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

MODELO DE DETECCIÓN DE FRAUDE EN CLIENTES DEL SERVICIO DE AGUA POTABLE DE UNA EMPRESA SANITARIA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

VICTORIA PATIÑO ESPINOZA

PROFESOR GUÍA:
RODOLFO URRUTIA URIBE

MIEMBROS DE LA COMISIÓN:
RICARDO SAN MARTÍN ZURITA
XIMENA SCHULTZ SORIANO

SANTIAGO DE CHILE
2014

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: VICTORIA PATIÑO ESPINOZA
FECHA: 06/01/2014
PROF. GUÍA: SR. RODOLFO URRUTIA URIBE

MODELO DE DETECCIÓN DE FRAUDE EN CLIENTES DEL SERVICIO DE AGUA POTABLE DE UNA EMPRESA SANITARIA

Aguas Andinas corresponde a la empresa sanitaria más grande del país, con alrededor de 1,8MM de clientes en la Región Metropolitana y una facturación anual de 509MM de m3 de agua potable, equivalente aproximadamente a $341MMM.

Todas las empresas sanitarias presentan cierto porcentaje de agua no facturada, vale decir, aquella agua potable que se produce, pero no logra cobrarse al consumidor final. Este índice es de alrededor del 30% y es explicado por 3 factores: pérdida técnica, pérdida por micromedición y pérdida por uso irregular. El presente trabajo de memoria pretende encargarse del tercer factor, relativo al agua no facturada por intervenciones que realizan las personas en sus medidores o red de agua potable, con el objeto de disminuir la lectura de m3. Este punto es el causante de una pérdida mensual estimada de 700.000 m3, lo que en dinero se traduce en $468MM.

Hasta ahora, la empresa ha utilizado ciertos criterios para detectar usuarios irregulares, como verificar cuáles de los clientes previamente visitados por personal técnico ha presentado una baja en su consumo (pues se sabe que algunos trabajadores gasfiteros ofrecen intervenir el medidor) o comprobar denuncias hechas por los clientes. Mas, se desaprovecha un sinfín de información con la que se cuenta y que podría aportar a una mayor detección de ilícitos, mejorándose la actual tasa de detección del 15%. Se tiene por cada 1% de mejora en dicha tasa aumenta la recaudación en $6MM.

El eje del trabajo lo constituye la metodología KDD, tendiente a extraer patrones útiles y coherentes de la información que se posee. Es así como se buscó obtener un modelo que señalara la probabilidad que tiene cada cliente de ser un irregular, con la mayor certeza posible, para lo cual se trabajó con datos de clientes irregulares y no irregulares, construyendo una base a la que se le aplicó 3 modelos de aprendizaje supervisado: regresión logística binaria, árbol de decisión CHAID y red neuronal, definiéndose el mejor modelo en base a los costos asociados a los errores tipo I y II.

El mejor método resultó ser el árbol de decisión, con la eficacia más alta igual a 75%, lo que conllevaba al costo por error más bajo. A la vez su eficiencia fue de 81,2%, por lo que en un escenario conservador, si se considera que la tasa de detección aumente de un 15% actual a un 40% con el nuevo método, se está ante una recuperación extra mensual de $150MM.

Finalmente, junto con la aplicación del modelo resultante de este trabajo se sugiere diseñar una estrategia orientada a mejorar la actual tasa de regularización del 55%, considerada baja, lo que aumentaría aún más la recuperación monetaria. Además se debe tener presente que este tipo de modelos posee un determinado ciclo de vida, vale decir, después de cierto tiempo de aplicación es altamente probable que la eficacia y eficiencia comiencen a decaer debido a cambios que vaya experimentando el universo bajo estudio. Por ello se recomienda retroalimentar cada cierto tiempo el modelo, utilizando toda la data que se vaya recolectando, de modo de mantener su calidad.

AGRADECIMIENTOS

Al final de esta larga etapa que comienza a finalizar sólo resta recordar a quienes significaron un impulso positivo tanto en la vida estudiantil como en la personal.

En primer lugar una mención especial para los miembros de la Unidad de Gestión de la Medición de Aguas Andinas, pues sin su apoyo y confianza incondicional desde el primer minuto, el trabajo no hubiera sido posible.

Un gran agradecimiento también para los profesores Rodolfo Urrutia y Ricardo San Martín, que siempre tuvieron expectativas en el tema desarrollado, y cuyos valiosos aportes y críticas ayudaron a construir los cimientos del presente trabajo.

No puedo dejar de dedicarles un párrafo especial a todos mis amigos y amigas ganados en la universidad, simplemente gracias por poder contar siempre con ustedes en los buenos y malos momentos, alegrándome las jornadas, soportándome en los trabajos grupales y siempre ayudándome a tener una visión más optimista del futuro.

Por último agradecer a mis otros grandes amigos ganados en el transcurso de la vida así como a mis seres más queridos, quienes me apoyaron en todo momento y muchas veces significaron mi cable a tierra al momento de tomar decisiones cruciales.

TABLA DE CONTENIDO

1. INTRODUCCIÓN

1.1. ANTECEDENTES GENERALES

1.1.1. Las Empresas sanitarias en Chile

Las empresas del sector sanitario corresponden a aquéllas que proporcionan al menos uno de los siguientes 5 servicios:[1]

- Servicio de agua potable:

i. Producción de agua potable: captación del agua bruta desde fuentes superficiales (embalses y ríos) y fuentes subterráneas (manantiales y pozos) para su posterior potabilización.
ii. Distribución de agua potable: suministro del agua potable a los consumidores a través de la red de abastecimiento.

- Servicio de alcantarillado:

iii. Recolección de aguas servidas: captación del agua utilizada por hogares, comercios e industrias a través de la red de alcantarillado.
iv. Tratamiento de aguas servidas: procesamiento y reciclaje del agua servida (no todas las sanitarias desarrollan esta parte del proceso).
v. Disposición de aguas servidas: reintegración del agua al medio natural (ríos, arroyos y mar) con el menor impacto posible para el entorno.

El sector sanitario se caracteriza por constituir un típico ejemplo de *monopolio natural*, vale decir, debido a los altos montos de inversión en infraestructura que se requieren para incurrir en la industria (caracterizada por su vida útil relativamente larga, sin uso alternativo y presencia de importantes economías de escala) es que se hace insostenible que entre una segunda firma a competir en el mercado. Debido a esta razón, sumada a la creciente participación de agentes privados en un sector antiguamente de carácter estatal y a las asimetrías de información presentes, es que se hace necesaria una estricta regularización y fiscalización del sector. Este rol en Chile es ejercido por la Superintendencia de Servicios Sanitarios (SISS), cuyas funciones se resumen en:

- Fijación de tarifas.
- Otorgamiento de zonas de concesión a las sanitarias para ejercer alguno(s) de los 5 servicios mencionados anteriormente.
- Fiscalización de servicio entregado por las sanitarias.
- Control de residuos industriales líquidos (riles).

En términos cuantitativos, a Diciembre de 2012 [1], se contabilizan 57 sanitarias a lo largo del país, logrando las siguientes cifras:

[1] Cabe destacar que la mayoría de las empresas chilenas hoy en día ejercen los 5 servicios mencionados, salvo contadas excepciones que, por ejemplo, poseen la concesión sólo para la producción de agua potable (Lago Peñuelas S.A. en V Región) o sólo la disposición y tratamiento de aguas servidas (Tratacal S.A. en Calama)

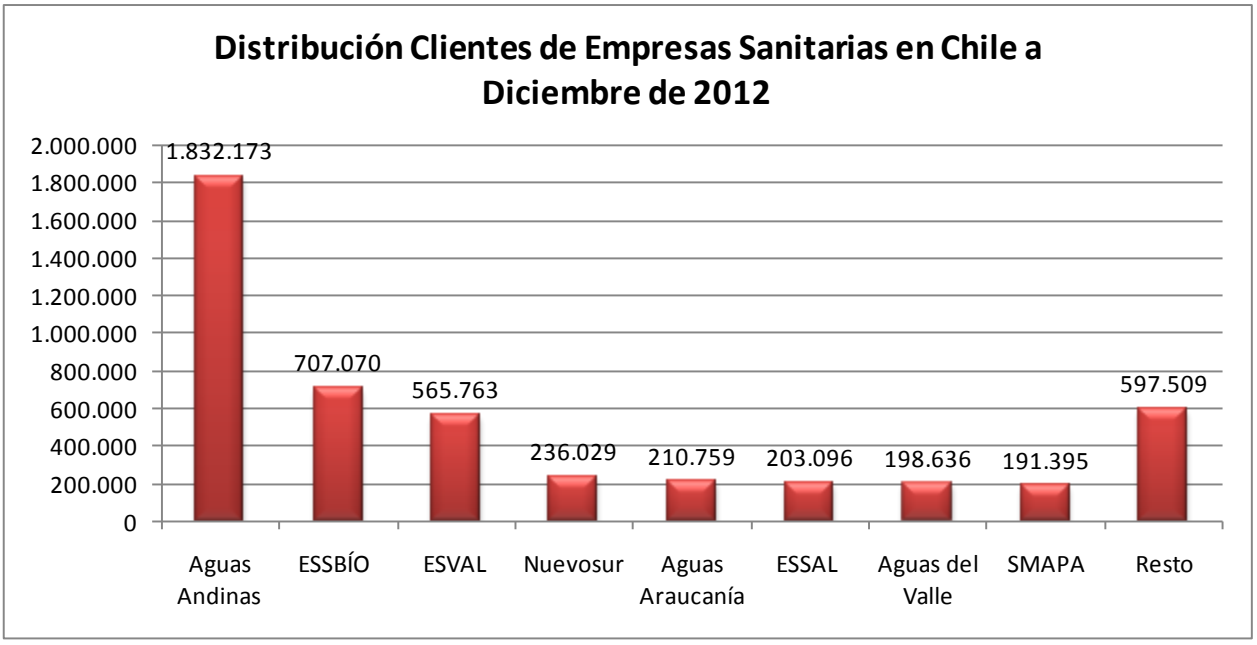
Tabla N°1: Cobertura de los servicios de las sanitarias en Chile a Diciembre 2012

Variable	Magnitud
Cobertura Urbana Agua Potable	99,9%
Cobertura Urbana Alcantarillado	96,3%
Cobertura Tratamiento Aguas Servidas[2]	99,8%

Fuente: SISS

Con respecto al número de clientes de las empresas sanitarias, a Diciembre de 2012 se estima en 4.742.430, distribuidos de la siguiente manera:

Gráfico N°1: Distribución clientes de las empresas sanitarias chilenas



Fuente: SISS

1.1.2. Aguas Andinas

Aguas Andinas, la empresa sanitaria más grande del país, posee presencia en:

- Gran Santiago y zonas periféricas: a través de las empresas Aguas Andinas, Aguas Manquehue y Aguas Cordillera (las 2 últimas atienden a determinadas zonas del sector nororiente de la capital).[3]
- Región de Los Lagos y Región de los Ríos: a través de la adquisición de ESSAL (Empresa de Servicios Sanitarios de Los Lagos).

[2]Porcentaje de cobertura sobre la población saneada, vale decir, la que cuenta con alcantarillado.
[3]En el presente informe, se entenderá por Aguas Andinas a las 3 empresas que la conforman: Aguas Andinas S.A., Aguas Manquehue S.A. y Aguas Cordillera S.A.

Su zona de concesión en la Región Metropolitana alcanza las 70.000 hectáreas, presentando las siguientes cifras de interés a Diciembre de 2012:

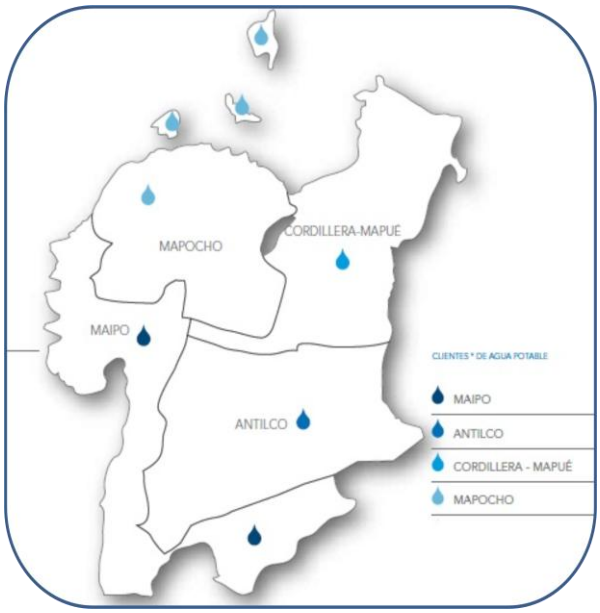
Tabla N°2: Cifras relevantes de Aguas Andinas

Variable	Magnitud
N° Clientes Agua Potable	1.831.803
N° Clientes Alcantarillado	1.786.749
N° Clientes nivel consolidado[4]	1.832.173
Población Urbana Abastecida	6.416.335
Metros Cúbicos Agua Potable Facturados año 2012	509.343.000 m3
Metros Cúbicos Alcantarillado Facturados año 2012	495.080.000 m3
Consumo Promedio mensual Agua Potable	23,2 m3
Consumo Promedio mensual Alcantarillado	23,1 m3

Fuente: Memoria Anual Aguas Andinas Año 2012

Para la eficiente administración y atención de la clientela, la empresa ha dividido la zona de concesión perteneciente a la Región Metropolitana en 4 zonales: Antilco, Maipo, Mapocho y Cordillera-Mapué[5].

Figura N°1: Distribución zonales de Aguas Andinas



Fuente: Memoria Anual Aguas Andinas Año 2012

[4]Considera clientes que reciben servicio de agua potable, alcantarillado, o ambos.
[5] Para conocer la distribución de comunas para cada zonal ver Anexo A.

Antilco, Mapocho y Cordillera-Mapué abarcan todas las comunas del Gran Santiago, mientras que Maipo cubre localidades pertenecientes a las provincias de Melipilla, Maipo, Talagante y Chacabuco.

Con respecto a la tarifa mensual a cobrar a los clientes, los principales ítems en que se desglosa son:

- Cargo fijo.
- Agua potable (m3) en periodo no punta
- Agua potable (m3) en periodo punta[6]
- Sobreconsumo de agua potable (m3) en periodo punta
- Servicio de alcantarillado de aguas servidas (m3)
- Tratamiento de aguas servidas (m3)

Cabe mencionar que las tarifas adquieren diferente valores para clientes de Aguas Andinas, Aguas Cordillera y Aguas Manquehue, y a su vez para los distintos grupos en que se han dividido los consumidores, según su ubicación geográfica, para la diferenciación de los precios a aplicar (usuarios de Aguas Andinas se han dividido en 5 grupos, Aguas Manquehue en 4, mientras que los de Cordillera conforman un solo grupo). El Grupo más numeroso es el número 1 de Aguas Andinas, que comprende la mayor parte del Gran Santiago y Pirque, por lo que se detallan a continuación sus tarifas [2]:

Tabla N°3: Tarifas aplicadas al principal grupo de consumidores

CONCEPTO	VALOR ($CLP)
Cargo Fijo ($/mes)	587
Agua Potable periodo no punta ($/m3)	299,21
Agua Potable periodo punta ($/m3)	296,72
Sobreconsumo agua potable periodo punta ($/m3)	746,84
Servicio alcantarillado aguas servidas ($/m3)	233,99
Tratamiento aguas servidas ($/m3)	136,12

Fuente: Reglamento Tarifario vigente desde Septiembre 2012, disponible en web de SISS

1.1.3. Agua no contabilizada

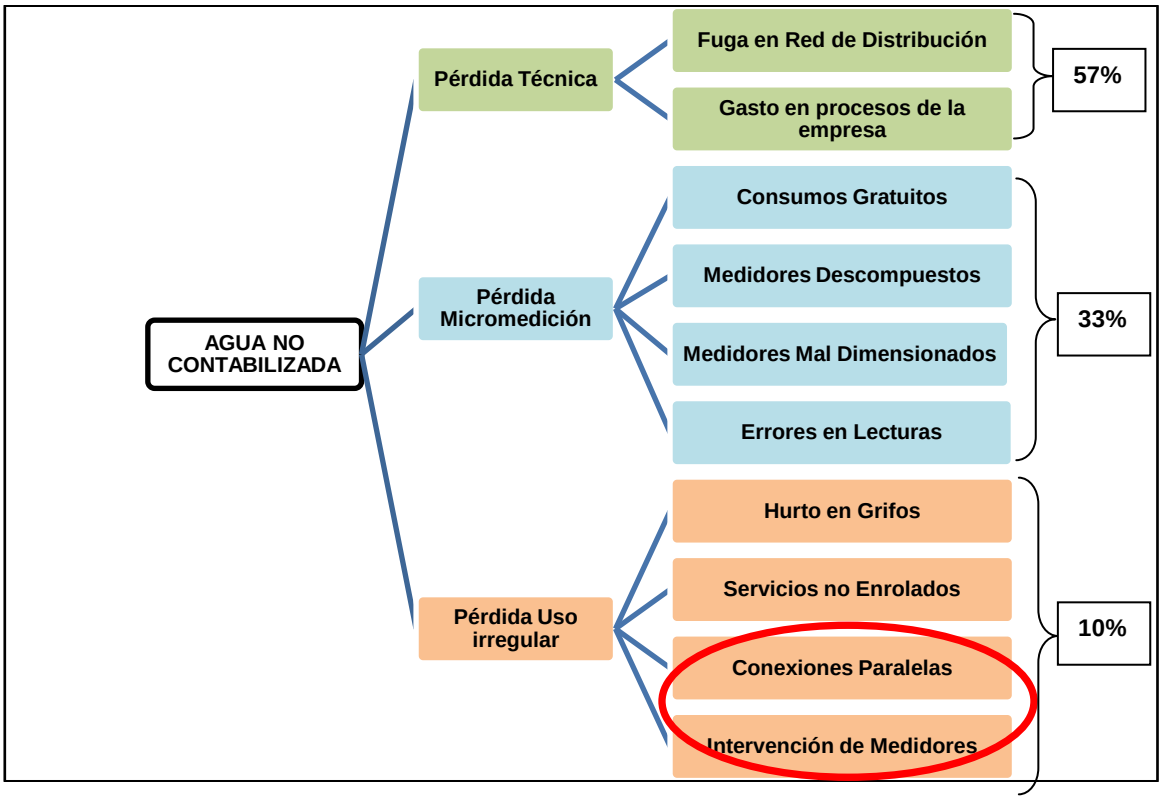
Como agua no contabilizada se define aquel volumen de agua producido (que ingresa a la red de distribución de agua potable), pero no facturado, vale decir, no cobrado a los clientes. Se representa por la fórmula:

$$\text{Índice } \boldsymbol{Agua\ no\ Contabilizada}\ (\%) = \frac{\boldsymbol{Agua\ Producida - Agua\ Facturada}}{\boldsymbol{Agua\ Producida}} * \mathbf{100}$$

Se tienen cierto número de factores que explican este hecho, los que se detallan en el diagrama siguiente (con su respectivo porcentaje de incidencia respecto al total de agua no facturada):

[6] Desde 1 de Diciembre a 31 de Marzo

Figura N°2: Factores que determinan el agua no contabilizada



Fuente: Elaboración propia

Todas las empresas sanitarias poseen un cierto porcentaje de agua no facturada, cuya magnitud dependerá de la calidad de sus instalaciones y cantidad de conexiones irregulares que se den. En este contexto cada una de las sanitarias presenta distinto grado de control sobre cada uno de estos factores, siendo posible en algunos casos desarrollar una adecuada gestión para reducir el nivel de agua no facturada, haciendo un apropiado balance entre costo y beneficio.

