

# Energizados: los beneficios de una herramienta basada en las metodologías de machine learning para facilitar la detección de robo eléctrico

María Cristina Giraldo  
Carlos Ríos  
Arturo Alarcón  
Virginia Snyder  
Carlos Echevarría  
Alexander Riobo  
Michelle Hallack  
Jose Luis Irigoyen

División de Energía

NOTA TÉCNICA N°  
IDB-TN-2444

Enero 2022

# Energizados: los beneficios de una herramienta basada en las metodologías de machine learning para facilitar la detección de robo eléctrico

María Cristina Giraldo  
Carlos Ríos  
Arturo Alarcón  
Virginia Snyder  
Carlos Echevarría  
Alexander Riobo  
Michelle Hallack  
Jose Luis Irigoyen

**Catalogación en la fuente proporcionada por la  
Biblioteca Felipe Herrera del  
Banco Interamericano de Desarrollo**

Energizados: los beneficios de una herramienta basada en las metodologías de machine learning para facilitar la detección de robo eléctrico / Cristina Giraldo, Carlos Ríos, Arturo Alarcón, Virginia Snyder, Carlos Echevarría, Alexander Riobo, Michelle Hallack, José Luis Irigoyen.

Incluye referencias bibliográficas p. cm. — (Nota técnica del BID ; 2444)

1. Machine learning-Latin America. 2. Energy Security-Latin America. I. Giraldo, Cristina. II. Ríos, Carlos. III. Alarcón, Arturo. IV. Snyder, Virginia. V. Echevarría, Carlos. VI. Riobo, Alexander. VII. Carvalho Metanias Hallack, Michelle. VIII. Irigoyen, José Luis. IX. Banco Interamericano de Desarrollo. División de Energía. X. Serie. IDB-TN-2444

Palabras clave: Digitalización, energía eléctrica, inteligencia artificial, machine learning, fraudes eléctricos, pérdidas no técnicas  
Códigos JEL: O13, Q49, N76

<http://www.iadb.org>

Copyright © [2022] Banco Interamericano de Desarrollo. Esta obra se encuentra sujeta a una licencia Creative Commons IGO 3.0 Reconocimiento-NoComercial-SinObrasDerivadas (CC-IGO 3.0 BY-NC-ND) (<http://creativecommons.org/licenses/by-nc-nd/3.0/igo/legalcode>) y puede ser reproducida para cualquier uso no-comercial otorgando el reconocimiento respectivo al BID. No se permiten obras derivadas.

Cualquier disputa relacionada con el uso de las obras del BID que no pueda resolverse amistosamente se someterá a arbitraje de conformidad con las reglas de la CNUDMI (UNCITRAL). El uso del nombre del BID para cualquier fin distinto al reconocimiento respectivo y el uso del logotipo del BID, no están autorizados por esta licencia CC-IGO y requieren de un acuerdo de licencia adicional.

Note que el enlace URL incluye términos y condiciones adicionales de esta licencia.

Las opiniones expresadas en esta publicación son de los autores y no necesariamente reflejan el punto de vista del Banco Interamericano de Desarrollo, de su Directorio Ejecutivo ni de los países que representa.



**Energizados: Los Beneficios De Una Herramienta Basada En Las Metodologías de  
Machine Learning Para Facilitar La Detección De Robo Eléctrico**

María Cristina Giraldo

Carlos Ríos

Arturo Alarcón

Virginia Snyder

Carlos Echevarría

Alexander Riobo

Michelle Hallack

Jose Luis Irigoyen

Enero 31, 2022

## **Tabla de Contenidos**

---

<b>RESUMEN .....</b>	<b>3</b>
<b>INTRODUCCIÓN .....</b>	<b>4</b>
<b>ESTADO DEL ARTE.....</b>	<b>6</b>
<b>ENERGIZADOS .....</b>	<b>11</b>
<b>MODELAMIENTO DE LA INFORMACIÓN .....</b>	<b>11</b>
<b>GRADIENT BOOSTING .....</b>	<b>11</b>
<b>AUTOENCODERS .....</b>	<b>12</b>
<b>REGLAS ESTADÍSTICAS .....</b>	<b>13</b>
<b>EXPERIMENTOS.....</b>	<b>14</b>
<b>CONFIGURACIÓN EXPERIMENTAL .....</b>	<b>15</b>
<b>DATOS .....</b>	<b>16</b>
<b>RESULTADOS.....</b>	<b>17</b>
<b>DISCUSION .....</b>	<b>19</b>
<b>CONCLUSIÓN.....</b>	<b>20</b>
<b>ANEXOS .....</b>	<b>22</b>
<b>REFERENCIAS.....</b>	<b>24</b>

## RESUMEN

---

Con la transformación digital y la cantidad de información que se está produciendo y almacenando por medidores inteligentes, el sector energético encuentra una oportunidad para afrontar y atacar un problema que ha sido preocupante durante muchos años, las pérdidas de energía eléctrica. Aproximadamente el 15% del total de la oferta de electricidad se pierde en la región, o, dicho de otra manera, más de 50 millones de viviendas de la región se podrían haber abastecido con electricidad durante todo el 2019 utilizando las pérdidas de los sistemas de transmisión y distribución.<sup>1</sup> En esta nota abordaremos el tema de las pérdidas no técnicas, es decir el robo o fraude de electricidad. El fraude energético es un tema de importancia dado que representa riesgos para las personas, pérdidas de materiales y económicas. El principal desafío y problema es quizá el de lidiar con el comportamiento humano.

Una de las posibles soluciones para enfrentar este problema del fraude eléctrico es implementar algoritmos sofisticados que sean capaces de detectar patrones anormales y así identificar dichas situaciones.

Para demostrar y resolver lo anterior se crea la herramienta “Energizados”. “Energizados” es una solución basada en aprendizaje automático el cual ayuda a detectar y disminuir las pérdidas no técnicas reduciendo tiempos de regularización e incrementando la precisión de identificación de fraudes. Energizados nace como una solución desarrollada por el Banco Interamericano de Desarrollo y hasta el momento se han obtenidos resultados prometedores indicando que los modelos aplicados podrían ser de utilidad para los países de la región, pues “Energizados” aumentó la captura de fraudes eléctricos en 1.65 veces. Dado que “Energizados” no se encuentra sesgado a una sola localidad de la ciudad esto ayudó a las cuadrillas a identificar

---

<sup>1</sup> <https://publications.iadb.org/en/electrokit-power-utility-toolkit-electricity-loss-reduction>

fraudes aledaños por “intuición”, lo cual es una ventaja comparada con el sistema actual de la compañía que tiende a remitir a las mismas localidades dentro de la ciudad. Aunque la herramienta propuesta obtuvo resultados prometedores la implementación de ella debería de estar acompañada de políticas que permita tomar acciones para la disminución de reincidencias del fraude eléctrico.

