Universidad Nacional de Misiones

Facultad de Ciencias Exactas Químicas y Naturales

Tesis de grado Licenciatura en Sistemas de Información

Predicción de consumos de agua potable utilizando Data Mining

Estudio de caso: Cooperativa de Agua San José (CASAJ)

Autor: Roberto Javier Kachuka

Tutor: Dra. Nancy Beatriz Ganz

Año: 2022

Dedicatoria

A Dios por guiar mi camino,
a mi familia por estar apoyándome
y a mis amigos por acompañarme durante esta carrera.

Resumen

El agua es un recurso primordial para todos los seres vivos de este planeta y, a través de los años, ha venido disminuyendo su disponibilidad debido a destrucción de fuentes naturales, contaminación, incremento poblacional y uso indebido.

La estimación correcta de la demanda de agua potable representa una condición indispensable para la planificación y el diseño de los sistemas de suministro, que en gran medida determina las inversiones necesarias para mejorar la calidad del servicio.

El presente trabajo propone predecir los consumos de agua potable de la CASAJ a través del *data mining*, utilizando como fuente de datos los consumos históricos de cada usuario. Se guía el proceso a través de una metodología denominada CRISP-DM (*Cross Industry Standard Process for Data Mining*), se trabajan los datos en crudo y se generan modelos de predicción basados en redes neuronales.

Los resultados de predicción obtenidos por los modelos satisfacen el enfoque del proyecto, pero específicamente se destaca uno, el modelo de red neuronal recurrente con memoria a corto plazo con una retención de información a 12 meses, el cual tiene una tasa de error por debajo del 8%.

De esta manera, el proyecto contribuye a tener visibilidad sobre consumos futuros para la planificación estratégica y aprovisionamiento de insumos.

Palabras claves: *Agua Potable, Minería de Datos, CRISP-DM, Redes Neuronales, LSTM.*

Abstract

Water is an essential resource for all living organisms on this planet and, over the years, its availability has decreased due to resource depletion, pollution, population growth, and improper use.

The correct estimation of drinking water demand represents an indispensable condition for the planning and design of supply systems, which to a large extent determines the investments needed to improve the quality of the service.

This work aims to predict drinking water consumption of CASAJ through data mining, using the consumption record of each user as data. The process is guided through a methodology called CRISP-DM (*Cross Industry Standard Process for Data Mining*). Raw data are processed, and prediction models based on neural networks are generated.

The results obtained by the models satisfy the focus of the project. Specifically, the recurrent neural network model with short-term memory stands out with 12-month information retention, and the error rate is below 8%.

In this way, this project contributes to strategic planning and supply of inputs for future water consumption.

Keywords: *Drinking Water, Data Mining, CRISP-DM, Neural Networks, LSTM.*

Reconocimientos

Quiero reconocer el gran trabajo de la Dra. Nancy Ganz, quien supervisó y guió el proceso de desarrollo de este trabajo.

A la CASAJ por brindar los datos y el tiempo para lograr el entendimiento del negocio.

Y finalmente, a la facultad por su calidad de enseñanza.

Índice

CAPITULO 1: Introducción 14
1.1 Objetivos 15
1.1.1 Objetivo general 15
1.1.2 Objetivos específicos 15
1.2 Estructura del documento 15
CAPITULO 2: Marco teórico 18
2.1 Conceptos relacionados al servicio de agua potable 18
2.1.1 Usuario 18
2.1.2 Predio 18
2.1.3 Conexión 18
2.1.4 Consumo 18
2.1.5 Lectura 18
2.1.6 Medidor 19
2.1.7 Planta de potabilización 19
2.1.8 Sistema de gestión comercial 19
2.2 Factores que inciden en el consumo de agua 19
2.3 Bases de datos 20
2.3.1 Definición 20
2.3.2 Objetivos 21
2.3.3 Sistemas de procesamiento de transacciones en línea 22
2.4 Descubrimiento de conocimiento en bases de datos 22

2.4.1 Definición....22
2.4.2 Etapas del Proceso....23
2.5 Minería de datos....24
2.5.1 Historia....24
2.5.2 Definición....25
2.5.3 Características y objetivos....25
2.5.4 Tipos de modelos....26
2.5.5 Técnicas....26
2.5.6 Metodologías....27
2.5.7 Estudios relacionados....31
2.6 Técnicas de predicción....32
2.7 Herramientas....34
2.7.1 RapidMiner....34
2.7.2 Python....34
2.8 Evaluación de desempeño de los modelos predictivos....36
2.8.1 Técnicas de evaluación....36
2.8.2 Métricas....37
CAPITULO 3: Materiales y métodos....42
3.1 Descripción del problema....42
3.1.1 Determinación de los objetivos....42
3.1.2 Evaluación de la situación....43
3.1.3 Determinación de los objetivos de *data mining*....44
3.1.4 Plan de trabajo....45
3.2 Comprensión de los datos....45

3.2.1 Recolección de datos iniciales 45
3.2.2 Descripción de los datos....47
3.2.3 Exploración de los datos 58
3.2.4 Verificación de la calidad de los datos....65
3.3 Preparación de los datos 65
3.3.1 Selección de los datos 65
3.3.2 Limpieza de los datos....67
3.3.3 Construcción, integración y formateo de los datos 68
3.4 Modelado....74
3.4.1 Selección del modelo 75
3.4.2 Plan de pruebas 75
3.4.3 Construcción de los modelos 75
3.5 Evaluación 79
3.6 Despliegue 87
3.6.1 Planificación....87
CAPITULO 4: Conclusiones y futuras líneas....90
4.1 Conclusiones 90
4.2 Futuras líneas de investigación....91
Bibliografía 93
Anexo 100

Índice de figuras

Figura 1. Fases de la metodología CRISP-DM 29
Figura 2. Fases de la metodología SEMMA 30
Figura 3. Modelo relacional de base de datos 48
Figura 4. Distribución de datos en la tabla categoria_socio 59
Figura 5. Distribución de datos en la tabla conexiones_situacion 59
Figura 6. Cantidad de estados por año 60
Figura 7. Cantidad de estados 0 agrupado por año y mes 61
Figura 8. Porcentaje de facturas agrupada por rango de consumo 62
Figura 9. Cantidad de facturas por rango de consumo 63
Figura 10. Datos de tabla facturacion_conceptos 64
Figura 11. Datos de tabla facturas_estados 64
Figura 12. Extracción de datos 69
Figura 13. Configuración de conexión a base de datos 69
Figura 14. Selección de atributos 70
Figura 15. Filtros de tabla conexiones y estados 71
Figura 16. Filtros tabla estados 72
Figura 17. Proceso de obtención del atributo consumo 73
Figura 18. Proceso de filtrado de consumos > 0 73
Figura 19. Formato final de datos 74
Figura 20. Modelo LSTM 77
Figura 21. Modelo backpropagation 77

Figura 22. Pérdida de error modelo 1 80
Figura 23. Predicción de consumos (modelo 1) 81
Figura 24. Pérdida de error modelo 2 82
Figura 25. Predicción de consumos (modelo 2) 82
Figura 26. Pérdida de error modelo 3 83
Figura 27. Predicción de consumos (modelo 3) 84
Figura 28. Pérdida de error modelo 4 85
Figura 29. Predicción de consumos (modelo 4) 86

Índice de tablas

Tabla 1. Atributos de tabla facturacion....49
Tabla 2. Atributos de tabla conexiones....50
Tabla 3. Atributos de tabla entidades....52
Tabla 4. Atributos de tabla estados....54
Tabla 5. Atributos de tabla categoria_socio....55
Tabla 6. Atributos de tabla conexiones_situacion....55
Tabla 7. Atributos de tabla zonas....56
Tabla 8. Atributos de tabla facturacion_tipo....56
Tabla 9. Atributos de tabla facturacion_conceptos....57
Tabla 10. Atributos de tabla facturacion_detalle....57
Tabla 11. Atributos de tabla facturas_estados....58
Tabla 12. Descripción del formato final de datos....74
Tabla 13. Parámetros modelo 1 (LSTM)....77
Tabla 14. Parámetros modelo 2 (LSTM)....78
Tabla 15. Parámetros modelo 3 (LSTM)....78
Tabla 16. Parámetros modelo 4 (backpropagation)....79
Tabla 17. Métricas modelo 1....80
Tabla 18. Métricas modelo 2....81
Tabla 19. Métricas modelo 3....83
Tabla 20. Métricas modelo 4....84
Tabla 21. Métricas por modelo....86

Capítulo 1

Introducción

CAPITULO 1: Introducción

La problemática de la alta demanda de agua potable no es ajena a ninguna localidad, provincia o estado, debido a temas económicos y operativos que pueden presentarse. A nivel mundial, todas las empresas del rubro de saneamiento tienen el reto de controlar y poder abastecer a toda la población. Además, existe una preocupación por los niveles de escasez de agua potable [1], [2].

El agua es un recurso primordial para todos los seres vivos de este planeta y, a través de los años, ha venido disminuyendo su disponibilidad debido a destrucción de fuentes naturales, contaminación, incremento poblacional y su uso indebido [3]. Por ende, cada vez existen más ríos secos y, a su vez, una disminución de la calidad del agua [4]. Esto lleva a pensar en el futuro y que sucedería si no se toman las decisiones correctas sobre el consumo responsable de agua para poder mejorar el panorama de este recurso natural.

La estimación correcta de la demanda de agua potable representa una condición indispensable para la planificación y el diseño de los sistemas de suministro, que en gran medida, determinan las inversiones necesarias para mejorar la calidad del servicio [5].

El *data mining* es una técnica que permite extraer conocimiento de las distintas fuentes de datos con las que cuenta una empresa y darle un sentido a la información almacenada [6], de esta forma, el foco de este proyecto está en explorar los datos históricos, explotarlos y predecir consumos futuros para permitir a la Cooperativa de Agua Potable San José tener un control y manejo adecuado del agua a distribuir. Además, podrá estar preparada para conocer en que periodos de tiempo requerirá de más o menos recursos, ya sean químicos, humanos o materia prima para satisfacer la demanda de agua potable, y también, permitirá a la empresa generar planes para controlar consumos excesivos y contribuir al cuidado del medio ambiente.

1.1 Objetivos

1.1.1 Objetivo general

El objetivo de este trabajo es predecir los consumos de agua potable comercial a partir de los datos históricos de la CASAJ haciendo uso de *data mining*.

1.1.2 Objetivos específicos

- Relevar los conceptos relacionados al *data mining*.
- Explorar las diversas aproximaciones disponibles para lograr la predicción sobre datos históricos.
- Definir el conjunto de datos a ser utilizado para llevar a cabo el proyecto.
- Asegurar la calidad de los datos relacionados a la conexión y consumos mensuales de agua de los usuarios de la CASAJ.
- Seleccionar un modelo de predicción que se ajuste a los datos obtenidos.
- Evaluar los resultados de predicción del modelo seleccionado.

1.2 Estructura del documento

La tesis se divide en cuatro capítulos y un anexo. A continuación, se describe brevemente el contenido de los mismos:

- El capítulo 1 presenta una introducción al proyecto junto a sus respectivos objetivos.
- El capítulo 2 describe un marco teórico con los conceptos utilizados a lo largo del presente trabajo.
- El capítulo 3 se enfoca en el proceso, herramientas y modelos utilizados para lograr los objetivos propuestos.

- El capítulo 4 plantea las conclusiones de la tesis y las futuras líneas de investigación que se pueden llevar a cabo.
- El anexo contiene información adicional necesaria para el entendimiento del proyecto.

Capítulo 2

Marco Teórico

CAPITULO 2: Marco teórico

En esta sección se presenta una revisión bibliográfica relevante de los diferentes temas de la problemática abordada. Iniciando por una introducción de los términos relacionados al servicio de agua potable y continuando con conceptos afines a la tecnología a utilizar.

2.1 Conceptos relacionados al servicio de agua potable

A continuación, se describen aquellos conceptos utilizados en el ámbito del servicio de agua potable.

2.1.1 Usuario

Es la persona física o jurídica que se beneficia con la prestación del servicio de agua potable. Éste puede ser el propietario del inmueble o bien el receptor directo del servicio [7].

2.1.2 Predio

Es el inmueble al cual se encuentra vinculado uno o muchos servicios de agua potable [8].

2.1.3 Conexión

Es el abastecimiento de agua a un determinado predio y vinculada a un usuario. Un usuario puede contar con múltiples conexiones. Cada conexión es identificada con un número único de identificación.

2.1.4 Consumo

Es la cantidad (volumen) de agua potable en metros cúbicos (m^3) que abastece a un predio en un periodo determinado [9].

2.1.5 Lectura

Es el proceso que tiene como objetivo recolectar los datos necesarios para determinar el consumo de agua en m^3 de los usuarios y

registrar las distintas incidencias detectadas durante su realización. Se puede realizar de manera automatizada a través de alguna aplicación móvil o bien de la manera tradicional con una planilla de estados en papel [10].

2.1.6 Medidor

Es el instrumento de precisión que mide, registra e indica el volumen de agua expresado en m^3 que pasa por él. Por lo general se encuentra situado frente al domicilio del inmueble y es propiedad de la empresa que brinda el servicio de agua potable [11].

2.1.7 Planta de potabilización

Lugar físico donde se realiza el proceso de potabilización del agua proveniente ya sea de pozos perforados o arroyos.

2.1.8 Sistema de gestión comercial

Es el sistema informático de la Cooperativa de Agua San José, donde se registra toda la información relacionada a las conexiones, predios y consumos de un determinado cliente, además de la operatoria diaria y mensual relacionada a la facturación y cobros [8].

2.2 Factores que inciden en el consumo de agua

El consumo de agua se encuentra determinado por diferentes factores que inciden a su uso. Éstos pueden ser [12]:

- Factor climático: vinculado a la temperatura, precipitación pluvial o humedad que presente el lugar de consumo.
- Factor social: referente a la cantidad de habitantes por vivienda, es decir, como se compone la familia, nivel de educación y estrato social de cada uno de los integrantes.
- Factor económico: depende del ingreso en montos de dinero con el que cuente la familia, el precio del agua en la zona donde se encuentren y el promedio de consumo histórico de la conexión.

- Factor cultural: hace referencia a la formación en cuanto a valores y normas, estilo de vida de las personas y su conducta asociada al cuidado ambiental.

El factor económico marca la diferencia debido a que, al aumentar el ingreso de dinero, aumenta el consumo de agua causado por elementos externos como ser piscinas, jardines o lavadoras. El factor climático también es destacable en el consumo de agua dado que se encuentra vinculado a los niveles de temperatura; a mayor nivel, mayor consumo, esto cambia significativamente en temperaturas bajas [12].

El agua es un recurso vital para todas las personas, y la Organización de las Naciones Unidas (ONU) pronostica que para el año 2.050, incremente en un 20 o 30% el consumo mundial de agua. Los motivos de este aumento son el incremento poblacional, el desarrollo económico y la forma en que las personas hacen uso de dicho recurso [13].

2.3 Bases de datos

2.3.1 Definición

Una base de datos es una colección de datos interrelacionados que contiene información clave y relevante para un individuo o compañía [14]. Puede verse como un gran depósito para almacenar, recuperar, modificar o eliminar datos.

A medida que avanza el tiempo van apareciendo nuevas tecnologías y entre ellas nuevas bases de datos, pero aquí haremos mención a las más utilizadas, entre ellas se encuentran las bases de datos relacionales, las orientadas a objetos, las multidimensionales y las documentales. Cada una presenta características particulares, pero la más destacada por su uso y popularidad es la base de dato relacional [15], [16].

Un sistema de gestión de bases de datos (SGBD) es un sistema computarizado que proporciona a los usuarios mecanismos para crear, eliminar y manipular bases de datos de manera práctica y eficiente. En otras palabras, un SGDB oculta a los usuarios la manera en la que realiza las operaciones a bajo nivel [17].

Un SGDB relacional proporciona dos tipos de lenguajes [14]:

- Lenguaje de definición de datos (LDD): utilizado para especificar los esquemas en las bases de datos.
- Lenguaje de manipulación de datos (LMD): utilizado para realizar consultas y modificaciones en las bases de datos.

En la práctica, ambos forman parte de un único lenguaje, como por ejemplo el Lenguaje de Consulta Estructurado (SQL[1]) ampliamente usado para la manipulación de los datos.

2.3.2 Objetivos

Una base de datos tiene como principal propósito organizar y almacenar datos para facilitar su manejo. Además, se enfoca en evitar la redundancia e inconsistencia de los datos, mejorar el funcionamiento de acceso concurrente a los datos, reducir los tiempos de respuesta y proporcionar mecanismos de control de acceso según el usuario que este accediendo [14].

En una compañía pueden existir distintas fuentes de datos y cada una con gran escala, al comienzo puede ser suficiente manejar estas fuentes de datos para cubrir las necesidades de almacenamiento y lectura de datos. Pero al pasar el tiempo será necesario aplicar otras técnicas de extracción

[1] *Structured Query Language*

de información que permitan el cruzamiento de los datos y que ayuden a la toma de decisiones [14].

2.3.3 Sistemas de procesamiento de transacciones en línea

Los sistemas de procesamiento de transacciones en línea (*OLTP - online transactional processing*) son bases de datos enfocadas al procesamiento de las transacciones, estas pueden ser inserción, modificación o eliminación de los datos. Estos sistemas permiten un óptimo de acceso a los datos, tanto para las tareas de lectura como escritura [18].

Los sistemas OLTP son utilizados por la mayoría de las compañías para el almacenamiento de las operaciones diarias, por ende, realizan miles de transacciones por segundo. Un gran problema se presenta cuando estos sistemas crecen, debido a que fueron pensados para introducir información y no para extraerla, en este sentido se pierde la eficiencia y sencillez para realizar consultas [18].

Las principales características de estos sistemas son:

- Gran número de usuarios accediendo simultáneamente.
- Su estructura generalmente es normalizada.
- Actualización de datos a diario.

2.4 Descubrimiento de conocimiento en bases de datos

2.4.1 Definición

El descubrimiento de conocimiento en bases de datos en inglés *knowledge discovery in databases* (KDD) hace referencia al proceso global de extraer conocimiento de grandes volúmenes de datos, cubriendo los aspectos de almacenaje y acceso a los datos, optimización de acceso a

dichos datos, interpretación y visualización de los resultados, y en general, la interacción entre el hombre y la máquina [6].

KDD puede verse como un conjunto de técnicas y herramientas aplicadas a un proceso no común de extraer y presentar conocimiento implícito, que previamente era desconocido para transformarse en algo útil y comprensible [19].

KDD hace uso de grandes conjuntos de datos, con objeto de predecir de forma automatizada, tendencias o comportamientos y descubrir modelos previamente desconocidos.

2.4.2 Etapas del Proceso

El proceso KDD es iterativo y altamente interactivo con el usuario, es decir, la mayoría de las decisiones las debe tomar el usuario [20].

