UNIVERSIDAD TÉCNICA DEL NORTE.

Facultad de Ingeniería en Ciencias Aplicadas.

Carrera de Ingeniería en Sistemas Computacionales.

DETECCIÓN DE PATRONES DE CONSUMO Y RECLAMOS EN EL SERVICIO DE AGUA POTABLE DE LA JUNTA DE AGUA DE SAN JUAN DE ILUMÁN, UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS PARA FORTALECER LA GESTIÓN ADMINISTRATIVA

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

Autor.

Jesús Eduardo Gonzales Conterón

Director.

PhD. Iván Danilo García Santillán

Ibarra – Ecuador

2023

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:	1003859400		
APELLIDOS Y NOMBRES:	GONZALES CONTERÓN JESÚS EDUARDO		
DIRECCIÓN:	OTAVALO – SAN JUAN DE ILUMÁN, EUGENIO ESPEJO Y MODESTO LARREA		
EMAIL:	jegonzalezc@utn.edu.ec – jesus2019edu@gmail.com		
TELÉFONO FIJO:	(06) 2946-310	**TELÉFONO MÓVIL:**	0980782879

DATOS DE LA OBRA	
TÍTULO:	DETECCIÓN DE PATRONES DE CONSUMO Y RECLAMOS EN EL SERVICIO DE AGUA POTABLE DE LA JUNTA DE AGUA DE SAN JUAN DE ILUMÁN, UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS PARA FORTALECER LA GESTIÓN ADMINISTRATIVA
AUTOR (ES):	GONZALES CONTERÓN JESÚS EDUARDO
FECHA: DD/MM/AAAA	06/06/2023
SOLO PARA TRABAJOS DE GRADO	
PROGRAMA:	■ **PREGRADO** ☐ **POSGRADO**
TITULO POR EL QUE OPTA:	**INGENIERÍA EN SISTEMAS COMPUTACIONALES**
ASESOR /DIRECTOR:	**PhD. Iván Danilo García Santillán**

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 06 días del mes de junio del 2023

EL AUTOR:

(Firma).....

Nombre: Gonzales Jesús

UNIVERSIDAD TÉCNICA DEL NORTE
FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

CERTIFICADO DEL DIRECTOR

En mi calidad de Tutor de Trabajo de Grado presentado por el egresado, **Gonzales Conterón Jesús Eduardo** para optar por el Título de Ingeniero en Sistemas Computacionales, cuyo tema es: **DETECCIÓN DE PATRONES DE CONSUMO Y RECLAMOS EN EL SERVICIO DE AGUA POTABLE DE SAN JUAN DE ILUMÁN, UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS PARA FORTALECER LA GESTIÓN ADMINISTRATIVA.** 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.

En la ciudad de Ibarra a los 06 días del mes de junio del 2023.

PhD. Iván García

DIRECTOR DE TRABAJO DE GRADO

Dedicatoria.

Por apoyarme en cada momento, María Fabiola Conterón de la Torre y Jaime Gonzales Guaján, quienes me han apoyado en cada momento, llenado de valores y me ha enseñado que nuestra única pasión en esta profesión es en seguir aprendiendo.

A mis hermanos, Anita Gonzales, Ángel Gonzales, Inti Gonzales, Tamia, Gonzales, Sisa Gonzales, Ñusta Gonzales, Hekaya Gonzales y Toa Gonzales, a mis cuñados Fausto Cando y Etelvina Cáceres, a mis sobrinos Mayte Cando, Leila Cando y Katherine Gonzales, que con su apoyo y consejos han estado pendientes en cada momento.

Jesús Eduardo Gonzales Conterón

Agradecimientos

El presente trabajo de tesis primeramente me gustaría agradecer a Dios por bendecirme para llegar hasta donde he llegado, porque ya es realidad la una de mis metas tan anheladas, a mi tutor PhD. Iván García Santillán y mis opositores PhD. Marco Pusdá, MSc. MacArthur Ortega Bustamante por todos sus conocimientos brindados, apoyo y trabajo para poder culminarlo exitosamente.

Gracias a todos quienes integran la carrera de Ingeniería en Sistemas Computacionales de quienes me llevo los más gratos recuerdos y a la Universidad Técnica del Norte por darme la oportunidad de estudiar y ser un gran profesional.

A la Junta Administrador de Agua Potable y Saneamiento Regional Ilumán por la apertura y apoyo brindando en el transcurso y desarrollo del proyecto.

Jesús Eduardo Gonzales Conterón

Contenido

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Los altos índices en reclamos y consumo de agua han venido afectando los servicios de la Junta de agua de San Juan de Ilumán, esta investigación pretende obtener patrones de consumo y reclamo relativo al servicio de agua potable de la Junta de Agua, aplicando técnicas predictivas y descriptivas de minería de datos, procesando registros históricos del sistema YAKUSOFT, desde el año 2017 a 2022. Con la metodología KDD y 13994 registros se logró la vista minable el cual permitió utilizar técnicas de agrupación y clasificación de datos mediante la herramienta Weka.

La calidad de los servicios como problemas de la Junta de agua da como objetivo primordial mejorar los servicios de agua potable de todos los sectores de San Juan de Ilumán; supervisando a la entidad JAAPYSR-I, en el desempeño de sus funciones. El tiempo de respuesta proporcionado por la junta de agua se utiliza como variable principal en dar atención de casos que se han acumulado por los usuarios. Estos datos vienen hacer la principal información en la implementación de la metodología KDD, la cual busca satisfacer a los usuarios y lograr este objetivo a través de un crecimiento relevante.

ABSTRACT

The high rates of claims and water consumption have been affecting the services of the San Juan de Ilumán Water Board, it is intended to obtain patterns of consumption and claim of the drinking water service of the Water Board, applying predictive and descriptive techniques of data mining, processing historical data registered in the YAKUSOFT system, from 2017 to 2022. With the KDD methodology and 13,994 records, the mineable view was achieved, which allowed the use of data grouping and classification techniques with the help of the Weka tool.

The quality of services with the problems of the Water Board has as its primary objective to improve drinking water services in all sectors of San Juan de Ilumán; supervising the entity JAAPYSR-I, in the performance of its functions as an independent entity to the sectors that use potable water and sewerage. It seeks the satisfaction of the users and through this objective to have a relevant growth, as the main variable the response time given by the water board is used, to provide the solution to different cases presented by the user's, said data is the raw material in the implementation of the KDD methodology.

INTRODUCCIÓN

En la provincia de Imbabura el uso de técnicas de minería de datos, para análisis de información, no se ha usado en ninguna de las parroquias o juntas de agua y alcantarillado, por lo tanto, los problemas existentes no han mejorado a lo largo de los años y la comunidad sigue viéndose afectada. El principal objetivo es detectar los patrones de consumo y reclamos en el servicio de agua potable de la Junta de agua de San Juan de Ilumán, utilizando técnicas predictivas / descriptivas de minería de datos e inteligencia de negocios que se utilizarán después de que se implementen los procesos de descubrimiento (KDD) logrando mostrar la vista minable con 16 atributos, valores numéricos, cadenas de carácter y con un total de 13994 exploraciones para su análisis en la herramienta Weka, usando algoritmos de árboles de decisión y agrupamiento de datos con métricas cuantitativas y cualitativas evaluados por modelos estadísticos y matriz de confusión.

- **Tema**

Detección de patrones de consumo y reclamos en el servicio de agua potable de la junta de agua de San Juan de Ilumán, utilizando técnicas de minería de datos e inteligencia de negocios para fortalecer la gestión administrativa.

- **Problema**

En la provincia de Imbabura el uso de técnicas de minería de datos, para análisis de información, no se ha usado en ninguna de las parroquias o juntas de agua y alcantarillado, por lo tanto, las comunidades aún se ven afectadas por problemas continuos, que no han mejorado con el tiempo, donde sus opiniones no son tomadas en cuenta.

En JAAPYSR-I, se ha evidenciado que no hay una técnica óptima que permita hacer el seguimiento a las quejas, solicitudes y reclamos de los usuarios, esto se ha venido presentando en cada periodo administrativo, generando altos índices de inconformidad en la mayoría de los habitantes. Este problema puede ser producido por la falta de

recursos, mala administración y compromiso con la gente, afectando la calidad de los servicios, por lo tanto, la implementación de un estudio de minería de datos permitirá agilizar los procesos de toma de decisiones y mejorar el servicio de agua potable y saneamiento de los usuarios.

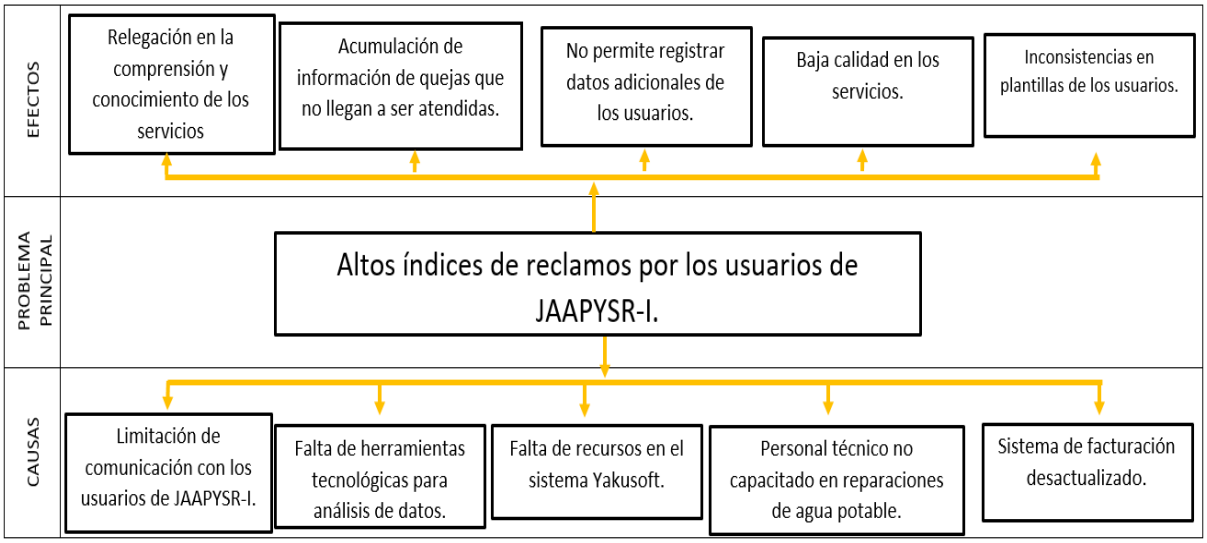


Fig. 1. Espina de pescado

- **Objetivos**

General

Detectar los patrones de consumo y reclamos en el servicio de agua potable de la junta de agua de San Juan de Ilumán, utilizando técnicas predictivas / descriptivas de minería de datos e inteligencia de mercados para fortalecer la gestión administrativa.

Específicos

- Sustentar las técnicas predictivas y descriptivas de minería de datos de acuerdo con bases teóricas de procesos en el descubrimiento de conocimiento en base de datos (KDD).
- Elaborar una data warehouse y vista minable, a partir de los datos registrados en el sistema de agua potable (YAKUSOFT), utilizando la Suite de Pentaho.

- Construir y evaluar los modelos predictivos / descriptivos para el análisis de los datos utilizando las herramientas Weka.

- **Alcance**

Mediante la investigación se obtendrá patrones de consumo y reclamos de los beneficiarios de agua potable de San Juan de Ilumán, con el propósito de tratar a tiempo las quejas de los usuarios, el cual da un enfoque a realizar un análisis con base a los datos históricos obtenidos en el sistema de cobro YAKUSOFT, a nivel de todas las comunidades y barrios de la parroquia, con ayuda de la metodología (KDD) implementando estrategias y técnicas para seleccionar, transformar y limpiar datos encontrados en la base de datos PostgreSQL, mediante la herramienta Pentaho Data Integration (PDI) el cual permite construir el data warehouse para generar la vista minable posteriormente aplicar en la herramienta Weka, algoritmos de predicción y descripción de métricas cuantitativas y cualitativas, finalmente validar los resultados con valores estadísticos de minería de datos en la herramienta Weka y Power BI.

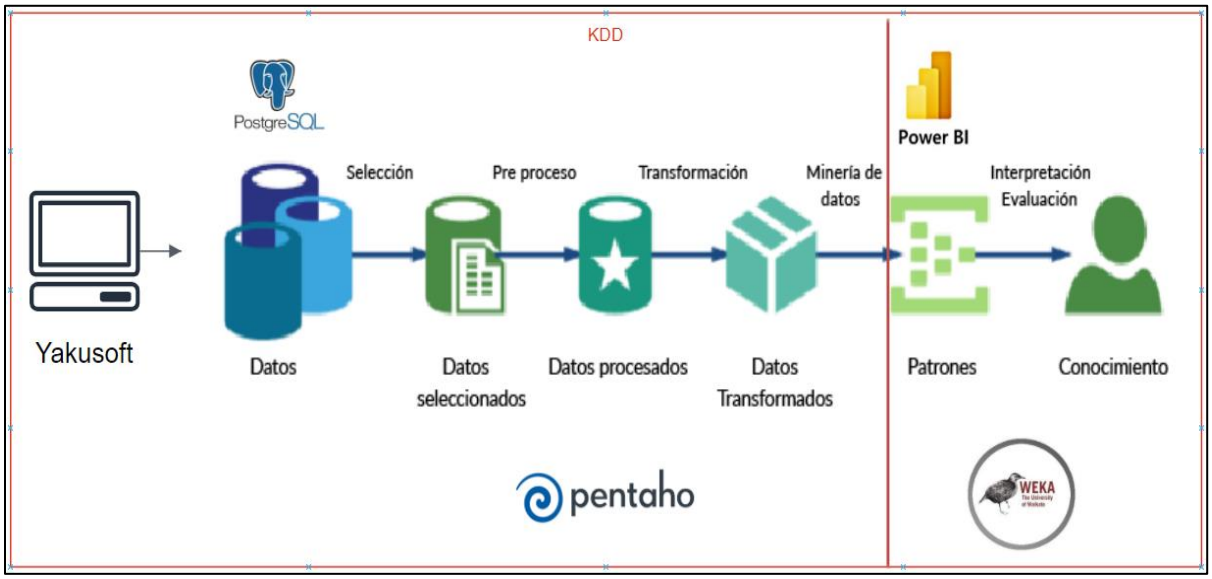


Fig. 2. Metodología KDD

CAPÍTULO I

MARCO TEÓRICO

1. Introducción (Data Mining)

Actualmente las instituciones públicas y privadas cuentan con repositorios para almacenar grandes datos históricos generados por las actividades a las que se dedican, tomando a la tecnología como herramienta principal en gestionar información, la minería de datos convierte enormes fuentes de datos en descubrimiento de conocimiento. Se convierten en retos computacionales orientados a descubrir conocimiento y aplicar estrategias de negocio a casos futuros.

1.1. Técnica de minería de datos

Es el proceso de adquirir conocimiento en una base de datos (Haro et al., 2018). Permite implementar estrategias en análisis de datos para obtener un nuevo conocimiento que proporciona una vista simple y clara de los datos, lo que facilita que su organización tome decisiones. Esta técnica permite extraer datos y construir modelos en base a indicadores estadísticos (leskovec et al., 2020).

Algunos mencionan que la minería de datos tiene un aprendizaje por entrenamientos con diferentes algoritmos y de forma automática. Se encargan de implementar estas estrategias con diferentes algoritmos como árboles de decisión, EM, K-means, RandomForest entre otros (Rajaraman et al., 2014).

1.2. Áreas relacionadas

Los progresos tecnológicos en la gestión de bases de datos permiten la integración entre diferentes disciplinas como se muestra en la Fig. 3. Además, se puede tener acceso a grandes volúmenes de datos para analizar con ayuda de distintas áreas de la tecnología.

a. Estadística

La estadística es una de las métricas que utilizan los algoritmos de minería de datos (Lara, 2014). Aplicar modelos matemáticos desarrollados en estadística a datos de minería para verificar resultados.

b. Bases de datos

Los datos históricos de la organización se almacenan en una base de datos. Existen innumerables herramientas de almacenamiento como: MongoDB, Oracle, MySQL, PostgreSQL etc. relacionadas por permitir almacenar valores numéricos, cadenas de carácter, Boleano entre otros y posteriormente llevarlos a un modelo cuantitativo y cualitativo.

c. Aprendizaje automático

El reconocimiento de forma automática al implementar metodologías de minería de datos presenta un modelo con eficiencia en resolver diferentes situaciones o estado actual en la que una institución se encuentra permitiendo en facilitar y comprender de manera sencilla gran cantidad de información de manera automática (Sierra, 2006).

d. Otras

Permite relacionarse con muchas áreas facilitando un plan estratégico de negocio que ayuda a visualizar con un modelo predictivo o descriptivo en la que se encuentra una institución, como también se incluye imágenes o símbolos para analizar.

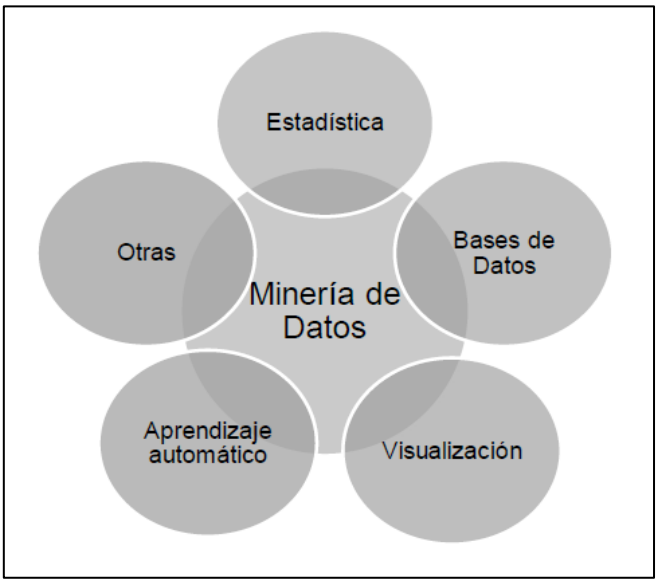


Fig. 3. Áreas relacionadas (Lara 2014).

Tipo de datos

El valor de los datos construye una comunicación por medio de caracteres, símbolos, letras, números etc. Mostrados en el ejemplo de la Tabla 1. Se caracterizan por transformar estos datos en información tomando en cuenta que son únicamente datos generados por modelos estadísticos los cuales no pueden generar ningún procedimiento que se dé (D'Ambrosio, 2018).

Tabla 1. Asociados a localidades

ATRIBUTO	DATO
NOMBRECIUDAD	Otavalo
CIUDAD	O
NUMEROHABITANTES	12340
NOMBREPARROQUIA	Ilumán
NOMBRECOMUNA	Pinsaquí

Fuente: Elaboración Propia

Cuantitativos

Muestran valores numéricos medibles.

a. **Discretos.** Conforman los números que su valor es limitado. Ejemplo: Número de estudiantes, número de usuarios.

b. **Continuos.** Conforman datos en los cuales es posible hallar valores medibles en lo posible. Ejemplo: Sueldo, peso, edad, altura de una persona.

Cualitativos

Muestran valores categóricos:

a. **Nominales.** Muestran una categoría que caracteriza una persona. Ejemplo: nivel académico, origen entre otros.
b. **Ordinales.** Tienen un orden jerárquico dentro de una categoría. Ejemplo: activo, inactivo, peso entre otros.

1.3. Bases de datos

En el caso de investigaciones que analizan la producción intelectual de una determinada organización, campo de estudio o área geográfica, las bases de datos científicas se han convertido en una herramienta importante (Vuotto et al., 2020).

1.3.1. Características generales

Según (Pulido Romero et al., 2019). Describa tres diferentes tipologías de peculiaridades de bases de datos.

- Imparcialidad de los datos. Por lo tanto, pueden ser utilizados por cualquier aplicación porque son independientes del programa.
- Sistemas menos redundantes. Nos referimos a la duplicación de datos como redundancia.
- Seguridad. La base de datos (BD) está protegida contra usuarios no autorizados.

1.3.2. Tipos de bases de datos

Relacional

La empresas e instituciones públicas y privadas cuentan con repositorios de base de datos para mantener seguros sus registros contables, con el fin de mantener de forma segura en bases de datos relacionales representadas en tablas. (Lara, 2014). Indicado en la Tabla 2.

Tabla 2. Ejemplo de bases de datos relacionales

Comunidad			Persona		
IdComunidad	nombreComunidad		idPersona	nombrePersona	IdComunidad
1	centro		1003859400	Juan	1
2	agualongo		1001913134	Pedro	4
3	Pinsaquí		1002800464	Lucas	3
4	centro		1004289456	Inti	1

Fuente: Elaboración Propia

1.4. Metodología (KDD)

Hendrickx define el Proceso de descubrimiento del conocimiento (KDD) como, una nueva estrategia con buenos veneficios a futuro útiles para detectar patrones ocultos dentro de una vista minable (Hendrickx et al., 2015). Se puede observar los procesos que adquiere los datos hasta obtener un conocimiento Fig. 4.

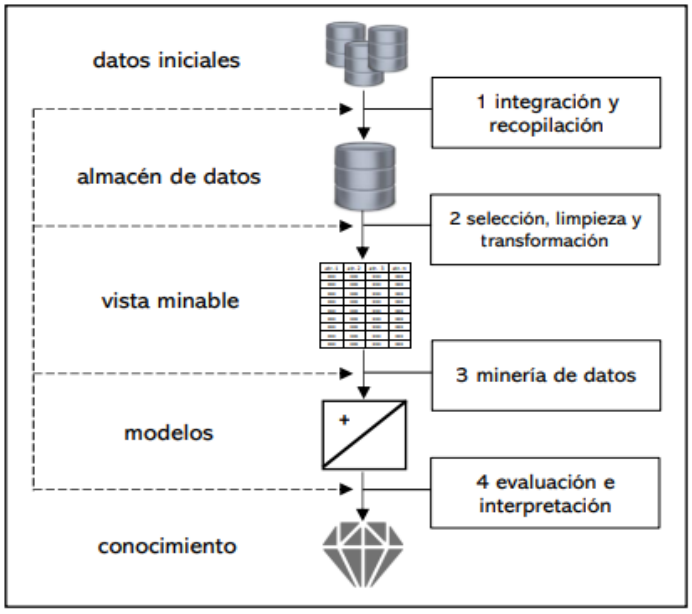


Fig. 4. Metodología (KDD) (Hendricks et al. 2015).

1.5. Etapas de la metodología KDD

Fase de integración y recopilación

Consiste en la obtención de datos históricos en diferentes fuentes y tener un almacén de datos general de información valiosa para su análisis de modo que la empresa o institución pueda facilitar en sus decisiones como muestra la Fig. 5.

Viene siendo un modelo complejo en almacenamiento de datos o conocidos como data warehouse, que provienen de diferentes fuentes de información constituyendo un nuevo almacén de base de datos donde se puede aplicar y seleccionar un algoritmo de acuerdo a las necesidades de las instituciones.

