Ingeniería Industrial/ISSN 1815-5936/Vol. XLII/No. 2/abril-junio/2021/4218

ARTÍCULO ORIGINAL
INFORMÁTICA EMPRESARIAL

Procedimiento para Índice Sintético de Gestión Ambiental: validación con minería de datos

Procedure for a Synthetic Index of Institutional Environmental Management: validation with data mining

Liz Pérez-Martínez [I]
http://orcid.org/0000-0001-6187-7875
Elayne Tápanes-Suárez [II]
https://orcid.org/0000-0003-16633-0560
Orlando Santos-Pérez [II]
http://orcid.org/0000-0003-2420-5732
Juan Alfredo Cabrera-Hernández [I]
https://orcid.org/0000-0002-2723-3619
Dianelys Nogueira-RiveraI
http://orcid.org/0000-0002-0198-852X

[I] Universidad de Matanzas. Matanzas, Cuba
correo electrónico: lizy.perez@umcc.cu, alfredo.cabrera@umcc.cu, dianelys.nogueira@umcc.cu
[II] Empresa de Proyectos de Arquitectura e Ingeniería de Matanzas (EMPAI)
correo electrónico: elayne-tapanes@empai.cu , orlando-santos@empai.cu

Recibido: 12 de abril del 2021.
Aprobado: 1 de junio del 2021.

RESUMEN
Los índices sintéticos tienen gran importancia para la toma de decisiones. El presente artículo tiene como objetivo la aplicación de un procedimiento para Índice Sintético de Gestión Ambiental Institucional validado mediante técnicas de minería de datos y de *Machine Learning*. Se aplicaron test de validación del procedimiento y se realizó una simulación que permitió: comprobar la factibilidad de su implementación como contribución a la gestión ambiental institucional; estructurar, de forma lógica, un grupo de herramientas para la evaluación de la gestión ambiental, así como adecuarlas a instituciones con diferentes objetos sociales. Como resultado, se ofrece información fiable y precisa del estado de la gestión ambiental institucional para el desarrollo de estrategias y planes de acción a partir de los indicadores de mayor incidencia en la gestión ambiental.

Palabras clave: índice sintético, gestión ambiental, minería de datos, toma de decisiones

ABSTRACT
Synthetic indexes are of great importance for decision making. The objective of this article is to apply a procedure for a Synthetic Index of Institutional Environmental Management validated

through data mining and Machine Learning techniques. Validation tests of the procedure were applied and a simulation was carried out to verify: the feasibility of its implementation as a contribution to institutional environmental management; to structure, in a logical way, a group of tools for the evaluation of environmental management, as well as to adapt them to institutions with different social objects. As a result, reliable and accurate information is offered on the state of institutional environmental management for the development of strategies and action plans based on the indicators with the greatest impact on environmental management.

Keywords: *synthetic index, environmental management, data mining, decision making.*

I. INTRODUCCIÓN

Los indicadores son instrumentos de medición analítica ampliamente utilizados en disímiles campos [1], por su acertada contribución a la toma de decisiones [2] y a la mejora continua de los procesos institucionales [3]. De esta forma, permiten reconocer el grado de cumplimiento de los objetivos, así como controlar y replantear estrategias, al expresar un resultado cuantificable que facilita la medición del avance de la organización.

Sin embargo, el listado de indicadores a medir puede tornarse engorroso en la práctica para los directivos de las organizaciones, lo que ha llevado a su reducción a un número más manejable, mediante el uso de herramientas accesibles a la gestión multiescala. Entre dichas herramientas se hallan los índices sintéticos, que permiten un mayor control y, por tanto, una toma de decisiones más efectiva.

Los índices sintéticos representan una medida que se obtiene mediante la agregación adecuada de conjuntos de indicadores por dimensiones en un determinado proceso en estudio [4]. Su concepción no resulta sencilla en la práctica, ya que los indicadores son de distinta naturaleza y provienen de diferentes fuentes de datos, de ahí la variabilidad de los métodos empleados con este fin. Entre los de mayor aplicación se encuentran [5]:

métodos de agregaciones simples, participativos, de análisis multivariante (Análisis de Componentes Principales, Análisis Factorial, Escalamiento Óptimo, Análisis Conjunto)

de análisis multicriterio (Teoría de la utilidad multiatributo Proceso Analítico Jerárquico (AHP)

métodos de sobreclasificación, procedimientos de agregaciones no compensatorias)

y basados en distancias.

La construcción de índices sintéticos en el último lustro ha sido aplicada en diversas áreas, tales como:

- evaluación del clima y la comunicación organizacional [6]
- alineamiento estratégico entre objetivos y procesos [7]
- nivel de servicio en instituciones de atención primaria de salud [8]
- requerimientos higiénico-sanitarios de los alimentos [9]
- gestión del conocimiento en organizaciones productivas [10]
- calidad de vida urbana [11], gestión de accesibilidad y movilidad en centros históricos [5], entre otros.

Escasos son los estudios relacionados con índices sintéticos dedicados a la gestión ambiental, y a elementos específicos de la misma. Las principales incursiones en este sentido están dirigidas a la creación de indicadores para la evaluación de la sostenibilidad ambiental [12], la selección de indicadores ambientales para la elaboración de una estrategia de evaluación ambiental [13], la evaluación del desempeño ambiental [14], y la evaluación de costos ambientales [15].

En los casos consultados, la construcción de los índices sintéticos se basa en criterios de expertos, sin tener en cuenta técnicas de análisis de datos basados en preceptos de la inteligencia artificial, como la minería de datos. El fin de la minería de datos es la realización de análisis automáticos o semiautomáticos de grandes cantidades de datos para extraer patrones desconocidos. Estos patrones resumen los datos de entrada para su empleo en el análisis adicional, aprendizaje automático y análisis predictivo.