Al desarrollar más en detalle los factores expuestos anteriormente [3]:

i. **Pérdida Técnica:** referente a la pérdida de agua producto de roturas y filtraciones presentes a lo largo de la red de distribución de agua potable, además del uso del recurso hídrico en la limpieza de los filtros de las plantas potabilizadoras.
ii. **Pérdida Micromedición:** toda aquella pérdida en la que se incurre producto de factores asociados al proceso mensual de medición que impiden la facturación del 100% de lo que el cliente consume, como la existencia de clientes con gratuidad de aguas (Bomberos y recintos de Aguas Andinas), medidores descompuestos o muy antiguos que sub-miden o simplemente no miden el flujo, medidores que poseen un diámetro mayor al requerido (medidores mayores sub-miden flujos de agua menores pues son menos sensibles a variaciones pequeñas del caudal) y errores en la lectura.

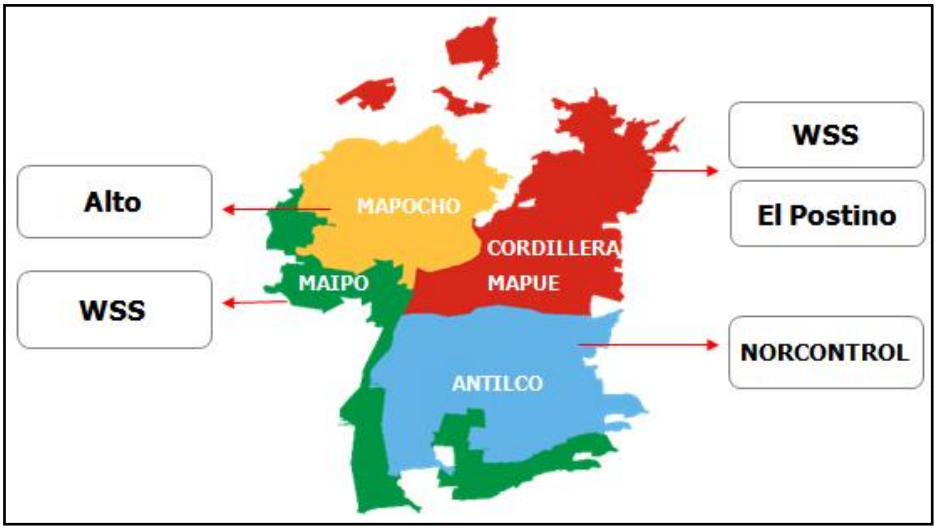
iii. **Pérdida uso Irregular:** toda aquella agua no facturada producto de prácticas ejercidas por las personas con el objeto de hurtar agua o pagar menos dinero por el servicio del que corresponde. Entre ellas se tiene la extracción de agua desde los grifos, conexiones a la red de agua potable sin estar registrado como cliente en la sanitaria (servicios no enrolados), clientes suspendidos que mediante conexiones irregulares siguen consumiendo el recurso, conexiones paralelas a los medidores de modo que sólo parte del caudal sea registrado por el medidor, e intervención del medidor para que su registro sea menor al que corresponde.

Es así como el trabajo de memoria pretende abarcar el tercer y cuarto punto del esquema concernientes a la "pérdida por uso irregular", vale decir, la detección de clientes que facturan un volumen menor al correspondiente a su consumo real, **ya sea por la existencia de una conexión paralela,** o **por una intervención de su medidor**(los conceptos encerrados en una elipse). Estos serán los ítems abarcados, pues para el análisis que se realizará se requiere contar con datos (catastrales, consumos históricos, etc.) de los potenciales clientes irregulares, y en los casos de robo de agua en grifos o servicios no enrolados no se tiene un medidor asociado al ilícito, por ende, no hay registro de datos de la persona que comete el hurto.

1.1.4. Metodología actual de detección de clientes irregulares en Aguas Andinas

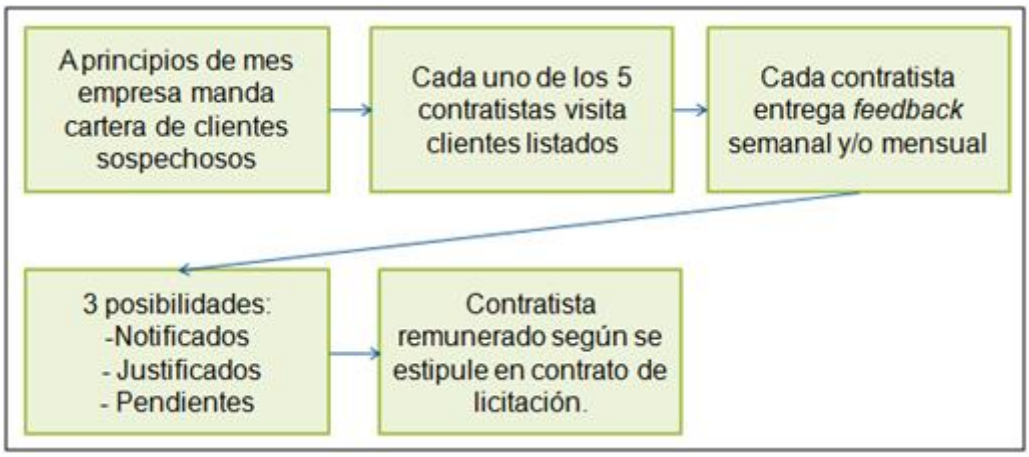
Actualmente, para combatir las prácticas irregulares relativas a **intervención de MAP (medidor de agua potable)** y **conexión paralela**, la empresa ha subcontratado el servicio de detección de clientes irregulares, asignando a cada una de las 4 zonales de Región Metropolitana un contratista encargado de fiscalizar a los clientes de los que se presume una conducta irregular. Conjuntamente, se trabaja con un quinto contratista, El Postino, enfocado en fiscalizar a aquellos clientes del segmento socioeconómico ABC1, los que geográficamente se sitúan en el sector nororiente de la Zona Cordillera-Mapué.

Figura N°3: Distribución contratistas según zonales



Fuente: Ledesma O., 2011, Presentación Proyecto Micromedición

El procedimiento a seguir es:



En resumen, la empresa es la responsable de enviar el listado de clientes de los que se presume una conducta irregular. El tamaño de la cartera mensual consolidada (para las 4 zonas) es de alrededor de **3.500 clientes**: un listado de **1.000** sospechosos para cada una de las 3 zonales más grandes (Antilco, Mapocho y Cordillera Mapué) y otro de **500** sospechosos para la zonal con menor población (Maipo). Se han establecido estas magnitudes para las carteras debido a que esas cifras son las que hoy en día alcanzan a cubrir los contratistas mensualmente, considerando que generalmente cada uno dispone de 2 móviles como recurso para realizar las inspecciones. Es así como los contratistas visitan estas residencias, realizando las pruebas de agua y medidores necesarias para descartar un hecho ilícito. De sus visitas hay 3 posibles resultados: cliente notificado, cliente justificado o cliente pendiente.

- **Cliente notificado:** se determinó que estaba incurriendo en una práctica irregular (intervención del MAP o conexión paralela) y por ende se le notifica,

informándole sobre los pasos a seguir para regularizar su situación. El monto a cobrar se desglosa en:

Monto a Cobrar = Costo reglamentación + Cobro consumo presunto

El costo de reglamentación incluye el cambio de medidor en caso de ser necesario, y el cobro por consumo presunto se refiere al monto retroactivo que el cliente debe reembolsar por todo el volumen de agua que dejó de facturar debido a la instalación ilegítima (volumen estimado por la empresa, la que analizando el gráfico de consumos del cliente infiere en qué momento comenzó el ilícito).

- **Cliente justificado:** se demuestra que su bajo consumo se debe a que la residencia está deshabitada, no se efectúa riego de área verde, se abastecen de un pozo, efectúan riego con agua de canal o hay un trabajo de demolición, entre otras explicaciones.

- **Cliente pendiente:** aquél al que no se le ha podido realizar una visita efectiva debido a que el lugar se encuentra sin personas al momento de concurrir, o el cliente se niega a recibir al personal. En el último caso se envía una carta dando detalles de las acciones en que se incurrirán en caso de persistir la negativa del cliente. Si el contratista no alcanza a concluir la inspección dentro del mes, deberá finiquitarla en el(los) próximo(s) mes(es), no incluyéndose estos casos en las nuevas carteras de sospechosos otorgadas posteriormente.

Hoy en día, para confeccionar la cartera mensual de presuntos irregulares, la sanitaria se ha basado en criterios como:

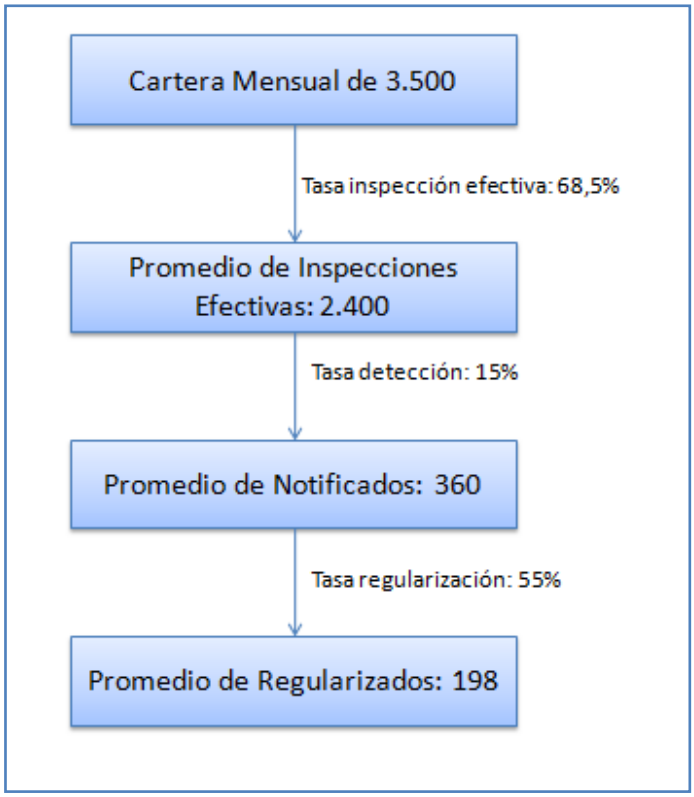
a. Análisis de clientes que hayan sido visitados por personal técnico de la empresa para trabajos de reparación y que registren a su vez un descenso en el consumo (pues se tiene antecedentes que ciertos trabajadores gasfiteros ofrecen realizar la intervención al MAP).
b. Inspección de clientes que hayan sido denunciados anónimamente por sus vecinos (existe una vía en que los propios vecinos pueden denunciar a los clientes irregulares por internet o vía telefónica). Este criterio tiene una alta tasa de efectividad (6 de cada 10 acusaciones son verídicas), pero son escasas las denuncias recibidas mensualmente (entre 150 y 200).
c. Barrido de zonas (ciertas manzanas y calles) que los propios contratistas proponen, pues sospechan de una alta concentración de intervenciones en ellas.

Entre los clientes notificados, históricamente se registran usuarios de las 3 categorías (comercial, industrial y residencial) así como de todo el espectro socioeconómico. A su vez, el tipo de intervención al medidor va desde las más simples como perforar la cúpula, prensar la cúpula para detener el giro de la aguja o invertir el medidor para que retroceda la lectura, hasta métodos más complejos como rebajar las aspas para hacer el giro más lento.

En términos cuantitativos, las carteras presentan en promedio un **68,5% de inspecciones efectivas** (vale decir, se realiza la fiscalización a 2.400 clientes aproximadamente), y un **15% de efectividad en la detección** sobre el universo

válidamente inspeccionado (cerca de 360 notificados al mes). En el capítulo siguiente se ahondará más en estas cifras.

Figura N°4: Cifras mensuales del actual método de detección de irregulares



Fuente: Elaboración propia

1.2. DESCRIPCIÓN DEL PROYECTO Y JUSTIFICACIÓN

Desde Septiembre de 2010 se está implementando en Aguas Andinas un proyecto denominado “Reducción de Pérdida por Agua no Facturada”, destinado a atacar cada uno de los factores detallados anteriormente (punto 1.1.3), que son los causantes de que se cobre menos agua de la producida.

Históricamente:

Tabla N°4: Evolución agua no facturada para Aguas Andinas

Año	Índice Agua no Facturada
2008	29%
2009	30%
2010	31%

Fuente: Ledesma O.,2011, Presentación Proyecto Micromedición

En efecto, en términos mensuales para el año 2010, se tuvo una producción promedio mensual de 58MM de m3, de los cuales, se facturó 40MM de m3, vale decir,

se dejó de facturar un **promedio de 18MM de m3 al mes.** Cabe destacar que en las empresas sanitarias de los países más desarrollados (mayoritariamente presentes en Europa y Asia), este porcentaje oscila entre el 10% y el 20%, mientras que en los países menos desarrollados, puede alcanzar el 60% [4].

Según estudios realizados por SUEZ[7], la repercusión que tiene cada factor en el agua no contabilizada es:

Tabla N°5: Desglose del agua mensual no facturada

Factor	Porcentaje	Magnitud en MM m3 al mes (2010)
Pérdida Técnica	57%	10,3
Pérdida Micromedición	33%	5,9
Pérdida Uso Irregular	10%	1,8
TOTAL MENSUAL	100%	18

Fuente: Ledesma O.,2011, Presentación Proyecto Micromedición

Como se mencionó, pérdida por uso irregular incluye hurto en grifos, servicios no enrolados, conexiones paralelas e intervención del MAP. Aguas Andinas ha estimado que de los 1,8MM m3 de agua potable perdidos al mes por uso irregular, cerca de **700.000 m3 corresponden a clientes enrolados con instalaciones irregulares** (conexión paralela o intervención de MAP). Si se considera que al no facturar dicho volumen, se deja de cobrar por concepto de agua potable, servicio de alcantarillado y tratamiento de aguas servidas, se habla de una **pérdida mensual que bordea los $468MM.**

Realizando un cálculo bastante general, si se ha estimado que un cliente regularizado aumenta su consumo, después de acabada la intervención, en **29 m3** en promedio, y se habla de 700.000 m3 perdidos por intervenciones, se está ante unos 24.000 potenciales clientes irregulares de la sanitaria, lo que equivale al 1,31% del total.

Hoy en día, la tasa de detección de los contratistas oscila en torno al **15%,** entendiéndose por tasa de detección:

$$\textbf{Tasa de detección clientes irregulares} = \frac{\textbf{Clientes Notificados}}{\textbf{Inspecciones Efectivas}}$$

En donde por inspecciones efectivas se entienden los clientes notificados más los clientes justificados (vale decir que efectivamente se hayan realizado las pruebas de agua pertinentes en el hogar).

Esta cifra del 15% resulta del promedio ponderado de tasas de detecciones experimentadas por las carteras enviadas a las 4 zonales durante el año 2011. La tendencia observada es que las tasas de Antilco y Mapocho son las más elevadas, mientras que las de Cordillera Mapué y Maipo las más bajas.

En el escenario actual, para determinar la cartera de clientes sospechosos no existe un modelo formal de detección, solamente se realiza una inspección a aquellos clientes de los que se presume una mayor probabilidad de realización de ilícito (pertenecientes

[7]Empresa multinacional francesa proveedora de servicios de agua, gas y electricidad, además de ser dueña del socio controlador de Aguas Andinas, Agbar.

a barrios donde es frecuente el fraude o visitados anteriormente por personal contratista) y se incluyen las denuncias de los propios clientes. En consecuencia, se desaprovecha la potencialidad que ofrece un sinfín de otras variables que caracterizan a los consumidores, a través de una gran cantidad de datos de la que se dispone. Debido a esto, la parte del proyecto "Reducción de Pérdida por Agua no Facturada" relacionada con intervenciones fraudulentas, pretende explotar todo el valor agregado que pueda aportar la base de datos con la que se cuenta en la detección de fraude.