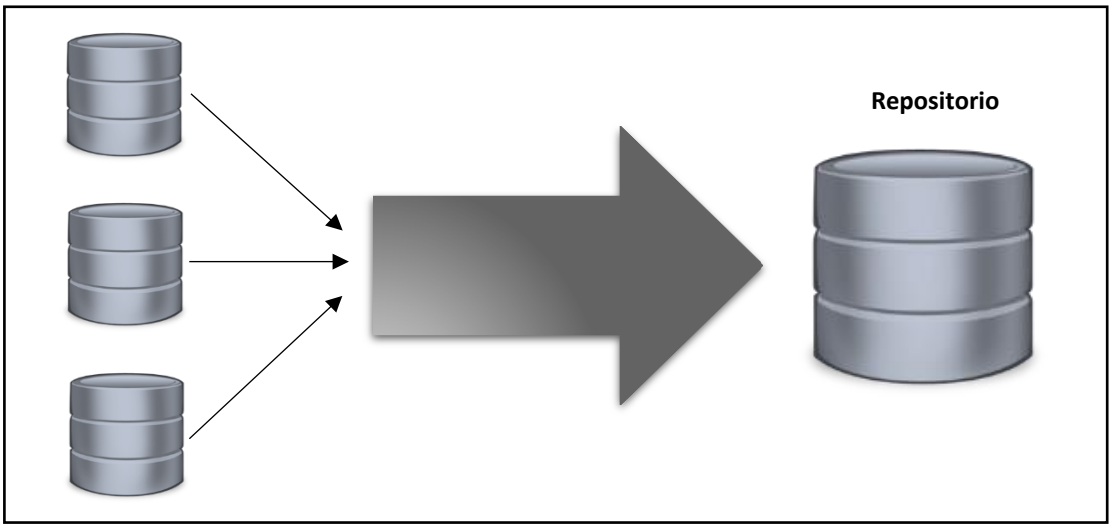


Fig. 5. Almacenamiento de datos (Santín and López 2007)

Fase de selección, limpieza y transformación

Además, se utilizan técnicas de minería de datos para lograr una buena calidad de análisis, por lo que se requiere selección, limpieza y transformación para obtener una imagen clara de la mina. (Pérez & Santín, 2006).

a. Selección

Para iniciar el proceso KDD, y una vez determinado la información más importante se procede a construir los datos a detectar que conjuntamente va de la mano con los objetivos de negocio el cual parte de los datos base y permite obtener un conocimiento (Timaran et al., 2016)**.**

b. Limpieza

Limpieza o filtrado de datos: en esta etapa, los datos se limpian o filtran de acuerdo con las necesidades y el algoritmo que se aplicará para eliminar los valores incorrectos o no identificados (Monjas, 1999).

Según (Gallardo, 2016). Menciona que, “Con la amplia gama de técnicas que existen para mejorar la calidad de los datos y dejarlos listos para la fase de modelado, esta tarea, que completa la anterior, es la que más tiempo lleva y requiere más trabajo”.

c. Transformación.

Las características útiles se buscan en esta etapa porque las representarán y dependerán de ellas de acuerdo con el resultado buscado del proceso. Usaremos técnicas de reducción de dimensiones para eliminar los datos más inconsistentes.

Fase de minería de datos

Al implementar técnicas de minería de datos, es muy importante considerar la toma de decisiones y tratar de identificar información que no afecte negativamente el estado actual de la institución para obtener el conocimiento necesario.

a. Elegir la estrategia de minería de datos que producirá el mejor modelo predicción o descripción.
b. Escoger los modelos principales de acuerdo con los datos obtenidos por ejemplo clustering, clasificación, regresión etc.
c. Elegir el algoritmo con mejores resultados el cual permitan facilitar la obtención de conocimiento.

Tarea de minería de datos

a. Tareas predictivas

Para crear reglas que se puedan aplicar para crear predicciones, el análisis predictivo requiere herramientas informáticas que puedan identificar patrones en los datos analizados (Espino & Martínez, 2017).

- **Clasificación**

La identificación de las características de un objeto o registro con el fin de categorizarlo en una clase o categoría predeterminada constituye una clasificación. Su objetivo es determinar si una instancia particular del conjunto de ocurrencias (conjunto de datos) se ajusta a una de las diversas clases que se han definido previamente. Para eso, necesitamos construir un modelo de clasificación. Valores discretos: las clases son

los componentes de la salida que se obtuvo. En el uso real, estas clases normalmente se definen a partir de valores particulares de variables o campos particulares(Gonzalez Marcos, 2007)**.**

b. Tareas Descriptivas.

Permite la agrupación de características preestablecidas de acuerdo con criterios de distancias o similitud.

- **Agrupamiento**

Esta es una tarea descriptiva destinada para derivar "agrupaciones naturales" de los datos (Chávez, 2017). A diferencia de la clasificación, donde los datos de muestra se etiquetan en clases, el objetivo aquí es encontrar "etiquetar variables" que aún no existen (Mining et al., 2016).

Formula Índice Dunn:

$$Índice\ Dunn = \frac{\min(Distancia\ interclúster)}{\max(Distancia\ interclúster)}$$

Fase de evaluación e interpretación

En esta fase se procede a la validación e interpretación de los resultados por medio de valores estadísticos, que permitan dar un resultado eficiente para la interpretación (Lara, 2014).

- **Evaluación**

Según Sierra, “La validación de un modelo predictivo mide su capacidad para predecir nuevos casos. Generalmente confirmado por comparación con valores estadísticos.” (Sierra, 2006).

1.6. Modelos de datos

Los modelos predictivos determinan que los datos analizados sean valores que establezcan diferentes casos a futuro (Sierra, 2006).

Clasificación

En minería de datos, la clasificación es una técnica supervisada donde busca determinar si los atributos pertenecen o no a un determinado concepto (3). Normalmente, un atributo de clase está presente. La capacidad de asignar un elemento de datos a una de varias clases predefinidas se conoce como clasificación. Las características (también conocidas como variables o atributos) de un objeto sirven como su descripción $X \rightarrow \{X1, X2, \dots Xn\}$. El objetivo de la tarea de clasificación es clasificar el objeto dentro de una de las categorías de la clase $C =\rightarrow \{C1, C2, \dots Ck\} F: X1xX2x \dots xXn \rightarrow C$ (Haro Rivera et al., 2018).

- **Árboles de clasificación**

Son algoritmos que organizan los datos en forma jerárquica, muestran sus predicciones siguiendo desde la raíz hasta la punta de sus hojas representadas en grafos (Hernández & Ramirez, 2004).

RandomForest: Crea datos aleatorios como un bosque relacionando todos los atributos de la data warehouse generados por el algoritmo RandomTree.

RandomTree: Crea datos aleatorios que parten desde la raíz hasta llegar a las hojas denominadas grafos sin eliminar ninguna hoja (Frank et al., 2016).

1.7. Aplicación de Normas (ISO/IEC 25012:2008)

El estándar ISO/IEC 25012 le permite lograr la calidad de los datos midiendo los sistemas de información en un entorno específico. Identificamos 15 características de calidad que se pueden evaluar desde dos perspectivas: intrínseca y dependiente del sistema (PERALTA et al., 2022).

Los sistemas automatizados permiten alcanzar el nivel de calidad del procesamiento de datos. Desde esta perspectiva, la calidad de los datos está determinada por el dominio técnico en el que se utilizan los datos y se logra a través de las capacidades de los componentes del sistema informático (Calabrese et al., 2019). Tales como:

- **Hardware:** Asistencia para lograr la recuperabilidad.
- **Software:** Lograr la portabilidad con herramientas de migración.

Los técnicos de sistemas suelen ser los encargados de este punto de vista (Calabrese et al., 2019). Mostrados en la Tabla 3.

Tabla 3. Características de calidad de datos

Característica	Inherente	Dependiente del sistema
Exactitud	X	
Completitud	X	
Consistencia	X	
Credibilidad	X	
Actualidad	X	
Accesibilidad	X	X
Conformidad	X	X
Confidencialidad	X	X
Eficiencia	X	X
Precisión	X	X
Trazabilidad	X	X
Comprensibilidad	X	X
Disponibilidad		X
Portabilidad		X
Recuperabilidad		X

Fuente: Elaboración Propia

La precisión determina el valor deseado en un contexto determinado. Completitud, datos requeridos están completos. La consistencia se refiere a datos que son claros y consistentes en un contexto dado.

En cambio, las características que determinan la calidad de los datos específicos y relacionados con el sistema son: Confidencialidad (en términos de seguridad de la información) describe el nivel de acceso a cada dato en un contexto particular (por ejemplo, personas que requieren tecnología de asistencia para ciertos tipos de discapacidades). accesibilidad (Calabrese et al., 2019).

1.8. Parroquia San Juan de Ilumán

1.8.1. Organización territorial.

El territorio parroquial cuenta con una superficie de 22,9 km2, situado a una altitud de 2.400 metros (río Ambi) a 4.650 metros (volcán Imbabura), tiene muchas capas ecológicas. (Autónomo, 2015). Población y superficie territorial mostrados en la Fig. 6 y Fig. 7.

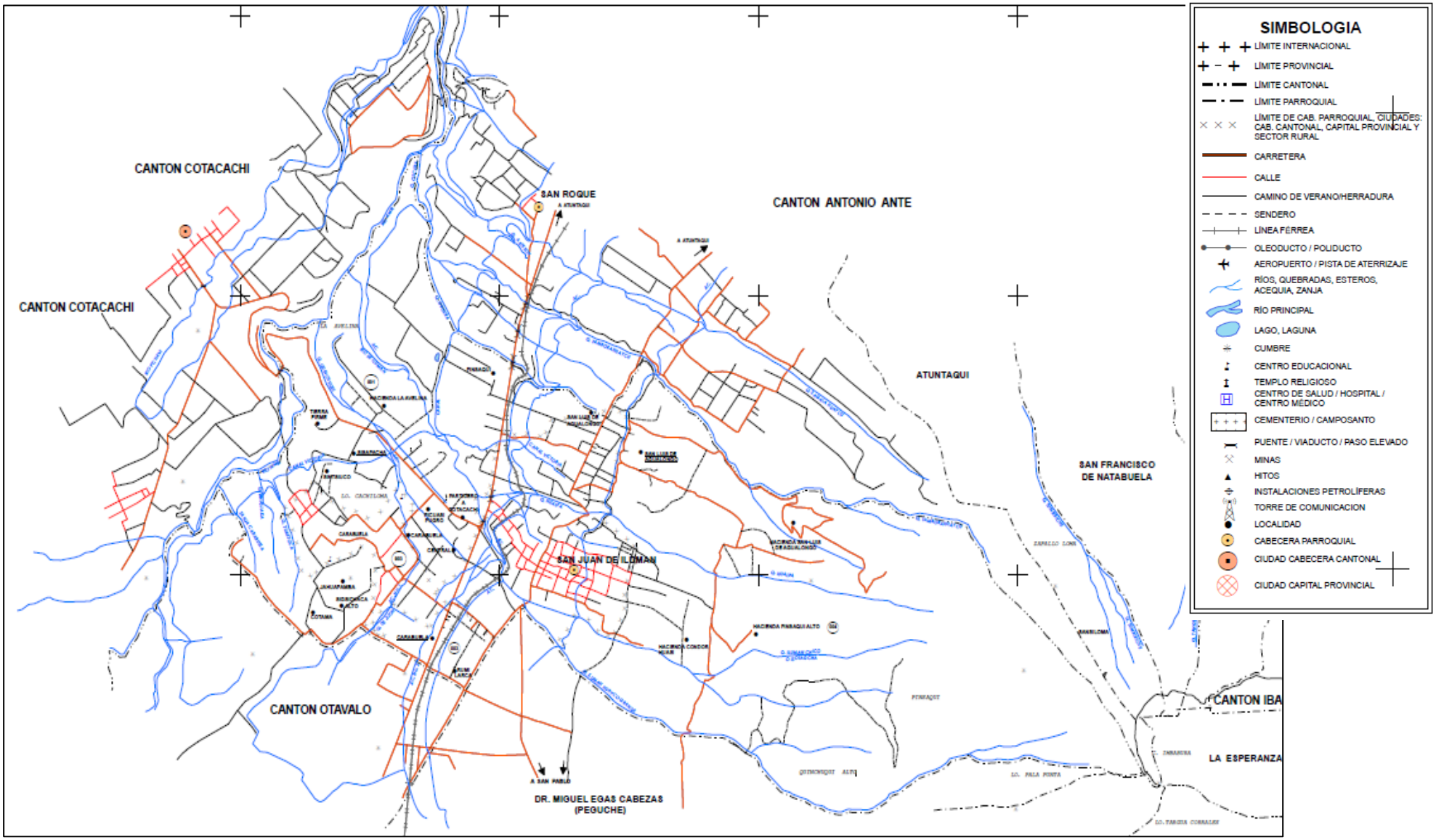


Fig. 6. Mapa de San Juan De Ilumán (INEC)

PARROQUIAS	SUPERFICIE km^2 Según IGM	SUPERFICIE Km2 Según GMO	POBLACIÓN Censo 1990	POBLACION Censo 2001	POBLACION Censo 2010
San Luis de Otavalo	82,1	74	35.889	44159	52753
Dr. Miguel Egas Cabezas (Peguche)	7,98	14	3.544	4231	4883
Eugenio Espejo	24,05	30	6.416	6004	7357
González Suárez	50,92	52	4265	5320	5630
San José de Pataquí	8,88	10	494	359	269
San José de Quichinche	89,71	118	4931	7318	8476
San Juan de Ilumán	21,82	21	5526	7225	8584
San Pablo	64,57	64	8833	9106	9901
San Rafael	19,58	18	2.559	4762	5421
Selva Alegre	137,86	178	2075	1704	1600
Total	**507,47**	**579**	**74.532**	**90188**	**104874**

Fig. 7. Superficie territorial (INEC)

El área total de la parroquia es de 21 km², de los cuales hay nueve comunidades (Autónomo, 2015) mostradas en la Fig. 8.

COMUNIDADES		BARRIOS SECTOR URBANO	
1	San Luis de Agualongo	1	Central
2	Ilumán Bajo	2	Santo Domingo
3	Ángel Pamba	3	Ilumán Alto
4	Pinsaquí	4	San Carlos
5	Carabuela	5	Hualpo
6	San José de Jahua Pamba	6	Rumilarka
7	Sinsi Uco	7	Rancho Chico
8	Capilla Centro	8	Santa Teresita
9	Picuasi	9	Guabo
		10	Cóndor Mirador
		11	Azares

Fig. 8. Comunidades y Barrios de San Juan de Ilumán (INEC)

1.8.2. Diagnóstico del sistema territorial.

De acuerdo al censo más reciente de noviembre de 2010, las cifras oficiales sitúan la población total en 8.584 (Autónomo, 2015) según las cifras del INEC Fig. 9.

Título
POBLACIÓN POR ÁREA, SEGÚN PROVINCIA, CANTÓN Y PARROQUIA DE EMPADRONAMIENTO

Provincia	Nombre del Cantón	Nombre de la Parroquia	ÁREA		
			URBANO	RURAL	Total
	OTAVALO		URBANO	RURAL	Total
		DR. MIGUEL EGAS CABEZAS	-	4.883	4.883
		EUGENIO ESPEJO (CALPAQUI)	-	7.357	7.357
		GONZALEZ SUAREZ	-	5.630	5.630
		OTAVALO	39.354	13.399	52.753
		PATAQUI	-	269	269
		SAN JOSE DE QUICHINCHE	-	8.476	8.476
		SAN JUAN DE ILUMAN	-	8.584	8.584
		SAN PABLO	-	9.901	9.901
		SAN RAFAEL	-	5.421	5.421
		SELVA ALEGRE	-	1.600	1.600
		Total	39.354	65.520	104.874

Fig. 9. Población territorial (INEC)

1.8.3. Agua para el consumo humano

El agua se obtiene directamente de manantiales en las faldas del volcán Imbabura en la parroquia y más allá. Uno de los suministros de agua proviene de la Zona de Araque y el otro de las faldas del cerro Imbabura, y llega a la vivienda mediante un sencillo proceso de recolección y transporte. (Autónomo, 2015) tal como se demuestra en el gráfico. Fig.10.

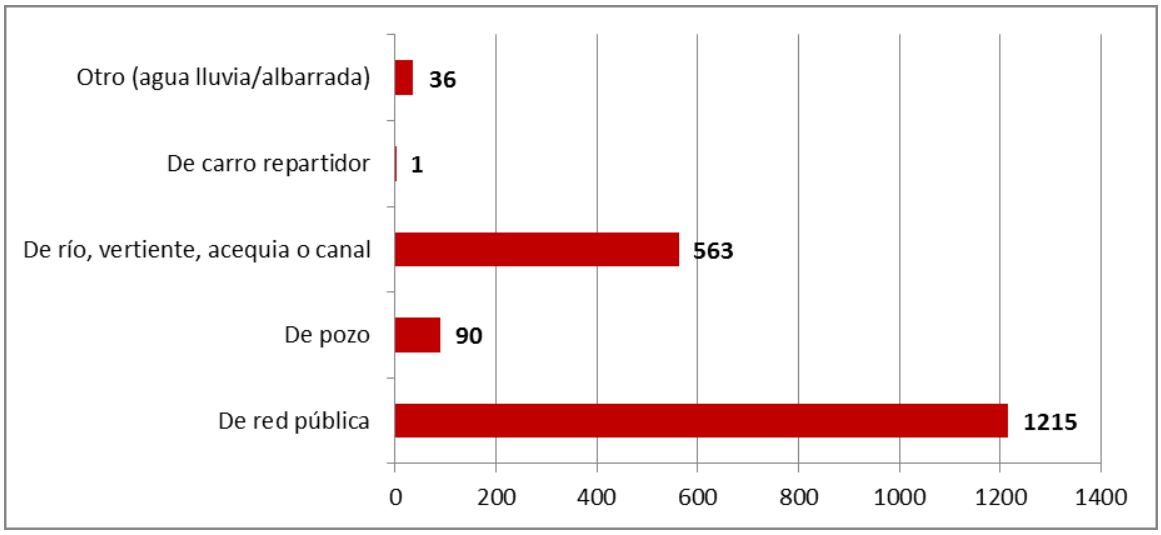


Fig. 10. Origen del agua para consumo humano

El régimen de agua se iniciaba con la captación de Araque llegando a las viviendas por simple tubería; no tiene problemas de distribución en las partes altas de la parroquia porque la distribución es por gravedad y tiene capacidad para llegar a un tercer nivel sin necesidad de bombeo. Si bien el área urbana tiene cobertura completa durante todo el año, es raro en el verano como muestra en los datos de la Fig. 11. Con una calidad aceptable, se carece de agua para riego.

Cuadro 14 Cobertura del sistema de agua potable

AGUA POTABLE								
Ilmuán	COBERTURA	CALIDAD	# FAMILIAS	CAUDAL	PRESION	FUENTE ABASTECIMIENTO	CONCESIÓN m3/seg	DIAMETRO TUBERIA
Urbano	100%	Buena	1336			Araque		
Rural	50%	Buena				Araque		

Fig. 11. Cobertura de sistema de agua potable

- **Agua Potable**

El área urbana está cubierta al 85%, con el problema de que el sistema ha colapsado por cumplir su vida útil y se hace necesario el cambio de tubería con mayor diámetro con una proyección de 30 años o más. Fig. 12.

Fig. 12. Sistema de agua potable de San Juan de Ilumán (JAAPYSR-I)

- **Alcantarillado**

El área urbana está cubierta al 100%, con el problema de que el sistema ha colapsado por cumplir su vida útil y se hace necesario el cambio de tubería con mayor diámetro con una proyección de 30 años o más, con el beneficio del alcantarillado no cuenta el área rural en un 100%. Son cuatro (4,00) km aproximadamente que forman parte del sistema de alcantarillado urbano que debe ser reemplazado Fig. 13.

Fig. 13. Sistema de Alcantarillado de San Juan de Ilumán (JAAPYSR-I)

1.9. Trabajos relacionados

La minería de datos ahora admite buscar automáticamente a través de grandes cantidades de información, descubriendo patrones recurrentes que revelan cómo se comportan los datos y de los cuales se pueden sacar conclusiones para ayudar a que el negocio se desarrolle y crezca. A continuación, se demuestra el uso de la minería de datos para situaciones con el consumo y reclamo de juntas administradoras de agua potable.

Según (Troncoso Espinosa et al., 2020) Utilizaron tres bases de datos correspondientes a los registros de consumo comercial, consumo de agua y fiscalización de 970.000 clientes en Concepción, Chile. Utilizaron 5 técnicas de minería de datos denominadas Decision Tree, Naive Bayes, Neural Networks, Support Vector Machines y KNN. De estos, los árboles de decisión lograron el mejor rendimiento con un 88 %, seguidos de las máquinas de vectores de soporte con un 77 % de precisión. Dice, además: “Identifica variables y patrones importantes para detectar el fraude en el consumo de agua potable. Se entrenaron y probaron varias técnicas de aprendizaje automático utilizando información histórica sobre consumo fraudulento y no fraudulento utilizando técnicas de descubrimiento de conocimiento de la base de datos KDD.

(Drogodependientes & En, 2018) Utilizaron la base de datos del sistema ERCO operada por el Departamento de Servicio al Usuario y el Módulo de Servicio al Cliente para realizar análisis de extracción de datos e investigaciones sobre los patrones de quejas de los usuarios. Si la herramienta Weka como técnica de predicción y la técnica de agrupamiento para obtener predicciones futuras concentran los datos en ciertas clases preestablecidas, la técnica de Asociación identifica posibles similitudes entre diferentes situaciones que ayudan a establecer. Una técnica de clasificación define un conjunto de clases posibles para diferentes casos. agrupados. Mediante la aplicación del algoritmo Árbol de decisión obtuvieron el 98.84% de precisión, con el algoritmo Naybe Bayes con el 97.273% de precisión y WEKA J48 con el 96.97% de precisión. Los modelos predictivos o patrones están en las diferentes soluciones Confirmar, Revocar y Revocar parcialmente.

(Humaid, 2017) Este trabajo aplica tecnología de minería de datos completos a este ESPACIO basado en el sistema de pago financiero. Para el consumo de agua en la ciudad de Gaza. Técnica seleccionada utilizada en el desarrollo de un modelo de detección de fraude. La eficiencia y precisión del modelo fueron probadas y evaluadas por un método científico y alcanzó una técnica aceptada. El modelo inteligente desarrollado en este estudio de investigación predice y selecciona clientes sospechosos para ser inspeccionado in situ por los equipos del departamento de lucha contra el robo de agua (DWTC) en el municipio de Gaza (MOG) para la detección de actividades fraudulentas. El modelo aumenta la tasa de aciertos de detección de 1-10 % de detección manual aleatoria a 80 % inteligente detección.

Basado en sus datos históricos de uso de agua. La técnica de clasificación de vectores de soporte (SVC) aplicada en este estudio utiliza información del perfil de carga del cliente para mostrar el comportamiento anómalo del perfil de carga del cliente.

(Al-Radaideh & Al-Zoubi, 2018) Las empresas y organismos de agua deben hacer frente a un problema grave: prácticas de consumo de agua fraudulenta. El mayor porcentaje de pérdidas no técnicas se debe a este comportamiento, lo que provoca una importante pérdida de ingresos. En los últimos años, ha habido una ola de investigación que ha desarrollado herramientas efectivas para identificar el fraude. Las empresas de agua pueden reducir las pérdidas al detectar estas actividades fraudulentas con la asistencia de técnicas inteligentes de extracción de datos. Este estudio investiga cómo identificar a los clientes de agua que pueden estar cometiendo fraude utilizando dos técnicas de clasificación (SVM y KNN). La principal fuerza impulsora detrás de esta investigación fue ayudar a Yarmouk Water Company (YWC) en la ciudad jordana de Irbid a superar la pérdida de ganancias. El enfoque basado en SVM utiliza los atributos del perfil de carga del cliente para exponer un comportamiento anormal que se sabe que está relacionado con actividades de pérdida no técnicas. La información fue recopilada a partir de los datos la institución. La precisión del modelo generado alcanzó una tasa de más del 74 %, que es mejor que los procedimientos de predicción manual actuales adoptados por YWC. Para implementar el modelo, se ha construido una herramienta de decisión utilizando el modelo generado. El sistema ayudará a la empresa a predecir los clientes de agua sospechosos que serán inspeccionados en el sitio.

