UNIVERSIDAD TÉCNICA DEL NORTE

Facultad de Ingeniería en Ciencias Aplicadas

Carrera de Ingeniería en Sistemas Computacionales

DETECCIÓN DE PATRONES DE CONTRABANDO PARA LA GESTIÓN DE APREHENSIONES Y RETENCIONES, UTILIZANDO TÉCNICAS PREDICTIVAS DE CLASIFICACIÓN Y REGRESIÓN DE MINERÍA DE DATOS.

Trabajo de grado previo a la obtención del título de Ingeniero en Sistemas Computacionales.

Autor:

Tommy Bryan Mancero Menoscal

Director:

PhD. Iván Danilo García Santillán

Ibarra – Ecuador

Abril, 2020

UNIVERSIDAD TÉCNICA DEL NORTE
BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

1. IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

DATOS DE CONTACTO			
CÉDULA DE IDENTIDAD:	1003973813		
APELLIDOS Y NOMBRES:	Mancero Menoscal Tommy Bryan		
DIRECCIÓN:	Ibarra, La Victoria, Calle Manuel Zambrano, Pasaje L, Casa 4-23		
EMAIL:	tbmancerom@utn.edu.ec		
TELÉFONO FIJO:	062615006	**TELÉFONO MÓVIL:**	0994658706

DATOS DE LA OBRA	
TÍTULO:	Detección de patrones de contrabando para la gestión de aprehensiones y retenciones, utilizando técnicas predictivas de clasificación y regresión de minería de datos
AUTOR (ES):	Mancero Menoscal Tommy Bryan
FECHA: DD/MM/AAAA	15/01/2021
SOLO PARA TRABAJOS DE GRADO	
PROGRAMA:	■ **PREGRADO** ☐ **POSGRADO**
TITULO POR EL QUE OPTA:	INGENIERÍA EN SISTEMAS COMPUTACIONALES
ASESOR /DIRECTOR:	PhD. Iván García

2. CONSTANCIAS

El autor (es) manifiesta (n) que la obra objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto, la obra es original y que es (son) el (los) titular (es) de los derechos patrimoniales, por lo que asume (n) la responsabilidad sobre el contenido de la misma y saldrá (n) en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, a los 15 días del mes de enero de 2021

EL AUTOR:

(Firma).....
Nombre: Mancero Menoscal Tommy Bryan

CUERPO DE VIGILANCIA ADUANERA

CENTRO DE FORMACION DE VIGILANCIA ADUANERA

CERTIFICA

QUE: El Sr. TOMMY BRYAN MANCERO MENOSCAL con cédula identidad 1003973813 estudiante de la Universidad Técnica del Norte de la Facultad de Ingeniería en Ciencias Aplicadas de la Carrera de Ingeniería en Sistemas Computacionales, ha desarrollado con los datos entregados por la Dirección Administrativa de la Aduana el Proyecto de Tesis **"DETECCIÓN DE PATRONES DE CONTRABANDO PARA LA GESTIÓN DE INFORMACIÓN DE APREHENSIONES Y RETENCIONES UTILIZANDO TÉCNICAS PREDICTIVAS DE CLASIFICACIÓN Y REGRESIÓN DE MINERÍA DE DATOS".**

QUE: El análisis del proyecto fue entregado al Centro de Formación de Vigilancia Aduanera acantonada en la ciudad de San Miguel de Ibarra el 14 de diciembre del 2020.

Es todo cuanto puedo certificar, facultando al Interesado hacer uso de este certificado como estime conveniente, excepto para trámites judiciales.

Ibarra, 14 de diciembre del 2020

Atentamente

Cnel. Mba. Edgar Duque Ch
Coordinador (E) Centro de Formación de Vigilancia Aduanera

CERTIFICADO TUTOR

En mi calidad de tutor de Trabajo de Grado presentado por el egresado **TOMMY BRYAN MANCERO MENOSCAL** para obtener Título de Ingeniería en Sistemas Computacionales cuyo tema es: **Detección de patrones de contrabando para la gestión de aprehensiones y retenciones, utilizando técnicas predictivas de clasificación y regresión de minería de datos.** Considero que el presente trabajo reúne los requisitos y méritos suficientes para ser sometido a la presentación pública y evaluación por parte del tribunal examinador que se designe.

En la ciudad de Ibarra, a los 15 días del mes de diciembre del 2020

Firmado electrónicamente por:
IVAN DANILO
GARCIA
SANTILLAN

PhD. Iván Danilo García Santillán

DIRECTOR DE TRABAJO DE GRADO

Dedicatorias

Este trabajo de titulación va dedicado a mi familia, la cual me ha apoyado en todo el transcurso de mi vida.

Agradecimientos

Mis más sinceros agradecimientos a mi familia y amigos por apoyarme durante todo el transcurso de titulación y por ayudarme a crecer como persona, de igual manera a los Ingenieros que conforman la carrera de Ingeniería en Sistemas Computacionales quienes han impartido conocimientos y valores éticos, ayudando a crecer profesionalmente y en el ámbito humano.

Tabla de contenido

Contenido

Índice de Figuras

Índice de Tablas

Resumen

El contrabando en el ecuador es un problema que afecta al comercio y a la economía ecuatoriana; teniendo como causas probables que ocasionan esta problemática pueden ser la falta de empleo, el alto precio de los productos, falta de educación y situación socioeconómica desfavorable. El propósito de esta investigación es tratar de reducir los niveles de ingreso de mercadería de contrabando, mediante la obtención de patrones de contrabando y los principales factores que contribuyen a esta problemática en el Ecuador, aplicando técnicas predictivas de minería de datos (clasificación y regresión), a datos históricos (2015-2020) provenientes de la entidad aduanera. Siguiendo el proceso KDD (Proceso de descubrimiento de conocimiento) el cual sirvió para la obtención de una vista minable, al cual se le pudo aplicar modelos de regresión y árboles de decisión en la herramienta KNIME y WEKA. Para la selección del mejor algoritmo se evaluó cuantitativamente cada uno de ellos, mediante las métricas estadísticas que muestra cada algoritmo como resultado, demostrando que el algoritmo J48 y Logistic Regression son los mejores algoritmos para compararlos y así obtener el conocimiento.

Palabras clave: minería de datos, descubrimiento de patrones, técnicas predictivas, aprehensiones y retenciones

Abstract

Smuggling in Ecuador is a problem that affects commerce and the ecuadorian economy; having as probable causes that cause this problem can be the lack of employment, the high price of the products, lack of education and unfavorable socioeconomic situation. The purpose of this research is to try to reduce the levels of entry of contraband merchandise, by obtaining smuggling patterns and the main factors that contribute to this problem in Ecuador, applying predictive data mining techniques (classification and regression), to historical data (2015-2020) from the customs entity. Following the KDD process (Knowledge discovery process) which served to obtain a mineable view, to which regression models and decision trees could be applied in the KNIME and WEKA tools. For the selection of the best algorithm, each one of them was quantitatively evaluated, using

the statistical metrics that each algorithm shows as a result, showing that the J48 algorithm and Logistic Regression are the best algorithms to compare them and thus obtain knowledge.

Keywords: data mining, pattern discovery, predictive techniques, apprehensions and retentions

Introducción

Antecedentes

El alto precio de los productos y la falta de empleo son algunos de los factores principales por los cuales ocurre el contrabando. Evadir controles y pago de impuestos para comercializar productos con altas ganancias es el clásico delito en las fronteras del Ecuador por la falta de sanciones (Chiriboga et al., 2015).

El contrabando es un fenómeno que se ha generalizado en la cultura ecuatoriana, del cual el Estado no recibe su parte correspondiente por este concepto (impuestos). El país se enfrenta a grandes grupos de poder que han intervenido de manera rutinaria en negocios de apariencia lícita, basados en el contrabando (Chiriboga et al., 2015).

Según el Servicio Nacional de Aduanas (SENAE) las recaudaciones aduaneras en el traspaso de 2015 a 2016 descendieron un 28% (SENAE, 2016).

El 60% de resoluciones tomadas en el Ecuador son de facilitación y simplificación de procesos. En América Latina del 100% de mercancía entrante, el 15% es retenido (impedir que salga) o aprehendida (capturar), mientras que el 85% pasa sin obstáculos. En Ecuador se retenía el 51% de la mercadería, pero ahora se ha reducido al 26% (Muñoz, 2019).

KDD (Knowledge Discovery in Databases) es una metodología genérica para encontrar información en un gran conjunto de datos y con ello generar conocimiento. Consta de varias etapas en las cuales funciona de forma iterativa (regresar entre etapas) e interactiva (Martínez, 2012).

Problema

El contrabando es un fenómeno que se ha generalizado y ha ido en aumento en nuestra cultura, en nuestro modo de vida, es legitimado en los mecanismos de intercambio comercial que sustentan

la cultura nacional (Chiriboga et al., 2015). El bajo grado de análisis de los datos provenientes del contrabando en el Ecuador, no son suficientes para encontrar patrones detallados de contrabando que ayuden a reducir este comportamiento.

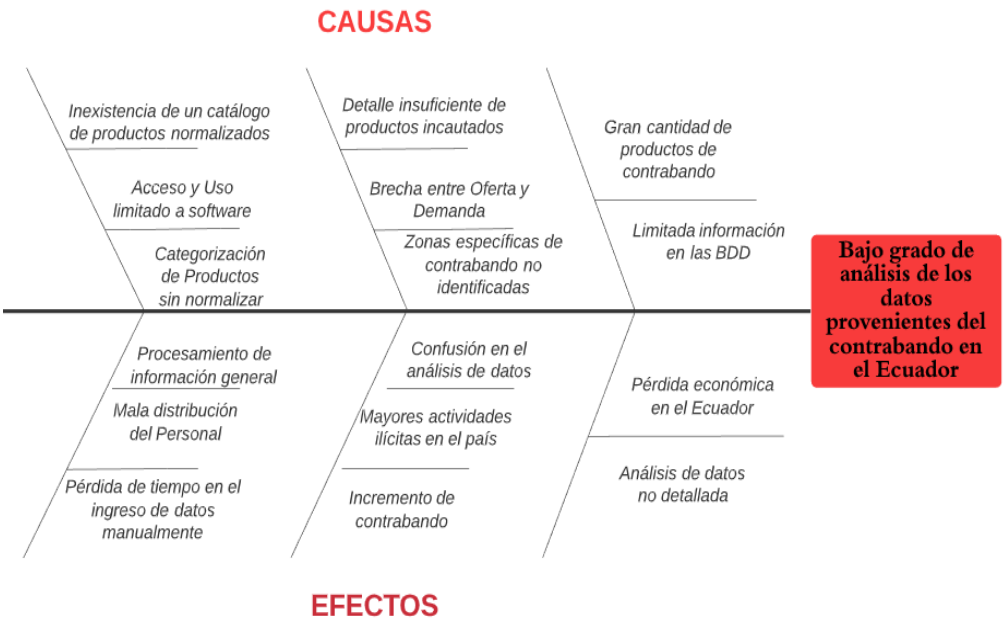


Figura 1 Árbol de Problemas (Elaboración Propia)

Objetivos

Objetivo General

Identificar patrones de contrabando para la gestión de aprehensiones y retenciones utilizando técnicas predictivas de clasificación y regresión en minería de datos.

Objetivos Específicos

Elaborar un marco teórico sustentando técnicas predictivas de minería de datos y el proceso de descubrimiento de conocimiento en base de datos (KDD).

Construir un conjunto de datos y vista minable a partir de datos de aprehensión y retención de un periodo de tiempo histórico basados en el proceso KDD.

Aplicar técnicas predictivas de clasificación y regresión a la vista minable utilizando la herramienta Weka.

Validar los resultados mediante las métricas estadísticas y la característica de Consistencia de la norma ISO/IEC 25012 para la calidad de los datos obtenidos.

Justificación

ODS

Para fortalecer Los Objetivos de Desarrollo Sostenible, en especial el # 16 “Plan, Justicia e Instituciones Sólidas” (UNODC, 2019), la meta #16.4 que hace referencia a reducir significativamente las corrientes financieras y de armas ilícitas, fortalecer la recuperación y devolución de los activos robados, además de luchar contra todas las formas de delincuencia organizada, entre estas el contrabando.

Justificación Tecnológica

Promover el uso de servicios de analítica (tecnologías emergentes) de grandes volúmenes de datos, que permitan la toma de decisiones para mejorar la gestión de los gobiernos, industrias, academias y ciudadanía (MINTEL, 2018).

Justificación Ambiental

El contrabando de productos comestibles que afectan al sector agrícola el cual se encuentra en una competencia infiel con productos que no cumplen las normas que se requieren para su ingreso al país (SENAE, 2018), identificando estos productos se evitara el desperdicio de material o productos que perjudiquen al medio ambiente. Los ambientes naturales del contrabando son

zanjas, trochas y caminos de verano no reconocidos (Chiriboga et al., 2015), con la identificación de estas zonas se disminuirá la alteración al medio ambiente.

Justificación Social

El traficante es quien dicta las reglas del comercio fronterizo, la extorsión, el sicariato y el secuestro para proteger su negocio (Chiriboga et al., 2015), identificar este comportamiento, permitirá a la organización de control de comercio interno actuar de una manera rápida y eficaz en la lucha contra los contrabandistas.

Identificar los comportamientos de contrabando más habituales en el País, significa un aporte importante para incrementar las aprehensiones y retenciones de contrabando en el país. Este es el motivo por el que el uso de minería de datos resulta fundamental, debido a que su análisis y aplicación ha tenido un impacto significativo en los últimos años, puesto que el uso de sus técnicas permite, entre otras cosas, prever cualquier hecho dentro del ámbito de investigación (Lara, 2014). En la siguiente investigación se realizará un estudio que nos permita identificar patrones de contrabando en el país, utilizando datos de un periodo de tiempo (2014 a 2018), otorgados por una entidad de control de comercio interno. Mediante técnicas predictivas de clasificación y regresión en minería de datos se obtendrá información que será utilizada por las organizaciones de control de contrabando encargadas de tomar decisiones acertadas con el fin de reducir de manera eficaz el contrabando y actividades ilícitas de este tipo en el país, reduciendo las pérdidas económicas para el estado ecuatoriano, mejorando la seguridad y la comercialización en el mercado local.

Alcance

Mediante la siguiente investigación se obtendrán patrones de contrabando que permitirán identificar de una manera más eficiente y rápida los comportamientos de contrabando más usados

en el Ecuador. El estudio consiste en realizar un análisis de datos de aprehensión y retención, aplicando el proceso de descubrimiento de conocimientos KDD, en el cual se realiza la extracción, transformación, y limpieza de los datos utilizando la herramienta open source Pentaho 7 en un ambiente local, posteriormente se ejecutará los algoritmos predictivos de clasificación y regresión utilizando la herramienta Weka 8.3 de igual manera en un ambiente local. El resultado obtenido será presentado en forma de reportes (detallando la ubicación, fecha, tiempo producto, entre otras) los cuales facilitarán la gestión de la información de aprehensiones y retenciones, se procederá a comparar los datos obtenidos mediante las métricas cuantitativas apropiadas para evaluar los algoritmos de minería de datos, y deducir cuál es el que mejor se acopla al estudio que permita obtener el conocimiento buscado. Se aplicará métodos estadísticos para evaluar los modelos generados y validar la información, con ayuda de la característica de consistencia de la norma ISO/IEC 25012 se analizará la calidad de los datos resultantes. Adicionalmente, se presentará un nuevo modelo de registro de los datos de aprehensiones y retenciones más detallado.

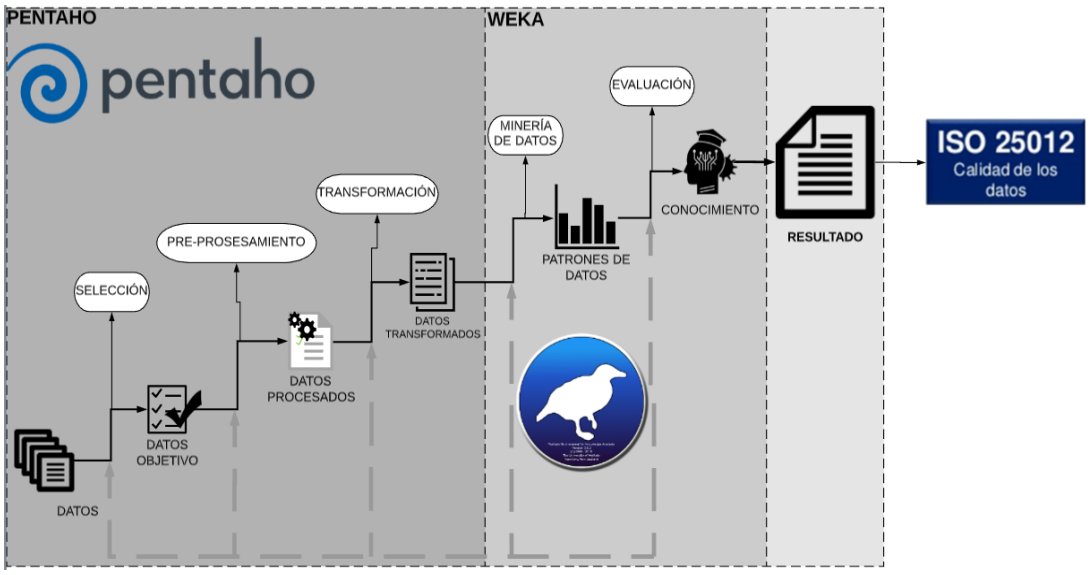


Figura 2 Adaptación Proceso KDD (Martínez, 2012)

CAPÍTULO 1

Marco teórico

1.1. Contrabando en el Ecuador

La palabra contrabando es de raíz hispánica y se compone por ‘contra’: en contra de y por ‘bando’: ley, disposición gubernamental o bando; se generó inicialmente para significar la producción y el comercio de mercancías que se realizaban en contra de las leyes emitidas por el Estado (Peña Cuervo et al., 2018).

El contrabando en el Ecuador se remonta a la época colonial, desde entonces ha ido evolucionando y modernizándose, debido a que quienes se encuentran inmersos en este aplican nuevas modalidades y estrategias de evasión de controles (Méndez, 2015).

El centro Aduanero tiene como fin la reducción del contrabando, facilitando el comercio exterior y ejercer el control de la entrada y salida de mercancía, unidades de carga y medios de transporte por las fronteras y zonas aduaneras de la República, así como quienes efectúen actividades directa o indirectamente relacionadas con el tráfico internacional de mercancías (Normativa & CÓDIGO, 2018).

En la Figura 3 se muestra los resultados de las aprehensiones realizadas por el SENAE en el año 2019 en millones de dólares causadas de la lucha contra el contrabando.

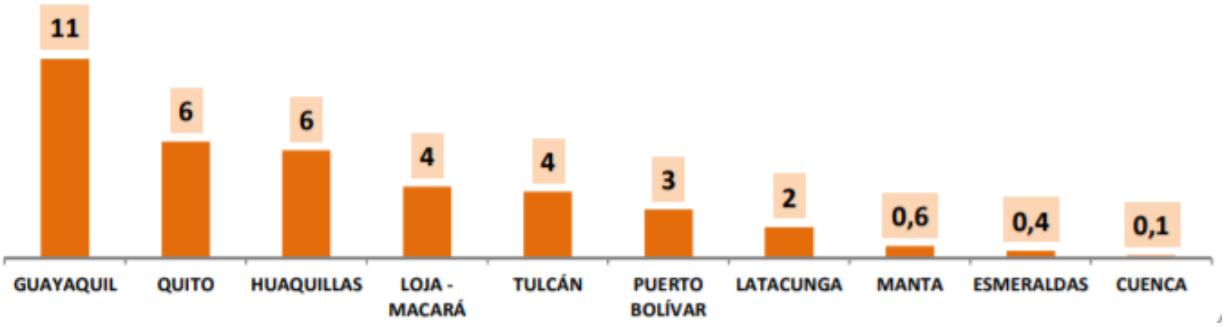


Figura 3 Aprehensiones Totales 2019 (SENAE, 2019)

En la Figura 4 se muestra en millones de dólares los principales productos aprehendidos en las diferentes zonas de control del ecuador.

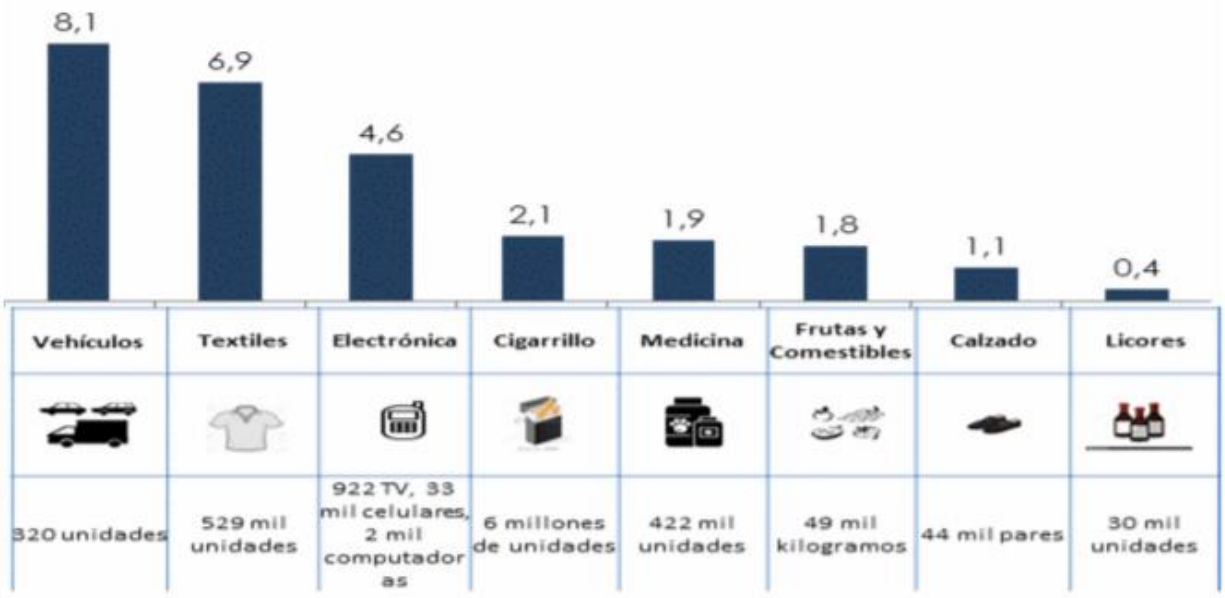


Figura 4 Principales productos aprehendidos 2019 (SENAE, 2019)

1.2. Introducción a la minería de datos

Debido al gran avance tecnológico que se ha venido dando al pasar el tiempo hasta la actualidad, las empresas han sido capaces de almacenar un gran volumen de datos los cuales puedes ser analizados para llegar a obtener conocimiento.

Fuera del ámbito informático la "minería" se refiere al arte de extraer material precioso de la corteza terrestre y "dato" hace referencia a un valor que toma una variable, atributo, etc., Existen disciplinas capaces de estudiar los datos, su transformación, obtención, almacenamiento, etc. El origen sucede desde 1960 debido a los estadísticos que utilizaron términos de *data fishing* o *data dredging* para el análisis de datos, en 1990 estos términos quedan atrás debido a que la concepción de los datos comenzó a cambiar, dando paso a la minería de datos que nace de la unión de ambos términos y es una disciplina que estudia el análisis de grandes cantidades de datos con el objetivo de obtener conocimiento (Lara, 2014).

Se denomina minería de datos al conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos (Perversi, 2007).

La minería de datos forma parte un proceso llamado KDD (Descubrimiento de conocimientos en bases de datos), en el cual la minería de datos se centra en la obtención de modelos (Luis Paulo Vieira Braga, Luis Iván Ortiz Valencia, 2009). Un modelo es una representación simplificada que intenta explicar un patrón/comportamiento en los datos (Landa, 2016)

Sí mismo, la minería de datos puede ser definida como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al analizar grandes cantidades de datos (López, 2007).

1.2.1. Relación con otras áreas

La minería de datos surge como la mezcla de conceptos procedentes de otras disciplinas, en la Figura 5 se detallan algunas de las disciplinas relacionadas con la minería de datos.

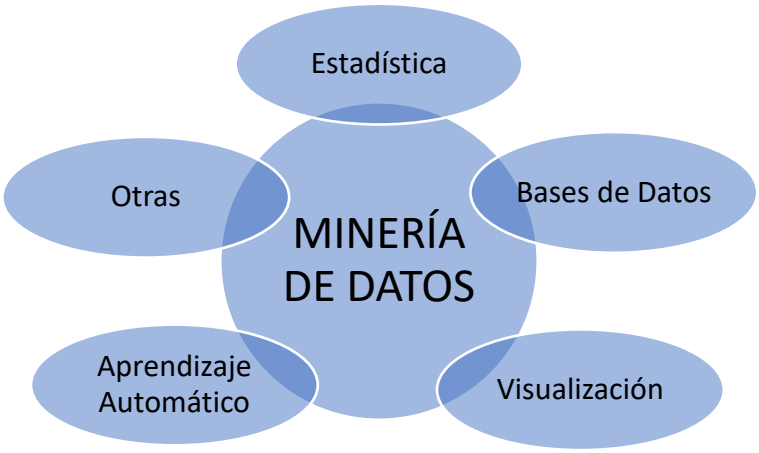


Figura 5 Relación de la Minería de Datos con otras disciplinas (Lara, 2014)

Estadística

Muchas técnicas y conceptos que utiliza la minería de datos tienen su base en la estadística, se podría decir que esta disciplina es la "madre" de la minería de datos (Lara, 2014). La estadística es una disciplina muy importante para la minería de datos debido al gran papel que cumplen sus conceptos y métodos en esta área.

Bases de Datos

Como ya se ha mencionado el proceso KDD es la obtención de conocimiento a partir de datos que se encuentran almacenados en bases de datos, los cuales son procesados para después ser analizados.

Visualización

Hace énfasis en la relación de la interfaz entre el usuario y los datos, y usuario y patrones (Barbosa, 2009). La visualización representa los resultados utilizando técnicas de visualización que permitan al usuario entender los datos resultantes.

Aprendizaje Automático

El aprendizaje automático o Machine Learning es un subcampo de la Inteligencia Artificial encargado de estudiar el problema de aprendizaje de las máquinas, su objeto de estudio es el problema de cómo las máquinas pueden adquirir el conocimiento que las capacite para resolver problemas determinados (Bello et al., 2008).

Persigue la obtención de modelos mediante mecanismos automáticos (Lara, 2014) y hace referencia al proceso de mejorar la habilidad de análisis de la máquina.

Otras

Debido al gran avance tecnológico y al gran volumen de datos con el que cuentan las empresas, la minería de datos o *data Mining* es utilizado en casi todas las disciplinas. Algunas áreas adicionales con las que está relacionada con sistemas de apoyo a la toma de decisiones, recuperación de información, tratamiento y procesamiento de señales, entre otras (Lara 2014).

1.2.2. Tipos de datos y bases de datos

Como se ha mencionado anteriormente, dato es una representación simbólica que toma una característica de un objeto o entidad el cual es almacenados en base de datos. En la Tabla 1 se presenta un ejemplo pequeño de datos que representan características de un equipo de futbol.

Tabla 1

Ejemplo de datos asociados a un Equipo de Futbol

ATRIBUTO	DATO
Nombre del Equipo	Barcelona
Ciudad de Origen	Guayaquil
Siglas	B.S.C.

Fuente Propia

Según Lara, 2014 los datos se dividen según la naturaleza del atributo que representan, a efecto de minería de datos se clasifican en:

1.2.2.1. Cuantitativos

Son datos que representan magnitudes o cantidades y se los puede denominar como:

- **Discretos:** pueden tomar cierta cantidad de valores.
- **Continuos:** un valor intermedio que se encuentra en un par de valores.

1.2.2.2. Cualitativos

Al contrario de los datos cuantitativos, los datos cualitativos representan una categoría y se los puede denominar en dos tipos:

- **Nominales:** son datos cuya asignación aleatoria de números o símbolos.
- **Ordinales:** son datos en los cuales se encuentran debidamente ordenados por una relación de orden existente.

1.2.2.3. Datos no Convencionales

La mayoría de las técnicas están pensadas para trabajar con datos convencionales como los presentados anteriormente; sin embargo, existen otros tipos de datos los cuales son:

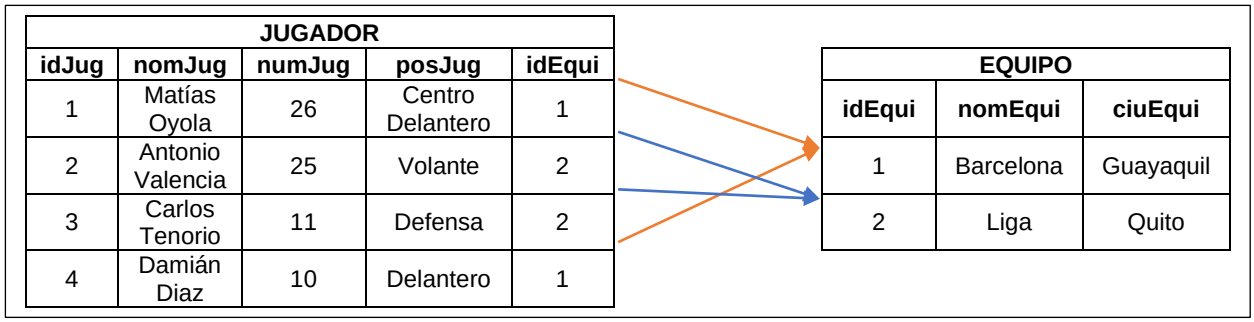
- **Series Temporales:** son sucesiones de valores evolucionados a lo largo del tiempo que toma una característica.
- **Datos Espaciales:** utilizados para la representación en 3D de un objeto.
- **Datos multimedia:** son muy conocidos debido a que son las imágenes, videos, audios, entre otros.
- **Documentos:** son descripciones textuales de un objeto
- **Datos procedentes de la web:** información encontrada sobre la estructura de los sitios web, patrones de navegación, entre otros (Lara, 2014).

1.2.3. Bases de Datos

Por lo general este tipo de datos se encuentra almacenado en bases de datos que habitualmente son relacionales. Sin embargo, la minería de datos distingue entre dos tipos de bases de datos las cuales son relacionales y desnormalizadas.

1.2.3.1. Bases de datos relacionales

Las bases de datos relacionales generalmente se basan en tablas relacionadas, que cuentan con filas (conjunto de instancias), las cuales pueden tener una clave primaria la que permitirá la relación con otras tablas, y columnas (atributos) que contienen datos relativos (Sanchez, 2004). En la Fig. 6 se presenta un pequeño en el que cada tabla posee un identificador (clave primaria) que identifica cada instancia y con la cual se va a establecer la relación entre ambas tablas.