Este proceso está formado por los siguientes pasos [21]:

- Comprensión del dominio: aquí se definen los límites y los objetivos de la aplicación. Se pretende llegar a un entendimiento completo del dominio relacionado.
- Selección de datos: esta etapa se enfoca en seleccionar el conjunto de datos objetivo o una muestra representativa del mismo. Se enfoca en obtener todos los datos que sean relevantes, ya que los mismos pueden provenir de distintas fuentes de datos.
- Preparación de los datos: tiene como objetivo mejorar la calidad de los datos, es por esto que los datos se someten a una limpieza, transformación, integración y reducción. Es muy frecuente que existan datos incompletos, que presenten ruido o sean inconsistentes, lo cual conlleva a que en la etapa siguiente se obtengan resultados inválidos.

- Minería de Datos: esta etapa tiene como objetivo la búsqueda y descubrimiento de patrones a través de distintos métodos para la generación de un modelo que represente el conjunto de datos.
- Interpretación: aquí se interpretan los patrones obtenidos y se utilizan diversas técnicas de visualización para lograr un mejor entendimiento. Por otro lado, se consolida el conocimiento generado para incorporarlo a otros sistemas o bien, para documentarlo y reportarlo. Desde este punto es posible retornar a cualquiera de los pasos anteriores.

Generalmente, cada una de las fases puede tener soporte de alguna herramienta que permita o facilite el proceso sobre el cual se esté trabajando, algunas pueden ser *RapidMiner*, *Metabase*, *Python*, *R*, *Matlab*, entre otras [22]–[24].

2.5 Minería de datos

2.5.1 Historia

El concepto de minería de datos, en ingles *data mining* (DM), comienza a presentarse en los años sesenta, junto a estadísticos de esa época que buscaban encontrar correlaciones sin una presunción válida en bases de datos con datos dispersos. En ese tiempo se utilizaban términos como *data fishing* o *data archaeology* para hacer alusión al *data mining* [25].

Ya a partir de los años ochenta, dicho término comenzó a consolidarse y para finales de esa década solo pocas empresas se dedicaban al rubro. A medida que avanzaba el tiempo, más empresas lo comenzaron a implementar, dado que el DM permitía extraer información de valor que se encontraba oculta en grandes volúmenes de datos [25].

Actualmente, el DM se ha incorporado a entornos gubernamentales, universidades, hospitales y diversas organizaciones. Se ha utilizado en distintos campos que van desde las ventas al por menor hasta la lucha contra el terrorismo [25].

2.5.2 Definición

El DM se considera como el proceso de descubrir conocimiento oculto a partir de grandes conjuntos de datos, para presentarla al usuario de manera comprensible y contribuir en la toma de decisiones [6].

La definición tradicional más conocida es la del autor Fayyad [21] que considera al DM como: "*Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos*".

De manera más abstracta puede considerarse como el proceso de construcción de un modelo que se ajusta a determinados datos para proporcionar conocimiento útil [6].

El DM muchas veces es utilizado como sinónimo de KDD, pero todo depende del contexto sobre el cual se esté trabajando, desde el punto de vista académico el DM es un paso particular en el proceso de KDD. Por otro lado, en entornos comerciales, ambos términos se utilizan de manera indistinta [26].

2.5.3 Características y objetivos

El DM se destaca por explorar grandes conjuntos de datos que tal vez son difíciles de asimilar o entender con los métodos de consulta tradicional. Puede trabajar con datos históricos, es decir, datos que se van acumulando a lo largo de la vida operativa de una empresa. Su foco está en aprovechar el valor de la información almacenada y lograr que los

directivos tengan un mejor conocimiento de su negocio para tomar decisiones acertadas [25], [27].

La información que se obtiene a través de un proceso de DM ayuda a una compañía a elegir los cursos de acción, tomar decisiones proactivas y definir estrategias competitivas [27].

Existen herramientas que automatizan el proceso de búsqueda y explotan los datos a través del uso de diferentes técnicas con el fin de correlacionarlos según las necesidades. Generalmente, estas herramientas tienen como objetivo encontrar datos extraños, patrones, tendencias o desviaciones [27].

2.5.4 Tipos de modelos

El DM puede generar distintos modelos según el enfoque deseado y las características de los datos [6]. Son los siguientes:

- Descriptivos o no supervisados: tiene como fin descubrir patrones y tendencias sobre el conjunto de datos sin tener ningún tipo de conocimiento previo de la situación a la cual se quiere llegar. Descubre patrones o relaciones describiendo los datos.
- Predictivos o supervisados: se basa en crear un modelo sobre los datos o resultados ya conocidos. Intenta predecir el valor de un atributo del cual se desconoce el resultado, estableciendo relaciones entre los ya conocidos.

2.5.5 Técnicas

Las técnicas de DM pueden distinguirse de la siguiente manera [21]:

- Modelado predictivo: se enfoca en predecir campos de una base de datos basado en otros campos. Si el campo que se predice es una variable numérica continua entonces es un problema de regresión. En cambio, si el campo es categórico entonces es un problema de

clasificación, algunos métodos pueden ser árboles de decisiones, máquinas de vectores de soporte, redes neuronales y K vecinos más cercanos (K-NN - *K-nearest neighbor*).

- *Clustering*: La agrupación en clústeres no especifica los campos que se van a predecir, sino los objetivos que separan los elementos de datos en subconjuntos que son similares entre sí. Algunos ejemplos de métodos que aplican esta técnica son: *K-means*, agrupamiento espacial basado en densidad de aplicaciones con ruido (*DBSCAN - density-based spatial clustering of applications with noise*), *Mean Shift*, entre otros.
- Resumen de datos: se enfoca en extraer patrones que describen subconjuntos de datos. Hay dos métodos, uno es tomar secciones horizontalmente y la otra es tomar secciones verticalmente; la primera es para casos específicos y la última, para campos.
- Modelado de dependencias: orientado a la creación de una estructura causal dentro de los datos. Los modelos de causalidad pueden ser probabilísticos o determinísticos.
- Detección de cambios y desviaciones: aquí es importante la información de la secuencia, ya sea de series de tiempo o algún otro orden. El orden de las observaciones es relevante y debe tenerse en cuenta.

2.5.6 Metodologías

Una metodología define una estructura a seguir a través de los proyectos de análisis de datos, y ayuda a los investigadores con una estructura sistemática para obtener mejores resultados. Se presentan las siguientes metodologías de *data mining* [28].

2.5.6.1 Cross industry standard process for data mining

CRISP-DM surge a fines de la década del 90 y proporciona una descripción normalizada del ciclo de vida de un proyecto de análisis de datos [29].

El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre cada una de ellas [30]:

- Entendimiento del negocio: esta fase se encarga de la comprensión de los objetivos del negocio y de la evaluación de la situación actual en la que se encuentra.
- Estudio y comprensión de los datos: se enfoca en la colección de datos inicial a fin de conocer su estructura y poder determinar la calidad de los mismos.
- Preparación de los datos: esta fase intenta construir el conjunto de datos que van a ser utilizados en la siguiente fase. Incluye tareas como selección de tablas, registros y atributos, así como la selección y limpieza de los datos.
- Modelado: aquí se seleccionan y aplican las técnicas de modelado que cubran los requerimientos del proyecto. Dependiendo de las características del conjunto de datos se puede optar por alguna de las técnicas descriptas en el apartado 2.5.5.
- Evaluación: aquí se evalúa el grado de acercamiento del modelo generado a los objetivos del negocio.
- Despliegue: se enfoca en propagar los resultados obtenidos a los usuarios finales.

La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario. En la Figura 1 podemos ver una ilustración del proceso.

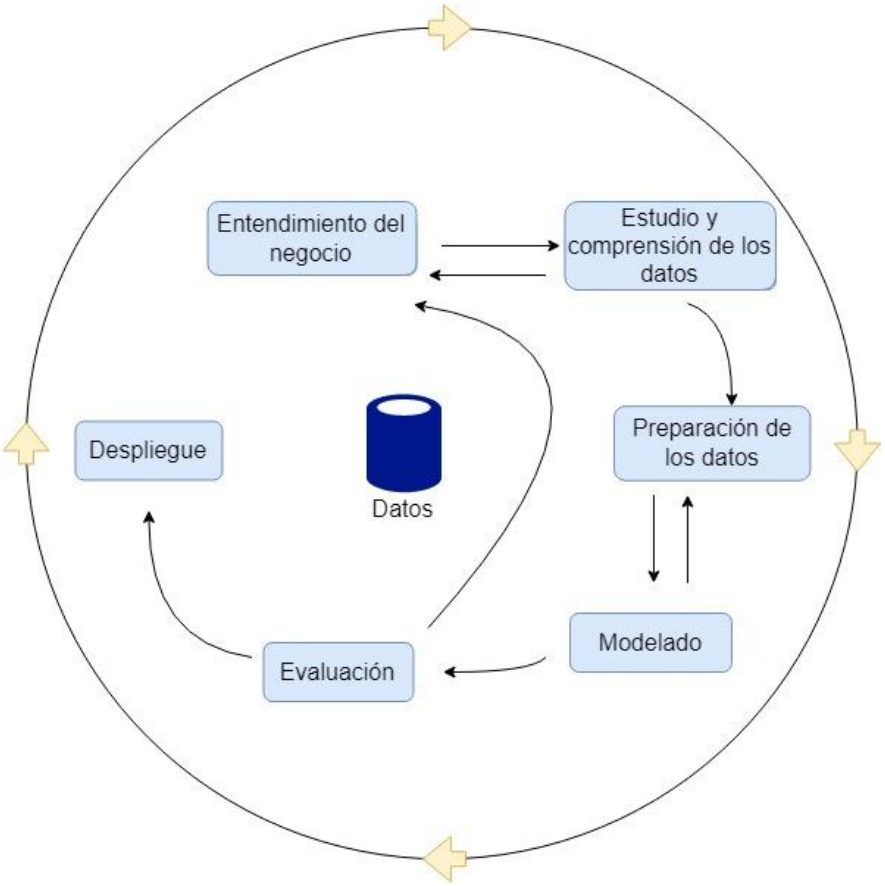

Figura 1. Fases de la metodología CRISP-DM[2]

2.5.6.2 SEMMA

Fue desarrollada por el *SAS Institute* y es definida como el proceso de selección, exploración y modelado de grandes volúmenes de datos con el fin de descubrir patrones de negocios desconocidos [30]. En la Figura 2 se describen las cinco fases

[2] *Imagen adaptada de [29]*

básicas que presenta el proceso.

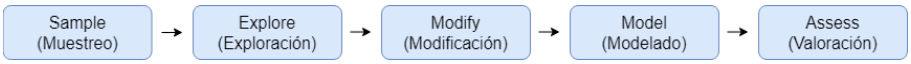

Figura 2. Fases de la metodología SEMMA[3]

Esta metodología se enfoca principalmente en los aspectos técnicos, excluyendo así las actividades de análisis y comprensión del problema [31].

2.5.6.3 P3TQ

Propuesta en el año 2003 por Dorian Pyle, su nombre viene de *Product, Place, Price, Time and Quantity* correspondiente a Producto, Lugar, Precio, Tiempo y Cantidad respectivamente [32].

Es una metodología que plantea la formulación de dos modelos, el modelo de negocios y el modelo de explotación de información [32].

El modelo de negocios se utiliza para identificar el problema y los requerimientos a cubrir, contempla si el proyecto no tiene definido el problema o la oportunidad de negocio, partiendo de un análisis de la cadena de valor organizacional, es decir, las relaciones precio/lugar/producto/tiempo/cantidad que son importantes para la empresa [30].

Por otro lado, el modelo de explotación de información, brinda una serie de pasos para la construcción y ejecución de modelos de *data mining* a partir del modelo previamente mencionado [30].

[3] *Imagen adaptada de [30]*

2.5.6.4 Elección de la metodología

Tanto CRISP-DM, SEMMA y P3TQ brindan una guía estructurada y contienen distintas fases que se relacionan entre sí para guiar un proyecto de *data mining*. SEMMA no comienza con un análisis de requerimientos y entendimiento del negocio, como si lo hacen CRISP-DM y P3TQ. Por otro lado, CRISP-DM y P3TQ fueron desarrolladas como metodologías neutras, de libre distribución, no es así el caso de SEMMA que se relaciona con los productos comerciales de la *SAS Institute* [30], [31] .

CRISP-DM y P3TQ presentan similitudes en cuanto al nivel de detalle con el que describen las tareas para cada fase del proceso. Al hablar de la evaluación de los resultados CRISP-DM es el único que lo realiza en base al modelo generado y a los objetivos del proyecto [30].

Entonces, teniendo en cuenta los aspectos previos mencionados, el presente trabajo utilizó la metodología CRISP-DM para guiar todo el proceso de desarrollo.

2.5.7 Estudios relacionados

Según un estudio realizado en Ecuador [33], debido a las distintas necesidades de consumo de agua potable y factores de escases de la misma, se llevó a cabo un trabajo para lograr pronosticar el consumo de agua potable con alta precisión a través de redes neuronales, utilizando un conjunto de datos genérico con muestras de 40 años, se plantearon distintos modelos del mismo tipo variando sus hiper parámetros y luego se seleccionó el mejor a partir de distintas métricas de error.

Otro estudio se enfoca en comparar la calidad de predicción de métodos estadísticos sobre modelos predictivos de *data mining* teniendo como base consumos de agua en formato de series de tiempo. La teoría se

basa en que si la serie temporal presenta una componente lineal entonces los métodos estadísticos arrojan buenos valores de predicción, en cambio, si se cuenta con una componente no lineal entonces las redes neuronales se presentan como una alternativa más precisa para predecir el comportamiento futuro de la serie [34].

Por otro lado, en la provincia de los Andes – Chile, se llevó adelante un estudio similar al antes mencionado, pero con un enfoque diferente. En este caso, la empresa sanitaria ya contaba con un pronóstico de agua mensual producida, entonces, la idea fue pronosticar el nivel de agua faltante no facturada. Esto fue posible gracias a que se contaba con los datos de agua total producida que generaba la planta de potabilización y se contrastó con el total facturado almacenado en el sistema de gestión comercial. Se utilizaron redes neuronales logrando tasas de error de predicción menor al 2% [35].

También, otro estudio, con objetivos similares a los de ésta tesis, pero con foco en los consumos energéticos, hizo uso del *data mining* para realizar un análisis completo de los datos históricos de 30 edificios y luego, a través de distintos métodos, lograr la predicción a corto y medio plazo. Se analizaron las facturas de consumo eléctrico de cada usuario de los edificios, para así, obtener el comportamiento de los puntos de facturación [36].

2.6 Técnicas de predicción

A continuación, se describen técnicas y modelos de predicción más utilizados [37]:

- Árboles de decisión: permiten evaluar mediante una representación gráfica los posibles resultados de una decisión compleja. Se utilizan algoritmos de aprendizaje supervisados para realizar la analítica. La estructuración es

similar a un árbol, puesto que se parte de un único nodo que después se ramifica en variables o alternativas [38].

- Redes neuronales: se denomina así por su similitud con las neuronas del cerebro. En esencia, se trata de varias capas que, a su vez, están relacionadas con elementos sencillos conectados entre sí, cada uno de estos elementos, es conocido como una neurona. Esta técnica ha ganado importancia los últimos años gracias al desarrollo de la inteligencia artificial. Al final, lo que buscan las nuevas tecnologías es asimilarse al patrón de acción del cerebro. Algunos tipos son:
    - Redes neuronales monocapa: es la más sencilla ya que se tiene una capa de neuronas que proyectan las entradas a una capa de neuronas de salida donde se realizan los diferentes cálculos [39].
    - Redes neuronales multicapa: es una evolución de la red neuronal monocapa, esta incorpora capas ocultas que permiten representar funciones no lineales. Está compuesto por una capa de entrada, una capa de salida y n capas ocultas entremedias [39]. Un algoritmo de los más usados es la propagación hacia atrás conocido como *backpropagation*, este algoritmo permite que los errores de predicción de la red sean propagados del final al comienzo, y de esta manera, cada neurona de la red podrá calcular que tan culpable fue y reajustarse en la próxima iteración [40].
    - Redes neuronales recurrentes: ésta no tiene una estructura de capas definida, sino que permiten conexiones arbitrarias entre las neuronas, incluso pudiendo crear ciclos, con esto se consigue crear la temporalidad, permitiendo que la red tenga memoria. Las más conocidas son las redes recurrentes simples (SRN [4]por sus siglas en

[4] *Simple Recurrent Network*

inglés) y las redes recurrentes con memoria a corto plazo (LSTM [5] por sus siglas en inglés) [41].

- Análisis de regresión: permite relacionar distintas variables entre sí. Cuenta con la siguiente clasificación, el análisis de regresión lineal y el análisis de regresión logística. El primero trabaja con variables dependientes, independientes y elementos aleatorios. Y la segunda, predice el resultado de una variable categórica con variables predictoras [42].

2.7 Herramientas

2.7.1 RapidMiner

Es una excelente herramienta para el análisis de datos, está basada en Java y es utilizada ampliamente en todas las fases del DM. Las versiones anteriores (v.5 o inferiores) eran de código abierto. La última (v. 6) es propietaria por ahora, con varias opciones de licencia (*Starter*, Personal, Profesional, Empresarial e incluso una Educativa). La versión *Starter* es gratuita con limitaciones en cuanto al tamaño máximo de la memoria (1 GB), los archivos de entrada y el tamaño máximo de registros al momento de obtener gráficas o generar reportes. La herramienta se ha hecho muy popular en los últimos años y cuenta con un gran apoyo de la comunidad [22].

2.7.2 Python

Python es un lenguaje de programación muy potente. Tiene estructuras de datos eficientes de alto nivel y un enfoque simple pero efectivo. Su sintaxis elegante y su tipificación dinámica, junto con su naturaleza interpretada, lo convierten en un lenguaje ideal para el rápido desarrollo de *scripts* y aplicaciones en muchas áreas. El éxito de Python se

[5] *Long Short-Term Memory*

debe a que, al ser de libre acceso, tiene una gran comunidad que desarrolla y mantiene diversos paquetes para distintas finalidades [43].

En el ámbito de DM, Jupyter es el *software* de libre acceso que brinda una interfaz amigable para poder programar y documentar código orientado a análisis de datos [44], también esta Spyder que es un entorno integrado de desarrollo y aparte de ejecutar código Python permite ejecutar grandes programas [45], y finalmente, Anaconda, que abarca un conjunto de aplicaciones y librerías orientadas 100% al análisis de datos [46].

Algunas de las librerías más utilizadas para el DM basadas en Python se describen a continuación:

2.7.3.1 NumPy

Es un paquete que es utilizado para el cálculo numérico, análisis y descripción de los datos con los cuales se esté trabajando. Está preparado para soportar grandes volúmenes de datos [47].

2.7.3.2 Pandas

Es una librería de código abierto que permite y facilita la manipulación y tratamiento de los datos. Brinda estructuras para la correcta y efectiva manipulación de los datos [48].

2.7.3.3 Keras

Es un paquete de Python que acelera la creación de redes neuronales, cuenta con diversos tipos de redes neuronales y, con tan solo importarlas y definir sus métricas, ya pueden ser utilizadas [49].

