UNIVERSIDAD NACIONAL DE COLOMBIA

Modelo para la detección de errores asociados a la liquidación de la factura de energía en Antioquia

Juan Sebastián Arboleda Restrepo

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Medellín, Colombia
2024

Modelo para la detección de errores asociados a la liquidación de la factura de energía en Antioquia

Juan Sebastián Arboleda Restrepo

Trabajo final de maestría presentada como requisito parcial para optar al título de:
Magíster en Ciencias - Estadística

Director:
Freddy Hernández Barajas
Doctor en Ciencias - Estadística

Linea de profundización:
Analítica

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Medellín, Colombia
2024

Agradecimientos

A Dios, a quien le debo todo.

A mi difunta esposa Erika, quien falleció transcurridos tres semestres de esta maestría y a quien le agradezco su amor, esfuerzo y tiempo para que yo pudiera materializar este logro. A su memoria que siempre estará presente en mi vida.

A mis adoradas hijas Luciana e Isabella, a quienes les quiero trasmitir un mensaje de optimismo y de amor, invitándolas a que pese a las muchas dificultades que sugiere la vida, no desistan en perseguir sus sueños, metas y anhelos.

A mi madre y a mi amiga Marina por su apoyo incondicional y su confianza permanente en mí.

A mi amigo y mentor Simonides Mauricio quien me invito hace unos años a conocer este mundo de los datos y la estadística.

A mi compañero de estudio Daniel, quien me acompaño en este maravilloso camino de aprendizaje y nunca me dejo solo.

Al profesor Freddy Hernández por su persistente labor, su rigor y conocimiento.

Resumen

Las empresas de servicios públicos tienen la responsabilidad de asegurar una facturación precisa de sus servicios a los clientes. En particular, la facturación de la energía eléctrica involucra diversos factores, como regulaciones, volumen de datos, cantidad de variables asociadas y restricciones temporales para la revisión, lo que complica el proceso. Las empresas a menudo no identifican de manera oportuna los errores de facturación, generando reclamaciones y clientes insatisfechos. El propósito de este estudio fue aplicar modelos estadísticos y de aprendizaje automático para detectar precozmente errores en el proceso de facturación eléctrica y mejorar el servicio al cliente. La regresión logística mostró una sensibilidad del 94.95 %, mientras que los árboles de clasificación alcanzaron una precisión del 75.36 %. La implementación de estos modelos puede perfeccionar el proceso de facturación, identificando errores de manera más eficiente. Además, la automatización y agilización del análisis de grandes conjuntos de datos facilita una detección rápida y oportuna de posibles fallos en las facturas.

Palabras claves: Servicio público domiciliario, modelo de clasificación, reclamo, facturación, sobremuestreo, submuestreo, widgets, función logística.

Abstract

Model for the detection of errors associated with the associated with the settlement of the energy energy bill in Antioquia.

Public utility companies are responsible for ensuring accurate billing of their services to customers. In particular, electricity billing depends on various factors such as regulatory requirements, data volume, number of associated variables, and limited time for review, making the process highly complex. Utility companies often fail to detect billing errors early, leading to numerous complaints and dissatisfied customers. The aim of this study was to apply statistical and machine learning models to detect billing errors in electricity billing processes early on, thereby improving customer service. Logistic regression achieved a sensitivity of 94.95 %, while classification trees achieved a precision of 75.36 %. The application of these models can enhance the billing process by identifying errors more efficiently. Moreover, automating and expediting the analysis of large datasets enables quicker and more timely detection of potential billing discrepancies.

Keywords: Residential public service, classification model, claim, billing, oversampling, undersampling, widgets, logistic function.

Contenido

Agradecimientos vii

Resumen ix

Lista de figuras xii

Lista de tablas xiv

1 Planteamiento del problema 1

2 Modelos de clasificación 3
2.1 Regresión logística simple . . . 3
2.2 Regresión logística múltiple . . . 5
2.3 Árboles de clasificación . . . 5
2.4 AdaBoost . . . 7
2.5 Selección de variables . . . 8
2.5.1 Selección exhaustiva en modelos lineales generalizados . . . 9
2.5.2 Selección hacia atrás (backward) . . . 9
2.5.3 Selección hacia adelante (forward) . . . 9
2.6 Balanceo de datos . . . 10
2.7 Paquetes R . . . 11

3 Descripción de los datos 12
3.1 Variables relacionadas con la información para facturar . . . 14
3.2 Información de reclamos . . . 17
3.3 Consolidación de los datos . . . 19
3.4 Balanceo de los datos . . . 20
3.5 Descripción de las variables en la base de datos consolidada . . . 20

4 Construcción de modelos y resultados 22
4.1 Modelo de regresión logística . . . 23
4.2 Modelo propuesto usando árboles de clasificación . . . 25
4.3 Modelo propuesto usando AdaBoost . . . 27
4.4 Evaluación de modelos propuestos . . . 28

5 Visualización de datos y dashboard **34**
5.1 Herramientas utilizadas 35
5.2 Dashboard 36

6 Conclusiones y recomendaciones **41**
6.1 Conclusiones 41
6.2 Recomendaciones 42

Lista de Figuras

2-1 Curva de la función logística, su dominio es el conjunto de todos los números reales y su rango está en el intervalo entre 0 y 1. Tomado de Wikipedia (s.f.) 4
2-2 Ejemplo de un árbol de clasificación usando la base de datos Iris. Elaboración propia. 6
2-3 Paso a paso del algoritmo AdaBoost. Elaboración propia. 8
2-4 Técnicas utilizadas para la selección de variables significativas, previo al proceso de modelamiento. Elaboración propia. 8

3-1 Flujo de facturación del servicio de energía en la empresa. Elaboración propia. 12
3-2 Porcentaje de datos faltantes para cada una de las variables de la base de datos. Valores en rojo indican porcentajes de valores faltantes $< 4\,\%$, valores en verde indican porcentajes entre $4\,\%$ y $15\,\%$, mientras que valores azules indican porcentajes $> 16\,\%$. Elaboración propia. 13
3-3 Diagrama de barras para la clasificación de las órdenes críticas de los consumos en estudio. Elaboración propia. 14
3-4 Diagrama de barras para la clasificación de las órdenes técnicas de los consumos en estudio. Elaboración propia. 15
3-5 Densidad para la variable "VALOR_ANTES_FIFA" dada en millones de pesos. Elaboración propia. 16
3-6 Boxplot para la variable "VALOR_ANTES_FIFA" diferenciando para cada uno de los niveles de la variable categoría del servicio. Elaboración propia. . 16
3-7 Densidad para la variable "VALOR_RECLAMO" dada en miles de pesos. Elaboración propia. 17
3-8 Diagrama de barras para la clasificación de los clientes en donde se presentó o no reclamos. Elaboración propia. 18
3-9 Nube de puntos con las razones por las cuales los clientes reclaman más. Entre mayor el tamaño de la palabra, mayor la frecuencia de clientes que reclaman por esa razón. Elaboración propia. 18
3-10 Esquema de consolidación de las bases de datos Facturación y Cliente. Las variables $X_1, X_2, X_3, X_4, Z_1, Z_2$ mostradas en el esquema son para ilustrar el proceso. Elaboración propia. 19

4-1 Representación gráfica del modelo árbol de clasificación. Elaboración propia. 26
4-2 Estructura de una matriz de confusión. Elaboración propia. 29

4-3 Curva ROC: evaluación de rendimiento del modelo árbol de clasificación. Elaboración propia. 31
4-4 Curva ROC: evaluación de rendimiento del modelo AdaBoost. Elaboración propia. 32
4-5 Curva ROC: evaluación de rendimiento del modelo regresión logística. Elaboración propia. 33

5-1 Pantalla de inicio del dashboard. 37
5-2 Pantalla consumos del dashboard. 38
5-3 Pantalla liquidación del dashboard. 39
5-4 Pantalla solicitudes del dashboard. 40

Lista de Tablas

4-1 Variables identificadas como importantes para el modelo de referencia (regresión logística) usando la técnica backward. Elaboración propia. 24
4-2 Tabla de resultados para el modelo de regresión logística luego del proceso de selección de variables. Elaboración propia. 24
4-3 Matriz de confusión del modelo regresión logística. Elaboración propia. . . . 25
4-4 Matriz de confusión para el árbol de clasificación. Elaboración propia. 27
4-5 Resultados de la matriz de confusión para el modelo AdaBoost. Elaboración propia. 28
4-6 Evaluación: resultados de los modelos ajustados. 30

1 Planteamiento del problema

De acuerdo a Penagos (1997) en el contexto general de servicios públicos, el producto de energía comprende el transporte de energía eléctrica desde las redes regionales de transmisión hasta el domicilio del usuario final, incluida su conexión, medición y actividades complementarias de comercialización, de transformación, interconexión y transmisión. Para efectos de este trabajo el problema está enmarcado principalmente en las actividades de comercialización asociadas a la facturación. Un producto de energía presenta error en su liquidación cuando el valor de cobro en su factura no corresponde a su realidad tarifaría y/o hay variaciones injustificadas en su consumo. La materialización de un error en la facturación genera inconformidades en los clientes que conllevan a reclamaciones y costos reputacionales para la organización.

Actualmente, en el proceso de facturación de la empresa de servicios públicos, existen varios controles que buscan evitar la presencia de estos errores en la liquidación. Por ejemplo, se tienen modelos de analítica descriptiva, automatizaciones de algunas revisiones con RPA (Robot Process Automation), listas de chequeo, etc. Todo ello está orientado a cubrir ciertas casuísticas ya identificadas. Sin embargo, al ser un proceso en el que la volumetría de información implica la revisión de 120 mil clientes de energía en promedio por día, y dado que las entidades regulatorias como la SIC (Superintendencia de Servicios Públicos) y la CREG (Comisión de Regulación de Energía y Gas) emiten constantemente decretos que implican cambios en la forma de liquidar, además de que entidades como la Fiscalía, la Contraloría y la Procuraduría vigilan el carácter público de la empresa, se hace necesario el uso de herramientas más especializadas y con una gran capacidad para desarrollar una fuerza de trabajo híbrida (empleado humano - empleado digital) que suplan las limitaciones operativas que se tienen.

Los reclamos en el servicio de energía están estrechamente relacionados con la liquidación, ya que muchos de los problemas reportados por los usuarios tienen que ver con errores o inconvenientes en la emisión de las facturas. Los errores de liquidación son una causa común de quejas por parte de los clientes, que pueden incluir cobros incorrectos, tarifas inadecuadas, lecturas erróneas de medidores y discrepancias entre el consumo real y el reflejado en la factura. Estos errores no solo generan molestias a los usuarios, sino que también pueden afectar negativamente la percepción de la empresa proveedora de energía y su reputación en el mercado.

Varios factores pueden contribuir a los reclamos relacionados con la facturación del servicio de energía. Uno de ellos es la complejidad del proceso de facturación, que involucra múltiples variables como tarifas, impuestos, cargos adicionales y descuentos, lo que aumenta la probabilidad de errores. Además, las empresas comercializadoras de energía manejan grandes volúmenes de datos, y cualquier fallo en la recolección, procesamiento o registro de la información puede resultar en facturas incorrectas.

En el presente trabajo final de maestría, se examinó la base de datos de la empresa de servicios públicos, focalizando la atención en el servicio de energía. Durante el año 2022, se gestionó un promedio de $3,900$ reclamos por mes, de los cuales 518 concluyeron favorablemente para el cliente. En ese mismo período, se registró un monto total de $6,700$ millones de pesos como resultado de las reducciones de valores aplicadas en respuesta a las reclamaciones.

2 Modelos de clasificación

Los modelos de clasificación suelen ser usados en diversos problemas como asignar un diagnóstico a un paciente, predecir si un cliente va a tomar un producto, si es apto para el pago de un crédito o si una factura posee o no errores.

Las técnicas de clasificación son una de las herramientas mas utilizadas en ciencia de datos para asignar una respuesta tipo etiqueta o categoría a un conjunto de datos previamente analizado, entendiendo patrones y relaciones (Smith y Johnson, 2020).

Según lo expuesto por Rongheng (2003), entre todas las herramientas de clasificación, tenemos que la regresión logística es una de las técnicas ampliamente utilizadas como un método para procesamiento de datos en términos de clasificación binaria y predicciones. En Soofi y Awan (2017), además de metodologías como la regresión logística, mencionan otras técnicas de aprendizaje automático como el algoritmo AdaBoost y árboles de clasificación.

Al entrenar un modelo de clasificación puede ocurrir que los datos están desbalanceados, esto quiere decir que hay un desequilibrio significativo en el número de muestras disponibles para cada clase. En un escenario de clasificación equilibrada o balanceada, se espera que todas las clases tengan una cantidad similar de datos. Esto permite que el modelo aprenda y se ajuste de manera adecuada a cada clase por igual.

Teniendo en cuenta que se hará uso de modelos supervisados y que los datos utilizados han evidenciado un desbalance importante en la variable objeto, se hace necesario complementar el análisis con el uso de algunas técnicas de balanceo de datos que permitan mejorar los insumos que utilizaran los modelos.

2.1. Regresión logística simple

Según Cramer (2003), la base de esta regresión logística fue inventada en el siglo XIX y explora los fundamentos teóricos del modelo logit, incluyendo la función de distribución acumulativa logística, que es utilizada para modelar la relación entre las variables independientes y la probabilidad de que la variable dependiente tome un valor específico.

La regresión logística es un enfoque que se basa en el modelo lineal generalizado (GLM) que permite modelar la probabilidad que una variable dependiente binaria Y tome el valor de 1 o el valor de 0. Esta variable Y se usa para estimar la probabilidad $P(Y = 1|X = x)$. Este modelo se puede representar de la siguiente forma:

$$P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}, \tag{2-1}$$

donde β_0 y β_1 son los coeficientes de regresión que se estiman a partir de los datos.

Según Yang y Li (2019) la regresión logística es el método principal para enfrentar la clasificación de datos en el campo de los grandes datos y el aprendizaje automático. Para estimar β_0 y β_1 se puede utilizar el método de máxima verosimilitud, haciendo uso del algoritmo de descenso de gradiente.

La idea principal de la regresión logística se utiliza la función logística es modelar procesos de crecimiento y saturación en los que los valores de salida deben estar limitados en un rango específico, es utilizar la función logística para no linealizar la regresión lineal multivariada, a fin de mejorar la capacidad de generalización del algoritmo. En la Figura **2-1** se muestra como la función logística permite mapear el predictor lineal $\beta_0+\beta_1 x$ a un rango probabilistico entre 0 y 1.

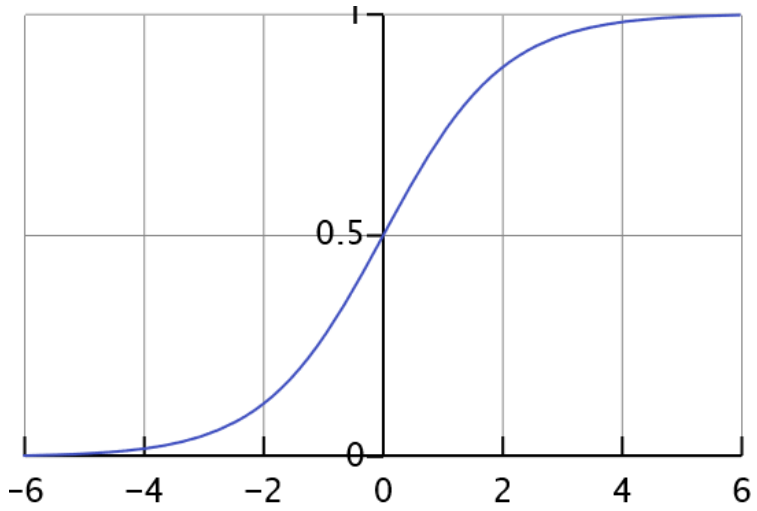

Figura 2-1: Curva de la función logística, su dominio es el conjunto de todos los números reales y su rango está en el intervalo entre 0 y 1. Tomado de Wikipedia (s.f.)

2.2. Regresión logística múltiple

La regresión logística múltiple es una extensión de la regresión lógistica simple que sirve para modelar la relación entre una variable dependiente binaria y múltiples variables independientes. Según Hosmer Jr, Lemeshow, y Sturdivant (2013), la regresión logística múltiple permite examinar el efecto conjunto de varias variables independientes en la probabilidad de que ocurra un evento.

Agresti (2015), define la ecuación de la regresión logística múltiple como:

$$P(Y = 1|X_1, X_2, \ldots, X_n) = \frac{e^{\beta_0+\beta_1 x+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x+\cdots+\beta_p x_p}}, \tag{2-2}$$

donde $P(Y = 1|X_1, X_2, \ldots, X_n)$ representa la probabilidad de que la variable dependiente sea igual a 1 dadas las variables independientes $X_1, X_2, \ldots, X_n$ y β_0, β_1, β_2, $\ldots$, β_p son los coeficientes de regresión que se estiman a partir de los datos.

2.3. Árboles de clasificación

Los árboles de clasificación son otra técnica de clasificación que se basa en la idea de dividir un conjunto de datos en subconjuntos más pequeños y homogéneos. Cada nodo del árbol representa una pregunta o una condición sobre un atributo, y las ramas representan las posibles respuestas o caminos que se pueden seguir (Breiman, 2017).

El proceso de construcción de un árbol de clasificación comienza con un nodo raíz que contiene todo el conjunto de datos. Se selecciona el atributo que mejor separa los datos en función de alguna medida como el índice de Gini. El árbol crece de manera recursiva dividiendo los datos en subconjuntos más pequeños, donde cada subconjunto se representa por un nodo y sus ramas correspondientes.

Según Gini (1912) el indice Gini es una herramienta estadística para medir la desigualdad económica. Aunque se han presentado desarrollos y refinamientos posteriores en la forma de calcular y aplicar el índice de Gini, el trabajo original de Gini proporciona el marco conceptual fundamental para su uso en la evaluación de la distribución de ingresos y riqueza en una sociedad.

En la Figura **2-2** se muestra una ilustración de un árbol de clasificación. En el lado izquierdo está el árbol y las ramas que conducen en la parte baja a las clasificaciones. En el lado derecho las particiones generadas con el árbol.

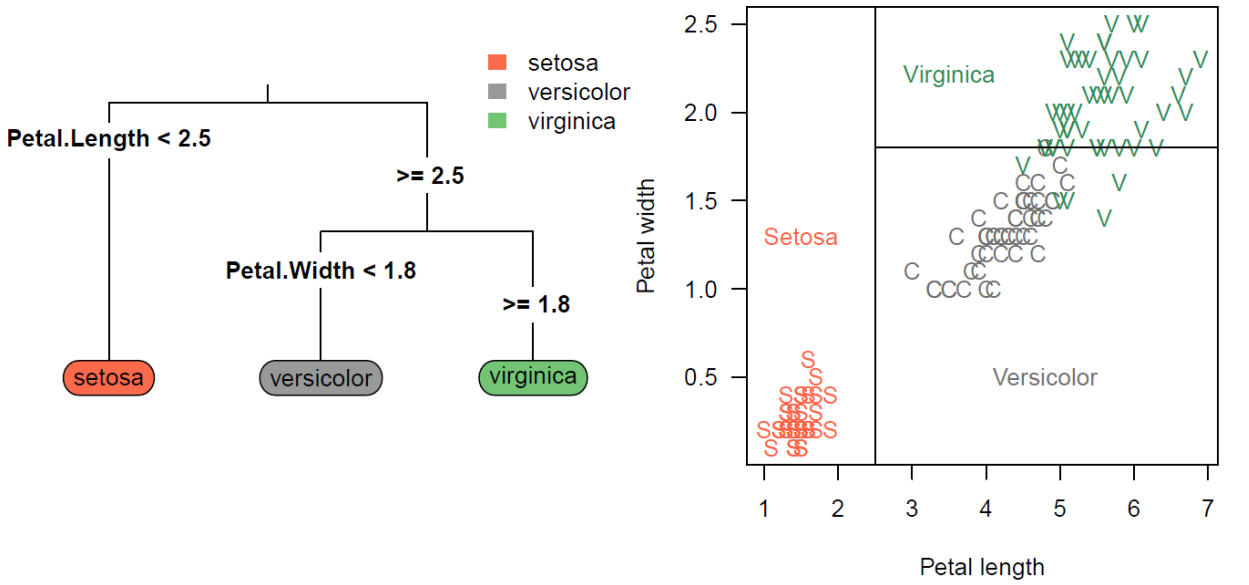

Figura 2-2: Ejemplo de un árbol de clasificación usando la base de datos Iris. Elaboración propia.

En Breiman (2001) se presenta la explicación del método asociado a un árbol de clasificación:

1. Construcción del árbol: el proceso de construcción de un árbol de clasificación comienza con un nodo raíz que contiene todo el conjunto de datos. Se selecciona un atributo y un valor de corte para dividir los datos en dos subconjuntos. Este proceso se repite recursivamente para cada subconjunto, dividiendo los datos en ramas adicionales hasta que se cumpla una condición de parada, como alcanzar una profundidad máxima o tener un número mínimo de datos en un nodo.

2. Criterio de división: en cada nodo del árbol, se utiliza un criterio para determinar qué atributo y valor de corte ofrecen la mejor división de los datos. En el caso de la clasificación, se busca maximizar la pureza de las clases en los subconjuntos resultantes. Para ello, se utilizan medidas como la ganancia de información o el índice de Gini para evaluar la calidad de la división.

3. Etiquetado de las hojas: una vez que se han construído todos los nodos y las ramas del árbol, las hojas representan las clases o categorías finales a las que se pueden asignar los datos de prueba. Cada hoja se etiqueta con la clase más común dentro de ese subconjunto de datos.

4. Poda de árboles: después de construir el árbol, se puede realizar un proceso de poda para evitar el sobreajuste. La poda implica fusionar nodos, simplificando la estructura

del árbol. Se utilizan técnicas como la poda de coste-complejidad, que equilibra la precisión del modelo con su complejidad.

5. Clasificación de nuevos datos: para clasificar un nuevo dato utilizando el árbol de clasificación, se sigue el camino desde la raíz hasta una de las hojas, tomando las decisiones basadas en los valores de los atributos. El dato se clasifica con la etiqueta de la hoja alcanzada.

2.4. AdaBoost

El término Adaboost viene de "adaptative boosting", de allí el término boosting se relaciona con un tipo de algoritmo cuya finalidad es mejorar la precisión de cualquier algoritmo de aprendizaje. Este método fue creado por Freund y Schapire (1997). El algoritmo AdaBoost es un método de aprendizaje automático que utiliza una combinación de árboles de clasificación débiles para construir un clasificador más fuerte. El concepto central detrás de AdaBoost es entrenar múltiples clasificadores débiles y combinar sus resultados ponderados para obtener una clasificación final más precisa.

Si quisiéramos explicar el paso a paso de este método tendríamos:

1. Inicie con un conjunto de entrenamiento (X, Y) con m observaciones denotadas como $(x_1, y_1), \ldots, (x_m, y_m)$ de tal manera que $x_i \in R^p$. Los valores de y deben ser -1 o 1 para aplicar el método.

2. Inicie con la distribución discreta $D_1(i) = 1/m$ que indica el peso de la observación i en la iteración 1.

3. Para $t = 1, \ldots, T$.
   - Construya un clasificador h_t definido así: $h_t : X \to \{-1, 1\}$.
   - Calcule el error asociado ϵ_t al clasificador $\epsilon_t = \sum_{i=1}^{m} D_t(i) \times \delta_i$, donde $\delta_i = 0$ si $h_t(x_i) = y_i$, es decir, si fue correcta la clasificación; caso contrario es $\delta_i = 1$.
   - Calcule la nueva distribución $D_{t+1}(i) = D_t(i) \times F_i/Z_t$, donde:
     - $F_i = \exp(-\alpha_t)$ si la clasificación fue correcta, es decir si $h_t(x_i) = y_i$.
     - $F_i = \exp(\alpha_t)$ si la clasificación fue incorrecta, es decir si $h_t(x_i) \neq y_i$.
     - $\alpha_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$.
     - Z_t es una constante de normalización de tal manera que $\sum_{i=1}^{m} D_t(i) = 1$.

4. Construya el clasificador final H_{final} como el promedio ponderado de los t clasificadores h_t, usando $H_{final} = sign(\sum_t \alpha_t h_t(x))$.

Para ver los pasos en acción y crear un Adaboost real, se recomienda ver un ejemplo en Hernández (2023). En la Figura **2-3** se describe el proceso que sigue AdaBoost:

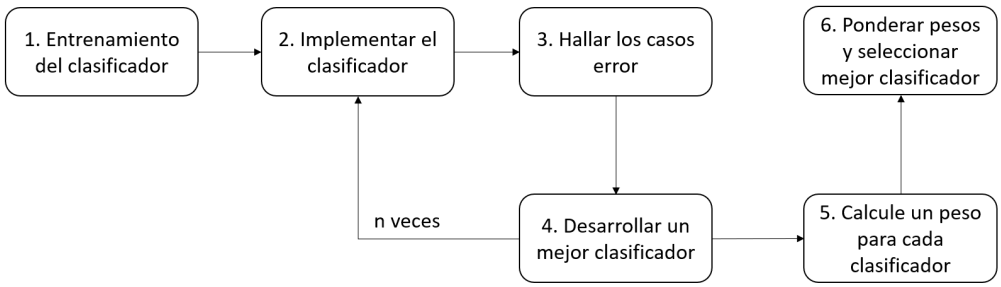

Figura 2-3: Paso a paso del algoritmo AdaBoost. Elaboración propia.

2.5. Selección de variables

La selección de variables es un paso importante en el análisis de datos y la construcción de modelos predictivos. Ayuda a identificar las variables más relevantes y eliminar las redundantes o irrelevantes, lo que puede mejorar la precisión y la eficiencia del modelo.

Antes de implementar cualquier modelo, es importante identificar cuáles variables son significativas y con ello contribuyen a la explicación de nuestra variable Y. Para lograr identificar la combinación óptima de variables para un modelo, existen muchas técnicas en la literatura (Guyon y Elisseeff, 2003). En los siguientes apartados se explican algunas de las técnicas disponibles en la literatura y que fueron usadas en este trabajo final de maestría.

En la Figura **2-4** se presentan un resumen de las técnicas más utilizadas, entre las cuales están las utilizadas en este trabajo final de maestría y otras adicionales:

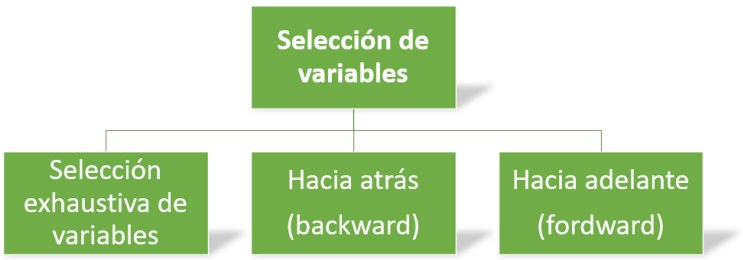

Figura 2-4: Técnicas utilizadas para la selección de variables significativas, previo al proceso de modelamiento. Elaboración propia.

2.5.1. Selección exhaustiva en modelos lineales generalizados

Para hacer selección de variables en modelos lineales generalizados hay una función en R muy útil que se llama **bestglm** del paquete **bestglm** McLeod, Xu, y McLeod (2020). Esta función realiza una búsqueda exhaustiva entre todas las combinaciones posibles de variables predictoras para encontrar el mejor modelo que minimice el criterio de selección especificado. Al modelo de referencia se le aplicó la función **bestglm**, sin embargo esta función tiene una restricción y es que solo puede manejar variables cualitativas de hasta 15 niveles. Esto significa que, si se tiene una variable categórica en el conjunto de datos con más de 15 categorías únicas, la función no podrá manejarla de manera directa y podría generar un error o un resultado inesperado. Para el caso de nuestro modelo implicaría descartar variables muy relevantes en el contexto de negocio, por esta razón se optó por utilizar otra metodología de selección de variables.

2.5.2. Selección hacia atrás (backward)

La técnica backward sirve para reducir la complejidad de un modelo utilizando como criterio de selección el BIC (bayesian information criterion). En lugar de calcular el BIC para cada modelo individual, la técnica "backward"comienza con un modelo que incluye todas las variables predictoras disponibles y luego, de forma iterativa, elimina una variable a la vez, evaluando cómo afecta esto el rendimiento BIC del modelo. En cada iteración se elimina la variable que tiene el menor impacto en el rendimiento del modelo o la que tiene el menor coeficiente de regresión, hasta que se cumplan ciertos criterios de parada predefinidos (Chen y Chen, 2008). Según Miller (2002), el proceso de eliminación hacia atrás es un enfoque en donde cada paso toma la decisión localmente óptima, eliminando la variable que tiene el menor impacto en el rendimiento del modelo en ese momento.

La función **stepAIC** en R opera realizando una búsqueda paso a paso para la selección de variables en un modelo estadístico R Core Team (2021). Su objetivo es encontrar un subconjunto óptimo de variables predictoras para un modelo, utilizando el criterio de información de Akaike como medida para evaluar y comparar diferentes modelos.

2.5.3. Selección hacia adelante (forward)

La técnica forward es un método utilizado en el análisis de datos y la modelización estadística para seleccionar un conjunto de variables predictoras relevantes para un modelo. Aquí hay una explicación clara paso a paso sobre cómo funciona (James, Witten, Hastie, Tibshirani, et al., 2013):

- Inicio con un conjunto vacío: comienzas con un modelo sin variables predictoras.

- Evaluar todas las variables: Calcular la relación entre la variable dependiente (la que estás tratando de predecir) y cada una de las variables independientes (las posibles variables predictoras) una a una.
- Seleccionar la mejor variable: de todas las variables independientes, eliges la que tenga la relación más fuerte con la variable dependiente según la medida de evaluación que hayas utilizado.
- Agregar la variable seleccionada al modelo: esta variable se convierte en la primera variable del modelo.
- Iteración: ahora, con esta primera variable en el modelo, vuelves a evaluar todas las variables restantes (incluida la variable inicial) junto con las variables restantes no incluidas, para seleccionar la segunda variable que mejore el modelo en términos de ajuste.
- Agregar la siguiente mejor variable: la segunda variable seleccionada se agrega al modelo junto con la primera.
- Repetir el proceso: continúas este proceso iterativo hasta que se cumpla algún criterio de detención.

Para el modelo de referencia con 14 variables se utilizó la función **StepAIC** del paquete **stats**, en esta ocación haciendo uso del atributo direction=**"forward"**.

2.6. Balanceo de datos

Según Wang y Sun (2021), el desequilibrio de clases es uno de los problemas más populares e importantes en el ámbito de la clasificación. Para solucionar el problema de los datos desbalanceados se han utilizado los siguientes métodos de reemuestreo:

- Submuestreo: este método consiste en disminuir la cantidad de datos para obtener una proporción de 50 a 50 en el nuevo conjunto de datos, es decir se le da prioridad a la clase minoritaria. (Donoho y Tanner, 2010)
- Sobremuestreo: a diferencia del método anterior, se generan registros adicionales que son artificiales, buscando equilibrar la clase minoritaria y obtener un conjunto de datos con una distribución 50 a 50 (Yap et al., 2014).
- Sobremuestreo minoritario sintético: es un método que genera muestras sintéticas de la clase minoritaria mediante la interpolación entre los datos existentes y sus vecinos más cercanos. Esto ayuda a abordar el desequilibrio de clases y mejorar el rendimiento de los modelos (Chawla, Bowyer, Hall, y Kegelmeyer, 2002).

2.7. Paquetes R

El problema de clasificación ha sido ampliamente estudiado en la literatura y fruto de ello son los diferentes paquetes que se encuentran implementados en el lenguaje de programación R. Algunos de los paquetes que se utilizarán en este trabajo final de maestría son:

- Adabag: es un paquete creado por Alfaro, Gámez, y García (2013) y ofrece funciones para entrenar modelos con diferentes configuraciones, como el número de iteraciones, el tipo de clasificador base y los parámetros específicos del clasificador base. También proporciona métodos para realizar predicciones utilizando el modelo entrenado y evaluar su rendimiento utilizando métricas de clasificación como la precisión, la sensibilidad y la especificidad.

- Rpart: es un paquete creado por Therneau y Atkinson (2022) y es útil para construir árboles de decisión y realizar tareas de clasificación. Permite la personalización de los parámetros de construcción del árbol y proporciona métodos para visualizar y evaluar el rendimiento del árbol resultante.

- DataExplorer: es un paquete creado por Cui (2020) y es una herramienta útil en R que facilita la exploración y visualización inicial de datos. Proporciona funciones que ayudan a resumir, visualizar y comprender rápidamente la estructura y distribución de los datos en un conjunto de datos.

- Caret: es un paquete creado por Kuhn (2008) y es una herramienta muy útil en R que se utiliza principalmente para entrenar y evaluar modelos de clasificación y regresión de manera eficiente. Caret proporciona una interfaz unificada para varios algoritmos de aprendizaje automático y simplifica el proceso de entrenamiento y validación cruzada de modelos.

- Themis: es un paquete creado por Hvitfeldt (2021) y ofrece una gran variedad de técnicas de remuestreo para el tratamiento de los datos desbalanceados.

- BestGLM: es un paquete creado por McLeod et al. (2020) y ofrece funciones para seleccionar automáticamente el mejor modelo de regresión lineal generalizado (GLM) a partir de un conjunto de variables predictoras.

- ROCR: es un paquete creado por Sing, Sander, Beerenwinkel, y Lengauer (2005) y es una herramienta que se utiliza para evaluar el rendimiento de modelos de clasificación binaria. Este paquete ofrece una variedad de funciones y gráficos que ayudan a analizar la calidad de un modelo de clasificación y su capacidad para discriminar entre dos clases.

3 Descripción de los datos

El proceso facturación tiene como objetivo, expedir oportuna y correctamente la facturación a los grupos de interés que corresponda, gestionando el cierre con lo generado por los negocios y por otras dependencias de la empresa, en cumplimiento de la normatividad legal vigente y los lineamientos aplicables para el cobro de la prestación de las soluciones. Los datos utilizados en este trabajo final de maestría provienen de un conjunto de archivos con registros históricos de facturación y de reclamos de la empresa de servicios públicos del año 2022.

El objetivo de este trabajo final de maestría es identificar los errores de la facturación proactivamente para el servicio de energía eléctrica en una importante empresa de servicios públicos. Utilizaremos la variable Valor_reclamo que contiene el histórico de errores, esta variable se denota por Y y cuando $Y = 1$ significa que la variable posee un valor en reclamación diferente de 0 (existe un reclamo a favor del cliente), en caso contrario $Y = 0$. En la Figura **3-1** se presenta el flujo de cómo operan las actividades comerciales:

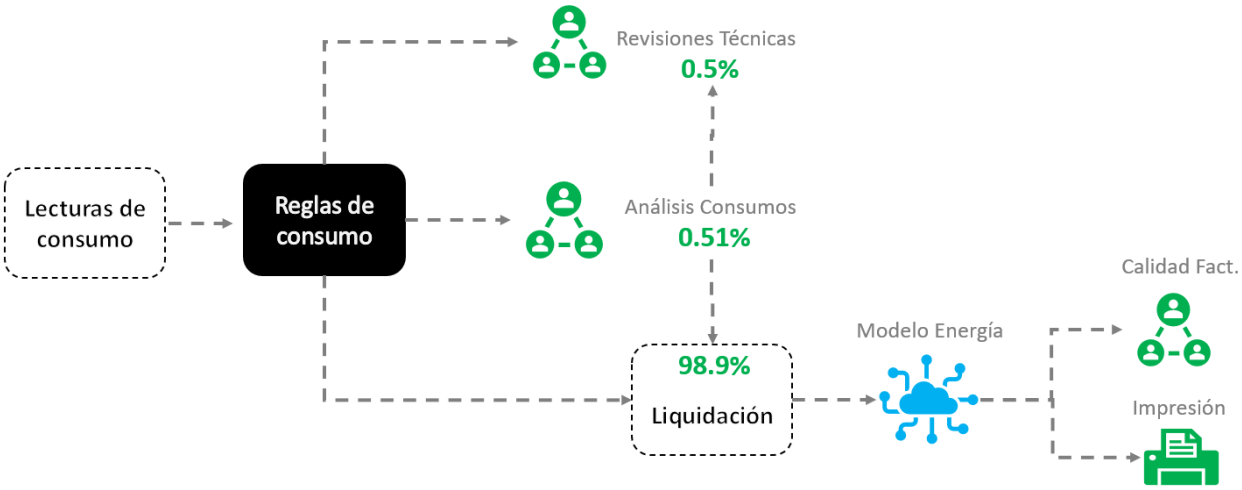

Figura 3-1: Flujo de facturación del servicio de energía en la empresa. Elaboración propia.

Usando la base de datos se llevó a cabo un análisis preliminar de las variables para explorar la completitud de los datos. En la Figura **3-2**, se observa que diecinueve variables estan en un 100 %, mientras que siete variables presentan un porcentaje de datos faltantes mayor al 10 %. Sin embargo las explicaciones desde las reglas de negocio avalan este comportamiento. La

variable "VALOR_RECUPERACIONES", es un ejemplo de una variable con gran número de observaciones faltantes, eso se debe a que no todos los clientes presentan recuperación de cobro de consumo.

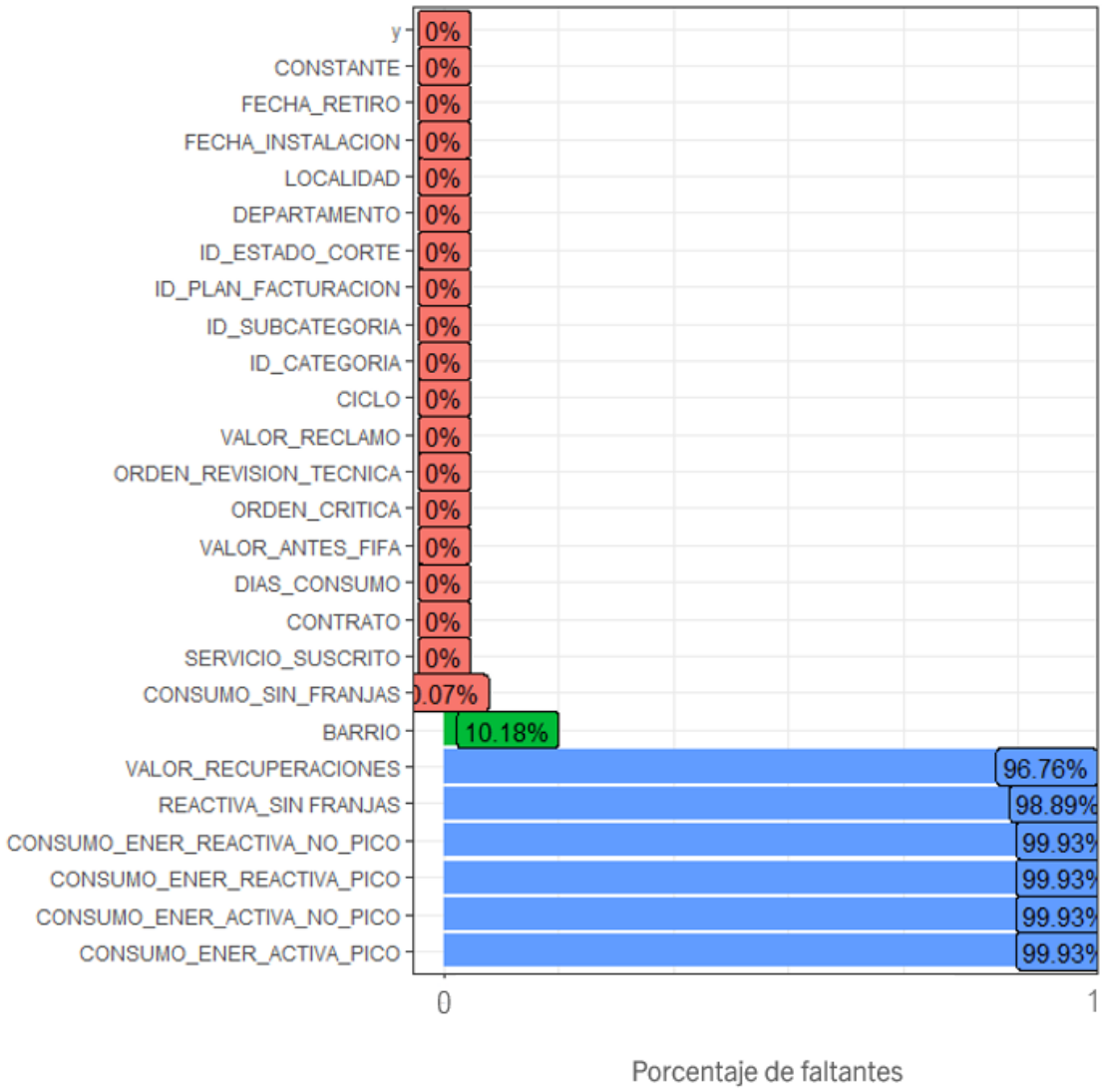

Figura 3-2: Porcentaje de datos faltantes para cada una de las variables de la base de datos. Valores en rojo indican porcentajes de valores faltantes $< 4\,\%$, valores en verde indican porcentajes entre $4\,\%$ y $15\,\%$, mientras que valores azules indican porcentajes $> 16\,\%$. Elaboración propia.

3.1. Variables relacionadas con la información para facturar

Este subconjunto de variables compone la estructura de información requerida para la facturación de un servicio de energía.

El proceso de generación de una factura implica una cadena de controles en su mayoría secuenciales. Antes del proceso de liquidación existe un subproceso que se llama "generación de ordenes de crítica" en donde se analiza el consumo, su comportamiento cliente a cliente, mes a mes, buscando identificar desviaciones de consumo. El resultado de este proceso es una clasificación en una de dos categorías, N si el consumo de este mes no presentó orden crítica y S si el consumo si presentó orden crítica. En la Figura **3-3** se hace una diferenciación del número de productos que tuvieron ese control. El gráfico muestra de manera visual la predominancia de la clase N = sin orden de critica, en el conjunto de datos, con solo una pequeña proporción de datos pertenecientes a la S = con orden de critica. Esto puede ser útil para comprender rápidamente la distribución de las clases en el conjunto de datos y su desequilibrio.

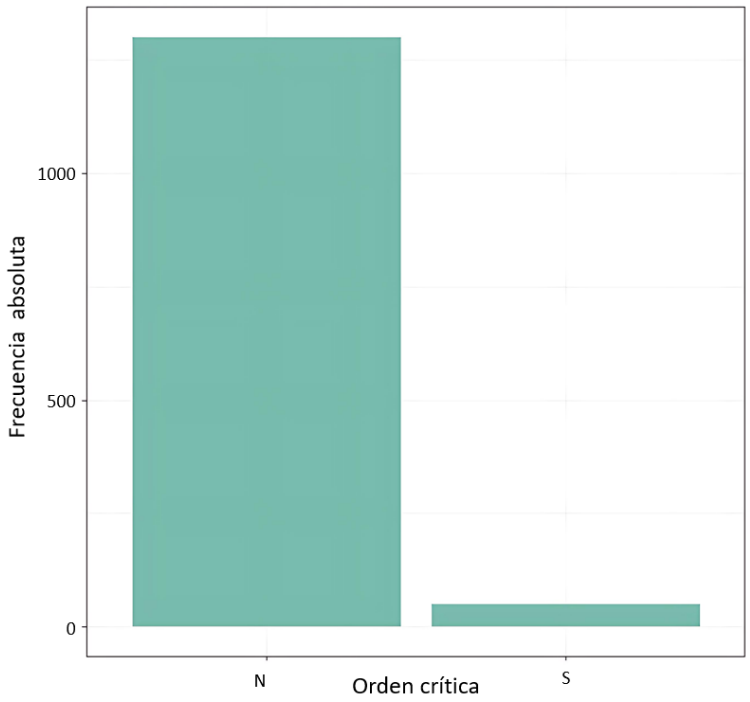

Figura 3-3: Diagrama de barras para la clasificación de las órdenes críticas de los consumos en estudio. Elaboración propia.

Existe otro subproceso llamado "generación de ordenes técnica" y que se ejecutan en el terreno al momento de hacer la lectura en el contador del inmueble. Si el consumo de inmubele

no tiene revisión técnica se denota por N, mientras que si hubo revisión técnica del consumo en el terreno se denota por S. En la Figura **3-4** se presenta el diagrama de barras para la variable orden técnica, de esta figura se observa que solo uno de los consumos analizados en el periodo Mayo a Junio del 2022 presentó revisión técnica, en este caso una visita en terreno este mes, producto de una desviación extrema de consumo. Esta es un variable dependiente que busca ayudar a priorizar aquellos casos en donde no se ha surtido ningún control previo.

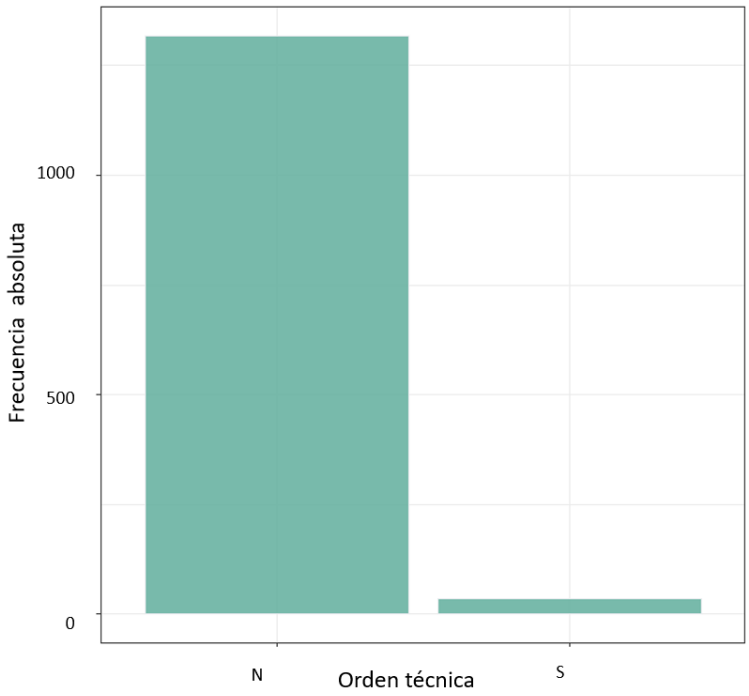

Figura 3-4: Diagrama de barras para la clasificación de las órdenes técnicas de los consumos en estudio. Elaboración propia.

En el presente estudio la información del valor facturado es sumamente relevante, esta variable en la base de datos se llama "VALOR_ANTES_FIFA", la palabra "FIFA" viene de "Forma de Impresión Factura". En la Figura **3-5**. Se puede evidenciar que los valores facturados en la mayoría de los clientes tienen valores en el servicio de energía por debajo de 1 millón de pesos. Esto denota una distribución asimétrica con un sesgo alto a la derecha. De acuerdo a lo anterior, es pertinente segmentar la variable "VALOR_ANTES_FIFA" en función de la variable categoria para identificar los valores extremos a qué tipo de clientes pueden estar asociados.

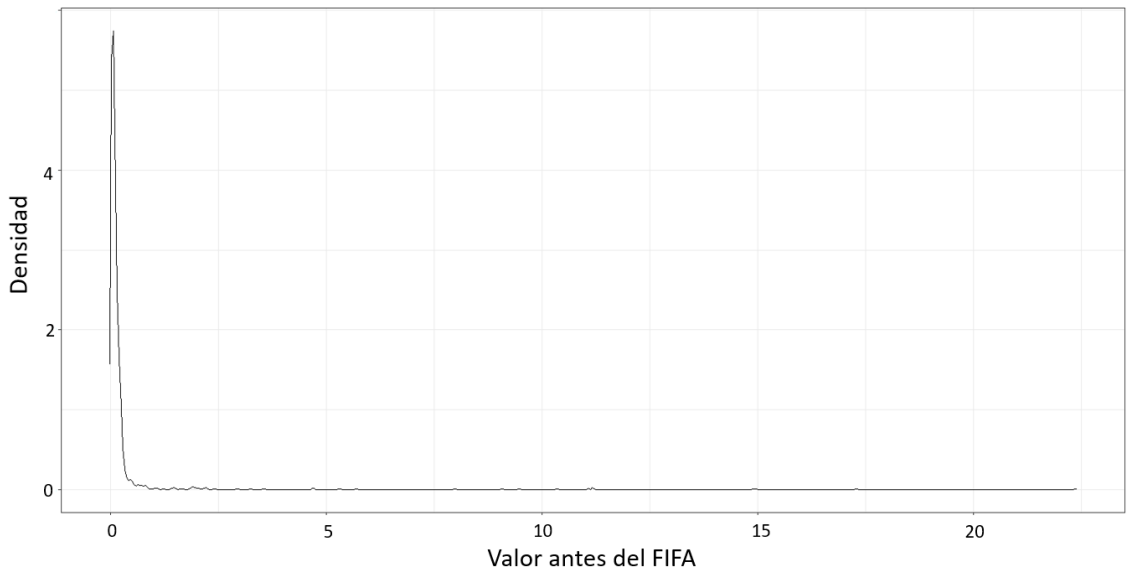

Figura 3-5: Densidad para la variable "VALOR_ANTES_FIFA" dada en millones de pesos. Elaboración propia.

En la Figura **3-6** se muestra el boxplot para la variable "VALOR_ANTES_FIFA" diferenciando por cada uno de los niveles de la variable categoría. De esta figura se observa que los valores más altos están asociados a la categoría sector industrial, lo cual tiene sentido, dado el consumo de energía que requieren para sus actividades:

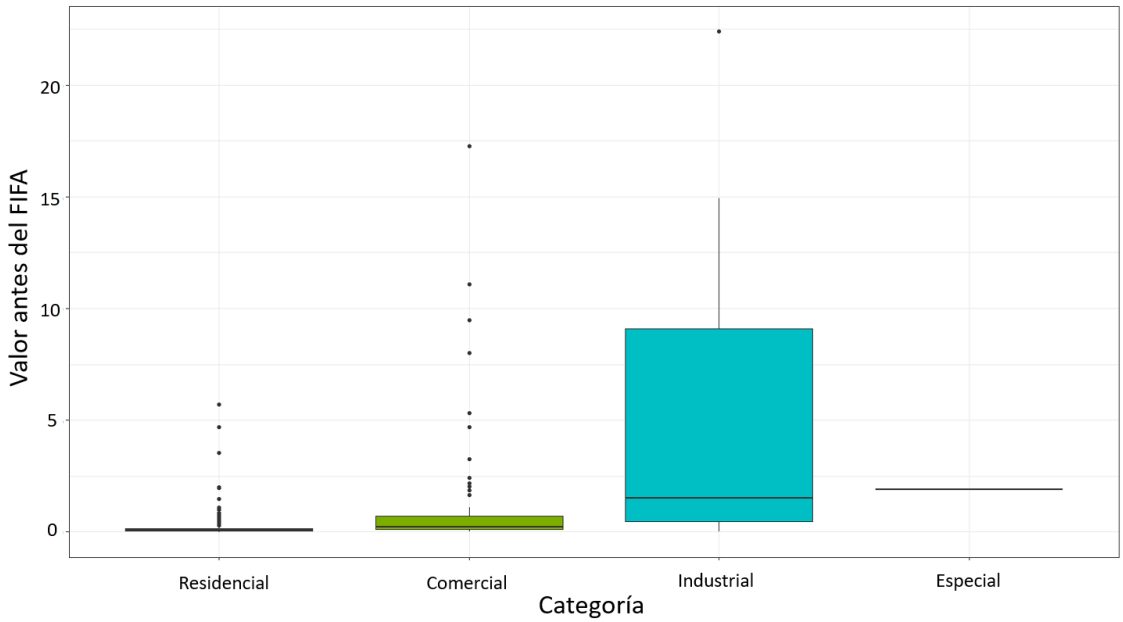

Figura 3-6: Boxplot para la variable "VALOR_ANTES_FIFA" diferenciando para cada uno de los niveles de la variable categoría del servicio. Elaboración propia.

3.2. Información de reclamos

La Ley 142 de 1994 Congreso de la República de Colombia (1994), establece unos mecanismos de defensa de los usuarios frente a la empresa de servicios públicos, uno de esos recursos es el reclamo. En el presente trabajo se consideró información histórica de reclamos relacionada con el servicio de energía eléctrica, la cual permitiera identificar los clientes que presentaron reclamaciones asociadas al periodo de facturación mayo y junio del 2022.

En la base de datos de reclamos, tenemos la variable "VALOR_RECLAMO", a partir de la cuál construimos nuestra variable Y. En la Figura **3-7** se tiene la representación de la densidad de esta variable y se identifica una distribución unimodal sesgada hacia la izquierda, con la mayor concentración de datos alrededor de los 2,757 pesos:

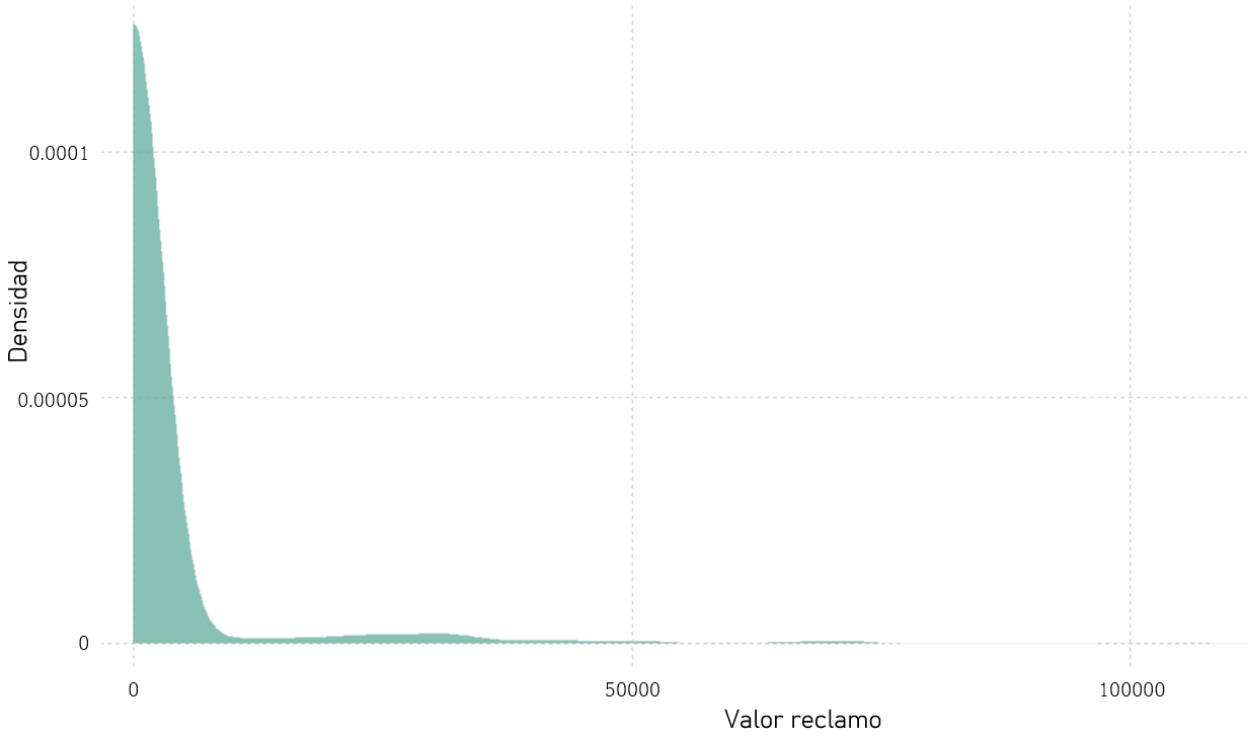

Figura 3-7: Densidad para la variable "VALOR_RECLAMO" dada en miles de pesos. Elaboración propia.

Ahora, se realizará una descripción sobre nuestra variable respuesta Y. En la Figura **3-8** se presenta un diagrama de barras en donde se representa la relación entre las dos posibles categorías de Y y sus respectivas frecuencias absolutas.

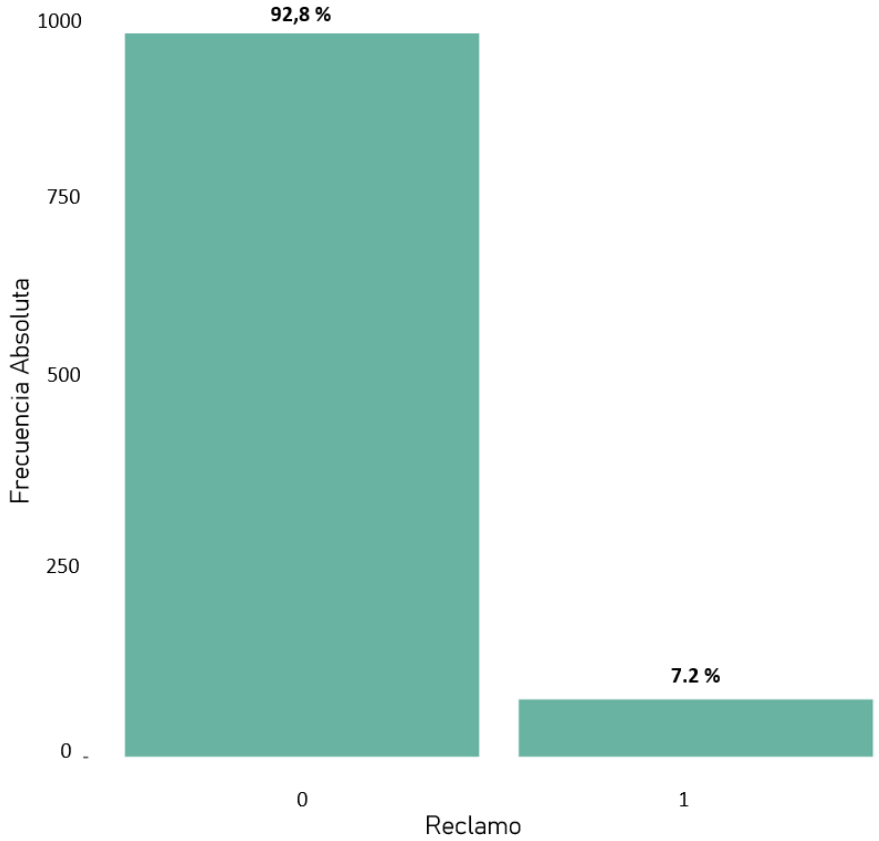

Figura 3-8: Diagrama de barras para la clasificación de los clientes en donde se presentó o no reclamos. Elaboración propia.

Finalmente, se realizó un análisis para explorar las razones sobre las cuales los clientes más reclaman y el resultado se muestra en la Figura **3-9**, usando una nube de palabras producto de un análisis de minería de texto.

Figura 3-9: Nube de puntos con las razones por las cuales los clientes reclaman más. Entre mayor el tamaño de la palabra, mayor la frecuencia de clientes que reclaman por esa razón. Elaboración propia.

De la figura anterior se observa que las tres razones por las cuales más reclaman los clientes son la lectura de medidores, las desviaciones significativas de consumo e inconformidad con los valores facturados.

3.3. Consolidación de los datos

Este proceso es necesario debido a que la extracción de los datos se hace por hilos y de diferentes fuentes de información. Por un lado tenemos la base de datos facturación de un mes típico y por otro la base de datos de reclamos. Lo primero que se realizó fue la identificación de las variables relevantes para la información de facturación. Luego se agregó a la sabana de datos la información de reclamos para así tener el conjunto de datos etiquetado con la variable "VALOR_RECLAMO", que es la variable binaria de interés en este trabajo final de maestría. En la Figura **3-10** se presenta las fuentes de información y los procesos de consolidación que se realizaron.

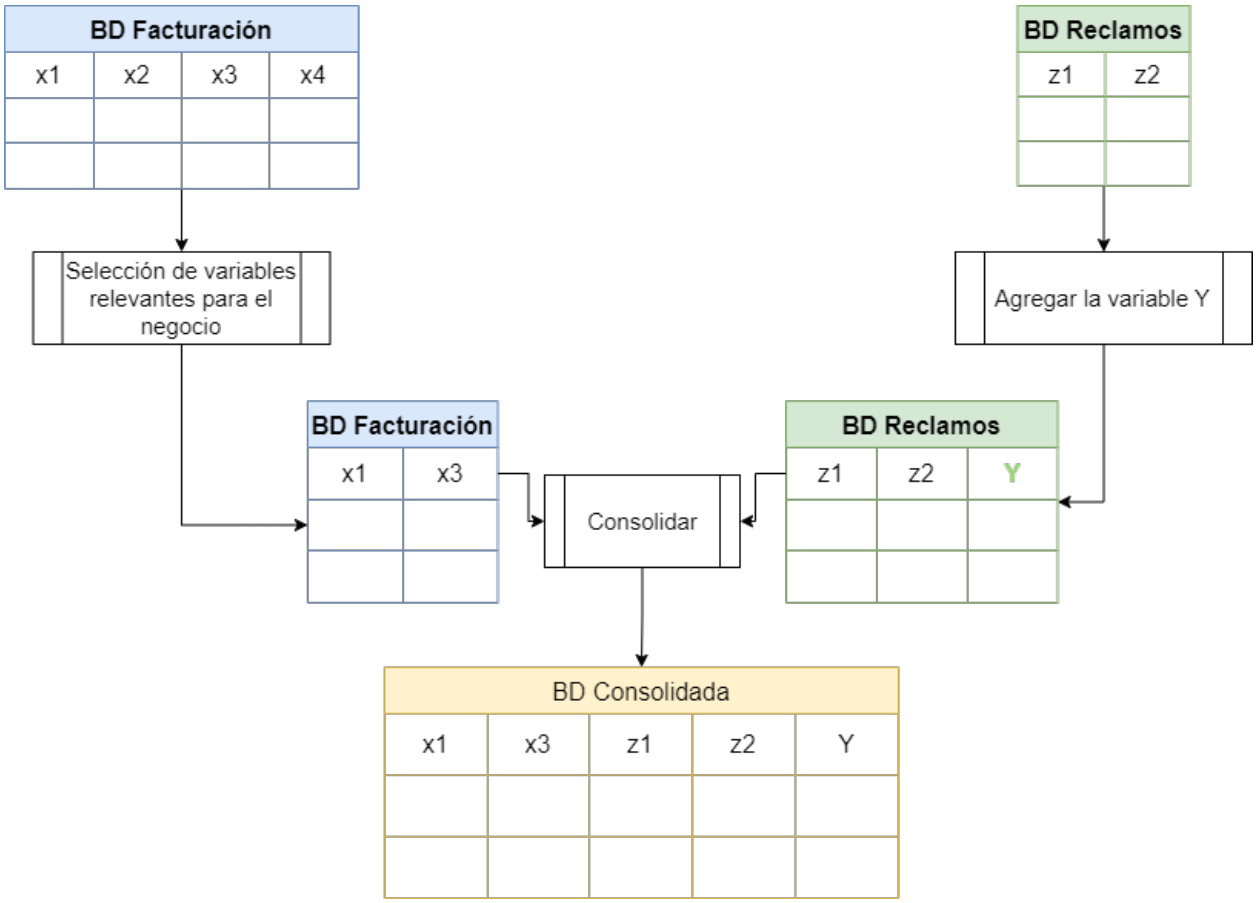

Figura 3-10: Esquema de consolidación de las bases de datos Facturación y Cliente. Las variables $X_1, X_2, X_3, X_4, Z_1, Z_2$ mostradas en el esquema son para ilustrar el proceso. Elaboración propia.

3.4. Balanceo de los datos

En la Figura **3-8** mostrada anteriormente, se observa que la variable respuesta Y reclamo está desbalanceada. De esa figura se observa claramente que hay un 7,2 % de reclamos a favor del cliente o valores de $Y = 1$, frente a un 92,8 % de reclamos no procedentes o valores $Y = 0$. El desafío de los conjuntos de datos desequilibrados radica en la necesidad de desarrollar estrategias efectivas para modelar y predecir las clases minoritarias (Fernández et al., 2018). En este trabajo final de maestría se utilizó el método de submuestreo para tener una base de entrenamiento balanceada.

3.5. Descripción de las variables en la base de datos consolidada

En esta sección se describen las variables de la base datos consolidada que se utilizarán para la fase de entrenamiento con el objetivo de ayudar a predecir si una reclamación será a favor del cliente ($Y = 1$) o si una reclamación será no procedente ($Y = 0$), usando información relacionada al servicio. Las variables de la base de datos consolidada son:

- DIAS_CONSUMO: representa el número de días que comprende el período de consumo que se está facturando al cliente. Esta variable proviene de la BD facturación.
- CONSUMO_SIN_FRANJAS: hace referencia al tipo de consumo que aplica sin restricción horaria. Esta variable proviene de la BD facturación.
- CONSUMO_ENER_ACTIVA_PICO: es un tipo de consumo atribuible al servicio de energía. Esta variable proviene de la BD facturación.
- CONSUMO_ENER_ACTIVA_NO_PICO: es un tipo de consumo atribuible al servicio de energía. Esta variable proviene de la BD facturación.
- CONSUMO_ENER_REACTIVA_PICO: es un tipo de consumo atribuible al servicio de energía. Esta variable proviene de la BD facturación.
- CONSUMO_ENER_REACTIVA_NO_PICO: es un tipo de consumo atribuible al servicio de energía. Esta variable proviene de la BD facturación.
- VALOR_ANTES_FIFA: representa el valor facturado para un cliente de energía. Esta variable proviene de la BD facturación.
- VALOR_RECUPERACIONES: comprende el valor que representa una recuperación de consumo, asociado a períodos de consumo anteriores. Esta variable proviene de la BD facturación.

- ORDEN_CRITICA: dentro del proceso de facturación existen actividades de revisión del consumo, este campo indica si el consumo asociado al cobro tuvo una revisión previa a la actividad de liquidación. Esta variable proviene de la BD facturación.
- ORDEN_REVISION_TECNICA: similar a la variable anterior, esta variable indica si hubo una revisión en terreno previa. Esta variable proviene de la BD facturación.
- ID_CATEGORIA: es una clasificación que se hace sobre los clientes, para diferenciar el cobro de la tarifa. Esta variable proviene de la BD facturación.
- ID_SUBCATEGORIA: contienen una subclasificación que define en el caso de la categoria residencial, el estrato del cliente. Esta variable proviene de la BD facturación.
- ID_ESTADO: es una clasificación que se hace sobre los clientes, para saber si están activos o suspendidos. Esta variable proviene de la BD facturación.
- ID_PLAN_FACTURACION: es una segmentación que se realiza sobre los clientes, según unos criterios de ley. Esta variable proviene de la BD facturación.
- RECLAMO (Y): variable respuesta Y que nos indica si una reclamación será a favor del cliente ($Y = 1$) o si una reclamación será no procedente ($Y = 0$). Esta variable proviene de la BD reclamos.

4 Construcción de modelos y resultados

La empresa de servicios públicos obtiene el 75 % de sus ingresos por medio de la generación, transmisión y comercialización de la energía eléctrica. Por lo anterior, es fundamental establecer controles que permitan identificar posibles errores en la facturación de este servicio. Con el tiempo se ha evidenciado que los modelos de clasificación son una metodología muy apropiada para lograr predecir estos errores. En este capítulo se busca mostrar los resultados obtenidos luego de ajustar tres modelos de clasificación que utilizan metodologías tradicionales y no tradicionales. Se presentarán las métricas de desempeño de cada modelo, para determinar cuál de los modelos genera los mejores resultados.

Para predecir si una factura tiene un error de facturación se tenían inicialmente 33 variables, sin embargo, 19 de ellas fueron descartadas para los análisis debido a que hubo un concenso en la organización de que esas variables tenían información única y de poco valor en términos del negocio. Con las restantes 14 covariables se utilizaron varias técnicas de selección de variables para reducir a un número menor de covariables que aporten realmente al desempeño del modelo.

Para la construcción de los modelos con el objetivo de predecir la variable respuesta binaria Y se procedió de la siguiente manera. Primero se ajustó un modelo estadístico de referencia, el modelo de regresión logística. Usando los resultados de este modelo, se procedió a aplicar un proceso de selección de variables para quedarnos con un modelo que sólo tuviera variables importantes o significativas. Luego con las variables seleccionadas se entrenaron los otros dos modelos de machine learning (árboles de clasificación y adaboost), de esta manera los tres modelos son comparables al tener el mismo conjunto de variables explicativas.

Para entrenar los modelos propuestos, se dividió la base de datos en dos partes: una para el entrenamiento y otra para la validación. En este enfoque, el 70 % de los datos se utiliza para entrenar el modelo, lo que permite que aprenda patrones y relaciones entre las variables. Mientras tanto, el 30 % restante se reserva para la validación, lo

que permite evaluar el rendimiento del modelo en datos no vistos y comprobar su capacidad para generalizar a nuevas observaciones. Esta división es crucial para evitar el sobreajuste del modelo a los datos de entrenamiento y garantizar que sea capaz de hacer predicciones precisas en situaciones del mundo real.

4.1. Modelo de regresión logística

El primer modelo o modelo de referencia creado para predecir la variable Y con dos niveles fue el modelo de regresión logística. Para la creación de este modelo se usó la función **glm** del paquete **stats** R Core Team (2021). Se utilizaron las 14 variables identificadas por la organización como variables aptas para el análisis. Los pasos realizados se muestran a continuación:

- Especificación del modelo: se especifica la fórmula que describe la relación entre la variable de respuesta binaria y las variables predictoras.
- Ajuste del modelo: la función glm utiliza algoritmos de optimización para ajustar los parámetros del modelo y encontrar los coeficientes que mejor se ajustan a los datos.
- Obtención de resultados: una vez ajustado el modelo, se obtienen los resultados, que incluyen los coeficientes estimados para las variables predictoras, sus p-valores y estadísticas de ajuste.
- Selección de variables: se aplicaron técnicas de selección de variables para identificar las variables que aportan en la predicción.
- Interpretación de resultados: los coeficientes estimados proporcionan información sobre la influencia de cada variable predictora en la probabilidad de que ocurra el evento de interés (la categoría positiva de la variable de respuesta).
- Predicciones y clasificación: después del ajuste, el modelo se puede utilizar para hacer predicciones en nuevos datos utilizando la función predict. Además, el modelo se puede utilizar para clasificar nuevas observaciones en función de las probabilidades obtenidas.

Luego de la aplicación de las técnicas de selección de variables, todas ellas identificaron el mismo subconjunto de covariables, representadas en la Tabla **4-1**. De esta tabla se observa que 7 de las 14 variables originales fueron identificadas como significativas.

Variables seleccionadas con la técnica backward	
Notación abreviada	Variables
X_1	$CONSUMO_SIN_FRANJAS$
X_2	$VALOR_ANTES_FIFA$
X_3	$VALOR_RECUPERACIONES$
X_4	$ORDEN_CRITICA$
X_5	$ORDEN_REVISION_TECNICA$
X_6	ID_ESTADO_CORTE
X_7	$ID_SUBCATEGORIA$

Tabla 4-1: Variables identificadas como importantes para el modelo de referencia (regresión logística) usando la técnica backward. Elaboración propia.

El modelo de regresión logística ajustado se resume en la Tabla **4-2**:

	Estimación	Error estándar	Valor z	Pr(>|z|)
(Intercept)	-0.0754	0.2606	-0.29	0.7723
CONSUMO_SIN_FRANJAS	0.0010	0.0008	1.29	0.1964
VALOR_ANTES_FIFA	0.0000	0.0000	1.06	0.2886
VALOR_RECUPERACIONES	0.0000	0.0000	1.01	0.3148
ORDEN_CRITICAS	3.0241	0.6910	4.38	0.0000
ORDEN_REVISION_TECNICAS	2.7937	1.1113	2.51	0.0119
ID_ESTADO_CORTE4	-14.4944	882.7434	-0.02	0.9869
ID_ESTADO_CORTE5	-3.4938	3.4319	-1.02	0.3087
ID_ESTADO_CORTE6	6.3925	882.7449	0.01	0.9942
ID_ESTADO_CORTE94	14.7700	882.7434	0.02	0.9867
ID_SUBCATEGORIA2	-0.8111	0.2723	-2.98	0.0029
ID_SUBCATEGORIA3	-0.6471	0.2775	-2.33	0.0197
ID_SUBCATEGORIA4	-0.7365	0.3746	-1.97	0.0493
ID_SUBCATEGORIA5	-0.0526	0.5101	-0.10	0.9179
ID_SUBCATEGORIA6	-8.2071	12.2045	-0.67	0.5013
ID_SUBCATEGORIA11	-3.0596	0.6262	-4.89	0.0000
ID_SUBCATEGORIA12	-0.2509	882.8179	-0.00	0.9998

Tabla 4-2: Tabla de resultados para el modelo de regresión logística luego del proceso de selección de variables. Elaboración propia.

Usando la información de la Tabla **4-2** se puede escribir la probabilidad estimada de que $Y = 1$, es decir, la probabilidad de que exista un error.

En la Ecuación (4-1) se muestra la probabilidad estimada.

$$\hat{P}(Y=1|x) = \text{logit}^{-1}(-0{,}0754 + 0{,}0010X_1 - 0{,}0000X_2 + 0{,}0000X_3 \\ - 3{,}0241X_4 \quad + 2{,}7937X_5 - 14{,}4944X_{6-4} + \cdots - 0{,}2509X_{7-12}). \quad (4\text{-}1)$$

Donde:

X_1: CONSUMO_SIN_FRANJAS
X_2: VALOR_ANTES_FIFA
X_3: VALOR_RECUPERACIONES
X_4: ORDEN_CRITICAS
X_5: ORDEN_REVISION_TECNICAS
X_{6-4}: ID_ESTADO_CORTE con nivel 4
$\vdots$
X_{7-12}: ID_SUBCATEGORIA con nivel 12

En la Tabla **4-3** al observar la diagonal principal se logra identificar que no posee los mayores valores a diferencia de los modelos de árbol de clasificación y AdaBoost2.

		Valores reales	
		Y=0	Y=1
Valores predichos	Y=0	232	137
	Y=1	2	35

Tabla 4-3: Matriz de confusión del modelo regresión logística. Elaboración propia.

El resultado del modelo de regresión logística representado en esta matriz de confusión entregó 232 verdaderos negativos (el modelo acertó), 137 falsos positivos (el modelo no acertó), 2 falsos negativos (el modelo no acertó) y 35 verdaderos positivos (el modelo acertó).

4.2. Modelo propuesto usando árboles de clasificación

Para construir el modelo bajo la metodología árboles de clasificación se utilizaron las variables de la Tabla **4-1** y la función **rpart**, la cuál es ampliamente utilizada en R para resolver problemas de clasificación y regresión en diversas áreas como ciencia de datos, aprendizaje automático y análisis estadístico Therneau y Atkinson (2022).

El método de particiones recursivas se basa en la idea de dividir el conjunto de datos en subconjuntos más pequeños de manera recursiva, de modo que la variabilidad dentro de cada subconjunto se minimice. Esto se realiza mediante la selección de variables predictoras y puntos de corte que mejor separan las observaciones en términos de la variable de respuesta. El resultado de aplicar árboles de clasificación a nuestro problema se ilustra en la Figura **4-1**.

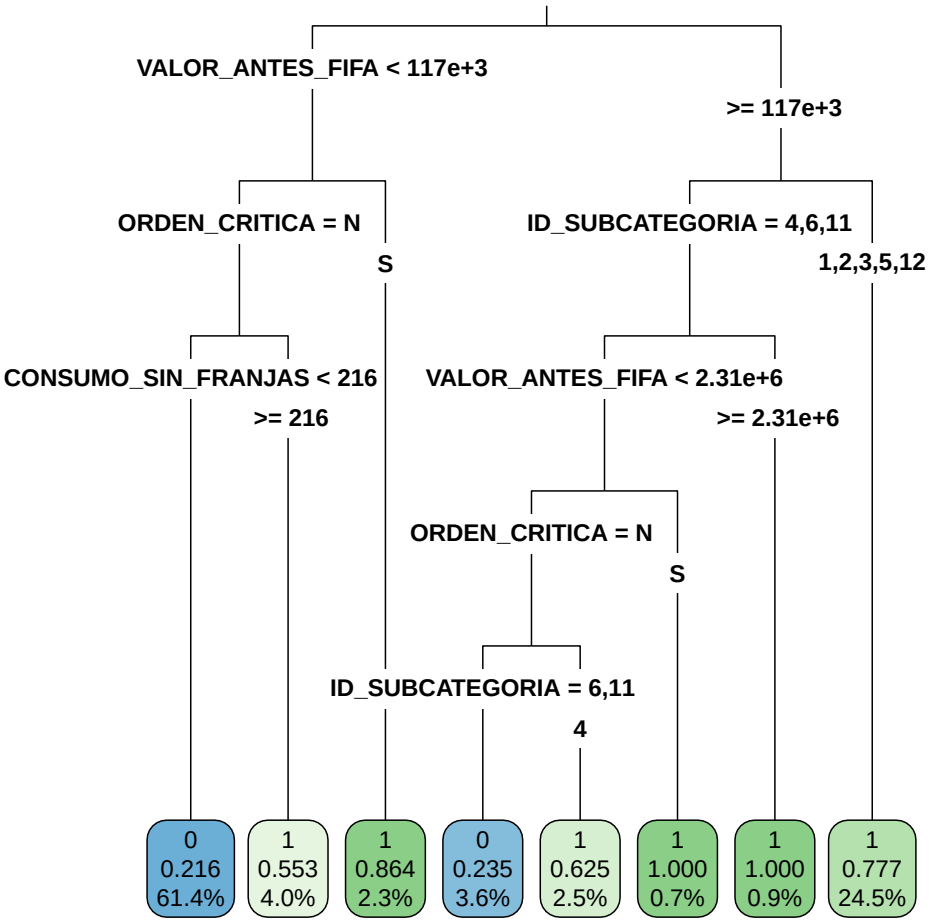

Figura 4-1: Representación gráfica del modelo árbol de clasificación. Elaboración propia.

El árbol de la Figura **4-1** tiene ocho hojas o nodos terminales, dos llevan a la predicción de $Y = 0$, es decir no hay error de facturación, mientras los seis nodos restantes llevan a la predicción de $Y = 1$, lo que significa que si hubo un error de facturación. De todas las variables usadas, sólo "VALOR_ANTES_FIFA", "CONSUMO_SIN_FRANJAS", "ORDEN_CRITICA" y "ID_SUBCATEGORIA" fueron importantes. Este árból de clasifi-

cación puede ser usado para predecir si una factura se puede catalogar como $Y = 0$ o $Y = 1$ usando las variables explicativas del árbol. El primer nodo del árbol se llama nodo raíz y contiene la variable que mejor divide los datos, en este caso la variable "VALOR_ANTES_FIFA". Podemos evidenciar que los nodos internos representan puntos de decisión, por ejemplo "ORDEN_CRITICA = N". Finalmente tenemos ocho nodos hoja, donde se muestra la predicción final.

Para explorar la utilidad de este modelo se usaron los registros u observaciones del conjunto de datos de prueba. Se tomaron las covariables y se hizo la predicción de $Y = 0$ o $Y = 1$ con el modelo. Luego, esas predicciones fueron comparadas con los valores reales de la variable respuesta error de facturación y se resumieron los resultados en la Tabla **4-4** que se conoce como matriz de confusión:

		Valores reales	
		$Y = 0$	$Y = 1$
Valores predichos	$Y = 0$	181	63
	$Y = 1$	37	125

Tabla 4-4: Matriz de confusión para el árbol de clasificación. Elaboración propia.

El resultado del árbol de clasificación representado en esta matriz de confusión entregó 181 verdaderos negativos (el modelo acertó), 63 falsos positivos (el modelo no acertó), 37 falsos negativos (el modelo no acertó) y 125 verdaderos positivos (el modelo acertó).

En una matriz de confusión se espera que las celdas de la diagonal principal tengan los mayores valores mientras que las celdas de la diagonal secundaria tengan valores pequeños, y lo más deseable es que sean valores de cero. Al observar la Tabla **4-4** se nota que la diagonal principal posee los valores mayores, lo cual es deseable.

4.3. Modelo propuesto usando AdaBoost

El algoritmo AdaBoost es un método de aprendizaje automático que se utiliza principalmente para problemas de clasificación, aunque también puede extenderse a problemas de regresión. La idea central detrás de AdaBoost es combinar múltiples clasificadores débiles para formar un clasificador fuerte y mejorar el rendimiento general del modelo. A diferencia de los árboles de clasificación y de la regresión logística, en AdaBoost no podemos crear una figura o una ecuación matemática que nos resuma el modelo creado. La única forma de resumir el modelo creado y su funcionamiento es

por medio de la matriz de confusión mostrada en la Tabla **4-5**.

		Valores reales	
		$Y = 0$	$Y = 1$
Valores predichos	$Y = 0$	170	59
	$Y = 1$	64	113

Tabla 4-5: Resultados de la matriz de confusión para el modelo AdaBoost. Elaboración propia.

En la Tabla **4-5** se presentan los resultados para este modelo luego de su estimación y posterior construcción de la matriz de confusión. Al observar la Tabla **4-5** se nota que la diagonal principal contiene los valores mayores, pero en menor proporción que los presentados por el modelo árbol de clasificación.

El resultado del modelo AdaBoost representado en esta matriz de confusión entregó 170 verdaderos negativos (el modelo acertó), 59 falsos positivos (el modelo no acertó), 64 falsos negativos (el modelo no acertó) y 113 verdaderos positivos (el modelo acertó)

4.4. Evaluación de modelos propuestos

En esta sección, se analizará el desempeño de los modelos de clasificación, que es una etapa crítica en el desarrollo de cualquier algoritmo de aprendizaje automático. La elección adecuada de métricas de desempeño es esencial para medir la capacidad predictiva de un modelo y entender su comportamiento en datos no vistos. En este contexto, se analizarán dos métricas ampliamente utilizadas: sensibilidad y exactitud, explicando sus conceptos, cálculos y sus mediciones a la luz de los modelos implementados. Utilizaremos inicialmente la matriz de confusión para evaluar el resultado de la clasificación. En la Figura **4-2** se muestra la estructura general de una matriz de confusión.

La evaluación de modelos de clasificación se basa en la comparación de las predicciones realizadas por el modelo con las etiquetas reales del conjunto de prueba. Existen diversas métricas que permiten cuantificar el rendimiento del modelo y la calidad de sus predicciones (Japkowicz y Shah, 2011).

- Sensibilidad: la sensibilidad, también conocida como recall o tasa de verdaderos positivos, es una métrica que mide la proporción de casos positivos que el modelo ha identificado correctamente. Los casos positivos se calculan como el cociente entre los verdaderos positivos y la suma de verdaderos positivos y falsos negativos. Una alta sensibilidad indica que el modelo es capaz de identificar la mayoría de

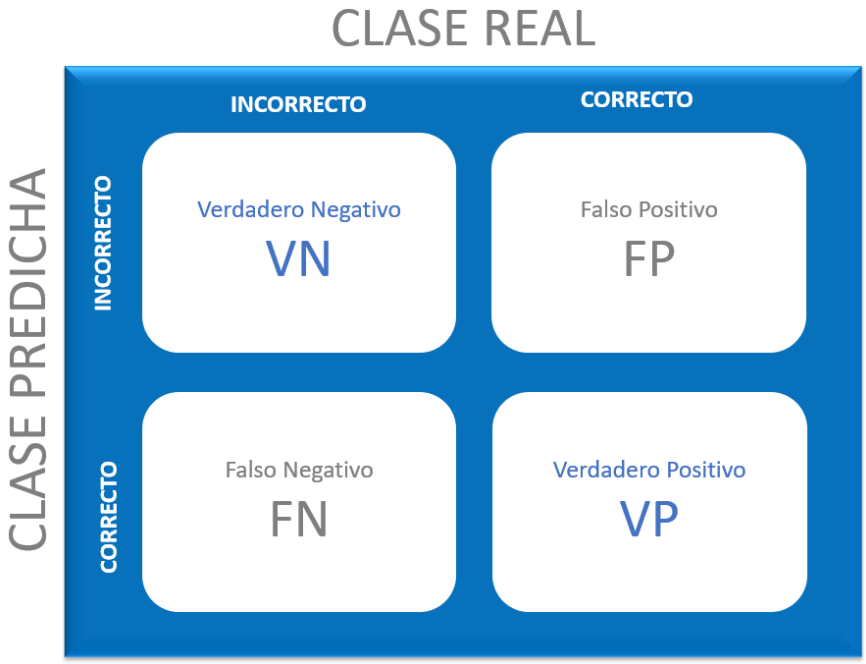

Figura 4-2: Estructura de una matriz de confusión. Elaboración propia.

los casos positivos. En la ecuación (4-2) se presenta la fórmula para calcular la sensibilidad:

$$Sensibilidad = \frac{VP}{VP + FN}. \tag{4-2}$$

- Exactitud: la exactitud es una métrica que mide la proporción de predicciones correctas del modelo en el conjunto de prueba. Se calcula como el cociente entre el número total de predicciones correctas (verdaderos positivos más verdaderos negativos) y el número total de ejemplos en el conjunto de prueba. La exactitud proporciona una visión general del rendimiento general del modelo. En la ecuación (4-3) se presenta la fórmula para calcular la exactitud:

$$Exactitud = \frac{VN + VP}{VN + FP + FN + VP}. \tag{4-3}$$

En la Tabla **4-6** se presentan los resultados de las métricas de los tres modelos ajustados, de esta tabla se observa que el modelo con la mayor exactitud fue el árbol de clasificación y que el modelo con la mayor precisión fue la regresión logística.

Comparación de modelos		
Modelo	**Exactitud**	**Sensibilidad**
Arbol de clasificación	**0.7536**	0.7716
Regresión Logística	0.6576	**0.9459**
AdaBoost	0.6970	0.6384

Tabla 4-6: Evaluación: resultados de los modelos ajustados.

Finalmente la fase de evaluación de los modelos con el uso de las curvas ROC (Receiver Operating Characteristic). En su artículo Zweig y Campbell (1993), destacan la utilidad de la curva ROC en la optimización de umbrales de decisión y su aplicación en la detección de enfermedades, lo que marcó el inicio de su amplio uso en la evaluación de modelos estadísticos.

Esta curva representa la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad) para diferentes umbrales de decisión. El área bajo la curva ROC (AUC) es un indicador ampliamente aceptado de la capacidad discriminativa de un modelo, donde un valor de 0.5 indica un rendimiento aleatorio y un valor de 1 representa un rendimiento perfecto. La curva ROC permite a los investigadores y profesionales de la estadística comparar y seleccionar modelos basados en su capacidad para distinguir entre clases. La curva ROC del modelo se traza conectando los puntos para diferentes umbrales de decisión. A medida que la curva se aleja de la línea de referencia hacia la esquina superior izquierda del gráfico, el rendimiento del modelo mejora. Esto significa que el modelo tiene una alta sensibilidad (un buen porcentaje de errores en la facturación se detecta correctamente) y una baja tasa de falsos positivos (pocos productos facturados se marcan incorrectamente como errores).

En la Figura **4-3** se presenta el área bajo la curva entregada por el modelo árbol de clasificación. Destacan algunos componentes importantes, por ejemplo el eje X de la gráfica ROC, en donde se representa la tasa de falsos positivos y el eje Y donde se representa la tasa de verdaderos positivos (sensibilidad). La forma y la posición de la curva ROC en relación con la línea de referencia indican el rendimiento del modelo. Cuanto más se curve hacia la esquina superior izquierda, mejor es el modelo para distinguir entre las clases, ya que logra un equilibrio óptimo entre la sensibilidad y la especificidad.

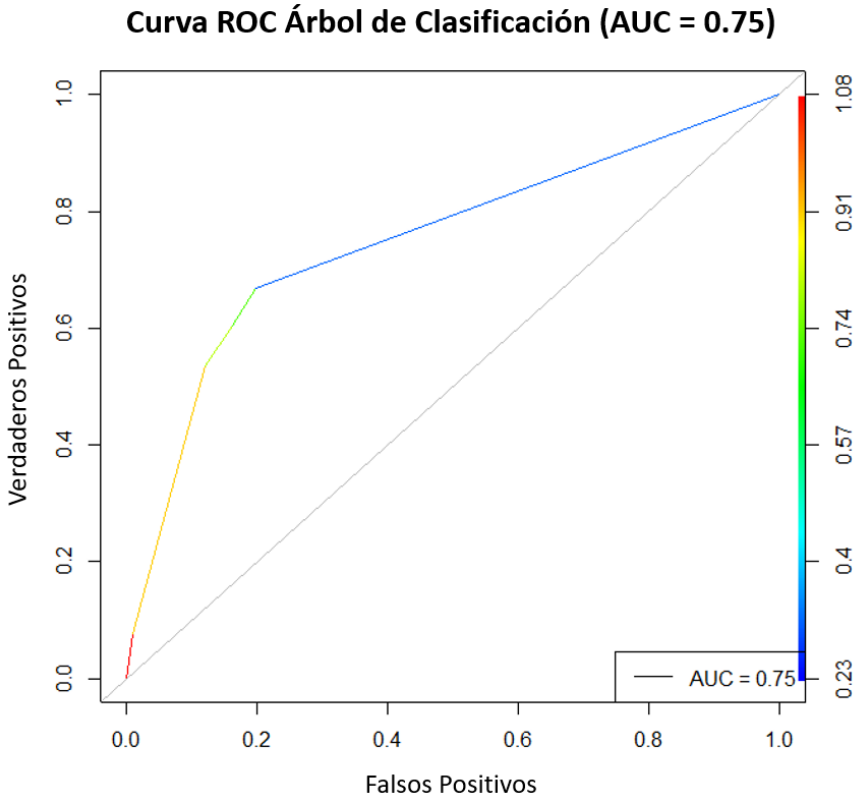

Figura 4-3: Curva ROC: evaluación de rendimiento del modelo árbol de clasificación. Elaboración propia.

Para el modelo AdaBoost se presenta una curva con un menor rendimiento. En la Figura **4-4** se logra evidenciar un resultado mas ajustado a la línea de referencia, la cual indica mayor aleatoriedad en el modelo. Para el caso de este modelo, se presenta un resultado de 0.62 en el área bajo la curva. De ello podemos inferir una menor capacidad de clasificación.

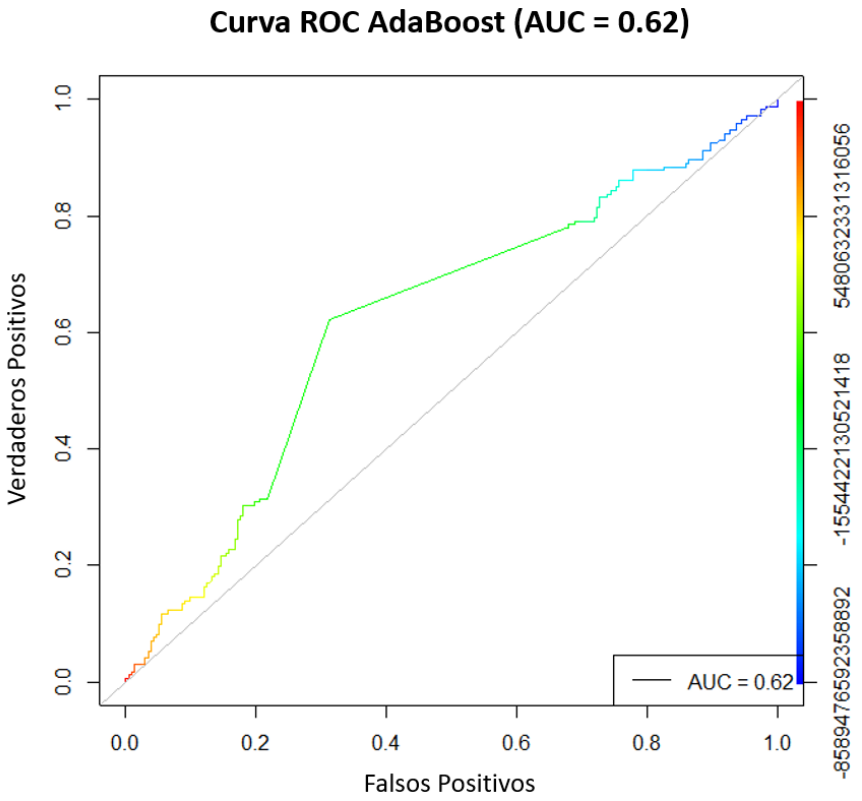

Figura 4-4: Curva ROC: evaluación de rendimiento del modelo AdaBoost. Elaboración propia.

Como lo hemos visto hasta el momento la curva ROC muestra cómo el equilibrio entre la sensibilidad y la especificidad cambia a medida que se ajusta el umbral de decisión. En general, hay un compromiso entre ambas métricas. El modelo para el cual mejor se ajusta lo anterior es el modelo de regresión logística.

En la Figura **4-5** se representa el resultado del área bajo la curva para ROC para el modelo de regresión logística. Recordemos que un modelo deseable tendrá una curva ROC que se aleja de la línea de referencia hacia la esquina superior izquierda y en donde es deseable un AUC cercano a 1.

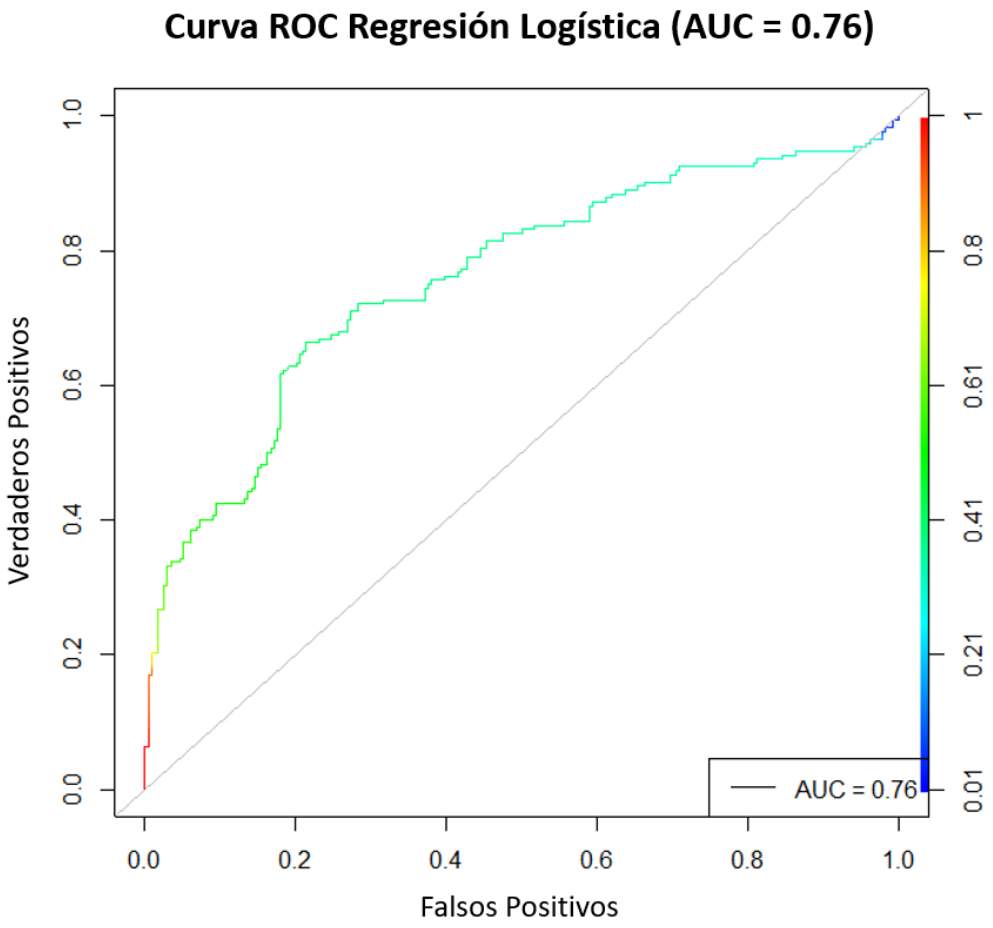

Figura 4-5: Curva ROC: evaluación de rendimiento del modelo regresión logística. Elaboración propia.

La regresión logística representa el modelo con mejor desempeño a la luz de esta metodología, entregando un valor AUC de 0.76, que se traduce en 26 puntos por encima de un sistema aleatorio ubicado en 0.5.

5 Visualización de datos y dashboard

Para cualquier área de negocio es fundamental el análisis y la comunicación de los datos, porque de esta manera se logran hacer claros los resultados y conclusiones. Existen varias formas de comunicar los hallazgos al realizar un análisis de datos y algunas a destacar son las presentaciones, los pósteres, los folletos y los dashboards. Estos últimos se convierten en herramientas imprescindibles para visualizar y entender esa información compleja de manera efectiva. Al reunir datos de diversas fuentes, los analistas pueden crear paneles interactivos que facilitan la exploración y el descubrimiento de conocimientos. Estos dashboards permiten a los usuarios interactuar con los datos, filtrarlos y personalizar la visualización según sus necesidades, lo que mejora significativamente la comprensión y el análisis de la información.

Según Few (2006), la visualización de datos es fundamental para facilitar la toma de decisiones en entornos empresariales y de análisis, de ahí el reto de lograr que las visualizaciones sean sencillas, precisas y fáciles de interpretar.

La visualización de datos es una herramienta poderosa para agregar contexto y narrativa a los datos, lo que los convierte en una parte esencial del proceso de contar historias. Al utilizar visualizaciones efectivas, los profesionales pueden hacer que los datos cobren vida y sean más comprensibles, lo que lleva a una mejor toma de decisiones (Knaflic, 2015).

La manera cómo el cerebro humano procesa y percibe la información visual, y cómo estas percepciones influyen en el diseño y la interpretación ayudan significativamente en la resolución de problemas y la toma de decisiones (Ware, 2010).

Villanueva Castillo y Reque Valqui (2018) crearon un dashboard para la cadena de farmacias Belén Farma. El Tiempo promedio de acceso a la información fue de 132.23 segundos antes del uso del Dashboard, después de la implementación del

Dashboard es de 32.16 segundos por lo que se obtiene una reducción de 100.07 segundos, que significa que se disminuyó en un 75.68 % el tiempo de acceso a la información.

Carrascal Pérez, Campanella Guerrero, y Regino León (2023) construyeron un dashboard para el control de calidad del servicio en un operador eléctrico en la Región del Caribe. Con este trabajo se logró entregar una visión clara de los patrones y tendencias en los informes de interrupciones, lo que les permitió planificar el mantenimiento, la inversión y la reorganización de recursos para fortalecer las áreas más afectadas.

Cachón Domínguez et al. (2022), Desarrollaron un dashboard para la monitorización de los datos de consumo y generación del sistema eléctrico español. Se automatizó la extracción, análisis y publicación de la información diaria, manteniendo al usuario permanentemente informado de la situación.

Granados Ostolaza (2023), crearon un dashboard en Power BI para el análisis y toma de decisiones del área de ventas en una empresa prestadora del servicio de acueducto. Se pudo determinar que la propuesta de la implementación de un dashboard permite al personal de ventas facilitar el análisis y la toma de decisiones del personal debido a la facilidad de manejo, así como también información más detallada.

En este capítulo se da a conocer el dashboard pantalla única de liquidación, herramienta que busca complementar el trabajo de priorización realizado por el modelo óptimo seleccionado en este trabajo final de maestría y consolidar la información requerida para que el analista de facturación pueda tomar mejores decisiones y con mayor oportunidad.

5.1. Herramientas utilizadas

Para este dashboard se utilizó la librería TKinter de Python que permite la creación de interfaces gráficas de usuario para aplicaciones de escritorio y se encuentra avalada por el área de tecnología de la empresa de servicios públicos. Algunos conceptos relacionados con esta librería:

- Widgets: se representa como un objeto de Python y representa una interfaz de usuario de Tkinter.

- Opciones de configuración: los widgets tienen opciones de configuración, que modifican su apariencia y comportamiento, como el texto a mostrar en una etiqueta o botón. Las diferentes clases de widgets tendrán diferentes conjuntos de opciones.
- Gestión de la geometría: los widgets no se agregan automáticamente a la interfaz de usuario cuando se crean, requieren un administrador de geometría.

Esta librería proporciona una variedad de herramientas para diseñar interfaces interactivas, lo que hace más eficiente el desarrollo de la aplicación. (Grayson, 2000).

5.2. Dashboard

A continuación se detallará la manera como el dashboard construido funciona, tomando como base los productos priorizados previamente por el modelo seleccionado en el Capítulo 4.

En la Figura **5-1** se muestra la pantalla inicial del dashboard. En la figura podemos ver como el analista de facturación deberá seleccionar el archivo de Microsoft Excel entregado por el modelo con el fin de ir analizando cada caso. También tiene la opción de consultar el producto a revisar digitándolo manualmente por alguno de los criterios definidos en el widget #1. En el segundo widget puede identificar información descriptiva del producto de energía.

Finalmente en el widget inferior se pueden identificar varios contenedores de información, para el caso de la Figura **5-1** se evidencia el primero de ellos, relacionado con la información del cliente. El objetivo de esta vista es ofrecer información muy general relacionada con el caso a revisar.

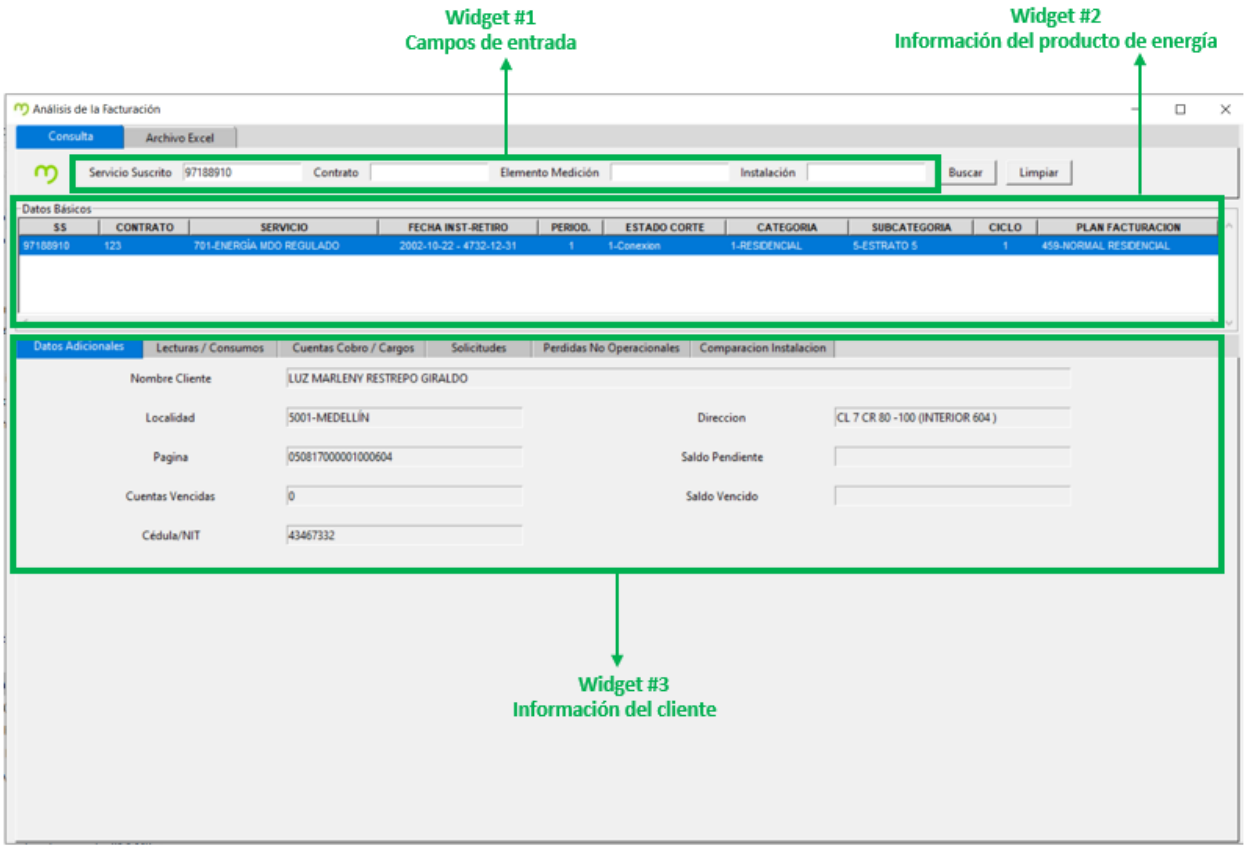

Figura 5-1: Pantalla de inicio del dashboard.

Dentro de los análisis más importantes que debe realizar el analista de facturación se encuentra la información de consumos. La información de consumo afecta la variación que pueda tener un producto de energía en su valor facturado, dado que el consumo es un componente de la liquidación junto con la tarifa. En la Figura **5-2** se muestra información sobre el histórico de consumo del cliente.

En la Figura **5-2** se puede observar un caso ilustrativo de un servicio en el cual se presenta la información relacionada con el histórico de las revisiones en terreno. Esta visualización permite identificar si han existido visitas técnicas anteriormente, entregando, informando valiosa para el análisis. Este componente del dashboard también entrega gráficos en donde el analista de facturación puede identificar tendencias.

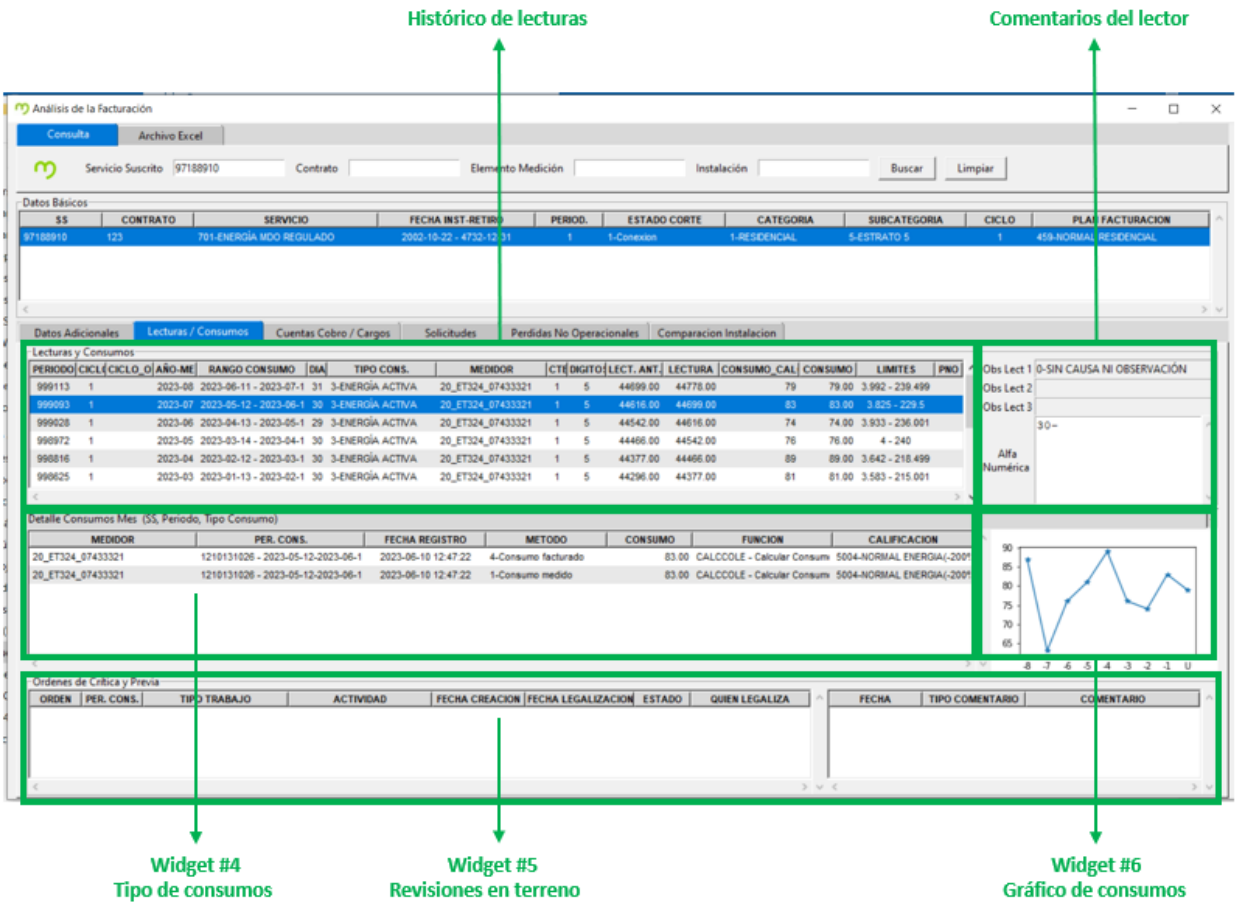

Figura 5-2: Pantalla consumos del dashboard.

La actividad de revisión de la calidad tiene como objetivo garantizar la calidad de la liquidación de los servicios y la correcta presentación de los datos en el formato de factura, así como su ajuste a la normatividad vigente, mediante la validación de la información de forma oportuna, de acuerdo con los tiempos establecidos en la programación de la facturación y previo a la generación de los archivos de impresión, identificando y corrigiendo las inconsistencias evidenciadas.

En la Figura **5-3** se encuentra la vista de la liquidación en donde convergen variables de las que depende mucho ese objetivo, como las tarifas y los conceptos de liquidación.

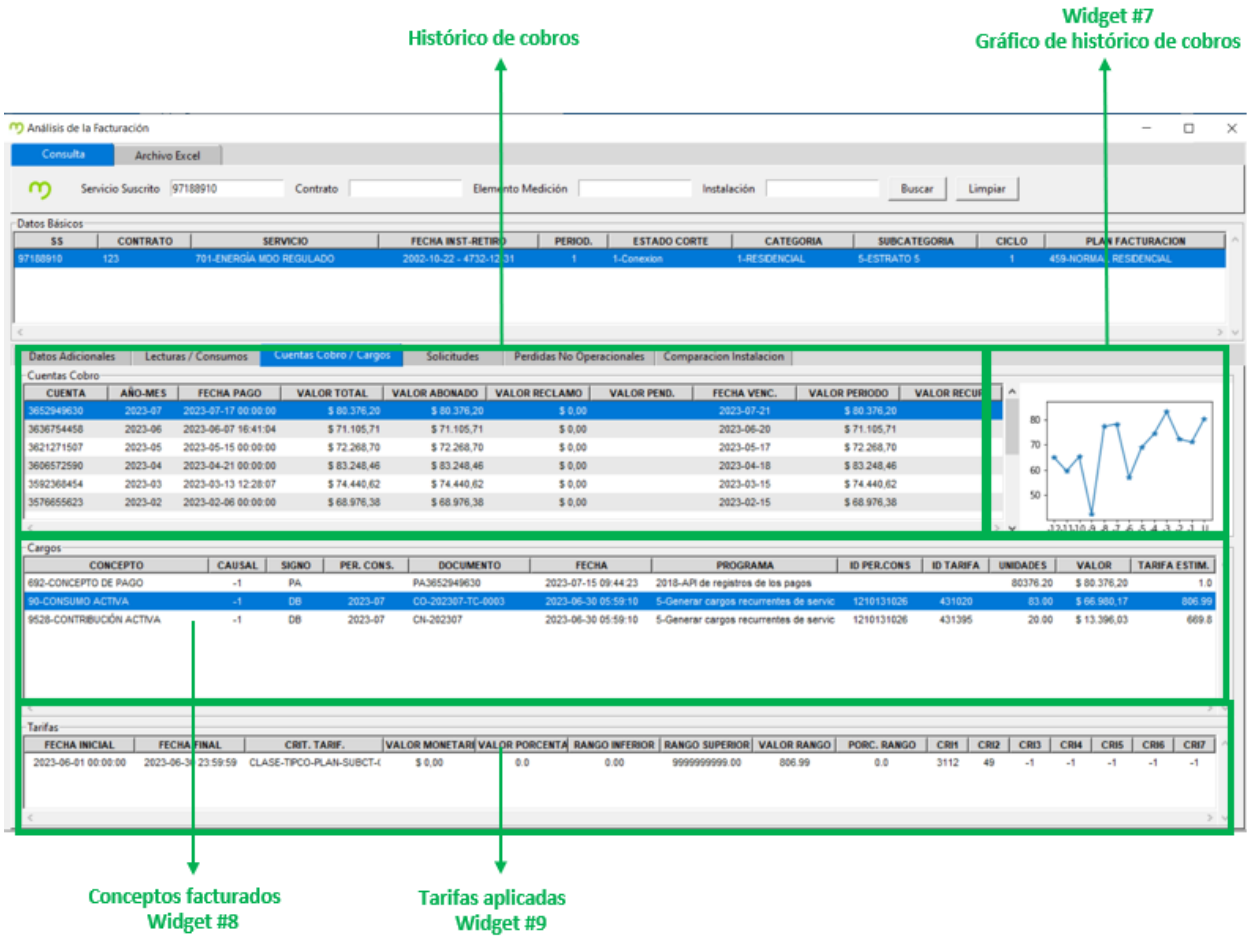

Figura 5-3: Pantalla liquidación del dashboard.

La meta con esta herramienta es que se revisen y corrigan los productos que presentan alguna inconsistencia y que deberán estar listas antes de la generación de cuentas de cobro definitivas; para ello se valida con los negocios y demás equipos de la "Unidad Facturación" involucrados en el proceso, dando solución a la causa raíz.

Un cliente puede tener varios tipos de solicitudes, por ejemplo, un trabajo producto de una reconexión del servicio. En la Figura **5-4** se presenta la historia de las solicitudes.

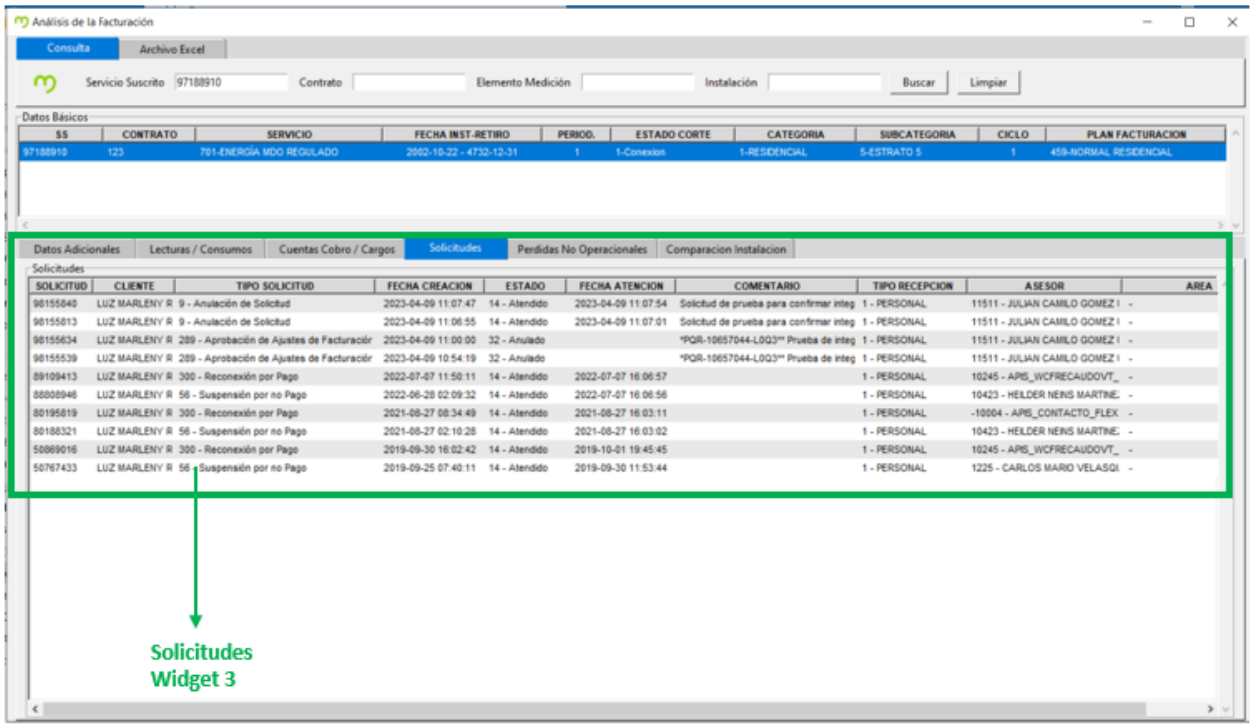

Figura 5-4: Pantalla solicitudes del dashboard.

El objetivo con este Dashboard, es que los analistas puedas contar con una herramienta que les brinde toda la información necesaria para revisar los casos priorizados. Sin esta herramienta el analista de calidad tendría que navegar por una gran cantidad de aplicaciones para poder analizar cada caso, es por esto que centralizar la información y darle un uso adecuado se vuelve sumamente necesario.

6 Conclusiones y recomendaciones

6.1. Conclusiones

1. Antes de tener el modelo y el dashboard el equipo de calidad se tomaba 8 horas para la revisión de un ciclo de facturación, en las pruebas realizadas con estas herramientas pasaron a 5 horas, lo que representa una reducción de 3 horas/ciclo de facturación en promedio en el tiempo de revisión. Lo anterior es relevante dado que se tienen 40 ciclos en un mes, para un ahorro/mes de 120 horas.
2. La eficiencia del proceso de facturación puede mejorar significativamente con el uso de modelos de clasificación para la identificación de errores en la facturación del servicio de energía. Estos modelos permiten automatizar y agilizar el análisis de grandes volúmenes de datos, lo que resulta en una detección más rápida y oportuna de posibles errores en las facturas.
3. Al utilizar modelos de clasificación, se logró identificar patrones y características específicas en los datos que pueden estar relacionados con errores en la facturación. Estos patrones pueden incluir fluctuaciones inusuales en el consumo de energía, discrepancias en los datos de medición, o comportamientos anómalos de los clientes.
4. La implementación de modelos de clasificación también puede ayudar a detectar posibles fraudes, lo que contribuye a mejorar la integridad y seguridad del proceso de facturación. Al identificar patrones sospechosos, las empresas de servicios de energía pueden tomar medidas preventivas para contribuir al indicador de pérdidas no comerciales.

6.2. Recomendaciones

1. Se sugiere revisar como el uso de estas herramientas impacta el indicador de reclamos en los meses posteriores a su puesta en producción. Lo esperado es que reduzcan los reclamos.
2. Se sugiere comenzar a desarrollar estrategias de calidad de datos sobre el modelo de datos comercial, dado que se identificaron algunas inconsistencias durante la construcción de este trabajo final de maestría.
3. Para evitar el sobreajuste y obtener estimaciones más precisas del rendimiento del modelo es crucial realizar una búsqueda sistemática de hiperparámetros para encontrar la configuración óptima del modelo. Se puede utilizar técnicas como búsqueda por cuadrícula (grid search) o búsqueda aleatoria (random search) para encontrar los hiperparámetros que mejor se adapten a los datos.

Referencias

Agresti, A. (2015). *Foundations of linear and generalized linear models*. Chicago: John Wiley & Sons.

Alfaro, E., Gámez, M., y García, N. (2013). adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software*, *54*(2), 1–35. Descargado de `http://www.jstatsoft.org/v54/i02/`

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L. (2017). *Classification and regression trees*. New York: Routledge.

Cachón Domínguez, P., y cols. (2022). Dashboard de análisis y monitorización del sistema eléctrico español. *Ediciones Universidad Oviedo*. Descargado de `https://digibuo.uniovi.es/dspace/handle/10651/64297`

Carrascal Pérez, R. D., Campanella Guerrero, G. , y Regino León, B. I. (2023). Diseño e implementación de un dashboard para el control de la calidad del servicio de energía eléctrica en operadores de red de energía del sistema de distribución local. *Ediciones Universidad Simón Bolívar*. Descargado de `https://bonga.unisimon.edu.co/handle/20.500.12442/12037`

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, J., y Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Congreso de la República de Colombia. (1994). *Ley 142 de 1994: Por la cual se establece el régimen de los servicios públicos domiciliarios y se dictan otras disposiciones.* Descargado 20 de abril de 2023, de `https://www.minenergia.gov.co/documents/9345/LEY142DE1994.pdf`

Cramer, J. S. (2003). The origins and development of the logit model. *Cambridge University Press Cambridge*, *2003*, 1–19.

Cui, B. (2020). Dataexplorer: Automate data exploration and treatment [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=DataExplorer` (R package version 0.8.2)

Donoho, D. L., y Tanner, J. (2010). Precise undersampling theorems. *Proceedings of the IEEE*, *98*(6), 913–924.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., y Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 10). Springer.

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Inc.

Freund, Y., y Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Gini, C. (1912). *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[fasc. i.]*. Tipogr. di P. Cuppini.

Granados Ostolaza, D. E. (2023). Aplicación de dashboards en power bi para el análisis y toma de decisiones en el área de ventas de la empresa distribuidora de equipos de tratamiento de agua. *Ediciones Universidad San Ignacio de Loyola*. Descargado de `https://repositorio.usil.edu.pe`

Grayson, J. E. (2000). *Python and tkinter programming*. Manning Publications Co.

Guyon, I., y Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182.

Hernández, F. (2023). *Electronic references.* Descargado 23 de Febrero del 2023, de `https://fhernanb.github.io/libro`$_m od_p red/adaboost.html$

Hosmer Jr, D. W., Lemeshow, S., y Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hvitfeldt, E. (2021). themis: Extra recipes steps for dealing with unbalanced data [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=themis` (R package version 0.1.4)

James, G., Witten, D., Hastie, T., Tibshirani, R., y cols. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Japkowicz, N., y Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, *28*, 1–26.

McLeod, A., Xu, C., y McLeod, M. A. (2020). Package 'bestglm'. *On line at https://cran. rproject. org/web/packages/bestglm/bestglm. pdf. Accessed*, *2*.

Miller, A. (2002). *Subset selection in regression*. CRC Press.

Penagos, P. A. R. (1997). *Regimen de servicios publicos domiciliarios*. Superintendencia de Servicios Publicos Domiciliarios.

R Core Team. (2021). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de

https://www.R-project.org/

Rongheng, S. (2003). *Applied mathematical statistics.* New York: Science Press.

Sing, T., Sander, O., Beerenwinkel, N., y Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, *21*(20), 7881. Descargado de http://rocr.bioinf.mpi-sb.mpg.de

Smith, J., y Johnson, M. (2020). A comparative study of classification algorithms for text sentiment analysis. *Journal of Machine Learning Research*, *21*(3), 123-145. doi: 10.5555/12345678

Soofi, A. A., y Awan, A. (2017). Classification techniques in machine learning: applications and issues. *Journal of Basic & Applied Sciences*, *13*(1), 459–465.

Therneau, T., y Atkinson, B. (2022). rpart: Recursive partitioning and regression trees [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=rpart (R package version 4.1.19)

Villanueva Castillo, D. H., y Reque Valqui, P. B. (2018). Desarrollo de un dashboard para la toma de decisiones estratégicas en la cadena de farmacias "belén farma"-áncash.

Wang, W., y Sun, D. (2021). The improved adaboost algorithms for imbalanced data classification. *Information Sciences*, *563*, 358–374.

Ware, C. (2010). *Visual thinking for design.* Elsevier.

Wikipedia. (s.f.). *Logistic function - wikimedia commons.* https://en.wikipedia.org/wiki/Logistic$_f$*unction*/*media*/*File* : *Logistic* − *curve.svg.*(30*deoctubredel*2023)

Yang, Z., y Li, D. (2019). Application of logistic regression with filter in data classification. En *2019 chinese control conference (ccc)* (p. 3755-3759). doi: 10.23919/ChiCC.2019.8865281

Yap, B. W., Abd Rani, K., Abd Rahman, H. A., Fong, S., Khairudin, Z., y Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. En *Proceedings of the first international conference on advanced data and information engineering (daeng-2013)* (pp. 13–22).

Zweig, M. H., y Campbell, G. (1993). Practical methodology of optimizing the receiver operating characteristic curve and its application to the detection of diabetic retinopathy. *Journal of Clinical Chemistry*, *39*(5), 1238–1248.