"Responsabilidad con pensamiento positivo"

UNIVERSIDAD TECNOLÓGICA ISRAEL

TRABAJO DE TITULACIÓN

CARRERA: INGENIERIA EN SISTEMAS INFORMÁTICOS

TEMA: Caracterización y modelación de la red de comercialización de combustible automotriz utilizando técnicas de inteligencia artificial.

AUTORA: ROSA MAGDALENA CHICAIZA ACOSTA

TUTOR: MG. JOSÉ RAÚL BARRERAS

AÑO 2013

DEDICATORIA

Dedico este trabajo con todo mi amor y cariño:

A Dios, por la inteligencia, sabiduría y fortaleza que me ha brindado para alcanzar un logro más en mi vida.

A mis hijos, por el amor y la comprensión que me han brindado por dejarlos solos para alcanzar mi meta.

A mi madre por su valioso apoyo, por todas las oraciones que hizo en mi nombre para que concluyera mi carrera.

A Byron por estar a mi lado en todo momento entregándome su apoyo incondicional en todas las etapas de mi carrera.

Magdalena.

AGRADECIMIENTO

Agradezco a las autoridades y profesores de la Universidad Tecnológica Israel por los conocimientos impartidos durante mi formación académica.

Un agradecimiento especial al Mg. José Raúl Barreras quien con sus conocimientos, orientación y dedicación, aportó valiosamente para el desempeño del presente trabajo investigativo.

La Autora.

UNIVERSIDAD TECNOLÓGICA ISRAEL

AUTORÍA DE TESIS

La abajo firmante, en calidad de estudiante de la Carrera de Sistemas Informáticos, declaro que los contenidos de este Trabajo de Titulación, requisito previo para la obtención del Grado de Ingeniera en Sistemas Informáticos, son absolutamente originales, auténticos y de exclusiva responsabilidad legal y académica de la autora.

Quito, diciembre del 2013

.....

Magdalena Chicaiza Acosta
C.C. 1713628137

UNIVERSIDAD TECNOLÓGICA ISRAEL

APROBACIÓN DEL TUTOR

En mi calidad de Tutor de Trabajo de Graduación certifico:

Que el Trabajo de Titulación “CARACTERIZACIÓN Y MODELACIÓN DE LA RED DE COMERCIALIZACIÓN DE COMBUSTIBLE AUTOMOTRIZ UTILIZANDO TÉCNICAS DE INTELIGENCIA ARTIFICIAL.”, presentado por Rosa Magdalena Chicaiza Acosta, estudiante de la carrera de Sistemas Informáticos, reúne los requisitos y meritos suficientes para ser sometido a la evaluación del Tribunal de de Grado, que se designe, para su correspondiente estudio y calificación.

Quito, diciembre del 2013.

TUTOR

……..

Mg. José Raúl Barreras

UNIVERSIDAD TECNOLÓGICA ISRAEL

APROBACIÓN DEL TRIBUNAL

Los miembros del tribunal de grado, aprueban el Trabajo de Titulación de acuerdo con las disposiciones reglamentarias emitidas por la Universidad Tecnológica Israel para títulos de pregrado.

Quito, diciembre del 2013.

Para constancia firman:

TRIBUNAL DE GRADO

……..

PRESIDENTE

……..

MIEMBREO 1

……..

MIEMBRO 2

Resumen

En este trabajo se propone una metodología dirigida al descubrimiento del conocimiento oculto en información que aparentemente no aporta valor; el objetivo es determinar en la red de comercialización de combustible automotriz consumos anómalos que puedan derivar en posibles delitos hidrocarburíferos. La metodología incluye la utilización de técnicas de inteligencia artificial.

La técnica seleccionada es la de minería de datos porque a través de esta se puede encontrar correlaciones útiles para algún proceso, esto permite que una persona con poca experiencia adquiera la comprensión sobre un tema específico de manera fácil y rápida.

El presente documento contiene los siguientes capítulos:

En el capítulo I se describen los fundamentos teóricos sobre los cuales se basa la investigación.

En el capítulo II se detalla las técnicas de investigación aplicadas para el cumplimiento de los objetivos.

En el capítulo III se especifica la metodología propuesta para categorizar y modelar la red de comercialización de combustible automotriz utilizando técnicas de inteligencia artificial.

Abstract

This thesis proposes a methodology oriented to discover the hidden knowledge in information which apparently does not summit any value, the objective is to establish in the automobile fuel commercialization scheme any inappropriate consuming activities that may derive in possible fuel felonies. The methodology includes the use of artificial intelligence techniques.

The chosen technique is data mining because through this you can find useful correlations for some process, which allows people with not much experience to gain understanding about specific subjects in an easy and fast way.

This document contains the following chapters:

Chapter I describes theoretical basis on which this research is sustained.

Chapter II explains the investigation techniques that were used to fulfill its aims.

Chapter III specifies the proposed methodology to categorize and model the automobile fuel commercialization scheme using artificial intelligence techniques.

Índice General

Índice de Tablas

Índice de figuras

Índice de Pantallas

INTRODUCCIÓN

Introducción General

En el año 1995, se crea las siglas KDD (Knowledge Discovery in Databases), *es el proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos.*(Fayyad, 1996)

En las organizaciones existe un gran crecimiento del almacén de datos, en esta gran masa de datos se esconde información valiosa que con la aplicación de técnicas de descubrimiento del conocimiento se la puede obtener para darle una utilidad.

La minería de datos es una técnica de inteligencia artificial que permite determinar patrones y correlaciones en los datos.

Con el desarrollo de herramientas y paquetes de "minería de datos", como, por ejemplo, *Clementine de SPSS, Intelligent Miner de IBM, Mine Set de Silicon Graphics, Enterprise Miner de SAS, DM Suite (Darwin) de Oracle, WEKA (de libre distribución), Knowledge Seeker, etc.* (Hernandez & y, 2005), se facilita el uso de técnicas de minería de datos a no especialistas.

Ente las técnicas de minería de datos están las de clasificación no supervisada, los algoritmos más conocidos son K-medias y mapas auto-organizados, estos algoritmos identifican clústeres en los datos mediante funciones de distancia que encuentran similitudes entre los datos.

Con la ayuda de estas técnicas se logrará el procesamiento de grandes cantidades de datos a través de procesos automatizados, se identificarán patrones principales para la predicción de tendencias o comportamientos.

Para la presente investigación la utilización de la técnica de inteligencia artificial basada en data mining brindará su apoyo en la toma de decisiones, pues, con la base de datos los despachos de combustible automotriz a las gasolineras existentes en el Ecuador, sometida a un proceso de extracción de conocimiento y con la ayuda del algoritmo de aprendizaje no supervisado K-medias se conseguirá contar con información valiosa para la investigación especializada de delitos hidrocarburíferos que afectan la economía de nuestro país.

Antecedentes

El Gobierno Nacional, con la finalidad de controlar el uso, distribución y comercialización de precursores químicos, combatir el tráfico ilegal, el almacenamiento, transportación y comercialización sin la debida autorización de hidrocarburos y sus derivados, ha venido expidiendo Decretos Ejecutivos que facilitan la gestión de las empresas públicas para el cumplimiento de este objetivo.

En tal razón, el Decreto Ejecutivo No. 1859 faculta a las Fuerzas Armadas (FFAA), Policía Nacional y Agencia de Regulación y Control Hidrocarburífero (ARCH), a realizar operativos de control en carreteras y fronteras ecuatorianas limitándose su trabajo a patrullajes a distintas horas del día lo que ha permitido identificar a pequeños y medianos infractores, sin embargo no se ha contado con una solución informática que permita identificar comportamientos anómalos en el consumo de las estaciones de servicio a nivel nacional, que sirvan de guía para la detección de posible ilícitos en este segmento de mercado.

Descripción del problema a resolver

La inexistencia del apoyo informático para la detección de posibles delitos hidrocarburíferos, ha derivado que las entidades de control actúen en base a denuncias formuladas por fuentes humanas o en base a programas de patrullajes, realizando esfuerzos cuyos resultados son de pequeña y mediana incidencia.

Objeto de Estudio

Aplicación de recursos tecnológicos basados en inteligencia artificial para el análisis masivo de información.

Campo de la Investigación

Técnicas y herramientas de inteligencia artificial aplicada a la extracción del conocimiento de las Bases de datos.

Objetivos

Objetivo General

Efectuar la investigación sobre metodologías y soluciones informáticas tendientes a extraer el conocimiento de las bases de datos utilizando técnicas de inteligencia artificial para identificar posibles ilícitos hidrocarburíferos.

Objetivos específicos

1. Realizar la investigación bibliográfica referente a las metodologías y técnicas de inteligencia artificial que permitan el desarrollo de la investigación mediante el método de investigación histórico-lógico.
2. Diagnosticar la problemática a resolver para analizar las posibles soluciones utilizando la opinión de expertos y herramientas computacionales.
3. Realizar el análisis de la red de comercialización de combustible automotriz utilizando técnicas de inteligencia artificial para identificar posibles ilícitos hidrocarburíferos.
4. Presentar los resultados obtenidos para su posterior validación en campo por expertos en casos de delitos hidrocarburíferos.

Idea a defender en el proceso investigativo

Caracterizando, modelando y presentando una metodología de análisis de la red de comercialización de combustible automotriz que permita identificar gasolineras con un anómalo consumo que podría derivar en la existencia de un posible ilícito.

CAPÍTULO I

1. MARCO TEÓRICO

1.1. Antecedentes investigativos

En la actualidad, la Inteligencia artificial se está aplicando a numerosas actividades realizadas por los seres humanos como por ejemplo: la robótica, la visión artificial, técnicas de aprendizaje y la gestión del conocimiento. Estas dos últimas aplicaciones de la inteligencia artificial son las que más se ajustan para el cumplimiento de los objetivos del presente trabajo de investigación, debido a que existe la necesidad de contribuir con una solución informática para incorporar el conocimiento sobre la identificación de posibles ilícitos que permitan a las autoridades competentes tomar acciones eficientes y oportunas frente al cometimiento de delitos hidrocarburíferos.

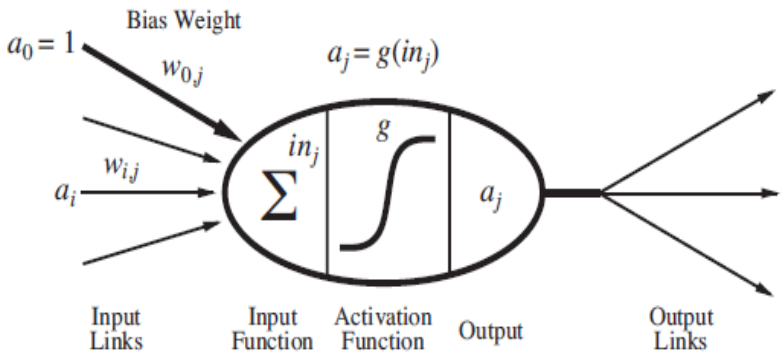
Las técnicas de inteligencia artificial toman la información cualitativa y a partir de ella se diseñan e implementan modelos estadísticos y computacionales que permiten realizar la resolución de diversos problemas empresariales.