2.7.3.4 Sk-learn

Es un paquete gratuito de Python con numerosos algoritmos de DM y funcionalidades para la obtención de métricas. El paquete sigue mejorando, aceptando valiosas contribuciones de la gran

comunidad que se encuentra detrás. Uno de sus principales puntos fuertes es una documentación en línea bien escrita para todos los algoritmos implementados. La documentación bien escrita es un requisito para cualquier colaborador y se valora más que un montón de implementaciones de algoritmos documentadas de forma deficiente [22].

2.7.3.5 Matplotlib

Es un paquete que permite trazar gráficos totalmente personalizados, entre los más comunes tenemos a los gráficos de barra, histogramas, gráficos de línea, diagramas de dispersión, entre otros [50].

2.8 Evaluación de desempeño de los modelos predictivos

La evaluación de desempeño es un punto clave para determinar la calidad de los modelos de predicción, sin esta, es imposible cuantificar o medir si tienen un impacto positivo o negativo.

2.8.1 Técnicas de evaluación

La efectividad de predicción de los modelos depende en gran parte del conjunto de datos utilizado para el entrenamiento y por otro lado del conjunto de datos usado para realizar las pruebas. El criterio de división y selección de estos datos impacta significativamente en los resultados, las técnicas más utilizadas son *holdout* y *k-fold* [51].

2.8.1.1 Holdout

La técnica de retención, *Holdout* en inglés, es la más sencilla y su objetivo es dividir los datos originales en dos conjuntos, uno de entrenamiento, utilizado para que el modelo pueda aprender las distintas características de los datos, y otro

conjunto de prueba para evaluar el rendimiento, el primer conjunto suele tener la mayor parte de los datos originales [51].

2.8.1.2 K-fold

Esta técnica consiste en dividir al conjunto de datos originales en K subconjuntos de aproximadamente el mismo tamaño, uno de los subconjuntos se utiliza como datos de prueba y los restantes grupos como datos de entrenamiento. El proceso es iterativo y se repite K veces utilizando un grupo distinto de entrenamiento en cada iteración, una vez finalizado, se generan K estimaciones del error cuyo promedio se utiliza como estimación final [51].

2.8.2 Métricas

Para evaluar el rendimiento de los modelos predictivos se requieren de métricas que permitan exponer el nivel de precisión de cada uno. A continuación, se describen las comúnmente más utilizadas [52].

2.8.2.1 Raíz del Error Cuadrático Medio

La raíz del error cuadrático medio (RMSE[6] por sus siglas en inglés) es la manera más habitual de evaluar un modelo de regresión. Mediante esta medida se calculan las diferencias entre los valores pronosticados por el modelo o un estimador y los valores reales a partir de los cuales se ha creado el modelo. El error cuadrático medio es una medida de la media de los cuadrados de los errores. Por error se entiende a la diferencia entre el valor estimado y el valor real [53]. El error cuadrático medio (Ecuación 1) se calcula de la siguiente manera:

[6] *Root Mean Square Error*

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - y'_j)^2}$$

Ecuación 1.

2.8.2.2 Error Absoluto Medio

El error absoluto medio (MAE[7] por sus siglas en inglés) es otra forma de evaluar la calidad en los modelos de regresión. Esta medida sirve para calcular la diferencia entre las predicciones hechas por un estimador y los valores reales. La diferencia entre la anterior surge del principal problema que tiene calcular el error cuadrático medio, y es que al elevar al cuadrado la diferencia se tiende a dar demasiado peso a los errores más extremos, afectando al resultado final. Utilizando el error absoluto medio se puede limitar este problema [53]. La fórmula para calcular el error absoluto medio (Ecuación 2) es la siguiente:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - y'_j\right|$$

Ecuación 2.

2.8.2.3 Error Cuadrático Medio

El error cuadrático medio (MSE[8] por sus siglas en inglés) es quizás la métrica más simple de calcular en evaluaciones de regresión. Mide el error cuadrático promedio de las predicciones y

[7] *Mean Absolute Error*
[8] *Mean Squared Error*

para cada punto calcula la diferencia cuadrada entre las predicciones y el objetivo, y luego promedia esos valores [53]. La fórmula para su cálculo (Ecuación 3) es la siguiente:

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - y'_i)^2$$

Ecuación 3.

2.8.2.4 Valor de Pérdida de Error

El valor de pérdida de error indica que tan bien funciona una red neuronal. Si el valor es alto, determina que la red tiene bajo desempeño, en cambio, si su valor es bajo, quiere decir que la red tiene gran capacidad de predicción. La manera de obtener este valor es la diferencia entre el dato esperado y el que la red neuronal retorna [33].

Capítulo 3

Materiales y Métodos

CAPITULO 3: Materiales y métodos

El fin de este capítulo es describir cada una de las fases de la metodología CRISP-DM. Comienza por explicar la problemática y el entorno de trabajo, luego, especifica los detalles de tratamiento y construcción de los datos que son entradas para la modelación de la solución. Finalmente, expone los resultados logrados y concluye con el plan de implementación.

3.1 Descripción del problema

A continuación, se describe cada una de las tareas con las que cuenta esta fase. Al finalizar esta sección se podrá tener un entendimiento del área que se está abordando.

3.1.1 Determinación de los objetivos

3.1.1.1 Contexto

La CASAJ presta su servicio en la localidad de San José – Misiones desde el año 1966. La empresa cuenta con tres abastecimientos, el primero y principal es un pozo perforado de 197 metros ubicado en la planta de potabilización, luego el Arroyo Pindapoy Grande ubicado aproximadamente a 580 metros de distancia de la planta y finalmente una perforación secundaria o de menor escala de 109 metros ubicado a 50 metros del Arroyo Pindapoy Grande.

Del pozo principal se extraen $40m^3$/hora, del pozo secundario $15m^3$/hora y del arroyo $30m^3$/hora, cada uno de ellos cuenta con una bomba que hace llegar el agua cruda (sin tratar) a la sección de potabilización donde se realiza su correspondiente tratamiento; luego el agua pasa al tanque principal que tiene una

capacidad de almacenamiento de 100.000 litros y finalmente es distribuida a lo largo de la red.

Según datos de la CASAJ, en temporadas de bajos consumos se utilizan aproximadamente 400.000 litros al día para abastecer a toda la ciudad, pero por pérdidas operacionales y comerciales entre 500.000 a 800.000 litros extras son requeridos. Por otro lado, en temporadas de alto consumo esto se duplica tanto en consumos como en pérdidas. También, la empresa hace mención que aproximadamente del 100% de agua potable producida, se factura entre el 55% al 70%.

Actualmente, es la única empresa de saneamiento de agua potable en dicha zona, abarcando así, gran parte de la población. El municipio de San José cuenta con una población aproximada de 7.000 habitantes [54].

Toda la información previa mencionada fue obtenida a través de una entrevista realizada a un personal de la CASAJ, ver Anexo – Entrevistas.

3.1.1.2 Objetivos a Alcanzar

Lo que se espera del proyecto es predecir los consumos para un determinado periodo de tiempo, basado en los registros históricos de los usuarios de la CASAJ, a través de un modelo pertinente que se adapte a los datos con los que se cuenta.

3.1.2 Evaluación de la situación

La CASAJ no cuenta con ningún mecanismo que facilite información acerca de los consumos totales generados mes a mes y una planificación adecuada sobre los consumos e insumos futuros que se producirán. Actualmente, presentan problemas tales como:

- Desconocimiento sobre los consumos totales generados, lo que conlleva a no ser transparentes con los usuarios finales.
- Problemas de potabilización de agua en determinados periodos del año.
- Faltante de insumos para el proceso de potabilización.

Además, la insuficiencia de agua en temporadas de consumos elevados es una realidad, así como el extremo cuidado de un bien natural de primer orden, entonces tener un control sobre los futuros consumos es una necesidad.

La CASAJ ha expresado la necesidad de predecir los consumos comerciales a través de la extracción de conocimiento de los datos históricos almacenados en su sistema de gestión.

3.1.2.1 Recursos Disponibles

Se cuenta con una base de datos que contiene los datos relacionados a los consumos históricos de los usuarios de la CASAJ desde el año 2015 al 2020, por lo que a priori se puede afirmar que se dispone de la cantidad de datos suficientes para afrontar la problemática.

3.1.3 Determinación de los objetivos de *data mining*

El objetivo principal de DM es predecir los consumos comerciales de agua potable partiendo de los datos históricos entre el año 2015 y 2020 de cada usuario de la CASAJ.

Para lograr esta predicción de los consumos mensuales, es necesario adaptar un modelo que pueda comportarse acorde a los datos y brindar buenas métricas de predicción que serán descriptas más adelante.

3.1.4 Plan de trabajo

El proyecto se dividirá en las siguientes etapas para facilitar su organización y estimación de duración:

#	Etapa	Estimación
1	Relevamiento bibliográfico e investigación de la situación actual.	35 días
2	Recolección de las fuentes de datos.	7 días
3	Análisis de la estructura de los datos y de la información del conjunto de datos.	30 días
4	Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la aplicación de técnicas de DM sobre ellos.	45 días
5	Elección de la técnica de modelado y ejecución de la misma sobre los datos.	45 días
6	Análisis de los resultados obtenidos.	20 días
7	Producción de informes con los resultados obtenidos en función de los objetivos.	10 días
8	Redacción del documento de tesis	150 días
9	Presentación de los resultados finales.	10 días

3.2 Comprensión de los datos

Esta fase incluye la recolección, descripción, exploración y verificación de la calidad de los datos. En esta instancia se procede a realizar un primer contacto con los datos en crudo con el fin de familiarizarse, identificar sus relaciones y averiguar en qué situación se encuentran.

3.2.1 Recolección de datos iniciales

Se cuenta con una base de datos que contiene registros del sistema de gestión a partir de enero del 2015 hasta abril del 2020 de los usuarios

de la localidad de San José – Misiones. En dicha base de datos se encuentra información relacionada a:

- Datos personales del usuario.
- Datos del predio donde se encuentra la conexión.
- Consumos mensuales.
- Facturación mensual de usuarios activos.
- Deudores.
- Intereses generados.
- Usuarios del sistema.
- Compras a proveedores, entre otros datos enfocados a la configuración del sistema de gestión.

Cabe mencionar que, para el presente trabajo, los datos personales de los usuarios no serán utilizados para llevar a cabo el análisis, únicamente se hará uso de algún identificador que lo distinga de manera unívoca.

Los aspectos que serán necesarios para continuar con el proceso de DM son los siguientes:

- Información de facturación mensual de los metros cúbicos de agua por cada usuario.
- Mediciones de estado por cada periodo.
- Información del tipo de conexión de cada usuario.

Para llevar a cabo el proceso de DM y lograr el éxito del proyecto es necesario que los datos satisfagan los requerimientos previos mencionados. A priori se desconoce que tablas serán realmente útiles. A continuación, se listan aquellas presentes en la base de datos dada:

- conexiones
- conexiones_situacion

- entidades
- categoria_socio
- facturacion
- facturacion_estados
- facturacion_detalle
- facturacion_conceptos
- facturacion_tipo
- estados
- zonas

Puede que en pasos posteriores no todas las tablas sean utilizadas en el proceso de DM, pero eso se determina en el punto 3.3.

3.2.2 Descripción de los datos

Los datos se encuentran en una base de datos PostgreSQL. Se utilizó DBeaver Community [55], que es una herramienta de libre acceso, para la generación del modelo relacional (Figura 3) teniendo en cuenta las tablas relevantes para el proyecto. Es importante recordar que los datos se encuentran en el rango de enero del 2015 hasta abril del 2020.

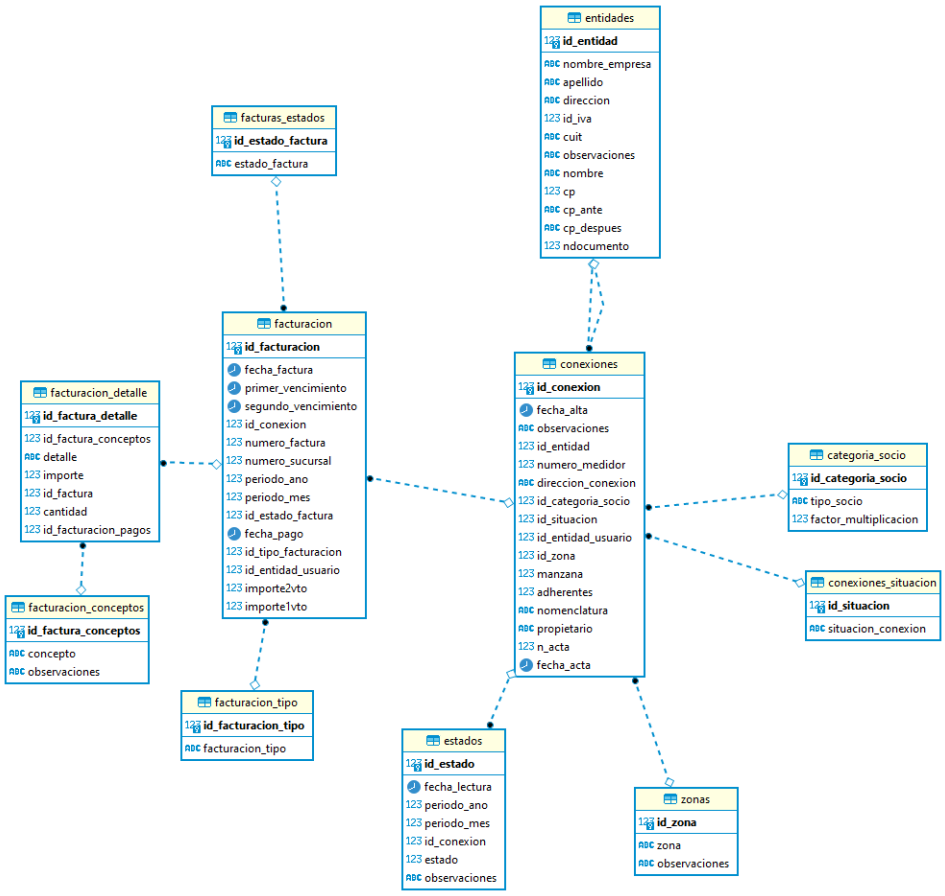

Figura 3. Modelo relacional de base de datos

3.2.2.1 Tabla facturacion

Esta tabla almacena los datos relacionados a la facturación por periodo de cada conexión. Contiene una cantidad de 75.898 registros. En la tabla 1 se describen cada uno de sus atributos.

Tabla 1. Atributos de tabla facturacion

#	Nombre	Tipo de Dato	Nulo	Referencias	Descripción
1	id_facturacion	bigint	False		Identifica de manera única a cada objeto de esta tabla
2	fecha_factura	date	False		Fecha en la que se emitió la factura
3	primer_vencimiento	date	False		Fecha del primer vencimiento de la factura
4	segundo_vencimiento	date	False		Fecha del segundo vencimiento de la factura
5	id_conexion	bigint	False	conexiones	Identificador de la conexión a quién corresponde la factura
6	numero_factura	bigint	False		Número de factura. Cada factura debe contar con un número único de identificación
7	numero_sucursal	bigint	False		Número de la sucursal que generó la factura
8	periodo_ano	integer	False		Representa al año de facturación
9	periodo_mes	integer	False		Representa al periodo de facturación. Este periodo toma en cuenta el mes en número
10	id_estado_factura	smallint	False	facturas_estados	Identificador del estado en el que se encuentra la factura.
11	fecha_pago	date	True		Indica si la factura ya ha sido pagada. Su campo nulo indica que aún no se ha realizado dicha acción

12	id_tipo_facturacion	smallint	False	facturacion_tipo	Identificador del tipo de facturación que se emitió
13	id_entidad_usuario	bigint	False		Identificador del usuario de la conexión. Hay casos en el que el titular de la conexión no es el usuario de la misma. Esto puede darse en caso de alquileres del predio
14	importe1vto	numeric(18, 2)	False		Monto expresado en pesos que el usuario debe abonar luego del primer vencimiento
15	Importe2vto	numeric(18, 2)	False		Monto expresado en pesos que el usuario debe abonar luego del segundo vencimiento

3.2.2.2 Tabla conexiones

Esta tabla almacena todas las conexiones dadas de altas en el sistema. Se pueden encontrar tanto conexiones activas como inactivas, cuenta con un total de 1.399 registros. En la tabla 2 se describen cada uno de sus atributos.

Tabla 2. Atributos de tabla conexiones

#	Nombre	Tipo de Dato	Nulo	Referencias	Descripción
1	id_conexion	bigint	False		Identifica de manera única a cada objeto de esta tabla
2	fecha_alta	date	False		Fecha en la que se dio de alta la nueva conexión
3	observaciones	character(150)	True		Información extra, por ejemplo faltante

					de algún papel a presentar
4	id_entidad	bigint	False	entidades	Identificador del titular de la conexión
5	numero_medidor	bigint	False		Representa el número que se encuentra en el medidor físico
6	direccion_conexion	character(100)	False		Dirección donde se encuentra el predio de la conexión, es importante mencionar que esta dirección no está normalizada
7	id_categoria_socio	smallint	False	categoria_socio	Identificador de la categoría a la que va a estar sujeta la conexión. Dependiendo la categoría, varía el monto de facturación
8	id_situacion	smallint	False	conexiones_situacion	Identificador del estado de la conexión
9	id_entidad_usuario	bigint	False	entidades	Identificador del usuario de la conexión. Existen casos donde el titular de la conexión no es el mismo que el usuario
10	id_zona	smallint	True	zonas	Identificador de la zona donde se encuentra el predio de la conexión
11	manzana	integer	True		Número de manzana donde se encuentra la conexión

12	adherentes	smallint	True	Representa la cantidad de personas de la familia adherentes a un servicio extra brindado por la CASAJ
13	nomenclatura	character(25)	True	Sin definir
14	propietario	character(1)	True	Sin definir
15	n_acta	bigint	True	Representa el número del acta física que da por aprobado el alta de la conexión.
16	fecha_acta	date	True	Representa la fecha de aprobación del acta

3.2.2.3 Tabla entidades

Representa a todas las personas físicas o jurídicas que pueden tener una conexión al servicio de agua o ser proveedores de productos/servicios de la CASAJ. Contiene una cantidad de 1.524 registros. En la tabla 3 se describen sus atributos.