En base a lo anteriormente expuesto, el proyecto destinado a crear un modelo formal de detección de fraude en clientes de la sanitaria, que **aproveche todos los datos de los que se dispone en la actualidad,** busca:

- Recuperar ingresos, al cobrar retroactivamente el agua no facturada al cliente irregular.
- Disminuir probabilidad de alteración del servicio a otros clientes (las intervenciones fraudulentas aumentan el riesgo de una disminución de la calidad del servicio a los vecinos cercanos).
- Disminuir costos por concepto de daño a infraestructura debido a intervenciones en redes de distribución.
- Disminuir consumo indiscriminado de agua.
- Asegurar las correctas facturaciones futuras del cliente irregular.
- Dar una potente señal al combatir estas malas prácticas, de modo de disuadir a clientes que se vean tentados a utilizar métodos fraudulentos.

Pese a que mes a mes, cifras como el número de clientes notificados o la magnitud de cartera de sospechosos presentan variaciones, la tendencia general durante el año 2011 fue:

Tabla N°6: Principales índices relacionados con plan clientes irregulares

Variable	Magnitud
Tamaño cartera mensual de clientes sospechosos	3.500
Inspecciones efectivas al mes	2.400
Tasa promedio detección	15%
Tasa promedio regularización[8]	55%

Fuente: Elaboración propia

Para valorar en términos monetarios lo que se recupera por cada cliente notificado regularizado, se define la **"recuperación por cliente notificado regularizado" (RPCNR)** como:

[8]Como cliente regularizado se define a aquel cliente, que después de ser notificado se acerca a Aguas Andinas S.A. a normalizar su situación, acordando una forma de pago con la empresa.

RPCNR = Monto retroactivo + Recuperación futura anual = $456.000
($223.000) + ($233.000)

En donde el **monto retroactivo** corresponde a los metros cúbicos cobrados que se estima dejó de facturar el cliente debido a la instalación irregular (además del pago por la instalación de un nuevo medidor si es necesario), y la **recuperación futura anual** es lo que la empresa estima que volverá a cobrar al cliente debido a la normalización del servicio (la convención es proyectar a un año).

Si para ambos conceptos se consideran los valores promedio que hasta ahora la compañía presenta; con un monto retroactivo de $223.000 y una recuperación mensual promedio post-normalización de 29 m3 (lo que se traduce en una recuperación futura anual de $233.000), se está frente a un RPCNR que bordea los $456.000.[9]

En consecuencia, cada aumento porcentual en la tasa de notificación, manteniendo el actual tamaño de cartera enviado mensualmente, se traduce en 24 nuevos notificados; y si de este número, un 55% regularizará, en términos monetarios significa que **un aumento del 1% en la tasa de detección implica una recaudación extra de $6.019.200 al mes.** Bajo esta perspectiva, sin duda resulta atractivo para la empresa contar con un modelo que logre mejorar su actual tasa de detección.

1.3. OBJETIVOS

Objetivo General

Construir un modelo enfocado en detectar clientes, de una empresa sanitaria, que presenten intervenciones irregulares en su red de agua potable, de modo de disminuir las pérdidas de la empresa por concepto de agua no facturada.

Objetivos Específicos

- Caracterizar la clientela de la empresa sanitaria en general, así como a los autores del ilícito de robo de agua potable en específico.
- Determinar el conjunto de datos que serán útiles en la construcción del modelo de detección de fraude.
- Obtener distintos modelos de detección de clientes irregulares a través de la utilización de determinados métodos de clasificación.
- Evaluar los métodos desarrollados y determinar cuál arroja mejores resultados para la problemática dada.
- Interpretar los resultados y generar una metodología tendiente a orientar a la empresa sanitaria en la aplicación del modelo seleccionado.

[9] Datos extraídos de todos los usuarios irregulares contabilizados desde Septiembre de 2010.

1.4. MARCO CONCEPTUAL

1.4.1. Conceptos útiles relativos a la empresa sanitaria

Cliente: es el inmueble que recibe el servicio sanitario de agua potable, alcantarillado, o ambos, quedando allí radicadas todas las obligaciones para con el prestador, derivadas del servicio. Por tanto, el cliente o usuario es el titular que habita o reside en dicho inmueble, sea persona natural o jurídica.

Cliente área verde: usuario cuyo MAP está destinado exclusivamente al regadío de cierta área verde.

Cliente estacional: clientes cuyos consumos presentan una clara tendencia al alza en época estival, debido a que la totalidad o gran parte de su predio se compone de área verde.

Clave de lectura: aquella clave determinada cada mes por el contratista, que refleja el estado de la lectura. La más usual es la "N", vale decir, que el proceso de lectura fue normal, sin embargo hay una gran cantidad de otros códigos, que pueden indicar desde casa deshabitada, hasta que el medidor está descompuesto.[10]

Clave irregular: un subconjunto del universo de claves de lectura, que dan indicio de la existencia de prácticas irregulares por parte del cliente (medidor invertido: S, instalación irregular: K, medidor intervenido: Q).

Tarifa: código que refleja el tipo de cliente, dependiendo si corresponde a un condominio, es un cliente corporativo, etc.

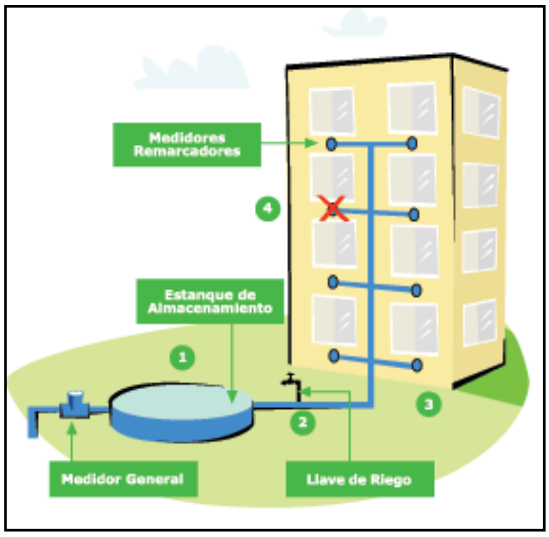
Tarifa 11: es la tarifa más común de servicios individuales, corresponde a la aplicada a las casas.

Tarifa 7: Medidor general o matriz con remarcadores, representa a los medidores que abastecen edificios y condominios.

Tarifa 77: Medidores remarcadores asociados a departamentos pertenecientes a un edificio, son abastecidos por el medidor matriz (tarifa 7) y se les cobra la diferencia entre el consumo del matriz y la suma de los remarcadores (según algún tipo de prorrateo).

[10] Para conocer el detalle de las claves, ver Anexo B.

Figura N°5: Esquema Tarifa 7 y Tarifa 77

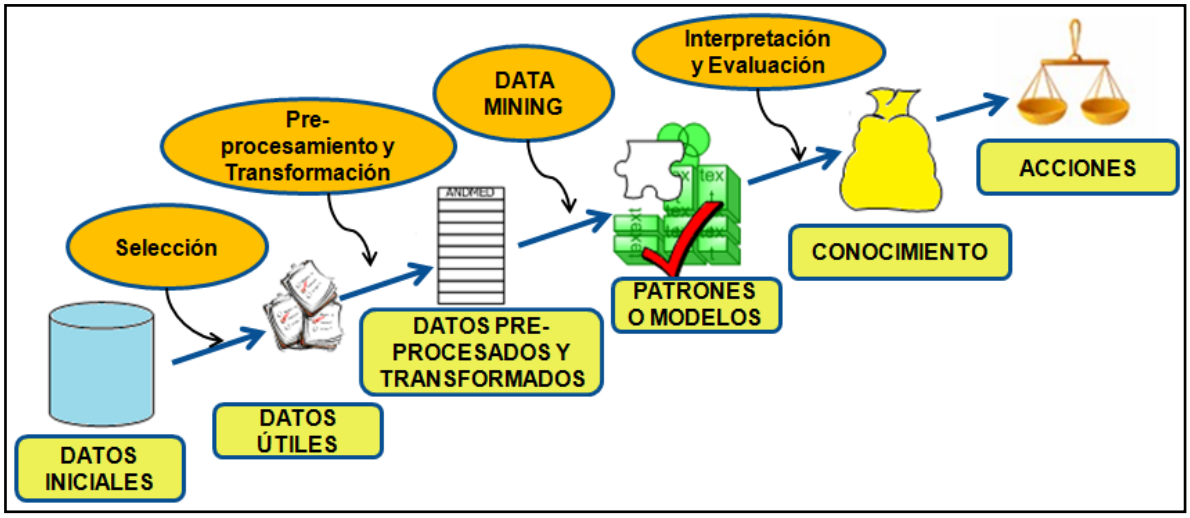


Fuente: Web Aguas Andinas

1.4.2. Conceptos relacionados al proceso KDD

El desarrollo del trabajo de memoria se sustentará en la metodología KDD (Knowledge Discovery in Databases), definida como aquella área interdisciplinaria enfocada en extraer conocimiento útil desde los datos, a través de la identificación de patrones válidos, novedosos y comprensibles [5]. A su vez, corresponde a un caso de ***aprendizaje supervisado***, pues existe un conocimiento a priori de la función objetivo que se desea deducir a partir de los datos (cliente fraudulento o no fraudulento en este trabajo), a diferencia de los casos de ***aprendizaje no supervisado*** en los que no se tiene conocimiento previo de la clasificación a obtener, de hecho eso es lo que se busca dilucidar. Las etapas que conforman esta metodología KDD se diagraman en el siguiente esquema:

Figura N°6: El proceso KDD



Fuente: Elaboración propia

A continuación se procede a describir cada una de las cuatro etapas, así como los métodos y modelos estadísticos asociados.

a. Selección de Datos

Usualmente se dispone de una enorme cantidad de datos, generalmente representada en muchas tablas, cada una con distintas observaciones (filas) y atributos (columnas). Esta etapa consta en poder discriminar qué información es potencialmente valiosa para el estudio, y cuál no, de modo de optimizar el tiempo de procesamiento en las etapas posteriores.

b. Pre-procesamiento y Transformación de Datos

El **pre-procesamiento** consiste en realizar una especie de “limpieza” con el fin de eliminar el mayor número posible de datos erróneos o inconsistentes que puedan disminuir la calidad de los resultados a obtener. Para ello se tratan los casos correspondiente a missings (datos faltantes), outliers (datos fuera de rango) y duplicados.

Luego se procede a la **transformación** de la data, llevando a cabo una conversión de los datos, dependiendo de los requerimientos de los modelos a aplicar, incluyéndose transformación de variables categóricas en dummies; normalización o estandarización de variables cuantitativas; y discretización de datos, en caso de ser necesario. A su vez se pueden crear nuevas variables a partir de las ya existentes.

Otro propósito que persigue esta etapa es la reducción de la magnitud de la data y la independencia entre variables, disminuyendo la dimensionalidad del problema a través de métodos de agrupamiento y elección de atributos [6]. Para ello se cuenta con diversos procedimientos, entre ellos:

b.1. Correlación entre variables cuantitativas: se utiliza el coeficiente de correlación de Pearson para comprobar la dependencia entre una variable numérica y otra:

$$\boldsymbol{Coef.de\ correlación = \rho(i,j) = \frac{\sigma i,j}{\sigma i\ \sigma j}}$$

Con i,j variables numéricas.

De este modo, con un valor de $|\rho_{i,j}| > 0{,}75$, se tiene dependencia fuerte entre una variable y otra, prosiguiendo a eliminar las redundantes. La limitación es que este método sólo identifica dependencias lineales.

b.2. Correlación entre variables categóricas: se utiliza el test χ^2 para independencia de 2 variables cualitativas, que utiliza tabla de contingencia.

Sea la variable A ={A_1, A_2, A_3…..A_r}, vale decir puede tomar r valores posibles, y sea la variable B ={B_1, B_2, B_3…..B_s}, vale decir puede tomar s valores posibles, se tiene una tabla de contingencia de dimensión rxs:

		B				
		B_1	B_2	B_j	B_s	
A	A_1	n_{11}	n_{12}	n_{1j}	n_{1s}	$n_{1.}$
	A_2	n_{21}	n_{22}	n_{2j}	n_{2s}	$n_{2.}$
	A_i	n_{i1}	n_{i2}	n_{ij}	n_{is}	$n_{i.}$
	A_r	n_{r1}	n_{r2}	n_{rj}	n_{rs}	$n_{r.}$
		$n_{.1}$	$n_{.2}$	$n_{.j}$	$n_{.s}$	n

En donde $n_{i,j}$ = N° de casos en que observación presenta A=A_i y B=B_j. La hipótesis de trabajo es:

H_0 = variables A y B independientes

H_1= variables A y B dependientes

Se define la "frecuencia esperada" como $\boldsymbol{e_{ij} = (n_{i.} \mathbf{x}\ n_{.j})/n}$

Luego se construye el estadístico χ:

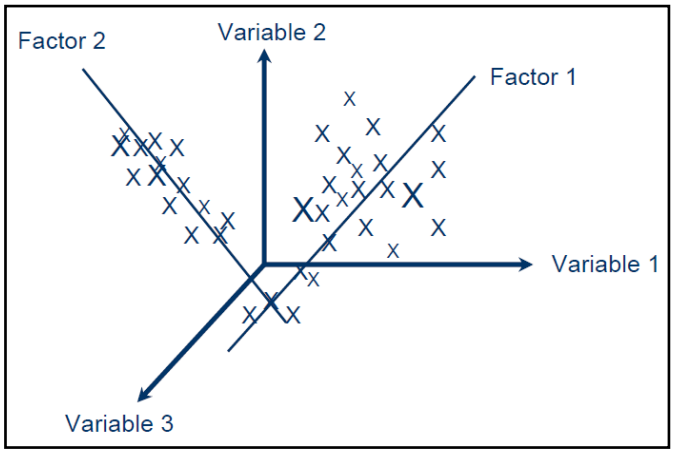
$$\boldsymbol{\chi^2 = \frac{(n_{11} - e_{11})^2}{e_{11}} + \frac{(n_{12} - e_{12})^2}{e_{12}} + \cdots + \frac{(n_{rs} - e_{rs})^2}{e_{rs}} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - e_{ij})^2}{e_{ij}}}$$

Comparándose este valor con el valor teórico $\boldsymbol{\chi^2_{(r-1)(s-1)}}$,extraído de tabla de distribución (pues bajo la hipótesis nula $\boldsymbol{\chi^2 \rightarrow \chi^2_{(r-1)(s-1)}}$), se tiene la siguiente regla de asociación:

Regla : - Si $\chi^2 > \chi^2_{1-\alpha}$ se rechaza H_0 (hay dependencia).
- Si $\chi^2 \leq \chi^2_{1-\alpha}$ se asume H_0 (no hay dependencia).

b.3. Método de agrupamiento ACP: el Análisis de Componente Principales ofrece la posibilidad de reducir la dimensionalidad de un conjunto de datos. Permite obtener de un conjunto de p variables otro conjunto de q componentes principales (las nuevas variables) ortogonales entre sí (independientes) a través de una transformación lineal de los datos originales (p>q), convirtiendo la información en un nuevo sistema de coordenadas tal que la mayor varianza se halle en la primera componente principal, la segunda mayor varianza en la segunda, y así sucesivamente [7]. El número de CP's a adoptar dependerá de la cantidad de varianza que expliquen. La limitación es que sólo encuentra relaciones lineales, además debe existir una interpretación lógica de las componentes obtenidas.

Gráfico N°2: Funcionamiento del ACP



Fuente: Terrádez M., "Análisis de Componentes Principales"

c. Data Mining

Corresponde a la etapa central del proceso KDD, en donde, una vez preparada la data, se buscar aplicar diversas técnicas para extraer toda la potencialidad de los datos, y en consecuencia, cumplir con el objetivo final del trabajo. En esta instancia se procede a aplicar el (los) método(s) del tipo aprendizaje supervisado o no supervisado, según corresponda. Como se indicó anteriormente, el presente trabajo corresponde a un caso de aprendizaje supervisado, que puede ser cubierto por los siguientes métodos de clasificación:

- Regresión Logística Binaria
- Árbol de Decisión
- Red Neuronal

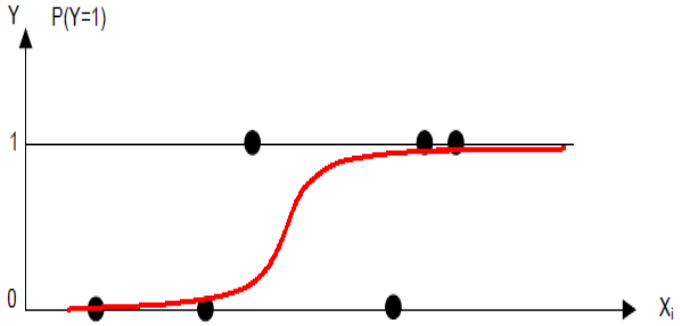
c.1. Regresión Logística Binaria:

La regresión logística es un método estadístico de clasificación utilizado para predecir el comportamiento de una variable dependiente dicotómica o para evaluar la relación que ésta tenga con otras variables independientes o de control, utilizando métodos de Máxima Verosimilitud para estimar los parámetros.
Lo que se pretende mediante la regresión logística es expresar la probabilidad de que ocurra el evento en estudio como función de ciertas variables, que se presumen relevantes o influyentes.

$$P(Y=1) = f(\alpha + \beta_1 X_1 + \cdots + \beta_k X_k)$$

La función f que se busca debe estar acotada por los valores 0 y 1, para niveles bajos de la variable independiente f debe aproximarse a 0, a medida que el valor de la variable dependiente crezca la probabilidad también lo haga, hasta en cierto punto disminuir su pendiente. Vale decir la función debe asemejarse a una curva "S":

Gráfico N° 3: Forma de la función asociada a la regresión logística



Se conocen 2 funciones que cumplen los mencionados requisitos, la logística (que da origen al modelo Logit) y la de distribución de una normal estándar (que da origen al modelo Probit).

