Universidad de Matanzas

Facultad de Ciencias Empresariales

Departamento de Industrial

Título: Procedimiento de *Machine Learning* para la evaluación de la gestión ambiental en la Empresa Comercializadora de Combustibles de Matanzas.

Trabajo de Diploma en opción al título de Ingeniero Industrial

Autor: Beatriz Pérez Cabrera

Tutor(es): MSc. Liz Pérez Martínez

MSc. Azucena González Verde

Matanzas, 2020

Pensamiento

"Lo importante es que seamos capaces de hacer cada día, algo que perfeccione lo que hicimos el día anterior"

Fidel Castro

Dedicatoria

A mamá y papá, son la razón de todos mis logros.

Agradecimientos

A mi mami por su amor diario, por apoyarme en todas mis decisiones sin importar lo difíciles que fueran, todos mis triunfos son para ella.

A mi papi que tanto me ama y cela, que cuida de toda la familia y nos enseña la importancia de mantenernos unidos. Te amo.

A abue, que me consiente tanto y es la mujer de corazón más noble que existe en mi mundo.

A mi numerosa familia que tanto quiero, a mi vecina que desde pequeña me ayudó en las labores de la escuela, y a quien en diez años de relación es mi hombro de apoyo, mi novio.

A mis amigas de estudio, risas y bailes,Yisel y Nayaris, siempre estarán conmigo a pesar de la distancia.

Por último y más importante de todos, sin ella esto no sería posible, a mi tutora Liz, que desde que la conocí me mostró su sonrisa y supe que estaría en buenas manos. También a la profesora Azucena, que me prestó su tiempo y conocimientos sin escatimar para la culminación de esta investigación.

Muchas gracias a todos.

Declaración de autoría.

Yo, Beatriz Pérez Cabrera, declaro que soy la única autora de este trabajo de diploma en opción al grado de Ingeniera Industrial, y autorizo a la Universidad de Matanzas sede “Camilo Cienfuegos” y alaEmpresa Comercializadora de Combustibles de Matanzas, a hacer uso del mismo, para la finalidad que estimen conveniente.

Y para que así conste, firmo la presente a los 26 días del mes de junio del año 2020.

_____ _____ _____

Firma del autor Firma de los tutores

Resumen

La presente investigación fue realizada en la Empresa Comercializadora de Combustibles Matanzas, ubicada en la zona industrial, en el litoral oeste de la bahía de igual nombre, con el objetivo de desarrollar un procedimiento de *Machine Learning* para la evaluación de la gestión ambiental en la misma. Se utilizan varios métodos y herramientas tales como: histórico-lógico, analítico – sintético y el inductivo – deductivo, además de diversos métodos empíricos como la aplicación de una lista de chequeo en la recolección de datos, realización de entrevistas, tormenta de ideas, grupo focal, todo ello se complementa con la consulta de documentos. Los *softwares* utilizados son: el paquete de Microsoft Office, el gestor bibliográfico *EndNote y RStudio* como herramienta de minería de datos. Como resultado se propone un procedimiento de *Machine Learning* basado en la minería de datos, que una vez aplicado provee solidez y rigor en el procesamiento de la información, lo cual ayuda a una gestión más eficaz de las variables de evaluación ambiental.

Summary

This research was carried out at the Matanzas Fuel Trading Company, located in the industrial area, on the west coast of the bay of the same name, with the aim of developing a Machine Learning procedure for evaluating environmental management in the same. Various methods and tools are used such as: historical-logical, analytical - synthetic and inductive - deductive, in addition to various empirical methods such as applying a checklist in data collection, conducting interviews, brainstorming, group focal, all this is complemented by the consultation of documents. The software's used are: the Microsoft Office package, the EndNote bibliographic manager and RStudio as a data mining tool. As a result, a Machine Learning procedure based on data mining is proposed that, once applied, provides solidity and rigor in the processing of information, which helps to more effectively manage the variables of environmental assessment.

Índice

Introducción 1
Capítulo 1. Marco teórico referencial 6
1.1 Contexto empresarial cubano 6
1.2 La empresa cubana y su vínculo con el Medio Ambiente 7
1.3 La gestión ambiental empresarial 10
1.3.1 La gestión ambiental en las empresas de la industria petrolera ... 13
1.4 Indicadores 16
1.4.1 Indicadores ambientales 18
1.4.2 Indicadores de gestión ambiental 19
1.5 Antecedentes de la investigación 21
1.5.1 Investigaciones y proyectos en Cuba 23
1.5.2 Antecedentes tecnológicos 24
1.6 Herramientas y tecnologías. 26
1.6.1 Minería de datos 26
1.6.2 Lenguajes de programación. 29
1.6.3 Herramientas de desarrollo. 30
Conclusiones parciales 31
Capítulo 2. Caracterización de la empresa y propuesta de procedimiento para la Evaluación de la Gestión Ambiental Empresarial 33
2.1. Caracterización de la Empresa Comercializadora de Combustibles Matanzas 33
2.2. Premisas 35
2.3. Elaboración del procedimiento 37
2.3.1. Descripción del procedimiento 40
Conclusiones parciales 53
Conclusiones generales 54
Recomendaciones 55
Referencias bibliográficas 56
Anexos .....

Introducción

En el seno de los países poderosos se encuentra el origen de la pobreza ambiental predominante en el mundo de hoy, al imponer a la humanidad los actuales patrones de desarrollo, donde ha predominado la ignorancia ambiental, junto a la avaricia, el egoísmo y la necedad propia de la especie humana. Pero no solo allí se hace urgente la protección de la vida natural, está en las manos del mundo salvar cada pedazo del planeta.

La preocupación por los problemas ambientales se hizo evidente a mediados del siglo XX, como consecuencia de la contaminación provocada por el acelerado desarrollo industrial. Comenzó entonces a difundirse una serie de ideas que cuestionaban el modelo de crecimiento económico imperante y sus implicaciones en la degradación del ambiente y la afectación de los recursos naturales (Pearce y Turner, 1995). Vino entonces la conferencia de las Naciones Unidas llamada Cumbre de la Tierra celebrada en Río de Janeiro en 1992 donde se inicia la etapa más importante por una mayor cultura y gestión ambiental. El logro más trascendental alcanzado radicó en que se creó una mayor conciencia acerca de los problemas ambientales y de los vínculos entre medio ambiente, economía y la sociedad (Cabrera Hernández, 2004).

La Organización Internacional de Normalización (ISO) unió a sus líneas la serie de normas ISO 14000 que incluye la Norma ISO 14001 que expresa cómo establecer un Sistema de Gestión Ambiental (SGA) efectivo (Hewitt y otros, 1999). En septiembre de 2015 se publica la última versión de esta norma, que trae consigo un enfoque hacia una gestión ambiental estratégica, el liderazgo como factor de éxito y la actitud proactiva en la protección del medio ambiente.

Se puede afirmar que, en Cuba, las concepciones de estrategia y gestión ambiental empresarial se vienen convirtiendo cada vez más en procesos enfocados a enfrentar asuntos importantes para las organizaciones, y ello constituye un aspecto esencial para que las mismas logren hacer efectiva la gestión ambiental y sus estrategias a nivel territorial y nacional.

El Decreto 281/2013 (Reglamento para la implantación y consolidación del Sistema de Dirección y Gestión Empresarial Estatal) (Consejo de Ministro, 2013), dedica su Capítulo VIII a los Sistemas de Gestión Ambiental, y en su

introducción se plantea que: “La incorporación de la Gestión Ambiental en los procesos productivos y de servicios, de las empresas que aplican el Sistema de Dirección y Gestión, tiene el propósito de prevenir, reducir, finalmente eliminar los impactos negativos que estos procesos causan al medio ambiente, y asegurar la protección y preservación de los recursos naturales sobre los cuales se sustentan la producción de bienes y servicios. Es una necesidad insoslayable de las empresas proteger el ambiente”.

En la Conceptualización del Modelo Económico y Social Cubano de Desarrollo Socialista se hace alusión particular al uso racional y la protección de los recursos y el medio ambiente. Por su parte, en el Plan Nacional de Desarrollo Económico y Social hasta el 2030 se enfatiza en el eje estratégico: Transformación productiva e inserción internacional, el fortalecimiento de la competitividad, diversificación y sostenibilidad del sector del turismo. Otro ejemplo es el referido a los recursos naturales y medio ambiente donde se alienta a promover e implementar modalidades de consumo y producción sostenibles. Uno de sus objetivos generales contiene la disminución de la vulnerabilidad del país ante los efectos del cambio climático mediante la ejecución gradual del Plan de Estado para el enfrentamiento al cambio climático, conocido como Tarea Vida. Tal y como queda referido en los documentos del 7mo. Congreso del Partido aprobados por el III Pleno del Comité Central del PCC el 18 de mayo de 2017 y respaldados por la Asamblea Nacional del Poder Popular el 1 de junio de 2017 (CC-PCC, 2017).

En los Lineamientos de la Política Económica y Social del Partido y la Revolución para el período 2016 – 2021 se expone la política para la preservación del medio ambiente. La nueva Constitución de la República de Cuba, aprobada de forma abrumadora en referendo popular el 24 de febrero de 2019 se refiere en su Capítulo II, a la promoción de la protección y conservación del medio ambiente y el enfrentamiento al cambio climático, y deja claro que el Estado protege el medio ambiente y reconoce su estrecha vinculación con el desarrollo sostenible de la economía y la sociedad. Por todo lo antes expuesto y a decir de Castro Felicori (2019), Cuba, como país, presta

especial atención al diseño e implementación de las estrategias ambientales y sus impactos y la adaptación al cambio climático.

En el caso específico de la provincia de Matanzas se avanza poco a poco en la difusión, evaluación y aplicación consecuente de la gestión ambiental, en lo que el sector profesional y académico ha desempeñado, y juega, un papel protagónico.

El desarrollo de la tecnología informática ha posibilitado la manipulación y el almacenamiento de gran cantidad de información en formato electrónico. Actualmente, esta información presenta un crecimiento exponencial, en su mayor parte impulsado por la Internet. De ahí la necesidad de procesarla automáticamente para facilitar la realización de varias tareas como la clasificación y la predicción de la información.

En el proceso de análisis e interpretación de dicha información, resulta de vital importancia la selección de los indicadores que permitan medir de forma más eficiente el resultado de la gestión en correspondencia a las particularidades que definen los procesos. Realizar una correcta predicción de los datos, posibilitará una valoración futura de las consecuencias de las acciones realizadas en la actualidad, y colaborar con la toma de decisiones.

El presente trabajo pretende proponer un procedimiento para la evaluación de la gestión ambiental mediante el uso de los indicadores ambientales frutos de la aplicación de técnicas de minería de datos, con el empleo de la herramienta RStudio. Dicha herramienta brinda, entre otras opciones, la posibilidad de crear gráficos, procesamiento estadístico mediante modelos lineales y no lineales, tests estadísticos y algoritmos de clasificación y agrupamiento.

En el sector empresarial cubano, destaca el caso de las empresas pertenecientes a la industria petrolera, que, dado su actividad productiva, se consideran una potencial amenaza para el medio ambiente. Actividades tales como: la sísmica, la perforación de pozos, emisiones atmosféricas, efluentes líquidos y desechos sólidos y peligrosos. Se hace imprescindible la implementación de planes de manejo ambiental y de un Sistema de Gestión Ambiental liderado. Debido a la especial situación de estas empresas, es vital la revisión y evaluación constante del cumplimiento de los parámetros y datos

medioambientales establecidos, con el objetivo de minimizar el impacto ambiental que ocasionan.

La Empresa Comercializadora de Combustibles de Matanzas perteneciente al Ministerio de Energía y Minas, ubicada en la Bahía de Matanzas, es una de las entidades que ha avanzado de forma meritoria en la implementación del Sistema de Gestión Ambiental, aunque no cuenta con un procedimiento, ni instrumentos para cumplir de forma certera y sistemática la evaluación de esta actividad, lo que trae consigo que no sea posible identificar correctamente o establecer un seguimiento de las debilidades y deficiencias del trabajo ambiental realizado en la entidad para la protección del medio ambiente, así como el establecimiento de un adecuado proceso de mejora continua de dicho trabajo.

De acuerdo a la **situación problémica** anteriormente descrita, se declara el **problema científico** siguiente:

¿Cómo contribuir a la evaluación de la gestión ambiental de la Empresa Comercializadora de Combustibles de Matanzas?

En la investigación se responde a las **preguntas científicas** siguientes:

1. ¿Cuáles son los fundamentos teórico-conceptuales sobre gestión ambiental empresarial, particularizar en la evaluación de la misma?
2. ¿Qué fundamentos metodológicos y qué procedimiento e instrumentos deben adoptarse para la evaluación de la gestión ambiental empresarial en la Comercializadora de Combustibles de Matanzas?

Entonces, el **objetivo general** de la investigación queda expresado como sigue:

Desarrollar un procedimiento de *Machine Learning* para la evaluación de la gestión ambiental en la Empresa Comercializadora de Combustibles de Matanzas.

Y se enuncian las **tareas de la investigación** siguientes:

1. Argumentación del marco teórico-conceptual de la investigación que evidencia la importancia de la gestión ambiental y su evaluación.

2. Selección de un procedimiento para la evaluación de la gestión ambiental en la Empresa Comercializadora de Combustibles de Matanzas, a partir de sus peculiaridades.

En el desarrollo de la investigación se utilizaron como métodos teóricos el histórico-lógico, analítico – sintético y el inductivo – deductivo, además de diversos métodos empíricos que se entrecruzan unos con los otros, como el análisis de indicadores, aplicación de una lista de chequeo en la recolección de datos, realización de entrevistas, tormenta de ideas, grupo focal, complementado todo ello con la consulta de documentos e imágenes satelitales. Los softwares utilizados son: el paquete de Microsoft Office y Visio, el gestor bibliográfico EndNote y RStudio como herramienta de minería de datos.

Entre los **aportes** de la investigación se destacan:

- El teórico-investigativo, al sentar bases para futuras investigaciones.
- El práctico, al proponer un procedimiento que asista a la manipulación de la información referente a la gestión ambiental empresarial.

El **resultado esperado** de esta investigación es contar con un procedimiento que permita evaluar la gestión ambiental empresarial en la Empresa Comercializadora de Combustibles de Matanzas. También contribuirá a la toma de decisiones, permitirá el monitoreo y control de indicadores y facilitará la realización de pronósticos y determinación de tendencias a partir de los datos históricos.

