DOI: 10.47460/uct.v24i104.366

PREDICCIÓN DE FRAUDES EN EL CONSUMO DE AGUA POTABLE MEDIANTE EL USO DE MINERÍA DE DATOS

Troncoso Espinosa, Fredy Humberto[1], Fuentes Figueroa, Paulina Gisselot[2],
Belmar Arriagada, Italo Ramiro[3]
ftroncos@ubiobio.cl[1], paulina.fuentesfi@gmail.com[2], italo.belmararriagada@gmail.com[3]
Departamento de Ingeniería Industrial, Universidad del Bío-Bío, Concepción-Chile.[1]
Bureau Veritas S.A Santiago-Chile.[2]
Planner en Head Chile, Concepción-Chile.[3]

Recibido (07/08/20), Aceptado (21/08/20)

Resumen: El comportamiento fraudulento en el consumo de agua potable es un problema importante que enfrentan las empresas de tratamiento de agua debido a que genera pérdidas económicas significativas. Caracterizar consumos fraudulentos es una tarea compleja, basada principalmente en la experiencia, y que presenta el desafío de la incorporación constante de nuevos clientes y la variación en el consumo mensual. En esta investigación, las técnicas de minería de datos se utilizan para caracterizar y predecir los consumos fraudulentos de agua potable. Para esto, se utilizó información histórica relacionada con el consumo. Las técnicas aplicadas mostraron un alto rendimiento predictivo y su aplicación permitirá enfocar eficientemente los recursos orientados a evitar este tipo de fraude.

Palabras Clave: Minería de datos, Machine learning, Agua potable, Detección de fraude.

PREDICTION OF FRAUD IN DRINKING WATER CONSUMPTION THROUGH THE USE OF DATA MINING

Abstract: The fraudulent behavior in drinking water consumption is a major problem faced by water treatment companies due to generates significant economic losses. Characterizing fraudulent drinking water consumption is a complex task, based mainly on experience, and which presents the challenge of the constant incorporation of new clients and the variation in monthly consumption. In this research, data mining techniques are used to characterize and predict fraud in the consumption of drinking water. For this, historical information on consumption was used. The techniques applied showed high predictive performance and its application will allow focusing efficiently resources oriented to avoid this type of fraud.

Keywords: Data mining, Machine learning, Drinking water, Fraud detection.

I.INTRODUCCIÓN

Las empresas sanitarias por lo general presentan cierto porcentaje de agua no facturada, es decir, aquella agua que se produce, pero no se logra cobrar al consumidor final. Las fallas en la infraestructura de las redes de las empresas sanitarias y los consumos ilegales o fraudulentos son los dos factores que explican esta perdida. En Chile esta pérdida corresponde a un tercio del agua potable que se produce [1], en donde los hurtos y conexiones clandestinas explican entre 8 y 10 por ciento de estas pérdidas.

Bureau Veritas S.A. es una empresa líder mundial en ensayos, inspección y certificación de agua potable [2]. Uno de sus principales clientes es la empresa sanitaria ESSBIO S.A. la cual es una de las empresas sanitarias más importantes en Chile [3]. Uno del servicio prestado por Bureau Veritas es la inspección para la detección de fraude en el consumo de agua potable residencial. Actualmente se inspeccionan en promedio 2500 servicios mensuales, de los cuales el 75 por ciento corresponde a inspecciones efectivas, es decir aquellas en las cuales tuvo acceso al medidor.

Del total de inspecciones efectivas cerca del 17 por ciento corresponden a ilícitos. Algunas de las variables que normalmente se consideran para la identificación de un consumo fraudulento son los descensos progresivos en el consumo, descensos bruscos en el consumo, consumo anormalmente bajo y ubicación geográfica [4].

Si bien las variables antes mencionadas y la experiencia de los inspectores son un factor importante en la detección de consumo fraudulento, hay variables que no son evidentes y que hacen complejo caracterizar el consumo fraudulento. Entre estas se encuentra el consumo máximo, mínimos, número de lecturas y estimadores de variabilidad [5]. Estas variables consideran el consumo histórico y Bureau Veritas cuenta con una base de datos que contiene los consumos históricos de ESSBIO incluidos aquellos que han cometido fraude de agua.

Por esta razón se hace necesario analizar con mayor profundidad los datos relacionados con el consumo en búsqueda de un patrón que permita predecir con un mayor nivel de asertividad un consumo fraudulento. La minería de datos permite encontrar estos patrones [6].

La minería de datos es una de las herramientas más eficientes en la detección de fraude [7] [8]. Las técnicas de minería de datos utilizadas en la detección de fraude son variadas. Dentro de las más utilizadas se encuentran técnicas de machine learning [9].