## INTRODUCCIÓN

---

Para una sociedad que depende altamente de la disponibilidad, eficiencia y confiabilidad de la electricidad, se hace indispensable para las empresas eléctricas del sector tener una buena administración de la producción y distribución de energía. Uno de los grandes problemas que preocupa a esta industria son las pérdidas eléctricas, que se produce mayormente en el segmento de distribución. En el transporte de energía, las pérdidas son la diferencia entre la electricidad que ingresa a la red y la que es entregada para el consumo final, y son reflejo del nivel de eficiencia de la infraestructura en transmisión y distribución.

El poder reducir las pérdidas es fundamental para incrementar la eficiencia de la distribución de energía, y en muchos casos puede apoyar incluso a mejorar la sostenibilidad financiera de las empresas de distribución. Datos de 2012 estimaban que las pérdidas de electricidad equivalían hasta el 0.3% del PBI de la región, entre unos 11 y 17 millones de dólares americanos.<sup>2</sup>

El concepto de pérdidas eléctricas incluye también la electricidad entregada pero no facturada, que se traduce directamente en pérdidas financieras y sirve como indicador del

---

<sup>2</sup> Electricidad perdida: dimensionando las pérdidas de electricidad en los sistemas de transmisión y distribución en América Latina y el Caribe / Raúl Jiménez, Tomás Serebrisky, Jorge Mercado. p. cm. — (Monografía del BID; 241)

desempeño operacional de las empresas eléctricas. Las pérdidas eléctricas se encuentran divididas en dos tipos: pérdidas técnicas y no técnicas. Las pérdidas técnicas son aquellas que se dan al transportar la energía eléctrica, debidas al calentamiento natural de los elementos por el paso de la corriente eléctrica, además de la magnetización. Las pérdidas técnicas pueden reducirse hasta cierto punto, con inversiones en infraestructura, no obstante, son inevitables, ya que se deben a procesos físicos.

Las pérdidas no técnicas, por su parte, pueden clasificarse en tres tipos; 1) Robo, hecho por el cual un usuario realiza conexiones ilegales a la red eléctrica y consume la electricidad sin pagarla; 2) Fraude, son aquellas causadas por mano humana que al manipular medidores y/o cableados reduce la lectura del consumo de energía dando como resultado una factura inferior por el servicio consumido y 3) Errores en facturación y medición, los cuales pueden ser causados por una lectura errónea del consumo por causas humanas, medidores antiguos o mal calibrados, falta de equipos de medición o por el mantenimiento deficiente de equipos causando una lectura inapropiada de los consumos o incluso por errores en los sistemas comerciales.

Tradicionalmente los robos de electricidad se identifican a través de inspecciones con cuadrillas en áreas donde se presume que hay tendencia al fraude, ya sea con base a información histórica (estadísticas de consumo), o mediciones realizadas a nivel de subestaciones, donde se evidencia descuadre entre consumo de energía y facturación. Algunas desventajas de estos métodos son: 1) alta dependencia humana en el proceso de detección de fraudes; 2) la incapacidad de procesar grandes cantidades y/o diferentes tipos de datos; 3) la baja tasa de asertividad en las inspecciones, debida a la poca precisión de las estimaciones, lo que se traduce en un alto costo para la detección de fraudes con cuadrillas.

Con el avance tecnológico han surgido nuevas tecnologías como los medidores electrónicos, el internet de las cosas (IoT), la inteligencia artificial (AI) y el aprendizaje de maquina



(Machine Learning - ML). Estas tecnologías están siendo utilizadas en diversas industrias para problemas similares. Por ejemplo, son usadas como herramientas en la detección de fraudes financieros. Dado el éxito en esta área, también se han implementado herramientas que usan dichas tecnologías en la industria eléctrica, específicamente, para la identificación del robo de energía.

El propósito de esta nota técnica es presentar un enfoque de detección de fraudes eléctricos utilizando consumos mensuales de energía y técnicas de aprendizaje automático. Así nace “Energizados” una solución basada en aprendizaje automático y visualizaciones interactivas que ayuda a detectar posibles fraudes con mayor precisión y permite planificar la fiscalización/regularización de manera más eficiente.

La nota técnica está organizada de la siguiente manera. La primera sección contiene la introducción del trabajo. La segunda sección contiene información sobre los trabajos relacionados. La tercera y cuarta sección analizan la información y metodologías implementadas para la identificación de robos. La quinta sección discute los resultados y la sección final contiene las conclusiones y el trabajo futuro.

## **ESTADO DEL ARTE**

---

El problema de detección de pérdidas no técnicas puede ser planteado como un análisis de anomalías para la identificación de fraudes. Es una tarea compleja que requiere de diferentes técnicas para hallar patrones subyacentes que ayuden a identificar este tipo de problema. La detección y reducción de pérdidas es un problema que representa altos costos para las compañías de servicios públicos (pérdidas económicas, infraestructura y humanas), en

consecuencia, durante los últimos años se han realizado diversas investigaciones y avances en este campo usando aprendizaje automático.

Durante el desarrollo de “Energizados” se analizaron diferentes estudios que corresponden al estado del arte e investigaciones relevantes aplicables a esta problemática. No obstante, debe mencionarse que esta es un área que es bastante activa, por lo que se continúan desarrollando algoritmos y aplicaciones para apoyar en la detección de pérdidas técnicas y no técnicas.

Los estudios analizados examinan el consumo eléctrico de los usuarios para extraer patrones que ayudan a entender e identificar anomalías. En general los investigadores optaron por algoritmos de clasificación ya que pudieron disponer de datos etiquetados, es decir, un distintivo en la información para identificar los consumos normales y los irregulares. Entre los algoritmos usados se encuentran modelos estadísticos, redes neuronales<sup>3</sup>, árboles de decisión<sup>4</sup> y *Supporting Vector Machine*<sup>5</sup> (SVM).