Modelo Logit: sea y la variable que se quiere predecir o modelar, y sea $\{X_i\}_{i=1}^{k}$ el conjunto de variables explicativas (independientes), la ecuación general o función logística es:

$$P[y=1] = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\cdots+\beta_k X_k)}} \quad \text{siendo} \quad f(z) = \frac{\exp(z)}{1+\exp(z)}$$

Donde $\{\beta_i\}_{i=1}^{k}$ son los parámetros entregados por el modelo

Modelo Probit: sea y la variable que se quiere predecir o modelar, y sea $\{X_i\}_{i=1}^{k}$ el conjunto de variables explicativas (independientes), la ecuación general o función logística es:

$$P[y=1] = \int_{-\infty}^{\alpha+\beta_1 X_1+\cdots+\beta_k X_k} \frac{1}{\sqrt{2\pi}} * \exp\left(-\frac{t^2}{2}\right) dt$$

$$\text{con} \quad f(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

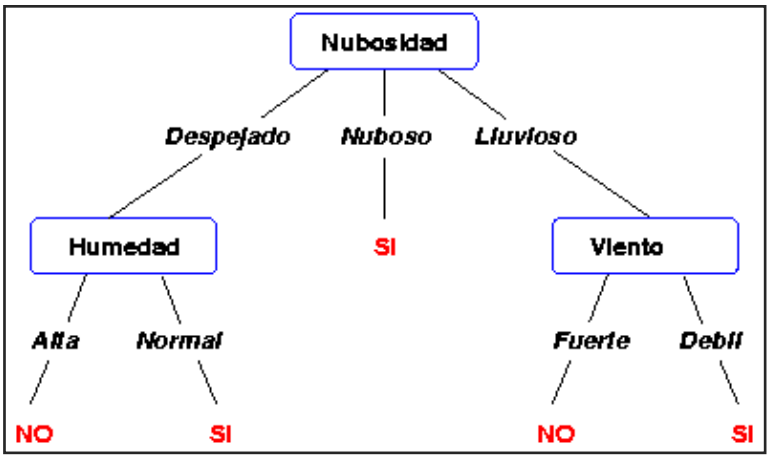
Donde $\{\boldsymbol{\beta}_i\}_{i=1}^{k}$ son los parámetros entregados por el modelo

c.2. Árbol de Decisión Clasificatorio

Los árboles de decisiones clasificatorios constituyen un método inductivo supervisado consistente en un conjunto de reglas, estructuradas en forma de árbol, destinadas a dividir a una población heterogénea en grupos más pequeños, en base a una variable objetivo (variable dependiente). Un árbol se compone de:

- Nodos Intermedios: contiene el nombre del atributo y engendran 2 ó más nodos (dependiendo del método empleado).
- Nodos terminales: aquellos que no pueden dividirse más.
- Ramas: arcos de conexión entre nodos, contienen posibles valores del atributo asociado al nodo.

Figura N°7: Árbol de decisión para el problema *"jugar tenis o no"*



Algunos tipos de árboles según el algoritmo empleado:

ID3 (InteractiveDichotomizer 3): uno de los algoritmos más populares, basado en la cantidad de información mutua entre cada variable explicativa y la variable de decisión. Utiliza el concepto de entropía, cuanto menor sea el valor de la entropía, menor incertidumbre y por ende, más útil el atributo para la clasificación pues se produce mayor ganancia de información. Los dominios de los atributos y la variable dependiente deben ser discretos, además matemáticamente se ha demostrado que el criterio de selección de atributos tiende a favorecer la elección de variables con mayor número de valores. Por otra parte, el algoritmo efectúa un test de independencia previo entre cada variable predictora X_i y la variable de clasificación C, de modo que para el proceso inductivo sólo se consideran las variables para las que se rechaza dicho test.

C 4.5: se planteó como una mejora al algoritmo ID3. Genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de profundidad-primero (depth-first). Antes de cada partición de datos, el algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y

selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información [9]. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma elatributo en los datos. Una de sus ventajas es que permite trabajar con variables continuas, además incorpora una poda al árbol, basada en la aplicación de un test de hipótesis destinado a responder la pregunta si vale la pena expandir determinada rama [8].

Entropía de nodo t = S(t) = $\sum_i \boldsymbol{p}\left(\frac{i}{t}\right) \boldsymbol{log_2}\, \boldsymbol{p}(\frac{i}{t})$

Con p(i/t) = proporción de registros del nodo t que pertenecen a la clase i.

CHAID (Chi-squared Automatic Interaction Detected): genera 2 ó más ramas a partir de un nodo, al segmentar el conjunto de datos utilizando test de chi-cuadrado para crear múltiples divisiones. No realiza una "post-poda", sino que en la misma fase de construcción del árbol se decide parar, deteniéndose en caso que el número de casos caiga por debajo de un límite específico o cuando la división no es estadísticamente significativa.

CART: genera 2 ramas a partir de un nodo, basándose en el índice de GINI, que mide el grado de impureza de un nodo. En este caso, los árboles que se van creando van creciendo en complejidad, por lo que para realizar la post-poda se utiliza un criterio de coste-complejidad.

c.3. Red Neuronal

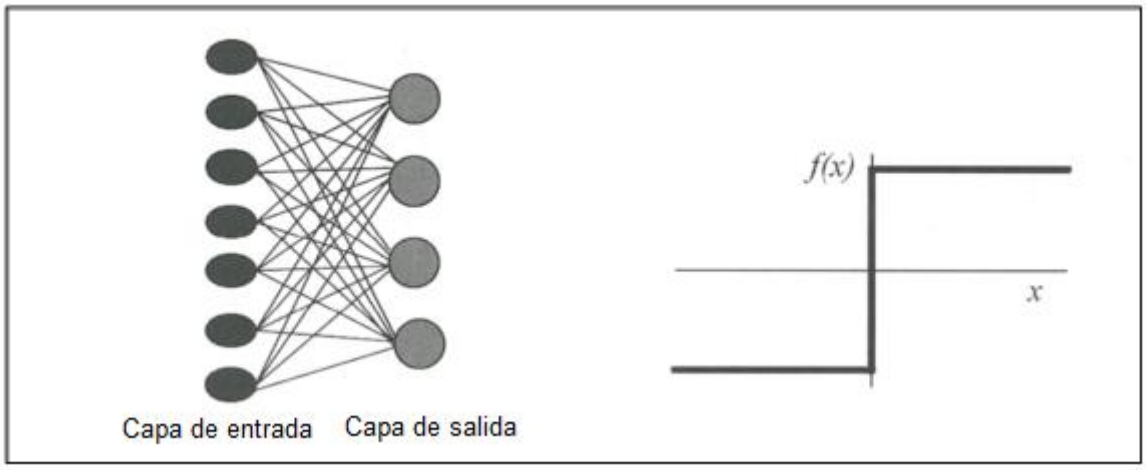
Corresponde a un modelo basado en los complejos sistemas nerviosos de animales y humanos, con su gran cantidad de interconexiones y nodos, que busca la detección de patrones basándose en la interconexión paralela de neuronas artificiales. Desde el punto de vista matemático se puede ver una red neuronal como un grafo dirigido (arcos unidireccionales, vale decir, la información se propaga en un único sentido) y ponderado (las conexiones tienen asociadas un peso), en donde cada unos de los nodos son neuronas artificiales, y los arcos que los unen son las conexiones sinápticas. Lo usual es que las neuronas se agrupen en capas, de manera que una red está formada por varias capas de neuronas, recibiendo las denominaciones [10]:

Capa de entrada: neuronas que reciben los datos que se proporcionan a la red para que los procese.
Capas ocultas: introducen grados de libertad adicionales en la red. El número depende del tipo de red que se está considerando. Realizan gran parte del procesamiento.
Capa de salida: proporciona la respuesta de la red neuronal y también realiza parte del procesamiento.

Algunos de los tipos que se observan:

Perceptrón simple: modelo unidireccional compuesto por 2 capas de neuronas, una de entrada (n neuronas) y otra de salida (m neuronas). La operación de un perceptrón simple se observa como:

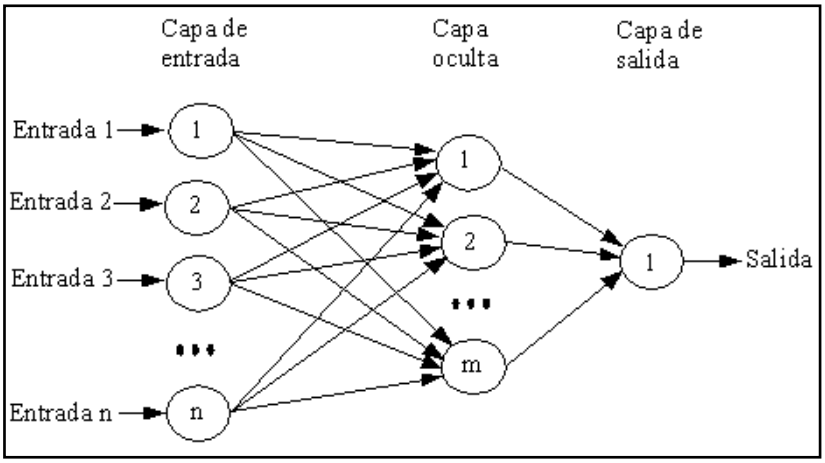
Figura N°8: Red Neuronal tipo perceptrón simple



Las neuronas de entrada no realizan ningún cómputo, únicamente envían la información (en principio se consideran señales discretas {0,1}) a las neuronas de salida. La función de activación de las neuronas de la capa de salida es del tipo escalón (ver figura anterior).

Red Neuronal Multicapa: a diferencia del anterior, incluye capas de neuronas escondidas (no se visualiza ni al input ni al output). Tiene la capacidad de separar inputs en múltiples funciones lineales, detectando patrones más complejos que redes de una función lineal como la vista anteriormente

Figura N°9: Red Neuronal tipo multicapa



d. Interpretación y Evaluación

Finalmente los patrones obtenidos son interpretados, de manera que se tenga certeza de que los resultados arrojados responden a la lógica. Luego se establece un parámetro destinado a evaluar y/o comparar la calidad de los patrones obtenidos, de modo de escoger aquel que aporte mayor valor agregado al problema.

1.5. METODOLOGÍA

1.5.1. Recolección y Selección de datos

Se trabajará con 2 tablas, extraídas de los repositorios de la empresa a través de consultas SQL. Una de ellas contiene todos los datos catastrales de los clientes, incluyéndose variables categóricas y numéricas, tales como dirección, comuna, marca medidor, claves de lectura, fecha instalación del medidor, si posee deuda, cuántos saldos hay pendientes, cuál es el monto que adeuda, categoría (comercial, residencial o industrial) entre otros. La otra tabla contiene datos referentes a los consumos históricos de los clientes, contándose con 5 años de historia. El número de observaciones corresponde a los cerca de 1.800.000 de clientes de la Región Metropolitana.

A su vez, se obtendrá información del histórico de clientes irregulares de la compañía, considerando los detectados desde Marzo de 2011 a Diciembre de 2011, datos que serán aportados por los contratistas y la propia sanitaria. Su magnitud bordea los 3.600.

1.5.2. Caracterización de clientela

Se buscará obtener relaciones entre clientes y variables, de modo de caracterizar al consumidor, y de paso ver la coherencia de los datos. Por ejemplo, en dónde se concentra el mayor número de clientes o en qué sectores el consumo es menor.

1.5.3. Pre-procesamiento de datos

Caso Missings: si una observación (cliente) posee un bajo número de valores faltantes, se procederá a imputar su valor. Las opciones son:

- Imputar por la moda.
- Imputar por el promedio.
- Imputar por aquel valor predominante de la variable en clientes de las mismas características.
- Ignorar la variable si presenta un alto porcentaje de missings.
- Ignorar la fila si presenta un alto porcentaje de missings.

Caso Outliers: si una observación (cliente) posee para cierta variable un valor no coherente que se aleja de la lógica, las opciones son:

- Reemplazar por la tendencia a través de una regresión. Esto deberá aplicarse con frecuencia en aquellos casos en que cliente sufre una fuga de agua, lo que

provoca una sobreestimación del consumo, por lo que habrá que "suavizar la curva". También en aquellos casos en que la casa estuvo deshabitada (si las claves de lectura así lo indican).
- Eliminar la variable si presenta muchos outliers.
- Eliminar la observación si presenta muchos outliers.
- Ignorar, aunque es probable que perjudique la calidad de los resultados arrojados posteriormente por el modelo.

1.5.4. Transformación de datos

Creación de variables: hay una gran disponibilidad de información que puede resultar útil, y sin embargo, no está expresada como variable. Ejemplos de variables a crear:
- Descenso de consumo: la idea es que un código (una macro confeccionada en Visual Basic) identifique si hay caídas bruscas de consumo comparando el antes y después de un cierto instante de tiempo. La variable tendrá el valor de la magnitud de dicha caída, de haberla; en caso contrario su valor será 0.
- Promedio consumo vecinos cercanos: la idea es calcular el promedio de consumo de los vecinos más cercanos de un cliente residencial. Se excluyen clientes comerciales e industriales pues su consumo es muy variable.
- Irregularidad vecindario: la idea es evaluar cada área geográfica, determinando si es un punto fuerte en casos de irregulares (pues muchos clientes notificados residen en el sector). Para este efecto se ocupará el software GIS, que permite mapear a los clientes.
- Consumo verano/Consumo invierno: para observar si hay variaciones en cuanto al consumo en ambos periodos del año.

Conversión de datos: se procederá a transformar los datos según lo requerimientos de cada modelo de clasificación (si se requiere variables estandarizada, normalizada, escalonada, entre otros).

Reducción de dimensionalidad: aplicando coeficiente de correlación de Pearson y test chi-cuadrado de independencia (en software SPSS) se buscará eliminar variables redundantes que posean un alto grado de dependencia.

1.5.5. Data Mining

Se procederá a aplicar los modelos de clasificación:
- Regresión Logística Lineal Modelo Logit
- Árbol de Decisión: CHAID
- Red Neuronal Multicapa

Se optó por considerar las técnicas de Regresión Logística y Árbol de Decisión debido a que son 2 de los modelos más populares en el área de la Minería de Datos, caracterizados por su sencilla aplicación e intuitiva interpretación de los resultados, además de estar disponibles en varios de los softwares estadísticos más populares y haber presentado resultados exitosos en numerosos problemas de Data Mining. Por

otra parte se incluyó la Red Neuronal Multicapa por su capacidad de modelar funciones no lineales que podrían eventualmente representar mejor el problema abordado.

Para esto, se emplearán los programas SPSS y Excel, no sólo para aplicar los modelos, sino también para verificar que el conjunto de datos cumpla con todos los requerimientos que implica cada uno de los métodos de clasificación.

1.5.6. Evaluación de modelos

Se debe establecer de forma analítica, cuál modelo resulta ser el más adecuado de los obtenidos en la etapa de *data mining*. Para ello se utilizará la matriz de confusión:

Matriz de Confusión		PREDICCIÓN	
		NI	I
REAL	NI	A	B
	I	C	D

NI: clientes no irregulares (normales).
I: clientes irregulares.
A: número de clientes clasificados por el modelo como “normales” y que en la realidad lo son.
B: número de clientes clasificados por el modelo como “irregulares”, pero que en la realidad son normales
C: número de clientes clasificados por el modelo como “ normales”, pero que en la realidad son irregulares.
D: número de clientes clasificados por el modelo como “irregulares” y que en la realidad lo son.

Con los indicadores:

-> EFICIENCIA = $\frac{D}{B+D}$

->EFICACIA = $\frac{D}{C+D}$

La eficacia está relacionada con cuántos irregulares puede detectar el modelo con respecto a la totalidad existente, la eficiencia por otro lado se define como cuántos de los clientes vaticinados como irregulares por el modelo, en verdad lo son.

Así, el mejor modelo se determinará según sea el que minimice la suma de los costos asociados a ambos errores:

- Error Tipo I: costo de mandar a inspeccionar un cliente que no es irregular (asociado a la eficiencia).

- Error Tipo II: costo de clasificar a un cliente como “normal” cuando en la realidad es irregular (asociado a la eficacia).

1.5.7. Interpretación de resultados

Se deducen patrones lógicos y coherentes no obtenidos antes, a partir de los parámetros del modelo. De resultar exitoso este paso, se confirmará la calidad de los resultados.

1.5.8. Confección de metodología

Se diseñará un instructivo destinado a orientar de la forma más clara y automatizada posible, a los miembros de la sanitaria, en la aplicación y mantención del modelo.

1.6. ALCANCES

- El modelo de detección de fraude abarcará sólo a los clientes que presenten tarifa 11 (residenciales, comerciales e industriales), pues este tipo de cliente es el más numeroso y además posee la ventaja que lo que factura es exactamente lo que consume (a diferencia de medidores remarcadores).
- Se excluyen clientes “áreas verdes”, pues hay un plan aparte dedicado a abarcar este subconjunto.

1.7. RESULTADOS ESPERADOS

Al finalizar el trabajo de memoria se espera:

- Contar con modelo robusto y coherente, que haga uso de la potencialidad que hoy en día contienen los datos de los que se dispone.
- Contar con nuevos patrones, obtenidos a partir del trabajo, que ayuden a explicar y conocer más al segmento de clientes irregulares.
- Disponer de un procedimiento tendiente a explicar detalladamente cómo aplicar el modelo desarrollado, facilitando la tarea lo más posible.