Entre las técnicas más destacadas de inteligencia artificial están los **sistemas expertos, las redes neuronales, los algoritmos genéticos, la lógica difusa y minería de datos**, estas técnicas pueden combinarse para obtener una solución más adecuada al problema de estudio.

1.2. Fundamentación Científico – Técnica

Los **sistemas expertos** tienen dos componentes principales: el primero es el conocimiento y las experiencias de los expertos en un determinado dominio representado por medio de símbolos y el segundo es el mecanismo que obtienen las conclusiones de la base de conocimiento mediante procesos de búsqueda.

*Las **redes neuronales** es sólo una colección de unidades conectadas entre sí; las propiedades de la red se determina por su topología y las propiedades de las neuronas. Un modelo matemático simple para una neurona está dado por* $a_j = g\,(\sum_i^n = 0\ w_{i,j}, a_i)$, *donde ai es la activación de la salida de la unidad i y wi,j es el peso en el enlace de la unidad i en la unidad* (Rusell, 2010).



Ésta se convierte en una herramienta tecnológica potente para el procesamiento de la información cuyos resultados pueden permitir la toma decisiones eficientes y oportunas.

Los **algoritmos genéricos** simulan la mecánica de la selección natural y de la genética utilizando la información histórica para encontrar nuevos puntos de búsqueda de una solución óptima, permitiendo obtener soluciones a un problema que por su complejidad no tiene ningún método de solución de forma precisa.

La **lógica difusa** *permite representar el ser miembro de un conjunto como una distribución de posibilidades*(Knight, 1994), esto permite el trabajo con información de difícil especificación.

La **minería de datos** es una fase de la metodología de KDD en la que se aplican técnicas de inteligencia artificial, estadística y aprendizaje automático con la finalidad de extraer el conocimiento de una base de datos y someterla a un proceso de transformación para su comprensión y uso posterior.

Para el desarrollo del presente trabajo se utiliza la técnica de Minería de Datos, para ello se evaluó las características de 2 sistemas de distribución libre, ver tabla 1.

Tabla 1.- Análisis comparativo de Sistemas de libre distribución.

Nombre del Sistema	Algoritmos	Plataforma sobre la cual trabaja	Tipos de archivos soportados.
Orange	Contiene técnicas de clasificación y evaluación automatizadas como por ejemplo: • Clasificación • Select atributes • Asociación • Clúster.	Requiere como base Phyton para su funcionamiento.	Funciona con varias tipos de archivos: *.tab, *.txt, *.data, *.dat, *.rda, *.rdo.
WEKA	Cuenta con una gran librería de técnicas de aprendizaje automático tanto supervisadas como no supervisadas, como por ejemplo: • Regresión • Algoritmos a priori • K-medias • Kohonen • Filtered • Clusterer	Esta desarrollado completamente en java por lo que basta tener una JVM disponible.	Funciona con varias tipos de archivos: *.arff, *.names, *.data, *.csv, *.dat, *.bsi, *.xrff, *.libsvm

Fuente: Páginas oficiales de Orange y Weka**. Elaborado por** la autora**.**

Del análisis se decide la utilización del software WEKA por ser más fácil de implementar tanto en Windows como en Linux.

Descripción del software WEKA

WEKA es una herramienta de aprendizaje automático y data mining desarrollado en Java por la Universidad de Waikato de Nueva Zelanda, que tiene una amplia gama de algoritmos de clasificación, regresión, clustering, asociación, adicionalmente cuenta con un modulo de visualización para una mejor interpretación de resultados.

WEKA es una herramienta de libre distribución y por estar desarrollada en Java no depende de una arquitectura especial, funciona en cualquier plataforma que cuente con un JVM (Java Virtual Machine) disponible.

Los instaladores de la herramienta están disponibles en la página oficial de la Universidad de Waikato (WEKA, The University of Waikato).

La JVM de Sun asigna un valor de 100 megas de memoria para realizar las operaciones en Weka, valor que para el desarrollo de este estudio resulta insuficiente. Para dar solución a la insuficiente asignación de memoria asignamos 1GB de memoria para el análisis de los datos mediante el siguiente comando:

javaw -Xmx1024m -classpath "C:\Archivos de Programa\Weka-3-6\weka.jar" weka.gui.Main

Dentro del ambiente gráfico de WEKA se encuentra el módulo exploración de datos (WEKA Knowledge Explorer) el cual permite pre procesar, clasificar, asociar y visualizar datos de una manera fácil e intuitiva.

WEKA cuenta con seis entornos de ejecución:

Preprocess.- permite seleccionar la fuente de datos y su preparación.
Classification.- permite aplicar esquemas de clasificación y regresión, entrenar modelos y evaluar su precisión.
Cluster.- Integra varios métodos de agrupamiento.
Associate.- Incluye algunas técnicas de reglas de asociación.
Select Attributes.- permite aplicar técnicas para la reducción del número de atributos.
Visualice.- permite estudiar el comportamiento de los datos mediante técnicas de visualización.

Descripción de los archivos .arff con los que WEKA trabaja.
Nativamente WEKA trabaja con un formato denominado ARFF, acrónimo de Attribute-Relation File Format, es un archivo de texto ASCII que describe una lista de instancias que comparten un conjunto de atributos.

Un archivo con este formato contiene meta-información, es decir, contiene el nombre y tipo de cada atributo y una descripción textual del origen de los datos. Está conformado por tres secciones:

a. **Cabecera.-** Se define el nombre de la relación bajo el siguiente formato:
   @relation <Nombre de la relación>
b. **Declaraciones de los atributos.-** se declaran todos los atributos contenidos en el archivo en forma secuencial de la siguiente forma:
   @attribute <Nombre del atributo> <Tipo>

El tipo de dato puede ser de cuatro clases: numérico, especificación nominal, string y date.

c. **Sección Datos.-** al iniciar esta sección se escribe la declaración @data y a continuación se declara los datos separándolos entre comas los atributos y con saltos de línea las relaciones.

Descripción del método K-medias utilizado.

El método de las K-medias, permite asignar a cada observación el clúster que se encuentra más próximo en términos del centroide (media), la distancia empleada es la euclídea $d_{ij}=\sqrt{\sum_{k-1}^{n}(x_{ik}-x_{jk})^2}$.

El Clustering o agrupamiento, es un procedimiento que se aplica para encontrar grupos y estructuras en los datos con características similares.

Pasos:

1. *Se toman al azar k clústeres iniciales, para el caso de estudio 4.*
2. *Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clústeres y se reasignan a los que estén más próximos. Se vuelven a recalcular los centroides de los k clústeres después de las reasignaciones de los elementos.*
3. *Se repiten los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo.*

Usualmente, se especifican k centroides iniciales y se procede al paso (2) y, en la práctica, se observan la mayor parte de reasignaciones en las primeras iteraciones.(UC3M)

La finalidad de formar clústeres es que los centroides estén lo más separado posible entre sí, además es importante la selección del número de clústeres puesto que se pueden presentar algunos problemas:

- *Si dos centroides iniciales caen por casualidad en un único clúster natural, entonces los clústeres que resultan están poco diferenciados entre sí.*
- *Si aparecen outliers (valores atípicos), se obtiene por lo menos un clúster con sus objetos muy dispersos.*
- *Si se imponen previamente k clústeres puede dar lugar a grupos artificiales o bien a juntar grupos distintos.*(UC3M)

Para enfrentar estos problemas se consideró varias posibilidades del número de clústeres y realizando una comparación de sus resultados.

CAPÍTULO II

2. METODOLOGÍA Y DIAGNÓSTICO DE LA INVESTIGACIÓN

2.1. Fuentes de información

Para el desarrollo del presente trabajo se toma como fuente de información los despachos efectuados por EPPETROECUADOR hacia las diferentes gasolineras del país, dicha información reside en la base de datos del Sistema de Comercialización de la Empresa.

2.2. Metodología de la investigación.

Para el desarrollo del trabajo se utiliza los métodos de investigación:

Inductivo-deductivo, para la elaboración de los objetivos general y específicos.

Histórico-lógico, en la realización del fundamentación teórica.

Analítico-sistémico, en el desarrollo de la propuesta.

Y, la metodología técnica propuesta por Usama Fayyad, Knowledge Discovery in Databases (KDD):

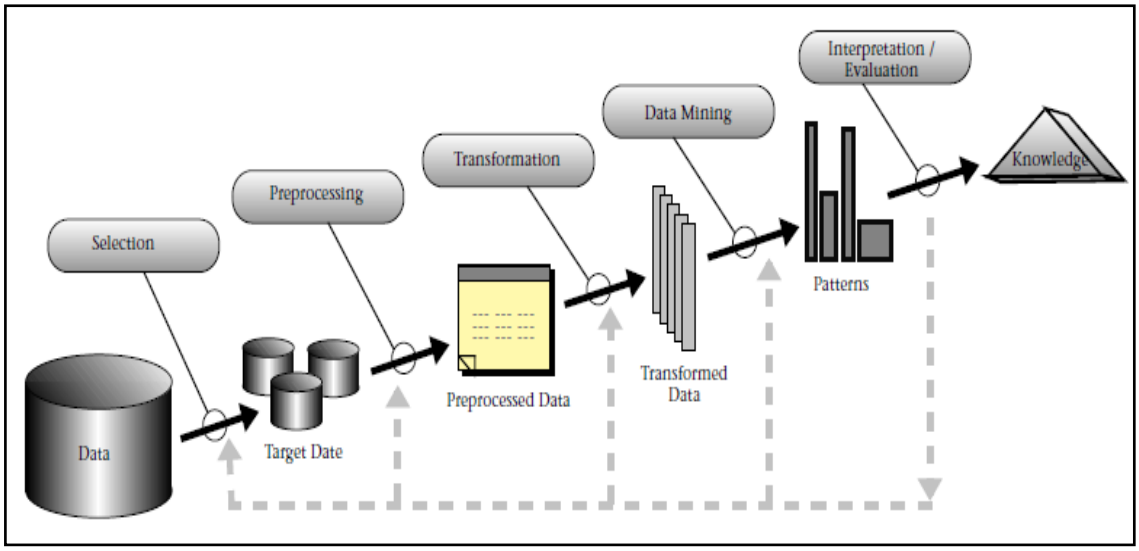


Figura 1.- Una visión general de los pasos que componen el proceso de KDD.
Fuente: From Data Mining to Knowledge Discovery in Databases, *Usama Fayyad*

Como se puede apreciar, para aplicar el KDD se requiere la realización de una serie de actividades previas dirigidas a la preparación de los datos de entrada puesto que pueden proceder de fuentes heterogéneas, no cuentan con un formato apropiado o compatible con el sistema a utilizar. Por esta razón se toma como base los pasos propuestos por Fayyad y se añade uno previo que tiene que ver con la determinación de objetivos, puesto que, sin conocer que es lo que se quiere conseguir no será útil iniciar un proceso de KDD. Los pasos a seguir son:

1. Determinación de objetivos.
2. Preparación de Datos que involucra:
   - Selección de las fuentes de información externas e internas.
   - Preprocesamiento de la información que implica el análisis de la calidad de los datos y determinación de las técnicas de KDD a aplicar.
3. Transformación de los datos.- donde se obtiene una evolución de los datos en un modelo adecuado para el análisis.
4. Minería de datos.- donde se aplica las técnicas automatizadas para explotar los datos seleccionando una combinación apropiada de algoritmos.
5. Análisis de Resultados.- interpretación de los resultados obtenidos en la etapa anterior, generalmente con la ayuda de una técnica de visualización.
6. Asimilación del conocimiento.- utilizar el conocimiento obtenido en función de los objetivos planteados.