Nagi et al., 2008, proponen el uso de SVM ya que es un algoritmo que es menos propenso a sobre ajustarse (overfitting<sup>6</sup>) para la detección de pérdidas no técnicas usando el consumo promedio diario y normalizando la información. El modelo está basado en detectar los cambios abruptos en los consumos, donde dicho cambio es considerado como una actividad fraudulenta.

En su trabajo, Ford et al., 2014 entrenaron una red neuronal simple llamada multilayer perceptron (MLP) <sup>7</sup> con datos históricos de consumo de energía para estimar el consumo

---

<sup>3</sup> Red de aprendizaje profundo o Red Neuronal: Algoritmo matemático que intenta simular el cerebro humano usando “neuronas” para aprender ayudando a encontrar la combinación de parámetros que mejor se ajusta para solucionar un problema.

<sup>4</sup> Un árbol de decisión es un mapa de los posibles resultados de una serie de decisiones relacionadas.

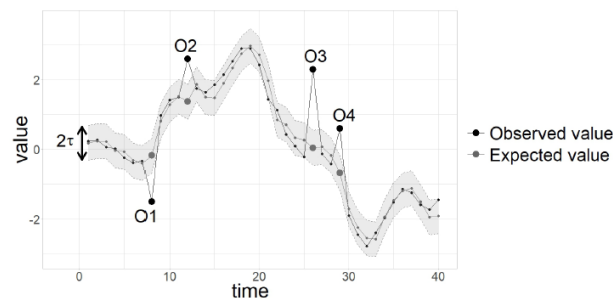
<sup>5</sup> SVM: Algoritmo supervisado que hace uso de un hiperplano para separar clases y clasificarlas correctamente.

<sup>6</sup> Sobreajuste(overfitting): Aumenta la capacidad de memorizar evitando que pueda generalizar con nuevos datos.

<sup>7</sup> Multilayer Perceptron (MLP): Es el tipo de red neuronal más simple el cual consta de entradas, una capa oculta y la capa de salida.

estimado y así poder detectar actividades fraudulentas en la red eléctrica. El sistema propuesto hizo uso de características como el identificador del medidor, consumo cada quince minutos y fecha. Los investigadores tuvieron resultados experimentales en escenarios simulados donde se pudieron capturar el 93.75% de los fraudes. sin embargo, 25% de los que no fueron fraudulentos también fueron clasificados como fraude.

Otra metodología propuesta se encuentra en el trabajo del estado del arte de Blázquez-García et al., 2020, el cual menciona que una forma de hallar patrones anómalos es usando técnicas de detección de anomalías en series de tiempo, el cual consiste en identificar puntos atípicos los cuales son datos que se comportan de manera inusual en un instante de tiempo específico en comparación con los otros valores en la serie de tiempo o sus puntos vecinos. En la Figura 1 se puede observar el concepto de una forma simple.



*Figura 1.* Detección de un punto como anomalía<sup>8</sup>

Para este tipo de detección Blázquez-García et al., 2020, sugieren entrenar un modelo para estimar el consumo esperado normal, luego contrastarlo con el consumo real y analizar la diferencia para decir si es un fraude o no. Por ejemplo  $x_t$  es el consumo real,  $\hat{x}_t$ , es el consumo

---

<sup>8</sup> Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2020, February 11). A review on outlier/anomaly detection in time series data. arXiv.org. Retrieved January 31, 2022, from <https://arxiv.org/abs/2002.04236v1> - pag 6 fig 8

normal estimado. Si la diferencia es mayor que el umbral ( $\tau$ ), entonces la probabilidad de la existencia de un fraude es alta<sup>9</sup>.

Con respecto a la metodología propuesta por Blázquez-García et al., 2020, se observa que la estimación del consumo real puede ser realizado a través de métodos tradicionales como análisis de series temporales y en métodos más novedosos como redes neuronales.

Asimismo, J, Jeyaranjani , & D, Devaraj . 2018, proponen usar modelos basados en similitudes, específicamente K-Means clustering y combinarlo con redes neurales. Los autores sugieren identificar el promedio (centroide) consumo de los usuarios no fraudulentos, luego verificar todos aquellos que estén más cercanos al centroide y clasificarlos como usuarios normales. Una vez se han identificado todos los usuarios normales se entrena una red neuronal para identificar los consumos fraudulentos. En esta investigación se usaron consumos cada 30 minutos en un periodo de cuatro semanas con un resultado teórico medio de 84% de precisión en la detección de fraudes (es decir que el 84% de lo que el modelo dijo que eran fraude realmente lo eran).

Por otra parte, Buzau et al., 2020, aplicaron un algoritmo híbrido de redes neuronales, en el cual se combina un multilayer perceptrón (MLP) y una red con memoria de corto plazo (LSTM)<sup>10</sup> para capturar patrones de una forma más eficiente. Los datos usados para este modelo fueron consumos diarios, datos demográficos y la estación del año. El resultado teórico final fue de un AUC <sup>11</sup>del 83%. Lo cual indica la capacidad discriminante del modelo.

---

<sup>9</sup> Una anomalía está definida por la siguiente ecuación:  $||x_t - \hat{x}_t|| > \tau$

<sup>10</sup> Red Neuronal Memoria de Corto Plazo (LSTM): red de aprendizaje profundo complejo con memoria de corto plazo que permite procesar imágenes y secuencias

<sup>11</sup> AUC (Área Bajo la Curva): Mide la capacidad de un clasificador para distinguir entre clases. Su valor se encuentra entre 0 y 1. Wikimedia Foundation. (2021, September 5). *Receiver operating characteristic*. Wikipedia. Retrieved September 9, 2021, from [https://en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic#Area\\_under\\_the\\_curve](https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve).

Adicionalmente, durante el proceso de investigación se identificaron soluciones de software ya existentes en el mercado, de forma comercial, que han sido implementadas por diferentes compañías. Por ejemplo, CPFL Energía detecta los fraudes energéticos potencializando el análisis de las variaciones en la red (por ejemplo, falta de energía en el medidor, pero sin interrupción del suministro de corriente o con corriente de alimentación, pero sin voltaje, etc.), a partir de esto se forman listas de criterios para identificar el fraude. Así por medio de la inteligencia artificial los algoritmos aprenden comportamientos fraudulentos haciendo que la identificación de los robos sea más asertiva (SIEMENS, 2021). Otro caso es el de la Compañía de Energía de Pernambuco (CELPE) el cual tiene un sistema que usa redes neuronales para identificar las pérdidas comerciales de acuerdo con grupos de usuarios con ciertas características similares conectados a un transformador. En este caso las detecciones de fraudes se realizan a nivel de transformador (Revista de P&D, 2015).

