UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA CIVIL

DETECCIÓN Y LOCALIZACIÓN DE FUGAS EN PARTE DE LA RED DE DISTRIBUCIÓN DE AGUAS DE SANTIAGO, UTILIZANDO UNA MÁQUINA DE VECTOR DE APOYO (SVM)

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL

BORIS ALEJANDRO GÁRATE MORALES

PROFESOR GUÍA:

YARKO NIÑO CAMPOS.

MIEMBROS DE LA COMISIÓN:

YOLANDA ALBERTO

NITZA MIRANDA VALLEJOS

SANTIAGO DE CHILE

2021

Powered@NLHPC: Esta investigación/tesis fue parcialmente apoyada por la infraestructura de supercómputo del NLHPC (ECM-02)

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: INGENIERO CIVIL CON MENCIÓN EN HIDRÁULICA, SANITARIA Y AMBIENTAL
POR: BORIS GÁRATE MORALES
PROFESOR GUÍA: YARKO NIÑO

DETECCIÓN Y LOCALIZACIÓN DE FUGAS EN PARTE DE LA RED DE DISTRIBUCIÓN DE AGUAS DE SANTIAGO, UTILIZANDO UNA MÁQUINA DE VECTOR DE APOYO (SVM)

En el marco de la escasa disponibilidad de agua por la que se atraviesa en Chile en la actualidad, además de las altas pérdidas en la distribución de agua potable existentes (33,8% promedio a nivel nacional en el caso de Chile). El presente trabajo de título consiste en la implementación de una metodología de detección y localización de fugas basado en un algoritmo de aprendizaje supervisado como lo es una máquina de vector de apoyo (SMV).

El objetivo general es la utilización de un algoritmo de aprendizaje supervisado, como la SVM, en la detección y localización de fugas para una red de distribución de gran tamaño. Para esto, se modela la red en el software OpenFlows WaterGEMS, de donde se obtienen los datos (presión y localización de la fuga) para el entrenamiento y la prueba de la SVM.

Dentro de los resultados del trabajo se tienen un total de cinco SVM que no tienen un buen comportamiento al momento de detectar y localizar las fugas, ya que como resultado arrojan un comportamiento sin fuga para la mayoría de los escenarios de prueba y entrenamiento. Además, se tienen dos SVM que detectan y localizan de buena forma fugas dentro del modelo hidráulico, en las que es posible ajustar uno de sus parámetros (C) para tratar de evitar el sobreajuste y obtener un correcto funcionamiento en la detección y localización de fugas.

Dedicatoria

A mi familia en general, pero en especial a mis padres y a mi compañera de vida, por ayudarme a no decaer, brindarme su apoyo incondicional y confiar siempre en mí.

Agradecimientos

Finalizando mi etapa en la universidad, me gustaría agradecer a todas las personas que me han acompañado a lo largo de toda mi vida que me han hecho ser la persona que soy.

Primero que todo, agradecerle a mi madre y mi padre, quienes siempre me han brindado su apoyo para cumplir las metas que me he propuesto, pero por sobre todo agradecer por todo el esfuerzo que han realizado a lo largo de toda mi vida para darnos el mejor pasar a mis hermanos y a mí. También me gustaría agradecer a mi hermana y mi hermano, quienes me han tenido que aguantar desde que soy chico y también me han apoyado en todo momento.

De igual forma, me gustaría agradecer a mi compañera de vida Estrella, quien ha sido mi sostén emocional y me ha apoyado en mis mejores y peores momentos.

Finalmente, me gustaría agradecer a todas las amistades que he realizado a lo largo de mi vida, ya que sin los momentos vividos con cada una de ellas no habría llegado a ser la persona que hoy soy.

Tabla de contenido

Índice de tablas

Índice de figuras

1. Introducción

1.1. Motivación

El agua es un recurso esencial para la vida humana y el desarrollo de nuestras actividades diarias, por lo que es necesaria una buena gestión y distribución de esta. En cualquier urbe, como en Santiago, el agua es distribuida mediante una red de distribución de aguas, la que puede experimentar un deterioro debido a diferentes factores, los que se pueden clasificar como internos o externos. Los internos se refieren a daños al interior de la misma tubería, causador por la corrosión, defectos de fábrica, antigüedad, mala mano de obra, etc. Los externos se refieren a los daños ocasionado por terceros como daños causados por carga excesiva, excavaciones, condiciones climáticas y el movimiento del suelo.

Para una buena gestión en la distribución de aguas es necesario poder localizar zonas con daños en la tubería y con fugas. Es por esto que existen diversos métodos para la detección de fugas en tuberías. Chan et al. (2018) realiza una revisión de los distintos sistemas de detección de fugas y concluye que en la actualidad los métodos existentes poseen bastantes limitaciones, como supuestos poco realistas, poca precisión y altos requerimientos en cuanto a sensores y datos.

En Chile, debido al incremento poblacional y la urbanización generalizada, las demandas de agua potable han aumentado en los últimos años, por lo que la buena gestión de esta es un tema crucial para no sobreexplotar la producción. Según el informe de gestión de la Super Intendencia de Servicios Sanitarios del 2018 (SISS, 2018), las aguas no facturadas en promedio a nivel nacional son del 33,8%. Por lo que, este estudio busca mejorar la gestión mediante la detección y localización de fugas utilizando la metodología desarrollada por Kemba et al. (2017), la que consta de utilizar un algoritmo de aprendizaje supervisado como una máquina de vector de apoyo (SVM) para detectar y localizar las fugas de un modelo hidráulico en el software EPANET. Esto se aplica a una red de mayor tamaño a la utilizada en el estudio de Kemba, para comparar los resultados obtenidos y comprobar si es posible extrapolar a redes de mayor tamaño las conclusiones obtenidas para redes pequeñas.

1.2. Objetivos

1.2.1. General

El objetivo principal del trabajo de memoria es la utilización de un algoritmo de aprendizaje supervisado, como la SVM, para la detección y localización de fugas en parte de la red de agua de una ciudad grande de Chile.

1.2.2. Específicos

Los objetivos específicos son los siguientes:

- Transformar un modelo hidráulico de la red de distribución de un sector de la ciudad de estado estático a estado dinámico.
- Programar máquina de vector de apoyo (SVM) para detectar fugas en redes de distribución de agua potable.
- Aplicar la metodología propuesta a la red de distribución modelada en estado dinámico.
- Comparar la efectividad de la metodología propuesta para la red modelada y compararlas con la obtenida por Kemba et al. (2017).

1.3. Metodología

1.3.1. Métodos de detección y localización de fugas

Se realiza una revisión bibliográfica acerca de los distintos métodos para la detección y localización de fugas. Con la finalidad de encontrar cuales son los datos de entrada necesarios para lograr la predicción o localización con las distintas metodologías. Se puso especial énfasis en los métodos mediante algoritmos de aprendizaje supervisado, que es el que finalmente se desarrolla en la memoria.

1.3.2. Recopilación de datos y creación de los modelos de la red de distribución de aguas de Santiago

Los datos fueron entregados por la empresa concesionaria del sector del estudio, quien decidió no seguir participando del mismo. Se recopila la información acerca de la red de distribución de aguas de la Ciudad tales como:

- Dimensiones
- Materialidad de tuberías
- Ubicación de nodos para la distribución
- Ubicación de elementos de la red como válvulas reductoras de presión, válvulas controladoras de flujo, estanques, etc.
- Condiciones de borde, características y consignas de los elementos nombrados en el punto anterior.
- Demandas del sector

Con esta información es posible crear un modelo para una parte de la red de distribución de la ciudad mediante el Software OpenFlows WaterGEMS.

1.3.3. Generar set de datos de simulaciones para entrenamiento y resolución.

Con el modelo en WaterGEMS se crea un set de datos de comportamientos normales (sin fuga) y anormales (con fuga) en la tubería. El set de datos constará de:

- Datos del comportamiento “normal” de la red para distintos caudales y presiones de flujo.
- Datos del comportamiento de la red para fugas de diferentes localizaciones y para distintos caudales y presiones de flujo.

1.3.4. Programar la SVM

Una SVM puede utilizarse para generar una regresión o una clasificación. En el presente estudio se utiliza como clasificador, donde la salida es una clase predictiva asociada a un patrón de entrada. En este caso las salidas corresponden a la ubicación del nodo más cercano del modelo en que se produce la fuga. Para la programación, la SVM puede utilizar diferentes funciones para el núcleo como lineal, polinómica y función de base radial, siendo esta última la que va a utilizarse en el presente estudio.

1.3.5. Entrenar la SVM para la clasificación de los estados de la red

Teniendo el set de simulaciones es posible entrenar la SVM con un porcentaje de estas para que el algoritmo reconozca la diferencia entre estados de la red mediante un patrón en los datos de presión y caudal entregados. Luego de entrenada la SVM, esta clasificará el resto de las simulaciones como comportamiento normal o anormal, entregando en caso de ser necesario la ubicación de la fuga o de la anomalía encontrada.

1.3.6. Analizar los resultados entregados por la SVM

Con los datos obtenidos para la red modelada, se compara los resultados obtenidos con Kemba et al. (2017), comparando la eficiencia en la detección y localización de fugas en redes de gran tamaño con redes de menor tamaño.

1.4. Organización por capítulo

1.4.1. Introducción

Se presenta la motivación del presente trabajo, los objetivos, y la metodología usada.

1.4.2. Estado del arte

Se exhiben las características de distintas metodologías actuales para la detección y localización de fugas.

1.4.3. Marco teórico

Se expone la teoría que constituye la base donde se sustentan los algoritmos y programas utilizados.

1.4.4. Modelo de distribución y calibración

Se detalla los procedimientos realizados para la construcción y calibración del modelo de distribución

1.4.5. Detección y localización de fugas utilizando una SVM

Se presenta detalladamente el paso a paso realizado para la confección del algoritmo de detección y localización.

1.4.6. Resultados

Se exponen los resultados del modelo calibrado y de la SVM.

1.4.7. Análisis de resultados

Se analizan los resultados entregados por la SVM.

1.4.8. Conclusiones

Se presentan las conclusiones de los resultados obtenidos, las implicancias de los resultados y sus limitaciones.

2. Estado del Arte: Detección y localización de fugas en tuberías en la actualidad

En la siguiente sección se realizará un repaso por las diferentes metodologías utilizadas para la detección y localización de fugas en sistemas de distribución de agua potable, las que se pueden categorizar como de predicción, clasificación, agrupación, análisis estadístico, basado en modelos y de señales transitorias (Chan et al. 2018).

2.1. Predicción

Este tipo de metodologías se basa en realizar una serie de predicciones de demandas y con las lecturas reales de las mismas se detecta alguna discrepancia causada por un flujo anormal.

Laucelli et al, (2016) proponen el paradigma de Regresión Polinómica Evolutiva (RPE), con el objetivo de reproducir el comportamiento de una red utilizando datos de presión y flujo en línea. Asumieron una función de densidad de probabilidad para representar estadísticamente el posible comportamiento de la red, donde, si el valor excede el umbral, basado en toda la probabilidad acumulativa del caudal máximo previsto para cada etapa de tiempo, se detecta una anomalía. Si se produce una demanda de agua inesperada, se clasificaría como un comportamiento anómalo. Dentro de las desventajas están que una fuga de gran magnitud podría provocar que otras de carácter menor pasen desapercibidas, por lo que, estas fugas pequeñas podrían solo detectarse posterior a la resolución de la fuga mayor. Esta metodología es dependiente de la experiencia de los operadores de los servicios de agua y del historial de consumo de la red.

2.2. Clasificación

En estas metodologías se entrena un clasificador basado en un conjunto de características que identifiquen de buena forma las particularidades de los diferentes tipos de eventos, para poder categorizarlos basándose en series temporales de presión y caudal.

Kang et al. (2017) utilizan una CNN (convolutional neural network) que extrae las características no lineales de las series temporales, pero éstas no son óptimas para la clasificación, mientras que una SVM (máquina de vector de apoyo) con una función de núcleo fija, es un clasificador eficaz cuando se aplican a características bien definidas. Propone también una mejora al rendimiento aplicando una arquitectura híbrida 1D-CNN-SVM. Para la localización modifica un algoritmo basado en gráficos de Srirangaran et al. (2013), donde se busca el nodo más cercano al lugar de la fuga y se impone una limitación del alcance de la búsqueda del nodo de fuga. Para calcular el tiempo de llegada, se utiliza una correlación cruzada generalizada con una función de ponderación de máxima

probabilidad y con la información de velocidad de onda y la longitud de las tuberías, es posible estimar la ubicación de la fuga. Los inconvenientes mostrados fueron que no era posible eliminar bien el ruido para la extracción de características, el tiempo de cálculo de la CNN es mayor que el de las SVM y la dependencia de una buena estimación de la velocidad de la onda para obtener localización precisa.

Por otra parte, Tao et al. (2014) propone una metodología de detección de roturas basada en una red AIS (artificial inmune system), la que simula el sistema inmunológico. La ubicación puede ser identificada mediante una Red Neuronal Aritificial de varios niveles; el primer nivel identifica la existencia de fuga y el segundo estima la magnitud y ubicación. La metodología expuesta fue:

i. Recopilación de datos.
ii. Procesamiento de datos.
iii. Extracción de características.
iv. Capacitación del modelo de clasificación.
v. Calibración del modelo con datos de roturas
vi. Ubicación de la rotura mediante el principio de los vecinos más cercanos

La metodología requiere de un modelo hidráulico bien calibrado y una gran cantidad de datos, la falta de uno de estos incide considerablemente en la exactitud de los resultados. Otras desventajas del método son imposibilidad de detectar pequeñas fugas, la localización se ve influenciada por la ubicación de los sensores, por lo que, solo sugiere una zona de fuga, requiriendo una inspección adicional para encontrar la fuga.

2.3. Agrupación

Es un tipo de metodologías no supervisadas. La idea es dividir la red de distribución o los datos en diferentes agrupaciones y en estas utilizar otra estrategia para identificar la posible zona de fuga.

Wu et al. (2016) propone un método que detecta las roturas dentro de una red de distribución de aguas. La metodología consiste en la transformación de datos para reducir el alcance de la variación de estos. Se detectan los valores atípicos diferenciando los inducidos por cambios de consumos inesperados y los inducidos por roturas. Se considera atípico un vector que no pertenece a ningún grupo. Sugiere que, si la media del valor atípico es más o menos de 0, con un número de elementos anormalmente grandes, se debía a una temporada cálida o fría respectivamente. Si no pertenece a estos, se considera fuga, es decir, si la medida identificada indica que la temporada es cálida, pero los datos pertenecen a la temporada fría, se considera una fuga. Los datos se extrajeron a las 6 de la mañana para una mínima variación del flujo, lo que hace que el método solo sea aplicable en la mañana. Presenta gran cantidad de falsos positivos, ya que un aumento inesperado en la demanda se clasifica como fuga. En Wu et al. (2018) se redujo la demanda de datos históricos para omitir el proceso de depuración de datos y tener en cuenta los cambios meteorológicos, las festividades y los cambios de

demandas periódicos. Se reducen los falsos positivos causadas por los cambios repentidos y siguieron sin poder abordar el problema causado por la demanda inesperada de agua. Sigue siendo incierta la utilidad de la metodología para detectar pequeñas roturas.

Por otro lado, Rajeswaran et al. (2018) propone un método que se enfoca en dividir el problema de la detección de fugas utilizando la partición gráfica y la resolución de subproblemas, utilizando el balance de agua. La partición gráfica determina que tuberías medir, de manera que la red se divide en subredes a las que aplicar el balance de flujo. Como criterio definieron que, si en flujo de entrada es igual al de salida, entonces no hay fuga. Asumen que la red de distribución se encuentra en un estado estable, que la topología de esta es conocida, que todos los caudales de oferta y demanda se miden continuamente y que las mediciones no poseen ruido. Dentro de las desventajas que tiene la metodología están que, debido a los supuestos adoptados, existe una alta probabilidad de falsos positivos y la necesidad de un gran número de sensores de flujo, además de que el método nunca fue validado en un sistema real.

2.4. Basado en modelos

Estos métodos implican el uso de funciones para representar o replicar el funcionamiento de una red de tuberías. Se puede determinar el lugar aproximado de la fuga comparando la medición de la presión con su estimación obtenida mediante el modelo de la red.

Perez et al. (2014) presenta una metodología que busca mejorar el aislamiento de las fallas con el uso de análisis de sensibilidad de éstas. La metodología consiste en:

i. la simulación de posibles fallas en los diferentes nodos de la red mediante un modelo hidráulico.
ii. generación de una matriz de sensibilidad de fugas.
iii. recolección de datos de presión obtenidos de los sensores instalados en la red.
iv. generación de residuos para comparar las mediciones de presión con un modelo sin fugas.
v. agregación y representación de resultados.
vi. señalar el nodo potencialmente defectuoso.

Dentro de las desventajas del método está que asumen que la fuga es en los nodos, siendo que lo más probable es que sea en las tuberías, también se requiere de muchas simulaciones y calibraciones por lo que el modelo es difícil de realizar.

Adedeji et al. (2017) propone un algoritmo de detección que incorpora un modelo de fugas en un modelo clásico de simulación hidráulica de una red de distribución. El proceso implica el análisis hidráulico de la red de distribución de agua y el cálculo de fugas. El algoritmo carga y lee los datos, realiza un análisis hidráulico mediante la modelación de la topología de la red de agua y la resolución del modelo resultante utilizando una metodología iterativa basada en el Método de Newton. El análisis hidráulico calcula el

caudal de fuga del nodo y comprueba si el caudal de fuga estimado del nodo es inferior a la tolerancia predefinida, si se confirma, informa que no hay fugas. En el caso contrario informa el número de nodos con fugas y busca todas las tuberías conectadas a esos nodos, en estas calcula el flujo de fuga de cada una y a partir de esto realiza la estimación de si la fuga es mayor o no a la tolerancia. Se etiquetan como tuberías críticas las que se encuentren sobre el rango de la tolerancia y recomienda un control de presión a lo largo de estas tuberías. Dentro de las desventajas se encuentra el desarrollo de los modelos hidráulico y de fugas, ya que estos deben ser muy precisos antes de ser utilizados. Por otro lado, esta metodología solo ha sido probada utilizando datos simulados.

2.5. Análisis estadístico

Esta metodología se basa en la teoría estadística para analizar los datos recopilados para identificar una fuga. Loureiro et al. (2016) realizó una metodología de cuatro pasos. Primero, se recopilan datos de flujo instantáneo (paso de tiempo menor a 15 minutos). Al disponer de ellos se recopilaron datos adicionales sobre las propiedades del medidor de flujo (diámetro, flujo mínimo y máximo), el sistema de adquisición y comunicación (frecuencia de lectura de y resolución) y las ordenes de trabajo (fecha de obra, fecha de reparación, tipo de obra) para validar los datos de flujo. Segundo, se validan, limpian y normalizan a un paso de tiempo regular los datos que implican la detección y corrección de anomalías. En las redes con múltiples medidores de flujo de entrada y salida, los conjuntos de datos normalizados se combinaron para estimar el consumo de la red. En las redes con grandes consumos continuos monitorizados, estos se normalizan y se extraen de las series temporales para permitir el enfoque en otros valores atípicos. Tercero, se utilizan métodos para detectar una anomalía en la serie temporal de flujos. Comentan que la categorización de los datos permite una detección más eficiente, ya que la mayoría de las redes muestran un comportamiento periódico durante el día. Se utilizó una curva ROC (Receiver Operating Characteristic) para determinar el umbral adecuado. Finalmente, se aplican métodos para detectar diferentes tipos de eventos atípicos. Las desventajas de la metodología es que no consideran los grandes consumidores o las demandas inesperadas de agua, por lo que, se pueden tener falsos positivos.

2.6. Basado en el análisis de señales transitorias

Se basa en la detección de la onda de presión transitoria. Zan et al. (2014) analiza en conjunto el tiempo y frecuencia de las señales transitorias de presión. Primero usa la transformación de ondas unidimensionales para eliminar el ruido de alta frecuencia de la señal transitoria de presión. Luego aplican la transformada rápida de Fourier de corta duración (STFT) para extraer las características de la fuga. Utilizan la Ventana Blackman para escoger la ventana óptima para la STFT y para minimizar el efecto del ruido en las señales de fuga. Calculan el espectrograma que muestra la energía de la señal y se utilizó la transformación de Gabor para eliminar una porción innecesaria del espectrograma. Después seleccionaron rangos de frecuencia óptimos para la detección de eventos

basándose en la interpretación del espectrograma. Calculan la intensidad media de la varianza y la desviación estándar de las distancias de la red utilizada y aplican una regresión lineal para aproximar la distancia entre los nodos del sensor y la fuga. Las desventajas o limitaciones del método es que la capacidad de detección depende de la resolución espectral del STFT y la localización depende de la resolución temporal. El ruido u otros pueden influir, además presenta un alto costo de inversión para la instalación de sensores y no es posible detectar fugas lentas o pequeñas, ya que no generan una señal distintiva. Mediante las transformaciones es posible que una señal pequeña cause un falso positivo. Sólo puede localizar la posible zona de fuga si los puntos de medición están lejos y como entrega una zona de fuga, por lo que, es necesaria una inspección adicional para la localización exacta.

Por otro lado, Kim et al. (2016) propone un método que solo utiliza la presión para superar la falta de robustez de los métodos en condiciones ruidosas, la baja precisión causada por la pérdida de la información en el tiempo y la ausencia de un límite de confianza. Primero, es medida la presión mediante un modelo hidráulico y un Flitro de Kalma. Segundo, se aplica una función de parte entera con tres parámetros y una función de curvatura, una vez obtenido el tiempo de ocurrencia, se aplican técnicas estadísticas para encontrar el segmento que contiene el punto de fuga con una confianza limitada. El algoritmo que proponen se valida utilizando las pruebas de detección de fugas y las pruebas de falsas alarmas. Obtienen una baja tasa de falsa alarma, pero no consideran la demanda inesperada de agua, por lo que resultará un falso positivo. Dentro de las desventajas se encuentra el requerimiento mínimo de 4 sensores para que la localización sea posible en una sola tubería, además la metodología solo fue validada utilizando datos modificados.

Por otra parte, Cristodoulou et al. (2017) proponen un enfoque basado en la vigilancia continua de las características de vibración de las tuberías para extraer un índice que sea sensible al inicio de las fugas e insensible a los ruidos del ambiente. Utilizan series cronológicas de consumo de agua a nivel macro para identificar los periodos de tiempo de interés y luego a nivel micro se acercan a las posibles anomalías. Se aplica la detección de anomalías, las que se identifican como la producción de uno o varios cambios en una serie temporal. Los métodos de detección de estos cambios se relacionan con la variación de la media, varianza, correlación, densidad o la pendiente de la señal.

2.7. Metodologías por considerar para el trabajo

La metodología propuesta para el desarrollo de este trabajo se sustenta en otras realizadas en el marco de estudios que se orientaron hacia objetivos de la misma índole que el de éste.