Atendiendo a lo planteado anteriormente, la tesis queda **estructurada** de la manera siguiente:

- Capítulo I “Marco teórico referencial”, en el que se presenta una exposición detallada de los referentes teóricos que argumentan la propuesta y permiten un acercamiento al objeto de estudio.
- Capítulo II “Se caracteriza la empresa objeto de estudio y se presenta el procedimiento de *Machine Learning* para la evaluación de la gestión ambiental en la misma.
- Conclusiones, Recomendaciones, Referencias bibliográficas y los Anexos.

Capítulo 1. Marco teórico referencial

El correcto enmarcado de la investigación a desarrollar determinará el desenlace exitoso de la misma, por lo que este capítulo tiene como objetivo un estudio de las bases teóricas existentes que sustentan la investigación referente a la propuesta de solución. Además, en el propio capítulo se describen las principales tendencias tecnológicas que se emplearán en la investigación. Se incluye también, el análisis de los antecedentes relacionados con la temática investigada, la interacción empresa - medioambiente, y la situación actual de los indicadores de gestion ambiental a nivel empresarial como medidores del contexto ambiental de la empresa.

1.1 Contexto empresarial cubano

Para Acosta (2017) el concepto de empresa revela un trasfondo filosófico que permite conocer la importancia que tienen las "personas" y sus "conversaciones" en el funcionamiento de toda empresa, además de las actividades que se realizan y los recursos que se utilizan. La empresa es una organización social que realiza un conjunto de actividades y utiliza una gran variedad de recursos (financieros, materiales, tecnológicos y humanos) para lograr determinados objetivos, como la satisfacción de una necesidad o deseo de su mercado meta con la finalidad de lucrar o no; y que es construida a partir de conversaciones específicas basadas en compromisos mutuos entre las personas que la conforman.

En el desarrollo normal de sus actividades, las empresas se ven influenciadas por ciertas condiciones de rigor extremo, estas condiciones están determinadas por variaciones internas, de la entidad misma, y variaciones externas, provenientes del entorno y normalmente fuera de control. Estos acontecimientos crean con urgencia la necesidad de una gestión integral en las organizaciones, volviendo a las empresas capaces de buscar en cualquiera de estos momentos la mejor solución para los problemas, disminuyendo gradualmente la improvisación y el riesgo en la toma de decisiones.

En Cuba se han adoptado medidas y transformaciones de notable repercusión en el funcionamiento del sistema empresarial, como parte de la actualización

del modelo económico; que, entre otros propósitos, buscan desatar viejas ataduras, otorgar mayores facultades y lograr más eficiencia y organización. En este contexto, la política ambiental de país desempeña un papel determinante.

1.2 La empresa cubana y su vínculo con el Medio Ambiente

La interacción empresa - medioambiente ha evolucionado de acuerdo al entorno en que la empresa se desarrolla y se interrelaciona. Este entorno se ha transformado de condiciones estables y con reglas fijas, funciona como un sistema cerrado, a otro turbulento y muy competitivo y se ejecuta como un sistema abierto. La empresa influye en el medio, proporciona productos y servicios de calidad para la mejora de la calidad de vida, genera bienes y servicios, empleo, dividendos, pero también consume recursos naturales escasos y genera contaminación y residuos. Los efectos que la empresa genera en su entorno han de clasificarse de carácter económico y social y medioambiental, siendo necesaria una visión más amplia de la definición de empresa como sistema abierto (Roffe, 1997).

Las empresas hoy en día buscan el logro de una imagen corporativa dentro de los márgenes legales de la sociedad unida a la calidad del producto requerida por el cliente y el costo de comercialización llega así a la competitividad en el mercado. Se hace necesario unir a esta búsqueda elementos ambientalistas que no son más que programas de ahorro y racionalización de recursos que además de la protección del medio ambiente obtiene ventajas económicas, aplicauna renovación tecnológica y resultados de coste-beneficio.

A nivel internacional en los comités de la ISO, está adquiriendo gran importancia la certificación de los productos con base en criterios medioambientales y de seguridad. Dentro de la empresa, se debe tener presente el grado de incumplimiento de las normas medioambientales y que la producción de un daño puede dar lugar a procesamiento entre el personal de la empresa, pago de multas importantes, indemnizaciones muy elevadas por la reparación del daño causado e incluso el cierre de la actividad contaminante de la empresa. El tema es de gran preocupación para toda empresa que busque generar beneficios a través de un perfil renovador y ambiental.

No existe un mercado específico para el medio ambiente, pues son bienes públicos, sin precio asignado, pero las acciones para mantenerlo sin contaminar si lo tienen; el hecho es que la sociedad está, en la práctica, concediendo un valor implícito a muchos de estos bienes desde el mismo momento en que se adoptan decisiones con impacto sobre el medio ambiente (Roffe, 1997).

El modelo de desarrollo que prevalece en el mundo que ha permitido avances importantes muestra, desde hace algunas décadas, manifestaciones inequívocas de crisis. Al respecto, la degradación ambiental y situaciones que desmejoran la calidad de vida de la población son preocupantes, de hecho, los problemas socioeconómicos y ambientales amenazan la sostenibilidad del propio proceso de desarrollo de la humanidad, a mediano y largo plazo (Ghul y Leyva, 2015). Es imprescindible realizar procedimientos normados a partir de la política trazada por cada país y en la misma rama empresarial.

Son numerosas las buenas prácticas que se han diseñado e introducido en los últimos años en el mundo empresarial, y entre ellas pueden mencionarse (Machín Hernández y Vazquez Santisteban**,** 2003): ecoeficiencia de procesos, producción más limpia, reconversión tecnológica, aprovechamiento de residuos como materia prima para otros procesos productivos, ahorros de energía, prevención de la contaminación, gestión del riesgo, monitorear y medir con regularidad, calibrar y dar mantenimiento y evaluar periódicamente la conformidad con la legislación y las regulaciones ambientales pertinentes

El proceso revolucionario cubano desarrollado en el país en los últimos 61 años, se ha caracterizado desde tiempos tempranos por apoyar la constante y creciente ola de acciones y medidas implementadas desde el año 1992 cuando fueron planteados mundialmente en la ‘’Cumbre de la Tierra” los problemas medio ambientales que se sucedían en el planeta. En Cuba la conservación del medio ambiente y la protección de los recursos naturales se realiza sobre bases científicas, se elaboran y aplican normas técnicas que contemplan la dimensión ambiental, se crean las bases para desarrollar los Sistemas de Gestión Ambiental Empresarial al diseñar procedimientos basados en las normas internacionales ISO 14000, se han desarrollado una serie de acciones

para introducir y comprometer a las empresas en el concepto de Producción Más Limpia (PML), se ha capacitado a los gestores ambientales e incentivado a los empresarios a que incorporen el componente ambiental como un elemento de competitividad en sus actividades económicas.

La comprensión de la necesidad de organizar las empresas para lograr una relación positiva y de mutuo provecho ha encaminado todos los esfuerzos hacia un futuro de beneficio y no de destrucción continua (Cabrera Guerra, 2011). Muchos son los momentos claves que marcan la evolución y aceptación del vínculo empresa-medioambiente en Cuba, algunos de ellos son: participación de Cuba en la Conferencia de las Naciones Unidas sobre el Medio Ambiente y Desarrollo (CNUMAD), conocida como la Cumbre de Río en 1992, a creación del Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA) en 1994, actualización de la Estrategia Ambiental Nacional para los períodos 2007-2010, 2011- 2015 y 2015-2020, cuyos objetivos yacen principalmente en indicar las vías más idóneas para preservar y desarrollar los logros ambientales alcanzados por la Revolución e identificar los principales problemas ambientales del país. A estos elementos es necesario añadirle los documentos con que cuenta el país para la integración de los conceptos en análisis, ellos son: las Evaluaciones de Impacto Ambiental (EIA), las Auditorías Ambientales e igualmente existe la Inspección Ambiental, con una visión más integral que las Auditorías Ambientales.

Para la autora las actividades socioeconómicas tienen que ir atadas al marco político ambiental junto al trabajo de instituciones, ministerios, centro de investigaciones y entidades ambientalistas. Las acciones inconscientes que el propio hombre ejecuta ocasionan daños mayormente irreversibles que terminan recayendo sobre la humanidad más temprano que tarde.

El deterioro acelerado y creciente del medio ambiente es hoy unos de los problemas más graves que enfrenta toda la especie humana en su conjunto. En lo que respecta a los países subdesarrollados es uno de los factores que agrava con más fuerza las condiciones de vida de cientos de millones de personas (Castro Ruz, 1992).

El estado cubano siempre ha sentido la responsabilidad de hacer suyo el cuidado del medio ambiente, hoy se puede afirmar que no todos los países les han dado la importancia que esto requiere, Cuba a pesar de ser un país subdesarrollado ha empleado disímiles recursos y esfuerzos para llevar a cabo esta actividad. A pesar de todo esto aún algunos sectores, especialmente el empresarial, enfocan su atención principalmente en alcanzar altos niveles de producción y pasan por alto el impacto que tienen sus actividades sobre medio ambiente, por ello el CITMA emitió en 2002 la Resolución No.111 la cual establecía el Sistema Nacional de Monitoreo Ambiental cuyo objetivo es valorar el estado del medio ambiente para contribuir a la toma de decisiones sobre la protección ambiental y el uso sostenible de los recursos naturales, a través de la realización del monitoreo ambiental[1].

1.3 La gestión ambiental empresarial

Las empresas se han caracterizado por la generación excesiva de residuos y por la contaminación del aire, el agua y el suelo. A ello se suma la demanda elevada de recursos naturales, altos consumos de energía y de insumos, de los cuales muchos son tóxicos, dañinos al medio ambiente y a la salud humana, así como la creación de escenarios de riesgos de accidentes y desastres. Todo lo anterior hace que la actividad empresarial sea la que más impactos negativos causa al medio ambiente, debido fundamentalmente a procesos de producción y servicios ineficientes, por las tecnologías y materias primas empleadas y los gastos de energía requeridos, lo que afecta la productividad, eficiencia y competitividad de las mismas. Todos los empresarios se preocupan por conocer su situación ambiental ya que buscan ser más competitivos ambiental, social y económicamente, y de esta manera poder lograr una certificación de calidad bajo las normas NTC – ISO, para ello es necesario la gestión de elementos ambientales que permitan conocer el estado ambiental de las mismas, además, de una evaluación integrada.

[1] Es la recolección sistemática de datos mediante mediciones u observaciones en series de espacio y tiempo de variables previamente identificadas, llamados indicadores, los cuales proporcionan un cuadro sináptico o muestra representativa del estado del medio ambiente nacional o territorial

La Gestión Ambiental Empresarial, según Trujillo (2010), surge como respuesta a una serie de dinámicas generadas por tendencias específicas de mercado como nuevos patrones de calidad y exigencias derivadas de la función social de las empresas, así como también, de las responsabilidades de la misma, no sólo a nivel interno (clientes y proveedores), sino también externas como el ambiente, las comunidades, otras empresas, entes territoriales y autoridades ambientales, enmarcadas dentro del contexto territorial en el cual se desempeña la empresa.

Hoy en día las empresas se encuentran sometidas a constantes cambios, debido a la incorporación de nuevos productos y tecnologías, obligados a resaltar impactos ambientales en los determinados procesos que afectan positiva o negativamente la gestión ambiental, principalmente en las áreas de manejos de residuos, emisión de gases, seguridad industrial y calidad de producto (Alonso, Duarte y Montes, 2006).

“La Gestión Ambiental puede considerarse como una tarea que comprende la evaluación, planificación, puesta en marcha, ejecución y evaluación del conjunto de acciones físicas, financieras, reglamentarias, institucionales, de participación, concertación, investigación y educación, con el fin de mejorar la calidad ambiental objeto de acción”(Estrada Latorre, 2000).

Para el PNUMA (1996) la gestión ambiental queda definida como “el conjunto de políticas, objetivos y programas en materia de medio ambiente que se establezcan y pongan en práctica a fin de contemplar el cumplimiento de todos los requisitos normativos correspondientes al medio ambiente y a la mejora continua y razonable de su actuación en ese sentido”.

La gestión ambiental en las organizaciones debe enfocarse, según (Rodríguez, 2001), “como la exigencia que adquiere mayor relevancia para la supervivencia de las empresas. Estas deben concentrarse en una planificación que involucre el establecimiento de normas, medidas preventivas, indicadores que puedan medir el control, siendo estas herramientas para que la gerencia pueda reducir la carga contaminante y obtener beneficios en la medida que trate de minimizar el impacto ambiental de sus actividades”.

La Ley No. 81/1997, en su Artículo 8 se refiere a la gestión ambiental, como el "conjunto de actividades, mecanismos, acciones e instrumentos, dirigidos a garantizar la administración y uso racional de los recursos naturales mediante la conservación, mejoramiento, rehabilitación y monitoreo del medio ambiente y el control de la actividad del hombre en esta esfera. La gestión ambiental aplica la política ambiental establecida mediante un enfoque multidisciplinario, teniendo en cuenta el acervo cultural, la experiencia nacional acumulada y la participación ciudadana" (Soler del Sol, 1997).

Para Cuba queda mucho camino por recorrer en cuanto a la gestión ambiental, sin embargo, se evidencian programas, normativas y acciones que reafirman el quehacer de las empresas en cuanto al tema. Se crearon las bases para desarrollar los Sistemas de Gestión Ambiental Empresarial al diseñar procedimientos basados en las normas internacionales ISO 14000, se creó el Decreto No.281, del 16 de Agosto de 2007 que registra el "Reglamento para la Implantación y Consolidación del Sistema de Dirección y Gestión Empresarial Estatal", se han desarrollado una serie de acciones para introducir y comprometer a las empresas en el concepto de Producción Más Limpia (PML), y se ha capacitado a los gestores ambientales e incentivado a los empresarios a que incorporen el componente ambiental como un elemento de competitividad en sus actividades económicas.

Para el logro de una gestión ambiental eficaz en Cuba hay que partir del reconocimiento de las condiciones concretas del país, de su modelo de desarrollo, sus logros en materia económica, social y ambiental y de los problemas ambientales existentes. El CITMA, en su condición de Organismo de la Administración Central del Estado rector de la política ambiental, es el encargado de desarrollar la estrategia y concertar las acciones encaminadas a mantener los logros ambientales alcanzados por el proceso revolucionario y contribuir a superar las insuficiencias existentes, con la garantía de que los aspectos ambientales se tienen en cuenta en las políticas, programas y planes de desarrollo a todos los niveles.