Las técnicas de machine learning aprenden el patrón general oculto en los datos y luego lo utilizan para generar una nueva predicción. Dentro de estas técnicas se encuentran las redes neuronales, support vector machine, naive bayes, árbol de decisión y k- nearest neighbor [10]. En cuanto a la aplicación de estas técnicas, destaca el empleo de redes neuronales para la detección de usuarios irregulares residenciales en el consumo de electricidad [11]. La red neuronal permitió reconocer los patrones de los consumos anormales de los usurarios permitiendo así identificar a posibles consumos fraudulentos. Las redes neuronales no sólo han sido utilizadas en el sector de distribución de electricidad, sino también en la detección de fraude en tarjetas de crédito [12] y en el fraude en suscripciones en telecomunicaciones [13]. El principal inconveniente de las redes neuronales es que no pueden dar como resultado una fórmula probabilística simple de clasificación [14]. Por otra parte, naive bayes ha sido utilizado para la detección de fraude en empresas de telecomunicaciones de telefonía móvil [15], en la detección de fraude de seguros [16], fraude en estados financieros [17] y fraude en transacciones de tarjetas de crédito [18] alcanzando un desempeño general de 81%. En cuanto a support vector machine se ha utilizado principalmente para la detección de fraude en telecomunicaciones [19]. El uso de support vector machine permitió clasificar correctamente a los suscriptores normales de los suscriptores de fraude con un desempeño de 99,06%. Por otra parte, support vector machine ha sido utilizado en la detección de fraude en tarjetas de crédito [20], y electricidad [21]. Arboles de decisión han sido utilizados en la detección de fraude en empresas de electricidad [5], en la detección de fraude en instituciones financieras [22] y en la detección de fraude en el comercio electrónico [23]. La literatura disponible relacionada con técnicas de clasificación en la detección de fraude en el consumo de agua es limitada en comparación con otros sectores, como el sector eléctrico, telecomunicaciones y financiero. Una investigación destacable utiliza técnicas de minería de datos para descubrir el consumo de agua fraudulento en la ciudad de Gaza [24]. El autor se centró en usar support vector machine y lo comparó con k-nearest neighbor y redes neuronales. En otra investigación se utilizó support vector machine y k-nearest neighbor para consumos de agua sospechosos con el objetivo de ayudar a Yarmouk Water Company (YWC) en la ciudad de Irbid en Jordania a superar la pérdida de ganancias [25]. Los experimentos realizados demostraron un buen rendimiento de support vector machine y los vecinos k-nearest neighbor con una precisión general de alrededor del 70% para ambos lo que mostro mejor desempeño que las inspecciones manuales aleatorias realizadas por los equipos de YWC con una tasa de impacto de alrededor del 1%. Respecto a árbol de decisión en [5] utilizó árbol de decisión en la detección de fraude en

ISSN 2542-3401/ 1316-4821 UNIVERSIDAD, CIENCIA y TECNOLOGÍA *Vol. 24, Nº 104 Septiembre 2020 (pp. 58-66)*

empresas de electricidad como medida complementaria a otras técnicas de machine learning, con el fin de obtener patrones adicionales de comportamiento gracias a su representación de red.

Dado los buenos resultados obtenidos en la detección de fraude, se propone la utilización de minería de datos para la identificación de consumos fraudulentos de agua potable. Para ello se utilizarán diversas técnicas de machine learning con el fin de determinar el de mejor desempeño. Se espera que la mejor técnica permita identificar y priorizar a los potenciales consumos fraudulentos, con el fin de guiar a los inspectores a una búsqueda más eficiente.

En el apartado II se analiza la metodología utilizada y en el III se muestra la aplicación a la base de datos. En el apartado IV se muestran los resultados que determinan la mejor técnica de machine learning, se obtiene el patrón general que caracteriza los consumos fraudulentos y se discuten los resultados.

II.METODOLOGÍA

La metodología utilizada es Knowledge Discovery in Databases KDD [26]. Esta metodología está compuesta por cinco etapas iterativas que tiene como objetivo principal la extracción de conocimiento oculto en bases de datos [27].

La primera etapa es la selección de datos, donde se determinan las fuentes de datos y el tipo de información a utilizar. Se deben conocer a cabalidad las variables involucradas y tener identificada la variable a predecir.

La segunda considera la limpieza de los datos, con el fin de tener información más confiable y que aporte mayor valor a la predicción. Esta limpieza incorpora el análisis de datos faltantes, de datos inconsistentes, y el análisis de datos fuera de rango.