A pesar del gran número de técnicas y herramientas que emplean inteligencia artificial, su uso aún continúa siendo insuficiente en algunas esferas, tales como: la ambiental, donde el empleo de las mismas podría dar solución a disimiles problemas de forma más eficiente, y prestar especial atención en la extracción de información.

El empleo de técnicas de minería de datos [16], vislumbra como una alternativa en la construcción de índices sintéticos. En la concluida década (2010-2020) esta variante de tratamiento de la información para generar conocimiento ha adquirido un mayor interés por parte de investigadores de diversas ramas del saber dada la objetividad que aporta a la concepción de estos instrumentos. Resaltan entre las técnicas de mayor aplicabilidad las provenientes de la estadística: análisis de varianza, regresión, análisis de agrupamiento, análisis discriminante y series de tiempo; y de la Inteligencia Artificial: Sistemas Expertos, Redes Neuronales Artificiales y Sistemas Inteligentes.
El presente artículo tiene como objetivo la construcción de un Índice Sintético de Gestión Ambiental Institucional (ISGAI) mediante el empleo de técnicas de minería de datos.

II. MÉTODOS

Aunque en la bibliografía nacional e internacional se constata la existencia de procedimientos para la concepción de indicadores de gestión ambiental, son limitadas las aplicaciones de técnicas de inteligencia artificial como la minería de datos para la construcción de índices sintéticos [17; 18; 19]. Resaltan, entre las escasas aplicaciones existentes, la construcción de un índice sintético para predecir el índice de calidad del agua de los ríos, con el empleo de las técnicas k vecino más cercano, árboles de decisión y redes neuronales artificiales [20]. Mide de manera estable y cuantitativa el impacto de las interferencias en los radares [21] y para construir el Índice Sintético de Desempeño Institucional municipal colombiano [22].
En muchos problemas de minería de datos, habitualmente, un especialista humano define las variables que son potencialmente útiles para caracterizar o representar a un conjunto de datos. Sin embargo, en muchos dominios es muy probable que no todas las variables sean importantes; algunas de ellas pueden ser variables irrelevantes o redundantes que no contribuyen de manera sustancial en tareas de clasificación o de análisis de datos. Por otro lado, existen muchas bases de datos en las que no se conoce la clase a la que pertenecen los objetos de estudio, en las cuales los algoritmos de clasificación supervisada no pueden ser aplicados. En estos escenarios surge la necesidad de emplear algoritmos capaces de clasificar datos, sin la necesidad de conocer la clase a la que pertenece cada objeto de la muestra. De hecho, se trata de encontrar los tipos o clases de objetos que existen en una muestra de datos.
El procedimiento parte de la gestión ambiental institucional, a partir de los datos históricos de las instituciones. El objetivo es realizar un procedimiento de clasificación el cual permita crear indicadores sintéticos a partir de la relación que pueda existir en un conjunto de datos ya sean indicadores o parámetros (figura 1). A partir de la integración y recopilación de estos se genera el almacén de datos, el que a través de la selección, limpieza y transformación origina una serie de datos seleccionados o vista minable. Mediante la aplicación de la minería de datos se identifican los patrones de comportamiento de los parámetros característicos del desarrollo del proceso, los cuales son evaluados e interpretados durante la generación de conocimiento, de cuya difusión y empleo resulta la construcción del Índice Sintético de Gestión Ambiental Institucional (ISGAI). La evaluación de dicho índice constituye un apoyo al proceso de toma de decisiones relacionado con la gestión ambiental institucional.

Fig. 1. Procedimiento para la construcción del Índice Sintético de Gestión Ambiental Institucional

De esta forma se propone un procedimiento para la construcción del ISGAI, a partir de la integración de herramientas de inteligencia artificial como la minería de datos.

III. RESULTADOS

Se procede a la aplicación del procedimiento para la construcción del Índice Sintético de Gestión Ambiental Institucional (ISGAI).

Etapa de integración y recopilación: consiste en la identificación del conjunto de datos de entrenamiento, validación y pruebas para el preprocesamiento. En esta etapa de seleccionan los datos iniciales que serán procesados y son almacenados en el sistema gestor de base de datos PostgreSQL.

Paso 1. Se seleccionó el conjunto de datos, tanto en lo referido a las variables objetivo (aquellas que se quiere predecir o clasificar), como a las variables independientes (las que sirven para hacer el cálculo o proceso). A partir de la aplicación del método *Hold Out*.

La recopilación y el preprocesamiento en *Machine Learning*, tuvo una influencia absoluta en el ajuste del procedimiento y su desempeño. Cuanto más y mejores fueron los datos obtenidos, mejor fue el rendimiento del procedimiento. Las fuentes de los datos fueron diversas: archivos *.csv*, hojas de cálculo, páginas webs, bases de datos, entre otros. Por tanto, se aplicaron disímiles técnicas para su recopilación.

Paso 2. Para comenzar a entrenar los algoritmos fue necesario transformar los datos a partir de la normalización de sus atributos de forma tal que sus valores estuvieran en un rango entre 0 y 1. Los datos fueron transformados a una estructura *data frame* para hacer posible su manipulación en el lenguaje R. Una vez realizado el preprocesamiento de los datos, estos quedaron normalizados y en formato *data frame*, para su uso en la construcción del procedimiento.

Paso 3. Se dividió el conjunto de datos mediante el empleo del método ***hold out***, que consiste en mantener aparte una porción de los datos como conjunto de datos de prueba, pues en el proceso de desarrollo, se entrena el procedimiento con la fracción restante de datos. Se ajustan sus parámetros con los datos de validación, y finalmente se evalúa su rendimiento con el conjunto de datos de prueba.

Etapa de selección, limpieza y transformación: consiste en la determinación de las técnicas de *Machine Learning* a emplear en la clasificación de los indicadores. En esta etapa de se procesan los datos iniciales hasta obtener un conjunto de datos minables, este tratamiento se realiza empleando métodos de minería de datos mediante la herramienta RStudio.