En resumen, los estudios analizados y descritos anteriormente muestran que se están desarrollando soluciones basadas en estas nuevas tecnologías y algoritmos (medidores inteligentes y Machine Learning) para la detección del fraude energético. Se observó en los diferentes artículos la tendencia al uso de datos de alta periodicidad los cuales son obtenidos por medidores inteligentes (Smart meter) y el uso de redes de aprendizaje profundo.

En lo que respecta a “Energizados”, es una propuesta de solución a la problemática de detección de pérdidas eléctricas para empresas de Latinoamérica y el Caribe, donde el uso de medidores inteligentes aún es incipiente. Por lo tanto, a diferencia de los trabajos mencionados, “Energizados” es un ejemplo de cómo este tipo de herramientas ayuda a la detección de fraudes, y como se puede construir una solución integral con un módulo de predicción y visualización que utiliza datos de consumo mensuales (en lugar de horarios, o por minutos) y otras variables como la clase, la actividad, el tipo de conexión, entre otras (esta información se encuentra más detallada en la sección “datos”).

## **ENERGIZADOS**

---

“Energizados”, es una solución que consta de dos módulos: 1) algoritmo de análisis basado en Machine Learning para obtener una probabilidad de fraude para cada unidad consumidora, y 2) interfaz con el usuario de “Energizados”, basado en visualización. El módulo de predicción combina tres modelos, un modelo supervisado<sup>12</sup>, un modelo semi supervisado <sup>13</sup>y un modelo de reglas analíticas. Esta combinación de modelos considera como datos de entrada los consumos mensuales de las unidades consumidoras, basado en datos etiquetados.

### **MODELAMIENTO DE LA INFORMACIÓN**

Para el desarrollo de la solución tres algoritmos fueron implementados para la detección de perdidas no técnicas, específicamente, técnicas de modelamiento que incluyen gradient boosting (modelo supervisado), autoencoders (modelo semi-supervisado) y modelos estadísticos (modelo de reglas analíticas). A continuación, se describen brevemente cada una de estas técnicas.

### **GRADIENT BOOSTING**

Este modelo supervisado usa datos etiquetados previamente. Para “Energizados” se usaron las etiquetas “fraudes” y “no fraudes” ya que la compañía tenía la información desagregada. Así, esta técnica de modelado aprende el comportamiento de aquellos que son fraudes y de aquellos que no lo son. Gradient Boosting <sup>14</sup>hace uso de una técnica de conjunto de árboles de decisión (ensamble). Es decir, las predicciones son hechas por un ensamble de

---

<sup>12</sup> Aprendizaje Supervisado: Aprende IA. <https://aprendeia.com/todo-sobre-aprendizaje-supervisado-en-machine-learning/>.

<sup>13</sup> Aprendizaje Semi-supervisado: Aprende IA. <https://aprendeia.com/todo-sobre-aprendizaje-no-supervisado-en-machine-learning/>.

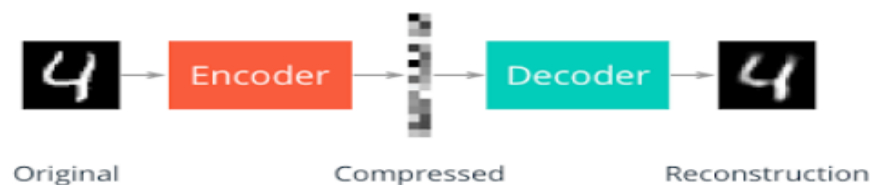
<sup>14</sup> Gradient Boosting: Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neuroinformatics*, 7. <https://doi.org/10.3389/fnbot.2013.00021>

estimadores sencillos. El objetivo del Gradient Boosting es fundamentalmente entrenar de forma simultánea un número alto de árboles de decisión, a este proceso se le llama boosting. Los resultados de esta técnica son mejores que el uso de árboles de decisión simples, que es otra técnica que podría utilizarse en este problema.

## AUTOENCODERS

Este modelo semi-supervisado, llamado autoencoders <sup>15</sup> es un tipo de red neuronal<sup>16</sup> en el cual sólo es necesario conocer los usuarios que no han cometido fraudes. Este tipo de red neuronal intenta reproducir los datos de entrada (los cuales en el caso de “Energizados” son los consumos eléctricos) en su salida final. Es decir, la entrada y la salida deben ser lo más parecida posible, por ejemplo, la Figura 2, muestra el ingreso de la imagen de un número y como resultado se obtiene la reconstrucción de la misma imagen.

En este proceso de reconstrucción el modelo estima un error (medido como la diferencia entre la entrada y la salida), el cual es utilizado para definir, en este caso, si existe o no un fraude. Si el error es bajo quiere decir que el modelo pudo reconstruir sin problemas la entrada. Por el contrario, si el error es alto quiere decir que el modelo no pudo reconstruir la entrada satisfactoriamente, lo que para “Energizados” indicaría una anomalía en los datos de entrada, y por ende un potencial fraude.



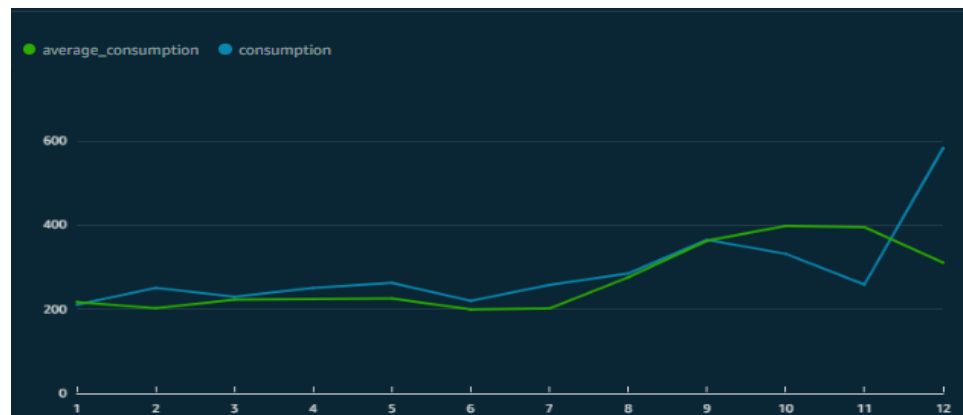
*Figura 2. Autoencoder*

<sup>15</sup> Autoencoder: Bank, D., Koenigstein, N., & Giryas, R. (2021, April 3). Autoencoders. arXiv.org. <https://arxiv.org/abs/2003.05991v2>.