Queda clara la importancia de la gestión ambiental como factor fundamental para las empresas y su acometer en el éxito de los retos del presente siglo XXI.

Se precisa de conciencia ambiental, realizar las actividades del país en función de la calidad y gestión ambiental. Es alentadora la situación actual de las tareas ambientales desarrolladas, mostrándose la gestión protegida de elementos socioeconómicos a partir de la innovación tecnológica e investigaciones científicas.

1.3.1 La gestión ambiental en las empresas de la industria petrolera

La industria petrolera en particular realiza numerosos procesos que generan consecuencias directas sobre el ambiente, en especial emisiones atmosféricas, contaminación y desechos sólidos y peligrosos. Por lo que se hace necesaria la acción inmediata de iniciativas que contrarresten estos efectos (Briggs, Tolliver y Szmerekovsky, 2012).

Todas las empresas que de una forma u otra intervienen en el proceso productivo-comercial de los hidrocarburos y sus derivados, están obligados a reforzar las medidas establecidas en cada una de ellas en cuanto a la protección medioambiental y de sus recursos humanos, puesto que en cada paso del proceso existen riesgos de contaminación que atentan contra el medio ambiente y la calidad de vida de las personas que interactúan con estos productos.

El modelo de desarrollo que prevalece en el mundo que ha permitido avances importantes muestra, desde hace algunas décadas, manifestaciones inequívocas de crisis. Al respecto, la degradación ambiental y situaciones que desmejoran la calidad de vida de la población son preocupantes, de hecho, los problemas socioeconómicos y ambientales amenazan la sostenibilidad del propio proceso de desarrollo de la humanidad, a mediano y largo plazo (Ghul y Leyva, 2015).

Las consecuencias de los efectos negativos ocasionados por el manejo y utilización de los hidrocarburos y sus derivados en la actualidad resultan obvias. La contaminación de los mares, los ríos y los suelos, incidiendo en la pérdida de la flora y la fauna, resulta una problemática alarmante que resulta, en gran parte, consecuencia directa de la industria petrolera, ya sea en la extracción, procesamiento, transportación o distribución de las materias primas

y los productos resultantes de la misma debido, entre otras causas, a la ocurrencia de derrames, mala manipulación y vertimientos.

Algunas de las situaciones problemáticas en el ámbito ambiental que se presentan en el sector petrolero de manera general son (Córdoba Durán, 2016):

- Contaminación de fuentes de aguas subterráneas.
- Manejo incorrecto de los fluidos y equipos de perforación.
- Fugas de oleoductos, tanques y pozos (Instalación incorrecta y mal mantenimiento).
- Liberación incontrolada de grandes volúmenes de petróleo.
- Partículas en la atmósfera debido a la alteración del suelo debido a las actividades de construcción y tráfico vehicular, además de la incineración de desechos y quema de gas.
- Emisión de hidrocarburos como el Metano (CH4), Monóxido de Carbono (CO), Dióxido de Carbono (CO2), Óxidos de Nitrógeno (NxOy) y Sulfuro de Hidrógeno (H2S) debido a fugas, derrames y desechos en la producción.
- Liberación incontrolada de gas, explosiones o incendios por reventón del pozo.
- Modificación de la topografía y eliminación de vegetación debido a la construcción de caminos, sitios de perforación e instalaciones de producción.
- Daño de áreas ecológicas frágiles, hábitats críticos de la fauna y especies amenazadas.

Se hace imprescindible destacar además otras consecuencias importantes como:

- Generación de aguas residuales durante actividades relacionadas con la explotación, además de la generada durante el mantenimiento y la limpieza de los equipos.
- Contaminación de fuentes hídricas por aceites, lodos y otros desechos.
- Enfermedades en especies y personas por inhalación de aire contaminado o consumo de aguas contaminadas.

- Pérdida de fertilidad en los suelos, y afectaciones en las actividades relacionadas con la agricultura.
- Contaminación por goteo o derrames generados durante actividades tales como: transporte y operación de maquinaria.
- Generación de residuos sólidos y peligrosos.

Estas consecuencias reafirman la necesidad de una correcta gestión ambiental en las empresas de la industria petrolera, con el firme propósito de disminuir los efectos negativos de sus procesos y establecer un plan de recuperación del medio ambiente.

Las evidencias de los impactos ambientales generados por la actividad petrolera, llevó a las empresas a proponer formas internas de gestión ambiental, los sistemas de gestión ambiental que actualmente se operan, consisten en procedimientos internos de manejo ambiental, los cuales han sido desarrollados por la industria para la industria, en respuesta de la creciente legislación ambiental, y del aumento del interés público por los temas ambientales, entre los que se destacan:

- Los Sistemas de Gestión Ambiental, mayormente se aplican las Normas ISO 14.000.
- Evaluación de Impacto Ambiental.
- Planes de Manejo enfocado en los recursos naturales y los productos peligrosos.
- Análisis de Riesgo Ambiental y Planes de Riesgo Ambiental.
- Planes de Contingencia.
- Monitoreos y actividades de ciencia, tecnología e innovación (CTI).
- Auditorías Ambientales.
- Planes de abandono y restauración en zonas de extracción.

Enfoque integral de dichos sistemas se proponen responder específicamente a:

- Responsabilidad ambiental: Se refiere a un manejo ambiental de los recursos naturales, que garantiza el uso adecuado y beneficios de manera permanente.

- Responsabilidad social: Implica una participación social directa en la planificación y toma de decisiones la organización.

En lo referente a entidades pertenecientes a la industria petrolera, se hace necesario el constante seguimiento y evaluación de la gestión ambiental, debido a los productos que son utilizados y los residuos que producen, considerados altamente contaminantes para el entorno (Vale CapdevilaI y Pérez Silvall, 2016).

La humanidad avanza hacia una era más desarrollada y con más exigencias energéticas y económicas, se hace innegable la necesidad de herramientas y procedimientos que contribuyan a la disminución de los efectos nocivos consecuencias de la industria petrolera, así como programas de recuperación de las zonas afectadas y planes de acciones para la mejora continua de la gestión ambiental de cada entidad.

1.4 Indicadores

Los indicadores son estadísticas seleccionadas por su capacidad de mostrar un fenómeno importante. Los indicadores, a menudo resultan de procesar series estadísticas en formas de agregación, proporción, tasas de crecimiento (entre otras), para poder mostrar el estado, la evolución y las tendencias de un fenómeno que interesa monitorear. Los indicadores se diseñan y producen con el propósito de seguir y monitorear algunos fenómenos o conjuntos de dinámicas que requieren algún tipo de intervención o programa. Los indicadores a menudo se presentan en forma contextualizada (se explica al usuario qué muestra el indicador, su importancia e implicancias), se representan en forma amigable y clara (se utiliza infografía, gráficos y mapas), y en general se publican como Sistemas de Indicadores (del tema en cuestión) como documento en papel y digital, y en forma de sitios Web para facilitar el acceso no experto a su contenido. Al igual que con las estadísticas, los indicadores deben ser respaldados por metadatos, que se conocen habitualmente como hojas metodológicas o fichas técnicas. Con los indicadores adecuados, quienes monitorean los procesos, pueden adelantar tendencias e intervenir antes de que se produzcan procesos indeseables o irreversibles (Martínez Quiroga, 2009).

Como establece González (2012), los indicadores deseables son variables que agregan, o de otra manera, simplifican información relevante, hacen visible o perceptible fenómenos de interés, y cuantifican, miden y comunican información relevante.

En general, un indicador corresponde a una o más variables combinadas, que adquiere distintos valores en el tiempo y en el espacio, y entrega señales al público y a los decisores acerca de aspectos fundamentales o prioritarios en el proceso de desarrollo, en particular respecto a las variables que afectan la sostenibilidad ambiental de dichas dinámicas. Un indicador es un tipo particular de estadística, es un variable que en función del valor que asume en determinado momento y en determinado territorio, despliega significados que no son aparentes inmediatamente, y que los usuarios decodificarán más allá de lo que muestran directamente, porque existe un constructo cultural y de significado social que se asocia al mismo. Un indicador despliega más significados de los que son inmediata o directamente aparentes, siempre y cuando se presenten adecuadamente contextualizados y descritos.

De ahí que no todas las estadísticas puedan ser consideradas indicadores, pues para entrar en esta última categoría, el indicador debe comunicar claramente una historia pertinente, debe ser una señal que alerta sobre lo que ocurre respecto de un fenómeno, problema, desafío o meta acordada, y debe decirlo en forma robusta, clara y contextualizada, sin lugar a dudas o interpretaciones encontradas. Los indicadores son variables, y no valores como a veces se establece.

Se hace preciso definir el concepto de datos puesto que ellos constituyen la materia prima del trabajo estadístico, aún no han sido descritos, validados, ni estructurados, sin ellos los indicadores carecerían de validez y exactitud.

Los datos son un conjunto de valores numerales que se observan, registran o estiman respecto de determinada variable en algún punto del espacio y del tiempo, que habitualmente resultan de la aplicación de algún tipo de levantamiento estadístico (como una encuesta o la explotación de un registro administrativo), medición en terreno u otra forma de medición u observación

como son por ejemplo los diversos instrumentos de percepción remota (Martínez Quiroga, 2009).

Los indicadores son necesarios para poder mejorar. Lo que no se mide no se puede controlar y lo que no se controla no se puede gestionar. Son necesarios para la supervisión, control y para la toma de decisiones, ya que definen cómo alcanzar mejores resultados productivos (Peteiro de Bureau, 2010).

1.4.1 Indicadores ambientales

Los indicadores ambientales corresponden a aquellos que se ocupan de describir y mostrar los estados y las principales dinámicas ambientales, es decir el estatus y la tendencia por ejemplo de: la biota y biodiversidad, la cantidad y calidad de agua, la calidad del aire respirable, la carga contaminante y renovabilidad de la oferta energética, la disponibilidad y extracción de algunos recursos naturales (bosques, pesca, agricultura), la contaminación urbana, la producción de desechos sólidos, el uso de agrotóxicos, la frecuencia e intensidad de los desastres naturales, etc (Martínez Quiroga, 2009).

Al igual que los económicos y sociales, estos indicadores, permiten que los distintos actores y usuarios puedan compartir una base común de evidencias e información cuantitativa, selecta, procesada, descrita y contextualizada. Los indicadores ambientales son los que capturan los principales estadios y dinámicas del medio ambiente en el territorio en cuestión, pudiendo ser presentados en solitario o bien como parte integrante correspondiente a la dimensión ambiental de los indicadores de desarrollo sostenible (González, 2012; Agencia Ambiental, 2002).

El uso de indicadores ambientales se ha extendido, no existe una definición única del concepto y éste varía de acuerdo a la institución y a los objetivos específicos que se persiguen. Una de las definiciones más conocida y aceptada proviene de la Organización para la Cooperación y el Desarrollo Económico (OCDE), que desde hace varios años utiliza un conjunto de indicadores como información base para realizar evaluaciones periódicas del desempeño ambiental de los diferentes países que integran la organización. Un indicador ambiental es un parámetro o valor derivado de parámetros que

proporciona información para describir el estado de un fenómeno, ambiente o área, con un significado que va más allá del directamente asociado con el valor del parámetro en sí mismo (Ebert, 1994).

Según el Florida Center forPublic Management 1998, institución que desarrolló un sistema de indicadores con el fin de asesorar a las dependencias ambientales de la Unión Americana, un indicador ambiental es un elemento que describe, analiza y presenta información científicamente sustentada sobre las condiciones y tendencias ambientales y su significado. En particular, precisa que los indicadores ambientales son estadísticas clave seleccionadas que representan o resumen un aspecto significativo del estado del ambiente, la sustentabilidad de los recursos naturales y su relación con las actividades humanas (Múnera Espinal, 2011).

Una de las principales ventajas de los indicadores ambientales es el hecho de que cuantifican importantes evoluciones en la gestión medioambiental de la empresa y las hacen comparables con el transcurso del tiempo. Si se determinan de una forma periódica, los indicadores medioambientales permiten detectar rápidamente tendencias opuestas y, por consiguiente, también pueden utilizarse como un sistema de alerta temprana (García Cespedes, 2014).

No existe ninguna institución que impulse el trabajo con indicadores ambientales, ellos son una alternativa más de las metodologías actualmente en desarrollo en el tema ambiental, de hecho, no todos los programas de un mismo estilo cuentan con una batería de indicadores ambientales. Un mismo tema implementado por distintos grupos de acción puede o no desarrollarse mediante un conjunto de indicadores ambientales (Pino Neculqueo, 2010).

De acuerdo con todo lo anterior la autora corrobora que la utilidad y validez de estos indicadores de gestión radica en el uso que se les pueda dar y en el grado de interacción con las variables y datos de las organizaciones involucradas en la evaluación y gestión. Los indicadores son una herramienta necesaria para alcanzar un estado deseado e ilustrarlo.

1.4.2 Indicadores de gestión ambiental

Una organización debería establecer indicadores ambientales medibles, dichos indicadores tienen que ser objetivos, verificables y reproducibles para así realizar una gestión enfocada en los procesos ambientales, además, de los procesos internos de las instituciones. Deberían ser apropiados para las actividades, productos y servicios de la organización, coherentes con su política ambiental, prácticos, eficaces en cuanto a costos y tecnológicamente viables.

Los indicadores de gestión ambiental resumen extensos datos medioambientales en una cantidad limitada de información significativa, para asegurar una rápida evaluación de las principales mejoras y de los puntos débiles en la gestión ambiental de la empresa (González, 2012).

Entre las principales funciones de los indicadores de gestión ambiental se encuentran las siguientes:

- Detectar potenciales oportunidades de mejora
- Obtener y perseguir metas medioambientales
- Identificar oportunidades de mercado y potenciales de reducción de costos
- Evaluar el impacto medioambiental de la empresa
- Proporcionar datos esenciales para informes y declaraciones medioambientales exigibles
- Favorecer la implementación de ISO 14001

Una vez definidos y establecidos los indicadores de gestión ambiental, la mayor utilidad de éstos se encuentra al permitir cuantificar aspectos relacionados con al menos los siguientes puntos:

- Ahorro de costos: mejora en el control de materias primas y energía; mejor posición para obtener préstamos y subvenciones; optimización de los costos de residuos y emisiones; reducción de los riesgos de accidentes y los costos de las reparaciones por daños al medio ambiente, etc.
- Ventajas de competitividad: buena imagen de la empresa; relaciones con los agentes externos; aumento de motivación de los empleados.
- Cumplimiento de la legislación vigente aplicable.