Paso 1. Se seleccionaron las técnicas de minería de datos y se diseñó el procedimiento predictivo y de clasificación, mediante la aplicación de los métodos Análisis de Componentes Principales (PCA), Árboles de Decisión, *Random Forest*, y *Receiver Operating Characteristic* (ROC).

CONSTRUCCIÓN DE ÍNDICE SINTÉTICO DE GESTIÓN AMBIENTAL INSTITUCIONAL CON MINERÍA DE DATOS

Paso 1. a) Análisis de Componentes Principales (PCA): para reducir la dimensionalidad del conjunto de datos, se transformó el conjunto de variables originales en otro conjunto de variables correlacionadas llamadas componentes principales. Se tomó el 70 % de los datos para el proceso de entrenamiento y validación, y el 30 % restante exclusivamente para pruebas. Para el cálculo de los componentes principales se empleó únicamente el conjunto de entrenamiento, a partir de esto se definió la matriz de transformación que posteriormente será aplicada a los datos de prueba. Esta técnica se aplicó para evitar el sobreajuste del procedimiento, pues PCA busca las correlaciones entre características, donde esta correlación implica que hay redundancia en los datos.

Paso 1. b) Árboles de decisión: los árboles de decisión son una serie de condiciones organizadas de forma jerárquica, a modo de árbol, en el que a los nodos terminales se les llaman hojas y a cada nodo no terminal del árbol se asocia un atributo y este a su vez a una condición, que determina cuáles datos de la muestra entran en esa rama, de tal manera que la decisión final se puede determinar a partir de las condiciones que se cumplen desde la raíz del árbol hasta algunas de sus hojas, lo que permite una fácil interpretación.

Paso 1. c) *Random Forest*: es una técnica que combina una cantidad grande de árboles de decisión independientes probados sobre conjuntos de datos aleatorios con igual distribución. En la fase de aprendizaje se crearon muchos árboles de decisión independientes, construyéndolos a partir de datos de entrada ligeramente distintos. Se alteró, por tanto, el conjunto inicial de partida, a partir de lo siguiente:

- Se seleccionó aleatoriamente con reemplazamiento un porcentaje de datos de la muestra total. Se incluyó un segundo nivel aleatoriedad, en el que se afectaron los atributos.
- En cada nodo, al seleccionar la partición óptima, se consideró solo una porción de los atributos, elegidos al azar en cada ocasión. Una vez que se generaron muchos árboles, la fase de clasificación se llevó a cabo de la siguiente forma: cada árbol se evaluó de forma independiente y la predicción del bosque fue el voto mayoritario sobre todos los arboles del bosque, es decir la clase con mayor voto.

Paso 2. *ROC (Receiver Operating Characteristic):* la curva ROC se empleó como representación gráfica del rendimiento del clasificador que muestra la distribución de las fracciones de verdaderos positivos y de falsos positivos. La fracción de verdaderos positivos (sensibilidad), arrojó la probabilidad de clasificar correctamente al individuo cuyo estado real fue definido como positivo. Mientras que la especificidad arrojó la probabilidad de clasificar correctamente a un individuo cuyo estado real fue clasificado como negativo.

Cada resultado de predicción representó un punto en el espacio ROC. El mejor método posible de predicción se situó en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, considerándose un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). Una clasificación totalmente aleatoria marcó un punto a lo largo de la línea diagonal, línea de no-discriminación. En definitiva, se considera un procedimiento inútil, cuando la curva ROC recorre la diagonal positiva del gráfico.

Etapa de minería de datos: consiste en la construcción del conjunto de datos para entrenamiento, validación y prueba a utilizar en el clasificador. En esta etapa se obtienen los patrones presentes en el conjunto de datos minables, a partir del análisis estadístico de las series temporales obtenidas mediante el empleo de métodos de minería de datos ajustados de las librerías de la herramienta RStudio.

Paso 1. Se analizaron las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos). Se transformó el conjunto de datos de entrada con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adaptó a los datos y al problema. A partir de la aplicación de los métodos Series Temporales, Descomposición estacional, Dickey-Fuller aumentada, Kwiatkowski-Phillips-Schmidt-Shin, Phillips-Perron.

El fragmento de código de la figura 2 muestra cómo se estableció la conexión a la base de datos y las consultas SQL (*Structure Query Language*) pertinentes para la obtención de los datos solicitados[1].

[1] Los datos que se utilizan en esta investigación son extraídos de la web Banco Mundial de Datos. Esta web ofrece datos de acceso abierto y gratuito sobre el desarrollo en el mundo. Fueron seleccionados para el entrenamiento de los procedimientos los datos Temperatura Mensual (˚C) – Cuba del Centro de Análisis de Información, División de Ciencias Ambientales del Laboratorio Nacional de Oak Ridge (Tennessee, Estados Unidos).

```
library(RPostgreSQL)
leer<-function(parametro="ph")
{
parametro<-paste("'",parametro, sep = "")
parametro<-paste(parametro,"'",sep = "")
con<-
dbConnect(PostgreSQL(),user="postgres",password="admin",dbname="obsam",port="5432"
)
consulta<-paste("SELECT id FROM parametro WHERE nombre=",parametro,sep = " ")
idParametro<-dbGetQuery(con,consulta)
consulta<-paste("SELECT id FROM relacion WHERE parametro=",idParametro$id,sep = " ")
idMediciones<-dbGetQuery(con,consulta)
valores<-dbGetQuery(con,"SELECT * FROM medicion")
return(data<-valores[valores$id== idMediciones,])
}
```

Fig. 2. Conexión con la base de datos y consultas SQL.

Una vez cargada correctamente la base de datos, en el programa se pasó a la transformación de los datos en una serie temporal.