La tercera etapa consiste en la transformación y selección de variables. Las variables se transforman para generar nuevas variables, que enriquezcan la información con la que se entrenará el modelo para que este tenga un mejor desempeño predictivo. Luego de esto se procede a identificar aquellas que mejor predicen la variable de interés [28].

La cuarta etapa es la de minería de datos donde se aplican las técnicas de machine learning. Para que estas técnicas puedan identificar el patrón y se pueda evaluar su desempeño se aplicará la técnica Hold Out [10]. Este método divide los datos aleatoriamente en dos conjuntos mutuamente excluyentes: conjunto de entrenamiento y conjunto de prueba. El conjunto de entrenamiento representa el 70% del total de los datos y el conjunto de prueba el 30%. Mediante el conjunto de entrenamiento la técnica de machine learning aprende el patrón que discrimina entre las clases. Mediante el conjunto de prueba, se mide el desempeño predictivo del modelo

La quinta y última etapa consiste en la evaluación de los resultados, que resumen en la Matriz de Confusión [29]. En la Matriz de confusión la clase 1 identifica un consumo fraudulento y la clase 0 que identifica un consumo no fraudulento. VP los verdaderos positivos que son los elementos de la clase 1 correctamente predichos por el modelo o verdaderos positivos y FN representa los elementos de la clase 1 incorrectamente predichos por el modelo o tasa falso positivo. TN representa los elementos de la clase 0 correctamente predichos por el modelo o tasa verdadero negativo y FP representa los elementos de la clase 0 incorrectamente predichos por el modelo o tasa falso positivo.

La Matriz de Confusión, permite obtener tres medidas de desempeño [30]. La primera es Accuracy que mide el desempeño general del modelo y representa la proporción total de predicciones que fueron correctamente clasificadas. Se obtiene la suma de VP y VN dividido por el total de datos en la matriz. La segunda medida es Recall que representa la tasa de elementos perteneciente a la clase 1 que fueron clasificadas correctamente y se obtiene al dividir VP entre la suma de VP y FN. La tercera medida es Precision que representa la tasa de elementos de la clase 1 entre el total de elementos predichos como clase 1. Se obtiene al dividir VP entre la suma de VP y FP. Otra medida de evaluación de los modelos es la técnica Gain Chart, la cual nos muestra una gráfica del ranking generado por cada machine learning [31] [32].

Luego de estas cinco etapas se obtiene el nuevo conocimiento que será aplicado al negocio. En la Figura 1 se muestran los procesos de datos en la metodología KDD.

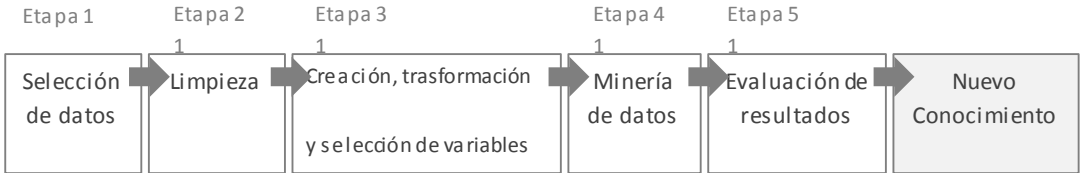


Fig. 1. Procesos de datos dentro de la metodología Knowledge Discovery in Databases KDD

ISSN 2542-3401/ 1316-4821

III.DESARROLLO

A.Selección de datos

Los datos proporcionados por ESSBIO comprenden 970.000 clientes. La información está contenida en tres bases de datos. La primera contiene la información comercial de los consumos. La segunda posee registros de los consumos de agua de los últimos 48 meses, considerando como último mes, noviembre 2014. La tercera contiene los registros de inspección e información histórica de las inspecciones realizadas durante los últimos 12 meses. De la base de datos se seleccionó los consumos con tarifa residencial y servicio normal. La base de datos final contiene 23.005 registros, donde 12.250 corresponden a consumos regulares y 10.755 a consumos fraudulentos.

B.Pre procesamiento del conjunto de datos

En esta etapa se identificó las variables con muy baja variabilidad, datos atípicos y datos faltantes. En cuanto a la variabilidad se eliminó la variable clase instalación, representada en un 93% por la instalación tipo 4 y que no explica la variable objetivo. Se eliminó las variables Tipo cliente, Tipo de servicio y Ruta ya que solo se consideró los clientes residenciales, normales y el identificador del cliente. La identificación de datos atípicos, se realizó para cada variable mediante la regla de tres sigmas [33]. Los datos faltantes y fuera de rango fueron reemplazados mediante el valor de la variable de un registro similar [34].