JUGADOR				
idJug	**nomJug**	**numJug**	**posJug**	**idEqui**
1	Matías Oyola	26	Centro Delantero	1
2	Antonio Valencia	25	Volante	2
3	Carlos Tenorio	11	Defensa	2
4	Damián Diaz	10	Delantero	1

EQUIPO		
idEqui	**nomEqui**	**ciuEqui**
1	Barcelona	Guayaquil
2	Liga	Quito

Figura 6 Ejemplo de tablas de bases de datos relacional (Fuente Propia)

1.2.3.2. Bases de datos desnormalizadas

Para realizar minería de datos no es necesario contar con esquemas que cuenten con las características anteriores (relacional), sino que se puede utilizar otro tipo de esquemas como las bases de datos desnormalizadas, las cuales contiene datos duplicados o redundantes (Lara, 2014). En la Figura 7 las tablas de jugador y equipo se unen lo que provoca información duplicada.

JUGADOR					
idJug	**nomJug**	**numJug**	**posJug**	**nomEqui**	**ciuEqui**
1	Matías Oyola	25	Centro Delantero	Barcelona	Guayaquil
2	Antonio Valencia	1	Volante	Liga	Quito
3	Carlos Tenorio	11	Defensa	Liga	Quito
4	Damián Diaz	10	Delantero	Barcelona	Guayaquil

Figura 7 Ejemplo de tablas de bases de datos desnormalizada (Elaboración propia)

1.2.4. Aplicaciones

En la actualidad la minería de datos tiene una gran variedad de aplicaciones, en distintas áreas como medicina, inteligencia artificial, detección de fraudes, entre otras (Asencios, 2004).

1.3. Proceso de descubrimiento del conocimiento (KDD)

El proceso KDD surge de la necesidad de analizar grandes volúmenes de información almacenada en bases de datos (SQL, MySQL, EXCEL, ACESS, entre otras). Proporciona la capacidad de descubrir información nueva y significativa usando los datos existentes y la minería de datos (García Molina, 2007).

La extracción no trivial de información potencialmente útil a partir de un gran volumen de datos, en el cual la información está implícita, donde se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones, para conseguirlo harán falta técnicas de aprendizaje, estadística y bases de datos (Asencios, 2004). KDD es el proceso de extracción automatizada de conocimientos a partir de grandes cantidades de datos (Lara, 2014).

KDD no es un campo aislado, sino la correlación de otros campos, las áreas principales apoyan el aprendizaje automático, las bases de datos y la estadística. Cada una de ellas aporta una serie de técnicas y herramientas, aplicadas correctamente dan como resultado un modelo de conocimiento (Tuya et al., 2007).

El proceso KDD consiste en emplear a una determinada base de datos operaciones de: selección, exploración, muestreo, transformación y métodos de modelado, para que posterior a este proceso se obtendrá patrones que serán evaluados y de igual forma obtener el conocimiento, el cual es el objetivo principal de dicho proceso (Cisneros, 2019).

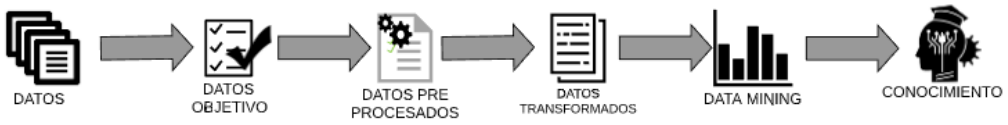


Figura 8 Adaptación Proceso KDD (Luis Paulo Vieira Braga, Luis Iván Ortiz Valencia, 2009)

El conocimiento adquirido al final de este proceso debe caracterizarse por ser válido, novedoso, potencialmente útil y comprensible (Montero Navarro, 2009).

La minería de datos, como se muestra en la Fig. 6, es una fase esencial de un proceso aun mayor llamado KDD el cual trata de descubrir conocimientos partiendo de cierta cantidad de datos, que serán pre procesados, transformados y analizados mediante técnicas de minería de datos y así obtener el conocimiento.

El proceso KDD se compone de 4 etapas; recopilación de datos, preprocesamiento de datos, data Mining o minería de datos e Interpretación y Resultados, las cuales funcionan de manera iterativa (recursiva) (UIAF, 2014).

1.3.1. Recopilación de datos e integración

Se selecciona las fuentes de información de interés y se procederá a integrarlas en un repositorio o almacenamiento de datos (Data warehouse) (Guil Reyes, 2009). La extracción de los datos se recopilará, analizará y tomará un rango para su estudio, de manera que permita conseguir información necesaria para obtener el conocimiento (García Molina, 2007).

La idea de integrar múltiples bases de datos, con sus respectivos formatos, identificadores, etc., es un reto significativo que ha dado lugar a los conocidos almacenes de datos o data warehouse, los cuales permiten crear un repositorio de bases de datos transaccionales provenientes de diferentes fuentes (Vila, 2019). Para integrarlas en un mismo repositorio es necesario orquestar un proceso que lea, limpie y adecue los datos a la estructura del datawarehouse (Lara, 2014).

Esta fase determina que las siguientes fases sean capaces de obtener conocimiento válido y útil a través de la información inicial (López, 2007). En esta fase se selecciona los datos de origen de diferentes fuentes para integrarlos en un mismo repositorio, el cual nos servirá de punto de partida para la obtención del conocimiento.

1.3.2. Selección, Limpieza y Transformación de Datos

En la selección se distingue el subconjunto de datos significativos y se descarga los datos que no aportan información necesaria, en la limpieza o preprocesamiento se elimina ruidos o anormalidades de los datos y en la transformación se busca encontrar una representación más adecuada de los datos (Tuya et al., 2007). Para mejorar la calidad de los datos se debe seleccionar y preparar un subconjunto del conjunto total de datos para formar la vista minable (Christopher et al., 2011). Este proceso previo es necesario porque se tardaría mucho tiempo en llegar a conclusiones si se trabajara con todos los datos. Al subconjunto de datos que se va a minar se denomina vista minable (Montero Navarro, 2009). Este proceso es de suma importancia para obtener conocimiento de una manera rápida y eficaz, debido a que elimina las inconsistencias existentes entre los datos, los transforma y selecciona la mejor manera de analizarlos.

Tabla 2

Adaptación Técnicas de Selección y Transformación

Técnicas de Selección	Descripción
Filtrado de Atributos	Filtración de datos que no aporten nada al desarrollo del conocimiento
Filtrado de Registros	Eliminar algunos registros
Técnicas de Transformación	
Numerización	Transformar un atributo cualitativo a cuantitativo
Discretización	Transformar un atributo cuantitativo a cualitativo ordinal
Creación de Características	Creación de un nuevo atributo
Normalización	Transformación de un rango de valores
Reducción de Dimensiones	Reducción del número de atributos

Fuente (Lara, 2014)

1.3.3. Minería de Datos (Data Mining)

La Data Mining es una etapa dentro del proceso completo del descubrimiento del conocimiento, este e intenta obtener patrones o modelos a partir de los datos recopilados

(Asencios, 2004). Es un proceso esencial donde se aplica métodos inteligentes para extraer patrones de datos.

Data Mining o Minería de Datos se refiere a una forma de análisis de información que encuentra patrones e irregularidades y descubre información no conocida (Cardosa M., 2006). Este es un proceso que consiste en la búsqueda de los patrones de interés que pueden expresarse como un modelo o simplemente que expresen dependencia de los datos (García-González et al., 2019). La validez y utilidad del modelo resultante depende en su mayoría de esta fase (Gorbea Portal, 2013).

En esta etapa la vista minable es sometida a una serie de algoritmos de extracción de conocimiento construyendo un modelo basado en datos (Montero Navarro, 2009).

1.3.3.1. Tipos de Minería de Datos (Data Mining)

Algunos autores distinguen la minería de datos en 2 tipos:

- **Aprendizaje Supervisado**

Utiliza técnicas predictivas, las cuales describen el conjunto de datos de una manera resumida y concisa. Las tareas que produces este tipo de modelos predictivos son la clasificación y regresión (Montero Navarro, 2009).

  - **Clasificación:** cada registro pertenece a una clase en específico, el objetivo es predecir una clase dados los valores de atributos.
  - **Regresión o estimación:** aprendizaje resultante de una función real asignado un valor numérico real, con el objetivo de predecir el valor de una clase.

- **Aprendizaje No Supervisado**

Utiliza técnicas descriptivas de inteligencia artificial, que construyen uno o varios modelos que realizan inferencias sobre el conjunto de entrenamiento para intentar predecir el comportamiento de nuevos datos. Las tareas que producen modelos descriptivos son el agrupamiento (clustering), las reglas de asociación secuenciales y el análisis correlacional (Montero Navarro, 2009).

- **Clustering o Agrupamiento:** utilizan varias metodologías para clasificar automáticamente datos similares en un grupo o clúster, utilizando medidas de asociación. Es utilizada para la descripción y clasificación de datos.
- **Reglas de Asociación:** explica relaciones implícitas entre atributos clasificados.
- **Análisis Correlacional:** comprueba la igualdad existente de valores de dos variables numéricas.

1.3.4. Interpretación y Evaluación

Se identifican patrones de interés que representan conocimiento aplicando técnicas de análisis estadístico y lenguajes de consultas (García-González et al., 2019), con ayuda de herramientas de software se analizará, interpretará y evaluará los resultados obtenidos en la fase de Data Mining (Guil Reyes, 2009).

Se presenta el conocimiento minado obtenido al usuario, de una manera que pueda ser legible para este, utilizando técnicas de visualización y representación del conocimiento (Gorbea Portal, 2013).

1.4. Técnicas de Minería de Datos

Los métodos o técnicas de la Minería de Datos permiten descubrir la información oculta en las bases de Datos y transformarlas en un valioso conocimiento tanto retrospectivo (histórico) como prospectivo (proyecciones) o comprensivos (entender lo que ocurre), lo cual resulta muy útil en la

toma de decisiones en los gobiernos, empresas u organizaciones, entre estas últimas las bibliotecarias y de la información (Gorbea Portal, 2013).

Las técnicas de minería de datos constituyen un enfoque conceptual que son implementadas por varios algoritmos, las cuales pueden clasificarse según la utilidad como técnicas de predicción, de clasificación, de asociación o de agrupamiento (clustering) (Aranda & Sotolongo, 2013).

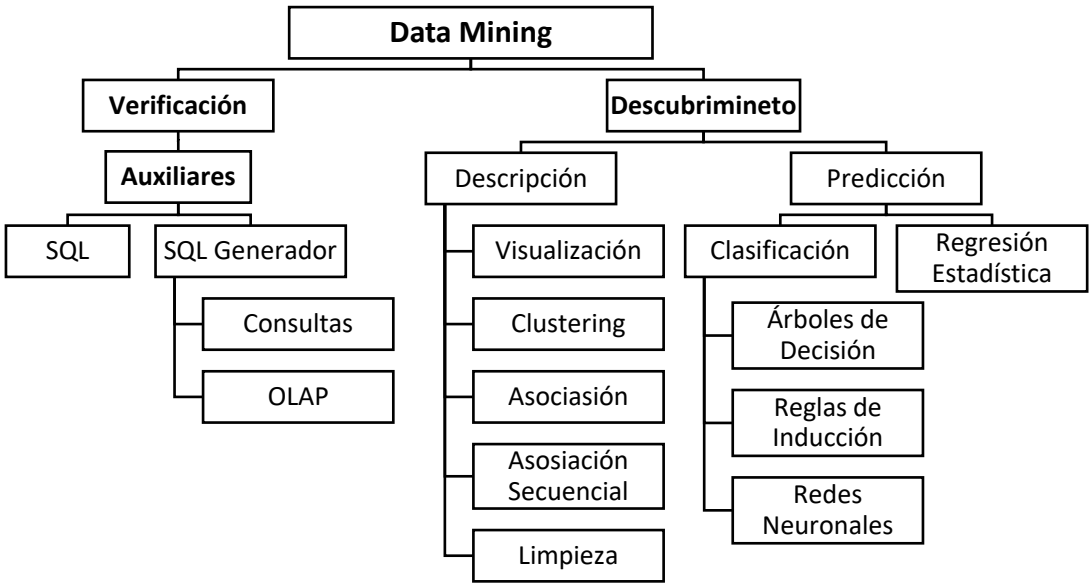


Figura 9 Técnicas de Minería de Datos (López, 2007)

En la Figura 9 se muestra la clasificación y subclasificaciones existentes de las técnicas de minería de datos.

Las tareas o problemas por resolver de la minería de datos más comunes se clasifican dependiendo del tipo de modelo que son capaces de generar (Lara, 2014).

Adicional a las técnicas descriptivas y predictivas existen técnicas de apoyo que son más superficiales y limitadas, basadas en técnicas estadísticas descriptivas, consultas e informes, enfocadas a la verificación (Pérez, 2015).

1.4.1. Tareas Predictivas

Especifican el modelo para los datos con base en un conocimiento teórico previo, dicho modelo debe diferir después de minería de datos antes de ser validado (López, 2007). Se utiliza para predecir el valor desconocido de uno o varios atributos para uno o varios registros de la vista minable (Lara, 2014). Además, de permitir obtener pronósticos de comportamientos futuros a partir de los datos recopilados (Aranda & Sotolongo, 2013).

Existen diferentes técnicas predictivas de minería de datos, entre las cuales se encuentran la de clasificación y regresión.

1.4.1.1. Clasificación

Aprende una función que mapea (clásica) un ejemplo de datos en una de varias clases categóricas predefinidas. En otras palabras, encuentra las propiedades comunes entre un conjunto de objetos y los clasifica en diferentes clases, de acuerdo con un modelo de clasificación. Se utiliza un conjunto de entrenamiento, en el que cada instancia consiste en un conjunto de atributos y el valor de la clase a la cual pertenece (Perversi, 2007).

Se apunta a identificar las características o atributos que hacen que un elemento se vincule a un grupo siguiendo un patrón de datos (Martínez, 2012), cuyo objetivo es predecir el valor desconocido de un atributo (Lara, 2014).

En caso de que no haya un conjunto de entrenamiento, entonces no hay conocimiento sobre los datos para clasificar. En tales casos, la técnica de agrupamiento se puede utilizar para dividir un conjunto de muestras desconocidas en grupos (Gandge & Sandhya, 2018). Es uno de los métodos de minería de datos más utilizados en el sector de la salud. Divide las muestras de datos

en clases objetivas (Ahmad et al., 2015). Algunos tipos de técnicas y algoritmos de clasificación que se indican a continuación son los principales de los varios que existen:

Árboles de Decisión

Es parecido a un diagrama de flujo en forma de árbol, donde cada nodo interno es una prueba sobre un atributo. Cada rama es la solución a la prueba y los nodos hoja identifican clases o distribuciones de clases (Mondragón, 2007). También es conocido como un conjunto organizado de manera jerárquica de condiciones, por tal motivo permite tomar la decisión final (Aranda & Sotolongo, 2013). Los árboles de decisión se utilizan desde hace siglos, y son especialmente apropiados para expresar (interpretabilidad) procedimientos médicos, legales, comerciales, estratégicos, matemáticos, lógicos, entre otros (Solarte Martínez G. R., 2009).

Técnicas Bayesianas o Naive bayes

Son técnicas basadas en el teorema de *Bayes*, que se aplica con el objetivo de calcular probabilidades condicionales de pertenencia de objetos a clasificar dentro de las clases (Kaur & Bawa, 2019). Es un clasificador estadístico, que predicen las probabilidades de membrecía de una clase particular. Ayuda a obtener de forma rápida al método de aprendizaje probabilístico, esta técnica juega un rol muy importante en este método (Ahmad et al., 2015). Se utiliza ampliamente para la clasificación y agrupamiento, pero su perspectiva para el modelado probabilístico de uso múltiple continúa siendo inexplotada (Kaur & Bawa, 2019).

Redes Neuronales Artificiales

Es una técnica popular utilizada en la minería de datos educativos, tienen la capacidad de obtener las posibles interacciones entre variables predictoras (Shahiri et al., 2015). Son estructuras de aprendizaje inspiradas en el sistema nervioso de los animales y humanos (Lara, 2014). Los

elementos que componen una red neuronal o nodos, están interconectados y trabajaran juntos dentro de la red para obtener funciones de salida (Ahmad et al., 2015). Una arquitectura muy utilizada es el MLP (Multi-Layer Perceptron) que es una variante del modelo original propuesto por Rosenblatt en 1950 y utiliza el algoritmo de entramiento de propagación hacia atrás (Ramchoun et al., 2016). Consta de tres o más capas, una entrada, una salida y una o más capas ocultas, utiliza la función de activación no lineal con las neuronas y capas que se encuentran conectadas con las siguientes. Se combinan varios perceptrones para crear límites de decisión con el uso de funciones de activación no lineal / lineal (Zhai et al., 2016).

Vecino más próximo

KNN (*k-nearest neighbors*) es conocido con diferentes nombres: aprendizaje basado en instancias, vecinos más cercanos a K, aprendizaje diferido, razonamiento basado en memoria. Es la técnica de clasificación más importante que nunca conoce el conocimiento previo sobre la distribución de datos (Kaur & Bawa, 2019). Es uno de los clasificadores más simples que descubre el punto de datos no identificado utilizando los puntos de datos previamente conocidos (el vecino más cercano) y los puntos de datos clasificados según el sistema de votación (Ahmad et al., 2015).

1.4.1.2. Regresión

Es una función que le asigna a un elemento un valor real, utilizando valores existentes para predecir datos futuros (Martínez, 2012). Similar a la tarea de clasificación con la única diferencia de que el atributo a predecir es cuantitativo (Guil Reyes, 2009).

Aprende una función que mapea un ejemplo de datos a un valor de una variable predictiva numérica. La regresión asume que los datos se ajustan a algún tipo conocido de función (por ejemplo, lineal, logística, etc.) y después determina la mejor función de este tipo que modele los

datos dados (Mondragón, 2007). Buscan la obtención de un modelo que permita predecir el valor numérico de alguna variable (Asencios, 2004).

La regresión logística es un tipo de análisis estadístico el cual está orientado a predecir una variable categórica en función de distintas variables consideradas como parámetros (Menes Camejo et al., 2015). Además, es un proceso cuantitativo para problemas donde la variable dependiente toma valores de un conjunto finito (Camilo Giraldo Mejía & Alberto Vargas Agudelo, 2016).

1.4.2. Tareas Descriptivas

Las variables no tienen un papel determinado, no son dependientes o independientes y no se supone un modelo previo de datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones (López, 2007). Generan modelos que de alguna manera describen los datos, es decir, tienen el objetivo de describir los datos existentes, más no de predecir algún dato (Lara, 2014).

Existen diferentes técnicas descriptivas de minería de datos, entre las principales se encuentran: agrupamiento (clustering), asociación y detección de atípicos.

1.4.2.1. Clustering o Agrupamiento

Identifica un conjunto finito de categorías o grupos que describen los datos (Mondragón, 2007), permitiendo agrupar un conjunto de datos basándose en las similitudes que presentan los valores de los atributos (Perversi, 2007). Tiene como objetivo la obtención de grupos o clases de objetos sin etiquetar (Guil Reyes, 2009), detallando los contornos del clúster, de manera que los datos que no tienen similitud alguna son considerados como datos atípicos (Alasadi & Bhaya, 2017).

Según la forma en que se aplica la similitud, los algoritmos de clúster se pueden dividir en clustering particional, jerárquico, basado en densidad y basados en grid (Sajana et al., 2016).

1.4.2.2. Asociación

Son un instrumento descriptivo, mediante la utilización de probabilidades de ocurrencia de que dos objetos encuentran relaciones significativas entre los datos (Martínez, 2012), cuyo objetivo es encontrar las relaciones no explicitas en los atributos, por medio de reglas de asociación (Vila, 2019). La obtención de reglas de asociación permite localizar relaciones de asociación o correlación entre un conjunto extenso de datos (Cisneros, 2019).

1.4.2.3. Detección de Atípicos

La detección de atípicos consiste en encontrar objetos que se diferencien notablemente por su comportamiento en medio de un grupo de registros de una vista minable (Vila, 2019). Los tipos de técnicas de detección de atípicos existentes se encuentran basados en aproximaciones estadísticas, proximidad, densidad y clustering (Lara, 2014).

1.5. Herramientas de Minería de Datos

1.5.1. SPSS CLEMENTINE

Es un potente y versátil banco de trabajo de análisis de datos y texto que ayudan en la construcción de modelos predictivos precisos de manera rápida e intuitiva, sin necesidad de programación (*SPSS Clementine Download Free Version (Clementine.Exe)*, 2020). Se centra en la entrega de inteligencia predictiva durante las operaciones de negocio diario integrando a data Mining con otros procesos y sistemas de negocio (Rodríguez Suárez & Amador, 2009).

1.5.2. SAS ENTERPRICE MINER

Proporciona información que incrementa una mejora en la toma de decisiones, optimizando procesos de minería de datos para el desarrollo rápido de modelos, comprende relaciones clave y encuentra patrones de mayor importancia (*Data Mining Software, Model Development and Deployment, SAS Enterprise Miner | SAS*, 2020). Se crean modelos precisos predictivos y descriptivos a partir de grandes volúmenes de datos de fuentes diferentes, mediante un proceso de transparencia, lo que ayuda a la colaboración, cuenta con una interfaz de usuario con diseño de SAS (Jaramillo & Paz-Arias, 2015).

1.5.3. KNIME

Fue creada para un acceso rápido, fácil e intuitivo a la ciencia avanzada de datos, ayudando a impulsar la innovación en las organizaciones (*KNIME | Open for Innovation*, 2020). Además, es una plataforma de código abierto de fácil uso y comprensible para integración de datos, procesamiento, análisis, y exploración. Permite crear flujos de datos de forma visual, ejecutar los pasos de análisis seleccionado, seguido estudiar los resultados, modelos y vistas interactivas (Jaramillo & Paz-Arias, 2015).

1.5.4. PENTAHO

Es una herramienta de Business Intelligence (BI) actualmente la más completa, a la cual se puede acceder vía web y se puede generar vistas de análisis e informes. Es una plataforma compuesta que satisface todos los requisitos de BI (Morales et al., 2016). Combina el análisis empresarial con la integración de datos, lo que permite a los usuarios empresariales tomar decisiones basadas en la información, a los científicos de datos crear modelos de datos robustos y a los administradores de TI ofrecer una plataforma segura y escalable para un amplio conjunto de usuarios (*7.0 - Pentaho Documentation*, 2020).

1.5.5. WEKA

Es un software de aprendizaje automático de código abierto al que se accede a través de una interfaz gráfica de usuario, aplicaciones de terminal estándar o una API de Java. Es muy utilizado para la enseñanza, la investigación y las aplicaciones industriales, contiene una gran cantidad de herramientas integradas para tareas de aprendizaje automático estándar (*Weka 3 - Data Mining with Open Source Machine Learning Software in Java*, 2020). Recopila un conjunto de algoritmos ML (Machine Learning) diseñados para DM (Data Mining), tiene un sistema de paquetes para ampliar su funcionalidad, con paquetes oficiales y no oficiales disponibles, lo que aumenta la cantidad de métodos DM implementados. Ofrece cuatro opciones para DM: interfaz de línea de comandos (CLI), Explorer, Experimenter y Knowledge Flow (Nguyen et al., 2019). Es una herramienta para el aprendizaje automático y minería de datos diseñado en Java, y es de distribución de licencia GNU-GLP. Contiene una colección de algoritmos para el análisis de datos y modelado predictivo, permite la visualización de datos, provee una interfaz gráfica (Jaramillo & Paz-Arias, 2015).

1.5.6. POWER BI

Es una colección de servicios de software, aplicaciones y conectores que trabajan en la conversión de orígenes de datos sin relación, en datos coherentes, interactivos y atractivos visualmente. Para ello utiliza datos de Excel o datos híbridos encontrados en una base de datos (Olcrod, 2019).

En este estudio se usará la herramienta Weka para algoritmos de clasificación porque presenta una gran variedad de algoritmos de fácil interpretación, rápida ejecución y permite el análisis de datos con varias variables, además, se utiliza la herramienta Knime para el análisis de algoritmos predictivos por su fácil interpretación, rapidez de ejecución y por la interfaz gráfica que presenta.

1.6. NORMA ISO/IEC 25012

La ISO/IEC 25012 es uno de los estándares fundamentales de Calidad de Datos, el cual describe un conjunto de características de calidad de datos, que todo conjunto de datos debería cumplir antes de su análisis (Manrique de la Cuadra, 2017). Forma parte de un conjunto de normas internacionales bajo el título general de Ingeniería de Software, Requerimientos y Evaluación de producto de software (SQueRe) (Cisneros, 2019). Primordialmente, la ISO/IEC 25012 se clasifica en 2 grandes grupos evaluando la calidad de los datos los cuales son **inherentes** y **dependientes** del sistema. En la Tabla 3 se da a conocer las características pertenecientes a la norma ISO/IEC 25012.

Tabla 3

Características de Calidad

Característica	Definición
Accesibilidad	Grado de acceso a los datos por parte de cualquier usuario.
Actualidad	Grado en que los datos tienen atributos con las fechas y tiempos correctos.
Conformidad	Grado de construcción de los datos conforme a estándares, convenciones o regulaciones.
Completitud	Grado en el que existen suficientes valores para todos los atributos necesarios para la representación de una entidad.
Comprensibilidad	Grado en el que los datos se expresan de manera que los usuarios puedan leerlos e interpretarlos correctamente.
Confidencialidad	Grado en el que los datos tienen atributos específicos que sólo pueden ser accedidos por usuarios autorizados
Consistencia	Grado en que los datos están libres de contradicción y son coherentes con el resto de los datos de su contexto de uso
Credibilidad	Grado en que los datos se consideran ciertos y creíbles por los usuarios.
Disponibilidad	Grado en el que los datos están disponibles para ser accedidos por usuarios y/o aplicaciones autorizados.
Eficiencia	Grado en el que los datos tienen atributos que pueden ser procesados y provistos dentro de los niveles de rendimiento esperados.
Exactitud	Grado en el que los datos tienen atributos que son exactos y precisos.

Portabilidad	Grado en el que los datos pueden ser alojados, reemplazados o movidos desde un sistema a otro.
Exactitud	Grado en que los datos tienen atributos que representan correctamente el valor de un atributo.
Recuperabilidad	Grado en el que los datos disponen de formas de mantener un nivel especificado de operabilidad incluso cuando se producen fallos
Trazabilidad	Grado en el que los datos tienen atributos que proveen información detallada sobre los cambios realizados en los datos.

Fuente (Manrique de la Cuadra, 2017)

1.6.1. Calidad de datos Inherentes

Se refiere al grado en que la calidad de los datos tienen el potencial de satisfacer necesidades, en cuanto se cumpla condiciones específicas (Vila, 2019).

1.6.2. Calidad de datos dependientes del sistema

Se refiere al grado de que la calidad de los datos es alcanzada y preservada dentro de un sistema computacional cuando la información es usada específicamente (Cisneros, 2019).

Tabla 4
Características del modelo de Calidad de Datos

CARACTERÍSTICA	INHERENTES	DEPENDIENTES DEL SISTEMA
Accesibilidad	X	
Actualidad	X	
Compleción	X	X
Completitud	X	
Comprensibilidad	X	X
Confidencialidad	X	X
Consistencia	X	
Credibilidad	X	
Disponibilidad		X
Eficiencia	X	X
Exactitud	X	X
Portabilidad		X
Precisión	X	
Recuperabilidad		X
Trazabilidad	X	X

Fuente (Manrique de la Cuadra, 2017)

1.7. Trabajos existentes

En la actualidad existen diferentes investigaciones similares al estudio a realizar, entre las más relevantes se encuentran las siguientes:

1.7.1. Identificación de patrones delictivos en Colombia durante el periodo 2010-2016 mediante el uso de Técnicas de minería de datos

Utilizando técnicas de minería de datos y técnicas de agrupamiento, se analizó 402.631 registros cada uno con 21 atributos sobre hurtos de celulares, vehículos, motocicletas y delitos sexuales ocurridos en Colombia. Para el procesamiento y análisis de estos datos se utilizó la metodología CRISP-DM y la técnica de agrupamiento o clustering "K-Modes", permitiendo la identificación de los barrios de Popayán en los que ocurre la mayoría de estos delitos (Yacup et al., 2018).

1.7.2. Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina

Mediante herramientas de minería de datos en el ámbito criminal se realizó el análisis de homicidios cometidos en la república de Argentina. Utilizando el algoritmo K-Means para clasificar en grupos los datos con el objetivo de disminuir la función de error cuadrático, el algoritmo ID3 para construir árboles de decisión simple, el algoritmo C4.5 para extender la categorización a numéricos y el algoritmo J48 permite funcionar atributos nominales y numéricos, de esta manera eliminar ramas innecesarias, de esta manera identificar y detectar patrones de homicidios (Valenga et al., 2008).

1.7.3. Modelado y simulación de robos y hurtos basados en redes SOM, TDIDT y Bayesianas, un caso de estudio.

Integra tecnologías orientadas a la obtención de conocimiento de minería de datos y GIS para identificar y clasificar clúster de robos y hurtos cometidos en Argentina en el 2017. Además de utilizar diferentes técnicas de minería de datos para interpretar los patrones delictivos, los cuales

vinculados con GIS sirvieron para identificar zonas de mayores actos delictivos (Flores et al., 2019). Para la obtención de estos patrones se utilizó la metodología CRISP-DM y la utilización del algoritmo Kohonen-SOM para descubrir grupos, C4.5 para caracterización de reglas de cada clúster y Redes Bayesianas para la ponderación de los atributos, obteniendo como resultado que el hurto a mano armada es el más cometido, además de identificar las horas, días, meses, zonas conflictivas y objetos que son mayormente hurtados.

1.7.4. Sistema De Predicción De Hechos Delictivos Para La Mejora Del Proceso De Prevención Del Delito En El Distrito De La Molina Utilizando Minería De Datos

Este sistema de predicción, mediante la utilización de minería de datos, tiene el objetivo de mejorar la prevención de delitos cometidos en 3 comisarías de jurisdicción, de manera que se mejore el proceso de salvaguardar y mantener el orden en la zona. Obtienes predicciones de delitos que podrían realizarse en un futuro con base en un registro de datos histórico (Jaulis & Vilcarromero, 2015). Utilizaron la metodología CRISP-DM, ingeniería inversa para el desarrollo del modelo relaciona y para el análisis se utilizó el algoritmo RNA (Redes Neuronales Artificiales) con la tipología perceptrón multicapa, obteniendo como resultado los delitos por hora, por mes, el tipo de delito y la zona en donde se realizaba este delito.

1.7.5. Impacto de la implementación de minería de datos en el mantenimiento y análisis de la información catastral en una municipalidad distrital

Utilizando la metodología CRISP-DM y los algoritmos K-Means y Konohen, permitiendo obtener un modelo el cual permite identificar variables catastrales más relevantes para cada área municipal, las autoridades pueden obtener información con valor para la realización de operativos estratégicos pertinentes (Antezana, 2018).

