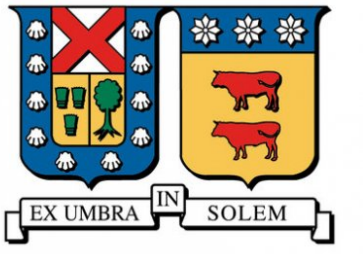
2019

DETECCIÓN DE ALOMALÍAS EN EL PAGO DE LA CUENTA DE AGUA

CATALÁN FAÚNDEZ, ANIBAL EDUARDO

Detección de anomalías en el pago de la cuenta de agua

Anibal Catalán



Valparaíso, 2019

Detección de anomalías en el pago de la cuenta de agua

Anibal Catalán

Universidad Técnica Federico Santa María
Departamento de Electrónica
Ingeniería Civil Telemática

Profesor Guía
Werner Creixell

Profesor Co-Referente
Ronny Vallejos

Índice general

Capítulo 1

Agradecimientos

Este trabajo representa crecer, es lograr un objetivo, que no hubiera podido ser realidad sino fuera por el apoyo de mi madre Rosa, mi padre Rafael y mis Hermanos Felipe y Rafael, que a pesar de las vicisitudes de nuestra historia, siempre han estado y estarán.

Quiero agradecer la compañía de los que caminaron este mismo sendero, muchas veces lejos de nuestras familias, y que vimos a la Universidad como un segundo hogar.

A la Generación de telemáticos 2009, que en los primeros años me hicieron sentir un curso como los del colegio y que a lo largo del camino fueron tomando distintos rumbos. De los pocos que quedamos, pudimos irnos conociendo bien a lo largo de los años y hasta el día de hoy tenemos, veo a un par que me gustaría mencionar.

A Gonzalo Sanchez y Daniel Veas por sus nobles corazones, apoyo, risas, juegos y horas de estudio compartido, a todos los que me ayudaron un par de días antes de alguna prueba a entender todo lo que entraba, con su enseñanza, de pares, supe nadar sobre todas las olas que vinieron.

Al equipo de Taekwondo y al profesor Pinochet, por las incontables horas de entrenamiento, campeonatos y la educación moral que conllevan las Artes Marciales, Disciplina y Amor es lo que recojo de esto, valores que me han ayudado siempre en cada momento.

A la carrera Telemática, por su particular personalidad, la unión en la diferencia, las risas, el apañe y corazón que tienen, que los diferencia de todas las demás y por haber tenido la oportunidad de contribuir en la construcción de esta comunidad, que al mismo tiempo me hizo crecer tanto.

Y por ultimo al profesor Milan Derpich, un maestro con un corazón enorme, por la luz que refleja en cada una de sus clases, que me hizo comprender que las ondas, y campos electromagnéticos, son parte esencial de la naturaleza y la vida.

Capítulo 2

Resumen

Las palabras claves de este trabajo son:

- **Perdidas Aparentes**: Corresponden a consumos de agua potable no autorizados o consumos con medición defectuosa.
- **Método de Clasificación**: Corresponde a un modelamiento matemático, que permite categorizar o clasificar una entidad con respecto a una o muchas variables (datos sobre la entidad) de comportamiento de la misma.
- **Data Mining**: Proceso que intenta descubrir patrones en grandes Volúmenes de conjuntos de datos.
- **Algoritmo de Clasificación**: Conjunto de acciones destinadas a separar un conjunto de datos en subconjuntos basados en criterios definidos por el usuario.
- **Matriz de Confusión**: Es una matriz que categoriza los resultados de los algoritmos. Los elementos de la diagonal principal de dicha matriz corresponde a las decisiones tomadas correctamente, mientras que lo elementos fuera de esta diagonal corresponde a aquellos elementos en donde nos equivocamos al momento de predicción.
- **Validación Cruzada**: Técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.

Capítulo 3

Introducción

En Chile las empresas sanitarias produjeron en su conjunto un total de 1.670 millones de metros cúbicos de agua potable de la cual un 33,65 % no fue facturada, de ese total estimaciones de la SISS establecen que el 74 % se originan por pérdidas físicas, mientras que el 26 % restante corresponde a pérdidas aparentes, esto equivale a cerca de 146 millones de metros cúbicos de agua potable, 8.749 % del total del agua potable producida.

Esto constituye un problema para las compañías de agua, que ven disminuidas sus ganancias en un 8.749 % a causa de este tipo de pérdida.

En el caso de las pérdidas aparentes, estas pueden ser producto de consumos no autorizados o mediciones defectuosas, para reducir las pérdidas por este concepto se han propuesto dos tipos de soluciones, medidores inteligentes y detección de anomalías basados en el consumo histórico de un cliente.

- **Medidores Inteligente**: Nueva generación de medidores conectados a internet que permiten una lectura a distancia 24/7, pudiendo detectar cambios en tiempo real.
- **Detección de consumos anómalos**: Modelo matemático que usa datos históricos de consumo de agua de los clientes, con el fin de detectar anomalías en estos, estas anomalías se pueden referir a un consumo no registrado o a una medición defectuosa. El modelo se basa en el supuesto de que un cliente mantiene un historial de consumo sin una gran desviación en el tiempo.

La detección de anomalías en el consumo de agua potable ha sido estudiado con anterioridad, sin embargo, el énfasis investigativo se ha inclinado por la detección de pérdidas físicas y en particular en lo referido a las filtraciones en lineas de distribución. Solo un trabajo de grado de la Universidad Islámica de Gaza ha usado un enfoque de análisis de datos.

La investigación se enfoco en crear un modelo que permita obtener la mayor precisión a la tarea de detectar pérdidas aparentes por parte de los usuarios que consumen agua a través de una compañía de servicios básicos. Se probaron distintos modelos los cuales entregaran un resultado basado en sistemas de clasificación, estos fueron categorizados,

ordenados y combinados con el fin de entregar el mejor resultado posible, comunicando a la compañía proveedora de servicios básicos una mejor idea de como el consumo de agua se comporta en un determinado sector.

La hipótesis principal es que el usuario de servicios básicos de agua tiene un consumo que no presenta una desviación muy grande entre uno y otro mes de distintos años, hasta que su medidor se avería y/o es modificado presentando desviaciones significativas que pueden ser detectadas. El modelo matemático es entrenado con los datos históricos de consumo que ya hayan sido detectados por las fiscalizaciones, es decir, al método de clasificación se le enseña que debe detectar, para después ser usado con datos de usuarios que no han sido detectados.

Las variables serán los datos proporcionados por la compañía de servicios básicos, estos serán datos históricos de consumo de 5 años. Estos datos pasaran por un proceso de limpieza y transformación de los mismos, para ser ocupados de forma correcta en los algoritmos de clasificación.

Los algoritmos usados mostraron resultados alentadores, pudiendo identificar hasta con un 92,28 % de exactitud clientes que están en situación de pérdidas aparentes.

Además la metodología para categorizar estos tipos de usuarios según la cantidad de algoritmos que los detecto, muestra ser una estrategia acertada, en el gráfico de dispersión del consumo se muestra que que entre mas algoritmos clasifiquen a un usuario como con posible pérdida aparente existen menos outlayers, pareciéndose mas al conjunto que ya se sabe que tiene pérdida aparente.

Comparado con el porcentaje de detecciones exitosas de hoy en día, 10 %, este trabajo demuestra teóricamente un aumento de por lo menos 70 % en el porcentaje usuarios que al ser fiscalizados efectivamente tengan pérdidas aparentes.

Capítulo 4

Estado del Arte

El problema tanto de la cuantificación como de la detección de pérdidas en sistemas de agua potable ha sido estudiado de manera consistente, sin embargo, el énfasis investigativo se ha inclinado por la detección de pérdidas físicas y en particular en lo referido a las filtraciones en líneas de distribución, de esta forma es posible encontrar estudios enfocados en la detección mediante el análisis de vibraciones en tuberías soterradas, utilización de algoritmos genéticos, etc.

4.1. Perspectivas generales de cuantificación de pérdidas, caso Kampala, Uganda.

Teniendo en consideración que los valores particulares de las pérdidas no son necesariamente extrapolables entre poblaciones de características disímiles, resulta valioso considerar el enfoque de cuantificación de pérdidas aparentes o comerciales utilizado en la ciudad de Kampala, Uganda (Mutikanga, Sharma, y Vairavamoorthy, 2010)

Los autores proponen separar el análisis en base a grupos causales de pérdidas:

- **Medidores imprecisos**: Bajo la premisa de que al ser un dispositivo mecánico, los medidores se ven sometidos a desgaste, en consecuencia sufren una merma en su precisión, y siguiendo métodos de muestreo estadístico previamente establecidos, se consideraron categorías de medidores de acuerdo a su antigüedad que fueron sometidos a pruebas con distintos niveles de flujo.
- **Errores en la lectura**: Dado que en Kampala, al igual que en Chile, la recolección de datos de consumo se realiza de manera manual por inspectores, existen errores humanos inherentes en la lectura.
- **Manejo de datos y errores de facturación**: Al comparar los datos entregados por los inspectores al realizar la lectura con los datos finales en sistema se detectan nuevos errores.

- **Consumos no autorizados**: Se considera la alteración de medidores, la instalación de bypass, reposición no autorizada de servicio y conexiones no autorizadas a la red. Una vez detectados se calcula el volumen de consumo no autorizado como un promedio de los consumos mensuales previos a la falta, lo anterior en base a la detección mediante auditorías en terreno.

De esta forma los autores concluyen que, para el caso de Kampala, la estimación de la desagregación de las pérdidas aparentes se expresan como:

- **Medidores imprecisos**: 22±2 % de la facturación.
- **Errores en la lectura**: 1,4±1 % de la facturación.
- **Manejo de datos y errores de facturación**: 3,5±0,5 % de la facturación.
- **Consumos no autorizados**: 10±2 % de la facturación.

La metodología se basa en considerar muestras representativas y extrapolar al tamaño total de la población.

4.2. Estrategias de prevención, remediación y disuasión.

Existen sistemas de grillas de medidores inteligentes que prescinden de una de las fuentes de error definidas anteriormente, pues no requieren la participación de inspectores para la lectura de medidores, sino que envían los datos de manera automática disminuyendo también pérdidas por manejo de datos. Sin embargo, estos sistemas son susceptibles de intrusión, ante esto surgen alternativas comerciales de protección de grillas (McCullough, 2010).

4.3. Modelos de detección.

Se han generado en los últimos años modelos de detección basados en el análisis de datos, en particular destaca el utilizado en la ciudad de Gaza, Palestina (Humaid & Barhoom, 2012), el que considera la utilización de distintas técnicas:

4.3.1. Data Mining

Se entenderá por Data Mining como un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

Los algoritmo que se ocupan son de dos tipos:

- **Algoritmos supervisados (o predictivos)**: Predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.

- **Algoritmos no supervisados (o del descubrimiento del conocimiento)**: Descubren patrones y tendencias en los datos.

Metodología del Data Mining

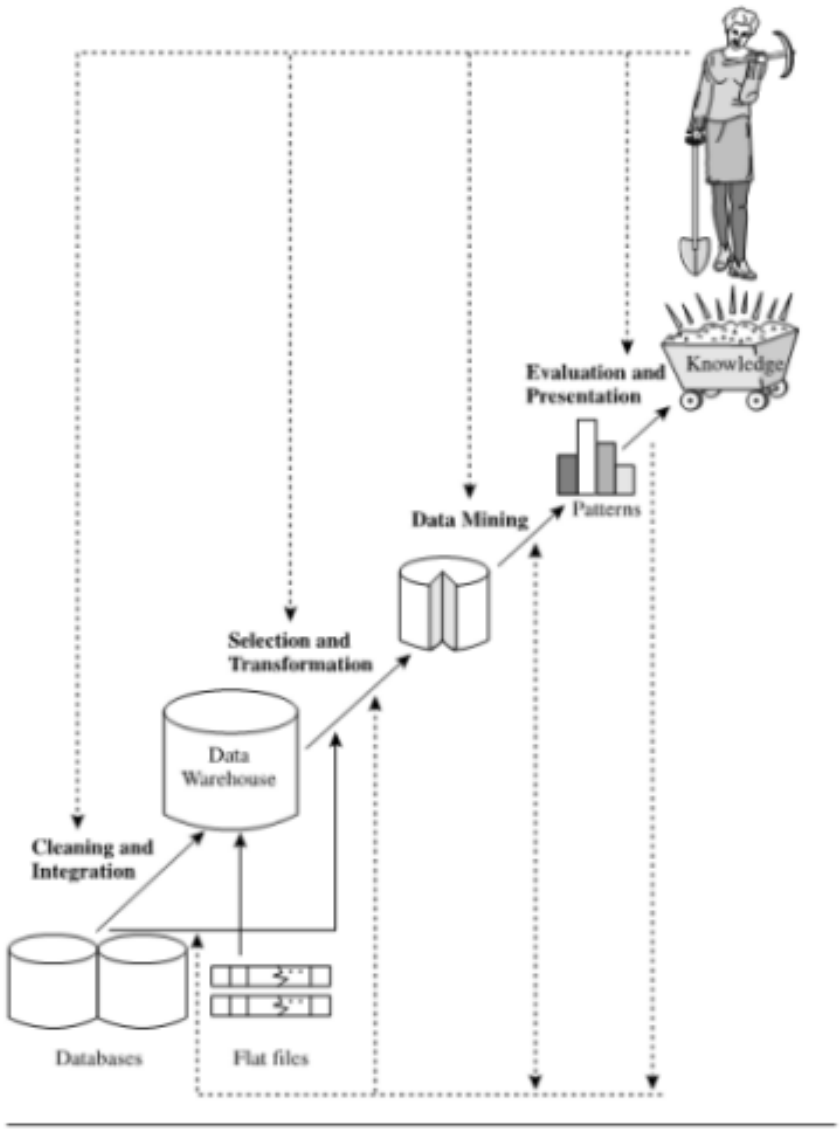


Figura 4.1: Proceso de data mining.