Tabla 3. Atributos de tabla entidades

#	Nombre	Tipo de Dato	Nulo	Referencias	Descripción
1	id_entidad	smallint	False		Identifica de manera única a cada objeto de esta tabla
2	nombre_empresa	character varying(50)	False		Nombre que representa a la empresa en el caso de ser una persona jurídica
3	apellido	character varying(30)	False		Apellido de la persona física

4	direccion	character varying(100)	False		Representa a la dirección real de la persona física o jurídica
5	id_iva	smallint	False	iva	Identificador de la categoría frente al Impuesto al Valor Agregado (IVA) que tiene la persona
6	cuit	character varying(14)	False		Número de CUIT de la persona jurídica o CUIL de la persona física.
7	observaciones	character varying(100)	True		Representa información extra relacionada a la persona
8	nombre	character(50)	True		Nombre completo de la persona física
9	cp	integer	False	codigo_postal	Es el identificador del código postal de la localidad donde vive la persona
10	cp_ante	character(3)	True		Sin describir
11	cp_despues	character(3)	True		Sin describir
12	ndocumento	bigint	False		Documento Nacional de Identidad (DNI) de la persona física

3.2.2.4 Tabla estados

Esta tabla contiene toda la información relacionada a las lecturas de estados que fueron realizadas en cada periodo desde el momento que una conexión se encuentra activa. Cada estado corresponde a lo que indica el medidor en un periodo dado. Para calcular cuánto es el consumo del usuario, se resta el estado del

periodo$_x$ menos el periodo$_{x-1}$, y se obtiene el consumo en m^3. En la tabla 4 se describen sus atributos.

Tabla 4. Atributos de tabla estados

#	Nombre	Tipo de Dato	Nulo	Referencia	Descripción
1	id_estado	bigint	False		Identifica de manera única a cada objeto de esta tabla
2	fecha_lectura	date	False		Representa la fecha en la que fue realizada la lectura del medidor
3	periodo_ano	integer	False		Representa el año en el que fue realizada la lectura. Se almacena en formato numero
4	periodo_mes	integer	False		Representa el mes en que fue realizada la lectura. Se almacena en formato numero
5	id_conexion	bigint	False	conexiones	Identificador de la conexión a quien corresponde la lectura de estado efectuada
6	estado	bigint	False		Representa lo que indica el medidor a la fecha que se realiza la lectura
7	observaciones	character(150)	True		Sirve para describir alguna información extra relacionada al proceso de lectura

3.2.2.5 Tabla categoria_socio

Esta tabla almacena información relacionada al tipo de categoría al cual va a estar sujeta una conexión. Este tipo de categoría influye en el cálculo de los montos de facturación. Cuenta con un total de 5 registros. La tabla 5 describe sus atributos.

Tabla 5. Atributos de tabla categoria_socio

#	Nombre	Tipo de Dato	Nulo	Referencia	Descripción
1	id_categoria_socio	integer	False		Identifica de manera única a cada objeto de esta tabla
2	tipo_socio	character(100)	False		Descripción del tipo de categoría de una determinada conexión
3	factor_multiplicacion	real	False		Valor de multiplicación utilizado en el cálculo de la facturación

3.2.2.6 Tabla conexiones_situacion

Almacena la información relacionada al estado en la que se encuentra una conexión. Una conexión puede encontrarse como activa o suspendida. Cuenta con un total de 4 registros. La tabla 6 describe sus atributos.

Tabla 6. Atributos de tabla conexiones_situacion

#	Nombre	Tipo de Dato	Nulo	Referencia	Descripción
1	id_situacion	integer	False		Identifica de manera única a cada objeto de esta tabla
2	situacion_conexion	character(50)	True		Define la situación en la que se puede encontrar una conexión

3.2.2.7 Tabla zonas

Aquí se almacena información referente a la zona donde se encuentra la conexión. Una zona puede ser un barrio, conjunto de los mismos, o determinada ubicación particular. Cuenta con un total de 20 registros. La tabla 7 describe sus atributos.

Tabla 7. Atributos de tabla zonas

#	Nombre	Tipo de Dato	Nulo	Referencia	Descripción
1	id_zona	integer	False		Identifica de manera única a cada objeto de esta tabla
2	zona	character(150)	False		Nombre de la zona
3	observaciones	character(150)	True		Representa algún aspecto extra con respecto a la zona

3.2.2.8 Tabla facturacion_tipo

Esta tabla almacena la información de los tipos de facturación. En la CASAJ pueden ser mensuales o especiales (dadas por algún caso en especial). Cuenta con un total de 2 registros. La tabla 8 describe sus atributos.

Tabla 8. Atributos de tabla facturacion_tipo

#	Nombre	Tipo de Dato	Nulo	Referencia	Descripción
1	id_facturacion_tipo	bigint	False		Identifica de manera única a cada objeto de esta tabla
2	facturacion_tipo	character(150)	False		Describe el tipo de facturación

3.2.2.9 Tabla facturacion_conceptos

Almacena todos los conceptos que son incluidos en la factura de un usuario. Contiene un total de 15 registros. La tabla 9 describe sus atributos.

Tabla 9. Atributos de tabla facturacion_conceptos

#	Nombre	Tipo de Dato	Nulo	Referencia	Descripción
1	id_factura_conceptos	bigint	False		Identifica de manera única a cada objeto de esta tabla
2	concepto	character(100)	False		Describe el concepto a incluir en la factura
3	observaciones	character(100)	True		Representa alguna información extra

3.2.2.10 Tabla facturacion_detalle

Es la tabla intermedia entre la facturación y los conceptos. Almacena toda la información referida a qué conceptos se deben incluir en una determinada factura y su correspondiente importe. Tiene un total de 596.429 registros. La tabla 10 describe sus atributos.

Tabla 10. Atributos de tabla facturacion_detalle

#	Nombre	Tipo de Dato	Nulo	Referencia	Descripción
1	id_factura_detalle	bigint	False		Identifica de manera única a cada objeto de esta tabla
2	id_factura_conceptos	smallint	False	facturacion_conceptos	Identificador del concepto que se ha facturado
3	detalle	character(50)	False		Detalle del concepto facturado. Por lo general se utiliza el mismo que la descripción del concepto
4	importe	numeric(18, 3)	False		Importe del concepto que se ha facturado

5	id_factura	bigint	False	facturacion	Identificador de la facturación al que corresponde
6	cantidad	numeric(18, 3)	False		Cantidad del concepto que se ha facturado
7	id_facturacion_pagos	bigint	True	facturacion_pagos	Identificador de la tabla de pagos de facturas

3.2.2.11 Tabla facturas_estados

Representa al estado en el que se encuentra la factura del usuario. Tiene un total de 4 registros. La tabla 11 describe cada uno de sus atributos.

Tabla 11. Atributos de tabla facturas_estados

#	Nombre	Tipo de Dato	Nulo	Referencia	Descripción
1	id_estado_factura	smallint	False		Identifica de manera única a cada objeto de esta tabla
2	estado_factura	character(50)	True		Describe el estado en el que se puede encontrar una factura

3.2.3 Exploración de los datos

Esta sección se enfoca en exponer la consistencia y completitud de los datos. Se utilizó la herramienta Metabase [56] en su versión de libre acceso, para relevar las propiedades de los datos a través de distintos métodos como tablas de frecuencia y gráficos de distribución. Las consultas SQL generadas por cada gráfica se encuentran en la sección Anexo – Consultas SQL, cada una con su referencia a la figura correspondiente.

La Figura 4 muestra la distribución de los datos de la tabla categoria_socio, como se puede ver cuenta con 5 registros que no se repiten

en ningún caso, además, no contiene ningún registro nulo. En este caso la categoría relevante para este estudio es la de socio.

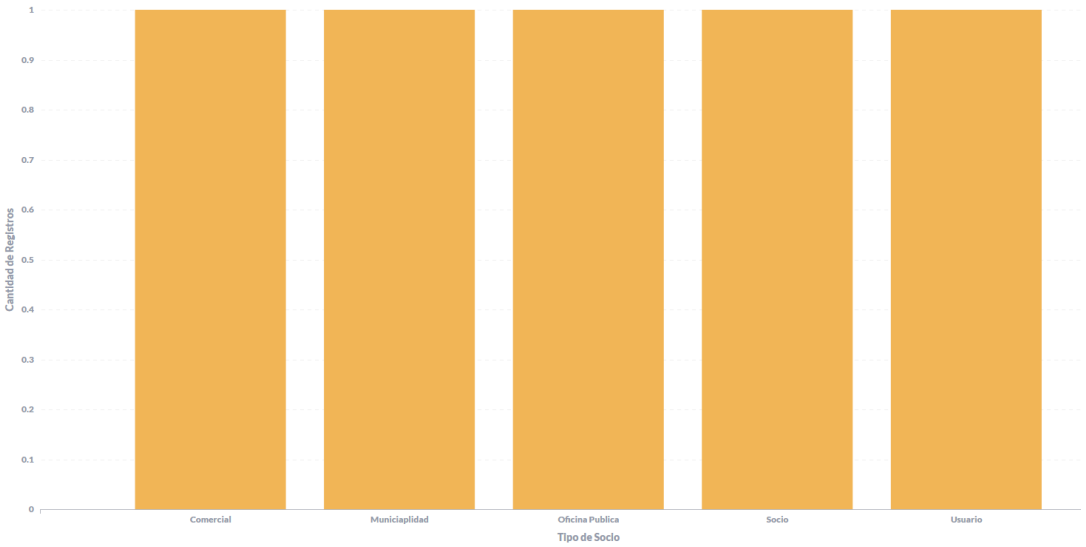

Figura 4. Distribución de datos en la tabla categoria_socio

La tabla conexiones_situacion cuenta con 3 registros repetidos con denominación Suspendido y solamente una aparición con denominación Activo (Figura 5). Para futuras fases esto no será un inconveniente debido a que solo se tomaran en cuenta aquellas conexiones activas.

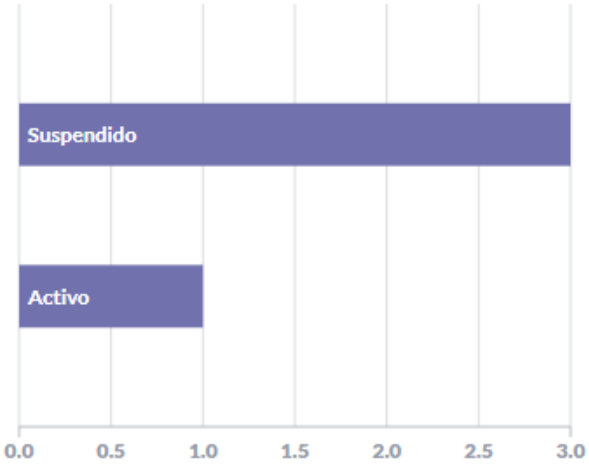

Figura 5. Distribución de datos en la tabla conexiones_situacion

La Figura 6 muestra el volumen de estados agrupados por el atributo periodo_ano de la tabla estados. Claramente se puede notar que a medida que avanza el tiempo, crece la cantidad de estados dado por las nuevas conexiones, excepto el año 2020 porque solo se tienen los datos de enero hasta el mes de abril de dicho año. El eje "x" muestra la cantidad de estados registrados y el eje "y" muestra los periodos en años.

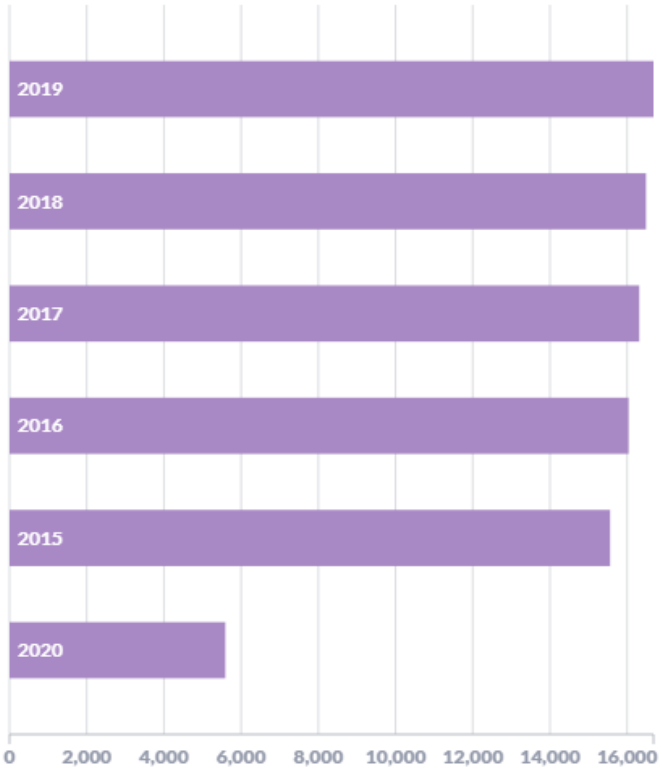

Figura 6. Cantidad de estados por año

La Figura 7 se centra en la tabla estados y tiene como fin visualizar la cantidad de estados que son 0 agrupado por mes y por año. Se puede notar que, por lo general, la cantidad de estados 0 entre mes y mes es menos de 200, excepto en el mes 4 del año 2020 que la mayoría son 0. Es probable que este periodo no aporte buenos resultados en el proceso de DM.

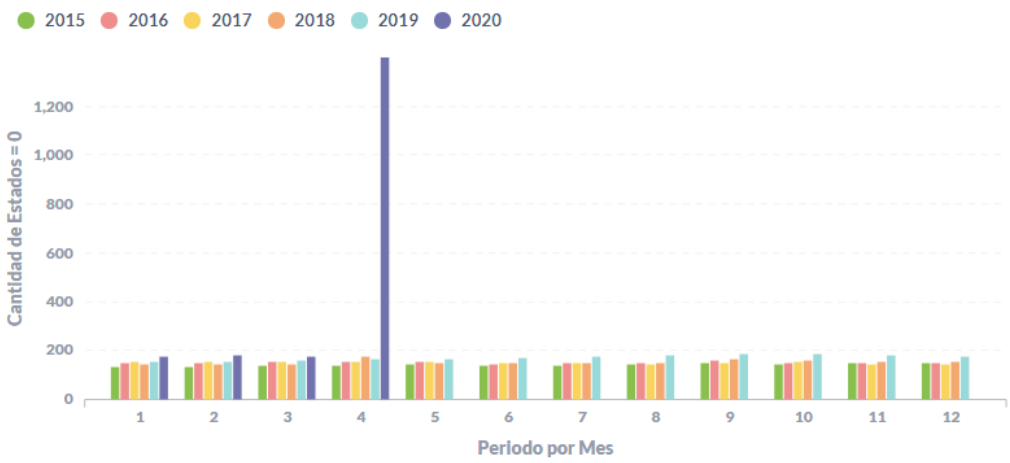

Figura 7. Cantidad de estados 0 agrupado por año y mes

La Figura 8 expresa la cantidad de facturas en porcentaje agrupadas en rango de consumo. Cada punto simboliza hasta que valor es incluido dentro del rango. Este gráfico de torta sirve para entender entre que rango de consumo, la CASAJ ha facturado más y tener una pre visualización del grupo de consumo con los que se cuenta. La etiqueta “otro” indica que el consumo se encuentra por encima de los 36 m^3 o incluso representa valores negativos. Existe un total de 596.429 facturas generadas.

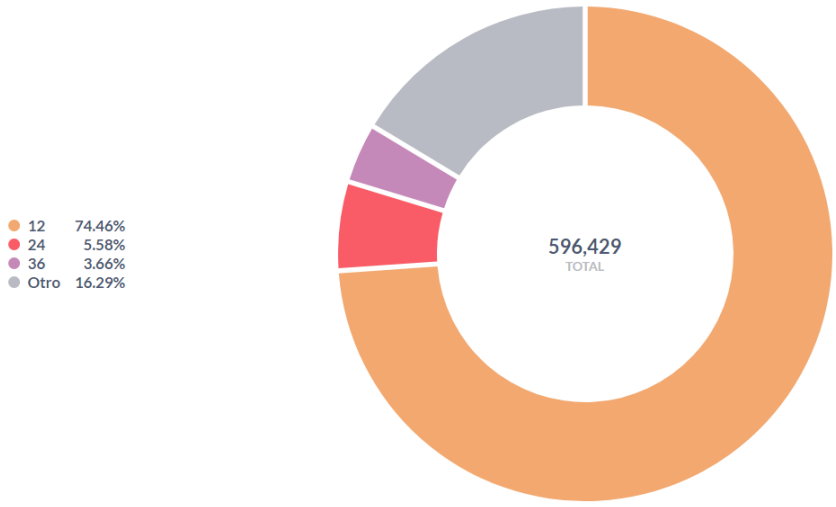

Figura 8. Porcentaje de facturas agrupada por rango de consumo

En contraste, la Figura 9 permite visualizar la cantidad de facturas que se encuentran en otros rangos fuera de lo normal, es decir, expone aquellas facturas que en la Figura 8 se clasificaban como “otro”. Esto se logró al cruzar la información de las tablas facturacion, facturacion_detalle y facturacion_conceptos. Es importante tener en cuenta que la tabla estados almacena información sobre la lectura realizada sobre los medidores, es por este motivo que las tablas de facturación previamente mencionadas juegan un papel importante dado que pueden brindar información acerca del consumo en m^3 por cada periodo. En la Figura 10 se muestran los conceptos que se facturan mensualmente a los usuarios, y el registro relevante para el estudio es el de *id* = 1, denominado Servicio Consumo.

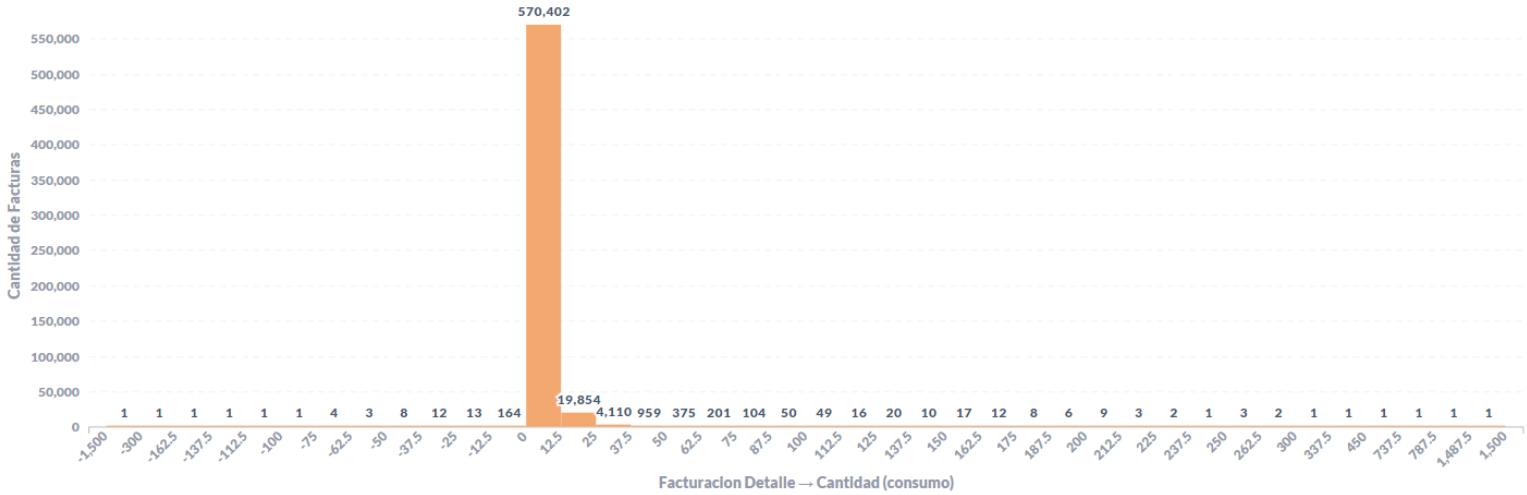

Figura 9. Cantidad de facturas por rango de consumo

coop_agua • Facturacion Conceptos

ID Factura Conceptos	Concepto	Observaciones
3	Iva	
4	Tasa Municipal/E.Subt.	
5	Gas de 10 kg	
6	Uso de salon de servicios multiples	
1	Servicio Consumo	.
2	Cuotas Sociales	.
7	Aporte Futuras obras	
8	Percepción	
9	Intereses por mora	
10	Reimpresion de Facturas	
11	Iva por rec.pago fuera termino	
14	Conexion Nueva	
12	AYUDA SEPELIO	
15	Regimen 14 RG DGR 29/15	impuesto provincial
13	Renovación Automática Contrat...	