Dentro de estos se encuentra la metodología propuesta por Salam et al. (2014), la que se basa en la utilización de los cambios de presión y velocidad que se manifiestan al ocurrir una fuga. Para esto, se modela una red de tuberías mediante el software EPANET

2.0 para el comportamiento “normal” (sin fugas) de una red de distribución, para diferentes presiones y demandas. Genera las fugas mediante nodos seleccionados como puntos de fuga, además, mediante la herramienta de emisor del software, se modelan las fugas en las tuberías. Se utilizan fugas de diferentes caudales, obteniendo un mayor número de datos para el entrenamiento de la SVM. Los datos de entrada de presión y velocidad de cada nodo y tubería, más los datos de salida (tamaño y localización de fuga), serán procesados por el núcleo de la Función de Base Radial (RBF) con el propósito de determinar el núcleo de ésta para facilitar el proceso de aprendizaje del SVM y determinar el vector de apoyo. Se tiene en cuenta esta metodología, ya que obtiene un promedio de precisión para la ubicación de la fuga del 76,14% y para el tamaño de esta de 86,68%.

Un estudio basado en el trabajo de Salam es el realizado para la red de distribución de agua de Tsumeb East por Kemba et al. (2017) en el que se utiliza las SVM para realizar las tareas de clasificación binaria, ya que son superiores a las redes neuronales artificiales (RNA) para realizar esta labor. Las SVM son menos sensibles a factores como la falta de generalización y la convergencia incontrolable, además de poder operar con éxito en espacios de entrada de alta dimensionalidad y son capaces de tratar con pequeñas muestras de entrenamiento y pruebas. Las SVM pueden establecerse asignando pocos parámetros, como la función del núcleo y el parámetro C de penalización. Modelan la red de distribución de 140 casas y una demanda mensual de 8,3 m3 por hogar, un tanque con un volumen de 17117 m3, 96 tuberías y 80 nodos. Se escogen 15 nodos y se generan 40 conjuntos de fugas, lo que implica 600 escenarios, que se dividieron en 540 conjuntos de entrenamiento y 60 conjuntos de prueba. Utilizan la RBF (Función de Base Radial) como función del núcleo para la SVM. Realizan una calibración de los parámetros γ de la RBF y C. Se experimenta con C entre 0,03 y 16.384, y γ entre 0,00005 y 16 para encontrar los mejores parámetros para este problema. Los que le resultaron en C = 128 y γ = 0,1 para la precisión más alta y C = 0,003 y γ = 0,00003 para la más baja.

Otro estudio es el ya nombrado de Kang et al. (2017) de arquitectura híbrida de CNN y SVM, el que aplica clasificadores duales heterogéneos (híbridos CNN-SVM). Además los mapas de características de la CNN se utilizan como entrada para cada clasificador. Se basa en las probabilidades de las clases de cada clasificador y el resultado final se deriva utilizando un esquema de combinación de probabilidades. El algoritmo de localización requiere información de la velocidad de onda (que depende del material), la posición del sensor, la frecuencia y la diferencia de tiempo de llegada. Proponen un algoritmo de búsqueda local basado en gráficos que puede determinar eficazmente la posición correspondiente de las fugas en redes de tuberías, utilizando un esquema de nodos virtuales para estimar el lugar de fuga más cercano, además de minimizar los errores de distancia. Concluye que, con una tasa de muestreo adecuada (alrededor de los 2000), señales preprocesadas y un conjunto de clasificadores duales heterogéneos capaces de reflejar las características de la señal objetivo a ser analizada, la precisión de la clasificación puede ser mejorada, en comparación con la extracción de características normales (extracción de información de cambios en información de entrada debido a las fugas) o con la arquitectura convencional de Deep Learning. El clasificador debe ser más robusto en cuanto a las variaciones de las condiciones de fuga, como la presión de la fuga, la cantidad de fuga y otras operaciones de mantenimiento.

En cuanto a la calibración de los parámetros de la SVM se revisa el estudio realizado por Mashford et al. (2009), donde utilizan EPANET con su diseño de grifos de incendio para modelar las fugas, el cual depende de un coeficiente de emisión. Desarrollan un método de análisis del SVM:

i. Determinan cuan eficazmente es una SVM utilizada como regresor para predecir los valores del coeficiente de emisión cuando un determinado nodo fijo tiene fugas. Varía el coeficiente de 0 a 0,3 con pasos de 0,001 generando 300 casos, los que se dividen entre 200 y 100 para entrenamiento y pruebas, respectivamente. Utilizan un núcleo de RFD para la SVM.
ii. Determinan la efectividad de usar un clasificador SVM para la localización de fuga. Considera 10 nodos de posible fuga. Fugas con coeficiente de 0 a 0,3 con paso de 0,002. Realiza 10 conjunto de datos de 150 casos cada uno, generando 1000 y 500 de entrenamiento y pruebas, respectivamente. Ajusta los parámetros de la SMV alcanzando precisiones de 76,8%.
iii. Utilizan otros 30 nodos de posible fuga, generando 6000 casos, con 4000 y 2000 de entrenamiento y prueba respectivamente, alcanzando una precisión de 57,2%.
iv. Aunque el sistema no predice correctamente el nodo de fuga real, en el 42,75% de los casos le interesa saber si el nodo de fuga previsto está cerca del real. El nodo previsto se encuentra a menos de 100 metros del nodo real en el 77,5% de los casos y a menos de 200 metros el 97,7% de los casos.

Mashford et al. (2009) realizan el método para bajas tasas de fugas, en el que establecen las limitaciones del procedimiento de detección de fugas pequeñas (menores a 100 l/hora). La sensibilidad de EPANET no fue suficiente para registrar pequeñas diferencias de presión, si no hasta que se tiene un coeficiente de emisión mayor a 0,0025, lo que equivale a una emisión de 90 l/hora. El análisis demostró que la predicción de la ubicación exacta tenía un índice de éxito del 35%. Se obtiene una tasa de éxito del 100% en 500 metros de diferencia entre el previsto y el real. La precisión a distancias más cortas podría mejorarse, pero esto sería a expensas de un mayor volumen de fugas. Con un coeficiente de emisión de 0,0005 produjeron una tasa de éxito del 56% para la predicción de la ubicación exacta.

En cuanto a los aspectos prácticos del método, depende de la vigilancia de los cambios de presión que son trazables en las fugas. Es necesario vigilar los cambios de presión entre 0,1 m y 0,7 m para detectar fugas en el rango de los 90 l/hora (coeficiente de emisión 0,0025). Se llegó al límite de EPANET en cuanto a los niveles de fuga más bajos, pero se puede disponer de otras herramientas de simulación hidráulica que proporcionen datos a nivel de fugas más bajos.

Lang et al. (2017) propone identificar las fugas mediante TWSVM (maquinas vectoriales de apoyo gemelas), las que generan 2 hiperplanos no paralelos, resolviendo 2 problemas de programación cuadrática (QPP) de menor tamaño, de modo que cada hiperplano esté más cerca de una clase y lo más lejos posible de la otra. Esto hace que el aprendizaje del TWSVM sea 4 veces más rápido que el del SVM tradicional. Para la extracción de características se generan múltiples datos experimentales, además de incluir un ruido

como se produciría en condiciones reales. Las señales se descomponen mediante descomposición media local (LMD) y luego las señales reconstruidas se desacoplan en base al análisis de ondas de presión. Se seleccionan las 4 capas de la reconstrucción de baja frecuencia. De esta manera, el cambio de la forma de la onda de presión es beneficioso para extraer las características y localizar el punto de fuga.

3. Marco teórico

En esta sección, se presentan antecedentes teóricos en los que se basa a este trabajo de título. En la primera parte se presentan algunos de los conceptos de pérdidas en redes de distribución de agua. Posteriormente, se entrega información sobre los modelos hidráulicos y la calibración de estos. Luego, una breve introducción al aprendizaje automático, con énfasis en los problemas de clasificación existentes. Finalmente, se introduce información sobre las máquinas de vector de apoyo y se plantea la hipótesis sobre la cual se trabajará

3.1. Conceptos de pérdidas en redes de distribución de agua

Para la internalización de conceptos de pérdidas en las redes se realiza una revisión a la Guía para la reducción de las pérdidas de agua de Ziegler et al (2011). de donde se extrae la siguiente información sobre las aguas no facturadas y las pérdidas

Las pérdidas de agua reales y aparentes, y el consumo no facturado conforman las aguas no facturadas (ANF) en las redes de distribución. Para una inspección rápida de estas, se realiza un balance hídrico que debe constar con los siguientes elementos:

i. Volumen de agua que ingresa al sistema.
ii. Consumo autorizado: en el que se incluyen los consumos autorizados facturados (medido facturado, no medido facturado, agua exportada).
iii. Agua facturada: volumen de agua que se entrega y factura al cliente de manera exitosa, el que genera ingresos para la empresa distribuidora.
iv. Agua no facturada: volumen que no genera ingreso para la empresa. Se puede expresar como la diferencia entre el volumen de ingreso y el de consumo autorizado facturado o como la suma de las pérdidas y el consumo autorizado no facturado.
v. Pérdidas de agua: volumen entre el punto de suministro y el medidor del cliente. Puede expresarse como la diferencia entre el volumen de ingreso al sistema y el consumo autorizado o por la suma de las perdidas aparentes y reales.

3.1.1. Pérdidas reales

Corresponden a los volúmenes de agua perdidos dentro de un determinado periodo de tiempo a través de los distintos tipos de fugas, roturas o rebose. Estas se pueden clasificar según su ubicación dentro del sistema y por su tamaño y tiempo de fuga.

La ubicación de la fuga se puede dividir en:

- Fugas desde las troncales de transmisión y distribución, las que pueden ocurrir en tuberías, uniones y válvulas, y usualmente con tasas de flujo de medianas a altas y tiempos de fuga cortos a medianos.

- Fugas desde conexiones de servicio hasta el punto del medidor, trayectos que son puntos débiles de las redes, ya que sus uniones y accesorios poseen altas tasas de fallas. Existe gran dificultad para su detección ya que las tasas de flujo son bajas, por lo que, tienen un tiempo de fuga largo.
- Fugas y rebases de estanques de almacenamiento, causadas por deficiencias o daños en los controles del nivel del tanque. Son de fácil detección, pero la reparación es complicada y de alto costo.

El tamaño y tiempo de fuga se dividen en:

- Fugas visibles: provienen generalmente de roturas de uniones en grandes tuberías de distribución. El agua aparece en la superficie rápidamente dependiendo de la presión del agua, del tamaño de la fuga y de las características del suelo.
- Fugas ocultas o no reportadas: poseen caudales mayores a 250 l/hora a 50 m de presión, pero debido a las condiciones no favorables no aparecen en la superficie. Se pueden identificar analizando tendencias en el comportamiento del consumo dentro de una zona definida.
- Fugas de fondo: Fugas muy pequeñas con caudales menores a 250 l/hora a 50 m de presión que no se pueden detectar mediante métodos acústicos, por lo que, se asume que muchas fugas de fondo no son detectadas ni reparadas hasta que se reemplaza un elemento de la parte defectuosa. Son el aporte principal para las perdidas reales de agua debido a las altas cantidades y el largo tiempo durante el que ocurren.

3.1.2. Pérdidas aparentes

No se deben a fugas físicas en la infraestructura, ya las causan otros factores tales como:

- Inexactitud de medición debido a contadores incorrectos de agua de los clientes o medidores de flujo incorrectos.
- Manejo de datos y errores de contabilidad, así como una mala rendición de cuentas de los clientes en los sistemas de facturación.
- Consumos no autorizados debido a robos de agua o a conexiones ilegales.

3.2. Modelos hidráulicos de redes de agua potable y su calibración

Los modelos hidráulicos se definen como la representación matemática de los elementos existentes en una red y según Coelho et al. (2006) estos están compuestos por:

- Un conjunto de datos descriptivos de las características físicas del sistema, de los consumos y de las condiciones de operación.
- Un conjunto de ecuaciones matemáticas que reproducen el comportamiento hidráulico de cada elemento del sistema.

- Algoritmos numéricos capaces de resolver el conjunto de ecuaciones del sistema.

Pueden ser clasificados en estáticos y dinámicos. Los primeros simulan el estado de la red en un único instante, mientras que los segundos simulan el comportamiento a lo largo del tiempo.

Según Coelho et al. (2006) las fugas, al ser un fenómeno hidráulico, pueden ser modeladas como una demanda dependiente de la presión. El enfoque más simple y utilizado para la modelación de fugas consiste en utilizar la siguiente ecuación de emisores de flujo:

$$q_{F,j} = K_j (P_j)^N$$

La que permite simular el flujo de salida a través de una tobera u orificio descargando a la atmósfera. Donde $q_{F,j}$ es el caudal de fuga en el nodo j, K_j es el coeficiente emisor del nodo j y depende del tamaño y la forma del orificio de fuga P_j es la presión de nodo j y N es el exponente de fugas. Varios autores han propuesto variaciones y mejoras a la ecuación teniendo en cuenta diferentes consideraciones.

Para la calibración de modelos hidráulicos para redes de distribución de agua se tiene en cuenta el procedimiento propuesto por Ormsbee y Lingireddy (2002) y consta de los siguientes pasos:

- Identificar el propósito del modelo: esto nos proporciona una guía de los requerimientos al nivel del detalle del modelo, tipo y calidad de los datos de campos a ser recolectados, además de la tolerancia que debe superar el error entre los datos observados y simulados.
- Determinar el valor inicial de los parámetros a estimar: valores básicos del modelo, como la rugosidad en tuberías, las demandas de cada nodo y su factor de modulación para las simulaciones de periodo extendido.
- Recolectar información de calibración: Información que permite evaluar los resultados preliminares del modelo, pueden ser recolectados mediante pruebas de campo, datos de telemetría y datos referentes a pruebas con trazadores químicos conservativos.
- Evaluar los resultados del modelo: esto para valorar la precisión del mismo, comparando los resultados de este con las mediciones realizadas.
- Realizar una calibración a nivel macro: se enfoca en la calibración de toda la red de distribución y consiste en identificar y corregir las fuentes de error que proporcionan las diferencias más significativas entre las mediciones y los resultados del modelo.
- Realizar un análisis de sensibilidad: se varían los parámetros del modelo para cuantificar el efecto sobre los resultados e identificar cuales generan un mayor impacto sobre estos.

- Realizar una calibración a nivel micro: se enfoca en la precisión de un área en particular del modelo. Los parámetros para ajustar en esta etapa son generalmente los coeficientes de rugosidad de las tuberías.

3.3. Aprendizaje Automático

El aprendizaje automático puede definirse como un junto de métodos que son capaces de detectar automáticamente patrones y estructuras en los datos, para luego utilizar estos patrones y estructuras características aprendidas para realizar predicciones o tareas de toma de decisiones, como la clasificación por inteligencia artificial en el marco de este trabajo. El aprendizaje automático es un subconjunto del campo de la inteligencia artificial, que se basa en gran medida en la teoría de la probabilidad, las técnicas estadísticas y las ciencias de la computación. Los modelos intentan aprender de los datos, por lo que, no es necesaria una programación explicita para resolver problemas complejos. Los tipos de problemas que se pueden resolver con el aprendizaje automático se pueden dividir en tres categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje de refuerzo (Murphy, 2012).

- Aprendizaje Supervisado: el objetivo principal de este tipo de problemas es aprender mediante un mapeo a las entradas (x) y salidas (y), dado un conjunto de datos que contiene un conjunto etiquetado de N pares de entrada-salida $D = \{(x_i, y_i)\}_{i=1}^{N}$. Todo el proceso se basa en encontrar la relación $f : x \rightarrow y$ que minimiza la discrepancia entre las salidas deseadas y las salidas que el modelo produce. La salida y_i depende del tipo de problema que se resuelve. Para problemas de clasificación, el objetivo es asignar a cada x una o más clases de un conjunto $C = \{c_1, c_2, \dots c_k\}$.
- Aprendizaje No Supervisado: el objetivo principal es encontrar patrones o estructuras definidas dentro de los datos. La base solo consiste en un conjunto de entradas $D = \{x_i\}_{i=1}^{N}$.
- Aprendizaje por Refuerzo: Se centra en los procesos de aprendizajes reglamentados, en los que se entregan algoritmos de aprendizaje automático con un conjunto de acciones, parámetros y valores finales. Mediante la definición de las reglas, el algoritmo intenta explorar diferentes opciones y posibilidades, monitoreando y evaluando los resultados para determinar el óptimo. Este sistema enseña mediante el ensayo y error.

3.4. Problemas de Clasificación con Inteligencia Artificial

Dentro de los problemas de clasificación con Inteligencia Artificial, existen principalmente cuatro categorías que dependen del número y la estructura de las clases presentes en el problema (Sokolova y Lapalme, 2009):

- Clasificación Binaria: aquí la entrada debe clasificarse en una de dos clases no superpuestas.
- Clasificación de Clases Múltiples: el objetivo aquí es clasificar cada entrada como una clase dentro de un conjunto k de clases no superpuestas.
- Clasificación con Etiquetas Múltiples: aquí el objetivo es clasificar cada entrada en una o varias clases no superpuestas.
- Clasificación jerárquica: aquí la entrada debe clasificarse en una clase que podría dividirse en subclases o agruparse en diferentes superclases.

3.5. Support Vector Machine

Para conocer la teoría detrás del funcionamiento de las SVM se revisa el trabajo de Nefedov (2016). El que parte explicando el problema de clasificación, donde existe un gran conjunto de objetos, que puede clasificarse en dos clases diferentes Y_i e Y_{i+1}. Usando los objetos del conjunto se intenta definir un algoritmo que los clasifique a todos con el mínimo error. En la mayoría de las aplicaciones prácticas el espacio vectorial a utilizar (E) es el espacio de coordenadas reales (R^n). En este espacio se tiene un vector o una matriz X que es el conjunto de m números reales x_{ij} (donde j corresponde al número de columna de la matriz y el total de columnas de la matriz corresponde a el número de dimensiones de ésta “n”), además, para cada x_{ij} se tiene un y_i con los que se generan muestras de entrenamiento. La función de decisión se basa en una que clasifique los vectores X como clases que se diferencian por la salida (Y).

El hiperplano es un conjunto de vectores de n-1 dimensiones que satisface cierta ecuación, el que divide el espacio coordenado en R^n en dos partes ubicada a los lados del hiperplano, llamadas semiespacio positivo y negativo. Se denomina vector de apoyo a los vectores que maximizan el margen de las muestras con el hiperplano.

El núcleo del SVM es una función que se utiliza en el problema de clasificación y en la función de decisión en lugar del producto interno. El uso de los núcleos permite definir que en lugar de hiperplanos utilizan una clase mucho más amplia de superficies de separación, por lo que, utilizando superficies no lineales es posible separar de mejor manera las clases dadas que utilizando un hiperplano.

Para un mejor entendimiento del funcionamiento de las SVM en la práctica se revisa el trabajo de Chih-Wei et al. (2003) donde propone un procedimiento para evitar cometer errores al no saber aplicar bien las SVM. El procedimiento consiste en:

- Transformar los datos al formato de un paquete SVM: requiere que los datos se representen como un vector de números reales. Por lo que, si hay atributos categóricos, primero hay que convertirlos a datos numéricos.
- Llevar a cabo un simple escalado de los datos: la principal ventaja del escalamiento es evitar que los atributos en rangos numéricos mayores predominen

a los de rango menores, además evita dificultades numéricas durante el cálculo. Recomienda escalar linealmente cada atributo en el rango [-1,1] ó [0,1].

- Considerar los diferentes núcleos (recomienda el núcleo de función radial RBF).
- Utilizar validación cruzada para encontrar los mejores parámetros C y γ: se divide el conjunto de entrenamiento en subconjuntos de igual tamaño para probarlos usando el clasificador, entrenado todos menos uno, repitiendo eso con cada subconjunto, de modo que la precisión de la validación cruzada es el porcentaje d datos que se clasificaron correctamente.
- Usar los parámetros que mostraron mejor resultados para entrenar a todo el conjunto de entrenamientos.
- Realizar las pruebas para generar el clasificador final.

Para el desarrollo de las SVM se utilizará PYTHON por lo que se revisa el trabajo de Lee (2019) Python machine learning, específicamente el Capítulo 8 Supervised Learning – Classification Using Support Vector Machines, en el que plasma una manera simple de visualizar la SMV consiste en trazar una línea entre dos o más clases de la mejor manera posible. Para poder predecir datos futuros con la definición de esta línea. Separa las clases en función de los márgenes más amplios, es decir, encontrar la mayor amplitud posible que pueda separar los dos posibles grupos. En la figura 3.1 se observa como el mejor hiperplano, considerando el de a y b, correspondería a b, ya que el margen que separa el hiperplano de b del punto más cercano es mayor que el margen del hiperplano de a (d2>d1).

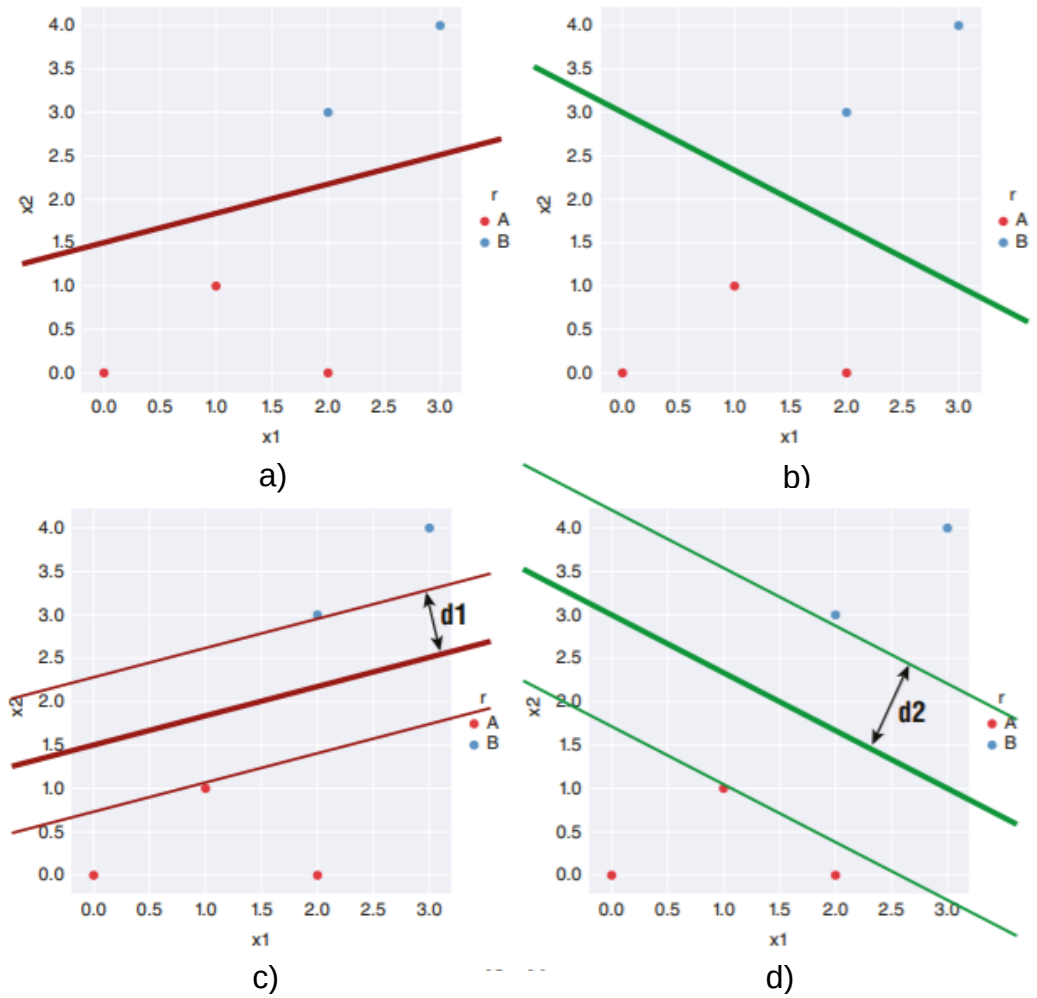

Figura 3.1: Ejemplo de hiperplano con un mayor margen. Fuente: Lee (2019).