1.5 Antecedentes de la investigación

En el estudio y uso de indicadores son muchas las iniciativas realizadas por distintos organismos internacionales. Entre los que se destacan: La Comisión para el Desarrollo Sostenible de Naciones Unidas (CDS), la Organización para la Cooperación y el Desarrollo Económico (OCDE), la Comisión de la Unión Europea y la Comisión Económica para América Latina (CEPAL), el Programa de Naciones Unidas para el Medio Ambiente (PNUMA), entre otros. A estos organismos se les suman las diversas instituciones de los países que trabajan en el área ambiental. Sin embargo, los métodos y herramientas han sido escasos y solo es posible mencionar algunos ejemplos exitosos como el de la OCDE con su Modelo de Presión-Estado-Respuesta (PER), que propone un marco de políticas internacionales y nacionales en base a la estadística ambiental; mientras que por otra parte, el caso de la Unión Mundial para la Conservación de la Naturaleza (UICN) que promueve el método MARPS (Mapeo Analítico, Reflexivo y Participativo de la Sostenibilidad) el cual se aplica a un nivel comunitario. Estas dos resultan ser las mejores experiencias en la detección y aplicación de criterios e indicadores ambientales y de sostenibilidad.

Existen tendencias mundiales de los mecanismos de sostenibilidad y criterios de indicadores ambientales, tales como: El Proceso de Montreal en 1997, Programa Frontera XXI México-Estados Unidos, el Tratado de Libre Comercio, el CIAT-UNEP, OCDE-UNEP. También se han tomado de base las experiencias de Seattle (Estados Unidos), Upper (Austria), Australia (Southwest, Western Australia, SouthernAdelaide, Far North Outland, Lower Hunter y Central COSAT, Gippsland en Victoria, y Huon Valley en Tasmania), las experiencias de la UICN en regiones de los países, tales como: Zimbabwe, Colombia y la India, Cairngorms (Escocia), Estado de Chihuahua (Estadísticas del Medio Ambiente) y de Bosque Modelo Chihuahua.

En el 2000 Nicaragua realizó un estudio junto a el SINIA (Sistema Nacional de Información Ambiental) para determinar indicadores sintéticos ambientales se desarrolla una metodología básica para calcularlos. Estos indicadores ambientales servirían como base para la planificación de los tomadores de

decisiones en el desarrollo económico – ambiental – social. Su metodología inicia con la formación de un grupo capacitado, la revisión de la información ambiental existente, y con ella, la elaboración de una ficha que presentara la información requerida en un formato claro y eficaz. Con toda esta información se realizó la selección de variables con las cuales era posible formular indicadores ambientales. Esta investigación presentó 52 indicadores agrupados en 10 temas en función de los intereses (PNUMA, 2001; Grupo Banco Mundial, 2000; Agencia Ambiental, 2002).

En el 2010 España presenta una investigación con una serie de indicadores sintéticos de turismo sostenible para los destinos turísticos de Andalucía. Su forma de obtención proporciona una visión de conjunto mediante procedimientos que reducen la subjetividad asociada y facilitan la interpretación de los resultados. Presentaron un procedimiento inspirado en la metodología formulada por los profesores Díaz Balteiro y Romero (2003) basada en la Programación por Metas, que permite obtener distintos indicadores sintéticos en función de cómo se utilice la información proporcionada por las variables de desviación. A cada una de las medidas sintéticas obtenidas le denominaron Indicador Sintético de Programación por Metas (IPM). Este estudio relevó dos indicadores que tienen un menor grado de subjetividad y con resultados más fáciles de interpretar: el indicador DCP (Distancia - Componentes Principales) y el indicador IPM, ambos dejan una base de datos propia para realizar análisis locales en función del desarrollo del turismo.

En una publicación del Ministerio de Medio Ambiente: indicadores ambientales, una propuesta para España, quedó evidenciada una propuesta de indicadores que sirvió de base para la versión inicial del Sistema inicial de Indicadores Ambientales. La obra clasifica los mismos por diferentes áreas de interés para el país, además, contiene una revisión de los indicadores ambientales empleados por los principales Organismos Internacionales y por diversos países con amplia trayectoria en el desarrollo de sistemas de indicadores. El proceso de consenso de la selección de indicadores se desarrolló en el foro del Grupo de Usuarios de la Red EIONET española, con participación de los

Centros Nacionales de Referencia y Puntos Focales Autonómicos. Se determinaron más de 60 indicadores en las diferentes áreas.

1.5.1 Investigaciones y proyectos en Cuba

A partir del triunfo revolucionario Cuba dio inicio al desarrollo de una conciencia ambiental, a todos los niveles, que ha evolucionado paulatinamente y siempre en aumento. Numerosas instituciones del país, vinculadas a los recursos naturales de mayor importancia, han dedicado esfuerzos en el establecimiento de sistemas de monitoreo ambiental de corte específico, donde se destacan el agua, suelo y aire.

Años atrás se comenzó a trabajar en la creación de un conjunto selectivo de datos e informaciones sobre medio ambiente, que a su vez constituyeran la base para el desarrollo de los indicadores a escala nacional. Esta recopilación de indicadores posibilitó que se pudiera iniciar el desarrollo de un Sistema de Datos e Informaciones sobre Medio Ambiente, al que se le denomino SIMARNA, el cual se caracterizó por ser una base de datos descriptiva sobre los componentes naturales del medio ambiente y que contemplaba a su vez información de índole económica, social y legal, la que estaba directamente relacionada con la gestión y manejo del medio ambiente y el uso racional de los recursos naturales. El desarrollo y utilización de este sistema permitió a las autoridades cubanas la elaboración de importantes documentos y diagnósticos de la situación ambiental vinculada con el desarrollo económico y social del país, a su vez contribuyó crear o identificar una serie de actividades de relevancia en esa esfera.

El Proyecto Estrategia Ambiental Nacional 2016-2020 plantea que se ha mantenido un comportamiento estable en cuanto a la generación de indicadores ambientales y que esto se refleja en las estadísticas oficiales del país. No obstante, aún se adolece de un Sistema de Información Ambiental, que sea capaz de articular: los sistemas de indicadores ambientales y de monitoreo ambiental, los datos e informaciones resultantes de la investigación científica, las evaluaciones e informes ambientales y la información regulatoria y geoespacial, incluyendo la infraestructura necesaria para garantizar la conectividad y la

disponibilidad de herramientas web para dar visibilidad a la información que el sistema gestiona.

El objeto de estudio de la investigación se establece a partir del proceso que se lleva a cabo luego de obtenidos los datos de un monitoreo ambiental, realizado en una determinada empresa, pues en dicho proceso la gestión de los datos posee poca organización debido a que la información no está centralizada lo que trae consigo una mala manipulación, falta de integridad y persistencia de los datos y como consecuencia de esto,en ocasiones la información emitida carece de veracidad.

A lo anterior se añade, la contratación a terceros para que realicen el monitoreo, debido a la indisponibilidad de recursos. Esta situación conlleva a un encarecimiento del proceso y largos periodos de tiempo de ejecución del mismo. Además de que la confiabilidad y seguridad en la interactividad entre los usuarios que manejan la información es pobre, ya que no existe ningún mecanismo que regule y garantice este proceso.

El proceso de toma de decisiones también se ve afectado, ya que la información no llega a los decisores en tiempo, ni con la veracidad requerida para que el proceso fluya adecuadamente. El monitoreo ambiental de las empresas, la determinación de los problemas ambientales y sus causas, así como las áreas más afectadas, dependen de la subjetividad de los evaluadores y no de herramientas confiables. La realización de pronósticos y determinación de tendencias a partir de los datos históricos no es posible ya que en numerosas ocasiones esta información no existe.

1.5.2 Antecedentes tecnológicos

Hasta el momento de esta investigación no existía en el país un sistema de gestión ambiental empresarial que se ajustara a las normas ambientales establecidas por el CITMA. Los estudios ambientales actualmente se realizan mediante pequeños sistemas de resultados muy limitados por sus características basadas en intereses específicos, algunos de estos sistemas estudiados son:

EGAM Medio Ambiente: es un software diseñado para ayudar a cualquier empresa u organización a implantar, mantener y certificar su sistema de gestión ambiental acuerdo a lo exigido en la Norma ISO 14001, que permite despreocuparse de las evidencias y procedimientos, eliminar la burocracia, aumentar la productividad y reducir el coste (eGAM, 2013).

DISPER 5.2 (contaminación atmosférica): Software para empresas de consultoría ambiental que permite evaluar la contaminación atmosférica en el medio ambiente: impacto ambiental, auditoría medioambiental y gestión ambiental en general (Fernández Debill, 2016).

CUSTIC 3.2 (contaminación acústica): software para empresas de consultoría ambiental que permite evaluar la contaminación sonora y el ruido: impacto ambiental del ruido, contaminación acústica, gestión ambiental del ruido y de la contaminación acústica en general (Fernández Debill, 2016).

DESCAR 3.2 (contaminación marina): Software para empresas de consultoría ambiental que permite evaluar la dispersión de contaminantes en el agua: Impacto Ambiental, Auditoría Medioambiental y Gestión Ambiental en general (Fernández Debill, 2016).

RADIA 2.1 (contaminación electromagnética): Software para empresas de consultoría ambiental que permite evaluar la contaminación electromagnética producida por las estaciones base de la telefonía móvil y sus posibles efectos sobre la salud: impacto ambiental, auditoría medioambiental y gestión ambiental en general (Fernández Debill, 2016).

eco2biz: es una plataforma tecnológica diseñada y desarrollada para proveerles una eficiente gestión, control y visibilidad de las actividades que realizan las empresas hoy en día, en el cumplimiento oportuno de sus compromisos con el medio ambiente, manejo de los residuos sólidos y que permita evitar los sobrecostos (Martínez Hals, 2015).

Como se puede apreciar, los software mencionados anteriormente son de propósitos muy específicos y no cumplen con lo establecido en las normas ambientales cubanas. En otros casos, son sistemas que se ejecutan bajo

licencias de software propietarias, lo cual va contra la política del país de migración a software libre. Dados estos motivos, es necesario desarrollar el SAPGAE permitiendo ajustarse al cumplimiento de las normas legales de gestión ambiental.

1.6 Herramientas y tecnologías.

1.6.1 Minería de datos

La minería de datos es un campo de la estadística y las ciencias de la computación referida al proceso de detectar la información procesable de los conjuntos grandes de datos. El término es un concepto de moda, y es frecuentemente mal utilizado para referirse a cualquier forma de datos a gran escala o procesamiento de la información. En el uso de la palabra, el término clave es el descubrimiento, comúnmente se define como "la detección de algo nuevo", para esto utiliza el análisis matemático para deducir los patrones y tendencias que existen. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional porque las relaciones son demasiado complejas o porque hay demasiados datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en conocimiento. Para entender su significado es imprescindible conocer los términos relacionados en esta:

- Datos: son cualquier hecho, número o texto que puede ser procesado por una computadora. Hoy en día, las organizaciones acumulan grandes cantidades, y cada vez mayores, en diferentes formatos y diferentes bases de datos
- Información: los patrones, asociaciones, o relaciones entre todos estos datos pueden proporcionar información. Por ejemplo, el análisis del punto de venta de datos de transacciones puede dar información sobre qué productos se venden y cuándo.
- Conocimiento: la información puede ser convertida en conocimiento acerca de los patrones históricos y las tendencias futuras. Por ejemplo, la información resumida sobre las ventas de supermercados minoristas puede ser analizada a la luz de los esfuerzos de promoción para facilitar el

conocimiento del comportamiento de compra del consumidor. Por lo tanto, un fabricante o distribuidor puede determinar qué elementos son los más susceptibles a los esfuerzos de promoción.

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces se refiere al conocimiento (Martínez Cohen, 2018).

Suscita cierta polémica el definir las fronteras existentes entre la minería de datos y las disciplinas análogas como la estadística y la informática. Pero a pesar de existir muchas similitudes, en la minería de datos se encuentran una serie de problemas y métodos específicos que la hacen distinta de otras disciplinas.

El hecho es que, en la práctica la totalidad de los modelos y algoritmos de uso general en minería de datos como redes neuronales, árboles de regresión y clasificación, modelos logísticos y análisis de componentes principales gozan de una tradición relativamente larga en otros campos.

Ciertamente, la minería de datos bebe de la estadística, de la que toma las siguientes técnicas (Maimon y Rokach, 2010):

- Análisis de varianza, mediante el cual se evalúa la existencia de diferencias significativas entre las medias de una o más variables continuas en poblaciones distintas.
- Regresión: define la relación entre una o más variables y un conjunto de variables predictoras de las primeras.
- Análisis de agrupamiento o *clustering*: permite la clasificación de una población de individuos caracterizados por múltiples atributos (binarios, cualitativos o cuantitativos) en un número determinado de grupos, con base en las semejanzas o diferencias de los individuos.
- Análisis discriminante: permite la clasificación de individuos en grupos que previamente se han establecido, permite encontrar la regla de clasificación

de los elementos de estos grupos, y por tanto una mejor identificación de cuáles son las variables que definan la pertenencia al grupo.

- Series de tiempo: permite el estudio de la evolución de una variable a través del tiempo para poder realizar predicciones, a partir de ese conocimiento y bajo el supuesto de que no van a producirse cambios estructurales.

Técnicas de minería de datos

Como ya se ha comentado, las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Las técnicas más representativas son:

- Redes neuronales. Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.
- Regresión lineal. Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.
- Árboles de decisión. Es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.
- Reglas de asociación. Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.
- Series temporales. Son utilizadas para el análisis de la relación causal entre diversas variables que cambian con el tiempo y se influyen entre sí.

En resumen, la minería de datos se presenta como una tecnología emergente, con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; por otro, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Además,

no hay duda de que trabajar con esta tecnología implica cuidar un sinnúmero de detalles debido a que el producto final involucra "toma de decisiones".