Cabe indicar que la metodología propuesta es totalmente iterativa, es decir de no obtener resultados con una primera selección de datos y algoritmos a utilizar se puede regresar al paso inicial.

2.3. Técnicas e instrumentos de recolección de datos.

La información se obtendrá del Sistema de Comercialización de EPPETROECUADOR a través de la herramienta Business Objetcs que la empresa mantiene para análisis y extracción de información.

2.4. Análisis e interpretación de resultados.

Para el análisis e interpretación de los resultados obtenidos luego de aplicar la metodología de trabajo se toma como muestra dos o tres elementos que permitan determinar en primera instancia si se consiguió los objetivos planteados, para ello se

utiliza el internet como medio de consulta para obtener información de las gasolineras que se agruparan dentro del marco de estudio, las paginas a consultar son: la del Servicio de Rentas Internas y la Superintendencia de Compañías.

2.5. Problemas y especificación de requerimientos.

Los problemas que se presentaron en el proceso de investigación fue el número de registros extraídos de la base de datos que al inicio superaban los 2 millones y medio de registros ya que se realizó una extracción diaria de despachos por gasolinera entre los años 2008 y 2013.

Luego de analizados y procesados los datos se determinó la no necesidad de contar con un dato tan especifico como despacho diario, reduciendo a mensual la temporalidad de los despachos.

Adicionalmente de detecta la necesidad de contar con atributos que describan de manera general a cada registro obtenido, como fue el caso de la creación del atributo “ubicación” que permitió clasificar a las gasolineras como fronterizas y no fronterizas.

2.6. Estudio de Factibilidad (Operativa, Tecnológica y Económica).

El presente trabajo, propone una nueva metodología para realizar análisis de información, esto involucrará la participación de un analista que ponga en práctica los pasos descritos en la propuesta.

Las herramientas tecnológicas a utilizar son: Business Object con licencia pagada con la que cuenta la empresa y WEKA software de libre distribución que funciona en cualquier plataforma; la persona a cargo de la ejecución deberá contar con conocimientos en el manejo de la herramienta Business Object y en la información almacenada en el sistema de comercialización.

Los costos requeridos para la ejecución del proyecto se reducen al costo de renovación de licencias de Business Objects y el costo hora/trabajada de un analista que corresponde a $8.38 USD.

CAPÍTULO III

3. PROPUESTA

3.1. Antecedentes de la propuesta

En la actualidad los entes de control FFAA, Policía Nacional y ARCH realizan operativos de control en base a programas de patrullaje o denuncias. De ahí nace la necesidad de la utilizar herramientas computacionales que permitan extraer información útil de datos aparentemente sin valor.

La Empresa Pública PETROECUADOR (EPPETROECUADOR) mantiene una base de información referente a los despachos efectuados diariamente a las 1.100 gasolineras existentes a nivel nacional, información a la que se aplicarán herramientas del campo de estudio denominado Knowledge Discovery in Databases, KDD *(*Extracción de Conocimiento en Base de Datos).

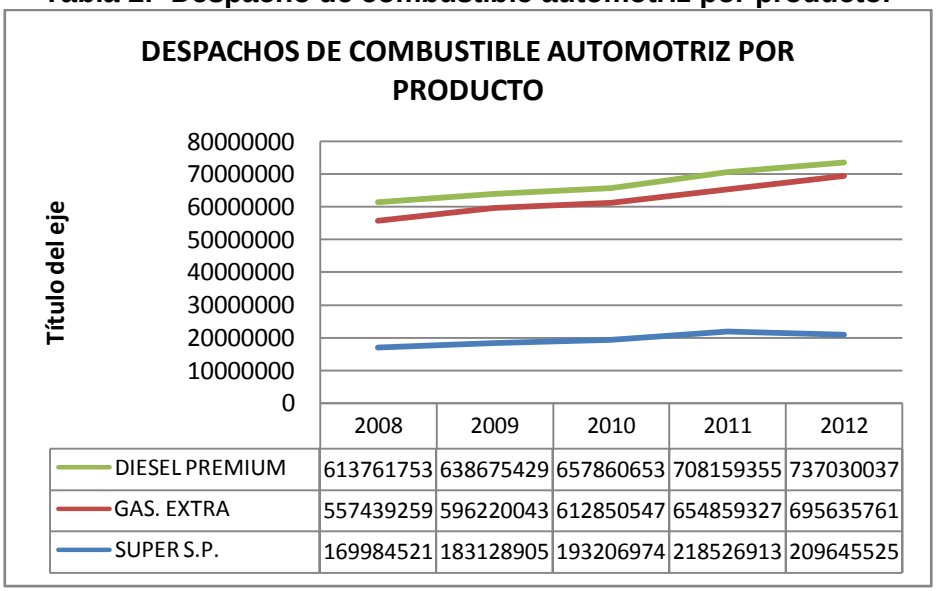
Con la Extracción de Conocimiento en Base de Datos se pretende encontrar patrones útiles de datos que permitan identificar posibles delitos hidrocarburíferos con mayor asertividad y de mayor incidencia.

Para mostrar un panorama macro del consumo inequitativo de combustible automotriz en el Ecuador se toma las siguientes variables para el análisis preliminar:

- Despachos de combustible automotriz a nivel nacional por producto.
- Despachos de combustible automotriz a nivel nacional por provincia.
- Número de habitantes por provincia.
- Número de vehículos por provincia.

Tomando como base los despachos de combustible automotriz a nivel nacional por producto desde el año 2008 hasta el año 2012, se identificó que el Diesel es el combustible más utilizado a nivel nacional y su consumo se encuentra en constante crecimiento.

Tabla 2.- Despacho de combustible automotriz por producto.



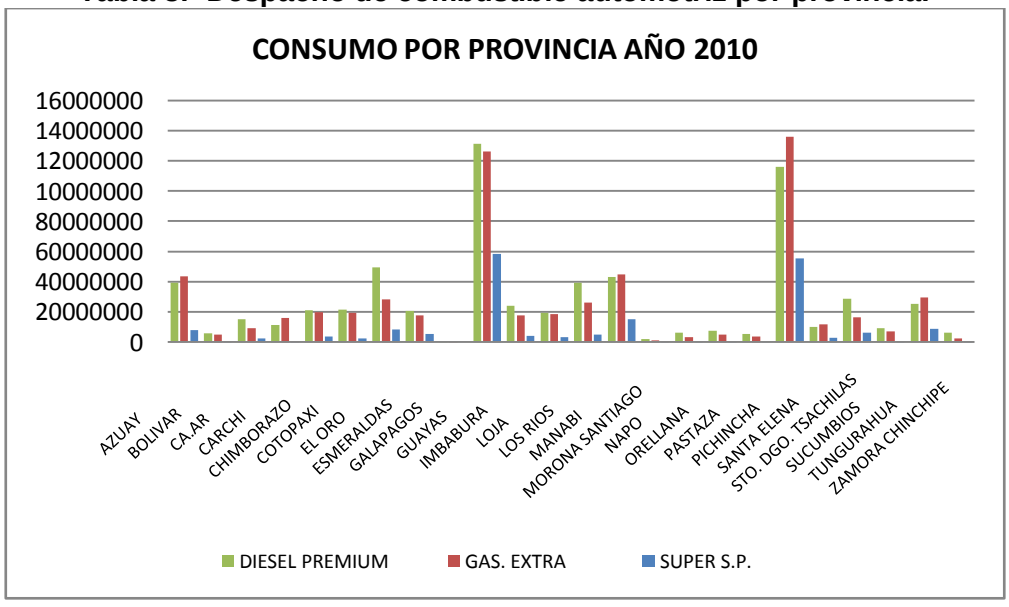
	2008	2009	2010	2011	2012
DIESEL PREMIUM	613761753	638675429	657860653	708159355	737030037
GAS. EXTRA	557439259	596220043	612850547	654859327	695635761
SUPER S.P.	169984521	183128905	193206974	218526913	209645525

Fuente: Sistema de Comercialización de EPPETROECUADOR

Con los despachos de combustible automotriz a nivel nacional por provincia del año 2010 se analizará la relación con el número de habitantes y vehículos por provincia.

Es importante indicar que se toma el año 2010 como referencia para el análisis porque que es el año en el cual el INEC publicó las cifras oficiales del CENSO de población y vivienda.

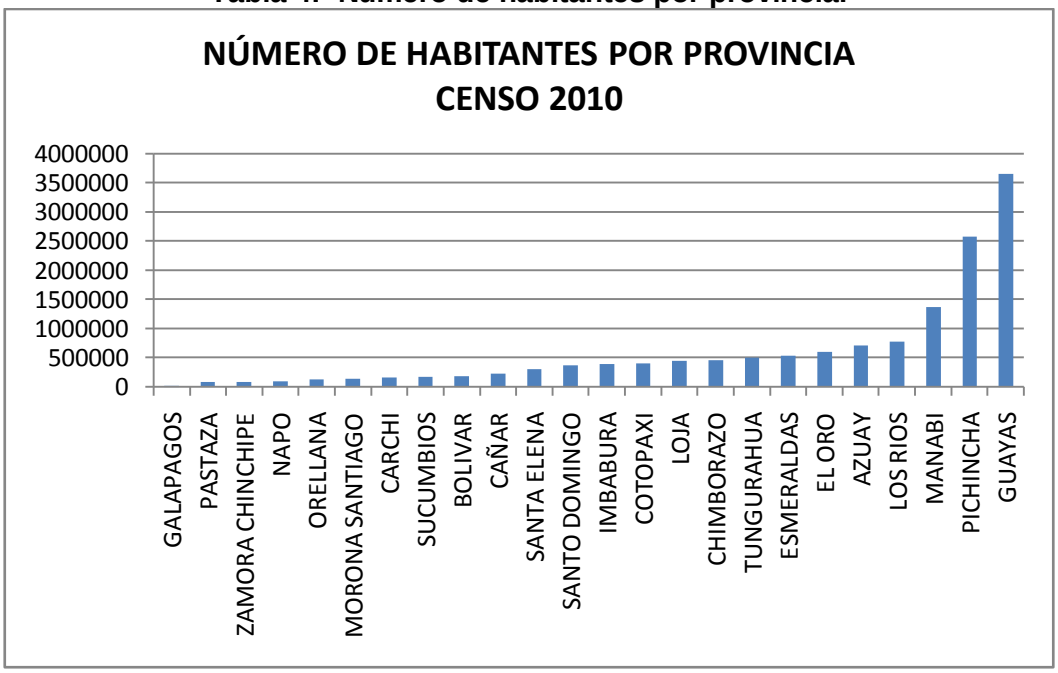
Tabla 3.- Despacho de combustible automotriz por provincia.