C.Creación y transformación de variables

La creación de nuevas variables es importante para identificar los patrones que caracterizan a los consumos fraudulentos de agua. Para esto se consideró elementos significativos para la detección de fraude como los descensos progresivos en el consumo, descensos bruscos en el consumo, consumo anormalmente bajo y ubicación geográfica de agua, similares a las mencionadas en [4] [35] [5]. Se crearon las variables que se muestran en la Tabla I.

TABLA I. Descripción de las variables creadas

A	Identificador único del cliente	I	Desviación últimos 3 meses
B	Cantidad de fraudes	J	Desviación últimos 6 meses
C	Diámetro del medidor (cm)	K	Desviación últimos 12 meses
D	Meses sin consumo últimos 12 meses	L	Consumo total últimos 3 meses
E	Consumo promedio últimos 3 meses	M	Consumo total últimos 12 meses
F	Consumo promedio últimos 6 meses	N	Consumo total últimos 24 meses
G	Consumo promedio últimos 12 meses	Ñ	Ubicación lectura
H	Consumo promedio últimos 24 meses	O	Fraude

Se identificó las variables con alta dependencia lineal y se dejó una de ellas pues las variables con alta dependencia lineal explicarán un fenómeno de manera similar. Se utilizó una matriz de correlación y se consideró una alta dependencia lineal cuando la correlación fue mayor o igual a +-0.9. Las variables eliminadas fueron el E, F, H, L, M, N y G. Se decide eliminar el variable Ñ (Ubicación lectura) para que el modelo creado no dependa de variables geográficas. La transformación de variables se llevó a cabo para la selección de las variables más importantes y para el entrenamiento de cada algoritmo de machine learning, según sus requerimientos.

D.Selección de variables

Para el entrenamiento y prueba de los modelos se consideró las variables con mayor poder predictivo [28]. Para esto se utilizó el estadístico Chi-cuadrado el cual indica que, a mayor valor, mayor es la dependencia entre una variable y la variable a predecir. La variable con mayor valor de Chi Cuadrado resulta ser la variable más importante. Para aplicar este método se requiere categorizar las variables numéricas. Se categorizó de acuerdo al número bajo de categorías y que maximiza su dependencia con la variable a predecir como se muestra en la Tabla II.

TABLA II. Categorías por variable y poder predictivo según estadístico Chi-cuadrado

Variable	Categorización	Chi-cuadrado
Cantidad de fraudes	3 Categorías: No tiene – [1,2] – >=3	13747
Meses sin consumo últimos 12 meses	3 Categorías: No tiene - [1,2] - >=3	900
Consumo promedio últimos 12 meses	4 Categorías: [0-5] - [6,11] - [12,15] - >15	431
Desviación últimos 6 meses	4 Categorías: [0,2] - [3,10] - [11,20] - >20	423

IV.RESULTADOS

A.Minería de datos

Se entrenó y probó cinco técnicas de machine learning utilizando los algoritmos incorporados en la librería Scikit-Learn de Python [36] llamados: Decision Tree, Naive Bayes, Neuronal Net, Support Vector Machine y KNN. De manera de optimizar el desempeño predictivo de cada algoritmo, se iteró en los distintos algoritmos y se ajustó los respectivos parámetros de cada algoritmo siguiendo el procedimiento que se muestra en el pseudocódigo de la Figura 2.

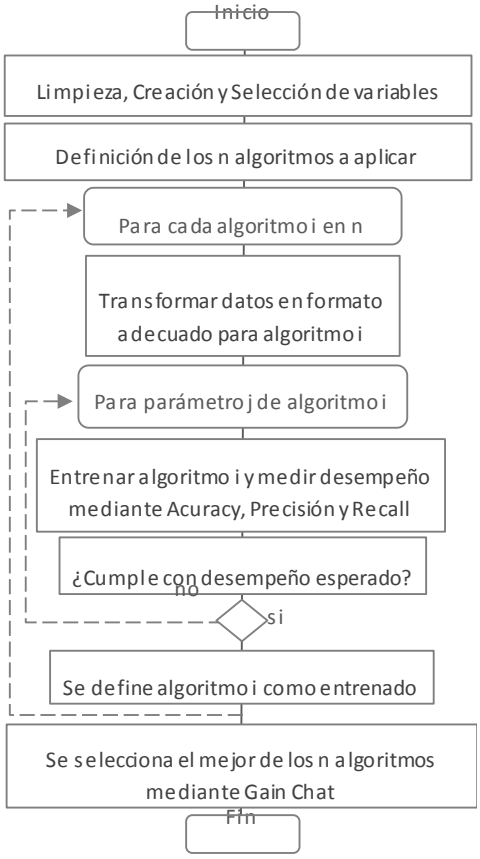


Fig. 2. Pseudocódigo para la evaluación y selección de los algoritmos de machine learning