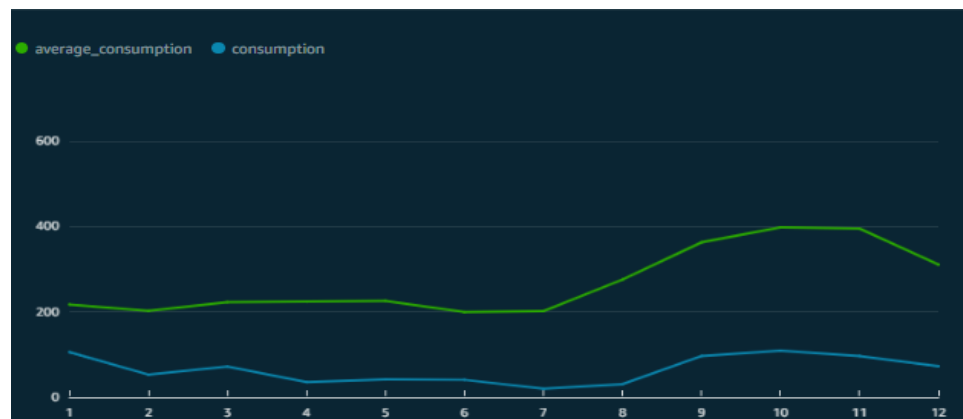
<sup>16</sup> Bishop, 1996 explica de una forma más extensa el funcionamiento de redes neuronales, la aplicabilidad y los tipos.

## REGLAS ESTADÍSTICAS

Este modelo es el más simple, y utiliza la analítica exploratoria para extraer reglas básicas basadas en estadísticas para detectar posibles fraudes. Por ejemplo, podemos obtener el consumo esperado en los datos históricos para luego ser usado como regla de predicción. Es decir, si el consumo real está por debajo del consumo esperado entonces existe un fraude. Estos análisis básicos permiten mejorar el asertividad de los algoritmos de ML y redes neuronales usados en la predicción de fraudes de “Energizados”.



*Figura 4.* Sin Fraude - Consumo promedio Real (azul) vs Consumo Promedio Esperado (verde)



*Figura 5.* Fraude - Consumo promedio Real (azul) vs Consumo Promedio Esperado (verde)



## EXPERIMENTOS

---

La evaluación de “Energizados” fue llevada a cabo con el diseño de experimentos en etapas de desarrollo y también con una prueba piloto en campo con la participación de una empresa privada de distribución de electricidad.

Actualmente la empresa eléctrica con la que se ejecutó la prueba piloto ya cuenta con un proceso para la identificación de fraudes, el cual, consiste en 1) a través de un software comercial, evalúan el consumo mensual de los usuarios, comparándolo con su histórico de consumo, y lanzan un reporte preliminar donde se identifican los usuarios con anomalías (cambios de tendencias de consumo, consumos en cero, etc.), 2) con base en el reporte anterior, realizan una revisión manual de los datos de consumo, para determinar cuáles medidores deben de ser revisados y 3) envían una cuadrilla a campo para la revisión física de las unidades potencialmente fraudulentas detectadas por el analista. Una debilidad de la solución actual es que no permite la retroalimentación, es decir, una vez se verifica si el potencial fraude es realmente fraude, esta información no es retroalimentada al sistema. Por otro lado, el método actual no contiene visualizaciones interactivas y la mayoría de las detecciones de fraudes ocurren en las mismas ubicaciones geográficas.

En comparación, “Energizados” identifica las unidades consumidoras con mayor probabilidad de fraude, con base a un número de unidades a analizar, que el usuario identifica. Estas unidades con alta probabilidad de fraude son luego verificadas en campo. Las ventajas respecto al método tradicional de la empresa es que: 1) Al utilizar la probabilidad dada por el modelo no hace falta una intervención de una persona para generar la lista a fiscalizar; 2) la detección de fraudes en campo permite retroalimentar los algoritmos, que van mejorando su asertividad; 3) “Energizados” cuenta con una interfaz visual, que permite un análisis geográfico

de los fraudes, identificando áreas y optimizando las inspecciones. Los resultados de “Energizados” han mostrado su capacidad para identificar fraudes en áreas que previamente no eran identificadas.

Para la prueba piloto en campo, unidades consumidoras de un municipio fueron evaluadas a través de la solución actual y la solución “Energizados”, luego los resultados fueron analizados. El experimento fue guiado por las dos siguientes preguntas:

- 1. Detección de Anomalías:** ¿Es posible detectar usuarios fraudulentos mejorando los resultados actuales? Con esta pregunta, se intenta demostrar que “Energizados” mejora la precisión en la detección de los fraudulentos. En comparación con el sistema actual de la empresa con la cual se hizo el piloto.
- 2. Intuición de Anomalías:** ¿Es posible que el analista en campo detecte otras unidades consumidoras fraudulentas por proximidad a las recomendadas por “Energizados”? Con esta pregunta se intenta demostrar que “Energizados” puede detectar patrones geográficos a recomendar zonas fraudulentas.

## CONFIGURACIÓN EXPERIMENTAL

Los experimentos computacionales fueron corridos en la nube usando AWS Sagemaker con 4 núcleos y 8 GPU<sup>17</sup>. El código fue escrito en lenguaje Python. Para todos los experimentos las librerías scikit-learn<sup>18</sup>, Tensorflow Keras<sup>19</sup> y TSFEL<sup>20</sup> fueron usadas.

---

<sup>17</sup> GPU: ES un chip que permite procesar grandes cantidades de datos simultáneamente lo que la hace un imprescindible en algoritmos que usan aprendizaje automático

(<https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html>)

<sup>18</sup> scikit-learn: <https://scikit-learn.org/stable/>

<sup>19</sup> Tensorflow: <https://www.tensorflow.org/>

<sup>20</sup> TSFEL: <https://tsfel.readthedocs.io/en/latest/index.html>

## DATOS

El análisis ejecutado en esta nota técnica está basado en los consumos de energía correspondientes a 36 meses de 265.954 unidades consumidoras pertenecientes a un estado de Brasil.