Como se aprecia el la Figura 4.1, el proceso de data mining consta de cinco etapas, estas son:

– **Limpieza de datos**: Datos que provienen de una o muchas fuentes, los cuales se ordenan y modifican para ser combinados.

- **Integración de datos**: Unión coherente datos de distintas fuentes previamente ordenados.
- **Transformación de datos**: Transformaciones aplicadas a los datos con el fin de realizar un análisis matemático de su conjunto.
- **Evaluación de patrones**: Evaluación cuantitativa y cualitativa del análisis realizado anteriormente.
- **Presentación de la información**: Traducción conceptual y practica de la información generada en el proceso.

4.3.2. Clasificación

Se definen dos tipos de clasificación:

- **Lineal**: Modelo en el cual se presentan relaciones lineales entre los coeficientes, gráficamente se ven como la Figura 4.2 y son del tipo: $a_1 * x_1 + a_2 * x_2 + \cdots + a_z * x_z = a_0$. Los modelos lineales son computacionalmente fáciles de implementar y son estables, sin embargo, en la realidad esto no sucede, por lo cual suponer un modelo lineal a un conjunto de dato que no lo es, introduce un sesgo que no puede corregirse en las predicciones.

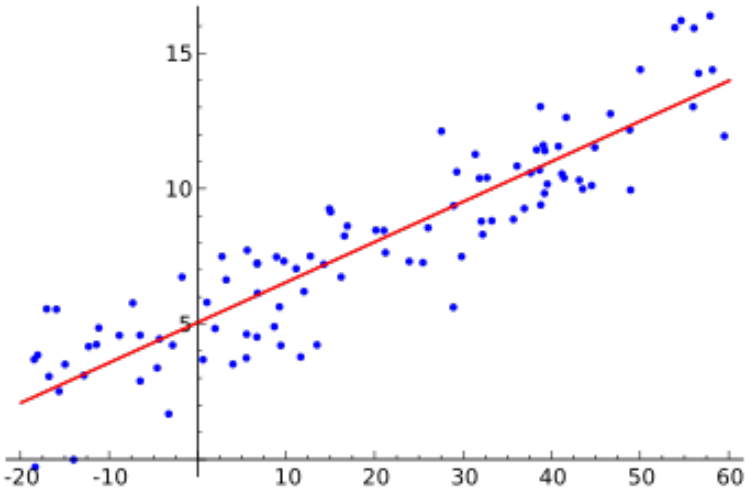


Figura 4.2: Modelo lineal.

- **No lineal**: Modelo en el cual se presentan relaciones no lineales entre los coeficientes, gráficamente se ven como la Figura 4.3 y son del tipo: $a_1 * x_1 + a_2 * x_2^2 + \cdots + a_z * x_z^z = a_0$. Los modelos no lineales son computacionalmente mas difíciles de implementar que los lineales, sin embargo, se acercan mejor a la realidad.

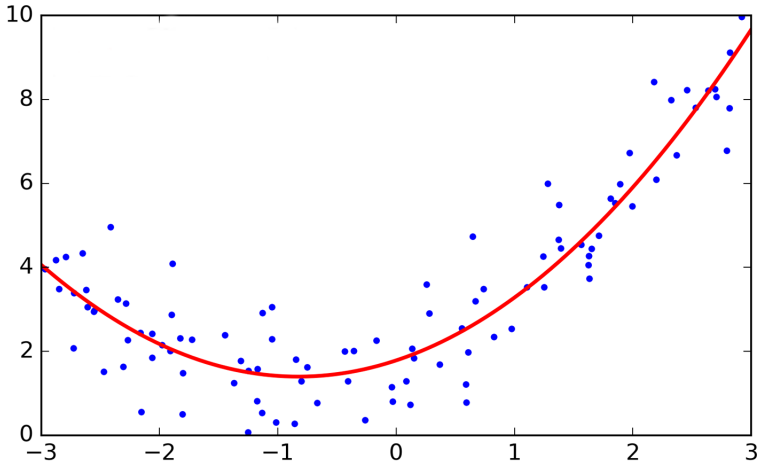


Figura 4.3: Modelo no lineal.

4.3.3. Métodos de clasificación

- **SVM (support vector machine)**
  Este método está propiamente relacionado con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Una SVM es un modelo que representa a los puntos de muestra en el espacio separando las clases a 2 espacios lo más amplios posibles, mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, mas cercanos al que se llama vector soporte. Las nuevas muestras se ponen en correspondencia con dicho modelo en función de los espacios a los que pertenezcan, siendo clasificadas en una u otra clase como se puede apreciar en la Figura 4.4 .

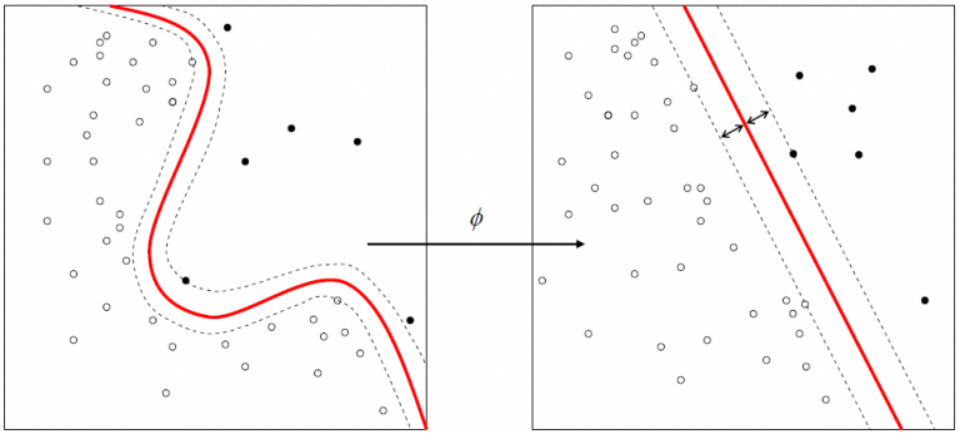


Figura 4.4: Clasificación en support vector machine.

- **Redes neuronales (Neuronal Metwork)**
  Las redes de neuronas artificiales (denominadas habitualmente como RNA o en inglés como: "ANN" son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso biológico. Se trata de un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida. En inteligencia artificial es frecuente referirse a ellas como redes de neuronas o redes neuronales. En la Figura 4.5 se muestra el diagrama general de una red neuronal multicapa.

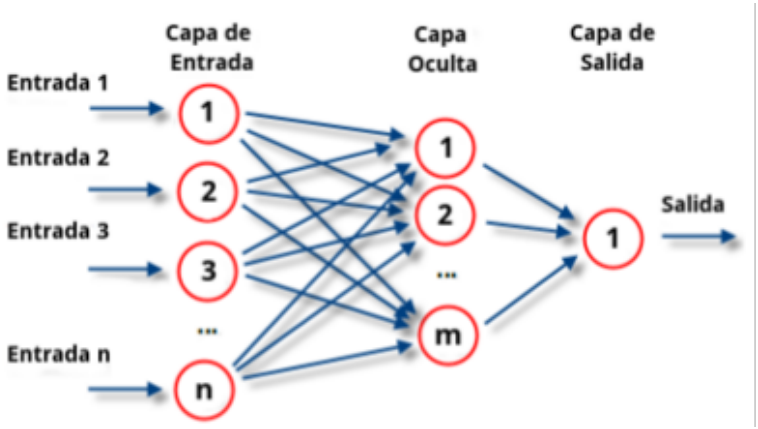


Figura 4.5: Diagrama de red neuronal multicapa.

- **K-Nearest-Neighbor (KNN)**
  Es un método de clasificación supervisada (Aprendizaje basado en un conjunto de entrenamiento y prototipos) que sirve para estimar la función de densidad $F(x/C_j)$ de las predictoras x por cada clase C_j. En la Figura 4.6 se muestra un ejemplo grafico.

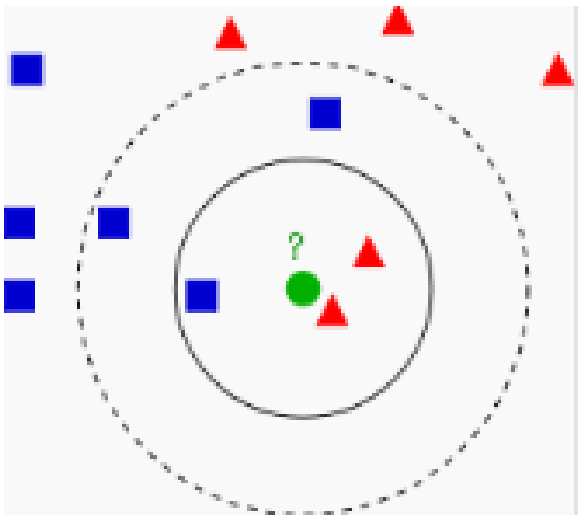


Figura 4.6: Ejemplo de implementación de KNN.

Este es un método de clasificación no paramétrico, estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j a partir de la información proporcionada por el conjunto de prototipos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

4.3.4. Medición del desempeño

Para este tipo de problemas es posible ocupar muchos modelos que clasifiquen el problema de buena forma, pero el objetivo es busca aquel que tiene un mayor grado de precisión, por lo que se debe encontrar una medida de comparación entre los modelos, para ello se ocuparán dos medidas de desempeño:

- **Matriz de Coincidencia**
  Como se muestra en la Figura 4.7, la matriz de coincidencia corresponderá a una matriz de NxN, donde N corresponde al número de clasificadores, en la cual los elementos de la diagonal principal de dicha matriz corresponde a las decisiones tomadas correctamente, mientras que lo elementos fuera de esta diagonal corresponde a aquellos elementos donde el modelo aplicado a fallado en su predicción.

Matriz de Confusión		RESPUESTA	
		SI	NO
SEÑAL	SI	INTENTO ACERTADO	INTENTO FALLADO
	NO	FALSA ALARMA	RECHAZO CORRECTO

Figura 4.7: Matriz de Coincidencia.

- **Validación cruzada (Cross Validation)**
  Como se muestra en la Figura 4.8, la validación cruzada es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar cómo de preciso es un modelo que se llevará a cabo a la práctica.1 Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados.

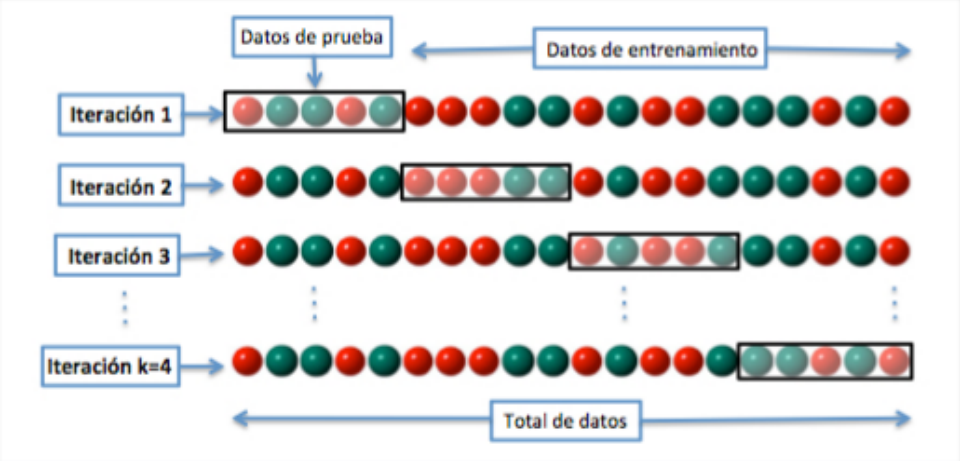


Figura 4.8: Validación cruzada.

A menudo el conjunto de prueba se toma como un tercio del conjunto total de datos, aunque hay criterios heurísticos que proponen elegir el 25 % de los datos, ya que con esto la validación cruzada mejora.

4.4. Comprensión del negocio

4.4.1. Análisis trabajo de grado de la Universidad Islámica de Gaza

Se obtuvieron datos por un período de 144 meses que representan consumos desde 03/2000 al 02/2012, comprendiéndose dos tipos de datos, que son: La información de facturación al cliente del sistema (perfiles de consumo) e irregularidades de agua de los clientes de datos (casos de fraude).

- **Entendiendo los datos**
  En la Figura 4.9 se muestran los tipos de datos obtenidos sin procesar.

Table 4.1 The attributes that extracted from historical water consumptions data.

Column	Description
Agreement_id	The customer account no
Service_type	The billing service type (water or electricity)
Meter_no	Customer water meter number
Reading_date	Water consumption Reading date
Previous_Reading	Meter Previous reading
Reader_id	The meter reader code to identify the meter reader name
Meter_status	To record if the meter normal or destroyed and othe states
Calc type	Record if the consumption automatic averaged or real quantity by reader
Current reading	Meter current reading
Batch_no	Reading batch or file no
Consumption_qty	Consumption quantity
Location_id	The location number to identify the building

Figura 4.9: Tipo de datos.

Los datos anteriormente nombrados se usaron para crear nuevos tipos de datos que se muestran en la Figura 4.10, estos permiten una mejor comprensión del problema.