La Tabla III muestra que el mejor desempeño de cada algoritmo de machine learning entrenado. El mejor desempeño general lo obtiene el Decision Tree, seguido del Support Vector Machine. La predicción específica de la clase fraude muestra un buen desempeño con recall sobre 77% y precisión sobre 88%. Esto implica que las técnicas en general identifican bien los consumos fraudulentos.

ISSN 2542-3401/ 1316-4821

TABLA IV. Desempeño predictivo de los algoritmos considerados

	Modelos de clasificación				
Medida de desempeño	Decision Tree	Naive Bayes	Neuronal Net	Suppot Vector Machine	KNN
Accuracy	88.16%	87.34%	87.96%	88.03%	86.05%
Recall	78%	79.86%	77.78%	77.44%	80.29%
Precision	95.92%	92%	95.66%	96.23%	88.79%

La definición de la mejor técnica considera un ranking de consumos fraudulentos mediante la gráfica Gain Chart que muestra la variación de la tasa verdadero positivo (consumos fraudulentos clasificados correctamente) en función del porcentaje de individuos dentro del ranking. En esta gráfica, como se muestra en la Figura 3, un mejor desempeño implica una curva más cercana al punto (0,1), lo que se asocia a un mayor número de consumos clasificados en las posiciones más alta del ranking. Al inicio, las curvas se interponen, sin embargo, en la parte superior se puede ver una leve diferencia entre los modelos. El mejor desempeño lo obtiene Neural Net seguido de Naive Bayes. Como complemento a Neural Net se generó un árbol para comprender las relaciones entre cada uno de las variables en el fraude de agua potable.

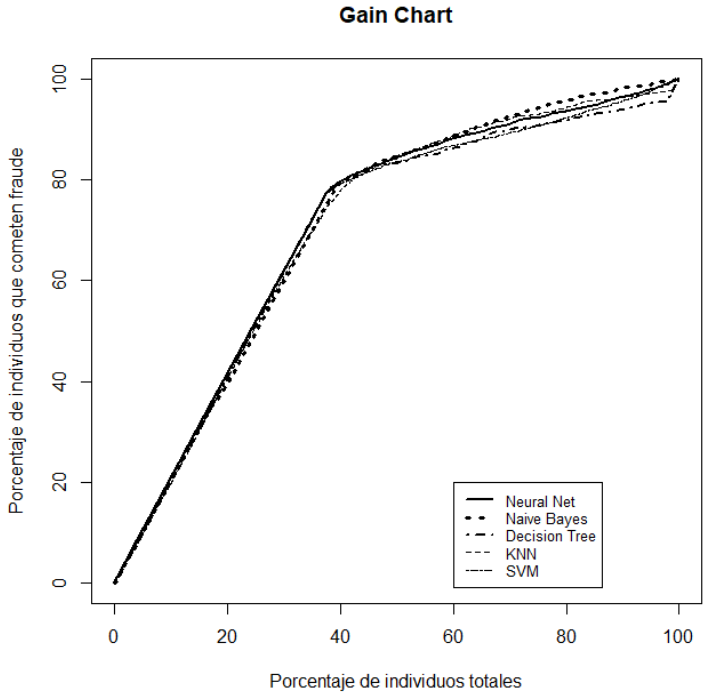


Fig. 3. Desempeño de cada algoritmo de machine learning considerado

La Figura 4 muestra el árbol de decisión obtenido a partir del resultado entregado por el algoritmo Decision Tree. Es posible observar que cuando se ha cometido al menos una vez fraude de agua se seguirá cometiendo fraude. Cuando no hay un historial de fraude y no existen meses sin consumo en los últimos 12 meses, no se comete fraude. Por otra parte, cuando existen algunos meses sin consumo durante los últimos 12 meses, el consumo fraudulento dependerá de la desviación entre los consumos de agua. Si las desviaciones en los últimos 6 meses son altas, se comete fraude. Si bien el árbol de decisión permite obtener las reglas generales de un consumo fraudulento, es necesario utilizar probabilidad entregada por la red neuronal para generar un ranking de que permita priorizar la inspección decidir las acciones a seguir.