El conjunto de datos contiene 202.730 observaciones etiquetadas como normales y 63.224 observaciones etiquetadas como unidades consumidoras fraudulentas. La periodicidad de la información es consolidada mensual (una por mes). Asimismo, la información en el conjunto de datos incluye la siguiente información: id del medidor, situación del medidor (conectado o desconectado), municipio, clase (comercial, residencial, industrial, rural, entre otros), fase (indica si el tipo de conexión es monofásico, bifásico o trifásico), subclase (indica la subclase de la clase, por ejemplo, restaurantes, bar, etc.), consumos. Finalmente, el indicativo de fraude donde cero “0” indica sin fraude y uno “1” indica fraude.

Cabe resaltar que la base de datos se le tuvo que aplicar una etapa de limpieza, es decir que se le tuvo que realizar un preprocesamiento de esta incluyendo, estandarización de caracteres y fechas, llenado de valores faltantes y generación de variables adicionales para el correcto modelado de la información.

Asimismo, dada la poca cantidad de características suministradas, se crearon nuevas variables, derivadas de las ya existentes, para detectar patrones engañosos más fácilmente. Estas variables adicionales pueden ser divididas en tres tipos: 1) Estadísticas: máximo, promedio, mínimo, mediana, son una muestra de los valores estadísticos calculados; 2) Espectrales derivadas de la serie de consumo: distancia de la señal<sup>21</sup>, pendiente de la señal,

---

<sup>21</sup> Señal: se refiere a la serie de consumo de energía por unidad consumidora.

varianza de la señal, etc.; 3) Temporales: autocorrelación entre las variables, entropía, centroides, entre otros.

Es de notar que dada la sensibilidad de la información y los aspectos legales que la atañen, la información usada para el modelamiento, no se encuentra en sitios públicos ni puede ser compartida.

## RESULTADOS

---

Normalmente en proyectos de ML los datos son particionados en dos conjuntos, entrenamiento y validación. Esto con el fin de simular el comportamiento en los datos reales y evaluar el funcionamiento del algoritmo. Por lo tanto, para el enfoque realizado con “Energizados” se seleccionaron el 80% y 20% para entrenamiento y evaluación, respectivamente. Asimismo, los modelos implementados en el prototipo fueron entrenados con unidades consumidoras fraudulentas y no fraudulentas.

Las métricas de evaluación en esta etapa fueron las siguientes: AUC<sup>22</sup>, Precisión<sup>23</sup>, Exhaustividad<sup>24</sup> y F1<sup>25</sup>. La tabla 1 muestra los resultados de los algoritmos en el conjunto de

---

<sup>22</sup> AUC o Área Bajo la Curva: Mide la capacidad de un clasificador para distinguir entre clases. Por tanto, cuanto mayor sea el AUC, mejor será el rendimiento del modelo para distinguir entre las clases fraude y no fraude.

<sup>23</sup> Precisión: Indica el porcentaje de casos positivos detectados. Lo que matemáticamente es  $p = TP / (TP + FP)$ . En otras palabras, de todas las positivas que se han predicho cuantas son realmente positivas.

<sup>24</sup> Exhaustividad: Es la fracción de instancias relevantes que han sido recuperadas. Matemáticamente es:  $R = TP / (TP + FN)$ . Indica, de todas las clases positivas cuantas se predijo correctamente.

<sup>25</sup> F1: Se utiliza para combinar la precisión y las medidas de exhaustividad en un solo valor y usa la media armónica para castigar los valores extremos. Matemáticamente es:  $F1 = 2 * (precisión * exhaustividad / (precisión + exhaustividad))$ . Esto es útil ya que facilita la comparación de resultados combinados de exactitud e integridad entre diferentes soluciones.

validación. Dichos resultados concluyen que los algoritmos pueden capturar los patrones anormales con una precisión del 61<sup>26o</sup>%.

<b>Tabla 1.</b> <b>Matriz de Resultados en Etapa de Desarrollo</b>	
AUC	74%
F1-Score	45%
Precisión	61%
Exhaustividad	35%

Para la etapa de validación en campo un municipio fue evaluado con “Energizados”. Para este municipio se analizaron aproximadamente 13.000 unidades consumidoras de las cuales 159 fueron identificadas como posibles fraudes, de esas el 28% fueron realmente fraudes. El desempeño del prototipo para la detección de fraudes y anomalías en campo se encuentran reportado en la tabla 2.

<b>Tabla 2.</b> <b>Matriz de Resultados por Municipio – Realmente Fraudes</b>	
Municipio 1 - Energizados	28%
Municipio 1 – Software Actual	17%-23%

Por tanto, “Energizados” reportó un incremento de cinco puntos porcentuales respecto al software usado por la compañía actualmente, el cual arroja una precisión entre el 17 y 23 por ciento. Consecuentemente, las cuadrillas obtuvieron también mayor intuición al detectar las anomalías. Es decir, dado que “Energizados” ayuda a detectar fraudes en diferentes puntos de

---

<sup>26</sup> 61%: resultado de evaluación en el conjunto de datos de prueba. Se muestra un resultado inferior a los reportados en la sección “trabajos relacionados”, esto se debe a 3 razones 1) la periodicidad de los datos usados en “Energizados” son menor a los estudiados; 2) los datos usados en “Energizados” son relativamente pocos y 3) los datos para entrenamiento vienen sesgados

la ciudad, esta característica ayudó a las cuadrillas a fiscalizar mayor número de unidades consumidoras anormales, lo cual hoy la herramienta usada por la compañía remite normalmente a los mismos lugares.

## DISCUSION

---

Como se pudo observar en los resultados obtenidos en el piloto, el enfoque realizado con “Energizados” sugiere ser promisorio para la disminución de perdidas no técnicas ya que ha contribuido a realizar a inspecciones de forma más eficiente lo cual induce a una disminución de los costos en las inspecciones. La asertividad de la detección de fraudes en la empresa aumento en 1.65 veces. Además, con la interfaz gráfica proveída a los usuarios finales, “Energizados” permite obtener una visión general que ayuda a la toma decisiones más asertivas para ir a realizar las inspecciones en campo.

También se pudo notar una diferencia entre los valores obtenidos en validación y la prueba real en campo, esto se debe principalmente a la calidad de los datos con la que se entrenó el modelo. Ya que se utilizó información histórica generada por el software actual usado para la detección de fraudes. Además, existe información recolectada en periodo de pandemia COVID-19 (esto está relacionado al concepto de “data drift”<sup>27</sup>). Una posible solución a este problema es el reentrenado periódico de “Energizados” con datos más confiables.