(Rebelo et al., 2022) Buscaron un modelo predictivo de ITSR de mezclas asfálticas utilizando varios parámetros que influyen en la sensibilidad al agua y para clasificar su importancia relativa. Para crear el modelo, se recopilaron y almacenaron en la base de datos 13 parámetros de 160 mezclas asfálticas diferentes. Las técnicas de minería de datos se utilizan para procesar los datos mediante regresión múltiple, redes neuronales artificiales y máquinas de vectores de soporte (SVM). Varias métricas probadas mostraron que SVM era el modelo de predicción de ITSR más preciso (sesgo absoluto medio 0,116, error cuadrático medio 0,150, coeficiente de correlación de Pearson 0,667). Los resultados del análisis de sensibilidad mostraron que el contenido de ligante (26%) fue el factor que más influyó en la sensibilidad al agua de la mezcla asfáltica. Las características de los agregados gruesos y finos (24,9%), ligante asfáltico (19,3%) y el uso de aditivos (10%) influyen simultáneamente en este comportamiento. De acuerdo con los resultados del análisis de sensibilidad, el parámetro que afecta la sensibilidad de la mezcla asfáltica al agua (26%) es el contenido de ligantes. Características del agregado grueso y fino (24,9%), el ligante asfáltico (19,3%) y el uso de aditivos (10%) afectan simultáneamente esta propiedad. Según los resultados de un análisis de sensibilidad, el parámetro que más afecta la sensibilidad al agua de una mezcla asfáltica (26%) es su contenido de ligante. Sin embargo, esta propiedad también está influenciada por otros elementos, incluyendo las características de los agregados gruesos y finos (24,9%), las propiedades del ligante asfáltico (19,3%) y el uso de aditivos (10%).

(Torres-Quezada, 2021). Utilizaron datos alojado de la página oficial (ANT 2020b) con 420 variables correspondientes a las categorías incluidas en una reserva policial determinada: Accidentes en Ecuador en 2020 por error humano con una probabilidad de 65,64 de que un ser humano sea el más influyente.

1.10. Herramienta para minería de datos.

1.10.1. Inteligencia de negocios

La información es crucial para que cada negocio analice, tome decisiones, forje nuevas direcciones y esté listo para actuar rápidamente. En este sentido, la inteligencia de negocios aparece como una estrategia de negocios porque contiene muchos datos y puede ser procesado y analizado. Este incluye métodos, prácticas y aplicaciones destinadas a crear, analizar y administrar información que incluye la identificación de

indicadores clave que permiten medir el progreso organizacional. (Mazón Olivo et al., 2018).

1.10.2. Pentaho Data Integration

Pentaho Data Integration (PDI, también conocido como Kettle) se encarga del proceso de extracción, transformación y carga de datos (ETL) (León, 2015). entre las características:

- Exportar información de la base de datos a un solo archivo.
- Llenar la base de datos con datos.
- Eliminar datos inconsistentes
- Integración de sistema.

1.10.3. Weka

Es una herramienta que permite realizar experimentos de ciencia de datos con cualquier conjunto de datos de usuario y evaluar los métodos de análisis de datos más adecuados, principalmente recolectando datos a través del aprendizaje automático (García & Álvarez, 2020).

1.10.4. Propuesta de minería de datos.

De acuerdo con la información recopilada de investigaciones relacionadas, formaron un modelo que muestra las estrategias que fueron utilizadas por diferentes autores, así como la implementación de algoritmos como en la Fig. 16. Este estudio plantea el uso de información histórica de consumo y reclamos utilizando los datos de los usuarios de agua potable de JAAPYSR-I, construyendo métricas descriptivas y predictivas para obtener información de reclamos de forma aleatoria y relacionando similitudes de cada sector de la parroquia.

Autor	Técnicas				Algoritmos																			Actividades																						Recursos								Conexión			Datos Académicos				Resultados del estudio
	Predictivas		Descriptivas																					Foros				Tareas				Questionarios				Wiki	Pruebas																															
	Clasificación	Regresión	Agrupamiento	Asociación	K-means	Expectation–maximization	Regresion lineal	Regresion multiple	Decision Tree J48	REP Tree	Random Forest	Logistic Model Tree	Hoeffding Decision Trees	Naive Bayes	Support Vector Machine (SVM)	C4.5	K-Nearest-Neighbor (KNN)	SMO	Apriori-SD	Acceso a foros	Tiempos de permanencia	Num palabras y Num Comentarios	Frecuencia de consulta	Tiempo de Entrega	Num. Tareas Enviadas	Total Tareas entregadas	Total Tareas	Num. de Vistas a questionarios	Frecuencia de acceso a questionarios	Intentos de resolver questionario	Questionarios de satisfacción	Num. Vistas	Num. Wikis Editadas	Num. Test iniciados	Num. Test Vistos	Num. Intentos	Num. Tests aprobados	Participaciones	Frecuencia acceso a material didáctico	Frecuencia de acceso a todos los glosarios	Num links vistos	Num Total clicks	Num paginas vistas	Num recursos vistos	Tiempo en contenidos	Num inicios de sesion	Total tiempo online	Max tiempo inactivo	Datos sociodemográficos	Nota Actividades y Examen final	Nota Final del curso	Nota final de los semestres anteriores	
Cerezo et al., 2016			X		X	X															X	X		X																				X							X	X			Determinar la relacion entre los clusteres y notas finales usando ANOVA
Helal et al., 2017				X															X	X				X												X	X				X	X								X				Realacion entre acceso a recursos-actividades y la tasa de éxito
Conijn et al., 2017		X					X															X				X	X			X	X	X	X	X	X			X	X	X		X	X	X				X	X			Prediccion de la nota final
Hasan et al., 2018	X								X	X	X	X	X	X				X																			X		X		X				X					X	X			Definió una categoría dentro de una escala de (Reprobado, Bueno, Promedio, Aprobado y Excelente)
Bharara et al., 2018			X		X															X										X						X	X				X		X				X			X				Determinar variables que afectan el rendimiento academico y agrupacion.
Bravo et al., 2021		X						X														X	X				X	X		X	X	X					X	X						X				X			X	X	Prediccion de la nota final y determinacion de factores de rendimiento academico.
			X		X																	X	X				X	X		X	X	X					X	X						X				X			X	X	Agrupación según comportamientos academicos
Calderon-Valenzuela etal., 2022			X		X															X				X			X										X	X	X										X				Agrupación de docentes asociados al uso de los recursos y actividades
				X															X	X				X			X										X	X	X										X				Identificar asociaciones de uso de las actividades y recursos
Ouatik F. et al,. 2022	X														X	X	X					X																					X				X	X	X			Modelo declasificacion para predecir el exit académico.

Fig. 14. Matriz de Investigaciones relacionadas (Aguagallo Leonardo 2022)

CAPÍTULO II

DESARROLLO

2. Metodología de investigación

En este capítulo se describen los métodos a utilizar y la recolección de información suficiente para las entrevistas como herramienta de procesamiento de la información para recolectar los datos más relevantes, asegurando así la claridad en el desarrollo de los métodos y el análisis final de los resultados a partir de la información tomado de entrevistas.

2.1. Tipo y método de investigación

Este estudio utilizó un enfoque cuantitativo porque uno de los objetivos era encontrar un modelo predictivo basado en datos cuyo comportamiento está determinado por los modelos; cualitativa, con documentos observables disponibles para el análisis; y un resumen basado en los datos obtenidos durante el estudio.

Este también es un estudio histórico que comienza con una investigación de los antecedentes del caso y dependerá del estado de facturación de cada consumidor y de cada reclamo tramitado por la JAAPYSR-I para encontrar una correlación con diversas variables que pueden afectar el resultado. y es un estudio en profundidad porque es un análisis especifico de minería de datos y como herramienta tecnológica en los negocios.

El método utilizado y desarrollado en el estudio es el método de síntesis inductiva, donde inicia un estudio que son los datos que generarán los resultados del atributo reclamo, que a su vez serán abordados de manera general en el reclamo del usuario; el análisis global define parámetros a seguir o comportamiento.

Entrevista estructurada.

Este modelo permite elaborar un guion de entrevista fijo y secuencial con el propósito de obtener información relevante y clara. En este estudio se formularon las preguntas de acuerdo a la jerarquía administrativa (Salazar, 2018)

2.2. Población muestra y muestreo.

Dado a la recolección de información para la revisión, diseño y pruebas son datos históricos del sistema YAKUSOFT, componentes principales para definir la investigación, población y muestra.

2.3. Herramienta para procesamiento de información.

En la Tabla 4 – 5. Se muestra los tipología y miembros claves del proyecto que se les realizó una entrevista previa para conocer el estado actual de la JAAPYSR-I.

Tabla 4. Participantes directos del proyecto

ENTREVISTADO	FUNCIÓN	CÓDIGO
Inti Córdova	Presidente JAAPYSR-I	E1
Luz maría Cáceres	Recaudadora.	E2
Marco Velásquez	Operario	E3

Fuente: Elaboración Propia

Tabla 5. Tipología de entrevista según criterios

CRITERIOS	TIPOLOGÍA DE ENTREVISTA
Según el momento	Inicial De desarrollo Final
Según el grado estructuración	Estructurada Semiestructurada Profundidad
Según el número de participantes	Individual Grupal

Fuente: (Bertomeu, 2018)

2.3.1. Guía de entrevista

Tabla 6. Modalidades de preguntas

TIPOLOGÍA DE PREGUNTAS	E1	E2	E3
Comparar	X	X	X
De hechos pasados	X	X	
De actos pasados	X		
De relación emocional			
De causa - efecto	X	X	X
Información adicional de aspectos, reacciones y eventos	X	X	X
Operaciones Condicionales		X	
Indagaciones	X	X	X

Fuente: (Bertomeu, 2018)

Preguntas para E1:

1. ¿Cuál es el nombre que recibe el procedimiento en reclamos en la JAAPYSR-I?
2. Describa de manera breve el procedimiento anterior.
3. ¿De qué forma el reclamo del usuario no procede a ser atendido?

Preguntas para E2:

1. ¿Cuál es el nombre que recibe el procedimiento en reclamos en la JAAPYSR-I?
2. Describa el procedimiento anterior.
3. ¿Cuáles son los resultados de los procesos que maneja diariamente?
4. Describa la pregunta ¿Cuáles son las causas para ese resultado?
5. ¿Cuál es su opinión al respecto a los casos obtenidos?

Preguntas para E3:

1. ¿Cuál es el tiempo que dura en atender un reclamo negado?
2. ¿Cuál es el porcentaje de reclamos negados y aceptados?
3. ¿Cuántos son los reclamos negados?

2.3.2. Análisis de resultados entrevista.

Para el informante E1, esto indica que esta persona tiene un papel de liderazgo importante y puede ser considerado como el primer filtro de decisión en los requisitos de JAAPYSR-I. Sus usuarios tienen que presentar varias solicitudes de entrada de quejas; por ejemplo: copias de certificados de ciudadanía, cartas a empresas e informes de decisiones del franquiciador u otras agencias. Uno de los puntos que menciona E1 es la verificación de los reclamos de los usuarios, y en este sentido considera importantes los siguientes parámetros:

¿Cuándo procede un reclamo?

- Consumo ilimitado.
- Incremento Extraordinario de consumo en planillas.
- Fugas internas y externas.

¿Cuándo no procede un reclamo?

- Consumo básico, (lectura actual – lectura anterior = consumo mensual m3).

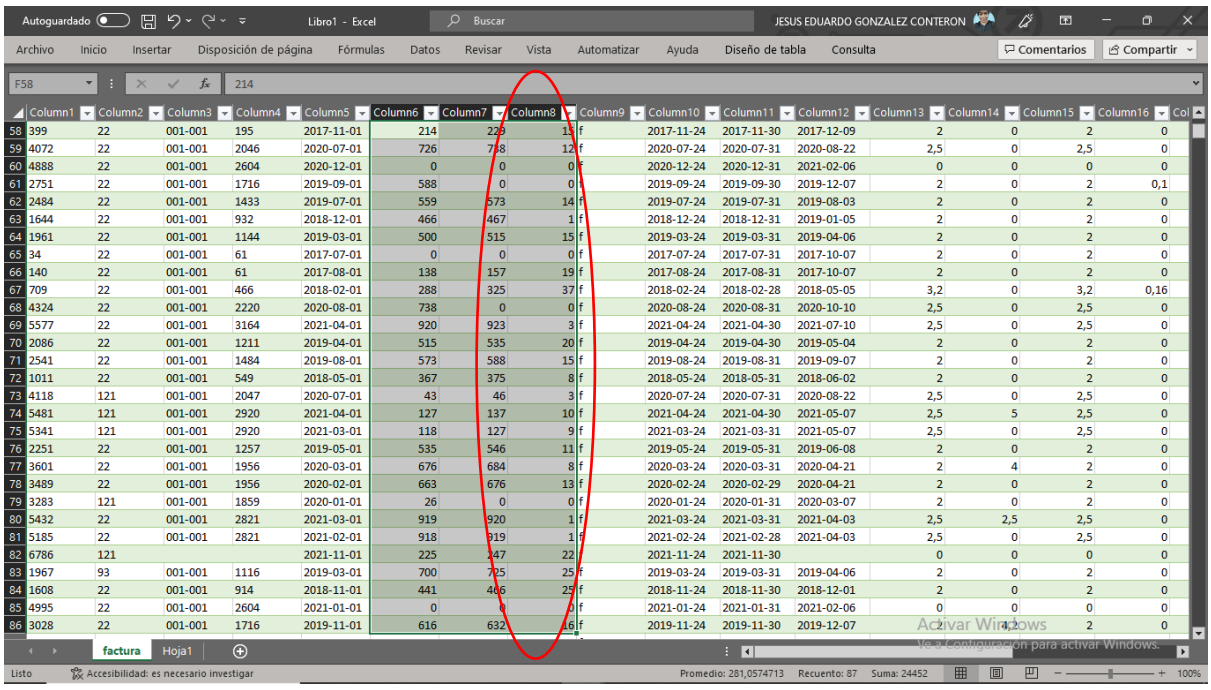


Fig. 15. Calculo para consumo básico

En las entrevistas con el informante E2 se mostró importante su rol en el proceso de derivación y respuesta para resolver conflictos entre los usuarios y la JAAPYSR-I. La decisión es una respuesta al tercer reclamo JAAPYSR-I presentado por el usuario,

y luego de ingresar el reclamo presentado por E1 en YAKUSOFT, ingresa a la etapa de análisis de acuerdo a los respectivos reclamos. La tarea de E2 es comprobar si el franquiciado ha infringido las normas de servicio interno y, en caso afirmativo, la reclamación es inmediatamente a favor del usuario.

Cuando hay facturas excesivas.

- lecturas y consumos
- Promedio de consumo

Hay un aumento anormal en el consumo de "IEC", es decir, la lectura mensual supera el promedio de consumo de los usuarios. Para ello, se aplica el proceso de crítica.

Confirmar con conciliación: El reclamo fue denegado según las reglas, pero el usuario recibió una encuesta para corregir la fuga con JAAPYSR-I.

Cabe señalar que actualmente en San Juan de Ilumán existe un alto porcentaje de casos rechazados en atención al cliente; a menudo, las solicitudes de los usuarios no se cumplen porque los usuarios no conocen estos procedimientos y el público no conoce los servicios de agua potable. Cuidado y buen uso, mantenimiento de las instalaciones internas, programa de divulgación mensual en busca de sanitarios y cañerías antiguas en sus locales, etc. Entre el 80% y el 100% de los casos rechazados llegan al servicio de atención al cliente todos los meses.

Las entrevistas con el informante E3 fueron clave para la aprobación de la decisión. El reglamento interno de la Gerencia del Servicio en materia de denuncias establece un plazo máximo de 35 días hábiles para apelar las resoluciones administrativas. Entre enero de 2017 y noviembre de 2017, el 54% de los casos recibidos por la JAAPYSR-I fueron rechazados y el 43% aprobados.

Es bien sabido que existe una falta de comprensión de las normas internas que rigen los servicios, incluidos las normas de los usuarios y operadores. Las reclamaciones rechazadas superan el 50% mensual porque muchas veces no son aplicables por fugas internas invisibles de agua y falta de cultura de mantenimiento preventivo en el domicilio del usuario.

2.4. Aplicación de metodología KDD.

Entregables del Proyecto

En la Tabla 7. Se clasifica en diferentes actividades de acuerdo a cada fase de la metodología KDD, los cuales se modificarán constantemente hasta conseguir modelos eficientes y de calidad.

Tabla 7. Entregables del proyecto

ENTREGABLE	FUENTE
Calidad de datos Data warehouse Vista minable Modelo Interpretación de Conocimiento	ISO/IEC 25012 Herramienta (PDI) Herramienta Weka Predictivo y Descriptivo Clasificación y agrupamiento

Fuente: Elaboración Propia

2.5. Organización de directores implicados

En la TABLA 8. Se designa a participantes de cada área.

Tabla 8. Directores de las áreas comprendidas

DEPENDENCIA	PARTICIPANTE	FUNCIÓN
COSOFT	MSc. Cosme Macarthur Ortega	Designar al director.
Departamento informático	Presidente Junta de Agua	Designar encargado en Bases de Datos y sistema YAKUSOFT
Dirección general de JAAPYSR-I	Sra. Diana de la Torre	Designar operario de junta de agua.

Fuente: Elaboración Propia

En la TABLA 9. Participantes del proyecto

Tabla 9. Participantes directores del proyecto

ROL	DEPENDENCIA	NOMBRE
Director	CISIC	PhD. Iván García
Administrador de bases de Datos	Junta administradora y saneamiento san juan de Ilumán.	Sr. Tauri Montalvo
Analista de Data Mining	CISIC	Sr. Gonzales Jesús

Fuente: Elaboración Propia

Roles y Responsabilidades

Se muestran en la TABLA 10. A directores de proyectos, administradores de bases de datos y analistas de sistemas YAKUSOFT. Con sus respectivos roles y responsabilidades.

Tabla 10. Roles y responsabilidades

ROL	RESPONSABILIDAD
Director	Es responsable en la toma de decisiones que tiendan al cumplimiento de los objetivos.
Administrador de bases de datos	Designa la administración de la base de datos JAAPYSR-I de las tablasConsumo, Parámetro, Tarifa, Contribuyente, Factura, Medidor, Sector, Reclamos.
Analista de Datos	Analizar los valores entregados por el administrador de base de datos de la junta de agua San Juan de Ilumán

Fuente: Elaboración Propia

2.6. Costos de proyecto

Estimaciones

En las Tablas 11–12. Se detallan el presupuesto y recursos estimados. Considerando que el costo es por el número de horas y recursos empleados.

Tabla 11. Talento humano del proyecto

DESCRIPCIÓN	N. DE HORAS	COSTO POR HORA ($)	COSTO TOTAL ($)
Horas de investigación	240	25.00	600.00
Horas de desarrollo	230	25.00	300.00
		TOTAL	900.00

Fuente: Elaboración Propia

Tabla 12. Recursos materiales del proyecto

DESCRIPCIÓN	COSTO REAL ($)	COSTO ACTUAL ($)
Hardware		
Laptop	720.00	00.00
Copiadora	245.00	00.00
Software		
Paquetes Office	00.00	00.00

Mendeley	00.00	00.00
Pentaho Data Integration (PDI)	00.00	00.00
Ultima versión Weka	00.00	00.00
Materiales de Oficina		
Epson EcoTank	50.00	50.00
Papel bon A4	03.50	03.50
Bolígrafos	01.00	01.00
CNT Internet	160.00	120.00
Memoria flash 16 GB	15.00	15.00
Investigación		
ISO IEC 25012:2008	80.00	00.00
TOTAL	1274.50	189.50

Fuente: Elaboración Propia

Tabla 13. Valor total del proyecto

DESCRIPCIÓN	COSTO ($)
Talento humano	900.00
Recursos materiales	1274.50
TOTAL	2174.50

Fuente: Elaboración Propia

2.6.1. Tiempo del proyecto

En la Tabla 14. Muestra el número de horas empleada a cada fase cumpliendo un número total de tiempo empleado en el proyecto.

Tabla 14. Horas implementadas al proyecto

FASE	DURACIÓN (HORAS)
Fase de recopilación de datos históricos YAKUSOFT	30
Aplicación de normas (ISO/IEC 25012:2008) en base de datos obtenido.	10
Fase de datos (selección, limpieza y transformación)	30
Fase de (Data Mining) y Data Warehouse	30
Fase de (Evaluación e Interpretación) de datos similares	50
Documentación del documento de investigación	170
Análisis de Resultados en herramienta Weka	50
Análisis de Impactos (estadística)	40
TOTAL	410

Fuente: Elaboración Propia

2.7. Proceso de integración y recopilación

Tipos de datos base

- **Sistema YAKUSOFT**

Los datos de consumo y reclamos de cada usuario de JAAPYSR-I se obtuvieron y recopilaron de la base de datos PostgreSQL y del sistema de recolección IAKUSOFT, de la junta de aguas de San Juan de Ilumán indicadas en las Fig. 18-19.

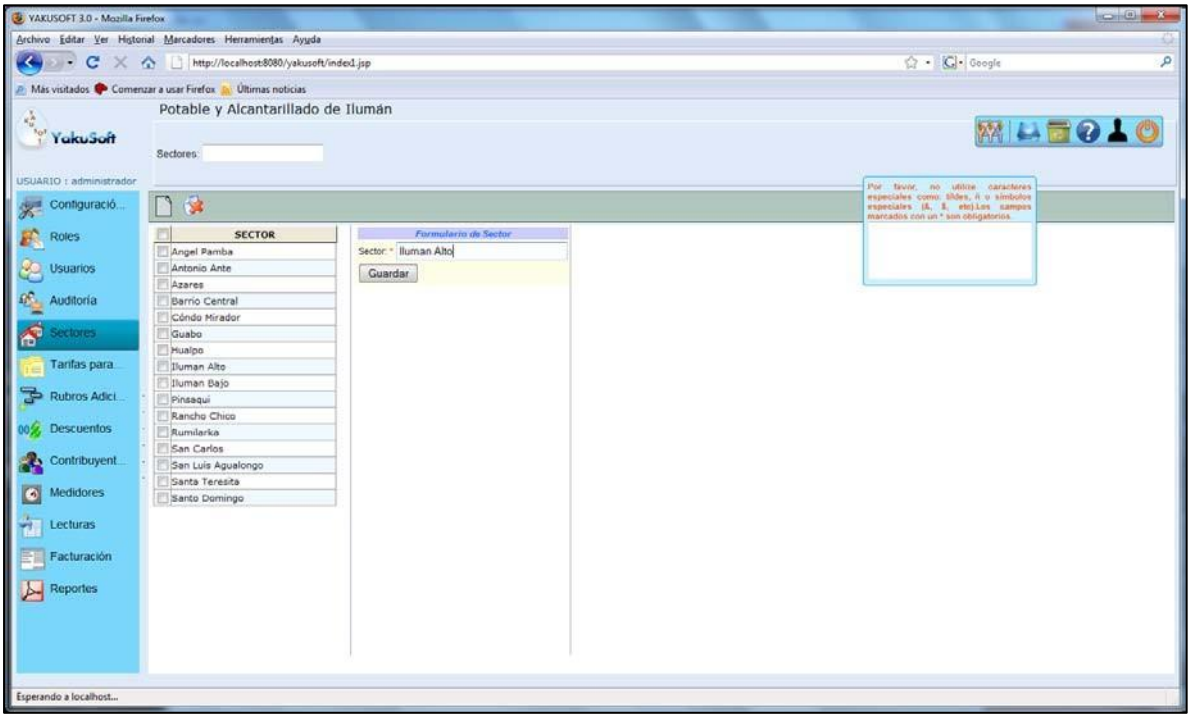


Fig. 16. Sistema YAKUSOFT (JAAPYSR-I).

- **Base de datos YAKUSOFT en PostgreSQL**

Se emplearon datos relevantes de la base de datos YAKUSOFT, con 22 tablas que corresponden a los registros de cada usuario: tbl_archivo, tbl_auditoria, tbl_configuracion, tbl_consumo, tbl_contribuyente, tbl_factura, tbl_factura_descuento, tbl_descuento_medidor tbl_factura_rubro, tbl_lectura_macro, tbl_rubro_medidor tbl_usuario, tbl_sector, tbl_tarifa tbl_parametro, tbl_descuento, tbl_medidor, tbl_medidor_novedad, tbl_pagina, tbl_privilegio, tbl_rol, tbl_rubro. Los cuales contienen valores numéricos, boléanos y cadena de caracteres, Fig. 19.

