UNIVERSIDAD DE GUANAJUATO

División de Ingenierías Campus Guanajuato

Departamento de Ingeniería Geomática e Hidráulica

TESIS:

"APLICACIÓN DE ALGORITMO k-NN PARA LA DETECCIÓN DE FUGAS EN LA RED DE AGUA POTABLE DEL SECTOR LAS HACIENDAS EN VALLE DE SANTIAGO, GTO"

Presenta:

Jairo Eduardo Hernández Ramírez

Que para obtener el grado de:

Ingeniero Hidráulico

Directora de tesis:

Dra. Elizabeth Pauline Carreño Alvarado

Codirector de tesis:

Dr. Gilberto Reynoso Meza, PUCPR, Brasil.

Agradecimientos

A mi familia quien ha confiado en mí, especialmente a mis padres, quienes me han apoyado siempre y que con mucho cariño les dedico cada uno de mis logros. A ti madre donde quiera que te encuentres te agradezco todo lo que me has enseñado y todo el amor que has dejado en mí, sin duda alguna me harás mucha falta el resto de mis días.

QEPD Alicia Ramírez Becerra

A mi novia, por haberme apoyado incondicionalmente desde el principio, por compartir conmigo días de estudio, por todos tus aportes en el área de diseño gráfico.

También expreso mi más sincero agradecimiento a mi tutora y guía de tesis, la Dra. Elizabeth Pauline Carreño Alvarado, por haberme brindado la oportunidad de trabajar con ella y tener la paciencia y vocación de transmitirme su conocimiento de forma incondicional y accesible, además de brindarme su apoyo profesional y personal en momentos difíciles. Al Dr. Gilberto Reynoso Meza quien me asesoró y brindó información en favor del desarrollo de la investigación, por su paciencia en mi incursión en el área de la programación y ciencia de datos, además de su total disponibilidad.

A la Universidad de Guanajuato, por haberme permitido desarrollar profesionalmente en el área que me apasiona.

A todos los docentes que me han apoyado y guiado en el crecimiento académico, sin duda alguna han sido parte fundamental de mi trayectoria.

Al SAPAM por haberme abierto las puertas incondicionalmente, especialmente al director general Ing. Arturo Castillo Serrano, al Coordinador Operativo Ing. Juan Manuel Cuevas Carrillo y la responsable de proyectos Ing. Thanya Yeraldin Rico Pérez, por todas las atenciones hacia nosotros en favor de la investigación.

A todos los compañeros y colegas de carrera, en quienes encontré buenos amigos por compartir todas las experiencias y hacer más ligera la carga de trabajo.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

DEFINICIONES Y SIMBOLOGÍA

k-NN = k-Nearest Neighbors (k-Vecinos más Cercanos)

RDA= Red de distribución de agua

O.O. = Organismo Operador

IA = Inteligencia Artificial

SAPAM = Sistema de Agua Potable y Alcantarillado Municipal

CNA = Comisión Nacional del Agua

INEGI = Instituto Nacional de Estadística y Geografía

SoftSensors = Sensores blandos

Q = Gasto (m^3/s)

lps: Litros por segundo

mca = metros de columna de agua

msnm = metros sobre el nivel del mar

has = hectáreas

” = pulgadas

m = metros

m^3 = metros cúbicos

t = tiempo

hr = horas

mm = milímetros

JUSTIFICACIÓN

Para México, en 1930 la población urbana representaba el 33 por ciento del total, en 2010 aumentó a 78 por ciento y se proyecta que para el 2050 esta cifra sea de una cantidad de 121 millones de personas que estén viviendo en metrópolis más concentradas y complejas (Salazar A. et Al. 2015).

Aunque las autoridades aseguran que más del 90% de la población tiene acceso al agua potable y que una parte un poco mejor tiene conexiones al alcantarillado, la realidad es que se está sufriendo grandes estragos por la inadecuada disponibilidad en calidad y cantidad (Barkin D. 2006).

Con este crecimiento en la población y las consecuencias del cambio climático, los usos del agua han ido cambiando de la misma forma en que las necesidades de los usuarios cambian, en estaciones cálidas se abusa del uso de agua para refrescarse o en actividades recreativas, aunado a todo lo que ya es un problema la epidemia de la covid-19 ha provocado un aumento en el volumen de agua consumida debido al obligado aumento en la limpieza (Camacho A. 2020); debido a estos cambios, la cultura del cuidado del agua juega un papel muy importante.

Si continuamos con estos hábitos de explotación de este recurso la posibilidad de continuar satisfaciendo las necesidades de los usuarios se verá cada vez más comprometida y comenzarán a surgir problemas cada vez más graves, si los O.O. (Organismos Operadores) hacen lo posible por conservar el agua, los consumidores tenderán a ser más cooperadores en otros programas de conservación (Zachaira M. 2009).

El agua es uno de los recursos más presentes en los seres vivos, vital para el desarrollo de diversos procesos y actividades; aunque el total del agua presente es constante, su disponibilidad no lo es, es por ello que se debe mantener clara la importancia de llevar una gestión adecuada y sostenible de su consumo que evite su agotamiento y el estrés hídrico.

En México la mayoría de los O.O. no cuentan con los medios que faciliten la detección de fugas, ya que estas representan una pérdida de energía, dinero y del recurso mismo, así que contar con métodos de detección de anomalías es muy importante para que puedan brindar un mejor servicio a los consumidores.

Llevar a cabo un mantenimiento preventivo ayuda a la conservación de la infraestructura, equipos e instalaciones, garantizando un buen funcionamiento y fiabilidad en sus operaciones, reduciendo las posibilidades de tener fallas; un O.O. con infraestructura en malas condiciones verá afectada su

capacidad de dotar de agua potable. En grandes ciudades los organismos llegan a contar con mejor infraestructura y dispositivos de regulación y control que optimizan la dotación del agua potable, sin embargo, también hay organismos de pequeñas ciudades que abastecen a la población con la mínima infraestructura y que por diversas circunstancias no tienen la posibilidad de mejorar.

Las tuberías antiguas, pobremente construidas, el inadecuado control de la corrosión, el mantenimiento pobre de válvulas y el daño mecánico son algunos de los factores contribuyentes a las fugas. En el caso de rupturas de tuberías tienen efectos importantes en el suministro de agua, además de la posible interrupción del servicio puede provocar efectos inesperados como sótanos inundados, derrumbes de carreteras o daños a instalaciones.

Contar con métodos de detección de anomalías que requieran datos que la misma red posee como la presión y caudal es una oportunidad para cualquier organismo, pero especialmente para aquellos que presentan problemas similares, contar con técnicas de detección temprana de anomalías trae muchos beneficios, tratándose de agua potable representa una oportunidad de hacer más eficiente la distribución.

Debido a la problemática detectada en el O.O. SAPAM, el presente trabajo está enfocado en la investigación de nuevos métodos y herramientas de detección de anomalías en redes de distribución de agua potable y su posterior aplicación, teniendo en cuenta que la situación de este organismo operador en cuestión de cantidad de datos e información de la RDA disponible son escasos y en algunos casos discontinuos.

OBJETIVOS

Objetivo General

Proponer un método de detección de fugas que utilice los datos que cualquier organismo operador pueda obtener de su red de distribución como pueden ser caudales y presiones; en efecto se utiliza un algoritmo de clasificación de *machine learning* que analiza diferentes modelos hidráulicos e identifica anomalías en el funcionamiento de estos, los ajustes para su aplicación dependen de la disponibilidad de información y condiciones en que se encuentra la RDA; aparte de ello los conocimiento científicos para evaluar y analizar que aporta la ingeniería hidráulica en esta investigación, son clave en el manejo, control y preservación de los recursos de agua.

Con lo anterior se pretende dar ejemplo para aquellos organismos que no cuentan con los recursos y necesiten satisfacer la demanda del usuario de forma eficiente; siendo esta investigación un precedente dentro del organismo para la continuidad en este sentido de mejorar la eficiencia física a través de nuevas técnicas de detección de anomalías.

Objetivos Específicos

1. Realizar el modelado hidráulico del caso de estudio y su calibración
2. Analizar el comportamiento del modelo y compararlo con el comportamiento de la red
3. Proponer soluciones a los problemas detectados en la red de distribución derivadas de la aplicación del método de detección de anomalías
4. Detectar mediante la modelación hidráulica las zonas que provocan problemas
5. Contribuir en la optimización del sector trabajado y ayudar en la mejora de la calidad del servicio
6. Emitir recomendaciones de operación para el organismo.

CAPÍTULO I

MARCO TEÓRICO

INTRODUCCIÓN

A pesar de la escasez de recursos hídricos, el crecimiento poblacional y los efectos del cambio climático, hoy en día hasta el 50% del agua producida se pierde en la distribución, debido a las deficiencias en infraestructura, o no es contabilizada ni se factura de forma adecuada, estas pérdidas representan una insolvencia financiera para los O.O. (Salazar A. et Al. 2015).

En algunos países, las fugas de los sistemas de abastecimiento de agua llegan a ser del 50% de la cantidad ingresada a la red para satisfacer sus necesidades hídricas; esto implica pérdidas económicas de importancia y un mal aprovechamiento de los recursos naturales. En Malasia el porcentaje de las fugas es del orden del 40%, en Brasil y Suecia del 25% y en México 39% (Mariles O. et Al. 2011).

Una fuga es una salida de agua no controlada en cualquiera de los componentes del sistema de distribución de agua potable; con mayor frecuencia ocurren en uniones de tuberías, codos, roturas de conductos y válvulas que en muchas ocasiones son provocadas por un golpe de ariete (un caso de particular interés en el tema de fugas). Abordar estas pérdidas debido a fugas es necesario para mejorar la eficiencia en los sistemas de abastecimiento.

El efecto principal de la fuga es la pérdida del recurso hídrico, en consecuencia, la reducción de la presión en los sistemas de abastecimiento. El elevar las presiones para compensar tales pérdidas incrementa el consumo de energía y ese aumento en presión empeora las fugas y tiene un impacto negativo sobre el medio ambiente (Zachaira M. 2009).

Sin duda alguna detectar y reparar las fugas de agua ha adquirido mayor importancia en los O.O., estas fugas no solo significan una pérdida de recurso financiero, sino que por otra parte la conservación de las fuentes de agua es indispensable si se quiere garantizar su disponibilidad en un futuro.

El objetivo principal de la detección de fugas es en primer lugar eliminar la fuga, además de mitigar los problemas antes mencionados, aumentar el rendimiento de los sistemas de abastecimiento y aumentar la calidad del servicio al cliente y disminuir el abatimiento de pozos provocado por el uso desmedido y volúmenes perdidos de agua. Para un O.O. detectar y atender fugas a tiempo, representa, una disminución de volúmenes de agua perdidos, que a su vez prolonga la vida útil de los cuerpos de agua de donde se abastece y así garantiza la dotación a futuro; además una

infraestructura en mejores condiciones, tiene mejor eficiencia, que genera un bienestar y satisfacción respecto al servicio para el usuario disminuyendo proporcionalmente las quejas y posiblemente las tomas clandestinas.

Muchas veces los puestos directivos en los O.O. son ocupados por personas que desconocen el sector y obtienen el puesto como premio a su lealtad política al ser cercanos a los presidentes municipales o gobernadores, estas deficiencias en el personal directivo y operativo conducen a un mantenimiento insuficiente de la red y malas prácticas en la atención al usuario que se manifiestan en una disminución de la calidad del servicio, pérdidas de agua y reducciones en la recaudación (Alejandro S. et Al. 2016).

La tarea de satisfacer el abastecimiento de agua, con la creciente demanda debido al crecimiento poblacional, ha hecho evolucionar a la ingeniería y a las empresas dedicadas a la distribución de agua potable, ya que siempre se ha visto que para el diseño y el cálculo de estos sistemas físicos se requerían complejos y tediosos métodos de cálculo hidráulico.

REDES DE DISTRIBUCIÓN

Se le denomina red de distribución al conjunto de elementos que se encarga del transporte del agua, desde los puntos de producción y/o almacenamiento hasta los puntos de consumo, por ejemplo, viviendas, comercio, industria, servicio público, entre otros.

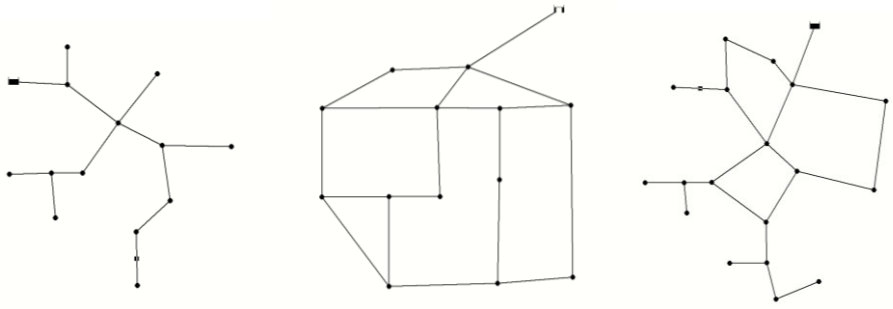
La red debe proporcionar este servicio todo el tiempo, en cantidad suficiente, con la calidad requerida y a una presión adecuada, esta se constituye principalmente de tuberías y cuenta con otros elementos especiales, accesorios y estructuras que deben de ser dimensionados y seleccionados adecuadamente para garantizar el correcto funcionamiento de la RDA (V. Tzarchkov. 2007).