Table 4.2 The list of calculated attributes (to be evaluated in feature selection phase).

Column	Description
Building_agr_count	Water consumption accounts counts in the same building
Persons_per_building	Number of persons per building
Persons_per_unint	Number of persons per one unit
Payment_count_pct	The percentage of monthly paid voucher according to number of invoices
Paid_voucehr_count_pct	The percentage of paid vouchers according to invoices count

Figura 4.10: Tipos de datos creados.

También se agregaron nuevos tipos de datos que se muestran el la Figura 4.11 estos provienen de la municipalidad de Gaza (CWBD). Son datos que recogen los últimos 12 años de clientes a los cuales se les ha detectado fraude.

Table 4.3 list the attributes of fraudulent cases.

Column	Description
Agreement_id	The customer account no.
Breach date	The date of water breach
Breach cost	The money needed to pay for that breach
Location_id	The building and street number

Figura 4.11: Tipo de datos de la municipalidad.

- **Procesamiento de los datos**
  La Figura 4.12 corresponde a un esquema de como los datos mencionados anteriormente fueron procesados, con el fin de ser ocupados por los algoritmos de clasificación.

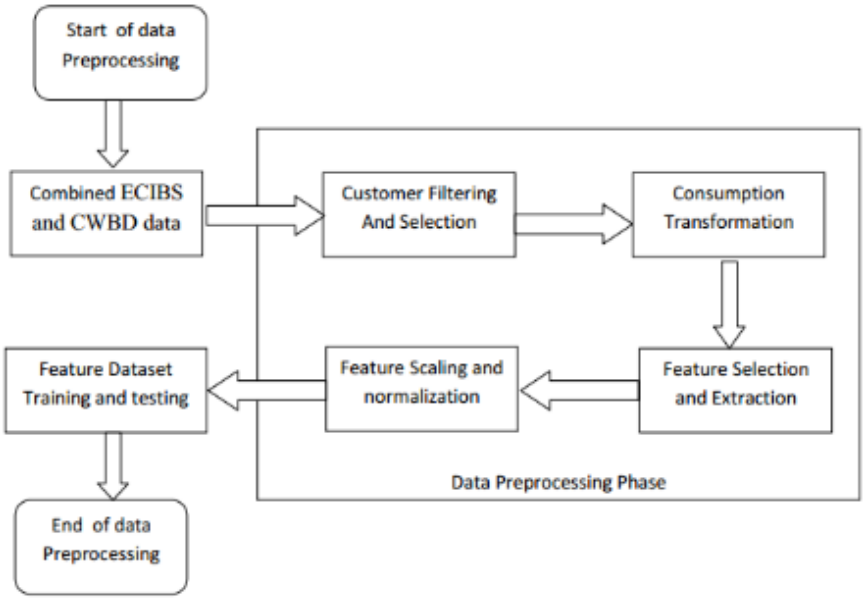


Figura 4.12: Esquema de procesamiento.

- **Análisis de los perfiles**
  Se trabajó con un conjunto de datos de 4774 individuos, donde 4114 corresponden a individuos con un perfil normal y los otros 660 corresponde a clientes con un perfil fraudulento, este tipo de muestra se tomó con el fin de entrenar adecuadamente un modelo SVM.

  A continuación en la figura 4.13, se presentan los dos perfiles de usuarios con sus respectivos consumos(en m^3). El gráfico superior muestra usuarios con pérdidas y el inferior usuarios normales.

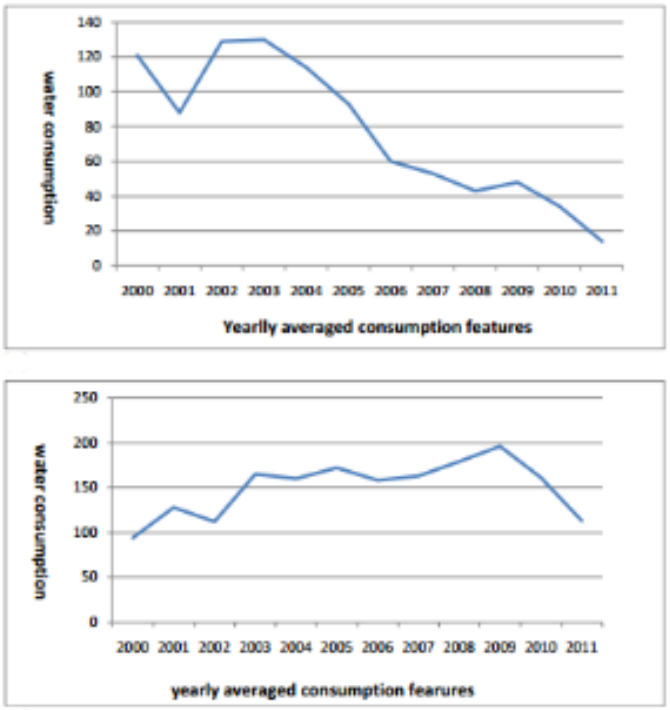


Figura 4.13: Perfiles de consumo.

- **Predicciones**
  La Figura 4.14 muestra los resultados de las predicciones a través de un análisis anual, estacional y mensual.

Data Set Type	Classifier	Performance					Data Set Status
		Accuracy	Recall		Precision		
			Yes	No	Yes	NO	
Yearly_DS (Load Profile)	SVM	**87.62**	**56.52**	92.56	54.69	93.05	Balanced
Yearly_DS (All Selected Attr.)	SVM	**85.98**	**61.06**	89.94	49.09	93.56	Balanced
Seasonally_DS (Load Profile)	SVM	**93.12**	**81.32**	95.14	74.06	96.76	Balanced
Seasonally_DS (All Selected Attr.)	SVM	**93.76**	**80.61**	95.87	75.78	96.86	Balanced
Monthly_DS (Load Profile)	SVM	**92.29**	**79.55**	93.81	67.14	96.65	Balanced
Monthly_DS (All Selected Attr.)	SVM	**91.86**	**80.45**	95.65	74.79	96.83	Balanced

Figura 4.14: Predicciones.

4.4.2. Modelos de detección en otros mercados

- **Ley de Benford para detección de fraudes en contabilidad**

Se usaron datos de las prácticas de contabilidad del hospital en la ciudad de Major en Albania, donde se detectaron anomalía de indicadores (Asllani & Naco, 2014).

La Ley de Benford se basa en la observación única que ciertos dígitos aparecen con mayor frecuencia que otros, con esta información se calcula la probabilidad de que el primer dígito sea uno o cero o distinto de cero y la probabilidad del mismo evento, pero para el segundo dígito.

De esta forma, si la distribución de los dígitos de los datos no sigue la distribución de probabilidades obtenidas, existe una razón para creer que los datos fueron manipulados por intervención humana, y estos datos serán investigados por posibles fraudes. Predpol.

Algoritmo de predicción de crímenes el cual necesita las siguientes entradas: tipo de crimen, donde ocurrió y la fecha exacta de dicho crimen, como respuesta, el algoritmo delimita las áreas de una ciudad donde es más probable a que se cometa un crimen.

Como se muestra en la Figura 4.15, la forma que se entrega la información trata de ser lo más amigablemente posible con las patrullas policiales, de tal forma que esto puedan leer los mapas de manera sencilla y eficaz.

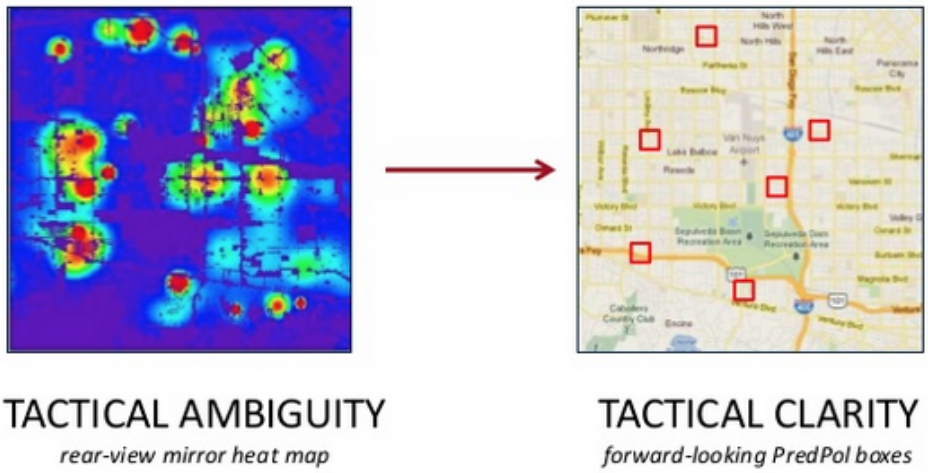


Figura 4.15: Interfaz de Predpol.

Respecto a los estudios realizados en donde se aplicó dicho software, el número de crímenes se redujo en un 19 %.

Capítulo 5

Planteamiento del Problema

5.1. Cobertura actual de servicios de aguas

Desde la llegada de la modernidad el acceso a los servicios básicos ha ido cambiando, por ejemplo la provisión de agua para beber comenzó siendo responsabilidad individual de los usuarios quienes acudían a diversas fuentes, posteriormente concurrían a pozos centralizados y regularizados, sin embargo la llegada de la energía eléctrica domiciliaria cambió la forma de acceso pues no era posible para un usuario individual generar su propia energía, teniendo que acudir a terceros que proveían el servicio generándose un monopolio natural.
El incremento de los requisitos básicos para el consumo de agua junto con el aumento de la población en los núcleos urbanos, hizo necesario que se normalizara y centralizara el acceso, dando paso a las empresas de servicios sanitarios.
En Chile la cobertura desde el año 1965 se ha incrementado de manera sostenida en el tiempo, como se aprecia en el Cuadro 5.1, llegando a un nivel de acceso del 99,9 % de la población (Superintendencia de Servicios Sanitarios, 2014), lo que lo coloca por sobre el promedio latinoamericano que es de un 93 % (Soulier Faure, Ducci, & Altamira, 2013), ubicándose a niveles comparables con países como Noruega o Dinamarca (OECD, 2007).

Año	Población Millones Habs.	Cobertura Agua Potable Urbana (%)	Cobertura Alc. Urbano (%)
1965	5.85	53.5	25.4
1966	6.01	56.3	26.0
1967	6.18	59.1	26.8
		...	
2012	15,7	99,9	96,3
2013	16,1	99,9	96,5
2014	16,5	99,9	96,7

Cuadro 5.1: Evolución de la cobertura de agua potable en Chile, Datos SISS

5.2. Tipos de consumos y pérdidas

En términos de consumos, la Asociación Internacional de Agua (IWA) en conjunto con la Asociación americana de trabajos en Agua (AWWA) define los siguientes tipos de consumos (American Water Works Association) que se muestran en el Cuadro 5.2:

<table>
<tr><td rowspan="8">Suministro de Agua Potable</td><td rowspan="4">Consumos Autorizados</td><td rowspan="2">Consumos Autorizados Facturados</td><td>Consumos Medidos Facturados a Clientes Registrados</td></tr>
<tr><td>Consumos No Medidos Facturados a Clientes Registrados</td></tr>
<tr><td rowspan="2">Consumos Autorizados No Facturados</td><td>Medidos</td></tr>
<tr><td>No Medidos</td></tr>
<tr><td rowspan="4">Pérdidas de Agua</td><td rowspan="2">Pérdidas Aparentes</td><td>Consumos No Autorizados</td></tr>
<tr><td>Consumos Con Medición Defectuosa</td></tr>
<tr><td rowspan="2">Pérdidas Físicas</td><td>Fugas en Redes</td></tr>
<tr><td>Fugas y Rebalses en Tanques de Almacenamiento</td></tr>
<tr><td colspan="2"></td><td>Fugas en puntos de Servicios</td></tr>
</table>

Cuadro 5.2: Componentes y Definiciones del Balance de Agua, IWA/AWWA

- Suministro de Agua Potable: Corresponde al volumen anual de agua potable inyectada al sistema.
- Consumos autorizados: Corresponde al volumen anual de agua medida y/o no medida entregada a clientes autorizados.
- Pérdidas de Agua: Diferencia entre el agua inyectada al sistema y los consumos autorizados.
- Pérdidas aparentes: Consumos no autorizados, todo tipo de imprecisiones de medición, y errores sistemáticos de manejo de datos.
- Pérdidas Físicas: El volumen anual de pérdidas mediante todo tipo de filtraciones, fugas, roturas o rebalse en redes, almacenamiento o puntos de servicio, hasta el punto de medición del cliente.

5.3. Cuantificación de las pérdidas

De acuerdo a los datos contenidos en el Informe de Gestión del Sector Sanitario 2014 presentado por la Superintendencia de Servicios Sanitarios (SISS), en Chile las empresas sanitarias produjeron en su conjunto un total de 1.670 millones de metros cúbicos de agua potable de la cual un 33,65 % no fue facturada, de ese total estimaciones de la SISS establecen que el 74 % se originan por pérdidas físicas, mientras que el 26 % restante corresponde a pérdidas aparentes, esto equivale a cerca de 146 millones de metros cúbicos de agua potable, 8.749 % del total del agua potable producida.

Este es un claro problema para las compañías de agua, que ven disminuidas sus ganancias en un 8.749 % a causa de este tipo de pérdida, esto origina las siguientes preguntas:

5.4. Cuestionamiento a partir del problema

¿Como podemos aumentar el porcentaje de aciertos en la detección de pérdidas aparentes?

¿Como podemos hacer mas eficiente el proceso de fiscalización?

¿Existe una relación entre los usuarios que se les detecta una pérdida aparente y su situación socio-económica?