Figura 10. Datos de tabla facturacion_conceptos

Los estados en los que se puede encontrar una factura se muestran en la Figura 11, y además, se expone la consistencia de la tabla facturas_estado. Para llevar a cabo este estudio fueron tomadas únicamente aquellas facturas con un estado pagado.

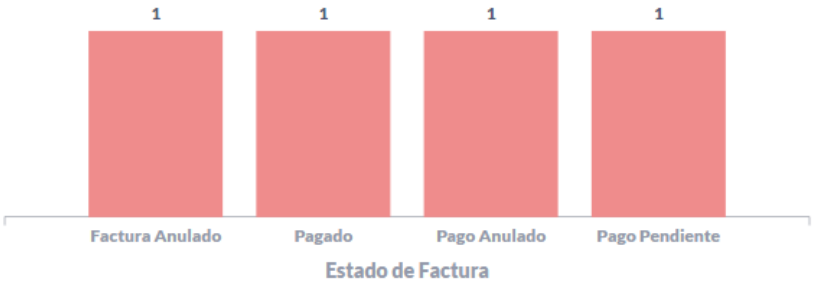

Figura 11. Datos de tabla facturas_estados

3.2.4 Verificación de la calidad de los datos

Luego de realizar el proceso de exploración de datos se puede afirmar que se poseen los datos necesarios para el proceso de DM, dado que la tabla relevante "facturacion" cuenta con datos completos. Además, la tabla "estados" cuenta con registros dentro del periodo que se pretende realizar el análisis.

Sin embargo, hay tablas que no son de tanta relevancia para el proyecto y también, como se mostró en la Figura 9, existen muchas facturas que contienen consumos fuera del rango normal (0 – 37 m^3), por lo tanto, los datos necesitan ser tratados y mejorados en la etapa de limpieza de los datos.

3.3 Preparación de los datos

En esta fase de la metodología se adecuaron los datos para facilitar su uso en la siguiente etapa. Esto implica seleccionar el subconjunto de datos a utilizar, limpiarlos para mejorar su calidad y finalmente, integrarlos y formatearlos según lo requerido.

3.3.1 Selección de los datos

A continuación, se listan aquellas tablas y atributos relevantes para llevar a cabo el análisis:

- Tabla estados:
  - periodo_ano
  - periodo_mes
  - id_conexion
  - estado
- Tabla conexiones:
  - id_conexion
  - fecha_alta

    - id_categoria_socio
    - id_situacion
    - id_entidad_usuario
- Tabla categoria_socio:
    - id_categoria_socio
    - tipo_socio
- Tabla facturacion
    - id_facturacion
    - id_conexion
- Tabla facturacion_detalle
    - id_factura_detalle
    - id_factura
    - id_factura_conceptos

Cada uno de los datos seleccionados permiten lograr los objetivos del proyecto. La tabla “estados” es importante para tener información relacionada a los estados registrados en cada periodo y a que conexión corresponden, se descartartó el atributo fecha_lectura debido a que no presenta consistencia y, además, contiene fechas inválidas. De la tabla “conexiones” es importante conocer su número de identificación, cuándo se dio de alta la conexión para saber si puede tener un consumo histórico relevante y si la conexión está activa o no. Es necesario tener la información del consumo facturado en cada periodo, por este motivo son relevantes las tablas “facturacion” y “facturacion_detalle”. La tabla “categoria_socio” permite filtrar por aquellas categorías que sean “socio”, dado que ésta es la categoría de interés para poder llevar a cabo el proyecto, las demás fueron descartadas del conjunto de datos.

Las tablas no mencionadas fueron descartadas del proceso debido a que no aportan información necesaria. Además, se aplicó la misma metodología para la selección de los atributos de cada tabla.

3.3.2 Limpieza de los datos

La base de datos podría contener información con valores nulos, ruido o datos que no son relevantes, es por esto que se lleva a cabo la limpieza de los mismos.

De la tabla “conexiones” se procedió a descartar todas aquellas que se encuentran inactivas o dadas de baja debido a que no se contaba con una fecha de baja para determinar el periodo de actividad y, además es irrelevante trabajar sobre conexiones que ya no siguen produciendo información. En la sección 3.2.3 se observa que la tabla “conexiones_situacion” (Figura 5) presentaba elementos repetidos con denominación Suspendido. Es por esto, que se descartaran todas las conexiones con el atributo id_situacion igual a 2, 3 y 4 (ver Anexo – Filtro Conexiones Activas). Luego de aplicar el filtro, queda un total de 1.234 conexiones activas.

También, se procedió a eliminar todas aquellas conexiones que, a lo largo del periodo en estudio o gran parte de él, cuentan con estados igual a 0, debido a que no son útiles para la etapa de DM. Por cada mes se genera un nuevo estado entonces, si se tiene en cuenta que el periodo para llevar a cabo el análisis va desde enero del 2015 a abril del 2020, se tiene un total de 64 estados por cada conexión. Por ende, es razonable pensar que al menos cada conexión debe contar con menos del 70% de los estados iguales a 0. Dado que, si se supera este porcentaje, se considera una normalidad de tendencia a 0. El 70% de 64 estados como totalidad es igual a 44,8, entonces se eliminaron todas las conexiones que tenían más de 45 estados igual a 0 (ver Anexo – Filtro Conexiones Estado 0).

Como se observó en la Figura 7 de la sección 3.2.3 el periodo del mes de abril del año 2020, cuenta con todos los estados igual a 0. Por consecuente, se eliminó todo ese periodo (ver Anexo – Filtro Conexiones Estado Periodo 04/2020).

Analizando la tabla “estados” no se encontró ningún registro nulo o que se encuentre incompleto que imposibilite el proceso de DM.

3.3.3 Construcción, integración y formateo de los datos

En esta sección se detalla el proceso para la generación del conjunto de datos limpio, íntegro y con el formato adecuado para facilitar el proceso de minería. Se utilizó la herramienta RapidMiner en su versión educativa [57]; esta herramienta abarca todas las fases del proceso de *data mining*, en este caso se emplea como soporte para la fase de preparación de datos.

3.3.3.1 Proceso

Comenzando por la lectura de datos, la Figura 12 muestra cómo se leen los mismos desde la base de datos PostgreSQL a través de su respectiva configuración (Figura 13), el elemento *Multiply* permite múltiples conexiones en simultáneo y luego se utilizan los elementos *Read Database* especificando que tabla se pretende leer según lo descripto en la sección 3.3.1

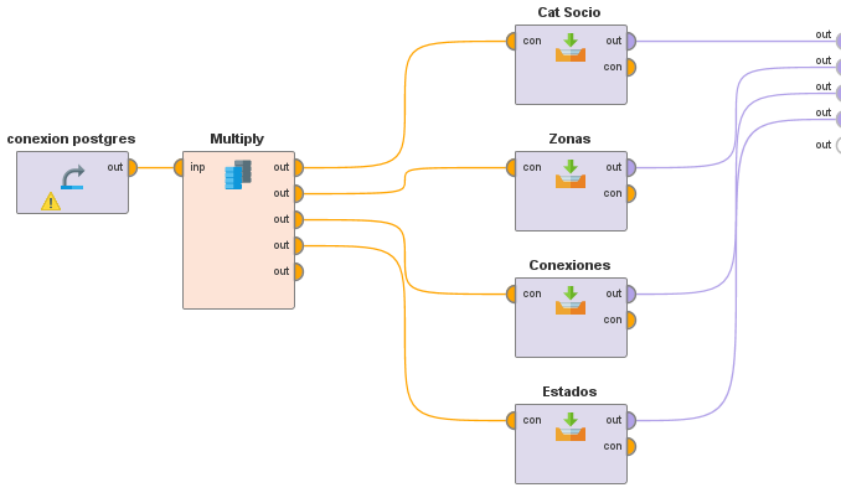

Figura 12. Extracción de datos

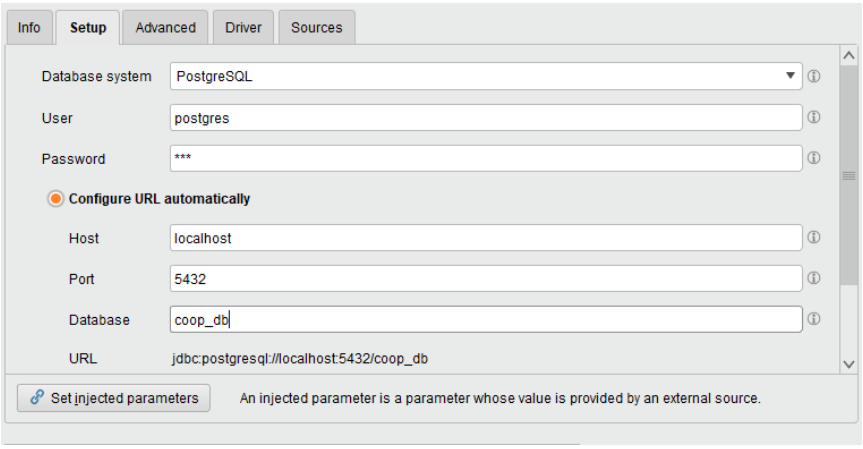

Figura 13. Configuración de conexión a base de datos

Lo mencionado anteriormente se encuentra dentro de un elemento *Subprocess* denominado “Extraccion” para trabajar de una manera más ordenada (Figura 14).

Cada una de las salidas generadas, fueron leídas por el elemento *Select Attributes* el cual permite que se seleccione

únicamente aquellas columnas que son relevantes para el proceso y que también se describieron en la sección 3.3.1

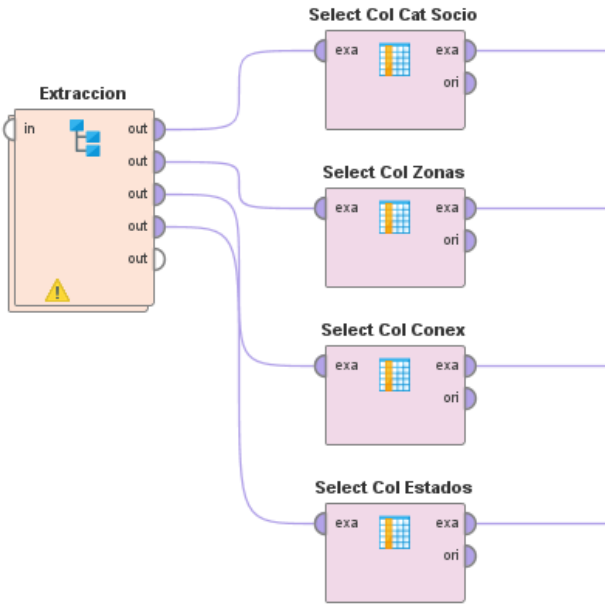

Figura 14. Selección de atributos

Luego, se aplicaron los filtros mencionados en la sección 3.3.2 y se procedió a unir los resultados para obtener únicamente aquellos estados de las conexiones activas (Figura 15):

- Filtro de conexiones activas.
- Filtro de periodo distinto a 04/2020.

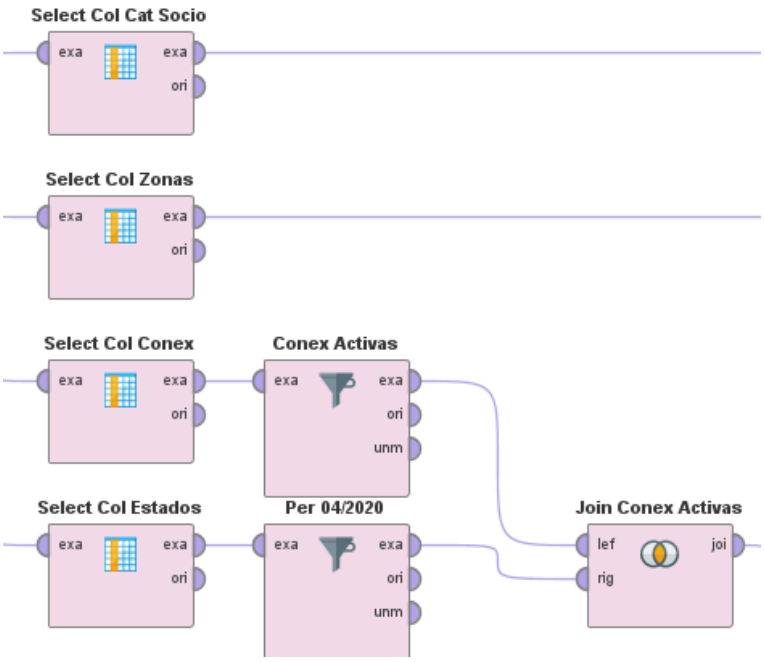

Figura 15. Filtros de tabla conexiones y estados

Posteriormente, sobre la tabla “estados” se aplicaron los filtros de estados iguales a 0 que se explicaron en mayor detalle en la sección 3.3.2 De los resultados obtenidos, se agruparon por el atributo id_conexion y se contaron la cantidad de estados iguales a 0, para luego filtrarlos por aquellas conexiones que tengan un resultado menor a 45, dado que ese es el número relevante para el objetivo de este trabajo. Una vez aplicados los filtros se volvieron a unir con la tabla original por el atributo id_conexion para obtener los datos completos. Por otro lado, del primer filtro de estados iguales a 0, se tomaron aquellos que no aplican para unirlos nuevamente con los datos originales y no perderlos. Finalmente, se realizó la unión entre los datos que satisfacen los requerimientos y se filtran aquellos repetidos, para seleccionar únicamente los atributos necesarios descriptos en la sección 3.3.1 (Figura 16).

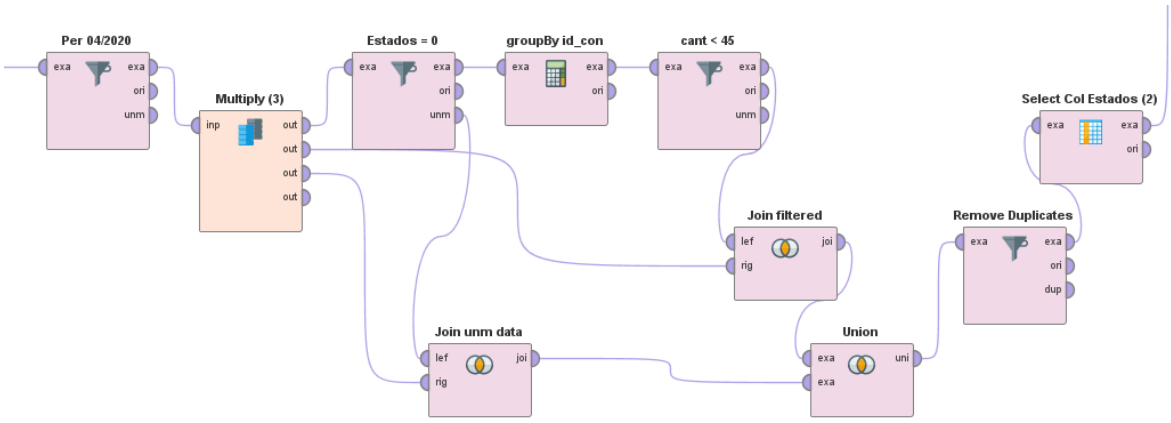

Figura 16. Filtros tabla estados

Para poder llevar a cabo el proceso de DM es necesario contar con un valor que es el consumo en m^3 por cada periodo, es por esto que se generó un nuevo atributo en la tabla "estados" denominado "consumo". El proceso que se llevó a cabo se muestra en la Figura 17. Se comenzó por la lectura de las tablas "facturacion" y "facturacion_detalle" y luego se filtró la última tabla mencionada por el atributo "id_factura_conceptos" igual a 1, que hace referencia al concepto de consumo. Seguidamente, se unieron ambas tablas y se seleccionaron los siguientes atributos:

- id_conexion
- periodo_ano
- periodo_mes
- cantidad

Los tres primeros atributos permiten la unión entre los resultados que se obtuvieron en la Figura 16 y el atributo cantidad contiene el consumo en m^3, por eso es renombrado a "consumo".

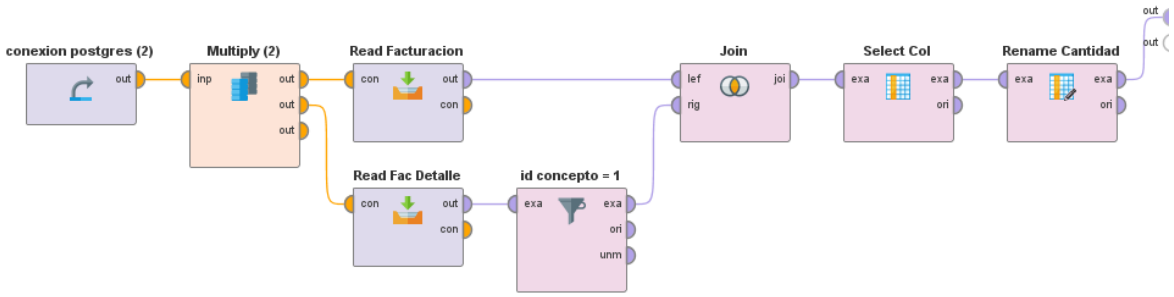

Figura 17. Proceso de obtención del atributo consumo

Teniendo en cuenta lo expuesto por la Figura 9 de la sección 3.2.3 se excluyeron todos aquellos consumos que sean menores a 0 dado que se consideran como datos erróneos. El proceso se muestra en la Figura 18, donde los datos llegan después de que el nuevo atributo “consumo” se haya generado y pasan a filtrarse aquellos consumos mayores a 0.