TIPOS DE REDES DE DISTRIBUCIÓN

Las redes de agua se pueden clasificar con base en diferentes criterios, según el uso de agua, la topología o por el sistema de alimentación:

1. Según el uso del agua: es común encontrar una única red que se encarga de cubrir todas las necesidades, sin embargo, existen casos en los cuales se tiene una red para un objetivo específico, su ventaja en algunos casos aprovechar otras fuentes de suministro de menor calidad y tener el control total del objetivo para la que fue hecha, algunos tipos de redes clasificadas según el uso de agua pueden ser:
    - Redes generales de suministro en zonas urbanas, la misma red se encarga del suministro doméstico, industrial, riego, entre otros
    - Redes de suministro para zonas residenciales
    - Redes de suministro en zonas industriales
    - Redes exclusivas de riego
    - Redes para uso exclusivo de extinción de incendios.

2. Según su topología: estas redes se clasifican en función de cómo están conectadas las tuberías entre sí, entonces podemos distinguir tres tipos de redes, el uso de algún tipo de estas redes dependerá de las condiciones topográficas de la población, cantidad de habitantes y distribución de las viviendas:
    - Abierta o ramificada
    - Cerrada o malladas
    - Mixtas.



Abierta Cerrada Mixta

Figura 1 Tipos de redes de distribución de agua

Según el sistema de alimentación: este se basa en la forma en que se adopta el sistema de alimentación a la RDA, estas se pueden alimentar desde:

- Depósitos a presión atmosférica, elevados o a nivel superficial
- Inyección directa mediante sistemas de bombeo constantes o intermitentes.

También podemos encontrar combinación de estos dos tipos, como los cárcamos de rebombeo o tanques secundarios que aportan caudal a zonas más alejadas.

FUNCIONAMIENTO DE UNA RED DE ABASTECIMIENTO

Una red de abastecimiento consiste en una serie de tuberías, codos, piezas especiales etc. unidos entre sí, comúnmente subterráneas, cuyo objetivo es entregar agua hasta la entrada de los predios de los usuarios.

Un sistema de abastecimiento de agua se compone de instalaciones para la captación, conducción, bombeo, tratamiento y distribución. Las obras de captación y almacenamiento permiten reunir las aguas aprovechables de ríos, manantiales y subterráneas; la conducción engloba a los canales y acueductos, así como las instalaciones complementarias de bombeo y piezas especiales para transportar el agua desde la fuente hasta el centro de distribución y finalmente realizar la distribución, que es dotar de agua al usuario final para su consumo (V. Tzarchkov. 2007).

A continuación, se enlistan los elementos que se encuentran comúnmente en una RDA:

- Fuente de agua: agua de origen de ríos, arroyos, lagos, embalses, manantiales y subterránea que proporciona agua a los suministros públicos de agua potable y a pozos de agua
- Tanque de regulación: depósito situado generalmente entre la captación y la red de distribución, cuyo objetivo es almacenar el agua proveniente de la fuente y guardar cierto volumen adicional de agua para las horas del día en que la demanda de la red sobrepasa el volumen suministrado
- Línea de alimentación: es el conjunto formado por tubos y su sistema de unión o ensamble también denominados nudos o uniones, de acuerdo con su función se puede dividir en primaria y secundaria

- Red Primaria: permite conducir el agua por medio de líneas troncales o principales y alimentar a las redes secundarias, generalmente se constituye de tubos de mayor diámetro y esta puede ir desde el tanque de regulación hasta donde inicie su distribución
- Red secundaria: distribuye el agua propiamente hasta las tomas domiciliarias, estas son las de menor diámetro y abarcan la mayoría de las calles de la localidad
- Piezas especiales: accesorios que se emplean para llevar a cabo ramificaciones, intersecciones, cambios de dirección, modificación de diámetro, uniones de diferente material
- Válvulas: accesorios para disminuir o evitar el flujo en las tuberías, estas se pueden clasificar en válvulas de aislamiento o seccionamiento y son usadas para separar o cortar el flujo del resto del sistema; y válvulas de control, usadas para regular el gasto o presión
- Tomas domiciliarias: conjunto de piezas y tubos que permite el abastecimiento desde la red de distribución hasta el predio, así como la instalación de un medidor
- Caja rompedora de presión: depósitos con superficie libre del agua y volumen pequeño, permite que el flujo de la tubería se descargue en este, eliminando presión hidrostática.

PROBLEMAS DE UN ORGANISMO OPERADOR

Muchos de los organismos operadores en México presentan deficiencias en la cobertura de los servicios, su eficiencia física y comercial, y en materia de autonomía técnica y financiera. Esto se debe a la combinación de una serie de factores que limitan su potencial (CONAGUA. 2012). Además de los aspectos técnicos y de gestión, los O.O. enfrentan retos adicionales relacionados con el crecimiento de la población, el incremento de la demanda del servicio y los conflictos sociales por el agua. Estos problemas pueden crear una posición endeble que afecta la fortaleza y estabilidad de los organismos para enfrentar y resolver la problemática que se presenta durante la prestación del servicio.

Las principales dificultades de un Organismo Operador son:

- Insuficiencia de recursos económicos
- Falta de continuidad de sus administraciones y profesionalización del personal
- Deficiencia en las gestiones de organización, técnica y comercial

- Rigidez en los esquemas de autorización de tarifas
- Endeudamiento excesivo por falta de liquidez
- Baja o nula disposición de pago por parte de los usuarios
- Politización de las decisiones y de los programas operativos
- Estructuras y niveles tarifarios que no reflejan los costos reales de operación para la prestación del servicio
- Estado crítico de la infraestructura
- Disminución en la disponibilidad de agua o un difícil acceso a las fuentes de agua cada vez mayor.

La debilidad y desventaja provocadas por los problemas a los que se enfrentan, ocasionan que los usuarios con toma domiciliaria no reciban con normalidad el servicio y en ocasiones a manera de protesta algunos usuarios no realizan su pago, agudizando de esta forma el efecto de un bajo desempeño y una mala calidad en la prestación de los servicios de abastecimiento de agua.

Económicamente las principales complicaciones a las que se enfrentan los O.O. es que los ingresos que tienen son muy inferiores a sus costos operativos, lo que ocurre debido a que entre el 25 y 30 por ciento del agua que se logra suministrar no se cobra, tienen más empleados de los necesarios para operar el sistema, además de las cuantiosas cantidades que se fugan y las tomas clandestinas (Alejandro S. et Al. 2016).

IMPORTANCIA DE LA DETECCIÓN DE FUGAS

Las fugas de agua afectan de forma inmediata y a corto plazo en la distribución de agua de los organismos, debido a que éstas provocan un desequilibrio en la red afectando de forma inmediata a los consumidores y que en las tomas domiciliarias no llegue con suficiente presión o en el peor de los casos, que se vea interrumpido el servicio.

Aunque en las redes de agua potable no se puede evitar que existan fugas y prevenirlas no es una tarea fácil, es necesario llevar a cabo acciones permanentes encaminadas a disminuir el número de éstas; una complicación en la detección de fugas es que en su mayoría no se encuentran visibles, por lo que es necesario contar con alguna herramienta para estimar su localización (Mariles O. et Al. 2011). Actualmente existen nuevas herramientas y equipos que ayudan a detectarlas, además,

es urgente que se comience a tomar medidas para disminuir el impacto que estas provocan sobre los cuerpos de agua de los que se abastece a la población, las sequías extremas son cada vez más frecuentes y muchos organismos no están preparados para enfrentar problemas de esta índole.

La detección y atención temprana de este tipo de anomalías, trae consigo varios beneficios para los O.O. como: ahorro económico, conservación de los cuerpos de agua, satisfacción en los usuarios por un servicio más eficiente usado en el abastecimiento.

Al tener un adecuado manejo de estas pérdidas físicas de agua debido a fugas, se evita (Montoya L. et Al. 2012):

- Bombear volúmenes suplementarios para satisfacer la demanda
- Introducción de aire debido a fugas, este provoca oxidación en el sistema e irregularidades en los medidores de caudal
- Daños en la cimentación de construcciones
- Disminución de presión en el servicio
- Aumento en la probabilidad de contaminación.

MÉTODOS Y TÉCNICAS DE DETECCIÓN DE FUGAS

Existen diferentes métodos y herramientas o equipo que se emplean en la detección de fugas en sistemas de distribución de agua. El uso de un método dependerá de diferentes factores como pueden ser: la situación en que se encuentre la red, su configuración, los materiales de ésta, la tecnología o equipo con la que cuente el O.O., por mencionar algunos.

Entre los métodos existentes podemos mencionar:

- Técnicas visuales en las que se puede identificar encharcamiento en la superficie del suelo o crecimiento anormal de vegetación
- La auscultación directa con aparatos de amplificación de sonido, micrófonos a tierra o geófonos
- Técnicas de radar
- Técnicas eletromagnéticas para identificar roturas en tuberías metálicas
- Inyección de gases trazadores

- Análisis de fotografías infrarrojas
- Balances de masa.

A continuación, se explica de forma breve en qué consisten algunas de las técnicas y métodos comúnmente utilizados en la detección de fugas:

Métodos

Detección de fugas por sonido (ruido)

Cuando el agua a alta presión sale por una grieta o junta de una tubería, genera ruido por el choque del agua con el material que le rodea, utilizar un equipo móvil de micrófonos es muy útil para encontrar el punto exacto de una fuga, no siempre suele ser un método muy efectivo como la monitorización continua, este método acústico de detección utiliza *Geófonos,* equipos de análisis de ruido de 30 a 3000 Hz.

Figura 2 Geófono

Medición de vibración

Las fugas además de provocar ruido, con la constante presión con la que sale el agua hace que la tubería vibre considerablemente, medir el nivel de vibración en los extremos de los tramos indica si ha ocurrido una fuga, un punto favorable de este método es que solo se requiere de un punto de monitorización al no calcular ninguna diferencia entre segmentos. Los cambios de presión causados por una fuga son acompañados por un cambio en la amplitud de la respuesta de vibración

de la superficie del ducto en un patrón específico. Las fluctuaciones de presión son proporcionales a la aceleración superficial del ducto y se pueden medir con un analizador de vibraciones.

Figura 3 Analizador de vibración

Visión termográfica

La visión termográfica es una tecnología muy útil para encontrar de forma certera el punto de la fuga en una tubería. Las cámaras termográficas son dispositivos que detectan y miden la energía infrarroja de los objetos, detectando en tiempo real la irradiación del agua cuando pasa por la tubería y así indica donde se encuentra la fuga. Sin embargo, no siempre existen las condiciones ideales para realizar este tipo de lecturas, por lo que es necesario tener en cuenta la capacidad limitada en la identificación fugas cuando estas no manifiestan la temperatura necesaria, además de los reflejos solares, estos pueden dar lecturas poco confiables.

Figura 4 Cámara termográfica

Nivel de humedad en la superficie

Existe la posibilidad de medir constantemente el nivel de humedad del suelo situado en la superficie de la tubería, aunque presenta complicaciones al tener que considerar la meteorología, y siendo la observación la forma más sencilla de detectar una fuga, suele utilizarse un medidor de humedad, de esta forma se introduce el instrumento en la zona que se cree que existe una fuga emergente, otra forma es por la observación del flujo de agua en la superficie del suelo, esto es indicador de que debajo existe una ruptura en la tubería.

Sin embargo, una desventaja es que, debido a la ubicación de la tubería y composición del suelo, el agua que se fuga se filtre y no sea perceptible la existencia de la fuga.

Figura 5 Medidor de humedad

Técnicas

Balance de masas

El balance de masas mide el volumen 'entrando' y 'saliendo' de la línea y cuando se instala cualquier transmisor de flujo en cualquier ducto, las medidas de flujo serán diferentes. Por este motivo, cuando se realiza la detección de fugas utilizando solamente el balance de masas siempre habrá una diferencia de flujo en el ducto.

Vin = Vout; el balance debe ser cero

En sistemas cerrados el principio de conservación de la masa se utiliza en forma implícita, ya que requiere que la masa del sistema permanezca constante durante un proceso. En el caso de volumen de control (VC), sin embargo, la masa no puede cruzar las fronteras, por lo que se debe seguir con

atención la cantidad de masa que entra y sale del volumen de control, a menudo, en la mecánica de fluidos, a la Ecuación [1] se denomina ecuación de continuidad.

$$\Sigma m_{en} - \Sigma m_{sal} = \Delta m_{VC} \quad [1]$$

Donde los subíndices *en, sal y VC* representan entrada, salida y volumen de control respectivamente.