5.5. Objetivos

Mediante el presente trabajo de memoria multidisciplinaria se pretende generar un modelo de análisis de datos de los consumos de servicios de agua, este permitirá a las empresas de servicios sanitarios detectar de manera temprana los puntos en que se genere pérdida aparente, ya sea por mal funcionamiento técnico del medidor de agua potable o por su manipulación por parte de los usuarios.
Lo anterior con un doble propósito, por una parte disminuir el total de pérdidas aparentes y por otra disminuir los costos asociados a la fiscalización al focalizarla a puntos específicos, además de cuantificar estos efectos.

5.6. Enfoque actual

Actualmente en Chile la fiscalización de este tipo de pérdidas se hace por lo general de manera aleatoria, este mecanismo suele ser muy ineficiente alrededor de un 10 % de éxito en la fiscalización, por lo que es razonable incorporar detección temprana y focalización de las fiscalizaciones a individuos y/o poblaciones con mayor "posibilidad"de estar en esta situación, con la finalidad de reducir las pérdidas aparentes.

Hasta el día de hoy este enfoque solo se ha usado en un lugar, en la Ciudad de Gaza, el modelo que desarrollaron incremento la detección de entre 1-10 % de detecciones aleatorias exitosas, a un 80 % con la detección ïnteligente".
La investigación se realiza a partir de una muestra de 30348 datos de consumos mensuales de agua por doce años, desde el 2000 al 2012, de los cuales 2700 son clientes que registran pérdidas aparentes, fue un trabajo hecho por Eyad Hashem S.Humaid de la Universidad Islámica de Gaza.

En el presente trabajo se aplicara un modelo similar en una geografía y cultura diferente, además se añadirá una categoría socio-económica para descubrir como estas afectan las clasificaciones.

También se incluirá la evaluación de los mas recientes métodos de análisis de datos, demostrando si estos pueden aumentar el porcentaje de aciertos y/o complementar los resultados obtenidos en Gaza.

Este modelo puede ser usado por las compañías de agua para crear procesos de detección y focalizar la fiscalización de pérdidas aparentes, disminuyendo sus pérdidas por este concepto y el costo operacional de los mismos procesos.

Detectar tempranamente consumos con medición defectuosa permitirá cumplir de mejor manera el acuerdo social que la compañía establece con sus usuarios, proveyendo una mejor calidad de servicio en un recurso tan importante como el agua.

En la actualidad las herramientas de análisis de datos son ampliamente usadas, en el estudio de mercado de Big Data Analytics del 2017, parte de la investigación de Wisdom of Crowds. La Figura 5.1 nos muestra que la tecnología de minería de datos y predicción esta dentro de las 7 primeras tecnologías estrategias para incorporar inteligencia de negocios en la empresa.

Figura 5.1: Tecnologías he Iniciativas Estratégicas para Inteligencia de Negocios.

En la siguiente Figura 5.2, se muestra que la adopción del "Business Inteligence"por parte de las empresas ya llego al 53 % (Columbus, L. 2017):

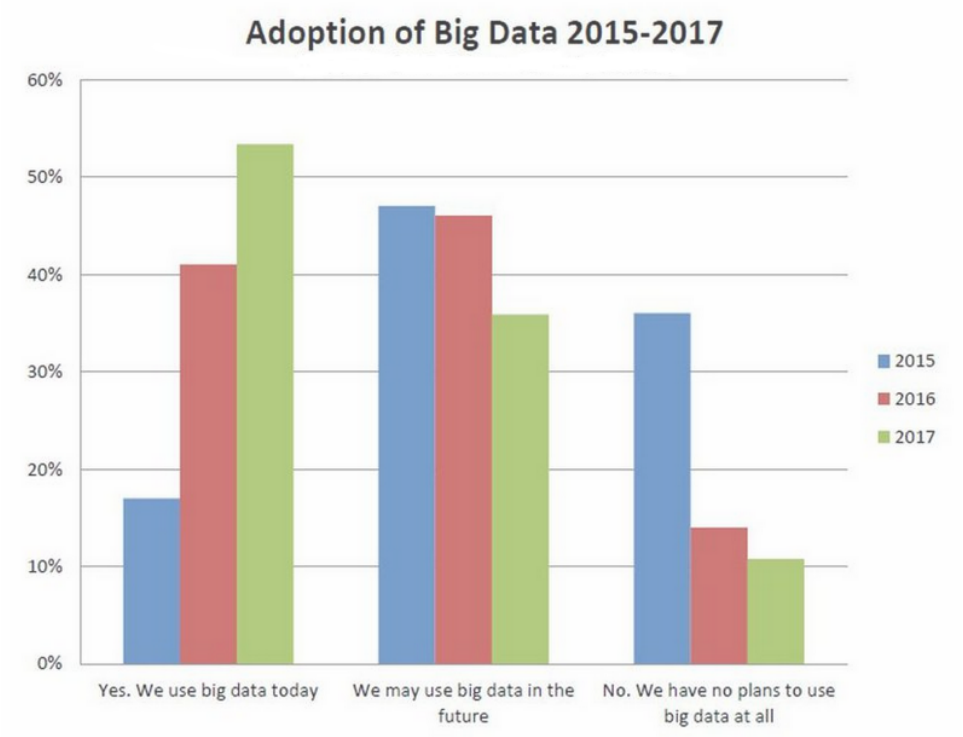


Figura 5.2: Adopción Big Data 2015-2017.

En la Figura 5.2 se infiere un creciente interés en esta tecnología, en el caso de servicios de agua solo existe un caso de investigación en este sentido, un trabajo del 2012 titulado "A Data Mining Based Fraud Detection Model for Water Consumption Billing System in MOG", hecho por Eyad Hashem S.Humaid de la Universidad Islámica de Gaza, que utilizo el algoritmos de aprendizaje supervisado Support Vector Machine.

5.7. Recursos necesarios

El elemento esencial para este trabajo, son los datos a analizar, estos fueron provistos por la empresa de aguas ESVAL S.A. Corresponden a consumos mensuales de 560000 clientes de la quinta región de Chile, desde 2010-2014, estos datos serán usados por los algoritmos de predicción, con el fin de detectar los clientes que actualmente supieran estar siendo objeto de pérdidas aparentes.

Además de los datos ya mencionados será necesario, un computador gamma media, procesador de 4 núcleos de 4GHZ, con 8 GB de memoria RAM, para la creación y prueba del modelo.

5.8. Idea de producto o solución

El modelo podrá ser ocupado por Empresas de servicio de aguas, que deseen una mejora en su proceso de detección de pérdidas y fiscalización a sus clientes, también entrega una metodología para ser probada en otros servicios con un modelo similar, como los servicios de luz.

El modelo puede ser envuelto en servicio de análisis de consumos, que necesite datos históricos y posición geográfica, y provea un dashboard con los posibles casos de pérdidas aparentes, distribuidos en el mapa de con diagrama de calor a través del cual se pueda identificar regiones de alta densidad de pérdida.

La empresa ZeCovery plantea este desafió, con la finalidad de ofrecer este servicio a su cliente Esval S.A. de los cuales provienen los datos históricos de consumo.

En términos del trabajo a desarrollar, se plantean desafíos importantes en términos de limpieza de datos, el establecimiento de relaciones entre las variables, la validación del modelo y la cuantificación de los efectos, entre otros.

Capítulo 6

Solución

6.1. Análisis

6.1.1. Alternativas de solución

Como solución a la reducción de pérdidas aparentes se han propuesto dos opciones que se definen a continuación:

- **Medidor Inteligente**: Nueva generación de medidores conectados a internet que permiten una lectura a distancia en intervalos de 15 minutos (T. Rafael, 2019), pudiendo detectar cambios en tiempo real.
- **Detección de consumos anómalos**: Modelo matemático que usa datos históricos de consumo de agua de los usuarios con el fin de detectar anomalías en éstos consumos, las anomalías pueden referir a un consumo no registrado o a una medición defectuosa.
  Éstos modelos se basan en el supuesto de que un usuario mantiene un historial de consumo sin una gran desviación en el tiempo.

6.1.2. Soluciones comparadas

- **Análisis cuantitativo y cualitativo**
  - **Medidores Inteligente**: Conocer los datos de consumo en tramos horarios posibilita una comunicación mas fluida con el usuario, haciendo posible entregar información en tiempo real y consejos de eficiencia, mejorando así la experiencia del usuario, además permite que la compañía reaccione oportunamente ante cualquier detección de anomalías.
    Su implementación conlleva un esfuerzo de inversión en infraestructura considerable, en personal y en gestión y control de la operación, además del tiempo que lleve cambiar los medidores antiguos.
  - **Detección de consumos anómalos**: Esta solución tiene una implementación sencilla y rápida, ocupa como insumo los datos que ya posee la com-

pañía y no necesita una gran infraestructura para su implementación. Se integra directamente a los procesos de fiscalización ya establecidos por la compañía, y significa una mejora de a lo menos 70 % en el éxito de estas fiscalizaciones (Humaid, E., & Barhoom, T. 2012).

- **Visión técnica**
  - **Medidores Inteligente**: La implementación de esta solución supone la integración de un sistema informático que reciba, almacene, ordene y procese los datos entregados por los medidores inteligentes.
    Estos medidores al estar conectados a internet y al ser objetos electrónicos, crean la necesidad de cubrir el costo y la logística del suministro a través de la red eléctrica y la red de comunicación informática.
    Además en una etapa media de implementación será necesario un modelo matemático que detecte anomalías, ya que estos medidores pueden ser intervenidos informáticamente, por lo que la compañía, tendrá que cubrir costos de ciber-seguridad.
    Todas estas nuevas necesidades implican un área nueva o al menos la remodelación del departamento de informática de la compañía, licencia de software y capacitaciones.
  - **Detección de consumos anómalos**: Esta solución puede ser envuelta en un servicio completamente desacoplado de la compañía, el cual, recibe los datos históricos y entrega el reporte con los usuarios que presenten comportamiento anómalo.
    Por lo que las barreras técnicas para su implementación recaen en el buen uso del servicio por parte del personal técnico de la compañía, pudiendo ser necesaria capacitación.
- **Visión económica**
  - **Medidores Inteligente**: Para la implementación de esta solución es necesaria una inversión considerable en infraestructura, logística y capacitación. Deben ser reemplazados los medidores antiguos esto supone su compra he instalación, además ya no habrá una medición mensual por medidor, sino que prácticamente en tiempo real, lo que genera un gran volumen de datos que deben ser debidamente almacenados y procesados, esto puede requerir software y personal adicional.
  - **Detección de consumos anómalos**: Esta solución es un servicio que puede ser provisto por una tercera compañía ad-doc a las necesidades de la empresa de servicios de agua. No supone una inversión inicial significativa.
- **Impacto ambiental**
  - **Medidores Inteligente**: Esta alternativa afecta al medio ambiente desde varios frentes, primero los recursos naturales necesario para la creación de los medidores inteligentes y desechos que deje este proceso de fabricación, luego al reemplazar estos nuevos medidores por los antiguos se genera "basura tecnológica"que si no es bien reutilizada provoca contaminación.

Se debe tener en cuanta el los costos ambientales de la generación de energía para abastecer estos medidores y también el reemplazo de los viejos medidores por éstos causará contaminación acústica y visual para los vecinos del sector.

- **Detección de consumos anómalos**: Esta solución es de software, por lo que un servicio en la nube podría soportar esta alternativa, además para la creación de la misma puede ser hecha en un computador personal de gamma media, existe un impacto al medio ambiente pero de una magnitud muy inferior al de los medidores inteligentes.

6.1.3. Elección

A continuación, en el Cuadro 6.1, se muestran las conclusiones de la comparación de ambas soluciones:

Perspectiva	Medidores Inteligentes	Detección de consumos anómalos
Análisis cuantitativo y cualitativo	Ambas soluciones tienen puntos fuertes y débiles	
Visión técnica	relativamente más compleja	relativamente menos compleja
Visión económica	significativamente más costosa	significativamente menos costosa
Impacto ambiental	impacto significativamente mayor	impacto significativamente menor

Cuadro 6.1: Comparación soluciones

Se toma en consideración la comparación anteriormente efectuada y se añade el argumento de que el modelo matemático es integrable a la solución de medidores inteligentes.
Este trabajo investigará la solución basada en un modelo matemático, que usa datos históricos de consumo de agua de los usuarios, con el fin de detectar anomalías.

6.1.4. Proceso

Se creará un modelo que permita obtener la mayor precisión al detectar pérdidas aparentes en los consumos de los usuarios de la compañía de servicios básicos de agua.
Se probarán distintos modelos que entregarán resultados basados en sistemas de clasificación, estos serán categorizados y ordenados con el fin de entregar el mejor resultado posible.
Comunicando a la compañía proveedora de servicios básicos una mejor idea del comportamiento del consumo de agua en un determinado sector.