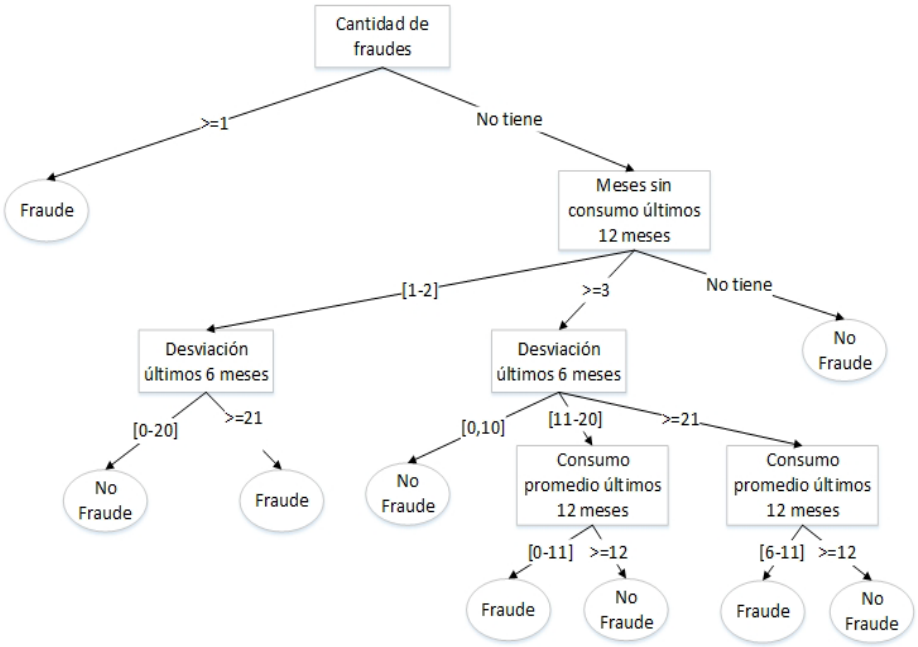


Fig. 4. Árbol de Decisión que caracteriza el consumo fraudulento.

B.Discusión

A través de los datos históricos fue posible extraer patrones de comportamiento de un consumo fraudulento. Las consideraciones obtenidas de [4] permitieron crear variables significativas para el modelo, las cuales concuerdan con estudios realizados anteriormente en la detección de fraude como ubicación geográfica y estimadores de variabilidad en los consumos [35] [5]. Sin embargo, a pesar de que la ubicación geográfica es una de las variables más importantes en la detección de fraude, se decidió eliminar y solo dejar las atribuibles a los consumos de manera de generar un modelo estándar y aplicable a cualquier empresa sanitaria o a diferentes sucursales. Las variables atribuibles a los consumos de agua son de fácil obtención y no dependen de factores geográficos.

El mejor desempeño predictivo logrado fue de un 88%, por lo que las variables creados mediante los consumos permiten la detección de consumos fraudulentos, sin depender de variables demográficas del consumo, información que es de difícil acceso para las empresas sanitarias. Sin embargo, es posible incorporar otras variables relacionadas al estado de las facturas que son expuestas por [25] en su estudio de detección de fraude agua, de manera de mejorar el desempeño predictivo.

El porcentaje de error en la predicción de consumos fraudulentos puede tener sus causas en el comportamiento que se percibe como aleatorio de los clientes como el hecho que dejen sus casas por un tiempo debido a vacaciones u otras actividades. Este patrón de comportamiento es difícil de detectar y altera variables como son los meses sin consumo, las desviaciones entre los consumos y los promedios de consumo. Sin embargo, el porcentaje de error general del modelo fue de un 12%, lo que se considera un error aceptable.

V.CONCLUSIONES

La aplicación de minería de datos mediante técnicas de machine learning permitió la identificación de variables importante y de patrones para la detección de fraude en el consumo de agua potable. Se entrenó y probó diversas técnicas de machine learning mediante información histórica de los consumos fraudulentos y no fraudulentos utilizando la metodología Knowledge Discovery in Databases KDD.

Las variables que maximizan el desempeño predictivo de los modelos entrenados fueron: Cantidad de fraudes, Meses sin consumo últimos 12 meses, Consumo promedio últimos 12 meses y Desviación últimos 6 meses. Estas variables son de fácil acceso para la empresa sanitaria por lo la implementación del modelo es altamente factible.

Considerando las medidas de desempeño Accuracy, Precision y Recall, la técnica de mejor desempeño predictivo fue Decision Tree Classifier. Sin embargo, mediante la utilización de la gráfica Gain Chart, que permite evaluar los algoritmos de acuerdo a un ranking de probabilidad de cometer fraude, el algoritmo Neural Net obtuvo el mejor desempeño.