Inteligencia artificial

Para muchas empresas la detección de anomalías es clave, su supervivencia puede ser determinada por la misma, en ocasiones la detección de anomalías tarda semanas, meses o hasta años, y al analizar balances e indicadores, estas anomalías ya pudieron haber causado estragos, es difícil saber qué pasa exactamente con los datos, y es allí donde entra la inteligencia artificial IA, el 85% de las empresas están de acuerdo en que la IA les ayudará a obtener una ventaja competitiva, ya que la IA se está convirtiendo en una tecnología cada vez más adoptada (Zaragoza G. 2020).

El aprendizaje de máquina (*machine learning)* es una disciplina del campo de la inteligencia artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo). En los últimos años las técnicas de *machine learning* se han hecho lugar dentro de los sistemas tradicionales de detección de anomalías debido a sus ventajas, entre ellas, la posibilidad de desarrollar algoritmos adaptativos a nuestra red y sus modificaciones (Estévez Pereira and Julio-Jairo, 2020), además ha demostrado que puede ser una gran herramienta, es un enfoque basado en datos bien procesados que puede encontrar relaciones y patrones complejos, después de un proceso de aprendizaje supervisado.

Estas técnicas se clasifican en supervisadas, no supervisadas y semisupervisadas, en el caso de las supervisadas el objetivo es entrenar un modelo sobre datos de entrada y salida conocidos para que pueda predecir resultados futuros, para el caso de las no supervisadas encuentra patrones ocultos en los datos de entrada y en las semisupervisadas son un término medio, donde, los datos se dividen en dos grupos, clasificados y no clasificados (Viera A. 2017).

Así pues, las máquinas de aprendizaje han demostrado ser una herramienta poderosa para diferentes propósitos, siendo su aprender patrones basado en datos conocidos de entrada y salida requieren un conjunto de entrenamiento para que ésta será capaz de identificar eventos anómalos de eventos normales. Cuando se utiliza para la clasificación el aprendizaje supervisado es la opción más popular, para ello los algoritmos se entrenan para clasificar datos sea binaria o multiclase (Carreño-Alvarado et Al. 2017).

Dentro del campo del *machine learning* podemos encontrar algoritmos de aprendizaje automático aplicados en la detección de anomalías, en este caso, se ha seleccionado el algoritmo k-NN vecinos más cercanos que pertenece a los algoritmos de aprendizaje supervisado.

El algoritmo k-NN (k-Nearest Neighbors), es un algoritmo de clasificación esencial en *Machine Learning,* catalogado como clasificador basado en instancias, es decir, para clasificar compara las instancias no vistas con aquellas etiquetas del conjunto de entrenamiento utilizando una función de similitud, generalmente la similitud es medida mediante una función de distancia (Maillo J, et Al. 2018). En otras palabras, memoriza las distancias de formación para usarlas en la fase de predicción.

Este método recolecta una serie de muestras etiquetadas que contienen mediciones de algunas características de los objetos o eventos que se desea clasificar, luego, para clasificar nuevas muestras se usan medidas de distancia para identificar cuál es la muestra del conjunto inicial, denominado conjunto de entrenamiento.

Para obtener estos dos conjuntos de datos, primero se divide el conjunto total de datos en dos, el conjunto número uno llamado conjunto de entrenamiento, en la fase de entrenamiento se almacenan las etiquetas de las clases de dichos elementos de entrenamiento, el conjunto número dos, llamado conjunto de prueba o *test*, es utilizado para la fase de predicción y se seleccionan los k elementos más cercanos, es decir, el modelo de *machine learning* ha aprendido el patrón que hay entre los datos de entrada y los resultados, se puede confiar en que el modelo va a funcionar bien con datos nuevos siempre que los datos nuevos vengan en una distribución similar.

Considerando un conjunto de dos categorías y se requiere clasificar un nuevo elemento que se encuentra en cierta región como se muestra en la Figura 6.

Tomando como ejemplo, si decidimos que K=3 es decir que busque los 3 vecinos más cercanos al nuevo elemento, el resultado sería con 2 vecinos de la clase 2 y 1 vecino en la clase 1. Por lo tanto,

al ser la clase 2 en la que más vecinos se encontraron cercanos al nuevo elemento, dicho elemento se asigna a la clase 2.

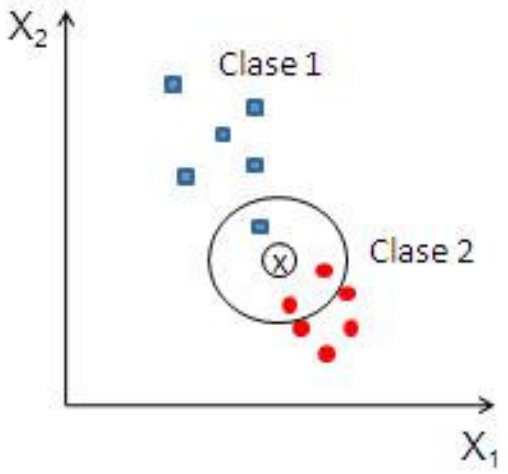


Figura 6 Nuevo elemento en el conjunto de datos

El algoritmo k-NN sigue los siguientes pasos para determinar a qué categoría pertenece el nuevo elemento en el conjunto de datos:

- Selecciona el número de *k* vecinos más cercanos
- Toma los *k* vecinos más cercanos al nuevo elemento de acuerdo con la métrica utilizada para medir la distancia entre elementos
- Entre los *k* vecinos, contar el número de elementos que pertenece a cada categoría
- Asigna el nuevo elemento a la categoría donde se encontraron más vecinos

Las reglas de clasificación por vecindad están basadas en un conjunto de prototipos de los *k* prototipos más cercanos al patrón a clasificar, se le conoce como mecanismo de aprendizaje perezoso (Cambronero, C. G., & Moreno, I. G. 2006).

La clasificación por vecindad más simple es la del vecino más cercano, o simplemente 1-NN. Se basa en la suposición de que la clase del patrón a etiquetar, X, es la del prototipo más cercano

en R (Conjunto de Referencia) al que se notará como X_{NN}.

En la fase de clasificación, el algoritmo recibe nuevas muestras de clase desconocida y debe asignarles una de las *q* clases de los datos de entrenamiento mediante algún proceso de inferencia. El algoritmo k-NN puede enmarcarse dentro de la teoría de decisión bayesiana, de modo que la clasificación de las nuevas observaciones se basa en hallar la clase con la mayor probabilidad a posteriori $P(c_j \mid \mathbf{x})$, j=1, 2, …… q, donde **x** es el vector de características de la muestra a clasificar y c_j representa la *j-ésima* clase (Santos et Al. 2019).

Existen diferentes distancias que se pueden utilizar para determinar los vecinos más cercanos, algunas de ellas se muestran a continuación:

- Distancia euclidiana
- Distancia Minkowski
- Distancia Cosenoidal
- Distancia Mahalanobis.

Distancia Euclidiana: La distancia euclidiana es la distancia ordinaria entre dos puntos en un espacio euclídeo, es decir, el espacio geométrico donde se satisfacen los axiomas de Euclides el cual se deduce del teorema de Pitágoras, que describe que la suma de los cuadrados de los catetos es igual al cuadrado de la hipotenusa.

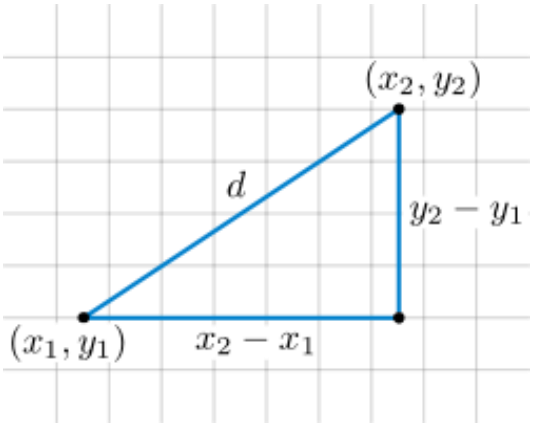


Figura 7 Distancia en un sistema de coordenadas cartesianas

$$d_E(P_1,P_2) = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2} \qquad [2]$$

La Ecuación [2] describe la distancia entre dos puntos ubicados sobre una recta como la raíz cuadrada de la suma de los cuadrados de la diferencia de sus coordenadas en X e Y.

Distancia Minkowski: Es la métrica de un espacio vectorial normado que puede considerarse una generalización de las distancias euclídea y Manhattan, es decir, es la distancia que se usa para medir la diferencia entre dos vectores, comúnmente aplicada en algoritmos de aprendizaje automático, la Ecuación [3] describe la distancia de Minkowsky:

$$d(x.y)=\left(\sum_{i=1}^{n}\left|x_i-y_i\right|^p\right)^{\frac{1}{p}} \qquad [3]$$

La distancia de Minkowsky calcula las diferencias entre los vectores X e Y resultando el vector (X – Y), y cada uno de los valores se elevan a p.

La distancia de Minkowsky se usa típicamente con p=1 o p=2, que corresponden a la distancia de manhattan y la euclidiana, en el caso de que p alcanzando el infinito se obtiene la **distancia Chebyshev.**

Distancia Cosenoidal: Es una medida de la similitud existente entre dos vectores en un espacio que posee un producto interior con el que se evalúa el valor del coseno del ángulo comprendido entre ellos, esta función trigonométrica proporciona el valor igual a 1 si el ángulo comprendido es cero, es decir, si ambos vectores se encuentran apuntando al mismo lugar.

Usa el coseno del ángulo entre dos vectores en un espacio vectorial como una medida de la diferencia entre dos individuos. La distancia cosenoidal se centra más en la diferencia en la dirección de los dos vectores que en la longitud (Sidorov et Al. 2014), se expresa de la siguiente forma:

$$\mathrm{Cos}(a,b)=\cos\theta=\frac{\vec{a}*\vec{b}}{\|a\|*\|b\|} \qquad [4]$$

Se supone que $\|A\|$, $\|B\|$ representan las 2 normas de los vectores A, B. La siguiente Figura 8, se representa gráficamente los vectores A, B:

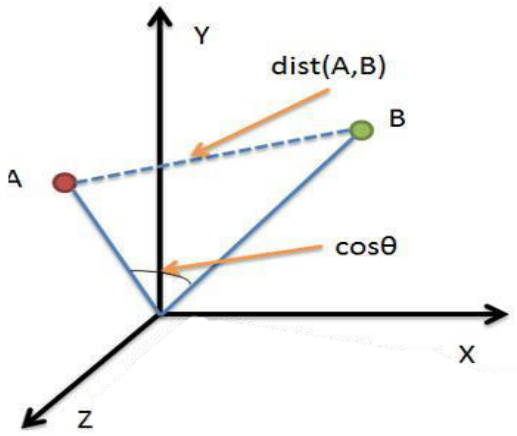


Figura 8 Representación de los vectores A y B, Cosθ y distancia (A, B).

Distancia Mahalanobis: Es una medida de distancia alternativa a las distancias euclídeas, su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales, teniendo en cuenta la correlación entre ellas (McLachlan. 1999). La distancia de Mahalanobis se define foralmente como:

$$d_m=(\vec{x},\vec{y})=\sqrt{(\vec{x}-\vec{y})^T\ \sum^{-1}(\vec{x}-\vec{y})} \qquad [5]$$

Es la distribución de probabilidad del vector $\vec{x}$ e $\vec{y}$ con la matriz de covarianza $\sum$, es decir, la distancia de $(\overrightarrow{x},\vec{y})$ es igual a la raíz de vector x menos vector y elevado a la T por la covarianza a la -1 por (x – y)

Matriz de confusión

La matriz de confusión es una tabla de contingencia que sirve como herramienta estadística para el análisis de observaciones emparejadas, esta ha sido adoptada como un estándar para informar sobre la exactitud temática de cualquier producto de datos derivados de la teledetección. El contenido de una matriz de confusión es un conjunto de valores que contabilizan el grado de semejanza entre observaciones emparejadas: un conjunto de datos bajo control y un conjunto de datos de referencia, para los que se ha establecido una clasificación (López et Al 2018).

La matriz de confusión también conocida como matriz de error, es una tabla que permite visualizar el rendimiento de un algoritmo, comúnmente uno del área de aprendizaje supervisado. Cada una de las filas de la matriz representa las instancias de la clase real y las columnas representan las predicciones o viceversa. El nombre de esta matriz tiene su origen en el hecho de que facilita ver si el algoritmo confunde dos o más clases, es decir, que al etiquetar lo datos de una clase los ha confundido con otra.

La plantilla para una matriz de confusión binaria utiliza cuatro tipos de resultados discutidos junto con las clasificaciones positivas y negativas, los resultados se pueden formular en una matriz de confusión de 2x2 como se muestra en la siguiente Figura 9:

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 9 Matriz de confusión binaria

En la detección de anomalías con un algoritmo de aprendizaje automático basado en clasificación, hay dos resultados de prueba posibles, un resultado de prueba positivo y uno negativo, debido a que hay dos posibles valores reales y dos posibles valores de predicción, a partir de estas opciones se crea la matriz de confusión con los siguientes 4 resultados posibles (Juan B. 2019):

Verdadero positivo: El valor real es positivo y la prueba predice un positivo.