Machine Learning

Machine Learning es un conjunto de técnicas que hacen parte de la inteligencia artificial, que basadas en algoritmos buscan el aprendizaje dentro de grandes conjuntos de datos. Una característica muy importante de estos algoritmos es la predicción de nuevos casos basándose en la experiencia aprendida del conjunto de datos utilizados para su entrenamiento, a esto se le conoce en la literatura como generalización (Fernández, 2003; Beltrán, 2008).

El aprendizaje en *Machine Learning* se divide usualmente en dos tipos, el aprendizaje “supervisado” donde cada uno de las observaciones o muestras del conjunto de datos tiene relacionado una variable o un dato que indica lo que sucedió, lo que pasó, es decir las entradas están etiquetadas. Este tipo de aprendizaje se subdivide en clasificación y regresión (Moreno García, 2007; Freitas, 2002).

El otro tipo de aprendizaje es el “No supervisado”, en este tipo de aprendizaje, en el conjunto de datos se disponen de datos para el entrenamiento, pero no se conoce o no se dispone de la salida o se conoce muy poco sobre esta, es decir no hay una variable objetivo y lo que se requiere es buscar patrones. Para determinar lo que se quiere predecir se pueden encontrar estructuras sobre los datos, dentro de estas estructuras se pueden mencionar el *clustering* (proceso de particionar un conjunto de datos en un conjunto de subclases significativas llamadas grupos) y la asociación (conjunto de características significativas).

1.6.2 Lenguajes de programación.

Lenguaje R

Creado en 1993, en la universidad de Auckland. Viene derivado de otros dos lenguajes, que son S y Scheme. Sus creadores son Ross Ihaka y Robert Gentleman. Es un lenguaje con licencia GNU, es decir, es libre, gratuito y abierto. En resumen, lo puede usar cualquiera y no es propiedad de nadie. R funciona con paquetes gratuitos, como las librerías en otros lenguajes, y

puedes descargar y usar esos paquetes. Algunas de sus características principales como lenguaje son:

- Posibilidad de crear gráficos, basado en LaTeX.
- Gran cantidad de herramientas estadísticas:
  - modelos lineales y no lineales.
  - tests estadísticos.
  - algoritmos de clasificación y agrupamiento.
- Posibilidad de crear tus propias funciones, además de objetos al ser su programación POO (orientada a objetos).
- Integración con distintas bases de datos.
- Puede tener un uso matemático.

Extensiones y paquetes

R forma parte de un proyecto colaborativo y abierto. Sus usuarios pueden publicar paquetes que extienden su configuración básica. Existe un repositorio oficial de paquetes cuyo número superó en otoño de 2009 la cifra de los 2000.Dado el enorme número de nuevos paquetes, estos se han organizado en vistas (o temas), que permiten agruparlos según su naturaleza y función. Por ejemplo, hay grupos de paquetes relacionados con estadística bayesiana, econometría, series temporales, etc.

Para facilitar el desarrollo de nuevos paquetes, se ha puesto a servicio de la comunidad una forja de desarrollo que facilita las tareas relativas a dicho proceso (The R Development Core Team, 2009).

1.6.3 Herramientas de desarrollo.

RStudio

Es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo. RStudio está disponible para Windows, Mac y Linux o para navegadores conectados a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux). RStudio tiene la misión de proporcionar el entorno informático estadístico R. Permite un análisis y desarrollo para que cualquiera pueda

analizar los datos con R. entre sus principales características se encuentran (RStudio, 2018):

- IDE construido exclusivo para R
- El resaltado de sintaxis, auto completado de código y sangría inteligente.
- Ejecutar código R directamente desde el editor de código fuente.
- Salto rápido a las funciones definidas.
- Potente autoría y depuración.
- Depurador interactivo para diagnosticar y corregir los errores rápidamente.
- Herramientas de desarrollo extensas.
- Autoría con Sweave y R Markdown.

Se decidió optar por RStudio por ser un IDE exclusivo para R. Este logra un entorno más agradable para trabajar y cuenta con extensas herramientas para el desarrollo de aplicaciones específicamente en este lenguaje. Además, Visual Studio en su versión gratuita cuenta con opciones de desarrollo muy básicas, aprovechadas principalmente por programadores principiantes, quedándose corto en muchas funcionalidades respecto a RStudio.

Conclusiones parciales

Luego de realizar un análisis del objeto de estudio de la investigación, los antecedentes y las principales tendencias tecnológicas a considerar, se arriba a las conclusiones siguientes:

La revisión bibliográfica realizada permite ratificar que la gestión ambiental es una necesidad real, y que las empresas deben adoptar y evaluar sistemas y mecanismos de gestión ambiental en pos de alcanzar un desarrollo sostenible, y enfatiza en la importancia de la evaluación ambiental mediante indicadores.

Dada la naturaleza contaminante de las materias primas y productos asociados a las empresas pertenecientes a la industria petrolera, como es el caso de la Empresa Comercializadora de Combustibles de Matanzas, se hace imprescindible prestar especial atención en su gestión medioambiental y por tanto establecer políticas, objetivos y programas en materia de gestión ambiental que ayuden a minimizar los daños que pudieran ocasionarse al medio ambiente.

La falta de datos y la incierta calidad de otros perjudican gravemente la evaluación ambiental integrada en los planos regional y mundial. La infraestructura de adquisición de datos y monitoreo de procesos en la mayor parte de los países en desarrollo tropieza con graves dificultades o no existe en absoluto debido a la limitación de recursos, personal y equipo.

Se definieron las tecnologías que mejor se ajustan a los requerimientos del problema detectado, comprobándose que el lenguaje R y el entorno de desarrollo RStudio, son las herramientas adecuadas para darle solución al mismo.

Capítulo 2. Caracterización de la empresa y propuesta de procedimiento para la Evaluación de la Gestión Ambiental Empresarial

La información es la columna vertebral y epicentro organizativo para la toma de decisiones organizacionales de cualquier entidad, la calidad de los datos, y el conocimiento que de allí se explote, guarda una importancia sin precedentes para el tratamiento de los datos en las instituciones, de ahí el valor de emplear modelos artificiales en el análisis de estos grandes volúmenes de datos , pues brindan un grupo de técnicas y herramientas informáticas para la extracción adecuada del conocimiento que encierran los mismos. Este hecho ha transformado el análisis de datos, orientándolos hacia determinadas técnicas especializadas, las cuales se encuentran englobadas bajo el nombre de minería de datos. La minería de datos hace uso del aprendizaje automático y tiene como objetivo utilizar datos y experiencias pasadas para resolver un problema que se plantee en la actualidad. Para ello se lleva a cabo un proceso de aprendizaje sobre un conjunto de datos, cuya clase ya se conoce (conjunto de entrenamiento), permitiendo así generar un modelo en base de relaciones, patrones o reglas, para poder clasificar nuevos elementos. La calidad de los datos es vital, puesto que de ellos depende el correcto análisis para obtener así una información útil.

2.1. Caracterización de la Empresa Comercializadora de Combustibles Matanzas

La UEB División Territorial de Comercialización de Combustibles Matanzas se encuentra ubicada en el kilómetro 4.3, Carretera Zona Industrial, Versalles, Matanzas. La misma tiene como misión la recepción, comercialización y almacenamiento de combustibles garantizan cantidad, plazo, calidad y uso racional. Sus principales mercados son las Termoeléctricas, la Unión del Cemento, la generación de electricidad por Grupos Electrógenos de Diésel y Petróleo Combustible, la Cadena de Servicentros, el Níquel y la red minorista.

En la **figura 2.1** se puede apreciar la fotografía donde se muestra la ubicación de la empresa Comercializadora de Combustibles, Matanzas.

Figura 2.1. Fotografía de la ubicación de la empresa Comercializadora de Combustibles, Matanzas.
Fuente: Empresa Comercializadora de Combustibles, Matanzas.

El **objeto social** de la empresa es comercializar hidrocarburos y sus derivados.

Misión: comercializar y brindar servicios especializados asociados al combustible y sus derivados en el territorio nacional, con estándares de calidad certificados y un capital humano calificado, con sentido de pertenencia, que asegure la competitividad, seguridad ambiental y satisfacción para los clientes.

Visión: ser reconocidos por la excelencia en la comercialización de combustibles y sus derivados mediante la implementación y mejora continua del Sistema Integrado de Gestión Empresarial, logran una posición innovadora con un eficiente trabajo en equipo que supere las expectativas de los clientes.

La UEB tiene la particularidad de ser la única del país donde se utilizan todas las vías de comercialización, la vía marítima, oleoductos y transporte terrestre (camiones cisternas y tanques de ferrocarril), ya sean a usuarios de la provincia como del resto del país. De estas vías, la más importante es la marítima pues a través de ella se comercializan los mayores volúmenes de combustibles. El 68 % del combustible que sale lo hace por esta vía.

La UEB cuenta con cinco áreas operacionales:

- Terminal 320 (T-320): en esta área se operan básicamente los productos blancos, y se utilizan los carros cisternas y el ferrocarril.

- Base de Crudo y Suministro (BCS): en la misma convergen los oleoductos de las principales productoras de crudo del país, se operan generalmente los productos negros. Además, se realizan mezclas para mejorar el crudo nativo.
- Base en Tierra (BT): es el área que abarca más instalaciones, porque a su vez está dividida en varias secciones. El objetivo de la misma es recibir, almacenar, y entregar productos combustibles como el petróleo crudo nativo, nafta y fuel oil.
- Planta Caribe o Planta de Gas Licuado del Petróleo (GLP): Se dedica a la recepción del gas, el llenado y distribución de botellones a toda la provincia, usa las vías terrestres por camiones cisternas y marítima.
- Área de los Muelles: cuenta con cinco muelles, entre ellos el más grande del país, donde pueden atracar buques de hasta 150 000 toneladas de peso muerto.

Esta instalación cuenta además con un laboratorio para el control de calidad en las operaciones, con un alcance de 20 ensayos acreditados, un taller para realizar operaciones de mantenimiento y una planta de tratamiento de residuales con parámetros de vertimiento seguros.

2.2. Premisas

Es usual que, dado un conjunto de datos, muchas veces se quiere saber cómo clasificar los datos nuevos en base a datos que ya se tienen. Sería interesante poder entrenar un modelo que dado unos análisis nuevos alerta sobre el comportamiento futuro de los mismos. 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. A esta área de investigación se le conoce como clasificación no supervisada, análisis de conglomerados, análisis clúster, o simplemente *clustering*.

A pesar de existir un 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. Por lo antes expuesto este trabajo se centra en la creación de un procedimiento de clasificación de indicadores donde se aplican técnicas de minería de datos para obtener mayor información de los mismos, de forma tal que el procedimiento, a partir del aprendizaje previo de estos, pueda definir qué tan relacionados están y la forma en la que influyen estos grupos en el resultado final, convirtiéndose el procedimiento en una herramienta de apoyo en la empresa para el proceso de toma de decisiones, ya sea tanto para la optimización y mejora de la producción como para minimizar el impacto de sus actividades en el medio ambiente.

Especial atención se pondrá en utilizar estas herramientas y recursos informáticos con un enfoque ambiental, que faciliten la labor de empresas e instituciones de manera que puedan alcanzar un desarrollo sostenible en equilibrio con el medio ambiente. Hay que atenerse al hecho evidente de que el avance incesante de las tecnologías no parece tener freno, el reto de estos centros radica en prepararse como institución y preparar a sus trabajadores para 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.

Las tecnologías en uso no garantizan con su sola presencia el éxito empresarial, es necesario diseñar con mucho cuidado donde será utilizada. Una de las mayores dificultades a vencer para la introducción y la utilización mismas en las organizaciones radica por lo general en la resistencia a los

cambios, a la hora de adaptarse y enfrentar los nuevos retos. Es necesario que en el ámbito empresarial se gane conciencia en que los empleos de estos medios no tradicionales imponen marcadas transformaciones en la configuración del proceso productivo, con cambios en los roles que han desempeñado los diversos actores del mismo.

2.3. Elaboración del procedimiento

Para la medición de resultados de la gestión empresarial a través de indicadores, se presenta el siguiente procedimiento basado en la aplicación de técnicas de minería de datos. Compuesto por fases con métodos asociados para cumplimentarlas. Según se muestra en la tabla 2.1.

El objetivo es desarrollar un procedimiento de *Machine Learning* para la clasificación de indicadores ambientales y la creación de indicadores sintéticos para la gestión ambiental en las instituciones, que brinde información oportuna a los diferentes niveles de dirección de la organización, y garantizar la elevación de la eficiencia y eficacia de la gestión empresarial.

Para alcanzar este objetivo es necesario transitar por las siguientes fases:

1. Identificar el conjunto de datos para el entrenamiento, validación y pruebas, para realizarles el preprocesamiento.
2. Establecer cuáles son las técnicas de *Machine Learning* que se pueden usar para realizar la clasificación de los indicadores.
3. Construir un conjunto de datos para entrenamiento, validación y prueba que pueda ser utilizado para determinar la técnica de *Machine Learning* a utilizar en el clasificador.
4. Construir, entrenar y validar los clasificadores basados en los procedimientos identificados, y escoger el que ofrezca los mejores resultados.
5. Probar el clasificador escogido con nuevos datos que permitan verificar la exactitud del mismo.

Tabla 2.1. Procedimiento para la medición del resultado de la gestión empresarial.

Fases	Objetivos	Descripción	Métodos
1	Identificar el conjunto de datos para el entrenamiento, validación y pruebas, para realizarles el preprocesamiento.	Selección del conjunto de datos, tanto en lo que se refiere 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), como posiblemente al muestreo de los registros disponibles.	HoldOut
2	Establecer cuáles son las técnicas de *Machine Learning* que se pueden usar para realizar la clasificación de los indicadores.	Selección de las técnicas de minería de datos, se diseña el procedimiento predictivo o de clasificación.	Análisis de Componentes Principales (PCA) Árboles de Decisión *Random Forest* Receiver Operating Characteristic (ROC)
3	Construir un conjunto de datos para entrenamiento, validación y prueba que pueda ser utilizado para determinar la técnica de *Machine Learning* a utilizar en el clasificador.	Análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos). Transformación del conjunto de datos de entrada, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema.	Series Temporales Descomposición estacional Dickey-Fuller aumentada Kwiatkowski-Phillips-Schmidt-Shin Phillips-Perron
4	Construir, entrenar	Extracción de conocimiento,	Clústeres

	y validar los clasificadores basados en los procedimientos identificados, y escoger el que ofrezca los mejores resultados.	mediante una técnica de minería de datos, se obtiene un procedimiento de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos procedimientos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.	Árbol de decisión Conjunto de reglas
5	Probar el clasificador escogido con nuevos datos que permitan verificar la exactitud del mismo.	Interpretación y evaluación de datos, una vez obtenido el procedimiento, se debe proceder a su validación y comprobar que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios procedimientos mediante el uso de distintas técnicas, se deben comparar los procedimientos en busca de aquel que se ajuste mejor al problema.	Test de Validación

Fuente: elaboración propia.