Fuente: Sistema de Comercialización de EPPETROECUADOR

Con los resultados de la tabla 3 se concluye que las provincias de mayor consumo son las que más habitantes tienen, lo que aparentemente se vería como un comportamiento normal de consumo de los combustibles, sin embargo, a continuación se presenta un análisis más detallado en el cual se toma los datos poblacionales por provincia.

Tabla 4.- Número de habitantes por provincia.



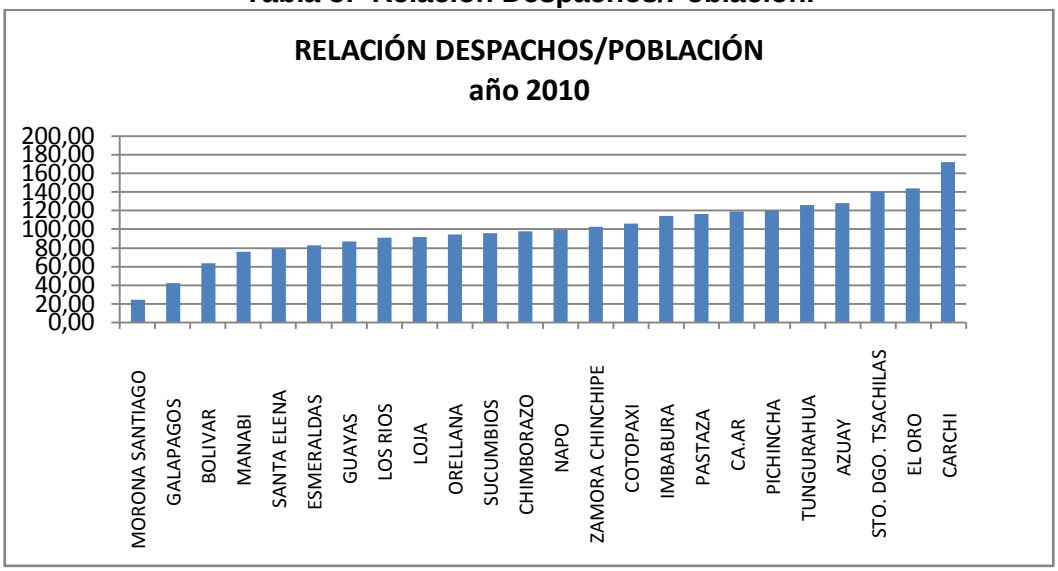
Fuente: Censo de población y vivienda año 2010 INEC

Con los datos poblacionales de la tabla 4 y los despachos de combustible automotriz por provincia de la tabla 3, se puede obtener una aproximación del consumo de combustible automotriz por habitante y por provincia, para lo cual se plantea la siguiente relación:

Despachos por Provincia / Número de habitantes por provincia

Aplicando esta relación, se determina que las tres provincias de mayor consumo por habitante son: Carchi, El Oro y Santo Domingo de los T-Sáchalas, aspecto que en parte concuerda con los controles a las provincias de frontera y de donde se obtiene el mayor número de operativos de control e incautaciones, no obstante la provincia de Santo Domingo no ha sido identificada como un problema mayor en la que se deba intensificar los controles.

Tabla 5.- Relación Despachos/Población.



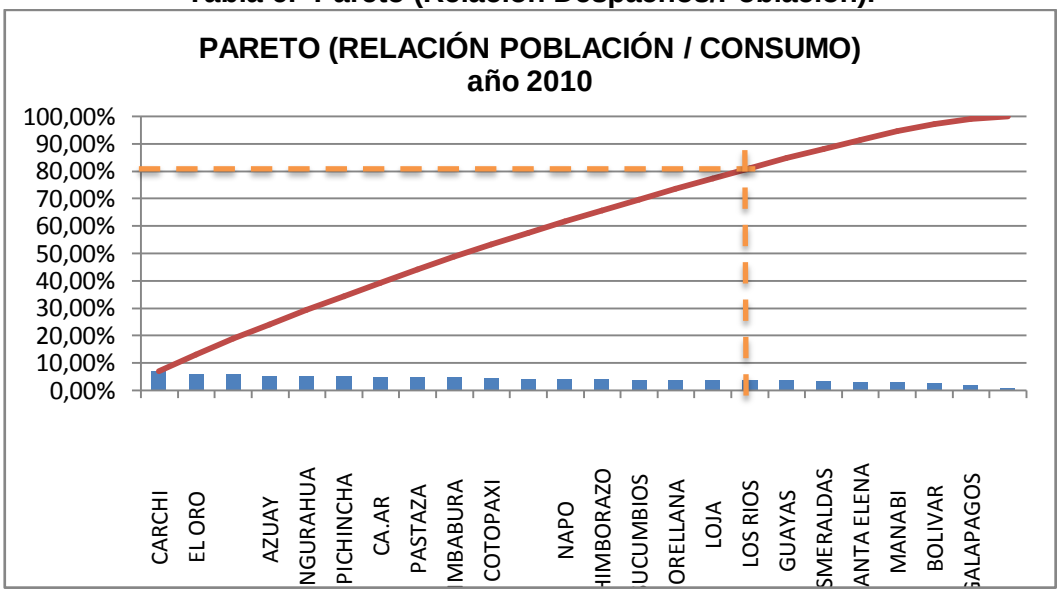
Fuente: Sistema de Comercialización de EPPETROECUADOR y Censo de población y vivienda año 2010 INEC

De acuerdo con los resultados obtenidos y mostrados en la tabla 5 se hace necesario realizar un análisis global de la información para esto se aplica la técnica de Pareto.[1]

Con la ayuda de la Ley de Pareto, se clasifican a las provincias, enmarcando los pocos vitales como las provincias de menor consumo y los muchos triviales como las provincias de mayor consumo, como se muestra en la tabla 6, enfocando el análisis a las 17 provincias del Ecuador consumen el 80% del combustible distribuido para el segmento automotriz, debiéndose considerar como prioridad al problema planteado.

[1] *El diagrama de Pareto, nace con un estudio realizado por el Economista italiano Vilfredo Pareto, quien realizó un estudio de cómo estaba distribuida la riqueza en Italia obteniendo como resultado que el 80% de las riquezas estaban en manos de solo el 20% de la población italiana, pero el Doctor Joseph Juran observó que la ley de Pareto era aplicable a problemas de calidad creando así su regla del 80/20 donde el 80% de los problemas representan los muchos triviales y el 20% los pocos vitales. Por lo que es mejor resolver los pocos vitales* (Delgado, 2008).

Tabla 6.- Pareto (Relación Despachos/Población).

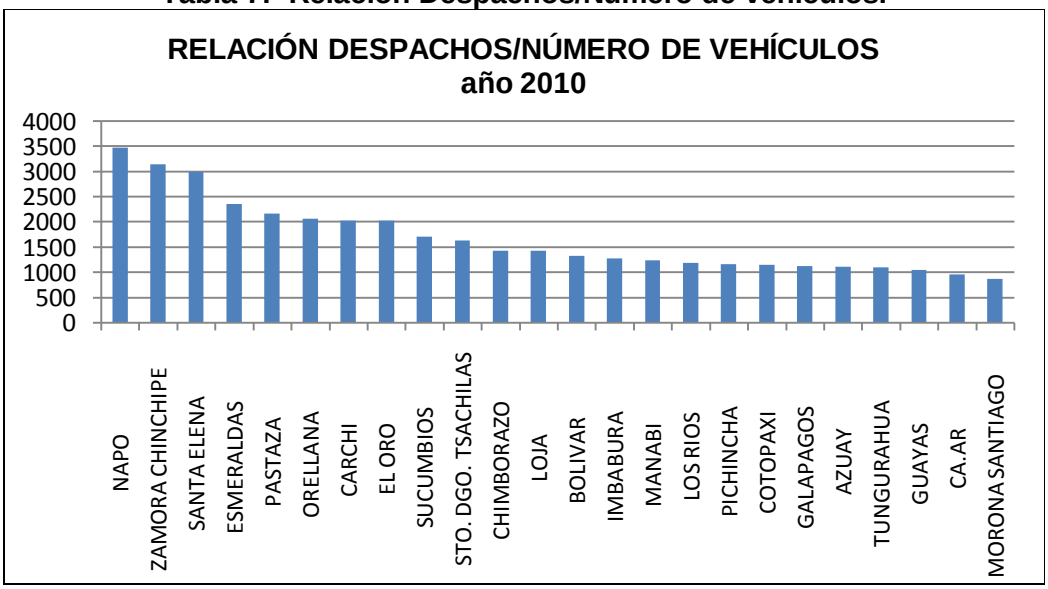


Por último, con la base de datos de los vehículos matriculados y los despachos de combustible automotriz por provincia, se puede obtener una aproximación del consumo de combustible automotriz por vehículo y por provincia, para lo cual se plantea la siguiente relación:

Despachos por Provincia / Número de vehículos matriculados

De la relación se obtienen los resultados de la tabla 7:

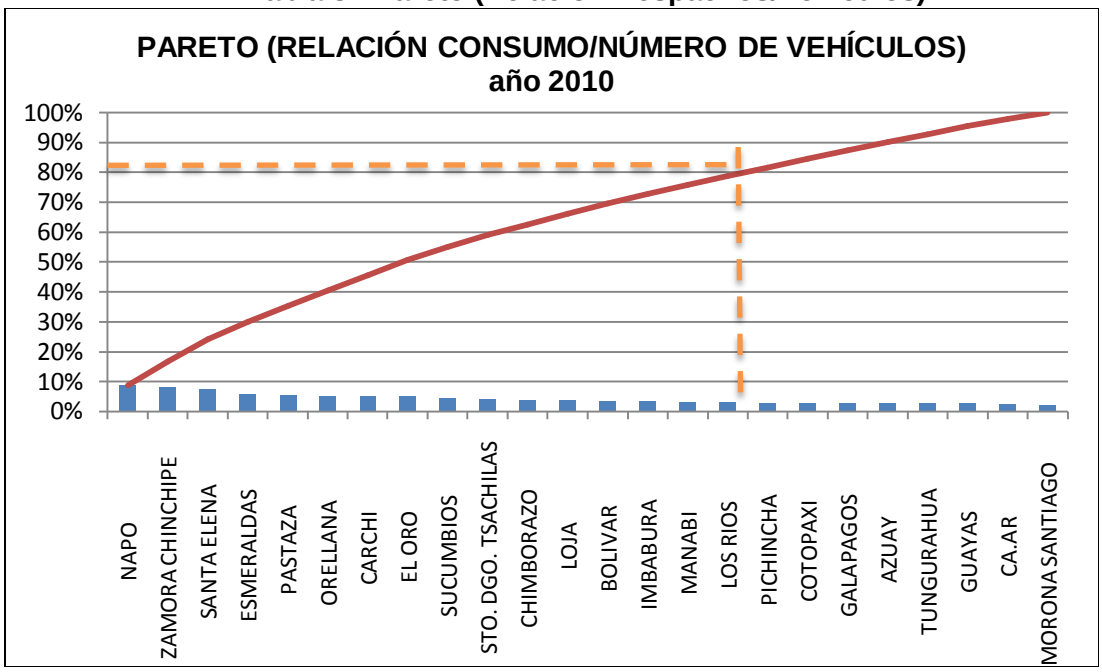
Tabla 7.- Relación Despachos/Número de vehículos.