Verdadero negativo: El valor real es negativo y la prueba predice un negativo.

Falso negativo: El valor real es positivo, y la prueba predice un negativo.

Falso positivo: El valor real es negativo, y la prueba predice un positivo.

La matriz de confusión no está limitada a la clasificación binaria se puede utilizar también en clasificaciones multiclase, la clasificación solo tiene la condición positiva y negativa. Un ejemplo

de una matriz de confusion multiclase se muestra en la Figura 10, donde se resume la comunicación de un lenguaje silbado entre dos hablantes.

vocal percibida / vocal producida	i	mi	a	o	tu
i	**15**		1		
mi	1		1		
a			**79**	5	
o			4	**15**	3
tu				2	**2**

Figura 10 Matriz de confusión multiclase

La ciencia de datos es un campo interdisciplinario que involucra estadísticas, análisis de datos, aprendizaje automático, entre otros. En este trabajo la ciencia de datos se aplica mediante el aprendizaje automático aplicando un algoritmo de clasificación con medida de distancia euclidiana k-NN; el método de k vecinos más cercanos utiliza la proximidad para hacer clasificaciones de datos y predicciones de un nuevo dato.

En el reconocimiento de patrones, se utiliza como método de clasificación de elementos basado en el entrenamiento mediante elementos cercanos. Este algoritmo requiere que tengamos dos conjuntos de datos, *Train y Test*, que se utilizan para entrenar al algoritmo y permiten que identifique con etiquetas cada uno de los datos, al introducir los datos de prueba este los clasificará y asignará la etiqueta del conjunto de datos que tenga más vecinos más cercanos a este nuevo elemento.

La elección del número de vecinos a analizar *k* depende esencialmente de los datos, valores grandes de *k* reducen el efecto de ruido en la clasificación, es decir, el algoritmo es más exacto.

CAPÍTULO II

DESARROLLO DEL CASO DE ESTUDIO

REGIÓN VALLE DE SANTIAGO

Valle de Santiago fue fundado oficialmente el 28 de mayo de 1607 por Pedro Rivera, Cristóbal Martínez, Francisco Gómez, Silvestre Aguirre, Luís de Fonseca, Antonio de Estrada, Andrés de Cuellar, Francisco de Santoyo y Juan Martínez Guerrero, acompañados de centenares de indígenas lugareños y algunos de ellos traídos de la lejana Taximoroa para ayudar a cavar los canales del Laborío en que participaría sus aguas el río Lerma.

La ciudad de Valle de Santiago, cabecera municipal, está situada a los 20° 23´ 34¨ de latitud norte y 101° 11' 29" de longitud oeste del Meridiano de Greenwich, en el declive oriental de la colina de la Alberca y el septentrional de las montañas de la conchita y La Batea. El municipio representa el 2.7 % del total de la superficie del estado con 82,010 has (820.1 km²) y la cabecera municipal se encuentra a 1,744 msnm.

A continuación, en la Figura 11 se muestra un mapa del estado de Guanajuato con división municipal, resaltando en este la ubicación del municipio de Valle de Santiago.

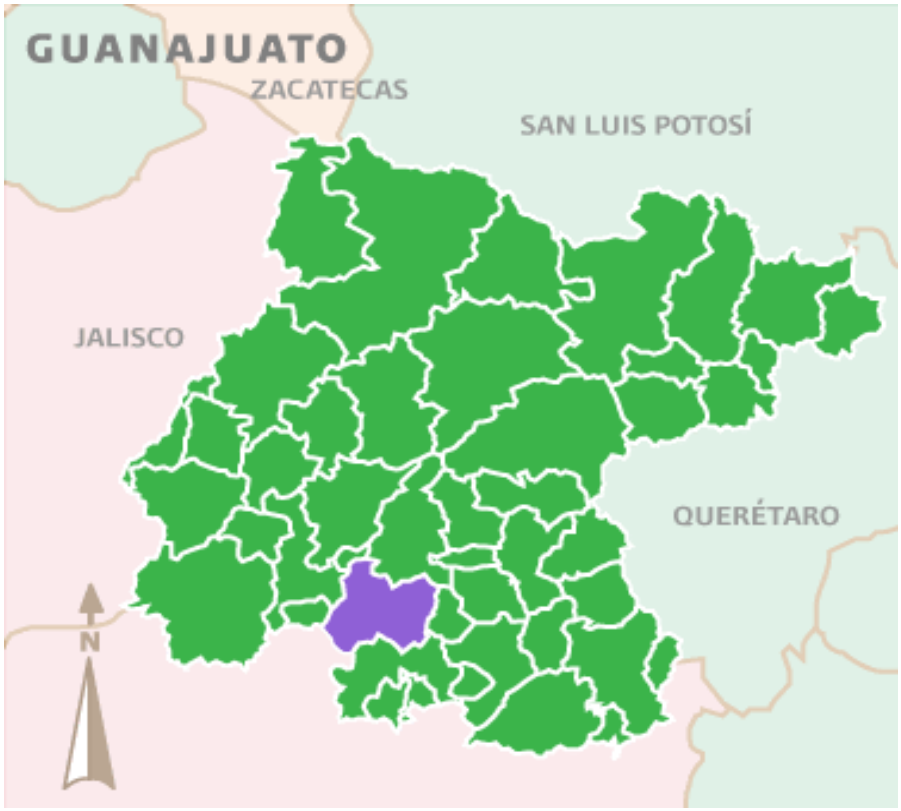


Figura 11 Mapa del estado de Guanajuato

Población

El municipio de Valle de Santiago tiene una población de 150,054 habitantes según el Censo General de Población y Vivienda 2020 que realizó el INEGI (décima posición a nivel estatal y 163 de 2,469 municipios del país). La población de la cabecera municipal es de 72,663 habitantes, lo que la ubica como la localidad número siete por población en el estado.

Los datos que arrojó dicho censo revelan que el municipio es, por primera vez, mayoritariamente compuesto por población urbana.

Actividad económica

De acuerdo con la Oficina de Información Municipal de Valle de Santiago, la industria en el municipio es la tercera actividad en importancia, la cual emplea el 22.31% de la población ocupada. Las principales actividades del ramo industrial son desarrolladas en los pequeños y medianos talleres de tejidos de lana, como gabanes y cobijas, cestería de carrizo; juguetes de cartón, y en pequeña escala piezas de cerámica.

El sector terciario es la actividad económica más importante dentro del municipio de Valle de Santiago, ya que emplea el 41.65% de la población ocupada municipal. Una manera de medir el comercio es mediante el número de usuarios y el volumen de las ventas de energía eléctrica del tipo comercial, el municipio aporta el 2.84% de los usuarios estatales y el 1.6% del volumen estatal.

En este municipio las principales actividades que se realizan son:

Agricultura: La actividad agrícola es la más importante en el municipio de Valle de Santiago, pues de acuerdo con el último registro estadístico del año agrícola 2005, se sembraron 66,507 hectáreas, de las cuales 53,517 hectáreas fueron sembradas con semilla mejorada y con una superficie mecanizada de 53,273 hectáreas, adicionalmente se reportan 735 hectáreas atendidas con servicios de Sanidad Vegetal. (INEGI, 2006).

Ganadería: El municipio de Valle de Santiago no destaca en el ámbito estatal por ser un municipio cuya población ganadera sea de un tamaño considerable, pero esta es una fuente importante de ingresos. Las principales especies que se crían son porcino, bovino y caprino.

Orografía

En Valle se localiza un grupo de volcanes que se compone de 13 cráteres, (algunos muy erosionados) situados en una superficie de 14 km² que comprende del cráter de Yuriria hasta el pie del cerro del Rincón de Parangueo. Y se encuentra en la provincia fisiográfica denominada eje neovolcánico, quedando el Municipio insertado en dos subprovincias, que son la del Bajío Guanajuatense y la de la sierra y bajíos Michoacanos, dentro de la primera sub- provincia queda el 42.70% del territorio Municipal, ubicada al norte; y en la segunda el 57.21%, que corresponde a la parte sur. Las elevaciones más importantes del Municipio son el cerro del Tule. El Picacho, El Varal, Cerro Blanco, La Batea, Los Cuates y el Cerro Prieto. La altura promedio de estos cerros es de 2,100 msnm

Clima

El clima es Húmedo y subhúmedo con lluvias en verano con temperatura máx. de 40º C y mínima de 5 °C y un promedio anual de 18.5ºC con evaporación de 2,371.8 mm anuales. La temperatura máxima que se ha registrado en el Municipio es de 43 °C y la mínima es de -7 °C, registrada en los años de 1964 y 1963 respectivamente. Es de clima templado con temperatura ordinaria entre 15 y 25 °C, la precipitación promedio anual es de 564.9 mm, el 85% de la superficie es semicálido subhúmedo con lluvias en verano, de humedad media, el 11.96% templado subhúmedo y el 2.08% es templado con menor humedad, los meses más cálidos son abril, mayo y junio.

Hidrografía

En su hidrografía cuenta al norte, y como límite, el Río Lerma además de pertenecer a la cuenca Río Lerma- Chapala y Lago de Pátzcuaro-Cuitzeo-Yuriria de la Región Hidrológica No. 12 denominada Lerma-Santiago sus aguas corren de oriente a poniente y drenan una superficie de 902.5 km, contando también con 887 pozos en total de 4",6"y 8" para uso agrícola y doméstico.

De acuerdo con la regionalización de Gerencia de Aguas Subterráneas Subdirección Técnica, Comisión Nacional de Agua (CONAGUA), en las inmediaciones del Área del Municipio de Irapuato subyace un sistema conformado por dos acuíferos, que se denomina en conjunto como

Irapuato - Valle de Santiago No. 1119, estos son aprovechados por los Municipios de Valle de Santiago, Salamanca e Irapuato.

El primer acuífero superficial, está conformado por depósitos de aluvión y de tobas que rellenan esta parte del Valle, estos cubren a otro acuífero en conglomerado poco empacado, riolitas y rocas basálticas fracturadas. La zona de recarga de estos dos sistemas se da por infiltración directa sobre los rellenos y en los afloramientos de roca, los que aportan agua a los rellenos en el ámbito subterráneo. El acuífero superficial del agua tiene una temperatura registrada de 24 °C y en el profundo es mayor de 34 °C. La superficie de este está calculada en 1,372 km de donde Valle de Santiago comprende el 29.06%, existen un total de 1,143 pozos para los usos de riego, agua potable, uso doméstico e Industrial, extracción que se hace a través de pozos profundos, norias y manantiales, aunque estos su captación es prácticamente superficial.

CASO DE ESTUDIO SECTOR “LAS HACIENDAS”

El sector “Las Haciendas” se encuentra ubicado al Noreste de la cabecera municipal del municipio de Valle de Santiago Gto., en la Figura 12 se muestra una fotografía aérea donde resalta la ubicación de la colonia Las Haciendas, además de una aplicación de esta en la que se aprecia de mejor forma su configuración.

Respecto a la infraestructura con la que cuenta el SAPAM, es una red compleja debido a las condiciones topográficas del municipio, además de que con los cambios de administraciones municipales no se había dado continuidad a la actualización catastral de la red, esto provoca que se instalara tramos de red con características desconocidas (diámetros y materiales), recientemente el SAPAM se ha dado a la tarea de renovar sectores en los que los equipos ya presentaban un deterioro notorio por el uso prolongado; por ejemplo, la zona centro fue beneficiada en el cambio de medidores de agua, habilitación de un tanque elevado y actualización catastral.

Esto nos refleja que, aunque se tenga una amplia red de distribución, si no se ha operado de forma eficiente genera problemas físicos cada vez más graves, sumado a ello se debe tomar en cuenta que el usuario es un factor importante que se debe considerar dada la problemática que implica en ocasiones entre otros problemas sociales que interfieren con las actividades cotidianas que los organismos realizan.

El caso de estudio corresponde a una sección de la red de abastecimiento de agua potable del SAPAM (Sistema Municipal de Agua Potable y Alcantarillado) del municipio de Valle de Santiago, Guanajuato. Donde se presentan diversos problemas relacionados al abasto de agua dicho sector es el denominado “Las Haciendas”, colonia que se encuentra en la cabecera municipal cuyo desarrollo tiene desde el año 2010.

El tanque que corresponde al pozo No. 15 identificado con el nombre de “Las Haciendas” cuenta con una línea de alimentación de PVC de 6” hacia la entrada de la colonia y conecta en la calle Hacienda de San Javier Sur con una línea de distribución de 3”.

En la zona, las viviendas y comercios están construidos con tabique de barro rojo recocido y techos de concreto, la mayoría de las construcciones son de un mismo modelo y medida, siendo de clasificación de interés social. En esta zona se tienen calles pavimentadas con concreto hidráulico y banquetas del mismo material, además de los servicios básicos, diferentes tipos de negocios como tortillerías, tiendas de abarrotes, carnicerías, etc. motivo por el cual se puede considerar un nivel de vida medio en esta zona.