2. APLICACIÓN MODELO KDD

2.1. DESCRIPCIÓN DE VARIABLES

Se trabajará con una base de datos compuesta por un 50% de clientes irregulares notificados (detectados desde Marzo 2011 a Diciembre 2011) y por un 50% de clientes considerados no irregulares. La base se ha compuesto bajo estos porcentajes, debido a que por ser un problema en que la variable predictora tiene 2 posibles valores (irregular o normal), ambos casos deberán tener la misma representatividad o "peso" para no obtener resultados sesgados. El tamaño corresponde a 7.256 registros (3.628 irregulares y 3.628 normales). A su vez, el 70% de la data conformará la *base de entrenamiento* (aquella que constituirá el input de los modelos y el 30%, la *base test* (utilizada para probar la calidad de los modelos generados).

Las variables de las que se dispone (de las tablas de datos de consumo y catastrales) son:

VARIABLE	TIPO	DESCRIPCIÓN
ID_CLIENTE	Categórica	Identificador del cliente
CATEGORÍA	Categórica	Tipo de cliente: residencial, comercial o industrial
SECTOR[11]	Categórica	Sector al que pertenece el cliente
ZONA	Categórica	Zonal a la que pertenece el cliente
RUTA	Categórica	N° identificador que caracteriza cliente en el proceso de toma de lectura. Los clientes vecinos presentan números de ruta correlativos (lo que no sucede con el "n° de cliente)
ESTADO	Categórica	Estado del suministro (normal, corte normal, corte especial, etc.)
NOMBRE	Categórica	Nombre del cliente
DIRECCIÓN	Categórica	Dirección del cliente
INFO_ADICION	Categórica	Información que complementa la dirección
COMUNA	Categórica	Comuna del cliente
RUT	Categórica	RUT asociado al cliente
TARIFA	Categórica	Tarifa del cliente
GIRO	Categórica	En caso de cliente comercial o industrial, giro del cliente.
TIPO DCTO	Categórica	Si cancela con boleta o factura
SALDOS	Numérica	N° de meses que adeuda el cliente
DEUDA	Numérica	Monto de dinero que adeuda cliente
TM	Numérica	Promedio consumo de m3 en últimos 6 meses
MARCA MED	Categórica	Marca del medidor
N° MEDIDOR	Categórica	N° que identifica al medidor
FECHA INS	Fecha	Fecha de instalación del medidor
DIÁMETRO	Numérica	Diámetro (en cm) del medidor

[11]El área de concesión ha sido dividida en 20 sectores geográficos, para la organización de la toma de lectura.

FECHA ENROL	Fecha	Fecha en que se registró el cliente
CLAVE LECT	Categórica	Clave de lectura de última lectura
LECT FACT	Numérica	Lectura registrada desde el medidor en último mes
POL. SOCIAL	Categórica	Si cliente reside en sector social vulnerable o no
C_i (i=1…48)	Numérica	Últimos 48 consumos
CL_j (j=1..12)	Categórica	Últimas 12 claves de lectura

Dado el conocimiento de la problemática, se debe discriminar entre aquellas variables que eventualmente pueden aportar valor en la generación de los modelos predictivos y aquéllas que no, a la vez de definir nuevas variables a partir de las existentes. Es así como las variables que caracterizarán a los clientes en la base de entrenamiento (tanto las originales como las elaboradas a partir de ellas) son[12]:

VARIABLE	TIPO	DESCRIPCIÓN
ID_CLIENTE	Categórica	Identificador del cliente
TIPO	Categórica	Corresponde a la variable dependiente. Señala si cliente es irregular o normal
SALDOS	Numérica	N° de meses que adeuda el cliente
DEUDA	Numérica	Monto de dinero que adeuda cliente
TM	Numérica	Promedio consumo de m3 en últimos 6 meses
CLAVE LECT	Categórica	Clave de lectura de última facturación
POLÍGONO_SOCIAL	Categórica	Si cliente reside en sector social vulnerable (1) o no (0)
CAÍDA	Numérica	Magnitud de la caída de consumo de mayor relevancia, en los últimos 48 meses.
CL_IRR_1	Categórica	Indica si última clave manifiesta indicio directo de irregularidad o no
CL_IRR_2	Categórica	Indica si última clave manifiesta indicio de anomalía en el medidor o no
CL_ANOR	Categórica	Indica si última clave correspondió a un proceso de lectura normal o no
CL_NO_HAB	Categórica	Indica si última clave señalaba al inmueble como deshabitado o no
CL_NORMAL	Categórica	Indica si última clave correspondió a lectura normal o no
N_CL_IRR_1	Numérica	N° de claves consideradas “irregulares” en últimos 12 meses
N_CL_IRR2	Numérica	N° de claves consideradas “anómalas” dentro de últimas 12 claves
N_CL_ANOR	Numérica	N° de claves consideradas “no efectivas” dentro de últimas 12 claves

[12]El detalle de la generación de nuevas variables se desarrolla en el apartado “creación de variables” tratado más adelante.

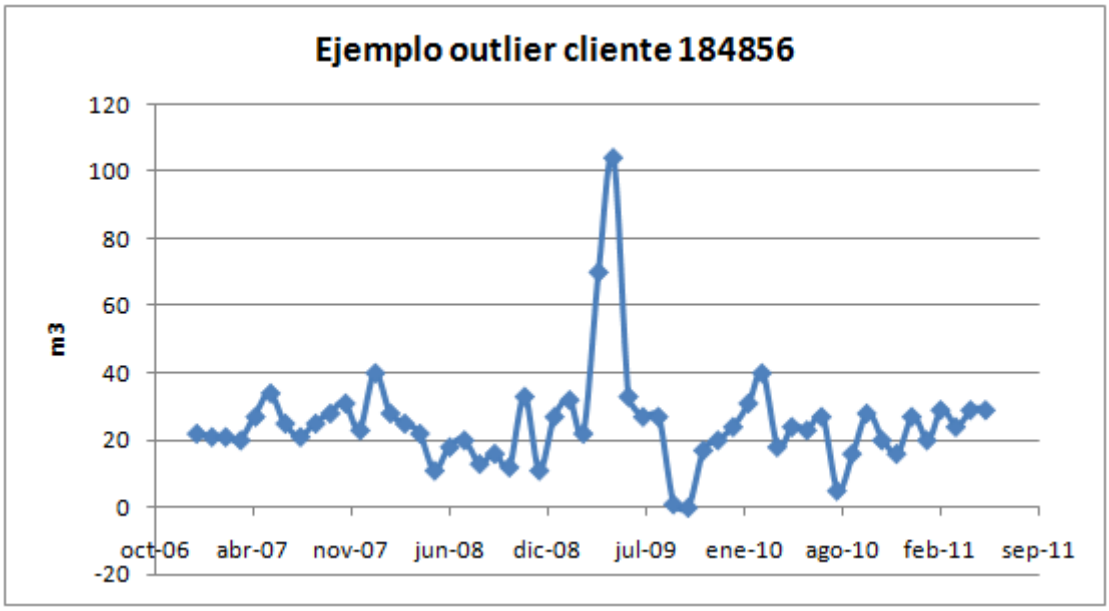
N_CL_DESH	Numérica	N° de claves consideradas “de inmueble deshabitado” dentro de últimas 12 claves
N_CL_NORM	Numérica	N° de claves consideradas “normales” dentro de últimas 12 claves
NCL_NO_NORMAL	Numérica	N° de claves distintas de “N” dentro de últimas 12 claves
DESV_CAIDA	Numérica	Magnitud de la caída de la σ de mayor relevancia, en los últimos 48 meses.
DESV_SUBIDA	Numérica	Magnitud del alza de la σ de mayor relevancia, en los últimos 48 meses.
TM_VECINO	Numérica	Promedio de consumo de últimos 6 meses de vecinos del cliente.
DENS_IRR	Numérica	Concentración de irregulares en el sector en que se sitúa el cliente
DESV_EST_1ANO_NORM	Numérica	Desviación estándar normalizada de los últimos 12 consumos mensuales
DESV_EST_6MES_NORM	Numérica	Desviación estándar normalizada de los últimos 6 consumos mensuales
CEROS_ULT12	Numérica	% de consumos nulos en últimos 12 meses

2.2. Pre-procesamiento de datos

2.2.1. Casos Outliers

En general la base de datos generada por el sistema de la empresa es bastante sólida, y por ende, no presenta errores en el valor de las variables. Sin embargo, con respecto a los consumos mensuales, hay ciertos valores que pueden distorsionar la realidad sobre los hábitos de consumo de un cliente. Tal es el caso de consumos que se elevan notoriamente sobre la tendencia general, correspondiendo generalmente a casos de fugas no controladas, como se ejemplifica a continuación:

Figura N° 7: Cliente presenta outliers en Abril y Mayo de 2009

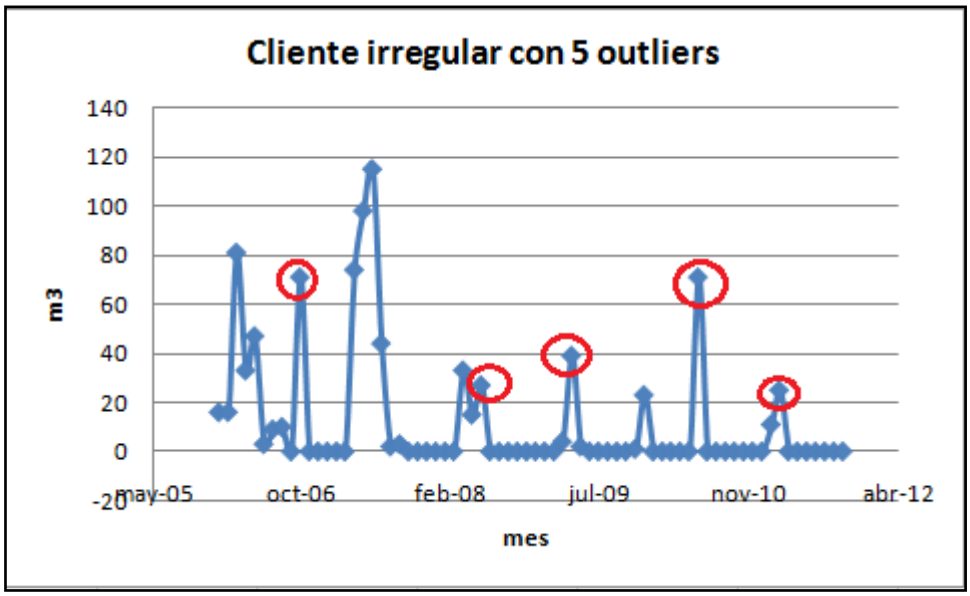


Para decidir la regla que se aplicará a estos casos se creó un pequeño programa [11] en Visual Basic destinado a detectar outliers. Un consumo se considerará "fuera de rango" si es superior en 8 veces al consumo promedio de los 6 meses aledaños.

El resultado luego de la aplicación fue un total de 335 clientes con al menos un outlier:

TIPO DE CLIENTE	N° DE CLIENTES CON OUTLIER EN CONSUMO
I	228
N	107
TOTAL	335

Como se aprecia en la tabla, del total de clientela que presentaba al menos 1 consumo fuera de rango, un 68% corresponde a irregulares. De esto se deduce que probablemente muchos consumos anormales pueden ser indicio de intervención, por lo que "emparejar" las cifras puede resultar perjudicial, pues se estaría eliminando información potencialmente útil. A continuación, un cliente irregular con diversos consumos "fuera de rango":



En aquellos casos en que el cliente posee 1 solo consumo "fuera de rango" dentro de su historial de facturaciones, es más probable que dicha situación se explique por una fuga de agua, por lo que se trataron sólo aquellos casos que presentaban únicamente 1 outlier en los últimos 48 meses, los que derivó en 210 detecciones. En estos casos, para "normalizar" el consumo, se procedió a reemplazar el valor fuera de rango por el promedio de los 6 consumos más cercanos a esa fecha.

2.2.2. Casos Missings

- Missings variable TM_VECINO: la variable que representa el consumo promedio de los vecinos más cercanos, presenta 588 casos con valores perdidos. Esta variable se confeccionó al promediar el consumo de los clientes vecinos incluidos dentro de una circunferencia imaginaria de 30 metros de radio trazada alrededor del cliente. Bajo este método, hubieron 588 clientes que no contaban con un vecino dentro de ese radio de alcance, lo que explica los "missings values". A estas entidades se les imputará a la variable TM_VECINO la magnitud de **23,2**, que corresponde al consumo promedio mensual de agua de un cliente de la sanitaria.

- Missings variable POLÍGONO_SOCIAL: esta columna presenta 3 entidades con valores ilógicos, se les imputará la moda que corresponde al valor 1 (sí es polígono social).

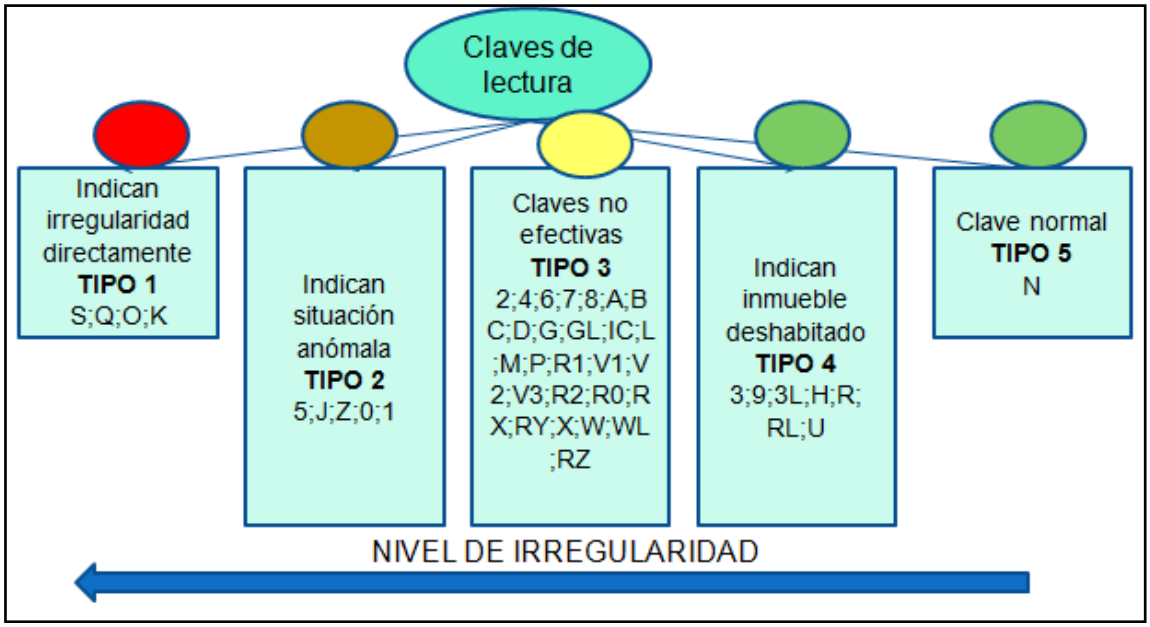
- Missings variable CL_i (Clave número i): se registran 4 entidades con valores ilógicos para estas variables. Se procederá a eliminarlas de la base.

2.2.3. Creación de variables

A partir de las 67 variables de las que se disponen, se procedió a crear otras que potencialmente pueden aportar en la generación del modelo de detección. A continuación se describe su proceso de generación.

- Variables relacionadas con últimas 12 claves de lectura.

Se procedió a dividir el universo de 44 claves en 5 grupos, de acuerdo a su grado de relación con la intervenciones fraudulentas; siendo las de tipo 1 la que indican directamente que se está ante un ilícito; las de tipo 2 y 3 aquellas que indican probable fraude o proceso no normal de lectura; las de tipo 4 que indican inmueble no ocupado y del tipo 5 que refleja proceso normal de toma de lectura.