Paso 1. a) Transformación de los datos en una serie temporal: una serie de tiempo es una lista de unidades de tiempo ordenadas tales como fechas, semestres o trimestres, cada una de las cuales se asocia a un valor. Las series de tiempo son un modo estructurado de representar datos. Visualmente, es una curva que evoluciona a lo largo del tiempo. El pronóstico de las series de tiempo significa que se extienden los valores históricos al futuro, donde aún no hay mediciones disponibles. Existen dos variables estructurales principales que definen un pronóstico de serie de tiempo, el período, que representa la frecuencia con la que se miden los datos y el horizonte, que representa la cantidad de períodos por adelantado que deben ser pronosticados.

Las series temporales se pueden definir como un caso particular de los procesos estocásticos, ya que un proceso estocástico es una secuencia de variables aleatorias, ordenadas y equidistantes cronológicamente referidas a una característica observable en diferentes momentos. El análisis de series temporales explica el hecho de que los puntos de datos tomados a lo largo del tiempo pueden tener una estructura interna (como la autocorrelación, la tendencia o la variación estacional) que debe tenerse en cuenta.

El empleo de series temporales en R se basó en la librería $tseries$, la cual proporcionó una gran cantidad de pruebas y funciones estadísticas que posibilitaron la implementación de un procedimiento predictivo.

La función ts se empleó para crear objetos de series temporales. Estos son vectores o matrices con una clase de "ts" (y atributos adicionales) que representan datos que se han muestreado en puntos equiespaciados en el tiempo. En el caso de la matriz, cada columna de los datos de la matriz contiene una única serie de tiempo (univariante). Las series de tiempo tienen al menos una observación, y aunque no necesitan ser numéricas, se empleó un soporte muy limitado para las

Esta base de datos posee la temperatura en Cuba desde 1901 hasta 2016, reporta un total de 1392 observaciones puesto que tiene una frecuencia mensual. El Banco Mundial de Datos brinda la posibilidad de descargas en diversos formatos como csv, xml y excel. Los datos fueron descargados en formato csv, e introducidos en una base de datos PostgreSQL.

Para cargar los datos desde Postgre fue necesaria la implementación del paquete RPostgreSQL, este permite la conexión de dicha base de datos con RStudio. Posteriormente se construyó la función leer (parámetro), que se le pasa el nombre del parámetro del cual se desea cargar sus mediciones.

series no numéricas. El valor de la frecuencia del argumento se usó cuando la serie se muestreó un número entero de veces en cada intervalo de unidad de tiempo. Por ejemplo, se empleó un valor de siete para la frecuencia cuando los datos se muestrearon diariamente, y el período de tiempo natural fue una semana, o 12 cuando los datos se muestrearon mensualmente y el período de tiempo natural fue un año. Se supone que los valores de 4 y 12 en (por ejemplo) métodos de impresión implican una serie trimestral y mensual, respectivamente.

Para la transformación de los datos se construyó la función $serieTemporal(datos, frecuencia)$, los parámetros de esta función fueron los datos que se transformaron y la frecuencia con la que fueron medidos, lo cual se observa en la figura 3.

```
serieTemporal<-function(datos,frecuencia)
 {
  fecha<-datos[order(datos$fecha),3]
  fechaInicio<-fecha[1]
  año<-as.numeric(format(fechaInicio,'%Y'))
  return(ts(datos[2],frequency = frecuencia,start = año))
 }
```

Fig. 3. Función para la transformación de los datos

En esta función se ordenaron las mediciones cronológicamente de manera ascendente y se tomó el año de la primera medición para dar inicio a la serie temporal. Plantear la fecha final de la muestra no es necesario cuando esta está en constante crecimiento, según la cantidad de datos proporcionados por la base de datos se calcula la fecha final a partir de la inicial.

Paso 1. b) Análisis de la serie temporal: la forma más sencilla de comenzar el análisis de una serie temporal es mediante su representación gráfica. El gráfico que se empleó para representar las series temporales es el de secuencia. Estos son diagramas de líneas en los cuales el tiempo se representa en el eje de abscisas (x), y la variable cuya evolución en el tiempo estudiamos en el eje de ordenadas (y). Para diagramas de dispersión simples, se usó $plot$. Sin embargo, existen métodos de trazado para muchos objetos R, incluidas funciones, marcos de datos y objetos de densidad. Esta función tiene un comportamiento especial, pues en función del tipo de dato que se definen como argumento, generará diferentes tipos de gráfica. Para cada tipo de gráfico, fue posible ajustar diferentes parámetros que controlan su aspecto, dentro de esta misma función. $plot(\ )$ siempre pide un argumento x, que corresponde al eje X de una gráfica. x requiere un vector y si no se especifica este argumento, se obtiene un error y no se creará una gráfica. El argumento más importante de $plot(\ )$ es y, el resto son opcionales. Este argumento también requiere un vector y corresponde al eje Y de la gráfica.

Si todas las variables aleatorias que componen el proceso están idénticamente distribuidas, independientemente del momento del tiempo en que se estudie el proceso, entonces la serie es estacionaria. Es decir, la función de distribución de probabilidad de cualquier conjunto de k variables (siendo k un número finito) del proceso debe mantenerse estable (inalterable) al desplazar las variables s períodos de tiempo, tal que, si $P(Yt+1, Yt+2, \dots, Yt+k)$ es la función de distribución acumulada de probabilidad, la ecuación 1:

$$P(Yt+1,Yt+2,\dots,Yt+k)=P(Yt+1+s,Yt+2+s,\dots,Yt+k+s), \quad \forall t,k,s \qquad (1)$$