Figura 12 Ubicación del Fraccionamiento “Las Haciendas” en el municipio de Valle de Santiago

El pozo No. 15 cuenta con un tanque elevado ubicado en el mismo domicilio del pozo; que cuenta con una capacidad de 120 metros cúbicos y una altura de 15 metros, del cual únicamente se abastece la colonia, por lo que es un sector delimitado. El pozo No. 15 opera de forma automática con un sensor de nivel en su tanque elevado, encendiéndose según la demanda de agua.

Tabla 1 *Características del pozo #15 (2010)*

POZO	LAS HACIENDAS
COORDENADAS	LONG 20.4038
	LAT -101.1772
HORAS DE OPERACIÓN	7
FLUJO PROM DIARIO (L/s)	16
DIÁMETRO	6"
PROFUNDIDAD	91 m
NIVEL DINÁMICO	52 m
NIVEL ESTÁTICO	50 m
PRESIÓN (Kg/cm2)	2
POTENCIA (HP)	20
TANQUE	LAS HACIENDAS

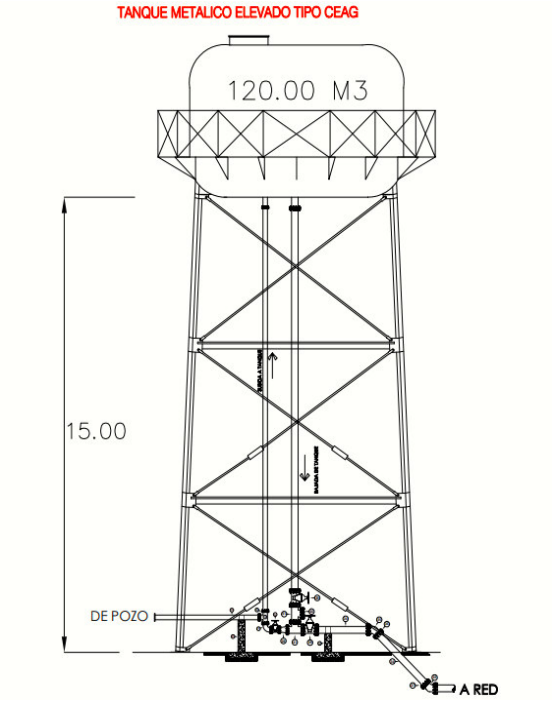


Figura 13 Plano del tanque elevado

En la Figura 14 se muestra una fotografía satelital del fraccionamiento "*Las Haciendas"* delimitado en color rojo, las líneas azules representan la línea de distribución y en color magenta se representa la línea de alimentación pozo-tanque y la línea de alimentación tanque-red principal, como ventaja se puede observar que este tiene una configuración simétrica respecto a la forma de las calles y las pendientes dentro del fraccionamiento son muy bajas.

La red de agua potable está compuesta por una línea de conducción de 99.86 metros lineales de tubería de PVC hidráulico tipo ANGER RD-32.5 de 8" de diámetro y una línea de distribución de 4,752.65 metros lineales de tubo de PVC hidráulico tipo ANGER RD-32.5 de 3" de diámetro

Figura 14 RDA "Las Haciendas"

DESCRIPCIÓN DE LOS PROBLEMAS DE LA RED

A continuación, se enlistan algunos problemas que el SAPAM ya tenía identificados, además de otros problemas que se detectaron durante recorridos al sector:

- Baja presión en tomas domiciliarias
- Fujas en líneas de producción
- Tomas domiciliarias clandestinas

- Sistemas de riego clandestinos instalados en áreas verdes
- Tipo de uso distinto al otorgado
- El hábito de los habitantes a conectarse a la red de manera clandestina cuando se tiene una construcción en proceso con la finalidad de no registrar la cantidad de agua utilizada para la obra
- La desconfianza que existe por parte de los usuarios hacia el organismo operador
- El surgimiento creciente de plantas purificadoras de agua, estos negocios se conectan a la red de forma clandestina o tienen una toma de uso doméstico y el agua extraída es purificada para su venta envasada (garrafones). Dichos modelos de negocio consumen grandes cantidades de agua, ya que el proceso requiere agua para diferentes actividades, además de purificar el líquido, se utiliza en la limpieza de equipo, envases, entre otros.

La desconfianza de la gente al O.O. se debe a que por largos periodos de tiempo muchos usuarios han tenido un servicio deficiente, o simplemente que no alcanza a satisfacer las necesidades de estos, especialmente con la intermitencia en el suministro o la baja presión en la red que provoca una disminución en la cantidad de agua. Al querer realizar un censo sobre la calidad del servicio o realizar la medición de presión en las tomas domiciliarias, los usuarios se niegan a que se realicen estos procedimientos en sus tomas, se detectó que muchos usuarios tienen la creencia que, al manipular los medidores de agua en sus tomas domiciliarias, se configuran de tal forma que el O.O. podrá realizar cobros excesivos. Las consecuencias de esta obstrucción al levantamiento de datos por parte de los usuarios al O.O., son la carencia de datos de funcionamiento y estado de la red, así como el estancamiento en su actualización y mejoras en el servicio.

SOLUCIÓN PROPUESTA

En el desarrollo de este trabajo nos vamos a enfocar en cómo disminuir fugas a través de una técnica de máquina de aprendizaje, como lo es un algoritmo de clasificación supervisada k-NN con distancia euclidiana (Ecuación [2]), utilizando los datos de caudal y presión.

Para lograr este objetivo, se requiere contar con datos e información que se adecuen al algoritmo k-NN; este algoritmo de clasificación requiere de una fase de entrenamiento donde realiza el reconocimiento de patrones y posteriormente ejecuta la fase de predicción y la nueva información

es clasificada en la clase que tenga más vecinos cercanos. Dicho de otra manera, en la fase de entrenamiento el algoritmo reconoce y aprende el comportamiento hidráulico ideal de la RDA modelada con los datos obtenidos de la red y realiza la comparación de esta con el comportamiento hidráulico de una red con una fuga de agua propuesta en diversos puntos del sector, finalmente se obtiene la matriz de confusión, donde se representan las clasificaciones de clases realizadas, visualizando las clases donde ubo equivocación en la predicción.

OBTENCIÓN DE DATOS

La aplicación de este tipo de métodos matemáticos donde se requiere analizar y tratar información resultó compleja desde la parte de obtención de datos, este problema ha sido un factor común en muchos organismos operadores, la discontinuidad de recolección de datos, datos faltantes y datos incorrectos son algunas de las causas por las que no es posible obtener una base de datos adecuada para aplicar estas técnicas de *machine learning*.

La información obtenida fue otorgada por el SAPAM, además de datos específicos de la RDA tomados en campo personalmente con la intención de visualizar el comportamiento del sector "Las Haciendas" como: presiones y caudal extraído, dentro de los datos obtenidos se enlistan los siguientes:

- Caudales (Q) de las tomas domiciliarias registrados por los medidores de caudal
- Caudales (Q) extraídos del "pozo 15" que abastece al fraccionamiento
- Presiones (m.c.a) obtenidas de una toma domiciliaria por medio del datalogger
- Pano topográfico y estructura de la RDA del sector "Las Haciendas"

En la obtención de datos de caudal (Q) del padrón de usuarios y volúmenes extraídos, el SAPAM proporcionó un gran compendio de información, cabe resaltar que dicha obtención fue interrumpida, dificultando la fluidez y rapidez de su revisión y ordenamiento. Además, por seguridad se requirió llevar vía remota la comunicación con el O.O. gran parte de la investigación debido a la pandemia por COVID-19, prolongando los tiempos de entrega por parte del SAPAM.

Los volúmenes de extracción se registran a través de un macro medidor instalado en la línea de alimentación pozo-tanque. Los volúmenes de agua consumidos por el usuario se obtuvieron mediante el padrón de usuarios proporcionado de la base de datos del SAPAM.

Las presiones se obtuvieron por medio del *Datalogger* de la marca **sebaKMT** situado en una caja de válvulas instalado por el O.O. previamente, además de realizar recorridos para la toma de presiones manual, en este caso no se logró concretar una base de datos completa debido a inconvenientes surgidos en campo.

Los planos topográficos del sector "Las Haciendas" fueron proporcionados por el SAPAM casi de manera inmediata, debido a que es un fraccionamiento relativamente nuevo (2010), se obtuvo toda la información respecto a las líneas de conducción, características del pozo, y características del tanque elevado, así como del funcionamiento de la red.

TRATADO DE DATOS

Debido a que los datos proporcionados contenían información innecesaria como: tarifa, No de medidor, giro, etc., se realizó una serie de filtros y procesos que se requieren para su manejo en la modelación hidráulica, en seguida se explica de manera breve los procesos realizados al padrón de usuarios del sector "Las Haciendas":

1.- El primer proceso se realizó para obtener el consumo en m^3 por mes de cada domicilio registrado en el padrón de usuarios, fue el filtro de las lecturas por toma de modo que se obtuviera una tabla de datos más clara únicamente con la información de domicilio y consumo mensual Jun-Ene.

2.- Se obtiene el volumen en m^3 por calle, resultado de la suma de los consumos obtenidos anteriormente por cada domicilio, se puede observar a detalle la información en el ***Anexo*** **Tabla 1**.

3.- Posteriormente a los datos de consumo mensual por calle se les aplicó la Ecuación [6] para convertir $\frac{m^3}{mes} \rightarrow lps$, esto con el objetivo de utilizar el consumo en el modelo de EPANET, obteniendo una tabla con los consumos en l/seg de cada mes por cada calle como se muestra en el ***Anexo*** **Tabla 2**.

$$\left(consumo\,\frac{m^3}{mes}\right) * \left(\frac{1000lt}{(30dias)(24hr)(60min)(60seg)}\right) = consumo\,\frac{lt}{seg} \qquad [6]$$

4.- Con el fin de simplificar la estructura del modelo en EPANET, se utilizaron los consumos en lps obtenidos anteriormente divididos por el número de nodos por calle, los datos de lps obtenidos para cada nodo de cada una de las calles se puede observar en el ***Anexo* Tabla3**.

La **Tabla 2** muestra el balance hidráulico Jun-Ene, se observa que no se obtuvo parte de la información de los meses de junio y enero.

***Tabla 2 Balance** Hidráulico JUN-ENE*

VOLÚMENES MENSUALES								
	JUN	JUL	AGO	SEP	OCT	NOV	DIC	ENE
VOL. EXTRAIDO	14,045	19,003	13,304	16,572	19,872	21,552	23,410	--------
VOL. FACTURADO	--------	6,317	5,744	6,516	5,453	5,607	6,348	5,230
VOL. NO CONTABILIZADO	--------	12,686	7,560	10,056	14,419	15,945	17,063	--------

La Figura 15 muestra el balance hidráulico del mes de septiembre, donde se puede observar la gran diferencia que existe entre el volumen facturado y el volumen no contabilizado, como es evidente el volumen que no es contabilizado representa una gran parte del que se está extrayendo del pozo.

Esta información será tomada como pérdida física de agua, ya sea que pueda ser una fuga o una extracción ilegal de la RDA, de forma que al realizar la modelación hidráulica se contempla dicho volumen como un consumo fuera del padrón de usuarios simulando una fuga de agua.

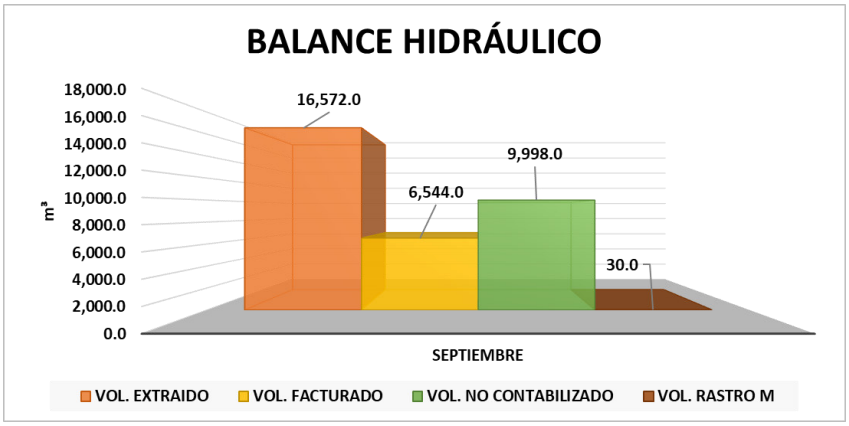


Figura 15 Balance hidráulico del mes de Septiembre

4.- Finalmente después de la modelación hidráulica, donde se simula la RDA agregando un nodo de consumo extraordinario en diferentes zonas de la red, se extrae la información de 24hr de simulación de presión y consumo. Estos datos se separaron en dos grupos *Test y Train*, agrupados en tablas con el formato y orden para el algoritmo k-NN, en este caso se incorporaron tres columnas: Hora, Caudal, Presión e identificador k-NN (asignado por la zona donde se ubicó el nodo extra).