Lee (2019) describe los puntos que se encuentran en los márgenes, si esos puntos estuviesen más lejos o cerca del hiperplano, esto definiría que el margen es mayor o menor respectivamente. Para encontrar los márgenes utiliza la biblioteca Scikit-learn de Python, que contiene el módulo de SVM, el que además tiene una serie de clases que implementan SVM para diferentes propósitos:

- svm.LinearSVC: Linear Support Vector Classification
- svm.LinearSVR: Linear Support Vector Regression
- svm.NuSVC: Nu-Support Vector Classification
- svm.NuSVR: Nu-Support Vector Regression
- svm.OneClassSVM: Unsupervised Outlier Detection
- svm.SVC: C-Support Vector Classification (Función a utilizar en el trabajo de título)
- svm.SVR: Epsilon-Support Vector Regression

Como por lo general los conjuntos de datos no son siempre linealmente separables, como lo muestra la Figura 3.2:

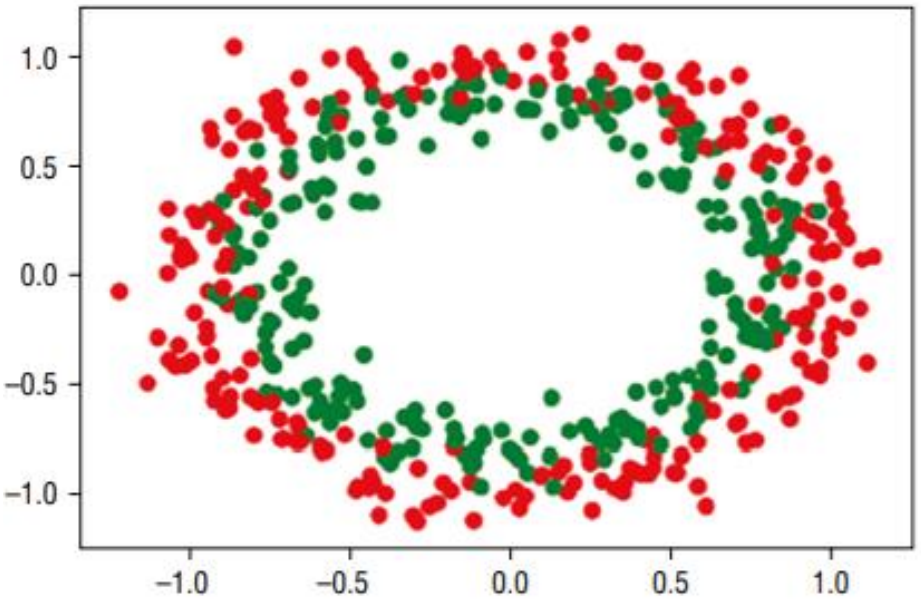

Figura 3.2: Conjunto de datos no linealmente separables. Fuente: Lee (2019).

Se manipula el conjunto de puntos para que sea linealmente separable mediante la técnica conocida como kernel trick, técnica de aprendizaje automático que transforma los datos en un espacio de dimensiones superiores para que después de la transformación se tenga un claro margen que divida las clases de datos. Si se añade una tercera dimensión a los datos mostrados en Figura 3.2 estos serán linealmente separables como lo muestra la Figura 3.3:

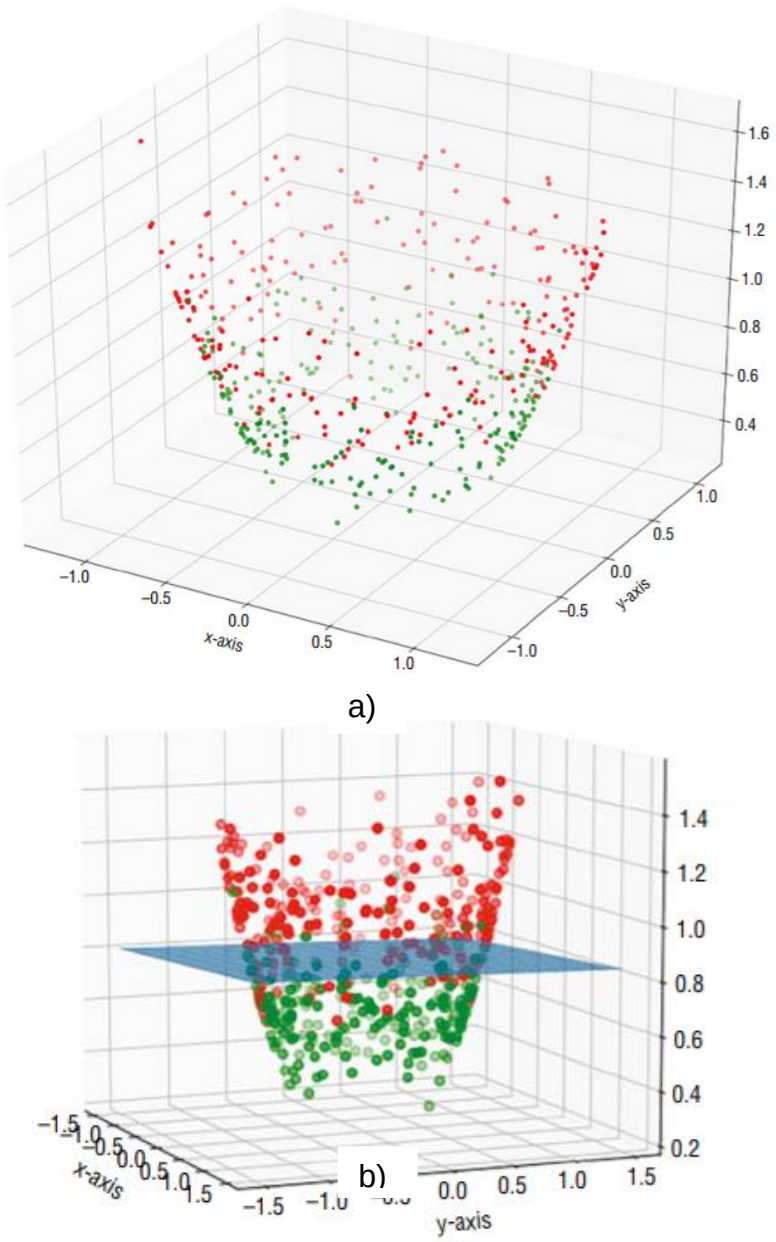

Figura 3.3: Transformación de datos a un espacio dimensional superior. Fuente: Lee (2019).

No siempre se puede utilizar este kernel trick, por lo que, es necesaria la utilización de las funciones de kernel para transformar los datos de espacios no lineales a espacios lineales, como el que muestra la Figura 3.4:

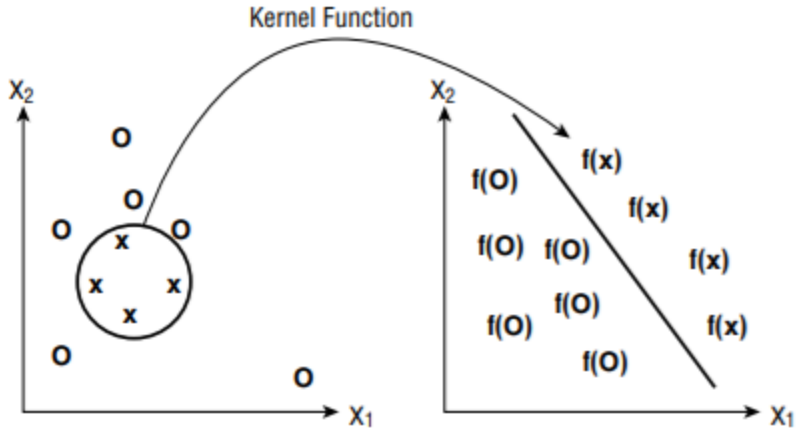

Figura 3.4: Utilización de función kernel para generar espacios lineales. Fuente: Lee (2019)

Existen diferentes funciones de kernel para la clasificación como los que se muestran en las Figuras 3.5 y 3.6 (lineal, polinómico y de base radial).

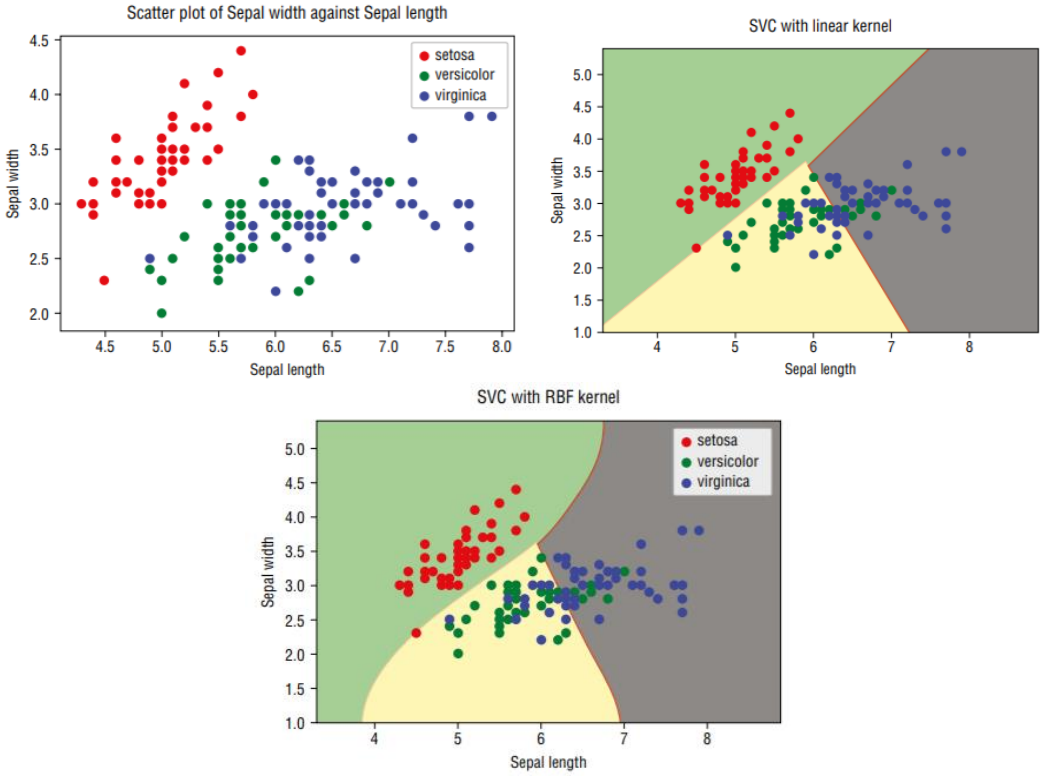

Figura 3.5: Diferentes kernels para la clasificación. Fuente: Lee (2019).

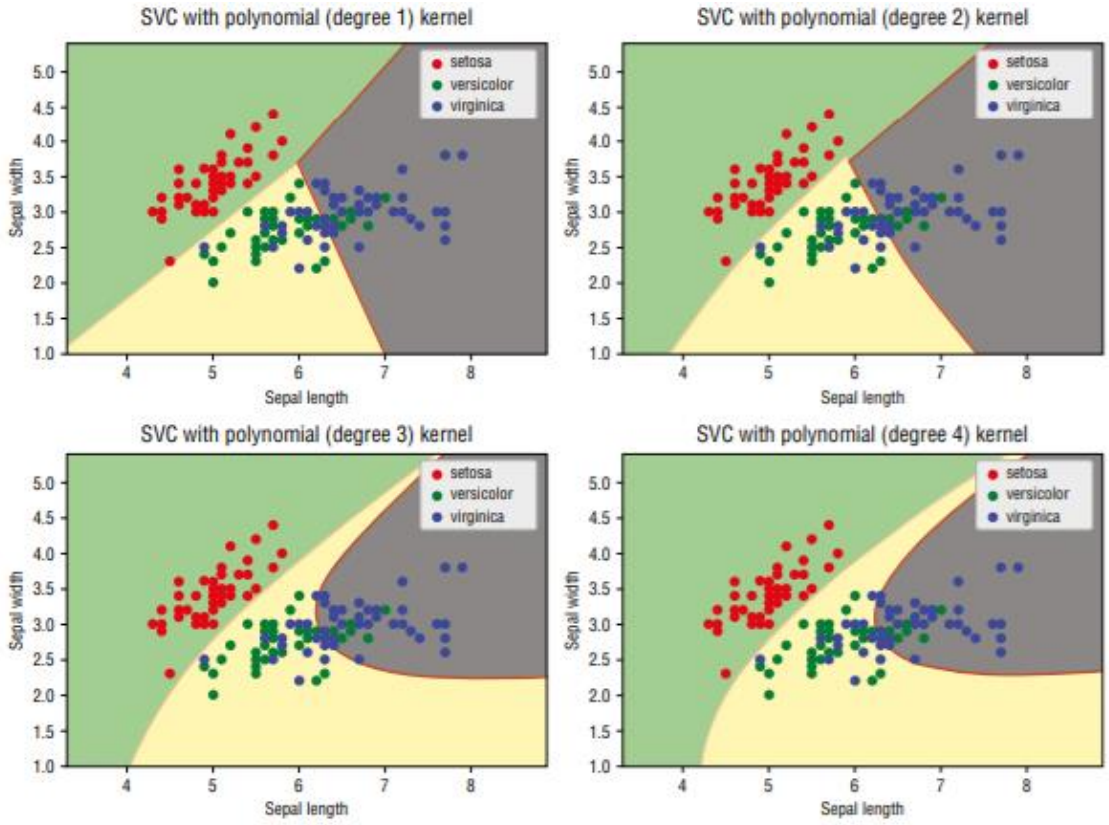

Figura 3.6: Diferentes kernels para la clasificación. Fuente: Lee (2019).

3.6. Hipótesis

De la revisión realizada se puede extraer que las metodologías existentes para la detección y localización de fugas son poco precisas y en muchos casos de un coste bastante alto, tanto por la necesidad de gran cantidad de equipo, de personas que manejen el equipo con la suficiente experiencia o de personas que manejen la estadística o los modelos hidráulicos de buena forma.

Por este motivo se plantea como hipótesis la posibilidad de detección y localización de fugas en una red de distribución mediante la utilización un modelo hidráulico de la red, construido con el software OpenFlows Watergems de Bentley, el que entregará como salida las presiones a lo largo de un periodo de 24 h en distintos puntos de la red de distribución a utilizar. Los que serán la entrada para el algoritmo de aprendizaje supervisado que buscará clasificar los comportamientos de la red por la existencia o no de una fuga y localizar el punto de fuga en caso de su existencia. Se utilizarán diferentes SVM entrenado con los diferentes ordenamientos que se les dé a los datos de salida del modelo hidráulico.

4. Modelo hidráulico de distribución de Agua Potable y su calibración

4.1. Modelación base y datos de entrada iniciales

El modelo hidráulico de distribución de Agua Potable con el que se inicia el trabajo pertenece a 4 comunas de la ciudad a utilizar en el trabajo las Figuras 4.7 y 4.8.

Figura 4.1: Área de distribución del modelo inicial.

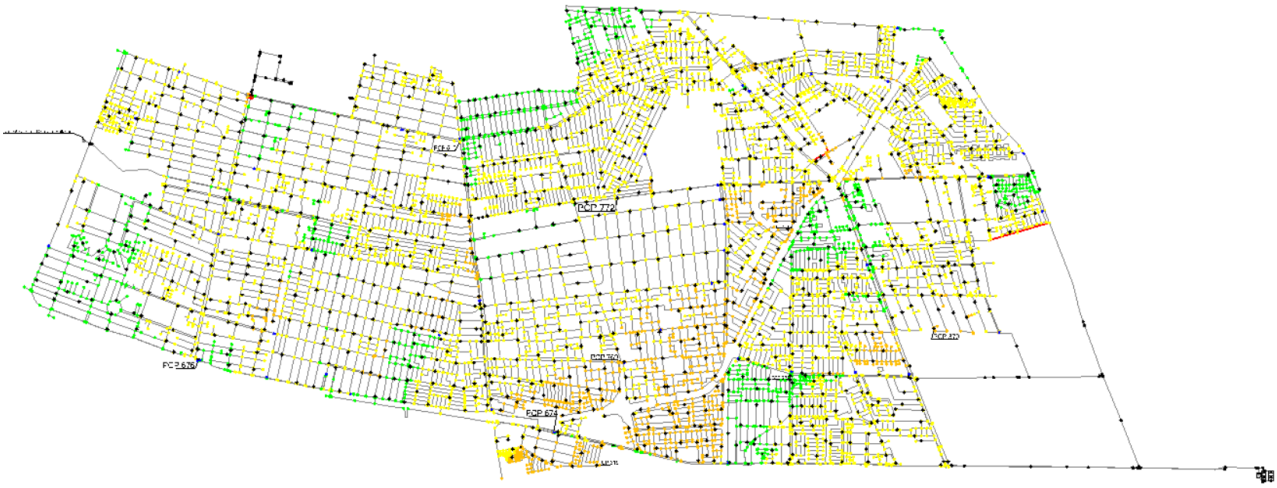

Figura 4.2: Red de tuberías de agua potable inicial.

Se aísla el sector con el que se trabajará eliminando los elementos constituyentes a la red de distribución de las comunas del sector derecho de las Figuras 4.1 y 4.2, quedando

la red que llega a las comunas del sector izquierdo de las Figuras 4.1 y 4.2, como lo muestra la Figura 4.3.

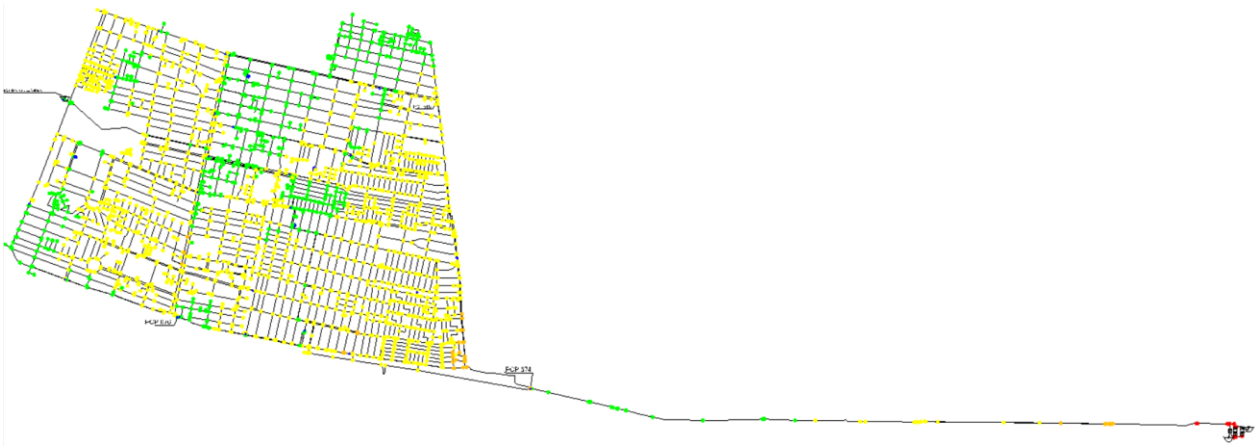

Figura 4.3: Sistema de distribución del sector a modelar.

Tras aislar la zona se procede a verificar los edificios actuales en el sector. Eliminando los que ya no existen y agregando los existentes, los que se presentan en la Figura 4.4, siendo edificios con arranque de 50 mm los de color rojo, edificios con arranque de 75 mm los de color azul y edificios de arranque de 150 mm los de color verde.

Figura 4.4: Edificios existentes en la zona.

El sistema de distribución queda con un total de 2.533 nodos, de los cuales 82 son edificios. Con una longitud de red de 232.783 m, 27 válvulas reductoras de presión (PVR) y 69 válvulas controladoras de flujo (FCV). En las Tablas 4.1, 4.2 y 4.3 se presentan las características del sistema de distribución.

Tabla 4.1: Elementos del sistema de distribución a modelar.

Elemento		
Nodos	n°	2533
Tuberías	m	232.167
PVR	n°	27
FCV	n°	69

Tabla 4.2: Materialidad de la red de distribución del sistema.

	Longitud (m)
Asbesto Cemento (AC)	182.715,9
Acero	5.653,9
Concreto	1.093
Fierro Fundido (FFD)	4.507,1
Hormigón	7.588,4
Polietileno de Alta Densidad (HDPE)	28.266
Policloruro de Vinilo (PVC)	2.342,7
Total general	232.167

Tabla 4.3: Caudales del sector

	[l/s]
Caudal máximo horario	1.193,55
Caudal máximo diario	795,70
Caudal mínimo horario	362,88

4.2. Transformar un modelo de estado estático a uno de estado dinámico

Un modelo en estado dinámico es obtenido basándose en la información de flujo de salida de los estanques que abastecen el sector, del que se tiene información, como la que se muestra en la Figura 4.5, para todos los meses de los años 2017, 2018 y 2019.

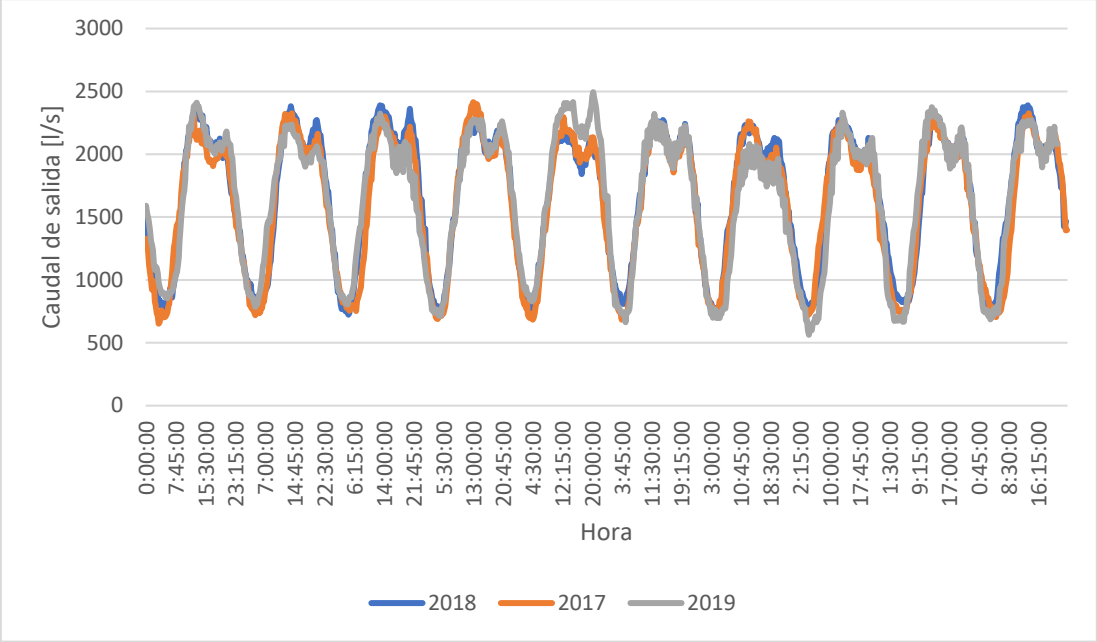

Figura 4.5: Información de caudales de salida 10 días del mes de febrero de los años 2017, 2018 y 2019

Observando en la Figura 4.5 el patrón repetitivo para los días del febrero para los diferentes años de análisis se decide trabajar con el mes de agosto debido a que se cuenta con toda la información de presiones del sector de este mes del 2020. Dividiendo el mes en 4 semanas, con sus respectivos fines de semana, se obtiene el caudal de demanda promedio esperado para la semana y el fin de semana del mes de agosto de los años 2017, 2018 y 2019. El que es mostrado en las Figuras 4.6 y 4.7.