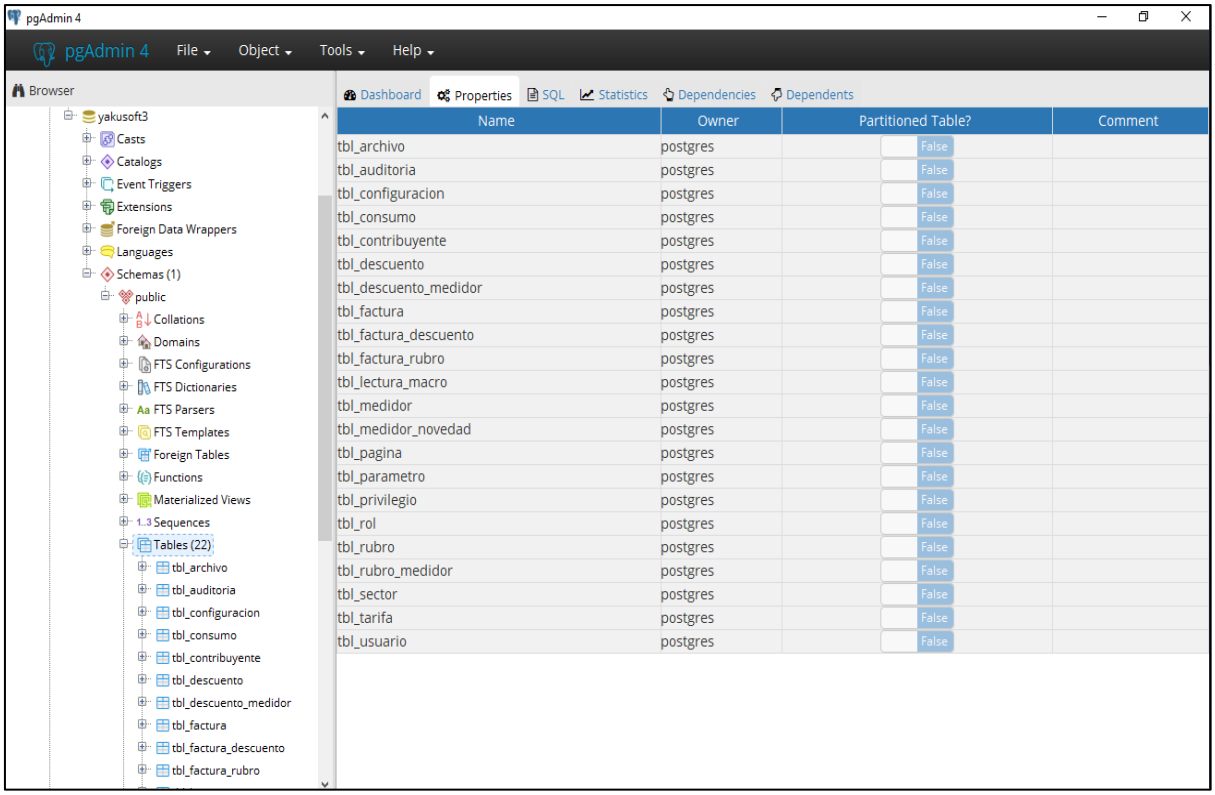


Fig. 17. Base de datos YAKUSOFT (JAAPYSR-I).

Los datos encontrados incluyen tipo de fecha, tipo entero, tipo numérico y tipo de cadena, que se dividen en dos tipos de datos: numéricos y discretos. A continuación, en las Tablas 12 –Tabla 19. Muestran los valores y atributos para su análisis.

- **Tipos de datos de la tabla Consumo**

Tabla 15. Categorización atributo Consumo

ATRIBUTO	TIPO DE DATO
ID_CONSUMO	Entero Grande
ID_PARAMETRO	Entero Grande
DESDE_M3	Numérico
HASTA_M3	Numérico
VALOR	Numérico

Fuente: BDD-YAKUSOFT3

- **Tipo de datos tabla contribuyente.**

Tabla 16. Categorización atributo contribuyente

ATRIBUTO	TIPO DE DATO
ID_CONTRIBUYENTE	Entero grande
CI_RUC	Entero grande
APELLIDO	Ristra de caracteres
NOMBRE	Ristra de caracteres
FECHA_NAC	Fecha
GENERO	Boleano
DIRECCION	Ristra de caracteres
ID_SECTOR	Entero grande
TIPO	Ristra de caracteres
ELIMINADO	Boleano

Fuente: BDD-YAKUSOFT3

- **Tipo de tabla Parámetros.**

Tabla 17. Categorización atributo parámetro

ATRIBUTO	TIPO (DATO)
ID_PARAMETRO	Entero Grande
ID_TARIFA	Entero Grande
FECHA_VIGENCIA_INI	Fecha
FECHA_VIGENCIA_FIN	Fecha
GASTO_ADMIN	Numérico
ALCANTARILLADO_VALOR	Numérico
ALCANTARILLAO_PORCENTAJE	Boleano
ALCANTARILLADO_DE	Cadena de caracteres
MULTA_MORA_VALOR	Numérico
ES_PORECENTAJE	Boleano
MULTA_MORA_DE	Cadena de caracteres
RECONEXION	Numérico
IVA	Numérico
LIMITE_DESCUENTO	Numérico
ELIMINADO	Boleano

Fuente: BDD-YAKUSOFT3

- **Tipo de tabla Factura.**

Tabla 18. Categorización atributo factura

ATRIBUTO	TIPO DE DATO
ID_FACTURA	Entero grande
ID_MEDIDOR	Entero grande
SERIE_FACTURA	Cadena de Caracteres
NUM_FACTURA	Entero grande
FECHA_LECTURA	Fecha
LECTURA_ANTERIOR	Numérico
LECTURA_ACTUAL	Numérico
CONSUMO	Numérico
ULTIMA	Boleano
MES_COBRO	Fecha
FECHA_PREFACTURA	Fecha
FECHA_EMISION	Fecha
TOTAL_CONSUMO	Numérico
MESES_IMPAGOS	Numérico
SUBTOTAL	Numérico
RECAR_MULTA	Numérico
RECAR_RECONEXION	Numérico
GASTO_ADMIN	Numérico
ALCANTARILLADO	Numérico
IVA	Numérico
DESCUENTOS	Numérico
DESC_3_EDAD	Numérico
OTROS_RUBROS	Numérico
TOTAL_PAGAR	Numérico
ES_FACT_IMPAGO	Boleano

Fuente: BDD-YAKUSOFT

- **Tipo de datos tabla Medidor**

Tabla 19. Categorización atributo medidor

ATRIBUTO	TIPO DE DATO
ID_MEDIDOR	Entero grande
ID_CONTRIBUYENTE	Entero grande
ID_TARIFA	Entero grande
CODIGO_MEDIDO	Cadena de caracteres
MARCA	Cadena de caracteres
MODELO	Cadena de caracteres
FECHA_CREACION	Fecha
DIRECCION_MEDIDOR	Cadena de caracteres

ID_SECTOR	Entero grande
MACRO	Boleano
ACTIVO	Boleano
ELIMINADO	Boleano

Fuente: BDD-YAKUSOFT3

- **Tipo de datos tabla Sector**

Tabla 20. Categorización atributo sector

ATRIBUTO	TIPO DE DATO
ID_SECTOR	Entero
SECTOR	Cadena de caracteres
ELIMINADO	Boleano

Fuente: BDD-YAKUSOFT3

- **Tipo de datos tabla Tarifa.**

Tabla 21. Categorización atributo tarifa

ATRIBUTO	TIPO DE DATO
ID_TARIFA	Entero grande
TARIFA	Cadena de caracteres
ELIMINADO	Boleano

Fuente: BDD-YAKUSOFT3

- **Tipo de datos tabla Reclamos.**

Tabla 22. Categorización atributo reclamo

ATRIBUTO	TIPO DE DATO
ID_MEDIDOR_NOVEDAD	Entero
ID_MEDIDOR	Entero grande
FECHA_REPORTE	Fecha
FECHA_SOLUCION	Fecha
NOVEDAD	Texto
OBSERVACION	Texto
RESUELTA	Boleano

Fuente: BDD-YAKUSOFT3

2.7.1. Normas de calidad con ISO/IEC 25012.

Se aplicaron diferentes modelos de calidad con respecto a las características de los datos obtenidos en la fase de extracción como se detalla en la Tabla 23.

Tabla 23. Valoraciones de calidad

Escala	Interpretación
Porcentaje % <= 10%	Datos que no pertenece a las reglas determinadas
Porcentaje % > 10% & Porcentaje % <=45%).	Pertenecen a las reglas, pero existen varios errores
Porcentaje % >45% & Porcentaje % <=85%	Muchos forman las reglasdefinidas
Porcentaje % Obtenido >85%	La integridad de los datos tiene ciertas reglas.

Fuente: adaptación de Calabrese et al., 2019

De acuerdo con la base de datos, el valor de los datos de cada atributo se considera para la evaluación y los resultados se muestran en la Tabla 23.

Tabla 24. Indicadores de calidad

ATRIBUTO	DATOS INCONSISTENTEMENTE	VALOR DE MEDICIÓN	PORCENTAJE%
CONSUMO	0	1	100,00
CONTRIBUYENTE	10	0,999902598	99,99
PARAMETROS	0	1	100,00
FACTURA	0	1	100,00
MEDIDOR	0	1	100,00
SECTOR	0	1	100,00
TARIFA	1343	0,962918952	96,29
RECLAMO	1233	0,987327963	98,73
ZONA	0	1	100,00

Fuente: Elaboración Propia

En la tabla, se puede ver las características consistentes de cada característica se encuentran en el rango de evaluación más alto, lo que significa que la información del conjunto de datos representa el 97,96% de todas las variables en promedio.

Construcción del Data Warehouse

Se ha construido un almacén de datos (Data Warehouse). utilizando el software (PDI), y aquí se completó la fase de extracción y transformación de datos.

- **Dimensión SECTORES**

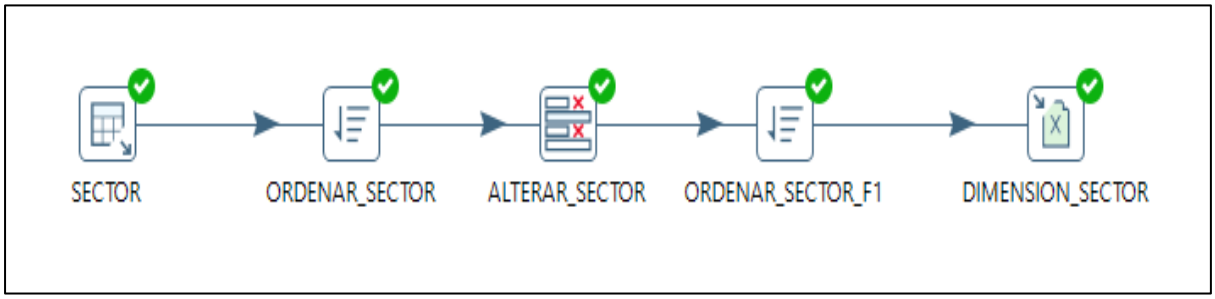


Fig. 18. Dimensión SECTORES Ilumán

- **Dimensión RECLAMOS**

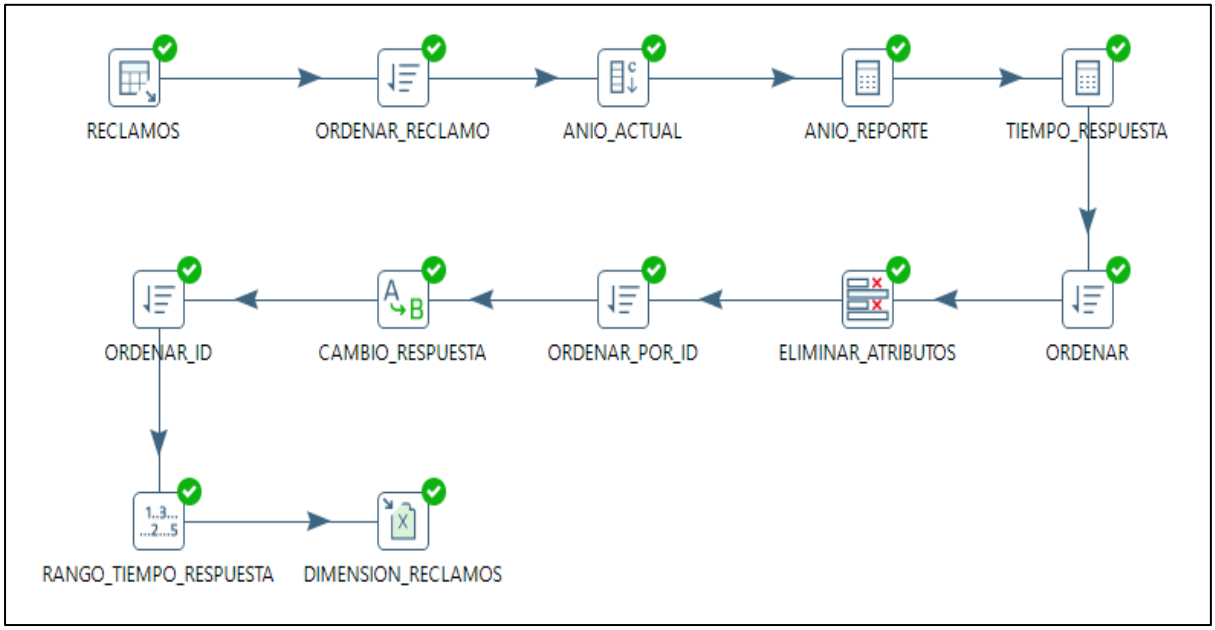


Fig. 19. Dimensión RECLAMO Ilumán

- **Dimensión MEDIDOR.**

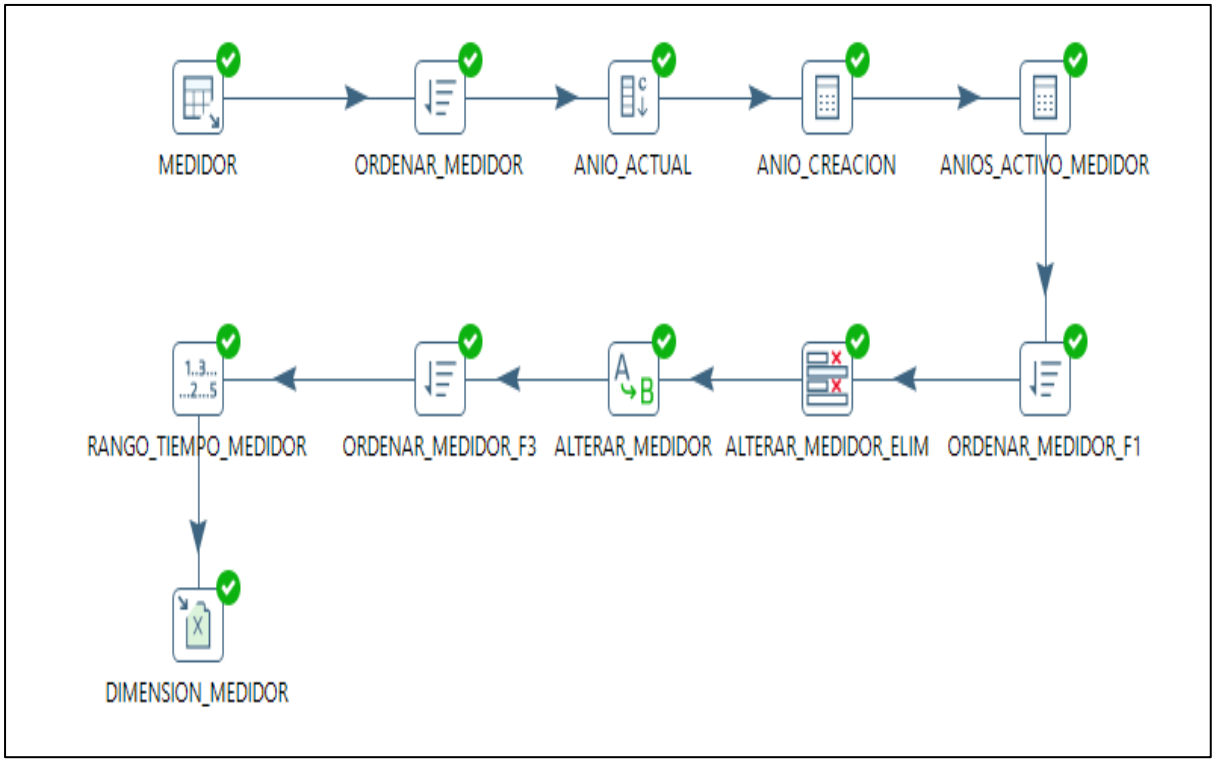


Fig. 20. Dimensión MEDIDOR Ilumán

- **Dimensión CONTRIBUYENTE**

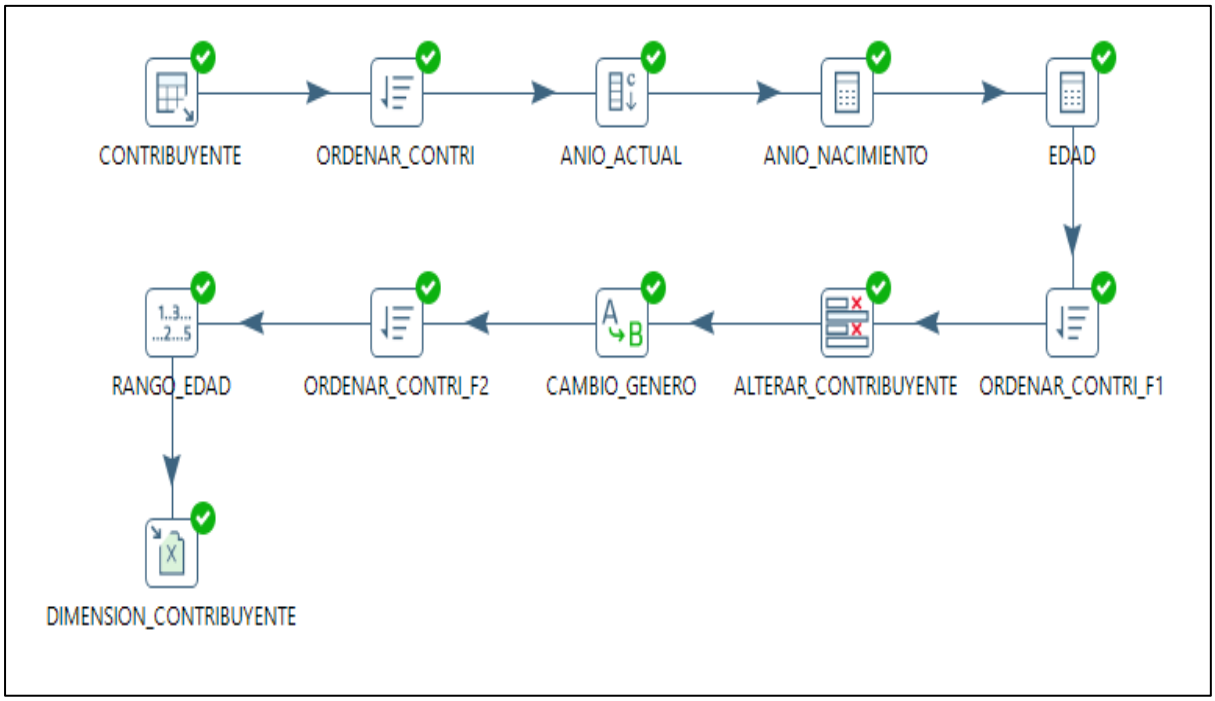


Fig. 21. Dimensión CONTRIBUYENTE Ilumán

- **Dimensión FACTURA**

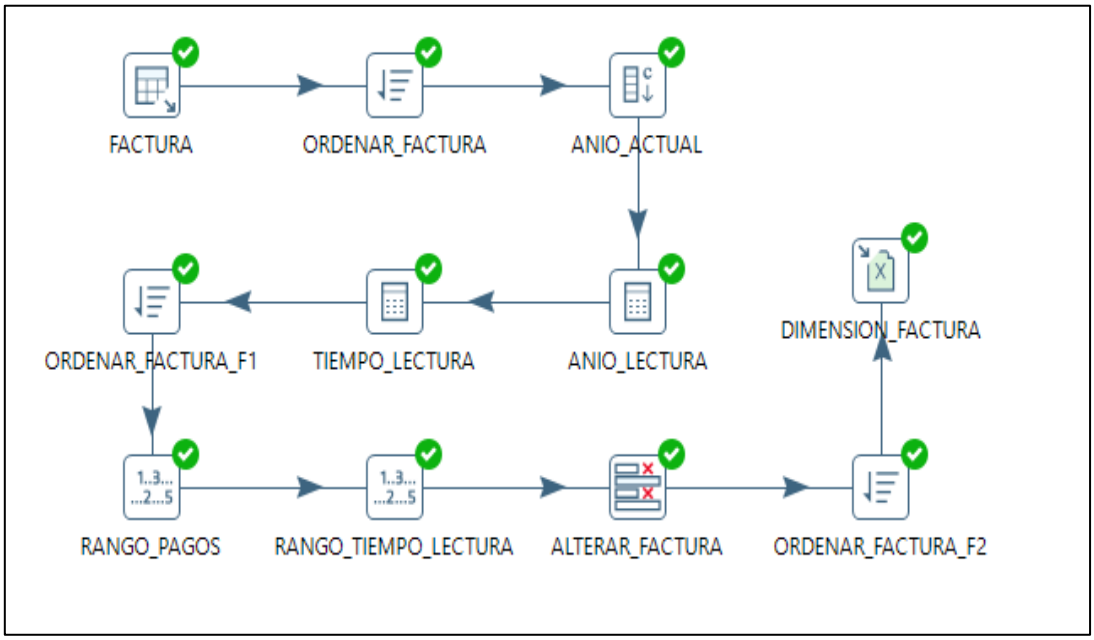


Fig. 22. Dimensión FACTURAS Ilumán

- **Dimensión DATA_WAREHOUSE**

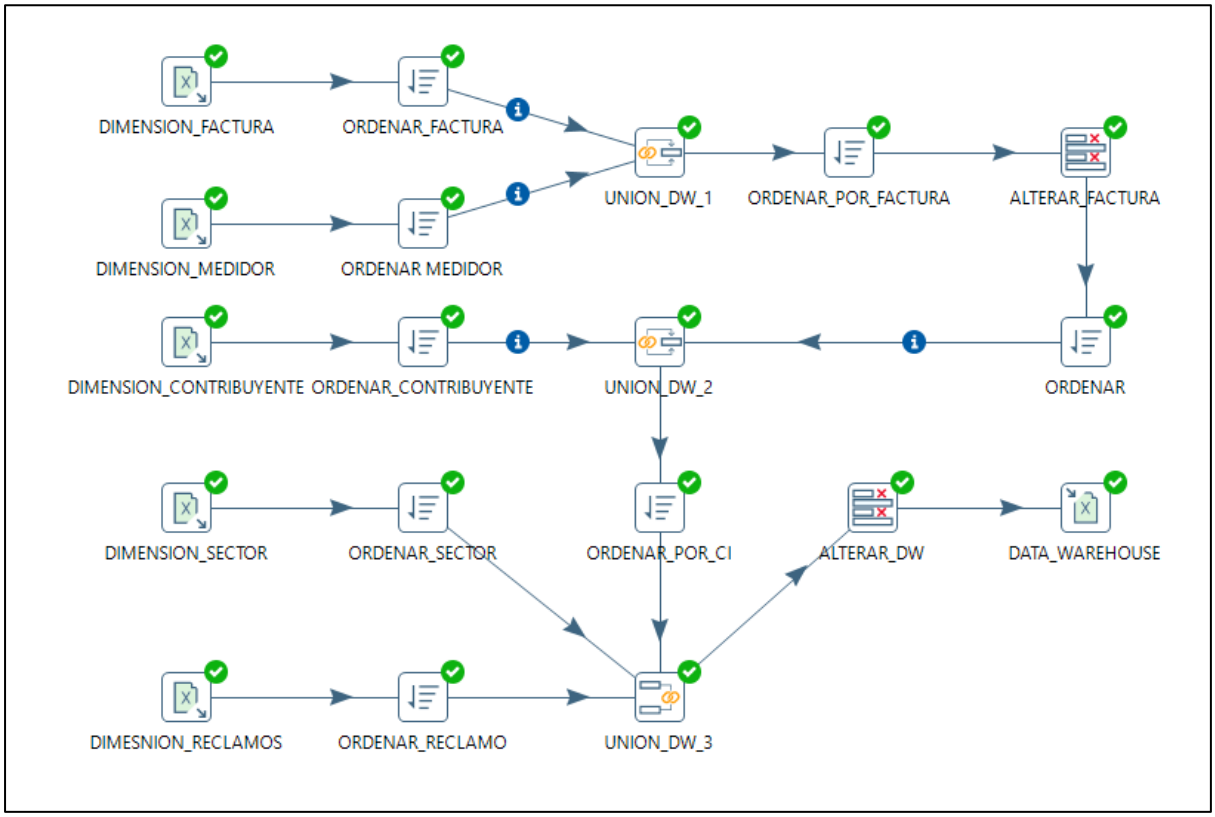


Fig. 23. Dimensión DATA WAREHOUSE Ilumán

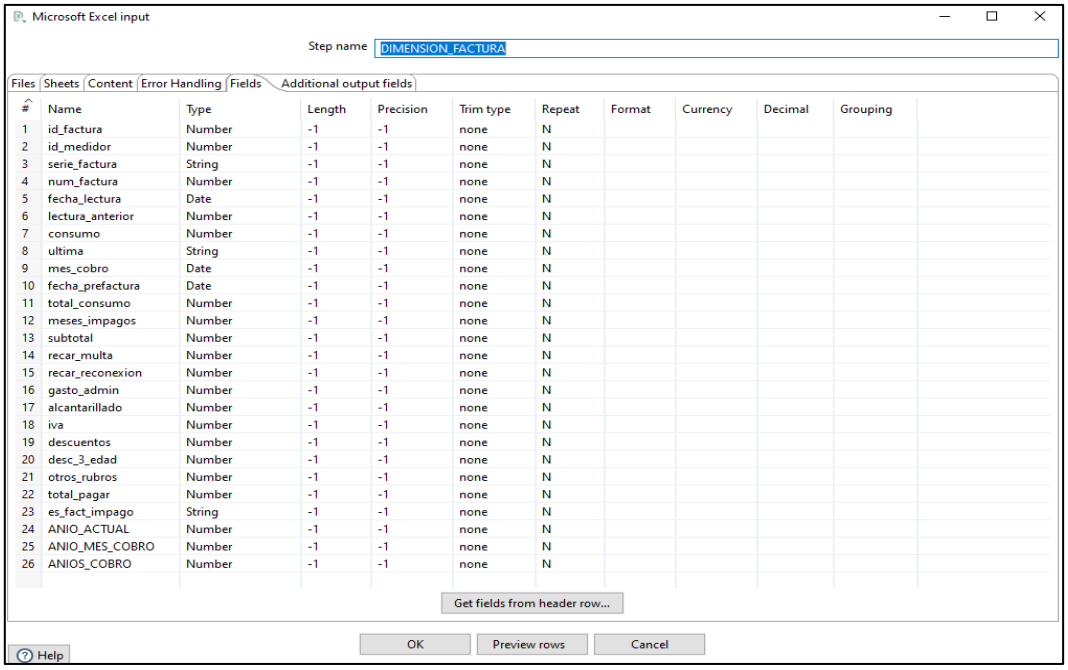
2.8. Proceso de Selección, limpieza y Transformación

