GRILLA: Grouping Recall Least Lazy Algorithm. Modelo Matemático para Cobranza Selectiva usando técnicas de aprendizaje automático

Juan Sevastian Moreno Zapata

Trabajo de grado

Olga Lucia Quintero Montoya

UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS
MAESTRÍA EN CIENCIAS DE LOS DATOS Y ANALÍTICA
MEDELLÍN
2020

Índice general

Índice de figuras

Capítulo 1

INTRODUCCIÓN

1.1. Grupo Konecta

Konecta es una organización contact Center española que inició sus operaciones en el año de 1999, prestando sus servicios de BPO (Business Process Outsourcing) a diferentes empresas en el mercado nacional e internacional (Konecta,2002). La compañía comenzó a operar en Bogotá Colombia en el año de 2010 abriendo 400 posiciones para atender la demanda de su cliente en los servicios de ADSL, Móvil Provisioning y OTC (Prezi,2020).

En la actualidad el Grupo Konecta tiene operaciones en el territorio nacional e internacional. En Colombia tiene presencia en las ciudades de Medellín, Bogotá, Cali, Montería y Barranquilla. A nivel internacional tiene instalaciones en los siguientes países: Argentina, Brasil, Chile, Colombia, Peru, México, Portugal, Marruecos y España (Konecta,2002).

1.2. Gestión de cobranza

Los aliados del Grupo Konecta son entidades financieras, las cuales poseen bases de clientes que han ingresado en mora, esta información es suministrada al área de cobranza de Konecta, con el fin de realizar los cobros a los clientes morosos. La gestión de cobro se efectúa por medio de llamadas directas con asesor, adicional, es reforzada con mensajes de texto y voz. Vale la pena reconocer que, si bien el servicio que es mayormente ocupado por los aliados de Konecta es el recogimiento de cartera, y el apoyo a la gestión de cobro, lo cierto es que, el modelo de negocio de la empresa basa su quehacer en el mejoramiento de la relación que el asociado tiene con el cliente, incluyendo eventualmente la realización de encuestas de satisfacción, concreción de modelos de cambio, adaptación de CRM (Costumer Relationship Management o Gestión de la Relación con el cliente) para las necesidades propias de cada empresa e inclusive fidelización del cliente. Es de esta manera que la empresa tiene ingresos directos a partir de

la compra de paquetes por sus aliados en donde se incluyen cierto número de llamadas por realizar o una cantidad de clientes por concretar, ello dependiendo del tipo de negociación que se realice.

En general Konecta tiene como objetivo organizacional incrementar la eficiencia y productividad de los procesos de negocio de los clientes, proporcionando un valor añadido la generación de flexibilidad para una mayor y más rápida adaptación a los cambios en el mercado.

1.3. Definición de caso de estudio

Considerando las actividades misionales que tiene la empresa, en el ejercicio de efectuar el cobro de la cartera asignada al BPO Konecta, se han identificado clientes que requieren mayor grado de esfuerzo para lograr el objetivo de recaudo u abono a la obligación, adicional, se han observado clientes que no requieren gestión efectiva o su grado de esfuerzo es mínimo para cancelar el valor vencido o efectuar abonos, ya que ellos mismos se encargan de realizar sus pagos en periodos de tiempo determinados.

En la actualidad el área de cobranza efectúa el proceso de recaudo a la cartera asignada sin un parámetro que indique cuales son los clientes que realizan sus pagos sin requerir gestiones catalogadas como efectivas, es decir, no necesitan ser contactados por asesor para cumplir con sus obligaciones, produciendo entonces un desgaste de tiempo del asesor de cobranza y un inconformismo por parte del cliente; por esta razón, se busca desarrollar un método matemático con técnicas de aprendizaje de máquinas que permita inferir que clientes cumplen con los criterios de auto cura.

Vale la pena aclara que el término auto cura se utiliza para los clientes que pagan o abonan a sus obligaciones en un rango menor o igual a quince días después haber ingresado en mora, sin tener gestiones efectivas, es decir, son clientes que se autogestionan, realizando sus pagos por los diferentes canales (Sucursal virtual, pagos físicos, débitos automáticos, entre otros).

El caso de estudio tomará la base histórica construida para realizar el análisis por cliente, e identificar diferentes patrones en la información entre los usuarios que requieren mayor esfuerzo en gestión y los que cumplen sus pagos con gestiones que no fueron efectivas por parte del área de cobranza; es decir, gestiones que no involucra contacto por parte de un asesor (mensaje de texto, correo electrónico, mensaje de voz, entre otros), estos últimos clientes serán denominados “clientes auto cura”.

Los parámetros para catalogar a los “clientes auto cura”, están fundamentados en la reconstrucción de la traza por cliente, sin embargo, aún no se tiene identificadas las variables con exactitud que definen el comportamiento de un

cliente auto cura. Dada la experiencia sobre el tema, se han seleccionado características y recalculado variables, con el fin de inferir los regresos explicativos del target (Clientes auto cura).

Aplicar el concepto de auto cura en el proceso de cobranzas del BPO Konecta, permitirá identificar los clientes que no requieren gestión con asesor, esto ayudará a la operación a enfocar sus esfuerzos en clientes que no califican con este criterio. Ante esta necesidad nace la siguiente pregunta:

¿Qué criterios tienen los humanos para el proyecto Auto Cura?

Con la ejecución de este análisis se generará etiquetas para entrenamiento supervisado que serán ocupadas en modelos de aprendizaje automático que permita inferir que cliente se pueden catalogar como "clientes auto cura".

Hipótesis del caso de estudio

Con el estudio de las diferentes variables de la base histórica, se podrá entrenar y probar diferentes modelos matemáticos y técnicas de aprendizaje de máquina, los cuales permitirán catalogar que clientes cumplen con el patrón de auto cura, esto beneficiara a la operación de cobro de la cartera asignada al BPO Konecta, ya que los esfuerzos con asesor serán enfocados a clientes que si requieren ser gestionados para obtener un pago u abono al valor vencido.

1.4. Objetivo General

Construir un modelo de Machine Learning el cual permita identificar a los clientes que poseen un comportamiento de pago en un rango menor o igual a quince días.

1.5. Objetivos Específicos

1. Construir una base histórica de la cartera de cobranzas, la cual permita conocer características importantes de las obligaciones de los clientes. Con el comportamiento de las obligaciones por cliente de la base histórica se creará una variable binaria (1,0), la cual detallará los clientes auto cura por obligación con uno y los demás con cero.
2. Realizar análisis descriptivo de las variables de la base histórica.
3. Seleccionar variables de mayor influencia para identificar los clientes auto cura con modelos estadísticos.

4. Entrenar y evaluar modelos de Machine Learning.

5. Evaluar diferentes métricas en los modelos de Machine Learning, tales como: ROC, Exactitud, Exhaustividad, Precisión y F1 score.

1.6. Metodología

1.6.1. Revisión teórica del caso de estudio

- Investigación de métodos utilizados por organizaciones del sector de cobranza para efectuar sus procesos de cobro.
- Estudio de métodos matemáticos y aprendizaje de máquina.

1.6.2. Integración de bases de cobranza

En esta fase se integrarán diferentes bases de datos de cobranzas, con el fin de construir la base histórica que permita identificar las características de los clientes.

- Base de Traslados: Esta base de datos posee la información de los clientes asignados para cobro por día.
- Gestiones: Esta base detalla las gestiones realizadas a cada uno de los clientes por su consecutivo..
- Tanque de pagos: Esta base posee la información de los pagos efectuados por los clientes por número consecutivo de la obligación.

1.6.3. Realización de Análisis Descriptivo de la base histórica

- Este análisis se efectuará para tener conocimiento previo de las variables de la base histórica, con el fin de identificar gráficamente el comportamiento de cada una. Esta visualización se desarrollará con el software R utilizando el complemento Mark Down.

1.6.4. Partición de base, Selección de variables y estandarización de variables

Esta fase es crítica para el desarrollo del proyecto, ya que se debe efectuar los siguientes pasos:

- Para probar los modelos, se ocupará una base transformada con una dimensión de 157.696 registros y 52 variables. Con la intención de garantizar la generalización de los modelos, se particionará la base histórica en entrenamiento, validación y testeo, con la siguiente proporción respectivamente:

(60 %, 20 % y 20 %), dicha partición se realiza aleatoriamente, garantizando que la proporción de la variable objetivo se mantenga en las tres bases resultantes.

- Debido a que las variables poseen diferentes unidades de medida, se debe proceder con la estandarización de estas, con el fin de eliminar las medidas de cada una de las características y así trabajar los datos sin dimensión (Garcia Polanco, 2019).
- Al realizar la selección de variables, se descartarán las características que no aportan valor a la generalización del caso de estudio, de esta forma se obtendrá el modelo de datos con el cual se desarrollarán las técnicas de aprendizaje de máquinas.
- Al transformar las variables categóricas de la base, se obtiene un conjunto de entrenamiento con 52 variables. Al generar los modelos con toda la dimensión de la base, se puede generar diferentes inconvenientes, los cuales son: incrementar el tiempo de cómputo del procesamiento de los modelos, colinealidad, alta correlación entre las variables y el más importante se genera en que el modelo seleccionado se va a aprender los datos de entrenamiento, obteniendo buenos resultados en las métricas de evaluación, sin embargo, al probar el modelo con la base de validación, su rendimientos van a estar por debajo a los resultados del entrenamiento, generando sobreajuste o alta varianza, lo cual no permite una correcta generalización al momento de la predicción.

  Para evitar el sobre ajuste de los modelos, se aplicarán las técnicas matemáticas de Ridge, Lasso y Elasticnet. Al aplicar estos métodos se debe ser muy cuidadoso, ya que al ingresar menos variables predictoras que el modelo requiera para realizar la predicción, se aumentara el error de sesgo, es decir, que en el entrenamiento el rendimiento de las métricas no son las deseadas, por ende, el modelo no generalizará con observaciones nuevas (Carrasco, 2016).

1.6.5. Entrenamiento y prueba de diferentes métodos matemáticos y aprendizaje de máquina.

- Dado que, la base histórica construida no posee la variable dependiente, esta se reconstruirá mediante aprendizaje no supervisado tomando el comportamiento de las obligaciones de los clientes, con el fin de obtener el grupo que describa el comportamiento del criterio auto cura, y así, obtener la característica dependiente o target.
- Luego de obtener la variable dependiente o target mediante el aprendizaje no supervisado, se procederá a experimentar con diferentes métodos de aprendizaje de máquinas supervisado, tales como: Regresión Logística, Radom Forest, Maquinas de soporte vectorial y K vecinos más cercanos (Knn) (Bishop, 2006).

1.6.6. Evaluación del entrenamiento y prueba de los diferentes métodos matemáticos de aprendizaje de máquina.

- Para evaluar el rendimiento de los métodos de aprendizaje supervisado se ejecutará la matriz de confusión la cual detallará el Target real y el valor predicho con los siguientes parámetros: verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, con base a estos resultados se calcularán las siguientes métricas (Izenman,2008):

- ROC.

- Accuracy.

- Recall.

- Precision.

- F1 Score.

Capítulo 2

MARCO CONCEPTUAL

2.1. Compañías BPO

Las organizaciones BPO (Business Process Outsourcing), son prestadoras de servicios a diferentes compañías del mercado, las cuales, a través de estrategias conjuntas entre las partes, mejoran los procesos comerciales utilizando capital humano especializado y tecnología de vanguardia, permitiendo obtener una ventaja competitiva a sus clientes frente a sus competidores (Schneider, 2012).

Las empresas BPO se han convertido en aliados claves para las compañías en la actualidad, ya que estas, aportan su experiencia y conocimiento a los procesos de sus clientes. Este trabajo en equipo permite maximizar los beneficios, disminuir los costos y explotar las ventajas competitivas. Al revisar la teoría económica se identifica que este es uno de los axiomas transcendentales con respecto a las empresas y su razón de ser en el mercado. El entorno empresarial exige que las empresas estén en continua evolución, mejorando sus procesos y hagan uso de colaboración en el desarrollo de las acciones de forma que avalen la permanencia y la competitividad en el medio (Duque, Gonzalez García, 2014).

2.2. Contac Center

El Contact center es una modalidad del BPO, la cual presta servicios de atención a los clientes a través de chat, redes sociales, email, líneas telefónicas, entre otros medios (Rangel, 2010).

El contar con el grupo de herramientas, anteriormente mencionadas, hace que se favorezca la combinación de tecnologías de comunicación e información, en pro de generar cada vez más disponibilidad para la realización de interacciones personalizadas con los clientes (Sanabria, 2015).

Las empresas contact center son bastante eficientes en la gestión comunicativa

con los clientes, lo cual se traduce para cualquier organización en la maximización de recursos y la reducción de costos, no obstante, su accionar se encuentra fundamentado en la generación de acercamiento con los clientes de la compañía. La actividad misional de los contact center es generar una mayor productividad en sus procesos, logrando altos índices de satisfacción y generando en el proceso que las relaciones cliente – empresa sean sostenibles en el tiempo (Sanabria, 2015).

2.3. Machine learning

La ventaja competitiva de las organizaciones está altamente correlacionada con la capacidad de convertir la información de sus clientes en conocimiento que le permita generar estrategias que marquen diferencia en la ejecución de sus procesos. La utilización de soluciones analíticas, elaboradas con el fin de explotar la información histórica de sus clientes, puede ayudar a entender rápidamente que está sucediendo en la ejecución de sus actividades, y así, tomar decisiones acertadas fundamentadas en los datos (TransUnion, 2018).

El aprendizaje automático de máquinas está fundamentado en modelos matemáticos y métodos estadísticos, lo cual permite que los algoritmos aprendan a reconocer patrones a partir de los datos, de acuerdo a este conocimiento adquirido se genera inferencia de los casos de estudio (Bishop, 2006).

Las técnicas de aprendizaje automático o machine learning aprenden directamente de los registros. Estos algoritmos mejoran su capacidad predictiva a razón de que se vayan incorporando nuevos datos al modelo existente, es decir, con los nuevos registros obtenidos se reentrena el modelo para que se ajuste a la nueva tendencia de los datos.

En las técnicas de aprendizaje automático existen tres tipos de metodologías, las cuales son: aprendizaje supervisado, no supervisado y seme supervisado, sin embargo, el caso de estudio se centra en el aprendizaje no supervisado y supervisado (MathWorks, 2019).