El proceso a seguir esta representado en la siguiente Figura 6.1:

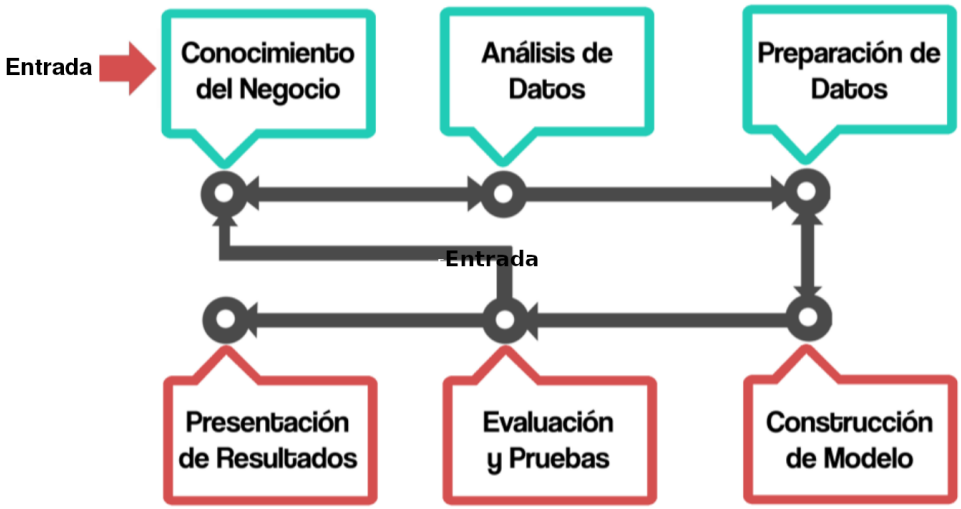


Figura 6.1: Proceso de creación.

6.1.5. Hipótesis

La hipótesis principal es que el usuario de servicios básicos de agua tiene un consumo que no presenta una desviación muy grande entre uno u otro mes de distintos años, hasta que su medidor se avería y/o es modificado, pudiendo ser detectado este cambio.
El modelo matemático es entrenado con los datos históricos de consumo de los usuarios que ya hayan sido detectados anteriormente por las fiscalizaciones, es decir, al método de clasificación se le enseña que debe detectar, para después ser usado con datos de usuarios que no han sido detectados.

6.1.6. Variables

Las variables serán los datos proporcionados por la compañía de servicios básicos, estos serán datos históricos de consumo mensual por 5 años. A los cuales se les agregaran los datos de los usuarios detectados en fiscalizaciones anteriores.
Estos datos pasaran por un proceso de limpieza y transformación de los mismos, para ser ocupados de forma correcta en los algoritmos de clasificación.

6.1.7. Antecedentes

Usuario sin pérdidas aparentes: La Figura 6.2 muestra el volumen de consumo de agua en metros cúbicos de 12 usuarios aleatorios sin pérdidas aparentes por 60 meses (5 años).

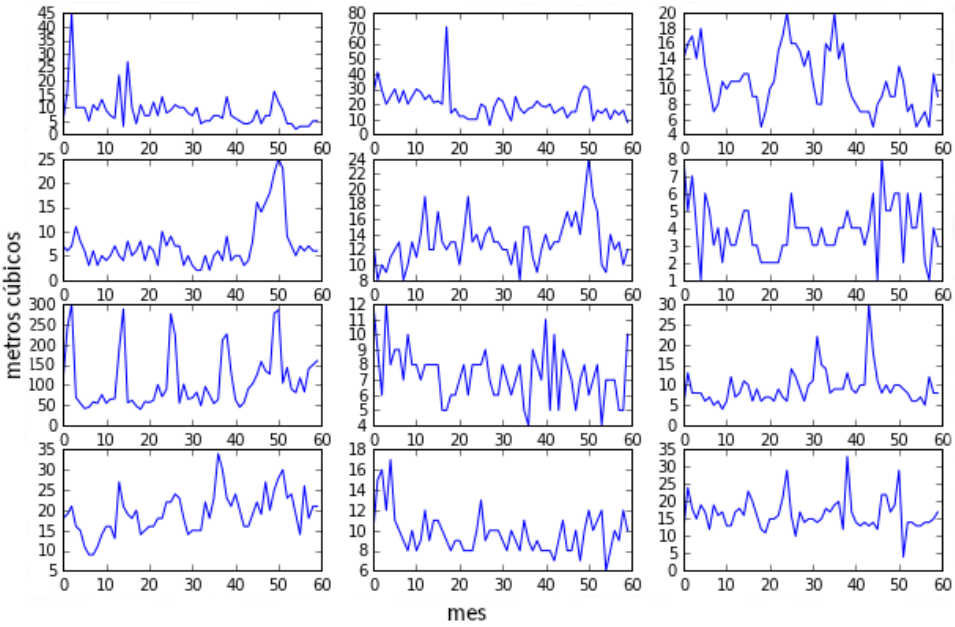


Figura 6.2: Consumo mensual sin pérdidas en m^3, Datos ESVAL S.A.

En la Figura 6.2 se pueden observar patrones comunes en los mismos meses del año, estacionalidad o un comportamiento esperado por mes.

A continuación, en la Figura 6.3 se muestra el volumen de consumo en metros cúbicos mes a mes comparado anualmente de un usuario característico, esto refleja mejor los patrones mencionado anteriormente.

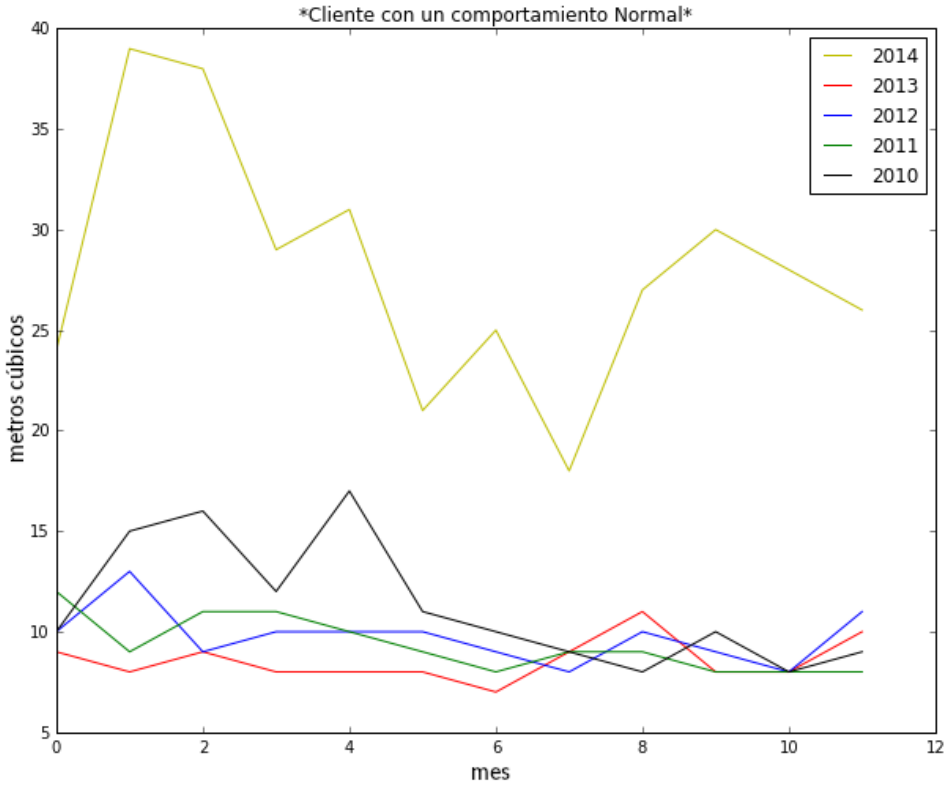


Figura 6.3: Comparación consumo mes a mes por periodos anuales en m^3.

Se aprecia un consumo mayor los primeros meses (correspondiente a los meses de Enero,Febrero y Marzo) y que el resto del año mantiene un comportamiento estacionario.

Usuario con pérdida aparente: La Figura 6.4 muestra el volumen de consumo de agua en metros cúbicos de 9 usuarios aleatorios con pérdidas aparentes por 60 meses (5 años).

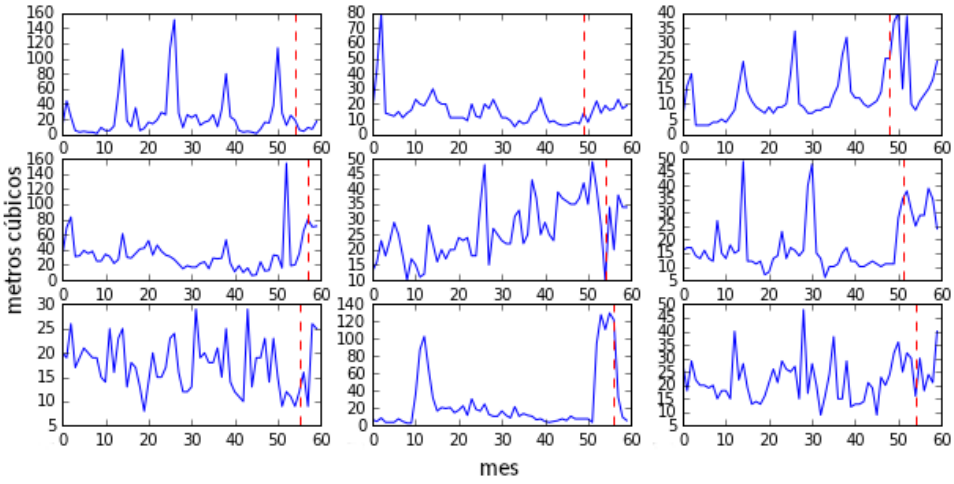


Figura 6.4: Consumo mensual con pérdidas en m^3, Datos ESVAL S.A.

Se observa en las lineas segmentadas que existe un comportamiento inusual en ese mes, que no coincide con el comportamiento de un usuario sin pérdidas aparentes del mismo mes en años anteriores.

A continuación, en la Figura 6.5 se muestra el volumen de consumo en metros cúbicos mes a mes comparado anualmente de un usuario característico, esto refleja mejor lo mencionado anteriormente.

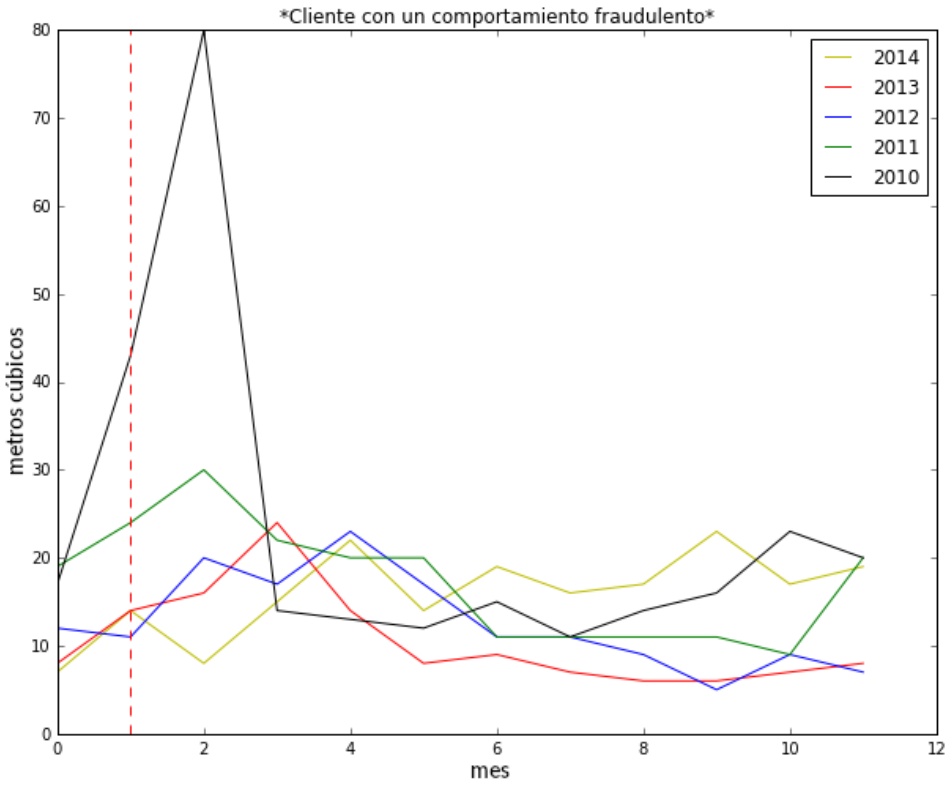


Figura 6.5: Comparación consumo mes a mes por periodos anuales en m^3, Datos ESVAL S.A.

Se observa que los usuarios con pérdidas aparentes presentan consumos menores en los primeros meses respecto a los usuarios normales, teniendo un comportamiento estacionario todo el año.

6.1.8. Viabilidad

Lo mostrado anteriormente mas la disponibilidad de las tecnologías justifican el quehacer de este trabajo, ya que la hipótesis es respaldada por el análisis previo de los datos.

6.1.9. Solución

La solución consiste en la implementación informática de un modelo matemático que detecta usuarios sospechosos de estar en situación de pérdida aparente.

Para lograr esto es necesario tener una fuente de datos ordenados, información que corresponde a consumos mensuales de usuarios de los servicios de agua potable por un periodo de 5 años. Esta información debe ser extraída, transformada para su uso

y cargada en los algoritmos de clasificación.

Los resultados de los algoritmos clasificarán a los usuarios en dos posibles opciones, usuarios con comportamiento anómalo, y usuarios con comportamiento normal.
Se usarán distintos algoritmos los cuales podrían coincidir o no en la clasificación de un mismo usuario.
Se construirá un sistema de categorías de sospechosos dependiendo de la cantidad de algoritmos que clasificaron a cierto usuario con comportamiento anómalo.

Se espera de esta manera que si un usuario es detectado por todos los algoritmos con comportamiento anomalo, tiene mayor probabilidad de serlo.

6.2. Diseño

6.2.1. Negocio

En Chile la cobertura desde el año 1965 se ha incrementado de manera sostenida en el tiempo como se aprecia en el Cuadro 6.2, llegando a un nivel de acceso del 99,9 % de la población (Superintendencia de Servicios Sanitarios, 2014). Chile esta sobre el promedio latinoamericano que es de un 93 % (Soulier Faure, Ducci, & Altamira, 2013), ubicándose a niveles comparables con países como Noruega o Dinamarca (OECD, 2007).