Sin embargo, la versión estricta de la estacionariedad de un proceso suele ser excesivamente restrictiva para las necesidades prácticas. Por lo anterior, se empleó un concepto menos exigente. El de estacionariedad en sentido débil o de segundo orden que se da cuando la media del proceso es

constante e independiente del tiempo, la varianza es finita y constante, y el valor de la covarianza entre dos periodos depende únicamente de la distancia o desfase entre ellos, sin importar el momento del tiempo en el cual se calculan.
Una serie puede ser no estacionaria por una variación en la media, una variación en la varianza o por la presencia de estacionalidad. Esto significa que si existe alguno de estos casos es necesario aplicar transformaciones en la serie.
Para el análisis de las series temporales se construyeron las funciones estacionalidad (serie, frecuencia) y estacionariedad (serie, frecuencia). Estas funciones contuvieron las principales pruebas que se realizaron a las series para conocer su composición, lo que se observa en la figura 4.

```
estacionalidad<-function(serie, frecuencia){
  if(frecuencia>1)
    {
    if(nsdiffs(serie)>0)
        return(TRUE)
     }
   return(FALSE)
 }
estacionariedad<-function(serie, frecuencia){
       if(adf.test(serie)$p.value<=0.05 & kpss.test(serie)$p.value>=0.05 & pp.test(serie)
$p.value<=0.05)
   {
    return(TRUE)
   }
         if(adf.test(serie)$p.value>0.05 & kpss.test(serie)$p.value<0.05 & pp.test(serie)
$p.value>0.05)
   {
    return(FALSE)
   }
   if(ur.za(serie, model = "both", lag = frecuencia)@teststat<=ur.za(serie, model = "both",
lag = frecuencia)@cval[2])
   {
    return(TRUE)
   }
   else
    return(FALSE)
 }
```

Fig. 4. Funciones estacionalidad y estacionariedad para el análisis de las series temporales

Paso 1. c) Método de descomposición: estos métodos son eminentemente descriptivos. Se emplearon para separar la serie en subseries correspondientes a la tendencia, la estacionalidad y el ruido (componente aleatorio). En ocasiones tendencia y estacionalidad se enmascaran, a veces una tendencia marcada puede no dejarnos ver la estacionalidad, y viceversa. Los métodos de descomposición estacional separan tendencia, estacionalidad y ruido, pero no predicen. Para predecir fue necesario combinarlos con métodos de ajuste de tendencia. De esta forma se realizó un ajuste de tendencia con el fin de obtener un procedimiento extrapolable, y se le añadió la estacionalidad.
Primero, se determinó cómo se combinan los componentes de la serie estacional. Se emplearon las combinaciones aditiva y multiplicativa. Se dice que hay presencia de una aditiva cuando a pesar del crecimiento de la tendencia, la varianza y la media se mantienen estáticas, en cambio las multiplicativas son cuando la varianza y la media varias en consecuencia de la tendencia. En una serie temporal Xt es una función que depende de cuatro componentes:

Componentes aditivas: Xt =Ct +Tt +St +Et (2)

Componentes multiplicativas: Xt =Ct ×Tt ×St ×Et (3)

Donde:

Tendencia (Tt),

Ciclo (Ct),

Componente estacional (St),

Componente irregular o ruido (Et).

Se empleó la librería *forecast* de R y esta a su vez con un método *decompose* que permitió la descomposición del gráfico para su análisis visual, para la aplicación de este método fue necesario procesar los datos tuvieron una frecuencia mínima dos. En caso de los datos con frecuencia menor que dos, se realizó el análisis por el método matemático, se realizaron pruebas de presencia de raíz unitaria, dado que en caso afirmativo esto implica la no estacionariedad y viceversa.

Las pruebas empleadas para este fin fueron:

- Prueba de Dickey-Fuller aumentada (ADF), es una versión aumentada de la prueba Dickey-Fuller para un conjunto más amplio y más complejo de procedimientos de series de tiempo. La estadística Dickey-Fuller Aumentada (ADF), utilizada en la prueba, fue un número negativo. Cuanto más negativo fuera el número seleccionado, más fuerte fue el rechazo de la hipótesis nula de que existe una raíz unitaria para un cierto nivel de confianza. Para la realización de esta prueba, se empleó la librería $tseries$, que cuenta con esta prueba estadística cuya función se llama $adf.test$, se refleja en la ecuación 4.

  adf.test(ts_temp) (4)

- Prueba de Kwiatkowski-Phillips-Schmidt-Shin, en base a su hipótesis nula de que no posee raíz unitaria. La función empleada fue $kpss.test$, de la ecuación 5.

  kpss.test(ts_temp) (5)

- Prueba de Phillips-Perron, cuya hipótesis nula es que posee raíz unitaria. Se basa en la **prueba de Dickey-Fuller**. Al igual que la prueba de Dickey-Fuller aumentada, la prueba de Phillips-Perron aborda la cuestión de que el proceso de generación de datos podría tener un orden superior de autocorrelación que es admitido en la ecuación de prueba. Mientras que la prueba de Dickey-Fuller aumentada aborda esta cuestión mediante la introducción de retardos de como variables independientes en la ecuación de la prueba, la prueba de Phillips-Perron hace una corrección no paramétrica a la estadística t-test. El ensayo es robusto con respecto a la **autocorrelación** y heterocedasticidad en el proceso de alteración de la ecuación de prueba. La función empleada en $tseries$ fue $pp.test$.

pp.test(ts_temp) (6)

Etapa de evaluación e interpretación: consiste en la construcción, entrenamiento y validación de los clasificadores basados en los procedimientos identificados, y escoger el que ofrezca los mejores resultados. En esta etapa se genera conocimiento de la información obtenida de procesar los datos iniciales.

Se extrajo el conocimiento mediante la minería de datos, se obtuvo un procedimiento que representó patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. A partir de la aplicación de los métodos Clústeres, Árbol de decisión y Reglas de asociación.