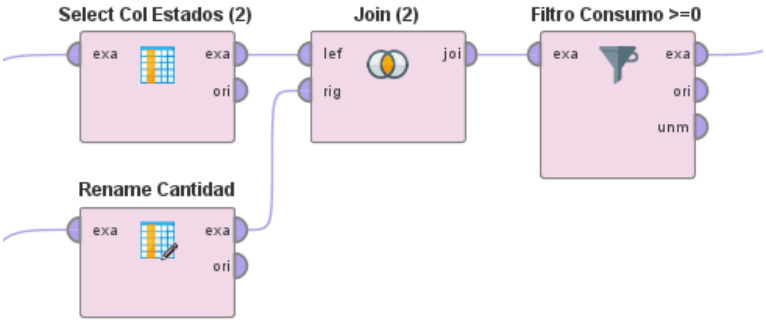

Figura 18. Proceso de filtrado de consumos > 0

Para obtener mejores resultados en la aplicación de las distintas técnicas, se realizó la transformación de los atributos “periodo_ano” y “periodo_mes” a un valor único denominado “fecha” el cual tiene el formato año-mes, y es utilizado como índice de los datos.

Finalmente, se llevó a cabo la generación centralizada de los datos limpios, para esto se utilizó el elemento "*write_csv*", el cual permite generar un archivo .csv con los datos anteriormente trabajados. En la Figura 19 se puede ver un resumen de los datos que fueron generados.

fecha	consumo
2015-01	3468
2015-02	3261
2015-03	2824
2015-04	2628

Figura 19. Formato final de datos

La siguiente tabla describe el significado de los datos resultantes.

Tabla 12. Descripción del formato final de datos

Dato	Detalle
fecha	Indica el período de consumo de agua en el formato año-mes.
consumo	Indica el consumo total de agua potable en el periodo dado. Su medida se expresa en m^3.

3.4 Modelado

En esta fase de la metodología se escogieron las técnicas más apropiadas para lograr los objetivos del trabajo de tesis. Luego, se generaron pruebas sobre los modelos producidos para evaluar si se cumple con lo requerido. La herramienta seleccionada para implementar los modelos fue Jupyter junto a Python y las librerías de libre acceso que permiten el análisis de datos, descriptas en la sección 2.7.2

3.4.1 Selección del modelo

El modelo seleccionado para llevar a cabo el proceso fue LSTM, dado que trabaja perfectamente con series de tiempo y además, permite retener información en la red ya sea a corto o largo plazo y, por ende, permite obtener mejores resultados para los datos resultantes de la sección anterior [58]. También, se creó un modelo de *backpropagation* para visualizar como los datos se ajustan sin la posibilidad de contar con información previa en cada iteración.

3.4.2 Plan de pruebas

Se generan distintos modelos LSTM con variaciones en sus parámetros. El procedimiento que se utiliza para probar la calidad y validez de los modelos generados son las medidas del error absoluto medio (*mean absolute error*), el error cuadrático medio (*mean squared error*), la raíz del error cuadrático medio (*root mean squared error*) y el valor de pérdida de error de cada modelo.

Para cada una de las métricas anteriormente mencionadas, su tendencia a 0 es lo que influye para determinar una buena calidad de modelo; en caso contrario, se dice que el modelo es deficiente [42].

Se dividen los datos de manera aleatoria, 75% para el entrenamiento de los modelos y 25% para validación, por ende, la técnica utilizada para este caso es *holdout* descripta en la sección 2.8.1

3.4.3 Construcción de los modelos

En esta sección se detallan las características de los modelos LSTM y como se ajustan los valores de cada parámetro para poder medir los resultados y ver como cada uno se comporta luego de su ejecución. También, se incluye los valores dados para el modelo de *backpropagation*.

Para cada uno de los casos se realizan 200 iteraciones para entrenar la red, se hace uso del optimizador Adam, el cual es muy eficiente para este tipo de casos [59], se usa la función de pérdida de error cuadrático medio y se utiliza la función de activación tangente.

Existen denominaciones que, de aquí en adelante, pueden ser llamadas diferentes, pero tienen el mismo significado. Para el caso del optimizador, su correspondencia es *optimizer*. La función de pérdida de error se puede encontrar como *loss*. Y, la función de activación, definida como *activation*.

Para los modelos LSTM es recomendado que los *lstm_units* se mantengan en un rango de 1 a 5, este valor indica a la red la cantidad de dimensiones internas con las que va a trabajar para lograr un mejor rendimiento dependiendo del conjunto de datos. El *batch_size* es el tamaño del conjunto de datos a tomar en cada iteración interna, un tamaño de entre 5 a 100 hacen al correcto funcionamiento de este tipo de arquitectura. El *input_shape* maneja los datos según el periodo de tiempo que se desea memorizar y es por este motivo que se generan modelos tomados de a 1 mes, 6 meses y 1 año respectivamente, es decir que al momento de ejecutar el modelo su tiempo de memorización de los datos dependerá de este parámetro *input_shape* [60]. Junto a cada modelo, se incluye su implementación en la herramienta seleccionada.

El código para crear los modelos LSTM es el siguiente (Figura 20). Está generalizado y se parametrizó el paso de tiempo con el que se van a tomar los datos.

```
def crear_modeloLSTM():
    model = Sequential()
    model.add(LSTM(5, input_shape=(1,PASOS), activation="tanh"))
    model.add(Dense(1))
    model.compile(loss='mean_squared_error',optimizer='Adam',metrics=["mse"])
    model.summary()
    return model
```

Figura 20. Modelo LSTM

Por otro lado, el código para generar el modelo utilizando *backpropagation* se puede ver en la Figura 21. Luego de ejecutar ambos casos, se podrán comparar y ver los resultados obtenidos para ambos tipos.

```
def crear_modeloBackpropagation():
    model = Sequential()
    model.add(Dense(1, input_shape=(1,1),activation='tanh'))
    model.add(Flatten())
    model.add(Dense(1, activation='tanh'))
    model.compile(loss='mean_absolute_error',optimizer='Adam',metrics=["mse"])
    model.summary()
    return model
```

Figura 21. Modelo backpropagation

3.4.3.1 Modelo 1

Tabla 13. Parámetros modelo 1 (LSTM)

Parámetro	Valor
tipo	LSTM
iteraciones	200
batch_size	5
optimizer	adam
loss	mean_squared_error
lstm_units	5
input_shape	(1, 1)
activation	tanh

3.4.3.2 Modelo 2

Tabla 14. Parámetros modelo 2 (LSTM)

Parámetro	Valor
tipo	LSTM
iteraciones	200
batch_size	5
optimizer	adam
loss	mean_squared_error
lstm_units	5
input_shape	(1,6)
activation	tanh

3.4.3.3 Modelo 3

Tabla 15. Parámetros modelo 3 (LSTM)

Parámetro	Valor
tipo	LSTM
iteraciones	200
batch_size	5
optimizer	adam
loss	mean_squared_error
lstm_units	5
input_shape	(1, 12)
activation	tanh

3.4.3.4 Modelo 4

Tabla 16. Parámetros modelo 4 (backpropagation)

Parámetro	Valor
tipo	Backpropagation
iteraciones	200
batch_size	5
optimizer	adam
loss	mean_squared_error
input_shape	(1, 1)
activation	tanh

3.5 Evaluación

Para evaluar la efectividad de los modelos se usaron los indicadores que se establecieron en el plan de pruebas, dichos indicadores fueron el error RMSE, MAE y MSE, también se tuvo en cuenta el valor de error generado por cada red. Se presentan las gráficas de predicción para poder visualizar cómo se comporta cada modelo. Los resultados de estas métricas se aprecian en las tablas 17, 18, 19 y 20. Como primer paso se muestran los resultados de cada métrica por cada modelo y luego se genera un cuadro comparativo entre todos los modelos.

De acuerdo con el plan de pruebas, el conjunto de entrenamiento está formado por el 75% de los datos y el conjunto de prueba por el 25% restante.

En la Tabla 17, se aprecian los valores de las métricas de rendimiento luego de la ejecución del modelo 1.

Tabla 17. Métricas modelo 1

Datos	Métrica	Resultado
Prueba	MSE	0.29
	RMSE	0.54
	MAE	0.40

Para este modelo se obtuvo una reducción del valor de pérdida de predicción, esto indica que el modelo fue aprendiendo en cada iteración. Sin embargo, el valor podría mejorar con una tendencia más cercana al 0. En la Figura 22 se puede observar cómo llega a un valor cercano a 0.10.

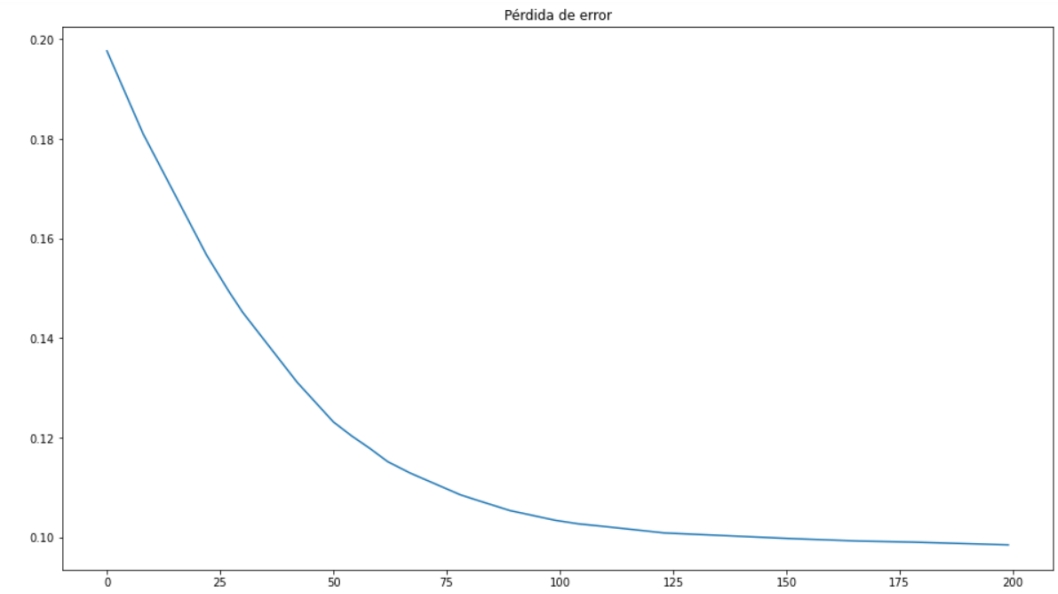

Figura 22. Pérdida de error modelo 1

Este modelo no tiene tanta exactitud en la predicción, pero, aun así, se puede ver en la Figura 23 como existe cierta tendencia de los resultados obtenidos en contraste a los valores originales.

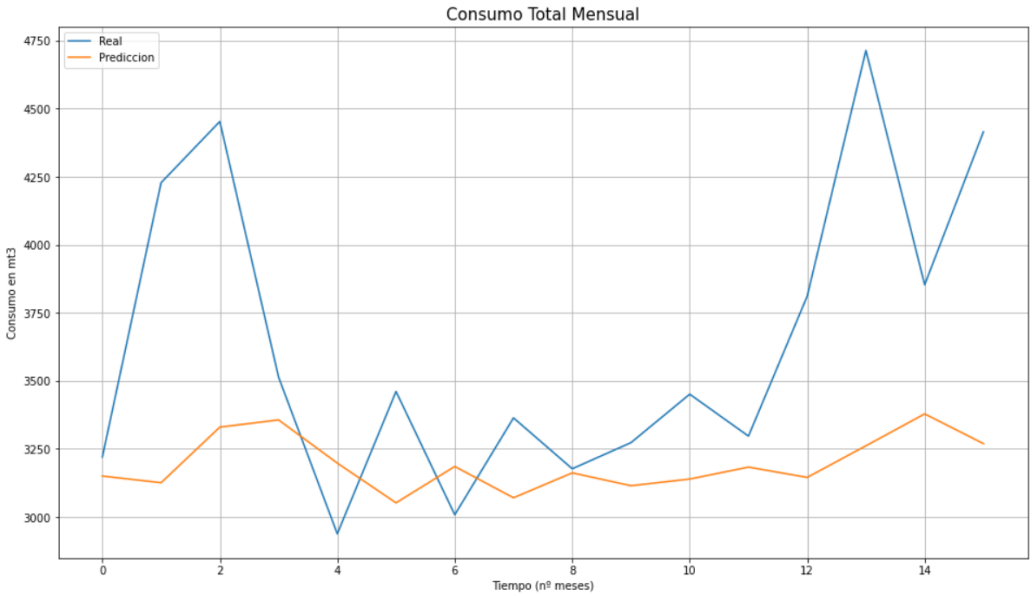

Figura 23. Predicción de consumos (modelo 1)

Siguiendo con el modelo 2, la Tabla 18 expone los resultados obtenidos luego de su ejecución.

Tabla 18. Métricas modelo 2

Datos	Métrica	Resultado
Prueba	MSE	0.23
	RMSE	0.48
	MAE	0.35

Para este caso, el valor de pérdida de error se comporta muy similar al modelo 1 (Figura 24), pero se puede notar que la red aprendió más rápido; dado que, para la iteración 50 el valor ya se encontraba por debajo de 0.11 y no fue así para el modelo previamente ejecutado.

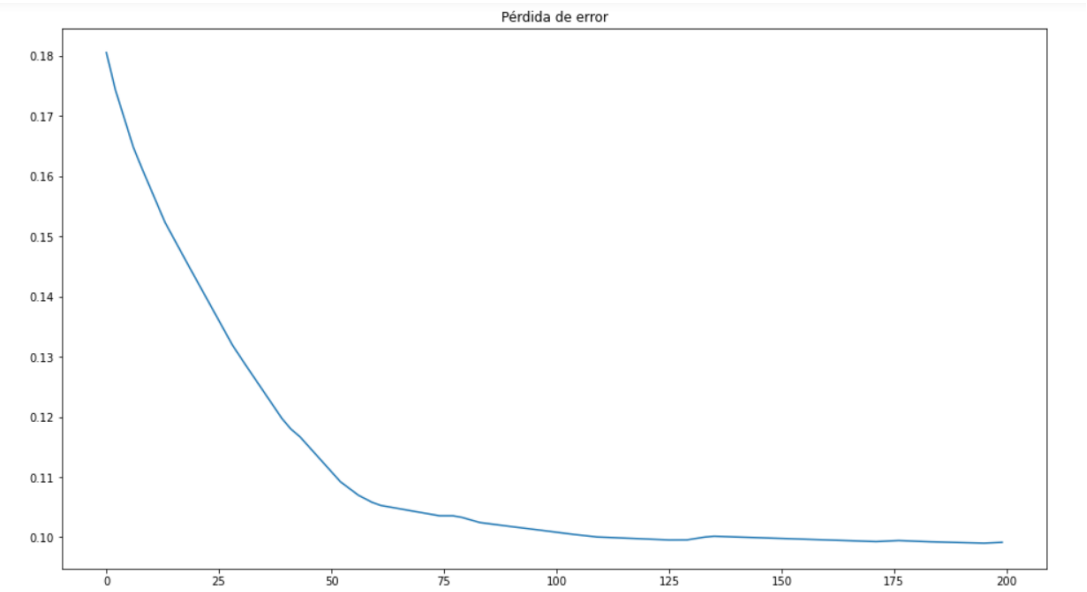

Figura 24. Pérdida de error modelo 2

El modelo 2 retuvo información de 6 meses en cada iteración, se puede observar en la Figura 25 que la gráfica no varía demasiado con respecto al modelo anterior, es decir, los resultados de predicción siguen sin ser tan efectivos.

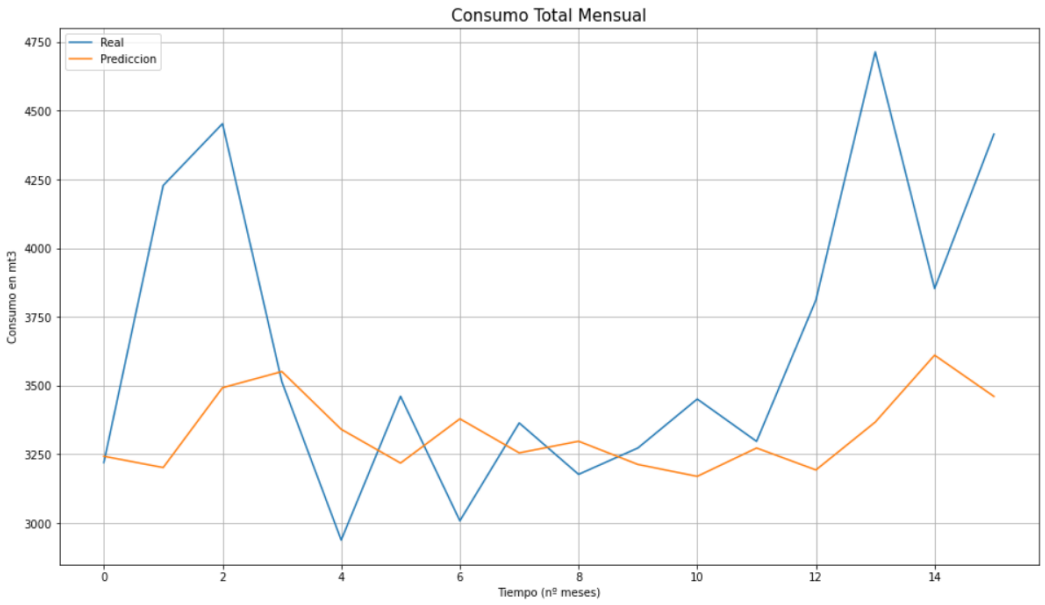

Figura 25. Predicción de consumos (modelo 2)

La siguiente tabla (Tabla 19), muestra las métricas de rendimiento para el modelo 3.

Tabla 19. Métricas modelo 3

Datos	Métrica	Resultado
Prueba	MSE	0.15
	RMSE	0.38
	MAE	0.28

En este caso, se puede observar gran diferencia en cuanto a la reducción del valor de pérdida de error (Figura 26). Este modelo se encuentra por debajo de 0.08, es decir, tiene gran capacidad de predicción, pudiendo así arrojar mejores resultados.

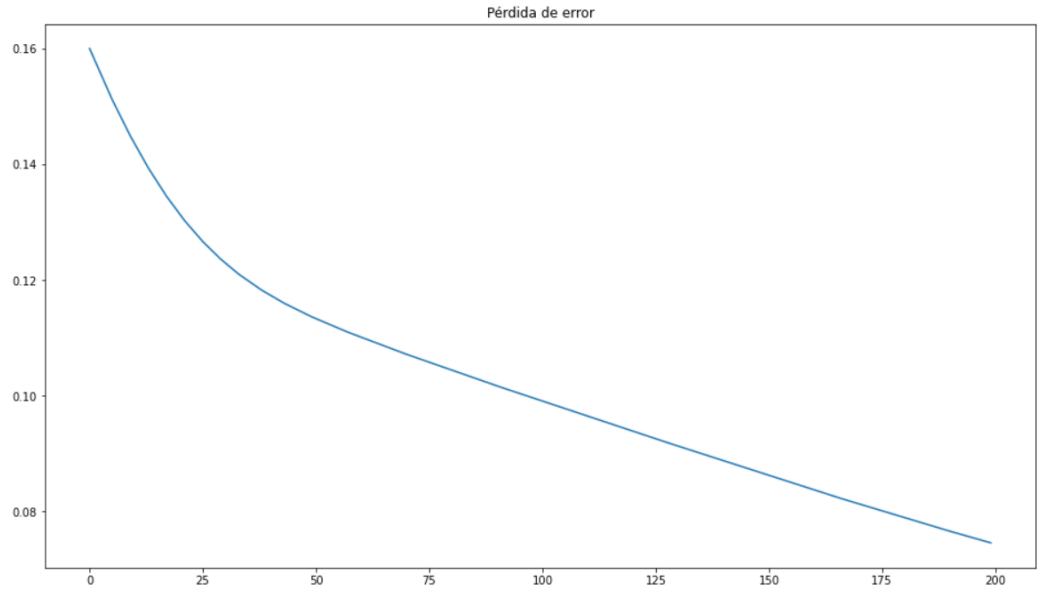

Figura 26. Pérdida de error modelo 3

La gráfica de predicción generada por este modelo (Figura 27) se aproxima bastante a los datos reales. Para este caso las métricas definidas en la sección 3.4.2 tienen una tendencia cercana a 0 y el nivel de error de la red está por debajo del 8%, esto permite que el modelo obtenga mejores resultados.