Año	Población Millones Habs.	Cobertura Agua Potable Urbana (%)	Cobertura Alc. Urbano (%)
1965	5.85	53.5	25.4
1966	6.01	56.3	26.0
1967	6.18	59.1	26.8
		...	
2012	15,7	99,9	96,3
2013	16,1	99,9	96,5
2014	16,5	99,9	96,7

Cuadro 6.2: Evolución de la cobertura de agua potable en Chile, Datos SISS

De acuerdo a los datos contenidos en el Informe de Gestión del Sector Sanitario 2014, presentado por la Superintendencia de Servicios Sanitarios (SISS), en Chile las empresas sanitarias produjeron en su conjunto un total de 1.670 millones de metros cúbicos de agua potable, de este volumen de agua un 33,65 % no fue facturada, de ese total la SISS estima que el 74 % se originan por pérdidas físicas, mientras que el 26 % restante corresponde a pérdidas aparentes, esto equivale a cerca de 147 millones de metros cúbicos de agua potable.
Esto se traduce en una perdida de $32.487.000.000 pesos para la industria, tomando como referencia la menor tarifa de agua en m^3 de Aguas Andinas 2018.

6.2.2. Datos

Los datos recabados corresponden a los usuarios de la Empresa de Servicios Sanitarios de Valparaíso (ESVAL). La información proporcionada tiene horizonte temporal histórico de 5 años (desde el año 2010 al 2014).
La primera etapa consistió en separar a todos los usuarios que poseen un consumo mensual durante estos 5 años, excluyendo aquellos usuarios que no registren consumo en al menos un mes.

- **Cantidad de Información**

  Debido a la gran cantidad de información que se dispone, 523.745 usuarios en la región de Valparaíso, es que se trabajara el problema por conjunto de datos más pequeños. Esto con el fin de hacer más rápido el análisis computacional y extender (de ser posible) los resultados al conjunto completo.

- **Tipo de variables**

  Puesto que la información del usuario se describe con variables numéricas (consumos) y categóricas (conducta del usuario), se estableció la definición del trabajo de la siguiente forma: Modelos de predicción basados solo en el historial de consumo; Modelos de predicción basados con todas las variables disponibles (categóricas y numéricas).

- **Análisis cualitativo**

  Dado que existe información de usuarios con fraude detectado (un tipo de pérdida aparente), la que incluye la fecha de detección, es posible contrastar el comportamiento de usuarios con fraude y usuarios con consumo normal, lo que se aprecia en la Figura 6.6:

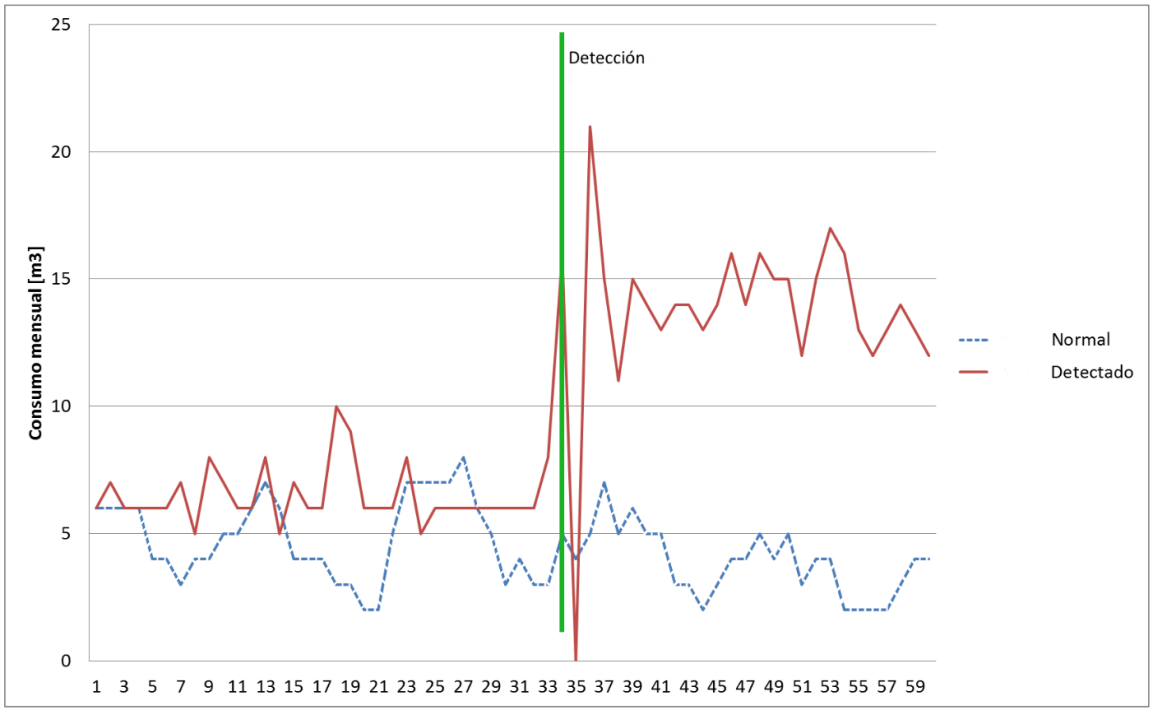


Figura 6.6: Consumo mensual: Usuario Normal vs Detectado, Datos ESVAL S.A.

Es posible apreciar dos fenómenos fundamentales:

- Una vez detectado el fraude, el consumo sube ostensiblemente.
- Consumos autorizados: El usuario con comportamiento normal presenta cierto grado de estacionalidad (aumento de consumo durante verano en relación a periodo de invierno).

En la siguiente Figura 6.7, se realizó análisis estadístico mediante plot-box, comparando la distribución en el tiempo de los usuarios con fraude detectado y los usuarios sin procesar.

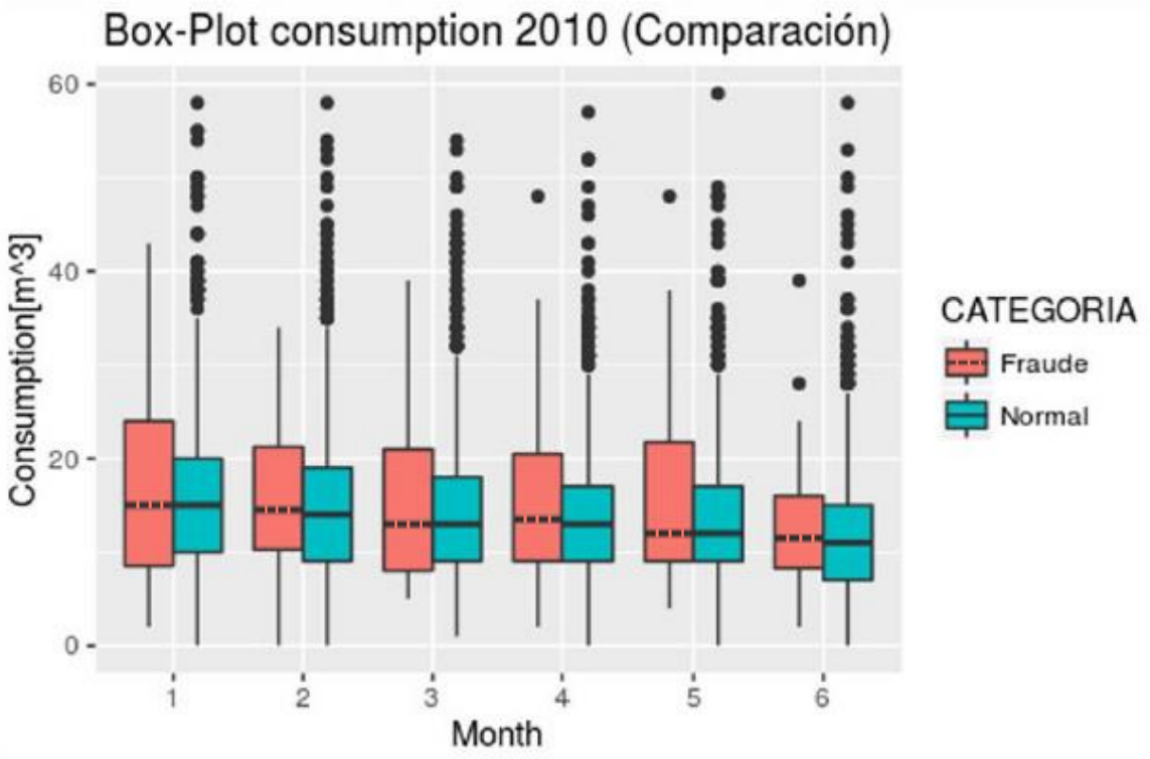


Figura 6.7: Distribución del consumo mensual: Usuario Normal vs Detectado, Datos ESVAL S.A.

El análisis mostró que entre los usuarios con consumo normal mensual existe una cantidad importante de outlayers, esto no se presenta entre los usuarios con pérdidas aparentes tienden a presentar consumos más planos.

En base a este análisis se concluye posible integrar algoritmos de clasificación para la detección de pérdidas aparentes.

6.2.3. Resultados Quilpue

Por capacidad de análisis y simplicidad se trabajó con la comuna de Quilpué, debido a que posee 51.892 usuarios equivalentes a aproximadamente el 10 % del total de usuarios de la región, lo que permite realizar los análisis de manera representativa.

A nivel de cantidad de datos la comuna de Quilpué posee 51.892 usuarios registrados, entre los cuales 865 usuarios presentan fraude detectado en el periodo de estudio (enero 2010 a diciembre 2014). Sin embargo, y dado que se estableció como restricción que los usuarios deben poseer 60 meses de consumo, el set de datos se redujo a un total de 22080 usuarios con un total de 330 usuarios con fraude detectado en el periodo.

- **Resultados clasificación**

  En base a lo anterior, y descartada la utilización de redes neuronales por sobre ajuste (ver Anexo pagina 53), se realizó la implementación de tres algoritmos, utilizando para su evaluación validación cruzada, generando los modelos con el 70 % de los datos y verificando su exactitud con el 30 % restante.

  En la siguiente Figura 6.8, se muestran los resultados obtenidos por los tres algoritmos: regresión logística(RL), support vector machine(SVM) y random forest (RF):

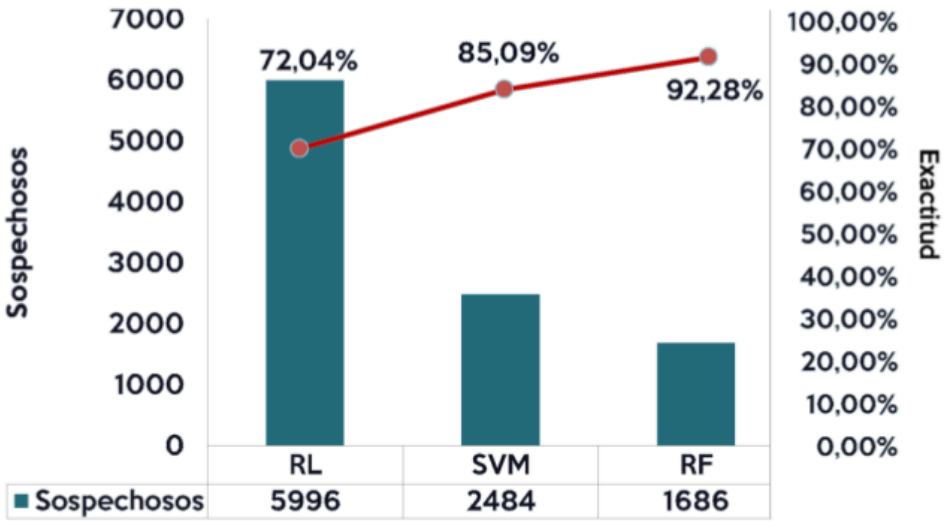


Figura 6.8: Resultados algoritmos usados.

- **Categorización usuarios**

  En base a lo obtenido por los algoritmos de clasificación, se clasifican los usuarios dependiendo del numero de algoritmos que etiquetan al usuario como sospechoso:

  - **Clase A**: aquellos que fueron detectados por solo un algoritmo, se detectan 4587 casos.
  - **Clase B**: aquellos que fueron detectados por 2 algoritmos, se detectan 2047 casos.
  - **Clase C**: aquellos que fueron detectados por los 3 algoritmos utilizados, se detectan 495 casos.