2.3.1. Aprendizaje de máquinas supervisado

El aprendizaje de máquinas supervisado se utiliza para crear modelos que generen predicciones a partir de los datos suministrados. Los registros de entrada poseen su respectiva respuesta, o etiqueta, a esta variable se conoce como objetivo o dependiente, con esta información se realiza el entrenamiento para efectuar predicciones al ingreso de nuevos datos (Bishop, 2006).

El aprendizaje automático supervisado posee dos metodologías primordiales, las cuales son: algoritmos de regresión y clasificación.

- Los algoritmos diseñados para regresión: en estos su predicción es continua (Bishop, 2006), algunos ejemplos para regresión son: predecir el cambio de la temperatura, la demanda de llamadas que recibirá un call center, predecir las ventas de un periodo, entre otros.

- Las técnicas supervisadas de clasificación realizan predicciones binarias o discretas, si tiene algunos ejemplos para clasificación: si una transacción financiera es real o fraude, si un cliente es apto para adquirir una tarjeta de crédito o no, entre otros ejemplos.

2.3.2. Aprendizaje de máquinas no supervisado

El aprendizaje de máquinas no supervisado se utiliza cuando los datos suministrados no poseen la variable objetivo (etiqueta), su principal función es hallar patrones en los datos. Este se aplica para realizar análisis de agrupaciones en los registros, donde los elementos pertenecientes a un grupo son similares entre sí y diferentes con respecto a elementos pertenecientes a otro grupo de datos. Esta técnica no tiene capacidad predictiva; para realizar el análisis exploratorio se utiliza el conjunto de datos completo (MathWorks, 2019).

2.4. Aplicaciones en cobranzas

En la gestión de cobranzas, se ha implementado IA Bots; esta tecnología utiliza la inteligencia artificial para elaborar conversaciones con los clientes. Los esfuerzos realizados hasta el momento se han focalizado en la digitalización de la experiencia del cliente, adicional, se han implementado en algunos procesos internos de solicitud de créditos, con el almacenamiento de datos y sus etiquetas, saldos y canales de pago. Las áreas de los BPO encargadas de realizar la recuperación de las obligaciones vencidas se encuentran lejos de los procesos de transformación, debido a que, las actuaciones en esta actividad, específicamente en las etapas de mora temprana, ocupan generalmente herramientas como: mensajes personalizados, mensajes de texto, o llamadas telefónicas, destinadas a recordar el pago de sus obligaciones (IBR Latam, 2018).

En la actualidad, la implementación de aprendizaje automático se encuentra dando sus primeros pasos en la gestión de cobro de obligaciones, sin embargo, se espera que en los próximos años la implementación de esta madure aportando cada vez más a la tarea mencionada. Uno de los principales objetivos que se tiene con la implementación de Machine learning en el proceso de cobranza es desde la recolección de información de los diferentes sistemas tipificar al cliente, con la intención de realizar una gestión de cobro inteligente. A partir de la clasificación de la información se puede ocupar el aprendizaje de máquinas supervisado y no supervisado con la intención de conocer el comportamiento de los clientes que presenten mora y desde este proceso orientar la toma de decisiones con respecto a estos (Morales, 2019).

La administración del área de cobranzas está enfocada en la asignación optima de recursos, por esta razón, es de vital importancia mapear la probabilidad de pago de los clientes, franjas de pago, entre otras variables para asignar de forma objetiva los agentes especializados en el cobro. Los modelos predictivos ayudan a priorizar la gestión de cobro y la asignación de recursos, por medio de un Score que unifique las variables que describan el comportamiento de los clientes (Martinez, 2014).

Esta actividad se hace prudente precisando que con la implementación de un sistema de manejo de grandes volúmenes de datos y el Machine learning, los BPO optimizan sus procesos de cobranzas. De hecho, en la investigación de Bonastre (2017) se ha comprobado que usar estas metodologías predice un aumento en los ingresos del 30 % con ahorros de costo operación del 25 %, adicionalmente, se disminuyen los litigios por mora.

Capítulo 3

ANÁLISIS DE LOS DATOS DE KONECTA

3.1. Descripción de las bases de datos

Cada una de las tres bases de datos contiene un histórico de seis meses de desarrollo del proceso de cobranzas, los cuales comprenden información sistematizada de octubre de 2019 hasta el mes de marzo 2020. En las tres bases se cuenta con dos identificadores del cliente, por el cual, se realizó el proceso de integración de las bases, obteniendo así, una fuente de información para desarrollar el caso de estudio.

En la base de traslados, se encontraron los clientes asignados al BPO para efectuar el proceso indicado, en dicha base, se tiene información transaccional por cliente, detallando las obligaciones, valores vencidos, productos, segmentos, días en mora, propensión de pago, entre otras características importantes para desarrollar las gestiones pertinentes. Siendo la propensión al pago realmente importante para el proceso de cobranza se encuentra que esta es se calcula teniendo en cuenta las llamadas o contacto que se debe tener con cliente antes de que este pague, las cuotas en mora y la cantidad de cuentas por pagar.

La base de gestiones contiene los registros de gestión efectuada por día a cada cliente, indicando si las gestiones realizadas fueron por medio de mensaje de texto, correo electrónico, agente virtual, mensaje de voz o por asesor, con estas métricas, se identifica si el contacto fue directo o indirecto. En la base de gestiones se tiene en cuenta la respuesta del cliente, que contiene en definitiva el resultado de la gestión, no obstante, no se tiene en cuenta los clientes que se autogestionan.

En el tanque de pagos, se observan los pagos realizados por los clientes detallado a que producto pertenece.

3.2. Integración de las bases de confianza

Al integrar la base de traslados, gestiones y tanque de pagos, se obtiene como resultado la base consolidada de información que permite conocer por cliente y obligación asignada al BPO en los seis meses de historia mencionados anteriormente. En la base consolidada se tiene la traza del número de gestiones realizadas y se tiene el detalle de: si estas gestiones se efectuaron por contacto directo o indirecto, así como, si la gestión fue efectiva convirtiéndose en un pago total o abono a la obligación, conservando por registro las métricas unificadas por obligación de la base de traslados que permitirán conocer el comportamiento de los clientes, con el fin de inferir si el registro cumple con el criterio de auto cura 3.1.

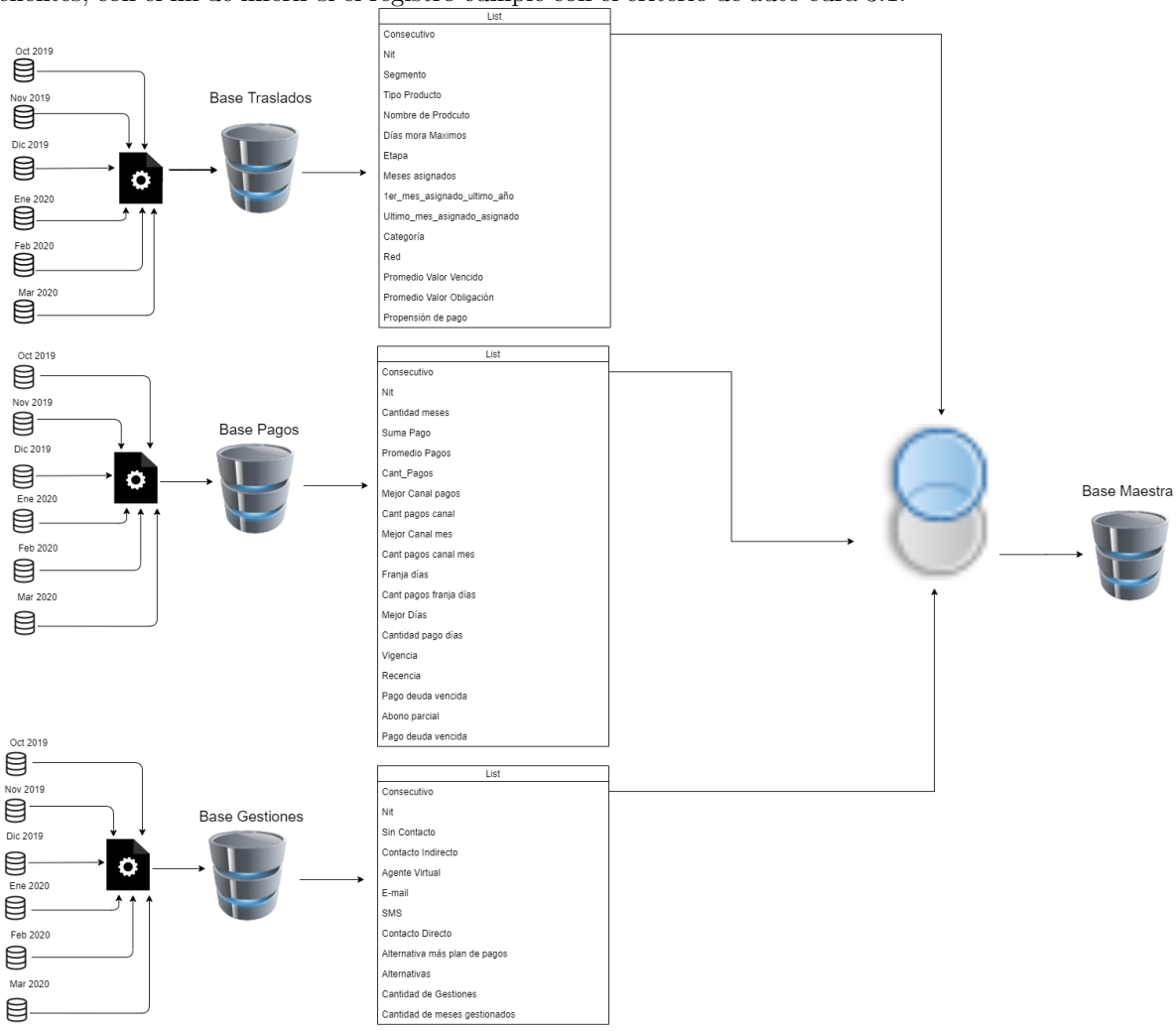

Figura 3.1: Base consolidada

3.3. Definición de variables

La base de datos consolidada (sin transformación de variables categóricas) posee una dimensión de 37 columnas por 157.696.

Variables	Definición
Número Consecutivo	Este código es único e identifica préstamo del cliente
Nit	Es el número de identificación del cliente
Segmento	Identifica la categoría a la cual pertenece el producto
Código de Producto	Es un indicativo único por producto
Producto	Es el tipo de préstamo que posee el cliente
Días de mora	Es el número de días que tiene el cliente sin cancelar la cuota
Etapa	Indica el estado de la cartera, es decir, si se pasó a cobro jurídico
Meses asignado	Es la cantidad de meses que el cliente fue asignado
Primer mes asignado	Primer mes que fue asignado al BPO
Ultimo mes asignado	Ultimo mes que fue asignado al BPO
Categoría	Indica si el clientes nuevo o antiguo
Red	Indica el tipo de negocio que pertece el cliente
Valor Vencido	Es el valor de la cuota pendiente de pago
Valor Obligación	Es el valor total adeudado por el cliente
Propensión de Pago	Es la probabilidad de pago que asigna el aliado
Suma de pagos	Total de pagos realizados
Promedio de pagos	Valor promedio de pagos realizados por cliente
Cantidad de pagos	Número de pagos efectuados
Mejor canal de pagos	Canal con mayor número de pagos
Franja en días	Rango en días que se realizaron los pagos
Moda franja en días	Franja con mayor número de pagos
Mejor día de pago	Días con mayor pago
Cantidad día	Cantidad de Días asignado al BPO
Primera aparición de pago	Primer pago realizado
Ultima aparición de pago	Ultimo pago realizado
Vigencia	Diferencia entre la ultima fecha de gestión y la fecha del reporte
Recencia	Diferencia entre la primera fecha de asignación y la fecha del reporte
Pago total	Pagos completos al valor vencido
Abono Parcial a Deuda	Pagos incompletos al valor vencido
Sin contacto	Número de gestiones que no se localizó al cliente
Contacto indirecto	Número de gestiones que se localizó al cliente sin asesor
Agente virtual	Número de gestiones con mensaje de voz
E-mail	Número de gestiones por correo electrónico
SMS	Número de gestiones por mensaje de texto
Contacto directo	Número de gestiones por medio de asesor
Alternativas	Número de alternativas ofrecidas
Cantidad de gestiones	Total de gestiones realizadas

Tabla 3.1: Defición de variables

Se aclara que por cuestión de simplificación de información se realiza el descarte de la información detallada de las gestiones, reemplazando esta información por el número de gestiones realizadas, dependiendo del tipo de contacto.

3.4. Análisis descriptivo

Para realizar el análisis exploratorio de la base de datos, se utiliza la estadística descriptiva, la cual brinda diferentes técnicas para la presentación y resumen de los registros de la base de datos (Fernández, Cordero, Córdoba, 2002). Se analiza las variables de interés individualmente, las cuales estan compuestas por: valor de obligación, valor vencido, propensión de pago, días en mora, recencia, gestiones, y pagos, ello acompañado de un estudio de las variables categóricas; se aclara que, las variables de interés fueron escogidas, considerando la experiencia en el proceso de cobranza. Esto se efectúa para encontrar registros con similitudes y adquirir conocimiento de los datos. Como parte del proceso descriptivo cada una de las variables se describe teniendo en cuenta la especificación de cuartiles calculados considerando la cantidad de datos encontrados dependiendo de las variables y la agrupación de los mismos.

3.4.1. Resumen valor obligación

La variable valor obligación describe el monto total que el cliente debe a la entidad financiera. Al realizar una descomposición de los registros con una tabla de frecuencia, se evidencia que el 86.9 % de los registros poseen un valor de obligación entre cero y diez millones, dado que este intervalo es amplio y embebe una cantidad importante de registros, se decide obtener los cuartiles, los cuales están comprendidos de la siguiente forma: valor mínimo, cuartil 1, mediana (cuartil 2), media y cuartil 3. Al revisar los cuartiles se identifica que el 25 % de los registros son menores o iguales a 569.588, por otra parte, se observa que el valor obligación posee una mediana de 1.779.312, un promedio de 6.238.041 y el 75 % de los registros están representados por el valor de 5.067.213.