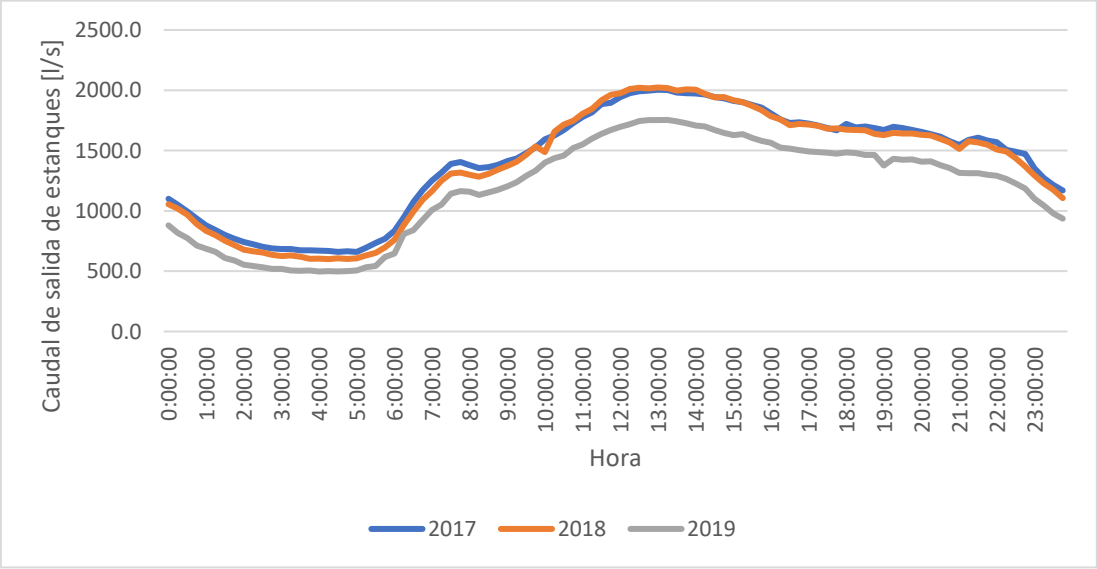

Figura 4.6: Caudal de demanda promedio para los lunes, martes, miércoles, jueves y viernes de agosto de 2017, 2018 y 2019.

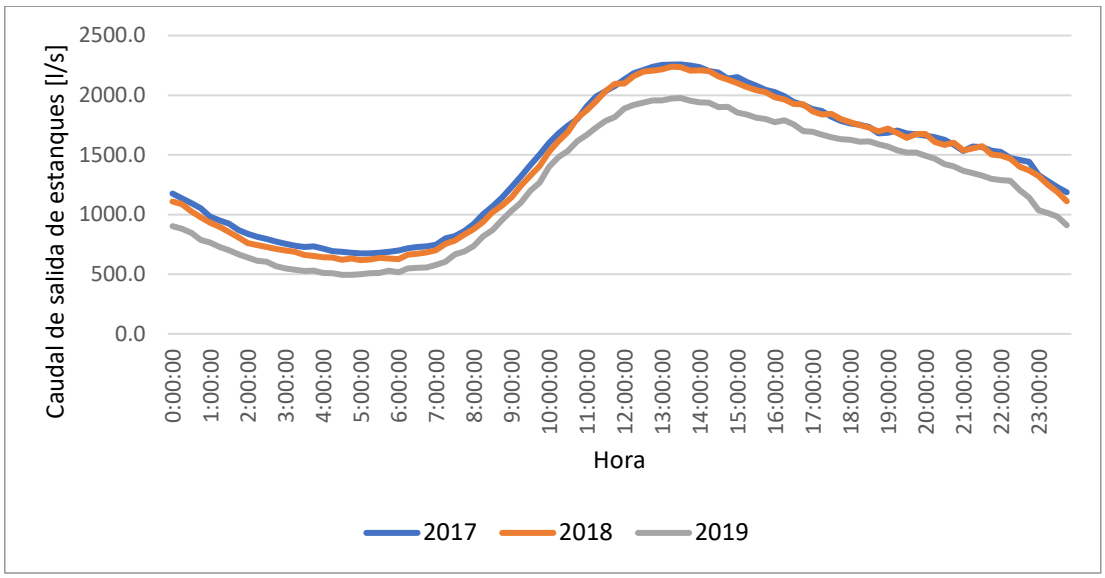

Figura 4.7: Caudal de demanda promedio para los sábados y domingos de agosto de 2017, 2018 y 2019.

De la Figura 4.6 y 4.7 se observa que el caudal para el mes de agosto año 2019 es aproximadamente entre 200 y 250 l/s menor que los caudales de agosto de los años años 2017 y 2018, esto podría deberse a razones como la reparación de alguna de rotura de la red o a un aumento de la eficiencia en el uso del agua ya sea en las aguas domiciliarias o en las de uso no domiciliario.

Normalizando los caudales por el máximo se obtiene el patrón de consumo de agua del sector para la semana y para el fin de semana para los años 2017, 2018 y 2019, promediando los patrones de los 3 años se obtiene el patrón promedio para la semana y fin de semana, que se utilizará como el esperado para agosto del 2020. Patrones que son mostrados en la Figura 4.8.

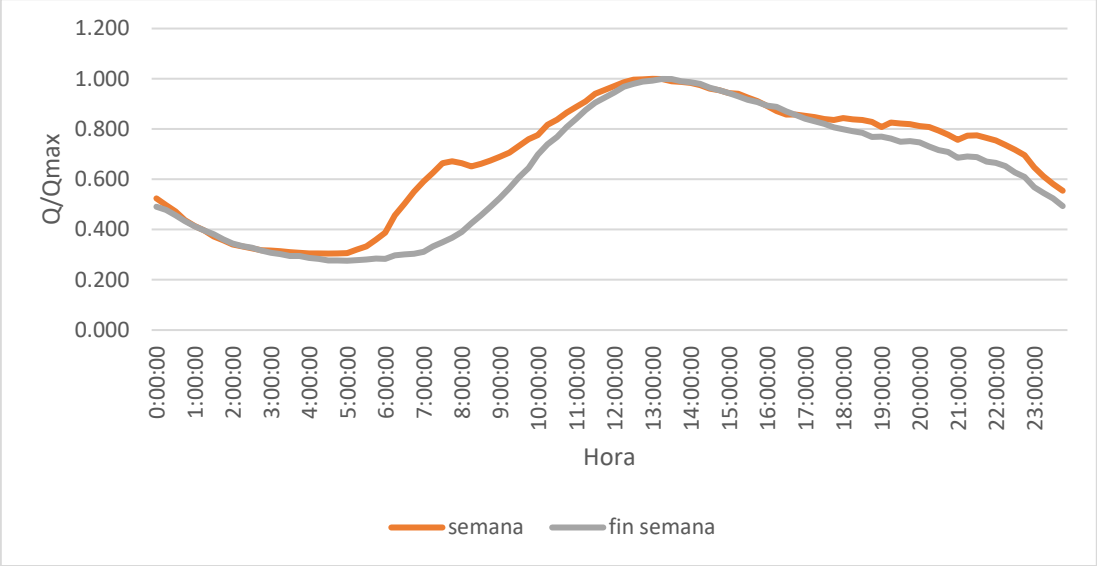

Figura 4.8: Patrón de consumo promedio para un día de la semana y para un día del fin de semana de septiembre.

Además del patrón de consumo, se obtiene el patrón diario para las válvulas reguladoras de presión existentes en el sistema. La información obtenida es un patrón en función de la menor presión observada en la Válvula Reductora de Presión (PRV) para agosto del 2020. En la Figura 4.9 se presenta el patrón tipo de la presión de salida de la PRV, en el caso de la Figura 4.9 es para la válvula ubicada en el punto 19.

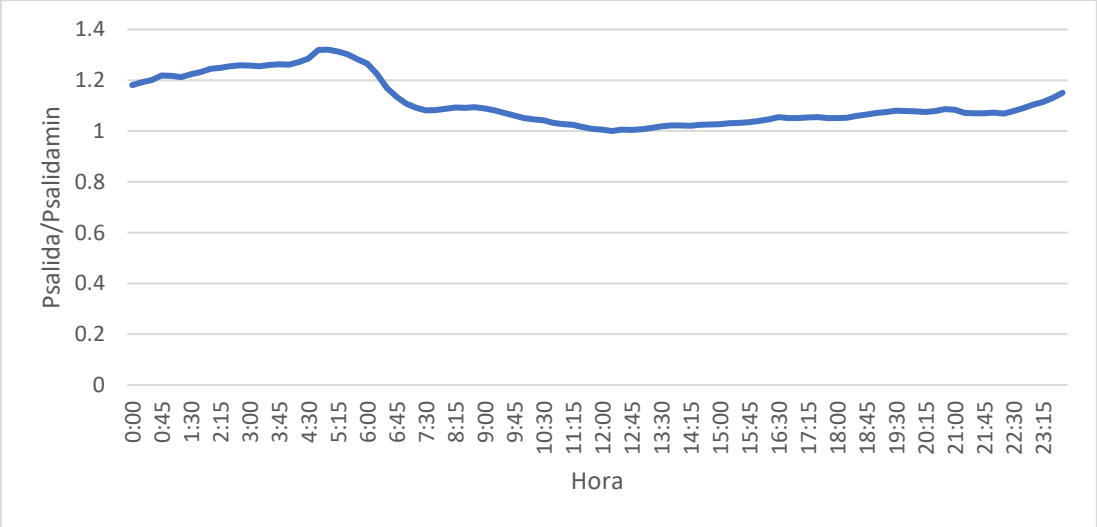

Figura 4.9: Patrón de presión de salida de la PVR ubicada el punto 19 para un día de la semana de agosto del 2020.

4.3. Calibración

En función de los procesos descritos en el punto 3.2 se realiza la calibración al modelo de distribución.

4.3.1. Identificar el propósito del modelo

El propósito del modelo es la detección y localización de fugas, por lo que, se requiere como resultado un modelo símil a la operación actual del sistema. En la Tabla 4.4 se muestran los criterios a seguir para la calibración del modelo para planificación y para operación, los que se usarán saber que tan bien calibrado está el modelo hidráulico.

Tabla 4.4: Criterios para la calibración utilizados por la empresa concesionaria del sector, según propósito del modelo.

Propósito	Error de presión aceptado (m)	% datos dentro del criterio
Planificación	3.5	100
Diseño	1.4	90
Operación	1.4	90

4.3.2. Determinar el valor inicial de los parámetros a estimar

En el punto 4.2 se presentan valores iniciales de los caudales para el modelo, el patrón de demanda de la zona y la materialidad de la red. La Tabla 4.5 muestra el coeficiente de Hazen-Williams para cada uno de los materiales existentes en la red.

Tabla 4.5: Coeficiente Hazen-Williams de los materiales en la red.

	Hazen-Williams
Asbesto Cemento (AC)	130
Acero	100
Concreto	110
Fierro Fundido (FFD)	110
Hormigón	110
Polietileno de Alta Densidad (HDPE)	140
Policloruro de Vinilo (PVC)	150

Además, en las Figuras 4.10, 4.11 y 4.12 se presenta el patrón de caudal demandado en el transcurso del día para los edificios de arranques de 50 mm, 75 mm y 150 mm.

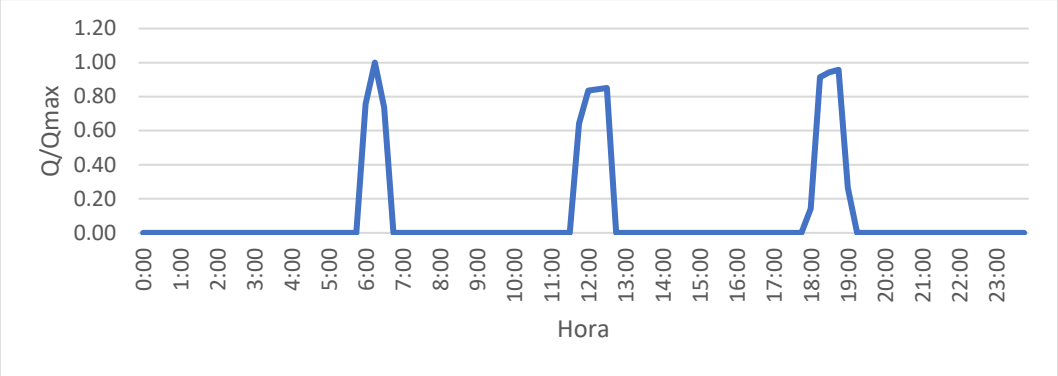

Figura 4.10: Patrón del caudal demandado en edificios con arranque de 150 mm.

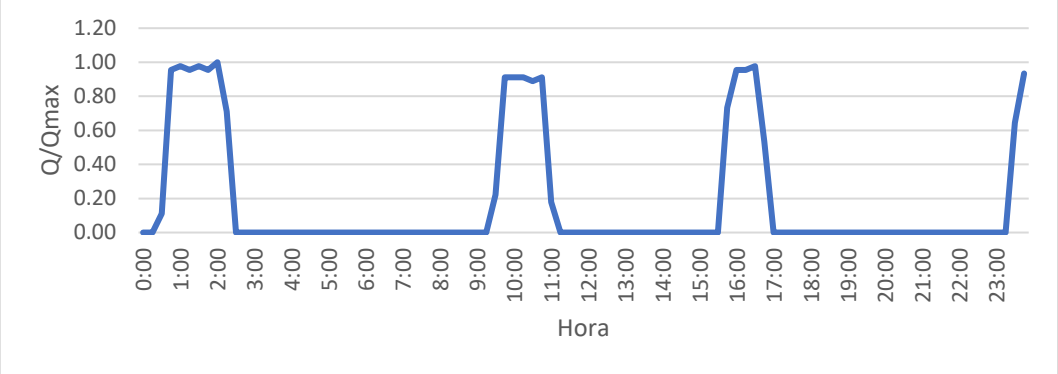

Figura 4.11: Patrón del caudal demandado en edificios con arranque de 80 mm.

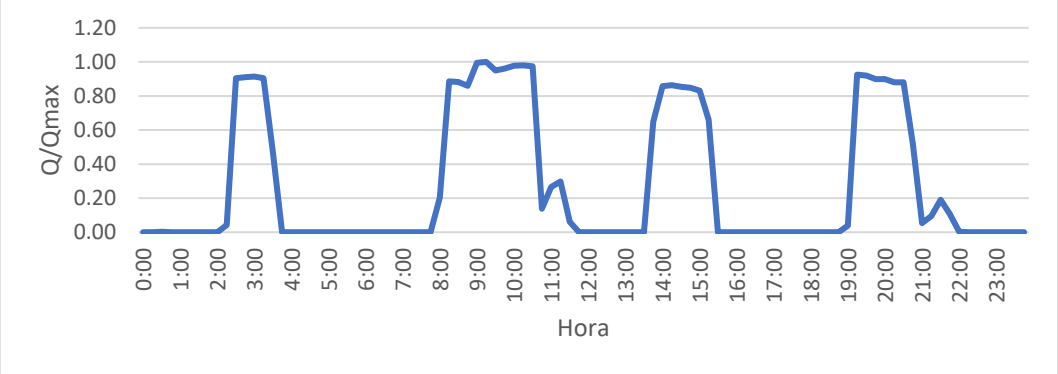

Figura 4.12: Patrón del caudal demandado en edificios con arranque de 50 mm.

Por otra parte, en el sistema de distribución existe una división del sector en sub y micro sectores para generar pisos de presión en estos y así evitar las presiones altas en el sector de menor elevación. Por lo que, para generar esta división existe una serie de válvulas de control de flujo, las que permiten distribuir el agua solamente por determinadas tuberías, generando así microsectores que se alimentan de la tubería que atraviesa todos los sectores distribuyendo el agua. En las Figuras 4.13 y 4.14 se muestran

los comportamientos que tendría el sistema si todas las válvulas de control de flujo se encontrasen abiertas para los caudales máximo diario y máximo horario.

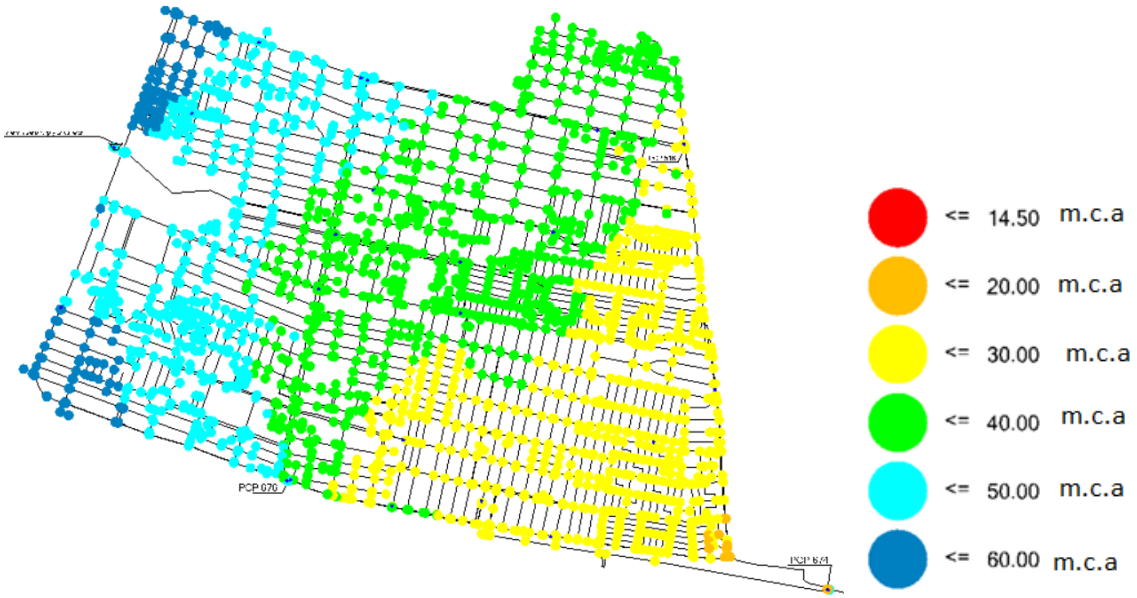

Figura 4.13: Presiones en el sistema con caudal máximo diario con las FCV abiertas.

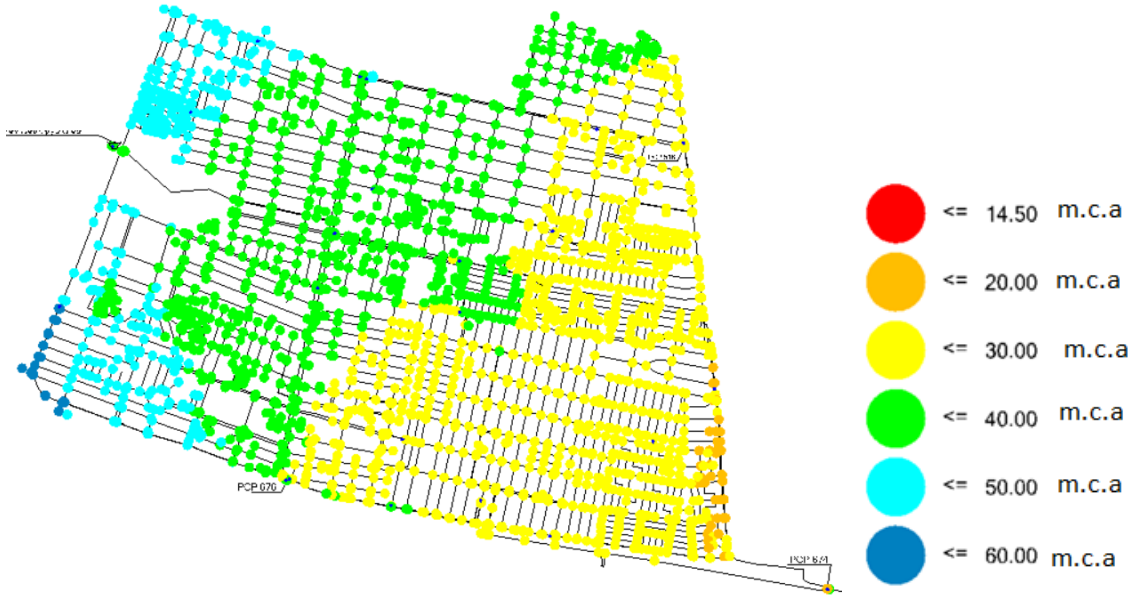

Figura 4.14: Presiones en el sistema con caudal máximo horario con las FCV abiertas.

En la Figura 4.15 se muestra la única zona de presión existente con las FCV abiertas, ya que la presión inicia en los 20-30 m.c.a. para el sector de mayor cota (sector derecho), ganando presión debido a que la cota disminuye a medida que se distribuye el agua hacia el sector de menor cota (sector izquierdo), terminando con presiones entre los 50-60 m.c.a para las menores cotas.

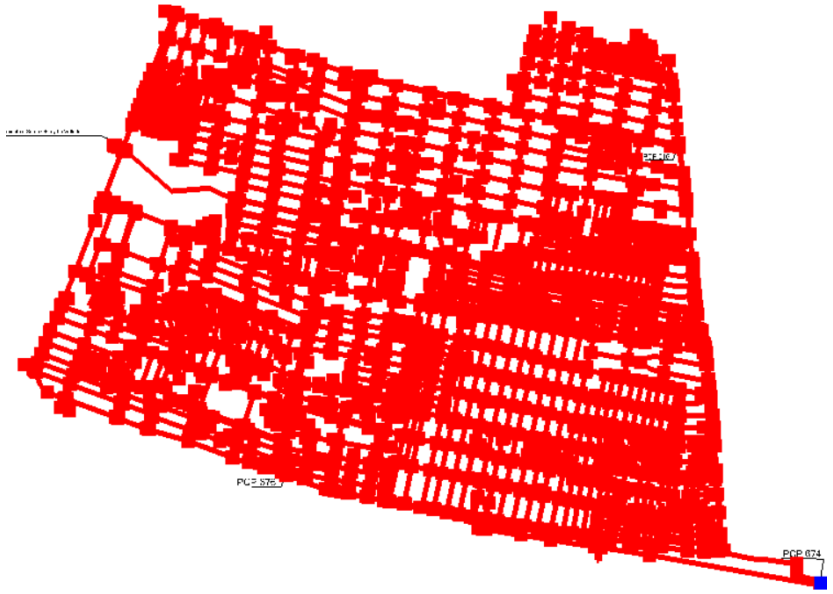

Figura 4.15: Única zona de presión existente con las FCV abiertas.