El árbol de decisión permitió identificar la relación

ISSN 2542-3401/ 1316-4821 UNIVERSIDAD, CIENCIA y TECNOLOGÍA *Vol. 24, Nº 104 Septiembre 2020 (pp. 58-66)*

existente entre las variables más importantes y como estas definen el patrón asociado al consumo fraudulento de agua potable. El patrón que más caracteriza el consumo fraudulento es que cuando se ha cometido fraude en el consumo de agua, existe una alta probabilidad de que vuelvan a cometer fraude nuevamente. Cuando no ha cometido fraudes, la probabilidad de cometer fraude dependerá del número de meses sin consumo durante el último año y de las variaciones los consumos de agua los meses anteriores. A mayor variación durante los últimos seis meses, mayor es la probabilidad de fraude.

La utilización de técnicas de machine learning permitirá mejorar la detección de consumos fraudulentos y focalizar los recursos involucrados en esta labor, al permitir concentrar el trabajo de inspección en aquellos consumos que muestre una mayor probabilidad de ser fraudulentos.

REFERENCIAS

[1]Centro de Investigación Periodística., «Producción y facturación de agua potable,» 30 Julio 2020. [En línea]. Available: https://ciperchile.cl/wp-content/uploads/gestion-siis-2014-pag88.pdf. [Último acceso: 30 Julio 2020].

[2]Bureau Veritas S.A., «https://www.bureauveritas.cl/es,» [En línea]. Available: https://www.bureauveritas.cl/es/bureau-veritas-lider-mundial-en-ensayos-inspeccion-y-certificacion. [Último acceso: 1 Junio 2020].

[3]Essbio S.A., «www.essbio.cl,» [En línea].

[4]I. Monedero, F. Biscarri, J. Guerrero, M. Peña, M. Roldán y C. León, «Detection of water meter under-registration using statistical algorithms,» Journal of Water Resources Planning and Management, vol. 142, nº 1, p. 04015036, 2016.

[5]I. Monedero, F. Biscarri, C. León, J. Guerrero, J. Biscarri y R. Millán, «Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees,» International Journal of Electrical Power & Energy Systems, vol. 34, nº 1, pp. 90-98, 2012.

[6]S. Wang, «A comprehensive survey of data mining-based accounting-fraud detection research,» de 2010 International Conference on Intelligent Computation Technology and Automation, New York, 2010.

[7]J. Bierstaker, R. Brody y C. Pacini, «Accountants' perceptions regarding fraud detection and prevention methods,» Managerial Auditing Journal, vol. 21, nº 5, pp. 520-535, 2006.

[8]C. Phua, V. Lee, K. Smith y R. Gayler, «A comprehensive survey of data mining-based fraud detection research,» arXiv preprint arXiv:1009.6119, 2010.

[9]S. Kotsiantis, I. Zaharakis y P. Pintelas, «Machine learning: a review of classification and combining techniques,» Artificial Intelligence Review, vol. 26, nº 3, pp. 159-190, 2006.

[10]J. Han, J. Pei y M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.

[11]Z. Zheng, Y. Yang, X. Niu, H.-N. Dai y Y. Zhou, «Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids,» IEEE Transactions on Industrial Informatics, vol. 14, nº 4, pp. 1606-1615, 2017.

[12]R. Patidar y L. Sharma, «Credit card fraud detection using neural network,» International Journal of Soft Computing and Engineering (IJSCE), nº 1, pp. 32-38, 2011.

[13]H. Farvaresh y M. M. Sepehri, «A data mining framework for detecting subscription fraud in telecommunication,» Engineering Applications of Artificial Intelligence, vol. 24, nº 1, pp. 182-194, 2011.

[14]I.-C. Yeh y C.-h. Lien, «The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,» Expert Systems with Applications, vol. 36, nº 2, pp. 2473-2480, 2009.

[15]P. Kang, «One-Class Naïve Bayesian Classifier for Toll Fraud Detection,» IEICE Transactions on Information and Systems, vol. 97, nº 5, pp. 1353-1357, 2014.

[16]R. Bhowmik, «Data mining techniques in fraud detection,» Journal of Digital Forensics, Security and Law, vol. 3, nº 2, p. 3, 2008.

[17]S. Viaene, R. Derrig y G. Dedene, «A case study of applying boosting Naive Bayes to claim fraud diagnosis,» IEEE Transactions on Knowledge and Data Engineering, vol. 16, nº 5, pp. 612-620, 2004.