Variable	min	Cuartil 1	Cuartil 2	media	Cuartil 3
Valor_Obligacion	0	569,588	1,779,312	6,238,041	5,067,213

Table: *Resumen Valor Obligacion*

Figura 3.2: Resumen Valor obligación

	Lower	Upper	Main	Frequency	Percentage	CF	CPF
1	0	10,000,000	5,000,000	740,016	86.9	740,016	86.9
2	10,000,000	20,000,000	15,000,000	56,402	6.6	796,418	93.5
3	20,000,000	30,000,000	25,000,000	20,280	2.4	816,698	95.9
4	30,000,000	40,000,000	35,000,000	10,642	1.2	827,340	97.1
5	40,000,000	50,000,000	45,000,000	7,754	0.9	835,094	98
6	50,000,000	60,000,000	55,000,000	4,943	0.6	840,037	98.6
7	60,000,000	70,000,000	65,000,000	3,154	0.4	843,191	99
8	70,000,000	80,000,000	75,000,000	2,117	0.2	845,308	99.2
9	80,000,000	90,000,000	85,000,000	1,664	0.2	846,972	99.4
10	90,000,000	100,000,000	95,000,000	1,273	0.1	848,245	99.6

Table: *Tabla de Frecuencia Valor Obligación*

Figura 3.3: Tabla de frecuencia Valor Obligación

3.4.2. Resumen valor vencido

El valor vencido informa el monto que el cliente no pago en la fecha correspondiente, es decir, cuando se ingresa en mora, este valor se afecta por los pagos o abonos que el cliente efectué a la entidad financiera. En la tabla de frecuencia se observa fácilmente que el 88.3 % de las obligaciones tienen un valor vencido $\leq$ 1,000,000, el 7.1 % de las obligaciones están entre el rango de 1,000,000 <Valor Vencido$\leq$ 2,000,000, el 2 % se encuentra entre 2,000,000 <Valor Vencido$\leq$ 3,000,000 y el 2.6 % de los datos poseen un valor vencido $>$ 3,000,000. Dado que el 88.3 % de las obligaciones se concentran en un solo intervalo, se decide explorar los registros con los cuartiles.

	Lower	Upper	Main	Frequency	Percentage	CF	CPF
1	0	1,000,000	500,000	750,631	88.3	750,631	88.3
2	1,000,000	2,000,000	1,500,000	60,551	7.1	811,182	95.4
3	2,000,000	3,000,000	2,500,000	17,326	2	828,508	97.4
4	3,000,000	4,000,000	3,500,000	7,863	0.9	836,371	98.3
5	4,000,000	5,000,000	4,500,000	4,372	0.5	840,743	98.9
6	5,000,000	6,000,000	5,500,000	3,076	0.4	843,819	99.2
7	6,000,000	7,000,000	6,500,000	1,738	0.2	845,557	99.4
8	7,000,000	8,000,000	7,500,000	1,268	0.1	846,825	99.6
9	8,000,000	9,000,000	8,500,000	905	0.1	847,730	99.7
10	9,000,000	10,000,000	9,500,000	768	0.1	848,498	99.8

Table: *Tabla de Frecuencia Valor Vencido*

Figura 3.4: Tabla de frecuencia Valor Vencido

En el análisis de cuartiles se identifica que el 25 % de las obligaciones poseen un valor vencido <= 82,644, la mediana de los datos es 197.898, con un promedio de 617.690 y el 75 % de los datos están representados por el valor de 500.082, como se observa en los cuartiles, hay valores que se salen de las medidas de tendencia central, los cuales son considerados datos atípicos.

Variable	min	Cuartil 1	Cuartil 2	media	Cuartil 3
Valor_Vencido	0	82,644	197,898	617,690	500,082

Table: *Resumen Valor Vencido*

Figura 3.5: Cuartil Valor Vencido

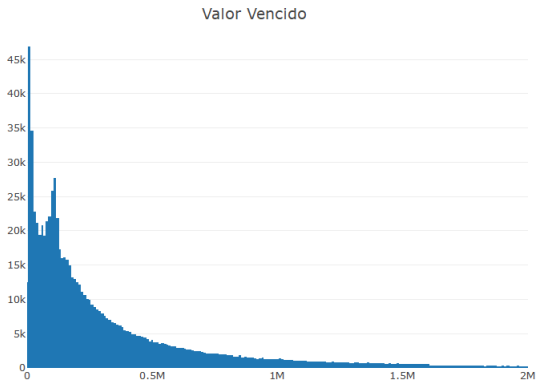

Figura 3.6: Histograma Valor Vencido

3.4.3. Resumen Propensión de pago

Este ítem contiene información de la probabilidad de pago de los clientes, esta probabilidad esta expresada en porcentaje, el cual fue calculado teniendo en cuenta la frecuencia de pago posterior al contacto con el cliente, y la totalidad de pagos realizados, como es de esperar entre mayor sea la probabilidad de pago, el cliente es más propenso en ejecutar el pago de sus obligaciones. En el histograma se observa que 67.6 % de los clientes asignados se concentran entre 80 % y 100 %, situación que se considera positiva para la empresa toda vez que, el sistema de cobranza es eficaz, y los clientes tienden a ponerse al día después de una llamada, mensaje o contacto con cualquier tipo de herramienta de comunicación.

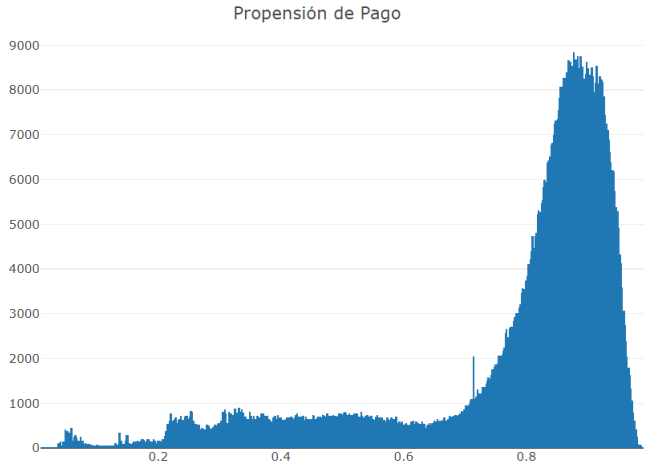

Figura 3.7: Histograma Propensión de pago

Para conocer la participación de datos de acuerdo a su propensión, se agrupan los registros en una tabla de frecuencia. Con la tabla de frecuencia se descubre que el 63.4 % de los clientes tienen una probabilidad de pago en el siguiente rango $80\,\% \leq P \leq 95\,\%$, adicional, la mediana de los clientes analizados es de 85.26 %.

	Lower	Upper	Main	Frequency	Percentage	CF	CPF
1	0.00%	5.00%	2.50%	1,137	0.1	1,137	0.1
2	5.00%	10.00%	7.50%	4,110	0.5	5,247	0.6
3	10.00%	15.00%	12.50%	2,094	0.2	7,341	0.9
4	15.00%	20.00%	17.50%	3,878	0.5	11,219	1.3
5	20.00%	25.00%	22.50%	13,054	1.6	24,273	2.9
6	25.00%	30.00%	27.50%	12,988	1.5	37,261	4.4
7	30.00%	35.00%	32.50%	18,643	2.2	55,904	6.6
8	35.00%	40.00%	37.50%	17,101	2	73,005	8.7
9	40.00%	45.00%	42.50%	16,757	2	89,762	10.7
10	45.00%	50.00%	47.50%	17,645	2.1	107,407	12.8
11	50.00%	55.00%	52.50%	18,252	2.2	125,659	14.9
12	55.00%	60.00%	57.50%	15,821	1.9	141,480	16.8
13	60.00%	65.00%	62.50%	13,282	1.6	154,762	18.4
14	65.00%	70.00%	67.50%	17,054	2	171,816	20.4
15	70.00%	75.00%	72.50%	33,366	4	205,182	24.4
16	75.00%	80.00%	77.50%	67,570	8	272,752	32.4
17	80.00%	85.00%	82.50%	138,227	16.4	410,979	48.9
18	85.00%	90.00%	87.50%	209,395	24.9	620,374	73.8
19	90.00%	95.00%	92.50%	185,709	22.1	806,083	95.8
20	95.00%	100.00%	97.50%	34,988	4.2	841,071	100

Table: *Tabla de Frecuencia Propensión de Pago*

Figura 3.8: Tabla propensión de pago

Variable	min	Cuartil 1	Cuartil 2	media	Cuartil 3
Propension_pago	0.91%	75.54%	85.26%	80.00%	90.25%

Table: *Resumen Propensión de Pago*

Figura 3.9: Cuartil propensión de pago

3.4.4. Resumen días en mora

Esta característica embebe los días en mora de cada cliente, es decir, cuantos días han transcurrido de la ficha limite que debía realizar el pago. Esta característica embebe los días en mora de cada cliente, es decir, cuantos días han transcurrido de la ficha limite que debia realizar el pago. Los registros de días en moras se particionaron en rangos, con el fin de conocer cómo se comporta esta variable en la base de datos, en el rango de 0 a 20 días en mora se aglomera el 56.7 % de los registros, entre 21 y 40 días se tiene el 13.4 %, entre 41 y 60 días posee el 5.5 % y mayor a 60 días se tiene el 24.4 % de la información.

Variable	min	Cuartil 1	Cuartil 2	media	Cuartil 3
Dias_en_mora	0	6	16	56.1	60

Table: *Resumen Días en Mora*

Figura 3.10: Resumen días en mora

	Lower	Upper	Main	Frequency	Percentage	CF	CPF
1	0	20	10	480,356	56.7	480,356	56.7
2	20	40	30	113,604	13.4	593,960	70.1
3	40	60	50	46,512	5.5	640,472	75.6
4	60	80	70	24,534	2.9	665,006	78.4
5	80	100	90	26,914	3.2	691,920	81.6
6	100	120	110	27,121	3.2	719,041	84.8
7	120	140	130	23,447	2.8	742,488	87.6
8	140	160	150	11,332	1.3	753,820	88.9
9	160	180	170	14,087	1.7	767,907	90.6
10	180	200	190	25,742	3	793,649	93.6

Table: *Tabla de Frecuencia Dias en Mora*

Figura 3.11: Tabla de frecuencia días en mora

3.4.5. Resumen Recencia

Esta variable mide el número de días que lleva asignado los clientes al BPO. Al revisar a profundidad los datos, se observa que el cuartil 1 los clientes están representados con una recencia de 33 días, adicional, se tiene una mediana de 71 y promedio de 80 días, en el cuartil 3 los clientes su valor es de 124 días. Esta variable no concentra gran cantidad de clientes en los rangos de frecuencia.

Variable	min	Cuartil 1	Cuartil 2	media	Cuartil 3
Recencia	0	33	71	80.1	124

Table: *Resumen Recencia*

Figura 3.12: Resumen Recencia

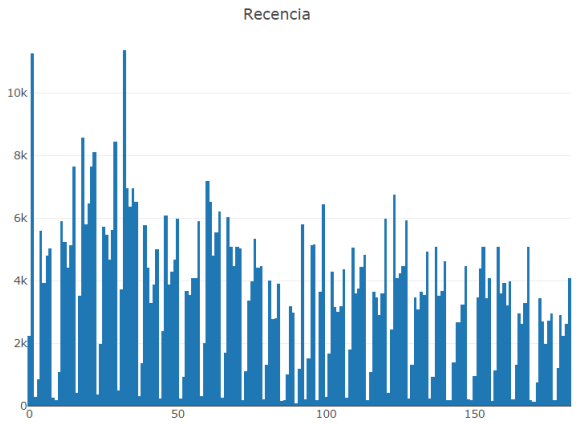

Figura 3.13: Histograma Recencia

	Lower	Upper	Main	Frequency	Percentage	CF	CPF
1	0	10	5	35,444	5.7	35,444	5.7
2	10	20	15	53,034	8.6	88,478	14.3
3	20	30	25	48,475	7.8	136,953	22.1
4	30	40	35	53,804	8.7	190,757	30.8
5	40	50	45	39,670	6.4	230,427	37.2
6	50	60	55	31,893	5.1	262,320	42.3
7	60	70	65	45,663	7.4	307,983	49.7
8	70	80	75	29,357	4.7	337,340	54.4
9	80	90	85	21,057	3.4	358,397	57.8
10	90	100	95	29,530	4.8	387,927	62.6

Table: *Tabla de Frecuencia Recencia*

Figura 3.14: Tabla de frecuencia Recencia

3.4.6. Resumen Gestiones

Esta métrica acumula el número de gestiones realizadas a cada cliente, se realiza la conglomeración de gestiones, debido a que cada gestión por separado acumulaba muy pocos datos poco significativos para el análisis. Al descomponer esta variable en cuartiles, se identifica que el cuartil 1 está representado por 5 gestiones, adicional, los datos de gestión tienen una mediana de 12 y un promedio de 23 gestiones, en el cuartil 3 cuenta con 30 gestiones acumuladas.

Variable	min	Cuartil 1	Cuartil 2	media	Cuartil 3
Cantidad_gestiones	1	5	12	23	30

Table: *Resumen Cantidad de Gestiones*

Figura 3.15: Resumen Gestiones

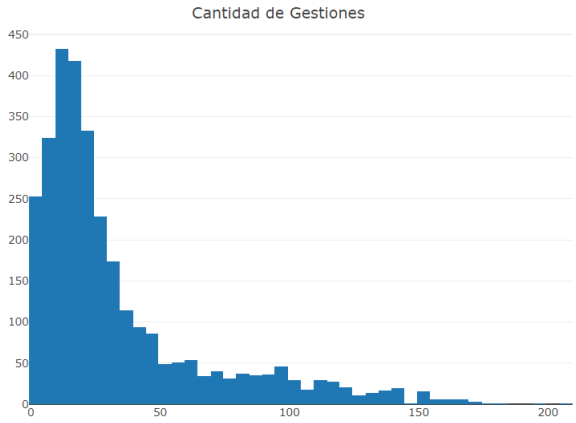

Figura 3.16: Histograma Gestiones

	Lower	Upper	Main	Frequency	Percentage	CF	CPF
1	0	5	3	591,028	95.4	591,028	95.4
2	5	10	8	23,068	3.7	614,096	99.1
3	10	15	13	3,555	0.6	617,651	99.7
4	15	20	18	1,042	0.2	618,693	99.9
5	20	25	23	410	0.1	619,103	99.9
6	25	30	28	178	0	619,281	100
7	30	35	33	112	0	619,393	100
8	35	40	38	49	0	619,442	100
9	40	45	43	34	0	619,476	100
10	45	50	48	24	0	619,500	100

Figura 3.17: Tabla de frecuencia Gestiones

3.4.7. Resumen Pagos Promedio

Son los pagos promedios que realizan los clientes a sus obligaciones, al realizar el análisis por cuartiles se observa que el cuartil 1 posee un valor de 66.665, adicional, esta variable tiene mediana de 178.500, un promedio de 515.918, y en el cuartil 3 está representado por un valor de pago promedio de 437.391. Se aclara que la distribución de pagos de acuerdo con la forma de pago y otras variables relacionadas con estos se presenta en el apartado siguiente.