Paso 1. Para crear un clasificador, el algoritmo analizó primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usó los resultados de este análisis en un gran número de iteraciones para determinar los parámetros óptimos para crear el procedimiento de minería de datos. A continuación, estos parámetros se aplicaron en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

El procedimiento de minería de datos empleado para crear el algoritmo a partir de los datos tomó diversas formas, que incluyeron:

- Un conjunto de clústeres que describió cómo se relacionan los casos de un conjunto de datos.
- Un árbol de decisión que predijo un resultado y que describió cómo afectan a este los distintos criterios.
- Un conjunto de reglas que describieron cómo se agrupaban los datos.

Paso 2. Una vez obtenido el conjunto de datos apto para iniciar el proceso de selección de la técnica de *Machine Learning* que se empleó para desarrollar el clasificador de indicadores, objeto de esta investigación, se realizó un análisis que consideró el número de variables y el número de ejemplos recolectados.

Cuando el conjunto de datos obtuvo una alta dimensionalidad, o sea posee más de 10 variables o atributos, se puede comprometer la eficiencia del clasificador escogido por tener un procedimiento con una complejidad alta que nos podría llevar a un *overfiting* (sobreajuste). Se puede correr el riesgo de tener atributos ruidosos que pueden tener el mismo peso que los atributos relevantes. En los casos en los que se trabajó con datos de alta dimensionalidad se aplicaron técnicas para la reducción de la misma, tales como: PCA, con el fin de seleccionar los atributos que recojan más información, y tener una descripción de los datos a un menor costo.

Se interpretaron y evaluaron los datos, una vez obtenido el procedimiento, se procedió a su validación para comprobar que las conclusiones que arrojó fueron válidas y suficientemente satisfactorias. En el caso de haber obtenido varios procedimientos mediante el uso de distintas técnicas, se compararon los procedimientos en busca de aquel que se ajustara mejor al problema.

Una vez realizada la lectura y particionamiento de los datos, estos fueron sometidos a un proceso de entrenamiento, validación y prueba, que se observa en la figura 5.

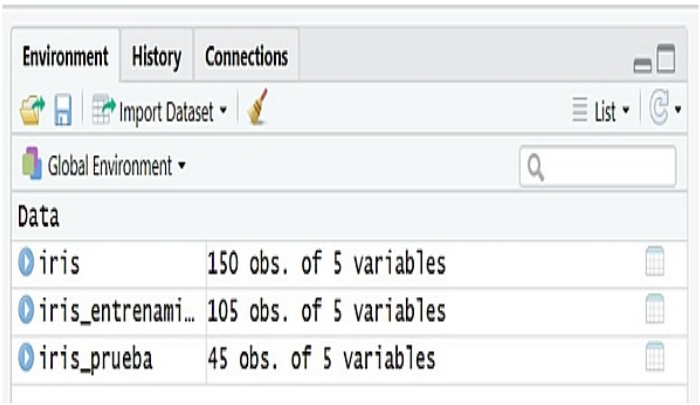

Fig. 5. Resultado del entrenamiento

El primero de los posibles procedimientos de clasificación estuvo basado en los árboles de decisión. Para ello se empleó la librería $rpart$ y se creó un procedimiento a partir de la función $rpart$ de dicha librería donde se declararon los atributos que fueron variables objetivo del procedimiento.
El resultado obtenido fue un árbol con un nodo raíz de 105 ítems. En la primera bifurcación se controla que la longitud del pétalo sea menor que 2.5 de modo que queden a un lado 36 casos y al otro 69. El árbol continuó ramificándose, y así hasta llegar a los nodos hojas, que fueron marcados con asteriscos, que se observa en la figura 6.

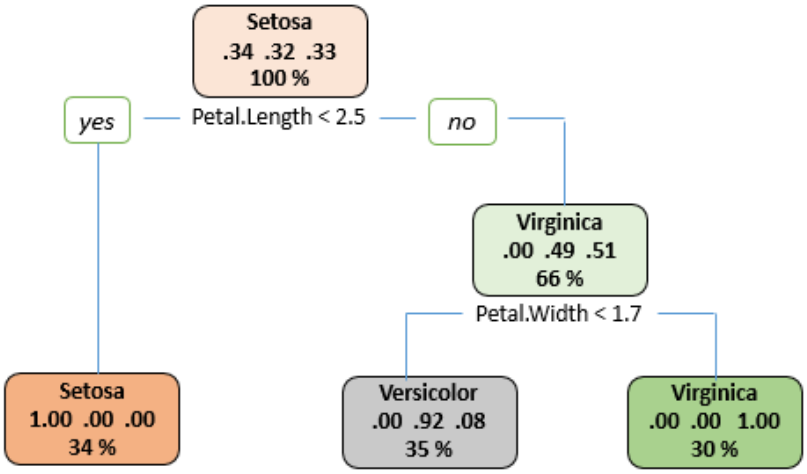

Fig. 6. Gráfico del árbol de decisión generado

Con el set de prueba se generó un vector con los valores predichos por el procedimiento entrenado. Se cruzó la predicción con los datos reales del set de prueba para generar una matriz de confusión, que se refleja en la figura 7.

Confusion Matrix and Statistics

	Reference		
Prediction	setosa	versicolor	virginica
setosa	14	0	0
versicolor	0	14	1
virginica	0	2	13

Fig. 7. Matriz de Confusión.
Fuente: salida de RStudio.

Etapa de difusión y empleo: en esta fase se comunicaron los resultados del análisis a los tomadores de decisiones, el objetivo fue proporcionar a los directivos de las instituciones información fiable y precisa del estado actual de su gestión y de sus proyecciones para el futuro.
A partir de esta información, deben trazarse planes de acciones y estrategias encaminadas a garantizar un correcto desempeño institucional respecto a la gestión ambiental, ya sea para mantener la situación o mejorarla, en dependencia de la interpretación y análisis de los indicadores y del ISGAI.
La información obtenida del análisis una vez comunicada a los tomadores de decisiones, es procesada y convertida en conocimiento; dicho conocimiento será trascendental para un mejor desempeño empresarial. Esto permite definir cuáles son los indicadores de mayor incidencia en la gestión ambiental, entre otras inferencias que se deducen del análisis.