Si el procedimiento final no superara esta evaluación el proceso se podría repetir desde el principio o, si el experto lo considera oportuno, a partir de cualquiera de los pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un procedimiento válido. Una vez validado el procedimiento, si resulta ser aceptable (proporciona salidas adecuadas y/o con márgenes de error admisibles) éste ya está listo para su explotación. Los procedimientos obtenidos por técnicas de minería de datos se

aplican incorporándolos en los sistemas de análisis de información de las organizaciones. La **Ilustración 2.1.** Esquema general de comportamiento del procedimiento propuesto, basado en técnicas de minería de datos.

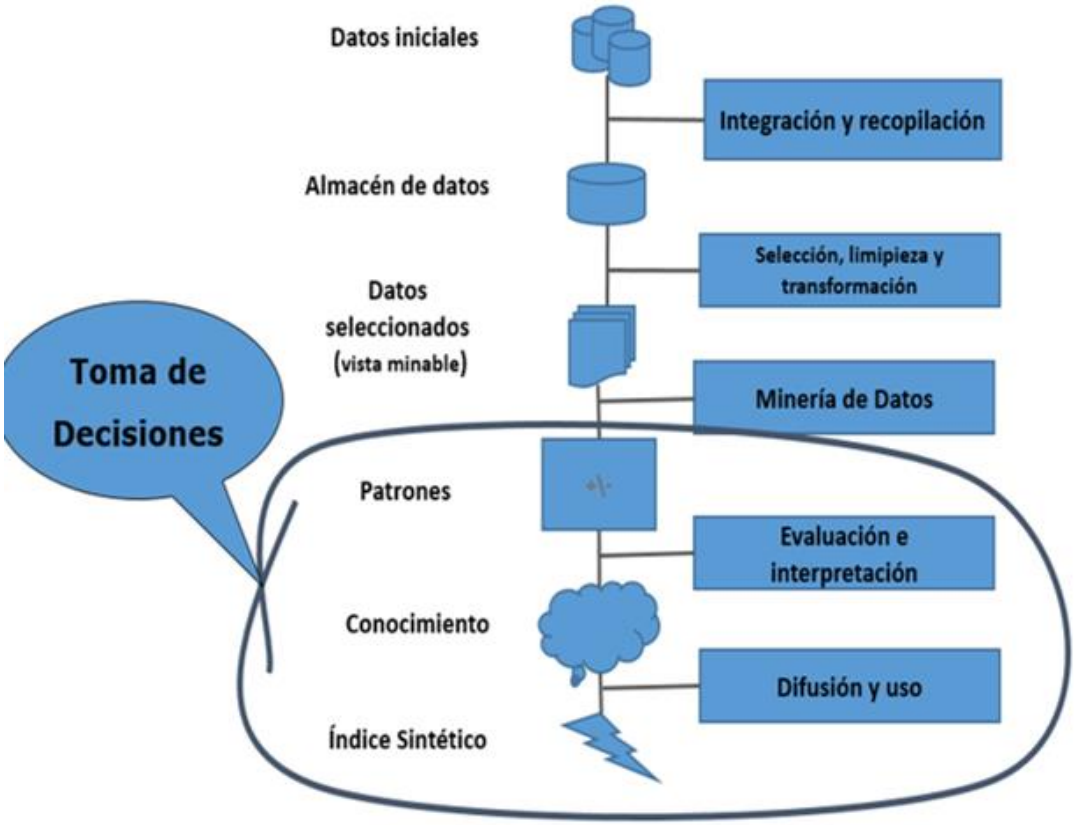

Ilustración 2.1. Esquema general de comportamiento del procedimiento propuesto, basado en técnicas de minería de datos.
Fuente: RStudio (2018).

2.3.1. Descripción del procedimiento

Conjunto de datos para el entrenamiento, validación y pruebas. Preprocesamiento

La recopilación y el preprocesamiento en *Machine Learning*, es un paso difícil con una influencia absoluta en cómo será de adecuado el procedimiento y su desempeño. Cuanto más y mejores datos obtengamos, mejor será el rendimiento del procedimiento, lo que en ocasiones es un punto crítico cuando se construyen los mismos. Las fuentes de los datos pueden ser diversas; archivos csv, hojas de cálculo excel, páginas webs, bases de datos entre otros, por tanto, hay disímiles técnicas para su recopilación.

Para comenzar a entrenar los procedimientos es necesario transformar los datos partiendo con la normalización de sus atributos de forma tal que sus valores estén en un rango entre 0 y 1. Los datos serán transformados a una estructura *data frame*para hacer posible su manipulación en el lenguaje R.

Una vez realizado el preprocesamiento de los datos quedan normalizados y en formato *data frame* (Ilustración 2.2. Datos normalizados (Primeras 20 observaciones), con estructura data frame), para su uso en la construcción del procedimiento. Es importante dividir los datos en varios conjuntos, siguiendo las buenas prácticas registradas por la literatura (Notas de la clase "Dimensionality Reduction - Advice for applying Principal Component Analysis", 2010) y (Assignment No. 3 del curso STA 414/2104: Statistical Methods for Machine Learning and Data Mining,, 2014), lo que se conoce como método "*holdout*" 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, ajustar sus parámetros con los datos de validación, y finalmente evaluar su rendimiento con el conjunto de datos de prueba que hemos dejado aparte.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	0.6410256	0.4358974	0.1666667	0.01282051
2	0.6153846	0.3717949	0.1666667	0.01282051
3	0.5897436	0.3974359	0.1538462	0.01282051
4	0.5769231	0.3846154	0.1794872	0.01282051
5	0.6282051	0.4487179	0.1666667	0.01282051
6	0.6794872	0.4871795	0.2051282	0.03846154
7	0.5769231	0.4230769	0.1666667	0.02564103
8	0.6282051	0.4230769	0.1794872	0.01282051
9	0.5512821	0.3589744	0.1666667	0.01282051
10	0.6153846	0.3846154	0.1794872	0.00000000
11	0.6794872	0.4615385	0.1794872	0.01282051
12	0.6025641	0.4230769	0.1923077	0.01282051
13	0.6025641	0.3717949	0.1666667	0.00000000
14	0.5384615	0.3717949	0.1282051	0.00000000
15	0.7307692	0.5000000	0.1410256	0.01282051
16	0.7179487	0.5512821	0.1794872	0.03846154
17	0.6794872	0.4871795	0.1538462	0.03846154
18	0.6410256	0.4358974	0.1666667	0.02564103
19	0.7179487	0.4743590	0.2051282	0.02564103
20	0.6410256	0.4743590	0.1794872	0.02564103

Ilustración 2.2. Datos normalizados (Primeras 20 observaciones), con estructura data frame.
Fuente: RStudio (2018).

.

Técnicas de Machine Learning

Análisis de Componentes Principales (PCA)

El análisis de componentes principales permite reducir la dimensionalidad de un conjunto de datos, transformar el conjunto de variables originales en otro conjunto de variables correlacionadas llamadas componentes principales. Se propone tomar el 70% de los datos para el proceso de entrenamiento y validación, y el 30% restante reservarlo exclusivamente para pruebas. Para el cálculo de los componentes principales se empleará únicamente el conjunto de entrenamiento, a partir de esto se define la matriz de transformación que posteriormente será aplicada a los datos de prueba. Esta técnica se aplica para evitar el sobreajuste del procedimiento, pues PCA busca las correlaciones entre características, donde esta correlación implica que hay redundancia en los datos.

Árboles de decisión

Los árboles de decisión son una serie de decisiones o 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 (Sutton Charani, Destercke y Denoeux, 2013), de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta algunas de sus hojas permitiendo una fácil interpretación (Santín González y Pérez López, 2006). Los árboles de decisión que se usan para predecir variables categóricas se llaman árboles de clasificación, mientras que los árboles de decisión que se utilizan para predecir variables continuas se llaman árboles de regresión (Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos, 2007).

RandomForest

RandomForest es una técnica que combina una cantidad grande de árboles de decisión independientes probados sobre conjuntos de datos aleatorios con igual distribución (*QuanDare*, 2019). La fase de aprendizaje consiste en crear muchos árboles de decisión independientes, construyéndolos a partir de datos

de entrada ligeramente distintos. Se altera, por tanto, el conjunto inicial de partida, haciendo lo siguiente:

- Se selecciona aleatoriamente con reemplazamiento un porcentaje de datos de la muestra total. Es habitual incluir un segundo nivel aleatoriedad, esta vez se afectan los atributos.
- En cada nodo, al seleccionar la partición óptima, se tiene en cuenta sólo una porción de los atributos, elegidos al azar en cada ocasión. Una vez que se generan muchos árboles, la fase de clasificación se lleva a cabo de la forma siguiente.

Cada árbol se evalúa de forma independiente y la predicción del bosque será la media de todos sus árboles en caso de que sea un problema de regresión, cuando se trate de un problema de clasificación realizara un voto mayoritario sobre todos los arboles del bosque es decir la clase con mayor voto (*Breiman*, 2019).

ROC (Receiver OperatingCharacteristic)

La curva ROC es una 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 se conoce como sensibilidad, sería la probabilidad de clasificar correctamente a un individuo cuyo estado real sea definido como positivo. La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea clasificado como negativo. Esto es igual a restar uno de la fracción de falsos positivos.

La curva ROC también es conocida como la representación de sensibilidad, cada resultado de predicción representa un punto en el espacio ROC. El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representan un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). Una clasificación totalmente aleatoria daría un punto a lo largo de la línea diagonal, que se llama también línea de no-discriminación. En definitiva, se considera un procedimiento inútil, cuando la curva ROC recorre la diagonal positiva del gráfico.

Construcción del conjunto de datos

Carga de datos en R

```
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,])
}
```

El fragmento de código muestra como se establece la conexión a la base de datos y las consultas SQL (*StructureQueryLanguage*) pertinentes para la obtención de los datos solicitados.

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).

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 queremos cargar sus mediciones.

Con la base de datos cargada correctamente en el programa se puede pasar a la transformación de los datos en una serie temporal.

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 extendemos 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 (RStudio, 2017).

Para el empleo de series temporales en R se empleará la librería *tseries*, la cual permite la creación de series temporales y cuenta con una gran cantidad de pruebas y funciones estadísticas que se utilizan para el análisis y estudio de la serie temporal siendo necesarias para la implementación de un procedimiento predictivo.

```
library(tseries)
```

La función *ts* se empleará 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, se supone que cada columna de los datos de la

matriz contiene una única serie de tiempo (univariante). Las series de tiempo deben tener al menos una observación, y aunque no necesitan ser numéricas, hay un soporte muy limitado para las series no numéricas. El valor de la frecuencia del argumento se usa cuando la serie se muestrea un número entero de veces en cada intervalo de unidad de tiempo. Por ejemplo, uno podría usar un valor de 7 para la frecuencia cuando los datos se muestrean diariamente, y el período de tiempo natural es una semana, o 12 cuando los datos se muestrean mensualmente y el período de tiempo natural es un año. Se supone que los valores de 4 y 12 en (por ejemplo) métodos de impresión implican una serie trimestral y mensual, respectivamente (Becker, Chambers y Wilks, 1988).

Para la transformación de los datos se construirá la función serie Temporal (datos, frecuencia), los parámetros de esta función son los datos que se desean transformar y la frecuencia con la que sen medidos los datos.

```
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))
 }
```

En esta función se ordenan las mediciones cronológicamente de manera ascendente y se toma el año de la primera medición para que dé 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 calculará la fecha final a partir de la inicial.

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 emplea 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 usará *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 dependiendo del tipo de dato que le demos como argumento, generará diferentes tipos de gráfica. Además, para cada tipo de gráfico, podremos 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 especificamos este argumento, obtendremos un error y no se creará una gráfica. El resto de los argumentos de *plot()* son opcionales, pero el más importante es y. Este argumento también requiere un vector y corresponde al eje Y de nuestra gráfica.

```
plot(ts_temp)
```

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,…,Yt+k) es la función de distribución acumulada de probabilidad (Parra, 2019)

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

Sin embargo, la versión estricta de la estacionalidad de un proceso suele ser excesivamente restrictiva para las necesidades prácticas. Es por ello que generalmente se conforma con un concepto menos exigente, el de estacionalidad en sentido débil o de segundo orden, la cual 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 (Parra, 2019).

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. A simple vista podemos observar que la serie no es estacionaria en media.

Para el análisis de las series temporales se construirán las funciones estacionalidad(serie, frecuencia) y estacionalidad(serie, frecuencia). Estas devuelven TRUE en caso de que la serie sea estacional o estacionaria respectivamente y FALSE en caso contrario. Estas funciones contienen las principales pruebas que se realizan a las series para conocer su composición.

```
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)
}
```

Método de descomposición

Los métodos de descomposición estacional son eminentemente descriptivos. Tratan de 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 es necesario combinarlos con métodos de ajuste de tendencia. De esta forma realizaremos un ajuste de tendencia con el fin de obtener un procedimiento extrapolable, y le añadiremos la estacionalidad.

El primer paso a seguir a la hora de descomponer una serie es determinar cómo se combinan sus componentes. Las combinaciones aditiva y multiplicativa son las más habituales. Decimos que estamos en 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:

```
Componentesaditivas: Xt =Ct +Tt +St +Et
Componentes multiplicativas: Xt =Ct ×Tt ×St ×Et
```

Donde:

Tendencia (Tt),
Ciclo (Ct),
Componente estacional (St),
Componente irregular o ruido (Et).

R cuenta con la librería *forecast* y esta a su vez con un método *decompose* que permite la descomposición del gráfico para su análisis visual, para la aplicación de este método es necesario que los datos tengan una frecuencia mínima dos.

En caso de que la frecuencia de los datos sea menor que dos, al análisis debe realizarse por el método matemático. Para este objetivo es necesario realizar pruebas de presencia de raíz unitaria, dado que en caso afirmativo esto implicaría la no estacionariedad y viceversa.