Fuente: Sistema de Comercialización de EPPETROECUADOR y Número de vehículos motorizados matriculados, por uso, según provincias año 2010 INEC

Aplicando la técnica de Pareto a esta información se encuentra que de las 24 provincias del Ecuador 16 tiene un elevado consumo respecto de sus vehículos.

Tabla 8.- Pareto (Relación Despachos/Vehículos).



Comparando los resultados mostrados en la tabla 8, con los obtenidos en la Relación tabla 6 se puede identificar que las provincias que se repiten en los dos análisis son: CARCHI, CHIMBORAZO, EL ORO, IMBABURA, LOJA, LOS RIOS, NAPO, ORELLANA, PASTAZA, STO. DGO. TSÁCHILAS, SUCUMBIOS y ZAMORA CHINCHIPE.

De las 12 mencionadas provincias 7 son fronterizas (CARCHI, EL ORO, LOJA, ORELLANA, PASTAZA, SUCUMBIOS y ZAMORA CHINCHIPE), 1 es considerada como fronteriza por la cercanía a límite con Colombia (IMBABURA), 2 pertenecen a lo se conoce como sierra central (CHIMBORAZO, STO. DGO. TSÁCHILAS), 1 corresponde al oriente ecuatoriano (NAPO) y la última LOS RIOS que pertenece a la Costa Ecuatoriana.

La información y análisis presentados muestran por qué las autoridades de control han aplicado su mayor esfuerzo en las provincias de frontera pero no se toma en cuenta a las provincias centrales porque al parecer no existen incentivos para el robo o desvío del combustible, por esta razón se justifica la necesidad de contar con la extracción de conocimiento de la Base de Datos de Despachos que mantiene EPPETROECUADOR y colaborar con las entidades de control en su labor de erradicar los delitos hidrocarburíferos de los que adolece el Ecuador desde hace muchos años.

3.2. Justificación

Sobre la base de los antecedentes de la propuesta, se advierte viable la solución, puesto que, se determina por medio de otras herramientas como el Excel y otras fuentes de información como el número de vehículos y número de habitantes obtenido del INEC, la existencia de comportamiento anómalos en la comercialicen de combustible automotriz, hecho que deberá comprobarse al aplicar la metodología propuesta sin la necesidad de depender de las fuentes externas de información.

3.3. Objetivos de la propuesta

3.3.1. General

Realizar el análisis de la red de comercialización de combustible automotriz utilizando técnicas de inteligencia artificial para identificar posibles ilícitos hidrocarburíferos.

3.3.2. Específicos

- Determinar los objetivos del análisis para delimitar el alcance utilizando el método de investigación inductivo-deductivo.
- Realizar la preparación de los datos seleccionando las fuentes de información y revisando la calidad de datos con herramientas de extracción y análisis de datos.
- Generar la transformación de los datos que permita obtener un modelo analítico con el sistema WEKA.
- Aplicar los algoritmos de agrupamiento para obtener el conocimiento oculto en los datos utilizando los recursos y técnicas del sistema WEKA.
- Analizar e interpretar los resultados visualizando los datos y comparándolos con deducciones previamente realizadas.
- Aplicar el conocimiento adquirido tomando muestras aleatorias y verificando a detalle la información de consumos, tributación, administración y ubicación.

3.4. Desarrollo de la Propuesta

De acuerdo con la metodología propuesta para el desarrollo del presente trabajo se realizaron los siguientes pasos:

Paso 1: Determinación de Objetivos

Objetivo General:

Realizar la caracterización y modelación de la red de comercialización de combustible automotriz utilizando técnicas de inteligencia artificial para identificar posibles ilícitos hidrocarburíferos.

Objetivos específicos:

- Categorizar automáticamente los datos de despachos a gasolineras.
- Clasificar gasolineras de acuerdo a su situación geográfica y volumen de consumo.
- Detectar consumos anómalos que podrían derivar en posibles delitos hidrocarburíferos.

Paso 2: Preparación de los datos

Identificación de las fuentes de información.

Por tratarse de un caso de estudio sobre la comercialización de combustible automotriz en el Ecuador se obtiene la información del Sistema de Comercialización de la Empresa Pública de Hidrocarburos del Ecuador sobre los Despachos de combustible automotriz realizados a las gasolineras a nivel nacional desde enero del 2008 hasta agosto del 2013. Los datos obtenidos para el análisis son: código de la estación de servicio, condición de la estación de servicio, nombre de la estación de servicio, ubicación geográfica de la estación de servicio (provincia y cantón), código del producto despachado, nombre del producto despachado, año/mes de despacho y volumen despachado.

Análisis de la calidad de los datos y determinación de los algoritmos de minería que se pueden utilizar.

Dentro del periodo de análisis de debió considerar una de las políticas gubernamentales que ejecuto EPPETROECUADOR en el año 2012 en lo referente al mejoramiento de los combustibles, lo que requirió transformar el código de producto despachado de 2 a 21 y nombre de producto despachado de DIESEL 2 A DIESEL PREMIUM, que en teoría se habla del mismo producto solo que con un mayor octanaje, índice que no es parte del caso de estudio pero si necesario para la uniformidad de los datos a analizar.

En el campo nombre del estación de servicio se determinó la presencia de caracteres especiales como las comillas (“”) y vocales tildadas, mismas que fueron reemplazadas por espacios en blanco y vocales no tildadas respectivamente.

En lo referente a los algoritmos de minería se consideró la utilización del algoritmo de APRENDIZAJE NO SUPERVISADO, puesto que se requiere identificar consumos anómalos que a priori se desconocen.

Paso 3: Transformación de los datos.- conversión de los datos a un modelo analítico.

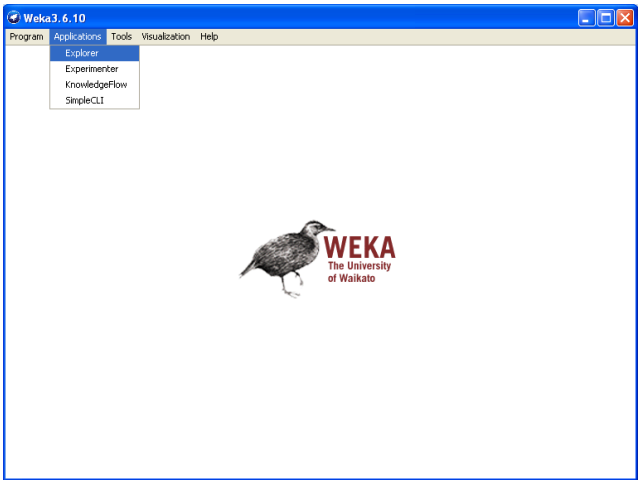
Para el caso de estudio se ha determinado que la utilización del software WEKA, esta herramienta trabaja con un tipo de archivo .ARFF el mismo que tiene la siguiente estructura:

```
@relation DatosSinDepositoCompletol

@attribute CodCliente numeric
@attribute CondicionCliente numeric
@attribute NombreEstacion {'ESPEJO / ESTACION DE SERVICIO   ','SAN ISIDRO /ESTACION DE SERV.   '........}
@attribute NomProvincia {'CARCHI             ','IMBABURA           '....}
@attribute NomCanton {'TULCAN             ','ESPEJO             '.....}
@attribute CODIGOPRODUCTO numeric
@attribute NOMBREPRODUCTO {'GAS. EXTRA         ','DIESEL PREMIUM     ','SUPER S.P.         '}
@attribute Ano/Mes {2008/01,2008/02,2008/03,2008/04......}
@attribute VolumenEntregado numeric

@data
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/01,32000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/02,30000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/03,30000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/04,30000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/05,30000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/06,30000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/07,34000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/08,35500
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/09,36000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/10,36000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/11,32000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2008/12,32000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2009/01,34000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2009/02,27500
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2009/03,32500
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2009/04,33000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2009/05,35000
1010001,3,'ESPEJO / ESTACION DE SERVICIO   ','CARCHI             ','TULCAN             ',1,'GAS. EXTRA        ',2009/06,34000
```

Figura 2.- Archivo .arff

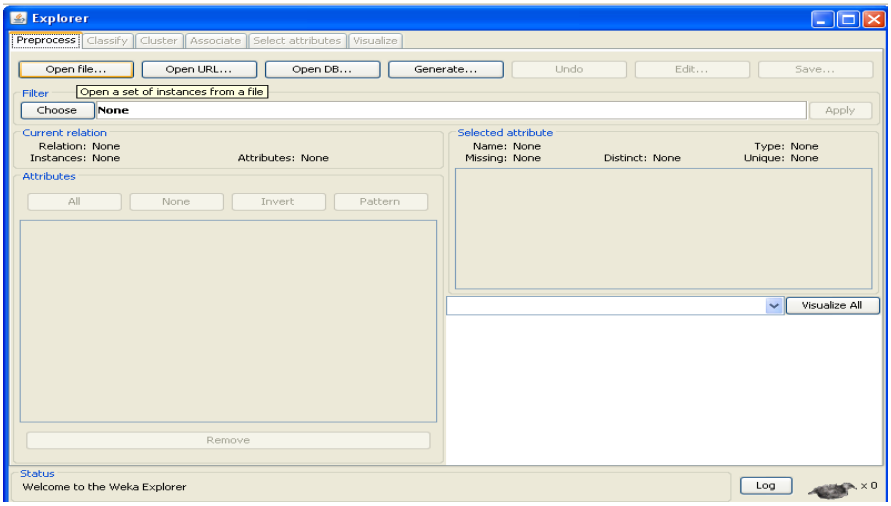
De acuerdo con los tipos de datos soportados por WEKA, se modificó el archivo Excel para que la información sea receptada y procesada; una vez obtenido el archivo.csv con la información proporcionada por EPPETROECUADOR, se sometió a la carga dentro del software WEKA el mismo que transformo la información a un archivo tipo .ARFF mediante los siguientes pasos:

a) Abrir el entorno WEKA



Pantalla 1.- Entorno WEKA

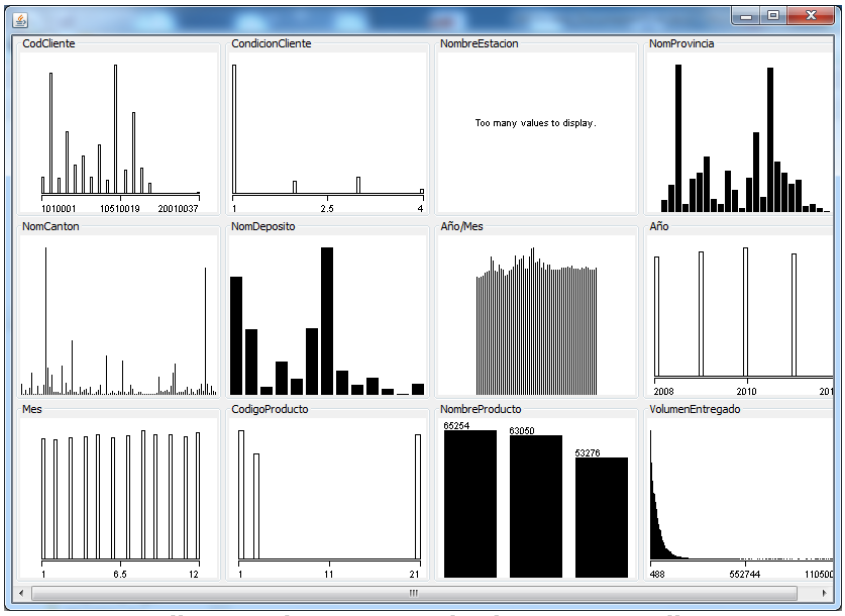
b) Con un clic en la pestaña Explorer, se abre el archivo .csv y lo convierte en .arff con la opción Save.



Pantalla 2.- Modulo Explorer

Paso 4: Minería de datos.- aplicar los algoritmos seleccionados.

Una vez creado el archivo .arff, se lo abre desde WEKA para visualizar la información de cada atributo, por ejemplo para el caso del campo nombre de producto se visualiza un histograma con relación al volumen despachado, es decir, se obtiene una estadística de cada atributo. En la parte izquierda nos da información sobre el número de datos, número de atributos y el nombre de los atributos. En la parte superior derecha aparece información estadística sobre los atributos (media, desviación típica, valores máximo y mínimo, “Unique” se refiere al número de valores que sólo aparecen una vez en ese atributo y “Distinct” al número de valores distintos.



Pantalla 3.- Primeros resultados con 9 atributos

De lo observado como primer resultado se pueden establecer características ya conocidas desde el inicio de la investigación. Con el atributo nombre de provincia se puede determinar a simple vista que las provincias de más alto consumo de combustibles en orden descendente son Pichincha, Guayas, Manabí, Tungurahua, Los Ríos, Chimborazo, El Oro, Cotopaxi, Azuay y Santo Domingo de los Tsáchilas. Así mismo, con el atributo Nombre de Deposito se corrobora que los terminales de EPPETROECUADOR más grandes son el Terminal Pascuales y El Beaterio, siguiéndoles el Terminal Santo Domingo y el Terminal Ambato.

Con el atributo Año/Mes, se observa un comportamiento casi uniforme de consumo por año, salvo unos picos presentado en agosto, septiembre y diciembre del 2008, agosto y diciembre del 2009, marzo, abril, mayo, agosto del 2010, diciembre del 2011 y diciembre del 2012, resultados que nos podrían suponer que el ingreso a clases en la región sierra y las fiestas de fin de año en todos los casos son de más alto consumo del combustible automotriz, no así en los meses de marzo , abril y mayo del 2010.

Con el atributo año se indica un ordenamiento mayor a menor en términos de despacho de la siguiente manera: 2010, 2009, 2011,2012 y 2008 siendo el año 2010 el de mayor consumo, así mismo los meses de mayor consumo en el periodo de estudio son los meses de agosto y diciembre.

Con el atributo “Nombre del Producto” se identifica que el combustible de mayor consumo es el diesel, seguido de la gasolina extra y en porcentaje menor la gasolina súper.

Continuando con el análisis, se requiere elegir los algoritmos de filtro y agrupamiento que permita identificar los consumos anómalos.

Inicialmente se realizó pruebas con diversas técnicas de clusterización con los 9 atributos del archivo Despachos Diesel 6 ATRIBUTOS.arff, pero los resultados fueron muy dispersos y requirió de un análisis adicional de la información con la finalidad de contar con un modelo mejor ajustado.

En este punto se regresa al Paso 3 Transformación de los datos, con la finalidad de obtener atributos cualitativos que permitan la aplicación de las técnicas de agrupamiento, tomando en cuenta las siguientes consideraciones:

- Sobre la base de que el mayor consumo por tipo de combustible es el DIESEL, se redujo la muestra a los despachos exclusivamente de este producto, así mismo de las estaciones de servicio a nivel nacional se sabe que tienen un tratamiento especial las cercanas a las fronteras por lo que se creó el atributo Ubicación en el cual se asigna a cada estación la condición de “fronteriza” y “no fronteriza”, ésta distribución se la realizó con ayuda del mapa político compilado por el Instituto Geográfico Militar, en el cual se marca sobre cada provincia el límite internacional con el signo convencional: [redacted] .

Figura 3.- Mapa Político del Ecuador

- El atributo condición cliente tiene datos numéricos que tienen un significado, es por ello que se optó por incluir el dato descriptivo de la condición, es así que:

Tabla 9.- Descripción de los códigos de la condición del cliente.

Condición Cliente	Descripción
1	ACTIVO
2	SUSPENDIDO
3	CESANTE
4	SUSPENDIDO POR NO RECIBIR DESPACHOS EN 3 MESES

Fuente: Sistema de Comercialización de EPPETROECUADOR

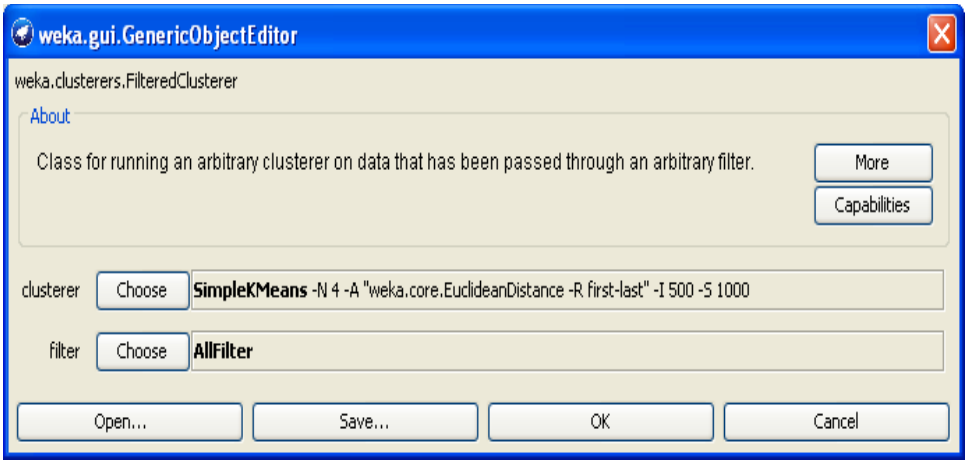
- El atributo “año/mes” se lo separo en “año” y “mes” con la finalidad de tomar el número de mes y transformarlo a nombre del mes, así entonces se trabajo con el campo “NomMes” al cual se le asigno las primeras tres letras de los nombres de cada mes, así por ejemplo en el mes de enero estará ENE.

Con estas transformaciones, se trabajó con los atributos “Condición del cliente”, “nombre de la estación de servicio”, “nombre la de provincia”, “ubicación”, “nombre del mes” y “volumen entregado” aplicándoles otras técnicas de filtro y clusterización pero esta vez se utilizó la opción de FilteredClusterer de la pestaña Clúster en el módulo Explorer de WEKA.

En este punto se retoma el paso 4 que tiene que ver con la aplicación del algoritmo seleccionado.

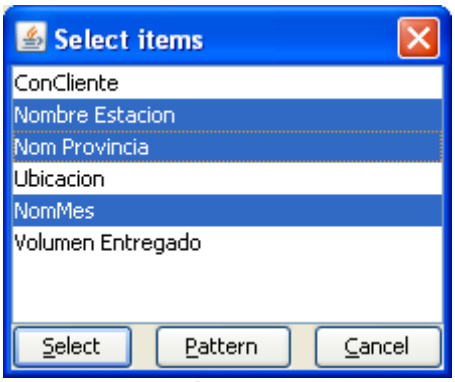
Con la opción de FilteredClusterer de WEKA se realiza la aplicación de filtros inmediatamente después de haberse aplicado la clusterización, evitando la aplicación manual de filtros y obteniendo resultados de filtros y clúster en un solo paso.

Los mejores resultados se los obtuvo con la siguiente configuración:



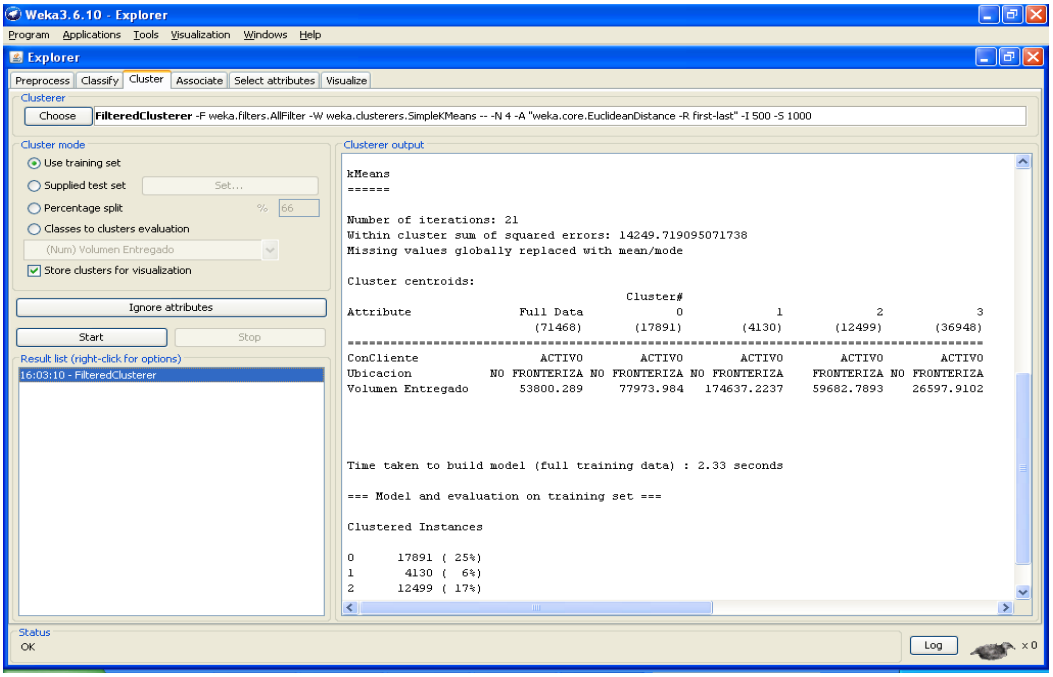
Pantalla 4.- Selección de los algoritmos de clusterización y filtro

Para el caso de la clusterización se aplicó el algoritmo SimpleKMeans y para el filtro la opción AllFilter como se muestra en la figura. Adicionalmente se excluyó los atributos de “Nombre Estación”, “Nombre Provincia” y “NomMes” para el procesamiento de la información, como se muestra a continuación.