Al cerrar ciertas FCV se generan los micro y subsectores del sistema de distribución, los que se muestran en el comportamiento de la presión. Se generan 6 zonas de presión, las que concuerdan con los subsectores del sistema. En las Figuras 4.16 y 4.17 se muestran los comportamientos que tiene el sistema con la configuración actual para los caudales máximo diario y máximo horario.

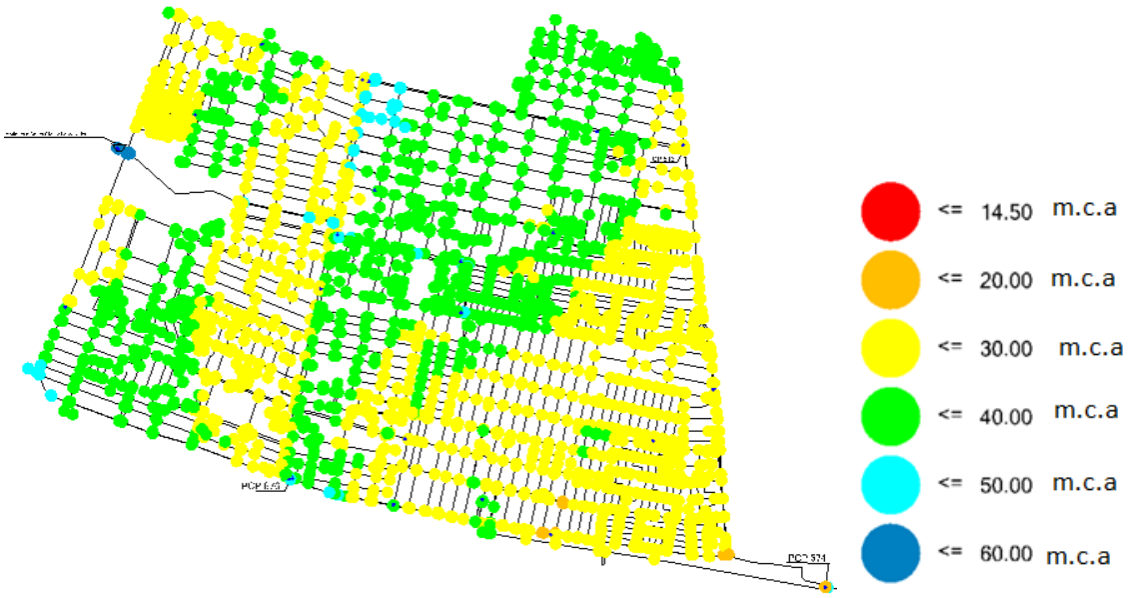

Figura 4.16: Presiones en el sistema para el caudal máximo diario al configurar las FCV.

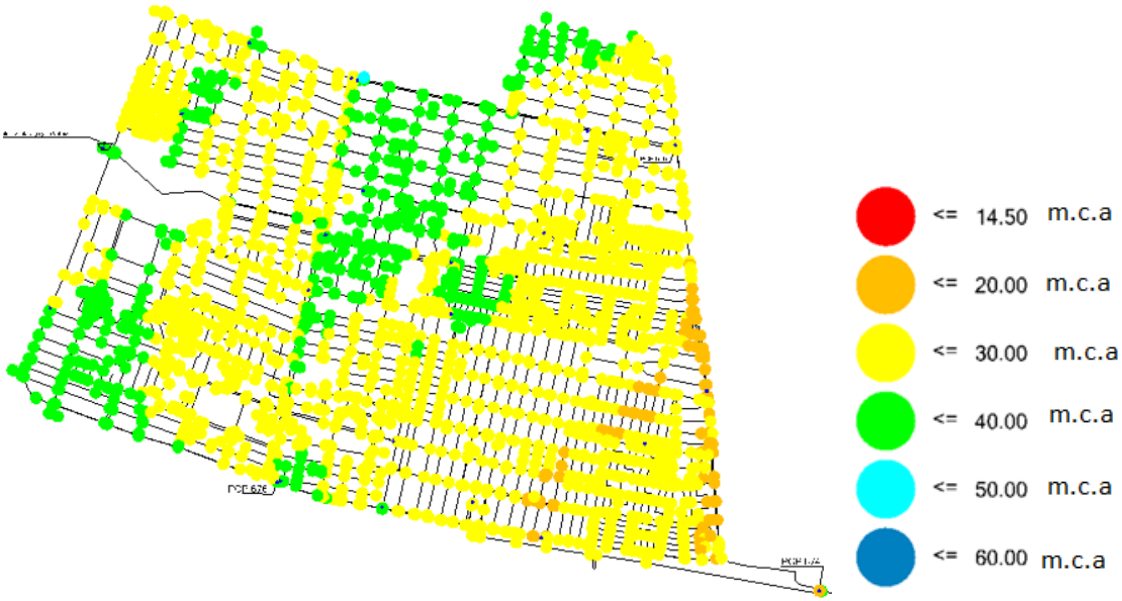

Figura 4.17: Presiones en el sistema para el caudal máximo horario al configurar las FCV.

En la Figura 4.18 se muestran en diferentes colores las zonas de presión generadas por el cierre de algunas de las FCV. La primera zona de presión cerrada por las FCV es la de color gris, que solo tiene conexión hidráulica con la zona de presión de color azul mediante determinadas tuberías, tal como la zona azul solo tiene conexión hidráulica con las zonas de presión de color rojo y magenta mediante determinadas tuberías. Esta división hidráulica de los subsectores genera que las presiones a lo largo de toda la zona de estudio se mantengan bajo los 40 m.c.a.

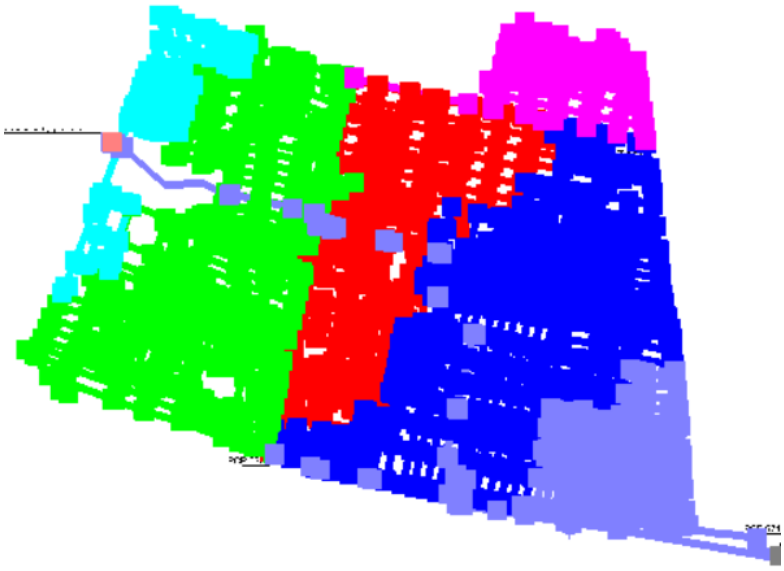

Figura 4.18: Zonas de presión existentes al configurar las FCV.

4.3.3. Recolectar información de calibración

Se utiliza información telemétrica de presión en diferentes puntos del sector, los que están ubicados como se muestra en la Figura 4.19.

Tabla 4.6: Puntos de medición telemétrica de la presión.

1	Punto 1
2	Punto 2
3	Punto 3
4	Punto 4
5	Punto 5
6	Punto 6
7	Punto 7
8	Punto 8
9	Punto 9
10	Punto 10
11	Punto 11
12	Punto 12
13	Punto 13
14	Punto 14
15	Punto 15
16	Punto 16
17	Punto 17
18	Punto 18
19	Punto 19
20	Punto 20
21	Punto 21
22	Punto 22
23	Punto 23
24	Punto 24
25	Punto 25
	Punto 26

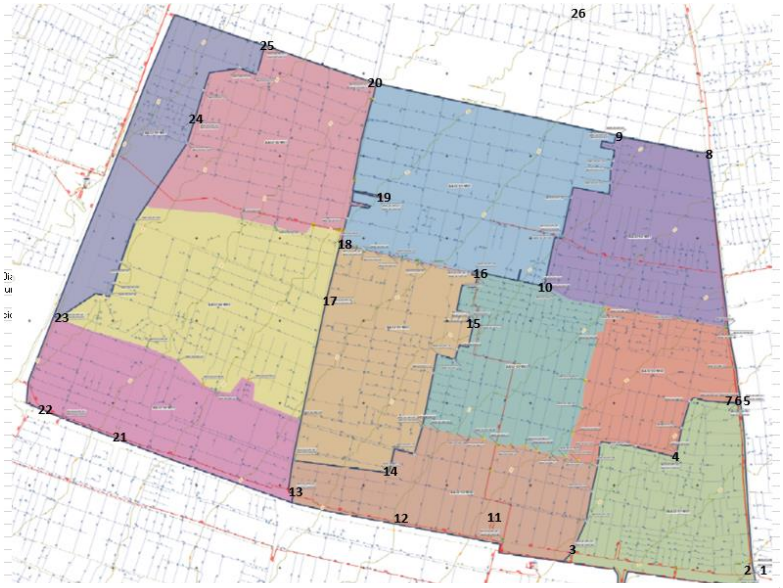

Figura 4.19: Puntos de medición telemétrica de la presión.

Se trabaja con la información del mes de agosto del 2020, obteniendo promedios del comportamiento en la semana y en el fin de semana de todos los puntos de medición mostrados en la Tabla 4.7. Se obtiene información de las presiones de entrada y salida de las válvulas reductoras de presión (PRV), como la mostrada en las Figuras 4.20 y 4.21, de la PRV ubicada en el Punto 9.

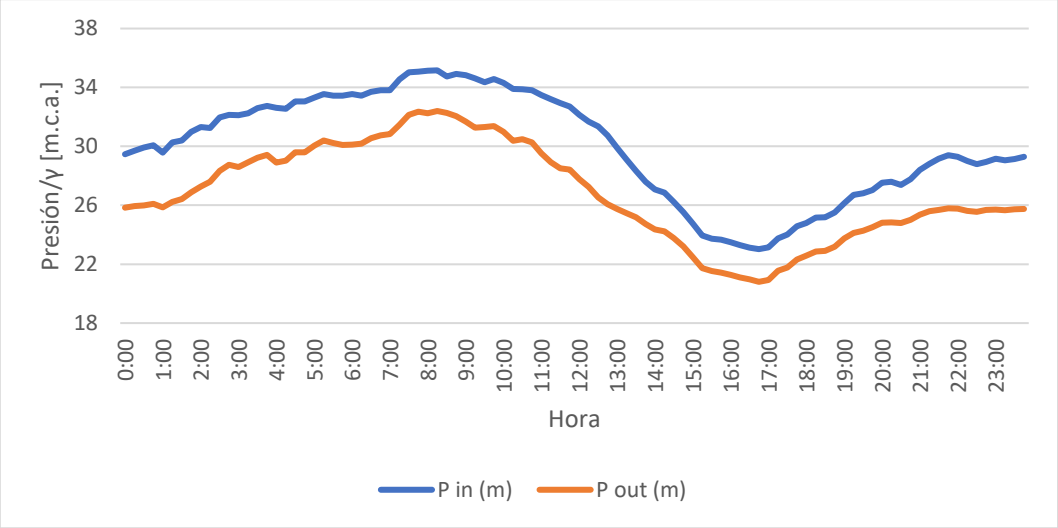

Figura 4.20: Comportamiento en la semana de la PVR ubicada en Alvear.

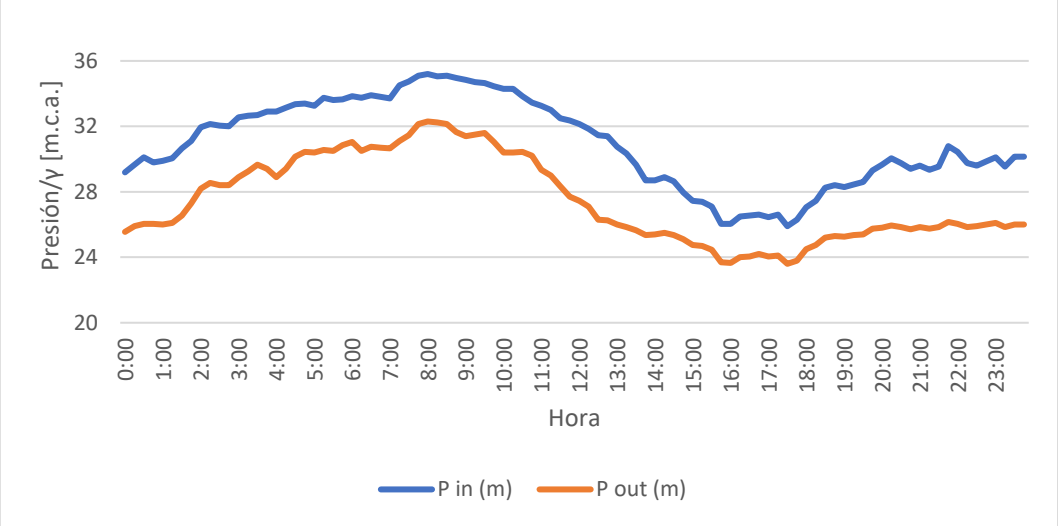

Figura 4.21: Comportamiento en el fin de semana de la PVR ubicada en Alvear.

4.3.4. Evaluar los resultados del modelo

Con los datos de telemetría recolectados se genera una planilla para evaluar la precisión de los datos de presión del modelo. Esta se construye en función de la salida del modelo, en este caso, las presiones de entrada y salida de las PRV y la presión en los nodos de medición. La Tabla 4.7 muestra el formato de salida de la planilla de precisión, donde las mediciones dentro del rango corresponden a la cantidad de mediciones que calza con la información telemétrica con un error aceptable en la presión de 3.5 m.c.a., coincidente con el criterio de planificación de la Tabla 4.4. Las mediciones fuera del rango corresponden a la cantidad de mediciones no coincidentes con la información telemétrica.

Tabla 4.7: Precisión de las presiones del modelo con la información telemétrica recolectada.

Punto de medición	Total de mediciones	Mediciones dentro del rango	Mediciones fuera del rango	Porcentaje de precisión
Punto 3	192	59	133	65.60%
Punto 4	192	144	48	49.50%
Punto 5	192	108	84	53.10%
Punto 6	192	34	158	56.30%
Punto 7	192	108	84	72.90%
Punto 8	96	0	0	79.20%
Punto 9	192	59	133	62.00%
Punto 10	192	87	105	26.60%
Punto 11	192	128	64	55.70%
Punto 12	192	134	58	57.30%
Punto 13	192	38	154	59.90%
Punto 14	192	0	192	73.40%
Punto 15	192	47	145	65.60%
Punto 16	192	46	146	80.70%
Punto 17	192	33	159	76.00%
Punto 18	192	29	163	46.40%
Punto 19	192	19	173	0.00%
Punto 20	192	49	143	51.00%
Punto 23	192	36	156	39.60%
Punto 24	192	0	192	78.10%
Punto 25	192	1	191	66.10%
Punto 26	192	108	84	51.00%
Punto 21	192	116	76	60.42%
Punto 22	192	106	86	55.21%
Precisión promedio final				57.57%

4.3.5. Calibración a nivel macro

Para la calibración a nivel macro se utilizan diferentes patrones en el consumo de agua de los edificios, además se programa el cierre de algunas de las PVR para las horas nocturnas, ya que durante la noche la operación del sistema de distribución cambia debido al caudal demandado.

Se obtienen mejoras en la precisión de las presiones del modelo llegando a los porcentajes mostrados en la Tabla 4.8.

Tabla 4.8: Precisión de las presiones del modelo con la calibración a nivel macro.

Punto de medición	Total de mediciones	Mediciones dentro del rango	Mediciones fuera del rango	Porcentaje de precisión
Punto 3	192	126	66	100.00%

Punto de medición	Total de mediciones	Mediciones dentro del rango	Mediciones fuera del rango	Porcentaje de precisión
Punto 4	192	95	97	51.04%
Punto 5	192	102	90	79.17%
Punto 6	192	108	84	64.06%
Punto 7	192	140	52	100.00%
Punto 8	192	152	40	96.35%
Punto 9	192	119	73	92.71%
Punto 10	192	51	141	90.63%
Punto 11	192	107	85	81.25%
Punto 12	192	110	82	84.38%
Punto 13	192	115	77	91.15%
Punto 14	192	141	51	31.25%
Punto 15	192	126	66	58.85%
Punto 16	192	155	37	100.00%
Punto 17	192	146	46	100.00%
Punto 18	192	89	103	95.31%
Punto 19	192	0	192	50.00%
Punto 20	192	98	94	48.96%
Punto 23	192	76	116	97.92%
Punto 24	192	150	42	62.50%
Punto 25	192	127	65	67.19%
Punto 26	96	49	47	43.75%
Punto 21	192	93	99	48.44%
Punto 22	192	60	132	31.25%
Precisión promedio final				73.59%

4.3.6. Realizar un análisis de sensibilidad

Se realizó un análisis de sensibilidad variando la presión de salida de la PRV de entrada a la red con un paso 0.5 m hasta restarle 3 m.c.a. y sumarle 3 m.c.a, obteniendo el gráfico de precisión promedio final del modelo hidraulico mostrado en la Figura 4.28. Se observa que la precisión promedio final se comporta como parábola, teniendo su peak en el escenario sin variación de la presión. Por otro lado, se ve que al aumentar o disminuir en 3 m la presión, se obtiene una disminución de aproximadamente el 10 % de la precisión promedio final.

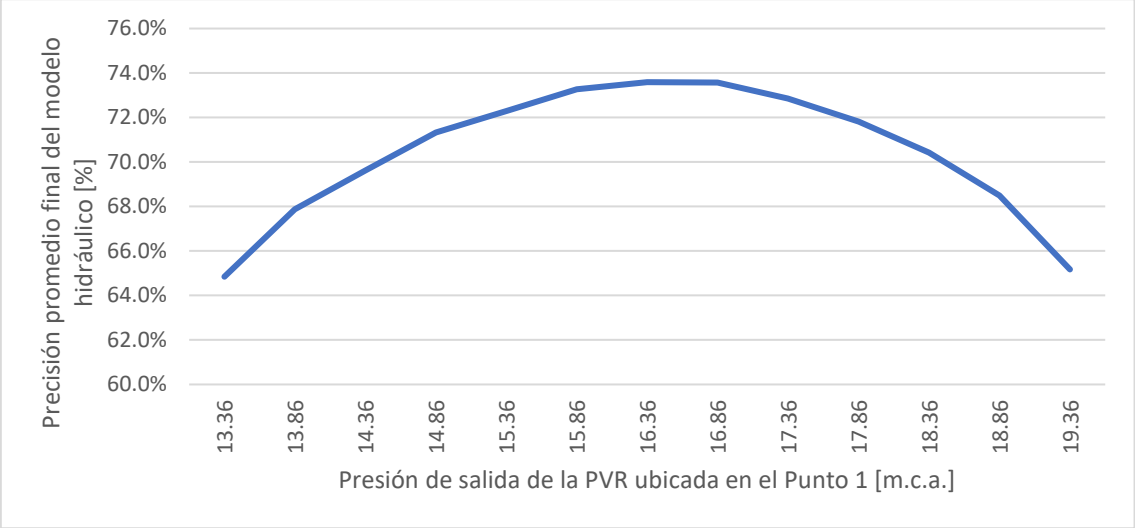

Figura 4.22: Análisis de sensibilidad variando la presión de entrada.

También se realizó un análisis de sensibilidad variando la demanda de agua, generando diferencias en los caudales que fluyen en la red. Se obtiene el gráfico de precisión promedio final del modelo hidráulico mostrado en la Figura 4.23. Se observa que la precisión promedio final también se comporta como parábola al variar el caudal demandado por la red, presentando su peak con 0.925 veces el caudal presentado en la Tabla 4.3. Además, si el caudal se encuentra entre 1.025 y 0.825 veces el caudal de la tabla 4.3, la precisión promedio final del modelo hidráulico no varía más que un 1 %, por lo que se opta por mantener el caudal de la Tabla 4.3 ya que el cambiarlo modificaría de gran manera el comportamiento hidráulico de la red y solo aumentaría en un 1 % la precisión promedio final. Por otro lado, también se ve que a partir de 1.05 veces el caudal la precisión promedio final del modelo hidráulico decae abruptamente.

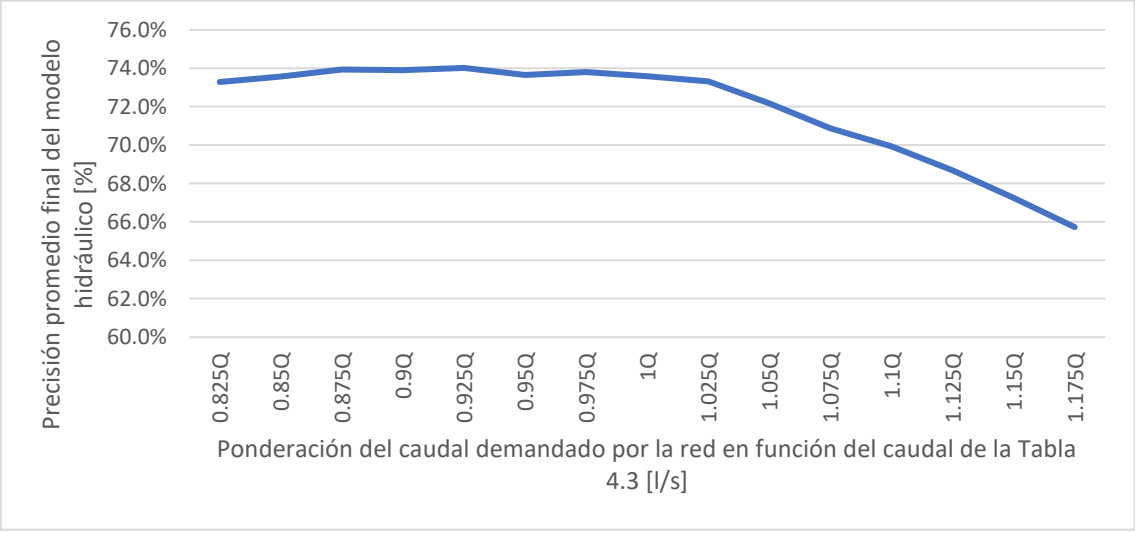

Figura 4.23:Análisis de sensibilidad variando la demanda.

4.3.7. Calibración a nivel micro

Se realizó una calibración a nivel micro modificando los coeficientes de rugosidad de las tuberías, pero para obtener mejoras en la precisión final del modelo era necesario modificar los coeficientes a un punto de dejar de ser coherente hidráulicamente, por ejemplo, la precisión final del modelo mejoraba al utilizar un coeficiente de Hazem-Williams (C) de 200 en PVC, pero el C de este material es de 150, por lo que, al utilizar un C de 200 estaríamos asumiendo que la pérdida por fricción en las tuberías es menor a la que ocurre en la realidad con una tubería de PVC recién instalada. Por esto, el trabajo continúa con el modelo de mejor precisión y teniendo en cuenta el análisis de sensibilidad realizado en el punto anterior. La tabla de precisión final del modelo hidráulico es presentada en la Tabla 4.9.

Tabla 4.9: Precisión del modelo con la información telemétrica recolectada.