Según las fuentes consultadas, las pruebas más utilizadas para este fin son:

- 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, es un número negativo. Cuanto más negativo es, más fuerte es el rechazo de la hipótesis nula de que existe una raíz unitaria para un cierto nivel de confianza. La librería tseries cuenta con esta prueba estadística cuya función se llama *adf.test*.

```
adf.test(ts_temp)
```

- Prueba de *Kwiatkowski-Phillips-Schmidt-Shin*. Su hipótesis nula es que no posee raíz unitaria. Esta función lleva el nombre de *kpss.test.*

```
kpss.test(ts_temp)
```

- 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. El nombre de esta función en *tseries* es *pp.test().*

```
pp.test(ts_temp)
```

Construcción del clasificador

Para crear un clasificador, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa 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 aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

El procedimiento de minería de datos que crea un algoritmo a partir de los datos puede tomar diversas formas, incluyendo:

- Un conjunto de clústeres que describe cómo se relacionan los casos de un conjunto de datos.
- Un árbol de decisión que predice un resultado y que describe cómo afectan a este los distintos criterios.
- Un conjunto de reglas que describen cómo se agrupan los datos.

Una vez obtenido el conjunto de datos apto para iniciar el proceso de selección de la técnica de *Machine Learning* que se usará para desarrollar el Clasificador de indicadores, objeto de esta investigación, se propone realizar un análisis teniendo en cuenta el número de variables y el número de ejemplos recolectados.

Cuando el conjunto de datos tiene 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 podría llevar a un *overfiting* (sobreajuste) y también se puede correr el riesgo de tener atributos ruidosos que pueden tener el mismo peso que los atributos relevantes. En caso de trabajar con datos de alta dimensionalidad se aplicarán 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.

Entrenamiento, Validación y Prueba

Una vez realizada la lectura y particionamiento de los datos, estos son sometidos a un proceso de entrenamiento, validación y prueba. La ***Ilustración 2.3****. Resultado de aplicar este proceso a un conjunto de datos "X"*, muestra el resultado de aplicar este proceso a un conjunto de datos "X", compuesto por cinco variables.

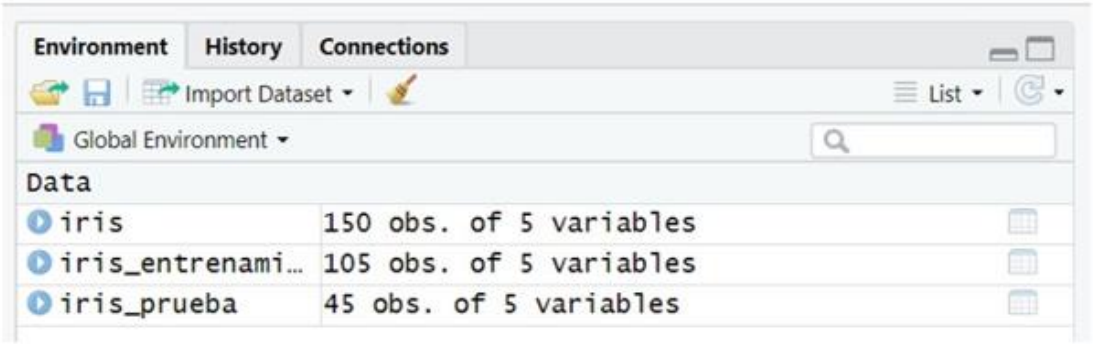

Ilustración 2.3. Resultado de aplicar este proceso a un conjunto de datos “X”.
Fuente: RStudio (2018).

El primero de los posibles procedimientos de clasificación estará basado en los árboles de decisión. Para ello cargamos la librería *rpart* y le indicamos que deseamos crear un procedimiento a partir de la función *rpart*de dicha librería donde se declara de los atributos como variable objetivo delprocedimiento.

Como se muestra en la **Ilustración 2.4.** a) Árbol de decisión generado. b) Gráfico del árbol de *decisión*, se obtuvo un árbol donde el nodo raíz nos indica que tiene 105 ítems. En la primera bifurcación mira que la longitud del pétalo sea menor que 2.5 y dejar a ese lado 36 casos y al otro 69. El árbol sigue ramificándose, y así hasta llegar a los nodos hojas, que son marcados con asteriscos.

a)

```
n= 105

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 105 69 setosa (0.34285714 0.32380952 0.33333333)
  2) Petal.Length< 2.5 36  0 setosa (1.00000000 0.00000000 0.00000000) *
  3) Petal.Length>=2.5 69 34 virginica (0.00000000 0.49275362 0.50724638)
    6) Petal.Width< 1.65 37  3 versicolor (0.00000000 0.91891892 0.08108108) *
    7) Petal.Width>=1.65 32  0 virginica (0.00000000 0.00000000 1.00000000) *
```

b)

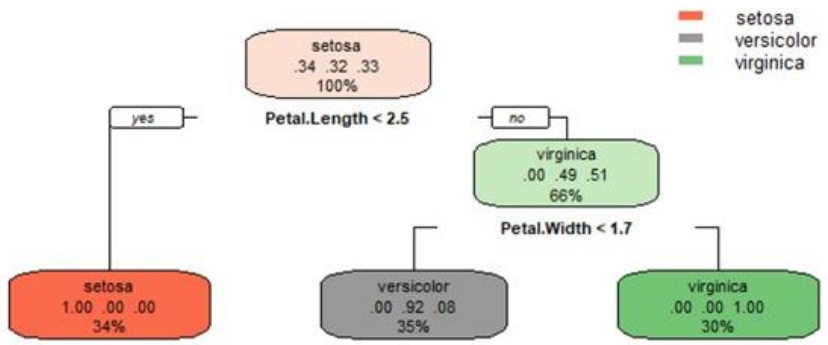

Ilustración 2.4. a) Árbol de decisión generado. b) Gráfico del árbol de decisión.
Fuente: RStudio (2018)..

Con el set de prueba se genera un vector con los valores predichos por el procedimiento entrenado. Se cruza la predicción con los datos reales del set de prueba para generar una matriz de confusión (