- **Análisis de categorías**

En la siguiente Figura 6.9, se realizo un análisis de dispersión de los consumos las distintas categorías mediante plot-box:

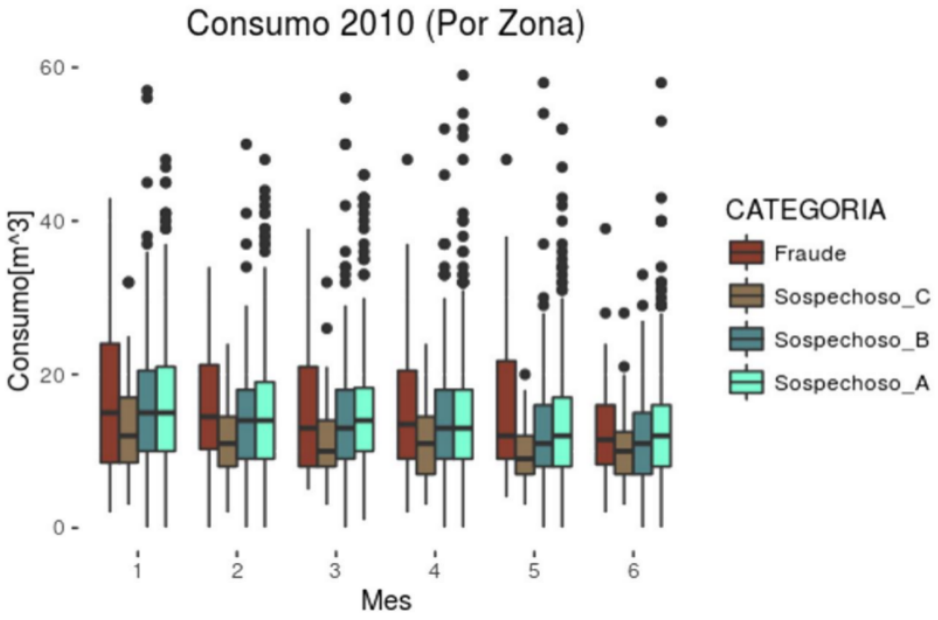
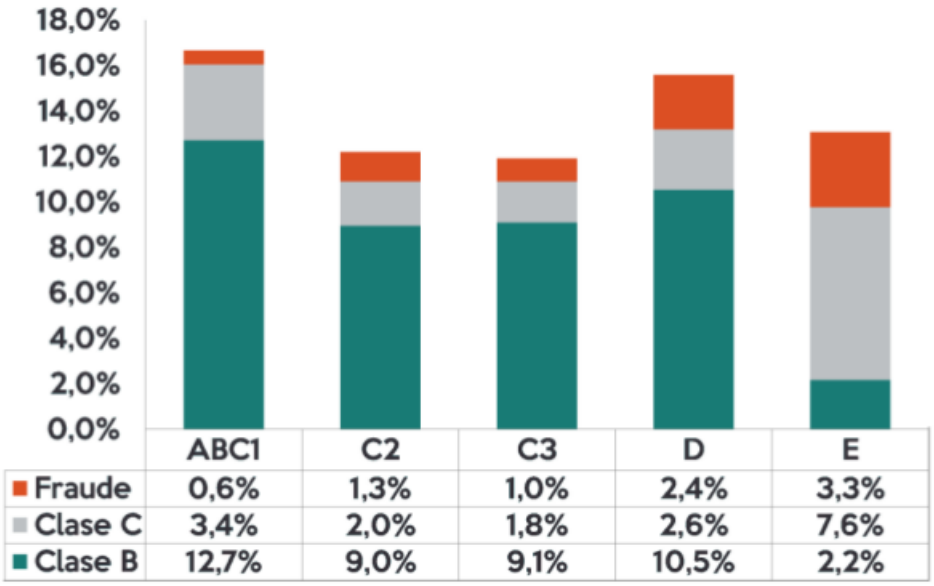


Figura 6.9: Dispersión de consumos mensuales por categoría.

Se aprecia que a medida que se avanza en la categoría, es decir, mas algoritmos detectaron a un sospechoso, la cantidad de datos outlayers disminuye, pareciéndose mas a los usuarios que si se les ha detectado pérdidas aparentes, siendo concordante con lo observado anteriormente.

- **Análisis socio-económico**

  En la siguiente Figura 6.10, se muestra un gráfico de el porcentaje de usuarios detectados, separados por su situación socio-económica.



	ABC1	C2	C3	D	E
Fraude	0,6%	1,3%	1,0%	2,4%	3,3%
Clase C	3,4%	2,0%	1,8%	2,6%	7,6%
Clase B	12,7%	9,0%	9,1%	10,5%	2,2%

Figura 6.10: Porcentaje de usuarios detectados, según situación socio-económico

Se observa que el nivel de sospechosos más la presencia de pérdidas aparentes es equiparable en todos los estratos, lo que permite concluir que el fenómeno de la pérdida aparente en el consumo de agua potable es transversal a la situación socio-económica del usuario.

- **Resumen de resultados**

  - De un total de 22080 usuarios. 495 fueron categorizados como sospechosos por los tres algoritmos, 2047 por dos algoritmos y 4587 por un algoritmo. El peor porcentaje de precisión entregado por un algoritmo es de 72 %.
  - La clasificación de usuario con comportamiento anómalo es transversal a la situación económica del usuario.
  - Hasta este momento la se realizan fiscalizaciones aleatorias que tienen una precisión de 10 %, la peor precisión entregada por los algoritmos es de un 72 %.
    Esto supone un ahorro para las compañías de agua de aproximadamente $20.141.940.000 pesos por concepto de detección temprana de usuarios con perdidas aparentes.

Capítulo 7

Conclusiones

7.1. Análisis satisfactorio

Los algoritmos usados muestran resultados alentadores, pudiendo identificar hasta con un 92,28 % de exactitud clientes que están en situación de pérdidas aparentes.

Además la metodología para categorizar estos tipos de usuarios según la cantidad de algoritmos que los detecto, muestra ser una estrategia acertada, en el gráfico de dispersión del consumo se muestra claramente que que entre mas algoritmos clasifiquen a un usuario como con posible pérdida aparente existen menos outlayers pareciéndose mas al conjunto que ya se sabe que tiene pérdida aparente.

Comparado con el porcentaje de hoy, 10 %, este trabajo demuestra teóricamente un aumento de por lo menos 70 % en el porcentaje usuarios que al ser fiscalizados, efectivamente tengan pérdidas aparentes.

7.2. Validación incompleta

Se reconoce que este trabajo tiene una validación incompleta, en el sentido de que no existe una verificación en terreno del prototipo, que permita comparar la precisión de los resultados obtenidos con fiscalizaciones en terreno.

7.3. Eficiencia en fiscalización

La compañía de servicios de agua al conocer exactamente los sospechosos de estar teniendo pérdidas aparente, podrán planea de mejor forma sus procesos de fiscalización, abriendo grandes oportunidades de mejora.

7.4. Eliminación de prejuicio

Se observa que si bien el nivel de fraudes detectados es bastante mayor en los niveles más bajos, sin embargo el nivel de sospechosos más la presencia de fraude es equiparable en todos los estratos, lo que permite concluir que el fenómeno de pérdidas aparentes en el consumo de agua potable es transversal. Por lo que no existe una relación entre situación socio-económica y ser clasificado como sospechoso de tener una pérdida aparente.

7.5. Trabajo futuro

7.5.1. Validación en terreno

El siguiente paso antes de la construcción de un producto o servicio, es verificar en terreno a través de fiscalizaciones a los sospechosos obtenidos de los resultados de este trabajo, pudiendo comparar resultados.

7.5.2. Retroalimentación

Para ir mejorando cada vez mas este modelo, es posible retroalimentarlo con nuevos datos de usuarios detectados, así este modelo puede ir evolucionando en el tiempo, mejorándose a si mismo.

7.5.3. Análisis de reincidencia

Al largo plazo cuando ya se hayan detectado mas usuarios que han incurrido en pérdidas aparentes mas de una vez, se podrían aislar y analizar su comportamiento, siendo posible detectar, posibles reincidentes en un futuro.

7.5.4. Aplicación en otros servicios básicos

Los servicios de que proveen luz eléctrica comparten varias características con los de agua potable, el usuario tiene un medidor que contabiliza el consumo del servicios, a través de esto se genera la factura. Por lo que se intuye que este trabajo es aplicable a los servicios de luz.

Capítulo 8

Anexos

8.1. Conocimiento del negocio

8.1.1. Tipo de empresa

En Chile existen dos tipos de empresas sanitarias:

Las situadas en áreas rurales no requieren concesiones por parte de la Superintendencia de Servicios Sanitarios(SISS), se organizan en cooperativas de acuerdo a los mandatos de sus socios-beneficiarios. Estos organismos no se encuentran sujetos a regulación tarifaría, pero requieren obtener derechos de aguas consuntivos. Su cobertura alcanza a aproximadamente un 11 % de la población nacional.

Las situadas en áreas urbanas e inmediaciones requieren concesiones por parte de la SISS, se encuentran sujetas a regulación tarifaría y dada su naturaleza corresponden a monopolios naturales, los que otorgan servicio a aproximadamente el 89 % de la población del país.

8.1.2. Tarificación

Para la regulación sanitaria debe tenerse en cuenta que a causa de la importante infraestructura que se requiere para entregar los servicios, los prestadores se constituyen como un monopolio natural donde los costos marginales (es decir, los de producir el agua potable y de tratar las aguas servidas) son inferiores a los costos medios (que incluyen los costos asociados al financiamiento de la infraestructura).
Esto lleva a que la tarificación logre que las empresas cubran la totalidad de estos costos para que existan incentivos para entrar en este mercado.
En la práctica ello presenta inconvenientes debido a la dificultad para determinar los costos socialmente óptimos. Es decir, los necesarios para proveer un servicio de acuerdo a los parámetros preestablecidos de calidad y cobertura, pero sin incluir

sobre-inversiones.

La SISS es el organismo encargado de llevar a cabo el proceso regulatorio del sector, recolectando para ello información de las empresas reguladas. Establece metas de calidad, fiscaliza cumplimientos y sanciona faltas.
En términos de establecimiento de tarifas tanto la SISS como la empresa regulada presentan un informe preliminar. Si no existen diferencias, se fijan los precios de acuerdo al estudio de la SISS.
Por el contrario, de encontrar la empresa que existen diferencias que la afectan, el estudio puede ser discutido mediante recurso ante una "Comisión de Expertos", formada por tres integrantes que debe pronunciarse en favor de cada una de las variables que determinan la tarifa entregada en uno u otro estudio sin posibilidades intermedias.

El conjunto de tarifas resultantes debe ser promulgado por el Ministerio de Economía que puede introducir comentarios.
Si incluso después de su dictamen la empresa no estuviera satisfecha, puede recurrir a los tribunales para zanjar temas de forma, pues el fondo es determinado en la etapa anterior.

8.1.3. Marco Legal

La provisión de servicios de agua potable y alcantarillado requiere de un elevado nivel de infraestructura, representada en redes de agua potable y alcantarillado, más las instalaciones destinadas a la captación y potabilización del agua, y aquellas para el tratamiento de las aguas servidas.

Estas instalaciones se constituyen en la práctica como costos hundidos en un porcentaje muy alto (es decir, tienen un bajo valor alternativo en otra actividad o en otras ubicaciones), con lo que la empresa que ya ha invertido en estos activos para prestar los servicios sanitarios tiene una muy elevada ventaja frente a potenciales competidores: en teoría, podría cobrar a sus clientes sólo los costos variables de producción, con lo que dejaría fuera a los competidores (suponiendo que es factible físicamente duplicar las redes).

Asimismo, el sector sanitario presenta importantes economías de escala, lo que implica que las empresas más grandes (que sirven a un mayor número de clientes) tienden a presentar menores costos medios por unidad de venta que las más pequeñas.
De lo anterior se deriva que la provisión de este tipo de servicios corresponda a un monopolio natural, resultando eficiente que sólo una empresa realice el desarrollo de estas actividades.

Con el fin de resguardar que la entrega de estos bienes y servicios se efectúe bajo estándares sanitarios predefinidos y adecuados, y para evitar que la empresa ejerza sobre sus clientes el poder que detenta en su área de operaciones, el Estado regula tanto la calidad del servicio como las inversiones mínimas a realizar y las tarifas que serán cobradas al público.

El Decreto con Fuerza de Ley No 382 de 1988 ("Ley General de Servicios Sanitarios") del MOP, conjuntamente con el Reglamento de la Ley (Decreto Supremo N°121 de 1992), otorgan en Chile el marco jurídico a la provisión de servicios por parte de empresas y fue modificado por última vez el 18 de diciembre de 2007 mediante la Ley 20.307.

Las tarifas del rubro, en tanto, tienen su marco normativo en el Decreto con Fuerza de Ley N°70 del 30 de diciembre de 1988 del MOP, complementado por el Decreto Supremo N°453 de 17 de enero de 1990 del Ministerio de Economía y que se constituye en el reglamento del DFL N°70.

8.1.4. Nuevas concesiones

Para extender su cobertura hacia nuevas áreas geográficas una empresa sanitaria debe solicitar otras concesiones a la SISS, por lo que, como ocurre en la práctica, puede operar varias concesiones, para cada una de las cuales se deberá establecer un menú de tarifas distintas (cargo fijo, agua potable, recolección y tratamiento de aguas servidas).

La solicitud puede ser comentada por el Ministerio de Vivienda y Urbanismo y por la municipalidad en que se ubique la concesión, debiendo ser publicada en un diario local, de modo que otras empresas también puedan participar compitiendo en el proceso.
Con posterioridad, el solicitante debe realizar un estudio de desarrollo para la concesión solicitada y también un plan de tarificación, siendo otorgada por la SISS a la empresa que cumpla con los requisitos establecidos por ella y que además cobre las menores tarifas, debiendo finalmente ser aprobada por el MOP.