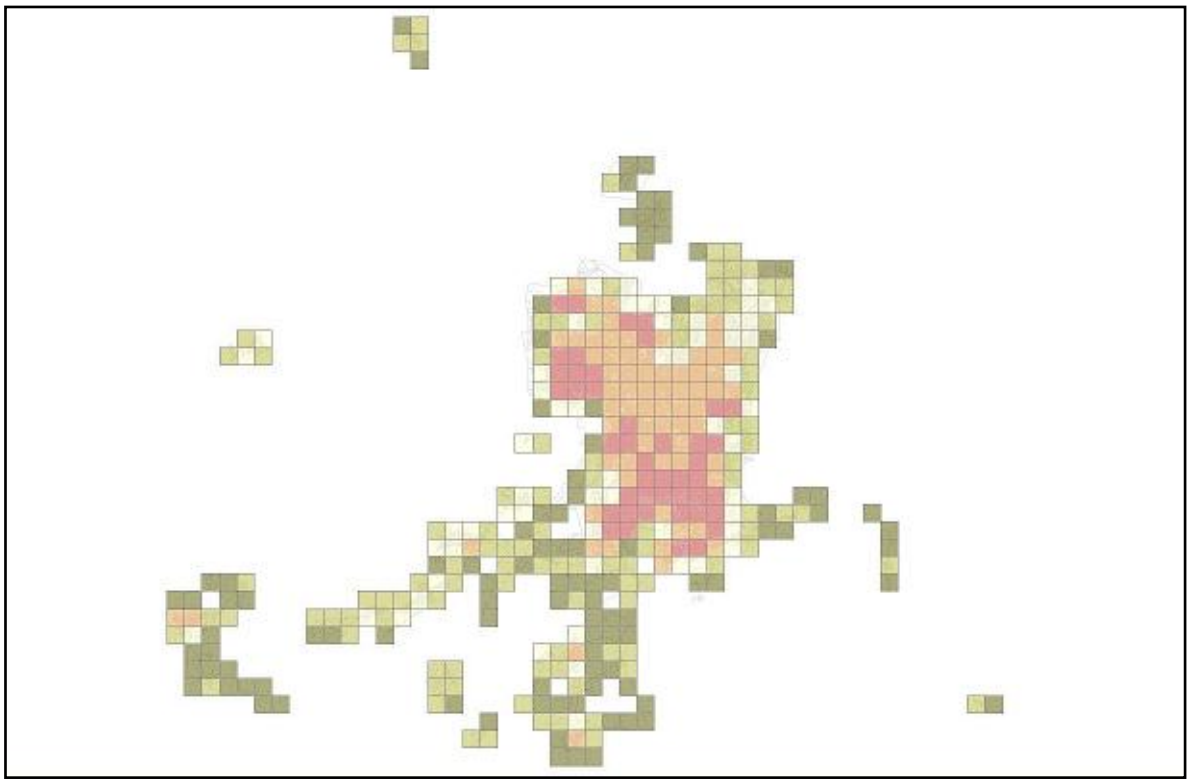
En base a esta clasificación se generaron los atributos:

- CL_IRR_1: del tipo binaria, indica si última clave es del tipo 1 o no.
- CL_IRR_2: del tipo binaria, indica si última clave es del tipo 2 o no.
- CL_ANOR: del tipo binaria, indica si última clave es del tipo 3 o no.
- CL_DESH: del tipo binaria, indica si última clave es del tipo 4 o no.
- CL_NORMAL: del tipo binaria, indica si última clave es del tipo 5 o no.
- N_CL_IRR_1: cuántas de las últimas 12 claves son del tipo 1.
- N_CL_IRR_2: cuántas de las últimas 12 claves son del tipo 2.
- N_CL_ANOR: cuántas de las últimas 12 claves son del tipo 3.
- N_CL_DESH: cuántas de las últimas 12 claves son del tipo 4.
- N_CL_NORM: cuántas de las últimas 12 claves son del tipo 5.
- N_CL_NO_NORMAL: cuántas de las últimas 12 claves no son N (normal).

- Variable DENS_IRR

Para obtener la densidad de irregulares (N° clientes notificados / N° total de clientes) de cada sector de la zona de concesión, se procedió a dividir el terreno en cuadrículas de 1.000 metros de ancho, mapeando en ellas (utilizando el software ARCMAP) los notificados históricos registrados desde el año 2006.

Figura N°10: Mapeo de concentración de irregulares por sector



Fuente: Software ARCMAP

De este procedimiento se obtuvo la concentración de irregulares para cada sector (zonas con tono rojizo presentan mayor densidad), pudiendo utilizar esta información para conocer el comportamiento fraudulento en cada sector en que se localizan los clientes de la base de datos.

- Variable TM_VECINO

Para conocer el consumo de los vecinos más cercanos a un determinado cliente, se trazó una circunferencia imaginaria alrededor del mismo (de 30 metros de radio), procediendo a promediar el consumo de todos los clientes que se incluyeran dentro de la figura geométrica (considerando sólo tarifas 11 y excluyendo áreas verdes). Se tomó un radio de 30 metros pues con esta distancia se consiguió en la mayoría de los casos incluir a los vecinos considerados “cercanos” (generalmente 6 ó 7) sin integrar demasiados clientes dentro de la circunferencia.

- Variables relacionadas con la desviación estándar de los consumos

Los atributos DESV_CAIDA (mayor descenso registrado de la σ), DESV_SUBIDA (mayor aumento registrado de la σ), DESV_EST_1ANO_NORM (σ de los últimos 12 meses normalizada por el promedio) y DESV_EST_6MESES_NORM (σ de los últimos 6 meses normalizada por el promedio) fueron generados a partir de códigos de programación en Visual Basic.

2.2.4. Reducción de variables

Se procedió a estudiar las correlaciones entre los distintos atributos cuantitativos, siendo las más relevantes:

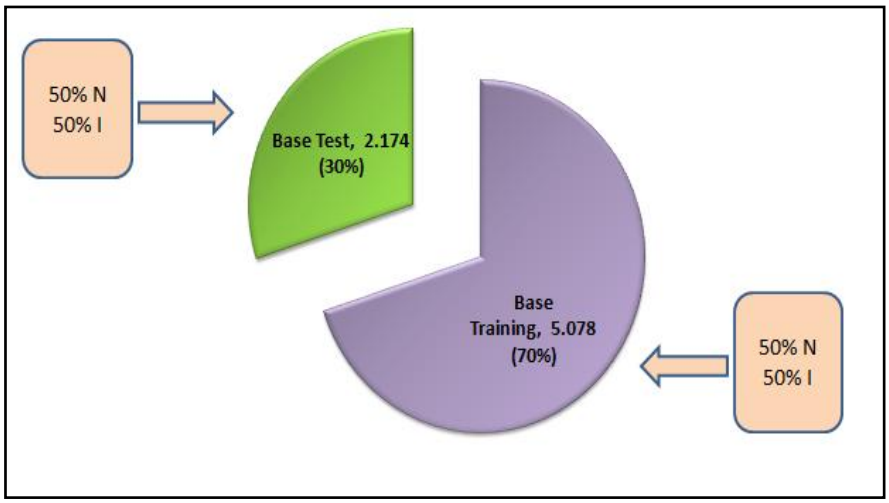
Variable 1	Variable 2	Correlación
Deuda	Término Medio	0,852
N_CL_NORM	N_CL_NO_NORMAL	-1
CL_IRR_ANOR	CL_NORMAL	-0,939
N_CL_NO_NORMAL	N_CL_ANOR	0,859

Debido a estas altas correlaciones, se decidió eliminar las variables DEUDA, N_CL_NORM, CL_NORMAL y N_CL_ANOR, debido a que son muy bien explicadas por las otras.

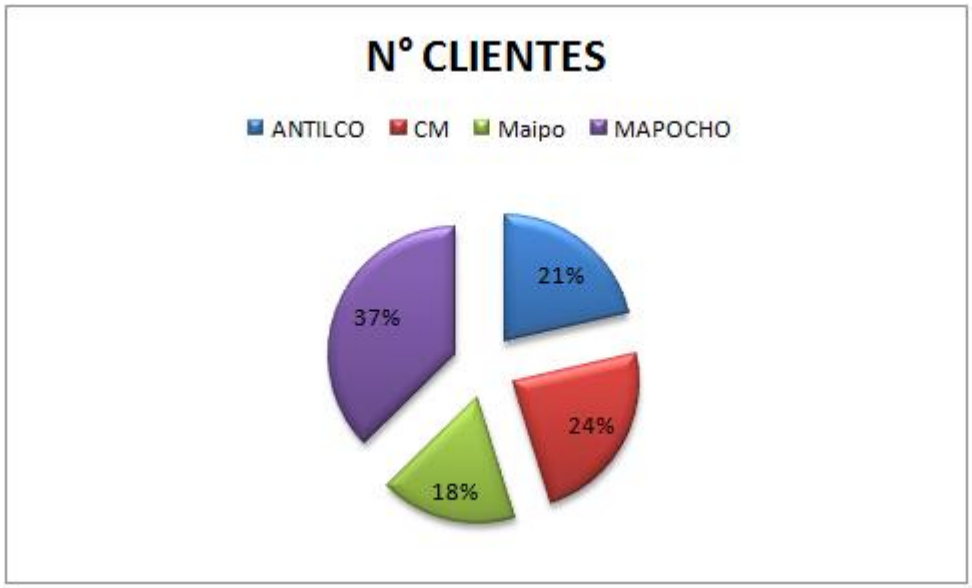
2.3. DESCRIPCIÓN BASE DE DATOS

a. Base Test y Base Training

Después del pre-procesamiento de datos **7.252** (3.626 normales y 3.626 irregulares) entidades conforman la base de datos, siendo distribuidas de la siguiente manera:



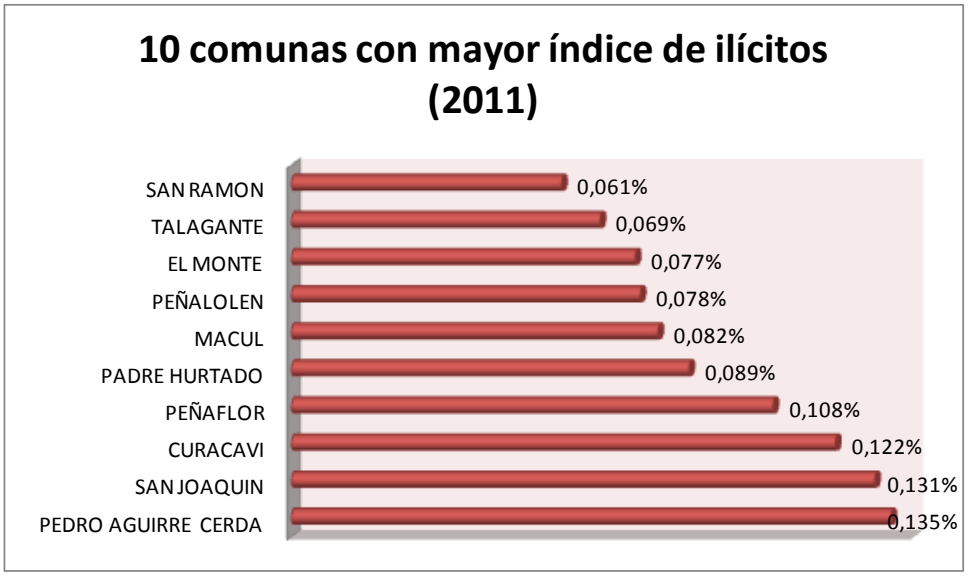
b. Representación de cada zona en Base de Datos



c. Comunas con mayor índice de ilícitos

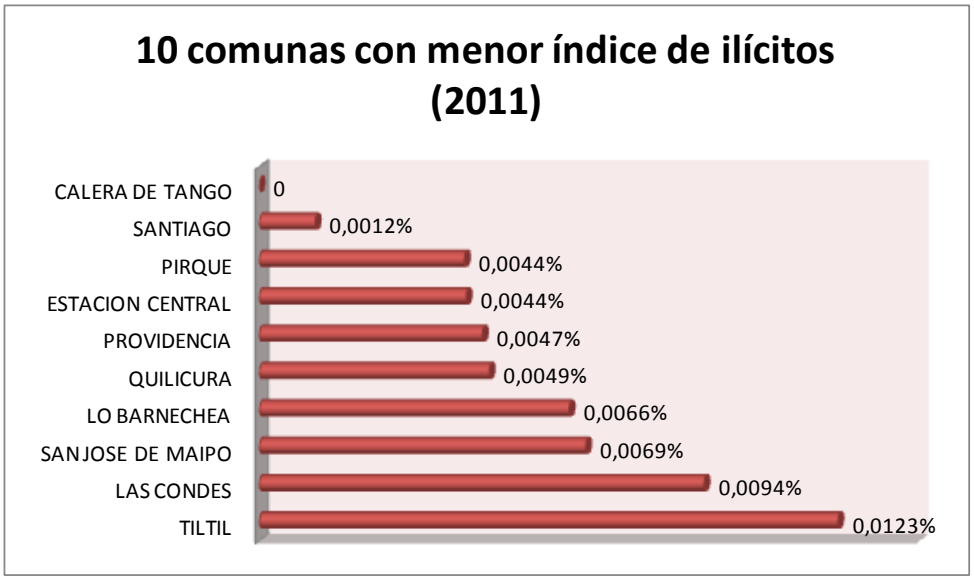
Si se entiende por índice de ilícito de determinada comuna el número de notificados dividido por la cantidad total de clientes enrolados (sólo considerando Tarifa 11), se obtiene que la comuna con el índice más alto de intervenciones es Pedro Aguirre Cerda (alrededor de 0,13%), seguida de San Joaquín y Curacaví.

$$Índice\ de\ ilícito = \frac{N°\ notificados}{N°\ total\ de\ clientes}$$



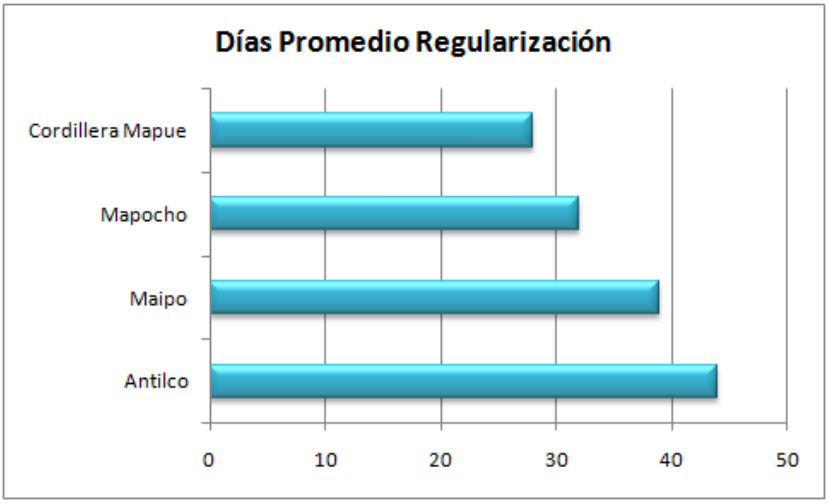
d. Comunas con menor índice de ilícitos

Siguiendo con el índice de ilícito, pero ahora analizando la cara opuesta, vale decir, las comunas con menor valor de este índice, se tiene que aquélla en que se ha detectado el menor grado de irregularidad corresponde a Calera de Tango (0%) seguida de Santiago (0,0012%) y Pirque (0,0044%).



e. Días promedio de regularización

El promedio de días que un cliente notificado tarda en normalizar su situación, para cada zona (es decir desde el día en que es notificado hasta el día en que firma el acuerdo de regularización en recintos de Aguas Andinas), es el siguiente:



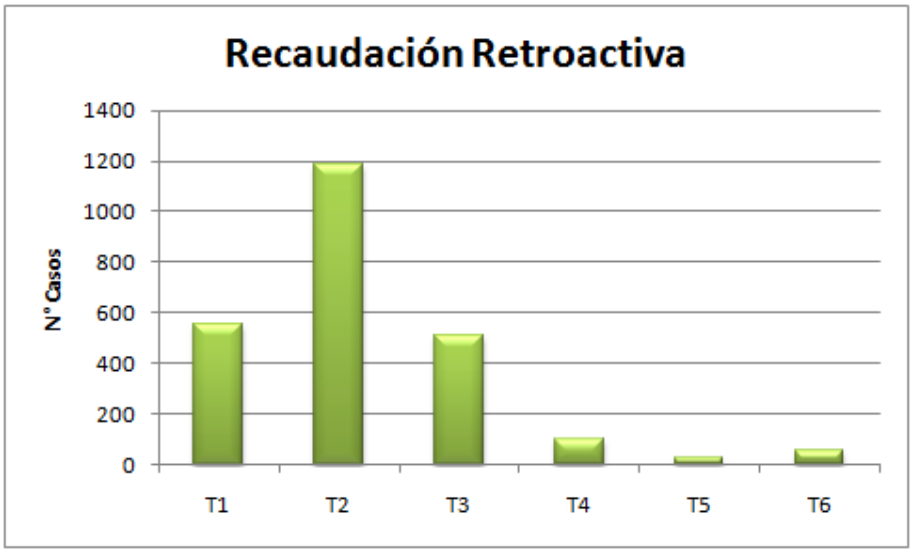
El promedio ponderado para toda la sanitaria es de **35,4 días.**

f. Recaudación retroactiva

Si se analiza el componente de la RPCNR correspondiente a la recaudación retroactiva, y dividimos los distintos niveles de pago en:

- T1 entre $0 y $100.000
- T2 entre $100.001 y $200.000
- T3 entre $200.001 y $300.000
- T4 entre $300.001 y $400.000
- T5 entre $400.001 y $500.000
- T6 mayor a $500.000

La tendencia se concentra en el tramo 2, como se aprecia en la figura:



2.4. OBTENCIÓN DE MODELOS DE APRENDIZAJE

Los modelos de aprendizaje supervisado que se aplicarán a la base training son:

- Regresión Logística Binaria
- Árbol de clasificación CHAID
- Red Neuronal Perceptrón multicapa

Estos 3 modelos se correrán en el software SPSS 19.

