UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Junio 2019

Detección de fraude de energía: comparación entre un modelo de lógica difusa y un MLG logit

Energy fraud detection: comparison between fuzzy logic model and GLM logit

Nasser Stefan De La Pava Roys[a]
nasserdelapava@usantotomas.edu.co

Gil Robert Romero[b]
gilromero@usantotomas.edu.co

Resumen

El fraude de energía es un fenómeno al cual se enfrentan las empresas prestadoras del servicio y que deja pérdidas millonarias para el sector, sin embargo, para mitigar estos fraudes se emplean diferentes técnicas desde el ámbito estadístico y las ciencias de la computación con el fin de tomar acciones preventivas frente a los posibles fraudes.

El presente trabajo tiene como objetivo la aplicación de un modelo de lógica difusa, que permita establecer la ocurrencia de posible fraude de energía eléctrica, el cual permita tomar acciones tempranas que mitiguen este fenómeno, el modelo de lógica difusa es comparado con un modelo logístico.

***Palabras clave*:** Estadística, Minería de datos, Aprendizaje Automático, Algoritmos, Patrones, Fraude, Lógica Difusa.

Abstract

Energy fraud is a phenomenon faced by the companies providing the service and that leaves millionaire losses for the sector, however, to mitigate these frauds are used different techniques from the statistical field and computer science in order to take preventive actions against possible fraud.

The objective of this work is to apply a fuzzy logic model that allows establish the occurrence of possible electric power fraud, which allows taking actions early that mitigate this phenomenon, the fuzzy logic model is compared to a model logistic.

***Keywords*:** Statistics, Data Mining, Machine Learning, Algorithms, Fraud, Patterns, Fuzzy Logic.

[a]Estudiante Estadística Universidad Santo Tomás
[b]Docente Estadística Universidad Santo Tomás

Introducción

La prestación del servicio de energía eléctrica en Colombia tiene sus orígenes a finales del siglo XIX, cuando miles de habitantes de la capital del país vieron cómo la luz alumbraba a un centenar de lámparas en las calles; en principio este servicio era privilegiado a clases sociales altas, funcionarios del gobierno y empresas de tipo industrial que podían costear dicho servicio. Según cifras del Sistema de Información Eléctrico Colombiano (SIEL) [1] para el año 2016 se estima un 97.02 % en cobertura del servicio eléctrico a nivel Colombia. Además, según estadísticas del Banco Mundial [2] el consumo de energía crece exponencialmente al mismo nivel de la población mundial.

El fraude de energía en Colombia es más común de lo que se podría pensar, es un fenómeno por el cual las empresas del sector energético pierdan miles de millones de pesos. No siempre el fraude implica que existan manos humanas detrás de estos eventos, este suceso también puede presentarse por personal no capacitado, defectos en medidores y componentes, logística en redes y calidad de su infraestructura entre otras.

A través de los años la *estadística* ha desarrollado técnicas robustas y óptimas que permiten tener la descripción de los patrones de un conjunto de datos y a partir de estos construir modelos probabilísticos con el fin de establecer la ocurrencia de un evento objeto de estudio, este tipo de técnicas no solo ayudan a establecer la probabilidad de ocurrencia, también se pueden establecer cuáles son las variables que mayor impacto tienen sobre el aumento o la disminución de la probabilidad de un evento, dentro de estos modelos encontramos los *lineales generalizados*, los cuales de acuerdo a la función de enlace y al tipo de variable objeto de estudio encontramos los modelos logit, probit y loglog.

Durante este último siglo, gracias a los avances computacionales y el aporte de las matemáticas aplicadas, se han venido desarrollando técnicas como el *automático* en inglés *machine learning*, las cuales a través de la aplicación de diferentes algoritmos buscan encontrar patrones dentro de conjuntos de datos que ayuden a identificar de una mejor manera comportamientos, en algunas ocasiones estos algoritmos pueden llegar a tener un mejor desempeño que los modelos desarrollados en estadística, algunos de estos algoritmos son las *Neuronales Artificiales*, *Máquinas de Soporte Vectorial*, *Extreme Gradient Boosting* y algoritmos desarrollados a partir de lógica difusa, entre otros.

Según Bolton & Hand (2002) [6] la estadística y el aprendizaje automático proporcionan información correlacionada y confiable; son de gran aplicación para detectar actividades fraudulentas como son el blanqueo de dinero, tarjetas de crédito, comercio electrónico, telecomunicaciones, suplantación y fraude en redes privadas entre otras.

Este documento está estructurado de la siguiente forma: en la sección 1, objetivos y justificación, en términos generales se busca comparar la eficiencia de un modelo de lógica difusa con un modelo clásico lineal generalizado logit. En la sección 2, marco teórico, conceptos fundamentales: proceso de la minería de datos, estadística, minería de datos, aprendizaje automático, evaluación y selección del modelo, componentes principales para datos categóricos, lógica difusa y modelos lineales generalizados. En la sección 3, la metodología donde se aplican las técnicas para la transformación del conjunto de datos objeto de estudio, el desarrollo e implementación de los modelos *logit* y *difusos*. En la sección 4, los resultados de los modelos *logit* y *difusos*, sus métricas de clasificación, evaluación y análisis. En la sección 5, se describen las conclusiones. En la sección 6, propuestas para trabajos futuros de investigación aplicando teoría de lógica difusa.

[1] `http://www.siel.gov.co`
[2] `https://datos.bancomundial.org/indicador/EG.ELC.ACCS.ZS?end=2016&start=1990&type=shaded&view=map`

1. Objetivo

1.1. Objetivo general

Identificar y detectar un fraude en consumo de energía eléctrica a un marco de clientes de una empresa de Colombia, a partir de un modelo de lógica difusa.

1.2. Objetivos específicos

1. Establecer las variables que tienen un mayor impacto en la detección de posible fraude en el consumo de energía.
2. Comparar la eficiencia con métricas de validación de aprendizaje automático para los modelos de lógica difusa frente a un MLG logit.

1.3. Justificación

Si bien los modelos estadísticos clásicos, han sido una alternativa eficiente y robusta para la detección de problemas similares al planteado en la sección 2, la actualidad y la dinámica mundial demandan hacer uso de nuevas técnicas a través de diferentes algoritmos y metodologías que puedan llegar a dar una mejor solución a la comúnmente planteada.

Teniendo en cuenta los antecedentes mencionados, con este proyecto se busca aplicar un modelo a partir de la lógica difusa y comparar los resultados contra un modelo logit, esto con el fin de encontrar alternativas más eficientes o complementarias para la detección de fraudes o anomalías en energía eléctrica.

2. Marco teórico

2.1. Conceptos fundamentales

2.2. Proceso de la minería de datos

Brett (2019) [1] explica el proceso CRISP-DM *Cross Industry Standard Process for Data Mining* por sus siglas en inglés como un modelo de procesos de la minería de datos que explican una serie de seis fases iterativos y bidireccionales no estrictos del cómo se pueden abordar e implementar (figura 1). Además, el descubrimiento de información a partir del proceso CRISP-DM a un conjunto de datos estructurados o no estructurados y de diferentes fuentes son abordados con metodologías estadísticas y procesos computacionales para la exploración, identificación, medición y análisis de patrones de observaciones inusuales.

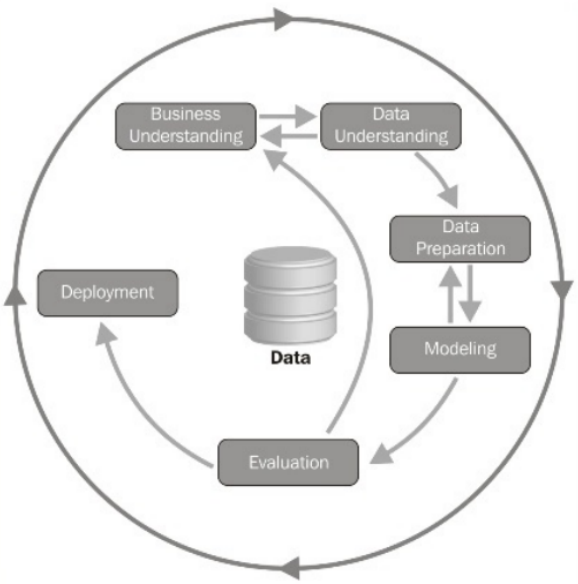

Figura 1: *Modelo de proceso CRISP DM* Brett (2019) [1]

A continuación, se describen cada una de las fases en que se divide CRISP-DM.

1. **Comprensión del negocio**: Esta tarea incluye la determinación de los objetivos del negocio, la evaluación de la situación actual, el establecimiento de objetivos de la minería de datos, y el desarrollo de un plan.
2. **Comprensión de datos**: esta tarea evalúa los requisitos de los datos e incluye la recopilación de datos inicial, descripción de los datos, exploración de datos y la verificación de la calidad de los datos.
3. **Preparación de datos**: una vez disponibles, los recursos de datos se identifican en el último paso. Luego, los datos deben ser seleccionados, limpiados y luego incorporados en la forma y formato deseados.
4. **Modelado**: la visualización y el análisis de conglomerados son útiles para el análisis inicial. Las reglas de asociación iniciales pueden desarrollarse aplicando herramientas como la inducción de reglas generalizadas. Esta es una técnica de extracción de datos para descubrir el conocimiento representado como reglas para ilustrar los datos en la vista de la relación causal entre los factores

condicionales y una decisión / resultado dado. Los modelos adecuados a la También se puede aplicar el tipo de datos.

5. **Evaluación**: los resultados deben evaluarse en el contexto especificado por los objetivos del negocio en el primer paso. Esto conduce a la identificación de nuevas necesidades y, a su vez, vuelve a las fases anteriores en la mayoría de los casos.

6. **Implementación**: la minería de datos se puede utilizar para verificar hipótesis mantenidas previamente o para conocimiento.

2.2.1. Estadística

La Estadística se encarga de la recolección, acopio y análisis de información para optimizar los procesos de toma de decisiones, utiliza un conjunto de funciones matemáticas que describen su función y distribución de probabilidad cuyos parámetros no son desconocidos. (González, 2018 [3])

2.2.2. Minería de datos

La minería de datos (en inglés data mining) es una técnica robusta de la estadística que emplea algoritmos de aprendizaje para explorar, extraer y descubrir patrones inusuales a partir de un determinado marco de información. Han et al. (2014) [10] define a un patrón interesante cuando representa una unidad de medida de escala ya sea objetiva o subjetiva, además, se pueden emplear para un nuevo proceso de explotación de conocimiento.

2.2.3. Aprendizaje automático

El aprendizaje automático (en inglés machine learning) investiga cómo las computadoras y la inteligencia artificial permite a desarrollar técnicas para aprender o mejorar su rendimiento en función de los datos, y optimizar su proceso en el descubrimiento de nuevos patrones dado a su aprendizaje. (Han et al, 2014 [10])

2.3. Evaluación y selección del modelo

El objetivo de esta etapa es evaluar el desempeño, las debilidades y fortalezas de los modelos probabilísticos, para ello, se utiliza una métrica para seleccionar el modelo probabilístico que predice los mejores resultados en función de criterios como: precisión, área bajo la curva roc y f beta score. (González, 2018 [3])

1. **Matriz de Confusión** : dada por un número de m clases ordenadas en filas y columnas, es simétrica por contener las mismas categorías en filas y columnas, las columnas corresponden a los resultados arrojados por el modelo de pronóstico mientras las filas representan la clasificación real de los individuos, sobre las casillas de la diagonal se identifican a los individuos bien clasificados por el modelo. La figura 2 muestra una matriz de confusión para un problema de clasificación binaria, la diagonal principal muestra los verdaderos positivos (VP) y verdaderos negativos (VN), la clasificación errada está dada por los falsos positivos (FP) y falsos negativos (FN).

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

Figura 2: *Matriz de Confusión*

a) Verdaderos Positivos (VP): Cantidad de casos No fraudulentos que fueron clasificados correctamente por el modelo.

b) Verdaderos Negativos (VN): Cantidad de casos Sí fraudulentos que fueron clasificados correctamente por el modelo

c) Falsos Positivos (FP): Cantidad de casos No fraudulentos que fueron clasificados como si fraudulentos.

d) Falsos Negativos (FN): Cantidad de casos Sí fraudulentos que fueron clasificados como no fraudulentos.

2. **Exactitud** : es un indicador evalúa la capacidad del modelo de clasificar correctamente los casos positivos y negativos las categorías, resultados que parte de los valores clasificados en la matriz de confusión, se calcula como los valores clasificados correcta mente de la diagonal principal o traza de la matriz verdaderos positivos y verdaderos negativos sobre el total de las categorías.

$$Exactitud = Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

3. **Precisión** : es la probabilidad promedio de recuperación relevante de información, esta estadística pretende proporcionar una indicación de que interesantes y relevantes son los resultados de un modelo, puede verse como una medida de exactitud o calidad, la alta precisión significa que un algoritmo que arrojó resultados sustancialmente más relevantes que los irrelevantes.

$$Precision = \frac{VP}{VP + FP} \tag{2}$$

4. **Curva Roc** : proviene de la teoría de detección de señales que se desarrolló durante la Segunda Guerra Mundial para el análisis de imágenes de radar. Una curva roc para un modelo muestra la compensación entre la tasa de verdaderos positivos (TVP) y la tasa de falsos positivos (TFP).

$$Sencibilidad = TVP = \frac{VP}{P} = \frac{VP}{(VP + FN)} \tag{3}$$

$$TFP = \frac{FP}{N} = \frac{FP}{(FP + VN)} \tag{4}$$

$$Especificidad = \frac{VN}{N} = \frac{VN}{(FP + VN)} = 1 - TFP \tag{5}$$

La figura 3 es una representación bidimensional que representa en su eje vertical la proporción de valores positivos de sensibilidad y el eje horizontal la proporción de valores falsos positivos de

especificidad, una recta de división desde el punto de origen hasta (1) en ambos ejes del plano; la curva roc o también llamada AUC área bajo la curva clasifica estadísticamente bajo hipótesis el mejor modelo que obtenga dado a su aprendizaje y entrenamiento, para poder interpretar estas señales se han establecido intervalos de calificación:

a) 0.50 - 0.60 No hay discriminación
b) 0.61 - 0.70 Pobre
c) 0.71 - 0.80 Aceptable
d) 0.81 - 0.90 Bueno
e) 0.91 - 1.00 Excelente

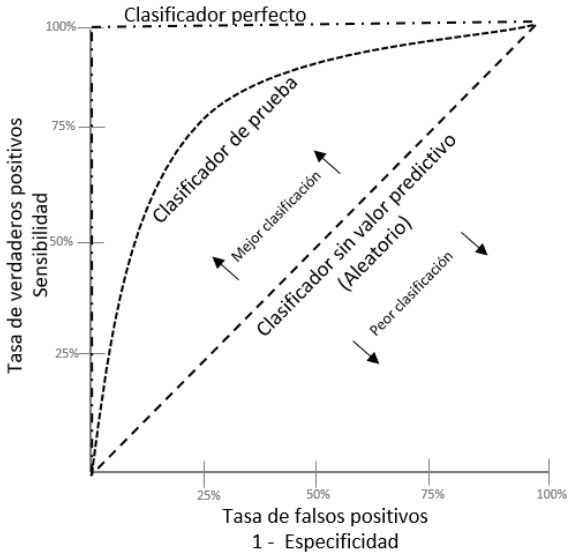

Figura 3: *Curva Roc, fuente propia*

5. **F Beta Score** : La estadística F_β es la media armónica de Precisión y Recall, La medida F_β es una medida ponderada de precisión y recuperación. Se asigna β veces más peso para recordar en cuanto a la precisión.

   La puntuación $F_{\beta=1}$ es el promedio armónico de la Precisión y Recall , donde un puntaje F_1 alcanza su mejor valor en 1 (precisión perfecta y recuperación) y el peor en 0.

   La Precisión y el Recall a menudo se fusionan en una única estadística, llamada F Beta Score o F - Measure por (*Rijsbergen, 1979*), dada por

$$F_\beta = (\beta^2 + 1)\ \frac{Precision\ Recall}{\beta^2\,(Precision + Recall)} \tag{6}$$

   Donde $0 \leqslant \beta \leqslant 1$, donde β es un número real no negativo, y controla la importancia relativa de Recall y la Precisión. (Torgo, 2011 [12])

De acuerdo con la literatura cada una de las medidas explicadas anteriormente pueden ser aplicadas para diferentes tipos de modelos de clasificación, bien sean de tipo binaria o clasificación múltiple es decir varias categorías. Sin embargo, de acuerdo con los objetivos planteados en este trabajo, las medidas que se van a tomar para la evaluación de los modelos logit y de lógica difusa son la exactitud en inglés accuracy y el área bajo la curva, esto con el fin de tener una visión global del rendimiento de cada uno de los modelos.

2.4. Componentes principales para datos categóricos

El método de análisis de componentes principales categóricos (ACP_{Cat}), al igual que su homólogo para variables continuas, es una técnica exploratoria de reducción de dimensiones de una base de datos, donde se incorporan variables ordinales y/o nominales de la misma manera que las numéricas. El (ACP_{Cat}) permite observar relaciones entre las variables originales, entre los casos (registros) y entre ambos (variables y casos). Además, se pueden analizar variables con su nivel de medición. Cuando existe relación no lineal entre las variables, pueden especificarse también otros niveles de análisis, de manera que estas relaciones pueden manipularse de manera más efectiva.

A continuación, se describe matemáticamente el análisis de componentes principales categóricos (ACP_{Cat}), según (Navarro et al. 2009) [7], se supone que se tiene una matriz de datos $H_{n\times m}$, la cual consiste en las puntuaciones observadas de n casos en m variables. Cada variable puede ser denotada como la $j-sima$ columna de $H; h_j$ como un vector $n \times 1, conj = 1, ..., m$. Si las variables h_j no tienen nivel de medición numérico, o se espera que la relación entre ellas no sea lineal, se aplica una transformación no lineal. Durante el proceso de transformación, cada categoría obtiene un valor escalado óptimo, denominado cuantificación categórica. ACP_{Cat} puede ser desarrollado minimizando la función de perdida mínima cuadrática en la que la matriz de datos observados H es reemplazada por una matriz $Q_n \times m$, que contiene las variables transformadas $q_j = \phi_j(h_j)$. En la matriz Q, las puntuaciones observadas de los casos se reemplazan por las cuantificaciones categóricas. El modelo ACP_{Cat} es igual al modelo del ACP, capturando las posibles no linealidades de las relaciones entre las variables en las transformaciones de las variables. Se comenzará explicando cómo el objetivo del ACP se alcanza por el ACP_{Cat} minimizando la función de pérdida, y por tanto se mostrará cómo esta función se amplía para acomodar las ponderaciones de acuerdo con los valores ausentes, ponderaciones por casos, y trasformaciones nominales múltiples.

A las puntuaciones de los casos en las componentes principales obtenidas a partir del ACP se le denominan puntuaciones de las componentes (puntuaciones de los objetos en ACP_{Cat}). ACP intenta mantener la información en las variables tanto como sea posible en las puntuaciones de las componentes. A las puntuaciones de las componentes, multiplicadas por un conjunto de ponderaciones óptimas, se les denominan saturaciones en componentes, y tienen que aproximar los datos originales tan cerca como sea posible. Usualmente en ACP, las puntuaciones de las componentes y las saturaciones en componentes se obtienen de una descomposición en valor singular de la matriz de datos estandarizada, o de una descomposición en valores propios de la matriz de correlación. Sin embargo, el mismo resultado puede obtenerse a través de un proceso iterativo en el que se minimiza la función de pérdida mínima cuadrática. La pérdida que se minimiza es la pérdida de la información debido a la representación de las variables por un número pequeño de componentes: en otras palabras, la diferencia entre las variables y las puntuaciones de las componentes ponderadas a través de las saturaciones en componentes. Si $X_n \times p$ se considera la matriz de las puntuaciones de las componentes, siendo p el número de las componentes, y si $A_m \times p$ es la matriz de las saturaciones en componentes, siendo su $j-sima$ fila indicada por a_j, la función de perdida que se usa en el ACP para la minimización de la diferencia entre los datos originales y las componentes principales puede ser expresada como:

$$L(Q, X, A) = n^{-1}\Sigma_j\Sigma_n(q_ij - \Sigma_s x_i s a_j s)^2 \tag{7}$$

La función de pérdida está sujeta a un número de restricciones. Primero, las variables transformadas son estandarizadas, a fin de que $q_j'q_j = n$, tal restricción se necesita para resolver la indeterminación entre q_j y a_j en el producto escalar $q_j a_j'$, esta normalización implica q_j que contenga z-scores y garantice que las saturaciones en componentes en a_j estén correlacionadas entre las variables y las componentes. Para evitar la solución trivial $A = 0$ y $X = 0$, las puntuaciones de los objetos se limitan y se requiere que:

$$X'X = nI \tag{8}$$

donde I es la matriz identidad. Se necesita también que las puntuaciones de los objetos estén centradas, por lo tanto:

$$1'X = 0 \tag{9}$$

donde el 1 representa el vector unidad.Esto implica que las columnas de X (componentes) son z-scores ortonormales: su media es cero, su desviación estándar es uno, y están incorrelacionadas. Para el nivel de escala numérica, $q_j = \Phi_j(h_j)$ implica una transformación lineal, o sea, la variable observada h_j es simplemente transformada en z-scores. Para los niveles no lineales (nominal, ordinal, spline), $q_j = \Phi_j(hj)$ denotan una transformación acorde con el nivel de medición seleccionado para la variable j.

2.5. Análisis de homogeneidad

El análisis de homogeneidad $HOMALS$ *homogeneity analysis by means of alternating least squares* por sus siglas en inglés, permite generar una cuantificación sobre las categorías presentes en las variables nominales, haciendo un análisis de homogeneidad y posterior aproximación por mínimos cuadrados alternantes. El principal objetivo de $HOMALS$ es describir las relaciones existentes entre dos o más variables nominales, creando un espacio de dimensiones pequeñas las cuales contienen las categorías de cada una de las variables, esta reducción de dimensión tiene un comportamiento similar al del análisis de componentes principales, los registros que pertenecen a la misma categoría son cercanos entre sí, mientras que los registros que pertenecen a diferentes categorías se representan alejados de los demás, cada registro está cercano a los puntos representados por las categorías a las que pertenecen.

El análisis de homogeneidad es más adecuado que el análisis de componentes principales típico cuando no se conservan las relaciones lineales entre las variables, o cuando las variables se miden a nivel nominal. Además, la interpretación del resultado es mucho más sencilla en $HOMALS$ que, en otras técnicas categóricas, como pueden ser las tablas de contingencia y los modelos loglineales. Debido a que las categorías de las variables son cuantificadas, se pueden aplicar sobre las cuantificaciones técnicas que requieren datos numéricos.

2.6. Lógica difusa

De acuerdo a la historia, la lógica difusa tiene sus inicios en la época de los filósofos Platón y Aristóteles, los cuales manifestaban que los comportamientos de una persona, sus ideas y las cosas en general no pueden ser del todo falso o verdadero, si no que existe una escala entre esos dos extremos, siglos después, en el siglo $XVIII$ los filósofos David Hume e Immanuel Kant, retoman estas ideas y concluyen que el razonamiento que adquiere una persona se genera en función de sus vivencias, lo que implica que una persona puede vivenciar experiencias similares sin necesidad de ser las mismas, y esto da como resultado una escala de sentimientos asociados a las vivencias, por ejemplo una vivencia puede ser encontrar dinero en la calle, la experiencia frente a ese primer evento está asociada al estado de ánimo de la persona en ese preciso momento, quizás pueda ser una persona que necesita dinero y el hecho de encontrarlo le puede generar felicidad, sin embargo esta experiencia se puede repetir nuevamente pero esto no implica que la persona se encuentre en el mismo estado de ánimo, quizás esta vez la persona está feliz y va ser indiferente frente al hecho de encontrar dinero y finalmente esta misma persona puede encontrar nuevamente dinero pero esta vez lo encuentra en medio de la noche y el sentimiento que va a tener puede ser el de preocupación, quizás por miedo a que le llegue a suceder algo malo; lo anterior es la conclusión a la que llegaron Hume y Kant, independientemente de que las vivencias se repitan, las experiencias son diferentes lo que hacer ver que la persona va a tener una escala de *calificación* frente a cada experiencia.

Durante el siglo XX, llegó el turno para los matemáticos, en principio el matemático Bertrand Russell expresó que la lógica produce contradicciones y realizó un estudio sobre las ambigüedades del lenguaje concluyendo con exactitud que la ambigüedad es un grado, paralelo a Russell, Ludwing Wittgenstein, estudió los diferentes sentidos que tiene una misma palabra. Éste llegó a la conclusión de que en el lenguaje una misma palabra expresa modos y maneras diferentes.

Para el año 1920 Jan Lukasiewicz desarrollo la primera lógica de ambigüedades, llegó a la conclusión, que los conjuntos tienen un posible grado de pertenencia con valores entre 0 y 1.

Finalmente, y aunque la teoría se sigue desarrollando, en el año 1965 el matemático Lofti Asier Zadeh publicó el ensayo *"Fuzzy Sets"* (Conjuntos Difusos) y es por esta razón que es considerado como el padre del término *borroso o difuso*. la tesis propuesta por Zadeh en 1965, está basada en los desarrollos y conclusiones a las que llegaron algunos de los filósofos y matemáticos mencionados anteriormente, básicamente plantea un formalismo para manejar de forma más eficiente en la imprecisión del razonamiento humano. En 1971 Zadeh publica *"Quantitative Fuzzy Semantics"* ensayo en el cual se observan los elementos formales que dan lugar a la metodología y aplicaciones de la *lógica difusa* tal y como se conoce actualmente. Posterior a la publicación de Zadeh, otros investigadores comienzan a aplicar la *lógica difusa* a diversos procesos y aparecen las primeras aplicaciones de la teoría desarrollada, siendo implementadas en la construcción de controladores para máquinas de vapor, eléctricas y electrónicas, en 1987 la empresa $OMRON$ inicia el proceso de fabricación de controladores en masa, este año es llamado como el *"Fuzzy Boom"*, gracias a la gran cantidad de elementos que se producen aplicando la teoría de la *lógica difusa*. Sin embargo, la aplicación de esta teoría no solo se dio en el área de la ingeniería, 1993 es el año donde la *estadística* entra a aplicar la teoría desarrollada por Zadeh y lo hace a través del *diseño de experimentos*, Fuji, desarrolló un método para el control de inyección química en plantas depuradoras de agua, de forma paralela Takagi y Sugeno desarrollan la primera aproximación para construir reglas difusas a partir de un conjunto de datos de entrenamiento.

En los últimos años, la investigación en el campo difuso ha tenido grandes desarrollos en la construcción de algoritmos matemáticos a partir de un conjunto de datos de entrenamiento, este avance se debe principalmente a la similaridad que tienen las reglas difusas y las redes neuronales artificiales, como se conoce estas últimas son algoritmos que se usan para predicción o clasificación de eventos, identificación de imágenes, entre otros. la relación entre estas dos técnicas ha dado como resultado los sistemas *neuro-fuzzy* los cuales usan métodos de aprendizaje basados en redes neuronales para identificar y optimizar los parámetros del algoritmo y así mismo mejorar las clasificaciones o predicciones de estos.

2.6.1. Conjuntos difusos

Los conjuntos difusos son una generalización de los conjuntos clásicos, estos últimos pueden ser representados mediante la función de pertenencia:

$$\mu_A(x) = \begin{cases} 1 & \text{Sí} \quad x \in A \\ 0 & \text{Sí} \quad x \notin A \end{cases} \tag{10}$$

Las operaciones de unión, intersección, diferencia, negación y complemento están definidas en los conjuntos difusos, además, están asociadas a una función de pertenencia que indica una medida. Los valores que toma un intervalo de lógica difusa son de $[0, 1]$ y se asigna a cada objeto un valor de ese intervalo; según Pérez (2014) [4] los conjuntos difusos pueden ser considerados como una generalización de los conjuntos clásicos como se muestra en la siguiente figura 4

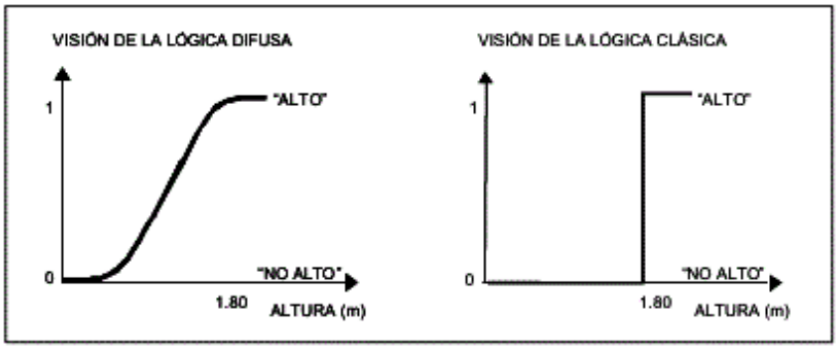

Figura 4: *Conjuntos lógicos, (a) visión de la lógica difusa (b) visión de la lógica clásica* (Pérez, 2014 [4])

2.6.2. Operaciones con conjuntos difusos

Las operaciones básicas entre conjuntos difusos son las siguientes:

a) La unión de dos conjuntos difusos A y B es un conjunto difuso $A \cup B$

b) La intersección de dos conjuntos difusos A y B es un conjunto difuso $A \cap B$ cuya función de pertenencia es $\mu_{A \cap B}(x) = min\{\mu_A(x), \mu_B(x)\}$

c) El conjunto complementario de un conjunto difuso A es el conjunto difuso $\bar{A}$ cuya función de pertenencia es $\mu_{\bar{A}}(x) = 1 - \mu_A(x)$

d) La igualdad de dos conjuntos difusos A y B se define como: $A = B \Leftrightarrow \mu_A(x) = \mu_B(x), \forall_x \in X$

e) La inclusión de un conjunto difuso A en otro B se define como: $A \subseteq B \Leftrightarrow \mu_A(x) \leq \mu_B(x), \forall_x \in X$

f) El α - corte se define como $A_\alpha = \{x | \mu_A(x) = \alpha, x \in X\}$

Estas operaciones cumplen las propiedades de asociatividad, conmutatividad y las leyes de Morgan, al igual que las mismas operaciones en la teoría de conjuntos clásica. Sin embargo, hay dos operaciones de la teoría clásica de conjuntos que no se cumplen en los conjuntos difusos: $A \cup \bar{A} \neq \Omega y A \cap \bar{A} \neq 0$

2.6.3. Funciones de pertenencia

La función de pertenencia o membresía, están dadas a su esperanza μ_x, es una curva que se define como un punto en el espacio de entrada y le es asignado un valor de membresía entre $[0, 1]$. Existen diferentes tipos de funciones de pertenencia, siendo así las más comunes y como se muestra en la siguiente figura 5

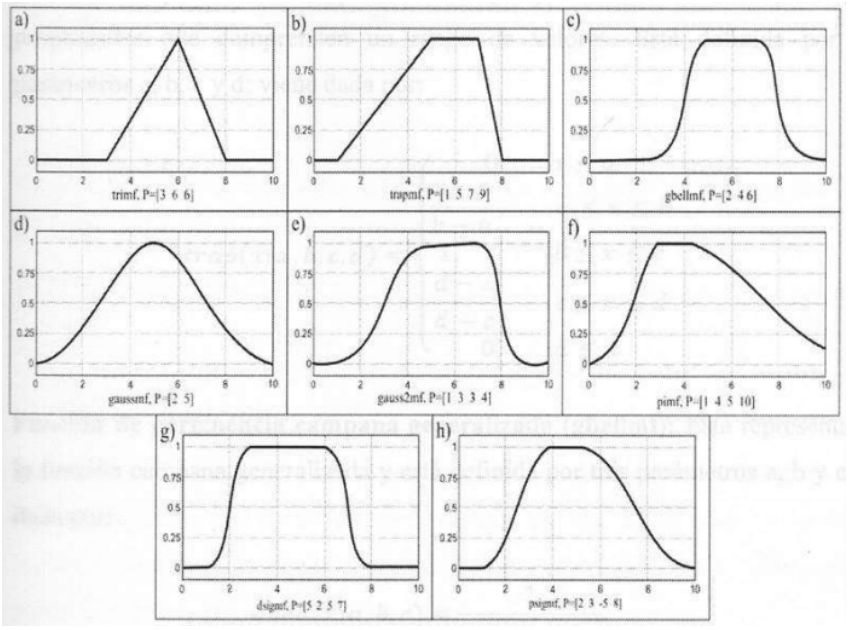

Figura 5: *Curvas de Pertenencia*
(a) Triangular (b) Trapezoidal (c) Campana Generalizada (d) Gaussiana (e) Gaussiana de Dos Caras (f) π (g) Diferencia Sigmoidal (h) Producto Sigmoidal (Pérez, 2014 [4])

a) *Función triangular*: esta función es apropiada para modelar propiedades con un valor de inclusión distinto de cero, para un rango de valores estrecho entorno a un punto b. Definida por tres parámetros: a, b y c. donde a es el límite inferior, b el límite superior y c el valor de la moda.

$$tri(x : a, b, c) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ \dfrac{c-x}{c-b} & b \leq x \leq c \\ 0 & c \leq x \end{cases} \tag{11}$$

b) Función trapezoidal: esta función es recomendada para modelar propiedades que se encuentran dentro de un rango de valores. definida por cuatro parámetros: a, b, cyd, donde a es el límite inferior, d límite superior y los límites de soporte inferior b y superior c

$$trape(x : a, b, c, d) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ 1 & b \leq x \leq c \\ \dfrac{d-x}{d-c} & c \leq x \leq d \\ 0 & d \leq x \end{cases} \tag{12}$$

c) Función de campana generalizada: definida por tres parámetros: a, byc, donde a es el límite inferior, c límite superior, b es el punto de inflexión, si este punto llegase a tener signo negativo la gráfica se invierte.

$$Gbell(x:a,b,c) = \frac{1}{1+|\frac{x-c}{a}|^{2b}} \tag{13}$$

d) Función gaussiana: diferente a la función de densidad normal, esta función de pertenencia tiene un valor máximo de 1, depende de dos parámetros σ y c, donde c es el valor de la media.

$$Gauss(x:\sigma,c) = e^{-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2} \tag{14}$$

e) Función gaussiana de dos caras: es una extensión de la función Gaussiana, más otros dos parámetros donde cada uno modifica un lado de la curva en mención.

f) Función π: tiene forma de campana, adecuada para conjuntos definidos en torno a un valor. y se representa a partir de:

$$\pi(x:c,l) = \begin{cases} S\left(x; c-\lambda, c-\frac{\lambda}{2}, c\right), & x \leq c \\ S\left(x; c, c+\frac{\lambda}{2}, c+\lambda\right), & x \geq c \end{cases} \tag{15}$$

g) Función diferencia sigmoidal: dada por la suma de funciones sigmoidales. depende de los parámetros a y c; y se representa a partir de:

$$Dsig(x:a,c) = \frac{1}{1+e^{-a(x-c)}} \tag{16}$$

h) Función producto sigmoidal: Es la multiplicación de dos funciones sigmoidales, cada una de estas representada por la expresión del literal (g), pero con valores distintos de los parámetros característicos.

2.6.4. Reglas difusas

Segun, Berlanga, 2005 [5] Una de las áreas de aplicación más importantes de la teoría de conjuntos difusos son los Sistemas Basados en Reglas Difusas ($SBRDs$) aplicados con gran éxito a campos tales como el control, el modelado o la clasificación. Tradicionalmente, el diseño de un $SBRDs$ considera como principal objetivo la mejora en el rendimiento o precisión, aunque algunos estudios recientes tienen también en cuenta su interpretabilidad.

Una regla difusa es una expresión lingüística que refleja una causa y un efecto. Su calidad de difusa radica en el hecho de que emplea adjetivos imprecisos y relativos: *Si X pertenece a A, entonces Y pertenece a B*, donde X y Y, son elementos; y, A y B, conjuntos difusos. Un ejemplo de regla difusa es: *Si la velocidad es ALTA, entonces la frenada es FUERTE.*

Para este caso la variable *velocidad* es partícipe en la causa, lo que puede verse como entradas al sistema y pudieran pertenecer a conjuntos difusos como *ALTO*, la variable *frenada*, es una *variable de salida* perteneciente, en cierto grado, al conjunto difuso *FUERTE*.

2.7. Métodos de aprendizaje difuso

a) *Métodos ad hoc*: método de aprendizaje automático de reglas difusas a partir de ejemplos que representan el comportamiento del problema.

b) *Métodos de algoritmos evolutivos*: se aplican principalmente para el aprendizaje o ajuste fuera de línea de reglas, semántica e inferencias. este tipo de algoritmo aprende mediante el procedimiento de simulaciones y evalúa cada una de las opciones de aprendizaje para construir la base de conocimiento del problema.

c) *Método de redes neuronales*: es una fusión entre el método por *algoritmos evolutivos*, donde el aprendizaje se ejecuta mediante simulaciones y evalúa cada una de las opciones de aprendizaje para construir la base de conocimiento, sin embargo, el proceso una vez construida la base de conocimiento entra inicia un proceso de profundización sobre la base construida.

El aprendizaje consiste en cambiar los parámetros que definen la base de conocimiento de forma iterativa hasta que el sistema difuso encuentre el ajuste que mejor relaciona a la información del sistema que se desea controlar. La información consiste en una fila de valores de la forma $x_1, x_2, ..x_n, y^r$, donde las variables x_n son valores de entrada para el sistema difuso y la variable y^r es respuesta esperada del sistema a dicha entrada. Los valores de entrada ingresan al sistema difuso y el sistema responde obteniendo en su salida un valor que puede ser de acuerdo con el problema objeto de estudio un resultado en términos de clasificación o predicción.

2.7.1. Des Difusor

A través de las funciones de pertenencia y de las reglas difusas planteadas, es decir del conjunto de inferencias difusas, el des difusor calcula los valores a predecir, en esta parte se obtiene un puntaje que permite establecer si el conjunto de datos que va a ser objeto de estudio presenta algún posible fraude de consumo de energía.

2.8. Modelos lineales generalizados

Según Rincón(2009) [2] un modelo lineal generalizado se origina cuando se interesa modelar un experimento en el cual la variable respuesta dependiente Y tiene una distribución que pertenece a la familia exponencial, y está asociada a un conjunto de variables explicativas independientes $X_1, ..., X_p$.

$$E(Y) = \mu = g^{-1}(X\beta)$$

Los modelos lineales generalizados están compuestos por tres componentes denominados:

1. **Componente aleatoria** :

   Está representada por un conjunto de variables independientes Y_i $i = 1, 2, ..., n$ cuya distribución para todo i pertenece a la familia exponencial, la función de densidad satisface.

   $$f(y_i; \theta_i; \phi) = exp\left(\frac{1}{a_i(\phi)}[y_i\theta_i - b(\theta_i) + c(y_i; \phi)]\right) \tag{17}$$

   Para algunas funciones $b(\cdot)$ y $c(\cdot)$ conocidas, y además

   a) $E(Y_i) = b'(\theta_i) = \mu_i$

   b) $V(Y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i)$

   c) $(\phi) = \dfrac{\phi}{w_i}$ siendo w_i un conjunto de valores o pesos.

2. **Componente sistemática**

   Está conformada por una matriz de variables independientes $X_1, ..., X_p$ y puede estar asociada a una componente sistemática a un modelo de rango completo o incompleto, para un diseño experimental con variables categóricas o de clasificación.

$$\eta_i = \sum_{i=1}^{p} x_{ij}\beta_j \quad \text{equivalente a } \eta = X\beta$$

3. **Función de Enlace**

   Es una función monótona, derivable que asocia o enlaza las componentes aleatoria y sistemática.

$$g(\mu_i) = \eta_i$$

2.8.1. Modelo logístico

El modelo logístico es un caso en particular del modelo lineal generalizado descrito antes con las siguientes tres componentes:

1. **Componente aleatoria** : Se asume que la variable respuesta U tiene distribución logística con parámetros μ y τ la función de densidad de probabilidad está dada por:

$$f(u;\mu;\tau) = \frac{1}{\tau}\frac{e^{(u-\mu)/\tau}}{\left(1+e^{(u-\mu)/\tau}\right)^2}$$

   Satisface $E(U) = \mu \quad V(U) = \dfrac{\pi^2\tau 2}{3}$.

   Reemplazando $\beta_1 = -\dfrac{\mu}{\tau}$; $\quad \beta_2 = \dfrac{1}{\tau}$, resulta

$$f(u;\beta_1;\beta_2) = \beta_2\frac{e^{\beta_1+\beta_2 x_i}}{\left(1+e^{\beta_1+\beta_2 x_i}\right)^2}$$

2. **Función de enlace** : Para este modelo se utiliza como función de enlace la función logit por lo que está definida por:

$$\eta_i = Logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) \tag{18}$$

3. **Predictor lineal** : Para el caso de una variable explicativa el predictor es $\eta_i = \beta_0 + \beta_1 x$ con la cual el modelo especificado es:

$$Logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i \tag{19}$$

   Que se transforma en el modelo

$$E(Y_i) = \mu_i = m_i\frac{\exp^{(\beta_1+\beta_2 x_i)}}{1+\exp^{(\beta_1+\beta_2 x_i)}} \tag{20}$$

4. **Algoritmo de estimación**: El proceso de estimación de los parámetros de un MLG y su método iterativo de *Newton-Raphson*, para la resolución de una ecuación basado en la aproximación de *Taylor*, para una función $f(x)$ de un punto de partida x_i,

$$f(x) = f(x_{i+1}) + (x - x_i)f'(x_i) = 0 \tag{21}$$

El proceso se repite como:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{22}$$

hasta alcanzar un valor suficientemente preciso. Según (Tjalling, 1995 [11] El método 21, y su extensión a la solución de sistemas de ecuaciones no lineales. La estimación de los coeficientes $\hat{\beta}_i$ del vector de parámetros del modelo, permite estimar en función $X_1, ..., X_p$, la probabilidad de que el suceso ocurra en función de π_i o el valor $E(Y_i)$.

2.9. Criterio de información Akaike

El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos, es una compensación entre la bondad de ajuste del modelo y la complejidad del modelo. Se basa en la entropía de información: se ofrece una estimación relativa de la información perdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos.

El AIC no proporciona una prueba de hipótesis, es decir no existe *hiptesisnula* que debe ser *rechazada* o *norechazada*, es decir, no puede decir nada acerca de la calidad del modelo en un sentido absoluto.

2.10. Selección de variables

En muchas situaciones se dispone de un conjunto de posibles variables regresoras, sin embargo, una pregunta que puede surgir es *¿todas las variables deben entrar en el modelo?*, en caso negativo, existen métodos para seleccionar las variables dentro de un modelo de regresión.

Algunos de estos métodos son:

a) *Eliminación progresiva* (*backward stepwise*): parte del modelo de regresión con todas las variables regresoras y en cada etapa se elimina la variable menos influyente según el contraste individual de la T o de la F hasta una cierta regla de parada. El procedimiento de eliminación progresiva tiene los inconvenientes de necesitar una alta capacidad de cálculo computacional, si se cuenta con un conjunto de datos grande y una cantidad considerable de variables, adicional puede llevar a problemas de multicolinealidad si las variables están relacionadas. Tiene la ventaja de no eliminar variables significativas.

b) *Introducción progresiva* ($fordward\ stepwise$): Este algoritmo funciona de forma inversa que el anterior, parte del modelo sin ninguna variable regresora y en cada etapa se introduce la más significativa. El procedimiento de introducción progresiva tiene la ventaja respecto al anterior de necesitar menos capacidad de cálculo computacional, pero presenta dos graves inconvenientes, el primero, que pueden aparecer errores de especificación porque las variables introducidas permanecen en el modelo, aunque el algoritmo en pasos sucesivos introduzca nuevas variables que aportan la información de las primeras. Este algoritmo también falla si el contraste conjunto es significativo pero los individuales no lo son, ya que no introduce variables regresoras.

c) *Regresión paso a paso* (*stepwise regression*): Este método es una combinación de los procedimientos anteriores, comienza como el de introducción progresiva, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer. Termina el algoritmo cuando ninguna variable entra o sale del modelo. El algoritmo paso a paso tiene las ventajas del algoritmo de introducción progresiva, pero lo mejora al no mantener fijas en el modelo las variables que ya entraron en una etapa, evitando de esta forma problemas de multicolinealidad. En la práctica, es un algoritmo bastante utilizado que proporciona resultados razonables cuando se tiene un número grande de variables regresoras.

3. Metodología

Este proyecto es de carácter comparativo, con el cual se busca comparar la eficiencia, precisión, confiabilidad y poder predictivo que tiene un modelo de *lógica difusa* frente a un modelo lineal generalizado *logit*. Para este trabajo de investigación y aplicación se busca a través de técnicas de minería de datos transformar y analizar una base de datos de consumo de energía eléctrica de una ciudad de Colombia con el fin de poder identificar fraude por consumo de electricidad.

Siguiendo la metodología $CRISP-DM$, planteada en la sección 2 del presente documento tenemos que:

a) *Entendimiento del negocio*: una empresa de energía eléctrica de Colombia, tiene un problema asociado al fraude que se presenta en la prestación del servicio para una ciudad, este comportamiento ha sido detectado gracias a la implementación de modelos probabilísticos que generan un score el cual permite clasificar los registros presentes en la base de datos, como se mencionó en la sección 1, este problema genera pérdidas millonarias para el sector, por esta razón la implementación de metodologías que permitan tomar medidas tempranas frente a este fenómeno son de gran interés para la empresa.

b) *Compresión de los datos*: El conjunto de datos a trabajar contiene la información de consumo de energía eléctrica de diferentes sectores económicos y estratos de una ciudad de Colombia con un total de 13025 registros y 14 variables, con una tasa de fraude legitima del 10.71 % por consumo de energía de diferentes sectores y estratos. El conjunto de datos contiene variables de entrada categóricas y numéricas, además, debido a problemas de confidencialidad por convenios empresariales no se proporciona la información de ciudad, empresa prestadora del servicio, identificación de los clientes y direcciones.

   El conjunto de datos está compuesto de las siguientes variables:

Variable	Tipo	Descripción
Registro reconstruido	nominal	Consumo de energía promediado (si o no)
Suministro en comisión de servicio	nominal	Servicio especial de energía (si o no)
Estado del medidor	nominal	Disponible, instalado, retirado, dado de baja
Energía activa	continua	Energía de entrada
Estrato	nominal	Estrato socioeconómico
Energía activa constante verificada	continua	Energía activa constante de entrada
Carga total cliente	continua	Carga de energía del cliente
Tipo de energía	nominal	Monofasica, bifasica o trifasica
Subsidio	nominal	Energía subsidiada (si o no)
Sector	nominal	Sector económico
Modelo del medidor	nominal	Referencia del modelo del contador
Voltaje	continua	Voltaje de entrada
Tipo de red	nominal	Aérea o subterranea
Fraude	nominal	Presencia de fraude (si o no)

Tabla 1: *Descripción y estructura*

La tasa de fraude legitimo corresponde al 10.71 % del total de registros, este comportamiento se puede observar en la tabla 2.

Fuente	Tasa (%)
No Fraude	89.29
Fraude	10.71

Tabla 2: *Tasa fraude*

La figura 6, muestra el comportamiento de fraudes legítimos por sectores económicos y tipo de red, los sectores con mayor presencia de fraudes son los sectores 4 y 81, así mismo, la red donde más se presenta fraude es la aérea, este comportamiento se debe a que las redes aéreas son de más fácil acceso y por otro lado también se pueden ver afectadas por cambios de clima.

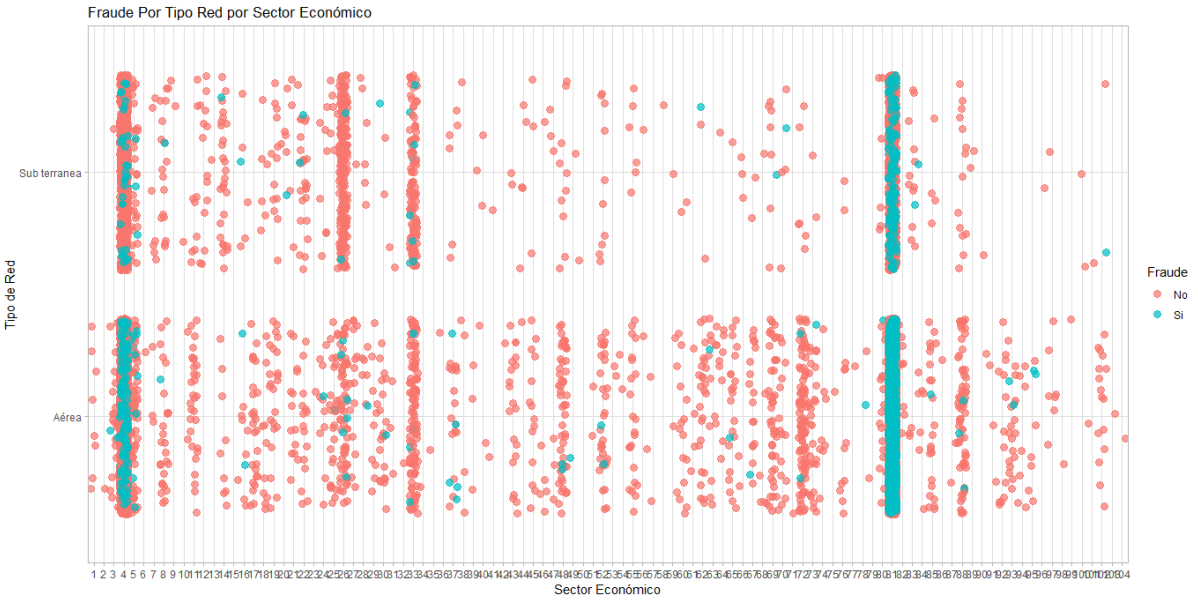

Figura 6: *Fraude por tipo de red y sector*

En la siguiente tabla 3, se observa la frecuencia de fraude por sector y tipo de red:

Código Sector	Tipo de Red	Frecuencia
81	Aerea	937
4	Aerea	172
81	Subterranea	162
4	Subterranea	41
5	Aerea	10
33	Subterranea	8

Tabla 3: Fraude por tipo de red y sector económico

En la figura 7, se observa cual es comportamiento de los fraudes legítimos por estrato y sector económico

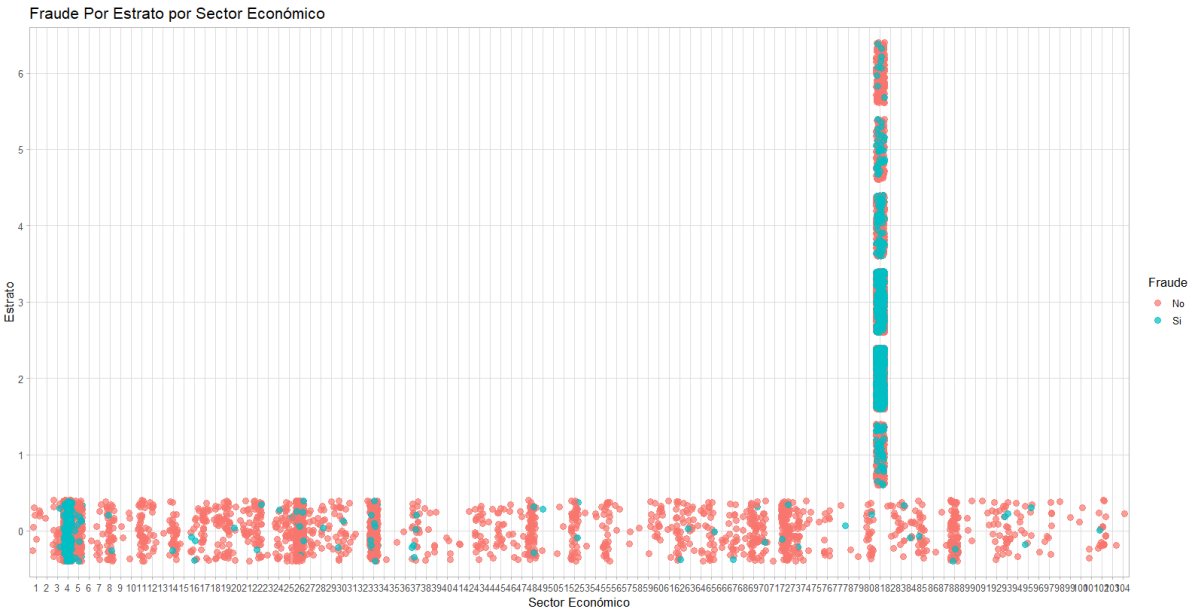

Figura 7: *Fraude por estrato y sector*

Se observa en la tabla 4 donde hay mayor presencia de fraude está en el sector económico 81 en los estratos 2 y 3:

Sector	Estrato	Frecuencia
81	Estrato 2	664
81	Estrato 3	268
4	Estrato 0	213
81	Estrato 4	75
81	Estrato 1	52
81	Estrato 5	31
5	Estrato 0	13
33	Estrato 0	13
81	Estrato 6	9
26	Estrato 0	8

Tabla 4: *Fraude por estrato y sector económico*

En la figura 8, se observa cual es comportamiento de los fraudes legítimos por estrato y subsidio, como se puede observar en la tabla 5 que los estratos con mayor tasa de fraudes legítimos son el estrato 0 no subsidiado con los estratos 2 y 3 subsidiados.

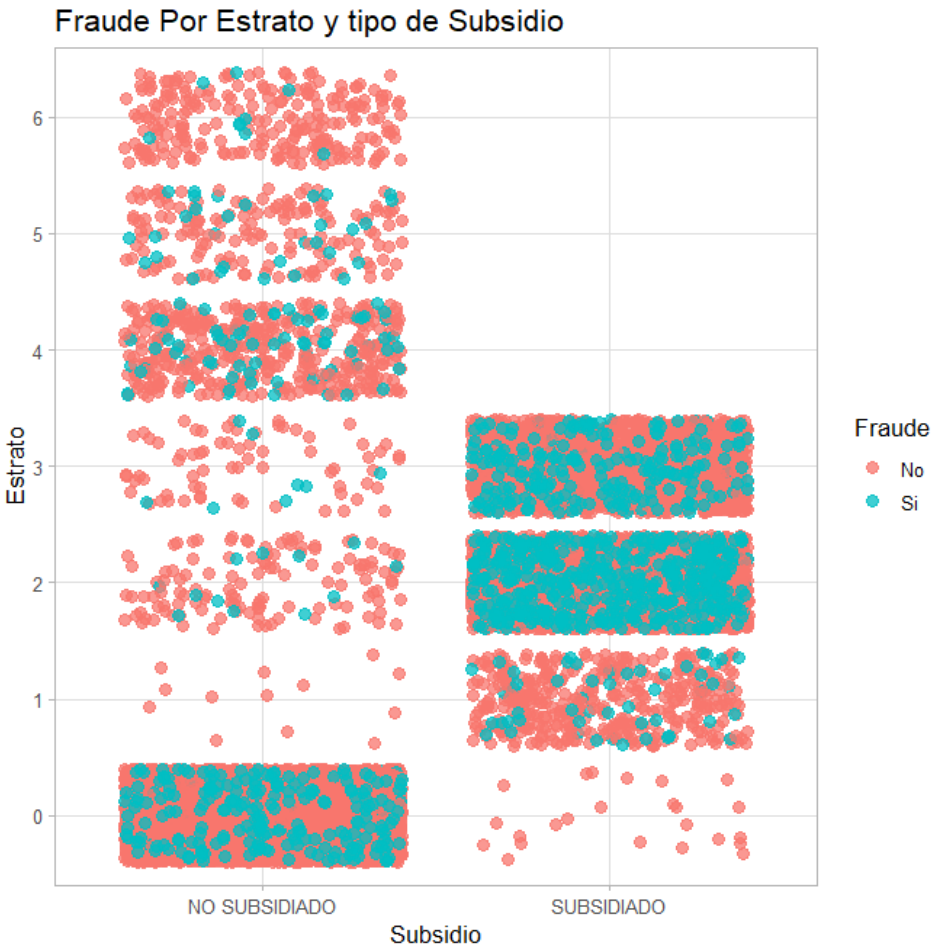

Figura 8: *Fraude por estrato y subsidio*

Estrato	Subsidio	Frecuencia
2	Subsidiado	652
0	No Subsidiado	296
3	Subsidiado	260
4	No Subsidiado	75
1	Subsidiado	52
5	No Subsidiado	31
2	No Subsidiado	12
6	No Subsidiado	9
3	No Subsidiado	8

Tabla 5: *Fraude por estrato y subsidio*

Una vez realizada y analizada la etapa de *comprensión de los datos*, se toma la decisión de construir tres modelos, en la siguiente tabla 6 se presentan los modelos con su respectiva tasa de fraude:

Modelo	Total registros	Tasa de fraude
Modelo global	13.025	10.71 %
Modelo sector 81	6.861	16.02 %
Modelo otros sectores	6.164	4.80 %

Tabla 6: Descripción fraude por modelo

c) *Preparación de los datos*: de acuerdo al tipo de modelos que se quieren construir y que están referenciados en los objetivos del presente trabajo, los datos originales deben ser transformados con el fin de garantizar el buen desempeño de los modelos; teniendo en cuenta que para la construcción del modelo de *lógica difusa* es necesario contar con variables de tipo numéricas, las variables categóricas son transformadas a variables de tipo continuas haciendo uso del *análisis de componentes principales categóricos* y el algoritmo de *HOMALS*, con estas transformaciones se garantiza que las variables del conjunto de datos fueron transformadas de una manera óptima sin alto riesgo de pérdida de información.

4. Resultados

4.1. Modelos

De acuerdo a los objetivos planteados, el proceso de desarrollo de los modelos inicia con la construcción de un modelo *logit*, el cual va a ser comparado con los diferentes modelos de lógica difusa para clasificación de variables binarias, para este proceso la base de datos original se divide en dos partes, siguiendo los criterios dispuestos por la teoría de la *minería de datos*, en el fin de tener una base de entrenamiento (70 %) y una de prueba (30 %), este proceso permite hacer una validación cruzada de los resultados de los modelos.

En la siguiente tabla 7, se observan los resultados de la partición de acuerdo con la tasa de fraude presente en cada base.

Fuente	No Fraude	Fraude
Entrenamiento	89.26 %	10.74 %
Prueba	89.35 %	10.65 %

Tabla 7: *Tasa de fraude*

4.2. Modelos logit

4.2.1. Modelo logit global

Se procede a realizar una simulación, con el fin de optimizar el modelo *logit* con el criterio *Akaike*, se hace uso del algoritmo de eliminación progresiva. Los resultados de la simulación como se muestran en la (gráfica 9) genera 4 modelos y por selección tomamos el de menor valor *Akaike*, con esto se garantiza que el modelo seleccionado es el de mejor calidad y de ajuste de bondad a la variable respuesta.

En la siguiente tabla 8 se observan los valores *Akaike* del modelo inicial y del modelo final.

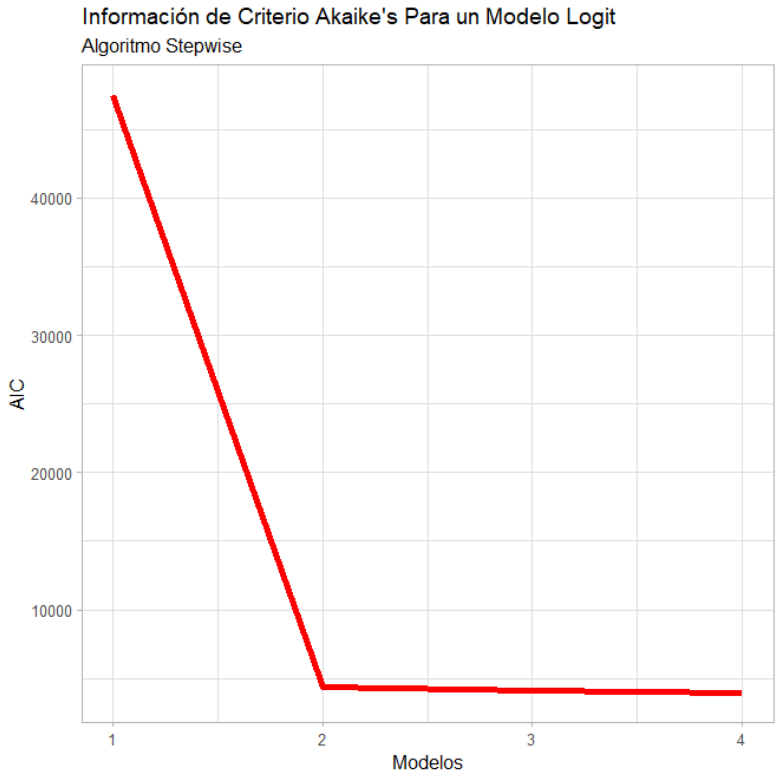

Figura 9: *Simulación del criterio Akaike global*

Fuente	Inicial	Final
AIC	47485	3934
No. Variables	13	8

Tabla 8: *Aic modelo logit global*

Los resultados de las variables del modelo final y sus coeficientes se muestran en la siguiente tabla 9.

Fuente	$\hat{\beta}$	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.733	360.85	-0.05	0.963
Registro reconstruido 2	2.626	0.096	27.5	<2e-16 ***
Suministro en comisión de servicio 2	1.385	0.1813	7.64	2.24e-14 ***
Estado del medidor 2	1.881	459.6055	0.00	0.9967
Estado del medidor 3	12.148	360.86	0.03	0.9731
Estado del medidor 4	14.1759	360.86	0.04	0.9687
Energía activa	-0.0001	0.0000	-2.75	0.00589 **
Estrato	0.1410	0.0343	4.11	3.92e-05 ***
Tipo de energía 2	0.2964	0.7204	0.41	0.6808
Tipo de energía 3	-0.1090	0.724	-0.15	0.8804
Subsidio 2	0.2017	0.1173	1.72	0.08564 .
Tipo de red 2	0.3163	0.1203	2.63	0.00856 **

Tabla 9: *Coeficientes modelo logit final global*

El modelo final es seleccionado para ejecutar el proceso de aprendizaje y posterior prueba, las métricas de evaluación del conjunto de entrenamiento se muestran en la tabla 10.

Fuente	Accuracy	Curva Roc
Entrenamiento	82.05 %	82.75 %

Tabla 10: *Métricas de clasificación modelo logit*

En la tabla 11, se observan las métricas de evaluación del conjunto de prueba.

Fuente	Accuracy	Curva Roc
Prueba	81.32 %	81.71 %

Tabla 11: *Métricas de clasificación modelo logit*

De acuerdo con los resultados, se puede decir que el modelo en base de entrenamiento tiene un buen proceso de aprendizaje, puesto que al aplicarlo en la base de prueba se obtienen unas buenas métricas de clasificación, en la figura 10, se observa el gráfico de curva *Roc* asociado a los resultados del modelo en base de prueba.

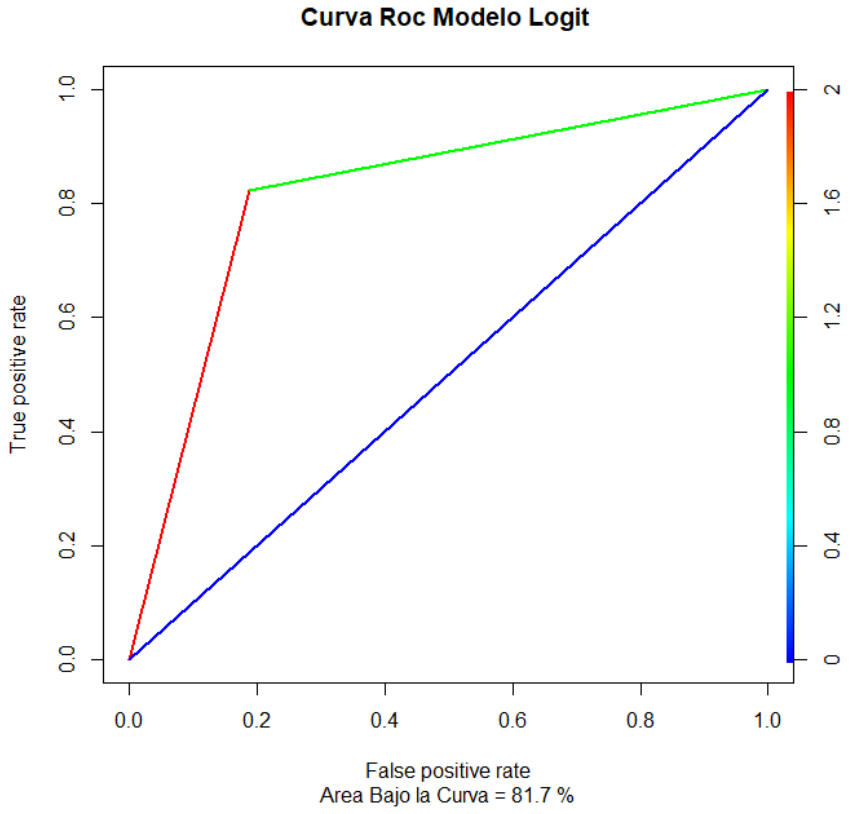

Figura 10: *Curva Roc modelo logit global*

4.2.2. Modelo logit sector económico 81

Bajo el mismo esquema y de acuerdo con los hallazgos encontrados durante el análisis descriptivo de los datos originales, se construye un modelo únicamente para el *sector económico 81*, en la siguiente tabla 12, se observa la tasa de fraude para el sector y se discrimina por la base de *entrenamiento* y *prueba*, estas últimas son calculadas siguiendo las mismas reglas del modelo global.

Fuente	No Fraude	Fraude
Entrenamiento	83.93 %	16.07 %
Prueba	84.11 %	15.89 %

Tabla 12: *Tasa fraude por sector económico 81*

A partir de la base de entrenamiento se busca encontrar el mejor modelo posible aplicando el algoritmo de *eliminación progresiva*, este modelo va a ser seleccionado de acuerdo con el criterio AIC, donde el mejor modelo es el que tenga menor criterio, en la siguiente tabla 13 se muestran los resultados.

Fuente	Inicial	Final
Akaike	4275	2794
No. Variables	12	5

Tabla 13: *Aic modelo logit sector económico 81*

El modelo final tiene las siguientes métricas de clasificación en el entrenamiento 14.

Fuente	Accuracy	Curva Roc
Entrenamiento	81.80 %	80.36 %

Tabla 14: *Métricas de clasificación modelo logit 81 entrenamiento*

Según la tabla 15, las métricas de clasificación para el modelo del sector económico 81, para la base de entrenamiento y pruebas son más homogéneas entre sí, comparado con las del modelo global.

Fuente	Accuracy	Curva Roc
Prueba	83.19 %	82.19 %

Tabla 15: *Métricas de clasificación modelo logit sector 81*

A continuación, la figura 11, muestra el área bajo la curva *Roc* del modelo para el sector 81.

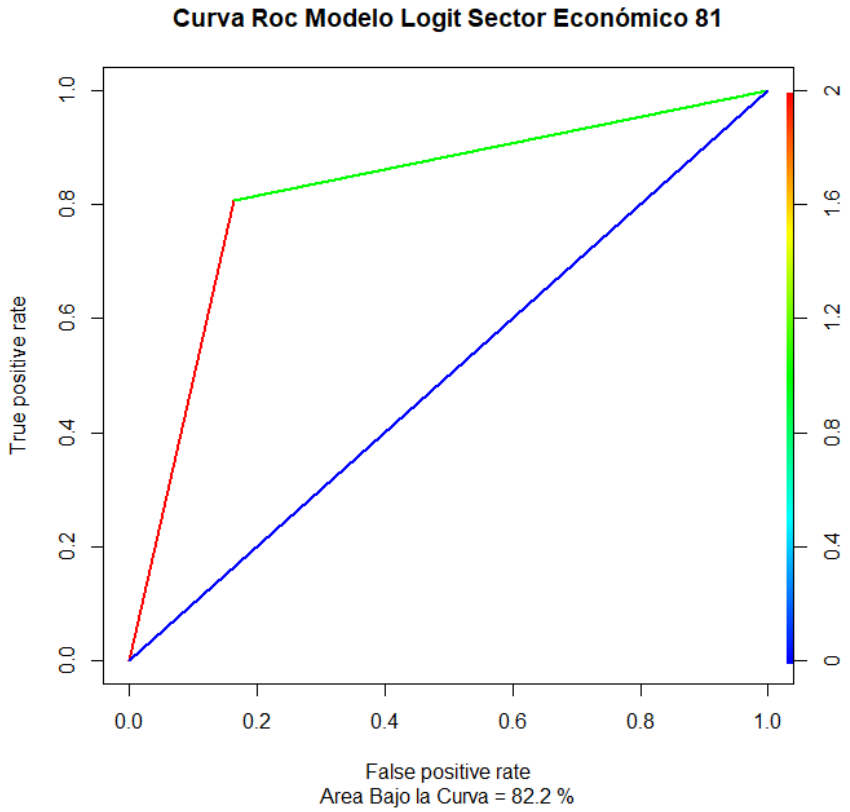

Figura 11: *Curva Roc modelo logit sector 81*

4.2.3. Modelo logit otros sectores

La construcción del modelo sin el sector económico 81, se ejecuta a partir de la base de entrenamiento, en la siguiente tabla 16, se observa el comportamiento de fraude presente en la base de entrenamiento y prueba.

Fuente	No Fraude	Fraude
Entrenamiento	95.48 %	4.52 %
Prueba	94.54 %	5.46 %

Tabla 16: *Tasa fraude sin sector económico 81*

El modelo de esta sección se evalúa a partir del criterio *Aic*, aplicando el algoritmo de *eliminación progresiva* para la selección de variables, en la tabla 17, se observan los resultados del criterio *Aic* para el modelo inicial y final.

Fuente	Inicial	Final
AIC	4754	1117
No. Variables	13	5

Tabla 17: *Aic modelo logit sin el sector económico 81*

De acuerdo con los resultados del modelo final, en la tabla 18, se observan las métricas derivadas de la evaluación del modelo, ejecutadas con la base de entrenamiento.

Fuente	Accuracy	Curva Roc
Entrenamiento	100 %	78.35 %

Tabla 18: *Métricas de clasificación de entrenamiento sin el sector 81*

Por otro lado, en la tabla 19, se visualizan las métricas del modelo final sin sector económico 81, ejecutadas con la base de prueba.

Fuente	Accuracy	Curva Roc
Prueba	68.25 %	76.68 %

Tabla 19: *Métricas de clasificación sin el sector 81*

El area bajo la curva roc, ejecutada con la base de prueba del modelo sin sector económico 81, se muestra en la siguiente figura 12.

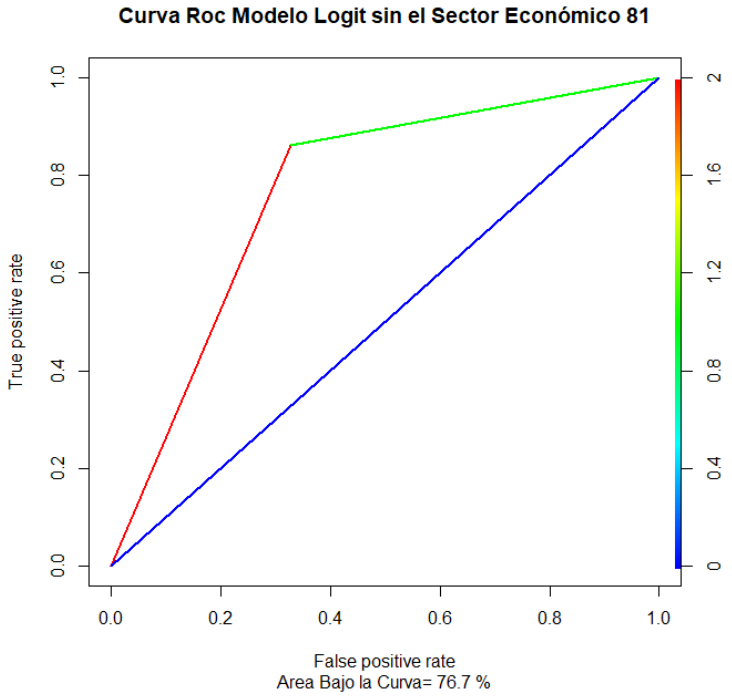

Figura 12: *Curva Roc modelo logit sin sector 81*

En términos generales los tres modelos propuestos,*modelo logit general, modelo logit sector económico 81 y modelo logit otros sectores*, tienen un buen desempeño de acuerdo a las métricas observadas, tanto en su fase de aprendizaje como de prueba, sin embargo, el más homogéneo entre estos es el *modelo logit sector económico 81.*

4.3. Lógica difusa

De acuerdo a la teoría consultada, los modelos de lógica difusa para clasificación de variables, requieren una transformación de datos, el cual consiste en trabajar con variables continuas; las variables categóricas usadas para la construcción del modelo son transformadas a partir de un análisis de componentes principales para datos categóricos y el algoritmo de *HOMALS*, el cual genera una puntuación para cada uno de los registros de acuerdo a cada una de las variables, con esto es posible calcular cada uno de los rangos asociados a las matrices dadas por las funciones de pertenencia.

En la siguiente tabla se puede observar los resultados de la transformación por componentes principales categóricos aplicada aplicada a los datos originales.

Registro	R.reconstruido	S.C.servicio	E.medidor	T.energia	Subsidio	Sector	M.medidor	T.red
1	1	1	4	3	2	81	40	1
2	1	1	4	3	1	46	311	1
3	2	1	4	2	2	81	91	1
4	1	1	3	2	2	81	40	1
5	1	1	4	3	1	81	112	1
13020	2	1	4	2	1	4	255	1
13021	2	1	4	2	2	81	40	1
13022	1	1	4	3	1	4	255	1
13023	2	1	4	2	2	81	255	1
13024	2	1	4	2	1	4	36	1
13025	1	1	4	2	2	81	255	1

Tabla 20: Tabla variables sin transformación

Registro	D1	D2	D3	D4	D5	D6	D7	D8
1	-0.61	-0.38	0.69	1.49	-0.96	0.95	0.04	-1.04
2	0.37	0.96	1.56	0.87	0.78	1.51	1.21	1.69
3	-1.67	0.18	1.37	-1.30	-1.57	-1.42	0.15	-0.09
4	-0.91	0.04	-1.25	0.14	-0.18	-0.04	-0.16	-0.29
5	0.25	-0.23	0.98	1.27	-0.91	0.67	-0.68	1.58
13020	-0.11	1.45	2.05	-2.03	-0.37	-1.90	0.03	0.66
13021	-1.67	0.18	1.37	-1.30	-1.57	-1.42	0.15	-0.09
13022	0.95	0.89	1.37	0.77	0.24	0.47	-0.08	-0.29
13023	-1.60	0.78	1.66	-1.51	-0.72	-0.99	1.09	-0.04
13024	-0.17	0.85	1.76	-1.82	-1.22	-2.32	-0.91	0.61
13025	-0.93	0.75	0.52	0.09	0.46	1.61	-0.04	-0.57

Tabla 21: Tabla variables transformadas ACPcat

Siguiendo la metodología planteada para la construcción de los modelos *logit* y de acuerdo con los hallazgos encontrados en la fase de comprensión de los datos, para la siguiente sección se plantea la elaboración los siguientes modelos:

a) *Modelo difuso global*
b) *Modelo difuso sector económico* 81

4.3.1. Modelo difuso global

En la siguiente tabla 22, se pueden observar las métricas de evaluación del modelo difuso global, de acuerdo con esto, podemos asegurar que el modelo, aunque cuenta con *Accuracy* relativamente alto, en

los diferentes modelos que se construyen la métrica de área bajo la curva *Roc*, indica que son modelos aleatorios, es decir no tienen buena capacidad de clasificación.

Fuente	Accuracy	Curva Roc	Método	$f(x)$ Pertenencia
Modelo 1	89.04 %	50.00 %	FRBCS.W	GAUSSIAN
Modelo 2	76.85 %	52.94 %	FRBCS.W	TRAPEZOID
Modelo 3	89.04 %	50.00 %	FRBCS.W	SIGMOID
Modelo 4	89.04 %	50.00 %	FRBCS.W	BELL

Tabla 22: *Métricas de validación*

A continuación, en la figura 13 se puede observar que los modelos globales basados en lógica difusa carecen de poder predictivo de acuerdo con la medida de área bajo la curva *Roc*.

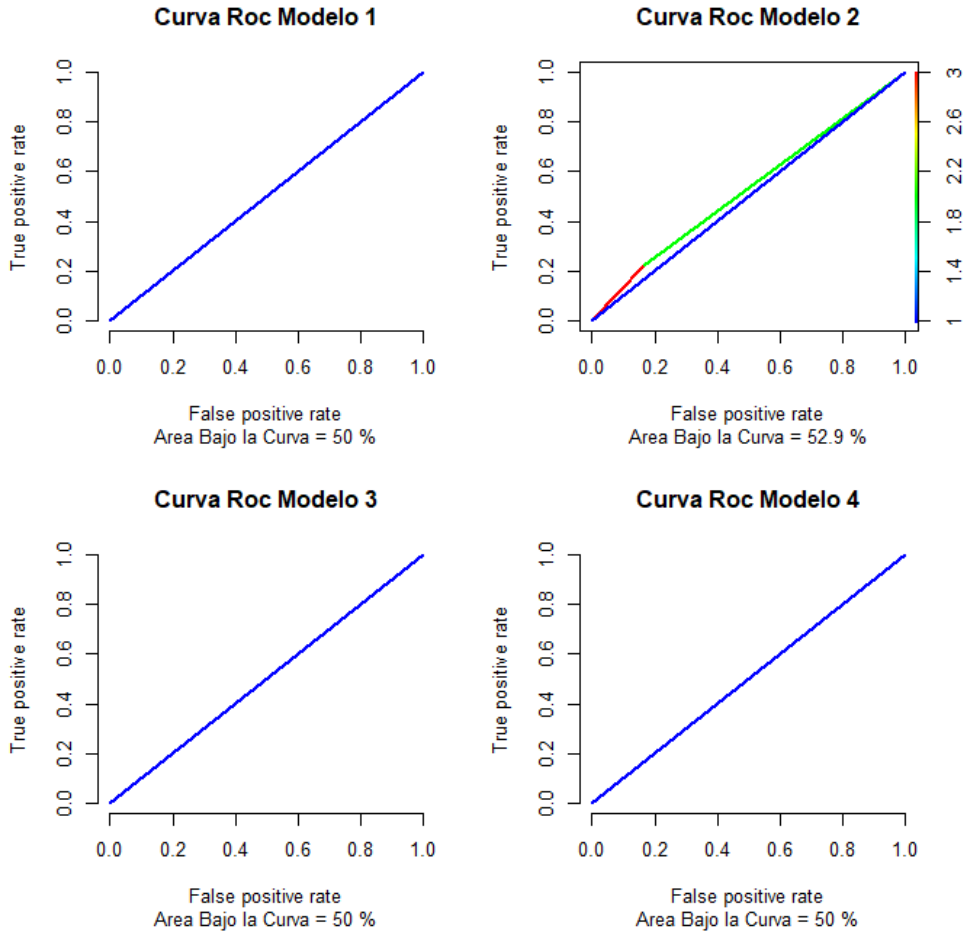

Figura 13: *Curva Roc modelo lógica difusa global*

En las siguientes figuras 14 y 15, se observan las reglas generadas para la construcción del modelo, para este modelo se generaron 290 reglas condicionales.

```
> data.frame(Fuzzy.1$rule[1:5,])
  X1 X2 X3      X4  X5 X6 X7      X8  X9 X10 X11     X12 X13 X14 X15     X16 X17 X18 X19     X20 X21 X22 X23     X24 X25 X26 X27     X28 X29 X30 X31     X32
1 IF D1 is v.1_a.1 and D2 is v.2_a.2 and  D3  is v.3_a.1 and  D4  is v.4_a.1 and  D5  is v.5_a.2 and  D6  is v.6_a.2 and  D7  is v.7_a.1 and  D8  is v.8_a.2
2 IF D1 is v.1_a.1 and D2 is v.2_a.1 and  D3  is v.3_a.1 and  D4  is v.4_a.2 and  D5  is v.5_a.1 and  D6  is v.6_a.2 and  D7  is v.7_a.2 and  D8  is v.8_a.2
3 IF D1 is v.1_a.1 and D2 is v.2_a.2 and  D3  is v.3_a.1 and  D4  is v.4_a.1 and  D5  is v.5_a.2 and  D6  is v.6_a.2 and  D7  is v.7_a.1 and  D8  is v.8_a.2
4 IF D1 is v.1_a.1 and D2 is v.2_a.2 and  D3  is v.3_a.1 and  D4  is v.4_a.1 and  D5  is v.5_a.2 and  D6  is v.6_a.2 and  D7  is v.7_a.1 and  D8  is v.8_a.2
5 IF D1 is v.1_a.2 and D2 is v.2_a.2 and  D3  is v.3_a.1 and  D4  is v.4_a.2 and  D5  is v.5_a.1 and  D6  is v.6_a.1 and  D7  is v.7_a.2 and  D8  is v.8_a.1
  X33            X34 X35     X36 X37     X38 X39      X40 X41                                    X42 X43      X44 X45                X46 X47      X48 X49
1 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.2 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.2 and
2 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.1 and
3 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.1 and
4 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.2 and
5 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.2 and
      X50 X51      X52  X53    X54 X55 X56
1 Voltaje  is v.13_a.1 THEN Fraude  is   2
2 Voltaje  is v.13_a.1 THEN Fraude  is   1
3 Voltaje  is v.13_a.1 THEN Fraude  is   1
4 Voltaje  is v.13_a.1 THEN Fraude  is   1
5 Voltaje  is v.13_a.1 THEN Fraude  is   1
```

Figura 14: *Reglas condicionales*

```
> data.frame(Fuzzy.1$rule[286:290,])
  X1 X2 X3      X4  X5 X6 X7      X8  X9 X10 X11     X12 X13 X14 X15     X16 X17 X18 X19     X20 X21 X22 X23     X24 X25 X26 X27     X28 X29 X30 X31     X32
1 IF D1 is v.1_a.1 and D2 is v.2_a.1 and  D3  is v.3_a.1 and  D4  is v.4_a.2 and  D5  is v.5_a.2 and  D6  is v.6_a.1 and  D7  is v.7_a.1 and  D8  is v.8_a.1
2 IF D1 is v.1_a.2 and D2 is v.2_a.2 and  D3  is v.3_a.2 and  D4  is v.4_a.2 and  D5  is v.5_a.1 and  D6  is v.6_a.2 and  D7  is v.7_a.2 and  D8  is v.8_a.2
3 IF D1 is v.1_a.2 and D2 is v.2_a.1 and  D3  is v.3_a.2 and  D4  is v.4_a.2 and  D5  is v.5_a.2 and  D6  is v.6_a.2 and  D7  is v.7_a.2 and  D8  is v.8_a.1
4 IF D1 is v.1_a.2 and D2 is v.2_a.2 and  D3  is v.3_a.2 and  D4  is v.4_a.2 and  D5  is v.5_a.2 and  D6  is v.6_a.1 and  D7  is v.7_a.1 and  D8  is v.8_a.1
5 IF D1 is v.1_a.1 and D2 is v.2_a.1 and  D3  is v.3_a.2 and  D4  is v.4_a.2 and  D5  is v.5_a.2 and  D6  is v.6_a.1 and  D7  is v.7_a.1 and  D8  is v.8_a.1
  X33            X34 X35     X36 X37     X38 X39      X40 X41                                    X42 X43      X44 X45                X46 X47      X48 X49
1 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.2 and
2 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.2 and
3 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.2 and
4 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.2 and
5 and Energia.activa  is v.9_a.1 and Estrato  is v.10_a.1 and Energia.activa.constante.verificada  is v.11_a.1 and Carga.total.cliente  is v.12_a.1 and
      X50 X51      X52  X53    X54 X55 X56
1 Voltaje  is v.13_a.1 THEN Fraude  is   2
2 Voltaje  is v.13_a.2 THEN Fraude  is   2
3 Voltaje  is v.13_a.1 THEN Fraude  is   1
4 Voltaje  is v.13_a.1 THEN Fraude  is   1
5 Voltaje  is v.13_a.2 THEN Fraude  is   1
```

Figura 15: *Reglas condicionales*

De acuerdo con los modelos generados según tabla 22, a continuación, se presentan las figuras correspondientes a las funciones de pertenencia: Gauss

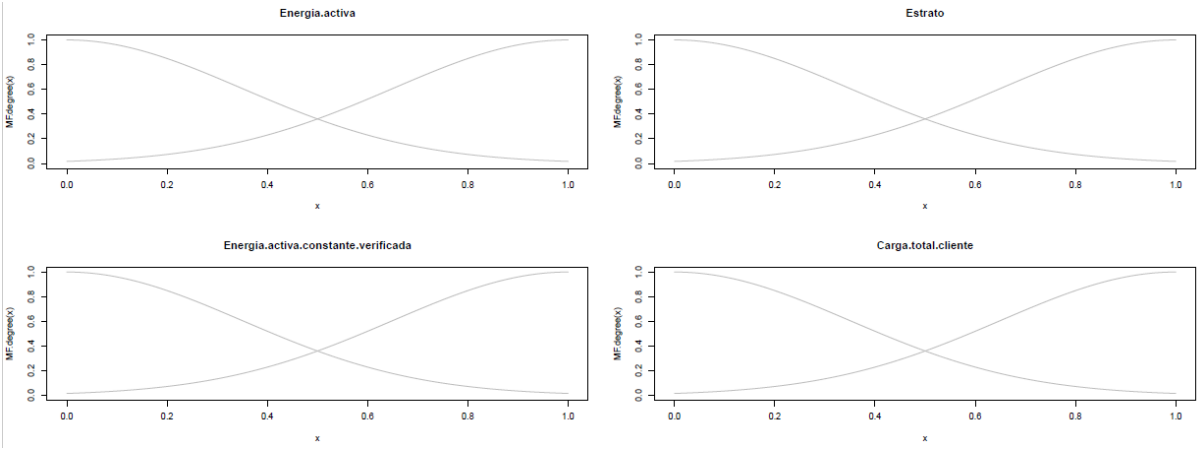

Figura 16: *Función gaussiana*

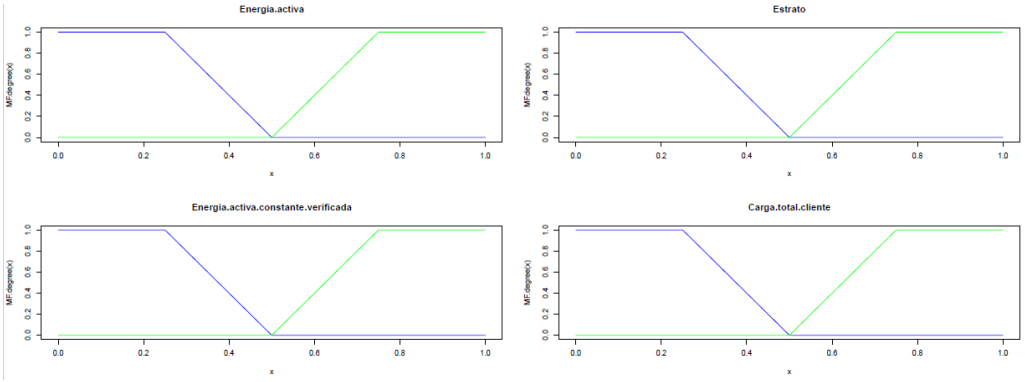

Figura 17: *Función trapezoidal*

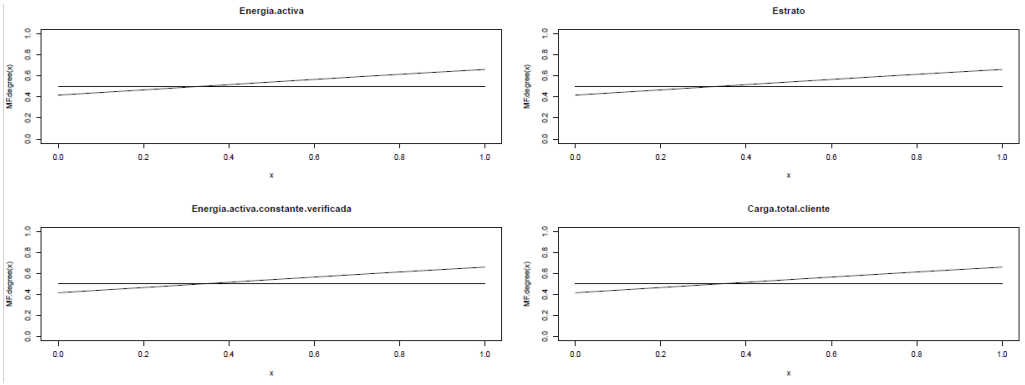

Figura 18: *Función sigmoidal*

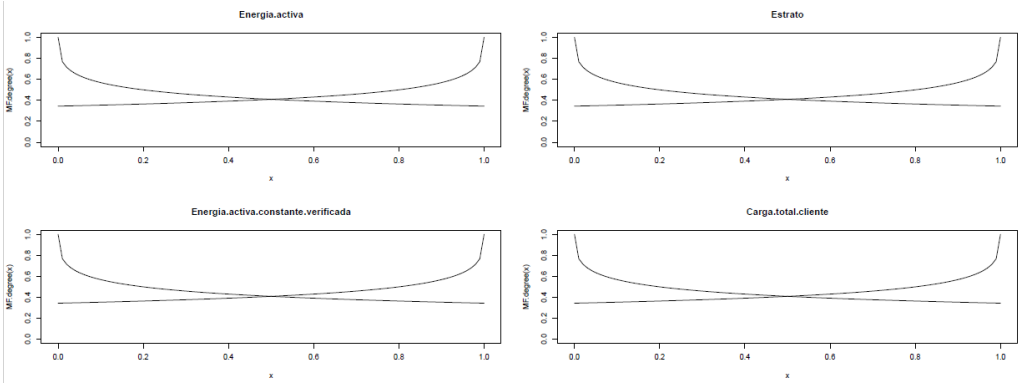

Figura 19: *Función campana*

4.3.2. Modelo difuso sector económico 81

El comportamiento del modelo por sector económico es aleatorio, es decir que, pese a que el Accuracy en todos los métodos evaluados está por encima del 80 %, sin embargo, el área bajo la curva *Roc*, indica que los modelos evaluados son aleatorios, es decir no están en la capacidad de predecir un resultado de manera óptima.

Fuente	Accuracy	Curva Roc	Método	$f(x)$ Pertenencia
Modelo 1	83.43 %	50.00 %	FRBCS.W	GAUSSIAN
Modelo 2	83.43 %	50.11 %	FRBCS.W	TRAPEZOID
Modelo 3	83.43 %	50.00 %	FRBCS.W	SIGMOID
Modelo 4	83.43 %	50.00 %	FRBCS.W	BELL

Tabla 23: *Métricas de validación sector económico 81*

4.3.3. Modelo difuso datos balanceados

Una vez evaluadas las métricas de los modelos difusos globales y por sectores, se procede a balancear la base datos de acuerdo a la tasa de fraude presente, además, contenga una tasa de fraude equilibrada; este proceso se ejecuta con el fin de crear un modelo base de entrenamiento equilibrada y realizar su clasificación con la base de prueba propuesta sin balancear; en la siguiente tabla 24, se observan los diferentes balanceos:

Tasa %	No Fraude	Fraude
50 50	50.38	49.62
60 40	39.10	60.90
70 30	29.66	70.34

Tabla 24: *Muestras balanceadas*

En la tabla 25, se observan las métricas de valoración de los modelos con sus funciones de pertenencia y las bases equilibradas obteniendo así resultados similares a las bases des balanceadas.

Tasa (%)	Accuracy	Curva Roc	Método	$f(x)$ Pertenencia
50-50	89.55	50.00	FRBCS.W	GAUSSIAN
60-40	90.32	50.01	FRBCS.W	GAUSSIAN
70-30	88.59	49.18	FRBCS.W	GAUSSIAN
50-50	88.13	50.01	FRBCS.W	TRAPEZOID
60-40	90.63	50.71	FRBCS.W	TRAPEZOID
70-30	89.16	49.98	FRBCS.W	TRAPEZOID
50-50	87.09	51.18	FRBCS.W	SIGMOID
60-40	82.19	50.11	FRBCS.W	SIGMOID
70-30	86.24	48.35	FRBCS.W	SIGMOID
50-50	91.04	51.16	FRBCS.W	BELL
60-40	90.17	50.20	FRBCS.W	BELL
70-30	89.71	50.43	FRBCS.W	BELL

Tabla 25: *Métricas de validación de muestras balanceadas*

Como se puede observar, el proceso de balanceo no mejora los resultados de clasificación dentro de cada uno de los modelos en fase de entrenamiento y prueba .

5. Conclusiones

Teniendo en cuenta el desempeño de los modelos logit y difusos, evaluados con las métricas planteadas se puede concluir:

1. La identificación de fraude de energía, a partir de un modelo de lógica difusa no tiene una clasificación adecuada, además, la comparación de eficiencia del modelo lineal generalizado logit y los de lógica difusa reportan diferencias considerables en sus resultados, esto con base en las métricas de accuracy y área bajo la curva roc, obteniendo así un mejor desempeño los modelos logit. Ver tabla 10

2. Teniendo en cuenta la métrica de área bajo la curva roc, se puede establecer que, para el caso de estudio, los modelos de lógica difusa no cuentan con valor predictivo. Ver tabla 22 y figura **??**.

3. Las variables que tienen una mayor contribución en función de la variable respuesta, se toman a través del criterio *AIC* el cual es de fácil interpretación dentro del modelo logit, sin embargo, en los modelos de lógica difusa no se encuentra un criterio que permita establecer la importancia de cada una de las variables en función de la variable respuesta. Ver tabla 8 y tabla 9.

4. El modelo logit obtuvo una mejor eficiencia en el proceso de aprendizaje, con menor costo computacional de cálculo y tiempo de espera. Por otro lado, es un modelo de fácil interpretación en la importancia de coeficientes significativos en función de la variable respuesta y de ajuste de bondad. ver tabla 9.

5. El costo computacional para el proceso de aprendizaje de los modelos de lógica difusa son altos en costo de procesamiento y tiempo de espera, los cálculos asociados a las diferentes matrices que se generan de acuerdo con las reglas difusas planteadas, dado a ello el modelo pierde interpretabilidad asemejándose a las llamadas cajas negras de los modelos de aprendizaje automático.

6. Trabajos futuros

Como trabajos futuros se recomienda que la teoría de lógica difusa sea aplicada a diferentes problemas y conjuntos de datos con el fin de ser evaluados frente a modelos convencionales estadísticos o algoritmos de aprendizaje automático:

1. Minería de texto a partir de conjuntos difusos.
2. Identificación de fraudes con redes neuronales difusas.
3. Estimación de riesgo financiero a partir de lógica difusa.
4. Técnicas de segmentación a partir de conjuntos difusos.

Índice de figuras

Índice de cuadros

[Tablas]

Referencias

[1] Brett Lantz (2013) *Machine Learning With R* First Edition, ISBN 978-1-78216-214-8

[2] Rincón Suárez Luis Francisco (2009) *Curso Básico de Modelos Lineales.* Universidad Santo Tomás

[3] González Martínez Edwin Fernando (2018) Trabajo de Grado Universidad Santo Tomás *Detección de Fraude en Tarjetas de Crédito Mediante Técnicas de Minería de Datos.*

[4] Pérez Dinibel (2014) Trabajo de Grado Universidad Central de Venezuela *Estudio Neuro difuso de isótopo de carbono 13 e isotopo de oxígeno 18 en el límite Triásico-Jurásico.*

[5] Berlanga Francisco José, Aprendizaje de reglas difusas mediante programación genética en problemas con alta dimensionalidad, Universidad de Jaén

[6] Bolton Richard J & Hand David J. (2002). *Statistical Fraud Detection: A Review.* Statistical Science, Vol. 17, No. 3 (Aug., 2002), pp. 235-249

[7] Navarro Céspedes Juan & Casas Cardoso Gladys & González Rodríguez Emilio. (2009). *Analisis de componentes Principales y Análisis De Regresión Para Datos Categóricos.* Revista de Matematica , ISSN: 1409-2433, pp. 199-230

[8] Vila María Sanchéz Daniel & Cerda Luis. (2004) *Reglas de Asociación Aplicadas a la Detección de Fraude con Tarjetas de Crédito.* XII Congreso Español Sobre Tecnologías y Lógica Fuzzy.

[9] Yanchang Zhao, Yonghua Cen, & Justin Cen (2013) *Data Mining Applications with R* Elsevier Science, ISBN libro electrónico 9780124115200

[10] Han Jiawei, Kamber Micheline & Pei Jian (2014) *Data Mining Concepts and Techniques* Third Edition, Elsevier Science, ISBN libro electrónico 9780123814807

[11] Tjalling J. Ypma, Historical development of the Newton Raphson method, SIAM Review 37 (4), 531?551, 1995. doi:10.1137/1037125

[12] Torgo Luis (2011) *Data Mining with R Learning With Case Studies* Chapman & Hall / CRC, ISBN 9781439810187