Variable	min	Cuartil 1	Cuartil 2	media	Cuartil 3
Pago_promedio	0	66,665	178,500	515,918	437,391

Table: *Resumen Pago promedio*

Figura 3.18: Resumen Pagos

Figura 3.19: Histograma Pagos

	Lower	Upper	Main	Frequency	Percentage	CF	CPF
1	0	1,000,000	500,000	554,635	89.6	554,635	89.6
2	1,000,000	2,000,000	1,500,000	37,700	6.1	592,335	95.7
3	2,000,000	3,000,000	2,500,000	11,516	1.9	603,851	97.6
4	3,000,000	4,000,000	3,500,000	5,269	0.9	609,120	98.4
5	4,000,000	5,000,000	4,500,000	2,964	0.5	612,084	98.9
6	5,000,000	6,000,000	5,500,000	2,011	0.3	614,095	99.2
7	6,000,000	7,000,000	6,500,000	1,182	0.2	615,277	99.4
8	7,000,000	8,000,000	7,500,000	785	0.1	616,062	99.5
9	8,000,000	9,000,000	8,500,000	582	0.1	616,644	99.6
10	9,000,000	10,000,000	9,500,000	544	0.1	617,188	99.7

Table: *Tabla de Frecuencia Pago promedio*

Figura 3.20: Tabla de frecuencia Pagos

3.4.8. Resumen Variables categóricas

El medio de pago más utilizado es otros con el 25.36 % de los registros, seguido por la sucursal virtual con 11.94 % y debito con el 11.53 %. El canal con el pago promedio mayor es la sucursal física con 845.779, seguido del medio de pago otro con 694.691.

La franja de días donde se realizo el mayor numero de pagos se encuentra en el intervalo de (25 a 30] con una participación de 16.74 %, los clientes que realizaron los pagos en esta franja contienen una probabilidad de pago promedio de 82.87 % y el valor del pago promedio fue de 467.681, la segunda franja en participación se encuentra en el intervalo de (10 a 15], la propensión de pago promedio es de 82,14 % y el valor medio pagado estuvo en 522.603.

mejor_canal_pagos	Cantidad	Participacion	Prom_cantidad_pagos	Pago_promedio
Sin_informacion	232,770	27.31%		
Otro	216,131	25.36%	1.65	694,691
Sucursal Virtual	101,782	11.94%	1.42	485,303
Debito	98,307	11.53%	2.25	230,223
Debito Manual	87,180	10.23%	2.32	269,794
Sucursal Fisica	71,012	8.33%	1.47	845,779
Debito Mora	34,840	4.09%	1.49	288,912
Debito Automatico	9,135	1.07%	1.6	354,399
Sucursal Teléfonica	1,171	0.14%	1.15	499,685

Table: *Resumen Canal de Pago*

Figura 3.21: Resumen canal de pagos

franja_dias	Cantidad	Participacion	Prom_cantidad_pagos	Prom_propension	Pago_promedio
Sin_informacion	232,770	27.31%		64.78%	
(25 a 30]	142,666	16.74%	1.52	82.87%	467,681
(10 a 15]	104,010	12.20%	1.44	82.14%	522,603
(20 a 25]	100,371	11.78%	1.35	83.57%	533,244
(15 a 20]	98,834	11.60%	1.44	82.32%	540,994
[1 a 5]	87,161	10.23%	1.35	82.41%	470,288
(5 a 10]	86,516	10.15%	1.38	82.40%	584,650

Table: *Resumen Franja de Pago*

Figura 3.22: Resumen franjas

3.5. Selección de variables por consenso de target empírico

Dado que en la base de datos no se cuenta con una variable que indique si un cliente tiene el criterio de auto cura, se aprovecha la experiencia del proceso para calcular dicho criterio. El cálculo efectuado se elaboró sobre las siguientes columnas: días en mora, contacto directo, abono parcial a la deuda y pago total.

Con estas variables se realizó el siguiente código en R para hallar la variable dependiente que describiera a los clientes auto cura:

Begin
For i=1 to nrow(base) do
IF diasenmora<= 15 and contactodirecto= 0 and abonoparcial> 0 and pagototal > 0 then
Tartget=1
Else
Target=0
End if
Next
End

Con el target elegido, se procede a investigar 3 métodos de selección de variables, ya que se debe encontrar el modelo adecuado de datos para entrenar las técnicas de aprendizaje supervisado.

3.5.1. Lasso (Least Absolute Shrinkage and Selection Operator)

Con el fin de hallar un método de regresión lineal que por medio de la reducción de las betas (Coeficientes) permita realizar selección de variables. Thibshirani propuso la técnica de regresión lineal regularizada de Lasso, que a partir de algunos valores del parámetro de complejidad realiza valoraciones iguales a cero para algunas betas y diferentes de cero para otros coeficientes, esto se efectúa a través de la norma L1, la cual genera vectores de caracteristicas cortos concibiendo la distribución de la información (Tibshirani,2011).

La técnica de Lasso soluciona el problema de mínimos cuadrados con restricción en L1 en el vector de betas:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

De esta fórmula se tiene que λ se concibe como parámetro de penalización por

complejidad, y por consiguiente deben tener valores mayores o iguales a cero. Para aplicar Lasso, las variables predictoras deben ser estandarizadas.

La regularización con la norma L1 genera vectores de características cortos, dado que la mayoría de los pesos de las variables tienden a cero. Efectuar el proceso de reducción de características es útil si se cuenta con una base de datos de alta dimensión, con muchas variables que no aportan valor a la predicción. La regularización L1 está dada por la siguiente ecuación (Raschka, Mirjalili, 2017):

$$L_1 : ||w||_1 = \sum_{j=1}^{m} |w_j|$$

El método de Lasso presenta diferentes limitaciones, la primera de las cuales se presenta cuando las variables (V) son mayores al número de observaciones (n) (datos de alta dimensión con pocos ejemplos), esto genera que Lasso elija como máximo n variables antes de saturarse. Otra limitación a considerar es que, si se tiene un conjunto de variables con alta correlación, se selecciona una variable del conjunto e ignora las demás características (Tibshirani, 2011).

Lasso no comparte la selección de variables agrupadas, debido a lo cual ignora la ordenación de las variables, en el escenario donde se tenga mayor número de variables que observaciones, poder efectuar la selección de variables agrupada toma gran importancia. Al revisar el estudio de Segal, Dahlquist y Conklin (Segal, Dahlquist, Conklin,2003) se evidencia que la ocupación de métodos de regularización es necesario para encontrar agrupamientos de datos.

3.5.2. Regresión Ridge

La regresión Ridge, inicialmente fue propuesta por Hoerl y Kennard, con el fin de evitar la colinealidad en un modelo lineal evaluado por mínimos cuadrados. Para el funcionamiento de la técnica de Ridge, el número de variables debe ser menor a la cantidad de observaciones. Las betas calculadas por Ridge son los valores que minimizan la siguiente ecuación:

$$\beta^{ridge} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

Donde lambda es mayor o igual a cero es el parámetro de contracción. La regresión Ridge contrae los coeficientes al incluir la penalización en la función objetivo, entre mayor sea lambda, mayor será la penalización, por ende, mayor la reducción de las betas.

El método Ridge realiza la reducción de coeficientes cercano a cero, esto implica que ninguno de ellos es igual a cero, por tal razón, no se genera selección de variables conservando la totalidad de los predictores. Esta se considera la principal desventaja pues si bien la penalización aplicada fuerza a que los coeficientes tiendan a cero, estos nunca llegaran a ser exactamente cero, debido a que lambda tiende a infinito (Hoerl, Kennard, 1970). El método, aunque consigue minimizar la influencia de los predictores menos relacionados con la variable respuesta, estos continúan presentes en el conjunto de datos, aunque ello no suponga un problema para la precisión del modelo, más si para su interpretación.

3.5.3. Elastic Net

La técnica Elastic Net (Red elástica) fue propuesta por Zou y Hastie como un método de regularización o selección de variables, la cual combina los beneficios de Ridge (Norma L2) y Lasso (Norma L1). Este método posee dos parámetros denominados λ y α, siendo λ una constante fija no negativa.La penalización se mueve en el rango de $0 \leq \alpha \leq 1$, para $\alpha = 0$, se utiliza el método de Ridge, es decir, norma L2 y para $\alpha = 1$, se utiliza la técnica Lasso normal L1, con $\alpha > 0$ y $\alpha < 1$, se obtiene una combinación de las normas L1 y L2. Esta técnica es útil cuando el número de variables es mayor a la cantidad de observaciones (Zou, Hastie, 2005).

$$s.a = \lambda \sum_{j=1}^{k} (\alpha|\beta_j| + (1-\alpha)\beta^2)$$

Elastic Net mejora las limitaciones expuestas de los métodos de Ridge y Lasso, el cual es utilizado para efectuar selección de variables y presenta buen rendimiento cuando se tiene más variables predictoras que observaciones.

Elección de los criterios de penalización (λ y α)

Como se observa en las metodologías anteriormente analizadas, la penalización depende de un criterio λ, el cual regula el peso de la penalización en el proceso de selección de variables. Entre mayor sea el parámetro, más rigurosa será la penalización en los β de regresión, acercando su valor a cero. Con la elección del criterio λ se afecta el sesgo y la varianza, por esta razón, la literatura recomienda utilizar una traza de Ridge para establecer el λ, esta propuesta consiste en graficar las betas estimados en función de λ y elegir el valor menor de los coeficientes. Para elegir los criterios de penalización de la técnica Elastic Net (λ, α), se utilizó el método GridSearchCV, el cual estima los criterios con validación cruzada (Carrasco, 2016).

El método GridSearchCV permite seleccionar los valores de los hiperparametros para un modelo o conjunto de datos. Para ellos se ha de crear un objeto GridSearchCV donde el modelo es el primer parámetro y un diccionario de los parámetros es el segundo (Rajnar Verma, Radhika, 2019). Al aplicar el método GridSearchCV se obtuvo los siguientes resultados: λ= 0.050118 y α=0.05.

Como se observa, los resultandos combinan la norma L1 y L2, predominando la norma L2. Al ejecutar Elastic Net con estos criterios, se disminuyó el 23.1 % de las variables, es decir, de 52 regresores se seleccionó un conjunto con 40 características.

3.5.4. Modelo de datos seleccionados

Modelo seleccionado con 40 variables

Variables
dias mora maximos
prop
cantidad meses
moda franjas
cantidad gestiones
cantidad dia
vigencia
recencia
sin contacto
agente virtual
e mail
contacto directo
alternativas
segmento negocios e indep
segmento mi negocio
segmento pymes
nombre de producto credito hipotecario
nombre de producto tarjetas de credito
nombre de producto cuenta corriente
nombre de producto credito de consumo
nombre de producto reestructuracion
nombre de producto prestanomina
nombre de producto microcredito
nombre de producto adelanto de ingresos
nombre de producto credito a la mano
nombre de producto venta digital
etapa administrativa
mejor canal pagos sucursal fisica
mejor canal pagos debito mora
mejor canal pagos sucursal virtual
mejor canal pagos debito
mejor canal pagos otro
mejor canal pagos debito manual
mejor canal pagos debito automatico
mejor canal pagos sucursal tel e9 fonica
franja dias 1 a 5
franja dias 10 a 15
franja dias 15 a 20
franja dias 20 a 25
franja dias 5 a 10

Capítulo 4

Grouping Recall Least Lazy Algorithm: GRILLA

4.1. Construcción de modelos matemáticos

4.1.1. Metodología

En la base de datos se identifican variables numéricas y categóricas. Las características de tipo numéricos son las que miden una cantidad como un número, estas pueden ser catalogadas como continuas o discretas, las variables categóricas describen una cualidad o característica de una unidad de un tipo de dato, al pertenecer a una categoría se excluye de las demás (Marketing-analítico,2016).

Las variables categóricas se convierten a formato numéricas con la técnica dummies, la cual consiste en transponer los datos en "columnas", donde la variable transformada se convierte en múltiples columnas nuevas (tantas columnas como categorías posea), las columnas generadas contienen datos binarios (cero y uno), la aparición de uno indica que el dato pertenece a la categoría y cero que no pertenece (RPubs, 2019).

Para el desarrollo del aprendizaje no supervisado, se utilizó el algoritmo Kmeans, el cual es particional de características duras. Al utilizar la técnica no supervisada, se debe garantizar que las variables cumplan con un criterio de independencia, es decir, no se encuentren altamente correlacionadas de forma positiva o negativa, para analizar el grado de correlación de las características, se usa el método de Pearson.

El coeficiente de Pearson puede tomar un valor entre -1 a 1, cuando el coeficiente es cercano a cero, se interpreta que las variables son independientes, si el coeficiente es cercano a 1, se entiende que las características poseen dependencia directa y si el valor es cercano a -1 las variables tienen dependencia

inversa, al aplicar esta técnica se excluyen las características cuyos coeficientes cumplan la siguiente condición: –0,75 <= P >= 0, 75, incluyendo con esto tres cuartiles que han sido analizados en la descripción de datos.

$$P_{x,y} = \sigma_x y \sigma_x \sigma_y = E[(X - \mu_x)(Y - \mu_y)]\sigma_x \sigma_y$$

Al eliminar las variables con alta correlación, se procede a estandarizar los datos. Dado que el método no supervisado K-means se le debe ingresar el número de clúster por el cual se quiere agrupar los registros, se utiliza el diagrama de codos, con las métricas de separación y cohesión, con el fin de hallar la cantidad de clúster necesarios para ejecutar el algoritmo de K-meas, se aclara que el método del codo es una técnica heurística que se ocupa para determinar el número de conglomerados en un conjunto de datos (Yan, 2005); luego se aplica la técnica de TSNE que es un algoritmo de reducción de dimensionalidad no lineal, ya que la base de datos posee una alta dimensión. Esta técnica transforma los datos multidimensionales a dos o más dimensiones, que permitan observar las diferencias de los registros (Olivon, Elie, Grelier, Roussi, Litaudon, Touboul y MetGem, 2018). La dimensión intrínseca se calcula por medio de la máxima verosimilitud, esto para establecer el número de variables en el espacio de baja dimensión.