IV. DISCUSIÓN

El procedimiento para ISGAI donde se aplican técnicas de minería de datos para obtener mayor información a partir del aprendizaje previo sobre el conjunto de datos, sus relaciones y la forma en la que inciden en el resultado final. Constituye una herramienta de apoyo para el proceso de toma de decisiones en las instituciones, ya sea tanto para la optimización y mejora de la producción como para minimizar el impacto de sus actividades en el medio ambiente.

El avance incesante de las tecnologías no parece que se frene, el reto radica en emplear estos medios para alcanzar una mejor producción y un mejor desempeño ambiental con un mínimo gasto de recursos humanos y materiales. Entre las claves fundamentales para el éxito está el lograr que el aprendizaje y la aplicación de estas herramientas sea un proceso natural y permanente.

La simulación realizada con el procedimiento propuesto permitió comprobar la factibilidad de su implementación como contribución a la gestión ambiental institucional. Esta herramienta metodológica permite estructurar de forma lógica un grupo de herramientas para la evaluación de la gestión ambiental, así como su adecuación a instituciones con diferentes objetos sociales.

Se realizaron diferentes test de validación al procedimiento propuesto, a través de la relación entre sensibilidad y 1-especificidad se aplicó la curva ROC, lo que arrojó como resultado que el algoritmo que mejor se ajustó a los datos fue el Bosque Aleatorio pues el valor del área bajo la curva es más cercano a 1. No obstante, se evidenció que ambos algoritmos arrojaron resultados similares, con una precisión total de 95.55 % y 93.18 %, para Bosque Aleatorio y Árbol de Decisión respectivamente.

En el ámbito de la gestión ambiental institucional, los decisores requieren de información oportuna, precisa y fiable acerca del medio ambiente y el desarrollo sostenible. Sobre esta base, los sistemas de indicadores poseen el potencial de constituir importantes herramientas en la comunicación de la información científica y técnica. Esto se sintetiza en un monitoreo permanente de las variables y sistemas ambientales, en aras de una evaluación de la sostenibilidad ambiental. A través de la selección de un conjunto de parámetros especialmente diseñados para obtener información específica, según objetivos predeterminados, de algún aspecto considerado prioritario de la relación institución - entorno natural.

El procedimiento propuesto proporcionará información fiable y precisa del estado de la gestión ambiental institucional. A partir de esta información deben trazarse planes de acciones y estrategias encaminadas a garantizar una correcta gestión institucional, ya sea para mantener la situación o mejorarla, en dependencia de la interpretación y análisis del ISGAI. La información obtenida del análisis es procesada y convertida en conocimiento; dicho conocimiento será trascendental para un mejor desempeño institucional. Esto permite definir los indicadores de mayor incidencia en la gestión ambiental, entre otras inferencias que se deducen del análisis.

V. CONCLUSIONES

1. El procedimiento propuesto para la evaluación de la gestión ambiental institucional, consta de cinco fases articuladas. Estas fases permiten: identificar el conjunto de datos a procesar, establecer las técnicas de *Machine Learning* para la clasificación de los indicadores, construir un conjunto de datos para entrenamiento, validación y prueba del clasificador, construir, entrenar y validar los clasificadores basados en los procedimientos identificados, y escoger el que ofrezca los mejores resultados y, en consecuencia, probar el clasificador escogido con nuevos datos que permitan verificar la exactitud del mismo. La obtención de indicadores mediante el empleo de técnicas de minería de datos provee solidez y rigor a la información obtenida, ya que es derivada de la simulación y no de la subjetividad de los investigadores.
2. Con la simulación realizada mediante la aplicación de la minería de datos se identifican los patrones de comportamiento de los parámetros característicos del desarrollo del proceso, los cuales son evaluados e interpretados durante la generación de conocimiento, de cuya difusión y empleo resulta la construcción del ISGAI. La evaluación de dicho índice constituye un apoyo al proceso de toma de decisiones relacionado con la gestión ambiental institucional.
3. Los test de validación realizados a los algoritmos aplicados para la construcción del ISGAI arrojaron resultados satisfactorios, con un porcentaje de precisión por encima de 90, lo que se determina para como alta precisión, por lo que es un procedimiento que emplea algoritmos altamente confiables.

VI. REFERENCIAS

1. Lara Galindo, E., Flores Domínguez, Á. D., & Zulaica, M. L. Evaluación de las condiciones de habitalidad de la ciudad de Puebla (México), mediante la construcción de un índice sintético. I+A Investigación + Acción. 2018; (21):23-42. ISSN 2250-818X.