Punto de medición	Total de mediciones	Mediciones dentro del rango	Mediciones fuera del rango	Porcentaje de precisión
Punto 3	192	126	66	100.00%
Punto 4	192	95	97	51.04%
Punto 5	192	102	90	79.17%
Punto 6	192	108	84	64.06%
Punto 7	192	140	52	100.00%
Punto 8	192	152	40	96.35%
Punto 9	192	119	73	92.71%
Punto 10	192	51	141	90.63%
Punto 11	192	107	85	81.25%
Punto 12	192	110	82	84.38%
Punto 13	192	115	77	91.15%
Punto 14	192	141	51	31.25%
Punto 15	192	126	66	58.85%
Punto 16	192	155	37	100.00%
Punto 17	192	146	46	100.00%
Punto 18	192	89	103	95.31%
Punto 19	192	0	192	50.00%
Punto 20	192	98	94	48.96%
Punto 23	192	76	116	97.92%
Punto 24	192	150	42	62.50%
Punto 25	192	127	65	67.19%
Punto 26	96	49	47	43.75%
Punto 21	192	93	99	48.44%
Punto 22	192	60	132	31.25%
Precisión promedio final				73.59%

5. Detección y localización de fugas utilizando una SVM

Como se mencionó anteriormente, el objetivo principal de este trabajo de título es la creación de un programa que sea capaz de detectar y localizar fugas en un sistema de distribución de agua potable, mediante la utilización de un algoritmo de aprendizaje supervisado, como lo es una SVM. Para este fin, se entrena la SVM con las salidas del modelo hidráulico de distribución calibrado.

5.1. Generar set de datos de simulaciones para entrenamiento y prueba.

Con el modelo calibrado dentro de los criterios definidos en el punto 4.3.1., se generan las simulaciones con fuga basándose en el comportamiento en las semanas y en los fines de semana de agosto del año 2019 (8 diferentes simulaciones, 4 semanas y 4 fines de semana). Las simulaciones con fuga constan de 24 (8x3) simulaciones de base, las que se basan en el caudal máximo horario demandado por la zona, ponderando para cada uno el caudal por un factor entre 0,9, 1 y 1,1. De estas 24 simulaciones base se les generan fuga en 16 puntos diferentes de la zona y a 6 horarios diferentes, lo que da un total de 96 combinaciones. Generando un total de 2.304 simulaciones con fugas. Para las simulaciones sin fugas se tiene los 8 escenarios de agosto 2019, variando en +0.5 la presión de salida de la PRV de entrada al sistema se tiene 16 escenarios en los que el caudal demandado por el sector se pondera entre 0.8 y 1.2 con un paso de 0.005, teniendo 1.296 escenarios sin fuga. Lo que da un total de 3.600 simulaciones las que se dividirán en 2700 (75%) de entrenamiento y 900 (25%) de prueba.

5.2. Programación de la SVM

Para la programación de la SVM se utiliza el entorno informático de “Jupyter Notebook”, el que es un entorno interactivo que permite desarrollar código de Python de manera dinámica. Se ejecuta de manera local y posibilita la ejecución de código y la escritura de texto, de manera que se pueda entender el código como un documento para la lectura.

Tabla 5.1: Librerías a utilizar para ordenar los datos.

Librería	Función
numpy	Realizar calculos estadísticos y realizar operaciones matemáticas
glob	Realizar la lecutara de nombres de lso archivos de salida del modelo
os	Modificación de los nombres de los archivos
pandas	Importar y manipular los datos de salida del modelo

Dentro del entorno de Jupyter Notebook se utilizan librerías de Python mostradas en la Tabla 5.1, como “pandas” para importar y manipular datos, “numpy” para calcular

estadísticos y realizar operaciones matemáticas, “glob” para la lectura de nombres de los archivos de salida del modelo y “os” para la modificación de los nombres de estos archivos. Con estas librerías se ordenan los datos de salida del modelo para poder ingresar a la SVM. Para poder generar más de una SVM se ordenan los datos de tres formas.

En el primer y segundo ordenamiento, cada escenario será una matriz donde las filas corresponderán a la hora del día con un paso de 15 minutos y las columnas representarán los nodos existentes en la red de distribución como se muestra en la Figura 5.30. La diferencia entre ambos ordenamientos es que, el primero contiene solo las columnas de los nodos y PRV que cuentan con información telemétrica y el segundo con las columnas de los nodos y PVR con una buena distribución espacial en el modelo (como muestran las Figura 5.1 y Figura 5.2 respectivamente).

Tabla 5.2: Matriz para los ordenamientos 1 y 2.

Hora	Av General Freire Pllegada	Av General Freire Psalida	Fernandez Albano Gran Avenida Pllegada	Fernandez Albano Gran Avenida Psalida	El Parron Gran Avenida Pllegada	El Parron Gran Avenida Psalida	PCP 674 Pllegada	PCP 674 Psalida				124564 P	129928 P	128473 P	128517 P	125869 P	127817 P	128744 P	label.1
0:00:00	37.14	33.01	54.8	31.31	35.91	30.23	52.22	17.71	.	.	.	22.25	22.06	22.29	22.29	22.32	21.83	20.61	0
0:15:00	37.67	33.34	54.15	31.65	36.44	30.56	52.2	16.75	.	.	.	22.31	22.08	22.37	22.37	22.41	21.91	20.69	0
0:30:00	37.75	33.6	53.93	31.76	36.53	30.64	52.16	16.37	.	.	.	22.34	22.09	22.39	22.38	22.43	21.93	20.71	0
0:45:00	37.19	34.06	53.71	30.64	36.45	29.32	51.54	16.43	.	.	.	22.15	22.02	22.05	22.05	22.01	21.57	20.35	0
1:00:00	37.05	34.06	53.84	30.66	36.31	29.34	51.59	16.46	.	.	.	21.89	22.02	21.99	21.98	21.9	21.49	20.27	0
1:15:00	36.83	33.88	53.99	30.43	36.1	29.11	51.57	16.55	.	.	.	21.9	22.01	21.96	21.96	21.87	21.46	20.24	0
1:30:00	35.8	34.18	53.59	28.94	35.66	27.38	50.83	16.75	.	.	.	21.68	21.93	21.6	21.6	21.48	21.07	19.86	0
1:45:00	35.98	34.45	53.82	29.17	35.87	27.58	50.89	16.87	.	.	.	21.71	21.94	21.63	21.63	21.52	21.11	19.89	0
2:00:00	37.71	34.78	54.84	31.31	36.98	29.93	51.92	16.89	.	.	.	22.08	22.07	22.15	22.14	22.1	21.67	20.45	0
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22:00:00	32.91	29.96	53.74	26.32	32.84	25.15	43.79	20.19	.	.	.	21.6	21.86	21.46	21.46	21.36	20.92	19.71	0
22:15:00	33.73	29.88	53.83	27.69	33.15	26.64	44.66	19.52	.	.	.	21.58	21.91	21.75	21.74	21.65	21.23	20.02	0
22:30:00	34.71	30.15	54.63	28.81	33.7	27.83	45.45	19.24	.	.	.	21.77	21.98	21.99	21.98	21.93	21.49	20.28	0
22:45:00	34.9	30.48	54.93	29.13	33.83	28.14	45.58	19.25	.	.	.	21.9	21.99	22.04	22.03	22.01	21.55	20.33	0
23:00:00	35.27	30.87	55.71	29.63	34	28.64	45.91	19.34	.	.	.	22.05	22.01	22.14	22.13	22.12	21.66	20.44	0
23:15:00	35.57	31.1	54.72	29.88	34.3	28.89	46.02	18.2	.	.	.	21.99	22	22.15	22.14	22.13	21.67	20.45	0
23:30:00	35.07	31.58	54.13	29.2	34.14	28.1	45.57	17.77	.	.	.	22.02	21.96	21.98	21.97	21.96	21.49	20.27	0
23:45:00	35.31	32.17	54.21	29.15	34.56	27.96	45.5	17.69	.	.	.	22.07	21.96	21.93	21.92	21.92	21.44	20.21	0
0:00:00	37.14	33.01	54.8	31.31	35.91	30.23	46.34	17.71	.	.	.	22.25	22.06	22.29	22.29	22.32	21.83	20.61	0

Figura 5.1: Cantidad de columnas en el ordenamiento tipo 1.

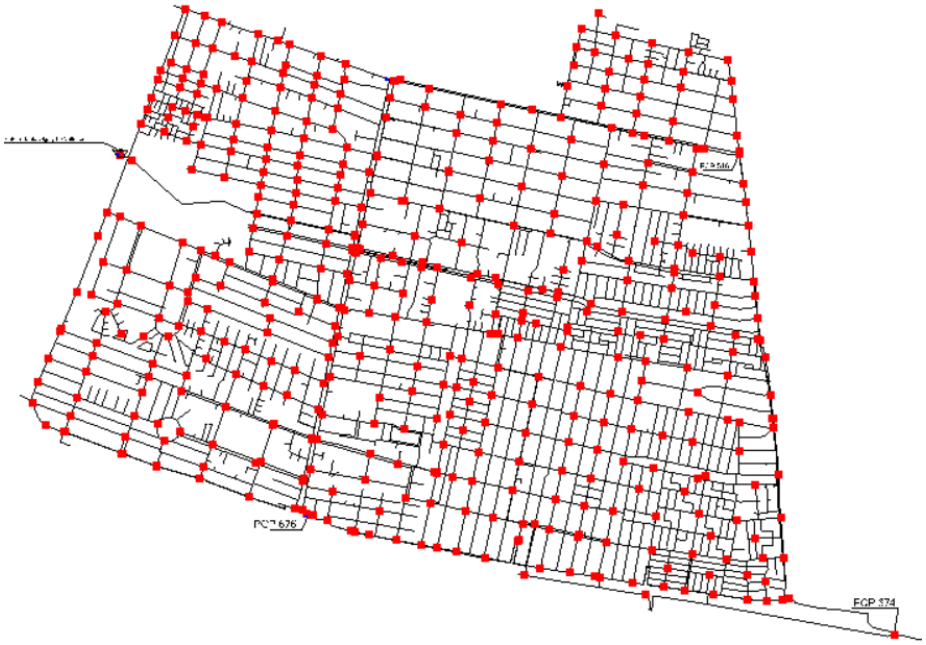

Figura 5.2: Cantidad de columnas en el ordenamiento 2.

En ambas matrices la última columna corresponde al elemento a ser clasificado con la SVM. Existen 17 clases del 0 al 16. La clase 0 corresponde a que no existe fuga a la hora de esa fila, la clase 1 corresponde a que existe fuga en la válvula reductora de presión ubicada en el Punto 25 de la Figura 4.19 o en el nodo “113228” de la red de distribución,

dependiendo de la matriz que se utilice para crear la SVM. En la Tabla 5.12 se presentan las clases de cada matriz.

Tabla 5.3: Tabla de las clases de fugas.

Clase	Matriz ordenamiento 1 (nodo)	Matriz ordenamiento 2 (PRV)
0	Sin fuga	Sin fuga
1	113228	Fuga en Punto 25 Figura 4.19
2	115822	Fuga en Punto 20 Figura 4.19
3	117271	Fuga en Punto 24 Figura 4.19
4	118656	Fuga en Punto 23 Figura 4.19
5	126278	Fuga en Punto 14 Figura 4.19
6	127118	Fuga en Punto 14 Figura 4.19
7	127938	Fuga en Punto 4 Figura 4.19
8	129129	Fuga en Punto 3 Figura 4.19
9	114633	Fuga en Punto 4 Figura 4.19
10	115905	Fuga en Punto 24 Figura 4.19
11	120135	Fuga en Punto 10 Figura 4.19
12	121124	Fuga en Punto 16 Figura 4.19
13	122037	Fuga en Punto 17 Figura 4.19
14	122746	Fuga en Punto 23 Figura 4.19
15	127033	Fuga en Punto 20 Figura 4.19
16	130102	Fuga en Punto 3 Figura 4.19

En el tercer ordenamiento (ordenamiento 3), cada escenario será una matriz donde las filas corresponderán a los nodos existentes en la red de distribución y las columnas corresponden a la hora del día con un paso de 15 minutos como se muestra en la Figura 5.33. En este caso las clases son 7 (a las 0, 6, 8, 10, 14, 15 y 18 hrs), donde cada una representa si hay fuga o no en el nodo o PRV de la fila. El 0 nuevamente corresponde a la ausencia de fuga y en el resto de las clases el número corresponde a la hora a la que ocurre la fuga.

Tabla 5.4: Matriz para los ordenamientos 1 y 2.

index	0:00:00	0:15:00	0:30:00	0:45:00	1:00:00	1:15:00	1:30:00	.	.	.	23:00:00	23:15:00	23:30:00	23:45:00	0:00:002	label
Av General Freire Pllegada	36.64	36.8	36.74	34.55	33.45	32.69	31.36	.	.	.	35.24	35.47	34.77	34.68	36.64	0
Av General Freire Psalida	33.01	33.34	33.6	34.06	33.45	32.69	31.36	.	.	.	30.87	31.1	31.58	32.17	33.01	0
Fernandez Albano Gran Avenida Pllegada	53.87	52.54	51.5	49.67	48.52	47.39	44.1	.	.	.	55.74	54.72	53.69	53.25	53.87	0
Fernandez Albano Gran Avenida Psalida	30.76	30.7	30.3	27.62	26.53	25.7	25.98	.	.	.	29.63	29.86	28.68	28.12	30.76	0
El Parron Gran Avenida Pllegada	35.61	35.89	36.04	35.1	34.4	33.81	32.44	.	.	.	33.99	34.21	34.03	34.28	35.61	0
El Parron Gran Avenida Psalida	29.64	29.53	29.05	25.99	24.89	23.89	22.44	.	.	.	28.64	28.86	27.53	26.82	29.64	0
PCP 674 Pllegada	51.62	51.17	50.31	48.77	47.86	47.08	45.26	.	.	.	45.92	46	45.2	44.73	45.75	0
PCP 674 Psalida	17.71	16.75	16.37	16.43	16.46	16.55	16.75	.	.	.	19.34	18.2	17.77	17.69	17.71	0
Paraguay Pllegada	33.63	32.55	31.92	31.02	30.6	30.26	29.45	.	.	.	35.32	34.22	33.39	33.08	33.63	0
Paraguay Psalida	23.11	23.11	23.12	23.13	23.14	23.14	23.15	.	.	.	23.11	23.1	23.09	23.1	23.11	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
124564 P	22.25	22.31	22.34	22.15	21.89	21.9	21.68	.	.	.	22.05	21.99	22.02	22.07	22.25	0
129928 P	22.06	22.08	22.09	22.02	22.02	22.01	21.93	.	.	.	22.01	22	21.96	21.96	22.06	0
128473 P	22.29	22.37	22.39	22.05	21.99	21.96	21.6	.	.	.	22.14	22.15	21.98	21.93	22.29	0
128517 P	22.29	22.37	22.38	22.05	21.98	21.96	21.6	.	.	.	22.13	22.14	21.97	21.92	22.29	0
125869 P	22.32	22.41	22.43	22.01	21.9	21.87	21.48	.	.	.	22.12	22.13	21.96	21.92	22.32	0
127817 P	21.83	21.91	21.93	21.57	21.49	21.46	21.07	.	.	.	21.66	21.67	21.49	21.44	21.83	0
128744 P	20.61	20.69	20.71	20.35	20.27	20.24	19.86	.	.	.	20.44	20.45	20.27	20.21	20.61	0

En los tres ordenamientos los distintos escenarios se agregan hacia abajo como filas quedando 3 matrices de distintas dimensiones. Con estas matrices y la librería “sklearn” es posible realizar la división de los datos para los conjuntos de entrenamiento y prueba, escalar los datos, crear una SVM para la clasificación, realizar una validación cruzada para obtener valores de los parámetros de la SVM y para generar diferentes métricas de la SVM y su matriz de confusión que es una herramienta que permite visualizar el desempeño de las SVM o cualquier método de aprendizaje supervisado.

5.3. Entrenamiento y prueba de la SVM para la clasificación de los estados de la red

Teniendo las matrices de las simulaciones es posible entrenar la SVM con un porcentaje de éstas, para que el algoritmo reconozca la diferencia entre estados de la red mediante un patrón en los datos de presión y caudal entregados. Luego de entrenada la SVM, esta clasificará el resto de las simulaciones como comportamiento normal o anormal, entregando en caso de ser necesario la ubicación de la fuga o de la anomalía encontrada.

5.4. Calibración de parámetros C y γ de la SVM

Como se vió en la sección 3.5, el parámetro C regula la maximización del margen de la SVM, por lo que a valores altos de C el margen es menor y se aceptan una menor cantidad de fallas, intentando clasificar de forma correcta todos los escenarios, por lo que, es fácil que ocurra sobreajuste en la SVM. A valores bajos de C el margen es mayor y se aceptan una mayor cantidad de fallas, por lo que puede existir una baja precisión en la clasificación. El parámetro γ representa la influencia de cada escenario del entrenamiento. Puede ser visto como el inverso del radio de influencia de los escenarios

seleccionados como vectores de apoyo. Los valores Scale y Auto de γ mostrados en la Tabla 5.5 utilizan las siguientes fórmulas:

$$\text{I.} \quad \text{si } \gamma = \text{Scale}\,, \gamma = \frac{1}{n°caracteristicas * Varianza(caracteristicas)} \quad (5.1)$$

$$\text{II.} \quad \text{si } \gamma = \text{Auto}\,, \gamma = \frac{1}{n°caracteristicas} \quad (5.2)$$

donde:

- n°caracteristicas: son el número de columnas de la matriz entregada, en el caso de los ordenamientos 1 y 2 son el número de PVR y nodos en la matriz y para el ordenamiento 3 son los datos de presión cada 15 minutos para cada nodo o PVR.
- Varianza(características): es la varianza de las características, es decir, la varianza de los valores en las columnas para cada fila.

Mediante validación cruzada se calibran los parámetros C y γ de la SVM. Los que se variarán como lo muestra la Tabla 5.5.

Tabla 5.5: Parámetros a utilizar para la validación cruzada.

Parámetro							
C	0.1	1	10	100	1000	10000	30000
γ	Scale	Auto					

5.5. Entrenamiento y prueba la SVM calibrada para la clasificación de los estados de la red

La validación cruzada arroja los parámetros para utilizar en la SMV, por lo que se programa la SVM con éstos y se realiza el entrenamiento y prueba con la SVM ya calibrada.

En la Figura 5.34 se muestran los puntos anteriores en el código de Python.

```
74
75  from sklearn.pipeline import Pipeline
76  from sklearn.model_selection import GridSearchCV
77  from sklearn.preprocessing import StandardScaler
78  from sklearn.svm import SVC
79  from sklearn.model_selection import train_test_split
80  #importar las funciones de la librería a utilizar
81
82  svm2 = SVC() # se define la SVM sin parámetro
83  pipe = Pipeline(steps=[('scaler', StandardScaler()), ('svm2', svm2)])
84  #para aplicar secuencialmente transformadas (escalar) y al final un estimador
85
86  X_train2, X_test2, y_train2, y_test2 = train_test_split(X1, Y1, stratify=Y1, random_state=1)
87  #dividir los datos para los conjuntos de entrenamiento y prueba
88
89  param_grid = {
90      'svm2__C': [10, 100, 1000, 10000, 20000],
91      'svm2__gamma': ['scale', 'auto']}
92  #parámetros a utilizar para realizar la validación cruzada
93
94  search2 = GridSearchCV(pipe, param_grid, n_jobs=-1, verbose = 1)
95  #se realiza la validación cruzada mediante la función GridSearchCV
96
97  search2.fit(X_train2, y_train2)
98  #se entrena la SVM con el conjunto de entrenamiento
99
```

Figura 5.3: Código para la SVM.

Además de las SVM para las 3 matrices antes mencionadas, se crean otras 3 SVM con el promedio de la presión por nodo para cada una de las 3 matrices ya creadas, por lo que se tiene un total de 6 SVM que presentan comportamientos diferentes, debido a que el ordenamiento de datos es diferente. Por último, se crea una última SVM para el primer ordenamiento (se creó solo en este ordenamiento porque si se hacía en los otros 2 ordenamientos el número de columnas aumentaría en demasía y la infraestructura de supercómputo del NLHPC de la facultad no permitía que corriera la programación). Este que consta de transformar la matriz de cada escenario en una sola fila (generando una fila de 97xn° de nodos columnas).

- Ordenamiento 1.1: Ordenamiento 1
- Ordenamiento 1.2: Ordenamiento 1 con la presión promedio
- Ordenamiento 1.3: Ordenamiento 1 con escenario en una fila
- Ordenamiento 2.1: Ordenamiento 2
- Ordenamiento 2.2: Ordenamiento 2 con la presión promedio
- Ordenamiento 3.1: Ordenamiento 3
- Ordenamiento 3.2: Ordenamiento 3 con la presión promedio

6. Resultados para la detección y localización de fugas

En este capítulo se presentan los resultados obtenidos con las diferentes SVM creadas. Dentro de los resultados que se muestran, se encuentran los mejores parámetros obtenidos por la validación cruzada, el tiempo que demoró en procesar la SVM y la matriz de confusión que, como se dijo anteriormente, es una herramienta que permite visualizar el desempeño de las SVM, donde cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa la clase que predice el algoritmo, por lo que, si se tiene una diagonal y solo ceros, el resultado es perfecto y el algoritmo posee un buen desempeño, cabe destacar que el resto de métricas no se condicen con la matriz de confusión y los resultados tendían a causar confusión al arrojar mejores eficiencias que las existentes, por lo que, se escoge la matriz de confusión como el elemento visual que mejor representa los resultados. Para el procesamiento de las SVM se utiliza la infraestructura de supercómputo del NLHPC.

6.1. SVM para el primer ordenamiento (Ordenamiento 1.1)

En la Tabla 6.1 se presentan los parámetros obtenidos por la función GridsearchCV mostrada en la Figura 5.3 (validación cruzada) y el tiempo que tomó en obtener el resultado. En la Tabla 6.2 se presenta la matriz de confusión de la SVM para los datos de prueba. En la Tabla 6.3 se presenta la matriz de confusión de la SVM para los datos de entrenamiento.

Tabla 6.1: Parámetros para la primera SVM (con ordenamiento 1.1) y tiempo de procesamiento.

Parámetro	Elección
C	10.000
γ	Scale
Tiempo (min)	384

Tabla 6.2: Matriz de confusión de la primera SVM (con ordenamiento 1.1) con datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	86488	1	0	0	0	0	0	2	0	0	5	0	7	0	4	0	0
1	29	3	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0
2	31	0	0	0	2	0	0	0	0	3	0	0	0	0	0	0	0
3	29	2	0	2	0	0	0	0	0	0	3	0	0	0	0	0	0
4	30	0	1	0	3	0	0	0	0	1	0	1	0	0	0	0	0
5	34	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
6	34	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
7	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0
8	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	29	0	6	0	0	0	0	0	0	1	0	0	0	0	0	0	0
10	28	0	0	2	0	0	0	0	0	0	6	0	0	0	0	0	0

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
11	31	0	1	0	3	0	0	0	0	0	0	1	0	0	0	0	0
12	30	0	0	0	2	0	0	0	0	0	0	0	3	1	0	0	0
13	34	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
14	32	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
15	31	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4	0
16	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla 6.3: Matriz de confusión de la primera SVM (con ordenamiento 1.1) con datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	259511	1	0	0	0	0	0	0	0	0	3	0	4	0	1	0	0
1	87	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	90	0	15	0	1	0	0	0	0	2	0	0	0	0	0	0	0
3	94	1	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0
4	94	0	1	0	12	0	0	0	0	0	0	0	0	1	0	0	0
5	97	0	0	0	0	7	4	0	0	0	0	0	0	0	0	0	0
6	93	0	0	0	0	2	13	0	0	0	0	0	0	0	0	0	0
7	96	0	0	0	0	0	0	2	0	0	0	0	0	0	0	10	0
8	105	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0
9	90	0	4	0	0	0	0	0	0	14	0	0	0	0	0	0	0
10	86	2	0	2	0	0	0	0	0	0	18	0	0	0	0	0	0
11	92	0	0	0	1	0	0	0	0	0	0	15	0	0	0	0	0
12	80	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0
13	94	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0
14	84	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0
15	91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0
16	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

6.2. SVM para el primer ordenamiento usando el promedio de la presión (Ordenamiento 1.2)

En la Tabla 6.4 se presentan los parámetros escogidos por la función GridsearchCV mostrada en la Figura 5.3 (validación cruzada) y el tiempo que tomó en obtener el resultado. En la Tabla 6.5 se presenta la matriz de confusión de la SVM para los datos de prueba. En la Tabla 6.6 se presenta la matriz de confusión de la SVM para los datos de entrenamiento.