Al tener construida la data warehouse nos permite seleccionar, transformar y hacer la limpieza de datos con la herramienta PDI.

a. Selección

- **Filtrado de atributos**

El IVA se eliminará ya que no utilizan después de la selección quedaron algunos atributos esenciales para cumplir el objetivo Fig. 26.



#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	id_factura	Number	-1	-1	none	N				
2	id_medidor	Number	-1	-1	none	N				
3	serie_factura	String	-1	-1	none	N				
4	num_factura	Number	-1	-1	none	N				
5	fecha_lectura	Date	-1	-1	none	N				
6	lectura_anterior	Number	-1	-1	none	N				
7	consumo	Number	-1	-1	none	N				
8	ultima	String	-1	-1	none	N				
9	mes_cobro	Date	-1	-1	none	N				
10	fecha_prefactura	Date	-1	-1	none	N				
11	total_consumo	Number	-1	-1	none	N				
12	meses_impagos	Number	-1	-1	none	N				
13	subtotal	Number	-1	-1	none	N				
14	recar_multa	Number	-1	-1	none	N				
15	recar_reconexion	Number	-1	-1	none	N				
16	gasto_admin	Number	-1	-1	none	N				
17	alcantarillado	Number	-1	-1	none	N				
18	iva	Number	-1	-1	none	N				
19	descuentos	Number	-1	-1	none	N				
20	desc_3_edad	Number	-1	-1	none	N				
21	otros_rubros	Number	-1	-1	none	N				
22	total_pagar	Number	-1	-1	none	N				
23	es_fact_impago	String	-1	-1	none	N				
24	ANIO_ACTUAL	Number	-1	-1	none	N				
25	ANIO_MES_COBRO	Number	-1	-1	none	N				
26	ANIOS_COBRO	Number	-1	-1	none	N				

Fig. 24. Fase de selección
Fuente: Software PDI

- **Filtrado de registros**

Se eliminan los registros que no tienen fecha establecidas ya que el análisis se lo realiza dese el año 2017, adema de sus campos vacíos.

b. Transformación

Para la transformación se tomaron en cuenta criterios específicos:

- **Clase UBICACIÓN.**

En la Tabla 25. Se muestra cómo se clasificó por medio de UBICACIÓN, ya que existen sectores que mediante una investigación de campo se encuentran tres tanques que son almacenadas de diferentes fuentes cercanas de la parroquia.

Tabla 25. Categorización atributo sector

CATEGORÍA	VALORES
SECTOR	NORTE (N) SUR (S) ESTE (E) OESTE (O)

Fuente: BDD-YAKUSOFT3

- **Clase GÉNERO**

Para categorizar la tabla GÉNERO, se tomó los datos del sistema YAKUSOFT, que muestra en la Tabla 26.

Tabla 26. Categorización atributo género

CATEGORÍA	VALORES
MASCULINO	M
FEMENINO	F

Fuente: Propia

- **Clase EDAD**

Se tomó la clase EDAD mediante una operación en la herramienta PDI, del atributo fecha_nacimiento de la tabla tbl_contribuyente, como indica en la Fig. 27. Para categorizar edades y conocer qué tipo de usuarios acuden a la junta de agua a realizar solicitudes de quejas como muestra la Tabla 27.

Fig. 25. Cálculo de edad contribuyente

Tabla 27. Categorización atributo edad

CATEGORÍA	VALORES
ADULTO JOVEN	18 a 25
ADULTO	26 a 39
MAYOR	40

Fuente: Propia

- **Clase CONSUMO**

Mediante la clase CONSUMO se realizó mediante el nivel de consumo medido en m3 como se muestra en la Tabla 28.

Tabla 28. Categorización atributo consumo

CATEGORÍA	VALORES
BAJA	120 m3 a 220 m3
MEDIA	230 m3 a 300 m3
ALTA	301 en adelante

Fuente: Propia

c. Limpieza

Se encontraron datos erróneos en la fase de transformación. detectando en el atributo EDAD_CONTRIBUYENTE, secciones nulas las cuales se eliminan con el propósito de obtener datos limpios para su implementación de algoritmos de clasificación y agrupamiento que corresponden al proceso de obtener un conocimiento.

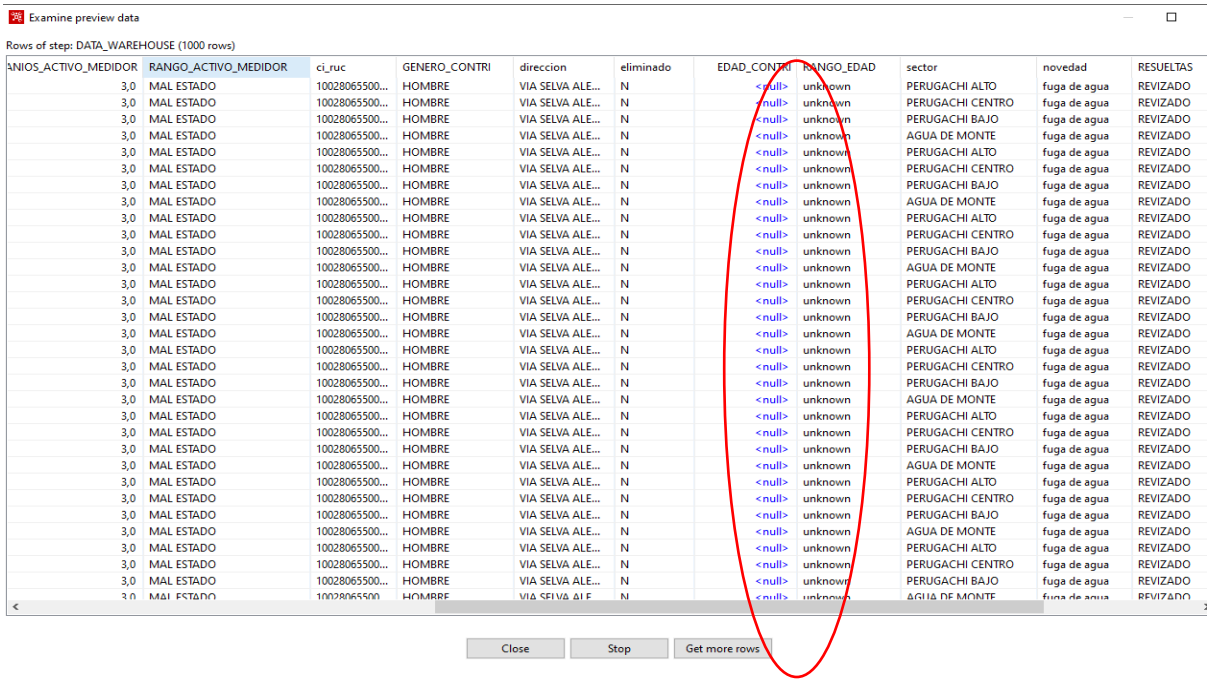


Fig. 26. Eliminación de datos inconsistentes

Se exportó los datos en fromato CSV y un modelado cuantitativo que permitiran ser entiendibles en la herramienta Weka para su análisis.

```
421;54;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;Y
500;83;31/3/2018 0:00;2018/05/05 00:00:00.000;PARCIAL;;N
559;131;31/3/2018 0:00;2018/08/04 00:00:00.000;FUGA INTERNA;;N
626;18;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;Y
717;24;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;N
878;76;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;N
925;1;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
1.049,00;152;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;Y
1.083,00;6;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;N
1.251,00;65;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;N
1.340,00;147;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;Y
1.460,00;32;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;Y
1.475,00;18;31/3/2018 0:00;2018/05/05 00:00:00.000;FUGA INTERNA;;N
1.614,00;40;31/3/2018 0:00;2018/05/05 00:00:00.000;FUGA EXTERNA;;Y
1.671,00;3;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;Y
1.788,00;159;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;N
1.793,00;8;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;Y
1.952,00;65;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;Y
1.965,00;43;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;N
2.105,00;148;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
2.137,00;192;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;Y
2.335,00;23;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;N
2.396,00;64;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
2.458,00;118;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;Y
2.558,00;3;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;Y
2.661,00;68;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
2.752,00;27;31/3/2018 0:00;2018/05/05 00:00:00.000;PARCIAL;;Y
2.768,00;53;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;Y
2.952,00;34;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
3.022,00;42;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;Y
3.042,00;67;31/3/2018 0:00;2018/06/02 00:00:00.000;FUGA INTERNA;;Y
3.146,00;45;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
3.267,00;174;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;N
3.379,00;12;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;Y
3.494,00;32;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
3.547,00;69;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
3.616,00;132;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;Y
3.688,00;145;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;Y
3.780,00;33;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA EXTERNA;;N
3.902,00;35;31/3/2018 0:00;2018/05/05 00:00:00.000;PARCIAL;;N
3.985,00;117;31/3/2018 0:00;2018/04/07 00:00:00.000;FUGA INTERNA;;Y
4.037,00;185;31/3/2018 0:00;2018/05/05 00:00:00.000;FUGA EXTERNA;;Y
4.170,00;44;31/3/2018 0:00;2018/04/07 00:00:00.000;PARCIAL;;N
```

Fig. 27. Formato CSV

2.8.1. Evaluación y selección de los algoritmos

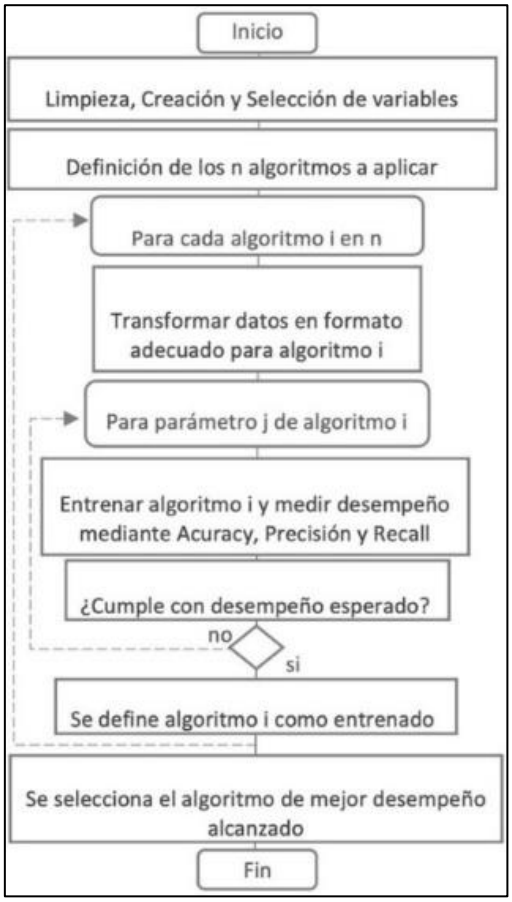


Fig. 28. Evaluación y selección de algoritmos (Lara 2013)

2.8.2. Tareas de clasificación de datos

Se utilizó el algoritmo RandomTree y RandomForest por su mayor valoración con respecto a las otras, y principalmente de acuerdo con la vista minable obtenida se evidenció atributos como: fecha_reporte y fecha_solucion para posteriormente crear la variable TIEMPO_RESPUESTA, de la fase de transformación, mediante el estatuto de la junta de agua muestra que antes de los 28 días pasa a ser un reclamo atendido y posterior a la fecha pasa a ser un reclamo no atendido lo cual se convierte en un modelo aplicable de este algoritmo.

Adicional el modelo cuenta con un atributo en el cual nos muestra el consumo en m3 y el costo de acuerdo con las lecturas del medidor en cada mes lo cual muestra la facturación y mediante la herramienta de Pentaho Data Integration se hizo el cálculo para identificar el incremento de pago en sus planillas cada mes, donde da inicio al

implementar el algoritmo Random Forest que nos permite encontrar ramificaciones que muestran los pagos y considerar el alto consumo y filtrado del agua.

2.8.3. Tareas de agrupamiento de datos

- **Algoritmo K-means**

Al implementar este algoritmo muestra de forma entendible a la institución dueña de los datos, facilitando en la toma de decisiones a futuro con nuevas propuestas de negocio. Aplicando técnicas de agrupamiento se buscó similitudes en las variables del atributo SECTOR con número de nodos 6 que hace referencia a los números de barrios de san Juan de Ilumán (Alexeis et al., n.d.).

- **Expectativa-Maximización EM**

Se aplicó para conocer las similitudes entre los diferentes clústeres asociados al consumo y reclamo de los usuarios formados por la validación cruzada y número de clústeres utilizados de acuerdo a los sectores donde existen mayores irregularidades (Huang & Chen, 2017).

CAPÍTULO III

RESULTADOS

3. Evaluación e interpretación.

3.1. Fase de interpretación de datos

Aplicando modelos cualitativos y cuantitativos se lograron analizar los algoritmos predictivos y descriptivos, basadas en matriz de confusión.

a. Tareas de Clasificación

- **Arboles de decisión.**

Random Tree

En la siguiente Fig. 31 – 32. Se muestran los algoritmos RandomTree con validación Cruzada 10 y 16 atributos conteniendo 13.994 instancias.

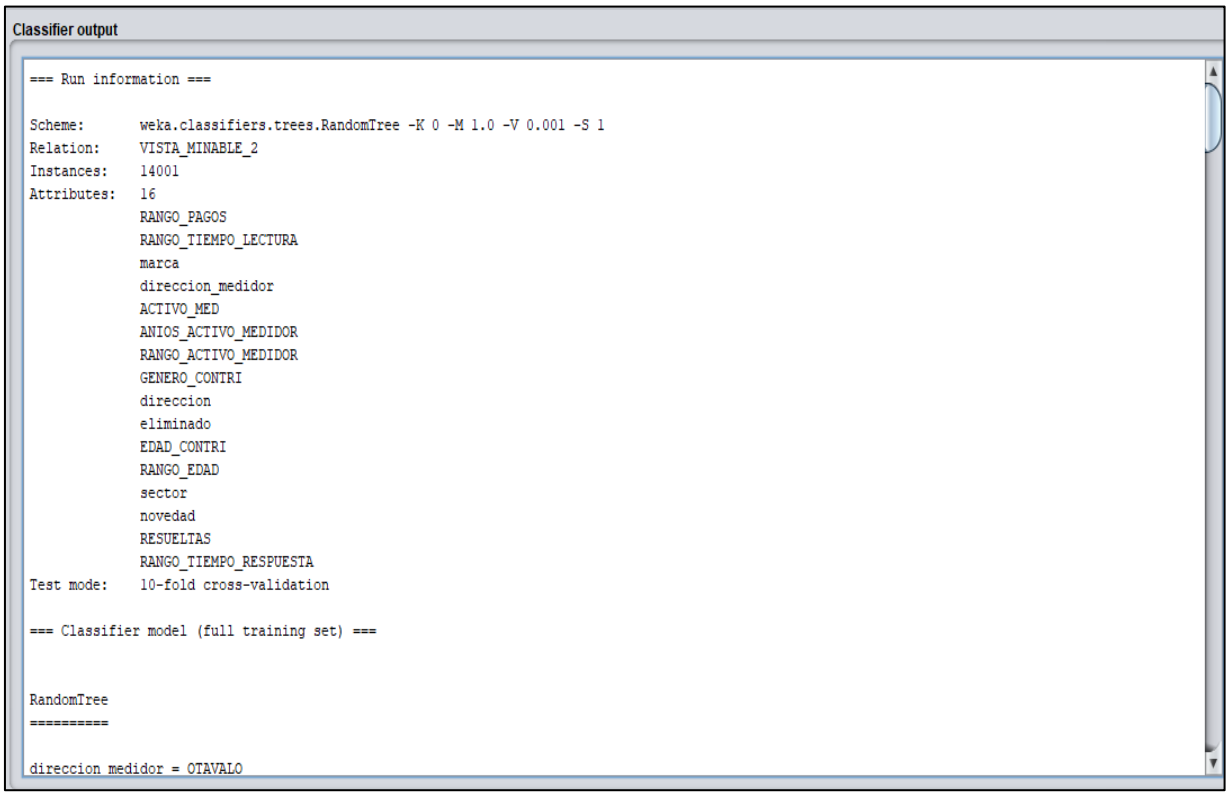


Fig. 29. Ejecución del algoritmo RandomTree (Parte 1)

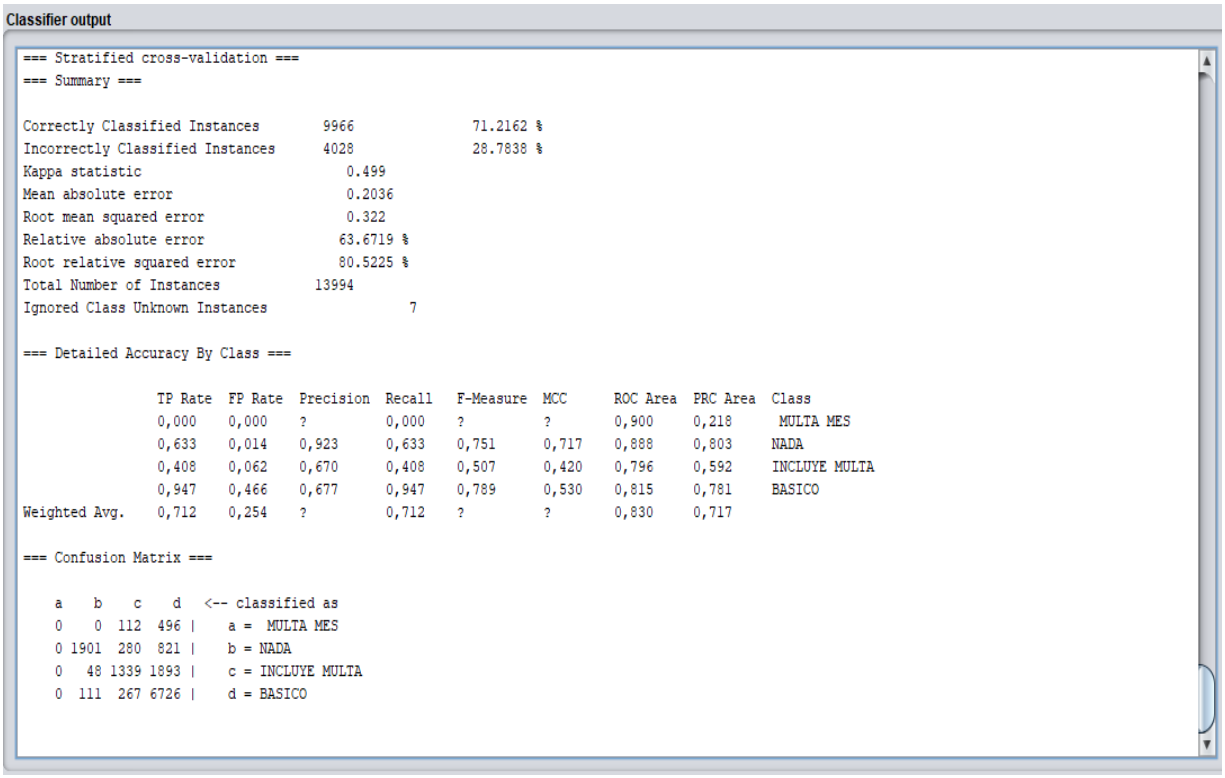


Fig. 30. Ejecución del algoritmo RandomTree (Parte 2)

En la Tabla 29. Muestra la matriz de confusión obtenida tras ejecutar el algoritmo RandomTree con 16 13994 atributos y 10 iteraciones de validación cruzada.

Tabla 29. Resultado del algoritmo RandomTree

		Clase predicha	
		R	NR
Clase verdadera	R	1140	356
	NR	670	136

Fuente: Resultados Weka

Donde:

- TP = 9650
- TN = 516
- FN = 276
- FP =660
- Número total de instancias = 13994

A continuación se muestra una descripción en la Tabla 30 índices de calidad Fig. 31 – 32, al ejecutar el algoritmo RandomTree.

Tabla 30. Índices de calidad de RandomTree

MEDIDA RT	VALOR RT
Tasa de error RT	12.08
Sensibilidad RT	95.38%
Especificidad RT	19.70%
Accuracy RT	87.97%
Coeficiente Kappa RT	0.17
Curva ROC RT	0.59
Precisión RT	85.7%
Recall RT	87.9%
TP Rate RT	87.9%
FP Rate RT	72.9%
F – Measure RT	86.7%

Fuente: Propia

Random Forest

En la Tabla 33 la matriz de confusión resultante se evalúa con 16 atributos, 13994 casos y 10 iteraciones de validación cruzada después de ejecutar el algoritmo de bosque aleatorio.