MODELADO HIDRÁULICO

Dentro de las nuevas herramientas para el diseño de redes de abastecimiento se encuentra el *software* libre EPANET, que permite realizar análisis hidráulicos de redes de tuberías a partir de las características físicas de las tuberías y dinámicas de los nudos (consumos) para obtener la presión y los caudales en nodos y tuberías respectivamente, así como el análisis de calidad de agua a través del cual es posible determinar el tiempo de viaje del fluido desde la fuente hasta los nodos del sistema.

EPANET fue desarrollado por la División de Recursos Hídricos y Suministros de Agua de la Agencia para la protección del Medio Ambiente de EE. UU. (USEPA), este es de dominio público y permite realizar análisis hidráulicos de redes de tuberías a partir de las características físicas de las tuberías y dinámicas de los nudos de consumo para obtener la presión y caudales en los nodos y tuberías, además EPANET permite realizar análisis de calidad de agua (Rossman L.A. 2001).

El uso de este *software* trae diversas ventajas como:

- Puede usarse para mejorar las características de la red hidráulica
- No existe límite en cuanto al tamaño de la red que puede procesarse
- Admite bombas, válvulas y depósitos de diversas características
- Ofrece resultados que pueden ser utilizados para el dimensionado y selección de componentes
- Evalúa el comportamiento y el consumo y costo de energía
- Puede modelar tomas de agua cuyo caudal dependa de la presión
- Considera diversas condiciones de demanda en la red

- Modela a lo largo de la red el comportamiento de químicos utilizados en la calidad de agua p.ej. cloro.

Siendo este *software* una herramienta versátil y de gran confiabilidad, no se han detectado desventajas en su aplicación, sin embargo, al requerir ciertas características y que funciona con la información y configuración que el usuario le otorga la restricción que este presenta es la habilidad del usuario para manipularlo.

En el sector "Las Haciendas" a distribución de agua se realiza por medio de la gravedad, iniciando en el tanque "Haciendas", por medio de la tubería de 203.2 mm (8)" de PVC hidráulico, de esta se desprende una red de distribución de 76.2 mm (3") de PVC hidráulico.

El modelo base se construyó simplificando la cantidad de nodos y líneas con la finalidad de simplificar su operación y visualización, esto derivo a realizar una estrategia para la configuración de la calibración e interpretación de datos. Este modelo consta de:

- 95 nodos
- 116 tuberías
- 1 reservorio
- 1 tanque
- 1 bomba.

El modelado hidráulico se realizó en EPANET, se llevó a cabo de la siguiente forma:

- Se exportó la imagen del plano topográfico al EPANET
- Se colocaron los nodos (cruces, bomba, tanque y reservorio)
- Se unieron los nodos con líneas de distribución
- Se introdujo la elevación a cada uno de los nodos
- Se configuraron las características del tanque, pozo y bomba
- Se introdujeron las características físicas de la tubería (rugosidad, longitud y diámetro)
- Se configuraron los patrones, de consumo para los nodos y de control para la bomba
- Se seleccionó la ecuación para el cálculo de pérdida de carga, unidades de flujo.

Para la simulación del sistema se consideró también la curva de operación de la bomba de extracción del pozo #15, no se tiene más información como el modelo u otro tipo de características, el O.O. proporcionó la curva de operación que consta de lo siguiente:

- Flujo 17 lps
- Altura 79 m.

En la Figura 16 se muestra el modelo base, este está construido con las características físicas antes mencionadas y que se especifican en los planos topográficos, posteriormente se complementa con la información obtenida y procesada del padrón de usuarios.

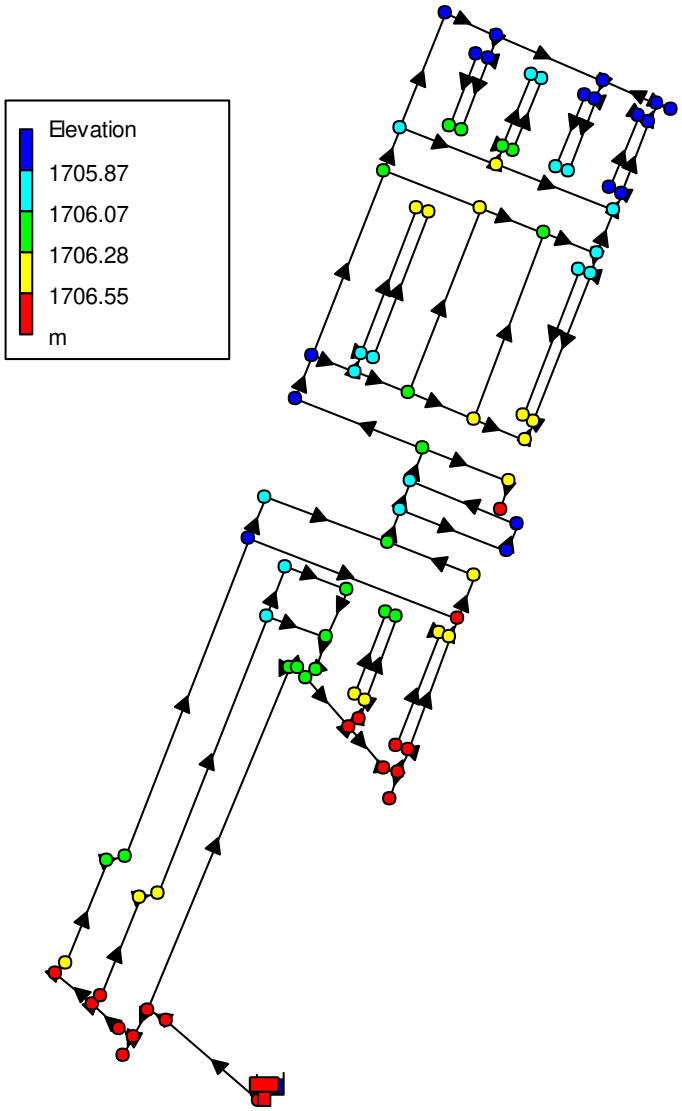


Figura 16 Configuración del sector "Las Haciendas" en EPANET

Se utilizaron los coeficientes de variación horaria para pequeñas comunidades (Manual de agua potable, alcantarillado y saneamiento, CONAGUA, 2019), tal como se muestra en la Figura 17**.**

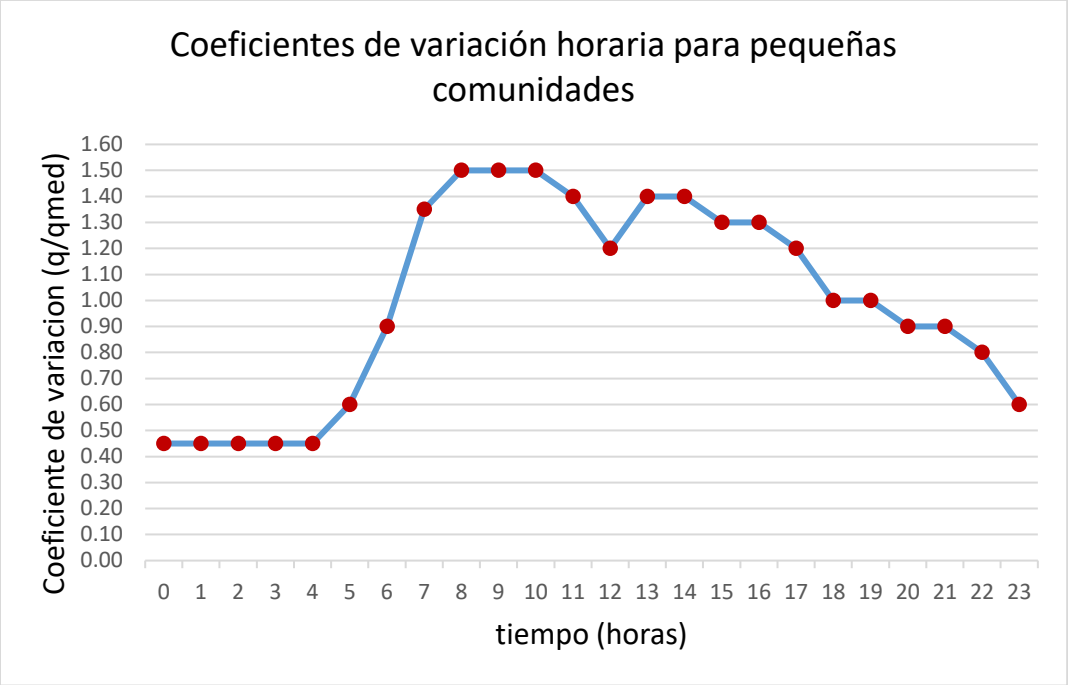


Figura 17 Coeficientes de variación horaria para comunidades pequeñas

Al tener el modelo configurado con la información de la red y del padrón de usuarios, con la finalidad de realizar una calibración en el modelo y poder obtener un comportamiento hidráulico real de la RDA, se coloca en el modelo los datos de presión obtenidos mediante el *Datalogger* en la ubicación conocida, las presiones se tomaron de un día al azar esto debido a que se detectó que tienen un comportamiento similar los mismos horarios.

La calibración del modelo no se logra completamente, debido a que se tienen datos de un solo punto, la Figura 18 muestra el gráfico de presiones obtenidas en campo y las calculadas en la modelación hidráulica, los datos obtenidos en campo son distintos a los obtenidos en el modelo, teniendo en cuenta que las características del modelo hidráulico son las reales y los datos de consumo fueron obtenidos de la RDA, al notar esta diferencia hace sentido que existen problemas en la RDA.

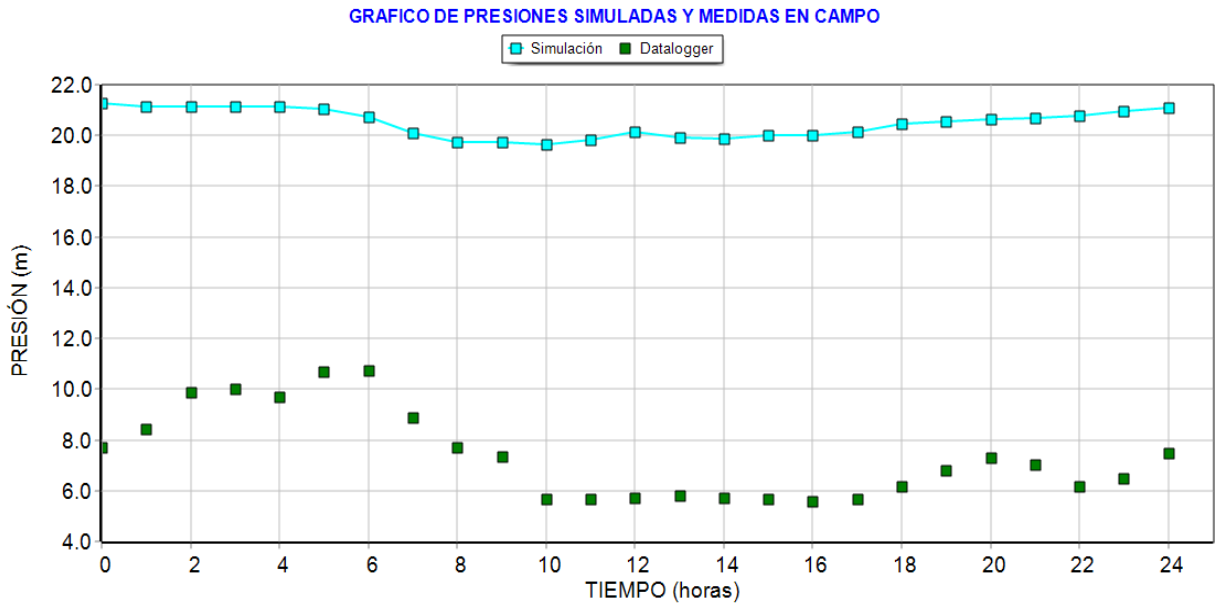


Figura 18 Gráfico de presiones

GENERACIÓN DE ESCENARIOS

Siendo en esta investigación el objetivo principal del algoritmo k-NN la detección de anomalías, se propone un algoritmo enfocado en el aprendizaje del comportamiento de una red de distribución con funcionamiento óptimo, es decir se le hará saber al algoritmo cuál es el comportamiento que se espera que tenga la RDA, esta red se construye con las características de proyecto de construcción otorgado por el O.O., después se le asignan los datos de consumo registrados por el mismo, posteriormente se compara con el comportamiento de la misma red con un comportamiento diferente.

El sector “Las Haciendas” es dividido en 5 zonas, cada zona contiene 5 ubicaciones donde es colocado un nodo de consumo, cada nodo de consumo será modelado por separado resultando un total de 25 escenarios. Dicha configuración de zonas en la que se dividió el sector “Las Haciendas” se muestra en la Figura 19, aquí se pueden observar las zonas 1-5 y las ubicaciones de cada nodo de consumo colocado para la simulación de las posibles fugas.