El cálculo de TSNE se fundamenta en la transformación de la distancia euclidiana de alta dimensión entre los registros de probabilidad condicional:

$$\rho_i|_j = exp(-||x_i - x_j||^2 2\sigma_i^2) \sum_{k \neq i} exp(-||x_i - x_k||^2 2\sigma_i^2)$$

Se tiene que σ es la varianza del vector en la función gaussiana centrada en el punto x_i.

TSNE utiliza SNE simétrico, que sustituye la probabilidad condicional con probabilidad conjunta entre los datos en el espacio de alta dimensión, la distribución de probabilidad de Gauss se usa en este mismo espacio, la distribución t con un grado de libertad se usa en bajas dimensiones. Esto está representado con las siguientes formulas:

$$\rho_{ij} = \frac{\rho_i|j + \rho_j|i}{2n}$$

$$q_{ij} = (1 + ||y_i - y_j||^2)^{-1} \sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}$$

Al obtener el resultado del aprendizaje no supervisado, se revisan los grupos resultantes por medio de las variables de días en mora, propensión de pago, valor obligación, valor vencido, cantidad de gestiones, recencia, vigencia, entre otras características, con el objetivo de encontrar los grupos o el grupo que describa a los clientes auto cura, y así, asignar la variable dependiente para ejecutar aprendizaje de máquinas supervisado.

4.1.2. Técnica de aprendizaje No Supervisado

Al utilizar las técnicas no supervisadas, se busca obtener información de los datos tales como: localización de anomalías, identificación de patrones, entre otros, con el fin de obtener diferentes grupos (Anil, 2010).

En el campo de clústeres se cuentan con dos tipos de agrupación, particionales y jerárquicos (Anil, 2010). El algoritmo jerárquico funciona hallando recursivamente los grupos anidados, ya sea de forma aglomerativo o divisivo. Los clústeres que funcionan de forma particional hallan los grupos en paralelo con una partición de los registros y no asignan un arreglo jerárquico, la complejidad algorítmica de la mayoría de los clústeres con estructura jerárquica es cuadrática o mayor en el número de puntos de datos (Anil, 2010), por ende, estos algoritmos no son recomendados para grandes volúmenes de datos, por otra parte, los algoritmos particionales poseen una complejidad menor de orden cuadrático.

Técnica No Supervisada K-means

En el aprendizaje no supervisado, el algoritmo de $Kmeans$ es uno de los más utilizados en agrupamiento particional (Anil, 2010). La utilización de $Kmeans$ se atribuye a tres aspectos, los cuales son: fácil de implementar, las métricas de inicialización, distancia y criterios de finalización se pueden modificar, por último, la complejidad algorítmica es de tiempo lineal en N, D y K, en general $D < N$ y $K < N$ (Bradley, Fayyad, 1998) donde N y D se definen como tiempos polinomiales.

Al revisar el algoritmo de $Kmeans$, se observan diferentes desventajas, tales como: esta técnica detecta clústeres compactos e hiperelípticos o sensible, esto se puede solucionar variando la métrica de distancia (Mao, Anil, 1996) al utilizar la distancia euclidiana, el algoritmo es sensible al ruido y a los datos atípicos, ya que estos registros pueden influir significativamente en las medias de sus respectivos puntos, este inconveniente se soluciona con la eliminación de los datos atípicos (Zhang, Leung, 2003).

Dado un conjunto inicial de K centroides, el algoritmo itera entre los dos siguientes pasos:

Se asigna cada observación al cluster con media más cercana, donde cada Xp va exactamente dentro de S_i^t

$$S_i^t = x_p : ||x_p - m_i^t|| \leq ||x_p - m_j^t|| \forall 1 \leq j \leq k$$

El siguiente paso se actualiza el cálculo de los nuevos centriodes:

$$m_i^{(t+1)} = 1|S_i^t| \sum_{x_j \in S_i^t} X_j$$

El ciclo termina cuando la asignación no genera cambios.

Métrica de Cohesión

La cohesión busca que los puntos pertenecientes a un grupo sea lo más cercano entre ellos, es decir, que no se encuentren a distancias lejanas.

$$SSE = \sum_{i=1}^{k} \sum_{X \in C_i} dist^2(m_i, x)$$

Siendo K el numero de grupos, x un dato del cluster Ci y mi el centroide del cluster Ci.

Métrica de Separación

La separación entre los grupos debe de ser amplia, ya que los puntos que contiene cada clúster deben de estar a mayor distancia a los demás clústeres.

$$SSB = \sum_{j=1}^{k} n_j dist^2(c_j - \bar{x})$$

k es el número de clústers, n_j es el número de elementos en el clúster j, c_j el centroide del clúster j y $\bar{x}$ es la media del data set.

4.1.3. Técnica de aprendizaje Supervisado

Con la variable objetivo definida, se procede a dividir los registros en tres conjuntos, los cuales son Train con el 60 % de los registros, validación y test con el 20 % cada uno, dicha partición se realiza aleatoriamente, garantizando que la proporción de la variable objetivo se mantenga en las tres bases resultantes, posterior a la partición, se procede a estandarizar las variables numéricas predictoras de cada una de las bases resultantes, ya que estas poseen diferentes dimensiones y no son comparables entre sí. El proceso de estandarización consiste en calcular la media de la variable y la desviación estándar, luego a cada dato de la característica se le resta la media y se divide por la desviación estándar, esto se efectúa en cada una de las variables regresoras (Alea, Jiménez, Muñoz, Torrelles, 2014).

Con el conjunto de datos estandarizado se procede a utilizar el método de regularización de variables Elastic net, ya que se requiere descartar las características que no son importantes para la predicción del target. Para elegir la técnica de Elastic Net, se revisó la teoría de los métodos de Ridge y Lasso.

Regresión Logística

Es una técnica de aprendizaje automático de máquinas, la cual se utiliza para resolver problemas de clasificación, donde la variable dependiente está representada por un número binario (0,1). La regresión logística calcula la relación entre la variable objetivos y el conjunto de características independientes por medio de la función sigmoide, la cual determina la probabilidad del target. Para interpretar la probabilidad obtenida se utiliza un valor umbral, dicho umbral asignara el valor de la predicción (Hastie, Tibshirani, y Friedman, 2009).

La función sigmoide esta dada por la siguiente ecuación:

$$\hat{y} = \sigma(Xw) = \frac{1}{1 + e^{-wX}}$$

La regresión logística parte de una regresión lineal de ecuación $\hat{y} = \beta_0 + \beta_1 X$ transformada con un logit de ecuación $\hat{y} = Log\ (\frac{P}{1-P})$. La regresión logística combina la regresión lineal y el logit igualando sus ecuaciones:

$$\hat{y}, queda\ Log\ (\frac{P}{1-P}) = \beta_0 + \beta_1 X$$

$$e^{Log\ (\frac{P}{1-P})} = e^{\beta_0 + \beta_1 X}$$

$$\frac{P}{1-P} = e^{\beta_0 + \beta_1 X}$$

$$P = (1-P)\ e^{\beta_0 + \beta_1 X}$$

$$P = e^{\beta_0+\beta_1 X} - Pe^{\beta_0+\beta_1 X}$$
$$P + Pe^{\beta_0+\beta_1 X} = e^{\beta_0+\beta_1 X}$$
$$P(1 + e^{\beta_0+\beta_1 X}) = e^{\beta_0+\beta_1 X}$$
$$P = \frac{e^{\beta_0+\beta_1 X}}{(1 + e^{\beta_0+\beta_1 X})}$$
$$P = \frac{1}{e^{-(\beta_0+\beta_1 X)}(1 + e^{\beta_0+\beta_1 X})}$$
$$P = \frac{1}{e^{-\beta_0-\beta_1 X}}$$
$$Y \text{ haciendo } \beta_0 + \beta_1 = w, \text{ entonces :}$$
$$P = \frac{1}{1 + e^{-wX}}$$

El valor umbral se mueve en el rango de $0 \leq Probabilidad \leq 1$

$$f(x) = \begin{cases} \text{si } x > 0{,}5 & 1 \\ \text{si } x \leq 0{,}5 & 0 \end{cases}$$

Para estimar la función de costo w, se utiliza la siguiente ecuación:

$$w = \sum_{i=1}^{N} (y_i \ln(\hat{y}) + (1 - y_1) \ln(1 - \hat{y}))$$

Árbol de decisión

La técnica de árboles de clasificación consiste en efectuar preguntas sistemáticas en cada paso, la pregunta que se realiza depende de la respuesta del paso anterior, al final de la secuencia se obtiene la predicción.

El árbol de decisión inicia la secuencia con el nodo Raíz, el cual embebe todo el conjunto de datos. A partir del nodo raíz, se ejecutan diferentes particiones en nodos intermedios, los cuales poseen un subconjunto de los datos, a estos nodos se pueden clasificar como nodo padre o no terminales, los nodos padres se dividen con una condición binaria o booleana, de estas particiones se derivan los nodos hijos o nodo terminal. De esta secuencia se obtiene la profundidad o número de ramas a partir del nodo Raíz (Izenman, 2008).

Los árboles de decisión pueden adoptar dos parámetros de medición, los cuales son: la perdidas de información (índice Gini) y la ganancia de información (Entropía).

El índice Gini mide la impureza, es decir, que tan a menudo un criterio elegido aleatoriamente del conjunto de datos, su predicción se realizó incorrectamente. El índice Gini está representado con la siguiente ecuación:

$$G = \sum_{k=1}^{c} p_k(1 - p_k)$$

Para calcular la impureza Gini en un conjunto, supóngase que k toma valores desde 1 hasta c. p_k es la probabilidad de un elemento de tener una clasificación determinada.

La entropía es la cantidad promedio de información que contiene una característica, las variables con menor índice de entropía son las que generan mayor información. La entropía está representada con la siguiente ecuación:

$$E = -\sum_{k} p_k \log(p_k)$$

Para calcular la entropía en un conjunto, supóngase que k toma valores desde 1 hasta c. p_k es la probabilidad de un elemento de tener una clasificación determinada.

Random Forest

Random Forest o bosque aleatorio, es una técnica de aprendizaje de máquinas, que consiste en combinar diversos árboles de decisión con la metodología de bagging o también conocida como "booststrap aggregationg", con el fin de reducir el sesgo y la varianza en las predicciones, adicional, evitar el sobre ajuste (overfitting).

El funcionamiento de random forest consiste en obtener N conjunto de datos aleatorios con reemplazo de variables en la base de entrenamiento, con estas muestras se entrenan los n árboles, al final cada árbol obtiene el valor predicho, para elegir la predicción final en problemas de clasificación se busca la clase que haya generado el mayor número de votos, es decir, la moda, con problemas de regresión, la predicción se estima con la media aritmética de las predicciones del grupo de árboles (Breiman, Friedman, Stone, y Olshen, 1984).

El algoritmo de random forest consiste en sacar el conjunto de árboles $\{T_b\}_1^B$,

donde B es el número de variables aleatorias, b es una variable en particular y T_b el árbol en el bosque con B árboles, donde b toma valores desde 1 hasta B. Para hacer una predicción al punto nuevo x, se calcula con: $\hat{f}_{RF}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$ para regresión, y $\hat{C}b(x)$ la clase predictora del b-ésimo árbol del random forest para clasificación (Friedman, Hastie y Tibshirani, 2001).

Antes de cada partición, se selecciona $n \geq p$ de las variables de entrada al azar como candidatos de división. Donde m es un número aleatorio de variables seleccionado de las p variables. Después de B árboles $\{T(x;\theta_b)\ \}_1^B$ crecen, el predictor de RF es $\hat{f}_{RF}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\theta_b)$ para regresión. Donde θ se explica a continuación.

Para el b-ésimo árbol se hace una generación de un vector aleatorio θ_b, con la misma distribución que todos los $\theta_1, ..., \theta_b$, pero independientes entre sí. Con este conjunto de entrenamiento y con el vector aleatorio, un árbol se desarrolla y resulta en una estimación, donde $T(x,\ \theta_b)$ es un vector de entrada, calculando así todos los árboles de decisión para la construcción del Random Forest (Merino y Chacón,2017).

Máquinas de soporte vectorial

Las máquinas de soporte vectorial, es un sistema de aprendizaje que en los últimos años ha tenido un desarrollo significativo. Este método apoya el aprendizaje en el uso de espacio de hipótesis de funciones lineales en un espacio de mayor dimensión inducido por un kernel. Este método es eficiente para problemas de regresión y clasificación. Esta técnica ha sido utilizada para clasificar imágenes, detección de proteínas, clasificación de patrones, entre otros (Resendiz, 2006).

La limitación computacional de las máquinas de soporte vectorial consiste en los R vectores de soporte libres y de nS, donde n es el número de muestras de entrenamiento y S es la cantidad de vectores de soporte. La complejidad del espacio depende de cuantas muestras de entrenamiento se almacenan en cada iteración. La complejidad de las máquinas de soporte vectorial depende de R^3 y de nS, por lo que el costo computacional de las máquinas de soporte vectorial tiene un componente cuadrático y uno cúbico. Crece parecido a n^2 cuando C es pequeño y crece como n^3 cuando C es grande (Bottou y Lin, 2007). Una alternativa para el problema de complejidad es utilizar la función del kernel, la cual es una alternativa para superar el problema mencionado, proyectando los datos de entrada a un espacio de mayor dimensión, esto aumenta la capacidad computacional del método.

$$F = \{\Phi(x)|x \in X\}$$

$$x = \{x_1, x_2, x_3, ..., x_n\} \longrightarrow \Phi(x) = \{\Phi(x)_1, \Phi(x)_2, \Phi(x)_3, ..., \Phi(x)_n\}$$

Las maquinas de soporte vectorial esta dada por la siguiente función:

$$f(x) = \langle w.x \rangle + b$$
$$f(x) = wx^T + b$$
$$f(x) = \sum_{i=1}^{n} w_i x_i + b$$

Knn (K vecinos más cercanos)

El funcionamiento del algoritmo de Knn se basa en clasificar cada registro nuevo en el grupo adecuado, esto se efectúa según tenga k vecinos más cercanos de un grupo o de otro. Para realizar esta asignación, el algoritmo Knn utiliza la función de distancia que permita medir la similaridad entre los puntos. Se tiene diversas formas de calcular la distancia, sin embargo, de forma tradicional, el algoritmo usa la distancia euclídea. Seleccionar la cantidad de vecinos en Knn, es fundamental, ya que esto determinara la generalización del modelo en los datos nuevos.

Elegir un valor de K alto reduce la injerencia por la variación generada por el ruido de los datos, sin embargo, esto genera sesgo; al elegir un K bajo, el modelo no tendrá capacidad de generalización.