8.1.5. Pérdidas

En términos de los consumos la Asociación Internacional de Agua (IWA) en conjunto con la Asociación americana de trabajos en Agua (AWWA) define los siguientes tipos de consumos (American Water Works Association):

Suministro de Agua Potable	Consumos Autorizados	Consumos Autorizados Facturados	Consumos Medidos Facturados a Clientes Registrados
			Consumos No Medidos Facturados a Clientes Registrados
		Consumos Autorizados No Facturados	Medidos
			No Medidos
	Pérdidas de Agua	**Pérdidas Aparentes**	**Consumos No Autorizados**
			Consumos Con Medición Defectuosa
		Pérdidas Físicas	Fugas en Redes
			Fugas y Rebalses en Tanques de Almacenamiento
			Fugas en puntos de Servicios

Cuadro 8.1: Componentes y Definiciones del Balance de Agua, IWA/AWWA

- **Suministro de Agua Potable**: Corresponde al volumen anual de agua potable inyectada al sistema.
- **Consumos autorizados**: Corresponde al volumen anual de agua medida y/o no medida entregada a clientes autorizados.
- **Pérdidas de Agua**: Diferencia entre el agua inyectada al sistema y los consumos autorizados.
- **Pérdidas aparentes**: Consumos no autorizados, todo tipo de imprecisiones de medición, y errores sistemáticos de manejo de datos.
- **Pérdidas Físicas**: El volumen anual de pérdidas mediante todo tipo de filtraciones, fugas, roturas o rebalse en redes, almacenamiento o puntos de servicio, hasta el punto de medición del cliente.

8.2. Lenguaje de Programación

Se uso el lenguaje de Programación Python version 3.5.0. Python es un lenguaje de programación interpretado multiparadigma, es decir; soporta hacer programación orienta-

da a objetos y programación imperativa.

La potencia y calidad de librerías es muy buena. Principalmente para hacer uso de algoritmos de Machine Learning las librerías en Python son mejores que las de R project, sobre todo las técnicas de Deep Learning.

8.3. Entorno de Desarrollo

IPython proporciona una arquitectura rica para la computación interactiva con:

- Un potente shell interactivo.
- Un núcleo para Jupyter.
- Soporte para visualización interactiva de datos y uso de herramientas de GUI.
- Intérpretes flexibles e integrables para cargar en sus propios proyectos.
- Fácil de usar, herramientas de alto rendimiento para computación paralela.

Notebook Jupyter es una aplicación web de código abierto que le permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo. Los usos incluyen: limpieza y transformación de datos, simulación numérica, modelado estadístico, visualización de datos y machine learning.

8.4. Bibliotecas

- **Pandas:** Código abierto con licencia BSD que proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento y fácil de usar.
  En este trabajo fue usada para la carga de los datos y ordenamiento de estos en una matriz con sus etiquetas respectivas a cada mes y la categoría del usuario, haciendo una separación en variables X para los consumos mensuales he Y para la categoría del usuario.

- **NumPy:** Código abierto con licencia BSD que contiene un poderoso objeto de matriz N-dimensional, sofisticadas funciones de transmisión, Herramientas para la integración de código C / C ++ y Fortran, Álgebra lineal útil, transformada de Fourier y capacidades de números aleatorios.
  En este trabajo fue usada para la transformación de los datos, concatenación, operación de matrices y normales con estos.

- **Matplotlib:** biblioteca de trazado 2D de Python que produce cifras de calidad de publicación en una variedad de formatos de papel y entornos interactivos en todas las plataformas.

puede generar gráficos, histogramas, espectros de potencia, gráficos de barras, gráficos de errores, diagramas de dispersión, etc.
En este trabajo fue usada para visualizar el consumo de los usuarios y poder obtener un análisis cualitativo de sus respectivos comportamientos, según categorías; usuario normal y con pérdidas aparentes

- **Scikit-Learn:** Biblioteca para aprendizaje de máquina de software libre. Incluye varios algoritmos de clasificación, regresión y análisis de grupos entre los cuales están Support vector machine, Random forest, Gradient boosting, K-means y DBSCAN. Está diseñada para interoperar con las bibliotecas numéricas y científicas NumPy y SciPy.
  En este trabajo se utilizo para la normalización de los datos, en la regresión logística, support vector machine, random forest y redes neuronales.

8.5. Algoritmos

En términos genéricos un algoritmo de clasificación es un conjunto de acciones destinadas a separar un conjunto de datos en subconjuntos basados en criterios definidos por el usuario. En particular para el presente análisis se intenta definir dos clasificaciones (Usuario Normal y Usuario con Pérdidas Aparentes).
Al comparar usuarios normales con usuarios a los cuales ya se les ha detectado pérdidas aparentes de anteriormente.

A continuación en el Cuadro 8.2, se muestra la matriz de confusión que entregará las medidas de eficiencia de cada algoritmo:

Usuario Normal	**Pérdida Aparente No Detectada**
Usuario Sospechoso de Pérdidas Aparentes	**Perdida Aparente Detectada**

Cuadro 8.2: Matriz de confusión

Para propósitos de este proyecto la sección de interés primario corresponde a los clientes calificados como sospechosos, estos presentarían pérdidas aparentes.

A continuación se describen brevemente los algoritmos utilizados durante la construcción del prototipo.

8.5.1. Regresión logística

Corresponde a un modelo de regresión para variables de tipo dependientes que busca determinar la probabilidad de un evento ocurriendo como función de los otros factores.

Se usa para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras.

La variable categórica corresponde al la categoría del usuario, si este tiene un consumo normal o presenta pérdidas aparentes en función de su consumo mensual por 5 años.

En particular en este caso se utiliza una regresión logística lineal.

A continuación en la Figura 8.1, se muestra un ejemplo gráfico de regresión logística lineal.

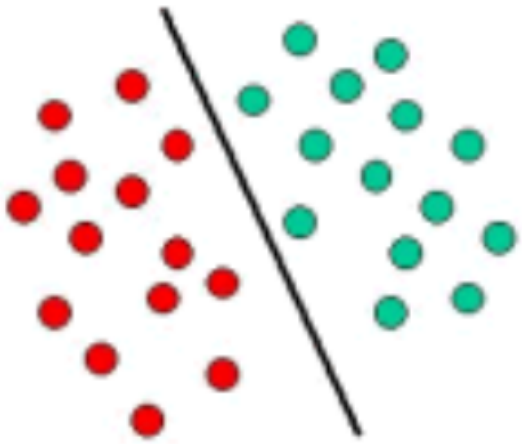


Figura 8.1: Ejemplo regresión logística lineal.

8.5.2. Support vector machine

Algoritmo basado en la noción de planos de decisión, que definen los límites de la decisión, posee cierta similitud al caso de la regresión logística aunque con una mayor complejidad, en este caso la separación no se hace mediante planos, sino mediante hiperplanos y vectores de soporte, a través de los cuales se transforma el conjunto de datos de modo de trasladarlos a un espacio donde puedan interpretarse los planos de decisión de manera más simple.

A continuación en la Figura 8.2, se muestra un ejemplo gráfico de support vector machine.

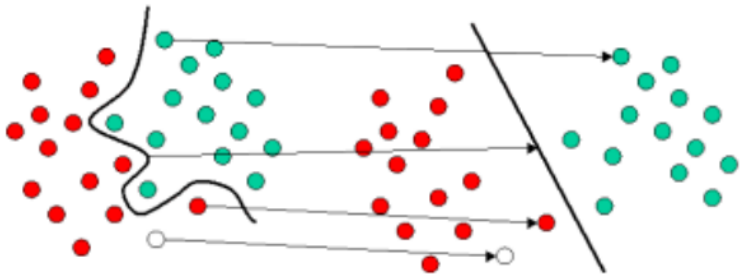


Figura 8.2: Ejemplo support vector machine.

8.5.3. Random forest

Es un algoritmo basado en la implementación y análisis de árboles de decisión, utilizando individuos al azar para generar diferentes set de datos, creando árboles de decisión con cada uno de los set de datos, de esta manera se utiliza el criterio de "voto mayoritario" para definir la clasificación de cada dato.

A continuación en la Figura 8.3, se muestra un ejemplo gráfico de random forest.

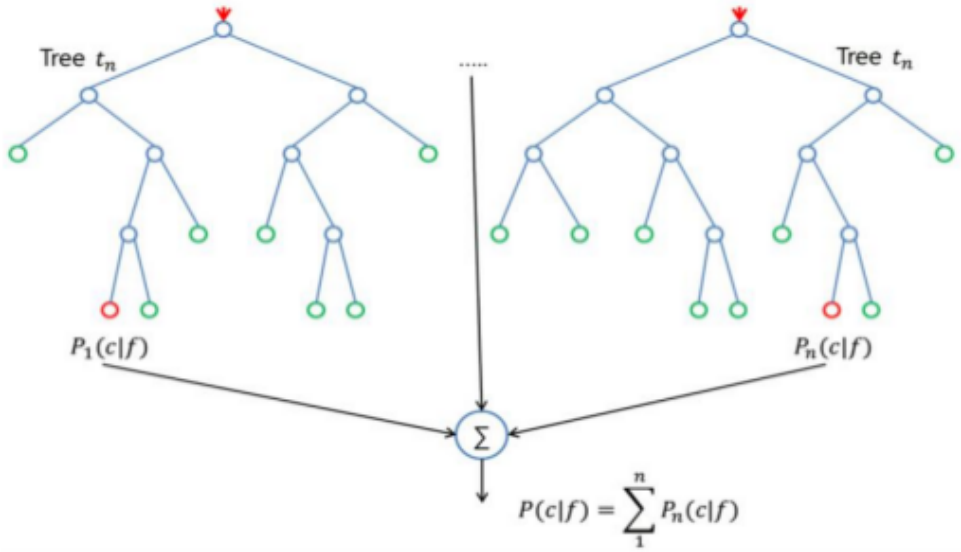


Figura 8.3: Ejemplo random forest.

8.5.4. Redes neuronales

Corresponde a un algoritmo que posee la capacidad de capturar y representar complejas relaciones entre los datos de entrada y salida, y posee dos características principales.

- El Algoritmo es capaz de adquirir conocimiento mediante aprendizaje al ser ejecutado.
- El conocimiento de la red es almacenado en las conexiones entre cada parte, en lo que se conoce como peso sináptico.

A continuación en la Figura 8.4, se muestra un ejemplo gráfico de redes neuronales.

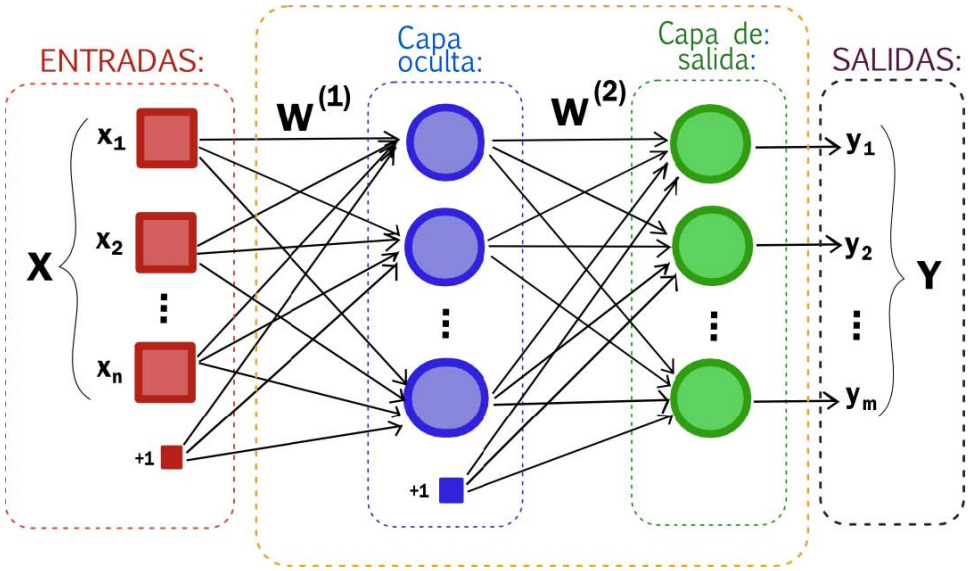


Figura 8.4: Ejemplo redes neuronales.

Este algoritmo fue utilizado en una primera parte, sin embargo fue desechado por dos razones fundamentales:

- Se aplicó en un set de datos pequeños y entregó resultados más discretos que el resto procesamiento mayor.
- Para su entrenamiento requiere de set de datos definidos, es decir que cada elemento está absolutamente etiquetado, en este caso perdida aparente o no perdida aparente, sin embargo, al no existir esta certeza el algoritmo tiene la tendencia a sobre ajustar, esto implica que no genera clientes sospechosos o los genera en muy baja cantidad.

Referencias Bibliográficas

American Water Works Association. (n.d.). IWA/AWWA Water Audit Method.

Asllani, A., & Naco, M. (2014). Using Benford's Law for Fraud Detection in Accounting Practices. Journal of Social Science Studies.

Humaid, E., & Barhoom, T. (2012). A Data Mining Based Fraud Detection Model for Water Consumption Billing System in MOG. Islamic University of Gaza.

McCullough, J. (2010). Deterrent and detection of smart grid meter tamering and theft of electricity, water or gas.

Mutikanga, H., Sharma, S., & Vairavamoorthy, K. (2010). Assessment of apparent losses in urban water systems. Water and Environment Journal, 327 - 335.

OECD. (2007). Financing water supply and sanitization in EECCA countries ans progress in achieving the water related millenium development goals.

Soulier Faure, M., Ducci, J., & Altamira, M. (2013). Agua Potable, Saneamiento y los Objetivos del Milenio en América Latina y el Caribe. Banco Interamericano de Desarrollo.

Superintendencia de Servicios Sanitarios. (2014). Informe de Gestión del Sector Sanitario.

Columbus, L. (2017). 53 % Of Companies Are Adopting Big Data Analytics. (`http://tiny.cc/as347y`)

Torres, R. (2019). Medidores Inteligentes, Distorsiones en las lecturas de consumo. (`http://tiny.cc/2qnp8y`)

Aguas Andinas. (2018). Tarifas Vigentes. (`http://tiny.cc/1mup8y`)