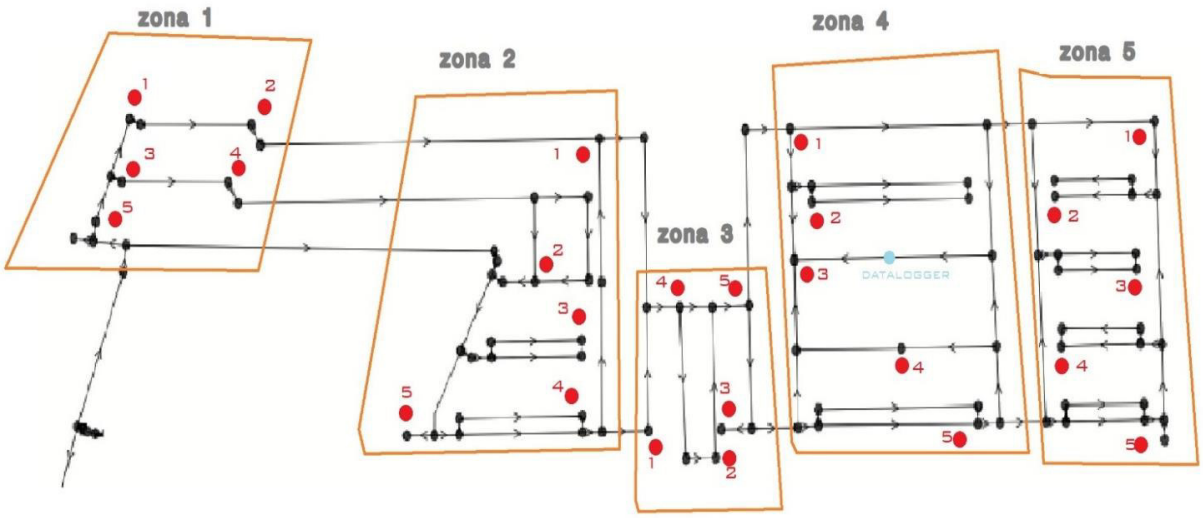


Figura 19 Zonas dentro de la RDA para simulación de fugas

Por ejemplo, para el punto 1 de la zona uno le corresponderá el escenario 1.1 y asi sucesivamente para cada punto en la red.

Una vez modelado cada escenario en EPANET, se necesita extraer los datos con los que trabaja el k-NN, para este caso los datos extraídos que se necesitan para la detección de anomalías fueron los de consumo (lps) y presión (m) del nodo 1 (ubicación del *Datalogger*) con la finalidad de poder ser comparado el comportamiento en las diferentes situaciones. Posteriormente los datos fueron organizados en dos grupos (*Test y Train*), los datos Train se utilizan para el entrenamiento, es decir ajustar la máquina de aprendizaje y los datos *Test* se utilizan para evaluar el desempeño final de la máquina (fase de predicción).

APLICACIÓN DEL k-NN

El algoritmo k-NN se realiza por medio de la plataforma de programación MATLAB, sistema que ofrece un entorno de desarrollo integrado con un lenguaje de programación propio, utilizado en la creación de algoritmos, análisis y visualización de datos y la creación de modelos.

A continuación, se describe el pseudocódigo utilizado en esta investigación:

1. Se establece la conexión de la interfaz con la carpeta que contiene el archivo de datos procesados previamente en formato .csv y se establece como carpeta de trabajo.
2. Se importa el archivo de datos que será utilizado como Train.csv
3. Se importa el archivo de datos que será utilizado para Test.csv
4. Se realiza la selección del algoritmo ML_KNN, número de vecinos y la distancia a calcular
5. Se selecciona la función *randperm* para que el algoritmo realice arreglos aleatorios en la selección de datos de entrenamiento y predicción
6. Se hace la predicción con el modelo ya entrenado, con la aleatoriedad antes seleccionada
7. Se crean las matrices de confusión.

RESULTADOS

En el sector "Las Haciendas" el clasificador k-NN con métrica euclidiana demostró un buen desempeño, la aplicación de este tipo de algoritmos pueden causar un alto impacto en el tema de la distribución de agua especialmente en la detección de anomalías que pueden representar una pérdida de agua, a pesar de solo contar con un punto de lectura de presiones en campo, el algoritmo logró clasificar y detectar posibles anomalías en dos zonas de la red donde empíricamente se había planteado la teoría de existencia de pérdida de agua en esas zonas de la red.

El modelado hidráulico funciona para una mejor visualización de los datos hidráulicos en el sector, en él se le agrega información obtenida en campo con la finalidad de observar un comportamiento hidráulico determinado que ayuda en la toma de decisiones, en este caso hace notar que las presiones obtenidas en un solo punto de la red no son suficientes para la detección de anomalías.

Siendo este un entorno de predicción donde se requiere estimar la precisión del modelo, se usa la aleatorización para la validación cruzada, resultando 25 matrices de confusión; el modelo toma aleatoriamente distintos valores para el entrenamiento de forma que la función de aproximación solo se ajusta con el conjunto de datos de entrenamiento y a partir de aquí calcula los valores de salida para el conjunto de datos de prueba *test*, en la Figura 20 se muestran las matrices multiclase 5x5 obtenidas de cada simulación.

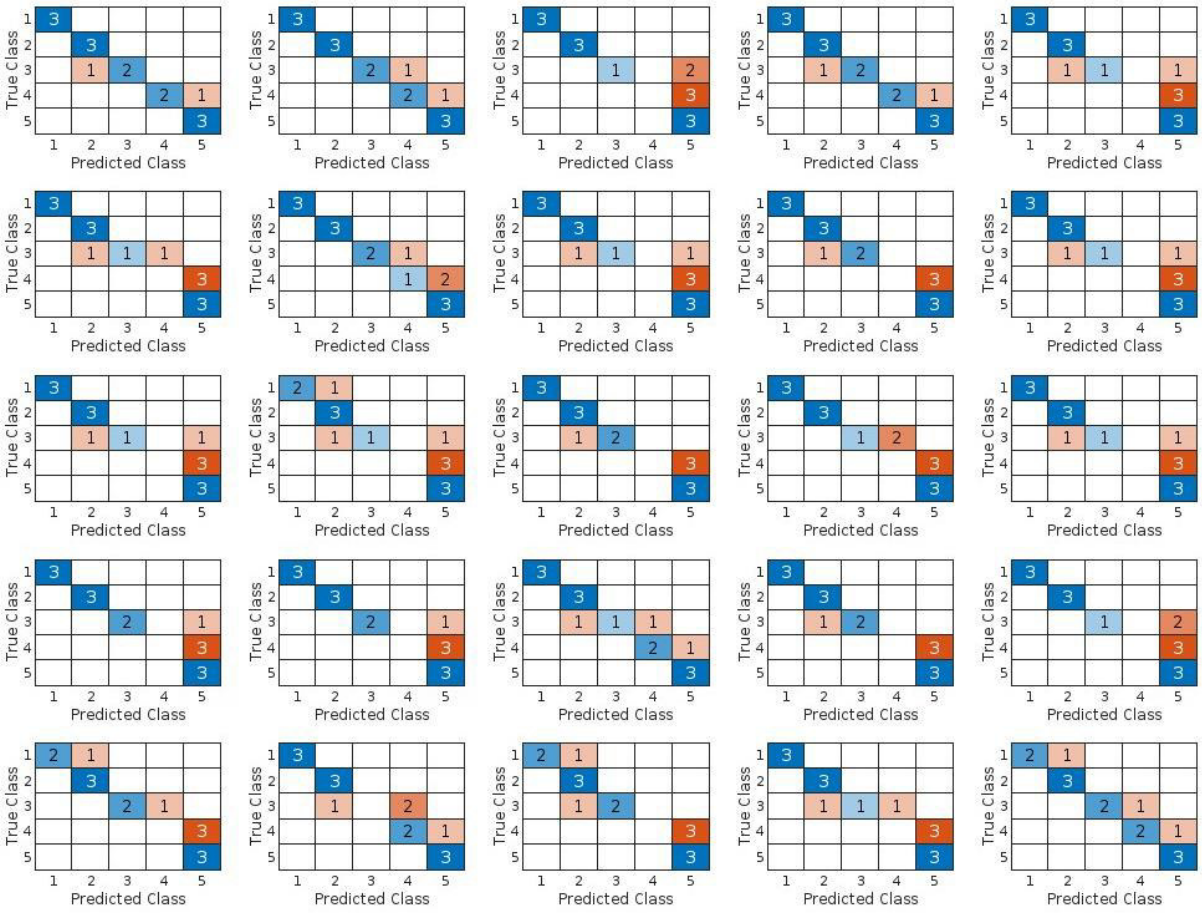


Figura 20 Matrices de confusión obtenidas

Cada una de las matrices de confusión mostradas representan un proceso de clasificación k-NN con una configuración de datos seleccionados para *Train y Test* diferente, es decir, en cada proceso utiliza datos de diferentes escenarios para realizar el proceso de clasificación, el objetivo de estas matrices es valorar qué tan bueno es el modelo de clasificación basado en aprendizaje automático, mostrando de forma explícita cuando una clase es confundida con otra, en este caso en qué zonas se ha equivocado en su clasificación.

Analizando las matrices de confusion se observa que en las zonas 1, 2 y 5, el 84% de las iteraciones no hay error en su clasificación de la clase real con la clase predicha, es decir, a cada clase le asigna en la etapa de predicción la clase real, sin embargo, en las zonas 3 y 4 existe un 100% de error en la clasificación de clases, estos errores en su clasificación nos dicen que en dichas clases existen

anomalías en los datos, siendo información de comportamiento hidráulico significa que en las zonas 3 y 4 el comportamiento hidráulico no es el correcto.

Debido a que no se tiene más información real del sector, no se puede precisar qué clase de anomalía está presente en las zonas que se detectaron problemas, sin embargo, esta técnica de detección de anomalías incrementa considerablemente la posibilidad de encontrar fallas en una red de abastecimiento.

CAPÍTULO III

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

De acuerdo con el análisis aplicado a la RDA del sector “Las Haciendas” de la ciudad de Valle de Santiago, se llega a las siguientes conclusiones:

La eficiencia con la que cuenta la RDA del sector “Las Haciendas” es baja, ya que al realizar el balance hidráulico comparando los volúmenes de agua extraídos contra los volúmenes consumidos registrados por el padrón de usuarios existe un volumen de agua no contabilizada que representa el 60.33% del total extraído, una cifra alarmante debido a que es una RDA relativamente nueva con solo 12 años de vida útil. Sin embargo, el SAPAM ha demostrado la iniciativa de continuar mejorando el servicio, beneficiando no solo a los usuarios, también ayudando a la preservación del vital líquido, desde la mejora continua en equipos de micro medición, recuperación de caudales y la detección oportuna de fugas.

Este caso en particular es ejemplo de que el no contar datos dificulta poder llevar a cabo algún proceso de análisis de una RDA, ya sea de funcionamiento hidráulico, o la detección de anomalías.

El k-NN desarrollado demuestra que aun teniendo los problemas anteriormente vistos se puede analizar una RDA y obtener buenos resultados, además de que puede seguir complementándose con más información y hacerse más poderoso en la detección de anomalías.

La combinación del programa EPANET en conjunto con el algoritmo resultó una gran herramienta de apoyo para la obtención y manejo de datos que ayudaron a realizar una comparación de escenarios y una clasificación más cercana a la realidad. La clasificación mediante un algoritmo significa un ahorro de tiempo comparándolo con un análisis in situ realizado manualmente, además significa el ahorro de recurso económico para el O.O. ya que no requirió hacerse de equipo para realizar una monitorización completa de la red.

El k-NN demuestra que es una herramienta muy útil tratándose del análisis de datos, para los O.O. el uso de la ciencia de datos y técnicas de *machine learning* para la detección de anomalías se convierte en una ventaja, ya que además de poder analizar problemas de funcionamiento hidráulico se puede enfocar en el análisis de calidad del agua que se brinda a la población, pasar por alto alguna anomalía negativa puede ser perjudicial en la salud pública.

El objetivo general de esta investigación se cumple en gran parte, se logró la aplicación de un algoritmo k-NN para el análisis del comportamiento hidráulico del sector y la posterior detección

de anomalías en diversas zonas del mismo, teniendo en cuenta que la falta de datos reduce la exactitud de la técnica, los resultados obtenidos demuestran que el k-NN es una técnica muy versátil en el campo de la hidráulica.

Dentro de los objetivos específicos se cumplen satisfactoriamente 5 de ellos, cabe mencionar que la calibración de un modelo hidráulico con las presiones obtenidas en campo no fue posible debido a que de la red analizada no cuenta con información suficiente, esto además de demostrar que existen problemas en la red provoca que se replantee la forma en que se utilizan los datos para su proceso.

RECOMENDACIONES

Basado en los resultados obtenidos, los problemas presentados y lo visualizado en los recorridos en campo, se presentan una serie de recomendaciones a distintos sectores:

A LOS O.O.