1.7.6. Using Data Mining for Intelligence-Led Policing and Crime Analysis

Utilizando algoritmos clásicos de clustering, asociación y clasificación, para la identificación de enlaces implícitos y ocultos entre objetos, basándose en el sistema RICAS (Real-time Intelligence crime analytics system) y con la utilización del algoritmo de búsqueda recursiva de relaciones y del algoritmo de búsqueda visual de enlaces se considera la construcción de un software correspondiente para el análisis criminal permitiendo ver una imagen predictiva y post factum (Uzlov et al., 2019).

1.7.7. Crime Hotspot Detection With Clustering Algorithm Using Data Mining

El objetivo de este trabajo es realizar un análisis sobre delitos existentes entre los años 2016-2018, utilizando algoritmos clustering y Streaming detectar e identificar delitos en la red criminal. En el estudio usa el algoritmo Streaming para realizar análisis en vivo y el algoritmo K-Means para resolver problemas de agrupamiento, además se usa la metodología de agrupamiento, la cual consiste en documentar el clúster, analizar los resultados implementando los algoritmos AP (permite obtener nuevos patrones) y RBF (produce resultados de los grupos obtenidos), permitiendo detectar los crímenes y facilitando el trabajo de la policía (Aarthi et al., 2019).

1.7.8. Integrating Game Theory and Data Mining for Dynamic Distribution of Police to Combat Crime

Haciendo uso de la minería de datos para modelar el entorno, la teoría del juego para modelar el juego entre partes y optimización para determinar la estrategia optima, se busca crear un modelo de distribución de recursos policiales para combatir el crimen en 3 zonas de mayor influencia de robo, robo con violencia y robo con fuerza en Santiago de Chile. Para cada uno de estos delitos se creó un modelo de regresión lineal, árbol de regresión y redes neuronales, permitiendo anticipar los niveles de crímenes en cada zona (Segovia & Smith-Miles, 2019).

CAPÍTULO 2

Desarrollo del Proceso KDD

El desarrollo del siguiente proyecto de minería de datos proporcionará información necesaria para una mejor toma de decisiones estratégicas, de parte de la Entidad Aduanera, con base en datos históricos almacenados en las bases de datos de la empresa. Este capítulo fue desarrollado en conjunto con el trabajo de titulación denominado "DETECCIÓN DE PATRONES DE CONTRABANDO PARA LA GESTIÓN DE APREHENSIONES Y RETENCIONES, UTILIZANDO TÉCNICAS DESCRIPTIVAS DE MINERÍA DE DATOS.", elaborado por Diana Carolina Rosero Rea, por lo que ciertas tablas y figuras son comunes a los dos trabajos.

2.1. Generalidades

- **Suposiciones**

Para la elaboración de este proyecto se dispone de datos históricos en bruto de la Entidad Aduanera para el análisis, obteniendo como resultado información de fácil entendimiento para las personas encargadas.

- **Restricciones**

El tiempo de desarrollo del proyecto no debe ser mayor a 6 meses, de tal manera, que la planificación de desarrollo se la realizó tomando en cuenta la naturaleza de las actividades a realizar.

El proyecto se desarrollará en conjunto con la institución acreedora del mismo, con el objetivo de dar resultados relevantes para esta.

El proyecto deberá cumplir los lineamentos encontrados en el acta de confidencialidad acordada entre ambas partes (Anexo 1).

La información entregada por la Entidad Aduanera es muy limitada y no se encuentra normalizada, de tal manera, se entregó un formato de ingreso de registros (Anexo 2).

2.2. Entregables del Proyecto

A continuación, se detalla los artefactos generados y a utilizar en el proyecto, en el transcurso del proyecto estos irán variando hasta culminar el proceso KDD, dando como fin la obtención de la vista minable.

- Documento de Excel: correspondiente a sugerencia de formato de ingreso de aprehensiones (Anexo 2).
- Data Warehouse: Correspondiente a la Fase de Integración de Datos.
- Vista Minable: Resultado de la fase de selección, limpieza y transformación de datos.
- Check List correspondiente a la aplicación de la Norma ISO/IEC 25012:2008.
- Modelos Predictivos de clasificación y regresión: correspondientes a la fase de minería de datos.
- Conocimiento: correspondiente a la fase de análisis e interpretación de los resultados.
- Documento de Word correspondiente al análisis estadístico de las preguntas del negocio (Anexo 3) establecidas con la entidad aduanera (Anexo 4).

2.3. Organización del Proyecto

2.3.1. Participantes del Proyecto

En la Tabla 5 se presenta el director de cada área implicada.

Tabla 5

Directivos de las áreas implicadas

Dependencia	Participante	Función
Coordinador Carrera de Ingeniería en Sistemas Computacionales	Msc. Pedro Granda	Asignar especialista en aprendizaje supervisado
Entidad de Control	Anónimo	Especialista en problemática

Fuente Propia

Se detallan los participantes directos del proyecto en la Tabla 6.

Tabla 6

Participantes del proyecto

Rol	Dependencia	Nombre
Jefe del Proyecto	Docente Carrera Ingeniería en Sistemas Computacionales	PhD. Iván García
Especialista de Negocio	Entidad de Control	Anónimo
Analista de Sistemas	Carrera de Ingeniería en Sistemas Computacionales	Tommy Mancero Menoscal

Fuente Propia

2.3.2. Roles y Responsabilidades

En la Tabla 7, se describe cada rol en función de sus responsabilidades y funciones.

Tabla 7

Participantes del proyecto

Rol	Responsabilidad
Jefe del Proyecto	Es el encargado de tomar decisiones que permitan el cumplimiento de los objetivos del proyecto. Se encarga de establecer comunicación con el usuario final de los entregables y el patrocinador, así como de gestionar los recursos empleados durante el proyecto, toma las decisiones necesarias para conocer en todo momento la situación actual en relación con los objetivos establecidos (Vila, 2019).
Especialista de Negocio	Es el encargado de establecer los requisitos para la elaboración del proyecto.
Analista de Sistemas	Aplicación de la ISO/IEC 25012:2008, desarrollo del proceso KDD, validación de resultados, documentación y análisis de impacto.

Fuente Propia

2.4. Gestión del Proceso

2.4.1. Estimaciones

A continuación, en las Tablas 8, 9 y 10 se detalla el presupuesto estimado y recursos involucrados en la realización del proyecto.

Tabla 8

Talento Humano

DESCRIPCIÓN	N. DE HORAS	COSTO POR HORA ($)	COSTO TOTAL ($)
Horas de investigación del proyecto	230	20.00	4600.00
Horas de desarrollo del proyecto	230	20.00	4600.00
		TOTAL	9200

Fuente: Propia

Tabla 9

Recursos Materiales

DESCRIPCION	COSTO REAL	COSTO ACTUAL
Hardware		
Computadora portátil	$1.100,00	$0,00
Impresora	$150,00	$0,00
Software		
Microsoft Excel	$0,00	$0,00
Pentaho 7	$0,00	$0,00
Weka	$0,00	$0,00
Materiales de Oficina		
Tinta de impresora	$50,00	$50,00
Hojas A4	$7,00	$5
Internet	$180,00	$180,00
Flash memory	$12,00	$0,00
INVESTIGACIÓN		
ISO/IEC 25012	$88,00	$0,00
SUBTOTAL	$2.087,00	$235,00
10% IMPREVISTOS	$208,70	$23,50
TOTAL	$2.295,70	$258,50

Fuente Propia

Tabla 10

Costo Total del Proyecto

DESCRIPCIÓN	COSTO TOTAL ($)
Talento Humano	9200
Recursos Materiales	2295,70
TOTAL	11495,70

Fuente Propia

2.4.2. Plan del Proyecto

La minería de datos consta de varias fases del proceso KDD, por lo que en la Tabla 11 se detalla la duración de cada una de las fases y la aplicación de la normalización.

Tabla 11

Distribución de Horas

Fase	Duración (H)
Fase de Integración y Recopilación	35
Fase de Selección, Limpieza y Transformación de Datos	30
Aplicación ISO/IEC 25012:2008	15
Fase de Minería de Datos	50
Fase de Evaluación e Interpretación	45
Documentos e Investigación	205
Análisis de Resultado	40
Análisis de Impacto	40
TOTAL	460

Fuente Propia

En la Tabla 12 se detalla cuando un hecho ha concluido.

Tabla 12

Distribución de Horas

HECHO	DESCRIPCIÓN
Obtención del conocimiento bibliográfico	Se obtiene el conocimiento necesario para realizar el análisis de los datos, determinar el proceso KDD y los modelos que se emplearan en la etapa de minería de datos, de acuerdo con la naturaleza del conocimiento que se pretenda obtener.

Obtención de los datos a analizar	Este hecho es importante para iniciar el análisis de los datos, el cual concluye cuando se tiene datos necesarios para realizar el proceso KDD.
Obtención de la data warehouse	Se limita los datos provenientes de la entidad aduanera.
Obtención Vista Minable	Se construye la vista minable, la cual se aplica las técnicas predictivas de minería de datos, en la cual los datos se encuentran cuantitativos o cualitativos.
Implementación ISO/IEC 25012:2008	Se obtiene el grado de calidad de los datos, de acuerdo con la característica de consistencia de la norma.
Obtención Modelos Predictivos	Se obtienen los modelos predictivos que permiten el análisis e interpretación del conocimiento.
Obtención del Conocimiento	Se determina si el modelo predijo de forma correcta y si se ajusta a lo requerido.
Obtención de Documentación	Se obtiene la documentación de cada una de las fases y actividades que se desarrollaron durante el proyecto.

Fuente Propia

2.5. Integración y Recopilación de Datos

2.5.1. Tipos de datos base

El presente proyecto de titulación presenta datos históricos (2014-2019) de contrabando, que se encuentran almacenados en documentos de Excel de la Entidad Aduanera. Los tipos de datos que se encuentran son enteros, decimales, reales, combinados, fechas, cadena de caracteres, sin embargo, se encuentran inconsistencias en los datos. Es necesario clasificarlos en numéricos y categóricos o discretos, para aplicar técnicas de minería de datos predictivas (Vila, 2019).

- **Tipos de datos**

Tabla 13
Estructura Base de Datos

N°	ATRIBUTO	TIPO DE DATO	ACTAS DE APREHENSIÓN					
			2014	2015	2016	2017	2018	2019
1	**Registro**	Numeric	X	X	X	X	X	X
2	**Numero**	Character	X	X	X	X	X	X
3	**Fecha**	DateTime	X	X	X	X	X	X
4	**NoExpediente**	Logical	X	X	X	X	X	X
5	**Cantidad**	Numeric	X	X	X	X	X	X

6	**Unidades**	Character	X	X	X	X	X	X
7	**Precio**	Numeric	X	X	X	X	X	X
8	**Cotización**	Numeric	X	X	X	X	X	X
9	**Total**	Numeric	X	X	X	X	X	X
10	**Grupo**	Character	X	X	X	X	X	X
11	**SubGrupo**	Character	X	X	X	X	X	X
12	**Descripcion**	Character	X	X	X	X	X	X
13	**Procedencia**	Character	X	X	X	X	X	X
14	**Bodega**	Character	X	X	X	X	X	X
15	**Peso**	Numeric	X	X	X	X	X	X
16	**Marca**	Character	X	X	X	X	X	X
17	**Status de la mercancía**	Character	X	X	X	X	X	X
18	**Lo que originó la aprehensión**	Character	X	X	X	X	X	X
19	**Sitio de la aprehensión**	Character	X	X	X	X	X	X
20	**Grupo Operativo**	Logical	X	X	X	X	X	X
21	**Distrito**	Logical	X	X	X	X	X	X
22	**Zonas**	Character						X
23	**Observaciones**	Character	X	X	X	X	X	X

Fuente Propia

2.5.2. Construcción Estructura ETL

Para la elaboración de la data warehouse se utilizó la herramienta Pentaho Data Integration (PDI), el cual permite realizar procesos de extracción, transformación y carga de información (ETL, Extract, Transfor and Load), lo cual nos servirá a lo largo del proyecto.

- **Selección Campos de Interés**

En la Fig. 10 se muestra la selección de campos de interés del documento "ACTAS DE APREHENSIÓN 2014"

Figura 10 Selección Datos 2014

En la Fig. 11 se muestra la selección de campos de interés del documento "ACTAS DE APREHENSIÓN 2015"

Figura 11 Selección Datos 2015

En la Fig. 12 se muestra la selección de campos de interés del documento "ACTAS DE APREHENSIÓN 2016"

Figura 12 Selección Datos 2016

En la Fig. 13 se muestra la selección de campos de interés del documento "ACTAS DE APREHENSIÓN 2017"

Figura 13 Selección Datos 2017

En la Fig. 14 se muestra la selección de campos de interés del documento "ACTAS DE APREHENSIÓN 2018"

Figura 14 Selección Datos 2018

En la Fig. 15 se muestra la selección de campos de interés del documento "ACTAS DE APREHENSIÓN 2019"

Figura 15 Selección Datos 2019

- **Agrupación de Datos**

En la Fig. 16 se muestra el agrupamiento de los datos de interés entre los documentos "ACTAS DE APREHENSIÓN 2014" y "ACTAS DE APREHENSIÓN 2015"

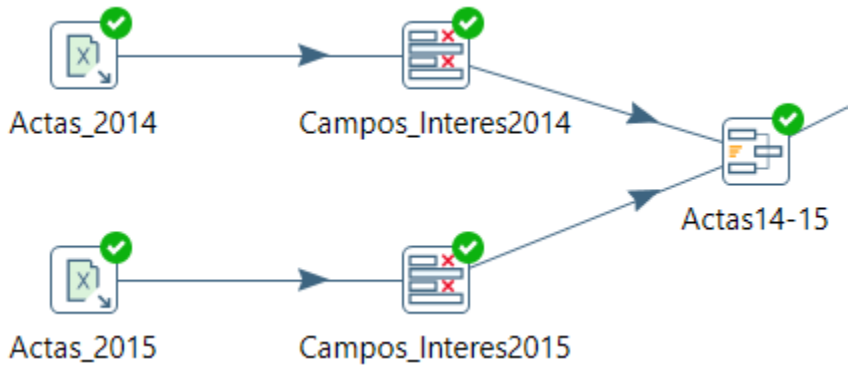


Figura 16 Agrupamiento Actas 2014-2015

En la Fig. 17 se muestra el agrupamiento de los datos de interés entre los documentos "ACTAS DE APREHENSIÓN 2016" y "ACTAS DE APREHENSIÓN 2017"

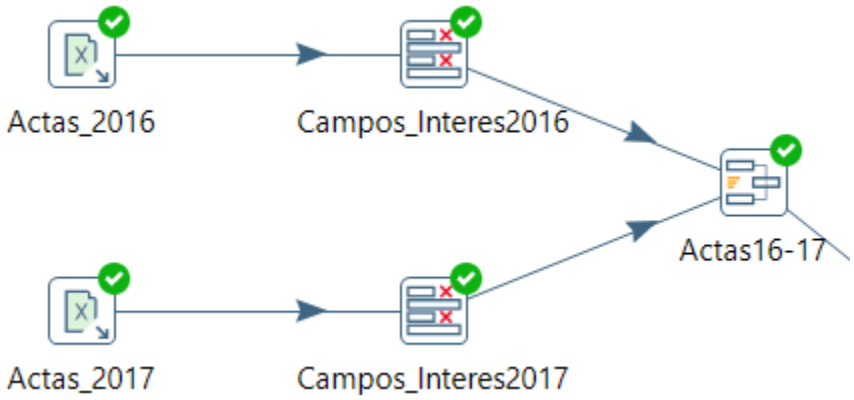


Figura 17 Agrupamiento Actas 2016-2017

En la Fig. 18 se muestra el agrupamiento de los datos de interés entre los documentos "ACTAS DE APREHENSIÓN 2018" y "ACTAS DE APREHENSIÓN 2019"

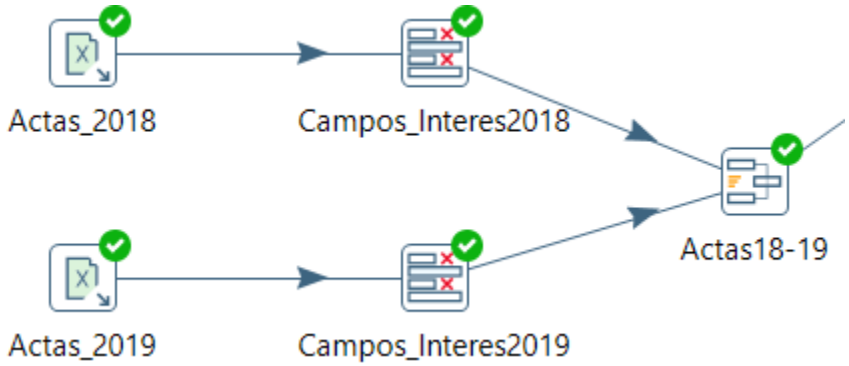


Figura 18 Agrupamiento Actas 2018-2019

En la Fig. 19 se muestra el agrupamiento de los datos de interés entre Actas16-17 y Actas 18-19.

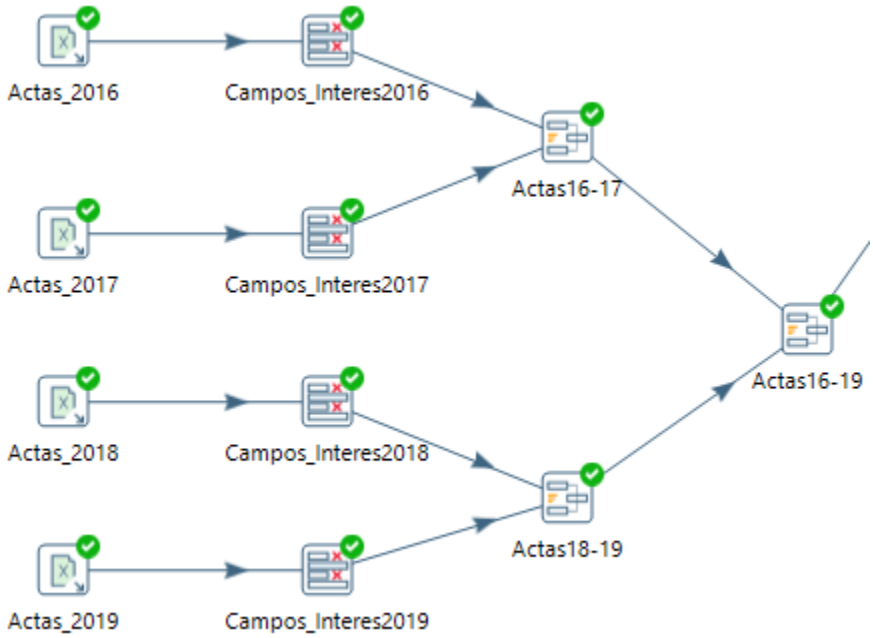


Figura 19 Agrupamiento Actas 2016 a 2019

En la Fig. 20 se muestra el agrupamiento total de los datos entregados por la institución de control aduanero.

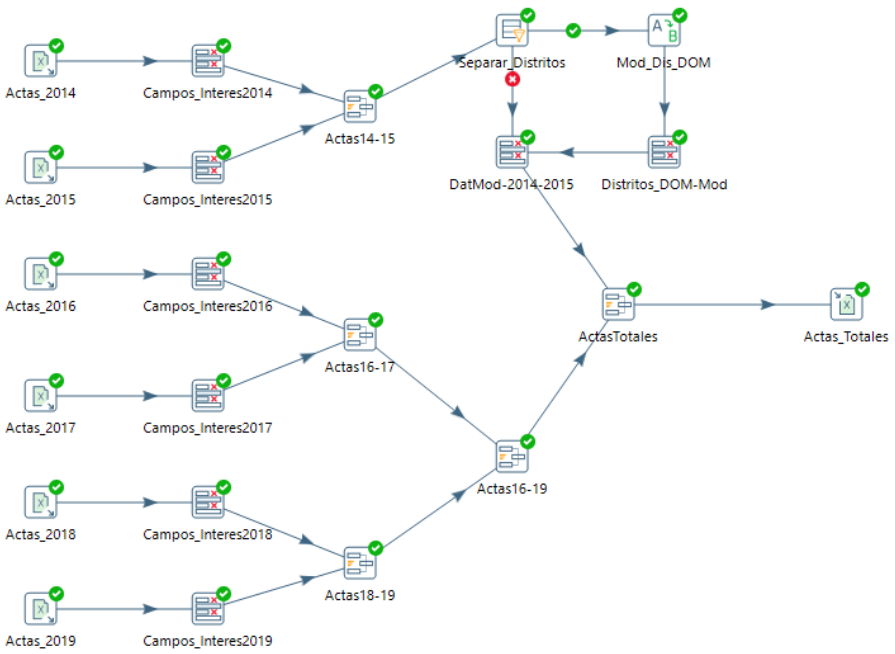


Figura 20 Actas Totales

En la Tabla 14 se muestra la estructura de los datos finales necesarios y los que se encuentran acorde al acta de confidencialidad firmada por ambas partes (Anexo 1), el cual cuenta con 102667 filas y un total de 16 columnas.

Tabla 14
Estructura Datos Finales

N°	ATRIBUTO	TIPO DE DATO
1	**Fecha**	DateTime
2	**Unidades**	Character
3	**Cantidad**	Numeric
4	**Precio**	Numeric
5	**Total**	Numeric
6	**Grupo**	Character
7	**SubGrupo**	Character
8	**Procedencia**	Character
9	**Bodega**	Character
10	**Marca**	Character
11	**Status**	Character
12	**Origen_Aprehension**	Character
13	**Sitio_Aprehension**	Character
14	**Grupo_Operativo**	Character
15	**Distrito**	Character
16	**Zonas**	Character

Fuente Propia

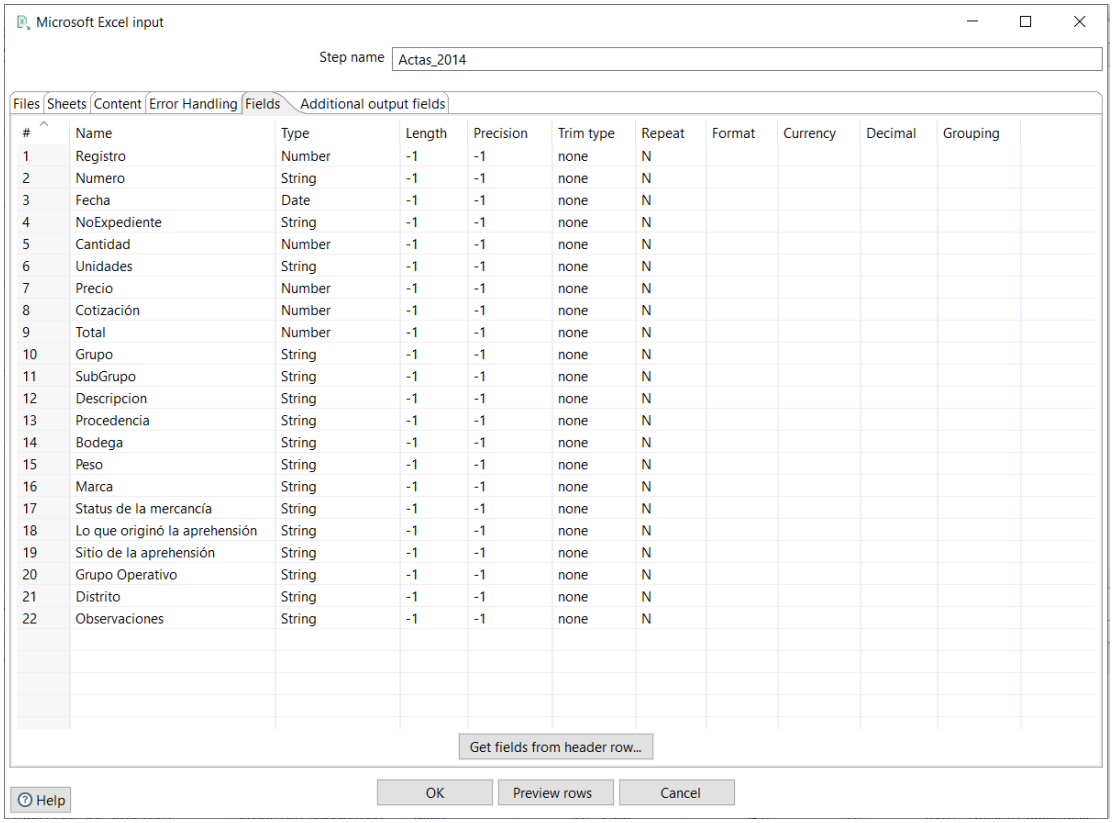
2.6. Selección, transformación y limpieza de los datos

Una vez consolidado los datos a usar en el proyecto se procede a seleccionar, limpiar y transformarlos. Para el análisis estadístico se empleará la herramienta PBI (Anexo 4).

2.6.1. Selección

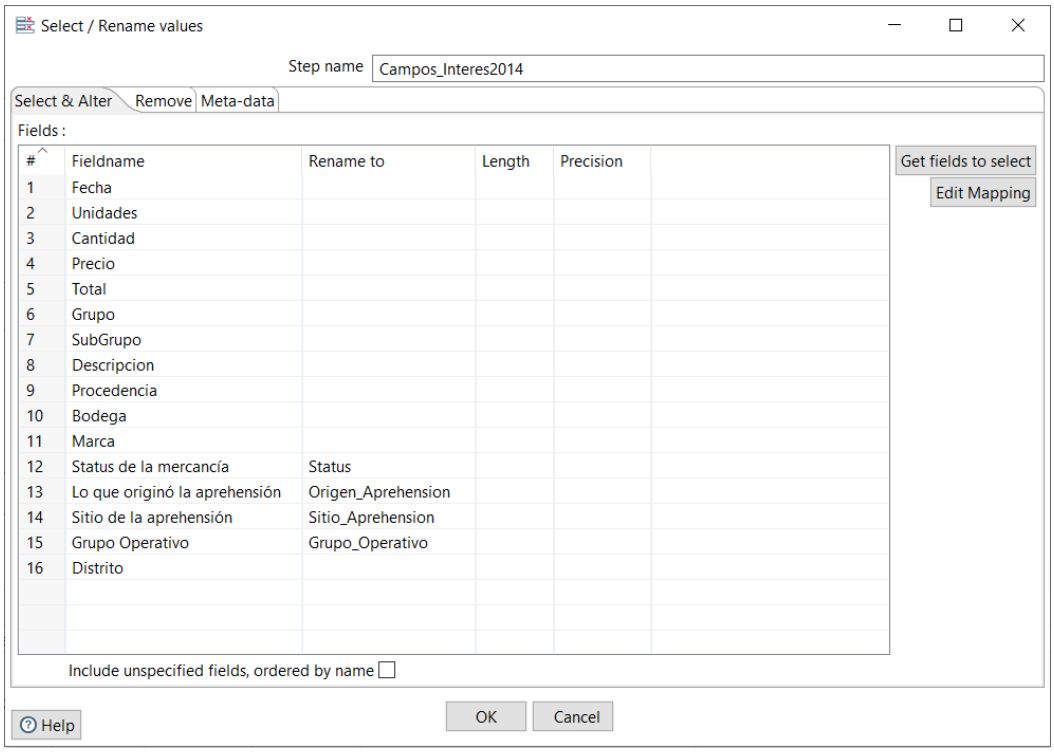
- **Selección de Atributos**

Esta etapa se inicia eliminando atributos no necesarios o que no aporten conocimiento para el desarrollo del proyecto, este proceso se aplica en todos los datos de los años 2014 a 2018, a excepción de datos del 2019, debido a que aumenta un atributo denominado Zonas. Por ejemplo, en el documento "ACTAS DE APREHENSIÓN 2014" inicialmente contaba con 22 atributos, después del proceso de selección quedaron 15 atributos de importancia para este estudio, sin mencionar campos que a lo largo del proceso puedan aparecer, los cuales se muestran en la Fig.21 y Fig. 22.



#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	Registro	Number	-1	-1	none	N				
2	Numero	String	-1	-1	none	N				
3	Fecha	Date	-1	-1	none	N				
4	NoExpediente	String	-1	-1	none	N				
5	Cantidad	Number	-1	-1	none	N				
6	Unidades	String	-1	-1	none	N				
7	Precio	Number	-1	-1	none	N				
8	Cotización	Number	-1	-1	none	N				
9	Total	Number	-1	-1	none	N				
10	Grupo	String	-1	-1	none	N				
11	SubGrupo	String	-1	-1	none	N				
12	Descripcion	String	-1	-1	none	N				
13	Procedencia	String	-1	-1	none	N				
14	Bodega	String	-1	-1	none	N				
15	Peso	String	-1	-1	none	N				
16	Marca	String	-1	-1	none	N				
17	Status de la mercancía	String	-1	-1	none	N				
18	Lo que originó la aprehensión	String	-1	-1	none	N				
19	Sitio de la aprehensión	String	-1	-1	none	N				
20	Grupo Operativo	String	-1	-1	none	N				
21	Distrito	String	-1	-1	none	N				
22	Observaciones	String	-1	-1	none	N				

Figura 21 Atributos del Documento Actas 2014-2015



#	Fieldname	Rename to	Length	Precision
1	Fecha			
2	Unidades			
3	Cantidad			
4	Precio			
5	Total			
6	Grupo			
7	SubGrupo			
8	Descripcion			
9	Procedencia			
10	Bodega			
11	Marca			
12	Status de la mercancía	Status		
13	Lo que originó la aprehensión	Origen_Aprehension		
14	Sitio de la aprehensión	Sitio_Aprehension		
15	Grupo Operativo	Grupo_Operativo		
16	Distrito			

Figura 22 Atributos se Interés

- **Creación Atributo Zona**

Esta etapa inicia mapeando valores por medio del atributo Distrito y la clasificación Zonal con la que cuenta la Entidad Aduanera, teniendo como resultado la inserción del atributo Zona. En la Fig. 23 se muestra el mapeo que se realizó en esta etapa.