Este algoritmo es simple de implementar, efectivo y su entrenamiento es rápido, sin embargo, la fase de clasificación es lenta, requiere de una capacidad alta en memoria y no produce un modelo, sin embargo, la complejidad sigue siendo lineal, es decir, cada registro que ingrese será comparado contra todo la base de entrenamiento, con el fin de determinar la predicción (RPubs,2020).

4.2. Selección de Target con aprendizaje no supervisado

4.2.1. Definición número óptimo de clústeres

Dado que el algoritmo de K-Means se le debe proporcionar el número de Clústeres con los que se desean separar los datos, se utiliza el diagrama de codos con las métricas de cohesión y separación, con el fin de inferir la cantidad de clústeres óptimos para realizar las agrupaciones.

Al revisar los diagramas de codos y seperación 4.1, se identifica que los grupos óptimos se encuentran entre 4 y 6, por ende, se procede a generar agrupaciones con este rango, posterior evaluar las métricas de separación, cohesión y la descripción de los grupos generados.

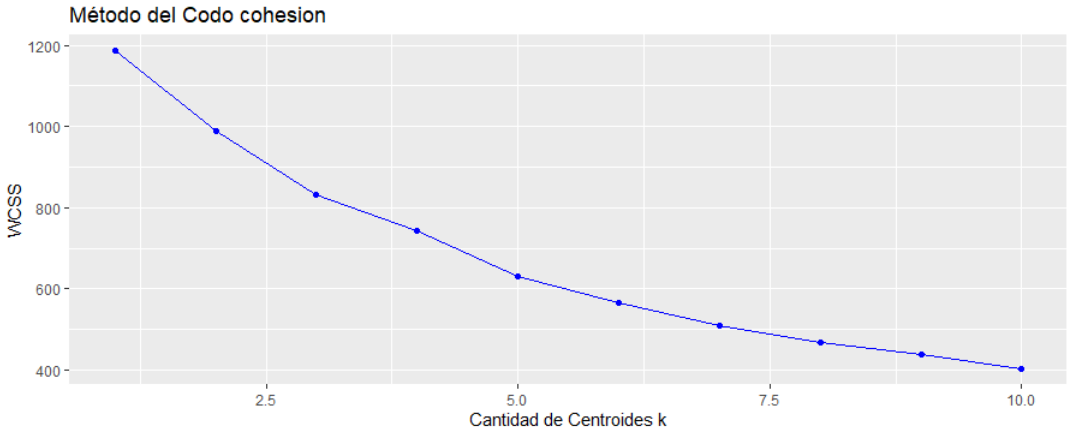

Figura 4.1: Métrica Cohesión

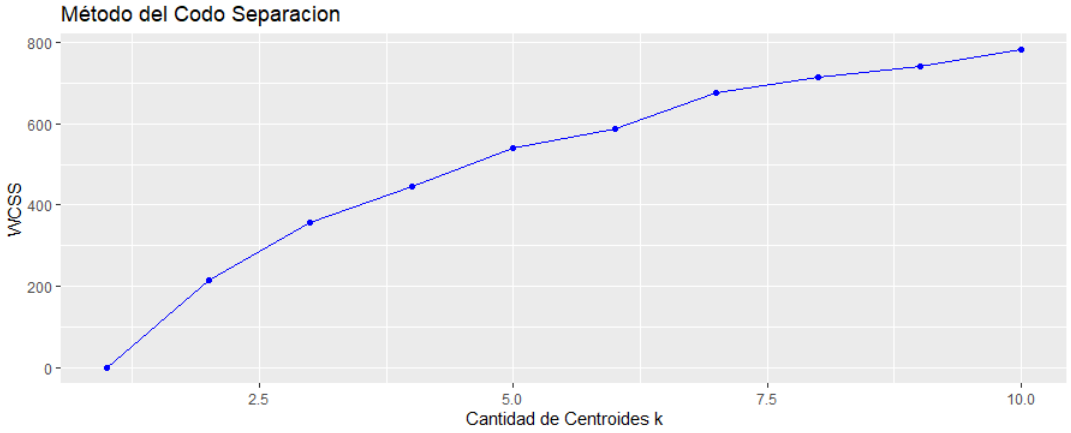

Figura 4.2: Métrica Separación

4.3. Generación de Laboratorio K-Means

Se generaron 3 laboratorios con la técnica de aprendizaje de máquinas no supervisada. Para utilizar el método no supervisado $K-means$, se utilizó disminución de dimensionalidad con $T-SNE$, con dimensión de 3 variables, una tasa de aprendizaje (etapa) de 200 y se itero 300 veces.

Laboratorio con 4 grupos

Al generar la agrupación con 4 clúster, se obtiene las siguientes métricas:
Separación: 983.757

Grupos	Cohesión	Tamaño	% Part
1	164,712	35,113	22%
2	217,244	37,810	24%
3	192,909	38,832	25%
4	269,131	45,941	29%

Figura 4.3: Métrica Cohesión 4 grupos

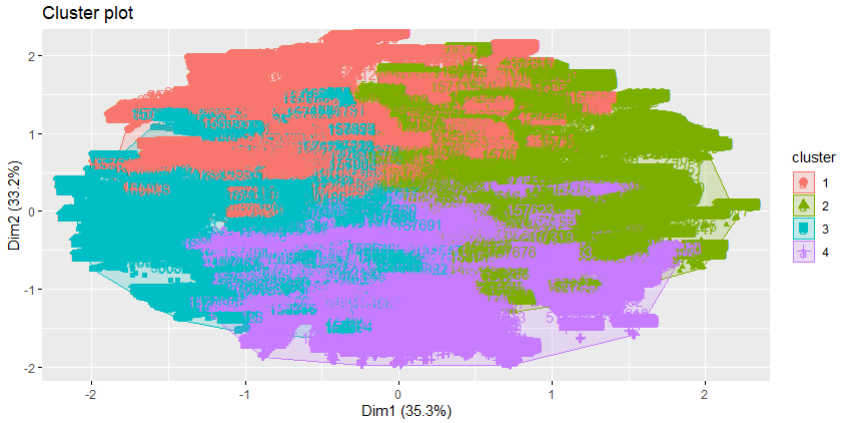

Figura 4.4: Kmeans 4 grupos

Al visualizar los grupos por las variables de interés medidas en sus promedios, se identifica que entre el grupo 2 y 3 sus características en días en mora, propensión de pago, valor vencido y gestiones, son similares, por ende, la generación del K-means con 4 grupos no logra separar los datos para identificar el criterio auto cura, se continua con la experimentación.

Grupos	Prom_Diasmora	Prom_Probpago	Prom_Valorobligacion	Prom_Valorvencido	Prom_Gestiones	Prom_Contacto	Prom_Vigencia	Prom_Recencia	Cantidad	%Part
1	20	80.6%	3,497,162	271,474	11.61	1.60	0.50	87	35,113	22%
2	17	84.8%	9,382,263	303,099	12.58	1.68	1.40	87	37,810	24%
3	17	83.1%	4,934,509	297,045	11.89	1.67	0.80	97	38,832	25%
4	23	78.3%	4,530,092	611,497	12.67	1.76	1.00	91	45,941	29%

Figura 4.5: Descripción 4 grupos

Laboratorio con 5 grupos

Al realizar la agrupación con 5 clúster, se obtiene las siguientes métricas:
Separación: 1.119.740

Grupos	Cohesión	Tamaño	% Part
1	108,736	27,221	17%
2	142,184	31,474	20%
3	137,233	31,991	20%
4	123,123	28,924	18%
5	196,737	38,086	24%

Figura 4.6: Métrica Cohesión 5 grupos

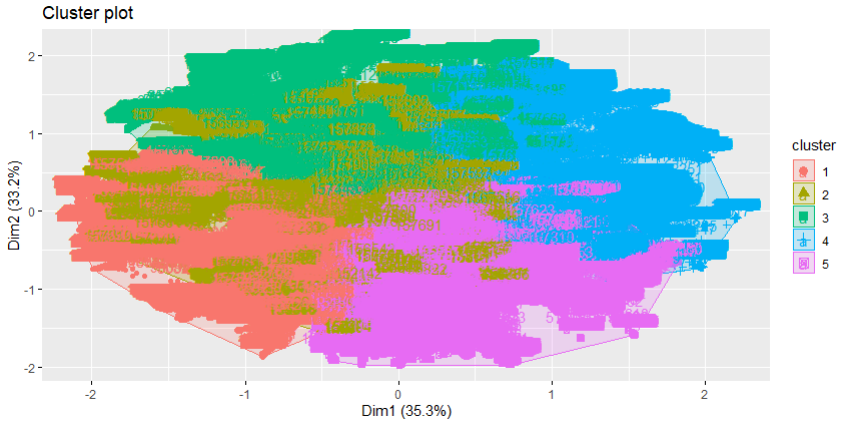

Figura 4.7: Kmeans 5 grupos

Al representar los grupos por las variables de interés medidas con sus promedios, se identifica que la agrupación con 5 clústeres, los grupos generados poseen diferencias entre sus características importantes, específicamente las siguientes variables: días en mora, valor obligación, valor vencido, gestiones y vigencia, adicional, la métrica cohesión, indica que los datos de este grupo están más cercanos que la de los demás grupos.

Grupos	Prom_Diasmora	Prom_Probpago	Prom_Valorobligacion	Prom_Valorvencido	Prom_Gestiones	Prom_Contacto	Prom_Vigencia	Prom_Recencia	Cantidad	%Part
1	15	81.8%	3,956,525	336,169	10.69	1.51	0.69	94	27,221	17%
2	17	83.6%	5,247,018	294,241	12.39	1.70	0.84	92	31,474	20%
3	20	80.5%	3,464,919	254,908	11.60	1.59	0.49	88	31,991	20%
4	17	84.9%	11,061,638	309,237	12.17	1.67	1.63	89	28,924	18%
5	26	78.1%	4,559,017	659,275	13.74	1.88	1.03	92	38,086	24%

Figura 4.8: Descripción 5 grupos

Laboratorio con 6 grupos

Al efectuar la agrupación con 6 clúster, se obtiene las siguientes métricas:
Separación: 1.226.230

Grupos	Cohesión	Tamaño	% Part
1	88,068	24,203	15%
2	96,007	25,282	16%
3	115,561	27,626	18%
4	104,253	27,985	18%
5	75,210	22,196	14%
6	122,423	30,404	19%

Figura 4.9: Métrica Cohesión 6 grupos

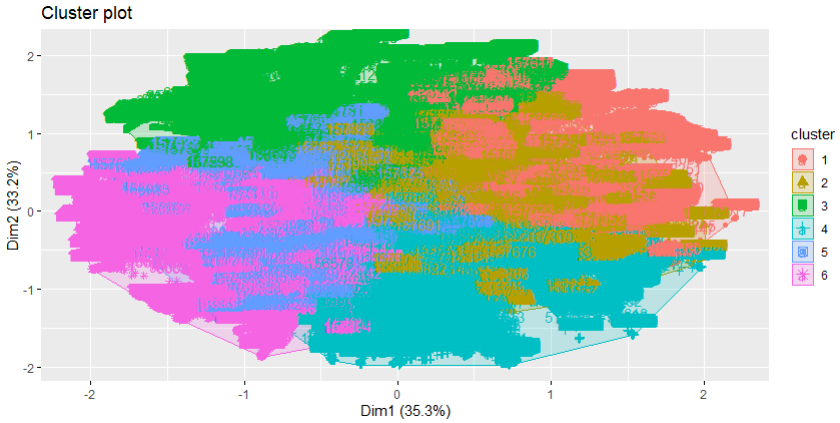

Figura 4.10: Kmeans 6 grupos

Grupos	Prom_Diasmora	Prom_Probpago	Prom_Valorobligacion	Prom_Valorvencido	Prom_Gestiones	Prom_Contacto	Prom_Vigencia	Prom_Recencia	Cantidad	%Part
1	18	81.1%	11,829,097	335,301	13.26	1.81	1.82	89	24,203	15%
2	16	84.1%	4,904,254	258,234	12.47	1.68	0.77	87	25,282	16%
3	21	83.3%	4,084,993	263,148	11.48	1.49	0.55	90	27,626	18%
4	29	76.5%	5,355,162	796,202	14.10	1.96	1.11	92	27,985	18%
5	14	81.6%	3,217,838	346,565	10.41	1.56	0.74	87	22,196	14%
6	17	82.8%	4,369,316	287,209	11.45	1.60	0.70	99	30,404	19%

Figura 4.11: Descripción 6 grupos

4.3.1. Resultados del Laboratorio

En la realización del laboratorio se inicializo la técnica no supervisada de K-Means, con 4, 5 y 6 grupos, con el fin de encontrar el grupo que describa las características de los clientes auto cura. Se identifica que K-Means con 5 clústeres, realiza la separación entre los grupos de forma óptima, adicional, los clientes que se encuentran en el clúster 1 poseen los criterios de un cliente auto cura, por ende, se procede a etiquetar los clientes del grupo 1 con 1 y los de más clientes se les asigna 0, de esta forma se halla la variable dependiente o target para proceder con el aprendizaje de máquinas supervisado.

4.4. Resultados del aprendizaje de máquinas supervisadas

Con las variables seleccionado con Elastic Net, se utilizó las bases de datos estandarizadas de Train con el 60 % de los registros, Validation con el 20 % y Test con el 20 % restante. Con la base Train se entrenaron los siguientes modelos supervisados: Regresión Logística, Arboles de decisión, Random Forest, Máquinas de soporte vectorial y K vecinos más cercanos (Knn), luego de obtener los resultados de entrenamientos de cada modelo, su capacidad de generalización se coloco aprueba con las bases de datos Validation y Test.

4.4.1. Métricas de medición de precisión, Recall, Accuracy, ROC y F1 Score

En primera instancia, debe aclararse que para las investigaciones es indispensable al momento de confirmar o rechazar una hipótesis, medir las magnitudes de los objetos o fenómenos que intervienen en sus estudios con la mayor fiabilidad posible. Por estos motivos, medir haciendo uso únicamente de los sentidos es un proceso poco fiable, carente de fundamentos, por ello nace la necesidad de hacer uso de instrumentos que faciliten esta tarea, dichos instrumentos de precisión reciben el nombre de instrumentos de medida. (Cohen, 1988)

En cuanto a las métricas de evaluación para valorar el rendimiento de un modelo, se tiene que esto es un componente integral de cualquier proyecto de ciencia de los datos y tiene como objetivo la estimación de la precisión de la generalización de un modelo sobre los datos futuros (no vistos/fuera de muestra).