[18]L. J. Santos y S. Ocampo, «Bayesian Method with Clustering Algorithm for Credit Card Transaction Fraud Detection,» Romanian Statistical Review, nº 1, 2018.

[19]R. Sallehuddin, S. Ibrahim, A. M. Zain y A. H. Elmi, «Detecting SIM box fraud by using support vector machine and artificial neural network,» Jurnal Teknologi, vol. 74, nº 1, pp. 137-149, 2015.

[20]Y. Şahin y E. Duman, «Detecting credit card fraud by decision trees and support vector machines,» de International MultiConference of Engineers and Computer Scientists, Hong Kong, 2011.

[21]J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed y M. Mohamad, «Nontechnical loss detection for metered customers in power utility using support vector machines,» IEEE transactions on Power Delivery, vol. 25, nº 2, pp. 1162-1171, 2009.

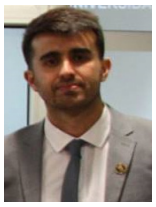
[22]. K. Zou, W. Sun, H. Yu y F. Liu, «ID3 decision tree in fraud detection application,» de 2012 International Conference on Computer Science and Electronics En-

gineering, Hangzhou, 2012.
[23]M. Lek, . B. Anadarajah, N. Cerpa y R. Jamieson, «Data mining prototype for detecting ecommerce fraud [Research in Progress],» de Global Co-operation in the New Millennium-The 9th European Conference on Information Systems, Bled, 2001.
[24]E. H. Humaid, «A data mining based fraud detection model for water consumption billing system in MOG,» PhD and MSc Theses, 2012.
[25]Q. Al-Radaideh y M. Al-Zoubi, «A data mining based model for detection of fraudulent behaviour in water consumption,» de 2018 9th International Conference on Information and Communication Systems (ICICS), Irbid, 2018.
[26]U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «Knowledge Discovery and Data Mining: Towards a Unifying Framework,» de KDD-96 Proceedings, 1996.
[27]R. Brachman y T. Anand, «The process of knowledge discovery in databases,» de Advances in knowledge discovery and data mining, 1996.
[28]I. Guyon y A. Elisseeff, «An introduction to variable and feature selection,» Journal of machine learning research, vol. 3, nº Mar, pp. 1157-1182, 2003.
[29]A. M. Hay, «The derivation of global estimates from a confusion matrix,» International Journal of Remote Sensing, vol. 9, nº 8, pp. 1395-1398, 1988.
[30]M. Sokolova y G. Lapalme, «A systematic analysis of performance measures for classification tasks,» Information processing & management, vol. 45, nº 4, pp. 427-437, 2009.
[31]S. H. Ha y S. H. Joo, «A hybrid data mining method for the medical classification of chest pain,» International Journal of Computer and Information Engineering, vol. 4, nº 1, pp. 33-38, 2010.
[32]A. Shen, R. Tong y Y. Deng, «Application of classification models on credit card fraud detection,» de International conference on service systems and service management, 2007.
[33]K. Lakshminarayan, S. Harp, R. Goldman y T. Samad, «Imputation of Missing Data Using Machine Learning Techniques,» de KDD, 1996.
[34]B. Nguyen , J. L. Rivero y C. Morell, «Aprendizaje supervisado de funciones de distancia: estado del arte,» Revista Cubana de Ciencias Informáticas, vol. 9, nº 2, pp. 14-28, 2015.
[35]C. León, F. Biscarri, I. Monedero, J. I. Guerrero, J. Biscarri y R. Millán, «Variability and trend-based generalized rule induction model to NTL detection in power companies,» IEEE Transactions on Power Systems, vol. 26, nº 4, pp. 1798-1807, 2011.
[36]F. Pedregosa, G. Varoquaux , A. Gramfort , V. Michel y B. Thirion, «Scikit-learn: Machine learning in Python,» Journal of machine Learning research, vol. 12, pp. 2825-2830, 2011.

RESUMEN CURRICULAR

Fredy Troncoso Espinosa, Doctor en Sistemas de Ingeniería, Universidad de Chile, Ingeniero Civil Industrial Universidad del Bío-Bío, Chile. Académico e Investigador Departamento de Ingeniería Industrial, Universidad del Bío-Bío. Concepción, Chile.

Paulina Fuentes Figueroa, Ingeniera Civil Industrial, Universidad del Bío-Bío, Chile. Jefe Control de Ingresos Bureau Veritas S.A,



Italo Belmar Arriagada, Ingeniero Civil Idustrial, Universidad del Bío-Bío, Chile. Planner en Head Chile.

ISSN 2542-3401/ 1316-4821