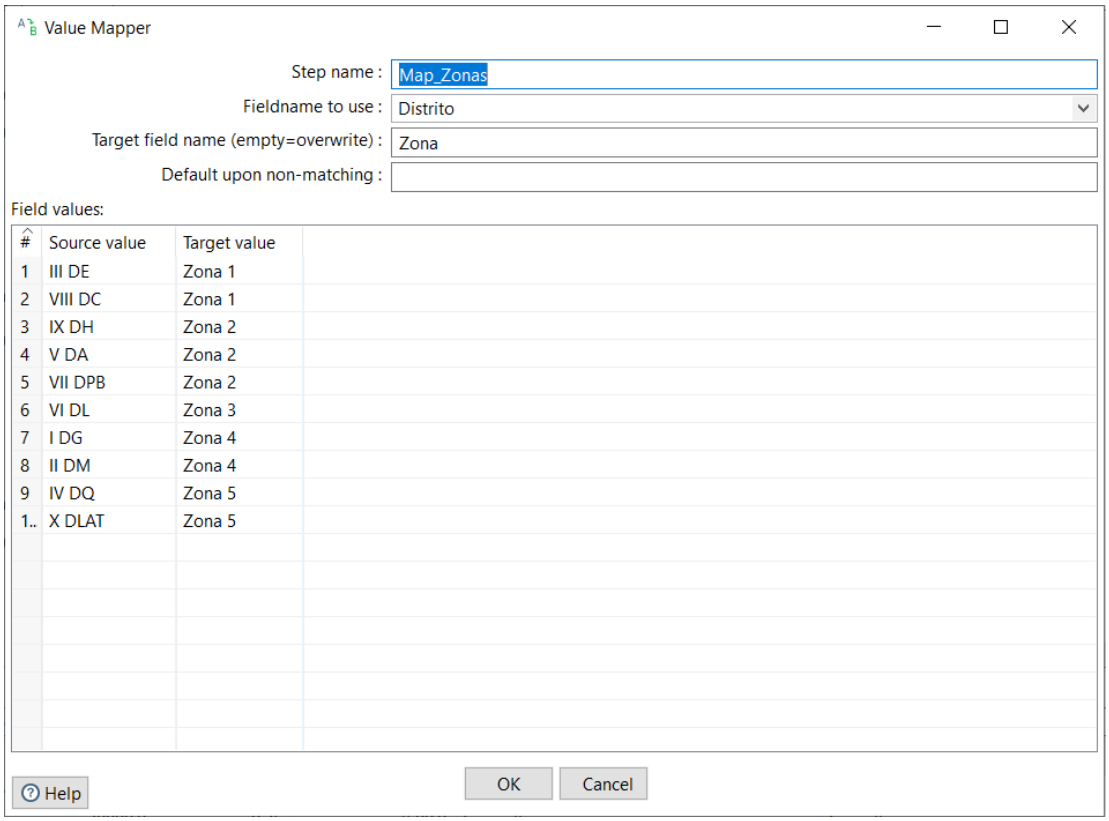


Figura 23 Mapeo Atributo Zona

2.6.2. Transformación

Para este proceso se categorizó los atributos que se ajustan a criterios específicos o similares.

- **Atributo BODEGAS**

En la Tabla 15 se muestra una parte de la categorización del atributo Bodegas por la gran cantidad de datos existentes y por el acta de confidencialidad existente, esto ayudará a reducir resultados dispersos.

Tabla 15
Categorización atributo BODEGAS

CATEGORÍA	VALORES
BODEGA CONTECON	BODEGA DE ADUANA - CONTECON BODEGA 9-1B EN CONTECON
BODEGA DE ADUANA - PTO. DE MANTA	BODEGA DE ADUANA - PTO, DE MANTA
BODEGA DE BOMASA	BODEGA DE CONTRABANDO BOMASA BODEGA DE RETENCIONES BOMASA DEL SENAE-GUAYAQUI

DIRECCIÓN DISTRITAL CUENCA	GERENCIA DISTRITAL DE CUENCA BODEGA DE CUENCA DIRECCION DISTRITAL CUENCA
DIRECCIÓN DISTRITAL ESMERALDAS	DIRECCION DISTRITAL ESMERALDAS INSTALACIONES DEL III DISTRITO ESMERALDAS BODEGA CAE-ESMERALDAS GERENCIA DISTRITAL DE ESMERALDAS
SALA DE ARRIBO INTERNACIONAL AEROPUERTO DE UIO	BODEGA DE SAI-SENAE AEROPUERTO MS UIO
UNIDAD DE LAVADO DE ACTIVOS DE LA PPNN	UNIDAD DE LAVADO DE ACTIVOS DE LA PPNN

FUENTE PROPIA

- **Atributo GRUPO_OPERATIVO**

Por motivo de confidencialidad en la Tabla 16 se muestra una pequeña parte de la categorización del atributo GRUPO_OPERATIVO, debido a que existen datos que son similares, lo que ocasionaría resultados dispersos.

Tabla 16
Categorización atributo GRUPO_OPERATIVO

CATEGORÍA	VALORES
ARCSA	ARCSA-UVA
CONTROL CONJUNTO INTERINSTITUCIONAL	CONTROL CONJUNTO INTERINSTITUCIONAL, POLICÍA NACIO PATRULLA CVA-DIP PATRULLA CVA-FFAA PATRULLA CVA-PPNN-FFAA PATRULLA UVA - CTE PATRULLA UVA - PPNN PATRULLA UVA -FFAA - PPNN PATRULLA UVA-COMITE OPERATIVO PROVINCIAL DEL AZUAY
CUERPO DE VIGILANCIA ADUANERA	PATRULLA AMAZONAS PATRULLA CATAMAYO PATRULLA CHACRAS UNIDAD DE CONTROL CONJUNTO DE CONTENEDORES
FUERZAS ARMADAS	EJÉRCITO ECUATORIANO PERSONAL DESTACAMENTO DEL EJERCITO LAURO GUERRERO PERSONAL DESTACAMENTO MILITAR SAUCILLO
INSPECTORIA DE PESCA	INSPECTORIA DE PESCA

FUENTE PROPIA

- **Atributo GRUPO**

En la Tabla 17 se muestra la categorización del atributo GRUPO que se realizó con motivo de reducir los campos similares y así disminuir resultados dispersos.

Tabla 17

Categorización atributo GRUPO

CATEGORÍA	VALOR
LICORES	WHISKY VARIAS MARCAS
PRODUCTOS QUIMICOS	PRODUCTOS DE LIMPIEZA
MEDICINA	INSUMOS MÉDICOS
ARTICULOS DE BAZAR	ARTESANIAS

FUENTE PROPIA

- **Atributo SUBGRUPO**

En la Tabla 18 se muestra la categorización del atributo SUBGRUPO, debido a que existen datos que son similares, lo que ocasionaría resultados dispersos.

Tabla 18

Categorización atributo SUBGRUPO

CATEGORÍA	VALOR
ARROZ	ARROZ AL GRANEL ARROZ GRANEL
MANDARINA	MANDARINA / TANGELO MANDARINA/TANGELO
MOROCHILLO/MOROCHO	MOROCHILLO
PLANTAS ORNAMENTALES/INDUSTRIALES/MEDICINALES ETC	PLANTAS ORNAMENTALES
GRANADILLA/MARACUYA/PITAHAYA	GRANADILLA/MARACUYA

PRODUCTOS NATURALES	PRODUCTOS NATURALES/VITAMINAS
CARTONES/CAJAS/GAVETAS_ BIDONES	CARTONES/CAJAS/GAVETAS
DULCES VARIOS	CARAMELOS
ALGODON/HILOS/ELASTICOS/ENCAJES	ALGODÓN/HILOS/ELÁSTICOS ALGODÓN/HILOS/ELÁSTICOS/ENCAJES
POLLITOS/POLLOS/GALLOS	POLLITOS
NARANJA	NARANJAS A GRANEL
DE VEHICULOS	DE VEHÍCULO
ARROCILLO	ARROZ QUEBRADO
VARIOS	

FUENTE PROPIA

- **Atributo PROCEDENCIA**

La categorización del atributo PROCEDENCIA se la realizó por motivo de que existen campos similares, de tal manera que se logre reducir resultados inconsistentes, lo cual se muestra en la Tabla 19.

Tabla 19

Categorización atributo PROCEDENCIA

CATEGORÍA	VALOR
PERU	PERUANA
DESCONOCIDA	

FUENTE PROPIA

- **Atributo SITIO_APREHENSION**

Para la categorización del atributo SITIO_APREHENSION, se tomó en cuenta un sitio específico al cual pertenezcan los registros existentes en este, por motivo de confidencialidad y la cantidad de registros existentes se mostrará una parte en la Tabla 20.

Tabla 20

Categorización atributo SITIO_APREHENSION

CATEGORÍA	VALOR
ALOAG-ECUADOR	VIA ALOAG – QUITO PEAJE – ALOAG
AMBATO-ECUADOR	AMBATO - MERCADO MAYORISTA AVENIDA BOLIVARIANA Y NELSON DUEÑAS PROV TUNGURAHUA
AMBUQUI-ECUADOR	AMBUQUI PEAJE DE AMBUQUI
ARENILLAS-ECUADOR	SECTOR SAN ANTONIO -ARENILLAS REDONDEL DE ARENILLAS SECTOR AEROPUERTO DE ARENILLAS SECTOR EL JOBO - CANTON ARENILLAS
CANTON COLTA-ECUADOR	CANTON COLTA PROVINCIA DE CHIMBORAZO SECTOR Y DE COLTA - RIOBAMBA SECTOR COLTA LA BALBANEDA PROVINCIA DEL CHIMBORAZO
EL TELEGRAFO-ECUADOR	EL TELEGRAFO SECTOR EL TELEGRAFO DESTACAMENTO MILITAR EL TELEGRAFO CONTROL MILITAR EL TELEGRAFO

FUENTE PROPIA

- **Atributo MARCA**

En la Tabla 21 se muestra una parte de la categorización del atributo MARCA, porque es un atributo con miles de registros entre los cuales se encuentran algunos similares, de esta manera se pretende reducir los resultados dispersos.

Tabla 21

Categorización atributo MARCA

CATEGORÍA	VALOR
AEROPOSTALE	AERO-ESPARTACUS AEROFOX
BIG_DIPPER	BIG_DIPPER_LM80 BIG_DIPPER BIG_DIPPER_LM108 BIG_DIPPER_LS90
CHEVROLET	CHEVROLET_AVEO

CHEVROLET_AVEO_COUPE_CXH-703
CHEVROLET_AVEO_EMOTION
CHEVROLET_CAPTIVA
CHEVROLET_CRUZE
CHEVROLET_LUV
CHEVROLET_OPTRA
CHEVROLET_SWIT
CHEVROLETH
CHEVROLETLIVA'S_ECUADOR_LEVI_STRAUSS&CO
CHEVROLET-TROOPER
MARCA_CHEVROLET_SWIFT

FUENTE PROPIA

- **Atributo CANTIDAD**

En la Tabla 22 se muestra la categorización del atributo CANTIDAD, siguiendo el orden del sistema de numeración, debido a que existen grandes cantidades en los datos.

Tabla 22

Categorización atributo Cantidad

CATEGORÍA	VALORES
UNIDADES	0 a 9.1
DECENAS	10 a 99.1
CENTENAS	100 a 999.1
MILES	1000 a 999999.99
MILLON	1000000 a 9999999.0

FUENTE PROPIA

- **Atributo PRECIO**

En la Tabla 23 se muestra la categorización del atributo PRECIO, la cual se basó en las normas del salario básico del Ecuador.

Tabla 23

Categorización atributo PRECIO

CATEGORÍA	VALOR
BAJO	0.0 a 399.99
MEDIO	400.0 a 3999.99
ALTO	4000.0 a 59999.99
MUY ALTO	60000.0 a 9.999999999E9

FUENTE PROPIA

- **Atributo TOTAL**

Para la categorización del atributo TOTAL se tomó en cuenta las normas del salario básico del Ecuador, como se muestra en la Tabla 24.

Tabla 24

Categorización atributo TOTAL

CATEGORÍA	VALOR
BAJO	0.0 a 400.0
MEDIO	400.0 a 4000.0
ALTO	4000.0 a 59999.99
MUY ALTO	60000.0 a 9.999999999E9

FUENTE PROPIA

2.6.3. Limpieza

En la Fig. 24 y Fig. 25 se procede a mostrar la fase de limpieza de datos en la que se procedió a eliminar o reemplazar caracteres especiales e incongruencias existentes en los registros, como faltas ortográficas, espacios en blanco, tildes, entre otros, que estaban presentes en la mayoría de los atributos de la tabla. En algunos casos el campo se encontraba vacío por lo que se le asignó un valor específico relacionándolo con otros atributos.

#	In stream field	Out stream field	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	Case sensitive
1	BODEGA		N	"		N		N	N
2	BODEGA		N	(		N		N	N
3	BODEGA		N	)		N		N	N
4	BODEGA		N	Á	A	N		N	N
5	BODEGA		N	É	E	N		N	N
6	BODEGA		N	Ó	O	N		N	N
7	BODEGA		N	Í	I	N		N	N
8	DESCRIPCION		N	,	_	N		N	N
9	DESCRIPCION		N	;	_	N		N	N
10	DESCRIPCION		N	:		N		N	N
11	DESCRIPCION		N	"		N		N	N
12	DESCRIPCION		N	(		N		N	N
13	DESCRIPCION		N	)		N		N	N
14	DESCRIPCION		N	•		N		N	N
15	DESCRIPCION		N	”	INCH	N		N	N
16	DESCRIPCION		N	\s	_	N		N	N
17	DESCRIPCION		N	Á	A	N		N	N
18	DESCRIPCION		N	É	E	N		N	N
19	DESCRIPCION		N	Í	I	N		N	N
20	DESCRIPCION		N	Ó	O	N		N	N
21	GRUPO		N	,	_	N		N	N
22	GRUPO		N	Á	A	N		N	N
23	GRUPO		N	É	E	N		N	N
24	GRUPO		N	Í	I	N		N	N
25	GRUPO		N	Ó	O	N		N	N
26	GRUPO_OPERATIVO		N	“		N		N	N
27	GRUPO_OPERATIVO		N	(		N		N	N
28	GRUPO_OPERATIVO		N	)		N		N	N
29	GRUPO_OPERATIVO		N	Á	A	N		N	N
30	GRUPO_OPERATIVO		N	É	E	N		N	N
31	GRUPO_OPERATIVO		N	Ó	O	N		N	N
32	GRUPO_OPERATIVO		N	Í	I	N		N	N
33	MARCA		N	"		N		N	N
34	MARCA		N	"		N		N	N
35	MARCA		N	.		N		N	N
36	MARCA		N	/		N		N	N
37	MARCA		N	*		N		N	N
38	MARCA		N	(		N		N	N

Figura 24 Limpieza de datos parte 1

#	In stream field	Out stream field	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	Case sensitive
39	MARCA		N	)		N		N	N
40	MARCA		N	,		N		N	N
41	MARCA		N	´		N		N	N
42	MARCA		N	"		N		N	N
43	MARCA		N		_	N		N	N
44	MARCA		N	\\s		N		N	N
45	MARCA		N	Á	A	N		N	N
46	MARCA		N	É	E	N		N	N
47	MARCA		N	Ó	O	N		N	N
48	MARCA		N	Í	I	N		N	N
49	ORIGEN_APREHENSION		N	"		N		N	N
50	ORIGEN_APREHENSION		N	Á	A	N		N	N
51	ORIGEN_APREHENSION		N	É	E	N		N	N
52	ORIGEN_APREHENSION		N	Ó	O	N		N	N
53	ORIGEN_APREHENSION		N	Í	I	N		N	N
54	PROCEDENCIA		N	(		N		N	N
55	PROCEDENCIA		N	)		N		N	N
56	SITIO_APREHENSION		N	"		N		N	N
57	SITIO_APREHENSION		N	"		N		N	N
58	SITIO_APREHENSION		N	(		N		N	N
59	SITIO_APREHENSION		N	)		N		N	N
60	SITIO_APREHENSION		N	Á	A	N		N	N
61	SITIO_APREHENSION		N	É	E	N		N	N
62	SITIO_APREHENSION		N	Í	I	N		N	N
63	SITIO_APREHENSION		N	Ó	O	N		N	N
64	SUBGRUPO		N	,	_	N		N	N
65	SUBGRUPO		N	""		N		N	N
66	SUBGRUPO		N	-	/	N		N	N
67	SUBGRUPO		N	.		N		N	N
68	SUBGRUPO		N	Á	A	N		N	N
69	SUBGRUPO		N	É	E	N		N	N
70	SUBGRUPO		N	Ó	O	N		N	N
71	SUBGRUPO		N	Í	I	N		N	N
72	UNIDADES		N	(		N		N	N
73	UNIDADES		N	)		N		N	N
74	UNIDADES		N	Ó	O	N		N	N

Figura 25 Limpieza de datos parte 2

2.6.4. Implementación Norma ISO

La calidad de los datos es un factor muy importante, debido a que el acierto de las decisiones tomadas por una organización depende en gran medida de la calidad de la información (Vila, 2019). Para el inicio de la aplicación de modelos predictivos se tomó en cuenta la característica de **consistencia** de los datos inherentes de la norma ISO/IEC 25012, que permite identificar la coherencia de los datos, se muestra en la Tabla 25.

Tabla 25

Evaluación ISO/IEC 25012

ATRIBUTO	Cantidad Datos no Coherentes	Valor Función de Medición	%
CAT_CANTIDAD	0	1	100
UNIDADES	10	0.99	99
CAT_PRECIO	0	1	100
TOTAL	0	1	100
CAT_TOTAL	0	1	100
GRUPO	3807	0.96	96
SUBGRUPO	1301	0.98	98
DESCRIPCION	0	1	100
PROCEDENCIA	11753	0.88	88
BODEGA	2949	0.97	97
MARCA	0	1	100
STATUS	0	1	100
ORIGEN_APREHENSION	19	0.99	99
SITIO_APREHENSION	0	1	100
GRUPO_OPERATIVO	106	0.99	99
DISTRITO	0	1	100
ZONA	0	1	100

Fuente Propia

2.7. Fase de Minería de datos

Una vez terminado el proceso de selección, limpieza y transformación, se obtiene la vista minable en formato *.xlsx, la cual se procede a transformarlo en un archivo *.csv (comma-separated values), debido a que en esta fase se emplea la herramienta Weka la cual reconoce archivos con esta extensión. En la Fig. 26 se observa una parte de los datos en formato *.csv.

```
ANIO,MES,DIA,CAT_CANTIDAD,CANTIDAD,UNIDADES,CAT_PRECIO,PRECIO,CAT_TOTAL,TOTAL,GRUPO,SUB
GRUPO,DESCRIPCION,PROCEDENCIA,BODEGA,MARCA,STATUS,ORIGEN_APREHENSION,SITIO_APREHENSION
,GRUPO_OPERATIVO,DISTRITO,ZONA
2015,1,14,DECENAS,70,UNIDAD,BAJO,20,MEDIO,1400,FRUTAS Y
COMESTIBLES,CEBOLLA,CEBOLLA,EXTRANJERA,DIRECCION DISTRITAL PUERTO
BOLIVAR,SM,BUENO,ACCIONES INTELIGENCIA Y PROTECCION - DIP,RIO SIETE-ECUADOR,CUERPO DE
VIGILANCIA ADUANERA,VIIDPB,ZONA2
2015,1,19,UNIDADES,1,UNIDAD,MEDIO,1100,MEDIO,1100,ELECTRONICA Y SUS ACCESORIOS,TELEVISORES,TV
DE 49,EXTRANJERA,DIRECCION DISTRITAL TULCAN,SONY,BUENO,CONTROL DE RUTINA,EL BARRIAL-
ECUADOR,CUERPO DE VIGILANCIA ADUANERA,VIIIDC,ZONA1
2015,1,19,UNIDADES,1,UNIDAD,MEDIO,900,MEDIO,900,ELECTRONICA Y SUS ACCESORIOS,TELEVISORES,TV DE
40,EXTRANJERA,DIRECCION DISTRITAL TULCAN,SONY,BUENO,CONTROL DE RUTINA,EL BARRIAL-
ECUADOR,CUERPO DE VIGILANCIA ADUANERA,VIIIDC,ZONA1
2015,1,19,UNIDADES,1,UNIDAD,MEDIO,400,MEDIO,400,ELECTRONICA Y SUS ACCESORIOS,ACC
ELECTRONICA,ASPIRADORA,EXTRANJERA,DIRECCION DISTRITAL TULCAN,ELECTROLUX,BUENO,CONTROL
DE RUTINA,EL BARRIAL-ECUADOR,CUERPO DE VIGILANCIA ADUANERA,VIIIDC,ZONA1
2015,1,19,UNIDADES,1,UNIDAD,MEDIO,800,MEDIO,800,ELECTRONICA Y SUS ACCESORIOS,TELEVISORES,TV DE
40,EXTRANJERA,DIRECCION DISTRITAL TULCAN,PANASONIC,BUENO,CONTROL DE RUTINA,EL BARRIAL-
ECUADOR,CUERPO DE VIGILANCIA ADUANERA,VIIIDC,ZONA1
2015,1,19,UNIDADES,1,UNIDAD,MEDIO,900,MEDIO,900,ELECTRONICA Y SUS ACCESORIOS,TELEVISORES,TV
DE 29,EXTRANJERA,DIRECCION DISTRITAL TULCAN,SAMSUNG,BUENO,CONTROL DE RUTINA,EL BARRIAL-
ECUADOR,CUERPO DE VIGILANCIA ADUANERA,VIIIDC,ZONA1
2015,1,19,UNIDADES,1,UNIDAD,MEDIO,400,MEDIO,400,ELECTRONICA Y SUS ACCESORIOS,TELEVISORES,TV DE
32,EXTRANJERA,DIRECCION DISTRITAL TULCAN,SONY,BUENO,CONTROL DE RUTINA,EL BARRIAL-
ECUADOR,CUERPO DE VIGILANCIA ADUANERA,VIIIDC,ZONA1
2015,1,19,UNIDADES,1,UNIDAD,MEDIO,800,MEDIO,800,ELECTRONICA Y SUS ACCESORIOS,TELEVISORES,TV DE
40,EXTRANJERA,DIRECCION DISTRITAL TULCAN,SONY,BUENO,CONTROL DE RUTINA,EL BARRIAL-
ECUADOR,CUERPO DE VIGILANCIA ADUANERA,VIIIDC,ZONA1
2015,1,21,UNIDADES,2,UNIDAD,BAJO,70,BAJO,140,LICORES,WHISKY VARIAS MARCAS,BOTELLAS DE
WHISKY,EXTRANJERA,DESTACAMENTO CHACRAS,JOHNNIE_WALKER,BUENO,CONTROL DE
RUTINA,CHACRAS-ECUADOR,CUERPO DE VIGILANCIA ADUANERA,IXDH,ZONA2
2015,1,21,UNIDADES,2,UNIDAD,BAJO,110,BAJO,220,LICORES,WHISKY VARIAS MARCAS,BOTELLAS DE
WHISKY,EXTRANJERA,DESTACAMENTO CHACRAS,JOHNNIE_WALKER,BUENO,CONTROL DE
RUTINA,CHACRAS-ECUADOR,CUERPO DE VIGILANCIA ADUANERA,IXDH,ZONA2
```

*Figura 26 Vista Minable en formato *.csv*

2.7.1.1. Algoritmo PCA

El análisis de componentes principales es una técnica de estadística para simplificar un conjunto de datos por matriz de rotación, de la cual la mayor varianza de cualquier proyección se encuentra en la primera coordenada, la segunda mayor varianza en la segunda y así sucesivamente (Li, 2018). Las ideas básicas del algoritmo PCA (Análisis de Componentes Principales) es tomar una serie de transformación lineal para la información original que cuenta con ciertas características de

correlación, luego a través del algoritmo PCA transformar la información en un conjunto de información con menor número de características (Han et al., 2016). En la Fig. 27 se muestra los resultados del algoritmo PCA aplicado en la herramienta WEKA con 17 atributos numéricos.

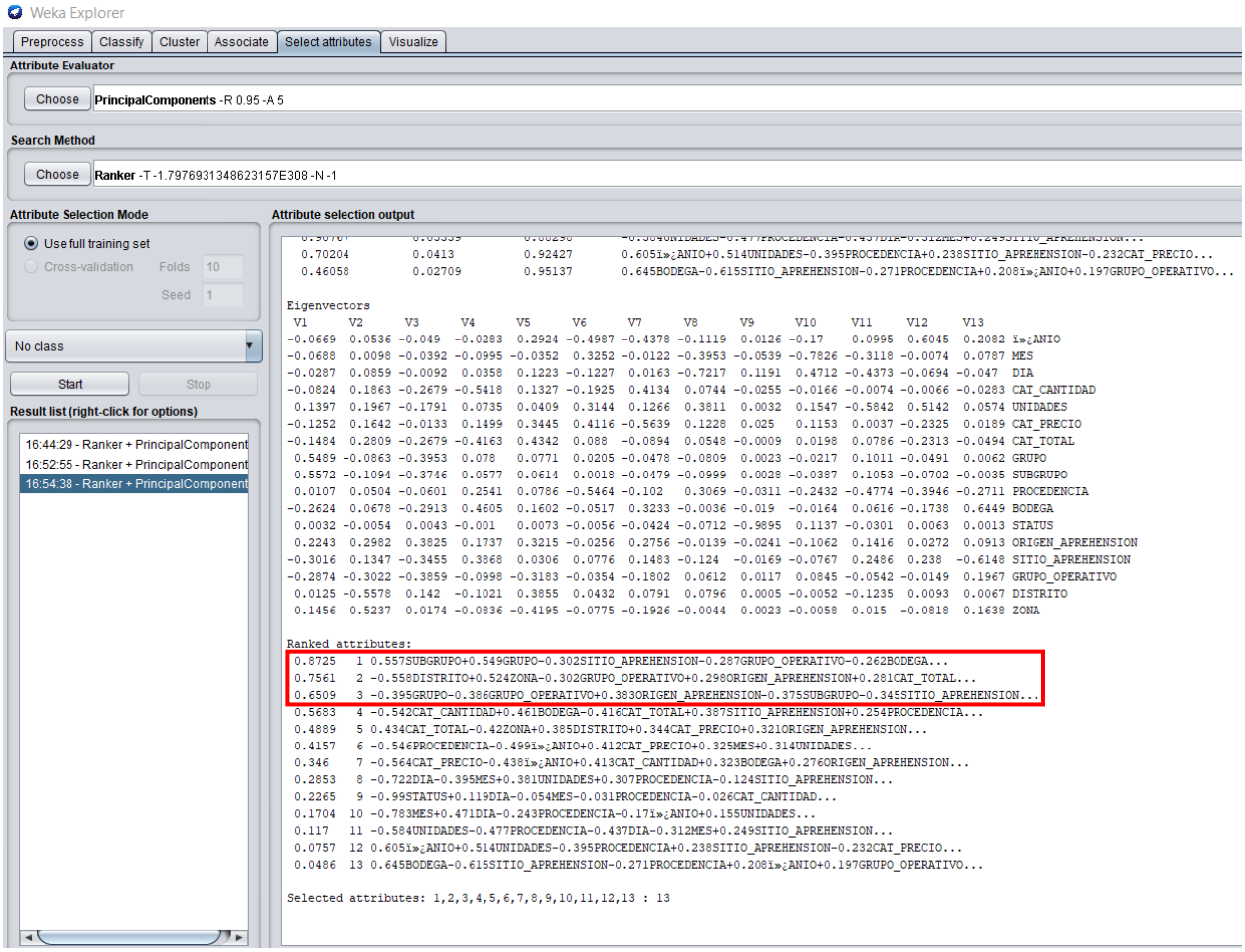


Figura 27 Resultado Algoritmo PCA

Del resultado mostrado en la Fig. 27 realizando el análisis con una variación de 0.95 y ninguna variable en concreto se procede a seleccionar los resultados más acertados, de tal manera que los conjuntos de datos más concordantes a esto se muestran en la Tabla 26.

Tabla 26

Resultado análisis de componentes principales (PCA)

N°	% Representativo	Atributos
PC1	0,873	SUBGRUPO GRUPO SITIO_APREHENSION GRUPO OPERATIVO BODEGA
PC2	0,756	DISTRITO ZONA GRUPO_OPERATIVO ORIGEN_APREHENSION CAT_TOTAL
PC3	0,651	SUB_GRUPO GRUPO_OPERATIVO ORIGEN_APREHENSION SUBGRUPO SITIO_APREHENSION

Fuente WEKA

2.7.2. Clasificación

Árboles de Decisión

Para realizar la clasificación se tomó en cuenta la técnica de árboles de decisión con el algoritmo *J48*, debido a la rapidez de ejecución del modelo y a la simplicidad de interpretación de los resultados, además de estudios realizados con este algoritmo, entre los cuales se encuentra Valenga (2008). Adicionalmente, se consideró los algoritmos *RepTree* y *RandomTree* también por su simplicidad del modelo y la rapidez de ejecución.

2.7.3. Regresión

Para aplicar la regresión al proyecto se utilizará una herramienta diferente a la especificada la cual es *Knime*, esta permite aplicar diferentes algoritmos de regresión como Linear/Polinomial Regression, Random Forest Regression, Logistic Regression y Gradient Boosting Regression. En este proyecto aplicaremos Logistic Regression y se detalla a continuación.

Logistic Regression

Se tomó en cuenta el algoritmo *Logistic Regression* por su facilidad y rapidez de procesamiento de los datos. Este algoritmo funciona con datos numéricos y cadenas de texto, en este caso se utilizará cadenas de texto por la facilidad de interpretación de los resultados.

CAPÍTULO 3
RESULTADOS

3.1. Fase de evaluación e interpretación

Para evaluar los algoritmos de predicción aplicados (clasificación y regresión) en este estudio, se tomó en cuenta las métricas cuantitativas de calidad derivadas de la matriz de confusión.

3.1.1. Evaluación de Algoritmos de Clasificación

Árboles de Decisión

PC1

- **Algoritmo J48**

Parte de los resultados de la aplicación del algoritmo J48, con validación cruzada de 10-hold, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 28 y Fig. 29.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-10,13-14,16-18,21-22
Instances:    102667
Attributes:   5
              GRUPO
              SUBGRUPO
              BODEGA
              SITIO_APREHENSION
              GRUPO_OPERATIVO
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

SITIO_APREHENSION = RIO SIETE-ECUADOR: CUERPO DE VIGILANCIA ADUANERA (2224.0/46.0)
SITIO_APREHENSION = EL BARRIAL-ECUADOR: CUERPO DE VIGILANCIA ADUANERA (356.0/70.0)
SITIO_APREHENSION = RUMICHACA-ECUADOR
|   SUBGRUPO = CEBOLLA: POLICIA NACIONAL (14.0/6.0)
|   SUBGRUPO = TELEVISORES: CUERPO DE VIGILANCIA ADUANERA (1259.0/98.0)
|   SUBGRUPO = ACC ELECTRONICA: CUERPO DE VIGILANCIA ADUANERA (57.0/24.0)
|   SUBGRUPO = PRENDAS DE VESTIR / NUEVAS: CUERPO DE VIGILANCIA ADUANERA (797.0/378.0)
|   SUBGRUPO = MEDIAS DE VESTIR_ NYLON_ PANTYS_ TOBILLERAS…: CUERPO DE VIGILANCIA ADUANERA (36.0/15.0)
|   SUBGRUPO = WHISKY VARIAS MARCAS: CUERPO DE VIGILANCIA ADUANERA (173.0/37.0)
|   SUBGRUPO = TEQUILA: CUERPO DE VIGILANCIA ADUANERA (10.0)
|   SUBGRUPO = CALZADO EN GENERAL: CUERPO DE VIGILANCIA ADUANERA (647.0/285.0)
|   SUBGRUPO = MERCADERIA SURTIDA: CUERPO DE VIGILANCIA ADUANERA (450.0/213.0)
|   SUBGRUPO = VODKA: CUERPO DE VIGILANCIA ADUANERA (2.0/1.0)
|   SUBGRUPO = LICORES VARIOS: CUERPO DE VIGILANCIA ADUANERA (38.0/10.0)
|   SUBGRUPO = LECHE: CUERPO DE VIGILANCIA ADUANERA (0.0)
|   SUBGRUPO = CARTERAS_ CINTURONES_ BILLETERAS_ MOCHILAS_ BOLSOS: CUERPO DE VIGILANCIA ADUANERA (73.0/20.0)
|   SUBGRUPO = VARIAS MARCAS: POLICIA NACIONAL (113.0/57.0)
|   SUBGRUPO = VINO: CUERPO DE VIGILANCIA ADUANERA (6.0)
|   SUBGRUPO = RON: CUERPO DE VIGILANCIA ADUANERA (8.0/1.0)
|   SUBGRUPO = JUEGOS PIROTECNICOS: CUERPO DE VIGILANCIA ADUANERA (14.0/1.0)
|   SUBGRUPO = JUGUETES: CUERPO DE VIGILANCIA ADUANERA (43.0/16.0)
```

Figura 28 Algoritmo J48-PC1 (Parte 1)

```
Size of the tree:            8199


Time taken to build model: 0.3 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       86372               84.1283 %
Incorrectly Classified Instances     16295               15.8717 %
Kappa statistic                          0.5122
Mean absolute error                      0.059
Root mean squared error                  0.1743
Relative absolute error                 60.5222 %
Root relative squared error             78.9455 %
Total Number of Instances           102667

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,966    0,527    0,856      0,966   0,908      0,542    0,849     0,932     CUERPO DE VIGILANCIA ADUANERA
                 0,492    0,016    0,768      0,492   0,600      0,584    0,912     0,666     CONTROL CONJUNTO INTERINSTITUCIONAL
                 0,417    0,022    0,733      0,417   0,532      0,508    0,845     0,580     POLICIA NACIONAL
                 0,528    0,000    0,824      0,528   0,644      0,659    0,849     0,453     DESCONOCIDO
                 0,149    0,001    0,647      0,149   0,243      0,308    0,798     0,202     FUERZAS ARMADAS
                 0,000    0,000    ?          0,000   ?          ?        0,333     0,000     INSPECTORIA DE PESCA
                 0,500    0,000    1,000      0,500   0,667      0,707    0,747     0,500     ARCSA
                 0,364    0,000    1,000      0,364   0,533      0,603    0,817     0,381     AGROCALIDAD
Weighted Avg.    0,841    0,407    ?          0,841   ?          ?        0,854     0,853

=== Confusion Matrix ===

a      b     c     d   e    f  g  h  <-- classified as
75825  1223  1374  12  51   0  0  0 |  a = CUERPO DE VIGILANCIA ADUANERA
4582   4904  460   0   21   0  0  0 |  b = CONTROL CONJUNTO INTERINSTITUCIONAL
7339   202   5414  0   19   0  0  0 |  c = POLICIA NACIONAL
45     5     0     56  0    0  0  0 |  d = DESCONOCIDO
767    49    136   0   167  0  0  0 |  e = FUERZAS ARMADAS
1      0     0     0   0    0  0  0 |  f = INSPECTORIA DE PESCA
1      1     0     0   0    0  2  0 |  g = ARCSA
3      1     3     0   0    0  0  4 |  h = AGROCALIDAD
```

Figura 29 Algoritmo J48-PC1 (Parte 2)

En la Tabla 27 se aprecia las métricas estadísticas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC1 es de 15,8%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.51, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.854, la precisión de las variables analizadas varía de 0.65 a 1, sin embargo la precisión total se muestra cómo "?", porque dos valores son desconocidos, el RECALL o cantidad de interpretación es del 0.84, el valor de verdaderos positivos (TP Rate) indica un valor de 0.84, en cuanto a los Falsos Positivos (FP Rate) de 0.407 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 27

Métricas Estadísticas Algoritmo J48-PC1

MEDIDA	VALOR
Tasa de error	15.8717%
Coeficiente KAPPA	0.5122
Curva ROC	0.854
Precisión	?
Recall	0.841
TP Rate	0.841
FP Rate	0.407
F-Measure	?

Fuente Propia

- **Algoritmo RepTree**

Parte de los resultados de la aplicación del algoritmo RepTree, con validación cruzada de 10-hold, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 30 y Fig. 31.

```
=== Run information ===

Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-10,13-14,16-18,21-22
Instances:    102667
Attributes:   5
              GRUPO
              SUBGRUPO
              BODEGA
              SITIO_APREHENSION
              GRUPO_OPERATIVO
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


REPTree
============

SITIO_APREHENSION = RIO SIETE-ECUADOR : CUERPO DE VIGILANCIA ADUANERA (1474/36) [750/10]
SITIO_APREHENSION = EL BARRIAL-ECUADOR : CUERPO DE VIGILANCIA ADUANERA (237/44) [119/26]
SITIO_APREHENSION = RUMICHACA-ECUADOR
|   SUBGRUPO = CEBOLLA : POLICIA NACIONAL (7/2) [7/4]
|   SUBGRUPO = TELEVISORES : CUERPO DE VIGILANCIA ADUANERA (827/57) [432/41]
|   SUBGRUPO = ACC ELECTRONICA : CUERPO DE VIGILANCIA ADUANERA (42/16) [15/8]
|   SUBGRUPO = PRENDAS DE VESTIR / NUEVAS : CUERPO DE VIGILANCIA ADUANERA (524/242) [273/136]
|   SUBGRUPO = MEDIAS DE VESTIR_ NYLON_ PANTYS_ TOBILLERASâ€¦ : CUERPO DE VIGILANCIA ADUANERA (25/11) [11/4]
|   SUBGRUPO = WHISKY VARIAS MARCAS : CUERPO DE VIGILANCIA ADUANERA (108/20) [65/17]
|   SUBGRUPO = TEQUILA : CUERPO DE VIGILANCIA ADUANERA (6/0) [4/0]
|   SUBGRUPO = CALZADO EN GENERAL
|   |   BODEGA = DIRECCION DISTRITAL PUERTO BOLIVAR : CUERPO DE VIGILANCIA ADUANERA (1/0) [0/0]
|   |   BODEGA = DIRECCION DISTRITAL TULCAN : CUERPO DE VIGILANCIA ADUANERA (353/150) [184/75]
|   |   BODEGA = DESTACAMENTO CHACRAS : CUERPO DE VIGILANCIA ADUANERA (7/2) [4/2]
|   |   BODEGA = DIRECCION DISTRITAL CUENCA : CUERPO DE VIGILANCIA ADUANERA (0/0) [0/0]
|   |   BODEGA = DIRECCION GENERAL SENAE : CUERPO DE VIGILANCIA ADUANERA (0/0) [0/0]
|   |   BODEGA = DESTACAMENTO YAHUARCOCHA : CUERPO DE VIGILANCIA ADUANERA (0/0) [0/0]
|   |   BODEGA = DIRECCION DISTRITAL MACARA : CUERPO DE VIGILANCIA ADUANERA (0/0) [0/0]
|   |   BODEGA = INSTALACIONES DEL VII DISTRITO EL ORO : CUERPO DE VIGILANCIA ADUANERA (1/0) [0/0]
|   |   BODEGA = INSTALACIONES DEL VI DISTRITO LOJA : CUERPO DE VIGILANCIA ADUANERA (0/0) [0/0]
```

Figura 30 Algoritmo RepTree-PC1 (Parte 1)

Size of the tree : 15063

Time taken to build model: 1.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	86561	84.3124 %
Incorrectly Classified Instances	16106	15.6876 %
Kappa statistic	0.5284	
Mean absolute error	0.056	
Root mean squared error	0.1716	
Relative absolute error	57.3665 %	
Root relative squared error	77.7119 %	
Total Number of Instances	102667	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,960	0,499	0,862	0,960	0,908	0,551	0,874	0,943	CUERPO DE VIGILANCIA ADUANERA
	0,535	0,021	0,737	0,535	0,620	0,595	0,913	0,667	CONTROL CONJUNTO INTERINSTITUCIONAL
	0,431	0,022	0,737	0,431	0,544	0,519	0,874	0,615	POLICIA NACIONAL
	0,528	0,000	0,824	0,528	0,644	0,659	0,867	0,447	DESCONOCIDO
	0,175	0,001	0,624	0,175	0,274	0,327	0,845	0,234	FUERZAS ARMADAS
	0,000	0,000	?	0,000	?	?	0,498	0,000	INSPECTORIA DE PESCA
	0,500	0,000	1,000	0,500	0,667	0,707	0,749	0,500	ARCSA
	0,545	0,000	1,000	0,545	0,706	0,739	0,818	0,555	AGROCALIDAD
Weighted Avg.	0,843	0,387	?	0,843	?	?	0,878	0,866	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	<-- classified as
75375	1553	1477	12	68	0	0	0	a = CUERPO DE VIGILANCIA ADUANERA
4207	5332	401	0	27	0	0	0	b = CONTROL CONJUNTO INTERINSTITUCIONAL
7070	287	5594	0	23	0	0	0	c = POLICIA NACIONAL
45	5	0	56	0	0	0	0	d = DESCONOCIDO
748	55	120	0	196	0	0	0	e = FUERZAS ARMADAS
1	0	0	0	0	0	0	0	f = INSPECTORIA DE PESCA
1	1	0	0	0	0	2	0	g = ARCSA
3	1	1	0	0	0	0	6	h = AGROCALIDAD

Figura 31 Algoritmo RepTree-PC1 (Parte 2)

En la Tabla 28 se aprecia las métricas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC1 es de 15,7%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.52, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.878, la precisión de las variables analizadas varía de 0.6 a 1, sin embargo la precisión total se muestra cómo "?", porque dos valores son desconocidos, el RECALL o cantidad de interpretación es del 0.84, el valor de verdaderos positivos (TP Rate) indica un valor de 0.84, en cuanto a los Falsos Positivos (FP Rate) de 0.387 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 28

Métricas Estadísticas Algoritmo RepTree-PC1

MEDIDA	VALOR
Tasa de error	15.6876%
Coeficiente KAPPA	0.5284
Curva ROC	0.878
Precisión	?
Recall	0.843
TP Rate	0.843
FP Rate	0.387
F-Measure	?

Fuente Propia

- **Algoritmo RandomTree**

Parte de los resultados de la aplicación del algoritmo RandomTree, con validación cruzada de 10-hold, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 32 y Fig. 33.

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-10,13-14,16-18,21-22
Instances:    102667
Attributes:   5
              GRUPO
              SUBGRUPO
              BODEGA
              SITIO_APREHENSION
              GRUPO_OPERATIVO
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


RandomTree
==========

SITIO_APREHENSION = RIO SIETE-ECUADOR
|  GRUPO = FRUTAS Y COMESTIBLES
|  |  SUBGRUPO = CEBOLLA
|  |  |  BODEGA = DIRECCION DISTRITAL PUERTO BOLIVAR : CUERPO DE VIGILANCIA ADUANERA (24/2)
|  |  |  BODEGA = DIRECCION DISTRITAL TULCAN : CUERPO DE VIGILANCIA ADUANERA (1/0)
|  |  |  BODEGA = DESTACAMENTO CHACRAS : CONTROL CONJUNTO INTERINSTITUCIONAL (1/0)
|  |  |  BODEGA = DIRECCION DISTRITAL CUENCA : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = DIRECCION GENERAL SENAE : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = DESTACAMENTO YAHUARCOCHA : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = DIRECCION DISTRITAL MACARA : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = INSTALACIONES DEL VII DISTRITO EL ORO : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = INSTALACIONES DEL VI DISTRITO LOJA : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = DIRECCION DISTRITAL GUAYAS : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = DIRECCION DISTRITAL ESMERALDAS : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = BODEGA HUAQUILLAS : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = DIRECCION DISTRITAL MANABI : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = DIRECCION DISTRITAL LATACUNGA : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = INSTALACIONES DEL V DISTRITO AZUAY : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = BODEGAS DE CALDERON  : CUERPO DE VIGILANCIA ADUANERA (0/0)
|  |  |  BODEGA = DIRECCION DISTRITAL QUITO : CUERPO DE VIGILANCIA ADUANERA (1/0)
```

Figura 32 Algoritmo RandomTree-PC1 (Parte 1)

```
Size of the tree : 90127

Time taken to build model: 0.23 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       86891               84.6338 %
Incorrectly Classified Instances     15776               15.3662 %
Kappa statistic                          0.5479
Mean absolute error                      0.0525
Root mean squared error                  0.1694
Relative absolute error                 53.872  %
Root relative squared error             76.7281 %
Total Number of Instances           102667

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,956    0,467    0,869      0,956    0,911      0,567    0,883     0,947     CUERPO DE VIGILANCIA ADUANERA
               0,570    0,022    0,736      0,570    0,642      0,615    0,908     0,676     CONTROL CONJUNTO INTERINSTITUCIONAL
               0,455    0,026    0,719      0,455    0,557      0,526    0,885     0,630     POLICIA NACIONAL
               0,500    0,000    0,791      0,500    0,613      0,629    0,884     0,498     DESCONOCIDO
               0,201    0,001    0,623      0,201    0,304      0,350    0,846     0,272     FUERZAS ARMADAS
               0,000    0,000    ?          0,000    ?          ?        0,498     0,000     INSPECTORIA DE PESCA
               0,500    0,000    1,000      0,500    0,667      0,707    0,750     0,500     ARCSA
               0,636    0,000    0,875      0,636    0,737      0,746    0,818     0,547     AGROCALIDAD
Weighted Avg.  0,846    0,362    ?          0,846    ?          ?        0,885     0,873

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
 75029  1647  1706    14    88     0     0     1 |     a = CUERPO DE VIGILANCIA ADUANERA
  3807  5678   455     0    27     0     0     0 |     b = CONTROL CONJUNTO INTERINSTITUCIONAL
  6733   323  5897     0    21     0     0     0 |     c = POLICIA NACIONAL
    48     4     1    53     0     0     0     0 |     d = DESCONOCIDO
   693    62   139     0   225     0     0     0 |     e = FUERZAS ARMADAS
     1     0     0     0     0     0     0     0 |     f = INSPECTORIA DE PESCA
     1     1     0     0     0     0     2     0 |     g = ARCSA
     2     1     1     0     0     0     0     7 |     h = AGROCALIDAD
```

Figura 33 Algoritmo RandomTree-PC1 (Parte 2)

En la Tabla 29 se aprecia las métricas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC1 es de 15,4%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.54, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.88, la precisión de las variables analizadas varía de 0.6 a 1, sin embargo la precisión total se muestra cómo "?", porque dos valores son desconocidos, el RECALL o cantidad de interpretación es del 0.84, el valor de verdaderos positivos (TP Rate) indica un valor de 0.84, en cuanto a los Falsos Positivos (FP Rate) de 0.362 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 29

Métricas Estadísticas Algoritmo RandomTree-PC1

MEDIDA	VALOR
Tasa de error	15.3662%
Coeficiente KAPPA	0.5479
Curva ROC	0.885
Precisión	?
Recall	0.846
TP Rate	0.846
FP Rate	0.362
F-Measure	?

Fuente Propia

PC2

- **Algoritmo J48**

Parte de los resultados de la aplicación del algoritmo J48, con validación cruzada de 10, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 34 y Fig. 35.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-8,10-17,19
Instances:    102667
Attributes:   5
              CAT_TOTAL
              ORIGEN_APREHENSION
              GRUPO_OPERATIVO
              DISTRITO
              ZONA
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

ORIGEN_APREHENSION = ACCIONES INTELIGENCIA Y PROTECCION - DIP: CUERPO DE VIGILANCIA ADUANERA (5440.0/417.0)
ORIGEN_APREHENSION = CONTROL DE RUTINA: CUERPO DE VIGILANCIA ADUANERA (63523.0/1206.0)
ORIGEN_APREHENSION = OPERATIVO CONJUNTO: CONTROL CONJUNTO INTERINSTITUCIONAL (7044.0/362.0)
ORIGEN_APREHENSION = ACTA DE ENTREGA - RECEPCION: POLICIA NACIONAL (15180.0/2287.0)
ORIGEN_APREHENSION = DENUNCIA: CUERPO DE VIGILANCIA ADUANERA (4826.0/175.0)
ORIGEN_APREHENSION = LLAMADA TELEFONICA
|   DISTRITO = VIIDPB: CUERPO DE VIGILANCIA ADUANERA (1.0)
|   DISTRITO = VIIIDC: CONTROL CONJUNTO INTERINSTITUCIONAL (87.0/38.0)
|   DISTRITO = IXDH: CONTROL CONJUNTO INTERINSTITUCIONAL (11.0/4.0)
|   DISTRITO = VDA: CONTROL CONJUNTO INTERINSTITUCIONAL (0.0)
|   DISTRITO = IDG: CUERPO DE VIGILANCIA ADUANERA (1.0)
|   DISTRITO = IVDQ
|   |   CAT_TOTAL = MEDIO: CUERPO DE VIGILANCIA ADUANERA (5.0/1.0)
|   |   CAT_TOTAL = BAJO: CONTROL CONJUNTO INTERINSTITUCIONAL (19.0/5.0)
|   |   CAT_TOTAL = ALTO: CUERPO DE VIGILANCIA ADUANERA (3.0/1.0)
|   |   CAT_TOTAL = MUY ALTO: CONTROL CONJUNTO INTERINSTITUCIONAL (0.0)
|   DISTRITO = VIDL: CUERPO DE VIGILANCIA ADUANERA (11.0/1.0)
|   DISTRITO = IIIDE: CUERPO DE VIGILANCIA ADUANERA (16.0/4.0)
|   DISTRITO = IIDM: CONTROL CONJUNTO INTERINSTITUCIONAL (0.0)
|   DISTRITO = XDLAT: CONTROL CONJUNTO INTERINSTITUCIONAL (0.0)
ORIGEN_APREHENSION = DESCONOCIDO: DESCONOCIDO (19.0/2.0)
```

Figura 34 Algoritmo J48-PC2 (Parte 1)

```
Size of the tree :      75


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       97495               94.9624 %
Incorrectly Classified Instances      5172                5.0376 %
Kappa statistic                          0.87
Mean absolute error                      0.0226
Root mean squared error                  0.1065
Relative absolute error                 23.2138 %
Root relative squared error             48.2376 %
Total Number of Instances           102667

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,975    0,086    0,974      0,975    0,974      0,891    0,958     0,978     CUERPO DE VIGILANCIA ADUANERA
               0,806    0,009    0,910      0,806    0,855      0,843    0,934     0,830     CONTROL CONJUNTO INTERINSTITUCIONAL
               0,996    0,026    0,849      0,996    0,916      0,907    0,985     0,844     POLICIA NACIONAL
               0,160    0,000    0,895      0,160    0,272      0,379    0,774     0,133     DESCONOCIDO
               0,000    0,000    ?          0,000    ?          ?        0,918     0,069     FUERZAS ARMADAS
               0,000    0,000    ?          0,000    ?          ?        0,433     0,000     INSPECTORIA DE PESCA
               0,000    0,000    ?          0,000    ?          ?        0,806     0,000     ARCSA
               0,000    0,000    ?          0,000    ?          ?        0,868     0,001     AGROCALIDAD
Weighted Avg.  0,950    0,070    ?          0,950    ?          ?        0,958     0,936

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
 76526   768  1189     2     0     0     0     0 |     a = CUERPO DE VIGILANCIA ADUANERA
  1905  8036    26     0     0     0     0     0 |     b = CONTROL CONJUNTO INTERINSTITUCIONAL
    41    17 12916     0     0     0     0     0 |     c = POLICIA NACIONAL
    84     3     2    17     0     0     0     0 |     d = DESCONOCIDO
    51     0  1068     0     0     0     0     0 |     e = FUERZAS ARMADAS
     0     0     1     0     0     0     0     0 |     f = INSPECTORIA DE PESCA
     0     3     1     0     0     0     0     0 |     g = ARCSA
     1     0    10     0     0     0     0     0 |     h = AGROCALIDAD
```

Figura 35 Algoritmo J48-PC2 (Parte2)

En la Tabla 30 se aprecia las métricas estadísticas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC2 es de 5,03%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.87, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.958, la precisión de las variables analizadas varía de 0.8 a 1, sin embargo la precisión total se muestra cómo "?", porque cinco valores son desconocidos, el RECALL o cantidad de interpretación es del 0.95, el valor de verdaderos positivos (TP Rate) indica un valor de 0.95, en cuanto a los Falsos Positivos (FP Rate) de 0.07 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 30

Métricas Estadísticas Algoritmo J48-PC2

MEDIDA	VALOR
Tasa de error	5.0371%
Coeficiente KAPPA	0.87
Curva ROC	0.958
Precisión	?
Recall	0.950
TP Rate	0.950
FP Rate	0.070
F-Measure	?

Fuente Propia

- **Algoritmo RepTree**

Parte de los resultados de la aplicación del algoritmo RepTree, con validación cruzada de 10, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 36 y Fig. 37.

```
=== Run information ===

Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-8,10-17,19
Instances:    102667
Attributes:   5
              CAT_TOTAL
              ORIGEN_APREHENSION
              GRUPO_OPERATIVO
              DISTRITO
              ZONA
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


REPTree
============

ORIGEN_APREHENSION = ACCIONES INTELIGENCIA Y PROTECCION - DIP
|   DISTRITO = VIIDPB : CUERPO DE VIGILANCIA ADUANERA (1077/53) [576/34]
|   DISTRITO = VIIIDC : CUERPO DE VIGILANCIA ADUANERA (666/113) [348/62]
|   DISTRITO = IXDH : CUERPO DE VIGILANCIA ADUANERA (226/13) [119/7]
|   DISTRITO = VDA : CUERPO DE VIGILANCIA ADUANERA (309/0) [138/0]
|   DISTRITO = IDG : CUERPO DE VIGILANCIA ADUANERA (446/53) [253/33]
|   DISTRITO = IVDQ : CUERPO DE VIGILANCIA ADUANERA (733/21) [389/18]
|   DISTRITO = VIDL : CUERPO DE VIGILANCIA ADUANERA (54/3) [26/0]
|   DISTRITO = IIIDE : CONTROL CONJUNTO INTERINSTITUCIONAL (3/0) [2/0]
|   DISTRITO = IIDM : CUERPO DE VIGILANCIA ADUANERA (40/0) [10/0]
|   DISTRITO = XDLAT : CUERPO DE VIGILANCIA ADUANERA (17/2) [8/0]
ORIGEN_APREHENSION = CONTROL DE RUTINA : CUERPO DE VIGILANCIA ADUANERA (42428/796) [21095/410]
ORIGEN_APREHENSION = OPERATIVO CONJUNTO : CONTROL CONJUNTO INTERINSTITUCIONAL (4748/245) [2296/117]
ORIGEN_APREHENSION = ACTA DE ENTREGA - RECEPCION : POLICIA NACIONAL (10122/1519) [5058/768]
ORIGEN_APREHENSION = DENUNCIA : CUERPO DE VIGILANCIA ADUANERA (3235/119) [1591/56]
ORIGEN_APREHENSION = LLAMADA TELEFONICA
|   DISTRITO = VIIDPB : CUERPO DE VIGILANCIA ADUANERA (0/0) [1/0]
|   DISTRITO = VIIIDC : CONTROL CONJUNTO INTERINSTITUCIONAL (55/25) [32/13]
|   DISTRITO = IXDH : CONTROL CONJUNTO INTERINSTITUCIONAL (6/2) [5/2]
|   DISTRITO = VDA : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   DISTRITO = IDG : CUERPO DE VIGILANCIA ADUANERA (1/0) [0/0]
|   DISTRITO = IVDQ
|   |   CAT_TOTAL = MEDIO : CUERPO DE VIGILANCIA ADUANERA (4/1) [1/0]
```

Figura 36 Algoritmo RepTree-PC2 (Parte 1)

```
Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       97491               94.9585 %
Incorrectly Classified Instances      5176                5.0415 %
Kappa statistic                          0.87
Mean absolute error                      0.0226
Root mean squared error                  0.1063
Relative absolute error                 23.1366 %
Root relative squared error             48.1494 %
Total Number of Instances           102667

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,975    0,086    0,974      0,975    0,974      0,890    0,958     0,979     CUERPO DE VIGILANCIA ADUANERA
                 0,807    0,009    0,910      0,807    0,855      0,842    0,935     0,833     CONTROL CONJUNTO INTERINSTITUCIONAL
                 0,995    0,026    0,849      0,995    0,916      0,907    0,986     0,844     POLICIA NACIONAL
                 0,160    0,000    0,895      0,160    0,272      0,379    0,782     0,137     DESCONOCIDO
                 0,000    0,000    0,000      0,000    0,000      -0,000   0,920     0,069     FUERZAS ARMADAS
                 0,000    0,000    ?          0,000    ?          ?        0,433     0,000     INSPECTORIA DE PESCA
                 0,000    0,000    ?          0,000    ?          ?        0,806     0,000     ARCSA
                 0,000    0,000    ?          0,000    ?          ?        0,868     0,001     AGROCALIDAD
Weighted Avg.    0,950    0,070    ?          0,950    ?          ?        0,959     0,937

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
 76518   775  1189     2     1     0     0     0 |     a = CUERPO DE VIGILANCIA ADUANERA
  1900  8041    26     0     0     0     0     0 |     b = CONTROL CONJUNTO INTERINSTITUCIONAL
    41    18 12915     0     0     0     0     0 |     c = POLICIA NACIONAL
    84     3     2    17     0     0     0     0 |     d = DESCONOCIDO
    51     0  1068     0     0     0     0     0 |     e = FUERZAS ARMADAS
     0     0     1     0     0     0     0     0 |     f = INSPECTORIA DE PESCA
     0     3     1     0     0     0     0     0 |     g = ARCSA
     1     0    10     0     0     0     0     0 |     h = AGROCALIDAD
```

Figura 37 Algoritmo RepTree-PC2 (Parte 2)

En la Tabla 31 se aprecia las métricas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC2 es de 5,04%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.87, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.959, la precisión de las variables analizadas varía de 0 a 1, sin embargo la precisión total se muestra cómo "?", porque cuatro valores son desconocidos, el RECALL o cantidad de interpretación es del 0.95, el valor de verdaderos positivos (TP Rate) indica un valor de 0.95, en cuanto a los Falsos Positivos (FP Rate) de 0.07 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 31

Métricas Estadísticas Algoritmo RepTree-PC2

MEDIDA	VALOR
Tasa de error	5.0415%
Coeficiente KAPPA	0.87
Curva ROC	0.959
Precisión	?
Recall	0.950
TP Rate	0.950
FP Rate	0.070
F-Measure	?

Fuente Propia

- **Algoritmo RandomTree**

Parte de los resultados de la aplicación del algoritmo RandomTree, con validación cruzada de 10, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 38 y Fig. 39.

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-8,10-17,19
Instances:    102667
Attributes:   5
              CAT_TOTAL
              ORIGEN_APREHENSION
              GRUPO_OPERATIVO
              DISTRITO
              ZONA
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


RandomTree
==========

ORIGEN_APREHENSION = ACCIONES INTELIGENCIA Y PROTECCION - DIP
|   DISTRITO = VIIDPB
|   |   CAT_TOTAL = MEDIO : CUERPO DE VIGILANCIA ADUANERA (510/43)
|   |   CAT_TOTAL = BAJO : CUERPO DE VIGILANCIA ADUANERA (911/16)
|   |   CAT_TOTAL = ALTO : CUERPO DE VIGILANCIA ADUANERA (219/27)
|   |   CAT_TOTAL = MUY ALTO : CUERPO DE VIGILANCIA ADUANERA (13/1)
|   DISTRITO = VIIIDC
|   |   CAT_TOTAL = MEDIO : CUERPO DE VIGILANCIA ADUANERA (427/97)
|   |   CAT_TOTAL = BAJO : CUERPO DE VIGILANCIA ADUANERA (414/42)
|   |   CAT_TOTAL = ALTO : CUERPO DE VIGILANCIA ADUANERA (170/36)
|   |   CAT_TOTAL = MUY ALTO : CUERPO DE VIGILANCIA ADUANERA (3/0)
|   DISTRITO = IXDH
|   |   CAT_TOTAL = MEDIO : CUERPO DE VIGILANCIA ADUANERA (129/7)
|   |   CAT_TOTAL = BAJO : CUERPO DE VIGILANCIA ADUANERA (181/8)
|   |   CAT_TOTAL = ALTO : CUERPO DE VIGILANCIA ADUANERA (34/5)
|   |   CAT_TOTAL = MUY ALTO : CUERPO DE VIGILANCIA ADUANERA (1/0)
|   DISTRITO = VDA : CUERPO DE VIGILANCIA ADUANERA (447/0)
|   DISTRITO = IDG
|   |   CAT_TOTAL = MEDIO : CUERPO DE VIGILANCIA ADUANERA (278/25)
|   |   CAT_TOTAL = BAJO : CUERPO DE VIGILANCIA ADUANERA (212/21)
|   |   CAT_TOTAL = ALTO : CUERPO DE VIGILANCIA ADUANERA (197/37)
|   |   CAT_TOTAL = MUY ALTO : CUERPO DE VIGILANCIA ADUANERA (12/3)
```

Figura 38 Algoritmo RandomTree-PC2 (Parte 1)

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       97478               94.9458 %
Incorrectly Classified Instances      5189                5.0542 %
Kappa statistic                          0.8695
Mean absolute error                      0.0213
Root mean squared error                  0.1036
Relative absolute error                 21.842  %
Root relative squared error             46.9354 %
Total Number of Instances           102667

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,975    0,087    0,973      0,975    0,974      0,890    0,978     0,990     CUERPO DE VIGILANCIA ADUANERA
                 0,804    0,008    0,911      0,804    0,854      0,842    0,967     0,877     CONTROL CONJUNTO INTERINSTITUCIONAL
                 0,995    0,026    0,849      0,995    0,916      0,907    0,992     0,931     POLICIA NACIONAL
                 0,151    0,000    0,889      0,151    0,258      0,366    0,783     0,162     DESCONOCIDO
                 0,002    0,000    0,400      0,002    0,004      0,026    0,952     0,153     FUERZAS ARMADAS
                 0,000    0,000    ?          0,000    ?          ?        0,493     0,000     INSPECTORIA DE PESCA
                 0,000    0,000    ?          0,000    ?          ?        0,747     0,012     ARCSA
                 0,000    0,000    0,000      0,000    0,000      -0,000   0,901     0,005     AGROCALIDAD
Weighted Avg.    0,949    0,071    ?          0,949    ?          ?        0,978     0,962

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
 76537   754  1188     2     3     0     0     1 |     a = CUERPO DE VIGILANCIA ADUANERA
  1928  8011    28     0     0     0     0     0 |     b = CONTROL CONJUNTO INTERINSTITUCIONAL
    43    19 12912     0     0     0     0     0 |     c = POLICIA NACIONAL
    85     3     2    16     0     0     0     0 |     d = DESCONOCIDO
    49     0  1068     0     2     0     0     0 |     e = FUERZAS ARMADAS
     0     0     1     0     0     0     0     0 |     f = INSPECTORIA DE PESCA
     0     3     1     0     0     0     0     0 |     g = ARCSA
     1     0    10     0     0     0     0     0 |     h = AGROCALIDAD
```

Figura 39 Algoritmo RandomTree-PC2 (Parte 2)

En la Tabla 32 se aprecia las métricas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC2 es de 5,05%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.869, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.978, la precisión de las variables analizadas varía de 0 a 1, sin embargo la precisión total se muestra cómo "?#, porque 3 valores son desconocidos, el RECALL o cantidad de interpretación es del 0.949, el valor de verdaderos positivos (TP Rate) indica un valor de 0.949, en cuanto a los Falsos Positivos (FP Rate) de 0.07 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 32

Métricas Estadísticas Algoritmo RandomTree-PC2

MEDIDA	VALOR
Tasa de error	5.0542%
Coeficiente KAPPA	0.8695
Curva ROC	0.978
Precisión	?
Recall	0.949
TP Rate	0.949
FP Rate	0.071
F-Measure	?

Fuente Propia

PC3

- **Algoritmo J48**

Parte de los resultados de la aplicación del algoritmo J48, con validación cruzada de 10, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 40 y Fig. 41.

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-10,13-17,21-22
Instances:    102667
Attributes:   5
              GRUPO
              SUBGRUPO
              ORIGEN_APREHENSION
              SITIO_APREHENSION
              GRUPO_OPERATIVO
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

ORIGEN_APREHENSION = ACCIONES INTELIGENCIA Y PROTECCION - DIP: CUERPO DE VIGILANCIA ADUANERA (5440.0/417.0)
ORIGEN_APREHENSION = CONTROL DE RUTINA: CUERPO DE VIGILANCIA ADUANERA (63523.0/1206.0)
ORIGEN_APREHENSION = OPERATIVO CONJUNTO: CONTROL CONJUNTO INTERINSTITUCIONAL (7044.0/362.0)
ORIGEN_APREHENSION = ACTA DE ENTREGA - RECEPCION
|   SITIO_APREHENSION = RIO SIETE-ECUADOR: POLICIA NACIONAL (7.0)
|   SITIO_APREHENSION = EL BARRIAL-ECUADOR: POLICIA NACIONAL (32.0)
|   SITIO_APREHENSION = RUMICHACA-ECUADOR: POLICIA NACIONAL (3014.0/267.0)
|   SITIO_APREHENSION = CHACRAS-ECUADOR: POLICIA NACIONAL (4078.0/1224.0)
|   SITIO_APREHENSION = YUNGUILLA-ECUADOR: POLICIA NACIONAL (0.0)
|   SITIO_APREHENSION = ARENILLAS-ECUADOR: POLICIA NACIONAL (74.0/2.0)
|   SITIO_APREHENSION = DURAN-ECUADOR: POLICIA NACIONAL (9.0)
|   SITIO_APREHENSION = GUAYAQUIL-ECUADOR: POLICIA NACIONAL (244.0/8.0)
|   SITIO_APREHENSION = ANTONIO ANTE-ECUADOR: POLICIA NACIONAL (20.0/9.0)
|   SITIO_APREHENSION = IBARRA-ECUADOR: POLICIA NACIONAL (1078.0/96.0)
|   SITIO_APREHENSION = MACHALA-ECUADOR: POLICIA NACIONAL (160.0/37.0)
|   SITIO_APREHENSION = EL GUABO-ECUADOR: POLICIA NACIONAL (1.0)
|   SITIO_APREHENSION = PUERTO BOLIVAR-ECUADOR: POLICIA NACIONAL (3068.0/54.0)
|   SITIO_APREHENSION = CANTON EL GUABO-ECUADOR: POLICIA NACIONAL (6.0)
|   SITIO_APREHENSION = CARCHI-ECUADOR
|   |   SUBGRUPO = CEBOLLA: POLICIA NACIONAL (0.0)
|   |   SUBGRUPO = TELEVISORES: FUERZAS ARMADAS (9.0/1.0)
|   |   SUBGRUPO = ACC ELECTRONICA: POLICIA NACIONAL (5.0)
|   |   SUBGRUPO = PRENDAS DE VESTIR / NUEVAS: FUERZAS ARMADAS (22.0/7.0)
|   |   SUBGRUPO = MEDIAS DE VESTIR_ NYLON_ PANTYS_ TOBILLERASâ€¦: FUERZAS ARMADAS (1.0)
```

Figura 40 Algoritmo J48-PC3 (Parte 1)

```
Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       97963               95.4182 %
Incorrectly Classified Instances      4704                4.5818 %
Kappa statistic                          0.8819
Mean absolute error                      0.0198
Root mean squared error                  0.1
Relative absolute error                 20.3106 %
Root relative squared error             45.2806 %
Total Number of Instances           102667

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,977    0,080    0,975      0,977    0,976      0,898    0,963     0,980     CUERPO DE VIGILANCIA ADUANERA
               0,822    0,007    0,926      0,822    0,871      0,859    0,936     0,835     CONTROL CONJUNTO INTERINSTITUCIONAL
               0,992    0,023    0,862      0,992    0,922      0,913    0,992     0,943     POLICIA NACIONAL
               0,160    0,000    0,895      0,160    0,272      0,379    0,783     0,133     DESCONOCIDO
               0,189    0,000    0,844      0,189    0,308      0,396    0,947     0,361     FUERZAS ARMADAS
               0,000    0,000    ?          0,000    ?          ?        0,477     0,000     INSPECTORIA DE PESCA
               0,000    0,000    ?          0,000    ?          ?        0,821     0,000     ARCSA
               0,455    0,000    1,000      0,455    0,625      0,674    0,904     0,477     AGROCALIDAD
Weighted Avg.  0,954    0,065    ?          0,954    ?          ?        0,964     0,953

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
 76664   637  1182     2     0     0     0     0 |     a = CUERPO DE VIGILANCIA ADUANERA
  1756  8190    21     0     0     0     0     0 |     b = CONTROL CONJUNTO INTERINSTITUCIONAL
    46    15 12876     0    37     0     0     0 |     c = POLICIA NACIONAL
    84     3     1    17     1     0     0     0 |     d = DESCONOCIDO
    49     0   859     0   211     0     0     0 |     e = FUERZAS ARMADAS
     0     0     1     0     0     0     0     0 |     f = INSPECTORIA DE PESCA
     0     3     1     0     0     0     0     0 |     g = ARCSA
     1     0     4     0     1     0     0     5 |     h = AGROCALIDAD
```

Figura 41 Algoritmo J48-PC3 (Parte2)

En la Tabla 33 se aprecia las métricas estadísticas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC3 es de 4,58%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.88, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.96, la precisión de las variables analizadas varía de 0,8 a 1, sin embargo la precisión total se muestra cómo "?", porque tres valores son desconocidos, el RECALL o cantidad de interpretación es del 0.95, el valor de verdaderos positivos (TP Rate) indica un valor de 0.95, en cuanto a los Falsos Positivos (FP Rate) de 0.06 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 33

Métricas Estadísticas Algoritmo J48-PC3

MEDIDA	VALOR
Tasa de error	4.5818%
Coeficiente KAPPA	0.8819
Curva ROC	0.964
Precisión	?
Recall	0.954
TP Rate	0.954
FP Rate	0.070
F-Measure	?

Fuente Propia

- **Algoritmo RepTree**

Parte de los resultados de la aplicación del algoritmo RepTree, con validación cruzada de 10, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 42 y Fig. 43.

```
=== Run information ===

Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-10,13-17,21-22
Instances:    102667
Attributes:   5
              GRUPO
              SUBGRUPO
              ORIGEN_APREHENSION
              SITIO_APREHENSION
              GRUPO_OPERATIVO
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


REPTree
============

ORIGEN_APREHENSION = ACCIONES INTELIGENCIA Y PROTECCION - DIP
|   SITIO_APREHENSION = RIO SIETE-ECUADOR : CUERPO DE VIGILANCIA ADUANERA (71/14) [39/7]
|   SITIO_APREHENSION = EL BARRIAL-ECUADOR : CUERPO DE VIGILANCIA ADUANERA (6/0) [1/0]
|   SITIO_APREHENSION = RUMICHACA-ECUADOR : CUERPO DE VIGILANCIA ADUANERA (163/15) [80/14]
|   SITIO_APREHENSION = CHACRAS-ECUADOR : CUERPO DE VIGILANCIA ADUANERA (81/5) [38/2]
|   SITIO_APREHENSION = YUNGUILLA-ECUADOR : CUERPO DE VIGILANCIA ADUANERA (92/0) [32/0]
|   SITIO_APREHENSION = ARENILLAS-ECUADOR : CUERPO DE VIGILANCIA ADUANERA (68/2) [36/2]
|   SITIO_APREHENSION = DURAN-ECUADOR
|   |   SUBGRUPO = CEBOLLA : CUERPO DE VIGILANCIA ADUANERA (2/1) [0/0]
|   |   SUBGRUPO = TELEVISORES : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = ACC ELECTRONICA : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = PRENDAS DE VESTIR / NUEVAS : CUERPO DE VIGILANCIA ADUANERA (21/5) [4/1]
|   |   SUBGRUPO = MEDIAS DE VESTIR_ NYLON_ PANTYS_ TOBILLERASâ€¦ : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = WHISKY VARIAS MARCAS : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [2/0]
|   |   SUBGRUPO = TEQUILA : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = CALZADO EN GENERAL : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [3/1]
|   |   SUBGRUPO = MERCADERIA SURTIDA : CUERPO DE VIGILANCIA ADUANERA (3/1) [1/1]
|   |   SUBGRUPO = VODKA : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = LICORES VARIOS : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = LECHE : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = CARTERAS_ CINTURONES_ BILLETERAS_ MOCHILAS_ BOLSOS : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = VARIAS MARCAS : CONTROL CONJUNTO INTERINSTITUCIONAL (0/0) [0/0]
|   |   SUBGRUPO = VINO : CONTROL CONJUNTO INTERINSTITUCIONAL (2/0) [0/0]
```

Figura 42 Algoritmo RepTree-PC3 (Parte 1)

```
Time taken to build model: 0.66 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       98398               95.8419 %
Incorrectly Classified Instances      4269                4.1581 %
Kappa statistic                          0.8931
Mean absolute error                      0.0159
Root mean squared error                  0.0915
Relative absolute error                 16.2845 %
Root relative squared error             41.4415 %
Total Number of Instances           102667

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,978     0,070     0,978      0,978    0,978      0,908    0,986     0,993     CUERPO DE VIGILANCIA ADUANERA
                 0,857     0,008     0,923      0,857    0,889      0,878    0,980     0,916     CONTROL CONJUNTO INTERINSTITUCIONAL
                 0,979     0,020     0,877      0,979    0,925      0,915    0,993     0,948     POLICIA NACIONAL
                 0,425     0,000     0,804      0,425    0,556      0,584    0,872     0,445     DESCONOCIDO
                 0,276     0,001     0,807      0,276    0,411      0,469    0,952     0,439     FUERZAS ARMADAS
                 0,000     0,000     ?          0,000    ?          ?        0,496     0,000     INSPECTORIA DE PESCA
                 0,000     0,000     ?          0,000    ?          ?        0,750     0,026     ARCSA
                 0,545     0,000     1,000      0,545    0,706      0,739    0,907     0,566     AGROCALIDAD
Weighted Avg.    0,958     0,057     ?          0,958    ?          ?        0,986     0,973

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
 76796   690   981    10     8     0     0     0 |     a = CUERPO DE VIGILANCIA ADUANERA
  1401  8546    20     0     0     0     0     0 |     b = CONTROL CONJUNTO INTERINSTITUCIONAL
   200    13 12696     1    64     0     0     0 |     c = POLICIA NACIONAL
    56     3     1    45     1     0     0     0 |     d = DESCONOCIDO
    39     1   770     0   309     0     0     0 |     e = FUERZAS ARMADAS
     0     0     1     0     0     0     0     0 |     f = INSPECTORIA DE PESCA
     0     3     1     0     0     0     0     0 |     g = ARCSA
     1     0     3     0     1     0     0     6 |     h = AGROCALIDAD
```

Figura 43 Algoritmo RepTree-PC3 (Parte 2)

En la Tabla 34 se aprecia las métricas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC3 es de 4,58%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.88, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.96, la precisión de las variables analizadas varía de 0,8 a 1, sin embargo la precisión total se muestra cómo "?", porque tres valores son desconocidos, el RECALL o cantidad de interpretación es del 0.95, el valor de verdaderos positivos (TP Rate) indica un valor de 0.95, en cuanto a los Falsos Positivos (FP Rate) de 0.06 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 32

Métricas Estadísticas Algoritmo RepTree-PC3

MEDIDA	VALOR
Tasa de error	4.1581%
Coeficiente KAPPA	0.8931
Curva ROC	0.986
Precisión	?
Recall	0.958
TP Rate	0.958
FP Rate	0.057
F-Measure	?

Fuente Propia

- **Algoritmo RandomTree**

Parte de los resultados de la aplicación del algoritmo RandomTree, con validación cruzada de 10, 5 atributos y 102667 instancias registradas, se muestran en las Fig. 44 y Fig. 45.

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:     VistaMinableFINAL-weka.filters.unsupervised.attribute.Remove-R1-10,13-17,21-22
Instances:    102667
Attributes:   5
              GRUPO
              SUBGRUPO
              ORIGEN_APREHENSION
              SITIO_APREHENSION
              GRUPO_OPERATIVO
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


RandomTree
==========

ORIGEN_APREHENSION = ACCIONES INTELIGENCIA Y PROTECCION - DIP
|   SITIO_APREHENSION = RIO SIETE-ECUADOR
|   |   SUBGRUPO = CEBOLLA : CUERPO DE VIGILANCIA ADUANERA (4/1)
|   |   SUBGRUPO = TELEVISORES : CUERPO DE VIGILANCIA ADUANERA (5/0)
|   |   SUBGRUPO = ACC ELECTRONICA : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = PRENDAS DE VESTIR / NUEVAS : CUERPO DE VIGILANCIA ADUANERA (47/2)
|   |   SUBGRUPO = MEDIAS DE VESTIR_ NYLON_ PANTYS_ TOBILLERASâ€¦ : CUERPO DE VIGILANCIA ADUANERA (6/1)
|   |   SUBGRUPO = WHISKY VARIAS MARCAS : CUERPO DE VIGILANCIA ADUANERA (1/0)
|   |   SUBGRUPO = TEQUILA : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = CALZADO EN GENERAL : CONTROL CONJUNTO INTERINSTITUCIONAL (3/1)
|   |   SUBGRUPO = MERCADERIA SURTIDA : CUERPO DE VIGILANCIA ADUANERA (2/1)
|   |   SUBGRUPO = VODKA : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = LICORES VARIOS : CUERPO DE VIGILANCIA ADUANERA (1/0)
|   |   SUBGRUPO = LECHE : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = CARTERAS_ CINTURONES_ BILLETERAS_ MOCHILAS_ BOLSOS : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = VARIAS MARCAS : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = VINO : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = RON : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = JUEGOS PIROTECNICOS : CUERPO DE VIGILANCIA ADUANERA (0/0)
|   |   SUBGRUPO = JUGUETES : CONTROL CONJUNTO INTERINSTITUCIONAL (1/0)
|   |   SUBGRUPO = PEPARACIONES ALIMENTICIAS DIVERSAS / ENLATADOS : CUERPO DE VIGILANCIA ADUANERA (0/0)
```

Figura 44 Algoritmo RandomTree-PC3 (Parte 1)

```
Time taken to build model: 0.22 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       98175               95.6247 %
Incorrectly Classified Instances      4492                4.3753 %
Kappa statistic                          0.8875
Mean absolute error                      0.0149
Root mean squared error                  0.0931
Relative absolute error                 15.2614 %
Root relative squared error             42.1787 %
Total Number of Instances           102667

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,977    0,076    0,977      0,977    0,977      0,902    0,979     0,988     CUERPO DE VIGILANCIA ADUANERA
                 0,848    0,009    0,912      0,848    0,879      0,867    0,965     0,892     CONTROL CONJUNTO INTERINSTITUCIONAL
                 0,973    0,019    0,880      0,973    0,924      0,914    0,993     0,951     POLICIA NACIONAL
                 0,425    0,000    0,726      0,425    0,536      0,555    0,866     0,421     DESCONOCIDO
                 0,322    0,001    0,752      0,322    0,451      0,488    0,932     0,466     FUERZAS ARMADAS
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,000     INSPECTORIA DE PESCA
                 0,500    0,000    1,000      0,500    0,667      0,707    0,750     0,500     ARCSA
                 0,545    0,000    1,000      0,545    0,706      0,739    0,863     0,578     AGROCALIDAD
Weighted Avg.    0,956    0,061    ?          0,956    ?          ?        0,979     0,968

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
 76690   794   973    14    14     0     0     0 |     a = CUERPO DE VIGILANCIA ADUANERA
  1488  8454    24     0     1     0     0     0 |     b = CONTROL CONJUNTO INTERINSTITUCIONAL
   241    12 12618     1   102     0     0     0 |     c = POLICIA NACIONAL
    56     3     1    45     1     0     0     0 |     d = DESCONOCIDO
    45     1   711     2   360     0     0     0 |     e = FUERZAS ARMADAS
     0     0     1     0     0     0     0     0 |     f = INSPECTORIA DE PESCA
     0     1     1     0     0     0     2     0 |     g = ARCSA
     1     0     3     0     1     0     0     6 |     h = AGROCALIDAD
```

Figura 45 Algoritmo RandomTree-PC3 (Parte 2)

En la Tabla 35 se aprecia las métricas de calidad derivadas de la matriz de confusión, en donde muestra que la tasa de error de este algoritmo con el PC3 es de 4,58%, el coeficiente KAPPA o nivel de concordancia muestra un valor de 0.88, la curva ROC o sensibilidad de los falsos positivos presenta un valor de 0.96, la precisión de las variables analizadas varía de 0,8 a 1, sin embargo la precisión total se muestra cómo "?", porque tres valores son desconocidos, el RECALL o cantidad de interpretación es del 0.95, el valor de verdaderos positivos (TP Rate) indica un valor de 0.95, en cuanto a los Falsos Positivos (FP Rate) de 0.06 y la combinación de la precisión y Recall o F-Measure es desconocida.

Tabla 33

Métricas Estadísticas Algoritmo RandomTree-PC3

MEDIDA	VALOR
Tasa de error	4.3753%
Coeficiente KAPPA	0.8875
Curva ROC	0.979
Precisión	?
Recall	0.956
TP Rate	0.956
FP Rate	0.061
F-Measure	?

Fuente Propia

3.1.2. Evaluación de Algoritmos de Regresión

- **Logistic Regression**

En este caso se aplicó el algoritmo a variables tipo texto más relevantes, de manera que sea fácil de interpretar:

Grupo Operativo

En la Fig. 46, se presenta el proceso de lectura de los datos y de predicción con respecto a la variable GRUPO_OPERATIVO.

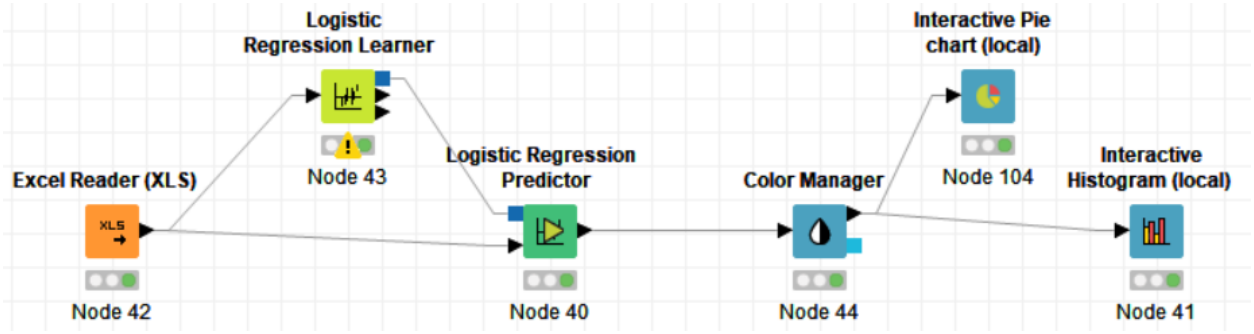


Figura 46 Proceso Algoritmo Logistic Regression - Grupo_Operativo

Distrito

En la Fig. 47, se presenta el proceso de lectura de los datos y de predicción con respecto a la variable Distrito.

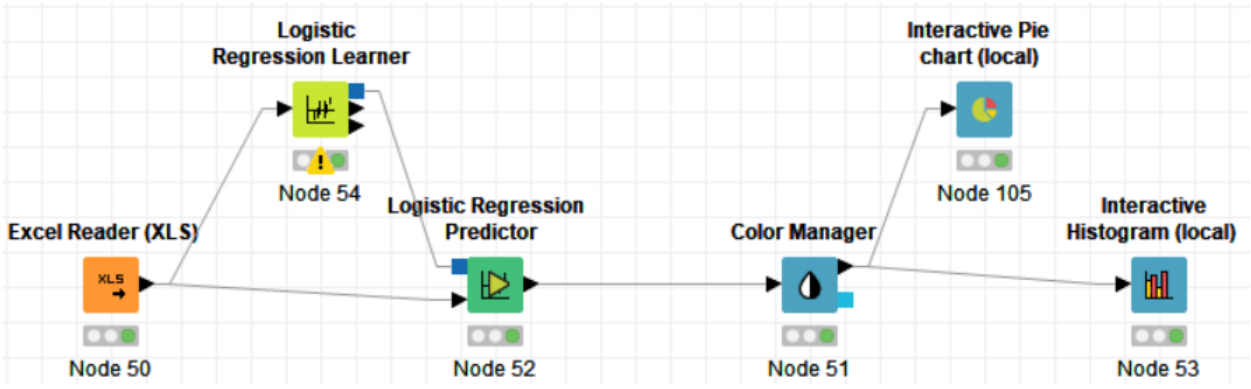


Figura 47 Proceso Algoritmo Logistic Regression – Distrito

Bodega

En la Fig. 48, se presenta el proceso de lectura de los datos y de predicción con respecto a la variable Bodega.

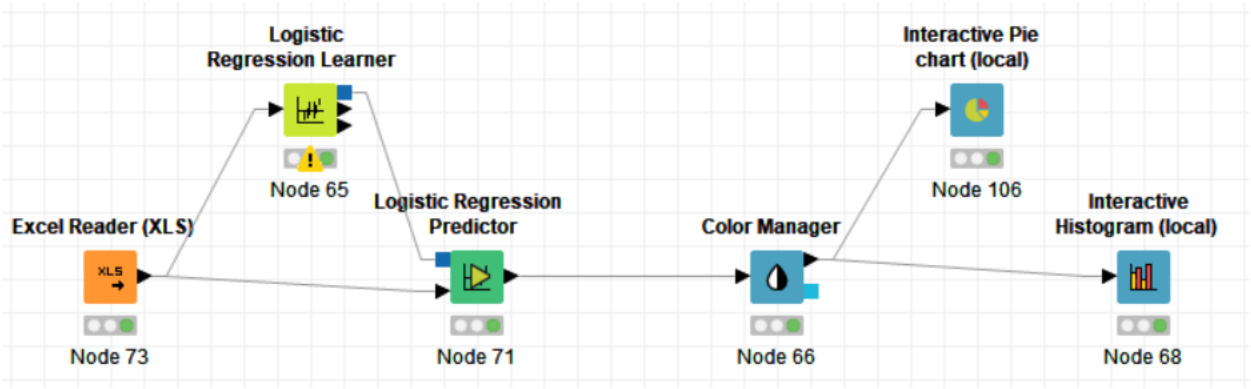


Figura 48 Proceso Algoritmo Logistic Regression – Bodega

Grupo

En la Fig. 49, se presenta el proceso de lectura de los datos y de predicción con respecto a la variable Grupo.

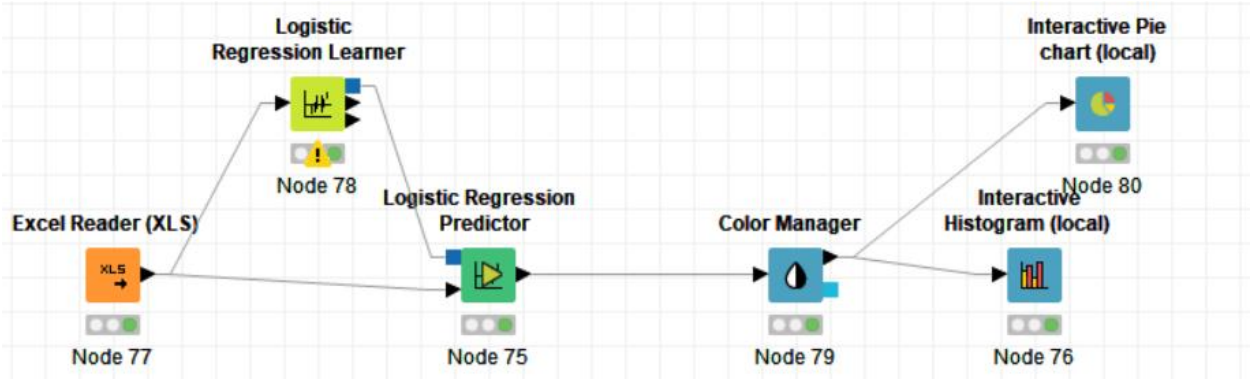


Figura 49 Proceso Algoritmo Logistic Regression – Grupo

3.2. Análisis e interpretación de resultados

3.2.1. Análisis e interpretación de resultados de algoritmos de clasificación

- **Análisis de Resultados**

Se procedió a evaluar cuantitativamente las métricas estadísticas presentadas en los resultados de los algoritmos, tales como tasa de error, coeficiente Kappa, curva ROC, precisión, entre otros; además, la facilidad de interpretación del algoritmo. En las tablas presentadas anteriormente los resultados son valores similares. Considerando esto y la facilidad de interpretación del algoritmo se toma en cuenta que el algoritmo J48 es el mejor clasificador para este caso. De tal manera, se presentarán los resultados obtenidos por este.

- **Interpretación de Resultados**

Una vez efectuado el análisis de resultados y determinado el mejor modelo de árbol de clasificación y reglas de decisión con el algoritmo J48. La interpretación se realizó por cada uno de los 3 principales componentes obtenidos por el algoritmo PCA, iniciando desde el nodo raíz, cuyo atributo es GRUPO_OPERATIVO con sus respectivos valores hasta llegar al fondo del árbol, presentando el resultado a continuación:

PC1

En este caso, específicamente se aplicó una simple regla de 3, para determinar que patrones son más importantes con base en el número de instancias analizadas y relevantes para el algoritmo. Por lo tanto, los patrones más relevantes son:

GRUPO_OPERATIVO = CONTROL CONJUNTO INTERINSTITUCIONAL

- Contrabando de Artículos electrónicos, localizados en Antonio Ante y almacenados en el Destacamento Yahuarcocha.
- Contrabando de Productos Medicinales Naturales, localizados en Ibarra y almacenados en la Dirección Distrital Tulcán.
- Contrabando de Relojes, localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.
- Contrabando de Prendas de Vestir / Nuevas, localizados en Otavalo y almacenados en el Destacamento Yahuarcocha.
- Contrabando de Prendas de Vestir / Nuevas, localizados en Yaguachi y almacenados en la Dirección Distrital Guayas.

GRUPO_OPERATIVO = CUERPO DE VIGILANCIA ADUANERA

- Contrabando de Prendas de Vestir / Nuevas, localizados en Puerto Bolívar y almacenados en el Destacamento Chacras.
- Contrabando de Artículos de Belleza_ Accesorios, localizados en Esmeraldas y almacenados en la Dirección Distrital Esmeraldas.
- Contrabando de ACC Celulares, localizados en Lago Agrio y almacenados en el Destacamento Amazonas.
- Contrabando de Cebolla, localizados en el Cantón Santa Rosa y almacenados en la Dirección Distrital Puerto Bolívar.
- Contrabando de Cigarrillos/Puros, localizados en Rumichaca y almacenados en la Dirección Distrital Tulcán.
- Contrabando de computadoras, laptops y tablets, localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.
- Contrabando de granadilla, maracuyá y pitajaya, localizados en Carchi y almacenados en la Dirección Distrital Tulcán.

- Contrabando de Prendas de Vestir / Nuevas, localizados en Lago Agrio y almacenados en la Dirección Distrital Quito.
- Contrabando de Cigarrillos, localizados en Rumichaca y almacenados en la Dirección Distrital Tulcán.
- Contrabando de Línea Blanca (refri, lavadora, cocina), localizados en Rumichaca y almacenados en la Dirección Distrital Tulcán.
- Contrabando de Cerámica, localizados en Rumichaca y almacenados en la Dirección Distrital Tulcán.
- Contrabando de Televisores, localizados en Puerto Bolívar y almacenados en la Jefatura del VII Distrito Carchi.
- Contrabando de Artículos de Belleza_ Accesorios, localizados en Rumichaca y almacenados en la Dirección Distrital Tulcán.
- Contrabando de Material Eléctrico, localizado en Carchi y almacenados en la Dirección Distrital Tulcán.
- Contrabando de Televisores, localizados en Catamayo y almacenados en el Destacamento Catamayo.

GRUPO_OPERATIVO = FUERZAS ARMADAS

- Contrabando de ACC Celulares, localizados en Lago Agrio y almacenados en la Dirección Distrital Quito.

GRUPO_OPERATIVO = POLICÍA NACIONAL

- Contrabando de ACC Celulares, localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.
- Contrabando de Medias de vestir nylon, pantys y tobilleras, localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.
- Contrabando de Calzado en General, localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.
- Contrabando de celulares, teléfonos y faxes, localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.
- Contrabando de Perfumes localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.

- Contrabando de Cebolla, localizados en el cantón Santa Rosa y almacenados en el Destacamento Chacras.
- Contrabando de uvas, ciruelos, kiwis y claudias, localizados en el cantón Santa Rosa y almacenados en la Dirección Distrital Puerto Bolívar.
- Contrabando de ACC Electrónica, localizados en CARCHI y almacenados en la Dirección Distrital Tulcán.
- Contrabando Arroz, localizados en Catamayo y almacenados en el Destacamento Catamayo.
- Contrabando Mercadería Surtida, localizados en Catamayo y almacenados en el Destacamento Catamayo.
- Contrabando de Cigarrillos, localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.
- Contrabando de Artículos de Limpieza, localizados en Ibarra y almacenados en el Destacamento Yahuarcocha.

PC2

GRUPO_OPERATIVO = CONTROL CONJUNTO INTERINSTITUCIONAL

- Mediante Llamada Telefónica se procede a realizar una aprehensión en el Cuarto Distrito Quito (IVDQ), catalogando el total de la aprehensión con una valoración en dólares Baja.
- Mediante Control Fijo se procede a realizar una aprehensión en el Noveno Distrito Huaquillas (IXDH), catalogando el total de la aprehensión con una valoración en dólares Alta.
- Mediante Control Fijo se procede a realizar una aprehensión en el Sexto Distrito Loja (VIDL), catalogando el total de la aprehensión con una valoración en dólares Media.
- Mediante Control Fijo se procede a realizar una aprehensión en el Sexto Distrito Loja (VIDL), catalogando el total de la aprehensión con una valoración en dólares Alta.
- Mediante Control Móvil se procede a realizar una aprehensión en el Séptimo Distrito Puerto Bolívar (VIIDPB), catalogando el total de la aprehensión con una valoración en dólares Alta.

GRUPO_OPERATIVO = CUERPO DE VIGILANCIA ADUANERA

- Mediante Llamada Telefónica se procede a realizar una aprehensión en el Cuarto Distrito Quito (IVDQ), catalogando el total de la aprehensión con una valoración en dólares Media.
- Mediante Llamada Telefónica se procede a realizar una aprehensión en el Cuarto Distrito Quito (IVDQ), catalogando el total de la aprehensión con una valoración en dólares Alta.
- Mediante Control Fijo se procede a realizar una aprehensión en el Noveno Distrito Huaquillas (IXDH), catalogando el total de la aprehensión con una valoración en dólares Media.
- Mediante Control Fijo se procede a realizar una aprehensión en el Noveno Distrito Huaquillas (IXDH), catalogando el total de la aprehensión con una valoración en dólares Baja.
- Mediante Control Fijo se procede a realizar una aprehensión en el Sexto Distrito Loja (VIDL), catalogando el total de la aprehensión con una valoración en dólares Baja.
- Mediante Control Móvil se procede a realizar una aprehensión en el Séptimo Distrito Puerto Bolívar (VIIDPB), catalogando el total de la aprehensión con una valoración en dólares Media.

PC3

GRUPO_OPERATIVO = CONTROL CONJUNTO INTERINSTITUCIONAL

- Mediante Llamada Telefónica se procede a incautar Dulces Varios en Ibarra.
- Mediante Llamada Telefónica se procede a incautar Prendas de Vestir / Usadas en Ibarra.
- Mediante Control Fijo realizado en San Agustín se procede a incautar Frutas y Comestibles.
- Mediante Control Fijo realizado en Urbina se procede a incautar aparatos electrónicos y sus accesorios.
- Mediante Control Fijo realizado en el Cantón Santa Rosa se procede a incautar Frutas y Comestibles.

GRUPO_OPERATIVO = CUERPO DE VIGILANCIA ADUANERA

- Mediante Control Fijo realizado en San Agustín se procede a incautar aparatos electrónicos y sus accesorios.
- Mediante Control Fijo realizado en Urbina se procede a incautar Textiles.

GRUPO_OPERATIVO = FUERZAS ARMADAS

- Mediante Acta de Entrega / Recepción en la ciudad de Carchi se procede a incautar Prendas de Vestir / Nuevas.

GRUPO_OPERATIVO = POLICÍA NACIONAL

- Mediante Acta de Entrega / Recepción en la ciudad de Carchi se procede a incautar Mercadería Surtida.
- Mediante Acta de Entrega / Recepción en la ciudad de Carchi se procede a incautar celulares, teléfonos, faxes e iPhone.
- Mediante Acta de Entrega / Recepción en la ciudad de Lago Agrio se procede a incautar Prendas de Vestir / Nuevas.
- Mediante Acta de Entrega / Recepción en la ciudad de Lago Agrio se procede a incautar Mercadería Surtida.

3.2.2. Análisis e interpretación de resultados de algoritmos de regresión

- **Análisis de Resultados**

Se procede a evaluar las variables cuantitativas más relevantes para el estudio y con menor número de registros únicos, tales como *GRUPO_OPERATIVO, ZONA, DISTRITO, BODEGA Y GRUPO*. Mediante la función *Interactive Histogram* de la herramienta *KNIME* se procederá a mostrar los resultados interactivos que presentan estas variables entre sí.

- **Interpretación de Resultados**

Considerando que la regresión logística en la herramienta KNIME presenta resultados de fácil interpretación con variables cuantitativas y relacionando la variable a analizar con las demás variables de la Base de datos, se presentan los siguientes resultados:

GRUPO OPERATIVO

En la Fig. 50 se muestra que los grupos operativos con mayor frecuencia serán CONTROL CONJUNTO INTERINSTITUCIONAL, CUERPO DE VIGILANCIA ADUANERA y POLICÍA NACIONAL.

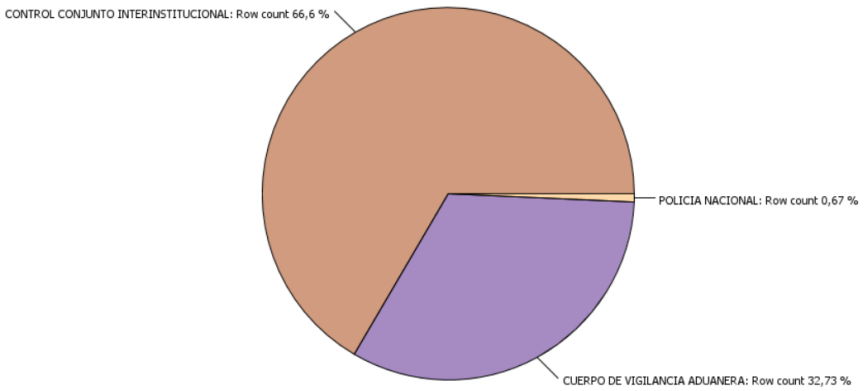


Figura 50 Resultado Algoritmo Logistic Regression - Grupo_Operativo

- **GRUPO**

El grupo operativo Control Conjunto Interinstitucional en relación con la variable grupo se muestra con mayor frecuencia en la categoría Textiles, Licores, Electrónica y sus Accesorios, Frutas y Comestibles y Calzado, tal como se muestra en la Fig. 51.

El grupo operativo Cuerpo de Vigilancia Aduanera en relación con la variable grupo se muestra con mayor frecuencia en la categoría Textiles, Electrónica y sus Accesorios, Artículos de Bazar y Cigarrillos, tal como se muestra en la Fig. 51.

El grupo operativo Policía Nacional en relación con la variable grupo se muestra con mayor frecuencia en la categoría Electrónica y sus Accesorios, tal como se muestra en la Fig. 51.

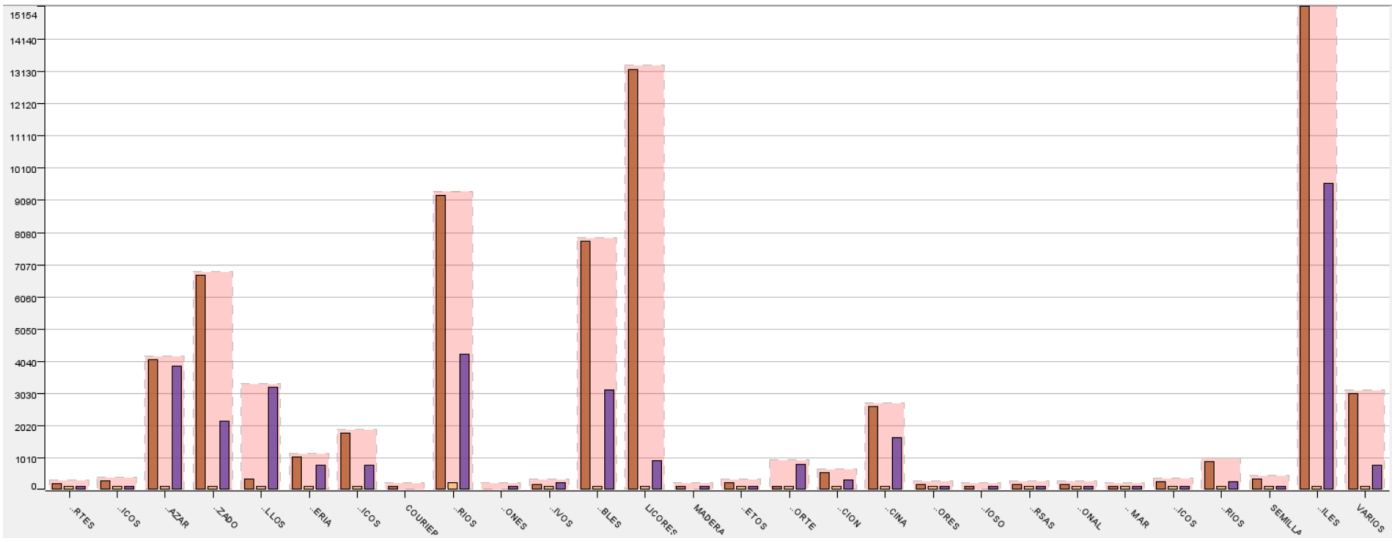


Figura 51 Relación Grupo Operativo - Grupo

- **BODEGA**

La variable GRUPO_OPERATIVO en relación con la variable Bodega, indica que la predicción para las bodegas con mayor número de almacenamiento son Destacamento Chacras, Bodega Huaquillas, Dirección Distrital Tulcán y Dirección Distrital Puerto Bolivar. Siendo el Destacamento Chacras el que presente mayor número de retenciones en todos los grupos operativos, como se presenta en la Fig. 52.

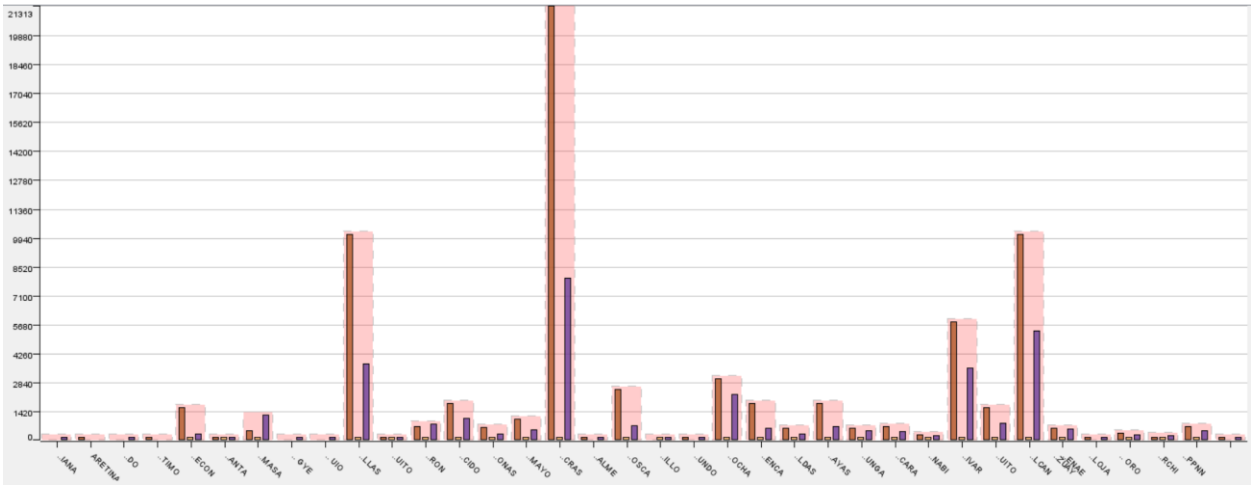


Figura 52 Relación Grupo Operativo – Bodega

- **ORIGEN DE APREHENSIÓN**

En la Fig. 53 se muestra las predicciones de Grupo Operativo con Origen de Aprehensión, de tal manera que los resultados muestran que se realizará mayor número de aprehensiones mediante Control de Rutina y Acta de Entrega – Recepción.

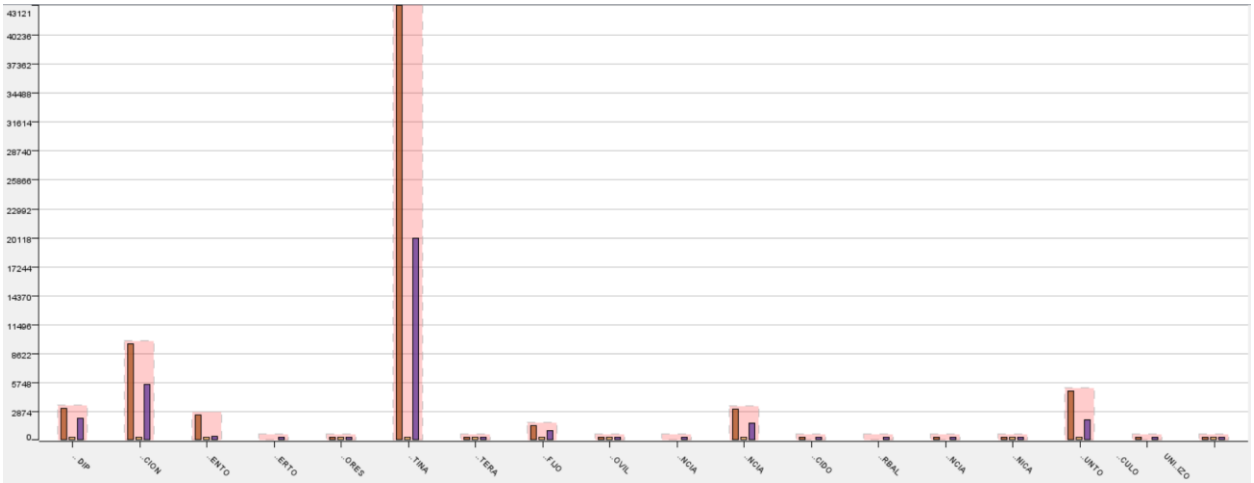


Figura 53 Relación Grupo Operativo - Origen de Aprehensión

- **DISTRITO**

Los grupos operativos que se encuentran en el noveno distrito Huaquillas (IXDH), el octavo Distrito Carchi (VIIIDC) y el cuarto Distrito Quito (IVDQ) son los que tendrán mayor trabajo por realizar; sin embargo, el grupo operativo Control Conjunto Interinstitucional es el que se encontrará mucho más presente en los operativos, tal como muestra la Fig.54.

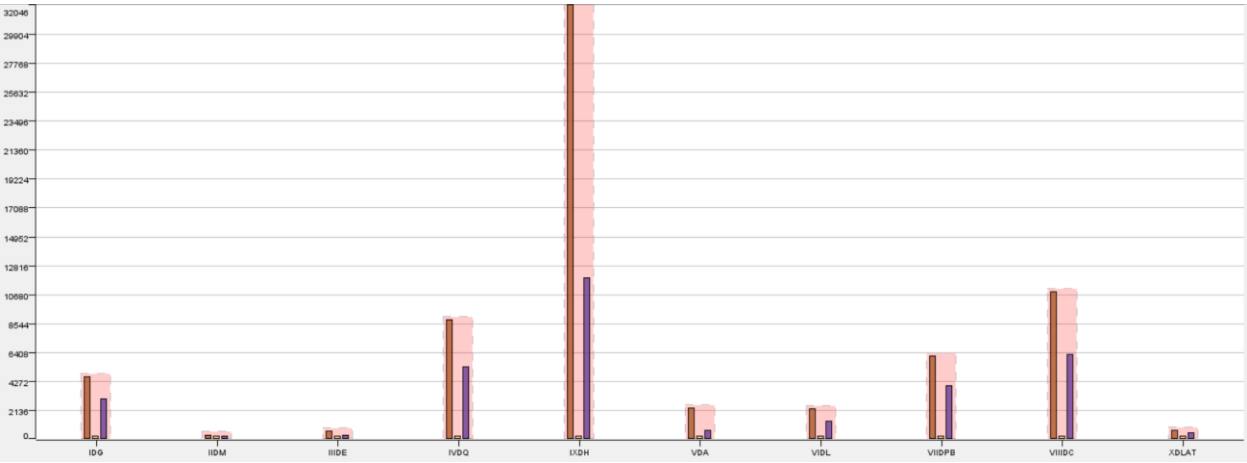


Figura 54 Relación Grupo Operativo - Distrito

- **ZONA**

Los grupos operativos asignados a la Zona 2 del Ecuador, tendrán una mayor labor que los grupos operativos asignados a otras zonas, tal como indica la Fig. 55.

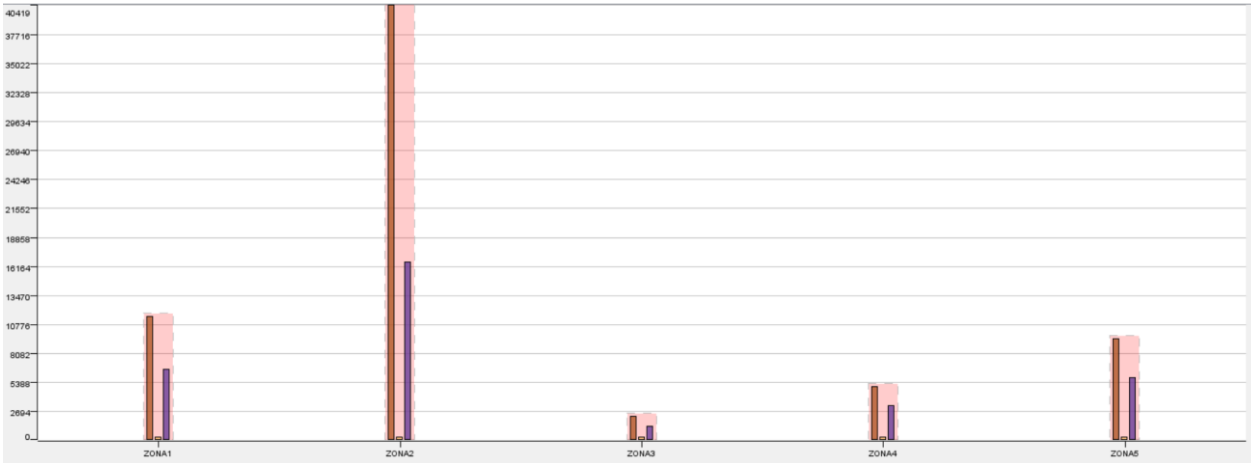


Figura 55 Relación Grupo Operativo - Zona

DISTRITO

Los Distritos que presentarán más aprehensiones son el Octavo Distrito Carchi (VIIIDC), Noveno Distrito Huaquillas (IXDH), Séptimo Distrito Puerto Bolívar (VIIDPB), Cuarto Distrito Quito (IVDQ) y el Décimo Distrito Latacunga (XDLAT), tal como se indica en la Fig. 56.

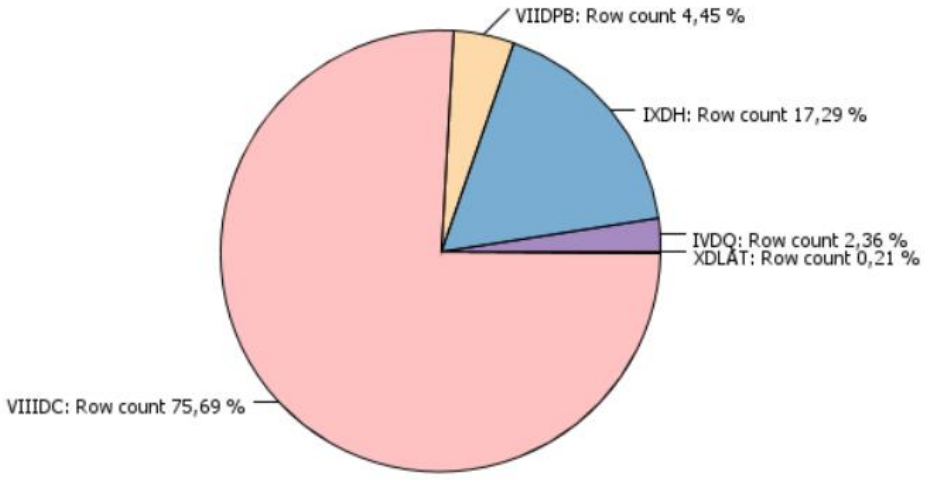


Figura 56 Resultado Algoritmo Logistic Regression – Distrito

- **GRUPO**

Los resultados del algoritmo con relación entre las variables grupo y distrito, indican que los distritos tendrán mayor número de aprehensiones con la categoría Textiles, Licores, Electrónica y sus Accesorios, Frutas y Comestibles, Artículos de Bazar y Calzado, como se presentan en la Fig. 57.

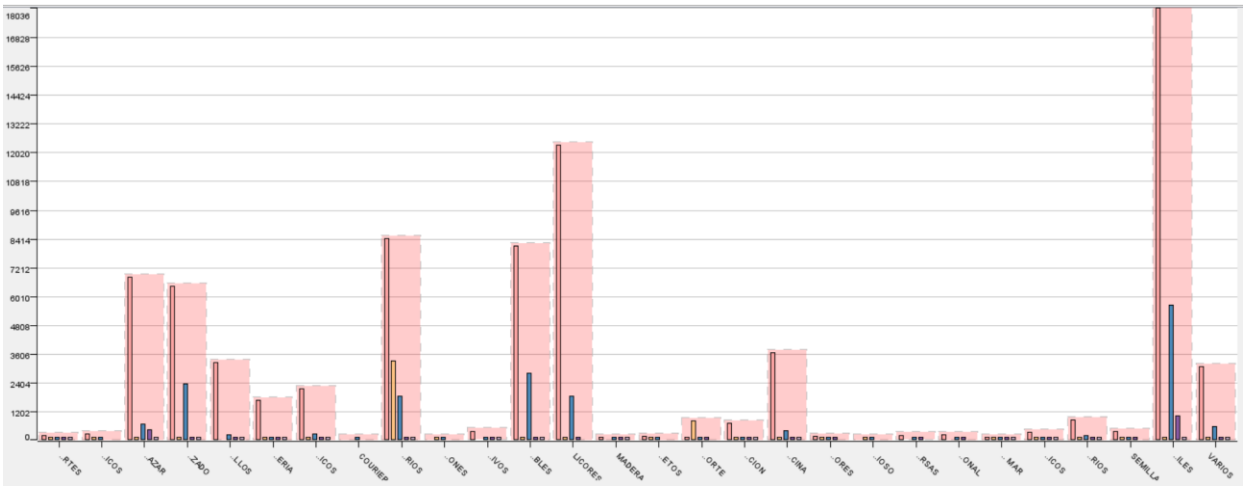


Figura 57 Relación Distrito - Grupo

- **BODEGA**

La variable Distrito en relación con la variable Bodega, indica que la predicción para las bodegas que estarán presentes con mayor frecuencia de aprehensiones son Destacamento Chacras, Dirección Distrital Tulcán, Bodega Huaquillas y Dirección Distrital Puerto Bolívar. Siendo el Destacamento Chacras el que presente mayor número de retenciones en todos los grupos operativos, como se presenta en la Fig. 58.

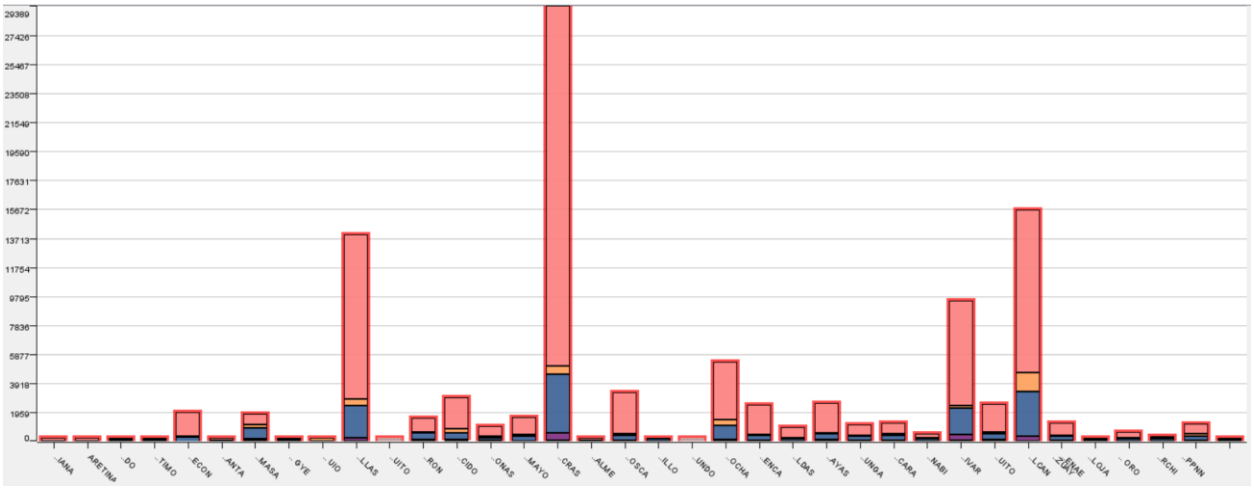


Figura 58 Relación Distrito - Bodega

- **ORIGEN DE APREHENSIÓN**

En la Fig. 59 se muestra las predicciones de Distrito con Origen de Aprehensión, de tal manera que los restultados muestran que se realizará un mayor número de aprehensiones mediante Control de Rutina y Acta de Entrega – Recepción.

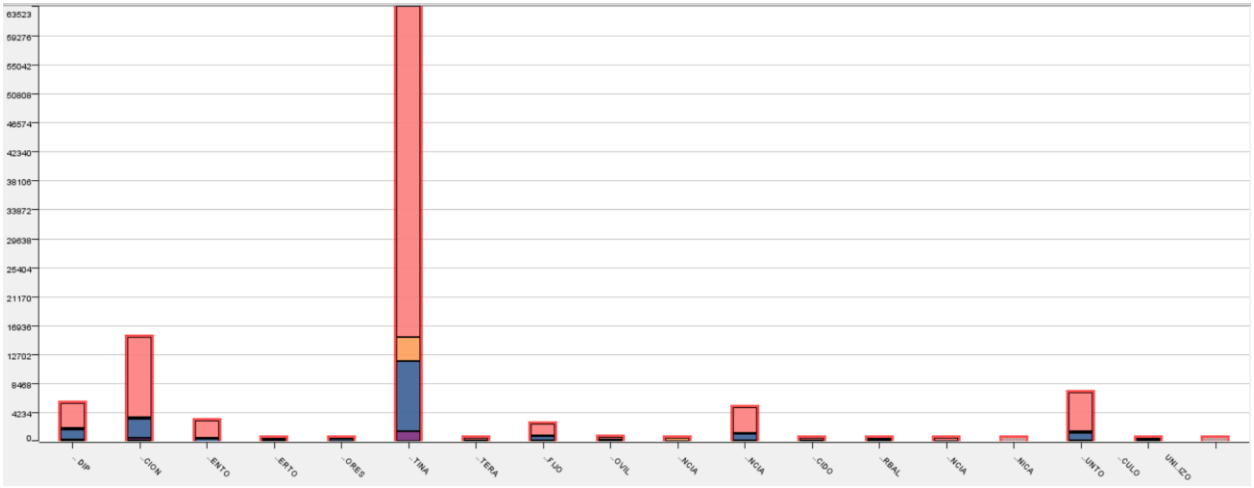


Figura 59 Relación Distrito - Origen de Aprehensión

BODEGA

Los Destacamentos Chacras, Huaquillas, Catamayo y Dirección Distrital Tulcán, son las bodegas que poseen mayor probabilidad de aparición en los informes, de tal manera que se muestra en la Fig. 60.

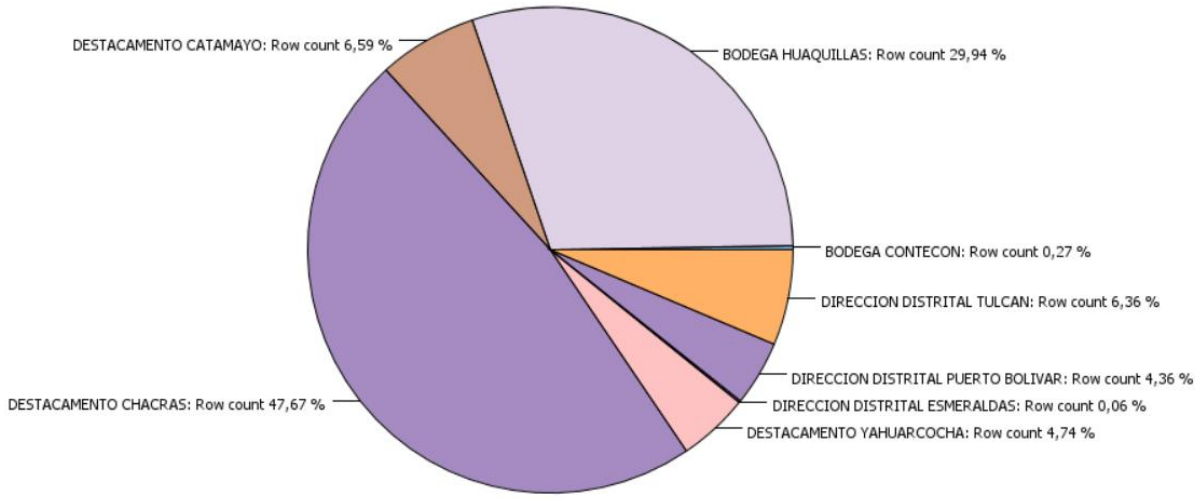


Figura 60 Resultado Algoritmo Logistic Regression - Bodega

- **GRUPO**

Los resultados del algoritmo entre las variables Bodega y Grupo, indican que las bodegas expuestas en la Fig. 61 tendrán mayor probabilidad de capturar mercadería perteneciente al grupo de textiles, licores, electrónica y sus accesorios, frutas y comestibles, calzado y artículos de bazar, perteneciendo la mayoría al destacamento Chacras.

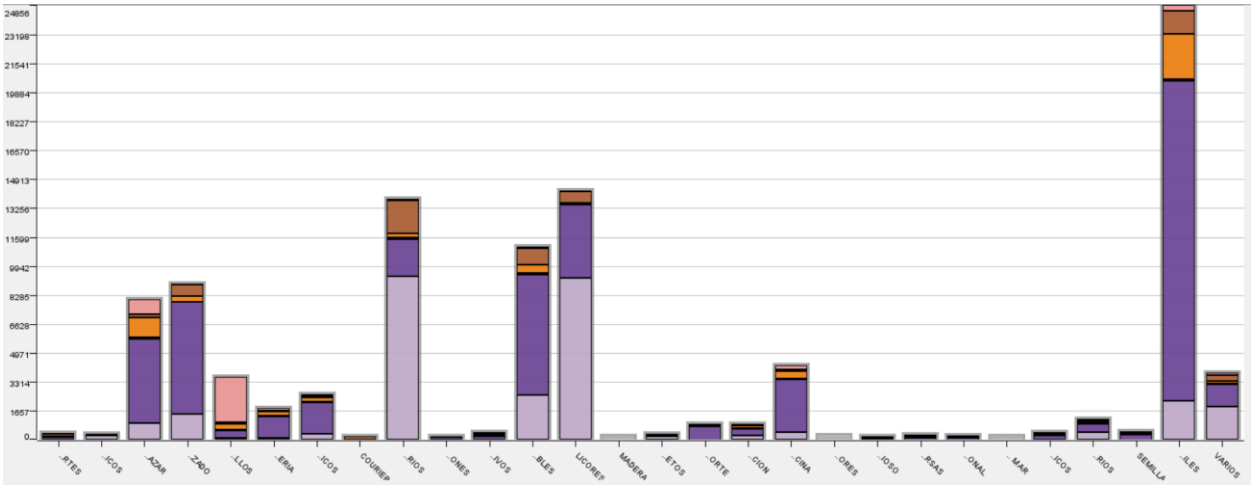


Figura 61 Relación Bodega – Grupo

- **ORIGEN DE APREHENSIÓN**

En la Fig. 62 se muestra las predicciones de Bodega con relación en Origen de Aprehensión, de tal manera que los resultados muestran que se realizará mayor número de aprehensiones mediante Control de Rutina, Acta de Entrega – Recepción, Operativo Conjunto, DIP y mediante denuncias.

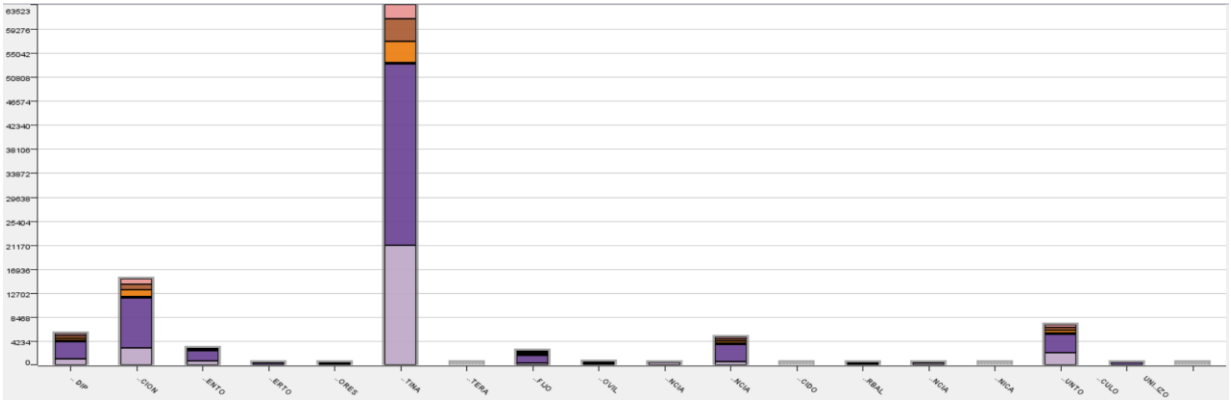


Figura 62 Relación Bodega- Origen de Aprehensión

GRUPO

En la Fig. 63 se indica que los contrabandos con mayor posibilidad de concurrencia pertenecen a los grupos Textiles, Licores, Electrónica y sus Accesorios y Cosméticos.

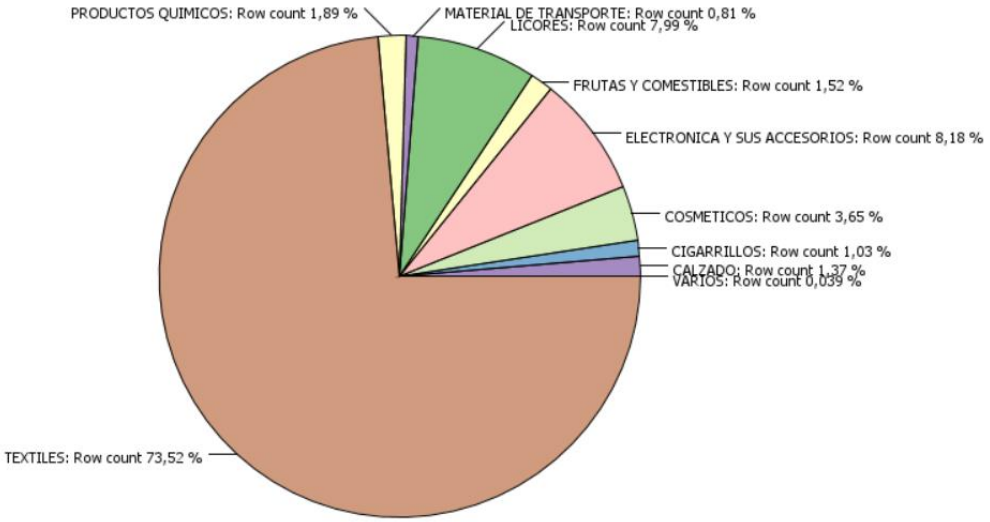


Figura 63 Resultado Algoritmo Logistic Regression – Grupo

- **ORIGEN DE APREHENSIÓN**

En la Fig. 64 se muestra las predicciones de Grupo con relación en Origen de Aprehensión, de tal manera que los resultados muestran que se realizará mayor número de aprehensiones mediante Control de Rutina, Acta de Entrega – Recepción, Operativo Conjunto, DIP y mediante denuncias.

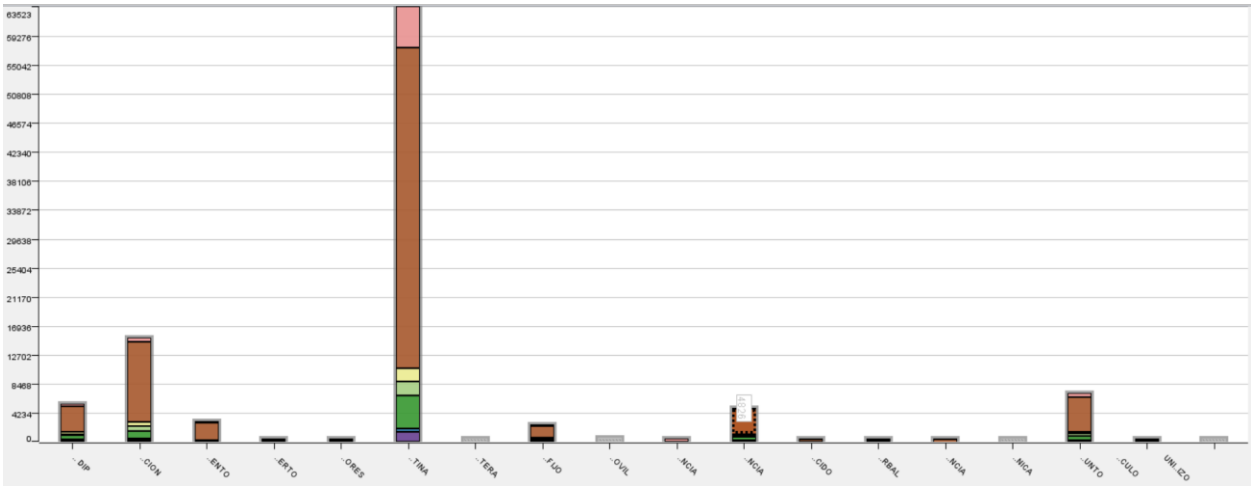


Figura 64 Relación Grupo- Origen de Aprehensión

3.3. Fase de Obtención del Conocimiento

Para la fase de obtención de conocimiento se realizó un análisis entre los resultados de algoritmos de clasificación “J48” y algoritmo de regresión “Logistic Regression”, verificando que los resultados tengan coincidencia entre ellos, obteniendo como resultado que los principales contrabandos tendrán las siguientes características:

- Contrabandos que realicen los grupos operativos Control Conjunto Interinstitucional y Grupo de Vigilancia Aduanera.
- Contrabandos que transcurran por los distritos Quito (IVDQ), Huaquillas (IXDH), Puerto Bolívar (VIIDPB) y Latacunga (VIDL).
- Contrabandos cuyo producto pertenezca a los grupos de Textiles, Licores, Electrónica y sus Accesorios y Frutas y Comestibles.
- Contrabando de prendas de vestir nuevas, cigarrillos, cerámica, material eléctrico, televisores y frutas.
- Contrabandos que se retengan en las bodegas del Destacamento Catamayo, Dirección Distrital Tulcán y Destacamento Chacras.
- Contrabandos que tengan como origen de aprehensión llamadas telefónicas, control de rutina, acta de entrega – recepción y control fijo.

3.4. Resumen Ejecutivo del Conocimiento Obtenido

Los resultados obtenidos, aplicando técnicas predictivas de clasificación, utilizando el algoritmo J48, indican que, mediante llamada telefónica, control fijo y control móvil, el grupo operativo:

- Control Conjunto Interinstitucional, procede a aprehender y retener en las ciudades de; Ibarra, productos medicinales naturales y relojes; en Antonio Ante, artículos electrónicos; en Otavalo y Yaguachi prendas de vestir nuevas.
- Cuerpo de vigilancia aduanera, procede a aprehender en las ciudades de; Puerto Bolívar Y Lago Agrio, prendas de vestir nuevas; en Rumichaca e Ibarra, cigarrillos y cerámica; en Carchi, frutas como granadillas, maracuyá y pitajaya; en Santa Rosa, cebolla; en Puerto Bolívar y Catamayo, televisores.