Ahora bien, en el campo de la inteligencia artificial y el aprendizaje automático una matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado enfocado a un caso de estudio de clasificación. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, o sea en términos prácticos permite ver qué tipos de aciertos y errores tiene un modelo a la hora de pasar por el proceso de aprendizaje con los datos (Barrios,

2019).

Dentro una matriz de confusión existen varios términos como los Verdaderos Positivos (TP) que son los casos en los que los datos reales son 1 (Verdadero) y la predicción también es 1 (Verdadero); Verdaderos Negativos (TN) son los casos en los que los datos reales con 0 (Falso) y el pronóstico también es 0 (Falso); Falsos Positivos (FP) corresponden a los casos en que los datos reales indican que es 0 (Falso) y la predicción indica que es 1 (Verdadero), es decir la predicción ha sido errónea. La palabra Falso es porque el modelo ha pronosticado incorrectamente y positivo porque la predicción ha sido positiva (1) y finalmente Falsos Negativos (FN) son los casos en que los datos reales indica que es 1 (Verdadero) y el pronóstico es 0 (Falso), ocasionando que la predicción ha sido incorrecta. La palabra Falso es porque el modelo ha predicho incorrectamente y negativo porque predijo que era negativa (0). (Sierra, 2020)

De acuerdo con (Santos, 2018) la matriz de confusión y sus métricas asociadas son parte fundamental de la caja de herramientas del científico de datos, puesto que permiten saber qué modelo funciona mejor para un determinado problema. Estas métricas son por una parte la exactitud y precisión y por otra sensibilidad y especificidad.

La Exactitud o Accuracy, Según (Santos, 2018) se representa por la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones, y se calcula mediante la ecuación:

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

Por otro lado, la Precisión para (Barrios, 2019) se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción entre el número de predicciones de verdaderos positivos y las demás predicciones catalogadas como positivas. La ecuación se representa de la siguiente forma:

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

En cuanto se refiere a la sensibilidad y la especificidad son dos valores que nos indican la capacidad del estimador para discriminar los casos positivos, de los negativos. La sensibilidad es la fracción de verdaderos positivos, mientras

que la especificidad, es la fracción de verdaderos negativos. (Santos, 2018).

La Sensibilidad o "Recall" también se conoce como Tasa de Verdaderos Positivos (True Positive Rate) o TP, es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo. Se calcula:

$$Recall = \frac{VP}{VP + FN}$$

La Especificidad también conocida como la Tasa de Verdaderos Negativos, ("true negative rate") o TN. Se trata de los casos negativos que el algoritmo ha clasificado correctamente. Expresa cuan bien puede el modelo detectar esa clase. Se calcula:

$$Especificity = \frac{VN}{VN + FP}$$

La precisión y la sensibilidad indican la relevancia de los resultados. Por ejemplo, un algoritmo muy exacto, (P alto) da muchos más resultados relevantes que irrelevantes, mientras que un algoritmo muy sensible, (TP alto), será el que detecte la mayoría de los resultados de interés.

EL F1 Score es otra métrica muy empleada porque resume la precisión y sensibilidad en una sola métrica, por ello es de gran utilidad cuando la distribución de las clases es desigual. Se calcula:

$$F1Score = \frac{2 * (Recall * Precision)}{(Recall + Precision)}$$

Conforme a estas métricas según (Barrios, 2019) se puede obtener cuatro casos posibles para cada clase:

- Alta Precisión y alto Recall: el modelo maneja perfectamente esa clase.
- Alta Precisión y bajo Recall: el modelo no detecta la clase muy bien, pero cuando lo hace es altamente confiable.
- Baja Precisión y alto Recall: El modelo detecta bien la clase, pero también incluye muestras de la otra clase.
- Baja Precisión y bajo Recall: El modelo no logra clasificar la clase correctamente.

En cuanto a las métricas de medición de la curva ROC, (Singh, 2020) considera que, medir el área bajo la curva ROC es también un método muy útil para evaluar un modelo, debido a que al trazar la tasa positiva verdadera (sensibilidad) frente a la tasa de falsos positivos (1 - especificidad), se obtiene la curva de Característica Operativa del Receptor (ROC), la cual permite visualizar el equilibrio entre la tasa de verdaderos positivos y la tasa falsos positivos

4.4.2. Resultados Regresión Logística

Al entrenar el modelo de regresión logística y poner a prueba su capacidad de generalización con las bases de validation y test, se obtuvo los siguientes resultados:

Resultados de entrenamiento

Matriz de Confusión

	Predicción 0	Predicción 1
Target 0	74,682	3,575
Target 1	10,051	6,309

Figura 4.12: Matriz de confusión entrenamiento Regresión logística

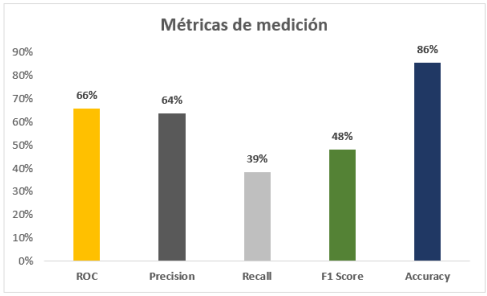

Figura 4.13: Métricas de medición entrenamiento Regresión logística

Resultados de validación

	Predicción 0	Predicción 1
Target 0	24,936	1,203
Target 1	3,300	2,101

Figura 4.14: Matriz de confusión validación Regresión logística

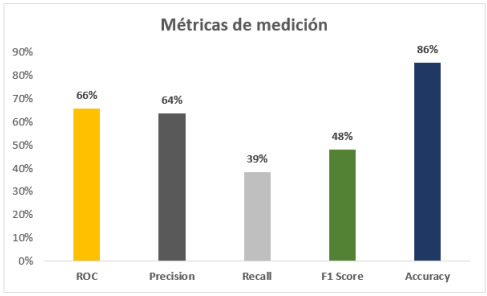

Figura 4.15: Métricas de medición validación Regresión logística

Resultados de testeo

	Predicción 0	Predicción 1
Target 0	24,841	1,238
Target 1	3,366	2,094

Figura 4.16: Matriz de confusión testeo Regresión logística

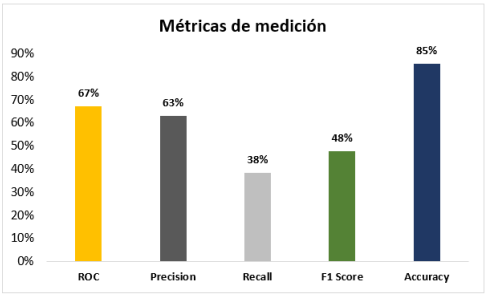

Figura 4.17: Métricas de medición testeo Regresión logística

Como se observa los resultados de las métricas de precisión, recall, F1 Score y accuracy, no presentaron variaciones importantes en la validación y testeo. Al tomar el resultado de las métricas del testo, se observa que el recall es del 38 % y la precisión estuvo en 63 %, es decir, que la regresión logística solo identifico el 38 % de los clientes auto cura y de los clientes auto cura que predijo, solo el 63 % realmente lo fueron.

4.4.3. Resultados Árbol de decisión

Al generar el entrenamiento del árbol de decisión y poner a prueba su capacidad de generalización con las bases de validation y test, se obtuvo los siguientes resultados:

Resultados de entrenamiento

	Predicción 0	Predicción 1
Target 0	76,780	1,477
Target 1	4,147	12,213

Figura 4.18: Matriz de confusión entrenamiento Árbol de decisión

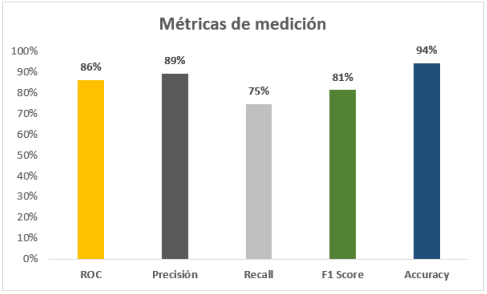

Figura 4.19: Métricas de medición entrenamiento Árbol de decisión

Resultados de validación

	Predicción 0	Predicción 1
Target 0	25,632	507
Target 1	1,394	4,007

Figura 4.20: Matriz de confusión validación Árbol de decisión

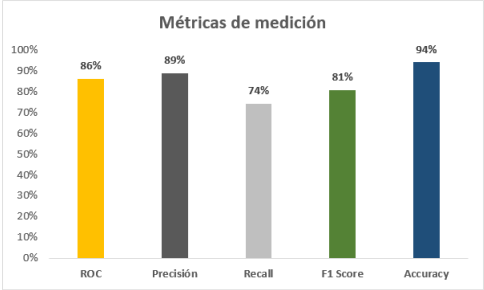

Figura 4.21: Métricas de medición validación Árbol de decisión

Resultados de testeo

	Predicción 0	Predicción 1
Target 0	25,512	567
Target 1	1,462	3,998

Figura 4.22: Matriz de confusión testeo Árbol de decisión

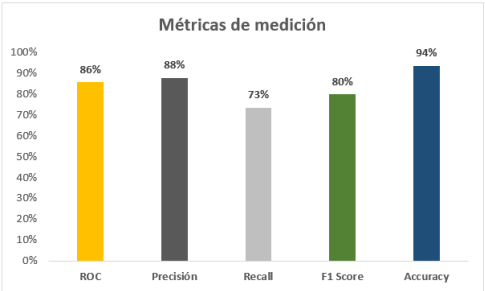

Figura 4.23: Métricas de medición testeo Árbol de decisión

Al revisar los resultados de árbol de decisión, se identifica que las métricas de precisión, recall, F1 Score, accuracy y ROC, no presentaron variaciones significativas en la validación y en el testeo. Al tomar el resultado del test, se obtuvo un recall del 73 % y una precisión del 88 %, es decir, que el árbol de decisión identificó el 73 % de los clientes auto cura y de los clientes auto cura que predijo, el 88 % realmente lo fueron.

4.4.4. Resultados Random Forest

Al realizar el entrenamiento de Random Forest y poner a prueba su capacidad de generalización con las bases de validation y test, se obtuvo los siguientes resultados:

Resultados de entrenamiento

	Predicción 0	Predicción 1
Target 0	77,737	520
Target 1	6,828	9,532

Figura 4.24: Matriz de confusión entrenamiento Random Forest

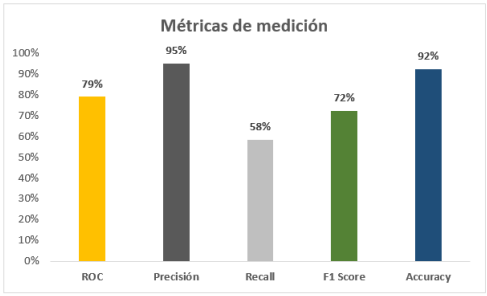

Figura 4.25: Métricas de medición entrenamiento Random Forest

Resultados de validación

	Predicción 0	Predicción 1
Target 0	25,938	201
Target 1	2,317	3,084

Figura 4.26: Matriz de confusión validación Random Forest

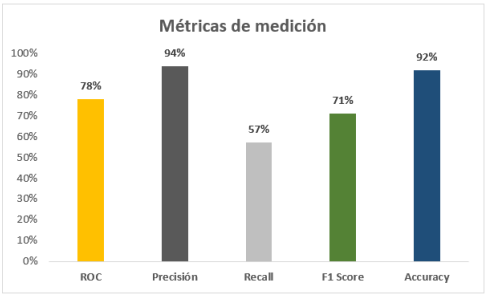

Figura 4.27: Métricas de medición validación Random Forest

Resultados de testeo

	Predicción 0	Predicción 1
Target 0	25,882	197
Target 1	2,337	3,123

Figura 4.28: Matriz de confusión testeo Random Forest

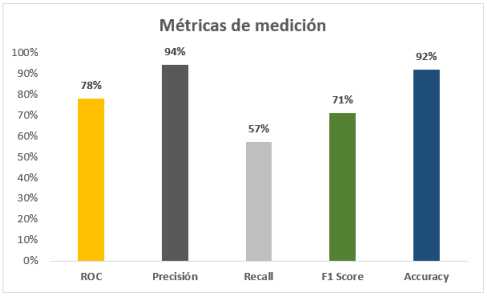

Figura 4.29: Métricas de medición testeo Random Forest

Los resultados de Random Forest en las métricas de precisión, recall, F1 Score, accuracy y ROC, son estables en la validación y en el testeo del modelo. Al revisar las métricas del test, el recall obtuvo un rendimiento del 57 % y la precisión del 94 %, es decir, respecto al rendimiento del recall de 10 clientes auto cura el modelo reconoció aproximadamente 6 clientes. Con respecto a la precisión, de 10 clientes que se predijeron como auto cura aproximadamente 9 son realmente auto cura.

4.4.5. Resultados Máquinas de soporte vectorial

En el entrenamiento de Máquinas de soporte vectorial y poner a prueba su capacidad de generalización con las bases de validation y test, se obtuvo los siguientes resultados:

Resultados de entrenamiento

	Predicción 0	Predicción 1
Target 0	74,229	4,028
Target 1	7,945	8,415

Figura 4.30: Matriz de confusión entrenamiento Máquinas de soporte vectorial

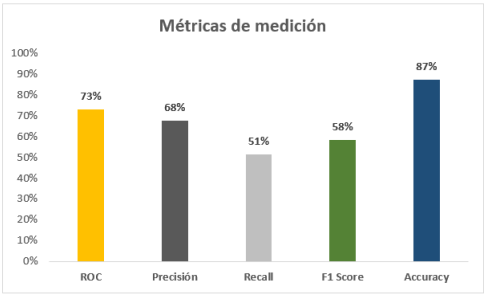

Figura 4.31: Métricas de medición entrenamiento Máquinas de soporte vectorial

Resultados de validación

	Predicción 0	Predicción 1
Target 0	24,787	1,352
Target 1	2,677	2,724

Figura 4.32: Matriz de confusión validación Máquinas de soporte vectorial

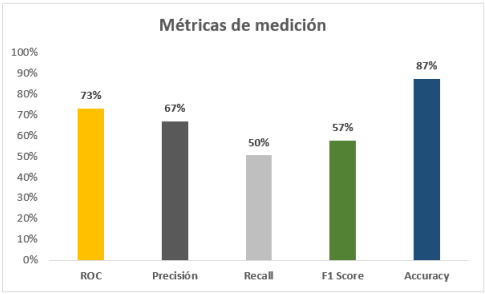

Figura 4.33: Métricas de medición validación Máquinas de soporte vectorial

Resultados de testeo

	Predicción 0	Predicción 1
Target 0	24,701	1,378
Target 1	2,675	2,785

Figura 4.34: Matriz de confusión testeo Máquinas de soporte vectorial

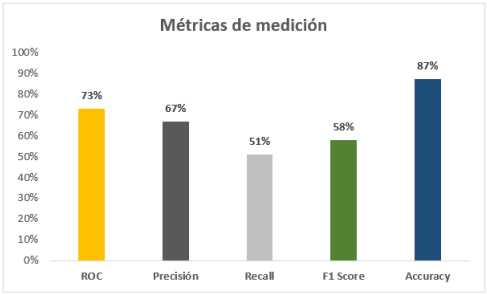

Figura 4.35: Métricas de medición testeo Máquinas de soporte vectorial

La máquina de soporte vectorial no presento variación significativa en las métricas de precisión, recall, F1 Score, accuracy y ROC en el entrenamiento, validación y testo del modelo. Al revisar las métricas del test, el recall obtuvo un rendimiento del 51 % y la precisión del 67 %, es decir, respecto al rendimiento del recall de 10 clientes auto cura el modelo reconoció aproximadamente 5 clientes. Con respecto a la precisión, de 10 clientes que se predijeron como auto cura aproximadamente 7 lo son realmente auto cura.