2.4.1. Aplicación de Regresión Logística Binaria

El output obtenido luego de aplicar el modelo es:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Saldos	,003	,009	,128	1	,721	1,003
	TerminoMedio	,000	,001	,507	1	,476	1,000
	PoligonoSocial	-,014	,070	,037	1	,847	,987
	CAIDA	2,937	,153	367,684	1	,000	18,853
	CL_IRR1	,529	1,148	,212	1	,645	1,697
	CL_IRR_2	3,664	,656	31,213	1	,000	39,026
	CL_IRR_ANOR	-,356	,090	15,515	1	,000	,700
	CL_IRR_DESH	1,298	,565	5,274	1	,022	3,663
	N_CL_IRR_1	-1,126	,264	18,190	1	,000	,324
	N_CL_IRR_2	-1,228	,127	92,964	1	,000	,293
	N_CL_DESH	,195	,064	9,395	1	,002	1,215
	N_CL_NO_NORMAL	-,214	,017	162,481	1	,000	,807
	DESV_Caida	,117	,097	1,452	1	,228	1,124
	DESV_SUBIDA	-,007	,003	5,050	1	,025	,993
	TM_BUFFER	-,002	,002	1,013	1	,314	,998
	DENS_IRR	-,705	1,219	,335	1	,563	,494
	DESV_EST_1ANO_NORM	,241	,186	1,687	1	,194	1,273
	DESV_EST_6MES_NORM	-,273	,181	2,278	1	,131	,761
	CEROS_ULT12	1,868	,445	17,621	1	,000	6,476
	N_OUTLIERS	,075	,164	,208	1	,648	1,078
	Constante	-1,194	,138	75,021	1	,000	,303

Analizando los p-valores arrojados para cada variable independiente, aquellos atributos significativos para el modelo (por tener un p-valor menor o igual a 0,05) son:

- CAIDA
- CL_IRR2

- CL_IRR_ANOR
- CL_IRR_DESH
- N_CL_IRR1
- N_CL_IRR2
- N_CL_DESH
- N_CL_NO_NORMAL
- DESV_SUBIDA
- CEROS_ULT12

Luego, la ecuación de regresión logística resulta ser:

$$P(cliente = normal) = \frac{1}{1 + \exp\left(1{,}194 - 2{,}937 * CAIDA - 3{,}664 * CL_{IRR2} + 0{,}356 * CL_{IRR_{ANOR}} - 1{,}298 * CL_{IRR_{DESH}} + 1{,}126 * NCL_{IRR1} + 1{,}128 * NCL_{IRR2} - 0{,}195 * NCL_{DESH} + 0{,}214 * NCL_{NONORMAL} + 0{,}007 * DESV_{SUBIDA} - 1{,}868 * CEROS_{ULT12}\right)}$$

Los signos negativos asociados a CAIDA, CL_IRR2, CL_IRR_DESH, NCL_DESH y CEROS_ULT12, indican que a mayor magnitud de estas variables, la probabilidad del cliente no sea irregular es mayor. Con respecto a la variable "caída", se debe pensar que se define como consumo post-caída/consumo pre-caída, por lo que la mayor magnitud de esta variable significa que el descenso de consumo es más bajo, por ende, mayor probabilidad que no sea irregular.

Si se establece a 0,5 como punto de corte (se clasifica al cliente como irregular si probabilidad es menor o igual a 0,5), las magnitudes asociadas a la base training son:

		PRONOSTICADO	
		N	I
REAL	N (2539)	2072	467
	I (2539)	666	1873

2.4.2. Aplicación árbol de decisión

Se aplicó el árbol con el algoritmo CHAID[13], resultando como variables importantes del modelo (situadas en los nodos superiores):

- CAIDA
- TERMINO MEDIO
- N_CL_NONORMAL
- N_CL_IRR2

La principal variable de clasificación es CAIDA (en el nodo 0), separando el conjunto de datos en 7 subconjuntos, siendo cada uno de ellos explicados a continuación:

- Subconjunto 1: CAIDA ≤ 0,225

Para este grupo de entidades que presentan la caída en consumo más alta, la mayoría es clasificada como irregular. Se observa que a mayor término medio, mayor probabilidad de que estén incurriendo en el ilícito, vale decir, que un cliente que generalmente consuma un alto volumen de m3 al mes y haya experimentado una baja súbita de consumo, es más probable que sea un irregular en comparación a aquel cliente que experimentó una caída brusca, pero que suele consumir en cantidades menores.

- Subconjunto 2: 0,225 < CAIDA ≤ 0,351

Para este grupo también se cumple que clientes con mayor término medio (mayor a 6 m3) presentan mayor probabilidad de tener conexión fraudulenta que aquellos con menor término medio. A su vez, de los servicios con mayor consumo, los más sospechosos resultan ser a su vez los que presentan 3 o más claves distintas de N (clave normal) en los últimos 12 meses, con una probabilidad del 98% de ser fraudulentos.

- Subconjunto 3: 0,351 < CAIDA ≤ 0,463

Dentro de esta categoría, los clientes con mayor término medio son más propensos a ser fraudulentos; los que no (con término medio menor o igual a 15 m3) son sub-clasificados de acuerdo a la cantidad de claves no normales en los últimos 12 meses, a mayor número de estas claves, aumenta la probabilidad.

- Subconjunto 4: 0,463 < CAIDA ≤ 0,575

En este caso los clientes con un TM menor o igual a 15 m3 no se pueden clasificar con certeza (pues poseen probabilidades cercanas al 50%), la mejor sub-clasificación que se logra es ver quiénes presentan deuda 0 (saldos 0), pues éstos son clasificados como “normales” con un 66,2% de certeza. Con respecto a aquellos clientes con un TM entre 15 y 24, quienes presenten un alza de desviación estándar relativa mayor a 1,88 son considerados irregulares con un 82% de confianza. Por último, de aquellos servicios con TM mayor a 24 m3, los que registren 4 ó más claves distintas a “N” constituyen casos de fraude con un 93% de confianza.

[13]Output disponible en Anexos C.

- Subconjunto 5: 0,575 < CAIDA ≤ 0,681

En este grupo ya no se discrimina según término medio, sino en base al número de claves distintas de N en los últimos 12 meses (variable N_CL_NO_NORMAL). Aquellos casos con sus 12 claves de lectura igual a N son “no irregulares” con un 79% de probabilidad. Aquellos cuya variable es mayor a 4, son irregulares con un 73% de probabilidad.

- Subconjunto 6: 0,681< CAIDA ≤ 0,771

Estos casos también son sub-clasificados en base a la variable N_CL_NO_NORMAL. Los casos más fiables son aquellos clientes que presentan a lo más 1 clave distinta de N, pues son “normales” con un 76% de certeza (y de ellos los que no poseen deuda lo son con un 93% de certeza).

- Subconjunto 7: 0,771< CAIDA

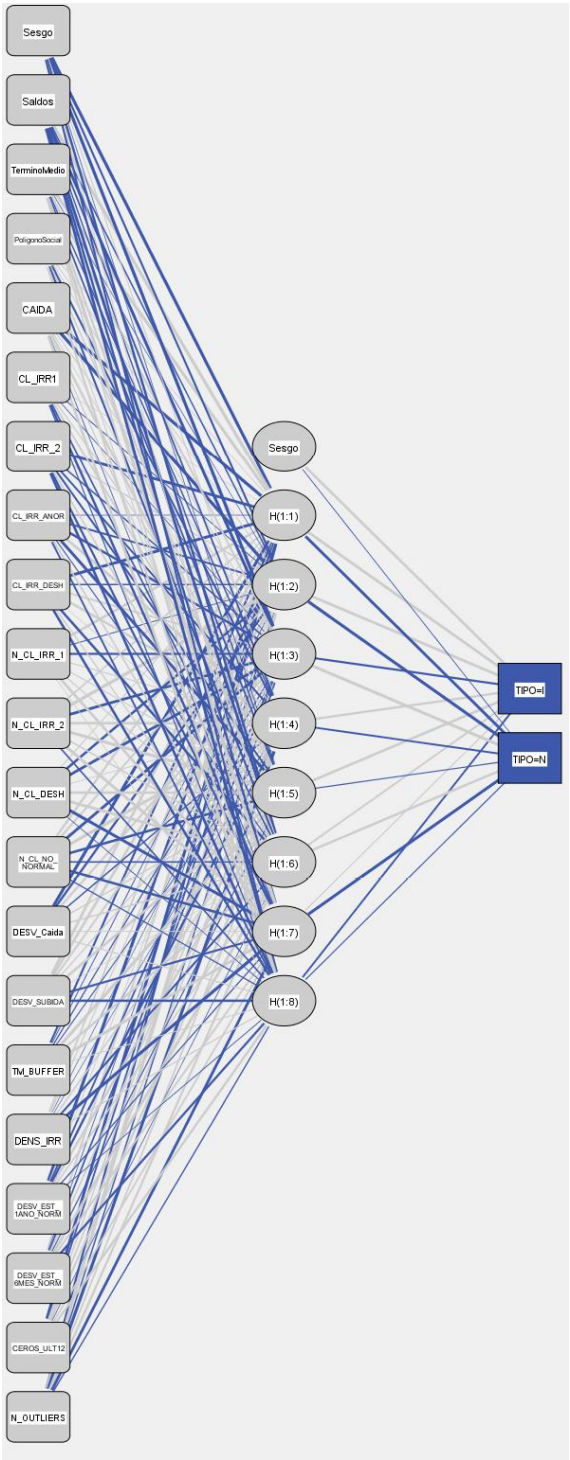
De estos clientes, que son los que poseen las menores caídas de consumo, los que tienen más probabilidad de ser fraudulentos son aquellos que registran 1 ó más claves TIPO 2 en el último año (con un 82% de certeza).

Manteniendo el punto de corte de 0,5 las magnitudes asociadas a la base training son:

		PRONOSTICADO	
		N	I
REAL	N (2539)	2078	461
	I (2539)	674	1865

2.4.3. Aplicación Red Neuronal

El output del programa fue:



Vale decir, una red con una capa oculta de 8 nodos más el sesgo, cuya función de activación corresponde a la ***tangente hiperbólica***

Los resultados fueron:

		PRONOSTICADO	
		N	I
REAL	N (2539)	2077	462
	I (2539)	612	1927

2.5. APLICACIÓN DE MODELOS A BASE TEST

Al aplicar los 3 modelos obtenidos a la base test, de 2.174 entidades, constituida por un 50% de clientes irregulares y un 50% de no irregulares, lo resultados obtenidos son:

a. Aplicación Regresión Logística

		PRONOSTICADO	
		N	I (1000)
REAL	N (1087)	865	222
	I (1087)	309	778

- Eficacia = qué % del universo real de clientes irregulares puede detectar = 71,6%
- Eficiencia= qué % del conjunto que predice como irregular, de verdad lo es = 77,8%

b. Aplicación Árbol de Decisión

		PRONOSTICADO	
		N	I (995)
REAL	N (1087)	907	180
	I (1087)	272	815

- Eficacia = qué % del universo real de clientes irregulares puede detectar = 75 %
- Eficiencia= qué % del conjunto que predice como irregular, de verdad lo es = 81,9%

c. Aplicación Red Neuronal

Los resultados sobre la base test fueron:

		PRONOSTICADO	
		N	I (984)
REAL	N (1087)	894	193
	I (1087)	296	791

- Eficacia = qué % del universo real de clientes irregulares puede detectar = 72,8%
- Eficiencia= qué % del conjunto que predice como irregular, de verdad lo es = 80,4%

2.6. ANÁLISIS DE RESULTADOS

Básicamente se tienen 2 tipos de costos asociados a un determinado tipo de error:

- Error Tipo I: Predecir un irregular como normal= $456.000 (se pierde la RPCNR).
- Error Tipo II: Predecir un normal como irregular = $6.000 (costo base de la justificación, monto cobrado por el contratista por realizar pruebas técnicas).

Claramente al contrastar las cifras se deduce que un tipo de error pesa mucho más que el otro, por lo que en términos monetarios conviene evaluar los modelos según su eficacia, vale decir, cuántos de los irregulares que en verdad lo son, es capaz de identificar. Este criterio es aplicable a esta problemática debido a que las eficiencias de los 3 modelos son bastante similares (bordea el 80%), por lo que se tendrá en los 3 casos un costo similar asociado a error tipo II, vale decir a inspecciones innecesarias, que en virtud del contexto, su costo se puede establecer como constante (marginalmente varía muy poco en comparación a la variación del error tipo II).

En base a este criterio, el mejor modelo corresponde al **árbol de decisión,** con un 75% de eficacia, si bien las cifras para los otros modelos son bastante similares (71,6% para la regresión logística y 72,8% para la red neuronal). Es así cómo se aprecia que

dicho modelo logra identificar la mayor cantidad de irregulares existentes en la base test (815 de un total de 1.087).

Con respecto a la elevada tasa de detección del árbol de decisión, de un 81,9%, se debe recalcar que ésta corresponde a una tasa teórica, vale decir, la que obtendría en condiciones ideales. En la práctica hay factores que seguramente menguaran este índice, tales como:

- Inspecciones no efectivas: hoy en día se cuenta con un 33% de inspecciones que no pueden ser concretadas por diversos motivos. Un supuesto válido es que muchos de esos clientes se niegan a ser fiscalizados debido a que han intervenido su suministro de agua potable, por ende, se pierden potenciales notificaciones exitosas.
- Calidad inspección: las pruebas de agua realizadas hoy en día son bastante acuciosas en la mayoría de los casos. Sin embargo, es altamente probable que algunos inspectores realicen mejor el análisis de las instalaciones que otros, detectando con mayor frecuencia aquellos casos en que la intervención es más elaborada (por ejemplo clientes con un *by-pass* cuidadosamente instalado bajo tierra).
- Potenciales conductas no éticas por parte de los trabajadores: si bien se ha depositado toda la confianza en los inspectores, no se puede tener la certeza de que éstos no incurran en prácticas poco legítimas como el llegar a un acuerdo con el cliente notificado. Esporádicamente se cae en conocimiento de casos puntuales de estas características, por lo que una visita que debió resultar en notificación, pasa a ser un caso “justificado”.

Para comparar en términos monetarios el modelo generado y el actualmente utilizado, se deben contrastar los índices de eficiencia de cada uno, pues el índice de eficacia del actual modelo es incierto. La eficiencia del modelo convencional es de un 15%, mientras que para el modelo generado se adquirirá una postura bastante conservadora considerando un índice del 40%. Esto debido a los factores mencionados anteriormente que pueden menguar la tasa teórica de 81,9%. Por otra parte se supondrá que se mantienen los tamaños actuales de cartera enviadas a los contratistas, así como la tasa de inspecciones efectivas y la de regularización vigentes. Bajo estas consideraciones, las potenciales ganancias proyectadas al mes son:

INDICADOR	MÉTODO ACTUAL	MÉTODO ÁRBOL DE DECISIÓN
Tasa Eficiencia	15%	40%
Inspecciones Efectivas	2.400	2.400
Clientes notificados	360	960
Clientes regularizados	198	528
Recuperación RPCNR	$90,3MM	$240MM

Al comparar ambos ingresos, se está ante una recuperación extra mensual de $150MM.

Con respecto a las variables consideradas por los métodos aplicados, se pueden analizar con propiedad aquellas de los casos relativos a la regresión lineal y árbol de decisión, no así la red neuronal, debido a que el output arrojado no permite discernir en forma explícita cuáles son los atributos considerados (por la naturaleza de la red neuronal de “caja negra”). Por ende, sólo revisando los outputs de la regresión y el árbol, se observa que los 2 consideran generalmente los mismos atributos, coincidiendo en el descarte de las variables TM_VECINO, DENS_IRR y POLÍGONO SOCIAL. Esta información resulta vital para concluir que las características del lugar en que se sitúa el cliente (el histórico de irregulares que presenta su manzana, el consumo de sus vecinos o la situación social) no influyen de manera determinante en la probabilidad de estar ante un fraude.

Por último, al relacionar el número de clientes irregulares que se estima existe actualmente (alrededor de 24.000 según el cálculo de la página 9) con la tasa de detección del modelo obtenido, que con la postura conservadora deriva en la detección de 960 casos mensuales, un cálculo a priori arrojaría que el total irregulares existente sería identificado en poco más de 2 años. Esto sin embargo no debe considerarse como la suposición de que en 2 años no habrá más clientes irregulares; el número de clientes fraudulentos hoy en día puede considerarse una unidad de stock, que cuenta con un flujo de entrada (nuevos servicios fraudulentos y reincidencia de clientes ya notificados) y un flujo de salida (servicios notificados y regularizados). Sumado a lo anterior, ningún modelo posee la eficacia del 100%, por lo que siempre quedarán conexiones irregulares que probablemente jamás serán detectadas por la empresa. Es así como se puede tener la certeza de que a lo largo del tiempo siempre habrá un volumen considerable de servicios intervenidos.

3. METODOLOGÍA DE APLICACIÓN

En base a los resultados obtenidos y ya contado con un modelo de detección teórico, ahora se proponen los siguientes pasos que debieran seguirse cada mes (considerando ciclos mensuales de aplicación de cartera) para su correcta aplicación.

- Paso 1: Elegir segmento de clientes a analizar

Debido al gran número de clientes de la sanitaria, actualmente se hace poco viable analizarlo computacionalmente en su totalidad, debido principalmente a que no existe una vía expedita y práctica para bajar antecedentes de cada cliente (la bajada de datos cada mes requiere bastante tiempo), tardando un día completo en obtener antecedentes de 200.000 clientes. Por ellos cada mes se debiera sugerir zonas y/o comunas de interés para confeccionar la cartera.

Experimentalmente se determinó, en base a la aplicación del modelo en 3 zonas escogidas al azar, que clasifica irregulares a alrededor del 5% del total de clientes utilizado como input, por lo que para mantener el actual tamaño de cartera solicitado por los contratistas (1.000 para Antilco, Mapocho y Cordillera; y 500 para Maipo) se debería analizar un total de 70.000 clientes tarifa 11 al mes.

- Paso 2: Adaptación de la base de datos.

Una vez bajados los datos debe crearse la base de datos estándar, con los nombres de las variables idénticos a los mencionados, y creando las variables descritas anteriormente por medio de las macros creadas. Al haber sido seleccionado el Árbol de Clasificación CHAID, las variables a extraer o crear son:

- CAIDA
- Término Medio
- N_CL_NO_CNORMAL
- N_CL_IRR_2
- N_CL_DESH
- CL_IRR_ANOR
- Saldos
- DESV_SUBIDA

A su vez, se deberán detectar missings y tratar los outliers relativos a consumos como se concluyó en los puntos 2.2.1. y 2.2.2.

- Paso 3: Uso del software

El modelo de árbol de decisión ha sido guardado para su fácil aplicación en SPSS, por lo que una vez importada la base basta con aplicar el algoritmo, y automáticamente se obtendrá la probabilidad que tiene cada cliente de ser un irregular.