---

<sup>27</sup> Data Drift: Shibsankar Das A Senior Data Scientist @WalmartLabs, Das, S., @WalmartLabs, A. S. D. S., & on, F. me. (2021, July 19). *Best practices for dealing with concept drift*. neptune.ai. Retrieved September 9, 2021, from <https://neptune.ai/blog/concept-drift-best-practices>.

## CONCLUSIÓN

---

El aprendizaje automático (Machine Learning) e Inteligencia Artificial son los términos más escuchados en los últimos años debido a su aplicabilidad en las diferentes industrias. Como es de esperarse, el sector eléctrico no es la excepción. Machine Learning provee ventajas como la automatización donde los algoritmos aprenden continuamente y la intervención humana es mínimamente requerida. Otra ventaja es el mejoramiento continuo, estos algoritmos con el tiempo aprenden a tomar mejores decisiones debido al crecimiento de la información. Por tanto, Machine Learning, al ser capaz de detectar patrones y tendencias en la información puede ayudar a solucionar diferentes problemas que van desde la reducción de costos, predicciones como consumos de energía, mantenimiento predictivo y detección de anomalías como lo hace “Energizados”.

“Energizados” como se mencionó a lo largo del documento es una aplicación que muestra como el problema del fraude podría ser atacado con un enfoque basado en el uso de Machine Learning.. La evaluación del enfoque desarrollado mostró que este alcanzó un mejor resultado ya que captura el 28% de los fraudes en comparación del 23% alcanzado por la herramienta usada en la contraparte.

Adicionalmente, el ejemplo desarrollado demostró que con el uso de datos mensuales es posible capturar patrones anormales que pueden ser señales de fraude. Además, al usar esta periodicidad de consumos hace que el enfoque sea de fácil implementación en países de Latino América y el Caribe.

Cabe recordar que este estudio se realizó específicamente para una compañía de la región por lo que en un futuro se planea implementar este tipo de enfoques en otras compañías pertenecientes a países de LAC.

## **Nota Técnica,** Energizados

Para trabajos futuros, se pretende implementar nuevos enfoques para experimentar con mayor volúmenes de información y periodicidades más altas, por ejemplo, con lecturas cada quince minutos provenientes de medidores inteligentes. Con este tipo de información se estima que los resultados y los hallazgos de las pérdidas no técnicas sean mayores y mejores.

Esta nota técnica muestra que es posible diseñar modelos que permitan procesar grandes cantidades de datos e incrementar la precisión en los resultados arrojados, proporcionando a las empresas que proveen energía una oportunidad de capturar fraudes y evitar riesgos que este delito puede ocasionar.

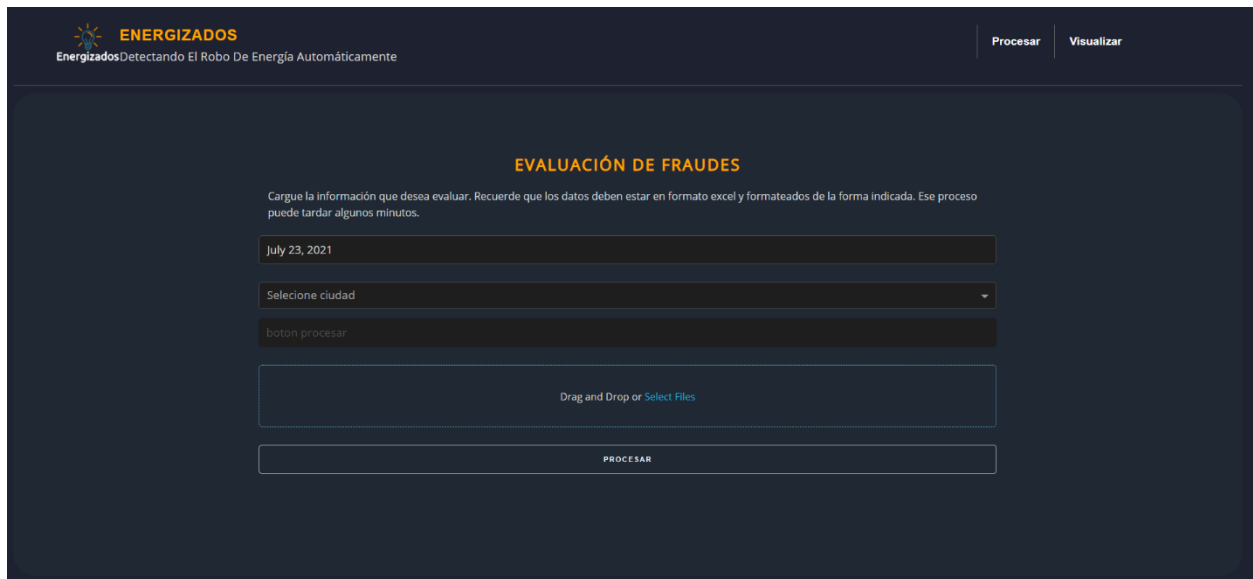


## ANEXOS

---

### HERRAMIENTA DE VISUALIZACIÓN

“Energizados” es una aplicación que permite no solo detectar las unidades consumidoras con un patrón fraudulento, si no, también a tener visualizaciones que ayudan a la toma de decisiones más asertivas y visuales. En la Figura 3 se puede observar cómo es el ingreso de información para analizar los datos y en la Figura 4 se puede observar el análisis de una forma visual he interactivo.



The screenshot shows the 'Energizados' application interface. At the top left is the logo 'ENERGIZADOS' with the tagline 'Energizados Detectando El Robo De Energía Automáticamente'. At the top right are two buttons: 'Procesar' and 'Visualizar'. The main section is titled 'EVALUACIÓN DE FRAUDES' in orange. Below the title is a text instruction: 'Cargue la información que desea evaluar. Recuerde que los datos deben estar en formato excel y formateados de la forma indicada. Ese proceso puede tardar algunos minutos.' There are three input fields: a date field containing 'July 23, 2021', a city selection dropdown menu labeled 'Seleccione ciudad', and a button labeled 'boton procesar'. Below these is a large rectangular area with the text 'Drag and Drop or Select Files'. At the bottom is a button labeled 'PROCESAR'.

*Figura 3. Evaluación de Fraudes*

## Nota Técnica, Energizados

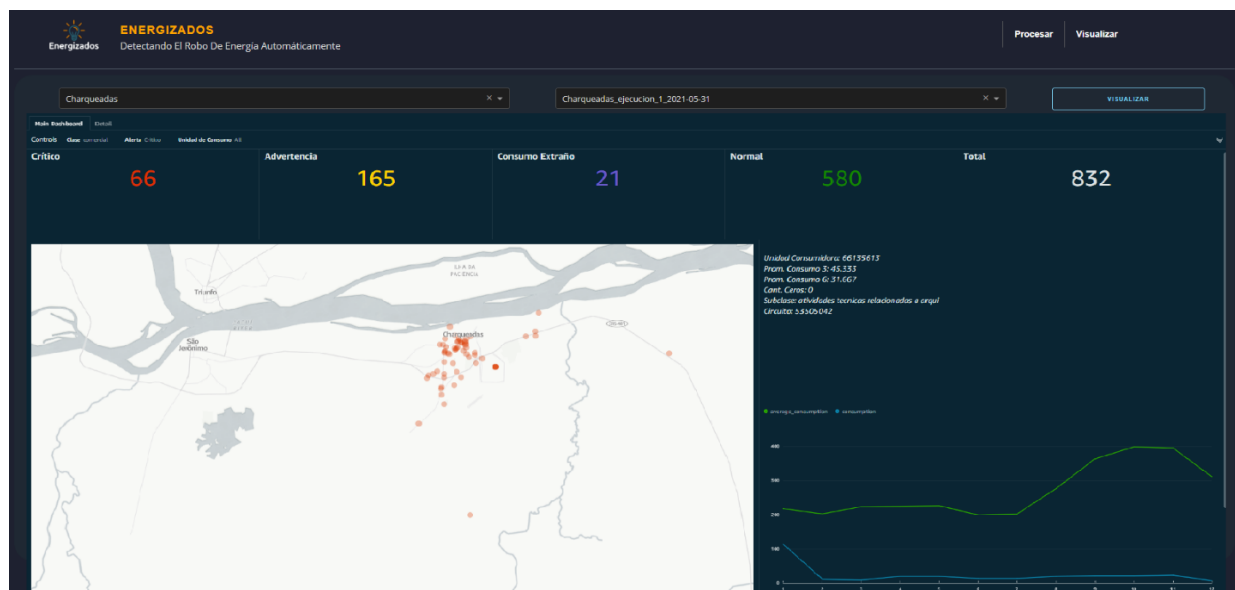


Figura 4. Tablero con Mapa Interactivo para Detección de Fraudes

## REFERENCIAS

---

- Jiménez, R., Serebrisky, T., & Mercado, J. (2014). *Power Lost Sizing Electricity Losses in Transmission and Distribution Systems in Latin America and the Caribbean*. iadb. <https://publications.iadb.org/publications/english/document/Power-Lost-Sizing-Electricity-Losses-in-Transmission-and-Distribution-Systems-in-Latin-America-and-the-Caribbean.pdf>.
- Nagi, J., Mohammad, A. M., Yap, K. S., Tiong, S. K., & Ahmed, S. K. (2008). Non-Technical Loss analysis for detection of electricity theft using support vector machines. *2008 IEEE 2nd International Power and Energy Conference*. <https://doi.org/10.1109/pecon.2008.4762604>
- Ford, V., Siraj, A., & Eberle, W. (2014). Smart grid energy fraud detection using artificial neural networks. *2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*. <https://doi.org/10.1109/ciasg.2014.7011557>
- Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2020). A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, 54(3), 1–33. <https://doi.org/10.1145/3444690>
- J, Jeyaranjani, & D, Devaraj. (2018). Machine Learning Algorithm for Efficient Power Theft Detection using Smart Meter Data. *International Journal of Engineering & Technology*, 7(3.34), 900-904. doi:<http://dx.doi.org/10.14419/ijet.v7i3.34.19585>
- SIEMENS, S. I. E. M. E. N. S. (2021, February 1). *Inteligência artificial e machine LEARNING para eliminar "gatos" NA REDE*. CanalEnergia. <https://www.canalenergia.com.br/noticias/53161937/inteligencia-artificial-e-machine-learning-para-eliminar-gatos-na-rede>.
- Revista de P&D, R. de P. D. (2015, November 25). *Revista de P&D: Celpe Identifica PERDAS DE ENERGIA com inteligência artificial - AGÊNCIA Nacional DE energia Elétrica - ANEEL*. Go to ANEEL. [http://www.aneel.gov.br/home?p\\_p\\_id=101&p\\_p\\_lifecycle=0&p\\_p\\_state=maximized&p\\_p\\_mode=view&\\_101\\_struts\\_action=%2Fasset\\_publisher%2Fview\\_content&\\_101\\_returnToFullPageURL=%2F&\\_101\\_assetEntryId=14555581&\\_101\\_type=content&\\_101\\_groupId=656877&\\_101\\_urlTitle=revista-de-p-d-celpe-identifica-perdas-de-energia-com-inteligencia-artificial&inheritRedirect=true](http://www.aneel.gov.br/home?p_p_id=101&p_p_lifecycle=0&p_p_state=maximized&p_p_mode=view&_101_struts_action=%2Fasset_publisher%2Fview_content&_101_returnToFullPageURL=%2F&_101_assetEntryId=14555581&_101_type=content&_101_groupId=656877&_101_urlTitle=revista-de-p-d-celpe-identifica-perdas-de-energia-com-inteligencia-artificial&inheritRedirect=true).
- Buzau, M.-M., Tejedor-Aguilera, J., Cruz-Romero, P., & Gomez-Exposito, A. (2020). Hybrid Deep Neural Networks for Detection of Non-Technical Losses in Electricity Smart Meters. *IEEE Transactions on Power Systems*, 35(2), 1254–1263. <https://doi.org/10.1109/tpwrs.2019.2943115>
- Yandex, Y. (2018, December 18). *CatBoost enables Fast gradient boosting on decision trees Using GPUs*. CatBoost. <https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus>.

Bishop, C. (1996, January 1). Neural Networks: A Pattern Recognition Perspective. Microsoft Research. <https://www.microsoft.com/en-us/research/publication/neural-networks-a-pattern-recognition-perspective/>.

Innovation, A. (2019). *Representación de Red Neuronal*. Atria Innovation. Atria Innovation. <https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/>. Qué son las redes neuronales y sus funciones