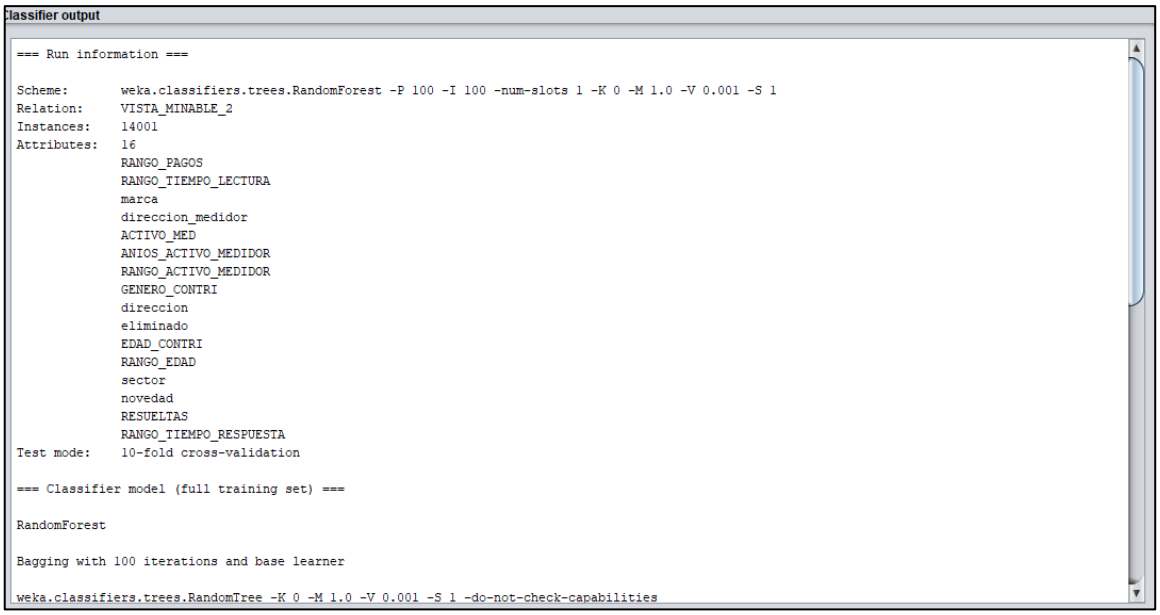


Fig. 31. Ejecución del algoritmo RandomForest (Parte 1)

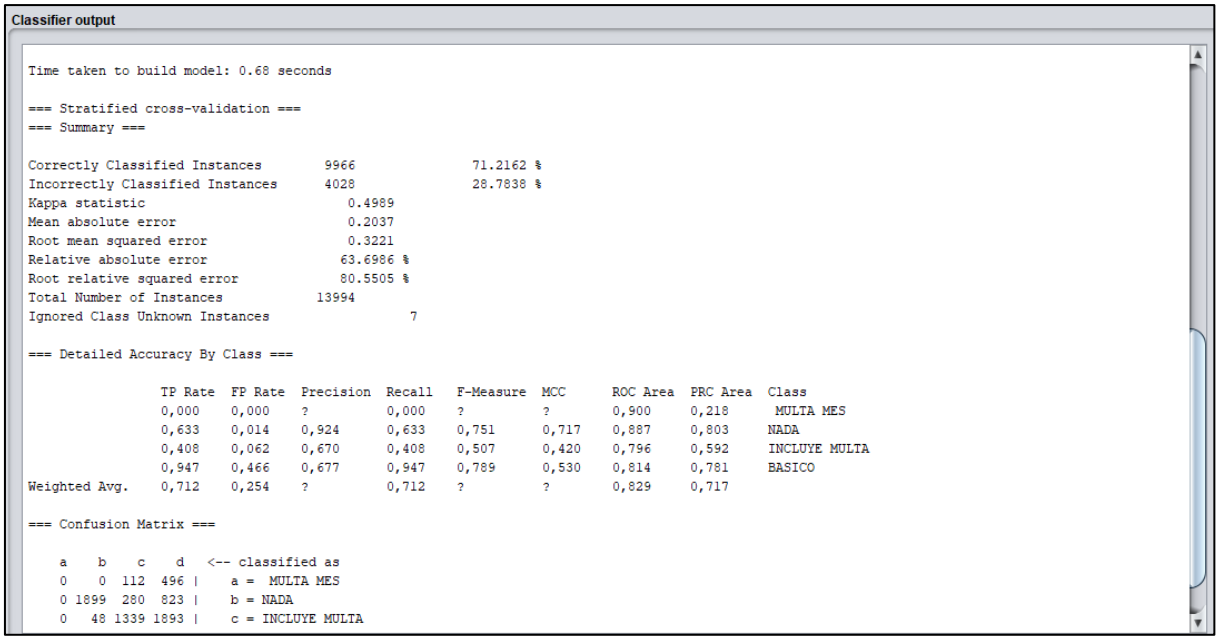


Fig. 32. Ejecución del algoritmo RandomForest (Parte 1)

En la Tabla 31. Se muestra la matriz de confusión que se obtuvo después de aplicar el algoritmo Random Forest Fig. 34 con 16 atributos, 13994 instancias.

Tabla 31. Resultados de la matriz de confusión del algoritmo RandomForest

		Clase predicha	
		R	NR
Clase verdadera	R	10104	240
	NR	1031	65

Fuente: Resultados del software Weka

Donde:

- TP = 13990
- TN = 45
- FN = 14
- FP = 1034
- Número total de instancias = 13994

En la siguiente Tabla 32. Se detalla las medidas de precisión adicionales de la matriz de confusión:

Tabla 32. Medidas estadísticas de RandomForest

MEDIDA RF	VALOR RF
Tasa de error RF	9.2143%
Sensibilidad RF	99.9901%
Especificidad RF	5.9307%
Accuracy RF	90.7857%
Coeficiente Kappa RF	0.1019
Curva ROC RF	0.748
Precisión RF	91.5%
Recall RF	90.8%
TP Rate RF	90.8%
FP Rate RF	84.9%
F – Measure RF	86.9%

Fuente: Propia

3.2. Tareas de Clasificación en (JASP)

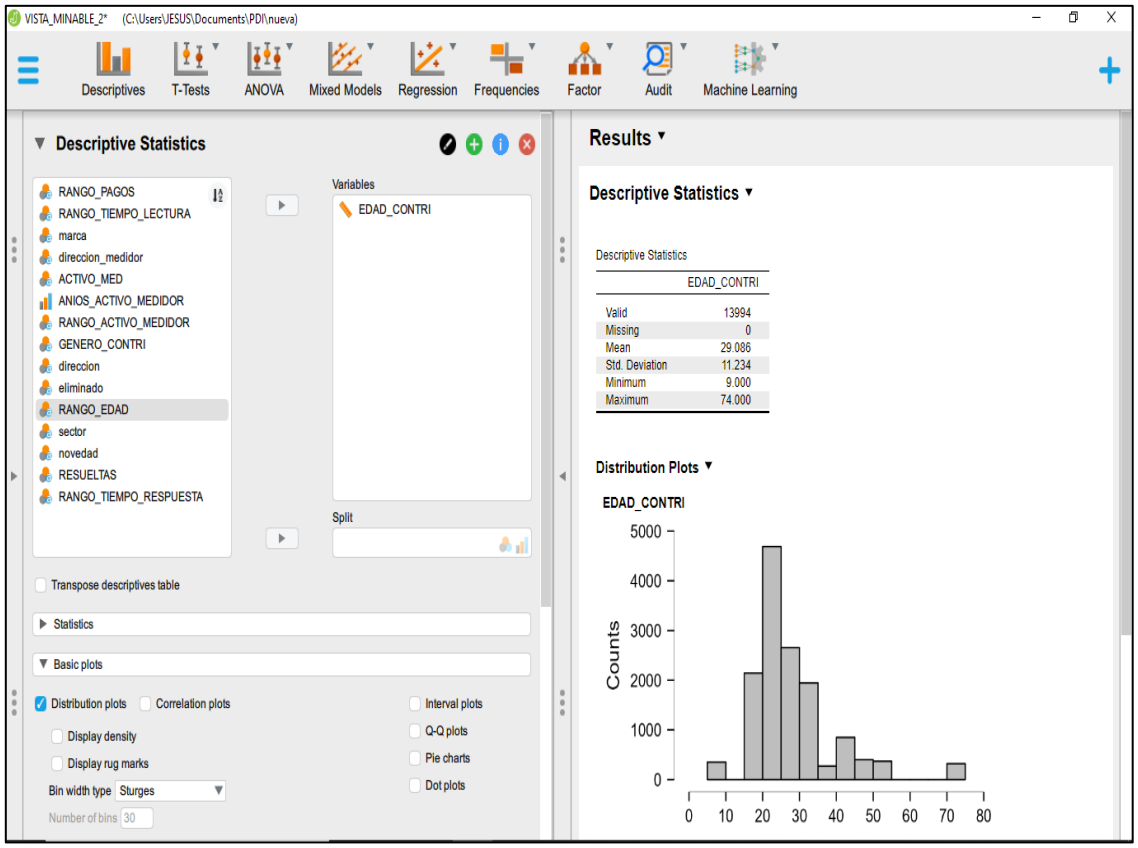


Fig. 33. Aplicación de técnicas descriptivas para EDAD_CONTRIBUYENTE

Descriptive Statistics

Descriptive Statistics

	EDAD_CONTRI
Valid	13994
Missing	0
Mean	29.086
Std. Deviation	11.234
Minimum	9.000
Maximum	74.000

Fig. 34. Datos Descriptivos

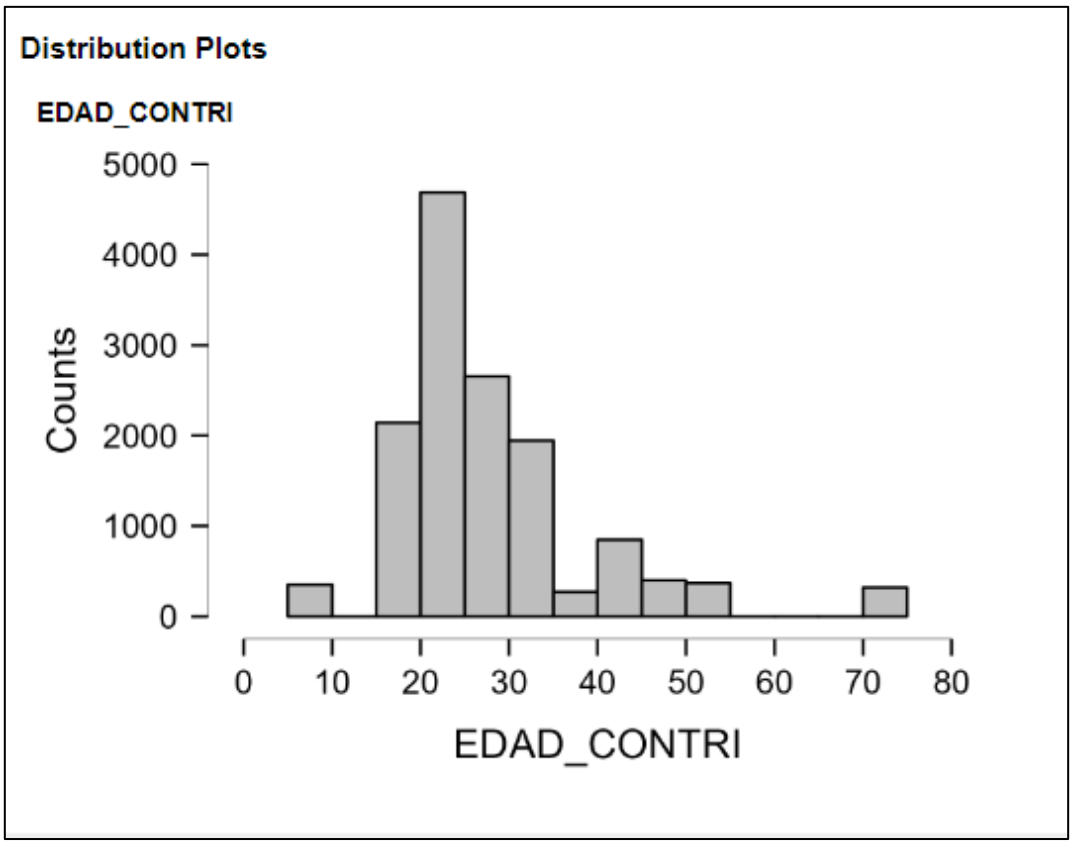


Fig. 35. Personas con mayor reclamo por edad

b. Tareas de Agrupamiento.

K-means

El algoritmo *K-means* utiliza una gran cantidad de datos existentes. Básicamente, el algoritmo le permite clasificar o segmentar información para identificar estructuras en conjuntos de datos que originalmente no tenían etiquetas ni categorías (Gironés Roig et al., 2017).

Para la aplicación del algoritmo se crearon 6 clústers Fig. 38 según los atributos de la dirección del medidor, lo que permitió identificar específicamente el conjunto de datos para cada sector con mayor nivel de quejas.

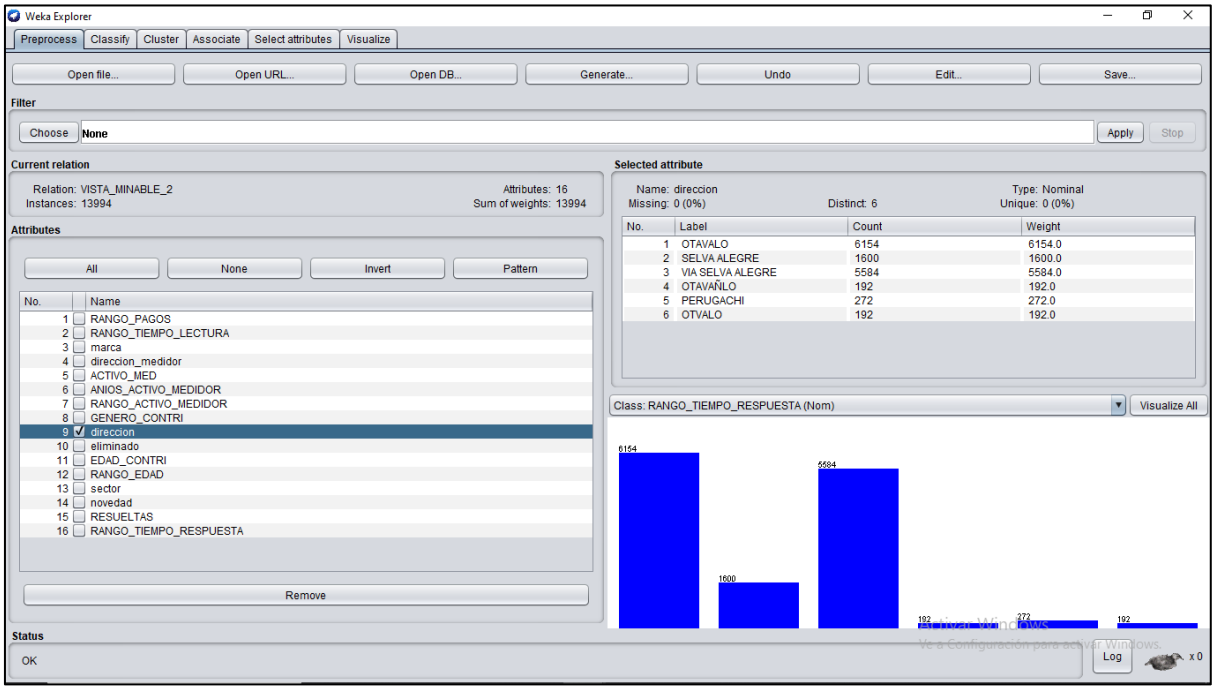


Fig. 36. Selección de atributos

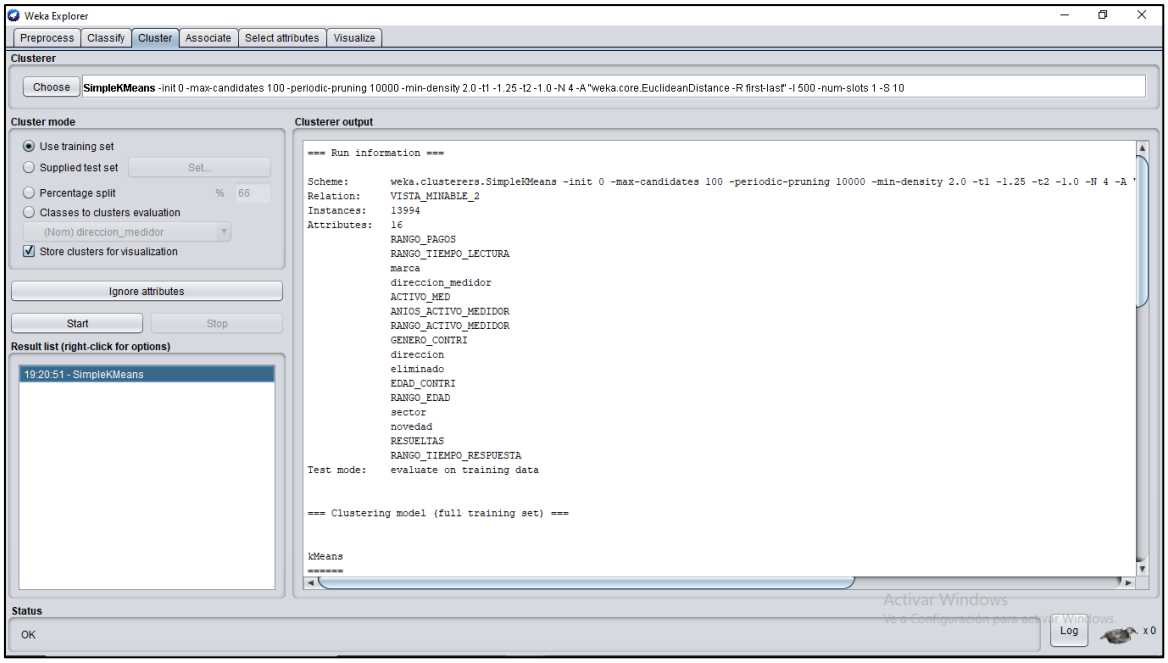


Fig. 37. Selección del algoritmo K-means (Parte 1)

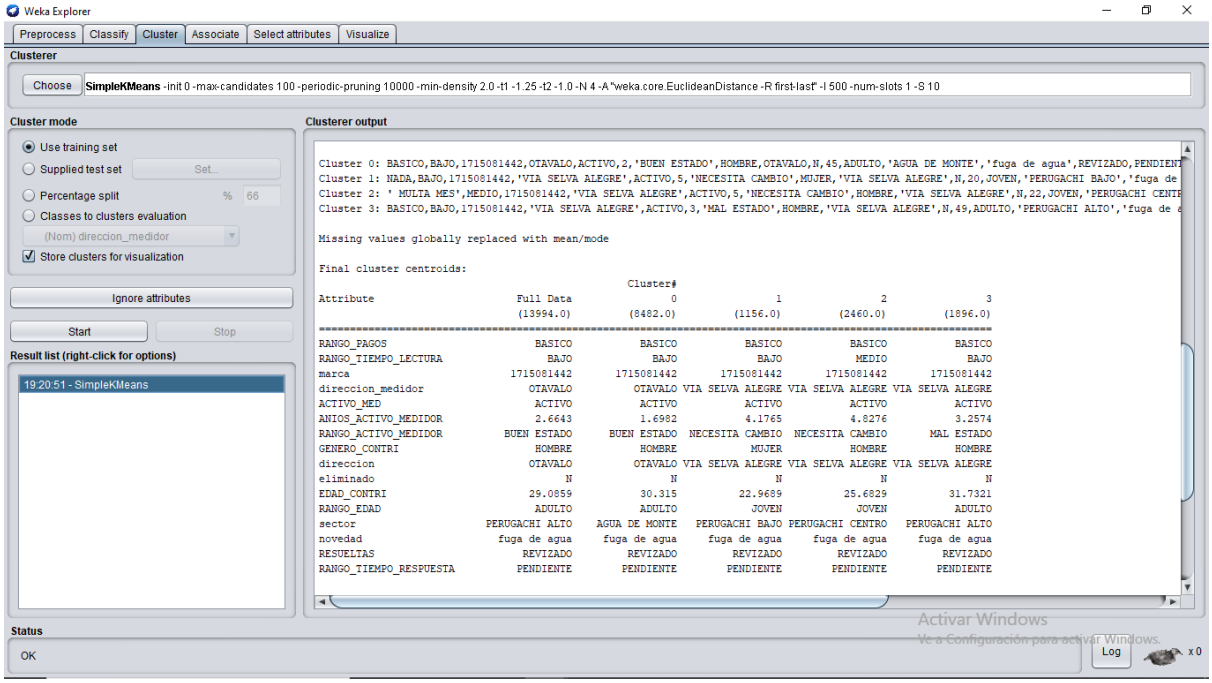


Fig. 38. Selección del algoritmo K-means (Parte 2)

Algoritmo EM

Se utiliza un algoritmo llamado EM o maximización de expectativas para determinar estimaciones de máxima verosimilitud que son fáciles de implementar y numéricamente sólidas (Huang & Chen, 2017). A cada instancia se le asigna una distribución de probabilidad que indica que pertenece a un clúster diferente creado de acuerdo la validación cruzada y según el número de clústeres a utilizar, hay 6 conjuntos basados en los atributos en la dirección de la información grupal. Los parámetros del algoritmo de ejecución se pueden ver en la Fig. 41.

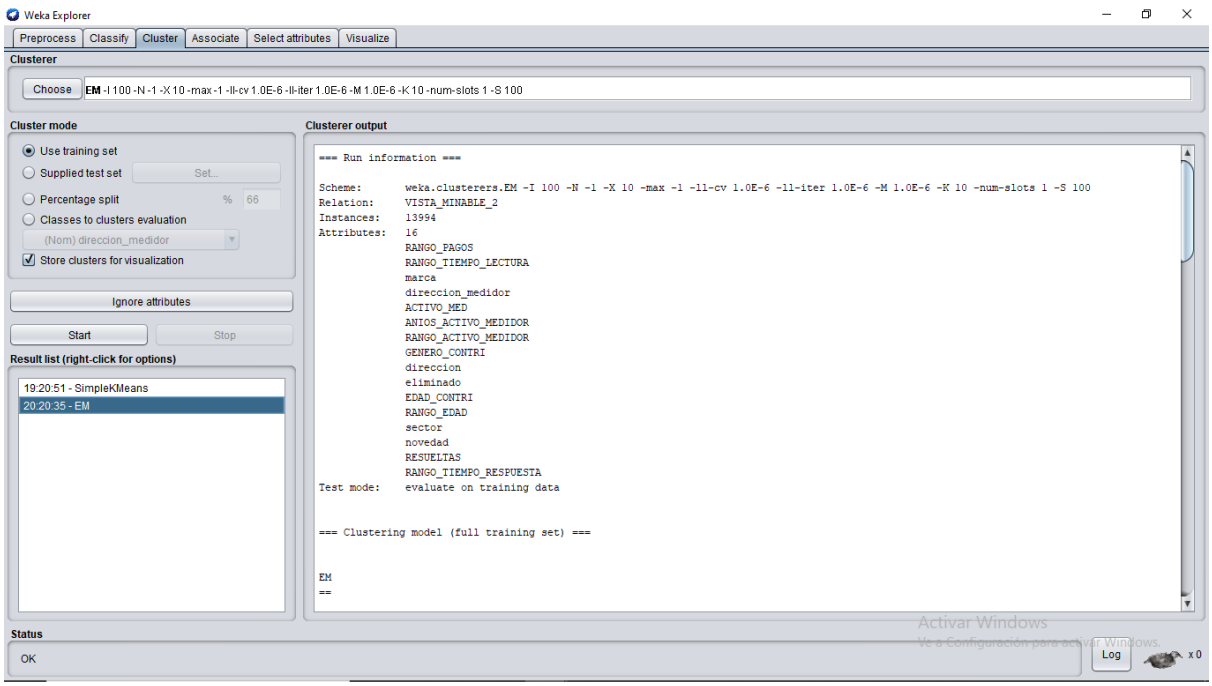


Fig. 39. Selección de algoritmo EM (Parte 1)

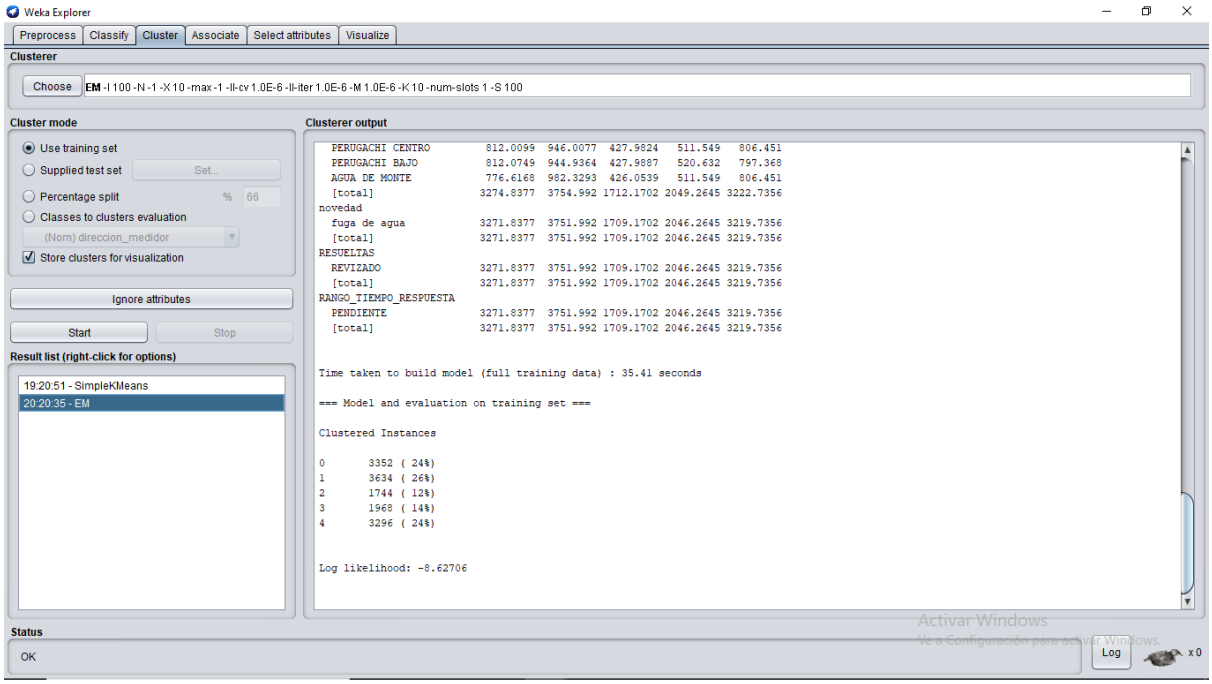


Fig. 40. Selección de algoritmo EM (Parte 2)

3.2. Evaluación e interpretación de datos

3.2.1. Evaluación del análisis e interpretación de clasificación

- **Análisis de resultados**

Para analizar el reclamo y consumo de agua se utilizó técnicas predictivas con el propósito de determinar las causas que produce un reclamo, estos modelos se validaron por métricas cuantitativas y con puntajes de precisión.

categorías:

- R = Reclamó
- NR = No reclamó

El modelo cuantitativo fue analizado con los algoritmos de clasificación con validación cruzada de 10-interaciones y de acuerdo a los valores que produce la ejecución del algoritmo RandomForest y RandomTree.

Se muestra las relaciones de los atributos que se generan al ejecutar el algoritmo de clasificación siguiendo un camino desde la raíz hasta las hojas conocidos como grafos de cuerdo a las Tablas 31 – 34 los indicadores muestran que existen diferentes categorías en las que dominan a otras en este caso se determinó elegir la variable RANGO_PAGOS y SECTOR para poder identificar los posibles barrios en los que

existan mayor irregularidad. En la categoría NR muestra 987 registro mientras tanto en la categoría R se analizaron 356 registros como se evidencia en la Tabla 29 y Tabla 31.

RandomForest

En la Fig. 43. Se aprecia los 13994 registros analizados con validación cruzada y matriz de confusión al ejecutar el algoritmo RandomForest en el software Weka.

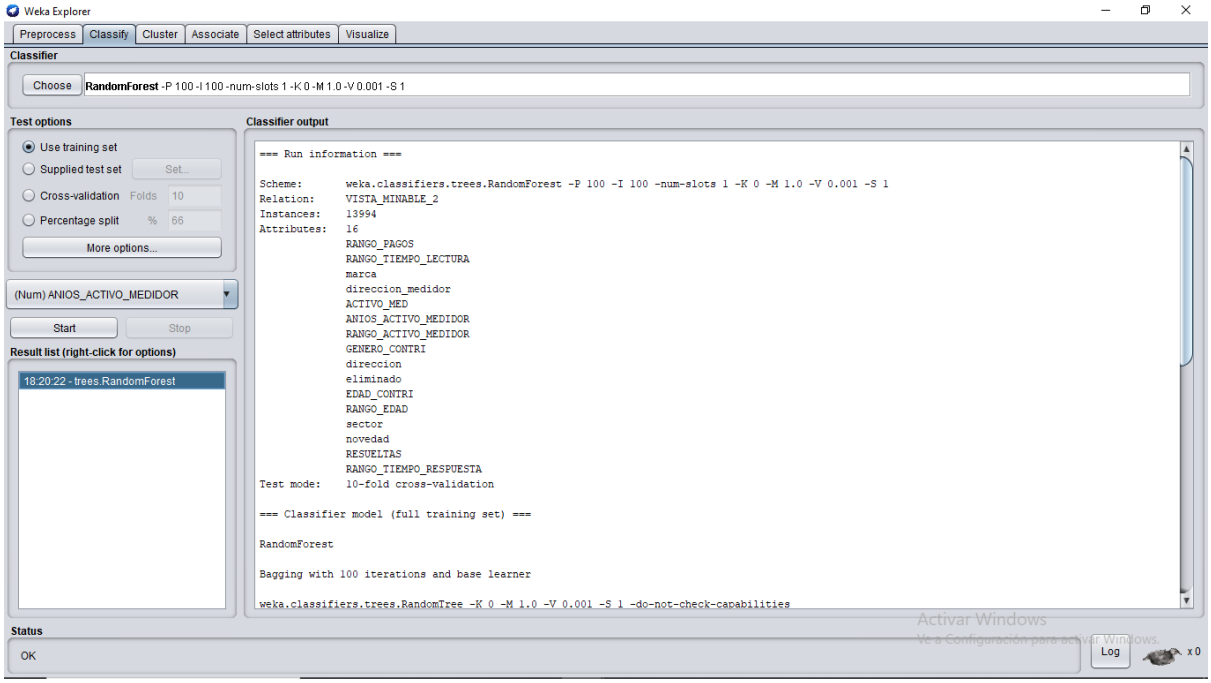


Fig. 41. Selección del algoritmo RandomForest (Parte 1)

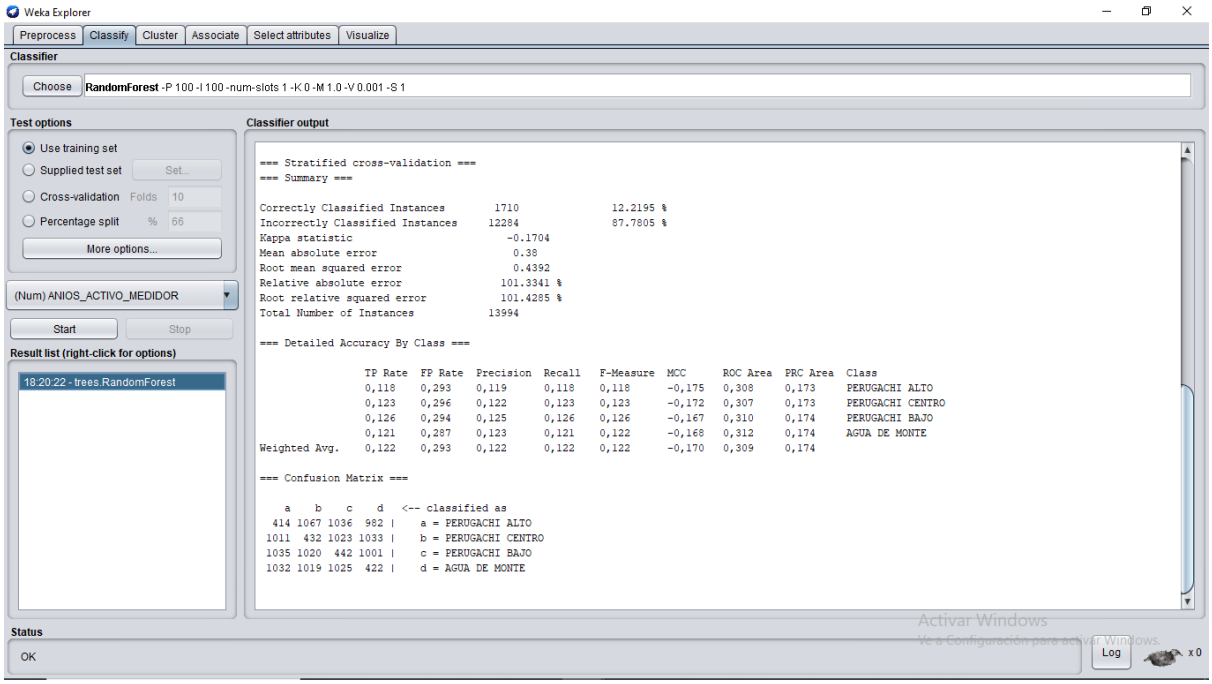


Fig. 42. Selección dell algoritmo RandomForest (Parte 2)

A continuación, los valores estadísticos se detallan en la Tabla 33, después de la ejecución de la matriz de confusión.

Tabla 33. Medidas estadísticas del modelo RandomForest

Medida	Valor
Accuracy RF	91.40%
Sensibilidad RF	99.94%
Precisión RF	95.46%
F1-score RF	93.45%
Área bajo la curva RF	0.78
Coeficiente Kappa RF	0.65

Fuente: Herramienta Weka

- **Interpretación de resultados**

Al finalizar la implementación de las faces que comprende la metodología KDD y obteniendo los mejores modelos de predicción con árboles de decisión RandomTree y RandomForest, siguiendo el camino desde la raíz hasta las hojas tomando como atributo clave ESTADO_MEDIDOR identificando los siguientes patrones:

- Reclamo por cambio de domicilio
- Reclamo por usuarios hombres de edad adulta

- Reclamo por usuarios de Otavalo
- Reclamo por cambio de medidor
- Reclamo por cambio de sector
- Reclamo por Usuarios que viven solos
- Reclamo por de la parte norte de Ilumán
- Reclamo por usuarios de Ángel Pamba
- Reclamo por usuarios de Ilumán Alto
- Reclamo por usuarios de Ilumán Bajo
- Reclamo por usuarios de Ilumán Centro
- Reclamo por usuarios de provienen del extranjero
- Reclamo por consumo excesivo de agua
- Reclamo por consumo mínimo de agua
- Reclamo por inconsistencia en planillas de Ilumán centro

Estos patrones provienen de la variable LOCALIDAD, ya que no existe un tanque que provee a este sector norte de Ilumán.

- Reclamo por usuarios de Ilumán con edad alta
- Reclamo por usuarios con edad baja por cambio de medidor.
- Reclamo del sexo femenino de Agualongo.
- Reclamos por usuarios de Ilumán bajo con edad alta.
- Reclamos por usuarios de Ilumán azares con edad alta.
- Reclamos por usuarios de Ilumán centro con edad alta.

3.2.2. Evaluación del análisis e interpretación de agrupamiento.

Algoritmo K-mean

Para aplicar la técnica de agrupamiento se tomó como atributo principal dirección medidor, los resultados se muestran en la Fig. 45.

```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 31061.282590828327

Initial starting points (random):

Cluster 0: BASICO,BAJO,1715081442,OTAVALO,ACTIVO,2,'BUEN ESTADO',HOMBRE,OTAVALO,N,45,ADULTO,'AGUA DE MONTE','fuga de agua',REVIZADO,PENDIENTE
Cluster 1: NADA,BAJO,1715081442,'VIA SELVA ALEGRE',ACTIVO,5,'NECESITA CAMBIO',MUJER,'VIA SELVA ALEGRE',N,20,JOVEN,'PERUGACHI BAJO','fuga de agua',REVIZADO,PENDIENTE
Cluster 2: ' MULTA MES',MEDIO,1715081442,'VIA SELVA ALEGRE',ACTIVO,5,'NECESITA CAMBIO',HOMBRE,'VIA SELVA ALEGRE',N,22,JOVEN,'PERUGACHI CENTRO','fuga de agua',REVIZADO,PENDIENTE
Cluster 3: BASICO,BAJO,1715081442,'VIA SELVA ALEGRE',ACTIVO,3,'MAL ESTADO',HOMBRE,'VIA SELVA ALEGRE',N,49,ADULTO,'PERUGACHI ALTO','fuga de agua',REVIZADO,PENDIENTE
Cluster 4: NADA,BAJO,1715081442,OTAVALO,ACTIVO,1,'BUEN ESTADO',MUJER,OTAVALO,Y,9,JOVEN,'PERUGACHI ALTO','fuga de agua',REVIZADO,PENDIENTE
Cluster 5: BASICO,MEDIO,1715081442,'VIA SELVA ALEGRE',ACTIVO,5,'NECESITA CAMBIO',MUJER,'VIA SELVA ALEGRE',N,20,JOVEN,'PERUGACHI ALTO','fuga de agua',REVIZADO,PENDIENTE

Missing values globally replaced with mean/mode
```

Final cluster centroids:

		Cluster#					
Attribute	Full Data	0	1	2	3	4	5
	(13994.0)	(5163.0)	(612.0)	(2264.0)	(1896.0)	(3499.0)	(560.0)
RANGO_PAGOS	BASICO	BASICO	NADA	BASICO	BASICO	NADA	BASICO
RANGO_TIEMPO_LECTURA	BAJO	BAJO	BAJO	MEDIO	BAJO	BAJO	MEDIO
marca	1715081442	1715081442	1715081442	1715081442	1715081442	1715081442	1715081442
direccion_medidor	OTAVALO	OTAVALO	VIA SELVA ALEGRE	VIA SELVA ALEGRE	VIA SELVA ALEGRE	OTAVALO	VIA SELVA ALEGRE
ACTIVO_MED	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO
ANIOS_ACTIVO_MEDIDOR	2.6643	1.8458	4.3203	4.8746	3.3418	1.4959	4.4714
RANGO_ACTIVO_MEDIDOR	BUEN ESTADO	BUEN ESTADO	NECESITA CAMBIO	NECESITA CAMBIO	MAL ESTADO	BUEN ESTADO	NECESITA CAMBIO
GENERO_CONTRI	HOMBRE	HOMBRE	MUJER	HOMBRE	HOMBRE	HOMBRE	MUJER
direccion	OTAVALO	OTAVALO	VIA SELVA ALEGRE	VIA SELVA ALEGRE	VIA SELVA ALEGRE	OTAVALO	VIA SELVA ALEGRE
eliminado	N	N	N	N	N	N	N
EDAD_CONTRI	29.0859	33.9682	23.1961	26.0742	32.1329	24.4944	21.0571
RANGO_EDAD	ADULTO	ADULTO	JOVEN	JOVEN	ADULTO	JOVEN	JOVEN
sector	PERUGACHI ALTO	AGUA DE MONTE	PERUGACHI BAJO	PERUGACHI CENTRO	PERUGACHI ALTO	PERUGACHI ALTO	PERUGACHI ALTO
novedad	fuga de agua	fuga de agua	fuga de agua	fuga de agua	fuga de agua	fuga de agua	fuga de agua
RESUELTAS	REVIZADO	REVIZADO	REVIZADO	REVIZADO	REVIZADO	REVIZADO	REVIZADO
RANGO_TIEMPO_RESPUESTA	PENDIENTE	PENDIENTE	PENDIENTE	PENDIENTE	PENDIENTE	PENDIENTE	PENDIENTE

Fig. 43. Algoritmo K-means

Algoritmo EM

Al implementar la técnica de agrupamiento de minería de datos se toma en cuenta el atributo sectores que permite dividir de forma específica y conocer los resultados del análisis y determinar las posibles soluciones con las tomas de decisiones.

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     VISTA_MINABLE_2
Instances:    13994
Attributes:   16
              RANGO_PAGOS
              RANGO_TIEMPO_LECTURA
              marca
              direccion_medidor
              ACTIVO_MED
              ANIOS_ACTIVO_MEDIDOR
              RANGO_ACTIVO_MEDIDOR
              GENERO_CONTRI
              eliminado
              EDAD_CONTRI
              RANGO_EDAD
              sector
              novedad
              RESUELTAS
              RANGO_TIEMPO_RESPUESTA
Ignored:
              direccion
Test mode:    Classes to clusters evaluation on training data

=== Clustering model (full training set) ===


EM
==
```

```
Number of clusters selected by cross validation: 6
Number of iterations performed: 3


                         Cluster
Attribute                     0         1         2         3         4         5
                         (0.11)    (0.18)    (0.25)    (0.24)    (0.13)    (0.08)
=============================================================================
RANGO_PAGOS
   MULTA MES                 97  110.7064     1.003       385    2.4208   17.8698
  NADA                  240.9996    3.4015  644.1376       337 1393.0513    389.41
  INCLUYE MULTA         304.9999 1450.9606  757.3447       593    3.0202  176.6746
  BASICO                928.9995 1015.0709   2111.37 1985.0006  467.7035  601.8555
  [total]               1571.999 2580.1394 3513.8553 3300.0007 1866.1958 1185.8098
RANGO_TIEMPO_LECTURA
  BAJO                 1008.9992 2577.1376  3050.098 1345.0002 1244.8825 1051.8825
  INDEFINIDO             96.9999    1.0018  461.7572       129  619.3133   99.9278
  MEDIO                      465         1         1 1537.0005         1   32.9995
  ALTO                         1         1         1       289         1         1
  [total]               1571.999 2580.1394 3513.8553 3300.0007 1866.1958 1185.8098
marca
  M056                         1    2.6562   21.0055         1  107.3383         1
  1715081442            1568.999 2575.4832 3490.8498 3297.0007 1756.8575 1182.8098
  [total]               1569.999 2578.1394 3511.8553 3298.0007 1864.1958 1183.8098
direccion_medidor
  OTAVALO                      1 1989.9643  2285.412         1 1673.0634   49.5603
  SELVA ALEGRE                 1   588.175 1034.4447         1   191.131  734.2493
  VIA SELVA ALEGRE      1568.999         1         1 3297.0007         1  401.0002
  OTVALO                       1         1  192.9986         1    1.0013         1
  [total]               1571.999 2580.1394 3513.8553 3300.0007 1866.1958 1185.8098
```

```
ACTIVO_MED
  ACTIVO                  1568.999 2577.1394 3510.8553 3297.0007 1863.1958 1182.8098
  [total]                 1568.999 2577.1394 3510.8553 3297.0007 1863.1958 1182.8098
ANIOS_ACTIVO_MEDIDOR
  mean                           3    1.9088    1.4507         5    1.7878    2.3369
  std. dev.                  1.415    0.2879    0.4976    0.0009    0.4088    0.4759

RANGO_ACTIVO_MEDIDOR
  BUEN ESTADO                    1 2577.1393 3510.8553         1 1863.1958  782.8096
  MAL ESTADO              1568.999         1         1    1.0007         1  401.0002
  NECESITA CAMBIO                1         1         1      3297         1         1
  [total]                 1570.999 2579.1394 3512.8553 3299.0007 1865.1958 1184.8098
GENERO_CONTRI
  HOMBRE                  1040.999 2318.3333 2893.5498 2529.0007 1734.5451 1179.5721
  MUJER                        529  259.8061  618.3055       769  129.6507    4.2377
  [total]                 1569.999 2578.1394 3511.8553 3298.0007 1864.1958 1183.8098
eliminado
  N                       1568.999 2573.3555 3258.7057 3297.0007 1751.1394 1182.7996
  Y                              1    4.7838  253.1496         1  113.0564    1.0102
  [total]                 1569.999 2578.1394 3511.8553 3298.0007 1864.1958 1183.8098
EDAD_CONTRI
  mean                     25.6429   33.8055   20.3346   25.6796   31.4192     55.18
  std. dev.                 1.6918    6.6542    4.1308    5.7685    6.5584   13.0066

RANGO_EDAD
  ADULTO                  512.9991 2568.3731   28.3095  849.0007 1674.5638  862.7539
  TERCERAEDAD                    1         1         1         1         1       321
  JOVEN                        529    1.4728 2812.3409      1569    5.1424    1.0439
  JOVEN,ADULTO                 529    9.2935  672.2049       881  185.4896     1.012
  [total]                 1571.999 2580.1394 3513.8553 3300.0007 1866.1958 1185.8098
```

```
sector
  PERUGACHI ALTO           392.9998  648.6953  877.3478  825.0002  468.9223  292.0346
  PERUGACHI CENTRO         392.9997  641.1672  879.5989  825.0002  465.1396  301.0944
  PERUGACHI BAJO           392.9998  644.9328  878.4561  825.0002  466.1896  296.4215
  AGUA DE MONTE            392.9998  645.3441  878.4525  825.0002  465.9442  296.2593
  [total]                 1571.999 2580.1394 3513.8553 3300.0007 1866.1958 1185.8098
novedad
  fuga de agua            1568.999 2577.1394 3510.8553 3297.0007 1863.1958 1182.8098
  [total]                 1568.999 2577.1394 3510.8553 3297.0007 1863.1958 1182.8098
RESUELTAS
  REVIZADO                1568.999 2577.1394 3510.8553 3297.0007 1863.1958 1182.8098
  [total]                 1568.999 2577.1394 3510.8553 3297.0007 1863.1958 1182.8098
RANGO_TIEMPO_RESPUESTA
  PENDIENTE               1568.999 2577.1394 3510.8553 3297.0007 1863.1958 1182.8098
  [total]                 1568.999 2577.1394 3510.8553 3297.0007 1863.1958 1182.8098


Time taken to build model (full training data) : 51.21 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       1568 ( 11%)
1       3008 ( 21%)
2       3568 ( 25%)
3       3296 ( 24%)
4       1514 ( 11%)
5       1040 (  7%)
```

```
Log likelihood: -8.18601


Class attribute: direccion
Classes to Clusters:

    0    1    2    3    4    5  <-- assigned to cluster
    0 2448 2576    0 1130    0 | OTAVALO
    0  544  320    0   96  640 | SELVA ALEGRE
 1568    0  320 3296    0  400 | VIA SELVA ALEGRE
    0    0  192    0    0    0 | OTAVAÑLO
    0    0  160    0  112    0 | PERUGACHI
    0   16    0    0  176    0 | OTVALO

Cluster 0 <-- No class
Cluster 1 <-- OTAVALO
Cluster 2 <-- OTAVAÑLO
Cluster 3 <-- VIA SELVA ALEGRE
Cluster 4 <-- OTVALO
Cluster 5 <-- SELVA ALEGRE