- Policía Nacional procede en la ciudad de Santa Rosa a incautar Cebollas y uvas, ciruelos, kiwis y claudias.

Los patrones obtenidos, mediante el uso del algoritmo Logistic Regression perteneciente a técnicas predictivas de regresión de minería de datos, determinan que los grupos Control conjunto interinstitucional, Cuerpo de vigilancia aduanera y Policía nacional tendrán mayor número de aprehensiones y retenciones en el octavo distrito Carchi (VIIIDC), Noveno Distrito Huaquillas (IXDH), Séptimo Distrito Puerto Bolívar (VIIDPB), Cuarto Distrito Quito (IVDQ) y el Décimo Distrito Latacunga (XDLAT), proceden a incautar por lo general textiles, licores, electrónica y sus Accesorios y Cosméticos.

Los resultados primordiales se consiguieron verificando coincidencias entre los patrones resultantes por ambos algoritmos, obteniendo aprehensiones y retenciones que realicen específicamente los grupos operativos Control Conjunto Interinstitucional y Grupo de Vigilancia Aduanera, los mismos que transcurran por los Distritos de Quito, Huaquillas, Puerto Bolívar y Latacunga, correspondientes a Textiles como prendas de vestir nuevas incautadas en Puerto Bolívar, Otavalo y Yaguachi, Licores incautados generalmente en Rumichaca, artículos electrónicos como televisores y material eléctrico aprehendidos y retenidos en Antonio Ante y Frutas y Verduras, adicionalmente cigarrillos y cerámicas en Rumichaca.

3.5. Análisis de Impacto

El análisis de impacto define las posibles consecuencias que se presentaran al momento de tomar decisiones estratégicas con base en patrones de aprehensión obtenidos; por tal motivo es necesario analizar el efecto de las decisiones tomadas, cualificando y cuantificando las ventajas y desventajas de acuerdo con ciertos indicadores (Vila, 2019).

Se utilizara un análisis de impacto prospectivo que permitirá identificar los aspectos positivos y negativos que presentara en un grupo o área específica, la aplicación de los patrones obtenidos(Cisneros, 2019), tal como se muestra en la Tabla 36.

Tabla 34

Niveles de Impacto

Niveles de Impacto	Ponderación
Impacto Alto Positivo	3
Impacto Medio Positivo	2
Impacto Bajo Positivo	1
Punto de Indiferencia	0
Impacto Bajo Negativo	-1
Impacto Medio Positivo	-2
Impacto Alto Positivo	-3

Fuente: Vila, 2019

Para este análisis se tomará en cuenta el impacto que tendrá el presente trabajo en el ámbito social y económico, que se aprecian en las Tablas 37, 38 y en la Tabla 39 se detalla el impacto general del proyecto.

3.5.1. Impacto Social

Tabla 35

Impacto Social

Indicador	Niveles						
	-3	**-2**	**-1**	**0**	**1**	**2**	**3**
Calidad de vida							X
Competencia Legal							X
Ambiente de Trabajo							X
Tasas de Empleo						X	
Total						2	9

$$Nivel\ de\ Impacto = \frac{\sum}{\text{Número de Indicadores}}$$

$$Nivel\ de\ Impacto = \frac{11}{4} = 2.75$$

Nivel de Impacto Social = Alto Positivo

Fuente Propia

El impacto social de este proyecto se considera alto positivo, por motivo de que la calidad de vida tanto de los comerciantes como de los agentes de la ley tendrá un impacto alto positivo, puesto que tendrán una mayor eficacia en su labor.

En cuanto el ambiente de trabajo tiene un impacto alto positivo debido a que el ambiente influye tanto en la calidad como en la cantidad de trabajo que una persona es capaz de realizar, mientras que la tasa de empleo tiene un impacto medio positivo, debido a que a menos contrabando más negocios surgen.

3.5.2. Impacto Económico

Tabla 36

Impacto Económico

Indicador	Niveles						
	-3	**-2**	**-1**	**0**	**1**	**2**	**3**
Productividad							X
Estabilidad Económica Estatal							X
Fuga de Capitales						X	
Total						2	6

$$Nivel\ de\ Impacto = \frac{\sum}{\text{Número de Indicadores}}$$

$$Nivel\ de\ Impacto = \frac{8}{3} = 2.6$$

Nivel de Impacto Económico = Alto Positivo

Fuente Propia

El impacto económico afectaría en gran medida a la productividad con un impacto alto positivo, de tal manera que los negocios no fraudulentos generen un mayor número de ingresos, de igual manera afectara a la estabilidad económica estatal.

En relación con la fuga de capitales el impacto generado es medio positivo, debido a que nos ayudaría a evitar perdida de dinero en el país.

3.5.3. Impacto General

Tabla 37

Impacto General

Indicador	Niveles						
	-3	**-2**	**-1**	**0**	**1**	**2**	**3**
Impacto Social							X
Impacto Económico							X
Total							6

$$Nivel\ de\ Impacto = \frac{\sum}{\text{Número de Indicadores}}$$

$$Nivel\ de\ Impacto = \frac{6}{2} = 3$$

Nivel de Impacto General = Alto Positivo

Fuente Propia

El nivel de impacto general del proyecto es alto positivo, de tal manera que el proyecto tiene altas expectativas para la toma de decisiones con base en los resultados obtenidos, beneficiándose el área comercial, la institución de control aduanero y la población en general del país.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

Se llevó a cabo el proceso de descubrimiento de conocimiento (KDD), con el objetivo de obtener patrones de contrabando que ayude a resolver la problemática de contrabando que presenta la Entidad Aduanera.

Los datos fueron procesados en la herramienta Pentaho Data Integration y Excel, obteniendo como resultado los patrones relevantes, que contribuirán a mejorar el control aduanero.

Se revisó la calidad de los datos obtenidos, mediante la característica de consistencia de la norma ISO/IEC 25012, dando resultados no relevantes para el análisis de los datos mediante minería de datos.

Para la implementación de técnicas predictivas se utilizó la herramienta WEKA y KNIME, las que permiten construir diferentes tipos de modelos, como árboles de decisión y regresión logística; además, identificar que las herramientas cuentan con una gran variedad de algoritmos para el análisis predictivo de datos.

La obtención del conocimiento se dio con el análisis de cada uno de los modelos de clasificación y regresión, dando como resultado una relación entre ambos; permitiendo obtener patrones relevantes de la investigación, de tal manera, con la ejecución por parte de los encargados de control aduanero, se podrá tomar decisiones enfocadas en los patrones obtenidos y así reducir la problemática.

Los resultados obtenidos demuestran que el contrabando de cebolla que se da por Santa Rosa y el contrabando de cerámica que se obtiene en Rumichaca, son patrones relevantes para este estudio, por tal motivo que resultan inesperados para los encargados de este estudio.

Existe la prohibición del contrabando, sin embargo, los productores de países fronterizos compiten con la producción local, que muy difícil se puede evitar el ingreso masivo de estos productos pese al control del personal asignado, por lo que este proyecto permitirá un mejor control aduanero y beneficiando el bienestar ciudadano en las fronteras.

Recomendaciones

El ingreso de información en la base de datos de la Entidad Aduanera, no se encuentra normalizado, presenta muchas inconsistencias, por lo tanto, se recomienda el uso del formato de Excel que se puso a disposición de la entidad, el cual permitirá mejorar la cantidad y calidad de la información.

Implementar este sistema de manera que se alimente en línea, con el fin de tener información actualizada y consistente, permitiendo una mayor agilización en el ingreso de registros y el análisis de los datos en vivo.

Para un trabajo futuro se recomienda otorgar información más relevante de tal manera que permita la aplicación de las técnicas de minería de datos a mayor profundidad, por ejemplo: datos de vehículos y de personas, como se muestra en el Anexo 2, permitiendo mejorar la calidad y cantidad de la información.

Verificar periódicamente que los datos ingresados se encuentren de manera correcta y no presenten inconsistencias.

Utilizar los patrones obtenidos para tomar decisiones relevantes con respecto a la distribución de personal de la Entidad Aduanera, permitiendo mejorar el control aduanero, promoviendo el pago de impuestos al introducir al país mercadería extranjera.

BIBLIOGRAFIA

7.0 - Pentaho Documentation. (2020). https://help.pentaho.com/Documentation/7.0

Aarthi, S., Samyuktha, M., & Sahana, M. (2019). Crime Hotspot Detection With Clustering Algorithm Using Data Mining. *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 401–405. https://doi.org/10.1109/ICOEI.2019.8862587

Ahmad, P., Qamar, S., & Qasim Afser Rizvi, S. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications*, *120*(15), 38–50. https://doi.org/10.5120/21307-4126

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. In *Journal of Engineering and Applied Sciences* (Vol. 12, Issue 16, pp. 4102–4107). https://doi.org/10.3923/jeasci.2017.4102.4107

Antezana, A. (2018). Impacto de la implementación de minería de datos en el mantenimiento y análisis de la información catastral en una municipalidad distrital. In *Universidad ESAN*. Universidad ESAN.

Aranda, Y. R., & Sotolongo, A. R. (2013). Integración de los algoritmos de minería de datos 1R, PRISM E ID3 A POSTGRESQL. *Journal of Information Systems and Technology Management*, *10*(2), 389–406. https://doi.org/10.4301/s1807-17752013000200012

Asencios, V. V. (2004). DATA MINING Y EL DESCUBRIMIENTO DEL CONOCIMIENTO (1)-9993 (electrónico). In *Revista de la Facultad de Ingeniería Industrial* (Vol. 7, Issue 2).

Barbosa, C. (2009). *MINERIA DE DATOS: 1.3 Relacion con otras disciplinas*. http://mindatos.blogspot.com/2009/09/13-relacion-con-otras-disciplinas.html

Bello, R., Arcos, L., & Magdaleno, D. (2008). *EL APRENDIZAJE AUTOMÁTICO EN LA*

GESTIÓN DEL CONOCIMIENTO.

Camilo Giraldo Mejía, J., & Alberto Vargas Agudelo, F. (2016). *Aplicación de la Técnica Regresión Logística de la Minería de Datos en el proceso de Descubrimiento de Conocimiento (KDD) en Bases de Datos Operativas o Transaccionales*.

Cardosa M., L. I. (2006). *Sistemas de Base de Datos II*. https://books.google.com.ec/books?id=wDL0VJNT4EkC&pg=PA143&dq=proceso+kdd&hl=es-419&sa=X&ved=0ahUKEwjpyr7n_sznAhVwuVkKHXsaA04Q6AEINjAC#v=onepage&q=proceso kdd&f=false

Chiriboga, G., Velasco, S., Argüello, S., Jaramillo, A., Carrión, F., & Enríquez, F. (2015). *"Economía política de la violencia en las regiones fronterizas de América Latina."* www.fiscalia.gob.ec

Christopher, E. I., Airam, E. M., López Plata, I., Batista Melián, B., & Moreno Vega, M. J. (2011). *Extraccióon de conocimiento en bases de datos*.

Cisneros, S. (2019). *DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS DESCRIPTIVAS DE AGRUPAMIENTO, ASOCIACIÓN Y ATÍPICOS EN MINERÍA DE DATOS PARA LA GESTIÓN ACADÉMICA EN LA UNIVERSIDAD TÉCNICA DEL NORTE.*

Data Mining Software, Model Development and Deployment, SAS Enterprise Miner | SAS. (2020). https://www.sas.com/en_us/software/enterprise-miner.html

Flores, L., Mariño, S., & Martins, S. (2019). *Modelado y simulación de robos y hurtos basados en redes SOM, TDIDT y Bayesianas. un caso de estudio*. 81–87.

Gandge, Y., & Sandhya. (2018). A study on various data mining techniques for crop yield prediction. *International Conference on Electrical, Electronics, Communication Computer Technologies and Optimization Techniques, ICEECCOT 2017, 2018-Janua*, 420–423. https://doi.org/10.1109/ICEECCOT.2017.8284541

García-González, J. R., Sánchez-Sánchez, P. A., Orozco, M., & Obredor, S. (2019). Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia. *Formación Universitaria*, *12*(4), 55–62. https://doi.org/10.4067/S0718-50062019000400055

García Molina, H. (2007). *Avances en Informática y Sistemas Computacionales Tomo II (CONAIS 2007) - Google Libros*. https://books.google.com.ec/books?id=rbCUa-nkXU8C&pg=PA44&dq=proceso+kdd&hl=es-419&sa=X&ved=0ahUKEwjpyr7n_sznAhVwuVkKHXsaA04Q6AEIKDAA#v=onepage&q=proceso kdd&f=false

Gorbea Portal, S. (2013). Tendencias transdisciplinarias en los estudios métricos de la información y su relación con la gestión de la información y del conocimiento. *Perspectivas Em Gestão & Conhecimento*, *3*(1), 13–27.

Guil Reyes, F. G. (2009). *Minería de patrones temporales basados en redes de restricciones. - Francisco Guil Reyes - Google Libros*. https://books.google.com.ec/books?id=FTZBAQAAQBAJ&pg=PA15&dq=proceso+kdd&hl=es-419&sa=X&ved=0ahUKEwjpyr7n_sznAhVwuVkKHXsaA04Q6AEIRTAE#v=onepage&q=proceso kdd&f=false

Han, X., Xu, L., Ren, M., & Gu, W. (2016). A Naive Bayesian network intrusion detection

algorithm based on principal component analysis. *Proceedings - 2015 7th International Conference on Information Technology in Medicine and Education, ITME 2015*, 325–328. https://doi.org/10.1109/ITME.2015.29

Jaramillo, A., & Paz-Arias, H. (2015). Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. *Revista Tecnológica ESPOL – RTE, 28*(1), 64–90.

Jaulis, J. J., & Vilcarromero, J. R. (2015). *SISTEMA DE PREDICCIÓN DE HECHOS DELICTIVOS PARA LA MEJORA DEL PROCESO DE PREVENCIÓN DEL DELITO EN EL DISTRITO DE LA MOLINA UTILIZANDO MINERÍA DE DATOS*.

Kaur, S., & Bawa, R. K. (2019). Review on data mining techniques in healthcare sector. *Proceedings of the International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2018*, 224–228. https://doi.org/10.1109/I-SMAC.2018.8653795

KNIME | Open for Innovation. (2020). https://www.knime.com/

Landa, J. (2016). *¿Qué es KDD y Minería de Datos?* –. http://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/

Li, B. (2018). A principal component analysis approach to noise removal for speech denoising. *Proceedings - 2018 International Conference on Virtual Reality and Intelligent Systems, ICVRIS 2018*, 429–432. https://doi.org/10.1109/ICVRIS.2018.00111

López, C. P. (2007). *Minería de datos: técnicas y herramientas*. https://books.google.com.ec/books?id=wz-D_8uPFCEC&pg=PA3&dq=introducción+a+la+minería+de+datos&hl=es-419&sa=X&ved=0ahUKEwj-

l72AqbbnAhUyTd8KHVB1B6MQ6AEIOzAC#v=onepage&q=introducción a la minería de datos&f=false

Luis Paulo Vieira Braga, Luis Iván Ortiz Valencia, S. S. R. C. (2009). *Introducción a la Minería de Datos*. https://books.google.com.ec/books?id=jIJEhHyESFsC&printsec=frontcover&dq=introducción+a+la+minería+de+datos&hl=es-419&sa=X&ved=0ahUKEwj-l72AqbbnAhUyTd8KHVB1B6MQ6AEIKDAA#v=onepage&q=introducción a la minería de datos&f=false

Manrique de la Cuadra, A. (2017). *Desarrollo de un catálogo de reglas de negocio referentes a datos, basado en ISO/IEC 25012 y SBVR*.

Martínez, C. (2012). *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA MEJORAR EL PROCESO DE CONTROL DE GESTIÓN EN ENTEL*. http://repositorio.uchile.cl/bitstream/handle/2250/112065/cf-martinez_ca.pdf?sequence=1

Méndez, D. S. (2015). *Los rostros del contrabando rutas fronterizas*.

Menes Camejo, I., Medina, G. A., Moreno Beltrán, P., & Carrillo, K. G. (2015). Performance of data mining algorithms in academic indicators: Decision Tree and Logistic Regression. *Revista Cubana de Ciencias Informáticas*, *9*(4).

MINTEL. (2018). *Libro Blanco de la Sociedad de la Investigación y del Conocimiento*.

Mondragón, R. (2007). *EXPLORACIONES SOBRE EL SOPORTE MULTI-AGENTE BDI EN EL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS*.

Montero Navarro, M. Á. (2009). *Extracción de conocimiento en bases de datos astronómicas*.

Morales, A., Cuevas, R., & Martínez, J. M. (2016). Analytical Processing with Data Mining.

RECI Revista Iberoamericana de Las Ciencias Computacionales e Informática, *5*(9), 22–43. http://www.reci.org.mx/index.php/reci/article/view/40/176

Nguyen, G., Dlugolinsky, S., Bobák, · Martin, Tran, V., Álvaro, ·, García, L., Heredia, I., Malík, · Peter, & Hluchý, · Ladislav. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, *52*, 77–124. https://doi.org/10.1007/s10462-018-09679-z

Normativa, D. N. J. D. de, & CÓDIGO. (2018). *CÓDIGO ORGÁNICO DE LA PRODUCCIÓN, COMERCIO E INVERSIONES*. 1–113.

Olcrod. (2019). *¿ Qué es Power BI?*

Peña Cuervo, J. J., Martínez Espinosa, L. F., & Peña Cuervo, L. A. (2018). *EL DELITO ADUANERO DE CONTRABANDO: IDENTIFICACIÓN DE LOS ELEMENTOS DE SU TIPO PENAL EN COLOMBIA*. https://doi.org/10.18359/prole.2944

Perversi, I. (2007). Aplicación de Minería de Datos para la exploración y detección de patrones delictivos en Argentina. In *... Aires: Instituto Tecnológico de Buenos Aires*. http://iidia.com.ar/rgm/tesistas/PERVERSI-tesisdegradoeningenieria.pdf

Ramchoun, H., Amine, M., & Idrissi, J. (2016). Multilayer Perceptron: Architecture Optimization and Training multi-criteria learning and nonlinear optimization View project. *Article in International Journal of Interactive Multimedia and Artificial Intelligence*, *4*, 1–26. https://doi.org/10.9781/ijimai.2016.415

Rodríguez Suárez, Y., & Amador, A. D. (2009). Herramientas de Minería de Datos Data Mining Tools. *Revista Cubana de Ciencias Informaticas*, *3*(3), 73–80. https://rcci.uci.cu/?journal=rcci&page=article&op=viewFile&path[]=78&path[]=70

Sajana, T., Sheela Rani, C. M., & Narayana, K. V. (2016). A Survey on Clustering Techniques for Big Data Mining. *Article in Indian Journal of Science and Technology*, *9*(3). https://doi.org/10.17485/ijst/2016/v9i3/75971

Sanchez, J. (2004). *Principios sobre Bases de Datos Relacionales*. www.jorgesanchez.net

Segovia, C., & Smith-Miles, K. (2019). Integrating Game Theory and Data Mining for Dynamic Distribution of Police to Combat Crime. *Proceedings - 2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018*, 780–783. https://doi.org/10.1109/WI.2018.00016

SENAE. (2016). *Informe de gestión senae i semestre 2016*.

SENAE. (2018). *Informe de Gestión 2015*. *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

SENAE. (2019). *Resumen Ejecutivo Informe de Gestión 2019*.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

SPSS Clementine Download Free Version (clementine.exe). (2020). https://spss-clementine.software.informer.com/

Tuya, J., Ramos Román, I., & Dolado Cosín, J. (2007). *Técnicas Cuantitativas para la Gestión en la Ingeniería del Software. - Google Libros*. https://books.google.com.ec/books?id=PZQoZ9KTNaEC&pg=PA247&dq=proceso+kdd&hl=es-419&sa=X&ved=0ahUKEwjpyr7n_sznAhVwuVkKHXsaA04Q6AEILzAB#v=onepage&q&f=false

UIAF. (2014). Técnicas de minería de datos para la detección y prevención del lavado de activos y la financiación del terrorismo (LA/FT). In *Minhacienda*.

UNODC. (2019). *ODS16*. https://www.unodc.org/mexicoandcentralamerica/es/romex/ODS16.html

Uzlov, D., Vlasov, O., & Strukov, V. (2019). Using Data Mining for Intelligence-Led Policing and Crime Analysis. *2018 International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T 2018 - Proceedings*, 499–502. https://doi.org/10.1109/INFOCOMMST.2018.8632122

Valenga, F., Fernández, E., Merlino, H., Rodríguez, D., Procopio, C., Britos, P., & García-Martínez, R. (2008). Minería de datos aplicada a la detección de patrones delictivos en Argentina. *7th Jornadas Iberoamericanas de Ingenieria de Software e Ingenieria Del Conocimiento 2008, JIISIC 2008*, *1*, 31–40.

Vila, D. (2019). *DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL UTILIZANDO TÉCNICAS PREDICTIVAS DE CLASIFICACIÓN Y REGRESIÓN DE MINERÍA DE DATOS, PARA LA GESTIÓN ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DEL NORTE*. UNIVERSIDAD TECNICA DEL NORTE.

Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (2020). https://www.cs.waikato.ac.nz/ml/weka/

Yacup, N. A., Antonia, M., Alcázar, W., Mauricio, Y., Pérez, N., & Castillo Landínez, S. P. (2018). *IDENTIFICACIÓN DE PATRONES DELICTIVOS EN COLOMBIA DURANTE EL PERIODO 2010-2016 MEDIANTE EL USO DE TÉCNICAS DE MINERÍA DE DATOS*.

Zhai, X., Ali, A. A. S., Amira, A., & Bensaali, F. (2016). MLP Neural Network Based Gas

Classification System on Zynq SoC. *IEEE Access*, *4*, 8138–8146. https://doi.org/10.1109/ACCESS.2016.2619181

Lara, J. (2014). *Minería de Datos (CENTRO DE ESTUDIOS FINANCIEROS).*

Solarte Martínez, G. R. (2009). *Técnicas de clasificación y análisis de representación del conocimiento para problemas de diagnóstico.*

ANEXOS

Anexo 1: Acta de Confidencialidad: https://bit.ly/36TGBKR

Anexo 2: Formato Macro Excel de Ingreso de Registros: https://bit.ly/33U0mQx

Anexo 3: Preguntas del negocio para análisis exploratorio (BI): https://bit.ly/2JX5agT

Anexo 4: Respuesta de preguntas del negocio: https://bit.ly/36U026e