- Realizar campañas de educación en el cuidado del agua, enfocados en los usos (industrial, agrícola, residencial)
- Incentivar a una toma de datos verificados y continuos
- Creación de un programa de mantenimiento constante a los micromedidores
- Crear un departamento encargado en la detección de anomalías, capacitándose en las áreas de ciencia de datos
- Realizar inspecciones enfocadas en el tipo de uso que los usuarios le dan al agua, debido a la detección de negocios como purificación de agua, lavanderías, autolavados, entre otros
- Continuar con las actualizaciones catastrales
- La creación de una base de datos que sea actualizada periódicamente
- Para la aplicación del clasificador k-NN, contar con la información necesaria para obtener resultados más precisos, el balance hidráulico de la red a analizar, las lecturas de consumo del padrón de usuarios, base de datos de presiones en varios puntos de la red; o en su caso, continuar trabajando en el clasificador con más datos de forma que permita la detección real.

AL USUARIO

- Al detectar una fuga realizar el reporte al O.O.
- Ser responsable en el pago de las cuotas por el servicio que ofrece el O.O.
- Contribuir en las acciones que el O.O. realiza en pro del mantenimiento de la red
- Comprender las situaciones y ser más accesible con el personal encargado de realizar los censos.

REFERENCIAS

Barkin, D. (2006). La gestión del agua urbana en México (pp. 1-45). Universidad de Guadalajara.

Barrios Juan Ignacio. (2019). La matriz de confusion y sus métricas. 15 septiembre 2021, de Healt Big Data Sitio web: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas

Camacho González, H., Casados Prior, J., Hansen Rodríguez, P., Cisneros Hermenegildo, A., & Ballinas González, H. (2017). Regulación de los servicios de agua potable y saneamiento en México.

Camacho Sánchez, A. (2020). La importancia del agua y su cuidado.

Cambronero, C. G., & Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. Intelgencia en Redes de Comunicación, Universidad Carlos III de Madrid, 23.

Carreño-Alvarado, E. P., Reynoso-Meza, G., Montalvo, I., & Izquierdo, J. (2017, July). A comparison of machine learning classifiers for leak detection and isolation in urban networks. In Congress on Numerical Methods in Engineering CMN.

Cárdenas Nelson Mauricio & Carlosama Gabriel Alejando (2008), Diseño de un sistema automatizado para la detección de fugas en tuberías inaccesibles. (Tesis de pregrado) Escuela Politécnica Nacional. Quito, Ecuador.

CONAGUA (2012). Memoria documental del programa para la modernización de organismos operadores de agua. México.

CONAGUA. (2019). Manual de Agua Potable, Alcantarillado y Saneamiento Datos Básicos Para Proyectos de Agua Potable y Alcantarillado, México.

Estévez Pereira, J. J. (2020). Detección de anomalías en la red empleando técnicas de machine learning.

Fuentes, O., Rodríguez, K., Jiménez, M., De Luna, F., (2004). Método para la detección de fugas en redes de distribución de agua potable usando el algoritmo genético simple, Valencia, España.

Hincapié, R. A., Porras, C. A. R., & Gallego, R. A. (2004). Técnicas heurísticas aplicadas al problema del cartero viajante (TSP). Scientia et technica, 10(24), 1-6.

Lahlou, Z. M. (2009). Detección de fugas y control de pérdida de agua. NATIONAL ENVIRONMENTAL SERVICES CENTER, 1-4.

López, F. J. A., Avi, J. R., & Fernández, M. V. A. (2018). Control estricto de matrices de confusión por medio de distribuciones multinomiales. Geofocus: Revista Internacional de Ciencia y Tecnología de la Información Geográfica, (21), 6.

López-Avila, Leyanis, Acosta-Mendoza, Niusvel, & gago-Alonso, Andrés. (2019). Detección de anomalías basada en aprendizaje profundo: Revisión. Revista Cubana de Ciencias Informáticas, 13(3), 107-123.

Maillo, J. L., Luengo, J., García, S., Herrera, F., & Triguero, I. (2018). Un enfoque aproximado para acelerar el algoritmo de clasificación Fuzzy kNN para Big Data. Asociación Española para la Inteligencia Artificial (AEPIA), 1143-1148.

Mariles O.A, Palma–Nava A. y Rodríguez–Vázquez K. (2011) Estimación y localización de fugas en una red de tuberías de agua potable usando algoritmos genéticos, Ingeniería, investigación y tecnología.

McLachlan, G. J. (1999). Mahalanobis distance. Resonance, 4(6), 20-26.

Montoya, L. J., & Montoya, R. D. (2012). Efecto de la presión sobre las fugas de agua en un sistema de tubería simple. Revista Ingenierías Universidad de Medellín, 11(20), 77-89.

Pineda S. Alejandro, Loera, E., Haro, N., Briseño, H., Flores, R., & Pérez, G. (2016). Fugas de agua y dinero: Factores político-institucionales que inciden en el desempeño de los organismos operadores de agua potable en México. El Colegio de Sonora.

Rodriguez, M., & Proyecto, I. D. E. L. (2015). Comparación de métricas de distancia en el algoritmo K-Vecinos Más Cercanos para el problema de Reconocimiento Automático de Dígitos Manuscritos. Pontificia Universidad Católica de Valparaíso, Facultad de Ingeniería, Escuela de Ingeniería Informática.

Rossman, L. A. (2001). Epanet 2 manual de usuario. US Environmental Protection Agency, Cincinnati, Ohio.

Salazar Adams, A., & Lutz Ley, A. N. (2015). Factores asociados al desempeño en organismos operadores de agua potable en México. Región y sociedad, 27(62), 05-26.

Santos-Ruiz, I., López-Estrada, F. R., Puig, V., Blesa, J., & Javadiha, M. (2019). Localización de fugas en redes de distribución de agua mediante k-NN con distancia cosenoidal.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas, 18(3), 491-504

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The annals of statistics, 35(6), 2769-2794.

Tzatchkov, V. (2007). Manual de agua potable, alcantarillado y saneamiento: datos básicos.

Viera, Á. F. G. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. Investigación bibliotecológica, 31(71), 103-126.

ANEXOS

Tabla 1. Volumen facturado mensual por calle

		VOLUMEN FACTURADO m³/MES						
		JUL	**AGO**	**SEP**	**OCT**	**NOV**	**DIC**	**ENE**
	SALVADOR ARREDONDO LOCAL 4	0	0	0	0	0	0	0
1	HDA DE PANTOJA	257	272	283	246	254	344	207
2	HDA SAN VICENTE DE GARMA	141	117	119	114	120	139	101
3	HDA SAN JOAQUIN	499	332	405	335	341	362	337
4	HDA CERRO GORDO	15	17	19	22	17	24	16
5	HDA DE CERRITOS	253	231	250	203	219	257	233
6	HDA DE LA IGLESIA	167	118	164	129	134	147	184
7	HDA DE QUIRICEO	234	227	269	218	216	251	226
8	HDA DE SAN IGNACIO	235	262	271	209	212	302	205
9	HDA SAN ANTONIO TERAN	242	216	266	189	192	179	161
10	HDA SAN ISIDRO	98	105	128	124	113	122	69
11	HDA SAN JAVIER NTE	1361	1275	1436	1228	1265	1340.5	1206.5
12	HDA SAN JAVIER SUR	1498	1425	1601	1366	1364	1620	1208
13	HDA SAN NICOLAS 11	565	426	464	375	458	513	397
14	HDA SAN RAFAEL DE SAUZ	252	222	246	225	234	256	240
15	HDA SANTA ANA	231	250	277	225	226	249	194
16	HDA SANTA BARBARA	269	249	288	245	242	237	243

Tabla 2. Consumo mensual por calle convertido a litros por segundo

VOLUMEN LPS POR CALLE								
		JUL	**AGO**	**SEP**	**OCT**	**NOV**	**DIC**	**ENE**
	SALVADOR ARREDONDO LOCAL 4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	HDA DE PANTOJA	0.0960	0.1016	0.1092	0.0918	0.0980	0.1284	0.0773
2	HDA SAN VICENTE DE GARMA	0.0526	0.0437	0.0459	0.0426	0.0463	0.0519	0.0377
3	HDA SAN JOAQUIN	0.1863	0.1240	0.1563	0.1251	0.1316	0.1352	0.1258
4	HDA CERRO GORDO	0.0056	0.0063	0.0073	0.0082	0.0066	0.0090	0.0060
5	HDA DE CERRITOS	0.0945	0.0862	0.0965	0.0758	0.0845	0.0960	0.0870
6	HDA DE LA IGLESIA	0.0624	0.0441	0.0633	0.0482	0.0517	0.0549	0.0687

7	HDA DE QUIRICEO	0.0874	0.0848	0.1038	0.0814	0.0833	0.0937	0.0844
8	HDA DE SAN IGNACIO	0.0877	0.0978	0.1046	0.0780	0.0818	0.1128	0.0765
9	HDA SAN ANTONIO TERAN	0.0904	0.0806	0.1026	0.0706	0.0741	0.0668	0.0601
10	HDA SAN ISIDRO	0.0366	0.0392	0.0494	0.0463	0.0436	0.0455	0.0258
11	HDA SAN JAVIER NTE	0.5081	0.4760	0.5540	0.4585	0.4880	0.5005	0.4505
12	HDA SAN JAVIER SUR	0.5593	0.5320	0.6177	0.5100	0.5262	0.6048	0.4510
13	HDA SAN NICOLAS	0.2109	0.1591	0.1790	0.1400	0.1767	0.1915	0.1482
14	HDA SAN RAFAEL DE SAUZ	0.0941	0.0829	0.0949	0.0840	0.0903	0.0956	0.0896
15	HDA SANTA ANA	0.0862	0.0933	0.1069	0.0840	0.0872	0.0930	0.0724
16	HDA SANTA BARBARA	0.1004	0.0930	0.1111	0.0915	0.0934	0.0885	0.0907

Tabla 3. Consumo lps por cada nodo por calle en el mes de septiembre

VOLUMEN LPS POR NODO			
		NODOS	lps por nodo
	SALVADOR ARREDONDO LOCAL 4	0	0.0000
1	HDA DE PANTOJA	1	0.1092
2	HDA SAN VICENTE DE GARMA	5	0.0092
3	HDA SAN JOAQUIN	5	0.0313
4	HDA CERRO GORDO	2	0.0037
5	HDA DE CERRITOS	4	0.0241
6	HDA DE LA IGLESIA	4	0.0158
7	HDA DE QUIRICEO	1	0.1038
8	HDA DE SAN IGNACIO	4	0.0261
9	HDA SAN ANTONIO TERAN	3	0.0342
10	HDA SAN ISIDRO	2	0.0247
11	HDA SAN JAVIER NTE	23	0.0241
12	HDA SAN JAVIER SUR	21	0.0294
13	HDA SAN NICOLAS	4	0.0448
14	HDA SAN RAFAEL DE SAUZ	2	0.0475
15	HDA SANTA ANA	1	0.1069
16	HDA SANTA BARBARA	2	0.05556

Tabla 4 configuración de datos para algoritmo

	Train	*Test*
ESCENARIOS	0	1.1
	1.2	1.3
	1.4	1.5
	2.2	2.1
	2.4	2.3
	3.2	2.5
	3.4	3.1
	4.2	3.3
	4.4	3.5
	5.2	4.1
	5.4	4.3
		4.5
		5.1
		5.3
		5.5

Sript k-NN

```
load WS
for i=0:10
    Inp_Train(i+1,:)=TRAIN(i*25+1:i*25+25,3)';
end

Tar_Train=[0; 1; 1; 2; 2; 3; 3; 4; 4; 5; 5];
for i=0:14
    Inp_Test(i+1,:)=TEST(i*25+1:i*25+25,3)';
End

Tar_Test=[1;1;1;2;2;2;3;3;3;4;4;4;5;5;5];
```

```
ML_kNN=fitcknn(Inp_Train,
Tar_Train,"NumNeighbors",3,"Distance","euclidean");
Out_kNN=predict(ML_kNN,Inp_Test);
DataSetFea=[Inp_Train; Inp_Test];
DataSetTar=[Tar_Train; Tar_Test];
for rodada=1:25
    Inp_Train=[];
    Inp_Test=[];
    Tar_Train=[];
    Tar_Test=[];
    for y=1:5
        yy=find(DataSetTar==y);
        yyy=randperm(5);
        Inp_Train=[Inp_Train; DataSetFea(yy(yyy(1:2)),:)];
        Inp_Test=[Inp_Test; DataSetFea(yy(yyy(3:5)),:)];
        Tar_Train=[Tar_Train; DataSetTar(yy(yyy(1:2)),:)];
        Tar_Test=[Tar_Test; DataSetTar(yy(yyy(3:5)),:)];
    end
    ML_kNN=fitcknn(Inp_Train,
Tar_Train,"NumNeighbors",3,"Distance","euclidean");
    Out_kNN=predict(ML_kNN,Inp_Test);
    [Tar_Test Out_kNN]
    subplot(5,5,rodada)
    confusionchart(Tar_Test, Out_kNN)
end
```