4.4.6. Resultados K vecinos más cercanos (Knn)

En el entrenamiento de K vecinos más cercanos y poner a prueba su capacidad de generalización con las bases de validation y test, se obtuvo los siguientes resultados:

Resultados de entrenamiento

	Predicción 0	Predicción 1
Target 0	77,195	1,062
Target 1	1,641	14,719

Figura 4.36: Matriz de confusión entrenamiento Knn

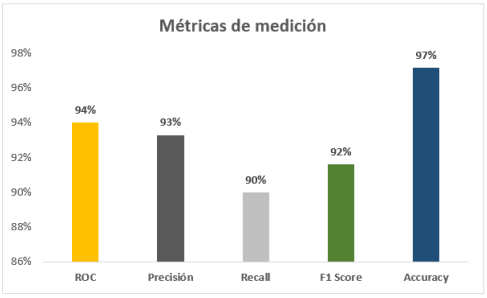

Figura 4.37: Métricas de medición entrenamiento Knn

Resultados de validación

	Predicción 0	Predicción 1
Target 0	25,733	406
Target 1	565	4,836

Figura 4.38: Matriz de confusión validación Knn

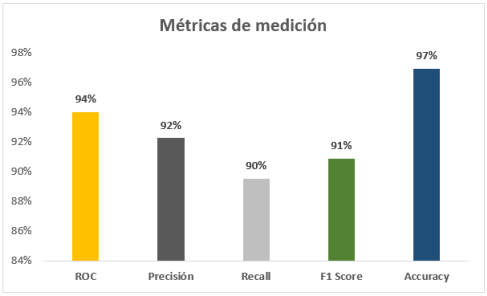

Figura 4.39: Métricas de medición validación Knn

Resultados de testeo

	Predicción 0	Predicción 1
Target 0	25,685	394
Target 1	576	4,884

Figura 4.40: Matriz de confusión testeo Knn

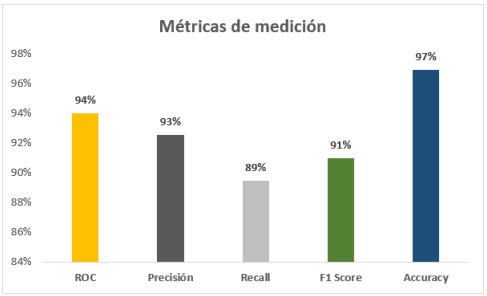

Figura 4.41: Métricas de medición testeo Knn

K vecinos más cercanos obtuvo resultados consistentes en las métricas de precisión, recall, F1 Score, accuracy y ROC en el entrenamiento, validación y testo del modelo. Al revisar las métricas del test, el recall obtuvo un rendimiento del 89 % y una precisión del 93 %, es decir, respecto al rendimiento del recall de 10 clientes auto cura el modelo reconoció aproximadamente 9 clientes. Con respecto a la precisión, de 10 clientes que se predijeron como auto cura, 9 lo son realmente auto cura.

4.4.7. Elección del modelo de machine learning

Para elegir el modelo que ayudara en la predicción de los clientes auto cura, se analizaron los resultados de las métricas del testeo. Como se observa en la imagen 4.42, el modelo que obtuvo mejor consistencia en los resultados de las métricas evaluadas (Precision, Recall, F1 score, ROC y Accuracy), es K vecinos más cercanos (Knn), por ende, se selecciona dicho modelo para afrontar el reto de identificación de los clientes auto cura.

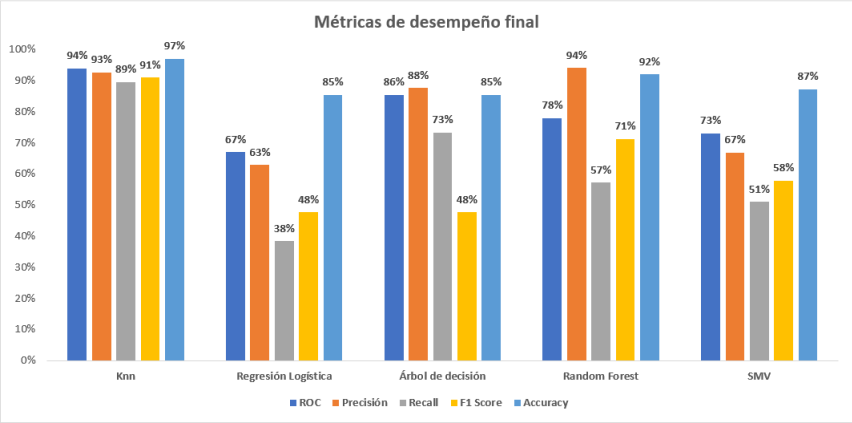

Figura 4.42: Desempeño final de los Modelos

Capítulo 5

CONCLUSIONES

Con el fin de recolectar la información en una base de datos consolidada que describa el comportamiento de los clientes auto cura de forma fácil y ágil, se implementó un código que integra las bases de traslados, tanque de pagos y gestiones, adicional, se construyó un visualizador que permitió conocer el comportamiento de las variables relevantes, y así, detectar patrones de cambio en ellas.

Dado que en la base de datos construida no posee las etiquetas que describen el comportamiento auto cura en los registros, se desarrolla un modelo utilizando aprendizaje automático no supervisado con el método de K-means, inicializando este método con 4, 5 y 6 clústeres. La separación entre los grupos y la cohesión interna en cada clúster optimo se consiguió con el modelamiento de 5 clústeres, esto permitió identificar un grupo que embebe el 17 % de los registros que describen el comportamiento auto cura, con esta acción se obtuvo la variable dependiente, asignando 1 a este porcentaje de registros y 0 al resto del conjunto de datos.

Con el fin de encontrar el conjunto de datos óptimo para la elaboración de técnicas de aprendizaje automático supervisado, se desarrolló selección de variables con el método de Elastic Net con los criterios de $\lambda = 0{,}050119$ y un $\alpha = 0{,}05$, esto permitió regularizar las variables, pasando de 52 variables a 40 características.

Se desarrollo un modelo de Machine Learning supervisado que permite reconocer a los clientes auto cura con un rendimiento en las métricas de evaluación de: 93 % en Precisión, 89 % en Recall, 91 % en F1 Score, 94 % en Roc y 97 % en Accuracy. Dados a los resultados obtenidos en las métricas del Recall, Precisión y demás indicadores, el negocio acepta el modelo, ya que posee capacidad de generalización y sus resultados fueron consistentes en el entrenamiento, validación y testeo. Con la implementación de este modelo, los esfuerzos en las gestiones se concentrarán en los clientes que requieren acompañamiento en sus pagos.

Al colocar en producción el modelo de Machine Learning, se debe realizar un seguimiento exhaustivo durante aproximadamente de 3 meses, con el fin de evaluar los rendimientos obtenidos en el proceso de gestión de cobranzas, y así, identificar si el modelo esta generalizando de acuerdo a las métricas evaluadas o requiere un ajuste.

Capítulo 6

REFERENCIAS

ALEA, V., JIMENEZ, E., MUÑOZ, MC., TORRELLES; E. VILADOMIU, N. (2014). Guía para el análisis estadístico con RCommander. Barcelona, Text docent 391UB

Anil. K Jain Data clustering: 50 year beyond K-measn Pattern Recognition Letters, 31(8) (2010), pp. 651-666, 264-323

Barrios, J. (2019). La matriz de confusión y sus métricas.
Obtenido de https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Singapur: Springer. Disponible en http://users.isr.ist.utl.pt/ wurmd/Livros/school/Bishop %20 %20Pattern % 20Recognition %20And %20Machine %20Learning %20- % 20Springer %20 %202006.pdf

Bonastre, R. (2017). La inteligencia artificial en la gestión de recuperación de deuda [en línea, 6 de noviembre]. Innovan.do.

Bottou, L., Lin, C. J. (2007). Support vector machine solvers. Large scale kernel machines, 3(1), 301-320. Disponible en https://innovan.do/2017/11/06/la-inteligencia-artificial-en-la-gestionde-larecuperacion-de-deuda/

Bradley, Fayyad (1998). Refining initial points for k-means clustering, In: Proc of the 15th int. conf. on machine learning, pp. 91-99

Breiman, L., Friedman, J., Stone, C., y Olshen, R. (1984). Classification and regression trees. California, Estados Unidos: Wadsworth, Inc.

Carrasco, M. (2016). TECNICAS DE REGULARIZACI ON EN REGRESÍON: IMPLEMENTACION Y APLICACIONES.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. L. Erl-

baum Associates.

Dunn J. C. (1973): Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 3257

Duque, J.L., González, C.H., García, Mónica. (2014). Outsourcing y Business Process Outsourcing desde la Teoría Económica de la Agencia. Entramado, 10(1), 12-29. FACULTAD DE MEDICINA HUMANA

Friedman, J., Hastie, T., Tibshirani, R. (2001). Random Forests. The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics (pp 587-592).

García, H.D, Polanco F. (2019) METODOLOGIA DE LA INVESTIGACION I. Licenciatura en Psicomotricidad, 9.

Hastie, T., Tibshirani, R. y Friedman, J (2009). The Elements of Statistical learning. 119 – 121

Hoerl, A.E. y Kennard, R.W. (1970). Ridge regression: biased estimation for non-orthogonal problems. Technometrics, 12(1): págs. 55-67.

IBR Latam (2018). Soluciones Multicanal. Sitio web ibrlatam.com/sitio2018

Izenman Alan J. (2008). Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning, 9-13, 281-288.

Konecta (2002). Konecta España: líderes en BPO y Relación Clientes. Disponible en https://www.grupokonecta.com/somos-globales/espana/

Mao J., Anil. K Jain, A self-organizing network for hyper ellipsoidal clustering (HEC) IEEE Transacations on Neural Networks, 7 (1) (1996), pp 16-29.

Marketinganalítico (2016). TIPOS DE VARIABLES EN EL ANALISIS DE DATOS Disponible en https://www.marketing-analitico.com/analitica-web/tiposde-variables-en-el-analisis-de-datos/

MathWorks (2019). Machine Learning. Tres cosas que es necesario saber [en línea]. Disponible en es.mathworks.com/discovery/machine-learning.html

Martínez, E. (2014) Scoring de cobranza preventiva en la gestión de riesgo [en línea, 5 de febrero]. Consumer Risk Analytics.
Disponible en https://edgarmartinezq.wordpress.com

Merino, R. F. M., Chacon, C. I. N. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases,

(10), 165-189.

Morales, P. (2019). Inteligencia artificial en procesos de cobranza de grandes empresas: más allá del Bot [e-Book]. Providencia, Chile: Digevo Corp. Disponible por descarga en http://soluciones.digevo.com/ebook-inteligenciaartificialen-procesos-de-cobranza-2

Olivon F., Elie N., Grelier G., Roussi F., Litaudon M., Touboul MetGem M. (2018) Software for the Generation of Molecular Networks Based on the t-SNE Algorithm

Prezi (2020). Konecta Multienlace BPO.
Disponible en https://prezi.com/hs8grx5rqzzc/konecta-multienlace-bpo/

Ranjan, G. S. K., Verma, A. K., Radhika, S. (2019, March). K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In 2019 IEEE 5th international conference for convergence in technology (I2CT) (pp. 1-5). IEEE.

Rangel L, R. M. (2010). Marketing comunidad.
Recuperado de http://www.marketingcomunidad.com/call-center-versus-contact center.html

Resendiz, J. (2006). Las máquinas de vectores de soporte para identificación en línea. Disponible en: https://www.ctrl.cinvestav.mx/ yuw/pdf/MaTesJAR.pdf

RPubs (2019). Variables dummy (one-hot encoding) con R
Disponible en https: https://rpubs.com/jboscomendoza/vairables dummy con r

RPubs (2020). Algoritmo k-Nearest Neighbors (kNN)
Disponible en: https://rpubs.com/Kataniss/636518

Raschka S., Mirjalili V (2017). Python Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow

Santos, P. d. (23 de enero de 2018). Machine Learning a tu alcance: La matriz de confusión. Obtenido de https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/

Sanabria, F.G. (2015). Estudio de caso desde la experiencia de empresa, del sector del transporte, en el contact center. Universidad Militar nueva Granada Facultad de estudios a distancia Administración de Empresas.

Schneider, B. (2012). OUTSOURCING, la herramienta de gestión que revoluciona el mundo de los negocios. Bogotá: Editorial Norma.

Segal, M.; Dahlquist, K. y Conklin, B. (2003). Regression approach for microarray data analysis. J. Computational Biology, 10(6): págs. 961-980.

Sierra, C. (2020). Matriz de Confusión.

Singh, N. (2020). Métricas De Evaluación De Modelos En El Aprendizaje Automático. Obtenido de https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso. A retrospective, Journal of the Royal Statistical Society: Series B (Methodological), 73(3): págs. 273-282.

TransUnion (2018). Agilice su proceso de toma de decisiones a través del poder de las soluciones analíticas [en línea]. TransUnion.
Disponible en https://www.transunion.mx/solucion/analiticas

Zhang J.S, Leung Y.W (2003) Robust clustering by pruning outliers

Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal Royal Statistical Society: Series B, 67(2): págs. 301 320.