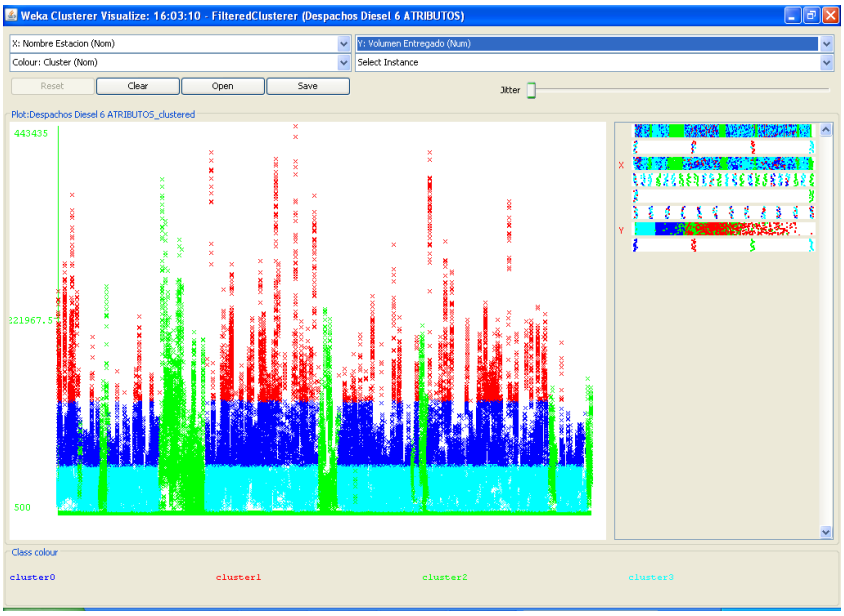
Pantalla 5.- Selección de atributos a ignorar

WEKA entrega un reporte de resultados luego del procesamiento del algoritmo.

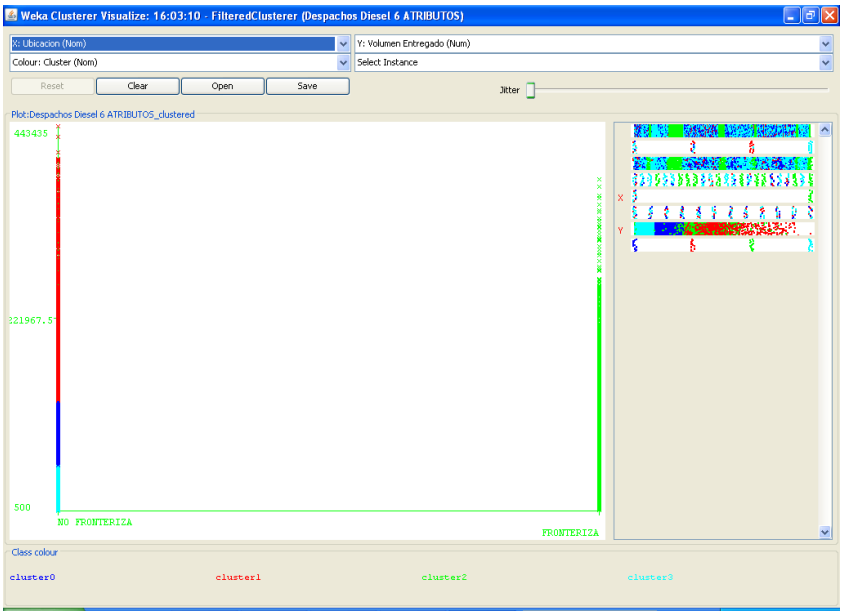


Pantalla 6.- Resultados numéricos de la clusterización

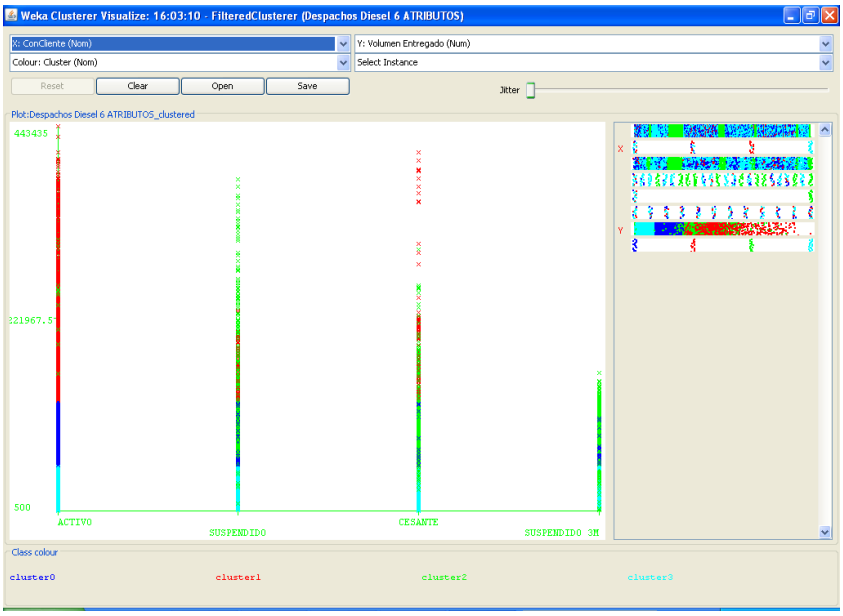
WEKA permite la visualización de los resultados obtenidos mediante diagramas como se muestra a continuación:



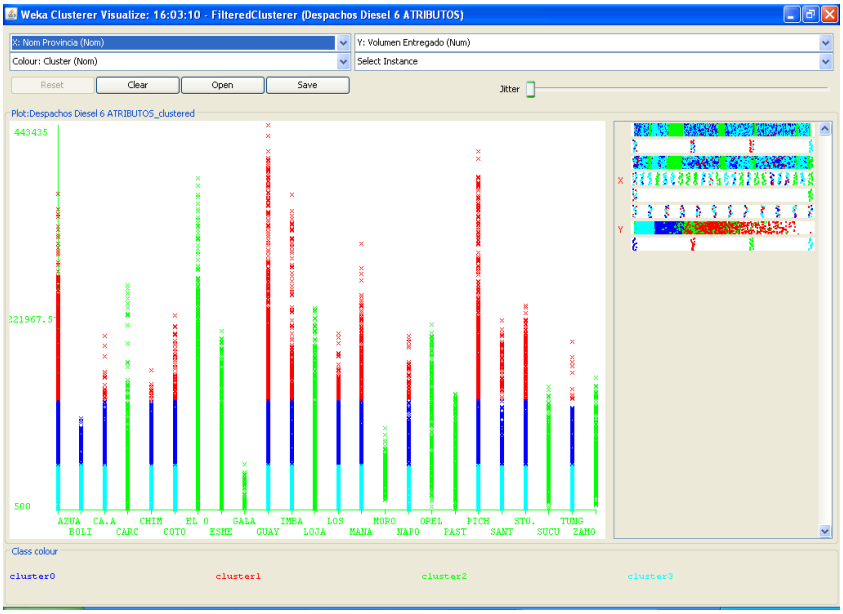
Pantalla 7.- Visualización de la relación Nombre de Estación vs Volumen Entregado luego de la clusterización



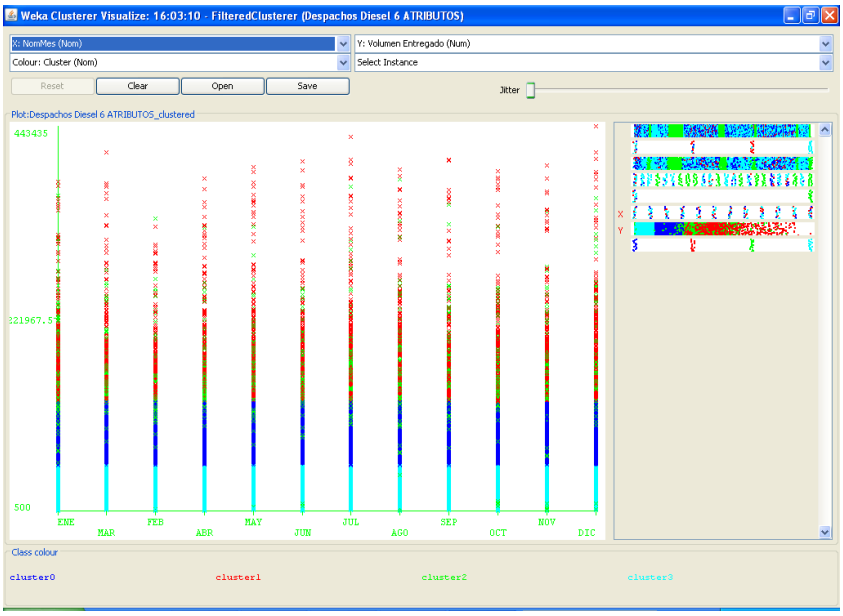
Pantalla 8.- Visualización de la relación ubicación vs Volumen Entregado luego de la clusterización.



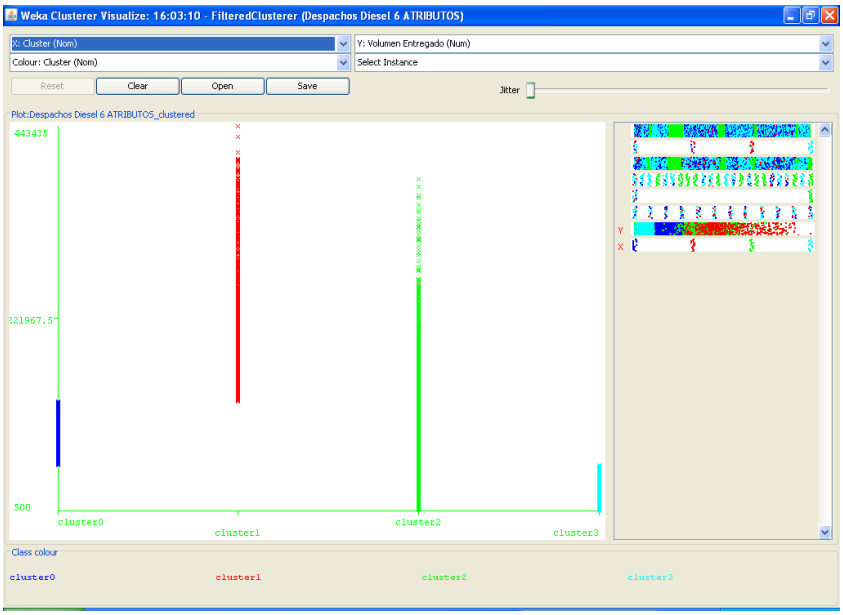
Pantalla 9.- Visualización de la relación Condición Cliente vs Volumen Entregado luego de la clusterización.