- Paso 4: Generación cartera

La cartera oficial de cartera de irregulares para cada zona estará compuesta por aquellos clientes con probabilidad mayor a 0,5 de ser irregulares, además de las denuncias de vecinos que poseen alta tasa de credibilidad (suman cerca de 200 al mes en total). Dicha cartera será otorgada a principios de mes a cada contratista.

- Paso 5: Retroalimentación

Será obligación para cada contratista tener a más tardar a principios del mes siguiente el reporte de todas las visitas realizadas. Se sugiere utilizar dichos resultados para mejorar cada 6 meses el modelo de aprendizaje, construyendo nuevamente el árbol CHAID con los antiguos datos y la nueva data recopilada.

Cabe mencionar que el responsable de estos 5 pasos deberá ser algún miembro de la Unidad de Gestión de la Medición, perteneciente a la Gerencia Corporativa del Servicio al Cliente.

4. CONCLUSIONES

El problema del agua no facturada por concepto de conductas ilícitas es una preocupación constante para todas las sanitarias del mercado. Si bien cada mes se detectan alrededor de 360 clientes irregulares, al mismo tiempo se van produciendo nuevas conexiones fraudulentas cada día.

Se hace urgente en la sanitaria formalizar y automatizar este proceso, debido a que cada mes las carteras no son generadas bajo un criterio fijo, y por ende las tasas de detección son variables e inciertas. Así, mejorar la gestión de la detección representa una gran oportunidad en cuanto a disminuir las pérdidas económicas que se producen por el fraude, teniéndose que una mejora del 1% en la tasa de detección implica un aumento de alrededor de $6MM en el monto mensual recaudado.

Con respecto a los resultados obtenidos, que los 3 modelos hayan arrojado indicadores de rendimiento tan similares, indica que se está ante un problema coherente cuyas características hacen que pueda tener una solución confiable en el área de *data mining*.

Con respecto al tamaño óptimo de la cartera mensual que se acuerde con el contratista se sugiere comenzar a generar listados de sospechosos con el modelo aquí escogido, bajo las indicaciones explicitadas en la metodología propuesta, y esperar a que se estabilice la tasa de detección, debido a que es esperable que en un principio las inspecciones exitosas aumenten (pues el modelo arrojará aquellos casos con mayor probabilidad), pero pasado un tiempo se comenzarán a agotar aquellos clientes con probabilidades más altas, experimentando un descenso la tasa de detección para estabilizarse en cierto punto.

Un aspecto importante a destacar es que, refiriéndose a las variables que utiliza el modelo escogido, aquellas de tipo “geográfica” (que hacen referencia al cliente y su entorno) no resultan ser preponderantes o influyentes en el modelo. Así las variables TM_VECINO (que muestra el consumo promedio de los vecinos incluidos en cierta circunferencia alrededor del cliente) y DENS_IRR (que señala el grado de fraude detectado en determinado cuadrante en que se sitúe el cliente) no resultaron ser útiles en el Árbol Clasificatorio, así como tampoco en el modelo obtenido por Regresión Lineal. Esto resulta positivo de cara a la implementación del modelo, debido a que la construcción de estas variables requiere un esfuerzo no menor al tener que mapear los datos en cierto software y calcular las magnitudes descritas. De este modo, mes a mes, se evitará incurrir en esa inversión de tiempo.

Sin duda se hace indispensable la revisión y mejora continua de estos modelos cada cierta cantidad de tiempo; el incorporar datos nuevos e ir retroalimentándose del feedback de su aplicación, sin duda enriquecerá los resultados.

Aunque el mejor modelo arroja un 81,9% de eficiencia, versus un 15% que presenta el método actual, ambas cifras no son del todo comparables debido a que el 15% es un valor real y comprobado en la práctica, en cambio el 81,9% refleja una cifra teórica que de seguro experimentará una baja cuando se comience a aplicar en terreno (debido a

que el contratista no fue capaz de identificar el fraude, las pruebas de agua fallaron, entre otros motivos). Sin embargo, al contrastar ambos valores, obviamente se espera que el nuevo método arroje resultados notoriamente mejores a los actuales.

En un futuro cercano (cuando se tenga una mayor cantidad de datos) resultaría bastante interesante confeccionar un modelo para cada zonal, de modo de tener parámetros más personalizados en cada caso. Esto, debido a que cada zonal posee ciertas características marcadas, por ejemplo, en Maipo abundan las zonas rurales, mientras que en Cordillera Mapué se caracteriza por concentrar las viviendas de clientes más acomodados. Dicha heterogeneidad entre zonales se manifiesta por ejemplo al visualizar datos históricos, de los que se deduce que constantemente Antilco y Mapocho presentan las tasas de detección más altas (alcanzando a veces el 25%), mientras que Cordillera Mapué y Maipo las menos altas (alcanzando algunos meses el 12%).

Una parte primordial del proceso consiste en la regularización por parte del cliente notificado, correspondiendo al momento en que éste acuerda con Aguas Andinas la forma de pago. Hoy en día la tasa promedio de regularización corresponde al 55%, vale decir, poco más de la mitad de los fraudes detectados. Ante esta magnitud se propone diseñar estrategias enfocadas en aumentar las tasas de regularización, que son bajas y disímiles entre zonales, de modo de aumentar la recaudación.

Para finalizar, es altamente recomendable que junto con mejorar la actual tasa de detección, parte de los esfuerzos vayan enfocados también en diseñar e implementar medidas orientadas a elevar la actual tasa de regularización (del 55%), pues a la larga hoy en día se deja de percibir prácticamente la mitad de la potencial recaudación por concepto de clientes irregulares detectados.

5. BIBLIOGRAFÍA Y FUENTES DE INFORMACIÓN

[1] Informe de Gestión del Sector Sanitario,<http://www.siss.gob.cl/577/articles-8852_recurso_1.xls>, [consulta: 01 enero 2013].

[2] Aguas Andinas – Tarifas Vigentes, <http://www.siss.gob.cl/577/articles-4625_aguas_andinas_g1_sep2012.pdf>, [consulta: 27 diciembre 2012].

[3] LEDESMA, O.,[2011], Reducción de Pérdida por Micromedición, [diapositiva].

[4] PÉREZ, L. 2004. Rango de Factibilidad Económica del Índice de Agua no Contabilizada. Memoria de Magíster en Ingeniería Civil. Bogotá, Universidad de los Andes. 99p.

[5] FAYYAD, U., PADHRAIC, S., PIATETSKY-SHAPIRO, G., [1996], From Data Mining to Knowledge Discovery in Databases. 18 p.

[6] ELISSEEFF, A., GUYON, I. [2003]. An Introduction to Variable and Feature Selection. 26p.

[7] TERRÁDEZ, M., Análisis de Componentes Principales. Universitat Oberta de Catalunya.

[8] LARRAÑAGA, P., Árboles de Clasificación. Universidad del País Vasco-Euskal. 6p.

[9] FIGUEROA, T. 2009. Modelo Predictivo de Quiebres de Stock en un Supermercado. Memoria Ingeniero Civil industrial. Santiago, Universidad de Chile. 54 p.

[10] ZHANG, P. [2007]. Avoiding Pifalls in Neural Network Research. 14 p.

[11] MORA, W. [2005] Programación VBA para Excel y Análisis Numérico. 30p.

6. ANEXOS

A. Distribución de comunas según zona de concesión.

Zona de Concesión	Comuna
MAPOCHO	Conchalí Renca Quilicura Quinta Normal Pudahuel Santiago Independencia Recoleta Huechuraba Cerro Navia Los Prado Estación Central Cerrillos
ANTILCO	La Cisterna San Bernardo La Granja Puente Alto La Florida San Ramón La Pintana Lo Espejo El Bosque
MAIPO	Calera de Tango Talagante Maipú Padre Hurtado Peñaflor El Monte Buin Paine Pirque San José de Maipo Curacaví Isla de Maipo Melipilla Tiltil
CORDILLERA MAPUÉ	San Miguel Providencia Ñuñoa La Reina Las Condes Peñalolén Macul

	San Joaquín Pedro Aguirre Cerda Las Condes Colina Vitacura Lo Barnechea

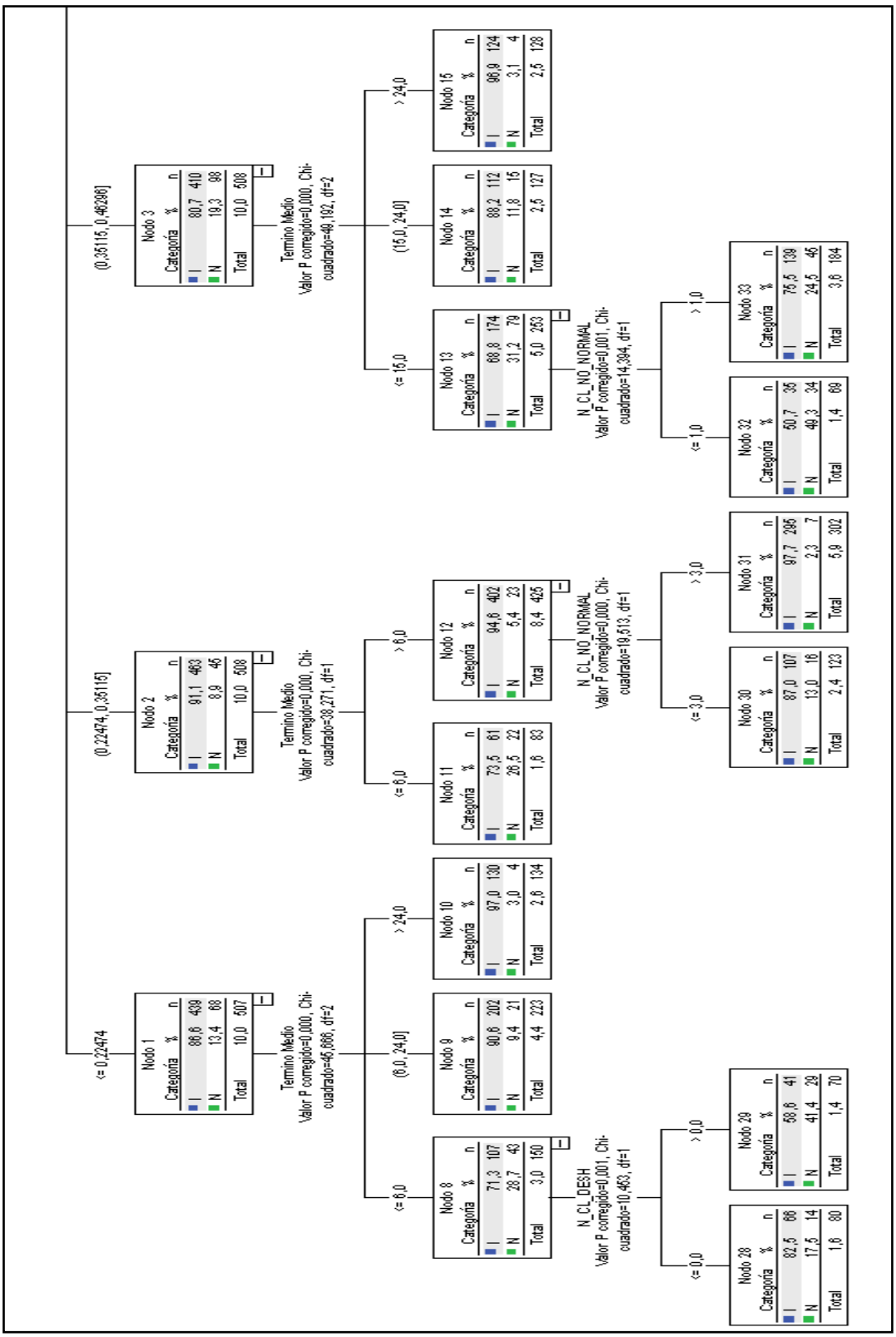
B. Descripción claves de lectura

CLAVE	DESCRIPCIÓN
0	MEDIDOR DESCOMPUESTO
1	MEDIDOR DESTRUIDO
2	CASA CERRADA
3	CASA DESOCUPADA
4	SERVICIO NO UBICADO
5	SERVICIO SIN MEDIDOR
6	DEMOLICION/CONSTRUCCION
7	MEDIDOR DIFICIL ACCESO
8	SERVICIO INACCESIBLE
9	JUSTIFICA SIN CONSUMO
3L	CASA DESOCUPADA C/LECT.
A	COBRA MINIMO
B	COBRA TERMINO MEDIO
C	RATIFICA LECTURA
D	FIJA CONS. MODIF.LECT.ACT.
G	INDUCE CONSUMO
GL	INDUCE CONSUMO FIJA LECTURA
IC	INFORMADA POR CLIENTE
J	MEDIDOR EMPANADO
K	INSTALACION IRREGULAR
L	ACCESO NO PERMITIDO
M	RATIFICA LECTURA (CONTRATISTA)
N	LECTURA NORMAL (INT)
P	SERVICIO FUERA RUTA
Q	MAP INTERVENIDO
H	SIN MEDIDOR SIN USO
Z	MEDIDOR DETENIDO
R	SITIO ERIAZO
R1	REFAC. CONSUMO 1 VEZ TM
R2	REFAC. CONSUMO 2 VEZ TM
R3	REFAC. CONSUMO 3 VEZ TM
RL	SITIO ERIAZO C/LECT
RX	REFC. CONSUMO BAJA 50% TM
S	MEDIDOR INVERTIDO
U	PROPIEDAD SINIESTRADA

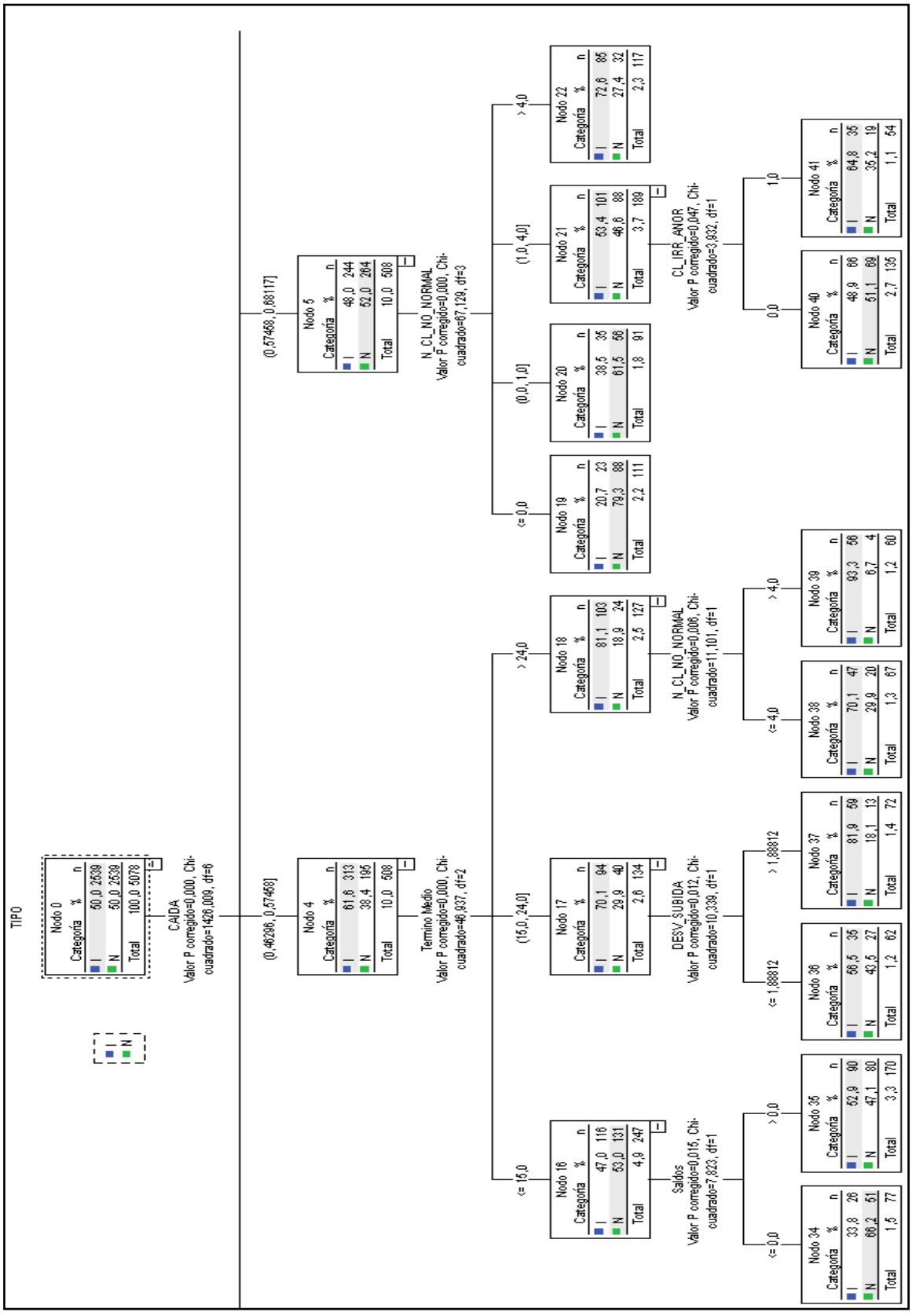
V1	CONSUMO 1 VEZ TM
V2	CONSUMO 2 VECES TM
V3	CONSUMO 3 VECES TM
W	CLIENTE NUEVO SIN LECTURA
WL	SERVICIO S/LECT, C/CORTE Y AL MENOS UNA FACTURA
X	SIN CORTE Y SIN LECTURA

C. Output Árbol de Decisión

Parte izquierda:



Parte Central:



Parte derecha: