UNIVERSIDAD NACIONAL AUTÓNOMA DE NICARAGUA, MANAGUA
UNAN - MANAGUA

Facultad de Ciencias e Ingeniería

Ingeniería Estadística

Tesis monográfica para optar al título de Ingeniero Estadístico

Tema:

Patrones de comportamiento inusuales en el consumo de agua de los usuarios de la Empresa Nicaragüense de Acueductos y Alcantarillados Sanitarios (ENACAL) del distrito V de Managua en el mes de julio del año 2017.

Autores:

Br. Gonzalo Alberto Lacayo Cuaresma.

Br. Julio David Martínez García

Tutor:

Msc. José David García.

Asesor metodológico:

Ing. Flor Ríos Laguna

Managua, noviembre 2018

Agradecimientos

Agradecemos a Dios por habernos dado la vida y por permitirnos llegar hasta este momento tan importante de nuestra formación académica.

Agradecemos a nuestros padres por el apoyo incondicional que nos han brindado, y reconocemos que han sido el pilar más importante en este proceso de estudios, que, sin sus sacrificios, nada de esto hubiese sido posible.

Agradecemos a todos nuestros maestros por las experiencias brindadas, por su paciencia y sus conocimientos. Les agradecemos por el esfuerzo que realizaron para mostrarnos el camino que debemos seguir para ser, cada día, mejores profesionales. Gracias por todas las vivencias creadas en nuestro entorno universitario, todo con el fin de formarnos como profesionales con ética, responsabilidad e innovación.

Resumen

La identificación de patrones de comportamiento inusual, no es un trabajo relativamente fácil, debido a la cantidad de factores que intervienen en los mismos. A como se analiza en este documento, el 95% de los usuarios de la Empresa Nicaragüense de Acueductos y Alcantarillados Sanitarios, entiéndase a partir de ahora como ENACAL poseen un servicio activo, a los cuales, se les puede asociar factores como la incidencia, problemas en la lectura de los medidores, el consumo, las horas de abastecimientos, entre otros; estos factores dan como resultado un perfil de comportamiento, donde no necesariamente es de riesgo. Un perfil de comportamiento se considera de riesgo, cuando posee características particulares definidas a nivel de criterio de experto basado en los resultados.

Los resultados de esta investigación se dividen en dos niveles. El primer nivel se enfoca directamente a identificar características similares en el universo de usuarios de la empresa ENACAL, y de esta manera lograr perfilar grupos de usuarios que tengan características inusuales, o poco frecuentes con respecto a los demás. El segundo nivel, es definir niveles de riesgos para cada grupo identificado, con base al criterio de experto.

Para lograr el primer nivel de esta investigación, se diseñó un modelo estadístico, que, considera los promedios de facturación de cada uno de los segmentos económicos, por lo cual, el modelo se divide en tres segmentos, cada uno dirigido a grupos de sectores económicos con promedios de facturación distintos; evitando que se sesgue por causa del volumen de consumo de cada usuario, y permitiendo valorar inusualidades, tanto en las personas naturales, como en las jurídicas.

Para el segundo nivel, se diseñó un modelo de árboles de decisión para caracterizar y valorar los niveles de riesgo establecidos en la segmentación. Esto permite que los tomadores de decisiones interpreten de forma más práctica y sencilla los resultados obtenidos.

Estos dos niveles de esta investigación se lograron realizar siguiendo la metodología CRISP-DM, metodología especializada en minería de datos para soluciones empresariales.

Índice

I. Introducción

En este trabajo se utiliza la metodología Cross Industry Standard Process for Data Mining, por sus siglas conocida como CRISP-DM, con el propósito de identificar patrones de comportamiento inusuales en el consumo de agua de los usuarios de la Empresa Nicaragüense de Acueductos y Alcantarillados Sanitarios (ENACAL) del distrito V de Managua en el periodo comprendido de julio a agosto del año 2017. Esta metodología surge como una iniciativa orientada principalmente a los negocios, para detectar riesgo, (Unidad de Información y Análisis Financiero [UIAF], 2014).

De esta forma la investigación se desarrolla en el contexto del análisis del comportamiento del consumo de agua potable, con la finalidad de descubrir características inusuales. Dentro de las principales anomalías se encuentran: medidores robados, consumo de agua potable desmesurado en comparación con el tipo de tarifa y el rubro que posee, esto sucede cuando no hay correspondencia entre el monto facturado y el volumen consumido. Otro escenario son los medidores trabajados, ya sea que tengan los sellos violados o con bypass. (Empresa Nicaragüense de Acueductos y Alcantarillados Sanitarios [ENACAL], 2006).

Para analizar esta problemática fué necesario llevar un orden metodológico en el que se utilizó técnicas estadísticas, mediante la descripción e identificación de grupos de usuarios que sean de interés para ENACAL, y consecuentemente se establecen perfiles que describieron las principales características de los individuos en cuestión, y concretamente proporcionar información para la correcta gestión por parte de la empresa.

El informe se organizó de la siguiente manera; primero, se presenta el respaldo teórico, es decir los conceptos más relevantes. Segundo, se muestra el detalle procedimental del estudio (metodología), seguido, se trabajó los resultados basados en las técnicas descriptivas de agrupación y predicción de minería de datos, y finalmente se exponen los resultados y anexos.

II. Antecedentes

El análisis de riesgo es una actividad de investigación muy importante y muchos estudiadores han realizado diversos análisis para lograr hacer un sistema de gestión del riesgo de la forma más adecuada. Entre los trabajos a nivel nacional tenemos los siguientes:

Chacón, Geraldine Judith. Método de clasificación para evaluar el riesgo crediticio: una comparación

Se comparan dos métodos clásicos de clasificación: Análisis de Regresión Logística y Árboles de Clasificación, con el método de Redes Neuronales. La comparación se realizó en base al poder de clasificación y predicción de los modelos obtenidos en la evaluación del Riesgo Crediticio, siendo Redes Neuronales el mejor método por tener mayor poder de clasificación y predicción. Para el análisis se utilizó una Base de Datos de Riesgo Crediticio. Asimismo, se establecen las ventajas y desventajas en el empleo de cada método.

Cruz Quispe, Lizbeth María. Detección de fraudes usando técnicas de clustering.

El fraude con tarjetas de crédito es uno de los problemas más importantes a los que se enfrentan actualmente las entidades financieras. Si bien la tecnología ha permitido aumentar la seguridad en las tarjetas de crédito con el uso de claves PIN, la introducción de chips en las tarjetas, el uso de claves adicionales como tokens y mejoras en la reglamentación de su uso, también es una necesidad para las entidades bancarias, actuar de manera preventiva frente a este crimen. Para actuar de manera preventiva es necesario monitorear en tiempo real las operaciones que se realizan y tener la capacidad de reaccionar oportunamente frente a alguna operación dudosa que se realice. La técnica de Clustering frente a esta problemática es un método muy utilizado puesto que permite la agrupación de datos lo que permitiría clasificarlos por su similitud de acuerdo a alguna métrica, esta medida de similaridad está basada en los atributos que describen a los objetos. Además, esta técnica es muy sensible a los outlier que se caracteriza por el impacto que causa sobre el estadístico cuando va a analizar los datos.

Ñaupas Caraza, Carol Maribel. Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias.

Propone a través de la aplicación de un proceso de descubrimiento de conocimientos en bases de datos, la generación de un modelo automático que permite clasificar las transacciones de la banca por internet y de la banca móvil de personas naturales de una entidad financiera, como fraudulentas o íntegras, mediante la aplicación de técnicas de minería predictiva basada en arboles de clasificación.

Flores Coaguila, Johanna Denise. Propuesta de modelo de detección de fraudes de energía eléctrica en clientes residenciales de Lima Metropolitana aplicando minería de datos

Desarrolla un modelo como propuesta para predecir potenciales situaciones de fraudes de energía eléctrica en clientes residenciales basado en aprender el comportamiento de clientes que anteriormente hurtaron. Para ello aplicaremos el proceso Minería de Datos para analizar, extraer y almacenar información de la base de datos, las cuales contienen la historia del consumo de energía. El modelo se propone apoyar a las empresas de distribución eléctrica, en especial a los técnicos eléctricos y verificar, de manera rápida y oportuna, los resultados obtenidos y que contribuya, de esta forma, en la toma de decisiones. Para la creación del modelo se utilizaron las redes neuronales por ser la de mejor desempeño en la detección. El modelo creado fue evaluado con datos de una empresa distribuidora de electricidad para el período 2009 y 2010. Las herramientas utilizadas para la creación del modelo fueron el Sql Server Management Studio (Database Engine para la base de datos y creación de procedimientos, Analysis Services para la creación de Estructuras y modelos) y como herramienta interactiva y de fácil entendimiento para el usuario el Complemento de Minería de Datos para Microsoft Excel.

Contreras Chinchilla, Leidys. Análisis del comportamiento de los clientes en las redes sociales mediante técnicas de minería de datos

En este documento, se describen los resultados de la utilización de técnicas de minería de datos para analizar el comportamiento de los clientes de una empresa de moda en la red social Instagram. La metodología utilizada fue CRISP-DM a través de la cual se evaluaron los modelos descriptivos utilizando las técnicas de reglas de agrupación y asociación. Los resultados muestran

que los modelos propuestos pueden proporcionar información útil para el diseño de estrategias de marketing apropiadas de acuerdo con las preferencias de los usuarios.

Maldonado Cadenillas, Rodrigo Ricardo. Proceso de extracción de patrones secuenciales para la caracterización de fenómenos espacio-temporales.

El objetivo de ese trabajo fué realizar un proceso de extracción de patrones secuenciales basado en la metodología Knowledge Discovery in Databases (KDD), empleando el algoritmo de minería de patrones secuenciales PrefixSpan para prever el comportamiento de fenómenos representados por eventos que cambian con el tiempo y el espacio. Estos tipos de fenómenos son llamados fenómenos espacio-temporales, los cuales son un conjunto de eventos o hechos perceptibles por el hombre. Además, están compuestos por un componente espacial (la ubicación donde sucede el fenómeno), un componente temporal (el momento o intervalo de tiempo en el que ocurre el fenómeno) y un componente de análisis (el conjunto de características que describen el comportamiento del fenómeno). En el mundo, se pueden observar una gran diversidad de fenómenos espacio-temporales; sin embargo, el presente se centra en los fenómenos naturales, tomando como caso de prueba el fenómeno espacio-temporal de la contaminación de los ríos en Reino Unido. Por lo tanto, con el fin de realizar un estudio completo sobre este fenómeno, se utiliza KDD (Knowledge Discovery in Databases) para la extracción del conocimiento a través de la generación de patrones novedosos y útiles dentro de esquemas sistemáticos complejos. Además, se utilizan métodos de Minería de Datos para extraer información útil a partir de grandes conjuntos de datos. Así mismo, se utilizan patrones secuenciales, los cuales son eventos frecuentes que ocurren en el tiempo y que permiten descubrir correlaciones entre eventos y revelar relaciones de "antes" y "después". En resumen, el presente trabajo se trata de un proceso para mejorar el estudio del comportamiento de los fenómenos gracias al uso de patrones secuenciales. De esta manera, se brinda una alternativa adicional para mejorar el entendimiento de los fenómenos espacio-temporales; y a su vez, el conocimiento previo de sus factores causantes y consecuentes que se puedan desencadenar, lo cual permitiría lanzar alertas tempranas ante posibles acontecimientos atípicos.

Giraldo Mejía, Juan Camilo. Aplicación de la Técnica Regresión Logística de la minería de datos en el proceso de Descubrimiento de Conocimiento (KDD) en bases de datos operativas o transaccionales.

El artículo presenta la caracterización de la técnica Regresión Logística de la Minería de Datos (Data Mining). Igualmente, se muestra la funcionalidad y aplicación de la técnica para apoyar al proceso de obtención de conocimiento (Knowledge Discovery in Databases o KDD), a encontrar información interesante a partir de Datos "ocultos". La funcionalidad de la técnica se ejemplifica con los resultados obtenidos en un trabajo de investigación que se realizó buscando encontrar el nivel de innovación y desarrollo tecnológico en algunas empresas de Colombia.

La finalidad es mostrar el proceso de obtención de conocimiento de un sistema de bases de datos Transaccionales u operativos para empresas de bienes y servicios. En él, se desarrollan los antecedentes conceptuales e investigativos y la caracterización de los conceptos fundamentales relacionados con el proceso de descubrir conocimiento, Minería de Datos y la Técnica de Regresión Logística.

Pizarro Solís, Pedro Arturo. Predicción del rendimiento académico en la Educación Superior usando minería de datos y su comparación con técnicas estadísticas

En la mayoría de las universidades se sigue un sistema de currículo flexible, esto significa que, a partir del segundo ciclo de estudios, los estudiantes universitarios pueden escoger los cursos a llevar, siempre y cuando se cumpla con el currículo y los reglamentos académicos correspondientes. Una gran dificultad en el proceso de inscripción es que el estudiante no tiene un sistema de ayuda o recomendación para tomar una buena decisión en la elección de los cursos a llevar, de tal manera que tenga la mayor probabilidad de salir airoso en su rendimiento académico. En este trabajo se aplican modelos predictivos (redes neuronales, regresión logística y regresión múltiple), que permitan al estudiante universitario predecir su rendimiento académico de cada curso en que desea inscribirse. El objetivo del estudio es predecir (a) si el alumno aprobará o no un curso o (b) la nota del curso. Para ello primero se realiza una selección de las variables predictores, en base a la experiencia de los autores en la cátedra universitaria y luego confrontando estas variables con las usadas en trabajos publicados relacionados al tema.

III. Justificación

El crecimiento vertiginoso de la población es proporcional a sus necesidades de abastecimiento de agua potable. Esto trae consigo una serie de desafíos que consisten en proporcionar un servicio eficiente.

Lo anterior, resalta la relevancia que resulta para la empresa ENACAL, contar con herramientas especializadas que permita identificar problemas potenciales con respecto al consumo de agua. Esto garantizaría un mejor servicio, lo que se traduce en un mejor aprovechamiento por parte de los usuarios y una mayor satisfacción.

En vista de que esta empresa no cuenta con un sistema de monitoreo que le permita detectar inusualidades. Se requiere de diversas metodologías que permiten desarrollarse en el campo de la minería de datos, estas resultan apropiadas para obtener resultados e información rápida y precisa, que ayuden a detectar patrones de comportamientos inusuales.

Todo ello facilitará la identificación y caracterización de estos usuarios potencialmente riesgosos y que a su vez representan pérdidas económicas para la institución, lo que se traduce en la reducción de costes de inversión y optimización de sus recursos.

Por tal razón, esta investigación es relevante por el impacto e importancia para aportar información relevante, la cual, permitirá desarrollar estrategias para la correcta gestión y toma de decisiones. Además, puede servir de base para la implementación de un sistema de monitoreo que pueda ser incorporado en el proceso de análisis.

IV. Planteamiento del Problema

De acuerdo al plan estratégico de desarrollo institucional de ENACAL [PEDI-ENACAL], (2013), a finales del año 2011, alrededor del 82.1 % de la población urbana en Nicaragua tiene acceso al suministro de agua potable, lo cual representa aproximadamente 2.8 millones de personas, que se distribuyen en 541,825 conexiones domiciliarias, de las cuales sólo el 52 % cuentan con redes en buenas condiciones.

El progresivo crecimiento de la población nicaragüense y el consecuente aumento de la demanda de agua potable ha ocasionado que muchos de los servicios se colapsen, y aumenten las exigencias para la institución. Induciendo al deterioro en la calidad y la posible exposición a múltiples situaciones de riesgo para la empresa.

Actualmente, la Empresa Nicaragüense de Acueductos y Alcantarillados Sanitarios no cuenta con un sistema de monitoreo eficiente que le permita identificar problemas potenciales asociados al consumo de agua del distrito V de Managua.

Todo ello conlleva a muchas problemáticas, dentro de las cuales se destacan la inadecuada gestión del servicio, debido a la falta de mantenimiento y por lo tanto obsolescencia de los medidores, lo que se traduce en medidores en mal estado, de igual forma existen daños en las redes, hay conexiones ilegales ocasionando que el usuario no se le mida el consumo, por lo tanto, no se tiene un control del agua no facturada, se presentan casos de fraude como la instalación de BYPASS, medidor aterrado, robos de objetos propiedades de ENACAL. De igual forma la baja recaudación que se traduce en usuarios morosos.

A partir de la delimitación del problema antes expuesto se plantea la pregunta principal de la presente investigación. ¿Cuáles son los perfiles asociados al comportamiento inusual en el consumo de agua, para los usuarios del distrito V de Managua en el mes de julio del año 2017?

V. Objetivos

V.1. Objetivo General:

Identificar patrones de comportamiento inusuales en el consumo de agua de los usuarios de la Empresa Nicaragüense de Acueductos y Alcantarillados Sanitarios (ENACAL) del distrito V de Managua en el mes de julio del año 2017

V.2. Objetivos específicos:

1. Describir el comportamiento en el consumo de agua en el distrito V de Managua.
2. Identificar grupos de usuarios con patrones de comportamiento inusuales en el consumo de agua, utilizando la metodología CRISP-DM.
3. Definir niveles de riesgo a partir de los perfiles de los usuarios.
4. Proponer perfiles de usuarios potencialmente riesgosos.

VI. Marco Teórico

VI.1. Concepción de ENACAL

La Empresa Nicaragüense de Acueductos y Alcantarillados, entiéndase adelante como ENACAL, es una empresa que se encarga de brindar el servicio de agua potable a las familias nicaragüenses.

Por lo tanto, "es la entidad pública encargada de implementar políticas… para el consumo humano de agua potable y el alcantarillado sanitario, el uso eficiente y racional de las fuentes de agua subterráneas y superficiales" (ENACAL, 2017).

VI.2 Historia de ENACAL

De acuerdo a la página oficial de esta empresa (ENACAL, 2017) (www.enacal.com.ni), se menciona su desarrollo histórico:

Como parte del proceso de la transformación de las empresas públicas en 1991, se inició el proceso de modernización de INAA, y en 1995 da comienzo el Programa de Reforma de las Empresas Públicas en el sector de Agua y Saneamiento.

En 1997 se produce a lo interno del INAA la separación de funciones: coordinación sectorial, fiscalización, regulación y operación de los servicios y para 1998 deriva la aprobación y promulgación de las leyes, decretos y reglamentos que conformarán el nuevo marco jurídico y legal del sector.

Amparados en las reformas de la Ley Orgánica del INAA publicadas en la Gaceta Diario oficial del 28 enero de 1998, Leyes No: 275 Y 276, se prosiguió el proceso de reestructuración del sector de agua potable y alcantarillado sanitario, a fin de dar cumplimiento a estas leyes, procediendo a establecer la separación de las actividades operativas, normativas y empresariales, constituyéndose dos órganos con subordinación directa de la Presidencia de la República, el INAA ente regulador y ENACAL empresa operadora de los servicios de agua potable y alcantarillado sanitario. Esta última se reestructura organizacionalmente para convertirse en una empresa operadora de los servicios de giro comercial.

VI.3 Servicios que brinda ENACAL

Como se menciona anteriormente ENACAL es una empresa de giro comercial, dentro de los servicios que ofrece se pueden destacar el servicio de alcantarillado, extracción de pozos y el de red de tuberías para el abastecimiento del agua potable.

VI.4 Situaciones de riesgo presentes en ENACAL

Lo cierto es que esta empresa se enfrenta a muchos retos, un artículo publicado en su página oficial deja claro que:

A finales del año 2007 existía, "… un total de 441,883 conexiones, el 45 % se encuentra en mal estado o sin medidor. La cobertura de ENACAL a finales de 2007 era muy limitada"

De igual forma hace un llamado exponiendo los principales problemas a los que se enfrenta, que suelen ser de gran riesgo para la empresa:

Dentro de las principales anomalías a las que se enfrenta ENACAL y que a su vez presenta factores de riesgo son: Medidores robados, consumo de agua potable desmesurado en comparación con el tipo de tarifa y el rubro que posee, esto sucede cuando no hay correspondencia entre el monto facturado y el volumen consumido. Otro escenario son los medidores trabajados, ya sea que tengan los sellos violados o con bypass.

Así mismo hace falta que las organizaciones ciudadanas y las instituciones de orden público colaboren más con ENACAL en la preservación de la red de distribución de agua y saneamiento, dado que la misma está siendo descapitalizada a causa de los robos, tanto de medidores como de tapas de manjoles y la destrucción de los hidrantes, todos ellos esenciales para prestar y medir los servicios. Esta situación adicionalmente provoca importantes pérdidas económicas a la empresa y pone en riesgo la vida de los transeúntes. (ENACAL, 2006).

Sin embargo, uno de los factores de principal interés y de mayor preocupación para esta empresa es la mora, esto es la evasión del pago del servicio de agua potable, para ello hay que hacer hincapié en los clientes morosos.

Usuarios morosos: Los usuarios morosos son aquellas personas que exceden del plazo de tiempo otorgado para pagar sus deudas. ENACAL considera que un cliente moroso es candidato a corte del servicio cuando posee dos o más facturas vencidas.

Por otra parte, buscando una definición más acertada, según el Diccionario de la Lengua Española, el concepto moroso "Se aplica al que se retrasa en un pago en la devolución de algo: arrendatario (o contribuyente) moroso". (Pere Brachfield, 2012)

Distrito V

Imagen 1. Mapa de los distritos de Managua

Fuente: Alcaldía de Managua

De acuerdo con el artículo publicado en el sitio web Índice de Managua, Nicaragua (MANFUF, s.f.) , establece:

El distrito V cuenta con una superficie de **82.6107** Kilómetros cuadrados. Además, cuenta con 156 Barrios de los cuales 39 son Residenciales, 38 Barrios Populares, 12 Urbanizaciones Progresivas, 61 Asentamientos Espontáneos y 6 Comarcas.

Por otra parte, este Distrito ha mostrado un importante crecimiento urbano en los últimos años, es aquí donde han surgido nuevas urbanizaciones dirigidas a las clases sociales de mayores ingresos económicos, sin mencionar que en este distrito se caracteriza por un comercio dirigido a la clase alta y media alta, se debe destacar que es la cara más moderna de la ciudad, formando corredores comerciales a lo largo de la carretera Masaya. Es en este territorio donde se han realizado las principales inversiones comerciales y de servicio en los últimos cuatro años por lo que se identifica como un lugar de gran potencial económico.

VI.5. Big Data

El siglo XXI es una época crucial, en donde estar informado es la clave fundamental para obtener mejores resultados en cada uno de los aspectos de la vida del ser humano. (Power Data, pág. 4)

En la actualidad se generan volúmenes de datos a cada instante que crecen en cantidades exponenciales. Esto se debe a que cada vez más, las personas son partícipes de herramientas tecnológicas, como los Smart phone, Tablet, computadoras en donde pueden registrar sus datos personales, ya sea en redes sociales o simplemente en servicios Online; y no solamente esto, sino que constantemente al hacer uso de algún servicio, ir de compras, conducir por la autopista, ir al médico, o simplemente por el hecho de ser un ciudadano de determinado país, se es una fuente de datos potencialmente relevante para determinados sectores.

El Big data se puede definir (Lucas, 2015) como el proceso de recolección y almacenamiento de grandes cantidades de datos, los cuales son difíciles de procesar mediante software y sistemas tradicionales.

También se puede considerar en términos sencillos como "Grandes volúmenes de datos de determinados individuos" Empresa Consultora de Tecnología, [CONSULTEC] (2017).

En efecto, estos conjuntos de datos son esenciales para una correcta gestión en cada uno de los procesos gerenciales. Y es que "es precisamente en ese tipo de datos donde las empresas han detectado que se encierra mayor valor…Para muchas empresas puede llegar a ser más importante detectar al cliente que más influye al resto de posibles compradores". (Power Data, pág. 4)

Según el analista Doug Laney de Gartner, 2001, (Natividad, 2016). El Big Data es, "El conjunto de técnicas y tecnologías para el tratamiento de datos, en entornos de gran volumen, variedad de orígenes y en los que la velocidad de respuestas es crítica"

VI.6 Características del Big Data

Según (CONSULTEC, Empresa Digital, pág. 8), el Big Data está relacionado con tres características:

Volumen: Grandes cantidades de datos provenientes de distintas fuentes. En la empresa de acueductos y alcantarillados se obtienen grandes volúmenes de datos a nivel nacional.

Variedad: Esta contiene, como se mencionó anteriormente información diversa, proveniente de distintas fuentes. Las empresas que se dedican al negocio de ventas de servicios, como ENACAL, tienen tablas de datos provenientes de distintos orígenes, ya sea información de clientes morosos, el registro de consumo de agua (lecturas en los medidores), horarios de abastecimiento, facturaciones, encuestas de calidad del servicio, entre otras., los cuales se almacenan en un Data Warehouse (almacén de datos).

Velocidad: En este aspecto, se relaciona con la frecuencia con la que se generan los datos, es decir, a un ritmo acelerado y en tiempo real. En la empresa mencionada diariamente se registra información proveniente de las facturaciones.

Por otra parte, uno de los enfoques del Big data, es establecer sistemas automatizados que generen, registren y almacenen periódicamente conjuntos grandes de datos. Actualmente, esto no representa un problema, ya que se cuenta con muchos recursos, como dispositivos con grandes capacidades de almacenamiento, software especializados, computadoras, entre otros.

VI.7. Minería de Datos (Data Mining)

La minería de datos, se puede resumir de forma coloquial como el interés por exprimir toda la información valiosa que se encuentra oculta en grandes montañas de datos.

Según Frawley (et. Al. 1992), citado por (Bernal, 2013, pág. 5) sugiere que la minería de datos es, "la tarea no trivial de extraer información implícita, previamente desconocida y potencialmente útil de bases de datos"

Por otra parte Jiawei Han, Micheline Kamber, (2001) citados por (Bernal, 2013, pág. 6), consideran el Data Mining como el proceso de descubrir conocimiento interesante de grandes cantidades de datos almacenadas en Data Warehouse u otros repositorios de información.

Para consolidar lo antes mencionado se puede citar a (Calderón, 2006, pág. 19) el cual afirma que:

La minería de datos es una nueva tecnología muy poderosa con un gran potencial para ayudar a las compañías a enfocarse en la información más importante en sus bases de datos o almacenes de datos. Las herramientas de minería de datos predicen comportamientos, permitiendo, la correcta gestión para la toma de decisiones.

Y sigue haciendo hincapié en que son:

"Un conjunto de recursos tecnológicos y técnicas estadísticas, los cuales tienen la finalidad de identificar patrones de comportamiento para extraer información interesante, novedosa y potencialmente útil de grandes bases de datos que puede ser utilizada como soporte para la toma de decisiones" (Calderón, 2006, pág. 19).

En otras palabras, la minería de datos es la sinergia de un conjunto de conocimientos que convergen en un objetivo principal, "torturar a los datos, para que confiesen" como lo expresa (Félix, 2002), en otras palabras, extraer información relevante.

En esta arquitectura intervienen los conocimientos avanzados en técnicas de manejo de datos a una escala multidimensional, software especializados y conocimiento en herramientas informáticas, el cúmulo de estos conocimientos se convierte en una herramienta adecuada, que permite encontrar patrones de comportamiento, perfiles de grupos de individuos, análisis de riesgo, ya sea para clientes morosos como para identificar posibles clientes para ofrecer determinada cartera de productos, segmentos rentables, modelos predictivos, todo ello dependerá de los objetivos pre-planteados.

Esto conlleva a tres aspectos básicos en los que se puede resumir Data Mining.

En primer lugar se inicia con un conjunto de datos, conocido como "la mínima unidad semántica, y se corresponden con elementos primarios de información que por sí solos son irrelevantes como apoyo a la toma de decisiones" (Sinnexus, 2007)

Posteriormente, al aplicar técnicas de análisis multivariante, esos datos tienden a tener sentido, resumiendo las grandes masas de datos en indicadores específicos.

Finalmente se encuentra el conocimiento, el cual es generado por la participación del experto en la materia de análisis, el cual le otorga sentido a la información para traer beneficios a la empresa.

VI.8. Técnicas Multivariantes de Minería de Datos

Actualmente, se cuenta con grandes sistemas de información que generan datos constantemente a diversos niveles de detalle, es allí donde se generan múltiples variables que describen las características de los individuos, sin embargo, se debe de prescindir de los recursos apropiados para potenciar los resultados que se obtengan de los insumos con los que se cuentan. Es por ello que, se necesitan técnicas y herramientas tanto matemáticas como informáticas que permitan responder a los objetivos de negocio de forma más acertada, evitando análisis someros.

Según (Quintanales, Moreno García, & García Peñalvo , 2001),

"La minería de datos ha dado lugar a una paulatina sustitución del análisis de datos dirigido a la verificación por un enfoque de análisis de datos dirigido al descubrimiento del conocimiento", por tanto, la aplicación de estos algoritmos de minería de datos permite detectar fácilmente patrones en los datos.

Antes de indagar acerca de las principales técnicas de análisis, es necesario conocer acerca del análisis multivariante.

VI.9. Análisis Multivariante de datos.

"El análisis multidimensional busca entregar al usuario final una manera fácil de representar la información, con varios componentes dimensionales o atributos, en una estructura común con la cual poder tomar decisión." (Álvarez, 2012)

En términos sencillos el análisis multivariante, consiste en analizar conjuntos de datos en los cuales interactúan simultáneamente muchas variables.

VI.9.1 Características

") citado por (Quintanales, Moreno García, & García Peñalvo , 2001)

Esto permite establecer que desde el punto de vista práctico y según la finalidad las técnicas de minería de datos se pueden resumir en dos aspectos esenciales. Técnicas descriptivas y técnicas predictivas

VI.9.2 Técnicas Descriptivas

Las técnicas descriptivas, se focalizan en encontrar patrones de comportamiento dentro de un conjunto de datos sin trascender al futuro, es decir su análisis son válidos para un momento determinado en el tiempo, lo que implica una toma de decisiones a corto plazo.

Según (Álvarez, 2012) se dice que estas técnicas,

> …no cuentan con un resultado conocido para poder guiar a los algoritmos, y por ello se conocen como modelos de aprendizaje no supervisado, donde el modelo se va ajustando de acuerdo a las observaciones o datos entregados, y se recurre muchas veces a argumentos heurísticos para evaluar la calidad de los resultados.

Dentro de las principales técnicas descriptivas se encuentran: los métodos de segmentación y de reducción de la dimensionalidad.

VI.9.3. Reducción de la dimensionalidad

Es común que, en estudios de minería de datos, se encuentren con un conjunto de escenarios que dificulten llevar a cabo los objetivos prefijados, en forma general, una de las situaciones características son la cantidad de variables con las que se pretende trabajar, pero el problema no solo radica aquí, sino en el hecho de que muchas de las variables no aportan información relevante o bien son redundantes, debido a que existen muchas variables que presentan fuertes correlaciones.

VI.9.4. Correlaciones

Las correlaciones se pueden considerar como el grado de asociación que tienen dos variables, ya sea cuantitativa o cualitativa, el grado de relación es posible medirlos mediante indicadores determinados.

Para las variables cuantitativas se utiliza la *correlación de Pearson* y para las variables cualitativas se utiliza medios visuales como los gráficos de asociación (malla) o bien los indicadores de *coeficiente de contingencia*, *Phi, V de Cramer y Tau b de Kendall*.

VI.9.5. Los modelos de Segmentación

Los modelos de segmentación tienen una característica peculiar, no exigen ninguna variable dependiente sobre la cual basar su análisis, por el contrario, su algoritmo consiste en detectar conjuntos con características similares para cada uno de sus elementos y completamente disjuntos entre todos los conjuntos formados. Dentro de los principales

modelos de segmentación se encuentran: las redes de *Kohonen*, la agrupación en clústeres de *K-medias*, la agrupación en *clústeres bi-etápicos.*

Para corroborar la información antes descrita se puede mencionar un apartado publicado por International Bussiness Machines [IBM] (2014), el cual dice que:

> Los modelos de agrupación en clústeres se centran en la identificación de grupos de registros similares y en el etiquetado de registros según el grupo al que pertenecen. No hay respuestas correctas o incorrectas para estos modelos. Su valor viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones. Estos modelos se usan con frecuencia para crear agrupaciones que se usarán posteriormente como entradas en análisis posteriores.

Modelo Bi-etápico

"El procedimiento de análisis conglomerado en dos fases, también llamado bi-etápico, es una herramienta de exploración diseñada para descubrir las agrupaciones naturales de un conjunto de datos, permitiendo así la generación de criterios de información" (Pérez, 2011) citado por (Rubio Hurtado & Vila-Baños, 2016, pág. 1)

El método de análisis de conglomerados en dos fases tiene unas características únicas respecto a otros métodos de segmentación tradicionales, que son las siguientes: un procedimiento automático del número óptimo de conglomerados, la posibilidad de crear modelos de conglomerados con variables tanto categóricas como continuas y la opción de trabajar con archivos de gran tamaño.

Modelo K-medias

"*K-medias* es un algoritmo no supervisado, que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza mimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster" (Uniovideo, S.f)

Dentro de las principales características es que trabaja únicamente con variables cuantitativas utilizando el algoritmo de las distancias cuadráticas para realizar los cálculos.

Modelo de Kohonen (Mapas auto-organizativos)

Estos modelos son una extensión de los modelos de segmentación, su característica singular es que son redes neuronales que trabajan bajo un procedimiento no supervisado. Se le denomina de esa forma porque se encontró que presentaban un comportamiento similar al del cerebro (Centro informático de Científico de Andalucía) [SAEM Thales-CICA] (s.f.).

Y continúa citando

"es un modelo de red neuronal con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro… de forma que la información captada… se presenta internamente en forma de mapas bidimensionales"

> Un modelo SOM está compuesto por dos capas de neuronas. La capa de entrada (formada por N neuronas, una equivalente por cada variable de entrada), se encarga de recibir y transmitir a la capa de salida la información procedente del exterior. La capa de salida (formada por M neuronas) es la encargada de procesar la información y formar el mapa de rasgos. Normalmente, las neuronas de la capa de salida se organizan en forma de mapa bidimensional. (Los mapas auto-organizados de Kohonen (SOM), S.f)

Técnicas Predictivas

El análisis predictivo permite encontrar las relaciones entre el conjunto de datos de entrada y la respuesta que tendrá determinado individuo cuando se cumplan ciertas características, por lo general se utiliza una base de datos para el entrenamiento, el cual consiste explicar mediante patrones de comportamiento una variable específica la cual indicará, donde se clasificarán las nuevas unidades de análisis, pero antes se deben validar mediante un conjunto de datos prueba, y establecer índices que prueben la eficiencia y precisión del modelo, debido a esto también se le conoce como técnicas de aprendizaje supervisado.

Dentro de las principales técnicas predictivas se encuentran: Los métodos de clasificación y de asociación.

Métodos de Clasificación

"Los modelos de clasificación usan el valor de uno o más campos de entrada para predecir el valor de uno o más campos de destino. Estos modelos ayudarán a las organizaciones a predecir un resultado conocido" International Bussiness Machines [IBM-SPSS Modeler] (2017)

Dentro de las técnicas más comunes son: árboles de decisión, máquinas de soporte vectorial y modelo discriminante.

Modelo discriminante

El análisis discriminante se basa primordialmente en entrenar un conjunto de datos, establecer reglas de validación para comprobar la robustez del modelo mediante un conjunto de entrenamiento y posteriormente la aplicación para predecir nuevos elementos.

Máquinas de soporte Vectorial

Para poder trabajar con una SVM, es necesario realizar un proceso de aprendizaje que permita encontrar un hiper-plano de separación de las clases del conjunto de datos. Otra de las características inherentes, es que, a diferencia del análisis discriminante, si puede trabajar con variables cuantitativas y cualitativas.

Árboles de decisión

Un árbol de decisión es el desenlace de un conjunto de reglas que permiten llegar a una conclusión para tomar una decisión.

De acuerdo a un artículo publicado en el 2013 (Alto Nivel, 2013), se llega a una conclusión:

"De acuerdo con los expertos, el árbol de decisión puede ayudarte a encontrar la mejor respuesta, de la mejor forma y bajo el mejor presupuesto"

Por lo tanto, se puede deducir que el objetivo principal es emplear distintos paradigmas que nos permitan ver el problema bajo distintas circunstancias y a su vez establecer una probabilidad asociada a cada posible decisión, lo que implica ver la situación desde distintas perspectivas, analizando todas las posibles soluciones.

Consideraciones generales

"La aplicación de los algoritmos de minería de datos requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada debido a que, en muchas ocasiones dichos datos proceden de fuentes heterogéneas, no tienen el formato adecuado o contienen muchos valores atípicos. (Cabena et al., 1998)".

Es importante tener en cuenta este aspecto, debido a que todo procedimiento requiere llevar una estructura que cumpla con los requerimientos estadísticos.

VI.10. Funciones de la minería de datos

Con el fin de mostrar el alcance de estos procesos se puede mencionar, de acuerdo a la referencia tomada de (Tecnologías de Información, 2016):

La minería de datos es el proceso de análisis de datos, de los cuales se pretende obtener información útil, con la finalidad de argumentar patrones de comportamiento en distintos campos de aplicación.

Lo que indica que tiene una gran relevancia en el análisis, tanto desde el punto de vista descriptivo como predictivo y una amplia aplicación en distintos campos de acción, por otro lado, se menciona que:

> La minería de datos se utiliza sobre todo hoy en día por las empresas con un enfoque fuerte en los consumidores, comunicaciones, comercio, finanzas y las organizaciones de comercialización. Esto permite que las empresas determinen las relaciones entre los factores externos… también de los internos…Por último, les permite profundizar en la información resumida para ver datos detallados...,(Calderón, 2006, pág. 19).

No cabe duda de que la minería de datos es cada vez más incipiente y primordial en la contribución de análisis de comportamientos inusuales y no solo eso, sino que es útil para identificar oportunidades financieras, estrategias de ventas, mercados potenciales, ect., existen muchas aplicaciones, pero de lo que si se está seguro es que la minería de datos es fundamental producir información relevante, que en conjunto con los criterios de los expertos se transformará en conocimiento.

VI.11. Metodologías de minería de datos

Dentro del análisis de grandes volúmenes de datos debe existir una estructura que permita guiar el proceso de investigación y determinación de conocimiento para la toma de decisiones.

Según un artículo publicado por la Universidad Nacional de Colombia, existen "tres metodologías dominantes para el proceso de la minería de datos, las cuales son*: KDD, SEMMA y CRISP-DM*. (Guzmán, S.f.). Es necesario mencionar que existen múltiples metodologías, pero estas son las más usadas:

VI.11.1. KDD

La Unidad financiera de Colombia (UIAF), define el KDD como "etapas… recursivas, es decir, que se retoma a ellas una y otra vez (proceso iterativo), a medida que se obtienen resultados preliminares que requieran replantear las variables iniciales." (UIAF, 2014).

Es una metodología propuesta por Fayyad en 1996 citado por (Guzmán, S.f.), propone 5 fases: Selección, pre procesamiento, transformación, minería de datos y evaluación e implementación. Este es un proceso iterativo e interactivo.

VI.11.2. SEMMA

"SEMMA es el acrónimo a las cinco fases: (Simple Muestreo, Explore, Modify, Model, Assess). La metodología es propuesta por SAS[1]" (Guzmán, S.f.)

[1] **SAS: Sistema de análisis estadístico (Statistical Analysis System)**

Estas fases se traducen como; muestreo, exploración, modificación, modelado y evaluación.

Institute Inc. citado por (Guzmán, S.f.), la define como "…proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones desconocidos…"

VI.11.3. CRISP DM

Los orígenes de CRISP-DM, se remontan hacia el año 1999 cuando un importante consorcio de empresas europeas tales como NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, SPSS, y Daimer-Chrysler, proponen a partir de diferentes versiones de KDD, el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining) (Arancibia, s.f)

CRISP-DM surge como una iniciativa de la metodología KDD, por lo cual se convierte en una generalización orientada principalmente a los negocios, para detectar riesgo.

Cross Industry Standard Process for Data Mining, comúnmente conocida por sus siglas CRISP-DM. Es una iniciativa promovida por la comunidad Europea, que plantea dos objetivos específicos según la (UIAF, 2014):

"-Fomentar la interoperabilidad de las herramientas a través de todo el proceso de minería de datos"

"-Eliminar la experiencia misteriosa y costosa de las tareas simples de minería de datos"

Lo anterior hace referencia a que todo el proceso lleva una secuencia lógica que facilita la planeación y gerencia de la investigación y a su vez, permite automatizar aquellos procesos que son por naturaleza mecánicos.

Las fases constitutivas de esta metodología son seis.

Comprensión del negocio: Es aquí donde se establecen los objetivos del negocio basado en el contexto o necesidades del negocio, para esto se requiere de conocer bien cada uno de las problemáticas y hacer una evaluación de la situación.

Comprensión de los datos*:* En esta fase el investigador debe familiarizarse con los datos teniendo en cuenta los objetivos pre-planteados, en otros términos, es la recopilación inicial de los datos y familiarización con la fuente de origen de los mismos.

Preparación de los datos*:* En esta etapa se debe realizar una serie de actividades como la limpieza de los datos, reformulación de las variables a partir de las ya existentes.

Modelado*:* Aquí se debe seleccionar la técnica de minería de datos que más se ajuste a los objetivos. Es necesario aclarar que no existe un modelo considerado como el mejor, sino que cada técnica se debe seleccionar de acuerdo con el tipo de datos, los objetivos y la experiencia del investigador.

Evaluación*:* Es entonces, cuando se debe examinar juiciosamente los resultados de las etapas anteriores, principalmente los resultados del modelado y verificar si es coherente con la realidad.

Distribución*:* Es aquí donde se utilizarán los resultados obtenidos para incorporarlas a las actividades de toma de decisiones, presentar un informe final y el perfilamiento de las unidades de análisis de interés, sobre los cuales se tomará acciones o monitoreos posteriores.

El proceso de la metodología CRISP DM es iterativo e interactivo en cada una de sus fases, es decir, se pueden volver a repetirse.

VII. Hipótesis

La identificación y caracterización de patrones de comportamientos inusuales de los usuarios del distrito V de Managua con respecto al consumo de agua potable, permitirá establecer factores de riesgos potenciales con una mayor precisión y de esta manera brindar mayor información para la creación de controles que mitiguen estos riesgos, logrando así reducir las pérdidas y aumentar la calidad del servicio y prestigio de la empresa.

VIII. Diseño Metodológico

El marco metodológico "se refiere al conjunto de procedimientos implícitos en el proceso de investigación; ofrece un procedimiento ordenado para que los resultados obtenidos sean consistentes y confiables" (L. Navarro, 2009).

Los aspectos que se deben tomar en cuenta son: Tipo te estudio, Definición de las variables, Población y muestra, Herramientas informáticas, Técnicas de análisis y procesamiento de datos.

VII.1. Tipo de estudio

Para definir el tipo de investigación que se desarrolló se tomaron tres aspectos: La profundidad de la investigación, el diseño y el propósito. La presente investigación, en base a su profundidad o alcance se define como investigación exploratoria, debido que es el primer acercamiento al fenómeno a estudiar, está estructurada siguiendo un diseño transversal, puesto que, los datos que se analizaron fueron recolectados en un punto en el tiempo y con respecto al propósito, se define como una investigación aplicada para la resolución de problemas y la toma de decisiones.

VIII.2. Definición de las variables

Nombre de variable	Definición	Fuente
Barrio	Barrio en el que habita el cliente	Estadísticas comerciales, ENACAL
Número de cuenta	Numero-Cuenta del cliente	
Economía	La economía se puede clasificar en varias categorías dentro de las cuales figuran: Asentamiento, Comercial, Domiciliar, Gobierno central, Gobierno descentralizado, Industrial, Multifactura Particular, Residencial, Sector Público y Urbanización social	
Consumo	El consumo de agua por cuenta, esta es medida en metros cúbicos	
Observaciones	Esta variable presenta los diversos tipos de situaciones en las que se puede encontrar un medidor. Ya sea en mal estado, con fuga o manipulado.	
Tipo de servicio	Esta variable indica si el cliente tiene servicio de agua potable o extracción de agua en pozos	

Incidencias	Son nuevos casos reportados con algún tipo de anomalías en cuanto al servicio de agua potable
Tarifa	La tarifa que establece ENACAL, está sumamente relacionada con el tipo de actividad económica a la que se vincula la cuenta
Rubro	Es el tipo específico de actividad económica a la que se dedica el cliente
Número de documentos vencidos	Es la cantidad de facturas vencidas
Mora	Es el monto moroso expresado en córdobas
Estado conexión alcantarillado	Expresa si la conexión al servicio de alcantarillado está activo o no
Estado actual de la conexión	Si el servicio de agua potable se encuentra activo o no

VIII.3. Población

La investigación se realizó en base a los registros de consumo de agua del mes de julio de la población de usuarios del distrito V de Managua, el cual consta de 45,930 registros. La Minería de Datos es el análisis de grandes volúmenes de datos, por esta razón, no se extrajo una muestra de la población, si no, que se analizó todos los registros.

VIII.4. Herramientas informáticas

Para el análisis y procesamiento de los datos se utilizó los paquetes informáticos como: Microsoft Office 2016 (Word, Excel, Power Point), SPSS Stadistic versión 22, SPSS Modeler versión 18. Los cuales son paquetes informáticos especializados para procesamiento de datos y presentación de informes.

VIII.5 Procesamiento de datos y técnicas de análisis

El procesamiento de los datos se realizó bajo la metodología Cross Industry Standard Process for Data Mining (CRISP-DM), metodología promovida por la comunidad europea, dirigida al análisis especializado para la detección de riesgos en negocios.

La metodología CRISP-DM establece o se estructura en seis faces: Comprensión del negocio, Comprensión de los datos, Preparación de los Datos, Modelado, Evaluación y Despliegue.

VIII.5.1. Comprensión del negocio: En esta primera etapa se realizó las siguientes actividades

- ✓ Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)
- ✓ Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio, entre otras actividades)
- ✓ Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)
- ✓ Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

VIII.5.2. Comprensión de los datos: En esta segunda etapa se realizó las siguientes actividades

- ✓ Recopilación inicial de datos
- ✓ Descripción de los datos
- ✓ Exploración de los datos
- ✓ Verificación de calidad de datos

VIII.5.3. Preparación de los datos: En esta tercera etapa se realizó las siguientes actividades

- ✓ Selección de los datos
- ✓ Limpieza de datos
- ✓ Construcción de datos
- ✓ Integración de datos
- ✓ Formateo de datos

VIII.5.4. Modelado: En esta cuarta etapa se realizó las siguientes actividades

- ✓ Selección de la técnica de modelado
- ✓ Diseño de la evaluación

- ✓ Construcción del modelo
- ✓ Evaluación del modelo

VIII.5.5. Evaluación: En esta quinta etapa se realizó las siguientes actividades

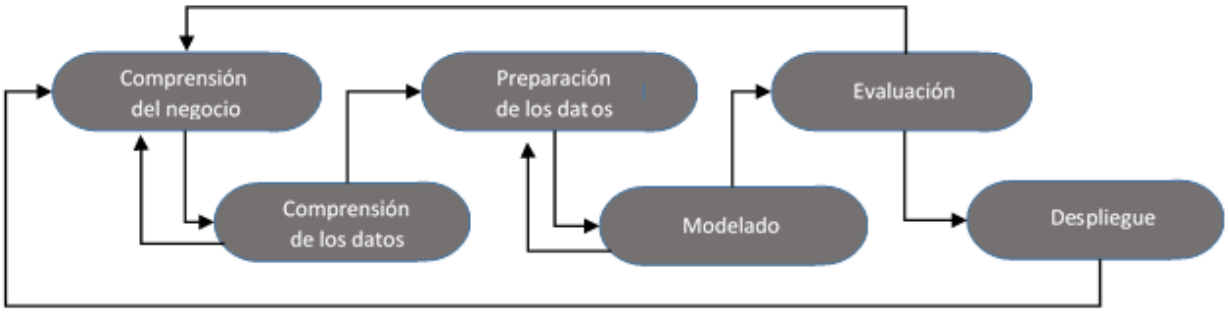
- ✓ Evaluación de resultados
- ✓ Revisar el proceso
- ✓ Establecimiento de los siguientes pasos o acciones

VIII.5.6. Despliegue: En esta última etapa se realizó las siguientes actividades

- ✓ Planificación de despliegue
- ✓ Planificación de la monitorización y del mantenimiento
- ✓ Generación de informe final
- ✓ Revisión del proyecto

Esta metodología tiene una estructura cíclica que es útil para los análisis de negocio y permite garantizar los mejores resultados para darle solución a los problemas o los fenómenos que se presenten.

Figura 2: Metodología CRISP-DM



VIII.6. Aspectos administrativos

Es importante analizar los recursos disponibles y necesarios para la realización de la investigación, debido a que influyen directamente en el alcance y duración de la misma.

Como aspectos administrativos es importante tomar en cuenta los recursos humanos necesarios, recursos económicos o monetarios y la disponibilidad de tiempo para la realización y la finalización de la investigación. Estos tres factores se describen a continuación:

VIII.6.1. Recursos humanos

Los recursos humanos son todas las personas que tienen actividades dentro de la investigación; las cuales se describen en la tabla siguiente:

Tabla 1

Recursos humanos necesarios para la realización de la investigación

Participante	Descripción de actividad	Objetivo de la actividad
Br. Gonzalo Alberto Lacayo Cuaresma	Análisis, Estructuración y diseño	Garantizar la calidad y finalización de la investigación
Br. Julio David Martínez García	Análisis, Estructuración y diseño	Garantizar la calidad y finalización de la investigación
Msc. José David Gutiérrez	Tutoría profesional y seguimiento	Verificar la calidad e integridad de la investigación
Msc. Leonardo Estradas	Asesoría profesional	Revisar los aspectos metodológicos involucrados
Lic. Alejandra Ramírez	Facilitador de datos	Dar acceso a las fuentes de datos
Lic. Melvin Hooker	Evaluación y criterio	

IX. Resultados de la Aplicación de la Metodología CRISP-DM

IX.1. Comprensión del negocio

La empresa ENACAL (Empresa Nicaragüense de Acueductos y Alcantarillados Sanitarios) es una empresa que tiene como objetivo garantizar un servicio de agua potable y un sistema de alcantarillado de calidad, brindarles a los usuarios posibilidades de mejoras continua de los sistemas de distribución.

Asimismo, ENACAL tiene el objetivo de mantener un sistema de monitoreo que garantice la detección de fraudes y de esta forma evitar debilidades en el servicio que brindan.

Actualmente, la empresa ENACAL cuenta con un sistema de monitoreo, el cual, está formado por un conjunto de reglas cualitativas basadas en la actividad económica, el tipo de consumo, el tipo de servicio y el segmento social al que pertenecen los usuarios. Este sistema de monitoreo ha demostrado ser eficiente para encontrar problemas de abastecimientos en los distintos barrios, pero, no es muy eficiente en la detección de fraudes.

Los analistas de ENACAL están interesados en conocer el impacto de la Minería de Datos en su sistema de monitoreo, por ende, les interesa diseñar propuestas de modelos basados en la misma.

Estos modelos tendrán como objetivo detectar el comportamiento con relación a la morosidad de los usuarios, y de esta manera detectar con anticipación grupos de usuarios que estén propensos a cometer fraudes.

Para la creación de estos modelos se utilizará la Metodología CRISP-DM y se procesarán con el programa estadístico SPSS-Modeler; También, se asignará un grupo de analistas que estarán a cargo de este proceso.

IX.2. Comprensión de los Datos

ENACAL consta de varias fuentes de recolección de datos, tales como: El chequeo de los medidores, la facturación mensual que realizan los usuarios y una encueta de opinión que se aplica anualmente a los mismos. Estas tres fuentes de recolección de datos, son las principales que se utilizaron para el análisis, las cuales, son recolectadas por los analistas de ENACAL e integradas en una sola base de datos y posteriormente se almacena en su servidor privado.

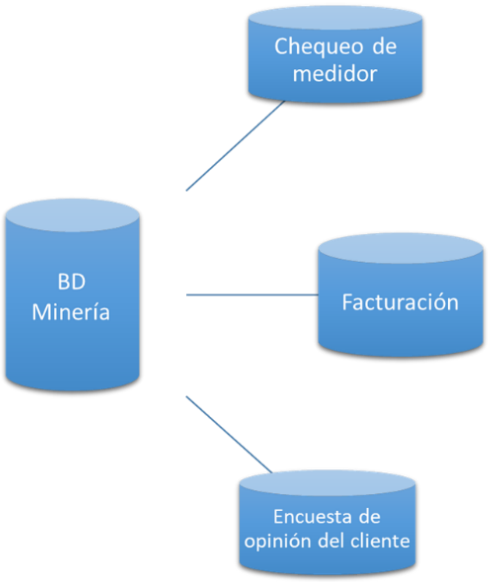


Figura 4: Fuentes de recolección de datos y su almacenamiento

En esta base de datos podemos encontrar 42 variables, tanto cualitativas como cuantitativas. Entre las 42 variables obtenemos, los códigos de barrio y de usuario, el segmento social al que pertenece cada usuario, la actividad económica, los montos de facturación y de mora, entre otras variables.

Con respecto a los montos de facturación podemos analizar que el 81% de los usuarios facturan mensualmente un monto menor a 500 córdobas, y solamente un 3% facturan un monto mayor a 2,000 córdobas. Estos datos se pueden observar a continuación en la tabla 3 y en el gráfico 1 en anexos.

Tabla 3

Montos facturados por los usuarios del Distrito V de Managua en el mes de julio del año 2017

Intervalos	Cantidad de Usuarios	%
Menor a 500	37,016	81%
500 - 1,000	5,849	13%
1,000 - 1,500	1,155	3%
1,500 - 2,000	547	1%
Más de 2,000	1,363	3%
Total	45,930	100%

Fuente: Base de datos ENACAL facturación julio 2017

Respecto a los años de antigüedad de los usuarios, podemos analizar que el 92% tienen menos de 5 años de antigüedad, el 5% entre 5 y 10 años, el 1% entre 10 y 20 años y solamente un 2% superan los 20 años de antigüedad. Con este dato podemos caracterizar a los usuarios como relativamente nuevos, lo que ocasiona que el riesgo de fraudes en general aumente considerablemente. Este dato se puede analizar en la tabla 4 a continuación y en anexos en el grafico 2.

Tabla 4

Años de antigüedad los usuarios del Distrito V de Managua en el mes de julio del año 2017

Intervalos	Cantidad de Usuarios	%
Menor a 5	42,034	92%
5- 10	2516	5%
10 – 20	325	1%
Más de 20	1055	2%
Total	45,930	100%

Fuente: Base de datos ENACAL facturación julio 2017

Para todo negocio, el impacto que causan los fraudes es lo más importante, y se debe analizar su evolución a través del tiempo, pero muchas veces este dato es relativamente imposible de obtener, por ende, este análisis se vuelve difícil de realizar. Por esta razón, una forma de tener una noción de la evaluación del impacto de fraudes, es controlando a través del monitoreo y análisis de los riesgos de que este se materialice. Una variable que se puede utilizar para monitorear el fraude, es la mora a la que se someten los usuarios al no cancelar sus facturas en tiempo.

Al corte de la facturación realizada por usuarios del distrito V de Managua en el mes de Julio 2017, se presentó una mora total de 199,659,674 córdobas, de la cual, se ha exonerado por jubilación 933,670 córdobas y las exoneraciones con el 100% acumulan 255,914 córdobas; en total se ha exonerado 1,189,584 córdobas, lo que representa 0.6% de la mora total. Este dato se puede analizar en la tabla 5, a continuación.

Tabla 5

Monto de mora por tipo de exoneración de los usuarios del Distrito V de Managua al corte de julio 2018

Tipo de Exoneración	Monto por Mora	% de usuarios
Sin Exoneración	198,470,091	97.7%
Jubilados	933,670	2.0%
Exoneración 100%	255,914	0.5%
Total	199,659,675	100%

Fuente: Base de datos ENACAL facturación julio 2017

También podemos analizar que al corte de julio 2017, el 79% de los usuarios tienen acumulado un monto de mora menor a 2,000 córdobas, el 16% entre 4,000 y 10,000 córdobas y solamente un 6% superan los 10,000 córdobas. Esto se puede analizar en la tabla 6 a continuación y en el grafico 3 en anexos.

Tabla 6

Monto de mora de los usuarios del Distrito V de Managua al corte de julio 2018

Intervalos	Cantidad de Usuarios	%
Menos de 2,000	36,120	79%
2,000 - 4,000	2,715	6%
4,000 - 6,000	1,736	4%
6,000 - 8,000	1,677	4%
8,000 - 10,000	807	2%
Más de 10,000	2,875	6%
Total	45930	100%

Fuente: Base de datos ENACAL facturación julio 2017

Las variables numéricas son muy importantes para realizar análisis de riesgos, asimismo, las variables cualitativas proporcionan gran cantidad de información que son muy útiles para la segmentación y monitoreo.

En este estudio se tomaron en cuenta, 9 variables cualitativas, las cuales son: Segmento económico, Tipo de servicio, Tipo de Conexión, Estado de la Conexión, Horas de abastecimiento, Estado del servicio, Incidencia, Tipo de consumo y El nivel de exoneración aplicado a ciertos usuarios.

Con respecto al segmento económico al que pertenecen los usuarios, podemos analizar que el 91.89% de los usuarios pertenecen a los segmentos de Domicilios, Residencias y Asentamiento, y solamente el 8.11% pertenecen a los demás segmentos. Esto lo podemos observar en la tabla 7 a continuación y en el grafico 5 en Anexos.

Tabla 7

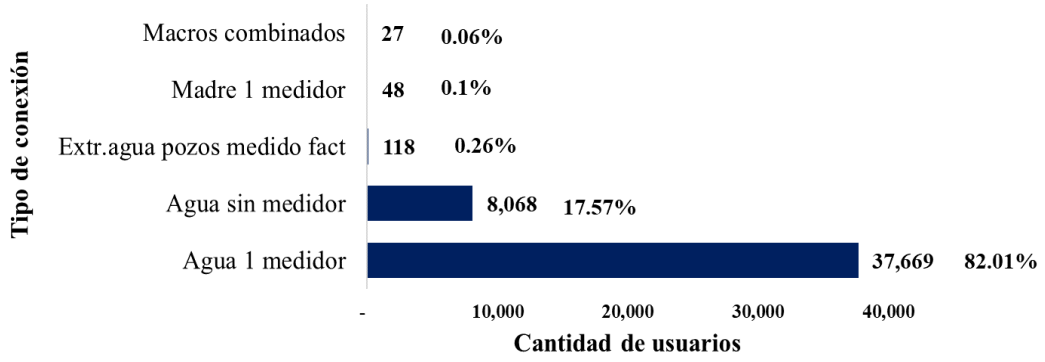
Segmento económico al que pertenece los Usuarios del Distrito V de Managua en el mes de julio 2017

Segmento Económico	Cantidad de Usuarios	%
Domiciliar	19,417	42.28%
Residencial	13,497	29.39%
Asentamiento	9,292	20.23%
Comercial	3,444	7.50%
Multifactura particular	91	0.20%
Gobierno central	80	0.17%
Sector publico	68	0.15%
Industrial	22	0.05%
Gobierno descentralizado	13	0.03%
Urbanización social	6	0.01%
Total	45,930	100.00%

Fuente: Base de datos ENACAL facturación julio 2017

También, con respecto al tipo de conexión del servicio que poseen los usuarios, podemos analizar que 82.01% de cuentan con una conexión de agua con un medidor y el 17.57% cuentan con una conexión de agua sin medidor. Esto se puede observar en el grafico 6 a continuación y en la tabla 8 en anexo.

Grafico 6 **Tipo de conexión que poseen los usuarios del distrito V de Managua registrado en el mes de julio del año 2017**



Fuente: Base de datos ENACAL facturación julio 2017

De estas conexiones de servicio, no todos los usuarios tienen la integridad de los medidores en condiciones adecuadas, el 5% de los usuarios poseen un medidor totalmente demolido. Este dato se puede observar en la tabla 9 a continuación y en el gráfico 7 en anexo.

Tabla 9

Estado de la conexión del medidor de los Usuarios del Distrito V de Managua en el mes de julio 2017

Estado	Cantidad de Clientes	%
Activo	43,592	95%
Demolido	2,338	5%
Total	45,930	100%

Fuente: Base de datos ENACAL facturación julio 2017

La condición de los medidores ocasiona que el estado del servicio varíe, y al corte de julio del año 2017, el 71% de los usuarios poseen un estado de servicio bueno, el 18% posee un servicio con cuota fija, el 7% tienen un servicio malo, pero sin arreglo, el 2% poseen un servicio sin medidor y con conexión directa y el 2% se encuentran con el servicio cortado. Este dato se puede observar en la tabla 10 a continuación y en el gráfico 8 en anexo.

Tabla 10

Estado del servicio que poseen los Usuarios del Distrito V de Managua en el mes de julio 2017

Estado	Cantidad de Clientes	%
Buenos	32,525	71%
Cuota Fija	8,068	18%
Malos	3,424	7%
Servicio Directo	997	2%
Cortado	916	2%
Total	45,930	100%

Fuente: Base de datos ENACAL facturación julio 2017

Una de las variables que es importante analizar es la hora de abastecimiento en la que los clientes reciben el servicio; este dato podría darnos una forma de como evaluar la satisfacción de los usuarios con respecto al servicio, y analizar el impacto que esto tiene a la posibilidad de fraudes.

Para el corte de julio 2017 el 72% de los usuarios del distrito V de Managua tienen un abastecimiento menor a 2 horas al día, el 5% de 3 a 5 horas al día, el 6% 6 a 20 horas al día y solamente el 17% de los usuarios tienen un abastecimiento mayor a las 21 horas al día. Esto se puede observar en la tabla 11 a continuación y en el grafico 9 en anexos.

Tabla 11

Horas de abastecimiento que poseen los Usuarios del Distrito V de Managua en el mes de julio 2017

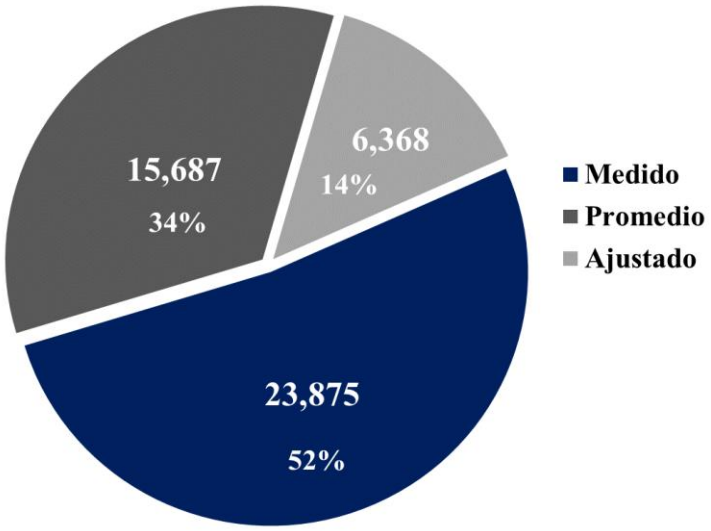
Periodos	Cantidad de Clientes	%
CERO A 2	33,251	72%
21 A MAS	7,812	17%
6 A 20	2,745	6%
3 A 5	2,122	5%
Total	45,930	100%

Fuente: Base de datos ENACAL facturación julio 2017

Debido a todas las variables que se analizaron en esta investigación, a los usuarios se les aplica un tipo de consumo que está ligado a todas las condiciones del medidor y del servicio en general. Para la facturación del mes de julio del 2017 al 52% de los usuarios se les aplicó un consumo medido, al 34% se les aplicó un consumo promedio y solamente al

14% se les aplicó un consumo ajustado. Esto se puede observar en el grafico 10 a continuación y en la tabla 12 en anexos.

Grafico 10 Tipo de consmo aplicado a los usuarios del distrito V de Managua en el mes de julio del año 2017



Fuente: Base de datos ENACAL facturación julio 2017

IX.3 Preparación de los datos

Antes de la creación de un modelo probabilístico, se debe realizar cierta preparación de los datos, como lo es, la estandarización de los datos, la eliminación de valores extremos, la eliminación de datos faltantes y revisión de correlación entre variables.

Las variables numéricas se estandarizaron de 0 a 100 utilizando la distribución uniforme (Figura 3). Este método de estandarización permite que las variables estandarizadas puedan interpretarse de forma más sencilla que utilizando la distribución normal estándar, esto es muy útil en Minería de Datos.

$$X_{Estandarizada} = \frac{X - MIN}{MAX - MIN}$$

Figura 5: Formula de estandarización de la Distribución

En toda base de datos encontraremos valores extremos, los cuales, se pueden tratar de varias maneras: por eliminación, sustituyéndolos por el promedio o sustituyéndolos por el máximo de los datos no extremos.

En una investigación de detección de inusualidades, los valores extremos representan parte de las inusualidades que se pueden presentar, debido a esta razón, no podemos realizar los tratamientos convencionales mencionados anteriormente, porque eliminaría la esencia de estos datos.

Una forma de tratar los valores extremos es realizando una estandarización combinada, que consiste, en una estandarización en dos etapas. La primera etapa de la estandarización combinada, es estandarizar todos los datos incluyendo los datos que consideramos extremos, la segunda etapa consiste en estandarizar los datos sin los datos extremos, luego estas dos etapas se combinan dando una única estandarización, la cual presenta un menor porcentaje de valores extremos. Este tratamiento de valores extremos fue diseñado por la metodología CRISP-DM e incluido en el SPSS-Modeler.

Otro problema que suele presentarse en las bases de datos, son los datos faltantes. La base de datos que analizamos, no presentó datos faltantes, esto permite que la base de datos sea de calidad.

En la figura siguiente (Figura 4), podemos observar que para todas las variables el 100% de los registros están completos, la base de datos posee 0 registros nulos, 0 registros vacíos y 0 espacios en blanco.

Campo	Medida	% Completo	Registros válidos	Valor nulo	Cadena vacía	Espacio en blanco	Valor vacío
ECONOMIA	Categórico	100	45930	0	0	0	0
TIPO_SERVICIO	Categórico	100	45930	0	0	0	0
TIPO_CONEX	Categórico	100	45930	0	0	0	0
ESTADO_CONEX	Categórico	100	45930	0	0	0	0
AÑOS_ANTIG	Continuo	100	45930	0	0	0	0
H_ABAST	Categórico	100	45930	0	0	0	0
MTO_FACTURA	Continuo	100	45930	0	0	0	0
ESTADO_SERVI	Categórico	100	45930	0	0	0	0
INCIDENCIA	Categórico	100	45930	0	0	0	0
NIVEL_EXONE	Categórico	100	45930	0	0	0	0
MEDICION	Categórico	100	45930	0	0	0	0
MTO_MORA	Continuo	100	45930	0	0	0	0
Q_MORA	Continuo	100	45930	0	0	0	0
TIPO_CONSUMO	Categórico	100	45930	0	0	0	0

Figura 6: Resultado del análisis con el nodo de auditoría de datos SPSS-Modeler

La última actividad indispensable que se debe hacer antes de crear un modelo probabilístico, es la revisión de las correlaciones entre las variables cuantitativas. En este análisis solo se incluyeron cuatro variables cuantitativas, de las cuales en la tabla 13 a continuación, se pueden observar que en la mayoría de las combinaciones de variables no existe alta correlación, excepto, la correlación entre el monto facturado y el monto mora, que a pesar de que es la correlación más alta entre las combinaciones de variables, esta no es lo suficientemente grande para que afecte a un modelo.

Tabla 13

Correlaciones de las variables cuantitativas		
Variables a contractar		Correlación
Años de antigüedad	Monto facturado	-0.003
Años de antigüedad	Monto facturado	0.006
Años de antigüedad	Monto de mora	0.138
Monto facturado	Monto de mora	0.243
Monto facturado	Cantidad de facturas en mora	-0.016
Monto de mora	Cantidad de facturas en mora	0.137

Fuente: Base de datos ENACAL facturación julio 2017

IX.4. Modelado

Las etapas de la metodología CRISP-DM que se desarrollaron anteriormente tienen como objetivo preparar todas las condiciones para diseñar un modelo probabilístico adecuado, que se ajuste a la realidad del fenómeno que estamos analizando y que cumpla con los parámetros de calidad.

El objetivo general de esta investigación, es detectar patrones de comportamiento de los usuarios del servicio de agua potable del distrito V de Managua, y los modelos de segmentación son herramientas muy útiles para lograrlo.

El SPSS Modeler nos facilita tres algoritmos para diseñar modelos de segmentación, el algoritmo de K-Medias, Kohonen y Bietápico; Estos algoritmos tienen características aplicables en tipos de variables específicos, por esta razón, los algoritmos que se ajustan a

nuestros datos son los algoritmos Bietápico y Kohonen, porque contamos con variables cualitativas y cuantitativas, pero por la facilidad de interpretación y de aplicación, utilizaremos el algoritmo Bietápico.

Este algoritmo posee distintas calibraciones, como lo es el número de clústeres, la medida de distancia y el criterio de agrupación de clústeres, por lo que, encontrar la calibración adecuada resulta tedioso; Pero gracias al SPSS Modeler, podemos crear modelos de forma simultánea, utilizando todas las calibraciones que se consideren conveniente y posteriormente, seleccionar el modelo que consideremos adecuado para estos datos. En esta investigación se ejecutaron un total de 72 modelos con distintas configuraciones, de los cuales, solo se seleccionaron 3:

1. El primer modelo es de 5 clústeres, con la medida de distancia Log-Verosimilitud y con un criterio bayesiano de Schwarz de agrupación de clústeres.
2. El segundo modelo es de 6 clústeres, con la medida de distancia Log-Verosimilitud y con un criterio bayesiano de Schwarz de agrupación de clústeres.
3. El tercer modelo es de 5 clústeres, con la medida de distancia Log-Verosimilitud y con un criterio bayesiano de Schwarz de agrupación de clústeres.

La razón por la que se eligieron tres modelos de segmentación, es porque, el promedio de facturación de los usuarios del servicio de ENACAL es distinto según el sector económico al que pertenece, por lo que, si realizamos un solo modelo, este se verá afectado por esta característica. En la tabla 14 podemos observar el comportamiento del promedio de facturación en los distintos sectores económicos, de lo cual, se puede observar que los usuarios en el sector del Gobierno Central y Descentralizado poseen los promedios de facturación más altos con respecto al resto, por esta razón, estos dos sectores conforman el primer grupo al que le corresponde el primer modelo de segmentación; también, se puede observar que en el Sector Público, Manufactura Particular, Industrial y Comercial, poseen promedios de facturación similares, por lo que, estos conforman el segundo grupo; el tercer grupo está conformado por los sectores económicos de Residenciales, Domicilios, Urbanizaciones Sociales y Asentamientos.

Cada uno de estos modelos actúan de forma independiente, pero, que dan resultados para un mismo objetivo, que es, la detección de patrones de comportamiento que nos ayudaron a detectar inusualidades que puedan ser potencialmente de riesgo.

Tabla 14

Promedio de facturación de los usuarios en los distintos sectores económicos del Distrito V de Managua en el mes de Julio 2017

Sector económico al que pertenece	Monto Promedio en la facturas	Segmento para modelo
GOBIERNO DESCENTRALIZADO	55609.9	1
GOBIERNO CENTRAL	29921.5	1
SECTOR PUBLICO	3347.7	2
MULTIFACTURA PARTICULAR	3106.1	2
INDUSTRIAL	1862.3	2
COMERCIAL	1646.1	2
RESIDENCIAL	544.3	3
DOMICILIAR	336.5	3
URBANIZACION SOCIAL	245.6	3
ASENTAMIENTO	183.5	3

Fuente: Base de datos ENACAL facturación julio 2017

Cada uno de estos modelos debe cumplir con ciertos parámetros, para poder definir si son adecuados; Los parámetros que se utilizaron para evaluar estos modelos son: Medida de Silueta, Importancia de los predictores y el Tamaño de los grupos.

Para el primer modelo se obtuvo una silueta de 0.557, lo que indica un ajuste en un rango bueno, de igual forma, en la figura 7 podemos observar que la variable "Monto de Mora" el modelo la evalúa como muy importante, seguido por la variable "¿Problemas en la lectura?"; las variables que el modelo las evalúa como menos importante son "Tipo de conexión" y "Monto Factura", el resto de las variables el modelo las evalúa como importancia regular.

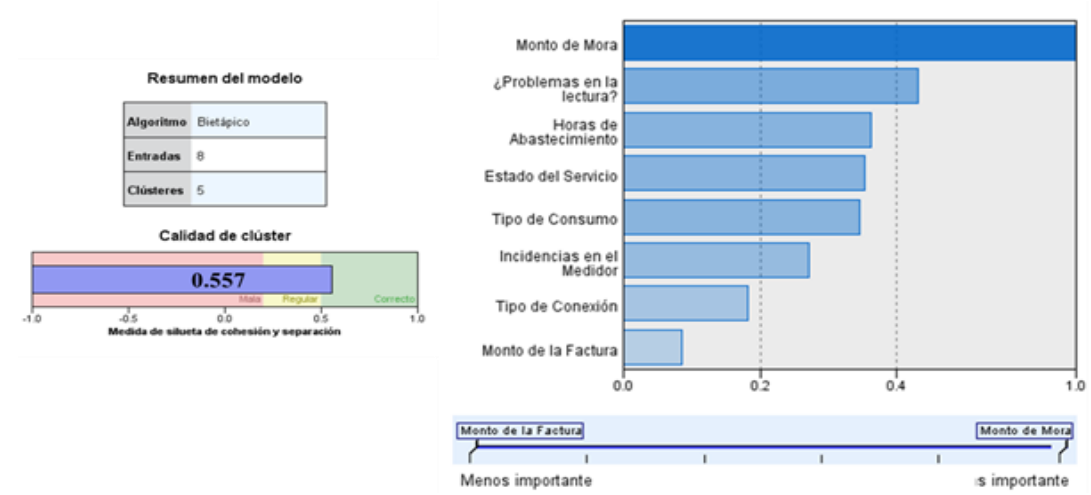


Figura 7: Visualización de parámetros de modelos de SPSS Modeler

En el modelo 2 se obtuvo una silueta de 0.504, lo que indica un ajuste en un rango bueno, de igual forma, en la figura 8 se puede observar que las variables "Incidencia en el Medidor", "Tipo de Consumo", "¿Problemas en la lectura?", "Estado del Servicio" y "Horas de Abastecimiento", el modelo las evalúa como muy importantes, seguido por las variables "Monto de Factura" y "Monto Mora".

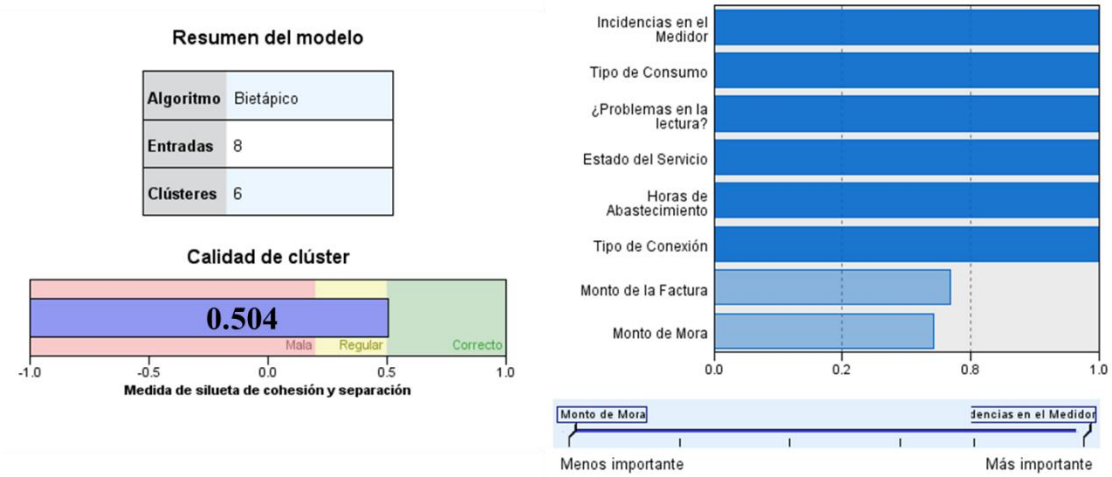


Figura 8: Visualización de parámetros de modelos de SPSS Modeler

Por último, en el tercer modelo se obtuvo una silueta de 0.561 lo que indica un ajuste en un rango bueno, de igual forma, en la figura 9 se puede observar que el modelo evalúa que todas las variables involucradas en el análisis son importantes.

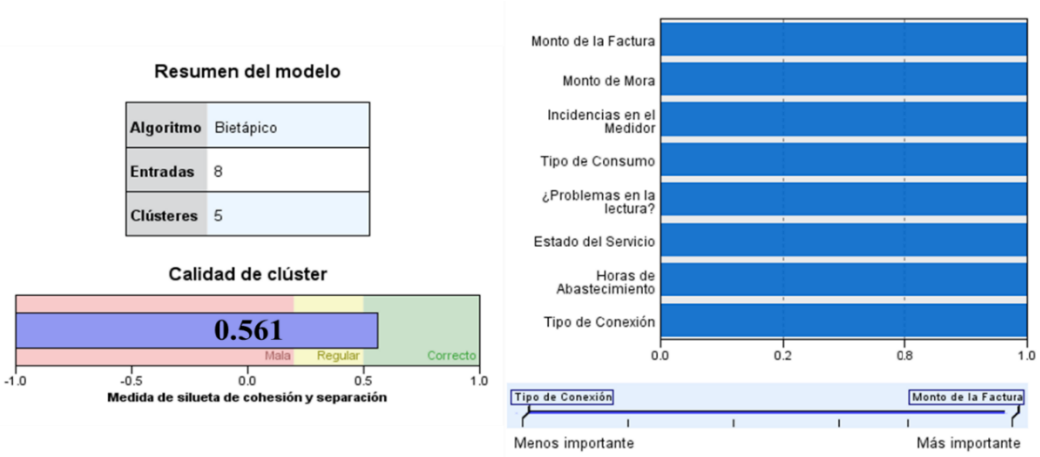
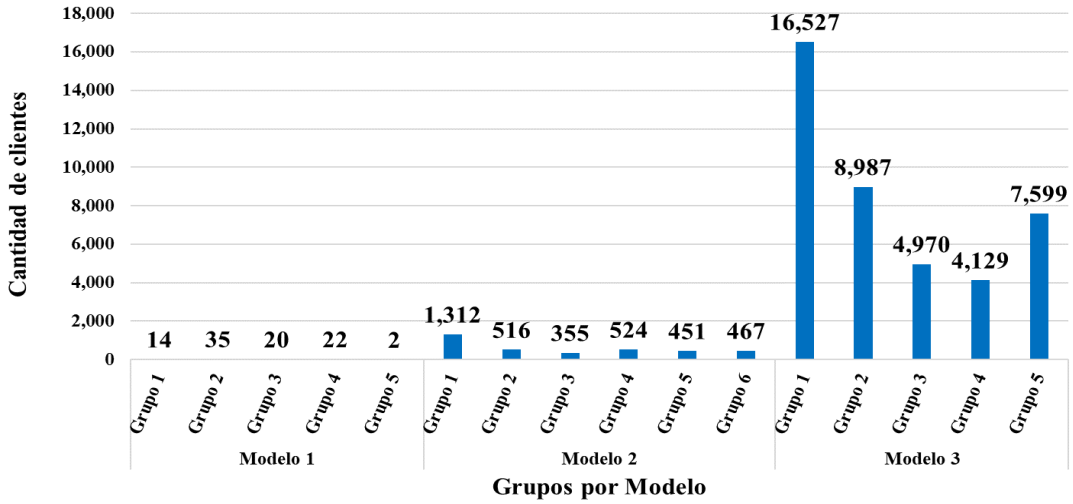


Figura 9: Visualización de parámetros de modelos de SPSS Modeler

Con respecto a los tamaños de los grupos en cada modelo, en la tabla 15 (Anexos) y en el grafico 11, podemos observar que la mayor parte de los usuarios se encuentran en el modelo 3, donde el grupo más pequeño en ese modelo es de 4,129 usuarios y el más grande es de 16,527 usuarios; Por otro lado, podemos observar que en el modelo 1 se encuentran la menor cantidad de usuarios, donde el grupo más pequeño es de 2 usuarios y el grupo más grande es de 14 usuarios; También, podemos observar que en el modelo 2 el grupo más pequeño es de 355 y el grupo más grande es de 1,312.

Grafico 11 Distribución de grupos en cada uno de los modelos de segmentacion de los clientes del Distrito V de Managua en el mes de Julio 2017



Fuente: Base de datos ENACAL facturación julio 2017

IX.5 Evaluación

En principio, se debe dejar claro que se trabajará con tres segmentos económicos, clasificados como: gobierno (central y descentralizado), empresas y personas naturales, este último comprende a la categoría de residencial, asentamiento y domiciliar.

Con respecto a las agrupaciones que realizó el modelo de segmentación, es importante mencionar que cada grupo tiene características que difieren del resto y a su vez reflejan el comportamiento de los usuarios con respecto al consumo de agua.

De lo anterior, se establece dos tópicos. Aquellos usuarios que presentan un comportamiento de riesgo o anómalo para la institución y aquellos que no, por lo cual, es imperativo dividir los resultados en dos fases, dentro de los que tienen un comportamiento inusual se destacan los usuarios morosos y aquellos que presentan alguna incidencia como: medidor aterrado, dañado, con Bypass o no registra consumo.

Es importante mencionar que aquellos usuarios que no tengan medidor no presenta ninguna incidencia, porque estos problemas tienen que ver con el medidor.

Primer segmento económico (Gobierno Central y Descentralizado):

EL grupo 5 se caracteriza principalmente por contener el mayor volumen de facturaciones y a su vez tiene la mayor cantidad de mora, problemas en la lectura y un horario de abastecimiento, de 3 a 5 horas diarias, el tipo de consumo es ajustado y no poseen exoneración. Sin embargo, no presenta ninguna incidencia en el medidor. Como se ve en la figura 10 en anexos

En cambio, según la figura 11 en anexos la mayoría de usuarios del grupo 1, tienen el medidor aterrado o dañado, por lo tanto, se les cobra un tipo de consumo promediado. Además, poseen un horario de abastecimiento bajo (de cero a dos horas diarias), por lo tanto, un mal servicio.

El resto de los grupos (2,3,4) se caracteriza por tener una facturación baja y no presentar mora o tener moras con montos bajos.

El tercer grupo (Figura 12) se caracteriza por que el 60 % de sus usuarios tiene un tipo de consumo promedio el restante se clasifica como ajustado. Mientras el cuarto presenta un 50% de consumo medido, el 27 % es ajustado y el 23 % es promediado. La diferencia

entre ambos radica en que el tercero no presenta problemas en la lectura y el cuarto sí. Esto se puede observar en la figura 13.

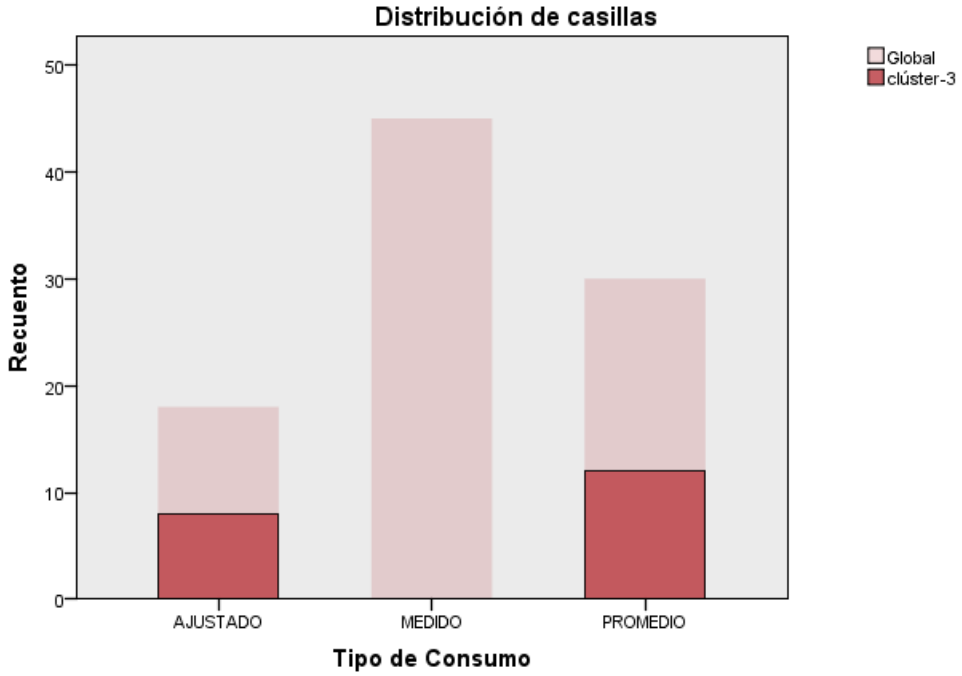


Figura 12: Tipo de consumo para el grupo 3

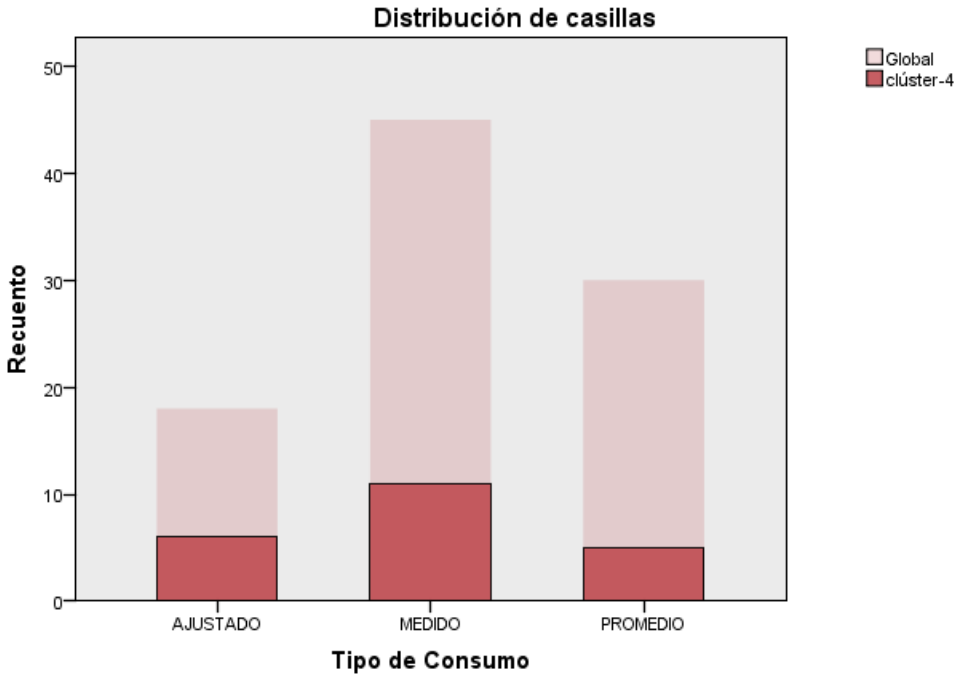


Figura 13: Tipo de consumo para el grupo 4

Por último, las características de la segunda agrupación son muy similares a las anteriores, añadiendo que el 100 % de los usuarios tiene un consumo medido.

Los grupos mencionados anteriormente, no presentan ninguna incidencia, exceptuando el cuarto que posee 10 % de usuarios con medidor dañado. Como se muestra en la figura 14.

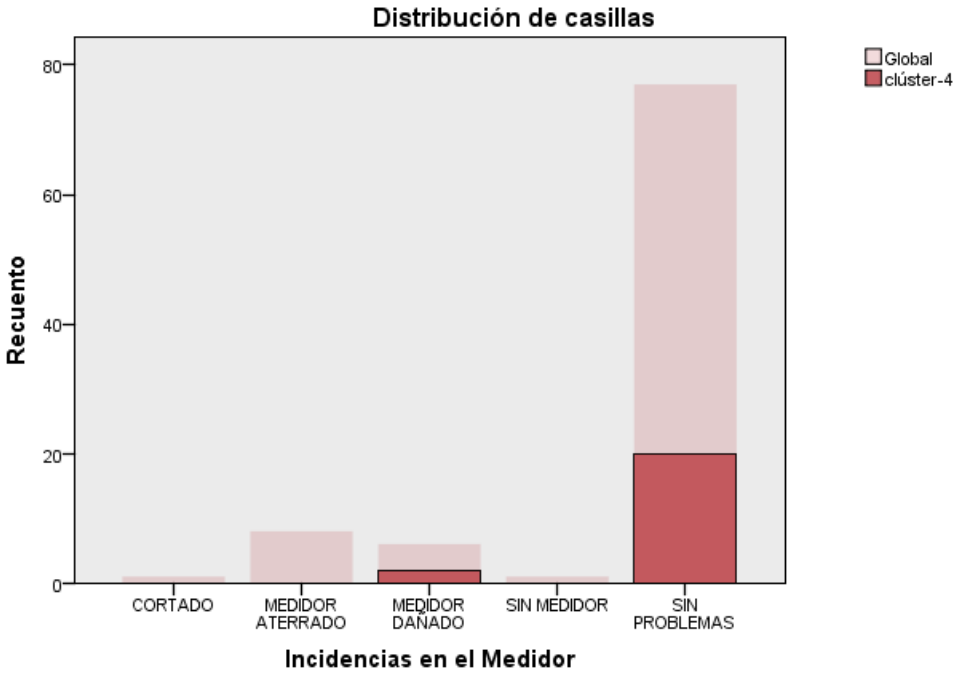


Figura 14: Incidencia en los medidores del grupo 4

Segundo segmento económico (Empresas):

El 86 % de los usuarios del grupo 3 no tiene problemas en la lectura. El 42 % tiene un excelente horario de abastecimiento, de 21 horas a mas por día, y un 37 % de cero a 2 horas, y en su mayoría, el (81 %), se le clasifica como consumo promediado. Por otro lado, es importante mencionar que el 100 % de los usuarios tiene alguna incidencia en el medidor, 55 % se distribuye en medidores aterrados, el restante tiene el medidor dañado o cortado, a como se refleja en la figura 15. Aunque este grupo presente una baja facturación, posee una gran cantidad de usuarios morosos, a como se puede ver en la figura 16 en anexos.

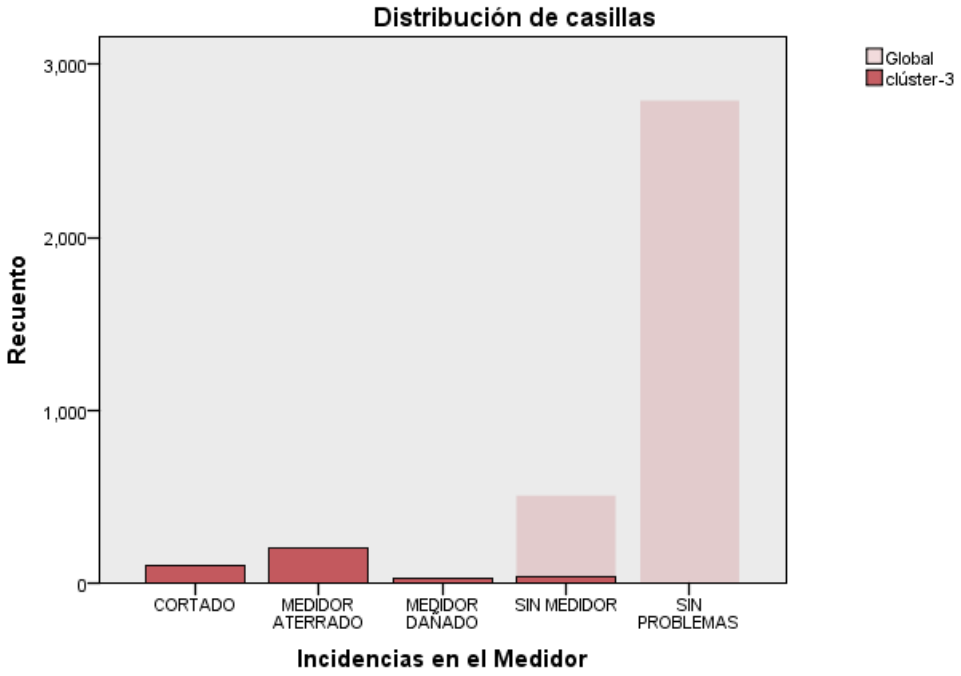


Figura 15: Incidencia en los medidores del grupo 3

En la figura 17 en anexos, se observa que el grupo 5 se caracteriza por que tiene un tipo de consumo ajustado, no presenta ninguna incidencia, ni problemas en la lectura, su horario de abastecimiento es de cero a 2 horas. El monto factura y moroso es sumamente bajo.

En vista de que los usuarios del grupo 6 no cuentan con un medidor, se les cobra una tarifa de cuota fija, en consecuencia, los montos facturados son relativamente bajos. Esto se logra apreciar en la figura 18 en anexos.

El grupo 2 posee un tipo de consumo medido, sin ningún problema al momento de realizar la lectura del medidor, sin embargo, un bajo horario de abastecimiento (cero a 2 horas). A pesar de que no presenta ninguna incidencia, es el que posee las mayores cantidades de facturación, así como de usuarios morosos. Esto se puede observar en la figura 19 en anexos.

Según la figura 20 y 21 en anexos, los últimos dos grupos el 4 y el 1, presentan aspectos similares, como, déficit en el horario de abastecimiento, de cero a 2 horas. Un tipo de consumo medido, sin problemas en el medidor, y facturación y mora sumamente baja. Sin embargo, el grupo cuatro presenta problemas en la lectura.

Tercer segmento económico (Personas Naturales):

En tercer lugar, se encuentran los usuarios clasificados como personas naturales, dentro de las cuales se encuentra la economía domiciliar, asentamiento y residencial.

Las características del grupo 4 se reflejan en la figura 22 en anexos, en la cual se puede observar que este grupo presenta un excelente horario de abastecimiento de 21 horas a mas, por lo tanto, un servicio bueno. El tipo de consumo es medido y no presenta problemas en el medidor, ni en la lectura, y los montos de moras son altos.

Los usuarios del grupo 3, poseen un horario de abastecimiento mayor 21 horas. Tienen los medidores aterrados, por lo tanto, el consumo es promediado. Y principalmente, son usuarios morosos. Esto se refleja en la figura 23 en anexos.

El grupo 5, se caracteriza porque sus usuarios no poseen medidor, a ellos se les clasifica dentro de la facturación de cuota fija, tienen un horario comprendido entre cero a 2 horas y por lo general tienen sus pagos atrasados, acumulando mucha mora; según la figura 24 en anexos.

El grupo 2, de acuerdo con la figura 25 en anexos, posee un bajo horario de abastecimiento, no presenta ninguna incidencia en el medidor, ni problemas en la lectura, tiene un tipo de consumo ajustado. Además, los montos facturados y morosos son bajos.

Por último, el grupo 1 al igual que el anterior, presenta un horario de abastecimiento de cero a dos horas por día, no tiene incidencias, ni problemas en la lectura, tiene facturaciones bajas. Sin embargo, el tipo de consumo es medido, a como se refleja en la figura 26 en anexos.

Niveles de Riesgo para los grupos de segmentación

La descripción de los grupos de segmentación, nos permiten determinar niveles de riesgo para cada uno de ellos. Para lo cual, se determina la importancia de las categorías de cada variable con respecto al riesgo, donde un conjunto de dichas categorías presentes en un mismo grupo de segmentación dará como resultado un nivel de riesgo alto. La importancia de las categorías se establece a nivel de criterio de experto, por ende, los niveles de riesgo se

definen de igual manera. Dicho de otra manera, los niveles de riesgo se definen por medio del análisis de los especialistas en el tema en estudio.

Para describir este nivel de riesgo con mayor eficiencia, se creó un árbol de decisión, el cual nos permite detallar las características principales de cada uno de los niveles de riesgos establecidos por los especialistas.

A continuación, se proporciona información sobre las especificaciones utilizadas al momento de crear el modelo, tales como: Profundidad del árbol, Importancia de los predictores, Matriz de confusión y el algoritmo utilizado.

El modelo se creó con un algoritmo C&R con una profundidad de 5. Con respecto a la importancia de los predictores, en la figura 27 se puede observar que el modelo define como más importante el tipo de consumo que se resume en medido, ajustado y consumo promedio, seguido por los montos morosos y los problemas que se presentan al momento de hacer la lectura del medidor.

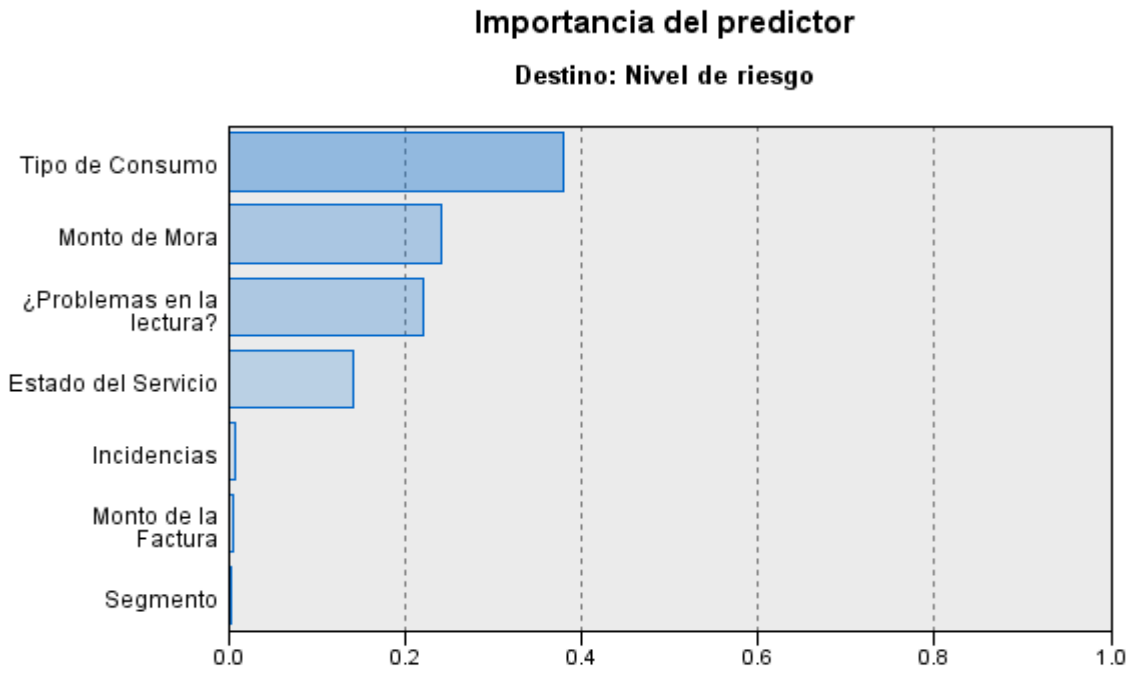


Figura 27: Importancia de los predictores del árbol de decisión.

De acuerdo con la matriz de confusión (Tabla 16 en Anexos), se puede observar la comparación entre los valores reales y los pronosticados por el modelo, lo cual nos permite tener la certeza de la calidad del modelo.

Para realizar este análisis de forma más específica, se calculan los índices de calidad, como la precisión global, precisiones específicas, el asertividad y los falsos positivos. Los cálculos de cada uno de estos indicadores se presentan a continuación:

Precisión Global

$$Precisión = \frac{1{,}770 + 3{,}305 + 3{,}546}{9{,}252} = \frac{8{,}736}{9{,}333} = 0.94$$

La precisión global del modelo fue de 94 %, lo que indica que la mayoría de las predicciones se realizaron correctamente.

Precisión Específica

$$Precisión_{Alto} = \frac{1{,}733}{2{,}046} = 0.85$$

Para los usuarios de alto riesgo 1,733, fueron pronosticados correctamente, esto representa un 85 %.

$$Precision_{Medio} = \frac{3{,}454}{3{,}610} = 0.96$$

Las predicciones del nivel medio fueron correctas en un 96 % de precisión.

$$Precision_{Bajo} = \frac{3{,}549}{3{,}677} = 0.97$$

Con respecto al nivel de riesgo bajo, el 97 % de las predicciones fueron correctas.

Falsos Positivos

Por otro lado, se aborda el problema tomando en cuenta los casos que fueron clasificados erróneamente. De esta manera se encuentra que el error global es de 6 %, es decir, las predicciones fueron incorrectas en ese porcentaje.

$$Falsos_{Alto} = \frac{313}{2{,}046} = 0.15$$

La cantidad de usuarios que fueron clasificados incorrectamente para el nivel de riesgo, alto, es de 313, el cual representa el 15 %.

$$Falsos_{Medio} = \frac{156}{3{,}610} = 0.04$$

El porcentaje de clasificación incorrecta para el nivel de riesgo, medio, es de 4 %.

$$Falsos_{Bajo} = \frac{128}{3{,}677} = 0.03$$

Para los usuarios con el perfil de riesgo bajo, el total de las predicciones incorrectas fue de 128, lo que representa 3 %.

Asertividad

$$Asertividad_{Alto} = \frac{1{,}773}{1{,}924} = 0.90$$

Las predicciones de alto riesgo, fueron acertadas para un 90 % de asertividad.

$$Asertividad_{Medio} = \frac{3{,}454}{3{,}640} = 0.95$$

El porcentaje de predicciones correctas para el nivel medio es de 95 %, equivalentes a 3,454.

$$Asertividad_{Bajo} = \frac{3{,}549}{3{,}769} = 0.94$$

De 3,769 predicciones que clasificaban al usuario en con un nivel de bajo riesgo, 3,549 fueron acertadas, representando un 94 %.

Conjunto de Reglas de decisiones.

El siguiente árbol de decisión (Figura 28), nos presenta distintas posibilidades para clasificar a un usuario según su nivel de riesgo, en alto, medio o bajo. El árbol toma como premisa las variables de tipo de consumo, resaltando los usuarios que poseen medidor de los que no, del primer caso, los divide si existe problema al momento de hacer la lectura o no, posteriormente están las variables de facturación y mora. Para el escenario donde no hay medidor los clasifica según el servicio que reciben, en directo, malo o cortado que es donde se presenta el riesgo, y en cuota fija, también, considera las variables de mora y segmento económico.

Es necesario mostrar la jerarquía que se presenta para cada una de las reglas. Es importante separarlas en tres grupos, según las categorías de la variable objetivo, y así establecer un perfil para cada una de ellas.

Primer escenario; **Nivel de riesgo alto**

Un usuario puede considerarse con un perfil de alto riesgo en las siguientes circunstancias:

Cuando el usuario posee medidor, además, no se presenta ninguna anomalía al momento de realizarse la lectura, y el monto de mora es mayor a C$ 903, entonces, la probabilidad de clasificación es de, 0.75.

Cuando el usuario no posee medidor, se le clasifica con un tipo de consumo ajustado o promedio, y dentro de ellos se presenta dos situaciones, primero cuando el servicio es directo, en mal estado o cortado, la probabilidad asciende a 0.99. Segundo, cuando el servicio es bueno, pero tienen moras superiores a los C$ 923, en este caso la probabilidad es de 0.83.

Ahora, es importante mencionar que un usuario con nivel de riesgo medio también puede resultar un problema para la empresa.

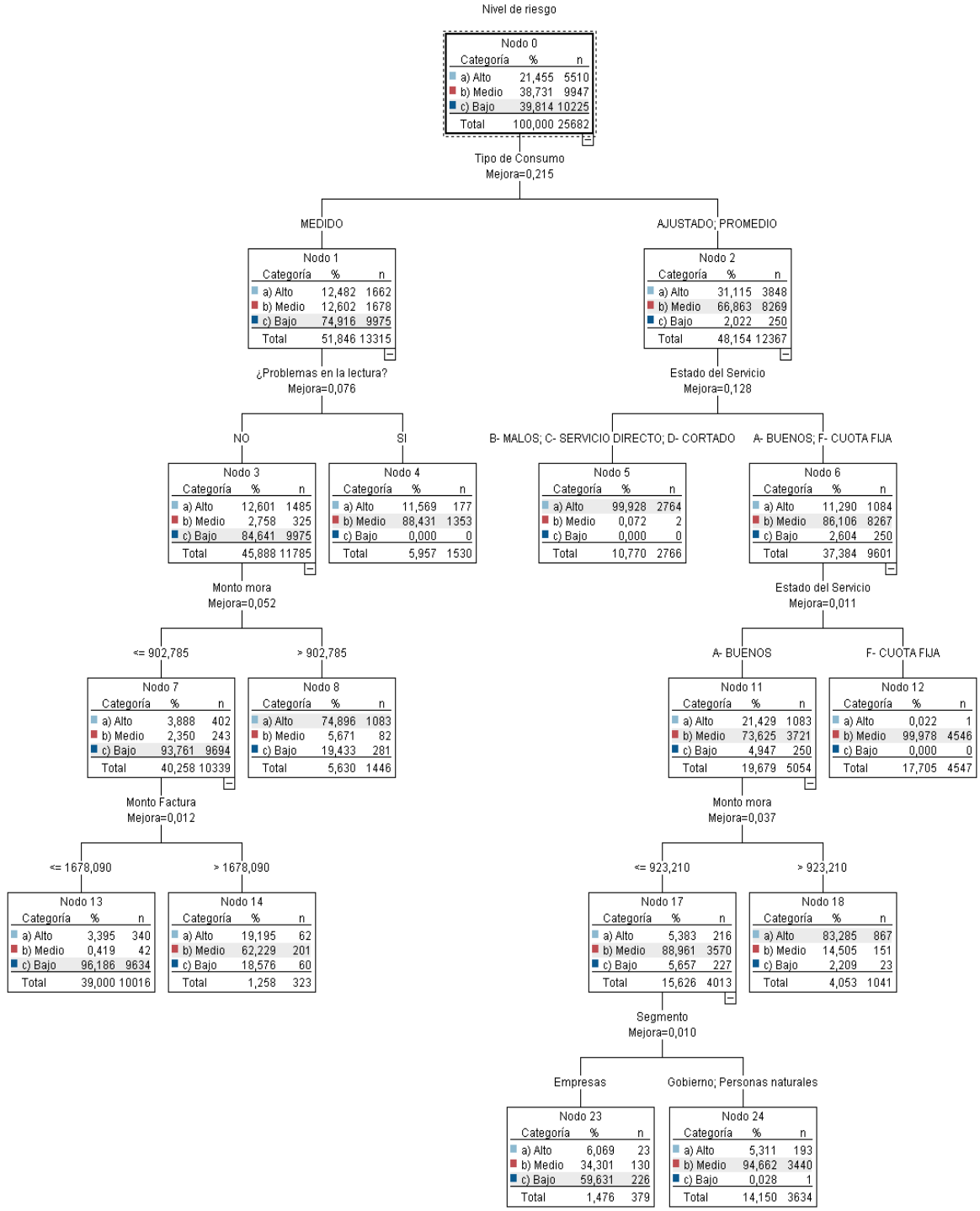


Figura 28: Dendograma (Conjunto de reglas de decisión)

Segundo escenario; **Nivel de riesgo medio**

El nivel medio, se describe con las siguientes características y sus probabilidades:

Cuando el usuario posee medidor, y presenta problemas en la lectura, con probabilidad de 0.88. Si no presenta problemas en la lectura y además poseen moras inferiores a C$ 903 y un monto facturado mayor a C$ 1,678, entonces, el usuario tiende a ser nivel de riesgo medio, con probabilidad de 0.62.

Si no posee medidor, y se clasifica con servicio de cuota fija, la probabilidad es de 0.99. Por otra parte, además de que no se le mida el consumo, y tiene un servicio excelente, es decir, el agua llega con regularidad y la red de tuberías se encuentra en perfecto estado, también poseen moras menores a C$ 923, y el segmento económico es gobierno y personas naturales, la probabilidad es de 0.95, si son empresas la probabilidad es, 0.34.

Y en tercer lugar se encuentran los usuarios con bajo nivel de riesgo.

Tercer escenario; **Nivel de riesgo bajo**

Si el tipo de consumo es medido, no hay problemas en la lectura y el monto de la mora y las facturaciones son pocas, inferiores a los C$ 903 y C$ 1,678, respectivamente.

De igual manera, cuando no posee medidor, el servicio es bueno, y los montos de moras son menor a C$ 923 y el segmento económico son empresas.

En la primera situación la probabilidad de que el usuario sea de bajo riesgo es de, 0.96, en la segunda regla es de 0.6.

X. Discusión de Resultados

Los modelos de segmentación son una herramienta muy útil al momento de identificar y caracterizar grupos, lo que las hace, una fuente de innovadora para la identificación de riesgos.

Cada uno de los modelos creados en esta investigación cumplen con los requisitos de calidad, lo que indica que los resultados son confiables. De estos resultados podemos mencionar que el índice de Silueta de cada modelo es superior a 0.5, dejándolos en un rango de calidad bueno con respecto a este indicador.

Una de las desventajas que poseen estos modelos de segmentación, es que están sujetos a modificaciones y reajustes en periodos de tiempo cortos, debido a que trabaja con fenómenos que pueden ser modificados o alterados por factores sociales no controlados.

Por otro lado, el modelo predictivo es una herramienta con mayor estabilidad con respecto a su ajuste de identificación de patrones de riesgos. Este tipo de modelos es efectivo en periodos más prolongados, debido a que detecta las probabilidades de que se materialice un perfil de riesgo ya identificado, lo que indica que esta herramienta es muy útil para mantener un monitoreo.

El modelo predictivo creado en esta investigación, cumple con los requisitos de calidad, como su precisión al predecir, su nivel de ajuste a los datos y la coherencia de sus resultados. De esta manera se logró caracterizar con precisión cada uno de los niveles de riesgos establecidos.

XI. Conclusión

Se logró determinar que existen usuarios con patrones de comportamientos que no están acorde a los comportamientos comunes en el servicio de agua que brinda la empresa ENACAL, lo que podría causarle a la empresa un impacto económico a corto o largo plazo.

Los niveles de riesgo de los usuarios establecidos en el servicio que brinda ENACAL y sus características, permite implementar un proceso de mitigación de riesgos potenciales para la empresa en periodos más cortos, con relación a los que actualmente se utilizan. Esto logra optimizar el proceso de monitoreo y mejorar los tiempos de respuesta.

Finalmente, los comportamientos de los usuarios con respecto al servicio que brinda ENACAL, nos permite evidenciar que el riesgo de fraudes y su impacto, están enfocados en la morosidad de los usuarios y en las incidencias que se presentan en los medidores. Esto indica, que, si la empresa invierte recursos para resolver estos dos factores, podría garantizar una mejoría en el servicio en general y ocasionaría un impacto social a gran escala.

XII. Recomendaciones

Debido a que estos modelos son válidos por cierto período, se recomienda ejecutarlos nuevamente con datos más actualizados y comparar los resultados obtenidos, para validar si es requerido modificarlos o reajustarlos.

De igual manera, se recomienda realizar una investigación de los usuarios que resultaron con nivel de riesgo alto, y realizar una depuración de falsos positivos, con el objetivo de identificar usuarios que ameriten acciones correctivas y preventivas para evitar futuros fraudes.

Con los resultados obtenidos después de la depuración, se recomienda integrarlos en un modelo predictivo para que se mantenga un monitoreo periódico para identificar nuevos casos de fraudes.

XIII. Bibliografía

Alto Nivel. (01 de Julio de 2013). *Árbol de decisión, una herramienta para decidir bien*. Recuperado el 25 de Septiembre de 2017, de Árbol de decisión, una herramienta para decidir bien: https://www.altonivel.com.mx/36690-arbol-de-decision-una-herramienta-para-decidir-correctamente/

Álvarez, C. A. (Octubre de 2012). *Aplicacion de Tecnicas de Mineria de Datos para Mejorar el Proceso de Control de Gestión en ENTEL.* Recuperado el 3 de 09 de 2017, de Aplicacion de Tecnicas de Mineria de Datos para Mejorar el Proceso de Control de Gestión en ENTEL: http://repositorio.uchile.cl/bitstream/handle/2250/112065/cf-martinez_ca.pdf?sequence=1

Arancibia, J. A. (s.f). *Metodología para el desarrollo de proyectos en Mineria de Datos CRISP-DM.* Obtenido de Metodología para el desarrollo de proyectos en Mineria de Datos CRISP-DM: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

Bernal, J. A. (2013). *Minería de datos.* Recuperado el 19 de 08 de 2017, de Minería de datos: https://ccc.inaoep.mx/~jagonzalez/AI/Sesion13_Data_Mining.pdf

Calderón, N. d. (4 de 2006). *Biblioteca USAC.* Recuperado el 19 de 08 de 2017, de Biblioteca USAC: http://biblioteca.usac.edu.gt/tesis/08/08_0307_CS.pdf

CONSULTEC, Empresa Digital. (s.f.). *Introducción al Big Data.* Recuperado el 20 de 08 de 2017, de Introducción al Big Data: http://www.spri.eus/euskadinnova/documentos/2343.aspx

EL BUHO ANALÍTICO. (2 de Febrero de 2012). *METODOLOGÍA SEMMA.* Recuperado el 1 de Octubre de 2017, de METODOLOGÍA SEMMA: http://elbuhoanaltico.blogspot.com/2012/02/metodologia-semma.html

ENACAL. (Diciembre de 2006). *ABC Sobre el recurso agua y su situación en Nicaragua.* Recuperado el 9 de Octubre de 2017, de ABC Sobre el recurso agua y su situación en Nicaragua: http://www.enacal.com.ni/media/imgs/informacion/ABCdelAgua1.pdf

ENACAL. (2017). *Antecedentes*. Recuperado el 15 de Octubre de 2017, de Antecedentes: http://www.enacal.com.ni/sobre-enacal/antecedentes.php.htm

ENACAL. (2017). *ENACAL.* Obtenido de ENACAL: http://www.enacal.com.ni/sobre-enacal/default.htm

Eumed.net. (2008). *Clientes Morosos.* Recuperado el 2017, de Clientes Morosos: www.eumed.net/libros-gratis/2008c/426/clientes%20morosos.htm

Félix, L. C. (2002). *Data mining: torturando a los datos hasta que confiesen[*]*. Recuperado el 11 de Septiembre de 2017, de Data mining: torturando a los datos hasta que confiesen[*]: http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

Gurrea, M. T. (S.f.). *ANÁLISIS DE COMPONENTES PRINCIPALES.* Recuperado el 13 de Septiembre de 2017, de ANÁLISIS DE COMPONENTES PRINCIPALES: https://www.uoc.edu/in3/emath/docs/Componentes_principales.pdf

Guzmán, E. L. (S.f.). *Minería de Datos.* Recuperado el 25 de Septiembre de 2017, de Minería de Datos: http://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf

IBM. (S.f.). *Análisis discriminante.* Recuperado el 20 de Septiembre de 2017, de Análisis discriminante: https://www.ibm.com/support/knowledgecenter/es/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/base/idh_disc.htm

IBM-SPSS Modeler. (s.f.). *Tipos de Modelos*. Recuperado el 12 de Septiembre de 2017, de Tipos de Modelos:

https://www.ibm.com/support/knowledgecenter/es/SS3RA7_16.0.0/com.ibm.spss.modeler.help/clementine/understanding_modeltypes.htm

Logicalis. (31 de Julio de 2014). *Velocidad, variedad y volumen, las 3 magnitudes clave de Big Data*. Recuperado el 10 de 09 de 2017, de Velocidad, variedad y volumen, las 3 magnitudes clave de Big Data: https://blog.es.logicalis.com/analytics/velocidad-variedad-y-volumen-las-3-magnitudes-clave-de-big-data

Los mapas auto-organizados de Kohonen (SOM). (S.f). Recuperado el 13 de Septiembre de 2017, de Los mapas auto-organizados de Kohonen (SOM): http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema5dm.pdf

Lucas, M. Á. (2015). *MiBloguel.* Recuperado el 20 de 08 de 2017, de MiBloguel: https://mibloguel.com/big-data-significado-y-su-utilidad-en-la-sociedad/

Manfut. (s.f.). *DISTRITO CINCO, MANAGUA, NICARAGUA* . Obtenido de DISTRITO CINCO, MANAGUA, NICARAGUA: http://manfut.org/managua/barrios/index.html

Martín, P. N. (s.f). *Aplicación de algoritmos de aprendizaje máquina para el análisis predictivo del mercado Forex.* Recuperado el 20 de Septiembre de 2017, de Aplicación de algoritmos de aprendizaje máquina para el análisis predictivo del mercado Forex: http://bibing.us.es/proyectos/abreproy/5730/fichero/Resumen+espa%C3%B1ol+PFC+Pablo+Nieto.pdf

Natividad, P. (2 de Agosto de 2016). *Conexión Esan*. Recuperado el 10 de 09 de 2017, de Conexión Esan: https://www.esan.edu.pe/conexion/actualidad/2016/08/02/la-importancia-del-big-data/

Pere Brachfield. (11 de Junio de 2012). *¿Qué es un moroso?* Recuperado el 2 de Octubre de 2017, de ¿Qué es un moroso?: http://perebrachfield.com/blog/morosos-y-pufistas/que-es-un-moroso/

Power Data. (s.f.). *¿Qué es Big Data?* Recuperado el 20 de 08 de 2017, de ¿Qué es Big Data?: http://cdn2.hubspot.net/hub/239039/file-359994269-pdf/docs/PowerData_-_Del_bit%E2%80%A6_Al_Big_Data.pdf

Quintanales, L. M., Moreno García, M., & García Peñalvo , F. (Enero de 2001). *Aplicación de Técnicas de Minería de Datos en la Construcción y Validación de Modelos Predictivos y Asociativos A Partir de Especificaciones de Requisitos de Software.* Recuperado el 21 de 08 de 2017, de Aplicación de Técnicas de Minería de Datos en la Construcción y Validación de Modelos Predictivos y Asociativos A Partir de Especificaciones de Requisitos de Software: http://ceur-ws.org/Vol-84/paper4.pdf

Rubio Hurtado, M. J., & Vila-Baños, R. (20 de Octubre de 2016). *El análisis de conglomerados bietápico o en dos fases*. Recuperado el 13 de Septiembre de 2017, de El análisis de conglomerados bietápico o en dos fases: http://revistes.ub.edu/index.php/REIRE/article/viewFile/reire2017.10.11017/20151

SAEM Thales-CICA. (s.f.). *El modelo de Kohonen*. Recuperado el 20 de Septiembre de 2017, de El modelo de Kohonen: http://thales.cica.es/rd/Recursos/rd98/TecInfo/07/capitulo6.html

Sinnexus. (2007). *Data Warehouse*. Recuperado el 10 de Septiembre de 2017, de Data Warehouse: www.sinnexus.com/business_intelligence/datawarehouse.aspx

Sinnexus. (2007). *Datos, información, conocimiento*. Recuperado el 11 de Septiembre de 2017, de Datos, información, conocimiento: http://www.sinnexus.com/business_intelligence/piramide_negocio.aspx

Tecnologías de Información. (2016). *Tecnologías de Información.* Obtenido de Tecnologías de Información: http://www.tecnologias-informacion.com/mineria-de-datos.html

UIAF. (2014). *Técnicas de minería de datos para la detección y prevención del lavado de activos y la financiación del terrorismo (LA/FT).* Recuperado el 25 de Septiembre de 2017, de Técnicas de minería de datos para la detección y prevención del lavado de activos y la financiación del terrorismo (LA/FT):

http://www.urosario.edu.co/observatorio-de-lavado-de-activos/Archivos_Lavados/Tecnicas-de-mineria-de-datos-para-la-prevencion-de.pdf

Uniovideo. (S.f). *El algoritmo k-means aplicado a clasificación y procesamiento de imágenes*. Recuperado el 13 de Septiembre de 2017, de El algoritmo k-means aplicado a clasificación y procesamiento de imágenes: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

WebMining Concultores. (Enero de 2011). *KDD: Proceso de Extracción de conocimiento.* Recuperado el 29 de Septiembre de 2017, de KDD: Proceso de Extracción de conocimiento: http://www.webmining.cl/2011/01/proceso-de-extraccion-deconocimiento/

Paredes Castillo, N. J. (2003). *Enfoque de planeamiento estratégico para la empresa gráfica de la UNMSM*. Universidad Nacional Mayor de San Marcos.)

Vigo Chacón, G. J. (2010). Método de clasificación para evaluar el riesgo crediticio: Una comparación. Universidad Nacional Mayor de San Marcos.

Cruz Quispe, L. M., & Rantes García, M. T. (2010). *Detección de fraudes usando técnicas de clustering*. Universidad Nacional Mayor de San Marcos.

Ñaupas Caraza, C. M. (2016). Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias. Universidad Nacional Mayor de San Marcos.

Flores Coaguila, J. D. (2014). *Propuesta de modelo de detección de fraudes de energía eléctrica en clientes residenciales de Lima Metropolitana aplicando minería de datos*. Universidad de San Martín de Porres, Lima, Perú.

Contreras Chinchilla, L., & Rosales Ferreira, K. (2016). *Análisis del comportamiento de los clientes en las redes sociales mediante técnicas de minería de datos*. Universidad Inca Garcilaso de la Vega.

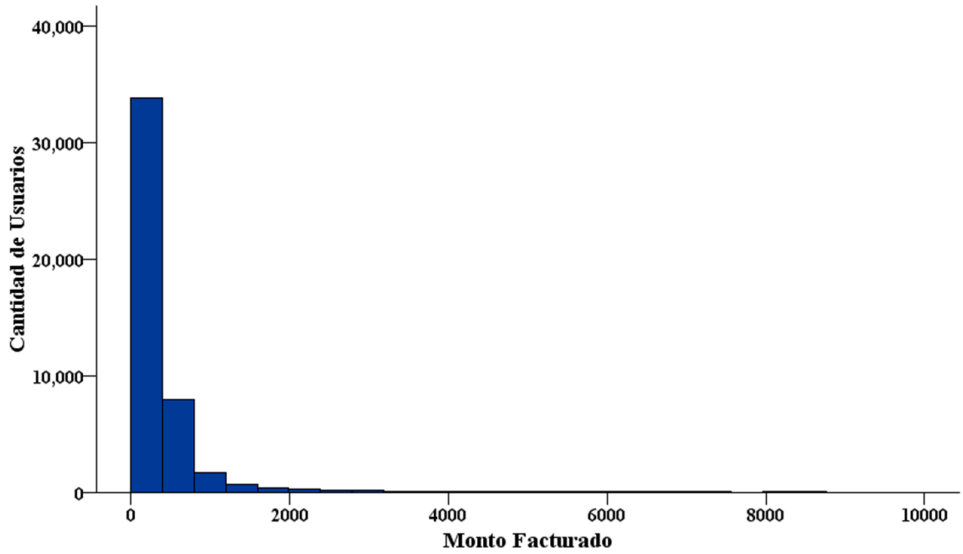
Maldonado Cadenillas, R. R. (2016). Proceso de extracción de patrones secuenciales para la caracterización de fenómenos espacio-temporales. Pontificia Universidad Católica del Perú.

Giraldo Mejía, J. C., & Vargas Agudelo, F. A. (2011). Aplicación de la Técnica Regresión Logística de la minería de datos en el proceso de Descubrimiento de Conocimiento (KDD) en bases de datos operativas o transaccionales. Universidad Inca Garcilaso de la Vega.

Pizarro Solís, P. A., & Acosta de la Cruz, P. R. (2011). *Predicción del rendimiento académico en la Educación Superior usando minería de datos y su comparación con técnicas estadísticas*. Universidad Nacional de Ingeniería.

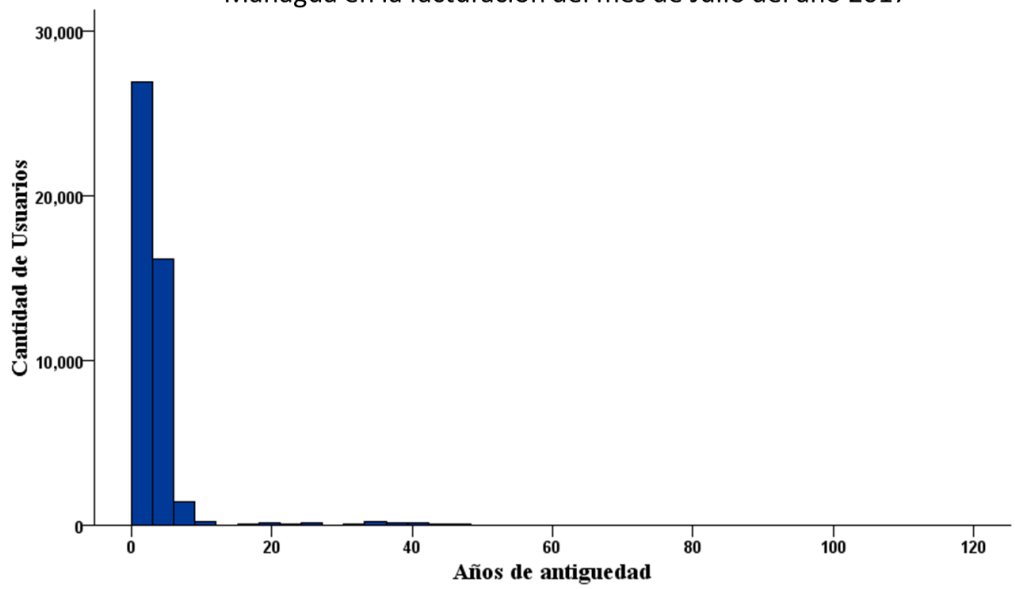
XIV. Anexo

Grafico 1 Histograma Monto facturado por los usuarios del distrito IV de Managua en el mes de Julio del año 2017



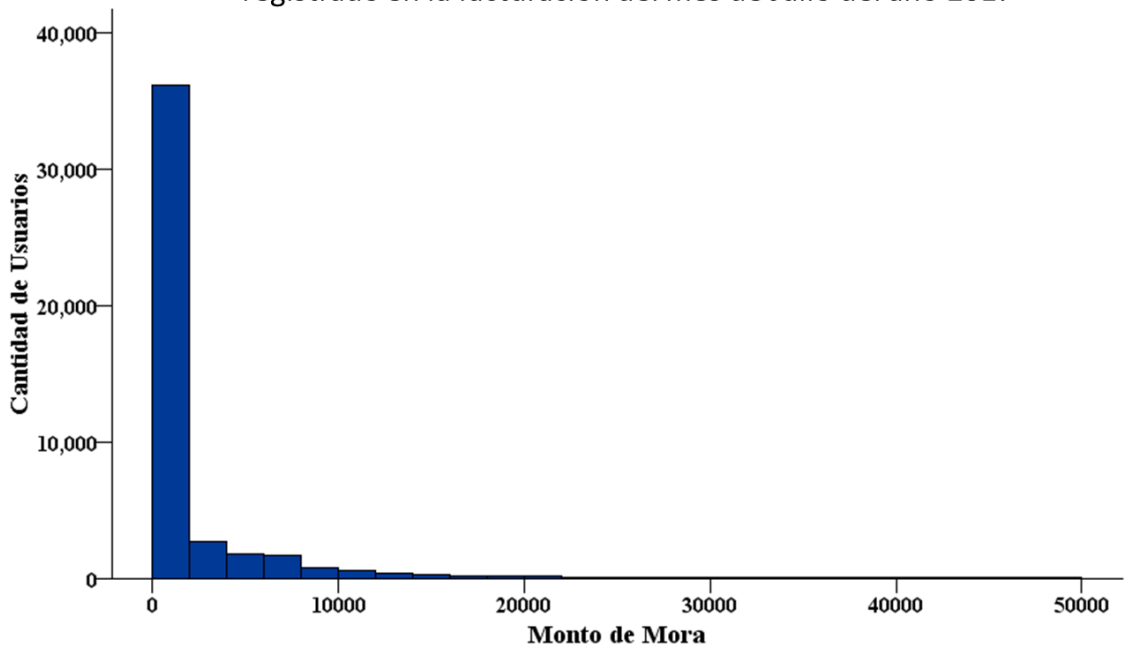
Fuente: Base de datos ENACAL facturación julio 2017

Grafico 2 Histograma de los años de antigüedad de los usuarios del distrito IV de Managua en la facturación del mes de Julio del año 2017



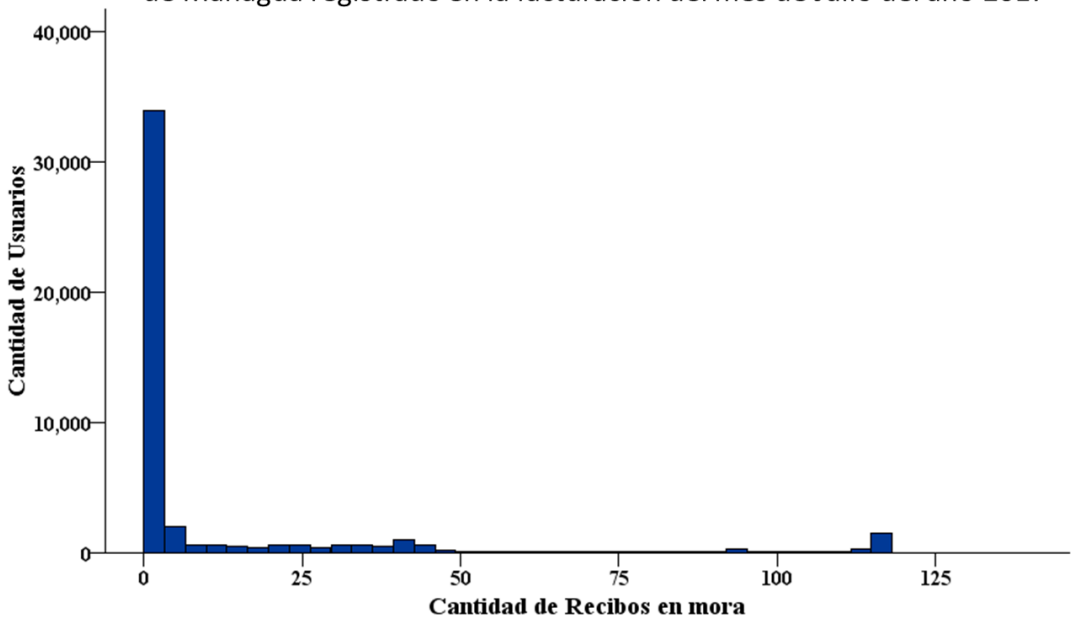
Fuente: Base de datos ENACAL facturación julio 2017

Grafico 3 Histograma monto de mora de los usuarios del distrito IV de Managua registrado en la facturación del mes de Julio del año 2017



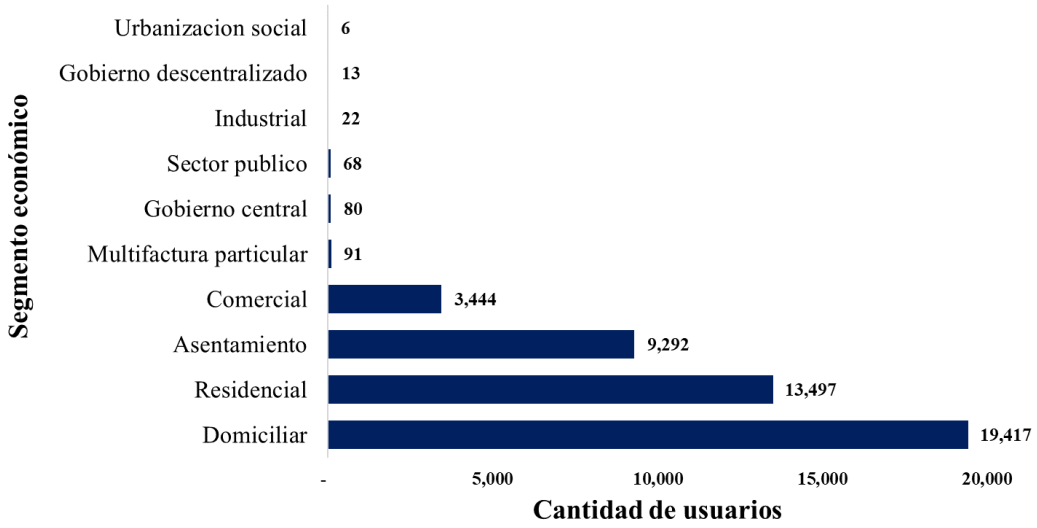
Fuente: Base de datos ENACAL facturación julio 2017

Grafico 4 Histograma de la cantidad de recibos en mora de los usuarios del distrito IV de Managua registrado en la facturación del mes de Julio del año 2017



Fuente: Base de datos ENACAL facturación julio 2017

Grafico 5 **Segmento económico de los usuarios del distrito V de Managua registrado en el mes de julio del año 2017**



Fuente: Base de datos ENACAL facturación julio 2017

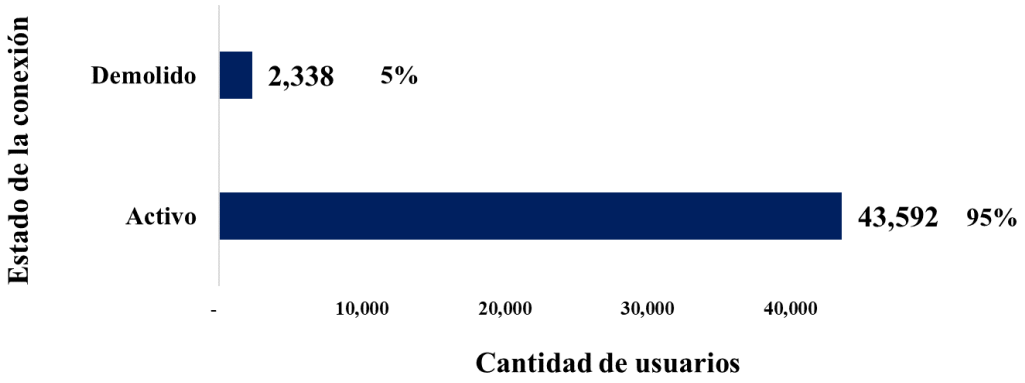
Tabla 8

Tipo de Conexión del Servicio de los Usuarios del Distrito V de Managua en el mes de Julio 2017

Tipo de Conexión	Cantidad de Clientes	%
Agua 1 medidor	37,669	82.01%
Agua sin medidor	8,068	17.57%
Extr.agua pozos medido fact	118	0.26%
Madre 1 medidor	48	0.10%
Macros combinados	27	0.06%
Total	45,930	100%

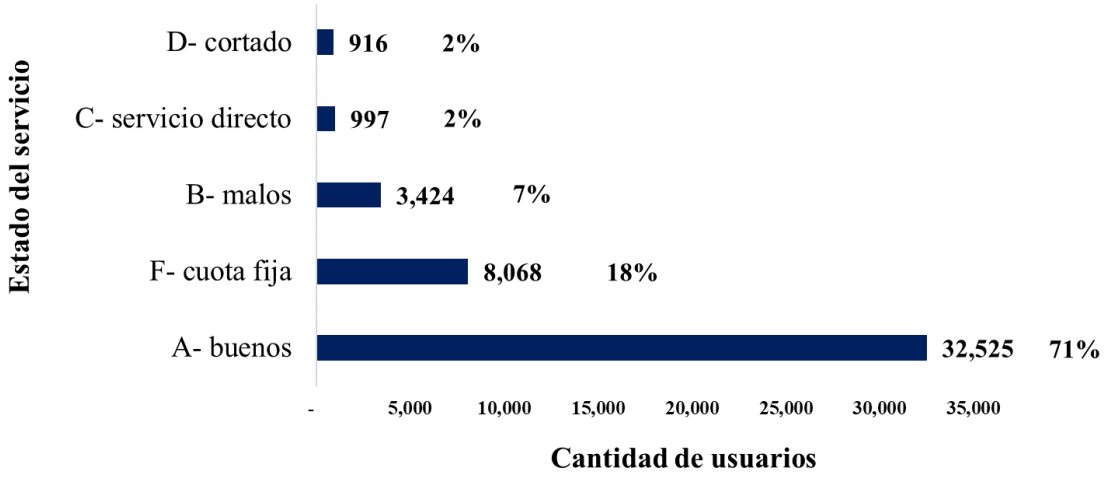
Fuente: Base de datos ENACAL facturación julio 2017

Grafico 7 **Estado de la conexión que poseen los usuarios del distrito V de Managua en el mes de julio del año 2017**



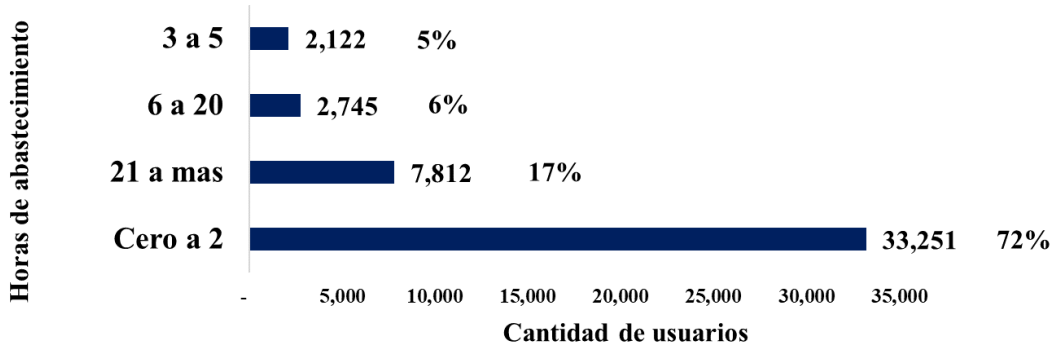
Fuente: Base de datos ENACAL facturación julio 2017

Grafico 8 **Estado del servicio de agua potable para los usuarios del distrito V de Managua en el mes de julio del año 2017**



Fuente: Base de datos ENACAL facturación julio 2017

Grafico 9 **Horario de abastecimiento que reciben los usuarios del distrito V de Managua en el mes de julio del año 2017**



Fuente: Base de datos ENACAL facturación julio 2017

Tabla 12

Tipo de consumo aplicado a los Usuarios del Distrito V de Managua en el mes de Julio 2017

Tipo de consumo	Cantidad de Clientes	%
Medido	23,875	52%
Promedio	15,687	34%
Ajustado	6,368	14%
Total	45,930	100%

Fuente: Base de datos ENACAL facturación julio 2017

Tabla 15

Distribución de grupos en cada uno de los modelos de segmentación de los clientes del Distrito V de Managua en el mes de Julio 2017

	Grupo	Cantidad de clientes	%
Modelo 1	Grupo 1	14	15.05%
	Grupo 2	35	37.63%
	Grupo 3	20	21.51%
	Grupo 4	22	23.66%
	Grupo 5	2	2.15%
	Total	93	100.00%
Modelo 2	Grupo 1	1,312	36.19%
	Grupo 2	516	14.23%
	Grupo 3	355	9.79%
	Grupo 4	524	14.46%
	Grupo 5	451	12.44%
	Grupo 6	467	12.88%
	Total	3,625	100.00%
Modelo 3	Grupo 1	16,527	39.15%
	Grupo 2	8,987	21.29%
	Grupo 3	4,970	11.77%
	Grupo 4	4,129	9.78%
	Grupo 5	7,599	18.00%
	Total	42,212	100.00%

Fuente: Base de datos ENACAL facturación julio 2017

Tabla 16

Matriz de confusión

		Valores pronosticados			
	Nivel de riesgo	a) Alto	b) Medio	c) Bajo	Total
Valor real	a) Alto	1,733	163	150	2,046
	b) Medio	86	3,454	70	3,610
	c) Bajo	105	23	3,549	3,677
	Total	1,924	3,640	3,769	9,333