Pantalla 10.- Visualización de la relación Nombre de provincia vs Volumen Entregado luego de la clusterización.



Pantalla 11.- Visualización de la relación Nombre de mes vs Volumen Entregado luego de la clusterización.



Pantalla 12.- Visualización de la relación Nombre de mes vs Volumen Entregado luego de la clusterización.

De los 4 clústeres conseguidos se analiza los centroides identificándose que en el clúster 1 se agruparon un 6% de las estaciones de servicio con condición de no fronterizas y con un volumen de 174637.2 que resultaría muy superior al del resto de clústeres, lo que nos presupone un buen agrupamiento de acuerdo a la teoría inicial presentada en la que

expone que las provincias de frontera cuentan con controles direccionados a erradicar el contrabando de combustible, no así en las provincias centrales.

Tabla 10.- Resumen de los clústeres y centroides resultantes.

	Clúster#			
Atributo	0	1	2	3
ConCliente	ACTIVO	ACTIVO	ACTIVO	ACTIVO
Ubicacion	NO FRONTERIZA	NO FRONTERIZA	FRONTERIZA	NO FRONTERIZA
Volumen	77973.984	174637.2237	59682.7893	26597.9102
Instancias	25%	6%	17%	52%

Fuente y Elaborado por: la autora.

Paso 5: Análisis de Resultados.- interpretar resultados

Tomando los resultados arrojados por WEKA se efectúa una comparación con los datos obtenidos con el análisis Pareto a las relaciones volumen despachado, población y vehículos, presentado inicialmente.

Tabla 11.- Comparativo entre análisis previos y los obtenidos con WEKA.

UBICACIÓN	PARETO DE CONSUMO/POBLACION	PARETO DE CONSUMO/VEHICULOS	RESULTADOS WEKA
Fronteriza	CARCHI	CARCHI	
	EL ORO	EL ORO	
		ESMERALDAS	
	LOJA	LOJA	
	ORELLANA	ORELLANA	
	PASTAZA	PASTAZA	
	ZAMORA CHINCHIPE	ZAMORA CHINCHIPE	
	SUCUMBIOS	SUCUMBIOS	
No Fronteriza	AZUAY		AZUAY
		BOLIVAR	
	LOS RIOS	LOS RIOS	LOS RIOS
	CAÑAR		CAÑAR
		MANABI	MANABI
	CHIMBORAZO	CHIMBORAZO	CHIMBORAZO
	COTOPAXI		COTOPAXI
	IMBABURA	IMBABURA	IMBABURA
	NAPO	NAPO	NAPO
		SANTA ELENA	SANTA ELENA
	STO. DGO. TSACHILAS	STO. DGO. TSACHILAS	STO. DGO. TSACHILAS
	PICHINCHA		PICHINCHA
	TUNGURAHUA		TUNGURAHUA
			GUAYAS

Fuente y Elaborado por: la autora.

De la comparación se determina que de las 13 provincias no fronterizas agrupadas por WEKA 12 se encontraban ya clasificadas con técnicas y atributos diferentes, lo que permite colegir que se consiguió cumplir con los objetivos de la investigación.

En lo referente a la provincia del Guayas se conoce de los análisis anteriores que sus consumos están relacionados al número de vehículos y al número de habitantes, y en cuanto a la provincia de Bolívar es de bajo consumo con respecto a las 11 clasificadas por lo que no fue posible su agrupamiento.

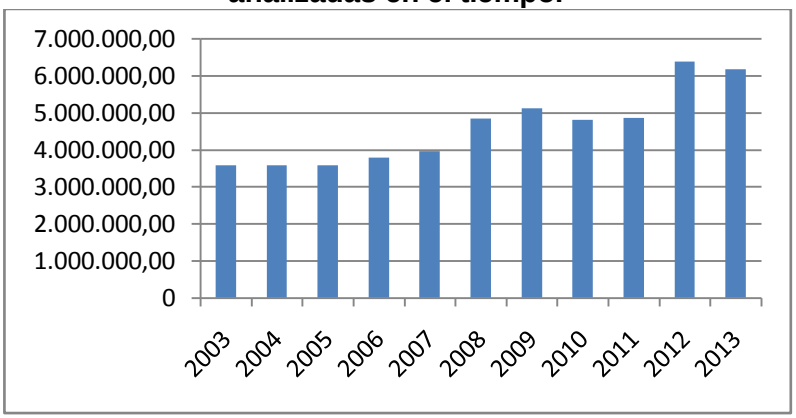
Paso 6: Asimilación del Conocimiento.- aplicación de los resultados

Analizando los resultados obtenidos a detalle se toma los datos de las gasolineras agrupadas en el cluster1 se detecta la existencia de 3 gasolineras con el mismo nombre lo cual llama la atención puesto que las tres están identificadas como de alto consumo.

Para conocer un poco más sobre estas gasolineras se obtuvo más información sobre ellas determinándose que 2 de las gasolineras se encuentran en el cantón Aloag y están registradas con el mismo RUC, la tercera gasolinera está ubicada en otro cantón y no tiene relación con las otras dos.

Se analiza los despachos a las dos gasolineras desde el año 2003, obteniéndose que en el tiempo a esta empresa se le han incrementando sus despachos aproximadamente al doble desde su establecimiento en el año 2003.

Tabla 12.- Comportamiento de despachos de combustible a las gasolineras analizadas en el tiempo.



Fuente: Sistema de Comercialización de EPPETROECUADOR

El presente análisis es factible de realizarlo a las 159 gasolineras agrupadas por la metodología propuesta, misma que inicialmente presupone contar con resultados óptimos.

Los resultados entregados en la presente investigación deberán ser comprobados posteriormente por los entes de control designados para el efecto.

CONCLUSIONES.

El desarrollo del presente trabajo ha contribuido de manera importante a la actividad de lucha contra el contrabando, puesto que al caracterizar y modelar la red de comercialización del segmento automotriz utilizando técnicas de inteligencia artificial, se ha logrado utilizar herramientas informáticas para la detección de posibles ilícitos.

La metodología aplicada permitió la categorización automática de los datos de despachos a las gasolineras con la ayuda del algoritmo de agrupamiento K-medias, porque de los cuatro clústeres obtenidos como resultado de la aplicación del algoritmo, en el clúster 1 se agruparon las gasolineras con mas alto consumo.

En la etapa de preparación de datos se tuvo que trabajar en la clasificación de las gasolineras de acuerdo a su ubicación en forma manual, debido a que, al trabajar con 9 atributos y en su mayoría numéricos los resultados fueron muy dispersos y dificultaban el análisis, por ello se tomó como guía el mapa político del Ecuador compilado por el Instituto Geográfico Militar, donde se marcan a las provincias fronterizas con una nomenclatura diferente, este esfuerzo adicional permitió cumplir con el objetivo.

En las etapas de análisis de resultados y asimilación del conocimiento se logra determinar que los datos obtenidos en el clúster 1 pueden clasificarse como de consumos anómalos, pues de acuerdo con el análisis pormenorizado realizado a tres gasolineras de este grupo se encontró que 2 de ellas tienen relación y son claves para el inicio de una investigación especializada a través de las demás entidades de control.

RECOMENDACIONES.

1. Es recomendable proponer la inclusión del campo ubicación en la base de datos del sistema de comercialización para garantizar la calidad de los datos.

2. Se recomienda la investigación especializada a las 159 gasolineras clasificadas con la metodología de este trabajo, para que de alguna manera se consigan resultados de mayor incidencia y se marque un precedente para evitar el mal uso y robo de los combustibles subsidiados que perjudican al país en grandes proporciones.

3. Se recomienda la aplicación de la metodología propuesta para el estudio de los consumos anómalos en otros segmentos del mercado que son críticos en el Ecuador por su condición de subsidiados como son el segmento Pesquero Artesanal, Naviero Nacional y GLP doméstico.

4. La metodología de extracción de conocimiento permite la selección de otras herramientas y algoritmos por lo sería adecuado realizar investigaciones similares para la obtención de mejores resultados.

5. Se recomienda dar continuidad a la presente investigación para lograr obtener un sistema experto que permita la total automatización de la extracción del conocimiento de la comercialización de combustibles en el Ecuador.

BIBLIOGRAFÍA DE CONSULTA

1. Delgado, R. (2008). El Diagrama de Pareto. *Revista Virtual de la Universidad Católica de Occidente* , 47,48,49.
2. Eumed.net. (s.f.). *Enciclopedia Virtual*. Recuperado el 5 de 11 de 2013, de http://www.eumed.net/diccionario/definicion.php?dic=1&def=137
3. Fayyad, P.-S. a. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine* , 40,41.
4. Hernandez, J., & y, o. (2005). *Introducción a la Minería de Datos.* España: Pearson Educación S.A.
5. IGM. (s.f.). *Instituto Geográfico Militar*. Recuperado el 12 de 10 de 2013, de http://www.igm.gob.ec/work/index.php?option=com_content&view=article&id=84%3Amapa-del-ecuador-uso-escolar-escala-14000000&catid=41&Itemid=99
6. INEC. (s.f.). *Instituto Nacional de Estadísticas y Censos*. Recuperado el 05 de 08 de 2013, de http://www.inec.gob.ec/estadisticas/
7. Knight, E. R. (1994). *Inteligencia Artificial.* España: McGraw-Hill.
8. Orange. (s.f.). *Donald Michie*. Recuperado el 15 de 07 de 2013, de http://orange.biolab.si
9. Rusell, S. (2010). *Artificial intelligence A Modern Approach, Third Edition.* New Jersey: PEARSON.
10. SC. (s.f.). *Superintendencia de Compañìas*. Recuperado el 5 de 11 de 2013, de http://www.supercias.gov.ec/consultas/inicio.html
11. SRI. (s.f.). *Servicio de Rentas Internas*. Recuperado el 5 de 11 de 2013, de http://www.sri.gob.ec/web/guest/home
12. UC3M. (s.f.). *Universidad Carlos III de Madrid*. Recuperado el 10 de 11 de 2013, de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema5am.pdf
13. WEKA. (s.f.). *Javadoc*. Recuperado el 21 de 10 de 2013, de http://weka.sourceforge.net/doc.stable/
14. WEKA. (s.f.). *The University of Waikato*. Recuperado el 05 de 07 de 2013, de http://www.cs.waikato.ac.nz/ml/weka/downloading.html.