2. López Palomeque, F., Torres Delgado, A., Elorrieta Sanz, B., Font Urgell, X., & Serrano Miracle, D. (2018). Gestión sostenible de destinos turísticos: la implementación de un sistema de indicadores de turismo en los destinos de la provincia de Barcelona. Polígonos. 2018; (77)428-461. ISSN 2444-0272. DOI: http://dx.doi.org/10.21138/bage.2547.
3. Villacreses Cajamarca, C. J. (2019). Evaluación de indicadores sintéticos de desarrollo sostenible para destinos turísticos consolidados caso Mindo (Pichincha, Ecuador). Facultad de Comunicación Social. [Tesis de Diploma]. Universidad Central del Ecuador, Quito. Disponible en: http://uce.cu/doodle.1222688822234895.pdf.
4. Actis di Pasquale, E. (2015). La elaboración de índices sintéticos de bienestar social. Validación teórica y empírica del método de agregación/ponderación. Artículo presentado al Congreso Nacional de Estudios del Trabajo. El trabajo en su laberinto. Viejos y nuevos desafíos. 5, 6 y 7 de agosto de 2015, Buenos Aires, Argentina. Disponible en: https://nulan.mdp.edu.ar.
5. Santos Pérez, O. (2020). Instrumento metodológico para la gestión de accesibilidad y movilidad en centros históricos. Aplicación en la ciudad de Matanzas. [Tesis de Doctorado]. Universidad de Matanzas, Matanzas, Cuba.
6. Jaquinet Espinosa, R. M. et al. Control de gestión: Facultad de Ciencias Económicas e Informática, Universidad de Matanzas. Revista Ingeniería Industrial. 2015;36 (1):270-81. ISSN 1815-5936.
7. González-Arias, M. et. al. Análisis de la calidad percibida por el cliente en la actividad hotelera. Revista Ingeniería Industrial. 2016;37(3):253-265. ISSN 1815-5936.
8. Rodríguez Sánchez, Y. et al. Análisis de la capacidad de un servicio de urgencia de la Atención Primaria de Salud, mediante simulación. Revista Med.Electrón.2020;42(5):2262-2276. ISSN 1684-1824.
9. García Pulido, Y. A. (2018). Contribución a la gestión de la inocuidad de los alimentos en servicios gastronómicos. [Tesis e de Doctorado]. Universidad de Matanzas, Matanzas.
10. Castillo-Zúñiga, J., Medina-León, A., Medina Nogueira, D., Medina Nogueira, Y. E. & Assafiri-Ojeda, Y. E. et al. Modelo de gestión del conocimiento para el cultivo de Cacao en Vinces. Revista Ingeniería Industrial. 2019; 40(1):48-58. ISSN 1815-5936.
11. Covas Varela, D. et. al. Evaluación de la calidad de vida urbana en la ciudad de Cienfuegos desde una dimensión subjetiva. Universidad y Sociedad. 2019; 9(2): 193-201. ISSN 2218-3620.
12. Shah, R. (2004). Assessment of Sustainability Indicators. Indicators of Sustainable Development: Recent Developments and Activities. Division of Sustainable Development, Department of Economic and Social Affairs, ASI Workshop, Prague, Czech Republic. [Citado: 12 de abril de 2021] Disponible en: http://www.un.org/esa/sustdev/natlinfo/indicators/scopepaper_2004.pdf.
13. Donnelly, A., Jones, M. B., O'Mahony, T., & Byrne, G. (2006). Selecting Environmental Indicators for Use in Strategic Environmental Assessment. Environmental Impact Assessment Review, 2007 (27), 161-175. ISSN: 0195-9255. Disponible en: https://www.elsevier.com/locate/eiar.
14. Miranda Cuéllar, R. L., Pell del Río, S. M., & Fernández Olivera, J. (2016). Proceso de evaluación del desempeño ambiental basada en indicadores sintéticos en Cuba. Revista Dilemas Contemporáneos: Educación, Política y Valores. 2016; IV(1):1-25. [Citado: 12 de abril de 2021] Disponible en: http://www.dilemascontemporaneoseducacionpoliticayvalores.com/index.php/dilemas/article/view/337/704.
15. Perera Conde, L., Nogueiras Valdés, A., & Alcober Álvarez, R. R. Indicadores para la medición de los costos ambientales en entidades de alojamiento turístico: Una necesidad ante la sostenibilidad del uso de los recursos. Explorador Digital. 2021;5(1):185-200.
16. Oded, M., & Lior, R. (2010). *Data Mining and Knowledge Discovery Handbook*. New York. [Citado: 12 de abril de 2021] Disponible en: https://www.springer.com/gp/book/9780387098227.

17. Chavarría Solera, F. Indicadores de gestión ambiental: Instrumento para medir la calidad ambiental de la Universidad Nacional de Costa Rica. Revista de Ciencias Ambientales (Trop J Environ Sci). 2016; 49(1):37-54. ISSN 2215-3896.
18. Acosta Gutiérrez, Z. G et. al. Integración de herramientas de gestión ambiental para reducir vulnerabilidades en áreas ganaderas. Revista Producción Animal. 2018;30(3):1-5. ISSN 2224-7920.
19. Isaac Godínez, C. L. et. al. La integración de herramientas de gestión ambiental como práctica sostenible en las organizaciones. Revista Universidad y Sociedad. 2017;9(4):27-36. ISSN 2218-3620.
20. Babbar, R., & Babbar, S. *Predicting river water quality index using data mining techniques. Environmental Earth Sciences. 2017;*76(14):1-15. Disponible en: https://www.researchgate.net/publications/318677791_Predicting_river_water_quality_index_using_data_mining-techniques.
21. Li, T., Wang, Z., & Liu, J. *Evaluation method for impact of jamming on radar based on expert knowledge and data mining. IET Radar, Sonar & Navigation. 2020;*14(9):1441-1450.
22. Ardila Delgado, A., & García Solano, D. J. Construcción de un índice sintético de desempeño institucional municipal en Colombia. Revista del CLAD Reforma y democracia. 2017; (67): 125-162.

Los autores declaran que no hay conflicto de intereses

Contribución de cada autor

Liz Pérez Martínez: Concepción de la investigación. Redacción de la versión inicial del artículo. Trabajo con software. Escritura de materiales y métodos y las conclusiones del trabajo.

Elayne Tápanes Suárez: Descripción de materiales y métodos. Recopilación, análisis y procesamiento de la información.

Orlando Santos Pérez: Dirección del proceso de investigación, el análisis y procesamiento de la información. Responsable de la escritura de materiales y métodos y las conclusiones del trabajo.

Juan Alfredo Cabrera Hernández: Investigación de los temas del trabajo. Revisión final del artículo.

Dianelys Nogueira Rivera: Investigación de los temas del trabajo. Revisión final del artículo. Adecuación al formato de la revista.