Ilustración 2.5. Matriz de confusión.

```
Confusion Matrix and Statistics

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

Ilustración 2.5. Matriz de confusión.
Fuente: RStudio (2018).

Para la construcción de otro posible procedimiento se utilizará el algoritmo *randomForest*, llamado de bosque aleatorio, una vez aplicada esta técnica se analizará la precisión de la misma.

Las predicciones de ambos algoritmos son bastantes similares para el mismo conjunto de datos. Se realizaron numerosas pruebas de validación al procedimiento propuesto, tal y como se muestra en el **Anexo 2. Validaciones realizadas al procedimiento a partir de *DataSets* aleatorios.**

Conclusiones parciales

Luego de describir el diseño de la solución propuesta al problema científico de esta investigación se concluye que:

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.

El hecho de que las clasificaciones de los indicadores sean muy cercanas a la realidad permite emitir criterios acertados para evaluar la situación de la empresa y en la toma de decisiones en su gestión ambiental, abriendo un mayor espectro para su uso a partir de sus propiedades estadísticas.

Se demuestra 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.

Conclusiones generales

Como resultado de esta investigación se dio cumplimiento a los objetivos trazados y permite arribar a las conclusiones siguientes:

1. El estudio realizado sobre los antecedentes, el estado actual de la temática, la bibliografía y documentos relacionados con el objeto de estudio, permitió contar con los elementos necesarios para dar solución a la problemática planteada.
2. 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.
3. El hecho de que las clasificaciones de los indicadores sean muy cercanas a la realidad permite emitir criterios acertados para evaluar la situación de la empresa y en la toma de decisiones en su gestión ambiental, abriendo un mayor espectro para su uso a partir de sus propiedades estadísticas.
4. Como resultado del procedimiento se evalúan los resultados de la gestión ambiental empresarial a través de indicadores lo que tributa al proceso de toma de decisiones de la entidad, y permite conocer el comportamiento que debe suceder en el futuro para condiciones similares concretas.
5. A pesar de existir un 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.

Recomendaciones

Desde el punto de vista del alcance del presente trabajo y teniendo en cuenta el tiempo para el desarrollo del mismo, se proponen las recomendaciones siguientes:

1. Evaluar la Gestión Ambiental en la Empresa Comercializadora de Combustibles de Matanzas mediante el procedimiento propuesto para formular un plan de acciones que permita la mejora continua de la gestión ambiental de la entidad.
2. Validar los resultados de la aplicación del procedimiento a partir del criterio de especialistas.
3. Realizar los estudios de adaptación en empresas pilotos de la provincia.

Referencias bibliográficas

Acosta, Alberto. (2017). *La Pequeña Empresa.* Rio de Janeiro: Direito e Práxis, Vol. 8.

Agencia Ambiental, Europea. (2002). Agencia Ambiental Europea. Retrieved: enero 18 de 2020 from: https://www.eea.europa.eu/es.

Alonso, S.; Duarte, C. y Montes, C. (2006). *Cambio global impacto de la actividad humana sobre el sistema tierra. Ed. CSIC y Cataratas, ISBN: 978-84-00-08452-3. 167p.* Madrid: Edición a cargo de Cyan, Proyectos y Producciones Editoriales, S.A.

Becker, R. A.; Chambers, J. M. y Wilks, A. R. (1988). *The New S Language.* Wadsworth & Brooks/Cole.ISBN: 978-o-534-09192-7, 702p.

Beltrán, R. (2008). *Complejidad de Modelos: Sesgo y Varianza. Notas de clases. 5p.*

Breiman, L. (2019). University of California, Berkeley - RANDOM FORESTS. [En línea] 26 de noviembre de 2019. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf..

Briggs, C A; Tolliver, D y Szmerekovsky, J. (2012). *Managing and mitigating the upstream petroleum industry supply chain risks: leveraging analytic hierarchy process.* International Journal of Business and Economics Perspectives,. Vol. VII.

Cabrera Hernández, Juan Alfredo**.** (2004). *Generalidades sobre el Medio Ambiente. Apuntes para un curso.* Cuba: Universidad de Matanzas, 2004.

Cabrera Guerra, Ricardo. (2011). *Gestión ambiental y salud en la provincia de Ciudad de La Habana.* La Habana.

Castro Felicori, Francisco Osvaldo. (2019). Proyecto de Ley. Texto de Nueva Constitución de la República de Cuba, aprobado en el 22 de diciembre de 2018, y que será sometida a referendo popular para su ratificación el próximo 24 de febrero de 2019. *Boletín Jurídico del Observatorio de Libertad Religiosa de América Latina y El Caribe.*

Castro Ruz, Fidel. (1992). *Cumbre de las naciones Unidas sobre el Medio Ambiente y el Desarrollo.*

CC-PCC. (2017). Lineamientos de la Política Económica y Social del Partido. *Documentos del 7mo. Congreso del Partido aprobados por el III Pleno del Comité Central del PCC el 18 de mayo de 2017 y respaldados por la Asamblea Nacional del Poder Popular el 1 de junio de 2017. Lineamientos de la Política Económica y Social del Partido.* La Habana: Tabloide, p.38.

Comité Ejecutivo del Consejo de Ministros. (2013). Decreto No. 281/2013: *Reglamento para la implantación y consolidación del Sistema de Dirección y Gestión Empresarial Estatal.* La Habana, Cuba: Gaceta Oficial No. 007 Ordinaria de 18 de febrero de 2013. pp. 312-315.

Cousera. (2010). *Notas de la clase "Dimensionality Reduction - Advice for applying Principal Component Analysis". C*urso *Machine Learning Stanford*.

Córdoba Durán, Verónica. (2016). *Implementación del proceso de gestión ambiental en la exploración de bloques evaluados para extracción de recursos minerales (gas natural y petróleo). Trabajo de Grado para Optar al Título de Ingeniero Geólogo.* Universidad Pedagógica y Tecnológica de Colombia.

Díaz Balteiro, L. y Romero, C. (2003). *La Valoración del Desarrollo Sostenible: Una Propuesta Metodológica.* Sevilla: Ecológica, Medio Ambiente,

Ebert, U. (1994). *The measurement of scientific and technological activities using patent data as science and technology indicators.* París.

eGAM. (2013). eGAM. Retrieved: febreo 01 de 2020 from http://www.egambpm.com/wiki/index.php?title=EGAM_Ambiental_ISO14001:_Documentaci%C3%B3n_oficial.

Estrada Latorre, Emilio. (2000). *Herramientas para la Participación en Gestión Ambiental.* Bogotá: Editorial Prisma Asociados Ltda.

Fernández, J. M. (2003). *Boosting Con Redes Neuronales RBF. Análisis Sesgo - Varianza en un problema de Clasificación.* VI Congreso Galego de Estatística e Investigación de Operación. Universidad de Vigo.

Fernández Debill, José. (2016). Canarina Software Ambiental. Retrieved: enero 08 de 2020 from https://zayriduran.wixsite.com

Freitas, A. A. (2002). *Data Mining and Knowledge Discovery With Evolutionary alghorithms.*

García Cespedes, Damarys. (2014*). Proposed methodology of environmental management for agro ecosystems health risks for chemical contamination*. Retrieved: febreo 01 de 2020 from http://scielo.sld.cu/scielo.php.

Ghul, Pablo y Leyva, Ernesto. (2015)**.** *La Gestión Ambiental en Colombia, 1994-2014:¿un esfuerzo sostenible?.* Colombia: Friedrich-Ebert-Stiftung. Primera edición.

González, Hugo. (2012). Indicadores de gestion ambiental. Retrieved: enero 14 de 2020 from Indicadores de gestion ambienta l https://calidadgestion.wordpress.com.

Grupo Banco Mundial. (2000). Banco Mundial. Retrieved: enero 18 de 2020 from https://datos.bancomundial.org/.

Machín Hernández, María Mercedes y Vazquez Santisteban, Mayelín. (2003). Desafíos y oportunidades de la gestión ambiental en el ámbito empresarial . Retrieved: febrero 02 de 2020 from http://www.monografias.com/trabajos15/gestion-ambiental.

Maimon, Oded Z. y Rokach, Lior. (2010). *Data Mining and Knowledge Discovery Handbook.* New York : Retrieved: febrero 05 de 2020 from https://link.springer.com
Parra, F. J. (2019). *Estadística y Machine Learning con R.* Bookdown. Ed. EAE. ISBN: 978-6202252164.

Martínez Cohen, Rafael**.** (2018). Tecnologías de Información. Retrieved: febrero 02 de 2020. From Tecnologías de Información: www.tecnologias-informacion.com.

Martínez Hals, Alberto. (2015). Aplicaciones para la gestion ambiental. Retrieved: febrero 2 de 2020 *from Aplicaciones para la gestion ambiental:* www.eco2biz.com.

Martínez Quiroga, Rayén. (2009). *Guía metodológica para desarrollar indicadores ambientales y de desarrollo sostenible en países de América Latina y el Caribe.* Santiago de Chile.

Moreno García, M. (2007). *Modelos de Estimación Software basado en técnicas de Aprendizaje Automático. Editor J. Tuya. Pp. 109-126.*

Múnera Espinal, Hernán Dario. (2011)**.** Indicdores de Gestion ambiental. Retrieved: enero 22 de 2020 from http://app1.semarnat.gob.mx.

Pearce, D. y Turner, R. (1995). *Economía de los Recursos Ambientales y del Medio Ambiente.* España: Ed. Celeste.

Peteiro de Bureau, Verita. (2010). Gestión del Conocimiento: el capital humano como pilar clave para la innovación en la empresa. Retrieved: febrero 02 de 2020 from http://www.catedrainnovacion.es.

Pino Neculqueo, Maria Eliana. (2010). *Los indicadores ambientales como parámetros clave de la sostenibilidad. Trilogía. Ciencia-Tecnología-Sociedad. Vol. 23, núm. 33, ISSN: 0716-03356.*

PNUMA. (1996). Propuesta de Ley de Evaluación de Impacto Ambiental para los países de América Latina y el Caribe, Serie de Documentos sobre Derecho Ambiental No.4.

PNUMA. (2001). *Report on environmental indicators and sustainability in Latin America and the Caribbean.* United Nations Environment Programme

QuanDare. (2019). Artificial Intelligence Random forest vs Simple tree. Retrieved: febrero 02 de 2020 from https://quantdare.com/random-forest-vs-simple-tree/.

Rodríguez, F. (2001). *Los Costos en el Sistema de Gestión medioambiental.* Argentina: IAPUCO.

Roffe, I. (1997). *Developing a dynamic in a learning innovation. Quality Assurance in Education.*

RStudio. (2018). RStudio. Retrieved: febrero 02 de 2020 from https://www.rstudio.com/products/rstudio/.

RStudio. (2017). RPubs. Retrieved: febrero 02 de 2020 from https://rpubs.com/palominoM/series.

Santín González, D. y Pérez López, C. (2006). *Data Mining Soluciones con Enterprise Miner.* ISBN: 978-84-7897-659-9. 576p.

Soler del Sol, Alfredo. (1997). Ley No. 81**:** *Del Medio Ambiente. Gaceta Oficial de la Replública de Cuba.* La Habana.

Sutton-Charani, N.; Destercke, S. y Denoeux, T. (2013). *«Learning Decision Trees from Uncertain Data with an Evidential EM Approach».* International Conference on Machine Learning and Applications.

The R Development Core Team. (2009). *The R Reference Manual- Base Package.*

Trujillo Davila, M.A. y Vélez Bedoya, R. (2010). Responsabilidad ambiental como estrategia para la perdurabilidad empresarial. *Revista Universidad & Empresa.* Vol.5, núm. 10, pp.291-308. ISSN: 0124-4639. Universidad del Rosario, Bogotá, Colombia.

Vale CapdevilaI, Rita María y Pérez Silvall, Rosa María. (2016). *Valoración del impacto ambiental en empresas de la industria petrolera.* Santiago de Cuba : Revista Cubana Química. Vol. 28, núm.2, ISSN 2224-5421.

Anexos

Anexo 1. Resumen de las librerías empleadas en RStudio.

Paquetes de R utilizados		
Paquete	**Función**	**Descripción**
	sample_frac ()	Facilita la selección de filas aleatorias de una tabla.
rpart	rpart()	El árbol del paquete R proporciona una reimplementación del árbol.
	rpart.plot ()	Esta función es un front-end simplificado para prp, con solo los argumentos más útiles de esa función y con diferentes valores predeterminados para algunos de los argumentos.
	predict()	Es una función genérica para predicciones a partir de los resultados de varias funciones de ajuste del procedimiento. La función invoca métodos particulares que dependen de la clase del primer argumento.
caret	confusionMatrix ()	Cree una matriz de confusión dado un límite específico.
randomForest	randomForest ()	Implementa el algoritmo de bosque aleatorio de Breiman (basado en el código Fortran original de Breiman y Cutler) para la clasificación y regresión. También se puede usar en modo no supervisado para evaluar las proximidades entre los puntos de datos.
pROC	roc()	Construye una curva ROC y devuelve un objeto "roc", una lista de la clase "roc". Este objeto se puede imprimir, trazar o pasar a las funciones auc, ci, smooth.roc y coords. Además, dos objetos roc se pueden comparar con roc.test.
corrplot	corrplot()	Una visualización gráfica de una matriz de correlación, intervalo de confianza. Se presta mucha atención a los detalles. También puede visualizar una matriz general.

Fuente: elaboración propia.

Anexo 2. Validaciones realizadas al procedimiento a partir de *DataSets* aleatorios.

Para ver cómo de bueno el procedimiento a través de la relación entre sensibilidad y 1-especificidad se aplicó la curva ROC, tal como se puede apreciar en los resultados (

Tabla A.2.Comparación de precisión de los procedimientos), lo que arrojó como resultado que el procedimiento que mejor ajustaba a los datos era el generado por algoritmo de Bosque Aleatorio pues el área bajo la curva posee un valor más cercano a 1. No obstante la**Error! Reference source not found.**, muestra como ambos métodos arrojan resultados similares.

Tabla A.2.Comparación de precisión de los procedimientos.

Algoritmos / Clases	Árbol de decisión		Bosque Aleatorio	
	Precisión por Clases	Precisión Total	Precisión por Clases	Precisión Total
Setosa	100%	93.18%	100%	95.55%
Versicolor	91.96%		93.75%	
Virginica	93.10%		96.67	

Fuente: elaboración propia.

Tabla A.3. Área debajo de la curva (ADC) ROC.

Procedimiento	Valor del área bajo la curva ROC
Árbol de decisión	**0.9861**
Bosque Aleatorio	**0.9375**

Fuente: elaboración propia.

Una vez seleccionado el procedimiento óptimo se generaron los grupos y se agruparon los datos según las clases identificadas (Ilustración A.1., A.2. y A.3.) obteniéndose tres clases para la distribución de los datos, para establecer la relación de las variables independientes se generaron graficas de correlación.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Predicción Arbol de Decisión	Predicción Bosque Aleatorio
1	5.0	3.4	1.5	0.2	setosa	setosa
2	5.4	3.7	1.5	0.2	setosa	setosa
3	4.8	3.4	1.6	0.2	setosa	setosa
4	4.8	3.0	1.4	0.1	setosa	setosa
5	5.4	3.4	1.7	0.2	setosa	setosa
6	4.8	3.4	1.9	0.2	setosa	setosa
7	5.2	3.5	1.5	0.2	setosa	setosa
8	5.2	3.4	1.4	0.2	setosa	setosa
9	4.8	3.1	1.6	0.2	setosa	setosa
10	5.5	4.2	1.4	0.2	setosa	setosa
11	4.9	3.1	1.5	0.2	setosa	setosa
12	5.0	3.2	1.2	0.2	setosa	setosa
13	5.1	3.8	1.9	0.4	setosa	setosa
14	5.1	3.8	1.6	0.2	setosa	setosa

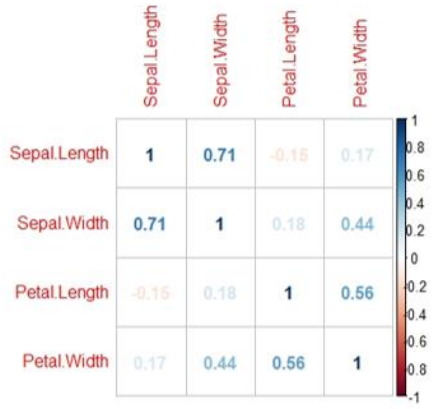

Ilustración A.1. a) Datos Clasificados Grupo1 y Gráfica de correlación de sus atributos
Fuente: elaboración propia.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Predicción Arbol de Decisión	Predicción Bosque Aleatorio
1	5.9	3.2	4.8	1.8	virginica	virginica
2	6.7	3.0	5.0	1.7	virginica	virginica
3	6.5	3.0	5.8	2.2	virginica	virginica
4	7.3	2.9	6.3	1.8	virginica	virginica
5	7.2	3.6	6.1	2.5	virginica	virginica
6	6.4	2.7	5.3	1.9	virginica	virginica
7	6.9	3.2	5.7	2.3	virginica	virginica
8	6.4	2.8	5.6	2.1	virginica	virginica
9	6.4	2.8	5.6	2.2	virginica	virginica
10	7.7	3.0	6.1	2.3	virginica	virginica
11	6.4	3.1	5.5	1.8	virginica	virginica
12	6.0	3.0	4.8	1.8	virginica	virginica
13	6.9	3.1	5.4	2.1	virginica	virginica
14	6.7	3.3	5.7	2.5	virginica	virginica
15	6.5	3.0	5.2	2.0	virginica	virginica

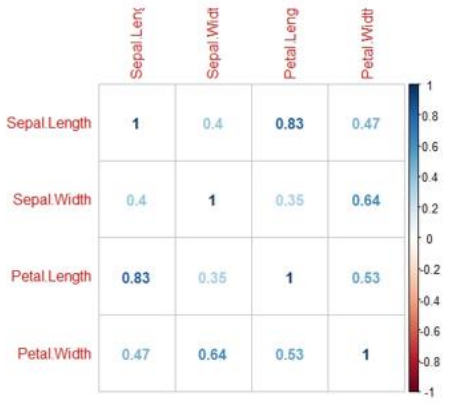

Ilustración A.2. b) Datos Clasificados Grupo 2 y Gráfica de correlación de sus atributos.
Fuente: elaboración propia.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Predicción Arbol de Decisión	Predicción Bosque Aleatorio
1	7.0	3.2	4.7	1.4	versicolor	versicolor
2	6.9	3.1	4.9	1.5	versicolor	versicolor
3	6.3	3.3	4.7	1.6	versicolor	versicolor
4	4.9	2.4	3.3	1.0	versicolor	versicolor
5	5.6	2.9	3.6	1.3	versicolor	versicolor
6	5.8	2.7	4.1	1.0	versicolor	versicolor
7	6.6	3.0	4.4	1.4	versicolor	versicolor
8	6.0	2.9	4.5	1.5	versicolor	versicolor
9	5.7	2.6	3.5	1.0	versicolor	versicolor
10	5.8	2.7	3.9	1.2	versicolor	versicolor
11	6.0	3.4	4.5	1.6	versicolor	versicolor
12	6.1	3.0	4.6	1.4	versicolor	versicolor
13	5.6	2.7	4.2	1.3	versicolor	versicolor
14	5.1	2.5	3.0	1.1	versicolor	versicolor
15	6.3	2.8	5.1	1.5	versicolor	virginica

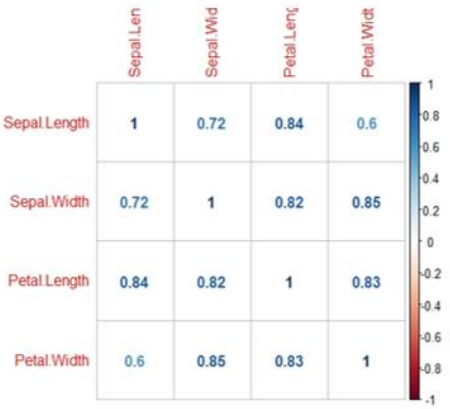

Ilustración A.3. c) Datos Clasificados Grupo 3 y Gráfica de correlación de sus atributos. **Fuente:** elaboración propia.

Anexo 3. Fragmentos de código fuente en R.

Árbol de Clasificación.

```
entrenar_procedimiento<-function (conjuntos, objetivo, predictores
=".") {

if (length (predictores >1)) {

predictores<-paste0(predictores, collapse ="+")

  }

mi_formula<-paste0(objetivo, " ~ ", predictores) %>%as.formula()


arbol<-list ()

arbol[["procedimiento"]] <-

rpart (data = conjuntos [["entrenamiento"]], formula =mi_formula,

control =rpart.control(cp = .01, xval =40, minsplit =2))

arbol[["prediccion"]] <-predict(arbol[["procedimiento"]], conjuntos
[["prueba"]], type ="class")

arbol[["referencia"]] <-conjuntos[["prueba"]][[objetivo]]

arbol

}

matriz_confusion<-function(arbol, objetivo) {

matriz<-list()

matriz<-confusionMatrix(data =arbol[["prediccion"]],

reference =arbol[["referencia"]])

matriz

}

clasificar<-function(datos, objetivo, predictores =".") {

resultado<-list()

resultado[["sets"]] <-dividir_datos(datos)

resultado[["arbol"]] <-entrenar_procedimiento (resultado[["sets"]],
objetivo, predictores)

resultado[["matriz"]] <-matriz_confusion(resultado[["arbol"]],
objetivo)

resultado

}
```

Bosque Aleatorio (*RandomForest*).

```
entrenar_procedimientoBA<-function (conjuntos, objetivo, predictores
=".") {

if (length (predictores >1)) {

predictores<-paste0(predictores, collapse ="+")

  }

mi_formula<-paste0(objetivo, " ~ ", predictores) %>%as.formula()


bosque<-list ()

bosque[["procedimiento"]] <-randomForest (formula= mi_formula, data =
conjuntos[["entrenamiento"]],importance=TRUE,proximity=TRUE, ntree =
200)

bosque[["prediccion"]] <-predict(bosque[["procedimiento"]], conjuntos
[["prueba"]], type ="class")

bosque[["referencia"]] <-conjuntos[["prueba"]][[objetivo]]

bosques

}

clasificar_BA<-function(datos, objetivo, predictores =".") {

resultado<-list()

resultado[["sets"]] <-dividir_datos(datos)

resultado[["bosque"]] <-entrenar_procedimientoBA (resultado[["sets
"]], objetivo, predictores)

resultado[["diagnostico"]] <-matriz_confusion(resultado[["bosque"]],
objetivo)

resultado

}
```