Tabla 6.4: Parámetros para la segunda SVM (con ordenamiento 1.2) y tiempo de procesamiento.

Parámetro	Elección
C	30.000
γ	Scale

Parámetro	Elección
Tiempo (min)	7

Tabla 6.5: Matriz de confusión de la segunda SVM (con ordenamiento 1.2) con datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	322	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	0	30	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	34	0	0	0	0	0	0	2	0	0	0	0	0	0	0
3	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	3	0	33	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	18	18	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	15	21	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	31	5	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	3
9	0	0	1	0	0	0	0	0	0	35	0	0	0	0	0	0	0
10	0	0	0	3	0	0	0	0	0	0	33	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	35	0	0	0	0	1
12	0	0	0	0	2	0	0	0	0	0	0	0	34	0	0	0	0
13	1	0	0	0	1	0	0	0	0	0	0	0	0	34	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
15	0	0	0	0	0	0	0	8	1	0	0	0	0	0	0	27	0
16	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30

Tabla 6.6: Matriz de confusión de la segunda SVM (con ordenamiento 1.2) con datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	966	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	0	105	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	104	0	0	0	0	0	0	4	0	0	0	0	0	0	0
3	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	88	20	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	26	82	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	83	25	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	1	101	0	0	0	0	0	0	0	5
9	0	0	3	0	0	0	0	0	0	105	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	108	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	108	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	107	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	108	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108	0	0
15	0	0	0	0	0	0	0	14	2	0	0	0	0	0	0	91	0
16	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	102

6.3. SVM para el primer ordenamiento con cada escenario en 1 fila (Ordenamiento 1.3)

En la Tabla 6.7 se presentan los parámetros obtenidos por la función GridsearchCV mostrada en la Figura 5.3 (validación cruzada) y el tiempo que tomó en obtener el resultado. En la Tabla 6.8 se presenta la matriz de confusión de la SVM para los datos de prueba. En la Tabla 6.9 se presenta la matriz de confusión de la SVM para los datos de entrenamiento.

Tabla 6.7: Parámetros para la tercera SVM (con ordenamiento 1.3) y tiempo de procesamiento.

Parámetro	Elección
C	10
γ	Scale
Tiempo (min)	2

Tabla 6.8: Matriz de confusión de la tercera SVM (con ordenamiento 1.3) con datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	323	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla 6.9: Matriz de confusión de la tercera SVM (con ordenamiento 1.3) con datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	967	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	107	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
3	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	107	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	107	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

6.4. SVM para el segundo ordenamiento (Ordenamiento 2.1)

En la Tabla 6.10 se presentan los parámetros obtenidos por la función GridsearchCV mostrada en la Figura 5.3 (validación cruzada) y el tiempo que tomó en obtener el resultado. En la Tabla 6.11 se presenta la matriz de confusión de la SVM para los datos de prueba. En la Tabla 6.12 se presenta la matriz de confusión de la SVM para los datos de entrenamiento.

Tabla 6.10: Parámetros para la cuarta SVM (con ordenamiento 2.1) y tiempo de procesamiento.

Parámetro	Elección
C	10.000
γ	Scale
Tiempo (min)	564

Tabla 6.11: Matriz de confusión de la cuarta SVM (con ordenamiento 2.1) con datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	86497	0	0	0	0	1	0	1	0	0	0	0	7	0	1	0	0
1	29	2	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
2	31	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	0
3	29	2	0	2	0	0	0	0	0	3	0	0	0	0	0	0	0
4	32	0	0	0	3	0	0	0	1	0	0	0	0	0	0	0	0
5	30	0	0	0	0	4	2	0	0	0	0	0	0	0	0	0	0
6	32	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0
7	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0
8	35	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
9	29	0	5	0	0	0	0	0	0	2	0	0	0	0	0	0	0
10	28	1	0	2	0	0	0	0	0	0	5	0	0	0	0	0	0
11	34	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
12	35	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	34	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
14	34	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
15	31	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4	0
16	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla 6.12: Matriz de confusión de la cuarta SVM (con ordenamiento 2.1) con datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	259520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	90	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	90	0	16	0	0	0	0	0	0	2	0	0	0	0	0	0	0
3	94	1	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0
4	99	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0
5	84	0	0	0	0	23	1	0	0	0	0	0	0	0	0	0	0
6	89	0	0	0	0	1	18	0	0	0	0	0	0	0	0	0	0
7	95	0	0	0	0	0	0	3	0	0	0	0	0	0	0	10	0
8	98	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0
9	92	0	1	0	0	0	0	0	0	15	0	0	0	0	0	0	0
10	93	2	0	1	0	0	0	0	0	0	12	0	0	0	0	0	0
11	92	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0
12	97	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
13	96	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0
14	87	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0
15	91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0
16	108	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

6.5. SVM para el segundo ordenamiento usando el promedio de la presión (Ordenamiento 2.2)

En la Tabla 6.13 se presentan los parámetros obtenidos por la función GridsearchCV mostrada en la Figura 5.3 (validación cruzada) y el tiempo que tomó en obtener el resultado. En la Tabla 6.14 se presenta la matriz de confusión de la SVM para los datos de prueba. En la Tabla 6.15 se presenta la matriz de confusión de la SVM para los datos de entrenamiento.

Tabla 6.13: Parámetros para la quinta SVM (con ordenamiento 2.2) y tiempo de procesamiento.

Parámetro	Elección
C	30.000
γ	Scale
Tiempo (min)	2

Tabla 6.14: Matriz de confusión de la quinta SVM (con ordenamiento 2.2) con datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	319	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
1	0	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	35	0	0	0	0	0	0	1	0	0	0	0	0	0	0
3	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	3	0	32	0	0	0	0	0	0	0	0	1	0	0	0
5	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	32	1	0	0	0	0	0	0	3	0
8	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	3
9	0	0	1	0	0	0	0	0	0	35	0	0	0	0	0	0	0
10	0	1	0	1	0	0	0	0	0	0	34	0	0	0	0	0	0
11	0	0	0	0	4	0	0	0	0	0	0	32	0	0	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	0	35	0	0	0	0
13	1	0	0	0	1	0	0	0	0	0	0	0	0	34	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
15	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	28	0
16	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28

Tabla 6.15: Matriz de confusión de la quinta SVM (con ordenamiento 2.2) con datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	965	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
1	0	107	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	105	0	0	0	0	0	0	3	0	0	0	0	0	0	0
3	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	108	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	106	2	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	106	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0	108	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	108	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	108	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	107	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	108	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108	0	0

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
15	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	95	0
16	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

6.6. SVM para el tercer ordenamiento (Ordenamiento 3.1)

En la Tabla 6.16 se presentan los parámetros obtenidos por la función GridsearchCV mostrada en la Figura 5.3 (validación cruzada) y el tiempo que tomó en obtener el resultado. En la Tabla 6.17 se presenta la matriz de confusión de la SVM para los datos de prueba. En la Tabla 6.18 se presenta la matriz de confusión de la SVM para los datos de entrenamiento.

Tabla 6.16: Parámetros para la sexta SVM (con ordenamiento 3.1) y tiempo de procesamiento.

Parámetro	Elección
C	10
γ	Scale
Tiempo (min)	23223 (16d3h3min)

Tabla 6.17: Matriz de confusión de la sexta SVM (con ordenamiento 3.1) con datos de prueba.

Clase	0	6	8	10	14	15	18
0	444957	0	0	0	0	0	0
6	96	0	0	0	0	0	0
8	96	0	0	0	0	0	0
10	96	0	0	0	0	0	0
14	96	0	0	0	0	0	0
15	96	0	0	0	0	0	0
18	95	0	0	0	0	0	0

Tabla 6.18: Matriz de confusión de la sexta SVM (con ordenamiento 3.1) con datos de entrenamiento.

Clase	0	6	8	10	14	15	18
0	1334864	1	2	0	3	1	0
6	280	8	0	0	0	0	0
8	279	0	7	0	0	0	0
10	272	0	0	6	0	0	0
14	268	0	1	0	19	0	0
15	280	0	0	0	0	6	0
18	281	0	0	0	0	0	7

6.7. SVM para el tercer ordenamiento promedio (Ordenamiento 3.2)

En la Tabla 6.32 se presentan los parámetros obtenidos por la función GridsearchCV mostrada en la Figura 5.3 (validación cruzada) y el tiempo que tomó en obtener el resultado. En la Tabla 6.33 se presenta la matriz de confusión de la SVM para los datos de prueba. En la Tabla 6.34 se presenta la matriz de confusión de la SVM para los datos de entrenamiento.

Tabla 6.19: Parámetros para la séptima SVM (con ordenamiento 3.2) y tiempo de procesamiento.

Parámetro	Elección
C	10
γ	Scale
Tiempo (min)	3.210 (2d5h30m)

Tabla 6.20: Matriz de confusión de la séptima SVM (con ordenamiento 3.2) con datos de prueba.

Clase	0	6	8	10	14	15	18
0	444957	0	0	0	0	0	0
6	96	0	0	0	0	0	0
8	96	0	0	0	0	0	0
10	96	0	0	0	0	0	0
14	96	0	0	0	0	0	0
15	96	0	0	0	0	0	0
18	95	0	0	0	0	0	0

Tabla 6.21: Matriz de confusión de la séptima SVM (con ordenamiento 3.2) con datos de entrenamiento.

Clase	0	1	2	3	4	5	6
0	1334871	0	0	0	0	0	0
6	288	0	0	0	0	0	0
8	288	0	0	0	0	0	0
10	288	0	0	0	0	0	0
14	288	0	0	0	0	0	0
15	288	0	0	0	0	0	0
18	285	0	0	0	0	0	0

7. Análisis de los resultados obtenidos

Luego de presentar los resultados se procede a realizar un análisis de estos.

7.1. SVM para el primer ordenamiento (Ordenamiento 1.1)

De la Tabla 6.1 de parámetros obtenidos de la validación cruzada, se puede apreciar que C es bajo, en consecuencia, la SVM posee un margen de un tamaño considerable, por lo que, se debe tener en consideración que, dentro de la clasificación, la SVM acepta errores. El tiempo de procesamiento también es un factor a tener en cuenta, ya que es poco menos de 7 horas.

De la Tabla 6.2 con la matriz de confusión de la primera SVM con los datos de prueba, se puede apreciar la tendencia de esta a indicar que no existe fuga en la red. La clase con un mayor número de verdaderos positivos es la 10 con un 16.67 %, mientras que existen 6 clases (2, 5, 6, 7, 8, 16) con 0 % de verdaderos positivos.

De la Tabla 6.3 con la matriz de confusión de la primera SVM con los datos de entrenamiento, se puede apreciar, nuevamente, la tendencia de esta a indicar que no existe fuga en la red. La clase con un mayor número de verdaderos positivos para este caso es la 12 con un 25.23 %, mientras que la única clase con 0% de verdaderos positivos es la 16, seguido de las clases 7 y 8 con un 1.95 % de verdaderos positivos.

Al ver que no existen muchos aciertos en la detección de la fuga y que la SVM tiende a decir que no existe fuga, por lo que, es posible que la SVM tenga un subajuste, generalizando el comportamiento sin fuga. Se podría mejorar disminuyendo el número de columnas de la matriz.

7.2. SVM para el primer ordenamiento promedio (Ordenamiento 1.2)

De la Tabla 6.4 de parámetros obtenidos de la validación cruzada, se puede apreciar que C es muy alto, en consecuencia, la SVM posee un margen de tamaño pequeño, por lo que, se debe tener en consideración que, dentro de la clasificación, la SVM acepta pocos errores y tiende al sobreajuste. El tiempo de procesamiento es bastante corto, de 7 minutos, siendo de los más cortos.

De la Tabla 6.5 con la matriz de confusión de la segunda SVM con los datos de prueba, se puede apreciar la tendencia a acertar, teniendo un alto porcentaje de verdaderos positivos en todas las clases, obteniendo un 100% para las clases 3 y 14, mientras que las clases con menos verdaderos positivos son las 5 y 6 con un 50.00 % y un 58.33 % respectivamente.

De la Tabla 6.6 con la matriz de confusión de la segunda SVM con los datos de entrenamiento, se puede apreciar, nuevamente, la tendencia de ésta a acertar, teniendo un mayor número de clases con un 100 % de verdaderos positivos (clase 3, 4, 10, 11, 12,

13, 14), mientras que las clases 6 y 7 con un 75.93 % y un 76.85% son las con menor número de verdaderos positivos.

Al ver que existe una gran cantidad de aciertos en la detección de las fugas, es posible que la SVM tenga un sobreajuste a los datos entregados, por lo que, el algoritmo más que aprender a notar las variaciones características en los promedios de las presiones al existir la fuga en los datos, recuerda el comportamiento de los escenarios y si se reproducen de manera idéntica los identifica como fuga. Es por esto que la matriz de confusión con los datos de entrenamiento tiene gran cantidad de clases con 100 % de verdaderos positivos. Se decide rehacer la SVM con un diferente C, en este caso se hace con C = 10000 y C = 1000, obteniendo las matrices de confusión que se muestran en las Tablas 7.1 y 7.2.

Tabla 7.1: Matriz de confusión de la segunda SVM con C = 10.000 con los datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	262	0	8	14	5	5	7	9	13	0	0	0	0	0	0	0	0
1	4	18	0	1	0	0	0	0	0	8	0	1	0	0	0	0	3
2	7	0	27	0	0	0	0	0	0	0	2	0	0	0	0	0	0
3	3	0	0	24	1	0	0	0	0	0	0	8	0	0	0	0	0
4	16	0	0	0	19	0	0	0	0	0	0	0	1	0	0	0	0
5	10	0	0	0	0	9	8	0	0	0	0	0	0	4	5	0	0
6	7	0	0	0	0	4	17	0	0	0	0	0	0	1	7	0	0
7	14	0	0	0	0	0	0	13	9	0	0	0	0	0	0	0	0
8	21	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0
9	1	10	0	5	0	0	0	0	0	12	3	2	3	0	0	0	0
10	13	0	8	0	0	0	0	0	0	0	14	1	0	0	0	0	0
11	5	0	0	11	0	0	0	0	0	0	3	17	0	0	0	0	0
12	9	0	0	0	9	0	0	0	0	4	0	0	13	1	0	0	0
13	2	0	0	0	0	6	2	0	0	0	0	0	2	15	9	0	0
14	2	0	0	0	0	6	7	0	0	0	0	0	0	3	18	0	0
15	17	0	0	0	0	0	0	2	1	0	0	0	0	0	5	10	1
16	21	3	0	0	0	0	0	0	2	0	0	0	0	0	0	3	7

Tabla 7.2: Matriz de confusión de la segunda SVM con C = 1.000 con los datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	281	0	4	15	4	3	3	7	6	0	0	0	0	0	0	0	0
1	8	16	0	3	0	0	0	0	0	8	0	0	0	0	0	0	0
2	10	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	10	0	0	25	0	0	0	0	0	1	0	0	0	0	0	0	0
4	29	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0
5	23	0	0	0	0	5	5	0	0	0	0	0	0	1	2	0	0
6	23	1	0	0	0	0	9	0	0	0	0	0	0	0	3	0	0
7	21	0	0	0	0	0	0	12	3	0	0	0	0	0	0	0	0

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
8	30	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
9	4	9	0	6	0	0	0	0	0	10	3	1	3	0	0	0	0
10	14	0	7	0	0	0	0	0	0	0	14	1	0	0	0	0	0
11	9	0	0	10	0	0	0	0	0	0	3	14	0	0	0	0	0
12	18	0	0	0	0	0	0	0	0	0	4	0	14	0	0	0	0
13	14	1	0	0	0	2	1	0	0	0	0	0	3	14	1	0	0
14	15	1	0	0	0	0	0	6	0	0	0	0	0	0	14	0	0
15	23	0	0	0	0	0	0	0	1	1	0	0	0	0	4	7	0
16	27	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5

De las matrices con diferentes C de esta SVM se observa que al disminuir el C los aciertos de los verdaderos positivos disminuyen, ya que aumenta el margen de aceptación de errores para la SVM, por lo que, existe un mayor número de falsos negativos. Para obtener una SVM que pueda detectar y localizar este tipo de fugas el C necesario es cercano al 30.000 para obtener una buena cantidad de verdaderos positivos y alejarse, aunque sea un poco, del sobreajuste. En los anexos se presentan las matrices de confusión de esta SVM, con los datos de entrenamiento, con C = 1.000 y 10.000 para apreciar que el sobreajuste existente con C = 30.000 disminuye.

7.3. SVM para el primer ordenamiento escenario en 1 fila (Ordenamiento 1.3)

De la Tabla 6.7 de parámetros obtenidos de la validación cruzada, se puede apreciar que C es muy bajo, en consecuencia, la SVM posee un gran margen, por lo que, se debe tener en consideración que, dentro de la clasificación, la SVM acepta una gran cantidad de errores. El tiempo de procesamiento es el menor, siendo de 2 minutos.

De la Tabla 6.8 con la matriz de confusión de la tercera SVM con los datos de prueba, se puede apreciar que clasifica todos los escenarios como sin fuga, por lo que, las clases de la 1 a la 16 tienen 0 % de verdaderos positivos

De la Tabla 6.9 con la matriz de confusión de la tercera SVM con los datos de entrenamiento, se puede apreciar que clasifica todos los escenarios como sin fuga, por lo que, las clases de la 1 a la 16 tienen 0 % de verdaderos positivos

Al ver que no existen aciertos en la detección de la fuga y que la SVM siempre dice que no existe fuga, es posible que, debido al ordenamiento, todo el escenario se encuentra en una sola fila de la matriz, existen una gran cantidad de columnas, por lo que, el comportamiento de la SVM disminuye su precisión. Se podría mejorar disminuyendo el número de columnas de la matriz.

7.4. SVM para el segundo ordenamiento (Ordenamiento 2.1)

De la Tabla 6.10 de parámetros obtenidos de la validación cruzada, se puede apreciar que C es bajo, en consecuencia, la SVM posee un margen de un tamaño considerable, por lo que, se debe tener en consideración que, dentro de la clasificación, la SVM acepta errores. El tiempo de procesamiento también es un factor a tener en cuenta, ya que es poco menos de 10 horas.

De la Tabla 6.11 con la matriz de confusión de la cuarta SVM con los datos de prueba, se puede apreciar la tendencia de esta a indicar que no existe fuga en la red. La clase con un mayor número de verdaderos positivos es la 10 con un 13.89 %, mientras que existen 3 clases (7, 12 y 16) con 0 % de verdaderos positivos.

De la Tabla 6.12 con la matriz de confusión de la cuarta SVM con los datos de entrenamiento, se puede apreciar, nuevamente, la tendencia de esta a indicar que no existe fuga en la red. La clase con un mayor número de verdaderos positivos para este caso es la 5 con un 21.30%, mientras que la única clase con 0% de verdaderos positivos es la 16, seguido de las clases 7 y 4 con un 2.78 % y un 8.33 % de verdaderos positivos respectivamente.

Al ver que no existen muchos aciertos en la detección de la fuga y que la SVM tiende a decir que no existe fuga, por lo que, es posible que la SVM tenga un subajuste, generalizando el comportamiento sin fuga.

7.5. SVM para el segundo ordenamiento promedio (Ordenamiento 2.2)

De la Tabla 6.13 de parámetros obtenidos de la validación cruzada, se puede apreciar que C es muy alto, en consecuencia, la SVM posee un margen de tamaño pequeño, por lo que, se debe tener en consideración que dentro de la clasificación la SVM acepta pocos errores y tiende al sobreajuste. El tiempo de procesamiento es bastante corto, de 7 minutos, siendo de los más cortos.

De la Tabla 6.14 con la matriz de confusión de la quinta SVM con los datos de prueba, se puede apreciar la tendencia a acertar, teniendo un alto porcentaje de verdaderos positivos en todas las clases, obteniendo un 100 % para las clases 1, 3, 5, 6 y 14, mientras que las clases con menos verdaderos positivos son las 15 y 16 con un 77.78 %.

De la Tabla 6.15 con la matriz de confusión de la quinta SVM con los datos de entrenamiento, se puede apreciar, nuevamente, la tendencia de esta a acertar, teniendo un mayor número de clases con un 100 % de verdaderos positivos (clase 1, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14), mientras que la clase 16 con un 92.59 % es la con menor número de verdaderos positivos.

Al ver que existe una gran cantidad de aciertos en la detección de las fugas, es posible que la SVM tenga un sobreajuste a los datos entregados, por lo que, el algoritmo no aprende a notar las variaciones características en los promedios de las presiones al existir la fuga en los datos, sino que recuerda el comportamiento de los escenarios y si se reproducen de manera idéntica los identifica como fuga. Es por esto que la matriz de confusión con los datos de entrenamiento tiene gran cantidad de clases con 100% de verdaderos positivos. Se decide rehacer la SVM con un diferente C, en este caso se hace con C = 10000 y C = 1000, obteniendo las matrices de confusión que se muestran en las Tablas 7.3 y 7.4.