Fuente: Base de datos ENACAL facturación julio 2017

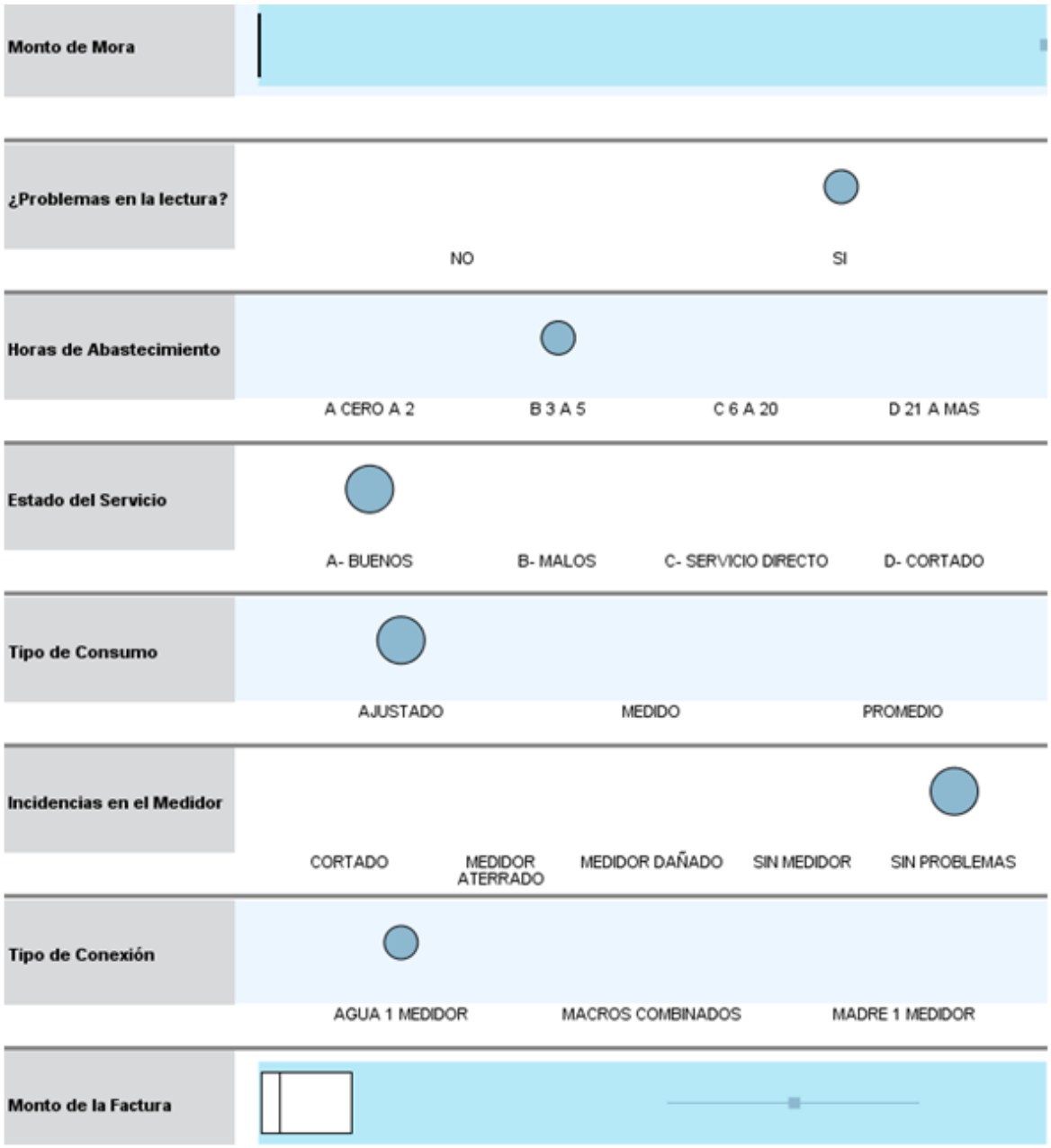


Figura 10. Características del grupo 5

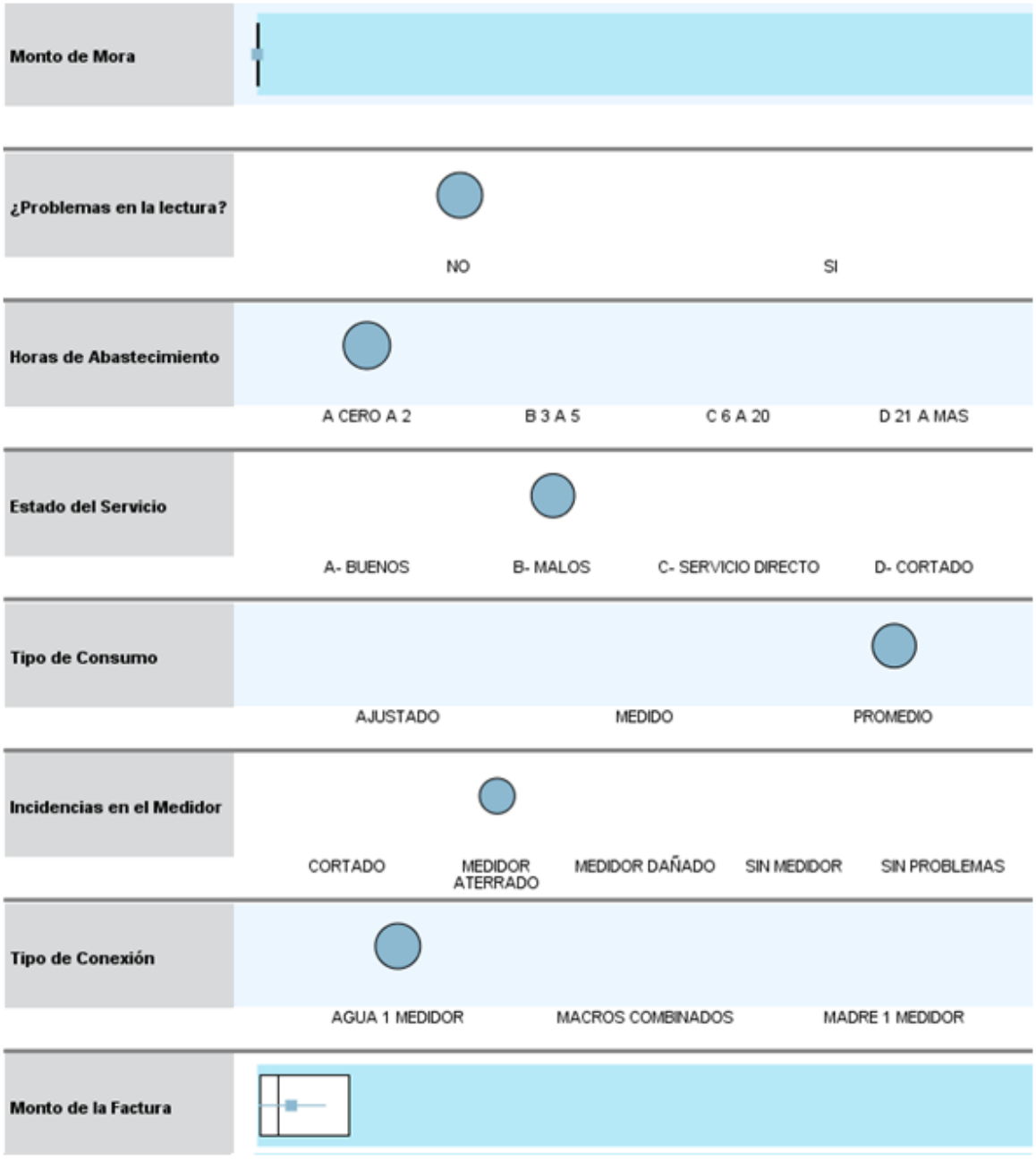


Figura 11. Características del grupo 1

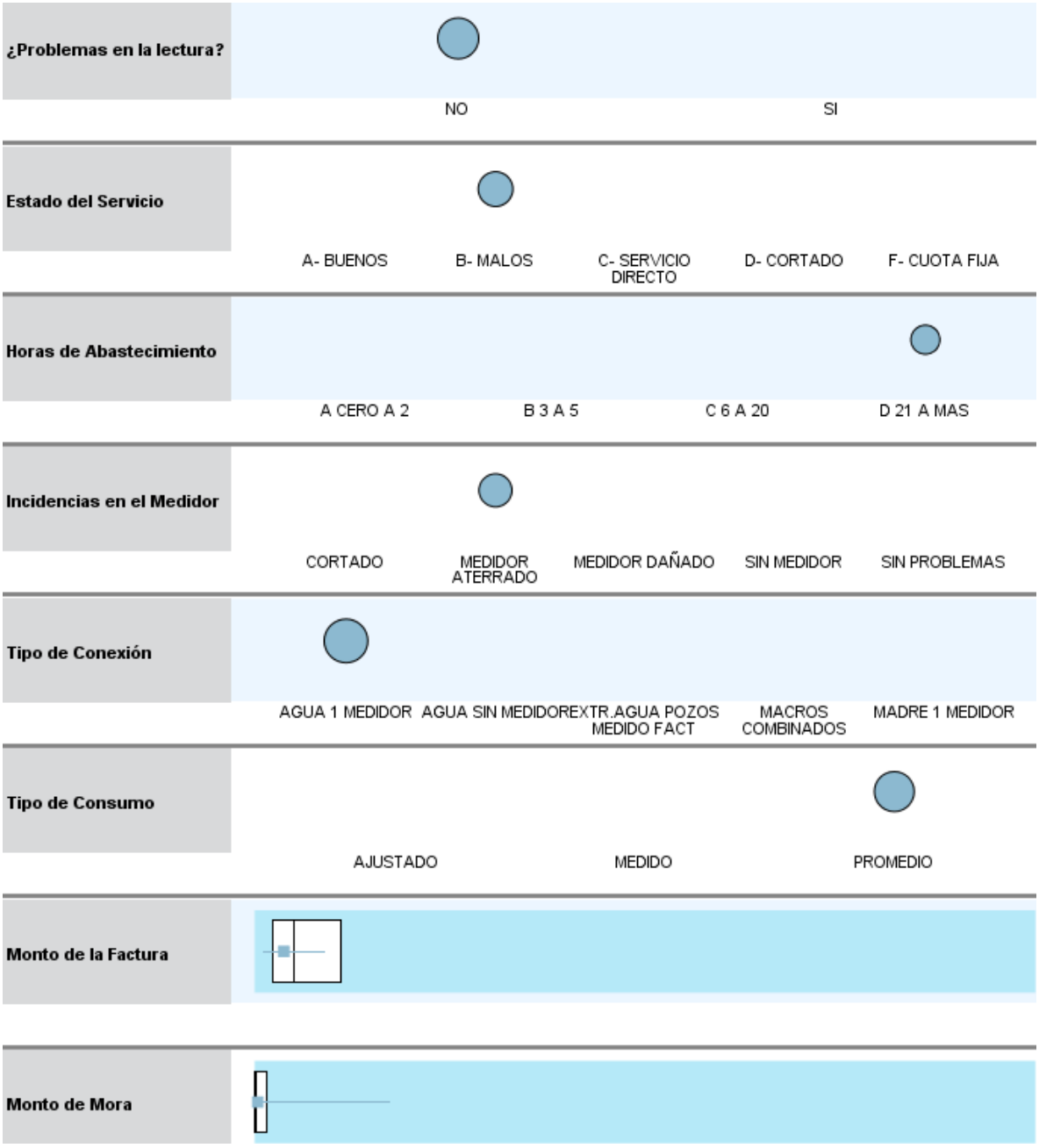


Figura 16: Características Grupo 3

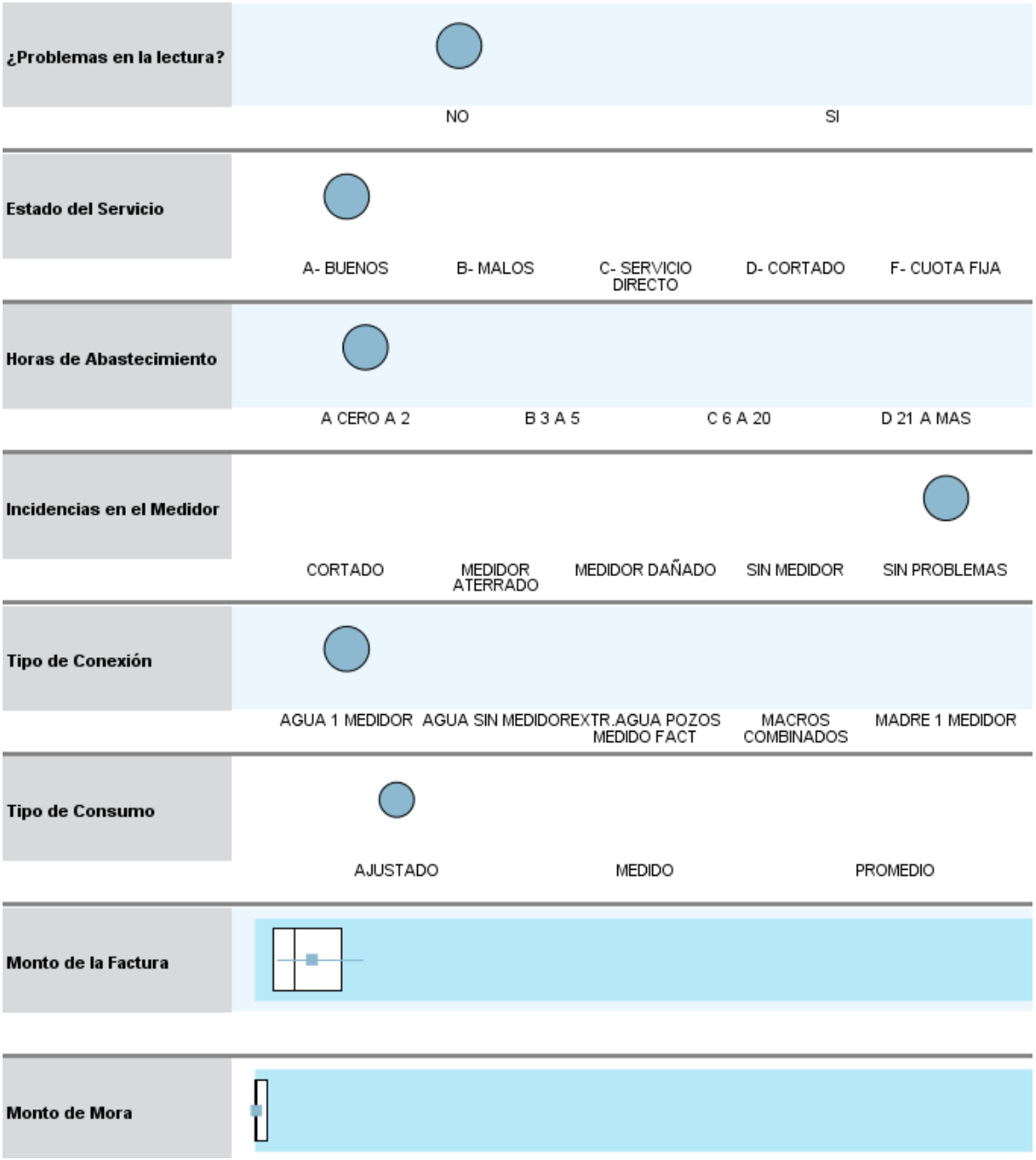


Figura 17. Características del grupo 5

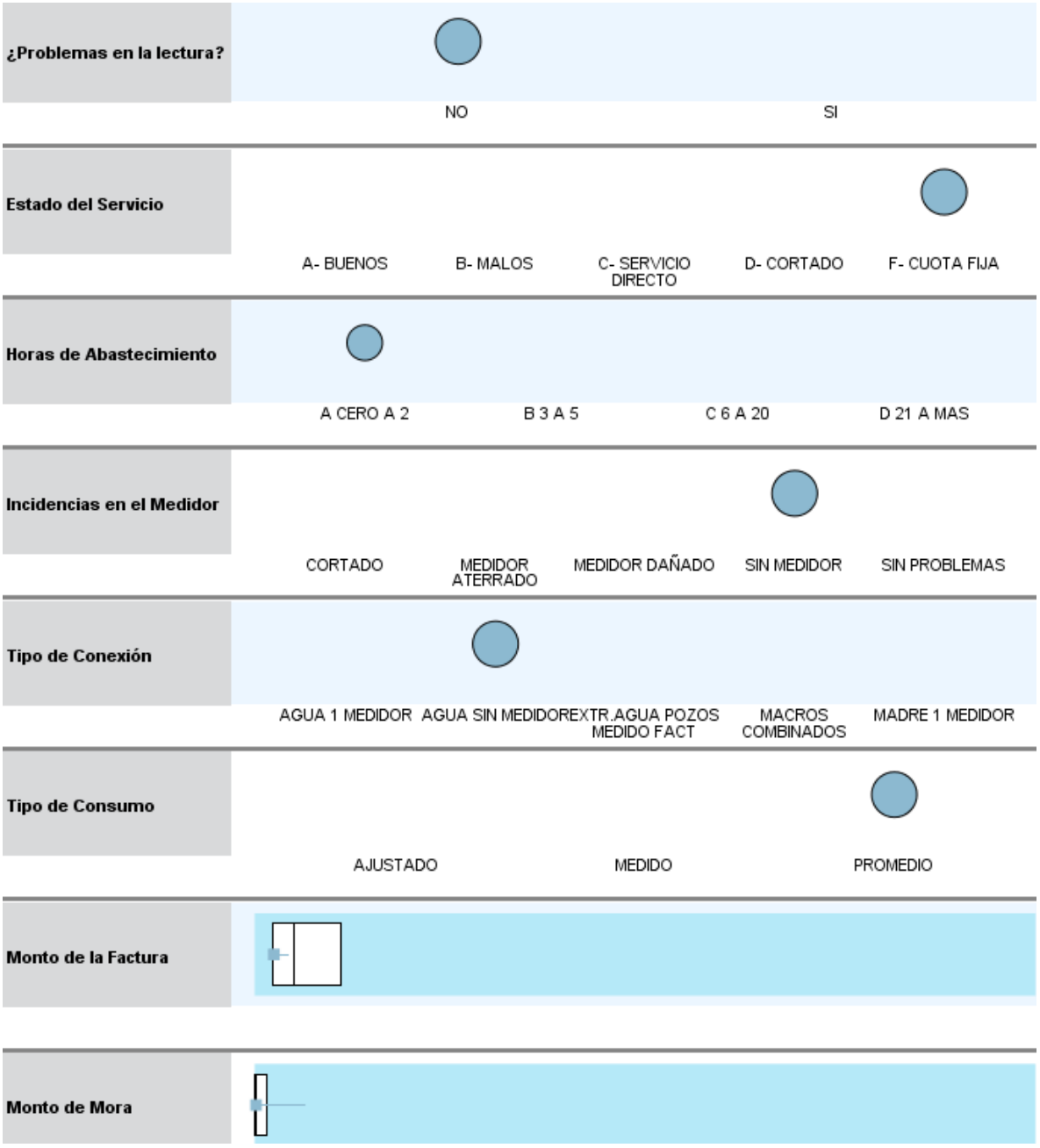


Figura 18. Características del grupo 6

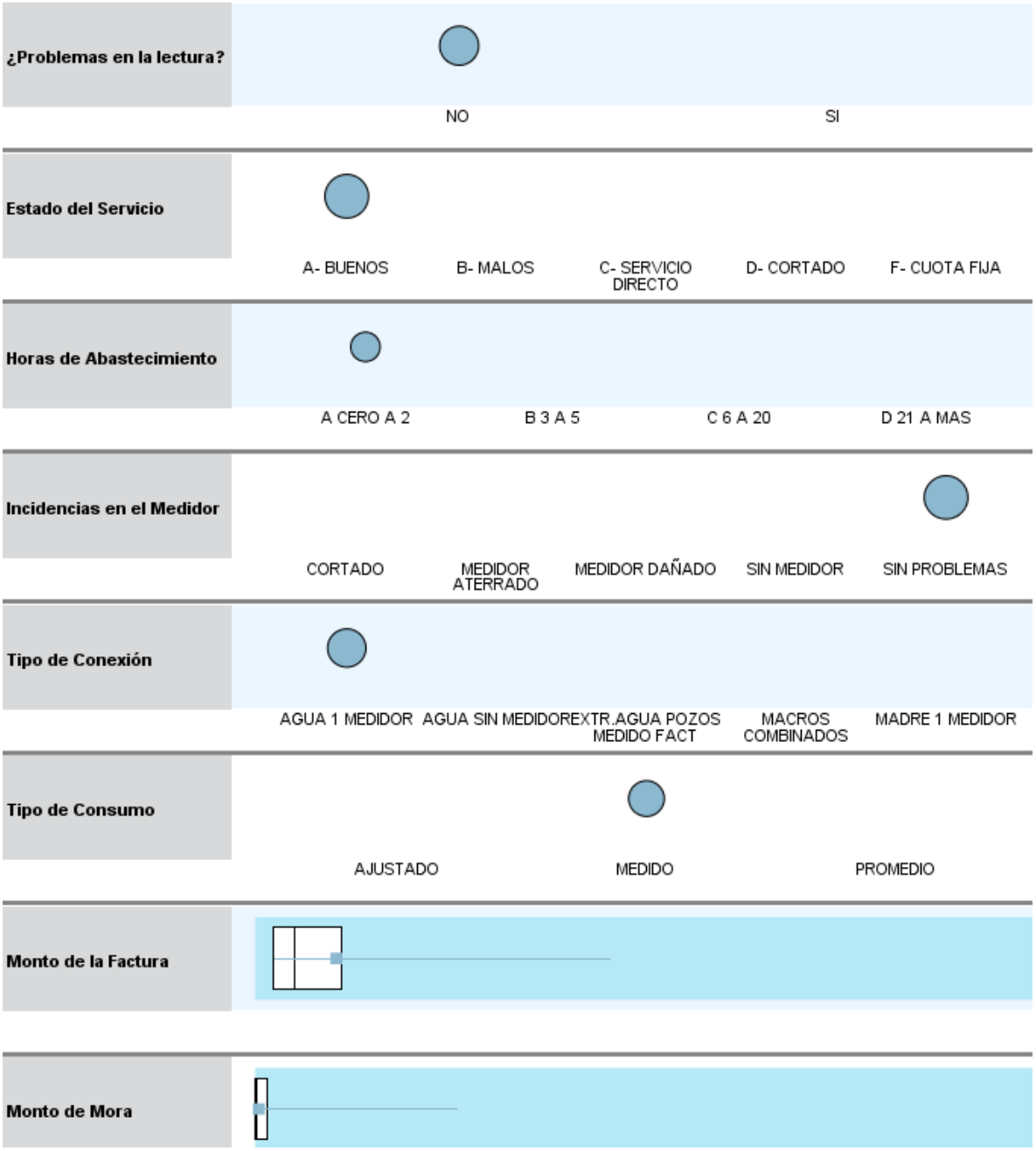


Figura 19. Características del grupo 2

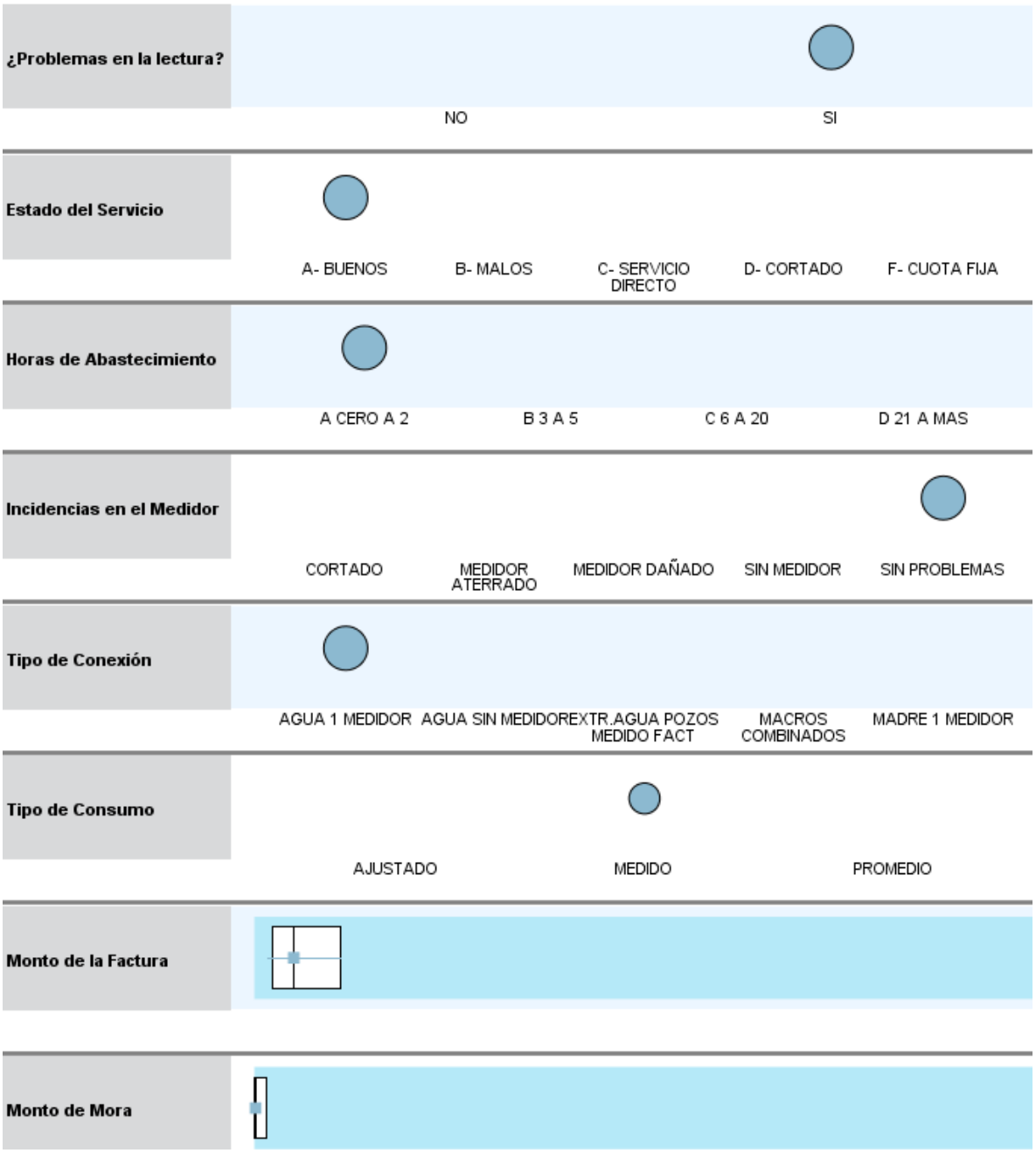


Figura 20. Características del grupo 4

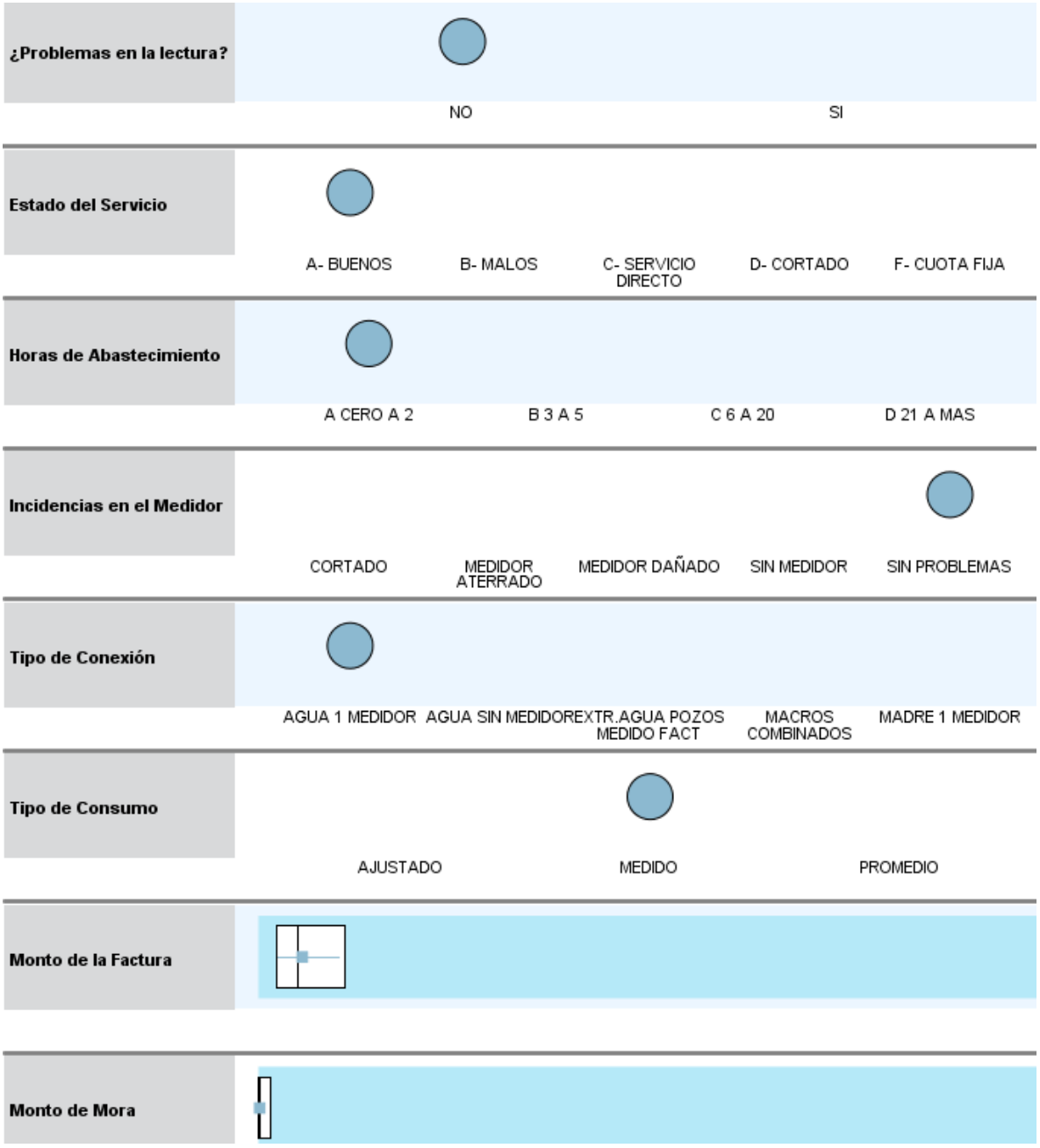


Figura 21. Características del grupo 1

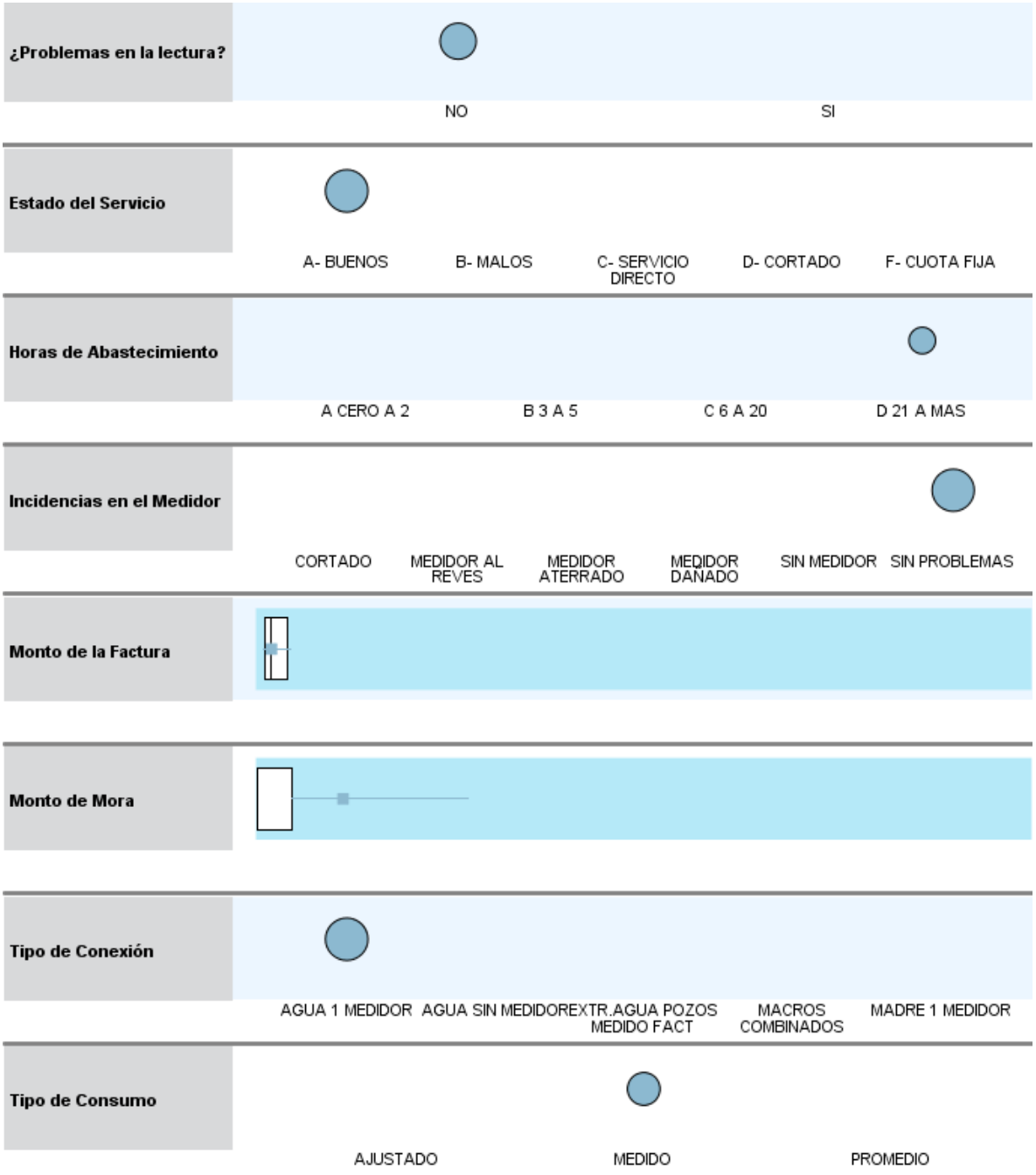


Figura 22. Características del grupo 4

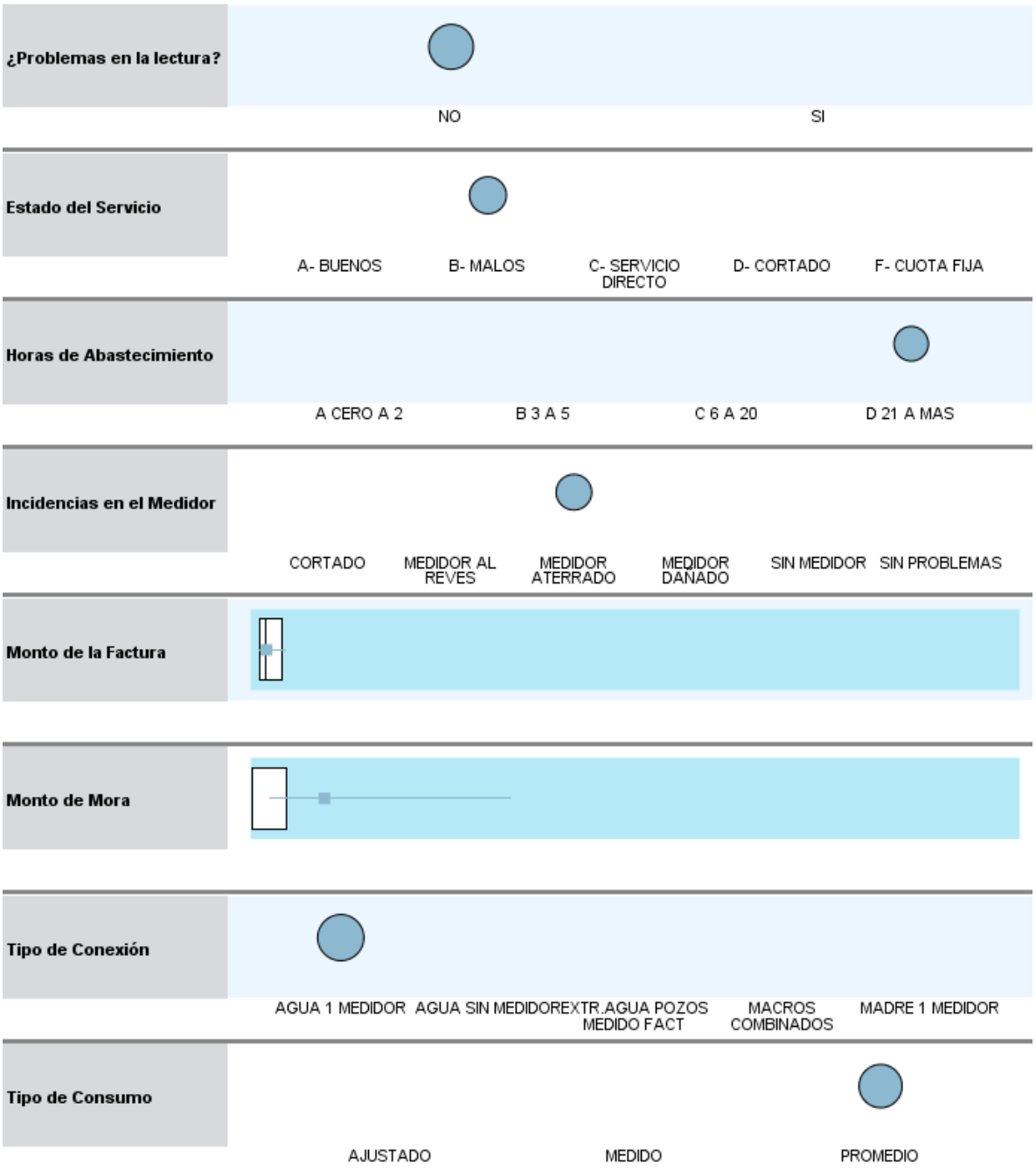


Figura 23. Características del grupo 3

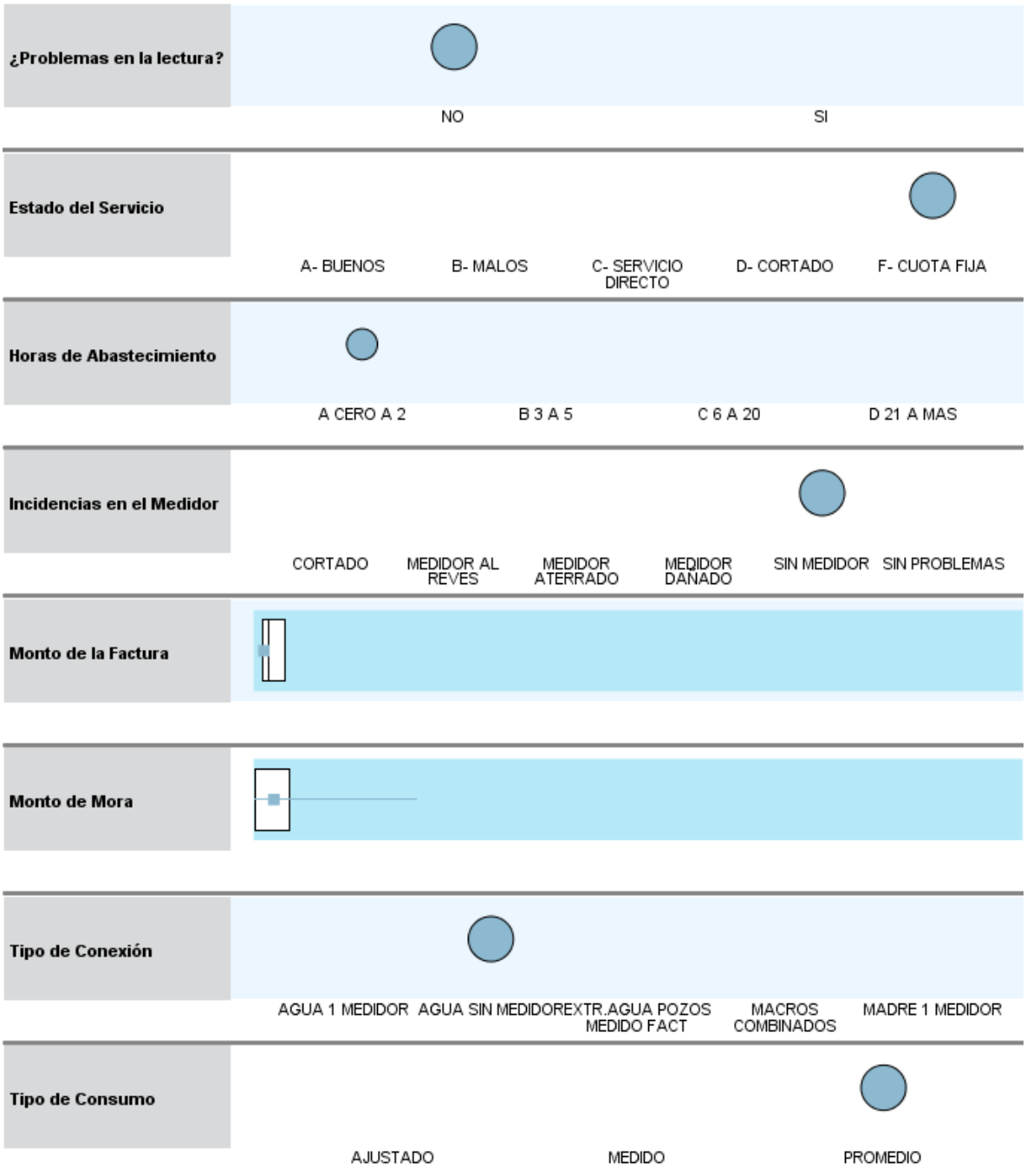


Figura 24. Características del grupo 5

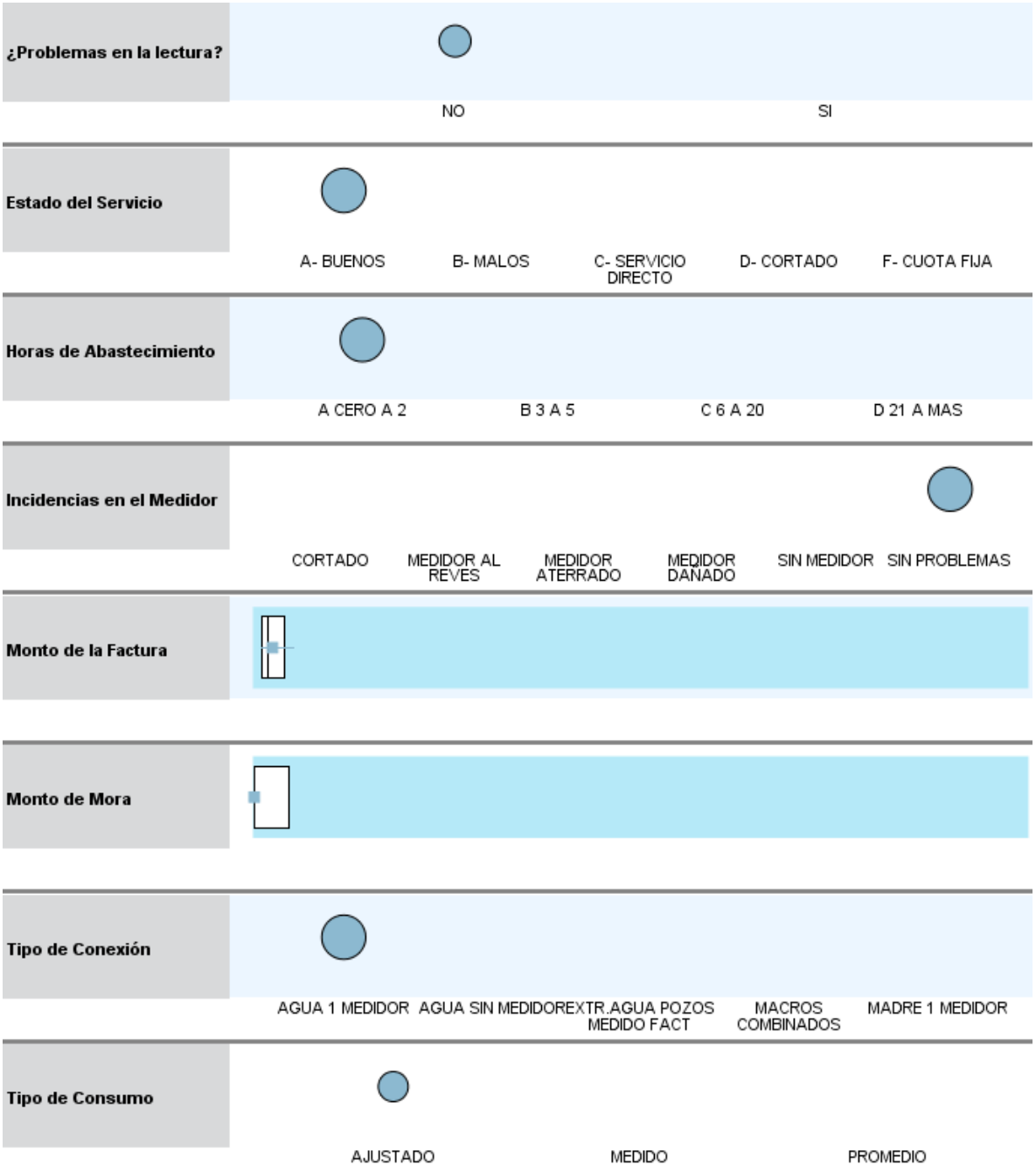


Figura 25. Características del grupo 2

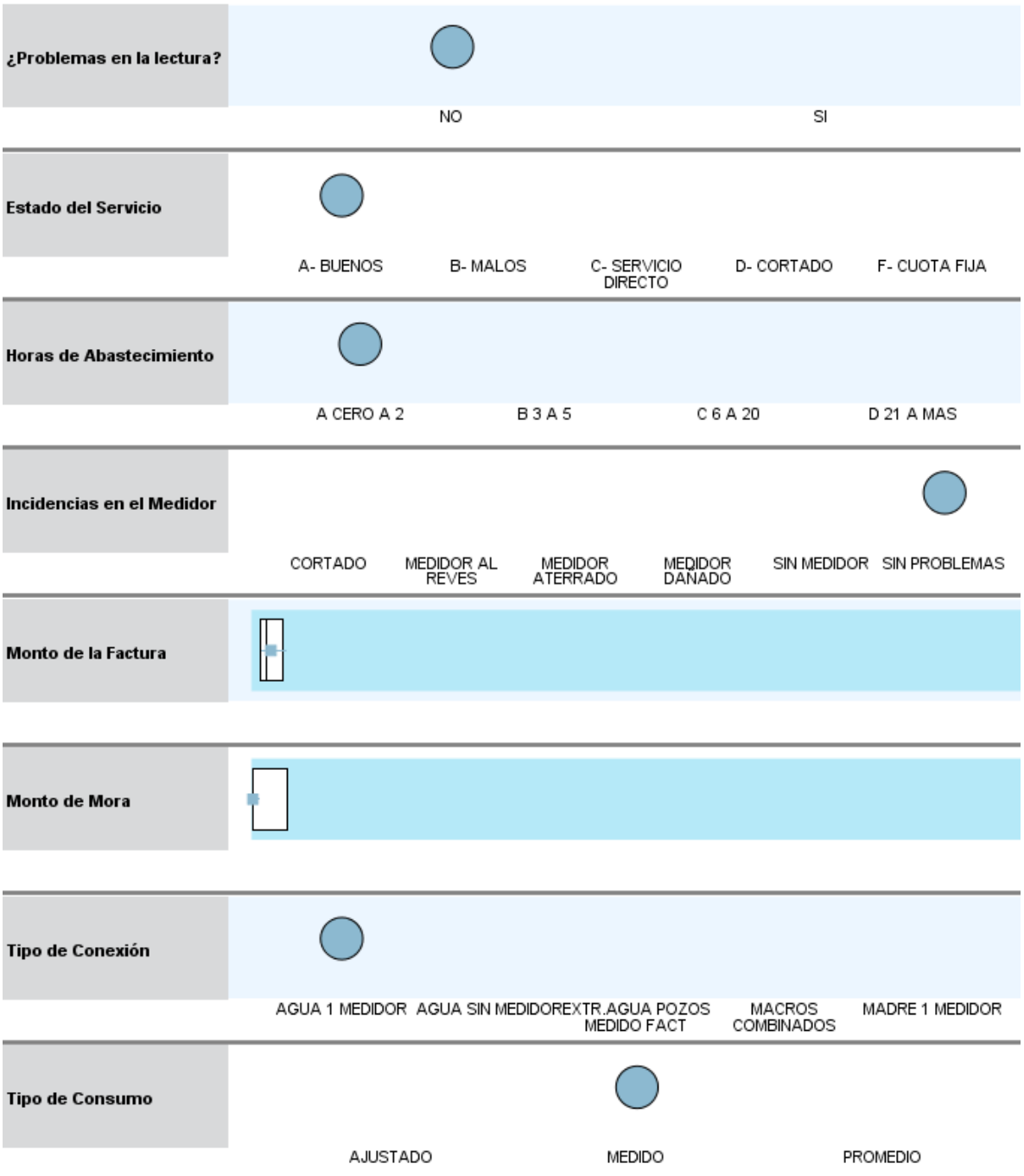


Figura 26. Características del grupo 1

Cronograma de actividades del proceso de investigación						
Etapa	Actividad	Agosto 2017	Septiembre 2017	Octubre 2017	Noviembre 2017	Diciembre 2017
Comprensión del negocio	Capacitación sobre administración y distribución del servicio de agua potable					
	Indagación sobre los conocimientos de expertos, intereses y objetivos					
	Revisión bibliográfica y documentación acerca del fenómeno a investigar					
	Evaluación de recursos administrativos					
Comprensión de los datos	Identificación de recursos de datos					
	Auditoria de datos					
	Definición y creación de variables de interés según criterio de experto					
	Análisis de la calidad de las variables					
Procesamiento de datos	Análisis descriptivo y tratamiento de las variables					
	Aplicación y evaluación de modelos					
	Evaluación, organización de resultados y criterios de experto					
	Resultados y análisis					
Finalización	Creación de informe final					
	Defensa de tesis					