Incorrectly clustered instances :       7242.0   51.7508 %
```

Fig. 44. Algoritmo EM

Con el modelo estadístico obtenido del algoritmo EM, nos permite identificar similitudes con otros atributos o clústeres a razón de su desviación estándar.

RANGO_PAGOS

En este atributo, los clústers 0 y 2 tienen categoría 1 y sus desviaciones estándar son valores muy similares, pero no muy altos, mientras que los demás valores contienen valores diferentes por igual y sus desviaciones estándar son muy similares, entendiendo los datos que este atributo puede La degeneración es mínima.

RANGO_TIEMPO_LECTURA

En este atributo, los clústers 0, 1 y 3 almacenan categorías indefinidas, medias y altas con desviaciones estándar por debajo de 3. Se entiende que los datos de cada conglomerado son diferentes, pero los clústers 3 y 5 tienen una categoría baja. cuyo valor de desviación estándar es diferente en comparación con otras categorías y el valor máximo de los datos agrupados se considera muy diverso.

Marca

En estos clústers de atributos 0,3 y 5, sus categorías MO56 mantienen una desviación estándar menor a 1, a excepción de los clústers 1, 2, 3 y 4, sus categorías también contienen 1715081442, lo que nos permite entender los datos.

ACTIVO_MED

Por esta característica, los conglomerados 0, 1, 2 y 4 han conservado la categoría ACTIVO en sus datos y muestran sus desviaciones estándar en valores altos para mostrar que los datos son muy variables.

ANIOS_ACTIVO_MEDIDOR

Para este atributo, los conglomerados 0, 1, 2 y 5 conservan categorías indefinidas, medias y altas, respectivamente, con desviaciones estándar menores a 1, lo que significa que los datos en cada conglomerado son iguales, mientras que los conglomerados 0, 3 y 5, que es una desviación estándar y el valor máximo para las categorías 2, 3 y 4 son muy diferentes de las otras categorías, y los datos agrupados pueden considerarse muy diferentes.

RANGO_ACTIVO_MEDIDOR

Para esta propiedad, en los conglomerados 0 y 3, almacenan la categoría "buen estado" se definen como el rango que clasifica a las poblaciones con una desviación estándar inferior a 1, lo que significa que sus datos no difieren mucho, mientras que para los conglomerados 1, 2, 3 y 4, tienen categorías MAL ESTADO con desviaciones estándar por debajo de 1, sabiendo que los datos no han cambiado mucho.

GENERO_CONTRIBUYENTE

Los conglomerados 0 y 6 contienen la categoría MUJER con un valor de desviación estándar mayor a 3, lo que significa que los datos son muy diferentes, mientras que los conglomerados 1, 2, 3, 4 y 5 contienen la categoría HOMBRE con valores de desviación estándar mayores de 3 desviaciones estándar fueron 2, lo que implica que los datos de estos conjuntos fueron similares.

eliminado

Para los conglomerados 0, 3 y 4, el conglomerado almacena Y categorías con desviaciones estándar entre 2 y mayores a 3, lo que indica una alta diversidad en los datos, mientras que para los conglomerados 2 y 3, esto corresponde a N. datos completamente diferentes.

EDAD_CONTRI

En este atributo, los conglomerados 0 y 3 almacenan 25 categorías y sus valores de desviación estándar están por debajo de 6, lo que significa que los datos son muy similares, mientras que los conglomerados 1, 2, 4 y 5 almacenan 20, 25, 30 y 40 Son 55 categorías. y las desviaciones estándar son mayores a 2, se puede entender que los datos de estos conglomerados son similares.

RANGO_EDAD

Para esta característica, los conglomerados 3 y 5 conservan la categoría de adulto, mientras que los conglomerados 0, 1, 2 y 3 corresponden a controles por ingresos monetarios de personas mayores.

3.3. Procesos de obtención de conocimiento

Clasificación.

En la tecnología de clasificación, aplique los algoritmos RandomForest y RandomTree, verifique si los datos se recorren de acuerdo con las principales variables y atributos de las características, el proceso de adquisición de conocimiento útil y clasifique de acuerdo con las páginas obtenidas.

- Reclamo que proceden de sectores urbanos, Ángel Pamba, Ilumán Bajo, San Luis de Agualongo o Barrió Centro
- Reclamo por usuarios que cambiaron de medidor
- Reclamo de usuarios hombres y mujeres con edad media
- Reclamo de usuarios con inconsistencias en planillas
- Reclamo por cambio de localidad
- Reclamo por usuarios de género femenino.
- Conocimiento obtenido de la data set completo:

Agrupamiento.

- **Clúster 0 (San Luis de Agualongo):** quejas de usuarios de género femenino mayores de 25 años por altas multas en las planillas, lecturas de medidor irregulares y en medidores del modelo M056.

- **Clúster 1 (Ángel Pamba):** quejas de usuarios de género masculino mayores de 33 años realizadas por fugas de agua y con tiempo de respuesta mayor a los 28 días mencionado por el E1.
- **Clúster 2 (Ilumán Bajo):** quejas de usuarios de género masculino jóvenes a partir de 20 años, cambio de medidor por mal estado del modelo 1715081443, dado que los años activos del medidor superan los 4 años con un consumo básico.
- **Clúster 3 (Pinsaquí):** quejas de usuarios de género femenino mayores de 25 años por altas multas en las planillas, lecturas de medidor irregulares y fugas de agua en medidores del modelo M056 y 1715081443 donde ha cumplido un periodo de vida útil lo cual los operarios de JAAPYSR-I han dado de baja de acuerdo con el E2.
- **Clúster 4 (Barrio Central):** quejas de usuarios de género masculino mayores de 31 años realizadas por fugas de agua y altos índices de consumo en sus planillas de agua con tiempo de respuesta mayor a los 28 días mencionado por el E1.
- **Clúster 5 (Rancho Chico):** quejas de usuarios de género femenino a partir de 31 años, por cambio de medidor por mal estado del modelo 1715081443, dado que los años activos del medidor superan los 4 años con un consumo básico.

3.4. Análisis de impactos

Nos permite determinar y evaluar los efectos de cualquier imprevisto que pueda afectar a la continuidad del negocio en base a la obtención de los patrones de consumo y reclamo. Motivo por el cual es primordial evaluar de acuerdo con ciertas dimensiones e indicadores. En la Tabla 34 detalla los niveles de impacto.

Tabla 34. Niveles de impacto (Posso,2013)

NIVELES DE IMPACTOS	VALOR
Alto Positivo	3
Medio Positivo	2
Bajo Positivo	1

Punto de Indiferencia	0
Bajo Negativo	-1
Medio Negativo	-2
Alto Negativo	-3

Fuente: (Posso)

Se tomaron en cuanta el ámbito económico, tecnológico y sociocultural para su impacto de análisis, como se detalla en las Tablas 35 – 38, finalmente se agrega la Tabla 35 que representa el impacto general que facilitara a directivos a la toma de decisiones.

3.4.1. Impacto Económico

Tabla 35. Resultados de impacto económico

INDICADOR	NIVELES						
	-3	-2	-1	0	1	2	3
Productividad de JAAPYSR-I							X
Presupuesto del Usuario							X
Presupuesto financiero.							X
TOTAL				0			9

$$Nivel\ de\ impacto = \frac{\Sigma}{Número\ de\ indicadores}$$

$$Nivel\ de\ impacto = \frac{9}{3} = 3$$

Nivel de Impacto Económico = Alto positivo

Fuente: (Posso, 2013).

De acuerdo con los resultados de la Tabla 35, se considera tener un impacto alto positivo, donde la Junta de agua de San Juan de Ilumán aumentará, Porque la información se almacena en la base de datos de YAKUSOFT, Tendrá como propósito mejorar la calidad de los servicios, se reflejará en él presupuesto para mantenimiento, atenciones a domicilio y movilidad durante el periodo del presidente de turno. Los presupuestos de los clientes no se verán afectados y podrán comprobar que no existen discrepancias en sus facturas de agua, evitando así el pago de impuestos acumulativos. La elaboración de un presupuesto económico tiene un efecto muy positivo, ya que proporciona una estimación preliminar de los ingresos y gastos que se producirán a lo largo de un determinado periodo de tiempo.

3.4.2. Impacto Tecnológico

Tabla 36. Resultados de Impacto tecnológico

INDICADOR	NIVELES						
	-3	-2	-1	0	1	2	3
Nuevos Servicios Tecnológicos							X
Nivel de respuesta						X	

$$Nivel\ de\ impacto = \frac{\Sigma}{Número\ de\ indicadores}$$

$$Nivel\ de\ impacto = \frac{5}{2} = 2.5$$

Nivel de Impacto Tecnológico = Medio positivo

Fuente: (Posso, 2013).

Los presupuestos de los clientes no se verán afectados y podrán comprobar que no existen discrepancias en sus facturas de agua, evitando así el pago de impuestos acumulativos. La elaboración de un presupuesto económico tiene un efecto muy positivo, ya que proporciona una estimación preliminar de los ingresos y gastos que se producirán a lo largo de un determinado periodo de tiempo.

3.4.3. Impacto sociocultural

Tabla 37. Resultados de Impacto sociocultural

INDICADOR	NIVELES						
	-3	-2	-1	0	1	2	3
Nivel de atención							X
Nivel de respuesta							X
Fuentes de financiación de otras instituciones							
Niveles de reclamo							X
Niveles de consumo						X	
TOTAL				0		2	9

$$Nivel\ de\ impacto = \frac{\Sigma}{Número\ de\ indicadores}$$

$$Nivel\ de\ impacto = \frac{11}{5} = 2.2$$

Nivel de Impacto Sociocultural = Medio positivo

Fuente: (Posso, 2013).

Se considera medio positivo al impacto Sociocultural, ya que los usuarios de JAAPYSR-I podrán contar con buenos servicios de agua potable, agilitando las peticiones de los usuarios reduciendo los indicadores en los niveles de reclamo y consumo.

3.4.4. Impacto General

Tabla 38. Resultados de Impacto general

INDICADOR	NIVELES						
	-3	-2	-1	0	1	2	3
Impacto Económico							X
Impacto Tecnológico							X
Impacto Sociocultural						X	
TOTAL				0		2	3

$$Nivel\ de\ impacto = \frac{\Sigma}{Número\ de\ indicadores}$$

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

Nivel de Impacto General = Medio positivo

Fuente: (Posso, 2013).

El impacto general es moderadamente positivo, lo que abre la esperanza de promover la toma de decisiones basadas en el conocimiento, ya que los usuarios y todos los habitantes de San Juan de Ilumán se benefician de la prestación de servicios de calidad, especialmente en las zonas más alejadas por las condiciones geográficas.

3.5. Discusión

El presente trabajo es continuación de la investigación realizada por (Drogodependientes & En, 2018). Donde se abordó esta problemática mediante el análisis de datos de consumo y reclamo de los usuarios de INTERAGUA Cía. Ltda. y EMAPAG EP, con registros de sus atributos de la data warehouse y algoritmo RandomTree. Los datos históricos de consumo y reclamo y entre otros se realizó con 17 variables aplicando técnicas de clasificación y algoritmos RandomForest y RandomTree.

Se obtuvieron similitudes con otras investigaciones de acuerdo a los resultados que muestran otros autores como el de (Drogodependientes & En, 2018). Encontrando semejanzas, donde los usuarios hacen reclamos por inconsistencias en sus planillas de agua. Considerando que se utilizó los mismos datos históricos de consumo, se tiene con mayor similitud directa con el reclamo por el sector y consumo,

para toma de decisiones por las autoridades de JAAPYSR-I, se diferencia de (Drogodependientes & En, 2018). Al encontrar que en el atributo GENERO se evidencia que hay sectores donde la parte femenina son quienes concurren en realizar una queja a JAAPYSR-I.

Según el trabajo de (Humaid, 2017). Afirman que los atributos de las personas, excepto la edad, no inciden en aplicar un reclamo a JAAPYSR-I, resultados que discrepan de la investigación de (Rebelo et al., 2022). Donde muestran que es necesario utilizar la mayor cantidad de datos reales para identificar de mejor forma los patrones de reclamo y consumo son las características personales de los usuarios. estado civil, edad y género principales características que construyen modelos más precisos.

3.6. Limitaciones

- La información básica de la Junta de Aguas es confidencial, ya que es un proyecto con fines académicos.
- El periodo de recogida de la información analizada es 2017-2022.
- El conjunto de datos proporcionado por la Junta de aguas no tiene muchas variables y no se pueden utilizar técnicas de minería de datos profunda.
- Esta información no incluye elementos clave para obtener conocimiento relevante adicional, en el caso de tal información, se considerará necesario registrar datos del usuario, su forma, especificaciones de materiales utilizados e información específica del sitio sobre su futura ubicación geográfica. lugar relevante.

CONCLUSIONES

Se obtuvo patrones de consumo y reclamo, realizado de acuerdo a las exigencias y faces de la metodología (KDD), tuvo como principal función de brindar soluciones al problema con los usuarios de los reclamos que se acumulaba en la parroquia de San Juan de Ilumán.

La información oculta de los registros acumulados de 2017 a 2022 se muestra utilizando una base de datos transaccional y estrategias de extracción de datos para proporcionar una base para el análisis de casos y la toma de decisiones posteriores.

Se obtuvo una vista minable con datos históricos del Sistema YAKUSOFT y una copia de seguridad, de la base de datos PostgreSQL, con un total de 13990 registros y procesados en la herramienta Pentaho Data Integration (PDI).

Según la entrevista, fijó un plazo máximo de 35 días para dar solución a los trámites emitidos por los usuarios de la JAAPYSR-I con base en las normas internas de gestión del servicio. E3 juega un papel importante en la aprobación de las solicitudes de registro.

RECOMENDACIONES

Mantener actualizada los almacenes de datos YAKUSOFT, Teniendo en cuenta que la solicitud de quejas de cada usuario son variantes durante un periodo de tiempo, mencionado esto los datos históricos obtenidos deben estar relacionados a la realidad y poder facilitar en el entendimiento de los resultados

Se recomienda la recopilación de datos pertinentes que permitan el acoplamiento a la metodología KDD y la aplicación de todas las herramientas necesarias para facilitar el proceso de cada fase para las investigaciones posteriores relacionadas con la detención del fraude. Por ejemplo, la base de datos relacional mantiene una dependencia con otras tablas y no se excluye del análisis.

El análisis de los antecedentes históricos (recursos) debe ser considerado como los puntos principales del plan de acción de la JAAPYSR-I en cuanto a los conocimientos adquiridos y la situación a preparar para cada punto principal.

Anexo.

JUNTA ADMINISTRADORA DE AGUA POTABLE Y SANEAMIENTO
REGIONAL ILUMAN JAAPYSR-I
Fundado el 29 de marzo de 1979
Decreto N° 3327, Registro oficial N° 802
Ilumán-Otavalo-Ecuador
RUC 1060024950001

CARTA DE AUTORIZACIÓN Y ACEPTACIÓN

Ilumán, a 30 de marzo del 2021

La Suscrita Junta Administradora de Agua Potable y Saneamiento Regional Ilumán JAAPYSR-I autoriza y acepta al Sr. Jesús Eduardo Gonzales estudiante de la carrera de Ingeniería en sistemas de la Universidad Técnica del Norte UTN, para que pueda realizar su tesis de grado en nuestra institución de acuerdo como lo solicita por medio del oficio N. UTN-FICA-CISIC-054-0

Puede dar el uso pertinente a su bien tuviere.

Atentamente,

DIRECTORIO 2019-2021

JUNTA ADM. DE AGUA POTABLE Y ALCANTARILLADO DE ILUMAN JAAPYSR-I * PRESIDENCIA * Telf: 2946129 RUC: 1060024950001

Taury Montalvo
PRESIDENTE JAAPYSR-I

JUNTA ADM. DE AGUA POTABLE Y ALCANTARILLADO DE ILUMAN JAAPYSR-I * SECRETARÍA * Telf: 2946129 RUC: 1060024950001

Diana de la Torre
SECRETARIA JAAPYSR-I

Dirección: Modesto Larrea y 12 de Noviembre Esq. **Telf.** 2946-129/2946-323
Correo: jaapysr.iluman19@gmail.com

BIBIOGRAFÍA

Agarwal, S. (2016). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Al-Radaideh, Q., & Al-Zoubi, M. (2018). *A data mining based model for detection of fraudulent behaviour in water consumption*. https://doi.org/10.1109/IACS.2018.8355440

Alexeis, I., Reyes, J. O., Arturo, I., García, O., Yovannys, I., Corales, S., Davila, I. F., & Iv, H. (n.d.). Componente web para el análisis de información clínica usando la técnica de Minería de Datos por agrupamiento Web component for the analysis of clinical information using the technique of clustering data mining. *Revista Cubana de Informática Médica*, *2014*(1), 5–16. Retrieved February 2, 2023, from http://scielo.sld.cu

Autónomo, G. (2015). *PLAN DE DESARROLLO Y ORDENAMIENTO TERRITORIAL DE LA PARROQUIA SAN JUAN DE ILUMAN*.

Beltrán Pascual, M., Muñoz Martínez, A., & Muñoz Alamillos, Á. (2014). Redes bayesianas aplicadas a problemas de credit scoring. Una aplicación práctica. *Cuadernos de Economía*, *37*(104), 73–86. https://doi.org/10.1016/J.CESJEF.2013.07.001

Bertomeu, P. (2018). La entrevista. *Pfolgueiras@ub.Edu*, 1–11. https://diposit.ub.edu/dspace/bitstream/2445/99003/1/entrevista pf.pdf

Calabrese, J., Esponda, S., Pasini, A., Boracchia, M., & Pesado, P. (2019). Guía para evaluar calidad de datos basada en ISO/IEC 25012. *XXV Congreso Argentino de Ciencias de La Computación*, 694–706.

D’Ambrosio, S. (2018). *El Concepto de Datos - Monografias.com*. https://www.monografias.com/trabajos14/datos/datos.shtml

Drogodependientes, P., & En, I. (2018). *MINERIA DE DATOS APLICADA A LA DETECCION DE PATRONES DE RECLAMOS DEL SERVICIO DE AGUA POTABLE EN LA CIUDAD DE GUAYAQUIL*.

Espino, C., & Martínez, X. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. *26/27*, *I*(Principio activo y prestación ortoprotésica), 67. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/caresptimTFG0117memòria.pdf

Frank, E., Hall, M. A., Witten, I. H., & Kaufmann, M. (2016). *WEKA Workbench Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques.”* https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Fuentes, M. D. C. (2013). Notas del curso: Bases de Datos. In *Universidad Autónoma Metropolitana*. http://www.cua.uam.mx/pdfs/conoce/libroselec/Notas_del_curso_Bases_de_Datos.pdf

Gallardo, J. (2016). *Modelos de proceso para proyectos de Data Mining (DM) CRISP-DM*. http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037

García, M., & Álvarez, A. (2020). Análisis de Datos en WEKA–Pruebas de Selectividad. *It.Uc3M.Es*, 1–9. http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

Gironés Roig, J., Casas Roma, J., Minguillón, J., & Caihuelas Quiles, R. (2017). *Minería de datos: modelos y algoritmos*. https://dialnet.unirioja.es/servlet/libro?codigo=868986&info=resumen&idioma=SPA

Gonzalez Marcos, A. (2007). *MINERÍA DE DATOS: HERRAMIENTA DE APOYO EN LA SELECCIÓN DE EQUIPOS DE PROYECTOS INFORMÁTICOS*.

Haro Rivera, S., Zúñiga Lema, L., Meneses Freire, A., Vera Rojas, L., & Escudero Villa, A. (2018). Métodos De Clasificación En Minería De Datos Meteorológicos. *Perfiles*, *2*(20), 107–113. https://doi.org/10.47187/perf.v2i20.40

Haro, S., Pazmiño Maji, R., Conde, M., & Peñalvo, F. (2018). Minería de datos para descubrir tendencias en la clasificación de los trabajos de titulación. *Congreso de Ciencia y Tecnología ESPE*, *13*(1), 125–128. https://doi.org/10.24133/cctespe.v13i1.739

Hendrickx, T., Cule, B., Meysman, P., Naulaerts, S., Laukens, K., & Goethals, B. (2015). Mining association rules in graphs based on frequent cohesive itemsets. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9078*(3), 637–648. https://doi.org/10.1007/978-3-319-18032-8_50

Hernández, J., & Ramirez, M. (2004). *Libro de Data Mining "Introducción a la Minería de Datos."* http://dmip.webs.upv.es/LibroMD/

Huang, W., & Chen, Y. (2017). *The Multiset EM Algorithm*. http://www.elsevier.com/open-access/userlicense/1.0/

Humaid, E. (2017). *A Data Mining Based Fraud Detection Model for Water Consumption Billing System in MOG*. 79.

http://library.iugaza.edu.ps/thesis/106989.pdf

Lara, J. (2014). *Fundamentos y Aplicaciones Prácticas del Descubrimiento de Conocimiento en Bases de Datos*. http://repositorio.cedia.org.ec/bitstream/123456789/965/1/Guia_docente.pdf

leskovec, J., Rajaraman, A., & Ullman, J. (2020). *Minería de conjuntos de datos masivos 3ª edición | Reconocimiento de patrones y aprendizaje automático | Prensa de la Universidad de Cambridge*. https://www.cambridge.org/es/academic/subjects/computer-science/pattern-recognition-and-machine-learning/mining-massive-datasets-3rd-edition?format=HB&isbn=9781108476348

Mazón Olivo, B., Jaramillo Paredes, M., Romero Hidalgo, O., Borja Herrera, A., Martha, A. B., & Contento Segarra, M. (2018). Tecnologías de Inteligencia de Negocios y Minería de Datos para el Análisis de la Producción y Comercialización de Cacao. *Revista Espacios*, *39*(32), 6–21.

Mining, D., Artificial, I., Score, C., & An, E. (2016). *Tema 1: Minería de datos y extracción de conocimiento*.

Monjas, Y. B. (1999). *Minería de datos*.

PERALTA, M., MERMA, J., SOTO, C., & JIMENEZ, W. (2022). *Evaluación de la calidad de datos en un Sistema de Gestión Académica de una Universidad Peruana basado en el estándar ISO/IEC 25000*. 1–6. https://www.iiis.org/CDs2022/CD2022Spring/papers/CB951LT.pdf

Pérez, C., & Santín, D. (2006). *DATA MINING. SOLUCIONES CON ENTERPRISE MINER. INCLUYE CD-ROM - CESAR PEREZ LOPEZ; DANIEL SANTIN GONZALEZ - 9788478976959*. https://www.agapea.com/libros/DATA-MINING-

SOLUCIONES-CON-ENTERPRISE-MINER-INCLUYE-CD-ROM--9788478976959-i.htm

Pulido Romero, E., Escobar Dominguez, O., & Nunez Perez, J. A. (2019). *Base de datos*. Grupo Editorial Patria. https://elibro.net/es/lc/utnorte/titulos/121283

Rajaraman, A., Leskovec, J., & Ullman, J. (2014). Mining of Massive Datasets. In *Mining of Massive Datasets*. https://doi.org/10.1017/CBO9781139058452

Rebelo, F. J. P., Martins, F. F., M.R.D. Silva, H., & Oliveira, J. R. M. (2022). Use of data mining techniques to explain the primary factors influencing water sensitivity of asphalt mixtures. *Construction and Building Materials*, *342*, 128039. https://doi.org/10.1016/J.CONBUILDMAT.2022.128039

Salazar, J. (2018). *El clima organizacional y su relación con la calidad del servicio en la empresa Mercredi S.A.* 1–33. http://repositorio.uees.edu.ec/bitstream/123456789/2923/1/SALAZAR ALCIVAR JANET KARINE.pdf

Sierra, B. (2006). *Aprendizaje automático : conceptos básicos y avanzados : aspectos prácticos ... - Basilio Sierra Araujo - Google Libros*. https://books.google.com.ec/books?id=BCzUAQAACAAJ

Timaran, S. R., Hernandez, I., Caicedo, S. J., Hidalgo, A., & Alvarado, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías*, *8*(26), 37–47. https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230-1?inline=1#:~:text=El Descubrimiento de conocimiento en,que el usuario los analice.

Torres-Quezada, Y. (2021). Minería de datos para determinar los factores más

influyentes en la ocurrencia de siniestros de tránsito en Ecuador en el año 2020. *Cedamaz*, *11*(2), 124–132. https://doi.org/10.54753/cedamaz.v11i2.1181

Troncoso Espinosa, F. H., Paulina Gisselot, F. F., & Belmar Arriagada, I. R. (2020). Predicción De Fraudes En El Consumo De Agua Potable Mediante El Uso De Minería De Datos. *Universidad Ciencia y Tecnología*, *24*(104), 58–66. https://doi.org/10.47460/uct.v24i104.366

Vuotto, A., Di Césare, V., & Pallotta, N. (2020). Fortalezas y debilidades de las principales bases de datos de información científica desde una perspectiva bibliométrica. *Palabra Clave (La Plata)*, *10*(1), e101. https://doi.org/10.24215/18539912e101