Tabla 7.3: Matriz de confusión de la quinta SVM con C = 10.000 con datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	322	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	0	34	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	29	0	1	0	0	0	0	6	0	0	0	0	0	0	0
3	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	2	0	32	0	0	0	0	0	0	0	0	2	0	0	0
5	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	26	3	0	0	0	0	0	0	5	1
8	2	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	5
9	0	0	4	0	1	0	0	0	0	31	0	0	0	0	0	0	0
10	0	1	0	3	0	0	0	0	0	0	32	0	0	0	0	0	0
11	0	0	0	0	3	0	0	0	0	0	0	33	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0
13	1	0	0	0	1	0	0	0	0	0	0	0	0	34	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
15	1	0	0	0	0	0	0	12	1	0	0	0	0	0	0	21	1
16	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21

Tabla 7.4: Matriz de confusión de la quinta SVM con C = 1.000 con datos de prueba.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	323	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	25	0	6	0	0	0	0	0	0	4	0	0	0	0	0	0
2	2	0	23	0	4	0	0	0	0	5	0	1	0	1	0	0	0
3	1	0	0	29	0	0	0	0	0	0	6	0	0	0	0	0	0
4	9	0	3	0	19	0	0	0	0	0	0	2	0	2	0	0	1
5	2	0	0	0	1	20	12	0	0	0	0	0	0	1	0	0	0
6	0	0	0	0	0	13	23	0	0	0	0	0	0	0	0	0	0
7	7	0	0	0	0	0	0	12	7	0	0	0	0	0	0	6	4
8	9	0	0	0	0	0	0	4	14	0	0	0	0	0	0	0	9
9	3	0	16	0	12	0	0	0	0	5	0	0	0	0	0	0	0
10	1	0	0	10	0	0	0	0	0	0	25	0	0	0	0	0	0

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
11	3	0	0	0	3	0	0	0	0	0	0	29	0	0	0	0	1
12	2	0	0	0	4	0	0	0	0	0	0	0	30	0	0	0	0
13	6	0	0	0	3	0	0	0	2	0	0	0	0	22	0	0	3
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
15	3	0	0	0	0	0	0	18	4	0	0	0	0	0	0	5	6
16	16	0	0	0	0	0	0	0	4	0	0	3	0	0	0	0	13

De las matrices con diferentes C de esta SVM se observa que al disminuir el C los aciertos de los verdaderos positivos disminuyen un poco, ya que aumenta el margen de aceptación de errores para la SVM, por lo que, existe un mayor número de falsos negativos. Para obtener una SVM que pueda detectar y localizar este tipo de fugas el C necesario es cercano al 10.000 para obtener una buena cantidad de verdaderos positivos y alejarse del sobreajuste. En los anexos se presentan las matrices de confusión de esta SVM, con los datos de entrenamiento, con C = 1.000 y 10.000 para apreciar que el sobreajuste existente con C = 30.000 disminuye.

7.6. SVM para el tercer ordenamiento (Ordenamiento 3.1)

De la Tabla 6.16 de parámetros obtenidos de la validación cruzada, se puede apreciar que C es muy bajo, en consecuencia, la SVM posee un gran margen, por lo que, se debe tener en consideración que, dentro de la clasificación, la SVM acepta una gran cantidad de errores. El tiempo de procesamiento ha sido el mayor de todos, siendo de 16 días 3 horas y 3 minutos.

De la Tabla 6.17 con la matriz de confusión de la sexta SVM con los datos de prueba, se puede apreciar que clasifica todos los escenarios como sin fuga, por lo que, todas las clases tienen 0% de verdaderos positivos.

De la Tabla 6.18 con la matriz de confusión de la sexta SVM con los datos de entrenamiento, se puede apreciar que clasifica todos los escenarios como sin fuga, por lo que, todas las clases tienen 0% de verdaderos positivos

Al tener un tiempo de procesamiento tan alto y no obtener un buen número de verdaderos positivos para todas las clases, se procede a desestimar esta SVM para la detección y localización.

7.7. SVM para el tercer ordenamiento promedio (Ordenamiento 3.2)

De la Tabla 6.19 de parámetros obtenidos de la validación cruzada, se puede apreciar que C es muy bajo, en consecuencia, la SVM posee un gran margen, por lo que, se debe tener en consideración que dentro de la clasificación la SVM acepta una gran cantidad de errores. El tiempo de procesamiento ha sido de los mayores, siendo de 2 días 5 horas y 30 minutos.

De la Tabla 6.20 con la matriz de confusión de la séptima SVM con los datos de prueba, se puede apreciar que clasifica todos los escenarios como sin fuga, por lo que, las clases de la 1 a la 16 tienen 0% de verdaderos positivos

De la Tabla 6.21 con la matriz de confusión de la séptima SVM con los datos de entrenamiento, se puede apreciar que clasifica todos los escenarios como sin fuga, por lo que, las clases de la 1 a la 16 tienen 0% de verdaderos positivos

Al ver que no existen aciertos en la detección de la fuga y que la SVM siempre dice que no existe fuga, es posible que debido al ordenamiento existen una gran cantidad de filas, por lo que, el comportamiento de la SVM disminuye su precisión y aumenta el tiempo de procesamiento. Debido a que con los ordenamientos 1 y 2 con el promedio de la presión diaria se obtuvo buenos resultados para las SVM se prueba aumentar el C para ver qué resultados se obtienen con esta SVM con el tercer ordenamiento promedio. Lamentablemente se obtiene los mismos resultados que para el C = 10. En los Anexos se presentan las matrices de confusión para esta SVM, con los datos de entrenamiento, con un C de 10.000 y 50.000.

8. Comentarios y conclusiones

La detección y localización de fugas en redes de agua potable es necesaria para buena gestión del recurso hídrico, ya que el deterioro de la red y las roturas inesperadas es inexorable, por lo que, las investigaciones que impliquen un avance en este tema permiten tener una idea de las metodologías existentes y los elementos necesarios para poder aplicar estas metodologías. Todo esto con el fin de poder reducir desastres generados por la rotura de las tuberías.

La metodología utilizada si bien toma bastante tiempo, tanto en la modelación hidráulica, en la generación de datos y en el procesamiento de la SVM, se logran obtener buenos resultados finales para las segunda SVM (con ordenamiento 1.2) y para la quinta SVM (con ordenamiento 2.2).

Se valida la implementación de algoritmos de aprendizaje supervisado para la detección de fugas en redes de gran tamaño, puesto que existen trabajos precedentes a este que muestran buenos desempeños en la detección de fugas, pero las redes utilizadas son de pequeño tamaño, por lo que, no se tenía conocimiento de cómo se comportarían estos para redes de mayor tamaño.

Para obtener un modelo con fines operacionales o de diseño es necesaria una gran cantidad de datos de la red, datos que no siempre se encuentran informados en la SISS, por lo que, es necesario entrar en contacto con la empresa concesionaria.

En cuanto a la modelación y calibración de la red hidráulica en primer lugar me gustaría comentar que en mi caso el depender completamente, en cuanto a información, de la empresa concesionaria, no fue una buena experiencia. Esto debido a que el tiempo en que se toma en obtener la información. En cuanto a mi caso personal, el flujo de información fue cortado en medio del trabajo de titulación, por lo que, este se vio afectado en que no se pudo obtener una mejor calibración.

Por otro lado, el no poder obtener una calibración con una presión mayor al 73 % se debe a que no se tiene información de la distribución diaria de la demanda diferenciada en el sector. Se tiene la información de salida de los estanques que alimenta el sector modelado y parte de otras dos comunas eliminadas de la Figura 4.2. De manera que, la forma en cómo se distribuyen los flujos de agua en la red a lo largo del día no se representa completamente en el modelo, ya que el consumo durante el día no es el mismo para todo el sector. También la información sobre los edificios del sector no era la óptima para obtener una buena calibración, ya que el número de edificios no estaba definido de un principio, además del uso de patrones de edificios de otras redes. Fuera de esto, un 73 % de precisión en el modelo parece aceptable para la investigación realizada. Ahora, si se quiere aplicar la metodología ocupada a un escenario en la realidad, la calidad y cantidad de los datos para obtener precisiones mejores es bastante alta, e inclusive en algunos casos no deben existir estos datos.

Con el modelo calibrado es necesario generar simulaciones con fugas y sin fugas para poder entrenar y probar el modelo En el caso del estudio se generaron simulaciones de

48 fugas diferentes a 6 horarios diferentes, pero para que la SVM pueda responder bien frente a una fuga en cualquier zona del sector modelado, es necesario crear una fuga en cada nodo (al menos los pertenecientes al segundo ordenamiento), a cada horario diferente lo que generaría alrededor de 15.000 escenarios, lo que aumentaría el tiempo de extracción y procesamiento de datos considerablemente.

En cuanto a la programación del modelo, luego de leer sobre el algoritmo de aprendizaje supervisado, debido a que la librería ya existe, es posible programarlo para alguien de nivel usuario y generar diferentes SVM variando el ordenamiento de los datos y los parámetros de éstas.

En cuanto a los parámetros de la SVM, la validación cruzada siempre escogió “scale” para el parámetro γ, por lo que este siempre representa el inverso del radio de influencia de los escenarios, siendo este último una multiplicación de las características insertadas en la SVM (presión cada 15 minutos) y la varianza de estas características. Por el lado de C, para los ordenamientos 1.1 y 2.2 base se tiene C de 10.000 y una gran cantidad de filas en la matriz (alrededor de 330.000), por lo que al utilizar la SVM, debido al rango de aceptación de ese C, no se obtienen buenos resultados para los verdaderos positivos al generalizar el comportamiento sin fuga, arrojando el comportamiento sin fuga para en muchos de los escenarios de las distintas clases. Para los ordenamientos 1.2 y 2.2 se tiene C de 30.000 y no muchas filas (alrededor de 4.000), por lo que al utilizar la SVM, debido al rango de aceptación de ese C, se obtienen buenos resultados para los verdaderos positivos. Al existir un sobreajuste para cada clase se obtienen buenos resultados de verdaderos positivos para los datos de prueba e inclusive se tiene mejores resultados para los datos de entrenamiento. En el caso del ordenamiento 1.3, debido a la alta cantidad de columnas, el algoritmo no reconoce las fugas e interpreta cada escenario como comportamiento sin fuga. Para corregirlo se podría disminuir el número de columnas de la matriz tal vez promediando la presión diaria por nodo previo a transformar el escenario completo en una columna.

Para la segunda y quinta SVM en los puntos 7.2. y 7.5. respectivamente, se comenta sobre el reentrenamiento de la SVM con diferentes C, donde de un acierto en la detección y localización de la fuga de alrededor del 90% debido al sobreajuste existente por el C de 30.000, se pasa a un acierto en la localización y detección de alrededor del 50% para la segunda SVM con un C de 10.000. Y para la quinta SVM pasa aproximadamente del mismo 90% a un 65% con un C de 1.000 y a un 80% con un C de 10.000. Por lo que, para el correcto funcionamiento de la detección y localización, tratando de evitar el sobreajuste, se recomienda la utilización de un C de 15.000 o 20.000 para la segunda SVM y uno de 5.000 para la quinta SVM.

La primera, cuarta, quinta, sexta y séptima SVM, se desestiman para la detección y localización de fugas, debido a que no se obtiene un buen número de verdaderos positivos sin importar los parámetros adoptados, además de poseer un tiempo de procesamiento bastante alto en comparación a las SVM que si dieron buenos resultados (segunda y quinta). La tercera SVM tampoco obtiene buenos resultados, aunque, su tiempo de procesamiento es igual de corto que las que dieron buenos resultados.

En síntesis, las SVM que arrojan mejores resultados en cuanto en la detección y localización de fugas, además de tener un tiempo de cómputo menor a 30 minutos son la segunda SVM y la quinta SVM, las con los ordenamientos 1.2 y 2.2. En cuanto al tiempo de procesamiento de éstas es de 2 y 7 minutos, respectivamente. Este tiempo es al procesar las SVM en la infraestructura de supercómputo del NHLPC de la facultad, por lo que, si se intentaran procesar en un computador de escritorio su tiempo de procesamiento aumentaría aproximadamente en 10 veces, es decir, la segunda SVM demoraría 20 minutos, mientras que la quinta SVM demoraría 70 minutos.

9. Bibliografía

Adedeji, K. B., Hamam, Y., Abe, B. T., & Abu-Mahfouz, A. M. (2017). Leakage detection and estimation algorithm for loss reduction in water piping networks. Water, 9(10), 773.

Chan, T. K., Chin, C. S., & Zhong, X. (2018). Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection. IEEE Access, 6, 78846-78867.

Christodoulou, S. E., Kourti, E., & Agathokleous, A. (2017). Waterloss detection in water distribution networks using wavelet change-point detection. Water Resources Management, 31(3), 979-994.

Coelho, S. T., Loureiro, D., & Alegre, H. (2006). Modelação e análise de sistemas de abastecimento de água. Lisboa: Laboratório Nacional de Engenharia Civil.

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for machine learning. Cambridge University Press.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan (www. csie. ntu. edu. tw/∼ cjlin/papers/guide/guide. pdf).

Kang, J., Park, Y. J., Lee, J., Wang, S. H., & Eom, D. S. (2017). Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems. IEEE Transactions on Industrial Electronics, 65(5), 4279-4289.

Kemba, J., Gideon, K., & Nyirenda, C. N. (2017, May). Leakage detection in Tsumeb east water distribution network using EPANET and support vector regression. In 2017 IST-Africa Week Conference (IST-Africa) (pp. 1-8). IEEE.

Kim, Y., Lee, S. J., Park, T., Lee, G., Suh, J. C., & Lee, J. M. (2016). Robust leak detection and its localization using interval estimation for water distribution network. Computers & Chemical Engineering, 92, 1-17.

Lang, X., Li, P., Hu, Z., Ren, H., & Li, Y. (2017). Leak detection and location of pipelines based on LMD and least squares twin support vector machine. IEEE Access, 5, 8659-8668.

Laucelli, D., Romano, M., Savić, D., & Giustolisi, O. (2016). Detecting anomalies in water distribution networks using EPR modelling paradigm. Journal of Hydroinformatics, 18(3), 409-427.

Lee, W. M. (2019). Python machine learning. John Wiley & Sons.

Lingireddy, S., & Ormsbee, L. E. (2002). Hydraulic network calibration using genetic optimization. Civil Engineering and Environmental Systems, 19(1), 13-39.

Loureiro, D., Amado, C., Martins, A., Vitorino, D., Mamade, A., & Coelho, S. T. (2016). Water distribution systems flow monitoring and anomalous event detection: A practical approach. Urban Water Journal, 13(3), 242-252.

Mashford, J., De Silva, D., Marney, D., & Burn, S. (2009, October). An approach to leak detection in pipe networks using analysis of monitored pressure values by support vector machine. In 2009 Third International Conference on Network and System Security (pp. 534-539). IEEE.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Nefedov, A. (2016). Support Vector Machines: A Simple Tutorial. svmtutorial. online. URL: https://svmtutorial. online.

Pérez Magrané, R., Sanz Estapé, G., Puig Cayuela, V., Quevedo Casín, J. J., Cugueró Escofet, M. À., Nejjari Akhi-Elarab, F., ... & Sarrate Estruch, R. (2014). Leak localization in water networks: a model-based methodology using pressure sensors applied to a real network in barcelona. IEEE control systems magazine, 34(4), 24-36.

Rajeswaran, A., Narasimhan, S., & Narasimhan, S. (2018). A graph partitioning algorithm for leak detection in water distribution networks. Computers & Chemical Engineering, 108, 11-23.

Salam, A. E. U., Tola, M., Selintung, M., & Maricar, F. (2014, November). A leakage detection system on the Water Pipe Network through Support Vector Machine method. In 2014 Makassar International Conference on Electrical Engineering and Informatics (MICEEI) (pp. 161-165). IEEE.

SISS, S. D. S. S. (2018). Informe de gestión del sector sanitario 2018. SISS. Santiago.

M. Sokolova y G. Lapalme. A systematic analysis of performance measures for classification tasks. Information Processing & Management, vol. 45, no. 4, pp. 427–437. 2009.

Srirangarajan, S., Allen, M., Preis, A., Iqbal, M., Lim, H. B., & Whittle, A. J. (2013). Wavelet-based burst event detection and localization in water distribution systems. Journal of Signal Processing Systems, 72(1), 1-16.

Tao, T., Huang, H., Li, F., & Xin, K. (2014). Burst detection using an artificial immune network in water-distribution systems. Journal of Water Resources Planning and Management, 140(10), 04014027.

Wu, Y., Liu, S., Wu, X., Liu, Y., & Guan, Y. (2016). Burst detection in district metering areas using a data driven clustering algorithm. Water research, 100, 28-37.

Wu, Y., Liu, S., Smith, K., & Wang, X. (2018). Using correlation between data from multiple monitoring sensors to detect bursts in water distribution systems. Journal of Water Resources Planning and Management, 144(2), 04017084.

Zan, T. T. T., Lim, H. B., Wong, K. J., Whittle, A. J., & Lee, B. S. (2014). Event detection and localization in urban water distribution network. IEEE Sensors Journal, 14(12), 4134-4142.

Ziegler, D., Sorg, F., Fallis, P., Hubschen, K., Happich, L., Baader, J., & Knobloch, A. (2011). Guía para la reducción de las pérdidas de agua: Un enfoque en la gestión de la presión. Deutsche Gesellschaft fur Internationale Zusammenarbeit (GIZ), Eschborn, Alemania, 1.

10. Anexos

Anexo A.

Matrices de confusión de la segunda y quinta SVM con C 1.000 y 10.000 con los datos de entrenamiento

A continuación, se presentan las tablas con las matrices de confusión de la segunda y quinta SVM con C igual a 1.000 y 10.000:

Tabla 10.1: Matriz de confusión de la segunda SVM con C = 1.000 con los datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	831	0	24	27	20	15	13	15	22	0	0	0	0	0	0	0	0
1	7	80	2	3	0	0	0	0	0	12	0	1	0	0	0	0	2
2	26	0	79	3	0	0	0	0	0	0	0	0	0	0	0	0	0
3	11	1	0	81	1	0	0	0	0	0	0	14	0	0	0	0	0
4	31	0	0	0	75	0	0	0	0	0	0	0	2	0	0	0	0
5	20	0	0	0	0	54	18	0	0	0	0	0	0	2	14	0	0
6	20	0	0	0	0	8	69	0	0	0	0	0	0	0	11	0	0
7	33	0	0	0	0	0	0	41	34	0	0	0	0	0	0	0	0
8	49	0	0	0	0	0	0	0	59	0	0	0	0	0	0	0	0
9	2	16	0	5	0	0	0	0	0	68	9	2	6	0	0	0	0
10	32	0	17	1	0	0	0	0	0	0	57	1	0	0	0	0	0
11	5	0	0	18	0	0	0	0	0	0	2	83	0	0	0	0	0
12	40	0	0	0	14	0	0	0	0	5	0	2	44	2	0	0	0
13	3	0	0	0	0	12	12	0	0	0	0	0	1	67	13	0	0
14	9	0	0	0	0	6	19	0	0	0	0	0	0	2	72	0	0
15	44	0	0	0	0	0	0	0	21	0	0	0	0	0	4	35	3
16	54	3	0	0	0	0	0	0	5	0	0	0	1	0	0	6	39

Tabla 10.2: Matriz de confusión de la segunda SVM con C = 10.000 con los datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	860	0	20	33	10	8	9	13	14	0	0	0	0	0	0	0	0
1	17	77	0	6	0	0	0	0	0	5	0	0	0	0	0	0	2
2	44	0	63	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	18	2	0	84	0	0	0	0	0	0	0	4	0	0	0	0	0
4	82	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0
5	67	2	0	0	0	26	9	0	0	0	0	0	0	0	4	0	0
6	56	1	0	0	0	0	43	0	0	0	0	0	0	0	8	0	0
7	59	0	0	0	0	0	0	34	15	0	0	0	0	0	0	0	0
8	78	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
9	9	17	0	14	0	0	0	0	0	49	9	3	7	0	0	0	0
10	36	0	15	0	0	0	0	0	0	1	55	1	0	0	0	0	0
11	20	1	0	22	0	0	0	0	0	0	2	63	0	0	0	0	0
12	54	0	0	0	0	0	0	0	0	4	0	2	47	0	0	0	0
13	44	0	0	0	0	5	4	0	0	0	0	0	4	50	1	0	0
14	46	0	0	0	0	0	13	0	0	0	0	0	0	0	49	0	0
15	63	0	0	0	0	0	0	0	11	0	0	0	0	0	4	28	1
16	73	1	0	0	0	0	0	0	0	0	0	0	1	0	0	4	29

Tabla 10.3: Matriz de confusión de la quinta SVM con C = 1.000 con los datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	965	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
1	0	102	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	103	0	2	0	0	0	0	3	0	0	0	0	0	0	0
3	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0	0	0
4	3	0	0	0	105	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	107	0	0	0	0	0	0	0	0	0	0
7	3	0	0	0	0	0	0	85	0	0	0	0	0	0	0	8	6
8	3	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0	7
9	0	0	11	0	1	0	0	0	0	96	0	0	0	0	0	0	0
10	0	0	0	7	0	0	0	0	0	0	101	0	0	0	0	0	0
11	2	0	0	0	0	0	0	0	0	0	0	106	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	107	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	108	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108	0	0
15	0	0	0	0	0	0	0	17	5	0	0	0	0	0	0	81	4
16	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86

Tabla 10.4: Matriz de confusión de la quinta SVM con C = 10.000 con los datos de entrenamiento.

Clase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	966	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	1	84	0	16	0	0	0	0	0	0	6	0	0	0	0	0	0
2	12	0	65	0	23	0	0	0	1	3	0	1	0	1	0	0	2
3	2	0	0	97	0	0	0	0	1	0	8	0	0	0	0	0	0
4	16	0	1	0	77	0	0	0	1	0	0	4	0	4	0	0	5
5	2	0	0	0	0	99	7	0	0	0	0	0	0	0	0	0	0
6	2	0	0	0	0	25	81	0	0	0	0	0	0	0	0	0	0
7	23	0	0	0	0	0	0	39	16	0	0	0	0	0	0	16	14
8	30	0	0	0	0	0	0	7	49	0	0	0	0	0	0	3	19

9	11	0	35	0	12	0	0	0	1	43	0	2	0	2	0	0	2
10	0	0	0	30	0	0	0	0	0	0	78	0	0	0	0	0	0
11	19	0	0	0	4	0	0	0	0	0	0	83	0	0	0	0	2
12	11	0	0	0	7	0	0	0	0	0	0	0	84	3	0	0	2
13	19	0	0	0	5	0	0	0	0	0	0	0	0	80	0	0	4
14	3	0	0	0	0	0	0	0	1	0	0	0	0	0	102	0	2
15	19	0	0	0	0	0	0	33	12	0	0	0	0	0	0	35	8
16	41	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	59

Tabla 10.5: Matriz de confusión de la séptima SVM con C = 10.000 con los datos de prueba.

Clase	0	6	8	10	14	15	18
0	444957	0	0	0	0	0	0
6	96	0	0	0	0	0	0
8	96	0	0	0	0	0	0
10	96	0	0	0	0	0	0
14	96	0	0	0	0	0	0
15	96	0	0	0	0	0	0
18	95	0	0	0	0	0	0

Tabla 10.6: Matriz de confusión de la séptima SVM con C = 10.000 con los datos de entrenamiento.

Clase	0	1	2	3	4	5	6
0	1334871	0	0	0	0	0	0
6	288	0	0	0	0	0	0
8	288	0	0	0	0	0	0
10	288	0	0	0	0	0	0
14	288	0	0	0	0	0	0
15	288	0	0	0	0	0	0
18	285	0	0	0	0	0	0

Tabla 10.7: Matriz de confusión de la séptima SVM con C = 50.000 con los datos de prueba.

Clase	0	6	8	10	14	15	18
0	444957	0	0	0	0	0	0
6	96	0	0	0	0	0	0
8	96	0	0	0	0	0	0
10	96	0	0	0	0	0	0
14	96	0	0	0	0	0	0
15	96	0	0	0	0	0	0
18	95	0	0	0	0	0	0

Tabla 10.8: Matriz de confusión de la séptima SVM con C = 50.000 con los datos de entrenamiento.

Clase	0	1	2	3	4	5	6
0	1334871	0	0	0	0	0	0

6	288	0	0	0	0	0	0
8	288	0	0	0	0	0	0
10	288	0	0	0	0	0	0
14	288	0	0	0	0	0	0
15	288	0	0	0	0	0	0
18	285	0	0	0	0	0	0