.608) and (i10 >= 99.08) and (i329 <= 239.05) and (i97 <= 55) and (i64 >= 14.19) and (i299 <= 8) => Eval_BajoRend=NoBajo (208.0/84.0) Donde:
  » Rendimiento Guía (i368)
  » Caña atrasada Total t (i329)
  » Miel B extraída t (i97)
  » Jugo Clarificado Brix % (i64)
  » Kilogramos de azúcar (i299)
  » Azúcar Alta Calidad a granel Pol % (i10)

Resulta necesario realizar con posterioridad a esta investigación, una validación y evaluación más profunda de los conjuntos de reglas obtenidas, para descartar factores conocidos que influyen en el rendimiento.

CONCLUSIONES

El trabajo permitió realizar a grandes rasgos una comprensión del negocio, una comprensión de los datos, así como una preparación de los mismos para realizar el modelado de diferentes técnicas.

El trabajo permitió comparar diferentes algoritmos de reglas que permiten identificar las que más se ajustan a los objetivos planteados, así como los indicadores que influyen en la cla- sificación del rendimiento industrial.

El trabajo constituye punto de partida para la evaluación más profunda de las reglas obte- nidas y su posterior validación.

REFERENCIAS

Beck, F., y Fürnkranz, J. (2021). *An Empirical Investigation Into Deep and Shallow Rule Lear- ning. Frontiers in Artificial Intelligence,* 4. Recuperado de https://bit.ly/3M0huZg

Concepción Cruz, E., Caraballoso Torrecilla, V., Nápoles Alberto, R. G., Morales Fundora, L., Cruz Coca, O., y Viñas Quintero, Y. (2015). *PROBLEMAS ASOCIADOS AL RENDI- MIENTO AGRÍCOLA DE LA CAÑA DE AZÚCAR EN LA COOPERATIVA POTRERI- LLO, PROVINCIA SANCTI SPÍRITUS: PROBLEMS ASSOCIATED TO THE AGRICUL- TURAL YIELD OF SUGARCANE IN THE POTRERILLO COOPERATIVE, PROVINCE OF SANCTI SPÍRITUS.* Centro Azúcar, 42(2), 83-92.

Coto Palacio, J., Jiménez Martínez, Y., y Nowé, A. (2020). Aplicación de sistemas neuroborro- sos en la clasificación de reportes en problemas de secuenciación. Revista Cubana de Cien- cias Informáticas, 14(4), 34-47.

Equipo Técnico de Krypton Solid. (2021, diciembre 28). *Examinando la plataforma de análi- sis de Knime para análisis de big data.* Recuperado 9 de enero de 2022, de Krypton Solid website: https://bit.ly/3vksF9d

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

Gordillo, J. J. T., & Rodríguez, V. H. P. (2009). *CÁLCULO DE LA FIABILIDAD Y CONCOR- DANCIA ENTRE CODIFICADORES DE UN SISTEMA DE CATEGORÍAS PARA EL ESTUDIO DEL FORO ONLINE EN E-LEARNING.* 27, 17.

Hernández Orallo, J., Ramárez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la minería de datos.* España: PEARSON EDUCACION. S.A.

Ian H., W., y Frank, E. (2011). *Data Mining: Practical Machine Learning Tools and Techni- ques.* Elsevier. https://doi.org/10.1016/C2009-0-19715-5

Iplus – Datazucar. (s. f.). Recuperado 11 de octubre de 2021, de Datazucar website: https://bit. ly/3Ig8Kv1

Martínez Heras, J. (2020, octubre 9). Precision, Recall, F1, Accuracy en clasificación. Recupe- rado 28 de abril de 2022, de IArtificial.net website: https://bit.ly/37PaLSE

Montequín, R., Teresa, M., Cabal, Á., Valeriano, J., Fernández, M., Manuel, J., y Valdés, G. (s. f.). *METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE DATA MINING.* DATA MINING, 9.

Núñez, V. B., Velandia, R., Hernández, F., Meléndez, J., y Vargas, H. (2013). *Atributos Relevan- tes para el Diagnóstico Automático de Eventos de Tensión en Redes de Distribución de Energía Eléctrica.* Revista Iberoamericana de Automática e Informática Industrial RIAI, 10(1), 73-84. https://doi.org/10.1016/j.riai.2012.11.007

Ortega, R. A. V., y Suárez, F. L. H. (2010). *EVALUACIÓN DE ALGORITMOS DE EXTRAC- CIÓN DE REGLAS DE DECISIÓN PARA EL DIAGNÓSTICO DE HUECOS DE TEN- SIÓN. 127.*

Peña-Ayala, A. (2014). *Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 41*(4), 1432-1462. https://doi.org/10.1016/j. eswa.2013.08.042

Pérez, F. M. (s. f.). *Estudio y análisis del funcionamiento de técnicas de minería de datos en conjuntos de datos relacionados con la Biología. 35.*

Ribas García, M., Consuegra del Rey, R., y Alfonso Alfonso, M. (2016). *ANÁLISIS DE LOS FACTORES QUE MÁS INCIDEN SOBRE EL RENDIMIENTO INDUSTRIAL AZUCA- RERO. 43*(1), 10.

Rivas Méndez, A. (2014). *Estudio experimental sobre algoritmos de clasificación supervisa- da basados en reglas en conjuntos de datos de alta dimensión.* Recuperado de https://bit. ly/3LoZcQR

Widmann, M. (2019, mayo 27). *From Modeling to Scoring: Confusion Matrix and Class Statistics.* Recuperado 20 de febrero de 2021, de KNIME website: https://bit.ly/3vw- jv9u

Wirth, R., y Hipp, J. (2000). CRISP-DM: *Towards a standard process model for data mining.* Proceedings of the 4th International Conference on the Practical Applications of Knowle- dge Discovery and Data Mining.