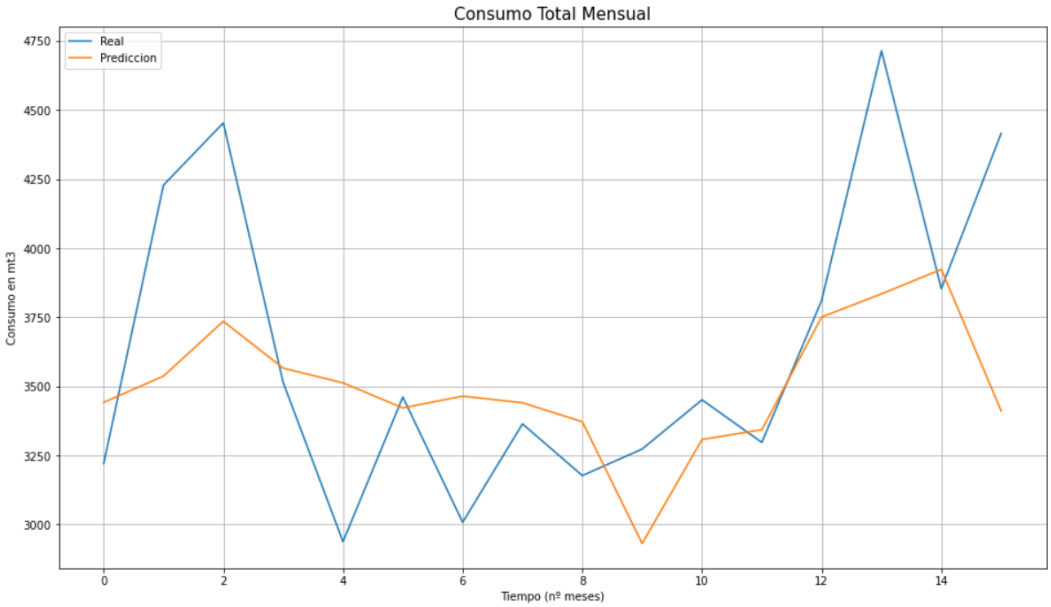

Figura 27. Predicción de consumos (modelo 3)

Finalmente, la Tabla 20 presenta las métricas obtenidas para el modelo de red neuronal tipo 4.

Tabla 20. Métricas modelo 4

Datos	Métrica	Resultado
Prueba	MSE	0.62
	RMSE	0.79
	MAE	0.63

Como se ve en la Figura 28, el valor de error se quedó estacionado en 0.24, es decir, de todos los modelos, es el que menor desempeño tiene.

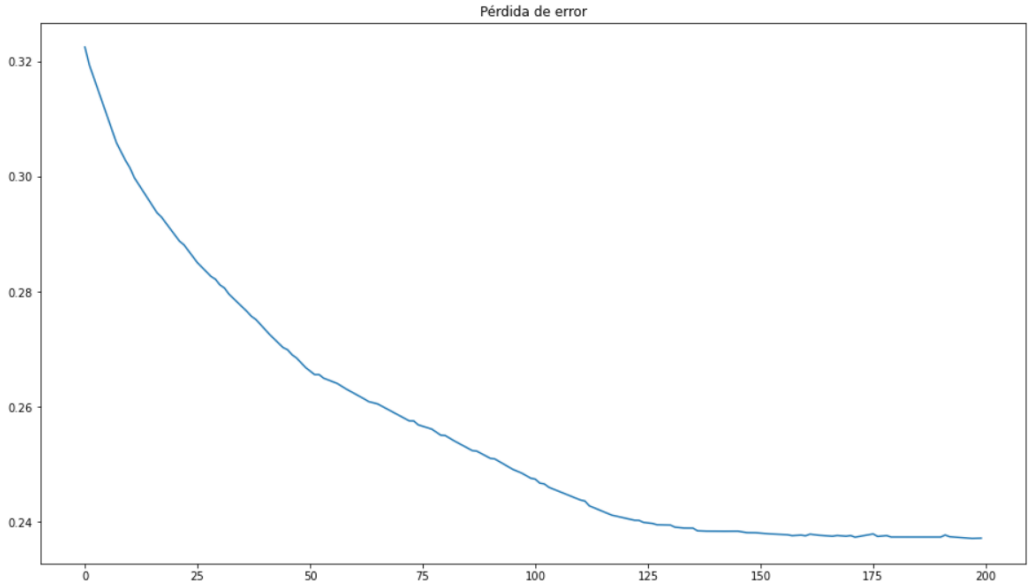

Figura 28. Pérdida de error modelo 4

Otro motivo por el cual este modelo no tiene el mejor rendimiento es debido a que las métricas de predicción tienen una tendencia cercana a 1, esto se ve reflejado en la Figura 29 como la gráfica de predicción se encuentra bastante alejada de los consumos reales.

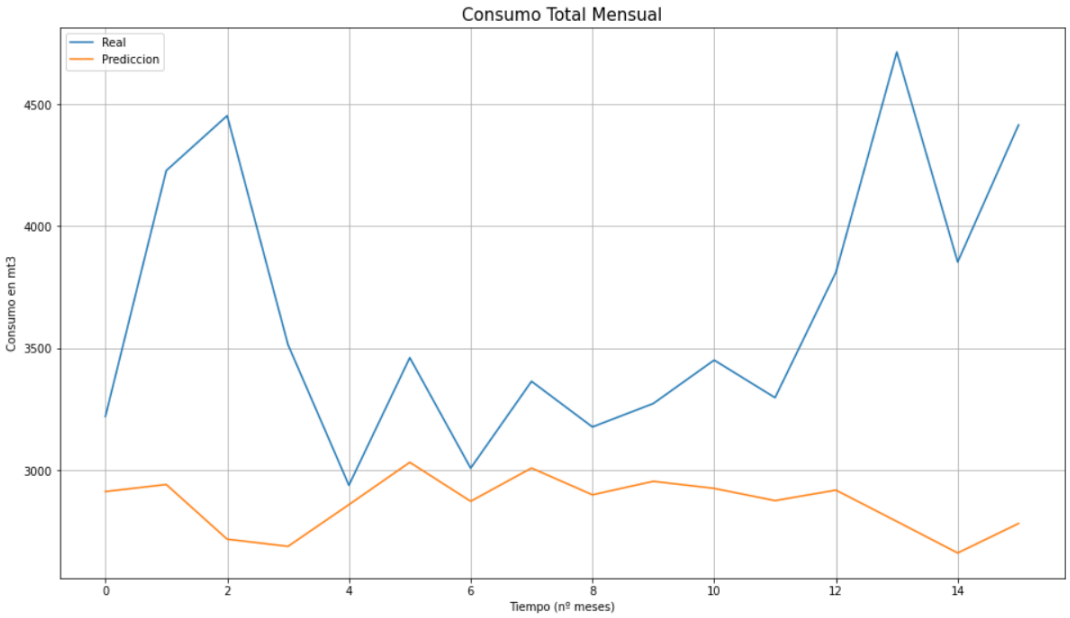

Figura 29. Predicción de consumos (modelo 4)

En la tabla 21 se compara cada métrica obtenida por cada modelo para los datos de prueba a fin de contrastar los diferentes resultados.

Tabla 21. Métricas por modelo

	MSE	RMSE	MAE	Error de la Red
Modelo 1	0.29	0.54	0.40	> 10% < 20 %
Modelo 2	0.23	0.48	0.35	> 10 % < 20%
Modelo 3	0.15	0.38	0.28	< 8%
Modelo 4	0.62	0.79	0.63	> 20%

Como se observa en la Tabla 21, los mejores resultados fueron obtenidos por el modelo 3, el cual mantiene en la red una memoria de 12 meses previos por cada iteración. En segundo lugar, se encuentra el modelo 2 el cual retiene en la red 6 meses previos; luego el modelo 1, donde sus métricas son las más altas y se alejan del valor 0, y finalmente el modelo 4 que se basa en una red simple de

backpropagation, es decir, esta última no retiene información previa y, por ende, se esperaba que los resultados de predicción fuesen menores a la de los modelos LSTM.

Para este conjunto de datos, el modelo 3 es el más eficiente y es el que permitirá a la CASAJ obtener mejores resultados para lograr predecir sus consumos totales de agua potable. Sin embargo, es importante mencionar que probablemente el modelo 3 pueda seguir mejorando su rendimiento, pero para el alcance de este trabajo, son satisfactorios los resultados obtenidos.

3.6 Despliegue

Esta es la última fase de la metodología y está orientada a exponer y lograr un fácil acceso a los resultados obtenidos. Permite que los usuarios que hagan uso de la información generada puedan tomar decisiones basándose en las mismas.

3.6.1 Planificación

Para lograr la implementación de este proyecto en el ambiente de trabajo real de la CASAJ será necesario acceder a la base de datos completa que contiene toda la información histórica desde los comienzos hasta la actualidad. Luego, los pasos a seguir serían los mismos descriptos en este trabajo, con la salvedad de que la etapa de comprensión y preparación de los datos quizás sea más costosa debido a la cantidad de datos que pueden presentarse. Por otro lado, será necesario adquirir una licencia empresarial de RapidMiner para no tener limitaciones con el producto y poder cubrir exitosamente las etapas previamente mencionadas. En cuanto a la etapa de modelado, no existiría ningún inconveniente para lograr los mismos resultados descriptos en la presente, dado que se trabaja con herramientas de libre acceso.

Para cubrir la necesidad de realizar el proceso de comprensión de los resultados y generación de los informes, se requiere de una persona con conocimientos básicos de Python y ejecución de código. De lo contrario, es necesario el desarrollo de un software que permita ejecutar el código de los distintos modelos y se comporte de una manera más amigable con el usuario, sin la necesidad de ver y entender código, que sea totalmente configurable y parametrizable, incluso que guarde y genere distintas clases de reportes.

Capítulo 4

Conclusiones y Futuras Líneas

CAPITULO 4: Conclusiones y futuras líneas

Esta sección expone el aporte de este trabajo y las conclusiones obtenidas del análisis de los resultados. Al final se muestran las futuras líneas de investigación que fueron identificadas a lo largo del desarrollo de la misma.

4.1 Conclusiones

El desarrollo de este trabajo aporta a la CASAJ la posibilidad de tomar decisiones basándose en lo que se observa en los datos, permite lograr una mejor planificación de los insumos que se requerirán en el futuro, lograr una visión de cuantos m^3 de agua van a facturar por mes y por ende estar preparados a nivel empresarial para poder potabilizar la cantidad demandada.

Por otro lado, a nivel sociocultural, es posible que pueda exponer y lograr la concientización de lo importante que es cuidar un recurso de primer nivel como es el agua, a través de la visualización de los picos de consumos dados en ciertos periodos de tiempo.

Teniendo en cuenta el proceso para poder llegar a los resultados, se considera que las etapas de entendimiento y preparación de los datos fueron encaradas exitosamente y exponen conocimiento relacionado a la estructura y significado de cada dato almacenado en la base de datos de la CASAJ, se destaca este punto debido a que la empresa no cuenta con esta información. La predicción de los consumos fue basada en modelos de redes neuronales y en un conjunto de datos reducido; es posible que, al contar con una mayor cantidad de datos, los resultados de predicción puedan mejorar e incluso sería interesante añadir otros atributos, como por ejemplo la estación del año en la que se encuentra o algún atributo generado, como la media mensual o el promedio anual.

El uso de distintas herramientas para cada etapa permitió exponer las ventajas y desventajas entre cada una. Metabase resultó ser una plataforma muy intuitiva y fácil de usar, no es así para RapidMiner, que, si bien facilitó mucho el proceso de preparación de datos gracias a que trabaja con elementos que son arrastrables y de fácil manipulación, es necesario leer la documentación de cada elemento para entender su correcto funcionamiento. Jupyter y Python trabajan perfectamente, pero demandan bastante tiempo debido a que se debe importar y codificar todo lo que se requiera y también, está muy ligado a la lectura de documentación y foros de la comunidad.

En cuanto a los resultados obtenidos, el modelo que más destacó fue LSTM por encima de *backpropagation*. Esto era de esperarse, debido a la capacidad de LSTM de retener información a corto plazo. Dentro del conjunto destacado, el que mejor se comporto fue aquel que retuvo información de 12 meses de consumos previos, manteniendo un error por debajo del 8% e indicadores de predicción con una tendencia cercana al 0, esto puede indicar que si se generan más atributos puede que la red trabaje mejor al contar con más información entre cada iteración.

Finalmente, el uso de la metodología CRISP-DM permitió tener un desarrollo más rápido y fluido debido a que especifica que debe contener cada etapa y lo necesario para poder avanzar sobre los siguientes pasos.

4.2 Futuras líneas de investigación

A partir de este proyecto surgen varias líneas de trabajo que serán enumeradas a continuación:

- Desarrollar un sistema que haga uso de los modelos de predicción de consumos de agua, orientado al usuario y con la posibilidad de seleccionar y parametrizar cada modelo a través de una interfaz amigable.

- Predecir consumos por usuario según las características propias a la conexión en donde se encuentre, esto permitiría la detección de posibles fraudes, la concientización del cuidado de agua y el ahorro de dinero.
- Comparar los métodos de predicción estadísticos contra los métodos de predicción basados en redes neuronales.

Bibliografía

[1] Iagua, “El impacto del Agua No Facturada y su solución a través de la tecnología | iAgua”, 2019. https://www.iagua.es/noticias/goaigua/impacto-agua-no-facturada-y-solucion-traves-tecnologia (accessed Jul. 26, 2022).

[2] D. M. García, A. Yully, D. Sánchez, and S. V. Ramírez, “Benchmarking Para La Gestión De Pérdidas De Agua En Sistemas De Acueductos”, *Encuentro Int. Educ. en Ing. ACOFI 2019*, p. 9, 2019, [Online]. Available: https://acofipapers.org/index.php/eiei/article/view/192.

[3] F. Gómez Isa, “Agua y privatización: un enfoque de derechos humanos”, *El derecho Hum. al agua situación actual y retos Futur.*, pp. 163–176, 2008, Accessed: May 16, 2022. [Online]. Available: http://dialnet.unirioja.es/servlet/articulo?codigo=2742768.

[4] M. M. L. L. y J. M. B. A. Cruz García Lirios, Javier Carreón Guillén, Jorge Hernández Valdés, “Actitudes, consumo de agua y sistema de tarifas del servicio de abastecimiento de agua potable”, *Polis Revista Latinoamericana*, pp. 1–33, 2013.

[5] V. G. TZATCHKOV and V. H. ALCOCER-YAMANAKA, “Modelación de la variación del consumo de agua potable con métodos estocásticos”, *Tecnologia y ciencias del agua*, 2016.

[6] J. C. Riquelme, R. Ruiz, and K. Gilbert, “Minería de Datos: Conceptos y Tendencias”, *Rev. Iberoam. Intel. Artif.*, vol. 29, pp. 11–18, 2006.

[7] Iagua, “El concepto de usuario en el ámbito de la gestión de la Administración hidráulica | iAgua”, 2014. https://www.iagua.es/blogs/ines-torralba/concepto-usuario-ambito-gestion-administracion-hidraulica (accessed Jul. 22, 2022).

[8] C. I. L. Plasencia, “Modelo de solución de Business Intelligence y Machine

Learning para el monitoreo y control de calidad de la medición del consumo de agua en el Centro de Servicios Breña.", *Orphanet J. Rare Dis.*, vol. 1, no. 1, pp. 1–9, 2020.

[9] Ambientum, "El consumo de agua en porcentajes - Enciclopedia Medioambiental." https://www.ambientum.com/enciclopedia_medioambiental/aguas/el-consumo-de-agua-en-porcentajes.asp (accessed Jul. 22, 2022).

[10] Selis, "Tomaestados, SELIS." https://www.selis.com/productos/87-tomaestados.html (accessed Jul. 22, 2022).

[11] JAPAC, "¿Qué es un medidor de agua? | JAPAC – Agua y Salud para todos", 2017. https://japac.gob.mx/2017/07/17/que-es-un-medidor-de-agua/ (accessed Jul. 22, 2022).

[12] S. Huaquisto Cáceres and I. G. Chambilla Flores, "Análisis Del Consumo De Agua Potable En El Centro Poblado De Salcedo, Puno", *Investig. Desarro.*, vol. 19, no. 1, pp. 133–144, 2019, doi: 10.23881/idupbo.019.1-9i.

[13] UNESCO, "Informe del 2019 – No dejar a nadie atrás", 2019. https://es.unesco.org/water-security/wwap/wwdr/2019#download (accessed Apr. 29, 2022).

[14] A. Silberschatz, H. F. Korth, and S. Sudarshan, *FUNDAMENTOS DE BASES DE DATOS*, 4ta ed. México: McGraw-Hill, 2002.

[15] V. N. Cabello, *Introducción a las Bases de Datos relacionales*. Vision Libros, 2010.

[16] J. Sánchez, "Principios sobre Bases de Datos Relacionales", *Creative Commons*, 2004.

[17] C. J. Date, *Introducción a los sistemas de bases de datos*, 7ma ed. Pearson Educación, 2001.

[18] I. A. Espinosa, "DATA WAREHOUSE PARA LA GESTIÓN DE LISTA DE ESPERA SANITARIA", Universidad Politecnica de Madrid, 2008.

[19] U. Fayyad, "Knowledge Discovery in Databases: An Overview", *Relational Data Min.*, pp. 28–47, 2001, doi: 10.1007/978-3-662-04599-2_2.

[20] H. O. Nigro, D. Xodo, G. Corti, and D. Terren, "KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario", *VI Work. Investig. en Ciencias la Comput.*, 2004.

[21] U. Fayyad and P. Stolorz, "Data mining and KDD: Promise and challenges", *Futur. Gener. Comput. Syst.*, vol. 13, no. 2–3, pp. 99–115, Nov. 1997, doi: 10.1016/s0167-739x(97)00015-0.

[22] A. Jović, K. Brkić, and N. Bogunović, "An overview of free software tools for general data mining", *2014 37th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2014 - Proc.*, no. May, pp. 1112–1117, 2014, doi: 10.1109/MIPRO.2014.6859735.

[23] R. Mikut and M. Reischl, "Data mining tools", *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 5, pp. 431–443, 2011, doi: 10.1002/widm.24.

[24] N. R. Zulyanti, "Data Visualization Application 'JIRA' at PT. Jati Piranti Solusindo Using the Metabase", *J. Mantik*, vol. 3, no. 4, pp. 444–450, 2020.

[25] L. C. Molina, "Data mining: torturando a los datos hasta que confiesen", pp. 1–11, 2000.

[26] S. J. Vallejos, "Minería de Datos", *UNNE*, 2006.

[27] M. I. Ángeles Larrieta and A. M. Santillán Gómez, "Minería de datos: Concepto, características, estructura y aplicaciones", 2004. [Online]. Available: http://www.ejournal.unam.mx/rca/190/RCA19007.pdf.

[28] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and

knowledge discovery process models and methodologies", *Knowl. Eng. Rev.*, vol. 25, no. 2, pp. 137–166, 2010, doi: 10.1017/S0269888910000032.

[29] IBM, "Conceptos básicos de ayuda de CRISP-DM - IBM Documentation", 2021. https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview (accessed May 20, 2022).

[30] I. Juan, M. Moine, D. S. Gordillo, D. Ana, and S. Haedo, "Análisis comparativo de metodologías para la gestión de proyectos de minería de datos", *Congr. Argentino Ciencias la Comput.*, 2011.

[31] C. Hernández and X. Dueñas, "Hacia una metodología de gestión del conocimiento basada en minería de datos", *Comtel*, pp. 80–96, 2009, [Online]. Available: http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/982/COMTEL-2009-80-96.pdf?sequence=1.

[32] J. M. Moine, A. S. Haedo, U. T. Nacional, and F. R. Rosario, "Una herramienta para la evaluación y comparación de metodologías de minería de datos", *XXI Congr. Argentino Ciencias la Comput.*, 2015.

[33] B. I. Chuchuca Alvarracin and J. D. Sicha Rodriguez, "Desarrollo e implementacion de un algoritmo de prediccion de consumo de agua potable y visualizacion de los datos de consumo por usuario y sectores mediante redes neuronales recurrentes dentro del proyecto cedia-tarpuq", Tesis de grado, Universidad Politecnica Salesiana, 2014.

[34] A. Rabasa, J. J. Rodríguez, L. Santamaría, J. F. Monge, and R. Neuronales, "Predicción sobre Series Temporales No-lineales con Redes Neuronales y modelos ARIMA", *Cent. Investig. Oper.*, vol. 23, no. 1576–7264, 2006.

[35] N. Rodríguez, "Pronóstico de demanda de agua potable mediante redes neuronales", Tesis de grado, Universidad Tecnica Federico Santa Maria, 2016.

[36] Morán Álvares Antonio, "Analisis y predicción de perfiles de consumo energético

en edificios públicos mediante técnicas de minería de datos", Tesis doctoral, Universidad de Oviedo, 2012.

[37] Universitat Carlemany, "Análisis predictivo: tipos, técnicas y herramientas | Universitat Carlemany." https://www.universitatcarlemany.com/actualidad/analisis-predictivo-tipos-herramientas (accessed May 16, 2022).

[38] Orea Valero Sergio, Vargas Salvador Alejandro, and Marcela García Alonso, "Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo el algoritmo de K vecinos más cercanos", pp. 4–9, 2005.

[39] D. J. Matich, "Redes Neuronales: Conceptos Básicos y Aplicaciones.", *Historia Santiago.*, p. 55, 2001.

[40] K. Shihab, "A Backpropagation Neural Network for Computer Network Security", *J. Comput. Sci.*, vol. 2, no. 9, pp. 710–715, 2006.

[41] C. Arana, "Redes Neuronales Recurrentes: Análisis de los Modelos Especializados en Datos Secuenciales", 2021. Accessed: May 18, 2022. [Online]. Available: www.cema.edu.ar/publicaciones/doc_trabajo.html.

[42] F. Soto-Bravo and M. I. González-Lutz, "Análisis de métodos estadísticos para evaluar el desempeño de modelos de simulación en cultivos horticolas", *Agron. Mesoam.*, 2019, doi: 10.15517/am.v30i2.33839.

[43] F. F. Medina, "Realtime Data Mining Aplicado a La Predicción De Índices De Bolsa Incluyendo Social Media Analytics", Tesis de máster, Universitat Politecnica Catalunya, 2017.

[44] E. J. Menke, "Series of Jupyter Notebooks Using Python for an Analytical Chemistry Course", *J. Chem. Educ.*, vol. 97, no. 10, pp. 3899–3903, Oct. 2020, doi: 10.1021/ACS.JCHEMED.9B01131/ASSET/IMAGES/LARGE/ED9B01131_000

1.JPEG.

[45] Spyder Doc Contributors, “Welcome to Spyder’s Documentation — Spyder 5 documentation.” https://docs.spyder-ide.org/current/index.html (accessed Jul. 28, 2022).

[46] Anaconda Inc, “Anaconda | The World’s Most Popular Data Science Platform.” https://www.anaconda.com/ (accessed Jul. 28, 2022).

[47] NumPy Developers, “What is NumPy? — NumPy v1.23 Manual.” https://numpy.org/doc/stable/user/whatisnumpy.html# (accessed Jul. 28, 2022).

[48] Pandas, “pandas documentation — pandas 1.4.3 documentation.” https://pandas.pydata.org/docs/ (accessed Jul. 28, 2022).

[49] Keras, “About Keras.” https://keras.io/about/ (accessed Jul. 28, 2022).

[50] J. Hunter, D. Dale, E. Firing, M. Droettboom, and Matplotlib development team, “Matplotlib documentation — Matplotlib 3.6.0.dev2729+g446de7bbb6 documentation.” https://matplotlib.org/devdocs/index.html (accessed Jul. 28, 2022).

[51] Z. Reitermanová, “Data Splitting”, *WDS’10 Proc. Contrib. Pap.*, 2010.

[52] Gobierno de España, “¿Cómo sé si mi modelo de predicción es realmente bueno? | datos.gob.es”, 2021. https://datos.gob.es/es/blog/como-se-si-mi-modelo-de-prediccion-es-realmente-bueno (accessed Aug. 15, 2022).

[53] P. S. Akhilendra, “Evaluation Metrics for Regression models- MAE Vs MSE Vs RMSE vs RMSLE”, 2019. https://akhilendra.com/evaluation-metrics-regression-mae-mse-rmse-rmsle/ (accessed May 30, 2022).

[54] Municipalidad Argentina, “San José en la region de Misiones - Municipio y gobierno municipal de Argentina.” https://www.municipalidad-argentina.com.ar/municipalidad-san-jose-n.html (accessed Aug. 13, 2022).

[55] DBeaver, "DBeaver Community | Free Universal Database Tool." https://dbeaver.io/ (accessed Aug. 07, 2022).

[56] Metabase, "Metabase." https://www.metabase.com/ (accessed Aug. 09, 2022).

[57] RapidMiner, "RapidMiner | Best Data Science & Machine Learning Platform." https://rapidminer.com/ (accessed Aug. 23, 2022).

[58] I. Ariel, A. Fierro, and F. Ronchetti, "Predicción de Series Temporales con Redes Neuronales", Universidad Nacional de la Plata, 2020.

[59] O. Surakhi *et al.*, "Time-lag selection for time-series forecasting using neural network and heuristic algorithm", *Electron.*, vol. 10, no. 20, pp. 1–22, 2021, doi: 10.3390/electronics10202518.

[60] C. Kühnert, N. M. Gonuguntla, H. Krieg, D. Nowak, and J. A. Thomas, "Application of LSTM networks for water demand prediction in optimal pump control", *Water (Switzerland)*, vol. 13, no. 5, Mar. 2021, doi: 10.3390/W13050644.

Anexo

Entrevista

1. **¿Hace cuánto está en funcionamiento la CASAJ?**
   Desde el año 1966
2. **¿La CASAJ cuenta con un sistema para gestión de consumos y facturación?**
   Si, el sistema actual permite gestionar clientes, consumos y facturación. Está en funcionamiento desde el año 2004 aproximadamente.
3. **¿Cuáles son las fuentes de abastecimiento de agua y la capacidad de cada una?**
   La CASAJ cuenta con tres fuentes de abastecimiento, un pozo perforado de 197 metros ubicado en la planta de potabilización, el Arroyo Pindapoy Grande ubicado aproximadamente a 580 metros de distancia de la planta y finalmente una perforación secundaria o de menor escala de 109 metros ubicado a 50 metros del Arroyo Pindapoy Grande.
4. **¿Cuál es la capacidad de producción de la planta de potabilización?**
   En temporadas de bajo consumo se producen unos 500.000 a 800.000 litros de agua potable al dia y en temporadas de alto consumo entre 1.000.000 a 1.600.000 litros, es decir, se duplica, y por ende la planta está en funcionamiento todo el día.
5. **¿Existen temporadas del año donde se requiera de más producción?**
   Si por lo general en el verano, hay casos donde la planta de potabilización funciona casi a su límite para poder satisfacer toda la demanda de agua.
6. **¿Todo lo que se produce es facturado?**
   No, generalmente se pierde entre el 30 y el 45 % de la producción en pérdidas de la red de agua.
7. **¿Existe otra empresa de potabilización en la ciudad?**
   No, somos la única empresa que brinda agua potable a San José.
8. **¿Se planifica y estima que insumos serán necesarios para próximos meses?**

Se realiza muy pocas veces y en base a la experiencia de los empleados más antiguos.

9. **¿Cuentan con algún registro de los litros de agua potable producidos?**
   Se cuenta con una planilla donde se suele anotar lo que se produce a diario, pero no se le da el seguimiento correspondiente, por ende, puede que no tenga datos o incluso este desactualizada.

Consultas SQL

Figura 9:

```
SELECT "public"."categoria_socio"."id_categoria_socio" AS
"id_categoria_socio", "public"."categoria_socio"."tipo_socio" AS
"tipo_socio", "public"."categoria_socio"."factor_multiplicacion" AS
"factor_multiplicacion"
FROM "public"."categoria_socio"
LIMIT 1048575
```

Figura 10:

```
SELECT "public"."conexiones_situacion"."id_situacion" AS "id_situacion",
"public"."conexiones_situacion"."situacion_conexion" AS
"situacion_conexion"
FROM "public"."conexiones_situacion"
LIMIT 1048575
```

Figura 11:

```
SELECT ((floor((("public"."conexiones"."manzana" - 0.0) / 20)) * 20) + 0.0)
AS "manzana", count(*) AS "count"
FROM "public"."conexiones"
GROUP BY ((floor((("public"."conexiones"."manzana" - 0.0) / 20)) * 20) +
0.0)
ORDER BY ((floor((("public"."conexiones"."manzana" - 0.0) / 20)) * 20) +
0.0) ASC
```

Figura 12:

```
SELECT "Zonas - ID Zona"."zona" AS "Zonas - ID Zona__zona", count(*) AS
"count"
FROM "public"."conexiones"
LEFT JOIN "public"."zonas" "Zonas - ID Zona" ON
"public"."conexiones"."id_zona" = "Zonas - ID Zona"."id_zona"
GROUP BY "Zonas - ID Zona"."zona"
ORDER BY "Zonas - ID Zona"."zona" ASC
```

Figura 13:

```
SELECT "public"."estados"."periodo_ano" AS "periodo_ano", count(*) AS
"count"
FROM "public"."estados"
WHERE ("public"."estados"."periodo_ano" = 2015
    OR "public"."estados"."periodo_ano" = 2016 OR
"public"."estados"."periodo_ano" = 2017 OR "public"."estados"."periodo_ano"
= 2018 OR "public"."estados"."periodo_ano" = 2019 OR
"public"."estados"."periodo_ano" = 2020)
GROUP BY "public"."estados"."periodo_ano"
ORDER BY "count" DESC, "public"."estados"."periodo_ano" ASC
```

Figura 14:

```
SELECT ((floor((("public"."estados"."periodo_ano" - 2010.0) / 0.5)) * 0.5)
+ 2010.0) AS "periodo_ano", "public"."estados"."periodo_mes" AS
"periodo_mes", count(*) AS "count"
FROM "public"."estados"
WHERE (("public"."estados"."periodo_ano" = 2015
    OR "public"."estados"."periodo_ano" = 2016 OR
"public"."estados"."periodo_ano" = 2017 OR "public"."estados"."periodo_ano"
= 2018 OR "public"."estados"."periodo_ano" = 2019 OR
"public"."estados"."periodo_ano" = 2020)
   AND "public"."estados"."estado" = 0)
GROUP BY ((floor((("public"."estados"."periodo_ano" - 2010.0) / 0.5)) *
0.5) + 2010.0), "public"."estados"."periodo_mes"
ORDER BY ((floor((("public"."estados"."periodo_ano" - 2010.0) / 0.5)) *
0.5) + 2010.0) ASC, "public"."estados"."periodo_mes" ASC
```

Figura 15:

```
SELECT ((floor((("Facturacion Detalle"."cantidad" - 0.0) / 12.5)) * 12.5) +
0.0) AS "Facturacion Detalle__cantidad",
sum("public"."facturacion"."numero_factura") AS "sum"
FROM "public"."facturacion"
LEFT JOIN "public"."facturacion_detalle" "Facturacion Detalle" ON
"public"."facturacion"."id_facturacion" = "Facturacion
Detalle"."id_factura"
WHERE ("Facturacion Detalle"."id_factura_conceptos" = 1
   AND "public"."facturacion"."fecha_factura" >= timestamp with time zone
'2015-01-01 00:00:00.000-03:00' AND "public"."facturacion"."fecha_factura"
< timestamp with time zone '2021-08-10 00:00:00.000-03:00')
GROUP BY ((floor((("Facturacion Detalle"."cantidad" - 0.0) / 12.5)) * 12.5)
+ 0.0)
ORDER BY ((floor((("Facturacion Detalle"."cantidad" - 0.0) / 12.5)) * 12.5)
+ 0.0) ASC
```

Figura 16:

```
SELECT ((floor((("Facturacion Detalle"."cantidad" - 0.0) / 12.5)) * 12.5) +
0.0) AS "Facturacion Detalle__cantidad", count(*) AS "count"
FROM "public"."facturacion"
LEFT JOIN "public"."facturacion_detalle" "Facturacion Detalle" ON
"public"."facturacion"."id_facturacion" = "Facturacion
Detalle"."id_factura" LEFT JOIN "public"."facturacion_conceptos"
"Facturacion Conceptos - ID Factura Conceptos" ON "Facturacion
Detalle"."id_factura_conceptos" = "Facturacion Conceptos - ID Factura
Conceptos"."id_factura_conceptos"
GROUP BY ((floor((("Facturacion Detalle"."cantidad" - 0.0) / 12.5)) * 12.5)
+ 0.0)
ORDER BY ((floor((("Facturacion Detalle"."cantidad" - 0.0) / 12.5)) * 12.5)
+ 0.0) ASC
```

Figura 17:

```
SELECT "public"."facturas_estados"."estado_factura" AS "estado_factura",
count(*) AS "count"
FROM "public"."facturas_estados"
GROUP BY "public"."facturas_estados"."estado_factura"
ORDER BY "public"."facturas_estados"."estado_factura" ASC
```

Figura 18:

```
SELECT "Facturas Estados - ID Estado Factura"."estado_factura" AS "Facturas
Estados - ID Estado Factura__estado_factura", count(*) AS "count"
FROM "public"."facturacion"
LEFT JOIN "public"."facturas_estados" "Facturas Estados - ID Estado
Factura" ON "public"."facturacion"."id_estado_factura" = "Facturas Estados
- ID Estado Factura"."id_estado_factura"
WHERE ("public"."facturacion"."fecha_factura" >= timestamp with time zone
'2015-01-01 00:00:00.000-03:00'
   AND "public"."facturacion"."fecha_factura" < timestamp with time zone
'2021-08-12 00:00:00.000-03:00')
GROUP BY "Facturas Estados - ID Estado Factura"."estado_factura"
ORDER BY "Facturas Estados - ID Estado Factura"."estado_factura" ASC
```

Filtro Conexiones Activas

```
SELECT "public"."conexiones"."id_conexion" AS "id_conexion",
"public"."conexiones"."fecha_alta" AS "fecha_alta",
"public"."conexiones"."observaciones" AS "observaciones",
"public"."conexiones"."id_entidad" AS "id_entidad",
"public"."conexiones"."numero_medidor" AS "numero_medidor",
"public"."conexiones"."direccion_conexion" AS "direccion_conexion",
"public"."conexiones"."id_categoria_socio" AS "id_categoria_socio",
"public"."conexiones"."id_situacion" AS "id_situacion",
"public"."conexiones"."id_entidad_usuario" AS "id_entidad_usuario",
"public"."conexiones"."id_zona" AS "id_zona",
"public"."conexiones"."manzana" AS "manzana",
"public"."conexiones"."adherentes" AS "adherentes",
"public"."conexiones"."nomenclatura" AS "nomenclatura",
"public"."conexiones"."propietario" AS "propietario",
"public"."conexiones"."n_acta" AS "n_acta",
"public"."conexiones"."fecha_acta" AS "fecha_acta"
FROM "public"."conexiones"
WHERE "public"."conexiones"."id_situacion" = 1
LIMIT 1048575
```

Filtro Conexiones Estado 0

```
SELECT "source"."id_conexion" AS "id_conexion", "source"."count" AS "count"
```

```
FROM (SELECT "public"."estados"."id_conexion" AS "id_conexion", count(*) AS
"count" FROM "public"."estados"
LEFT JOIN "public"."conexiones" "Conexiones - ID Conexion" ON
"public"."estados"."id_conexion" = "Conexiones - ID Conexion"."id_conexion"
WHERE (("public"."estados"."periodo_ano" = 2015
    OR "public"."estados"."periodo_ano" = 2016 OR
"public"."estados"."periodo_ano" = 2017 OR "public"."estados"."periodo_ano"
= 2018 OR "public"."estados"."periodo_ano" = 2019 OR
"public"."estados"."periodo_ano" = 2020)
   AND "public"."estados"."estado" = 0)
GROUP BY "public"."estados"."id_conexion"
ORDER BY "public"."estados"."id_conexion" ASC) "source" WHERE
"source"."count" < 45
LIMIT 1048575
```

Filtro Conexiones Estado Periodo 04/2020

Regla de Trasformación de Periodos

if(periodo_mes == 1,1, if(periodo_mes == 2,2,if(periodo_mes == 3,3,if(periodo_mes == 4,4,if(periodo_mes == 5,5,if(periodo_mes == 6,6,if(periodo_mes == 7,7,if(periodo_mes == 8,8,if(periodo_mes == 9,9,if(periodo_mes == 10,10,if(periodo_mes == 11,11,if(periodo_mes == 12,12,0)))))))))))