

Universidad Cenfotec

Maestría en Tecnología de Bases de Datos

Documento final de Proyecto de Investigación Aplicada 2

Aplicación de métodos predictivos de minería de datos para determinar la rentabilidad de clientes nuevos en el Banco Costa Rica

Ing. Ruiz Cid Milton Eduardo

Marzo, 2018

Declaratoria de derechos de autor

Yo, **Milton Eduardo Ruiz Cid**, número de identificación **5-0341-0369**, estudiante de la Universidad CENFOTEC, de la carrera Maestría en Tecnología de Bases de Datos, por este medio autorizo a la Universidad CENFOTEC y a sus estudiantes activos a que el presente documento pueda ser consultado única y exclusivamente con fines académicos y en presencia de un representante de la misma universidad.

Milton Eduardo Ruiz Cid

Agradecimientos

En primer lugar, agradezco a **Dios** por permitirme culminar este proyecto de maestría, porque solo él sabe que el esfuerzo ha sido formidable, el tiempo de trabajo que he dedicado a este proyecto ha sido arduo para mí y mi familia, pero hoy siento una gran tranquilidad y felicidad, además de un sentimiento gratificante porque lo di todo para obtener los mejores resultados.

A mi esposa **Marilyn Chaves Cordero**, a quien le dedico mucho este título porque ha sacrificado tiempo juntos en familia, pero aun así ha estado conmigo dándome fuerzas, apoyo e impulsándome a finalizar la maestría porque es importante para nuestra familia.

A mis hijos **Fabiana Ruiz Chaves y José Paulo Ruiz,** a quienes les he quitado tiempo de calidad conmigo, pero me esforcé porque ellos son mi ilusión y mi motor de cada día, quienes me inspiran a terminar, ser mejor persona y profesional, para darles un buen ejemplo en la vida.

A mi madre **Enid Ruiz Cid**, a quien también le dedico enormemente este título, porque desde niño me enseñó buenos valores y hoy me siento orgulloso del hombre que soy, porque soy el reflejo y fruto de su educación.

A mi abuela **Enna Cid Baltodano**, quien me enseñó la importancia del estudio; me enseñó a leer, escribir y realizar operaciones matemáticas a los 4 años.

A mi papá **Jefrey Provenzano**, quien me ha enseñado a ser puntual, esforzado y perseverante, su forma de ser me ha permitido entender que nada es imposible, que los límites están en nuestras mentes, me ha inspirado a trabajar fuerte, para un día ser un profesional independiente y darle lo mejor a mi familia.

A mi tutor **José Cabezas Jaikel,** por su enorme paciencia, dedicación, tiempo, esfuerzo y porque con su experiencia me ayudó a realizar un excelente trabajo.

A mi lector externo **Luis Naranjo Zeledón,** por su tiempo, aportes, disposición en todo momento para revisar mi trabajo y aconsejarme con su experiencia para afinar más mi proyecto.

A las personas que fundaron y trabajan en la **Universidad CENFOTEC**, por crear esta maestría que me permitió especializarme en lo que me apasiona.

A mis **compañeros**, quienes hoy son mis amigos, porque con sus experiencias y conocimientos me inspiraron a esforzarme mucho para ser igual de buenos profesionales que ellos.

Documento de aprobación final del Tribunal

Abstract o resumen ejecutivo

El Banco de Costa Rica es la segunda institución financiera semiestatal más grande de Costa Rica, en la actualidad busca generar nuevas oportunidades de negocio para sus clientes actuales y futuros.

La institución tiene claro lo importante que es contar con la información correcta en el menor tiempo posible, para ello busca digitalizar muchos de sus procesos y negocios, con el fin de mejorar los tiempos de respuesta de sus clientes y, a su vez, ayudarles a las gerencias a tomar mejores decisiones en el menor tiempo posible.

La organización está mentalizada en la digitalización de la banca en diferentes áreas, especialmente está muy interesada en la creación de un *Big Data* para manipular diversos tipos de información, como datos estructurados, no estructurados y semiestructurados.

Como primer paso, se quiere aplicar reglas de minería de datos en distintas áreas, una de ellas es la posibilidad de predecir si sus clientes nuevos son rentables para la organización, mediante la búsqueda de patrones en sus clientes actuales, que permitan saber dónde debe enfocarse para mejorar sus negocios, servicios y productos.

Hay claridad en que deben crearse modelos exploratorios y predictivos que le posibiliten determinar si un cliente nuevo puede ser potencialmente rentable para la entidad bancaria o no, luego a partir de esa información tomar decisiones del trato que debe darse al cliente en cuanto a productos y servicios, para saber qué ofrecerle sin exponer las utilidades del Banco.

Al crear el modelo predictivo personalizado para el Banco, se deben identificar en conjunto con un grupo interdisciplinario de gobierno de datos y de inteligencia financiera, cuáles son las variables predictoras correctas que se van a utilizar durante el proceso de calibración de los algoritmos y mediante la aplicación de validación cruzada se escoge el algoritmo óptimo para analizar los datos y realizar la predicción; siempre es necesario considerar que los métodos predictivos requieren ser revisados y ajustados porque los datos varían con el tiempo.

Palabras claves: minería de datos, rentabilidad, patrones, modelos predictivos, calibración, validación cruzada.

Tabla de contenido

Capítulo I Introducción 12
Generalidades 12
Antecedentes del problema 12
Definición y descripción del problema 12
Justificación 13
Viabilidad 13
Viabilidad técnica. 13
Viabilidad operativa. 14
Viabilidad económica. 14
Objetivos 18
Objetivo general. 18
Objetivos específicos. 18
Alcances y limitaciones 19
Alcances 19
Limitaciones. 20
Marco de referencia organizacional y socioeconómico 21
Historia. 21
Tipo de negocio y mercado meta. 22
Misión, visión y valores 23
Políticas institucionales. 24
Estado de la cuestión 24
Capítulo II Marco teórico 30

Cliente 30
Rentabilidad de cliente 30
Integración de datos 31
Limpieza de datos 32
Calidad de datos 33
Minería de datos 33
Machine learning 38
Evaluación y selección de modelos 42
Métricas para evaluar el rendimiento del clasificador. 42
Método de retención y submuestreo aleatorio. 45
Bootstrap. 46
Validación cruzada. 47
Árboles de decisión 50
K-means 51
K-medias. 52
K-medoides. 53
Máquinas de soporte vectorial (SVM) 53
Redes neurales (NN) 54
Bosques aleatorios (Random Forests) 56
Potenciación (Boosting) 57
Capítulo III Marco metodológico 59
Tipo de investigación 59
Alcance investigativo 59

Enfoque 59
Diseño 64
Población y muestreo 64
Instrumentos de recolección de datos 65
Instrumentos de análisis de la información 65
Estrategia de desarrollo de la propuesta 65
Capítulo IV Análisis del diagnóstico 66
Capítulo V Propuesta de solución 70
Fases de CRISP-DM 71
Fase I. Comprensión del negocio. 71
Fase II. Comprensión de los datos. 72
Fase III. Preparación de los datos. 72
Fase IV. Modelado. 72
Fase V. Evaluación. 72
Fase VI. Implementación. 73
Puesta en marcha de las fases CRSP-DM en el proyecto 73
Comprensión del negocio. 74
Comprensión de los datos. 74
Preparación de los datos. 79
Modelado. 92
Evaluación. 99
Implementación. 106
Capítulo VI Conclusiones y recomendaciones 108

Conclusiones 108

Recomendaciones 111

Capítulo VII Reflexiones finales 113

Capítulo VIII Trabajos a futuro 114

Glosario 116

Bibliografía 120

Apéndices 125

Índice de tablas

Tabla 1. Precios de las licencias RStudio 15

Tabla 2. Viabilidad económica 16

Tabla 3. Costo de desarrollo del proyecto 17

Tabla 4. Costo de desarrollo del proyecto 17

Tabla 5. Matriz de confusión 44

Tabla 6. Matriz de confusión, determina si los correos son *spam* o no 44

Tabla 7. Matriz de confusión, determina si es fraude o no 44

Tabla 8. Fórmulas para evaluar modelos 45

Tabla 9. Ejemplo completo para determinar si los correos son *spam* o no 45

Tabla 10. Criterios de evaluación de facilidad de implementación. 62

Tabla 11. Clasificación de los datos 78

Tabla 12. Archivos HTML creados en el desarrollo del proyecto 125

Índice de figuras

Figura 1. Integración de datos 32
Figura 2. Limpieza de datos 32
Figura 3. La minería de datos como un paso en el proceso de descubrimiento de conocimiento 35
Figura 4. Minería de datos, confluencia de múltiples disciplinas 38
Figura 5. Aprendizaje semisupervisado 42
Figura 6. Estimación de precisión con el método retención 46
Figura 7. Ontología de algoritmos de predicción 61
Figura 8. Modelo de referencia, fases de CRISP-DM 71
Figura 9. Exploración de datos físicos 76
Figura 10. Carga de datos físicos 80
Figura 11. Elimina datos NA 81
Figura 12. Estandarizar datos 82
Figura 13. Cambio de observaciones en variables 83
Figura 14. Análisis de calidad en las variables 84
Figura 15. Variables por eliminar en el primer análisis 85
Figura 16. Crea las variables producto 86
Figura 17. Desnormalizar datos 87
Figura 18. Guardar el archivo preseleccionado 88
Figura 19. Discretización por clúster 89
Figura 20. Análisis de componentes principales 90
Figura 21. Gráfico de correlación de variables 91

Figura 22. Validación cruzada cinco veces 92
Figura 23. Validación cruzada seis veces 92
Figura 24. Validación cruzada 15 veces 93
Figura 25. Validación cruzada cuatro veces 93
Figura 26. Validación cruzada con K-Folds 94
Figura 27. Creación de los clústeres para usar paralelismo 95
Figura 28. Paralelismo en R 96
Figura 29. Gráficas del modelo con validación cruzada 98
Figura 30. Evaluación del modelo de validación cruzada 99
Figura 31. Nuevos datos predichos 100
Figura 32. Monto de rentabilidad de clientes por segmento 101
Figura 33. Total de clientes rentables por sexo 101
Figura 34. Colocación y captación de clientes rentables por segmento 102
Figura 35. Colocación y captación de clientes rentables por provincia 103
Figura 36. Total de clientes rentables por segmento 104
Figura 37. Total de clientes rentables por categoría de riesgo 105
Figura 38. Serie de tiempo por rentabilidad 106
Figura 39. Arquitectura propuesta para la implementación 107
Figura 40. Curva ROC sobre correos spam 117

Capítulo I Introducción

Generalidades

Uno de los principales objetivos de la minería de datos es crear un proceso automatizado que ayude en la toma de decisiones a partir de los datos, los cuales pueden provenir de una bodega de datos o de otros insumos, por ejemplo, archivos.

A partir de lo anterior, este proyecto consiste en plantear un método personalizado y automatizado de predicción sobre un conjunto de datos que se divide en datos de prueba y entrenamiento, aplicando el método de validación cruzada para descubrir cuál es el algoritmo más óptimo para el Banco Costa Rica, con el fin de establecer la rentabilidad de los clientes actuales y futuros.

Antecedentes del problema

En la actualidad la institución recibe muchos clientes en todas sus sucursales bancarias, pero no tiene un modelo ágil y eficaz que le permita saber si cada uno es o no un buen cliente o, como esta lo denomina, un cliente rentable.

Definición y descripción del problema

Según el Banco, desde el punto de vista de negocio, un cliente rentable es aquel que perdura por más tiempo en la institución aportando no solo dinero, sino también potencial de crecimiento para ambas partes, experiencia y nuevos contactos o clientes.

Hoy la rentabilidad de un cliente no debe medirse solo en función del dinero que él produce o de la inversión que realiza la empresa para atenderlo, sino también deben analizarse otros factores

externos, por ejemplo, los comentarios que se hacen en las redes sociales, con el propósito de saber el grado de satisfacción de los clientes y determinar si los servicios y productos que se brindan son los esperados por ellos.

Por estas razones, la institución pretende crear un modelo predictivo ágil, que pueda ser utilizado por todos los colaboradores que atienden al público, el cual analice todos los factores, relaciones, correlaciones, tendencias, clústeres y segmentos que posibiliten encontrar patrones ocultos usando medios automatizados, algorítmicos, matemáticos o estadísticos para conocer la rentabilidad de los clientes actuales y nuevos.

Justificación

El Banco de Costa Rica tiene la necesidad de establecer un modelo predictivo que le ayude a saber si sus clientes actuales y nuevos son rentables.

Por tal motivo, este proyecto surge con la oportunidad de aplicar reglas de minería de datos para crear un modelo automatizado, obteniendo información desde las fuentes ya existentes como lo son el DWHC, lago de datos y cualquier otra fuente necesaria como archivos y planos en formatos txt, csv y Excel, con el objetivo de ser un apoyo para la toma de decisiones.

Viabilidad

Viabilidad técnica.

Después de hacer una evaluación tecnológica del equipo de la institución, de las licencias del *software* y de la preparación técnica de un grupo de colaboradores del mismo banco, se determina que está en proceso la licitación para adquirir la infraestructura tecnológica y el licenciamiento adecuado para el desarrollo y puesta en producción del método o los métodos

predictivos que sean construidos durante la realización del trabajo, por lo tanto de momento no se deben adquirir equipos nuevos ni licencias de *software*.

Viabilidad operativa.

Esta solución es factible aplicarla a nivel operacional, una vez que se cuente con la factibilidad técnica requerida para la realización y puesta en producción del modelo. Además, se cuenta con el equipo técnico y profesional capacitado para el desarrollo, también existe un fuerte apoyo de la organización para que produzca este modelo, por ende, no hay resistencia al cambio; al contrario, hay mucho interés para que sea implementado en el menor tiempo posible, con el fin de obtener los resultados a corto plazo.

Viabilidad económica.

El proyecto es factible económicamente porque innova al Banco y le da una ventaja competitiva sobre el resto de entidades financieras al crear una solución mediante la aplicación de algoritmos de predicción, que le permite mejorar la toma de decisiones de la organización en menor tiempo.

Lo anterior facilita tener un acercamiento con los clientes y que sean analizados adecuadamente, para conocer cuáles son las ventajas económicas que van a traer a la institución y cuáles productos nuevos se les pueden ofrecer o crear con el objetivo de aumentar los ingresos.
Por otra parte, posibilita que las diferentes áreas del negocio puedan mejorar la toma de decisiones con mayor eficacia y fluidez, lo que disminuye sus tiempos de respuesta y les ayuda a brindar soluciones, ofrecer mejores productos y servicios a los clientes en menor tiempo y con menor

costo, gestionar mejor el tiempo de los colaboradores y personalizar la atención de los clientes más importantes, así como abandonar aquellos que son un riesgo e implica más recursos mantenerlos.

El *hardware*, el espacio físico, las licencias de *software* y el personal que apoya el proyecto son aportados por la institución, por lo tanto el costo es asumido por la misma. En la tabla 1 se aprecia el costo de las licencias de las diferentes herramientas de *software* que debe adquirir el Banco para poner el proyecto en producción.

Tabla 1. Precios de las licencias RStudio

Producto	Descripción	Precio
RStudio Server Pro y RStudio Professional IDE	Precio por año por servidor (usuarios ilimitados), es para desarrollar a escala.	$ 19 990
RStudio Connect	Precio por año, se usa para publicar todo el trabajo que sus equipos crean en R.	$ 74 995
RStudio Shiny Server Pro	Precio por año, se usa para publicar en la web y apps, además permite tener 150 usuarios concurrentes.	$ 24 990
RStudio Shinyapps.io	Precio por año, es para hacer uso de las publicaciones y de los apps; la edición es profesional.	$ 3 300
	Total	$ 123 275

Fuente: Elaboración propia

No obstante, a continuación se presenta un cuadro con los costos que tiene el desarrollo de este proyecto sin recursos internos y usando una contratación externa (tabla 2). Cabe indicar que los siguientes números son reales y obedecen a un estudio de mercado llevado a cabo durante el mes de noviembre del año 2017 con dos de las empresas más grandes del mercado: HortonWorks y Cloudera.

Tabla 2. Viabilidad económica

Ítem	Descripción de la tarea	Cantidad de horas	Precio $ x hora	Valor total
1	**Servicio de soporte.**	640	125	$ 80 000
2	**Desarrollo evolutivo para carga y trasformación de los datos, evaluación e implementación.**	360	125	$ 45 000
3	**Desarrollo evolutivo, para la preparación y creación de modelos predictivos.**	686	150	$ 102 900
4	**Mantenimiento consultivo, correctivo y preventivo.**	160	125	$ 20 000
			Total	$ 247 900

Fuente: Elaboración propia

Todas las horas señaladas son en horario 5x8 (5 días hábiles de la semana) en horario de 8:00 a.m. hasta las 5:00 p.m. El mantenimiento consultivo comprende la atención de consultas en los aplicativos, originados a raíz de necesidades técnicas o funcionales; así como la generación de la documentación respectiva que la institución requiera. El mantenimiento preventivo abarca las actividades referentes a la atención y corrección de problemas o inconsistencias presentados e identificados en la plataforma y en los casos de negocio implementados.

Después de conocer los costos que tiene el desarrollo de este proyecto con una empresa externa, se puede ver que esto le provoca un ahorro de $ 147 900 que corresponden al desarrollo del proyecto (tabla 3).

Tabla 3. Costo de desarrollo del proyecto

Ítem	Descripción de la tarea	Cantidad de horas	Precio $ x hora	Valor total
1	**Desarrollo evolutivo para carga y trasformación de los datos, evaluación e implementación.**	360	$ 125	$ 45 000
2	**Desarrollo evolutivo, para la preparación y creación de modelos predictivos.**	686	$ 150	$ 102 900
	Total	1046		$ 147 900

Fuente: Elaboración propia

En la siguiente tabla 4 se observa más a detalle el costo de las tareas generales del proceso de desarrollo del proyecto bajo la metodología CRISP-DM.

Tabla 4. Costo de desarrollo del proyecto

Ítem	Descripción de la tarea	Cantidad de horas	Precio $ x hora	Valor total
1	**Comprensión del negocio**	70	125	$ 8 750
2	**Comprensión de los datos**	250	125	$ 31 250
3	**Preparación de los datos**	480	150	$ 72 000
4	**Modelado**	150	150	$ 22 500
5	**Evaluación**	56	150	$ 8 400
6	**Implementación**	40	125	$ 5 000
			Total	$ 147 900

Fuente: Elaboración propia

Objetivos

Existen diversas taxonomías para la definición de los objetivos como:

- Bloom, 1956.
- Bloom revisada, 2001.
- Marzano y Quendal, 2006.
- Era digital de Churches, 2008.

No obstante, los objetivos son formulados a partir de la taxonomía de Bloom 1956, ya que a pesar de haber pasado más de 50 años, hoy continúa siendo la herramienta fundamental para establecer objetivos de aprendizaje en forma exitosa, además los verbos que utiliza son los adecuados para la definición del proyecto.

Objetivo general.

Aplicar métodos predictivos de minería de datos para determinar la rentabilidad de los clientes actuales y futuros del Banco de Costa Rica.

Objetivos específicos.

Identificar las variables representativas que se van a utilizar en el modelo de minería de datos.

Descubrir la correlación de las variables disponibles para determinar cuáles inciden más sobre la rentabilidad del cliente.

Identificar el mejor algoritmo predictivo al crear y ejecutar el modelo de validación cruzada con todos los algoritmos.

Probar el algoritmo obtenido en la validación cruzada con un conjunto de datos que cumpla las mismas características del set de datos usado en el proceso.

Determinar los segmentos de clientes y zonas geográficas donde se concentran los clientes rentables.

Alcances y limitaciones

Alcances.

Se deben aplicar métodos predictivos de minería de datos, de acuerdo a las variables que tiene la empresa como insumos, con el objeto de desarrollar un modelo que calcule la rentabilidad de los clientes actuales y nuevos.

El método debe posibilitar la incorporación de nuevas variables para enriquecer el modelo. Asimismo, se debe tener bien claro que la identificación correcta de los ingresos y gastos y la calidad de la información suministrada determinan y sustentan la veracidad de la información.

El establecimiento de la rentabilidad de los clientes nuevos está apoyado en la experiencia y calidad de los datos actuales, con los cuales se identifica hoy a los clientes rentables.

Es necesario resaltar que al aplicar métodos predictivos de minería de datos para conocer la rentabilidad de los clientes, estos deben ser calibrados adecuadamente para determinar los mejores algoritmos que permitan establecer con claridad la mejor asertividad en los datos y disminuir el margen de error en las clases y el margen de error global de los modelos.

Los métodos predictivos son modelos probabilísticos, es decir, no son seguros por completo, tienen un margen de error y pueden tender a equivocarse, máxime si no son bien calibrados, a pesar de que los datos utilizados sean reales y se estén aplicando las mejores prácticas para descubrir cuál es el mejor algoritmo de predicción para el modelo de rentabilidad de clientes.

Limitaciones.

La calidad de los datos que existen en los sistemas fuentes y alimentan el almacén de datos corporativo puede afectar el resultado final de la información.

El cambio frecuente de las fuentes de datos que quieren incorporar las áreas de negocio, puede afectar los tiempos de entrega del proyecto, porque se requieren nuevos planes de trabajo.

Los cambios de fuentes pueden afectar la calibración de los métodos predictivos creados, ya que repercuten en las variables que apoyan el modelo de predicción cuando se apliquen los algoritmos.

Por políticas de seguridad de los datos, el nombre de la empresa y la información de los clientes de la institución no pueden ser expuestos ante terceras personas, sino que se encuentran bajo la custodia de algún representante oficial de la Universidad, para que mantenga la confidencialidad del documento.

Se presentan problemas de rendimiento en la computadora personal donde se pretende ejecutar el modelo, debido a la cantidad de información que se consulta en la base de datos. Por lo tanto, se acuerda solo usar una muestra de la misma, compuesta por ochenta y cinco mil trescientos sesenta (85 360) registros físicos; en el caso de los clientes jurídicos, se decide no considerarlos.

El conocimiento de las estructuras de la base de datos facilita la ubicación de la información; sin embargo, esta base tiene problemas de calidad por las fuentes que alimentan el DWHC, lo cual repercute en el manejo y preparación de los datos, porque en un principio se quiere obtener 107 variables para alimentar el modelo, pero la mayoría de ellas tiene inconsistencias como datos ausentes y nulos.

No se establece una conexión directa con la base de datos que posibilite consultar la información necesaria para el modelo, pues aún no se cuenta con el equipo de *hardware* ni con las licencias para el *software* RStudio Server; el proceso de licitación está en proceso.

Marco de referencia organizacional y socioeconómico

Historia.

De acuerdo con el Banco de Costa Rica (s.f.), la institución financiera es fundada el 20 de abril de 1877 con el nombre de Banco de la Unión, el cual mantiene hasta 1890, cuando lo cambia por el actual. Nace con el propósito de ser una nueva opción bancaria entre las ya existentes y tiene como funciones iniciales el prestar dinero, llevar cuentas corrientes, recibir depósitos, efectuar cobranzas, entre otras. A solo siete años de su fundación, la institución financiera se convierte en el único emisor de dinero y el primer administrador de las rentas públicas, mediante un contrato que se denomina Soto-Ortuño y que tiene vigencia hasta el año 1896.

Otros acontecimientos económicos suceden con los años hasta que en 1928 descentraliza sus servicios, con la creación de sus primeras sucursales en los puertos de Limón y Puntarenas.

En 1948 la Junta Fundadora de la Segunda República decreta la nacionalización de la banca, por lo que la institución financiera se integra a ella hasta la fecha. Luego, se define como institución autónoma de acuerdo con el artículo 189 de la Constitución Política de la República de Costa Rica de 1949.

En los últimos años se preocupa además por modernizar e innovar sus servicios y atención al público, procurando mayor agilidad y comodidad, mediante el uso y aplicación de su moderna tecnología.

Tipo de negocio y mercado meta.

Durante muchos años la institución enfoca sus giros de negocio hacia el público empresarial, básicamente en aquellas compañías que buscan ampliar su capital, invertir en maquinaria y ampliar su negocio.

No obstante, con la llegada del nuevo gerente, los tipos de negocio y mercados meta varían día tras día. La visión de la institución tiene cambios pequeños y ahora busca incursionar al negocio de la banca de consumo, por ello hoy está enfocada en la creación de diferentes marcas para atender distintos tipos de clientes según las necesidades que se encuentran mediante los comentarios de los clientes actuales y de personas que quieren ser clientes, también por medio de encuestas realizadas por el Área de Mercadeo.

Desde hace poco más de año y medio se trabaja en la creación de nuevas marcas con el fin de tener una mejor cercanía con el cliente, atendiendo sus necesidades y facilitándoles servicios; entre las cuales están Kristal, Esfera y Tucán.

La marca Kristal busca atender exclusivamente a las mujeres y refuerza el cambio de estrategia de esta entidad pública: un enfoque en la banca *retail*, pero sin descuidar el nicho empresarial.

La marca Tucán pretende tener una mayor cercanía con el público en general, proporcionando una plataforma de servicios bancarios que se brindan a clientes desde diferentes lugares como la pulpería del barrio o la ferretería de la esquina.

La marca esfera se orienta al desarrollo de fideicomisos para obra pública. El nombre hace referencia a las emblemáticas esferas de piedra del Diquís en Osa de la provincia de Puntarenas, patrimonio de la humanidad desde el 2014.

En camino se encuentran otras marcas, por ejemplo está la marca enfocada a los niños que hace más de una década se crea como banca infantil, pero que no se le da el seguimiento adecuado para fortalecerla, sin embargo está próxima a retomarse al igual que otras marcas enfocadas a *millennials*, *centennials* y estudiantes.

Todos estos cambios ocurren debido al mercado competitivo en que se encuentra esta entidad y todas las del país. Hoy si un banco quiere subsistir, debe replantear sus negocios para competir agresivamente con prácticas modernas de *retail*. Por ello la institución replantea la forma de hacer negocio y atraer clientes.

Misión, visión y valores.

Misión.

"Impulsar el desarrollo social, la competitividad y la sostenibilidad de Costa Rica al ofrecerles a sus clientes un conglomerado financiero público de excelencia e innovación en todos sus servicios" (Banco de Costa Rica, s.f.).

Visión.

"Ser la opción preferida de Costa Rica por ofrecer a sus clientes estándares mundiales de calidad, innovación y eficiencia en servicios financieros" (Banco de Costa Rica, s.f.).

Valores.

- "**Servicio**: Brindar un trato amable, respetuoso, oportuno y eficiente.
- **Excelencia**: Hacer las cosas bien desde la primera vez.
- **Compromiso**: Sentir como propios los objetivos y metas institucionales.

- **Integridad**: Mostrar coherencia entre lo que dice y lo que se hace.
- **Solidaridad**: Ayudar sin esperar a cambio recompensa o reconocimiento.
- **Innovación**: Generar y gestionar ideas que contribuyan a mejorar la ventaja competitiva de la organización" (Banco de Costa Rica, s.f.).

Políticas institucionales.

No existe una política institucional sobre los datos, no obstante se está comenzando a realizar una política de gobierno de datos, pero está desarrollándola un grupo de personas de diferentes áreas del Banco; hay colaboradores de TI y del área de Negocio, en especial del Área de CIE, ya que ellos pasan a ser los dueños de los datos del DWHC, lago de datos y de todos los datamart.

Estado de la cuestión

Desde hace muchas décadas la rentabilidad de clientes cobra interés en las empresas (Sevin, 1965). Aunque es muy extraño que se lleven a cabo análisis de este tipo (Mellman, 1963), a inicios de los años 80 se le comienza a dar su valor principalmente en el sector financiero (Ahern y Bercher, 1982; Knight, 1975), cuando la revolución del sector financiero de Estados Unidos provoca que los bancos busquen desarrollar procesos de análisis de rentabilidad de las cuentas para asegurarse que los balances son adecuados y las necesidades de los clientes más rentables se atienden de la mejor manera (Knight, 1975).

Antes de ahondar más en este tema, se debe tener claro qué es un cliente y qué es rentabilidad porque dentro de todas las áreas de las empresas, los conceptos pueden variar. Por lo

tanto, es necesario definirlos con el objetivo de tener claro qué se quiere analizar y cómo se va a hacer.

McNab (2005) explica que un cliente puede ser un titular de una cuenta, el originador de una transacción, un individuo, un hogar hasta un solicitante de un crédito o una empresa que requiere iniciar negocios o servicios. Algunas empresas financieras también manejan el concepto de cliente como cualquier persona que se acerca a efectuar una gestión de consulta sobre los servicios que se ofrecen y a estos individuos los tipifican como posibles futuros clientes potenciales.

Existen investigaciones donde se indica que para conocer la rentabilidad de los clientes basta con saber cuáles son los ingresos y los gastos, pero puede establecerse por simplicidad, como la diferencia entre los ingresos que aporte a la empresa menos los gastos que produzca la generación de tales ingresos. Otros señalan que no solo se deben analizar estas dos variables, sino asimismo otras variantes como las opiniones de los clientes en las redes sociales.

Al respecto, McNab (2005) comenta que este cálculo no es tan sencillo pues depende de la tipificación de los gastos e ingresos, del momento particular que vive el negocio y de las características propias de la empresa y el mercado en el que se circunscribe. Cada empresa tiene su propia tipificación de los ingresos y gastos, por lo tanto es fundamental conocer cuáles son los tipos dentro de la entidad financiera. Por ello en este trabajo se propone crear un modelo propio de rentabilidad de clientes que permita reflejar las diferencias particulares de la organización, así como sus requerimientos específicos, tanto actuales como futuros.

Los productos de colocación o captación generan tanto ingresos como costos operacionales y transaccionales, esto sumado a los procesos de cálculo de las tasas de costo o rendimiento y de los intereses pagados o cobrados, posibilita conocer la rentabilidad de un cliente.

A mediados de los años 80 surgen los sistemas de análisis de costos basados en actividades (CBA) como una forma de medir el costo de la demanda de recursos por parte de productos, servicios y clientes. Los sistemas CBA asignan los costos respectivos a las actividades y procesos, y luego a los productos, servicios y clientes. Este enfoque apoya el desarrollo de la administración basada en actividades (por sus siglas en inglés ABM *Activity-Based Management*), que es el conjunto de acciones que pueden ser tomadas utilizando la información de costos fundamentada en actividades (Kaplan y Cooper, 1998).

En el presente estudio es de interés particular el ABM estratégico, en el que se pretende "hacer las actividades correctas" cuando se intenta alterar la demanda de actividades para incrementar la rentabilidad, asumiendo que la eficiencia de la actividad permanece constante. El ABM estratégico modifica la combinación de la demanda de las actividades y, de este modo, las separa de las aplicaciones no rentables mediante la reducción de los factores de costo. Así, por ejemplo, la organización puede determinar si está operando en un punto donde los ingresos producidos por un cliente son menores que los costos de generar tales ingresos; como también refieren Kaplan y Cooper (1998).

El análisis de rentabilidad de clientes (por sus siglas en inglés CPA *Analysis of Customer Profitability*) es una de las aplicaciones típicas de un enfoque de ABM estratégico. Kotler y Keller (2012) mencionan que una vez establecidos los costos, directos e indirectos, variables y fijos, de aquellas actividades relacionadas con la atención de cada cliente, y cuando se identifiquen los clientes menos rentables, la "compañía debería alentarlos a preferir a los competidores". El enfoque ABM es más que un proceso o herramienta de mejora continua porque le facilita a la gerencia adquirir nuevos conocimientos sobre el desempeño y rendimiento organizacional de la administración (CIMA, 2001).

Si el enfoque ABM ayuda a conocer lo que pasa con los clientes de la empresa, el enfoque de valor de vida del cliente (por su denotación en inglés CLV) “describe el valor presente neto del flujo de ganancias que se espera recibir por las futuras compras de un cliente” (Kotler y Keller, 2012). El cálculo del CLV le posibilita a la empresa planificar estrategias de largo plazo, pero depende de un cálculo fiable de ingresos y costos.

Con el crecimiento del poder de cómputo de los años recientes, se incrementan además el registro de datos provenientes de clientes y sus transacciones. Adicionalmente, el número de actividades para generar negocios y los factores de costo asociados aumentan, haciendo que el análisis de la rentabilidad de los clientes presente un alto nivel de granularidad, que lo hace cada vez más complejo. Esto impulsa que la segmentación de clientes en el análisis de rentabilidad sea mandatorio.

Aunque los enfoques de ABM estratégico y de CLV se basan en la segmentación de clientes, utilizando diversos criterios para hacer dicha segmentación (localización geográfica, edad, ingresos, etc.), los mismos son artificiales y no muestran la naturaleza intrínseca en la mezcla de clientes de la empresa.

Por esto, autores como Adkins (2012) plantean el uso de modelos de rentabilidad de clientes para altos volúmenes, que buscan compensar las limitaciones de los modelos basados en CBA. Bajo esta premisa, entre algunas aplicaciones analíticas que utilizan herramientas de minería de datos están:

- Árboles de decisión, para determinar el comportamiento transaccional.
- Regresiones logísticas y redes neuronales, para establecer la ocurrencia de eventos ocasionales en el análisis de rentabilidad, como por ejemplo la emisión de una tarjeta de crédito nueva.

- Clústeres para el análisis de rentabilidad, que permiten la determinación de estrategias en las diferentes áreas de la empresa, una vez encontrados los agrupamientos naturales que presente el análisis.

Buscando emparejar los objetivos del patrocinador de este estudio con el uso de enfoques analíticos y herramientas de minería de datos formulados por Adkins (2012), se aplican a los datos de rentabilidad de clientes varias técnicas de minería de datos, como árboles de decisión, clústeres, K-medias, máquinas de soporte vectorial, redes neuronales, bosques aleatorios, potenciación y series de tiempo, con el fin de encontrar el mejor modelo para esta empresa y así predecir los clientes más rentables.

Muchas de las investigaciones que existen están en función de ventas y de *marketing*, algunas son desarrolladas bajo la tecnología de ERP y CRM, porque la idea principal es comercializar bien un producto o servicio, pocos profundizan más allá con otra visión y objetivo, como es el caso de Haenlein, Kaplan y Beeser (2007), quienes desarrollan un modelo basado en una combinación de la primer cadena de Markov en conjunto con árboles de decisión y regresión lineal.

Los clientes buscan de los proveedores servicios especiales, ideas, innovación y liderazgo intelectual, por ello las empresas están cambiando sus funciones mediante el uso de herramientas de minería de datos e inteligencia empresarial para entender el comportamiento de sus clientes, sus preferencias, hábitos de compra y grupos de afinidad de clientes, posibilitándoles identificar patrones o rasgos diferenciadores que les ayuden a tomar acciones para que los clientes reciban ganancias de sus inversiones y las empresas mejoren sus prontas respuestas en productos y servicios, pero más focalizadas, al centrarse en lo más importante, eliminar la complejidad, descubrir los patrones que estén ocultos y predecir qué características y comportamientos de los

clientes y productos separan a los clientes relativamente más y menos rentables y qué tan propenso puede ser un cliente ahora o en el futuro y así mantener aquellos individuos generadores de mejores rendimientos para la organización (Cokins y Charles, 2011).

Medir la rentabilidad de los clientes es un gran reto para todas las empresas desde hace muchos años. Desde hace medio siglo, no es un cálculo fijo, está en constante cambio porque las necesidades de los clientes, los servicios y productos que ellos requieren varían en el tiempo, implicando una constante revisión del proceso para predecir cuáles son las mejores técnicas de minería de datos que ayudan a prever el comportamiento de los clientes y sus necesidades. En relación con esto, hoy gracias a los avances tecnológicos, el poder computacional posibilita realizar este tipo de análisis en menor tiempo, accediendo a grandes cantidades de datos de diferentes fuentes y en distintos formatos (Sevin, 1965).

Capítulo II Marco teórico

En este capítulo se expone la información relevante utilizada en el desarrollo del proyecto.

Cliente

Según la Real Academia Española (2017), la palabra "cliente" se deriva del latín *clients, entis* y es una persona que compra en una tienda, o que utiliza los servicios de un profesional o empresa. También es una persona que está bajo la protección o tutela de otra.

Las empresas manejan sus conceptos para definir un cliente, algunas hasta difieren dentro de sus propias áreas; por ejemplo dentro de este banco, algunas áreas consideran que un cliente es una persona que tiene un producto activo, otras manejan el mismo concepto, pero además le agregan un valor más, que es tener más información sobre ese cliente, por ejemplo, datos personales tales como dirección, teléfono, ingresos, estado civil.

Rentabilidad de cliente

Kotler y Keller (2012) explican que un cliente rentable es una persona, hogar o empresa que a lo largo del tiempo genera un flujo de ingresos que excede por una cantidad aceptable el flujo de los costos en que incurre la empresa para atraerlo, venderle y atenderlo. El énfasis está hecho en el flujo de ingresos y costos vitalicios del cliente, y no en el beneficio de una transacción específica.

Los especialistas en *marketing* pueden evaluar la rentabilidad del cliente de manera individual, por segmento de mercado o por canal. No obstante, en ocasiones muchas empresas miden la satisfacción del cliente, pero pocas evalúan la rentabilidad individual del mismo.

Los bancos aseguran que esta es una tarea difícil, pues cada cliente utiliza diferentes servicios financieros y sus transacciones quedan registradas en distintos departamentos, lo cual conlleva un problema más complejo si no existe un lugar único para obtener los datos unificados, ya que ocasiona pérdidas para la organización.

Integración de datos

La integración de datos es un conjunto de procedimientos, técnicas y tecnologías usado para diseñar y construir procesos que extraen, reestructuran, mueven y cargan datos en los almacenes de datos operativos o analíticos, ya sea en tiempo real o en *batch* (Giordano, 2010).

PowerData (2017) agrega que es el proceso que combina datos heterogéneos de diversas fuentes, en la forma y estructura de una única aplicación. Este proceso de integración de datos facilita que varios tipos de datos, tales como matrices de datos, documentos y tablas, ver la figura 1, sean fusionados por usuarios, organizaciones y aplicaciones para un uso personal, de procesos de negocio o de funciones.

La integración de datos es muy importante dentro de todas las organizaciones para tener una adecuado gobierno de datos; cuando hay distintos orígenes de datos de fuentes heterogéneas, este proceso es vital para reducir y evitar redundancias e inconsistencias en la información (Han, Kamber y Pei, 2011).

Realizar procesos de integración de datos en minería es muy común para conseguir extraer la información que no es evidente, pero tiene mucha utilidad y es relevante para las organizaciones, permitiéndoles generar nuevas oportunidades de negocio.

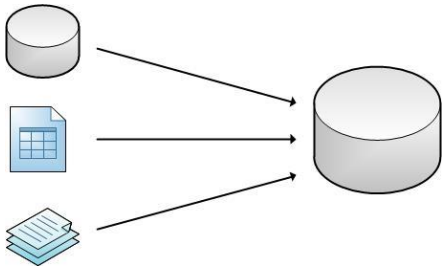


Figura 1. Integración de datos
Fuente: Elaboración propia

Limpieza de datos

La limpieza de datos es un procedimiento que requiere mucho trabajo y tiempo, consiste en tomar los datos de las diversas fuentes aplicando diferentes reglas para intentar completar o rellenar los valores perdidos y suavizar el ruido al identificar valores atípicos y corregir las inconsistencias.

Este paso consume mucho tiempo en una investigación de minería de datos, por lo general toma la mayor parte del proceso y algunas veces provoca que las personas se den por vencidas. Cuando los usuarios tienen conciencia de los problemas de calidad de los datos, es muy probable que no lleguen a confiar en los resultados de cualquier proceso que se haya aplicado (Han et al., 2011).

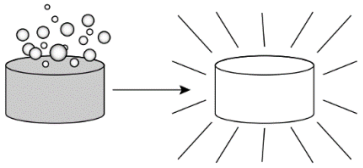


Figura 2. Limpieza de datos
Fuente: Han et al., 2011, p. 87

Calidad de datos

La calidad de los datos se refiere a aquellos que cumplen con las reglas, requisitos o propósitos del negocio; son datos consistentes, no ambiguos, precisos, íntegros, coherentes, que dan credibilidad y, por ende, generan información para la organización y la toma de decisiones.

Thomas (1992) sugiere que la calidad de los datos puede obtenerse comparando dos fuentes de datos: "Un dato o colección de datos ‘X’ es de mayor calidad (mejor) que un dato o colección de datos ‘Y’, si ‘X’ satisface las necesidades del cliente mejor que ‘Y’". Asimismo, define la calidad de los datos como "la capacidad de los datos para ser utilizados de forma económica y rápida para informar y evaluar las decisiones".

Minería de datos

La generación de información es habitual, desde tiempos remotos los seres humanos producen datos e información, no obstante, con los enormes avances tecnológicos se crean numerosos sistemas de información instalados en equipos locales y otros en la nube, todos ellos originan enormes cantidades de datos que en algunas ocasiones son consistentes y en otras no.

Desde que se llega al lugar de trabajo y se registra la marca de entrada, se produce información, pero muchas veces no se conoce el valor de la misma ni qué tratamiento o provecho se les puede dar a esos datos que no son tan visibles y pueden ser esenciales en la obtención de información provechosa para sí mismo u otras personas.

Por ejemplo, Google Trends utiliza términos de búsqueda específicos, como los indicadores de la actividad de la gripe. Gracias a esto encuentra una relación estrecha entre la cantidad de personas que buscan información relacionada con la gripe y la cantidad de personas que realmente tienen síntomas de gripe. Con esa información, utilizando datos agregados de

búsqueda de Google, Flu Trends puede estimar la actividad de la gripe hasta dos semanas más rápido que los sistemas tradicionales. Este es un ejemplo claro que demuestra cómo la minería de datos puede convertir una gran colección de datos en conocimiento que puede ayudar a enfrentar un desafío global actual (Han et al., 2011).

La minería de datos es el resultado de la evolución natural de la tecnología de la información. La base de datos y la industria de gestión de datos evolucionan en el desarrollo de varias funcionalidades críticas, recopilación de datos, creación de bases de datos, gestión de datos (incluido el almacenamiento y recuperación de datos y el procesamiento de transacciones de bases de datos) y análisis de datos avanzados (que involucran *data warehousing* y *data mining*); ver la figura 3. Las herramientas de minería de datos proporcionan clasificación de datos, clústeres, detección de anomalías y caracterización de cambios en los datos a lo largo del tiempo (Han et al., 2011).

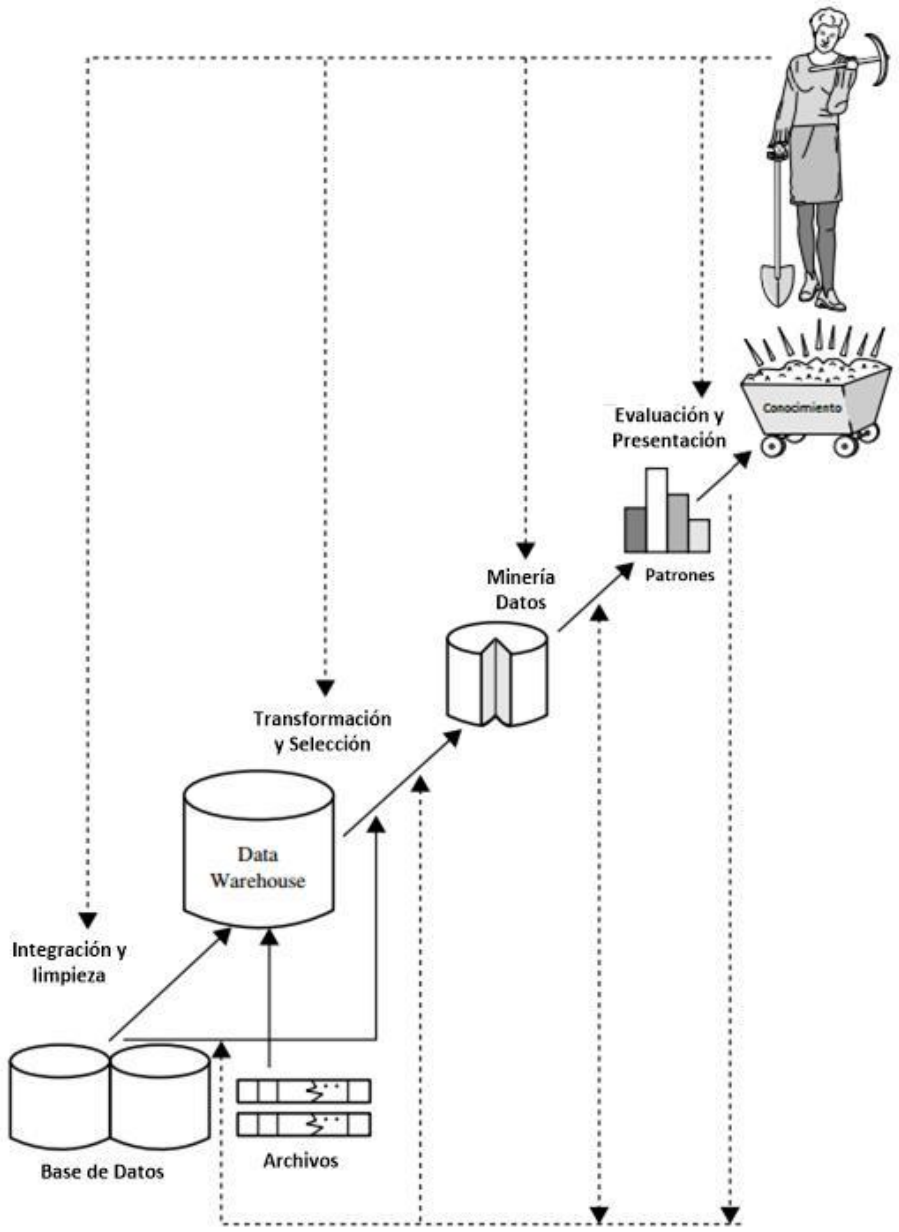


Figura 3. La minería de datos como un paso en el proceso de descubrimiento de conocimiento
Fuente: Han et al., 2011, p. 7

Muchas personas tratan la minería de datos como un sinónimo de otro término utilizado a nivel popular: el descubrimiento de conocimiento a partir de datos, o KDD; mientras que otros consideran la minería de datos como un paso esencial en el proceso de descubrimiento de conocimiento (Han et al., 2011).

La minería de datos es el proceso de descubrir patrones y conocimientos a partir de grandes cantidades de datos. Las fuentes de datos pueden incluir bases de datos, almacenes de datos, la web, otros repositorios de información o datos que se transmiten al sistema de modo dinámico (Han et al., 2011).

La minería de datos se define como el proceso de descubrir patrones ocultos en los datos. El proceso debe ser automático o (con más frecuencia) semiautomático. Los patrones descubiertos deben ser significativos ya que conducen a alguna ventaja, por lo general económica. Los datos están invariablemente presentes en cantidades sustanciales. Los patrones útiles hacen predicciones no triviales sobre nuevos datos. Las personas emplean con frecuencia la minería de datos para obtener conocimiento, no solo predicciones (Witten, Frank y Hall, 2011), el secreto está en poder producirlo y conseguirlo.

La extracción de datos es la extracción de información implícita, desconocida en un principio y potencialmente útil de los datos. La idea es crear programas de computadora que examinen las bases de datos de modo automático, buscando regularidades o patrones. Es probable que los patrones fuertes, si se encuentran, se generalicen para hacer predicciones precisas sobre datos futuros (Witten et al., 2011).

La búsqueda de patrones se realiza siempre porque no solo hace predicciones futuras, sino que también genera nuevos conocimientos; tal y como indican Witten et al. (2011), los cazadores buscan patrones en la conducta de migración de los animales, los agricultores en el crecimiento de los cultivos, los políticos en la opinión de los votantes y los amantes en las respuestas de sus parejas.

El trabajo de un científico es dar sentido a los datos, descubrir los patrones que gobiernan cómo funciona el mundo físico y encapsularlos en teorías que pueden usarse para predecir lo que

va a suceder en situaciones nuevas. La labor del emprendedor es identificar oportunidades, es decir, patrones de comportamiento que pueden convertirse en un negocio rentable, y explotarlos.

Otro ejemplo o caso práctico sobre la usabilidad de la minería de datos se vincula con nuevas oportunidades de negocio para el cliente y el crecimiento de la empresa. Según Witten et al. (2011), mediante el análisis de datos ya presentes en las bases de datos, se busca resolver problemas, como el de la lealtad del cliente en un mercado altamente competitivo.

Una base de datos de opciones del cliente, junto con los perfiles de los clientes, es la clave de este problema. Los patrones de comportamiento de los clientes anteriores pueden analizarse para identificar las características distintivas de aquellos que probablemente cambien de productos y aquellos que sigan siendo leales. El primer grupo puede ser objeto de un tratamiento especial, demasiado costoso para aplicarlo a la base de clientes en general.

De manera más positiva, las mismas técnicas se pueden utilizar para determinar a los clientes que pueden sentirse atraídos por otro servicio que la empresa ofrece, para dirigirse a ellos con ofertas especiales que promocionen este servicio. En la economía actual, muy competitiva, centrada en el cliente y orientada al servicio, los datos son la materia prima que impulsa el crecimiento del negocio, si se logra explotar.

Sintetizando lo expuesto, en la actualidad se produce gran cantidad de información -algunas veces de forma inconsciente-, desde datos estructurados hasta semiestructurados y no estructurados; estos dos últimos requieren un mayor tratamiento, pero contienen una enorme riqueza de información dentro de ellos.

Ahora bien, gracias a los grandes avances tecnológicos y el poder computacional que hoy existe, se pueden tratar estos datos en menor tiempo y con menor costo, mediante la aplicación de técnicas de minería de datos que facilitan descubrir reglas, relaciones, asociaciones, correlaciones

y/o dependencias en una gran cantidad de datos, gracias a estadísticas, matemáticas y reconocimiento de patrones, generando herramientas que apoyan la toma de decisiones de las personas dentro de una empresa de forma más apropiada y teniendo claro que no sustituye las tareas que requieren la intervención humana, pero que el fin es transformar datos en información. La minería de datos es casi como una nueva ciencia, algunos la denominan como ciencia de datos y requiere de la interacción de diferentes conocimientos y disciplinas, tal y como se aprecia en la figura 4.



Figura 4. Minería de datos, confluencia de múltiples disciplinas
Fuente: Elaboración propia

Machine learning

Machine learning es un área de la inteligencia artificial (AI) que trata sobre cómo escribir programas que puedan aprender. De acuerdo con Clarke, Fokoue y Zhang (2009), el aprendizaje automático se refiere al uso de estructuras formales (máquinas) para hacer inferencia (aprendizaje). Esto incluye lo que los científicos empíricos quieren decir con la construcción de modelos -

proponer expresiones matemáticas que encapsulan el mecanismo por el cual un proceso físico da lugar a observaciones-, pero también mucho más.

En particular, incluye muchas técnicas que no se corresponden con el modelado físico, siempre que procesen datos en información. Aquí, la información por lo general significa cualquier aspecto que ayude a reducir la incertidumbre. Entonces, por ejemplo, una distribución posterior representa "información" o es un "aprendiz" porque reduce la incertidumbre sobre un parámetro.

Otra definición es de Han et al. (2011), quienes señalan que el aprendizaje automático investiga cómo las computadoras pueden aprender (o mejorar su rendimiento) en función de los datos. Un área de investigación principal es que los programas de computadora aprendan automáticamente a reconocer patrones complejos y tomar decisiones inteligentes basadas en datos. Por ejemplo, un problema atípico de aprendizaje automático es programar una computadora para que pueda reconocer de forma automática los códigos postales escritos a mano en el correo después de aprender de un conjunto de ejemplos. El aprendizaje de la máquina es una disciplina de rápido crecimiento.

Witten et al. (2011) ejemplifican el uso y aplicación de este tipo de técnicas en la fertilización humana *in vitro,* la cual consiste en recolectar varios óvulos de los ovarios de una mujer que, después de la fertilización con esperma asociado o donante, producen varios embriones. Algunos de estos son seleccionados y transferidos al útero de la mujer. El desafío es seleccionar los "mejores" embriones para usar los que tienen más posibilidades de sobrevivir.

La selección se basa en alrededor de 60 características registradas de los embriones, que caracterizan su morfología, ovocito y folículo, y la muestra de esperma. El número de funciones es lo bastante grande como para dificultar que un embriólogo las evalúe todas al mismo tiempo y correlacione los datos históricos con el resultado crucial de si ese embrión da como resultado o no

un niño vivo. En un proyecto de investigación en Inglaterra, el aprendizaje automático investiga una técnica para hacer la selección, utilizando registros históricos de embriones y sus resultados como datos de capacitación.

Según Han et al. (2011), el aprendizaje automático se puede clasificar en cuatro tipos: el supervisado, no supervisado, semisupervisado y activo:

El **aprendizaje supervisado** es un sinónimo de clasificación. La supervisión en el aprendizaje proviene de los ejemplos etiquetados en el conjunto de datos de entrenamiento. Por ejemplo, en el problema del reconocimiento del código postal, se utiliza un conjunto de imágenes de códigos postales escritos a mano y sus correspondientes traducciones legibles por máquina como ejemplos de entrenamiento, que supervisan el aprendizaje del modelo de clasificación.

El **aprendizaje no supervisado** es un sinónimo de agrupamiento. El proceso de aprendizaje no está supervisado ya que los ejemplos de entrada no están etiquetados por clase. Por lo general, se puede usar el agrupamiento para descubrir clases dentro de los datos. Por ejemplo, un método de aprendizaje no supervisado puede tomar, como entrada, un conjunto de imágenes de dígitos escritos a mano. Si se encuentran, por ejemplo, 10 grupos de datos, estos clústeres pueden corresponder a los 10 dígitos distintos de 0 a 9, respectivamente. Sin embargo, debido a que los datos de entrenamiento no están etiquetados, el modelo aprendido no puede decir el significado semántico de los grupos encontrados.

El **aprendizaje semisupervisado** es una clase de técnicas de aprendizaje automático que utilizan ejemplos etiquetados y no etiquetados cuando se aprende un modelo. En un enfoque, los ejemplos etiquetados se usan para aprender modelos de clase y los ejemplos no etiquetados se usan para definir los límites entre las clases. Para un problema de dos clases, se puede pensar en el

conjunto de ejemplos que pertenecen a una clase como ejemplos positivos y aquellos que pertenecen a la otra clase como ejemplos negativos.

En la figura 5, si no se consideran los ejemplos no etiquetados, la línea punteada es el límite de decisión que mejor divide los ejemplos positivos de los ejemplos negativos. Usando los ejemplos sin etiqueta, se puede refinar el límite de decisión a la línea continua. Además, se puede detectar que los dos ejemplos positivos en la esquina superior derecha, aunque están etiquetados, son probablemente ruido o valores atípicos.

El **aprendizaje activo** es un enfoque de aprendizaje automático que permite a los usuarios desempeñar un papel activo en el proceso de aprendizaje. Un enfoque de aprendizaje activo puede pedirle a un usuario (por ejemplo, un experto de dominio) que marque un ejemplo, que puede ser de un conjunto de ejemplos no etiquetados o sintetizados por el programa de aprendizaje. El objetivo es optimizar la calidad del modelo mediante la adquisición activa de conocimiento de los usuarios humanos, dada la limitación de la cantidad de ejemplos que se les puede pedir etiquetar.

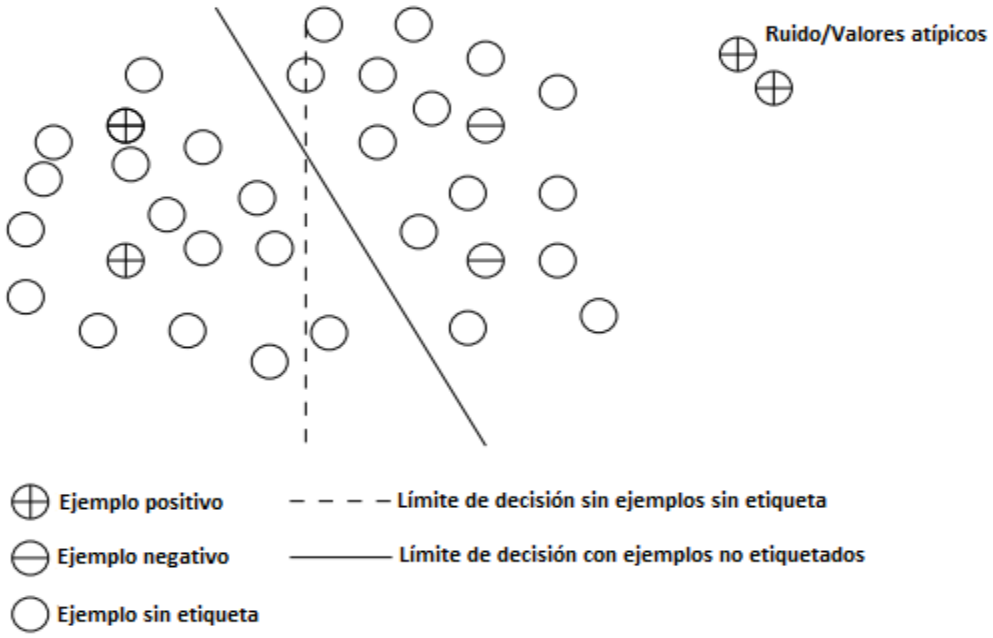


Figura 5. Aprendizaje semisupervisado
Fuente: Han et al., 2011, p. 25

Evaluación y selección de modelos

Existen varias métricas de evaluación para la precisión predictiva de un clasificador. A efectos de este trabajo, se elige el método de validación cruzada para evaluar y seleccionar el modelo o algoritmo más óptimo que resuelva la necesidad del caso en estudio. Sin embargo, se explica cada una de las técnicas que se pueden utilizar. Por lo general, para hacer la evaluación se emplean de forma aleatoria datos de entrenamiento y prueba, ambos para hacer la estimación de los modelos, con el objetivo de encontrar el de mejor precisión.

Métricas para evaluar el rendimiento del clasificador.

Han et al. (2011) citan cuatro términos que se utilizan para calcular medidas de evaluación, que ayudan a comprender el significado de las diversas medidas:

- **Verdaderos positivos** (TP): Se refieren a las tuplas positivas que son correctamente etiquetadas por el clasificador.
- **Verdaderos negativos** (TN): Son las tuplas negativas que el clasificador etiqueta correctamente.
- **Falsos positivos** (FP): Son las tuplas negativas que se etiquetan incorrectamente como positivas.
- **Falsos negativos** (FN): Son las tuplas positivas que son mal etiquetadas como negativas.

A continuación se muestran ejemplos con el fin de entender cómo se evalúa la calidad de un modelo a partir de estas medidas, sin embargo estos índices pueden ser engañosos si no se analiza bien la relación de las clases por evaluar.

- **La matriz de confusión o matriz de error**. Esta contiene información acerca de las predicciones hechas por un método o sistema de clasificación, comparando para el conjunto de individuos de la tabla de aprendizaje o de *testing*, la predicción dada versus la clase a la que estos en realidad pertenecen. La suma se denota por a+b (tabla 5).

  En la tabla 6 hay un ejemplo de la matriz de confusión aplicado a un caso que busca predecir si los correos recibidos se pueden considerar como *spam* o no.

  Continuando con la importancia de tener claridad en la relación de las clases, se puede ver el ejemplo de la detección de fraude en seguros, en este caso se quiere determinar si un seguro es fraude o no. Entonces para este ejemplo se puede suponer que se tiene una serie de datos y el cálculo en la matriz de confusión da como resultado una precisión de un 94.09 %, lo cual parece un excelente modelo, pero es engañoso porque no detecta ningún fraude y esto es lo que se busca.

  Por lo tanto, los valores con que se cuenta no son los más óptimos y necesitan de un trato especial para ser calibrados y encontrar los casos de fraude que ayuden a darle más validez al modelo. En la tabla 7 se puede observar el resultado del ejemplo.

- La **precisión** (P) de un modelo de predicción es la proporción del número total de predicciones que son correctas respecto al total. Se determina utilizando la ecuación: $P = (a+d)/(a+b+c+d)$.

- La **precisión positiva (PP)** es la proporción de casos positivos que son identificados correctamente, tal como se calcula usando la ecuación: $PP = d/(c+d)$.

- La **precisión negativa (PN)** es la proporción de casos negativos que son identificados correctamente; se calcula empleando la ecuación: PN = a/(a+b).
- **Falsos positivos (FP)** es la proporción de casos negativos que son clasificados incorrectamente como positivos; se calcula utilizando la ecuación: FP = b/(a+b).
- **Falsos negativos (FN)** es la proporción de casos positivos que son clasificados incorrectamente como negativos; se calcula empleando la ecuación: FN = c/(c+d).

Tabla 5. Matriz de confusión

		Predicción	
		Negativo	Positivo
Valor real	Negativo	a	b
	Positivo	c	d

Fuente: Elaboración propia

Tabla 6. Matriz de confusión, determina si los correos son *spam* o no

		Predicción		
		email	*spam*	Valor total x Valor real
Valor real	email	**813**	19	832
	spam	40	**508**	548
	Valor total x Predicción	853	527	

Fuente: Elaboración propia

Tabla 7. Matriz de confusión, determina si es fraude o no

		Predicción		
		Fraude	No fraude	Valor total x valor real
Valor real	Fraude	**0**	19	19
	No fraude	3	**350**	353
	Valor total x Predicción	3	369	

Fuente: Elaboración propia

En la tabla 8 se puede encontrar un resumen de las fórmulas para evaluar modelos.

Tabla 8. Fórmulas para evaluar modelos

Nombre	Ecuación
Precisión P	P = (a+d)/(a+b+c+d)
Precisión positiva (PP)	PP = d/(c+d)
Precisión negativa (PN)	PN = a/(a+b)
Falsos positivos (FP)	FP = b/(a+b)
Falsos negativos (FN)	FN = c/(c+d)
Asertividad positiva (AP)	AP = d/(b+d)
Asertividad negativa (AN)	AN = a/(a+c)

Fuente: Elaboración propia

En la tabla 9 se puede observar un resumen de las fórmulas, así como su aplicación, en el ejemplo donde se busca detectar si los correos son *spam* o no.

Tabla 9. Ejemplo completo para determinar si los correos son *spam* o no

Nombre	Ecuación	Resultado
Precisión P	P = (a+d)/(a+b+c+d)	95.72 %
Precisión positiva (PP)	PP = d/(c+d)	92.70 %
Precisión negativa (PN)	PN = a/(a+b)	97.72 %
Falsos positivos (FP)	FP = b/(a+b)	2.28 %
Falsos negativos (FN)	FN = c/(c+d)	7.30 %
Asertividad positiva (AP)	AP = d/(b+d)	96.39 %
Asertividad negativa (AN)	AN = a/(a+c)	95.31 %

Fuente: Elaboración propia

Método de retención y submuestreo aleatorio.

Han et al. (2011) explican que en este método los datos dados se dividen de manera aleatoria en dos conjuntos independientes, un conjunto de entrenamiento y un conjunto de prueba. Normalmente, dos tercios de los datos se asignan al conjunto de entrenamiento, y el tercio restante

al conjunto de prueba. El conjunto de entrenamiento se usa para derivar el modelo. La precisión del modelo luego se estima con el conjunto de prueba (figura 6). La estimación es pesimista porque solo una parte de los datos iniciales se utiliza para derivar el modelo.

El submuestreo aleatorio es una variación del método de retención en el que este último método se repite k veces. La estimación de precisión global se toma como el promedio de las exactitudes obtenidas de cada iteración.

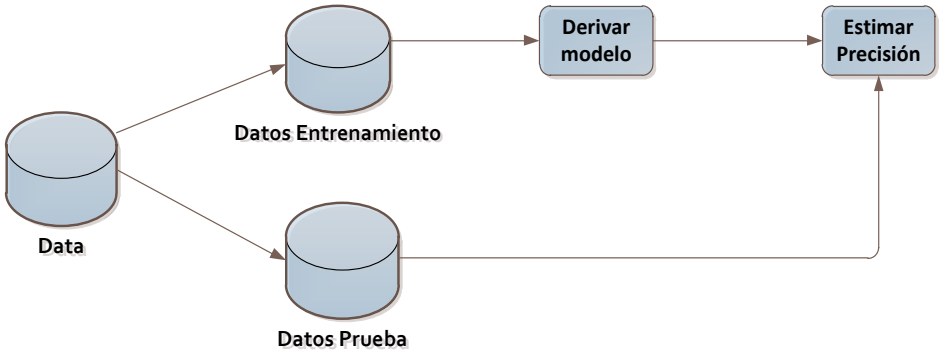


Figura 6. Estimación de precisión con el método retención
Fuente: Han et al., 2011, p. 370

Bootstrap.

El bootstrap es inventado por Efron en 1979 y es uno de los primeros y más poderosos logros de la inferencia estadística intensiva en computación. Rápidamente se convierte en un método para establecer regiones de confianza aproximadas en las estimaciones cuando se desconoce la distribución subyacente (Clarke et al., 2009).

A diferencia de los métodos de estimación de precisión que se acaban de mencionar, el método bootstrap muestrea las tuplas de entrenamiento proporcionadas de forma uniforme con el reemplazo. Es decir, cada vez que se selecciona una tupla, es muy probable que se seleccione de nuevo y se vuelva a agregar al conjunto de entrenamiento. Por ejemplo, si se tiene una máquina

que elige tuplas al azar para un conjunto de entrenamiento. En el muestreo con reemplazo, la máquina puede seleccionar la misma tupla más de una vez en la muestra (Han et al., 2011).

Validación cruzada.

Según Clarke et al. (2009), la validación cruzada es inventada por Stone (1959) en el contexto de la regresión lineal. Pretende equilibrar el beneficio de utilizar la mayor cantidad de datos posible para construir el modelo contra el falso optimismo creado cuando los modelos se prueban con los mismos datos que se usan para construirlos. Existen varias versiones, está el algoritmo Leave-One-Out Cross-Validation y K-Fold, este último es el más común.

El proceso consiste en dividir la muestra de datos en subconjuntos para realizarles un análisis. Esos subconjuntos se dividen en datos de prueba (*testing*) y de entrenamiento (*training*). Bien lo indican Witten et al. (2011) cuando explican que se decide sobre un número fijo de pliegues o particiones de los datos. Por ejemplo, si se utilizan tres, los datos se dividen en tres particiones aproximadamente iguales; cada uno a su vez se usa para las pruebas y el resto para el entrenamiento. En otras palabras, se emplean dos tercios de los datos para el entrenamiento y un tercio para las pruebas, y se repite el procedimiento tres veces para que al final cada instancia se haya utilizado una vez para la prueba.

Muchos expertos recomiendan como estándar para predecir la tasa de error de una técnica de aprendizaje dada una muestra de datos única y fija, el uso de la validación cruzada multiplicada por 10 veces (Clarke et al., 2009; Han et al., 2011).

Los datos se dividen de modo aleatorio en 10 partes en las que la clase se representa en aproximadamente las mismas proporciones que en el conjunto de datos completo. Cada parte se ofrece por turno y el plan de aprendizaje se entrena en las nueve décimas restantes; entonces su

tasa de error se calcula en el conjunto de retención. Por lo tanto, el procedimiento de aprendizaje se ejecuta un total de 10 veces en varios conjuntos de entrenamiento (cada conjunto tiene mucho en común con los demás).

Por último, las 10 estimaciones de error se promedian para producir una estimación de error global (Han et al., 2011). Sin embargo, se debe tener claro que no hay nada mágico en el uso del número exacto 10: la validación cruzada de cinco o 20 veces es casi tan buena como la de 10 (Witten et al., 2011).

Leave-One-Out Cross-Validation.

Dejando uno por fuera, por sus siglas y nombre en inglés *Leave-one-out* (LOOCV), es un caso especial de k-fold cross-validation, donde k se establece en el número de tuplas iniciales. Es decir, solo se omite una muestra a la vez para el conjunto de prueba (Han et al., 2011). LOOCV implica el uso de una única observación de la muestra original como el conjunto de datos de validación, y las observaciones restantes como el conjunto de entrenamiento (Clarke et al., 2009).

En primer lugar, se utiliza la mayor cantidad posible de datos para el entrenamiento en cada caso, lo que presumiblemente aumenta las posibilidades de que el clasificador sea preciso. En segundo lugar, el procedimiento es determinista: no se trata de un muestreo aleatorio. No tiene sentido repetirlo 10 veces o repetirlo: se obtiene el mismo resultado cada vez. Frente a esto se encuentra el alto costo computacional, ya que todo el proceso de aprendizaje debe ejecutarse n veces, y esto por lo general no es factible para grandes conjuntos de datos. Sin embargo, parece ofrecer la posibilidad de sacar el máximo de un pequeño conjunto de datos y obtener una estimación lo más precisa posible (Witten et al., 2011).

Esto significa que el uso de LOOCV implica realizar una separación de los datos de tal manera que para cada iteración solo se tenga un dato de prueba y todo el resto de los datos sean

utilizados para entrenamiento, asimismo el error se calcula obteniendo el promedio de todos los errores cometidos.

Algunas ventajas de este algoritmo son que la estimación del error no es muy variable dependiendo de cuáles datos hayan quedado en la tabla de aprendizaje y de prueba, permitiendo obtener una estimación del error más estable y además el error no se tiende a sobrestimar como sucede en otros métodos donde el error en la tabla de entrenamiento es mayor que cuando se usa toda la tabla de datos.

Las desventajas primordiales están en el tiempo de procesamiento computacional, al respecto Clarke et al. (2009) mencionan que suele ser costoso, pues todo el proceso requiere ajustar el modelo n veces, o sea la cantidad de repeticiones *n* que se debe ejecutar el modelo repercute máxime si es una cantidad enorme de datos.

K-Fold Cross-Validation.

En la validación cruzada k veces, los datos iniciales se dividen al azar en k subconjuntos mutuamente excluyentes o "pliegues", D1, D2, …, D_k, cada uno de similar tamaño. El entrenamiento y las pruebas se realizan k veces (Han et al., 2011).

Aquí las observaciones se eliminan en grupos. Por ejemplo, si la muestra original está dividida en K grupos de aproximadamente el mismo tamaño, para cada k = 1,…, K, uno se ajusta al modelo utilizando todos los datos y luego se calcula el error de predicción del modelo ajustado empleando los datos en el grupo. Este proceso se repite K veces, con cada grupo utilizado una vez como conjunto de validación. Los errores de predicción K se promedian para producir una única estimación (Clarke et al., 2009).

En otras palabras, aquí los datos se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento, se repite durante K iteraciones, con cada uno de los posibles subconjuntos de datos de prueba y por último el error se calcula como la media aritmética de los errores de cada iteración para obtener un único resultado.

Árboles de decisión

La estructura del árbol se utiliza en muchos campos diferentes, como la medicina, la lógica, la resolución de problemas y la ciencia administrativa. También es una estructura informática tradicional para organizar datos. Un árbol de decisión usa esta estructura tradicional. Por lo general se presenta el árbol boca abajo, con la raíz en la parte superior y las hojas en la parte inferior. Comenzando desde la raíz, el árbol se divide del tronco único en dos o más ramas. Cada rama en sí misma puede dividirse en dos o más ramas, lo que lleva a otros nodos. Esto continúa hasta alcanzarse una hoja, que es un nodo que no está más dividido. Se hace referencia a la división de una rama como un nodo del árbol. La raíz y las hojas asimismo se conocen como nodos. Los nodos de hoja contienen las "decisiones" (Graham, 2011).

Su atractivo radica en la simplicidad del modelo resultante, donde un árbol de decisiones (al menos uno que no sea demasiado grande) es bastante fácil de ver, comprender y, lo que es más importante, explicar. Los árboles de decisión no siempre ofrecen el mejor rendimiento, y representan una compensación entre el rendimiento y la simplicidad de la explicación. La estructura del árbol de decisiones puede representar modelos de clasificación y regresión (Graham, 2011).

El árbol de decisiones es una representación del conocimiento muy conveniente y eficiente. En general, los modelos expresados en un idioma se pueden traducir a otro idioma y lo mismo

ocurre con un árbol de decisiones. Una traducción simple y útil es un conjunto de reglas (Graham, 2011).

K-means

Según Witten et al. (2011), K-means es la técnica clásica de agrupamiento. Clarke et al. (2009) señalan que es la técnica de agrupamiento más común, la cual se basa en centroide y es inventada por MacQueen en 1967. Conceptualmente, el centroide de un clúster es su punto central (Han et al., 2011). Se fundamenta en asignar puntos a cualquier centroide que esté más cerca de ellos, volver a calcular el centroide y repetir el procedimiento, esperando que la agrupación converja. Esto utiliza una desemejanza, por lo tanto, puede considerarse jerárquico; sin embargo, también comienza con una asignación inicial de la agrupación, que va refinando; por lo que el agrupamiento inicial es más relevante para el procedimiento.

Primero, se debe especificar cuántos clústeres se buscan: este es el parámetro k. Luego, se eligen k puntos al azar como centros de agrupamiento. Todas las instancias se asignan a su centro de clúster más cercano de acuerdo con la métrica de distancia euclidiana ordinaria. Luego, se calcula el centroide de las instancias en cada grupo: este es el punto "medio". Estos centroides se toman como nuevos valores centrales para sus respectivos grupos. Por último, todo el proceso se repite con los nuevos centros de clúster. La iteración continúa hasta que se asignan los mismos puntos a cada grupo en rondas consecutivas, en cuya etapa los centros de clúster se estabilizan y se mantienen iguales para siempre (Witten et al., 2011).

Existen numerosas variaciones en el tema K-means, con frecuencia basadas en cambiar la desemejanza o el centrado. La característica común es calcular un conjunto de estadísticas de

resumen para ubicaciones y quizás una matriz de varianza. Dos de ellas son K-medias y K- medoid (Clarke et al., 2009).

K-medias.

En este caso, el error al cuadrado se reemplaza por el error absoluto, y la media se reemplaza con un objeto similar a la mediana que surge de la optimización (porque hay cierta ambigüedad acerca de cómo definir una mediana en dos o más dimensiones). La característica definitoria es que los centros de clúster en este caso minimizan la distancia L^1 global a cada punto. La matriz de covarianza a menudo se ignora porque no escala el error absoluto de los vecinos más cercanos de la misma forma que escala el error al cuadrático. Existe evidencia que este procedimiento es más resistente a los valores atípicos. No obstante, al igual que todos los métodos basados en centroide, funciona mejor cuando los conglomerados son convexos (Clarke et al., 2009).

En el algoritmo de K-medias, se observa que al elegir los centros iniciales con cuidado, se puede ser capaz de acelerar no solo la convergencia del algoritmo, sino además garantizar la calidad al final. El algoritmo k-means++ es una variante de K-means, que elige los centros iniciales de la siguiente manera. Primero, selecciona un centro uniformemente al azar de los objetos en el conjunto de datos. Iterativamente, para cada objeto $\boldsymbol{p}$ que no sea el centro elegido, elige un objeto como el nuevo centro. Este objeto se elige al azar con una probabilidad proporcional a $\text{dist}(\boldsymbol{p})^2$, donde $\text{dist}(\boldsymbol{p})$ es la distancia desde $\boldsymbol{p}$ hasta el centro más cercano electo. Así la iteración continúa hasta que se seleccionen k centros (Han et al., 2011).

K-medoides.

El algoritmo K-medoid, al igual que el K-means, intenta minimizar un criterio de error cuadrático, pero más bien en lugar de tomar una media, el centro del clúster elige uno de los puntos de datos. El medoid de un clúster es el elemento del clúster cuya diferencia de promedio con todos los objetos del clúster es mínima. Es el punto de datos prototípicos del clúster porque es el más céntrico.

De nuevo, se debe dar K y un agrupamiento inicial, y el procedimiento básico es similar al de K-medias, pero iterativamente reemplaza cada uno de los medoides por uno de los puntos no medoides para buscar un error más pequeño. Es decir, después de la selección de los puntos medoides de K, cada punto de datos está asociado con su medoide más cercano. Luego se selecciona al azar un punto de datos no medoide y se usa para reemplazar uno de los medoides con el fin de establecer si se reduce el error. Si es así, los dos puntos se intercambian; si no, se prueba otro punto no medoide. Esto se repite hasta que no haya cambios en el medoide. Ahora bien, esto no se adapta bien a grandes conjuntos de datos, pero es más robusto al ruido y valores atípicos que K-means (Clarke et al., 2009).

Máquinas de soporte vectorial (SVM)

Una máquina de soporte vectorial (SVM) busca los llamados vectores de soporte que son observaciones ubicadas en el borde de un área en el espacio que presenta un límite entre una de estas clases de observaciones (por ejemplo, los cuadrados) y otra clase de observaciones (por ejemplo, los círculos). En la terminología de SVM, se habla del espacio entre estas dos regiones como el margen entre las clases. Cada región contiene observaciones con el mismo valor para la variable objetivo (o sea, la clase).

Los vectores de soporte, y solo los vectores de soporte, se usan para identificar un hiperplano (una línea recta en dos dimensiones) que separa las clases. Se busca el margen máximo entre las clases separables. Esto luego representa el modelo. Por lo general, es bastante raro que se puedan separar los datos con una línea recta (o un hiperplano cuando se tienen más de dos variables de entrada). En otras palabras, los datos con frecuencia no se distribuyen de tal forma que sean linealmente separables. Cuando este es el caso, se usa una técnica para combinar (o reasignar) los datos de distintas maneras, creando nuevas variables para que las clases sean más fáciles de separar linealmente por un hiperplano (o sea, con los nuevos datos dimensionales hay una brecha entre las observaciones en las dos clases) (Graham, 2011).

Las máquinas de vectores de soporte funcionan bien en problemas que son no lineales, dispersos y de gran dimensión. Una desventaja es que el algoritmo es sensible a la opción de ajuste (por ejemplo, el tipo de transformaciones que se deben hacer), lo cual hace que sea más difícil de usar y requiere mucho tiempo para identificar el mejor modelo. Otra desventaja es que las transformaciones realizadas pueden ser computacionalmente caras y se llevan a cabo tanto al construir el modelo como al anotar nuevos datos (Graham, 2011).

Una ventaja del método es que el modelado solo trata con estos vectores de soporte en lugar de con el conjunto de datos de entrenamiento, por lo que el tamaño del conjunto de entrenamiento no suele ser un problema. Además, como consecuencia del uso exclusivo de los vectores de soporte para construir un modelo, el modelo se ve menos afectado por valores atípicos (Graham, 2011).

Redes neurales (NN)

El campo de las redes neuronales es iniciado por psicólogos y neurobiólogos que buscan desarrollar y probar análogos computacionales de neuronas. En términos generales, una red

neuronal es un conjunto de unidades de entrada/salida conectadas en el que cada conexión tiene un peso asociado. Durante la fase de aprendizaje, la red aprende ajustando los pesos para predecir la etiqueta de clase correcta de las tuplas de entrada. El aprendizaje de redes neuronales también se conoce como aprendizaje conexionista debido a las conexiones entre unidades (Han et al., 2011).

En un principio las redes neuronales matemáticas tienen la intención de modelar el aprendizaje y el reconocimiento de patrones realizados por las neuronas fisiológicas. Se dice en el modelo de Hebb que una sinapsis es un enlace desde la salida de un nodo a la entrada de otro nodo, con una ponderación relacionada con la correlación en su actividad. Por lo tanto, los grupos de neuronas pueden vincularse y modelarse como la activación de tales ensamblajes.

Otros pioneros como Rosenblatt continúan con el modelo de Hebb, centrándose en cómo pueden desarrollarse los enlaces de neurona a neurona; en particular, propone el modelo matemático básico que todavía se usa para redes neuronales artificiales (NNs). Su unidad básica se llama perceptrón (ahora denominado "nodo"), que al recibir una señal responde o no, dependiendo de si una función excede un umbral. Aunque este modelo no es considerado aplicable por los neurofisiólogos desde la década de 1970, la estructura matemática sigue siendo de gran importancia en los contextos de clasificación y regresión (Clarke et al., 2009).

Las redes neuronales implican largos tiempos de entrenamiento y, por lo tanto, son más adecuadas para aplicaciones donde esto es factible. Requieren una serie de parámetros que por lo general se determinan mejor de manera empírica, como la topología de red o la "estructura". Las redes neuronales son criticadas por su escasa capacidad de interpretación. Por ejemplo, es difícil para los humanos interpretar el significado simbólico detrás de los pesos aprendidos y de las

"unidades ocultas" en la red. Estas características en un inicio hacen que las redes neuronales sean menos deseables para la minería de datos (Han et al., 2011).

Las ventajas de las redes neuronales, sin embargo, incluyen su alta tolerancia a los datos ruidosos y su capacidad para clasificar los patrones en los que no son entrenados. Se pueden usar cuando hay poco conocimiento de las relaciones entre los atributos y las clases. Son adecuadas para entradas y salidas de valor continuo, a diferencia de la mayoría de los algoritmos de árbol de decisión. Tienen éxito en una amplia gama de datos del mundo real, incluido el reconocimiento de caracteres manuscritos, patología y medicina de laboratorio, y en el entrenamiento de una computadora para pronunciar el texto en inglés.

Los algoritmos de red neuronal son inherentemente paralelos; las técnicas de paralelización se pueden usar para acelerar el proceso de cálculo. Además, varias técnicas son de reciente desarrollo para la extracción de reglas de redes neuronales capacitadas. Estos factores contribuyen a la utilidad de las redes neuronales para la clasificación y predicción numérica en la extracción de datos (Han et al., 2011).

Bosques aleatorios (Random Forests)

La construcción de un árbol de decisión único proporciona un modelo simple del mundo, pero a menudo es demasiado simple, o bien, específico. Durante muchos años de experiencia en minería de datos, se determina que muchos modelos que trabajan juntos son mejores que un modelo que lo hace todo. Por esta razón surge la idea de combinar múltiples modelos (como árboles de decisión) en un solo conjunto de modelos (para construir un bosque de árboles) (Graham, 2011). Lo anterior da lugar a tomar mejores decisiones en comparación con los modelos que tienen una única experiencia.

De acuerdo con Graham (2011), la idea de construir múltiples árboles surge desde el principio con el desarrollo del algoritmo de aprendizaje inductivo múltiple (*Multiple Inductive Learning* MIL) propuesto por Williams en 1987. Cuando se construye un árbol de decisión único, se observa que a menudo hay muy poca diferencia al elegir entre variables alternativas. Por ejemplo, dos o más variables pueden no ser distinguibles en términos de su capacidad para dividir los datos en conjuntos de datos más homogéneos. El algoritmo MIL construye todos los modelos buenos de forma igual y luego los combina en uno solo, obteniendo un mejor modelo general.

Potenciación (Boosting)

El algoritmo Boosting es un enfoque eficiente, simple y fácil de usar para construir modelos. La popular variante llamada AdaBoost (una abreviatura para refuerzo adaptativo) se describe como el mejor clasificador comercial del mundo (atribuido a Leo Breiman por Hastie en el año 2001) (Graham, 2011).

Los algoritmos Boosting construyen múltiples modelos a partir de un conjunto de datos utilizando algún otro algoritmo de aprendizaje que particularmente no necesita ser un buen aprendiz. Boosting asocia ponderaciones con observaciones en el conjunto de datos y aumenta (impulsa) los pesos para aquellas observaciones que son difíciles de modelar con precisión. Se construye una secuencia de modelos, y después que se construye cada modelo, los pesos se modifican para dar más peso a aquellas observaciones que son más difíciles de clasificar.

De hecho, los pesos de tales observaciones por lo general oscilan hacia arriba y hacia abajo de un modelo al siguiente. El modelo final es entonces un modelo aditivo construido a partir de la secuencia de modelos, cada producto del modelo está ponderado por un puntaje. Se requiere poco ajuste y se asume poco sobre el aprendiz utilizado, excepto que debe ser un aprendiz débil. La

potenciación puede no funcionar si no hay datos suficientes o si los modelos débiles son demasiado complejos. También es susceptible al ruido (Graham, 2011).

Se puede decir que son similares a los bosques aleatorios en que se construye un conjunto de modelos y luego se combinan para ofrecer un modelo mejor que cualquiera de los modelos constituyentes. La característica distintiva básica del enfoque de Boosting es que los árboles se construyen uno tras otro y el refinamiento se basa en los modelos construidos con anterioridad. El concepto es que después de construir un modelo, se potencian las observaciones que están clasificadas de modo incorrecto por ese modelo. Una observación impulsada recibe esencialmente más importancia en el conjunto de datos, lo cual hace que la única observación esté sobrerrepresentada. Esto tiene el efecto de que es más probable que el próximo modelo clasifique de manera correcta esa observación. Si no, entonces esa observación se incrementa de nuevo (Graham, 2011).

El algoritmo de Boosting se presenta con frecuencia en términos de árboles de decisión como su forma primaria para la representación del conocimiento. La extensión clave para la representación del conocimiento se encuentra en la manera en que se combinan las decisiones que toman los "expertos" o modelos individuales. Para potenciar, se utiliza una puntuación ponderada, y cada uno de los modelos del conjunto tiene un peso correspondiente a la calidad de su experiencia (por ejemplo, la precisión medida de cada árbol) (Graham, 2011).

Capítulo III Marco metodológico

Tipo de investigación

El desarrollo de la investigación está orientado a la modalidad de proyecto aplicado, donde se emplean los conocimientos adquiridos a la realidad de la empresa para resolver de forma expedita los problemas que enfrenta.

Este estudio pretende evaluar los distintos algoritmos de minería de datos con el propósito de seleccionar el más apropiado para la predicción de la rentabilidad de los clientes de la institución y así generarles beneficios tangibles a las áreas solicitantes de la información.

Alcance investigativo

La investigación realizada es de un alcance exploratorio, dado que se busca "examinar un tema o problema de investigación poco estudiado" (Hernández, Fernández y Baptista, 2004). Esto consiste en incursionar en el uso de técnicas de minería de datos, mediante la utilización de algoritmos que permitan predecir la rentabilidad de los clientes del Banco de Costa Rica para apoyar la toma de decisiones.

Este tipo de investigación posibilita "obtener información sobre la posibilidad de llevar a cabo una investigación más completa sobre un contexto particular (…), identificar conceptos o variables promisorios" (Hernández et al., 2004) y cumplir los objetivos planteados.

Enfoque

En este proyecto de investigación, a la hora de definir el enfoque, se utiliza un abordaje alternativo, tomando como referencia las ideas expresadas por Chavarría (2011) y Padrón (1992),

quienes mencionan que no hay producción de conocimientos que solo aborden cantidades excluyendo las cualidades ni viceversa, por tanto es necesario aprender a reconocer que los enfoques cualitativos y cuantitativos no pueden existir de manera aislada.

El mundo contiene aspectos cuantitativos indisolublemente mezclados con aspectos cualitativos, de tal modo que resulta imposible reconstruirlo atendiendo solo a uno de ambos aspectos. La única diferencia está en el lenguaje (Padrón, 1992). Asimismo, no tiene sentido hablar de enfoques mixtos pues estos intentan mezclar de manera disimulada situaciones o hechos que nunca están separados.

Por ello en este trabajo se explican las dimensiones epistemológica, ontológica y axiológica de la investigación, sin hacer alusión explícita a un enfoque en particular.

En primer lugar, la dimensión epistemológica, al tratarse del estudio de algoritmos existentes y contrastar resultados entre ellos haciendo uso de la técnica de validación cruzada desde una visión de ejecutor, implica que se debe sostener una postura de observación del fenómeno. El investigador debe poner atención a los detalles de eficacia, eficiencia y facilidad de implementación. Esto implica que las mediciones se deben efectuar en condiciones similares para obtener los resultados óptimos durante el proceso de ejecución que permitan generar la información confiable y ordenada mediante el análisis de los datos para originar el modelo de predicción.

Para describir la dimensión ontológica, se parte del concepto expresado por Gruber (1993): "(…) donde una ontología es una explicación sistemática de la existencia. Para los sistemas de inteligencia artificial, lo que ‘existe’ es lo que puede representarse". La representación del conocimiento requiere de mecanismos formales para expresar con precisión la idea que se desea comunicar. En ciencias de la computación, se entiende por ontología un conjunto de términos

básicos y relaciones entre sí. La representación ontológica del tipo de implementación algorítmica que se propone se puede apreciar en la figura 7.

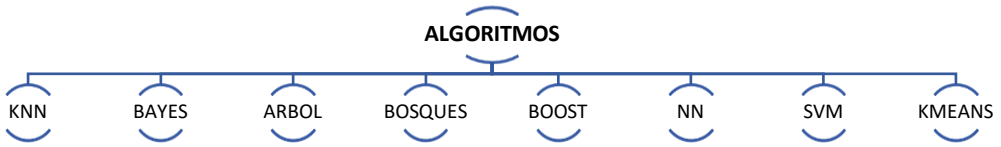


Figura 7. Ontología de algoritmos de predicción

Fuente: Elaboración propia

La implementación de la técnica que se propone para los algoritmos se puede programar de forma plana, sin recurrir a técnicas de paralelismo; no obstante, por la cantidad de datos y el número de ejecuciones que debe realizarse bajo las mismas condiciones con cada algoritmo, repercute en los tiempos de respuesta. Por tanto, se realiza la programación utilizando paralelismo, en este caso, usando las capacidades del computador personal. Si se contara con todo el equipo de cómputo distribuido, es más eficiente su ejecución y, por ende, los resultados son más rápidos. Estos algoritmos son ampliamente estudiados y utilizados para diferentes análisis de predicción. Por ello se considera que el estudio propuesto tiene una posibilidad clara de representación del conocimiento que se desea producir.

La dimensión axiológica requiere especial tratamiento en cuanto a la facilidad de implementación del algoritmo. La eficacia y la eficiencia únicamente demandan de los investigadores un esfuerzo por realizar mediciones y documentarlas de manera adecuada. Sin embargo, al hablar de algoritmos de predicción se debe tener claridad de lo que esto implica respecto a la certeza o aseguramiento de un hecho o evento. Tal y como se menciona en el párrafo

5 de alcances y limitaciones, estos algoritmos de predicción son probabilísticos. La facilidad de implementación, será evaluada de acuerdo a la escala de valores expuesta en la tabla 10.

Tabla 10. Criterios de evaluación de facilidad de implementación.

		Valores de aceptabilidad		
Criterio	Peso	Baja	Media	Alta
Comprensión del negocio	5%	2% si hay mucha complejidad para definir la necesidad del negocio	3% si no hay mucha complejidad para definir la necesidad del negocio	5% si no hay complejidad para definir la necesidad del negocio
Comprensión de los datos	10%	3% si es difícil entender y ubicar dónde están los datos	6% si se entiende y ubican los datos, pero se hacen varias extracciones para revisar	10% si se entiende con facilidad dónde están los datos y cuál es el significado de ellos
Preparación de los datos	30%	10% si la preparación de los datos lleva menos de 20 horas	15% si la preparación de los datos lleva entre 20 – 30 horas	30% si la preparación de los datos lleva más de 30 horas
Modelado	25%	5% si se usa alguna técnica de programación	10% si se usa una programación parametrizada	20% si se usan técnicas de paralelismo,

		plana que requiere de muchas líneas de código y más tiempo máquina en procesar	para la ejecución de los algoritmos	parametrización y validación cruzada de forma simultánea, requisiendo menos tiempo máquina para procesar
Evaluación	15%	5% si para evaluar los resultados no se tienen mediciones claras y precisas de los resultados	10% si los resultados no exponen con claridad lo planteado en los objetivos del proyecto	15% Si los resultados cumplen con los objetivos planteados en el proyecto
Implementación	15%	2% Si no se tiene el equipo de hardware, software y técnico	5% Si no se tiene el equipo de hardware, software, pero sí el equipo técnico	5% Si se tiene el equipo de hardware, software y técnico, con el conocimiento necesario para poner en desarrollo cualquier proyecto

Fuente: Elaboración propia

La ponderación total es 100%, donde la nota mayor indica que tan bueno resulta implementar la construcción del modelo de minería de datos para resolver el problema planteado. Para considerar que el modelo propuesto es óptimo y se debe implementar, se debe tener un valor por encima del 80%, un modelo aceptable con un porcentaje igual o mayor al 50% y un modelo no recomendable si se obtiene una nota por debajo del 50%.

Diseño

El diseño de esta investigación es evaluativo. Correa, Puerta y Restrepo (2002) consideran que la evaluación se aplica a la ejecución de una tarea, la calidad de la educación, las condiciones del clima organizacional, la calidad del material didáctico, el mercado potencial de un producto, el sentido de pertenencia, o bien un determinado programa, proyecto o institución, etc. (p. 17).

Correa, Puerta y Restrepo (2002) indican que todos los métodos evaluativos se refieren a la investigación y la apreciación de la eficiencia, eficacia, efectividad y relevancia social de una institución, programa o proyecto, para lo cual se requiere disponer de información sobre los insumos, los procesos y los productos o resultados. Cuanto más válidas, confiables y representativas sean estas informaciones, mayor es la posibilidad de reorganizar los fines y los medios de un programa o proyecto, de tal forma que este arroje resultados óptimos. Obtener y valorar estas informaciones es lo que se denomina evaluación (p. 29).

Se evalúa la eficiencia y eficacia de un algoritmo para predecir la rentabilidad de los clientes.

Población y muestreo

Para efectuar la recolección de datos, no es necesario llevar a cabo entrevistas, solo se recibe la indicación de parte del área experta de la empresa acerca de cuáles son las estructuras de

base de datos que se deben utilizar para la extracción de los datos que faciliten la creación y ejecución del modelado.

Instrumentos de recolección de datos

Por el tipo de investigación, no se realizan entrevistas ni otro elemento como observación, solo se hace una reunión con el departamento solicitante de la necesidad y las áreas expertas en datos, para que establezcan cuáles conjuntos de datos se deben considerar para el proceso de validación del modelo de predictivo.

Instrumentos de análisis de la información

El análisis de los datos se lleva a cabo en la plataforma de *software* profesional de código abierto de R y RStudio, dedicada a la computación estadística, mediante un conjunto de herramientas integradas y robustas que permiten trazar y visualizar por medio de gráficos los resultados cuando así se requiera.

Estrategia de desarrollo de la propuesta

Se pretende solucionar el problema al aplicar las fases de la metodología CRISP-DM, para tener un entendimiento completo del negocio y de los datos, con la finalidad de darles una estructura adecuada a estos últimos y generar un punto estable de los mismos, posibilitando efectuar el análisis correcto al usar todos los algoritmos de predicción mediante la técnica de validación cruzada en la fase de modelado, que ayude a obtener los resultados óptimos para escoger el mejor algoritmo por utilizar en el estudio de predicción que se propone en los objetivos del proyecto.

Capítulo IV Análisis del diagnóstico

En la actualidad el Banco de Costa Rica no tiene proyectos en el área de minería de datos, esta va a ser la primera experiencia en el desarrollo de un proyecto enfocado en este tema. Algunas personas y alrededor de tres o cuatro departamentos del Banco de Costa Rica tienen conocimientos o han escuchado sobre el concepto y las ventajas que se pueden obtener si aplican esta técnica.

Hoy a nivel mundial muchas entidades financieras utilizan procesos de minería de datos para la detección de fraudes en tarjetas de crédito; algunos de estos son redes neuronales, árboles de decisión y redes de creencia bayesiana en modelos predictivos, que emplean sistemas de aprendizaje autónomo para el reconocimiento de patrones basados en hechos históricos, con el propósito de generar alertas sobre la probabilidad de un fraude, identificando patrones de comportamientos atípicos, fraudulentos y/o sospechosos de las transacciones u operaciones que realizan los clientes.

El departamento dueño y pionero de este proyecto en el Banco de Costa Rica sufre una serie de transformaciones a nivel de nombre y de funciones en los últimos tres años, con anterioridad se denomina como el Área de Finanzas y luego pasa a llamarse Inteligencia Financiera, posteriormente se fusiona con el Departamento de Inteligencia de Negocios para formar lo que hoy se conoce como Centro de Información Estratégica, al que en adelante se le nombra por su siglas CIE.

Este departamento hoy se encarga de suministrar información de valor a las gerencias del Banco para tomar decisiones importantes a nivel de negocio, ya sea para eliminar o crear nuevos productos o servicios, o bien, para cerrar o crear nuevas oficinas comerciales para la atención de

los clientes. La información que se procesa permite medir financieramente a las oficinas y los clientes, con el objetivo de establecer si son rentables para la institución.

El departamento recibe una serie de archivos de datos de diversas oficinas del conglomerado del Banco de Costa Rica; algunos los procesa en un sistema informático y otros de forma manual para obtener la información que se les remite a los gerentes del Banco y a partir de ahí estos se apoyen en la toma de decisiones.

Para medir la rentabilidad de los clientes, la oficina procesa de modo manual alrededor de 65 archivos. El formato de los mismos es variable, muchos son txt, csv, xls, xlsx y Dbf. El proceso es asignado a un grupo de cinco personas que dura alrededor de 15 en la carga, procesamiento y validación de la información. Para los gerentes del Banco, el proceso es muy lento, porque no les posibilita ejecutar acciones preventivas, sino que más bien son lo que ellos llaman “apaga incendios”.

Para tales efectos, en paralelo se inicia el proceso de extracción, carga y transformación de los datos en el DWHC. A pesar de que se baja el tiempo de procesamiento, los resultados del proyecto aún no son los esperados, por la poca disponibilidad de los colaboradores de esa área para validar los resultados de las cargas y procesamiento de los datos.

En este proyecto el Departamento del CIE propone un archivo de datos con una serie de información del cliente tanto personal como de sus productos, más otros datos calculados con anterioridad mediante la información que suministran las oficinas del conglomerado BCR, que ayudan a conocer la rentabilidad de los clientes.

El objetivo de esa área no es solo saber la rentabilidad de los clientes, sino también predecirla para los clientes nuevos y actuales cuando estos cambian sus condiciones. Entonces, conociendo cuáles variables o criterios considera este departamento como determinantes para

calcular la rentabilidad de los clientes, se parte como un hecho que ya existe una experiencia sobre los datos y el análisis de estos; y mediante su análisis, se realizan cálculos matemáticos por medio de la aplicación de una fórmula que genera como resultado un valor que ellos denominan rentabilidad de un cliente BCR.

El modelo de rentabilidad de un cliente para este departamento tiene su fundamento en la identificación de los ingresos menos los gastos asociados a un cliente y sobre todo a los productos. Además, orientado a la valoración con y sin el efecto del diferencial cambiario (tipo de cambio del dólar). Sin embargo, el departamento encargado del proceso no cuenta con la totalidad de los datos de ingresos y gastos de sus clientes debidamente estructurados, este es un proceso que sigue en mejora continua.

A efectos del proyecto, ellos identifican una serie de datos que creen adecuados para poner en marcha el proceso de minería de datos, los cuales son tomados del DWHC y van a ser validados contra las fuentes de información que ellos tienen; se trata de un proceso en paralelo por un periodo aproximado de ocho meses.

Otro objetivo es que el modelo se pueda enriquecer de forma paulatina conforme se lleven a cabo las validaciones y evaluaciones de los datos, cálculos y resultados generados. Esta parte es importante para darle veracidad a la información que se obtenga del proceso predictivo, porque cuando se hacen cambios en las variables que se utilizan como ingresos para el modelo, ya sea eliminando o agregando, impactan los resultados e inclusive la selección del algoritmo predictivo resultante de la técnica de aprendizaje explicada en el capítulo II, en la sección “Validación cruzada”.

El área dueña del proceso identifica un conjunto de variables para alimentar el modelo de minería de datos, en total son 102 variables; no obstante, se les hace saber que no todas son

necesarias en el proceso, por el contrario al aplicar la solución puede ser que muchas sean irrelevantes y se descarten para obtener mejores resultados.

Capítulo V Propuesta de solución

En este capítulo se explica a detalle la solución empleada para resolver el problema planteado. Para el desarrollo del proyecto, se hace uso de la metodología CRISP-DM.

The CRISP-DM consortium (2000) indica que el ciclo de vida de un proyecto de minería de datos consta de seis fases, las cuales se muestran en la figura 8. La secuencia de las fases no es rígida o estricta. Siempre se requiere avanzar y retroceder entre las diferentes fases si fuese necesario. El resultado de cada fase determina qué fase o tarea particular debe realizarse a continuación. Las flechas señalan las dependencias más importantes y frecuentes entre las fases.

El círculo exterior en la figura 8 simboliza la naturaleza cíclica de la minería de datos en sí misma. La extracción de datos no termina una vez que se implementa una solución. Las lecciones aprendidas durante el proceso y a partir de la solución utilizada pueden generar nuevas preguntas empresariales, a menudo más enfocadas. Los procesos de minería de datos posteriores se benefician de las experiencias de los anteriores.

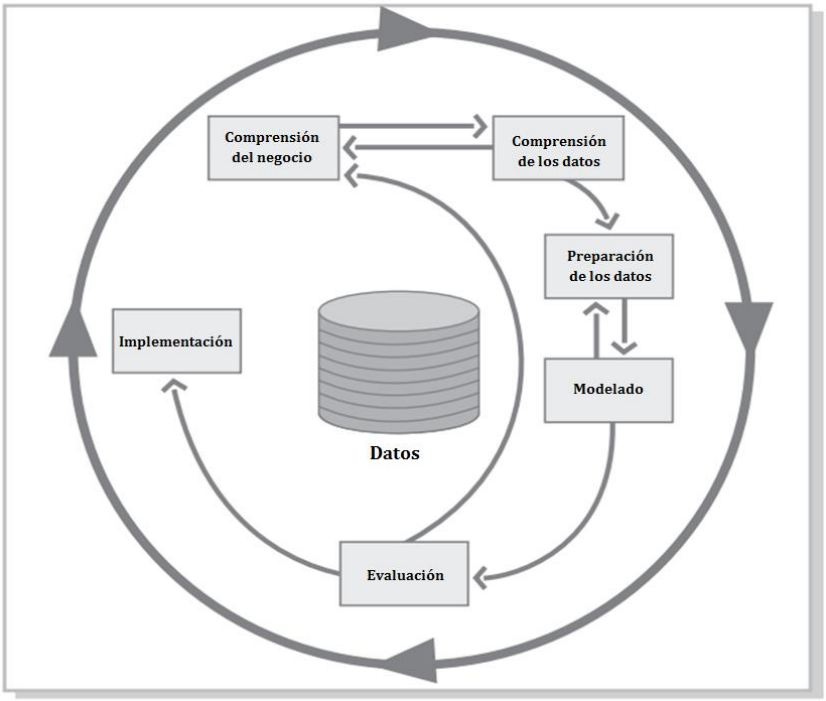


Figura 8. Modelo de referencia, fases de CRISP-DM
Fuente: The CRISP-DM consortium, 2000

Fases de CRISP-DM

En los siguientes puntos se hace una breve descripción de cada una de las fases que se ilustran en la figura 8. Todas las definiciones son tomadas de The CRISP-DM consortium (2000).

Fase I. Comprensión del negocio.

Esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva comercial, y luego convierte este conocimiento en una definición de problema de minería de datos y un plan preliminar diseñado para lograr los objetivos. Es decir, aquí se define la necesidad del cliente.

Fase II. Comprensión de los datos.

La fase de comprensión de datos comienza con la recopilación de datos inicial y continúa con actividades que permiten familiarizarse con los datos, identificar problemas de calidad de datos, descubrir conocimiento preliminar sobre los datos y/o detectar subconjuntos para formar hipótesis con respecto a la información oculta.

Fase III. Preparación de los datos.

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final [datos que se incorporan a la herramienta o herramientas de modelado] a partir de los datos brutos iniciales. Es probable que las tareas de preparación de datos se realicen varias veces y no en ningún orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de datos para las herramientas que hacen el modelado.

Fase IV. Modelado.

En esta fase se seleccionan y aplican diversas técnicas de modelado de acuerdo al problema que se analiza, y sus parámetros se calibran a valores óptimos. Por lo general hay varias técnicas para el mismo tipo de problema de minería de datos, algunas tienen requisitos específicos sobre la forma de los datos. Por lo tanto, volver a la fase de preparación de datos a menudo es necesario.

Fase V. Evaluación.

En esta etapa del proyecto se crea un modelo (o varios modelos) que parece tener alta calidad desde una perspectiva de análisis de datos. Antes de proceder al despliegue final del

modelo, se debe evaluar a fondo y revisar los pasos ejecutados para crearlo, con el propósito de asegurarse que el modelo logre los objetivos del área de negocio correctamente. Un objetivo clave es determinar si hay algún problema de negocio importante que no se haya considerado de manera adecuada. Al final de esta fase, se debe alcanzar una decisión sobre el uso de los resultados del proceso de análisis de datos de minería.

Fase VI. Implementación.

La creación del modelo por lo general no es el final del proyecto. Incluso si el objetivo del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido debe organizarse y presentarse de modo que el cliente pueda utilizarlo. A menudo implica la aplicación de modelos "en vivo" dentro de los procesos de toma de decisiones de una organización, por ejemplo, la personalización en tiempo real de las páginas web o la puntuación repetida de las bases de datos de *marketing*.

Según los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como emplear un proceso repetible de minería de datos en toda la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva a cabo los pasos de implementación. Sin embargo, incluso si el analista realiza el esfuerzo de despliegue, el cliente debe comprender por adelantado qué acciones deben efectuarse para utilizar los modelos creados.

Puesta en marcha de las fases CRSP-DM en el proyecto

Una vez definidas con claridad las fases que abarcan un proyecto de minería de datos, a continuación se procede a ejecutarlas en el proyecto.

Comprensión del negocio.

Al inicio de este proyecto, en el capítulo I, se abarca con detalle este punto, por consiguiente no se profundiza sobre el tema de comprensión del negocio en esta parte. Para tales efectos, se definen los objetivos y requisitos del proyecto entendiendo la problemática del área involucrada y la necesidad de aplicar la minería de datos. Esto permite alcanzar en un 5 % el primer valor de aceptabilidad del proyecto indicado en la tabla 10. No obstante, Se hace la aclaración y se cita este punto para ver el uso correcto del empleo de las fases de CRISP-DM.

Comprensión de los datos.

Una vez comprendida la necesidad del negocio y los objetivos formulados, para obtener los datos que se utilizan en el modelo se realiza la extracción del conjunto de datos desde el DWHC. El objetivo aquí es establecer el primer contacto con el problema, familiarizándose con la información para identificar la calidad y determinar los pasos por seguir.

Recolección de datos iniciales.

La primera tarea en esta segunda fase del proceso de CRISP-DM es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. El proceso de conocimiento de las estructuras de la base de datos facilita la ubicación de la información requerida en el DWHC, ya que al tener la indicación de parte del área experta de cuáles son las variables representativas para esta, se extrae la información por utilizar en el modelo.

Se procede a crear una consulta para ejecutarla en la base de datos Oracle, haciendo uso de la instrucción (*inner join*) para unir las tres tablas que contienen los tipos de información de los

clientes: desde datos personales, sus productos hasta de su rentabilidad. Los problemas presentados en este punto se citan en el párrafo 5 del apartado “Limitaciones”.

Descripción de los datos.

En total se cuenta con un millón ochocientos treinta y seis mil quinientos dieciocho registros físicos (1 836 518), sin embargo se acuerda con el área de negocio tomar una muestra de los datos, exportándola a un archivo de extensión csv, separado por punto y coma.

Conexión a la base de datos.

Se descarta establecer una conexión directa con la base de datos, por las razones indicadas en el párrafo 6 de la sección “Limitaciones”.

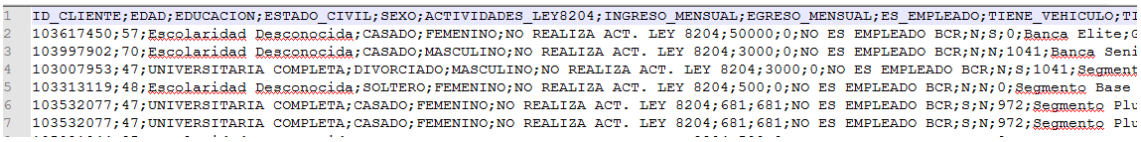
Extracción de los datos.

Se toma solo una muestra de datos físicos (figura 9) porque se requiere analizar por aparte las dos clasificaciones de los clientes que hace la empresa. La muestra del archivo tiene 85 360 registros con 75 columnas (tabla 11). Hablando en términos de minería de datos, se denomina a las columnas como ***variables*** y a las filas como ***individuos***.

La idea de la separación surge como una recomendación técnica debido a que las características que poseen ambas clasificaciones de clientes varían por su concepto, por ejemplo un cliente físico por lo general tiene un patrono, pero uno jurídico no; por lo tanto para evitar descartar datos que pueden ser relevantes para la clasificación de clientes físicos, se les aconseja hacer el análisis de los datos por separado, lo cual les parece muy acertado pues de lo contrario se pueden obtener resultados imprecisos y un análisis sesgado.

Exploración de datos.

En este proceso se revisan los datos para corroborar los tipos de variables (tabla 11), la codificación y el separador de listas y decimales, con el objeto de encontrar una estructura general previo a ser cargados en el s*oftware* RStudio para efectuar el análisis de calidad y la preparación para el modelo predictivo.



```
ID_CLIENTE;EDAD;EDUCACION;ESTADO_CIVIL;SEXO;ACTIVIDADES_LEY8204;INGRESO_MENSUAL;EGRESO_MENSUAL;ES_EMPLEADO;TIENE_VEHICULO;TI
103617450;57;Escolaridad Desconocida;CASADO;FEMENINO;NO REALIZA ACT. LEY 8204;50000;0;NO ES EMPLEADO BCR;N;S;0;Banca Elite;C
103997902;70;Escolaridad Desconocida;CASADO;MASCULINO;NO REALIZA ACT. LEY 8204;3000;0;NO ES EMPLEADO BCR;N;N;1041;Banca Seni
103007953;47;UNIVERSITARIA COMPLETA;DIVORCIADO;MASCULINO;NO REALIZA ACT. LEY 8204;3000;0;NO ES EMPLEADO BCR;N;S;1041;Segment
103313119;48;Escolaridad Desconocida;SOLTERO;FEMENINO;NO REALIZA ACT. LEY 8204;500;0;NO ES EMPLEADO BCR;N;N;0;Segmento Base
103532077;47;UNIVERSITARIA COMPLETA;CASADO;FEMENINO;NO REALIZA ACT. LEY 8204;681;681;NO ES EMPLEADO BCR;S;N;972;Segmento Plu
103532077;47;UNIVERSITARIA COMPLETA;CASADO;FEMENINO;NO REALIZA ACT. LEY 8204;681;681;NO ES EMPLEADO BCR;S;N;972;Segmento Plu
```

Figura 9. Exploración de datos físicos
Fuente: Datos del Banco de Costa Rica

Información de los datos					
	N°	Variable	Acepta nulos	Es llave	Descripción
1	ID_CLIENTE	Cualitativa, nominal	No	Sí	Identificación del cliente
2	ID_MES	Categórico	No	No	Identificador del mes
3	ID_ANIO	Categórico	No	No	Identificador del año
4	MON	Categórico	No	No	Identificador de la moneda
5	TIPO_PROD	Cualitativa, nominal	No	No	Identificador de las tipificaciones del producto
6	ID_PRODUCTO	Cualitativa, nominal	No	Sí	Identificador del producto
7	DESC_PROD	Cualitativa, nominal	No	No	Descripción del producto
8	A	Cuantitativa	No	No	Colocación (crédito, tarjetas de crédito, líneas de crédito, sobregiros cuentas corrientes)
9	B	Cuantitativa	No	No	Captación (cuentas corrientes, ahorros, CDPS)
10	D	Cuantitativa	No	No	Ingresos financieros (tarjetas de crédito, intereses y comisiones de formalización de créditos
11	E	Cuantitativa	No	No	Ingresos por servicios (comisiones)
12	H	Cuantitativa	No	No	Captación (gastos financieros)
13	I	Cuantitativa	No	No	Costos operativos indirectos (tarjetas, conectividades)
14	K	Cuantitativa	No	No	Costos operativos directos

15	M	Cuantitativa	No	No	Riesgo por productos
16	RFF	Cuantitativa	No	No	Rendimiento financiero por fondeo
17	CFF	Cuantitativa	No	No	Costo financiero por fondeo
18	TIPO_CLI	Cualitativa	Sí	No	Tipo de cliente
19	PROVINCIA_ESTADO	Cualitativa	Sí	No	Provincia del cliente
20	CANTON	Cualitativa	Sí	No	Cantón del cliente
21	DISTRITO	Cualitativa	Sí	No	Distrito del cliente
22	ID_SEXO	Cualitativa	Sí	No	Identificador del sexo (1= hombre, 2 = mujer)
23	SEXO	Cualitativa	Sí	No	Descripción del sexo
24	EDAD	Cualitativa	Sí	No	Edad del individuo
25	ID_ESTADO_CIVIL	Cualitativa	Sí	No	Identificador del estado civil
26	ESTADO_CIVIL	Cualitativa	Sí	No	Descripción del estado civil
27	TOTAL_HIJOS	Cuantitativa	Sí	No	Total de hijos del cliente
28	ES_DIFUNTO	Cualitativa	Sí	No	Indicador para saber si el cliente físico falleció
29	TIENE_VEHICULO	Cualitativa	Sí	No	Indicador para saber si el cliente tiene vehículo propio
30	TIENE_CASA_PROP	Cualitativa	Sí	No	Indicador para saber si el cliente tiene casa propia
31	TIENE_TELEF_CASA	Cualitativa	Sí	No	Indicador para saber si el cliente tiene teléfono en la casa
32	TIENE_CEL	Cualitativa	Sí	No	Indicador para saber si el cliente tiene teléfono móvil
33	TIENE_TELEF_TRAB	Cualitativa	Sí	No	Indicador para saber si el cliente tiene teléfono en el trabajo
34	DESC_SEGMENTO	Cualitativa	Sí	No	Descripción del segmento al que pertenece el cliente
35	PROF_LINEA	Cualitativa	Sí	No	Identifica si la profundidad de línea está activada, 1:sin profundidad de línea 0:con profundidad de línea
36	ES_EMPLEADO	Cualitativa	Sí	No	Indicador para saber si el cliente labora dentro de la empresa
37	TIENE_PATRONO	Cualitativa	Sí	No	Indicador para saber si el cliente tiene patrono o no
38	EDUCACION	Cualitativa	Sí	No	Nivel de educación del cliente
39	DESC_PROFESION	Cualitativa	Sí	No	Descripción de la profesión del cliente
40	OCUPACION	Cualitativa	Sí	No	Ocupación del cliente
41	INGRESO_MENSUAL	Cuantitativa	Sí	No	Ingreso mensual del cliente
42	EGRESO_MENSUAL	Cuantitativa	Sí	No	Egreso mensual del cliente
43	ACTIVIDADES_LEY8204	Cualitativa	Sí	No	Indicador para saber si el cliente realiza actividades relacionadas con lavado
44	ID_CLASIFICACION_SUGEF	Cualitativa	Sí	No	Identificador para la clasificación de SUGEF
45	DESCRIPCION_SEGMENTO	Cualitativa	Sí	No	Descripción del segmento del cliente dentro del Banco
46	DESC_MON_INGRESO	Cualitativa	Sí	No	Descripción del monto de ingreso del cliente
47	RECIBE_PAGO_PLANILLA	Cualitativa	Sí	No	Indicador para saber si recibe pago en planilla o no
48	RAZON_SOCIAL	Cualitativa	Sí	No	Razón social, aplica para jurídicos
49	NATURALEZA_NEGOCIO	Cualitativa	Sí	No	Naturaleza del negocio, aplica para jurídicos
50	OTROS_INGRESOS	Cuantitativa	Sí	No	Otros ingresos que tiene el cliente
51	COD_GR	Cualitativa	Sí	No	Número de la gerencia regional
52	NOM_GR	Cualitativa	Sí	No	
53	COD_UE	Cualitativa	Sí	No	Número de la oficina a la que pertenece el cliente
54	NOM_UE	Cualitativa	Sí	No	

55	COD_GL	Cualitativa	Sí	No	Número de la gerencia local
56	NOM_GL	Cualitativa	Sí	No	
57	SDO_PROM_MENS	Cuantitativa	Sí	No	Es el saldo promedio mensual del producto
58	SDO_FIN_MES	Cuantitativa	Sí	No	Es el saldo promedio mensual del producto
59	CANTIDAD	Cuantitativa	Sí	No	Identificador interno de DWH para la cantidad de productos
60	SDO_ACTUAL	Cuantitativa	Sí	No	Saldo actual del producto
61	SDO_CORTE	Cuantitativa	Sí	No	Es el saldo del producto al final del día
62	LIM_CREDITO	Cuantitativa	Sí	No	Es el límite de crédito para tarjetas de crédito
63	SDO_ACTUAL_DISPONIBLE	Cuantitativa	Sí	No	Es el saldo con los créditos y débitos en tránsito aplicados
64	SDO_FIN_MES_DISPONIBLE	Cuantitativa	Sí	No	Es el saldo fin de mes con los créditos y débitos en tránsito aplicados
65	MTO_RCLI	Cuantitativa	No	No	Es el monto de la rentabilidad por cliente
66	OPERA	Cuantitativa	Sí	No	Número de la operación
67	PRODU	Cuantitativa	Sí	No	Identificador del producto
68	TIMONEDA	Cuantitativa	Sí	No	Identificador de la moneda
69	ESTADO	Cuantitativa	Sí	No	Estado del producto
70	MTOORIGINAL	Cuantitativa	Sí	No	Monto original del producto de colocación
71	MTOCUOTA_TOTAL	Cuantitativa	Sí	No	Monto total de la cuota de pago sobre el producto de colocación
72	INGRESOBRUTO	Cuantitativa	Sí	No	Ingreso bruto del cliente
73	TOTALAVALUO	Cuantitativa	Sí	No	Monto total del avalúo
74	CATEGORIARIESGO	Cualitativa	No	No	Es el indicador de la categoría de riesgo del cliente
75	RCLI	Binario: "SÍ", "NO"	No	No	Identificador para determinar si el cliente es rentable o no a partir de la fórmula aplicada por el negocio

Tabla 11. Clasificación de los datos
Fuente: Datos del Banco de Costa Rica

Verificación de la calidad de los datos.

En la exploración se efectúan verificaciones sobre los datos para determinar la consistencia de los valores, la cantidad y los valores nulos y se revisa si existen valores fuera de rango con el propósito de saber las acciones por seguir para la completitud y limpieza de los datos. Al respecto, se llevan a cabo las siguientes tareas:

- Se definen las variables más importantes y representativas de los datos y el lugar de donde se extraen.

- Una vez exportados a un archivo, se exploran los datos con el fin de encontrar una estructura general y saber qué cambios y transformaciones se deben realizar.
- El set de datos extraído carece de información en múltiples variables por los problemas de calidad de datos que hay en las fuentes de información que alimentan el DWHC. Por consiguiente, para evitar errores en el procesamiento e inconsistencias en los resultados, es necesario hacer el proceso de limpieza de datos, utilizando técnicas de minería de datos para que se obtengan los mejores resultados al aplicar los algoritmos sobre los datos limpios.

El proceso de comprensión de los datos se realiza con claridad, se tiene un total entendimiento de la información extraída; a raíz de esto, se realizan las recomendaciones indicadas sobre los conjuntos de datos. Además se alcanza el segundo valor de aceptabilidad del proyecto, en su totalidad 10 %, señalado en la tabla 10.

Preparación de los datos.

Después de haber realizado la comprensión y recolección de los datos, se ejecutan varias técnicas para hacer la lectura, limpieza y corrección de la fuente con la finalidad de obtener los datos más apropiados para alimentar el modelo de minería de datos. Todo el código generado y las tareas efectuadas se ejecutan con el *software* R.

Filtrar los datos.

Se hace el filtrado sobre la columna PROVINCIA_ESTADO, usando solo los datos de las provincias de Costa Rica, dado que en esa columna hay muchos valores y caracteres inapropiados que afectan la calidad de la información. Para ello, en la figura 10 se observa que usando la función

de R “***filter”*** se hace la selección de los datos de Costa Rica y se descartan los valores erróneos. Además con la función ***“dim”*** se muestra la dimensión de los datos cargados.

```
#cargar los datos
fisicos <- read.csv("RENTABILIDAD_CLIENTES.csv",header=T, sep = ";",dec = ".") #Leo los datos del
archivo
dim(fisicos)[1]
```

```
## [1] 80473
```

```
# Filtro los datos únicamente de CR
fisicos <- filter(fisicos, PROVINCIA_ESTADO == 'SAN JOSE' | PROVINCIA_ESTADO == 'ALAJUELA' | PROVI
NCIA_ESTADO == 'HEREDIA' | PROVINCIA_ESTADO == 'CARTAGO' |PROVINCIA_ESTADO == 'PUNTARENAS' | PROVI
NCIA_ESTADO == 'LIMON' | PROVINCIA_ESTADO == 'GUANACASTE')

#Total de individuos después de la carga y filtro
dim(fisicos)[1]
```

```
## [1] 80360
```

Figura 10. Carga de datos físicos
Fuente: Elaboración propia en R

Eliminar datos NA.

R tiene muchas ventajas por la gran variedad de funciones ya diseñadas que posee. En el caso de eliminar datos con valores NA, se usa la función ***“na.omit”*** que provee la aplicación, lo cual permite borrar con rapidez una gran cantidad de información del set de datos original que no posee valores. Lo anterior provoca como resultado un conjunto de datos con 37 118 registros, tal y como se observa en la figura 11.

```
#Total de individuos después de la carga y filtro de provincias
dim(fisicos)[1]
```

```
## [1] 80360
```

```
#Eliminamos los datos NA
fisicos <- na.omit(fisicos)

#Total de individuos después del filtro de NA's
dim(fisicos)[1]
```

```
## [1] 37118
```

Figura 11. Elimina datos NA
Fuente: Elaboración propia en R

Estandarizar los datos

Al crear una función para dolarizar, se estandariza la moneda de todas las variables que corresponden a montos o saldos, para manejar un solo tipo de moneda en los diferentes productos. Además se convierte el set de datos a dataframe. En la figura 12 se observa el uso de la función y cómo se le envía por parámetros las columnas por convertir y la columna con el tipo de moneda que posee el registro. A partir de ese valor, la función determina si es necesario hacer la conversión o si debe mantenerse el valor original del dato.

```
################ Dolarizo todas las columnas que tienen datos en colones ################
fisicos[,var.cuantit] <- convierte.tc( fisicos[,var.cuantit], fisicos[,"TIMONEDA"], mto.tc)

#Convertimos a dataframe el set de datos
df.fisicos <-  as.data.frame(fisicos)

#Vemos los tipos de datos
str(df.fisicos)
```

```
## 'data.frame':    37118 obs. of  35 variables:
##  $ ID_CLIENTE            : int  103617450 103997902 103007953 103313119 103532077 103532077 105
021044 109296641 103555232 104976120 ...
##  $ EDAD                  : int  57 70 47 48 47 47 35 83 82 81 ...
##  $ EDUCACION             : Factor w/ 7 levels "Escolaridad Desconocida",..: 1 1 6 1 6 6 1 3 1 6
```

Figura 12. Estandarizar datos
Fuente: Elaboración propia en R

Cambio de observaciones en las variables.

Se reemplazan los valores internos de algunas variables para estandarizarlos. En la figura 13 se muestra un ejemplo de cómo se reemplazan las observaciones en la variable ***ES_EMPLEADO***, por valores como “N” y “S”. Así se lleva a cabo en varias variables con el objeto de mejorar su visualización e interpretación.



Figura 13. Cambio de observaciones en variables
Fuente: Elaboración propia en R

Análisis de los datos.

Por medio de una función en esta etapa se revisa la cantidad de valores nulos, NA's, ceros, el mínimo, el máximo, la media, la varianza y la desviación estándar de cada variable cuantitativa existente en el set de datos. En la figura 14 se nota un ejemplo de cómo se analiza y revisa la calidad de las variables, además en el caso de las variables ***"E, H, I, K y RFF"***, el porcentaje de valores con cero está por encima del 40 %, lo cual es casi la mitad de la información. Por esto se toma la decisión de eliminar del análisis todas aquellas variables con un porcentaje de nulidad, de ceros o de los demás puntos por revisar, que sean iguales o mayores a dicho valor porcentual.

	E	H	I	K	M	RFF	CFF
TotalNulos	0	0.0	0	0	0	0	0.0
TotalNa	0	0.0	0	0	0	0	0.0
TotalCeros	46	96.0	100	100	100	93	6.7
TotalNans	0	0.0	0	0	0	0	0.0
Minimo	-7920	0.0	0	0	0	0	0.0
Maximo	6164501	32033.3	0	0	0	14779463	1500059.6
Media	1537	3.4	0	0	0	4557	13370.1
Varianza	3189991921	47095.2	0	0	0	20413648247	3312347643.5
DesviacionEstandar	56480	217.0	0	0	0	142876	57553.0

Figura 14. Análisis de calidad en las variables
Fuente: Elaboración propia en R

Eliminar y seleccionar variables.

Gracias a los resultados presentados por el proceso de análisis de datos, este proceso de eliminación y selección de variables se realiza durante varias etapas del proceso de clasificación y estandarización de los datos, hasta obtener las variables más representativas. Básicamente sucede porque en el análisis de variables se encuentran muchos valores con falta de información que no agregan valor a los datos, pues muchas tienen altos porcentajes de nulidad y de ausencia de información, generando ruido a los datos y, en consecuencia, afectando los resultados finales.

En la figura 15 se observa que primero se crea un conjunto de datos con los nombres de las variables por eliminar; luego, haciendo uso de la función de R ***"which"***, ***"%in%"*** y ***"names",*** se eliminan del set de datos original todas las variables encontradas y se vuelve a asignar al dataframe original el nuevo conjunto de datos sin las variables con mayores problemas de calidad de información.

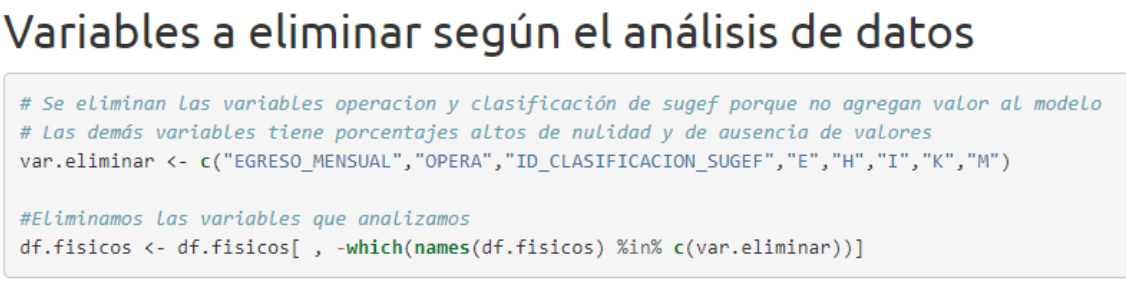


Figura 15. Variables por eliminar en el primer análisis
Fuente: Elaboración propia en R

Estructurar los datos.

Los datos que se cargan están unidos por cliente y por producto, pero se requiere analizar la rentabilidad del cliente, por lo cual es necesario obtener un solo registro por individuo y no varios como está al momento de la carga. En la figura 11 se observa que una vez realizada la eliminación, se obtiene un total de 37 118 individuos, pero aún están por producto y no por cliente, como se busca obtener la información.

Para ello, en primer lugar se convierten los productos -es decir, las observaciones- en variables, para indicar si el cliente tiene o no un producto. En la figura 16 se muestra el uso de las funciones ***"mutate"*** *y* ***"%>%",*** donde se crean las nuevas variables tomando como punto de partida los datos insertados dentro de cada registro.

```
df.fisicos <-
        df.fisicos %>%
        mutate(CANT_PROD=1)%>%
        mutate(PRODU_0=ifelse(PRODU==0,1,0)) %>%
        mutate(PRODU_2=ifelse(PRODU==2,1,0)) %>%
        mutate(PRODU_3=ifelse(PRODU==3,1,0)) %>%
        mutate(PRODU_5=ifelse(PRODU==5,1,0)) %>%
        mutate(PRODU_6=ifelse(PRODU==6,1,0)) %>%
        mutate(PRODU_7=ifelse(PRODU==7,1,0)) %>%
        mutate(PRODU_10=ifelse(PRODU==10,1,0)) %>%
        mutate(PRODU_11=ifelse(PRODU==11,1,0)) %>%
        mutate(PRODU_12=ifelse(PRODU==12,1,0)) %>%
        mutate(PRODU_13=ifelse(PRODU==13,1,0)) %>%
        mutate(PRODU_14=ifelse(PRODU==14,1,0)) %>%
        mutate(PRODU_15=ifelse(PRODU==15,1,0)) %>%
        mutate(PRODU_16=ifelse(PRODU==16,1,0)) %>%
        mutate(PRODU_18=ifelse(PRODU==18,1,0)) %>%
        mutate(PRODU_20=ifelse(PRODU==20,1,0)) %>%
        mutate(PRODU_21=ifelse(PRODU==21,1,0)) %>%
        mutate(PRODU_22=ifelse(PRODU==22,1,0)) %>%
        mutate(PRODU_23=ifelse(PRODU==23,1,0)) %>%
        mutate(PRODU_28=ifelse(PRODU==28,1,0)) %>%
        mutate(PRODU_99=ifelse(PRODU==99,1,0))
```

Figura 16. Crea las variables producto
Fuente: Elaboración propia en R

Desnormalizar los datos.

Una vez ejecutado el paso anterior, se efectúa la desnormalización de la información, para obtener un solo registro por individuo, en este caso se usan varias funciones propias del CRAN de R como sum, min, max, group_by, summarise y otras diseñadas para el proyecto como son:

- **getFirst**: Permite escoger el primer registro encontrado para un individuo cuando recibe por parámetro el nombre de la variable por evaluar.
- **revisaBanderas**: Recibe un arreglo y lo revisa, para determinar si en alguno de los elementos existe el valor pasado como valorBuscado; si logra encontrarlo, devuelve valorBuscado, de lo contrario devuelve valorNoEncontrado

- **revisaCantidadPRODUs:** Revisa la cantidad de productos que tiene un individuo, si la cantidad es mayor a cero, le pone una S; de lo contrario, una N. Se usa para marcar si el individuo tiene un producto.
- **summarise**: Es una función propia de R que recibe un *dataframe* y en combinación con la función ***group_by*** posibilita obtener resultados agrupados de las observaciones en sus respectivas variables.

La figura 17 presenta claramente el uso correcto de todas las funciones mencionadas.

```
df.fisicos <-
  df.fisicos %>%
  group_by(ID_CLIENTE) %>%
  summarise(  EDAD  = min(EDAD),
              EDUCACION =   getFirst(EDUCACION),
              ESTADO_CIVIL  =   getFirst(ESTADO_CIVIL),
              SEXO  =   getFirst(SEXO),
              ACTIVIDADES_LEY8204 = revisaBanderas(ACTIVIDADES_LEY8204,'S','N'),
              INGRESO_MENSUAL = max(INGRESO_MENSUAL),
              ES_EMPLEADO = revisaBanderas(ES_EMPLEADO,'S','N'),
              TIENE_VEHICULO    =   getFirst(TIENE_VEHICULO),
              TIENE_CASA    =   getFirst(TIENE_CASA),
              DESCRIPCION_SEGMENTO  =   getFirst(DESCRIPCION_SEGMENTO),
              PROVINCIA_ESTADO = getFirst(PROVINCIA_ESTADO),
              ESTADO    =   getFirst(ESTADO),
              MTOORIGINAL   =   sum(MTOORIGINAL),
              MTOCUOTA_TOTAL    =   sum(MTOCUOTA_TOTAL),
              INGRESOBRUTO  =   sum(INGRESOBRUTO),
              TOTALAVALUO   =   sum(TOTALAVALUO),
              CATEGORIARIESGO   =   getFirst(CATEGORIARIESGO),
              A =   max(A),
              B =   max(B),
              D =   max(D),
              RFF   =   max(RFF),
              CFF   =   max(CFF),
              MTO_RCLI  =   max(MTO_RCLI),
              RCLI = getFirst(RCLI),
              CANT_PROD=sum(CANT_PROD)
```

Figura 17. Desnormalizar datos
Fuente: Elaboración propia en R

- **Guardar los datos preseleccionados**: seleccionados los datos para el siguiente paso, se guardan tal y como se ve en la figura 18, donde se aprecia el nombre del archivo y los separadores de cada variable.

```
#Almaceno el archivo para realizar la discretizacion de las variables y proceder con la predicción
write.csv(df.fisicos,file="RENT.CLIENTES.SELECC.csv",sep=";")
```

Figura 18. Guardar el archivo preseleccionado
Fuente: Elaboración propia en R

Discretización.

Para entender mejor los datos y su clasificación, se lleva a cabo un proceso que se conoce como discretización. Existen tres tipos de clasificaciones de este método de reducción o suavizado de los datos. En primer lugar, la discretización por contenedores consiste en dividir desde arriba hacia abajo, basándose en un número de contenedores; por análisis de histogramas, es una técnica de discretización no supervisada porque no utiliza información de la clase y la otra es la discretización por clúster, árbol de decisión y análisis de correlación.

A efectos del proyecto, se usa la de clúster, aplicando el método de kmeans a las variables cuantitativas y estandarizando los datos a valores más fáciles de comprender, lo cual facilita categorizarlos, esto crea agrupaciones o clústeres y posibilita obtener un conjunto de datos de mayor calidad para alimentar el modelo de validación cruzada que arroja la rentabilidad de los clientes.

Al aplicar el método de kmeans a cada variable cuantitativa, junto con el método de validación cruzada 50 veces, se obtiene el mejor k en cada una de ellas, el proceso se ejecuta con cada uno de los algoritmos de kmeans, determinando de forma automática dónde se estabilizan los datos y generando la cantidad óptima de clúster que se debe tener en cada variable.

En la figura 19 se ilustra una gráfica del método con todos sus algoritmos y debajo de la imagen se encuentra el nombre del mejor algoritmo y el mejor k detectado computacionalmente, con la ejecución de la validación cruzada para el método de kmeans en todos sus algoritmos. Según el resultado generado para la variable cuantitativa de la figura 19, el mejor algoritmo es MacQueen y el mejor k es 11; o sea, la variable debe ser clusterizada en 11 grupos.

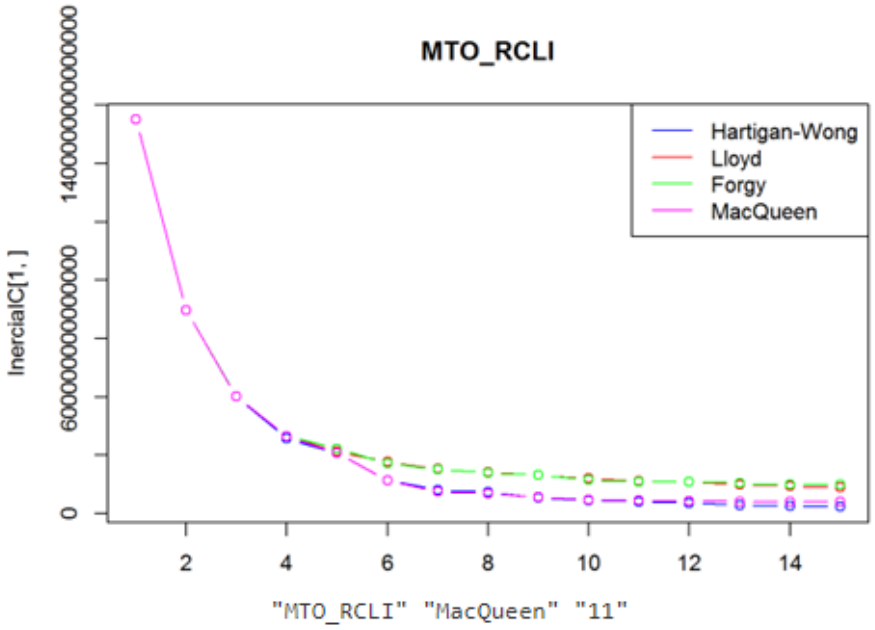


Figura 19. Discretización por clúster
Fuente: Elaboración propia en R

Análisis de componentes principales (ACP).

Durante el proceso de preparación de los datos, también se hace una reducción de la dimensión. Cuando se aplica esta técnica, se eliminan individuos y variables mal representados, usando coseno cuadrado mayor al 10 %. Esto ayuda a buscar la mejor representación de los datos, con el objetivo de sintetizar la información; o sea, si se tiene "x" cantidad de variables, con este método se va a expresar "x-1, x-2.. x-n" así sucesivamente, hasta encontrar una pequeña cantidad

de variables, pero con la suficiente cantidad de información. En la figura 20 se muestra el uso de la función coseno cuadrado del 10 %.

```
#Elimine individuos y variables mal representados (coseno cuadrado menor al 10 %).
modelo.fisicos<-PCA(fisicos.PCA, scale.unit=TRUE, ncp=5, graph = FALSE)

#Individuos y variables mal representados

#Grafica y Elimina Individuos mal representados: para tener una mejor representación en el gráfico
plot(modelo.fisicos, axes=c(1, 2), choix="ind",col.ind="red",new.plot=TRUE,select="cos2 0.1")
```

```
#Grafica y Elimina Variables mal representadas:
plot(modelo.fisicos, axes=c(1, 2), choix="var",col.ind="blue",new.plot=TRUE,select="cos2 0.1")
```

Figura 20. Análisis de componentes principales
Fuente: Elaboración propia en R

Correlación de las variables.

En la figura 21 se muestra cómo se realiza el análisis de correlación de las variables cuantitativas para determinar cuáles tienen correlaciones altas positivas, altas negativas o si del todo no hay ninguna relación; para decidir cuáles se deben mantener en el conjunto de datos final.

```
MC <- cor(fisicos.PCA)

corrplot(MC, method="circle", tl.cex = 0.5, cl.cex = 0.5)
```

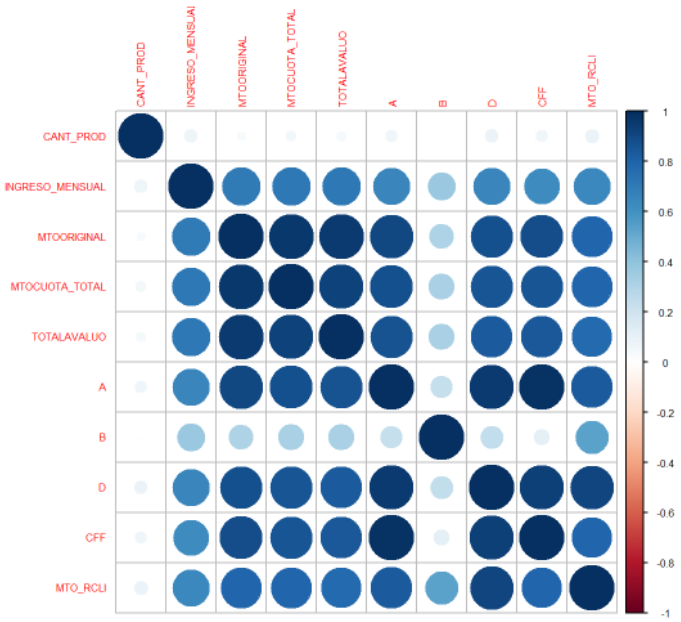


Figura 21. Gráfico de correlación de variables
Fuente: Elaboración propia en R

Cabe aclarar que los métodos aplicados para esta muestra del conjunto de datos utilizados no son una regla específica ni son los pasos exclusivos para cualquier otro conjunto de datos de otras investigaciones, todo depende y varía de acuerdo al tipo de información que se suministre y los objetivos y necesidades del área de negocio involucrada. Empero, este trabajo es uno de los más complejos porque requiere de muchas horas, pruebas y técnicas para obtener los datos precisos para el proceso de modelado. El total de horas invertidas en esta fase para este proyecto es mayor a 30 horas, aproximadamente se trabaja entre 40 y 60 horas, con lo cual se logra alcanzar en la totalidad el tercer valor de aceptabilidad del proyecto, indicado en la tabla 10.

Modelado.

Después de terminar la fase de preparación de los datos, se obtiene como resultado el conjunto de datos para el proceso de validación cruzada, donde se tiene un registro por cada individuo y la relación con sus productos, su calificación de riesgo y su rentabilidad; estos últimos se encuentran dentro de cada variable y corresponden a las características que se desean detectar para predecir la rentabilidad de los clientes.

En este etapa se hace el uso de la técnica de validación cruzada con diferentes pruebas, para encontrar en qué punto el error se estabiliza. Entre las pruebas de calibración realizadas, se ejecuta el proceso de distintas maneras; a continuación se enumera y detalla cada una:

1. Se repite la ejecución cinco veces, con 10 grupos, ver la figura 22.
2. Se repite la ejecución seis veces, con 10 grupos, ver la figura 23.
3. Se repite la ejecución 15 veces, con 10 grupos, ver la figura 24.
4. Se repite la ejecución cuatro veces, con 10 grupos, ver la figura 25.

```
# Validacion cruzada 5 veces
tiempo.paralelo <- system.time(
  # Hacemos Validación cruzada 5 veces para ver que el error se estabiliza
    for (i in 1:5) {
        grupos <- createFolds(1:n, 10) # Crea los 10 grupos
        no.svm <- 0
        no.knn <- 0
```

Figura 22. Validación cruzada cinco veces
Fuente: Elaboración propia en R

```
# Validacion cruzada 6 veces
tiempo.paralelo <- system.time(
  # Hacemos Validación cruzada 5 veces para ver que el error se estabiliza
    for (i in 1:6) {
        grupos <- createFolds(1:n, 10) # Crea los 10 grupos
        no.svm <- 0
        no.knn <- 0
```

Figura 23. Validación cruzada seis veces
Fuente: Elaboración propia en R

```
# Validacion cruzada 15 veces
tiempo.paralelo <- system.time(
  # Hacemos Validación cruzada 5 veces para ver que el error se estabiliza
    for (i in 1:15) {
        grupos <- createFolds(1:n, 10) # Crea los 10 grupos
        no.svm <- 0
        no.knn <- 0
```

Figura 24. Validación cruzada 15 veces
Fuente: Elaboración propia en R

```
# Validacion cruzada 4 veces
tiempo.paralelo <- system.time(
  # Hacemos Validación cruzada 5 veces para ver que el error se estabiliza
    for (i in 1:4) {
        grupos <- createFolds(1:n, 10) # Crea los 10 grupos
        no.svm <- 0
        no.knn <- 0
```

Figura 25. Validación cruzada cuatro veces
Fuente: Elaboración propia en R

Al final se decide ejecutarlo con cuatro veces porque los resultados mostrados en cada ejecución en los gráficos de la sección “Gráfica del modelo de validación cruzada” evidencian con claridad que con cuatro veces ya el error se estabiliza, entonces no es necesario aumentar la cantidad de ejecuciones de la validación cruzada pues ocasiona más procesamiento del equipo de cómputo, más duración en cada ejecución y al final se obtienen los mismos resultados.

El proceso desarrollado consiste en que computacionalmente, en otras de forma automática, se seleccione del conjunto de datos original dos subconjuntos, uno para prueba y otro para aprendizaje; todo se realiza dentro de la lógica programada en la función. Esto permite probar la calidad del modelo predictivo porque para todos los algoritmos va a usar las mismas condiciones; es decir, si se usa el 30 % de los datos para prueba y el 70 % de los datos para entrenamiento, todos los algoritmos utilizan la misma cantidad y los mismos datos en esa ejecución y así es cada vez

que se ejecuta, pero con la variante de que va a utilizar diferentes datos de prueba y aprendizaje por cada ejecución.

La figura 26 ilustra un ejemplo de cómo se toman las particiones de los datos para prueba y entrenamiento en los distintos grupos, con la finalidad de alimentar las ejecuciones de los algoritmos dentro del método de validación cruzada. Al final de la ejecución se utilizan todos los datos, lo cual permite garantizar la integridad y veracidad del modelo con la muestra completa de información.

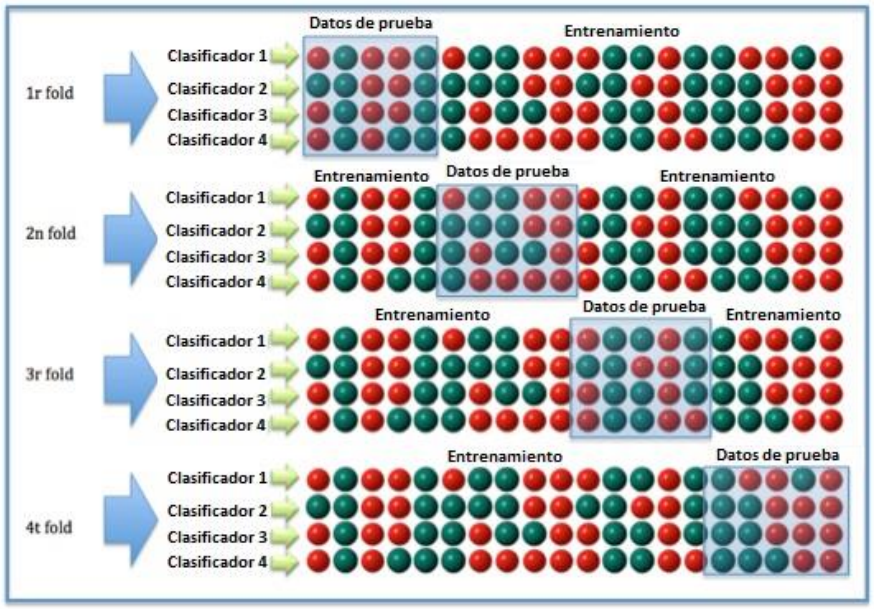


Figura 26. Validación cruzada con K-Folds
Fuente: Wikipedia, 2018

Aplicación de paralelismo.

También se desarrolla el uso de otra funcionalidad de R denominada ***"makecluster"*** del paquete ***snow***, lo cual paraleliza procesos, o sea dejarle una tarea a cada procesador; en este caso es aprovechar el procesador de la computadora para distribuir las tareas. El objetivo es que

mediante la paralización del código se mejore el tiempo de ejecución y de los recursos de un equipo, lo cual es muy utilizado en problemas reales de *Big Data*, donde se tienen conjuntos grandes de información y para procesarlos más rápido se recurre a este tipo de funcionalidades, aprovechando las ventajas a nivel de *hardware* y *software* como en el caso de R.

El paralelismo en R se trabaja básicamente bajo una arquitectura de jefe/peón, se crea una cantidad de clústeres según se definan en el código –en este caso se establecen ocho como se muestra en la figura 27–, posibilitando que los algoritmos se ejecuten en paralelo y se comuniquen a través de *sockets*.

En este caso se crean los indicados en la figura 27 por la capacidad del equipo donde se ejecuta el método de validación cruzada, el cual es un corei7, con 32 GB de RAM y un disco duro de estado sólido de 1TB. El código de validación cruzada con la muestra de datos utilizada en una ejecución normal sin el uso de paralelismo, dura entre ocho y 10 horas, pero usando esta técnica la ejecución se realiza en una hora y media.

```
suppressWarnings(suppressMessages(library(snow)))
clp<-makeCluster(8,type="SOCK")
# Constructor del cluster
ignore <- clusterEvalQ(clp, {
      suppressWarnings(suppressMessages(library(e1071)))
      suppressWarnings(suppressMessages(library(formula.tools)))
```

Figura 27. Creación de los clústeres para usar paralelismo
Fuente: Elaboración propia en R

Cuando se crean los clústeres de cada peón, se les pasan los mismos parámetros para la ejecución; por lo tanto, reciben la ruta, la cantidad de datos fragmentada para cada uno, la función y las bibliotecas que cargan los paquetes para cada modelo. No obstante, usar esta técnica es

complicado si el equipo donde se ejecuta no tiene suficientes recursos para atender todo el procesamiento, o bien, si alguno de los peones tiene un problema durante la ejecución, el proceso se detiene y lanza el error al nodo máster para que suspenda todo. Cada proceso en cada núcleo funciona de manera única, con lo cual se asegura que no se va a ser afectado por otro, ni presentar problemas de calidad ni tener alterados los resultados; el detalle es que al ser único, es más lento.

La figura 28 muestra cómo funciona el paralelismo dentro de R, donde primero crea un nodo máster y luego los nodos trabajadores para distribuir el trabajo, pasándole la muestra y los datos a cada uno para iniciar el proceso. Es necesario cerrar el clúster una vez terminado el proceso, para que los recursos como la memoria se devuelvan al sistema operativo.

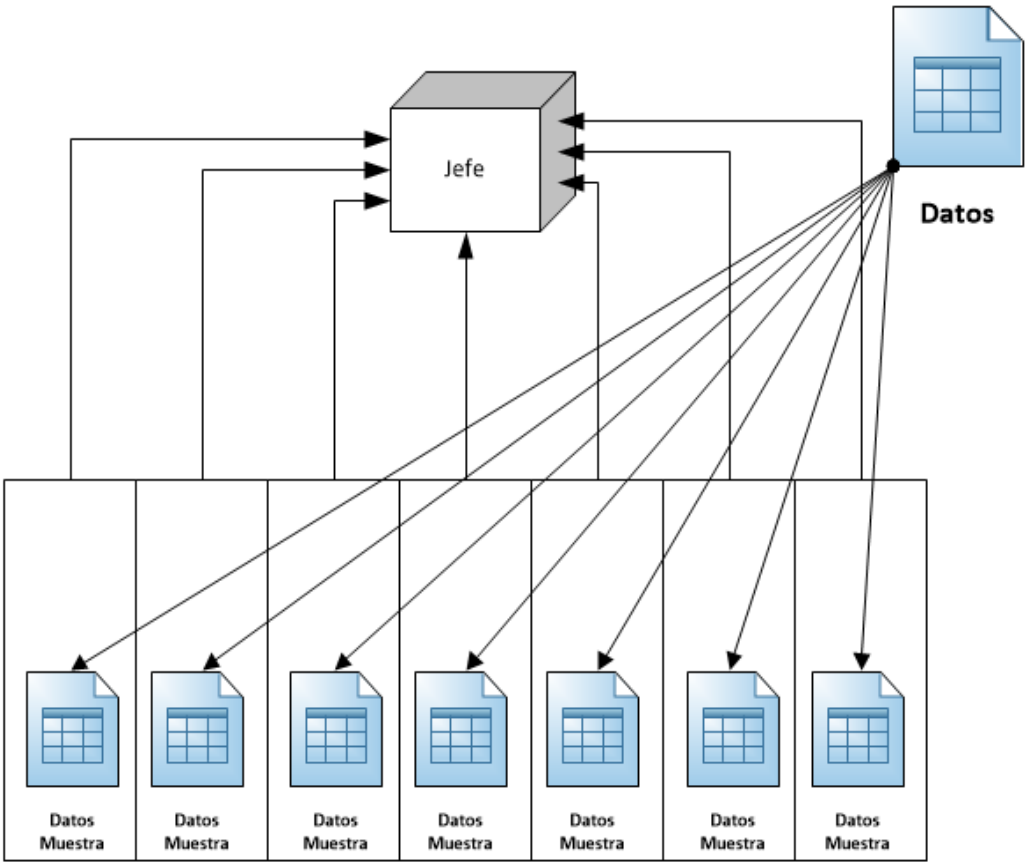


Figura 28. Paralelismo en R
Fuente: Elaboración propia

Gráfica del modelo de validación cruzada

En la figura 29 se observa que luego de ejecutar el proceso de validación cruzada con distintas iteraciones desde cuatro hasta 15, los resultados obtenidos siempre son los mismos. En este caso se puede concluir que según el conjunto de datos y las variables utilizadas de la muestra de información extraída del DWHC, el mejor algoritmo para predecir si un cliente no es rentable es el de red neuronal, ya que muestra mayor estabilidad en el proceso de validación cruzada en comparación con los demás algoritmos. Empero, si se requiere cambiar algún elemento como agregar o quitar variables, se debe efectuar de nuevo la validación cruzada con todos los algoritmos para determinar si el método que se tiene como válido continúa siéndolo o no.

Gráfico con cuatro iteraciones

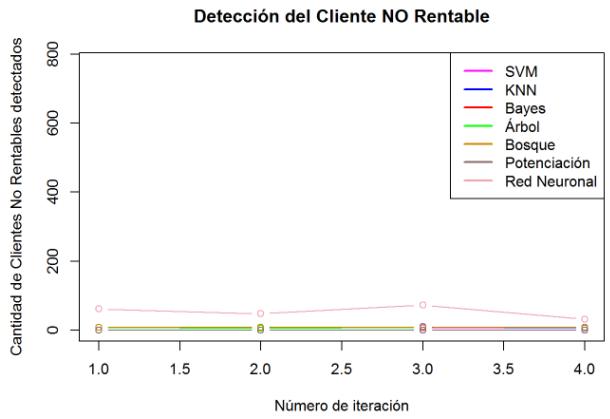


Gráfico con cinco iteraciones

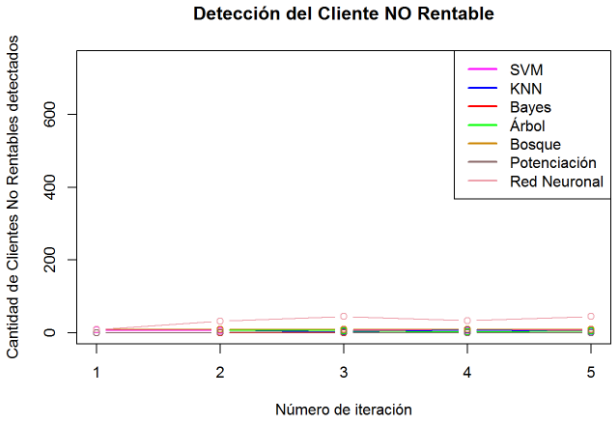


Gráfico con seis iteraciones

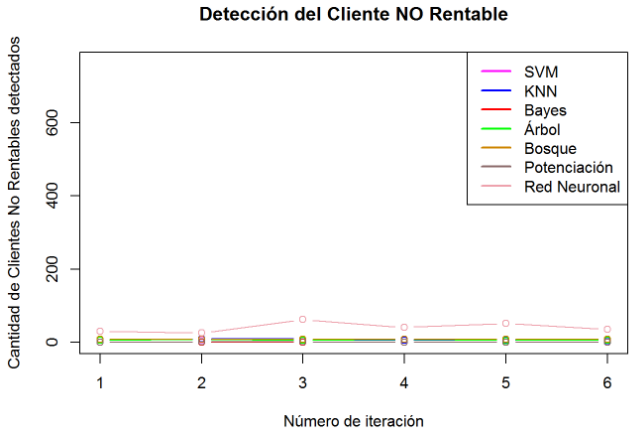


Gráfico con 15 iteraciones

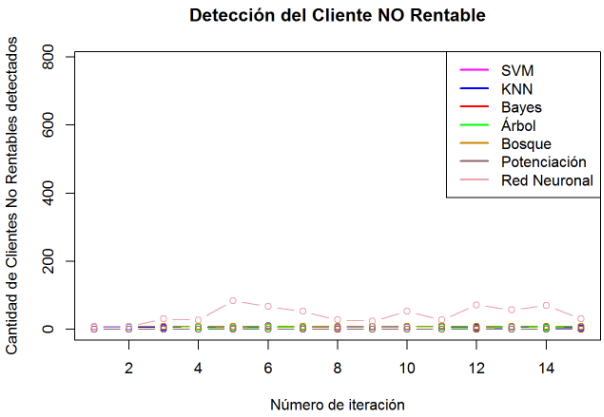


Figura 29. Gráficas del modelo con validación cruzada
Fuente: Elaboración propia en R

En esta fase todo el proceso se desarrolla parametrizando las entradas, usando técnicas de paralelismo y validación cruzada, haciendo más eficiente tanto los resultados como el tiempo de ejecución y los recursos del equipo. Además, permite alcanzar en un 20 % el cuarto valor de aceptabilidad del proyecto, señalado en la sección “Enfoque” del capítulo 3.

Evaluación.

Para la interpretación de los resultados se pueden usar múltiples herramientas, en este caso se sigue utilizado la funcionalidad de R para generar automáticamente archivos de tipo HTML con gráficos y análisis.

Una vez descubierto el mejor modelo predictivo para el conjunto de datos empleado en este proyecto, se procede a hacer la evaluación del mismo con un conjunto de datos con las mismas características de los datos de prueba y entrenamiento utilizados. Para tales efectos, como se aprecia en la figura 30, se leen los datos y se indica el algoritmo con el que se va a predecir, cuál es el conjunto que va a alimentar la tabla de aprendizaje y cuáles son los datos por predecir, con el propósito de establecer si es funcional según el estudio y la analítica avanzada previamente hecha en toda la investigación.

Evaluación del modelo predictivo

```
predecir <- read.csv("predecir.csv",header=T, sep = ";",dec = ".") #Leo los datos del a
rchivo
MODELADO <- read.csv("RENT.CLIENTES.MODELADO.csv", sep = ";", header = T)
vp <- dim(predecir)[2]
modelo<-nnet(RCLI~.,data=MODELADO,size=50,rang=0.1,decay=5e-4,maxit=300,trace=FALSE)
prediccion<-predict(modelo, predecir[,-vp], type="class")
Nuevo<-predecir[,-vp]
PruebaModeloRCLI <-cbind(Nuevo,prediccion)

#guardamos el resultado
write.csv(PruebaModeloRCLI,file="PruebaModeloRCLI.csv", sep = ";")
```

Figura 30. Evaluación del modelo de validación cruzada
Fuente: Elaboración propia en R

En la figura 31 se muestra que la aplicación a los nuevos datos del análisis predictivo para detectar la rentabilidad de los clientes es efectiva, pero más importante aún es que se cumplen las mismas características -en cuanto a variables se refiere- que el set de datos original.

Muestra de información

ModeloRCLI

MTOCUOTA_TOTAL <int>	TOTALAVALUO <int>	A <int>	B <int>	D <int>	C... <int>	MTO_RCLI <int>	CATEGORIARIESGO <int>	prediccion <fctr>
1	1	1	1	1	1	3	2	SI
1	1	1	1	1	1	3	1	SI
1	1	1	1	1	1	3	1	NO
1	1	1	1	1	1	3	1	SI
1	1	1	1	1	1	3	1	SI
1	1	1	1	1	1	3	1	SI
1	1	1	1	1	1	3	1	SI
8	5	7	1	6	7	5	1	SI
3	2	5	1	5	5	5	1	SI
1	1	1	1	1	1	3	1	SI

Figura 31. Nuevos datos predichos
Fuente: Elaboración propia en R

Con los datos finales limpios y estructurados, de la muestra extraída del DWHC se pueden llevar a cabo diversos tipos de análisis por medio de los gráficos. A efectos del proyecto se muestran algunos, alineados a los objetivos y que apoyan el desarrollo de las conclusiones del proyecto.

En la figura 32 se aprecia que los clientes con mayores niveles de rentabilidad se concentran en los segmentos uno, tres y siete, además en estos se encuentra una pequeña cantidad de clientes con niveles de rentabilidad menor a cero, es decir que a esos segmentos se les debe poner atención para determinar qué los hace tan rentables con respecto al resto. Otro segmento bastante rentable es el cinco porque ahí todos los clientes tienen buenos niveles de rentabilidad y no existe ninguno no rentable.

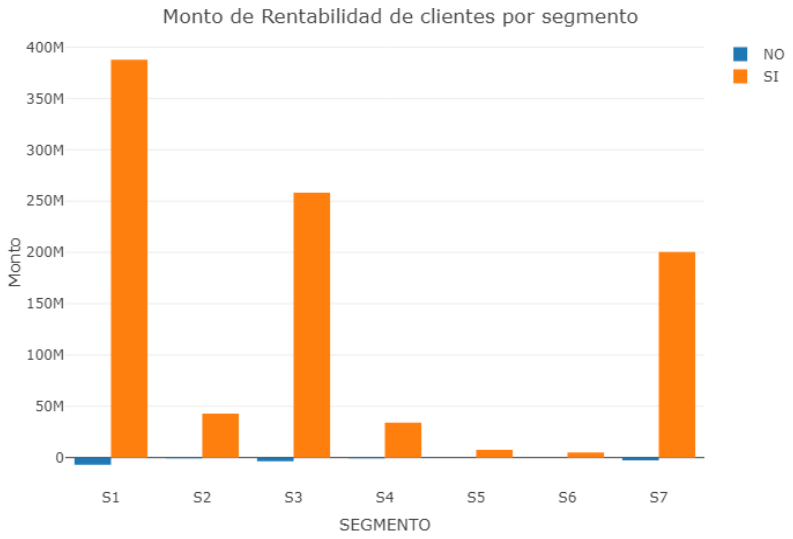


Figura 32. Monto de rentabilidad de clientes por segmento
Fuente: Elaboración propia en R

En la figura 33 se aprecia el total de clientes rentables por la clasificación de sexo, donde predominan mayormente los hombres sobre las mujeres, de la misma manera se ve una pequeña diferencia entre la cantidad de mujeres no rentables con respecto a los hombres.



Figura 33. Total de clientes rentables por sexo
Fuente: Elaboración propia en R

En la figura 34 se observa que la totalidad de productos de colocación y captación de la entidad financiera se concentran básicamente en los segmentos uno, tres y siete. Además, al realizar una comparación con el gráfico de la figura 32, se aprecia con claridad que los montos de rentabilidad de esos segmentos son mayores al resto, porque es donde hay mayor concentración de los productos y de los saldos, no obstante todo se centraliza sobre colocación, mas no así sobre captación, lo cual se convierte en un indicador por ser considerado en las campañas de mercadeo acerca del ofrecimiento de productos.

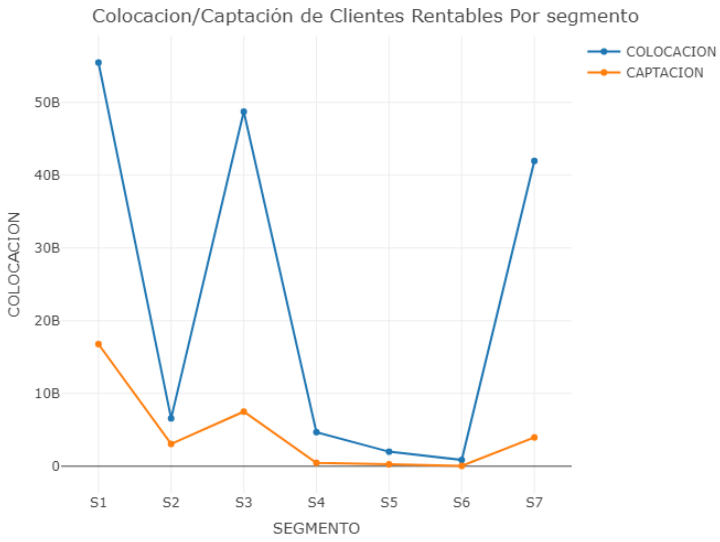


Figura 34. Colocación y captación de clientes rentables por segmento
Fuente: Elaboración propia en R

En el gráfico de la figura 35 se observa un gran predominio de la colocación y captación de productos sobre las provincias una, dos, tres y cuatro. En el resto de provincias es muy baja y tiene una mayor hegemonía S1 sobre el resto, esto quiere decir que se tiene un menor apoyo o aporte de parte del Banco en ciertos territorios, o bien, el poder adquisitivo de la población no es

el mismo en esas otras partes del país. En todas predomina la colocación, pero no la captación, por lo cual se deben mejorar las campañas para captación de dinero.

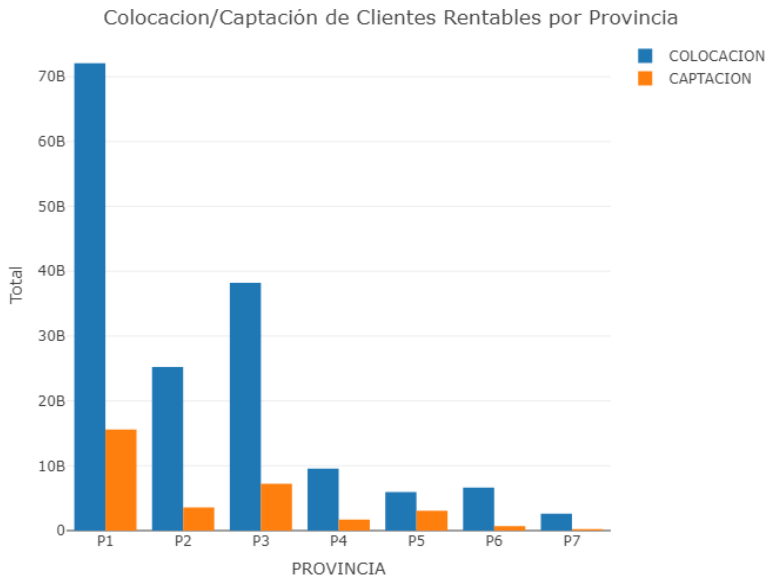


Figura 35. Colocación y captación de clientes rentables por provincia
Fuente: Elaboración propia en R

Según los gráficos de la figura 36, hay segmentos donde la mayoría son clientes no rentables, lo cual repercute mucho en los costos e imagen de la institución, porque en este momento ellos reciben la misma atención que los clientes rentables, haciendo más caro su mantenimiento y atención; además no generan un valor agregado al negocio en ningún ámbito. Esos segmentos requieren ser revisados con atención para conocer si al momento de otorgarles productos, la recopilación de información se hace correctamente o si es necesario mejorar los procesos.

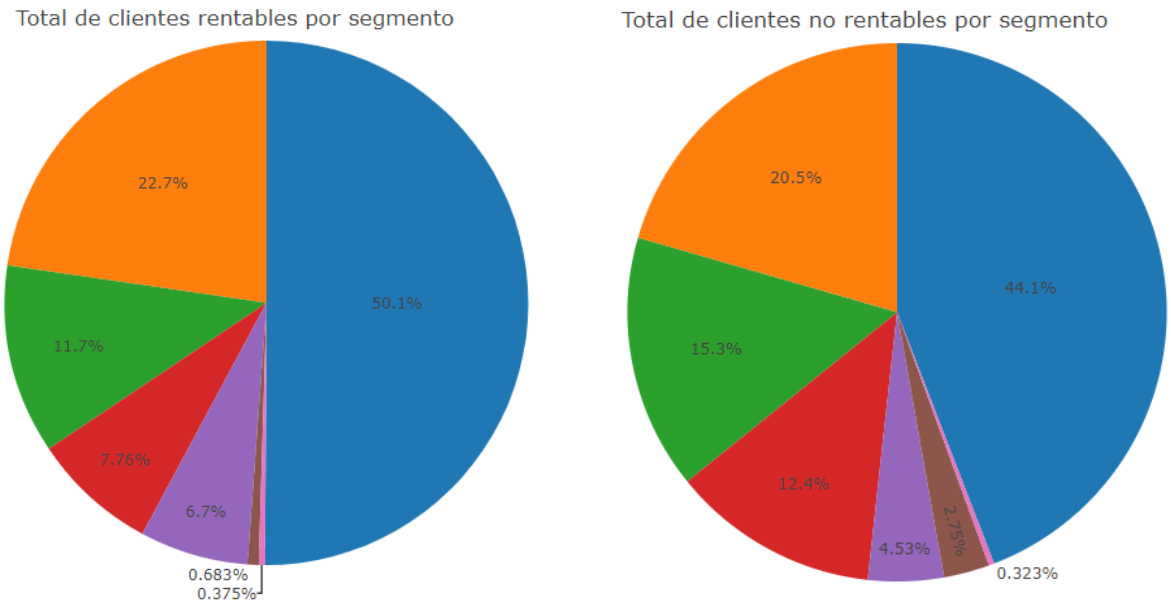


Figura 36. Total de clientes rentables por segmento
Fuente: Elaboración propia en R

En la figura 37 se muestran las cantidades porcentuales de clientes según las categorías de riesgo, donde los clientes rentables predominan con la mejor categoría en el primer gráfico y los no rentables en la peor categoría en el segundo gráfico. Lo anterior implica mucha vigilancia porque se puede estar dirigiendo mal la colocación de los productos hacia estos grupos de clientes, lo cual afecta los indicadores financieros y la imagen de la institución.

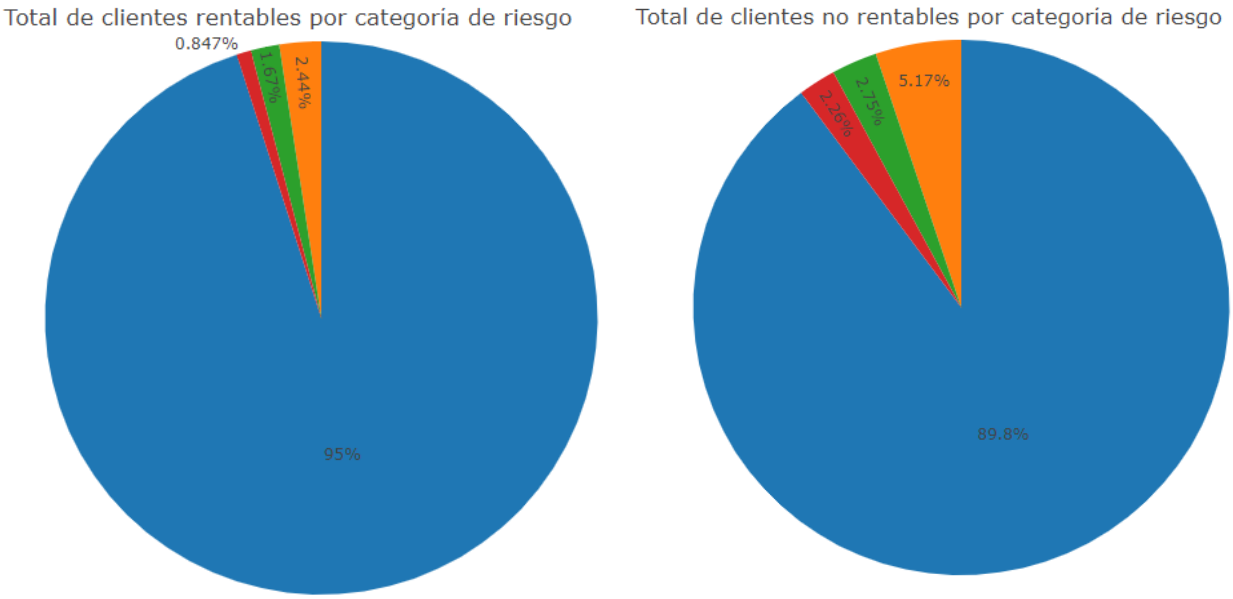


Figura 37. Total de clientes rentables por categoría de riesgo
Fuente: Elaboración propia en R

En la figura 38 se realiza un análisis usando la función de R ***"predict"***, en este caso se utiliza la historia de los datos y se efectúa la predicción del comportamiento de los datos hacia adelante, buscando predecir una semana. Así, se puede indicar cómo se pueden llegar a comportar los datos, dejando claro que es una predicción, por lo tanto se puede observar en el gráfico que la línea de color naranja tiene una sombra arriba y abajo, la cual indica que en ese rango puede estar la predicción y por ahí pueden moverse los datos. Lo señalado permite de manera proactiva tomar decisiones y acciones con tiempo para evitar que la rentabilidad tienda a caer en ese período, otorgándole una ventaja competitiva a la empresa para gestionar a sus clientes.

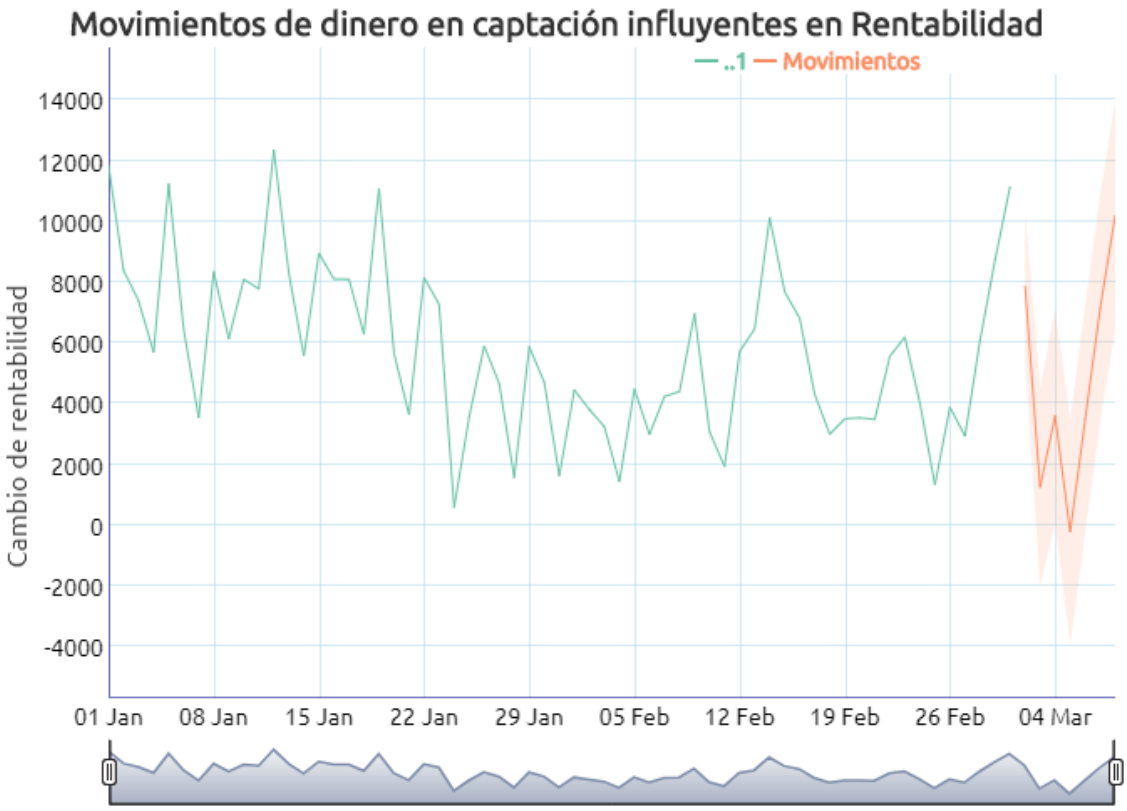


Figura 38. Serie de tiempo por rentabilidad
Fuente: Elaboración propia en R

La fase de evaluación logra los objetivos planteados en el proyecto y permite cumplir con el quinto valor de aceptabilidad, indicado en la tabla 10.

Implementación.

Para la implementación se requiere la adquisición del equipo de *hardware* adecuado para este tipo de proyectos y las licencias de *software* de Rstudio Server para la puesta en producción, considerando que se debe contar con un sitio alterno donde se realicen los respaldos y que también se brinde la posibilidad de contar con alta disponibilidad en caso de alguna falla o siniestro. No

obstante, sí se cuenta con el equipo técnico para poner en marcha proyectos de este tipo por lo que el último valor de aceptabilidad se alcanza en un 5 %, indicado en la tabla 10 del capítulo 3.

Ahora bien, suponiendo que se cuenta con lo descrito, entonces se recomienda una arquitectura como la que se muestra en la figura 39.

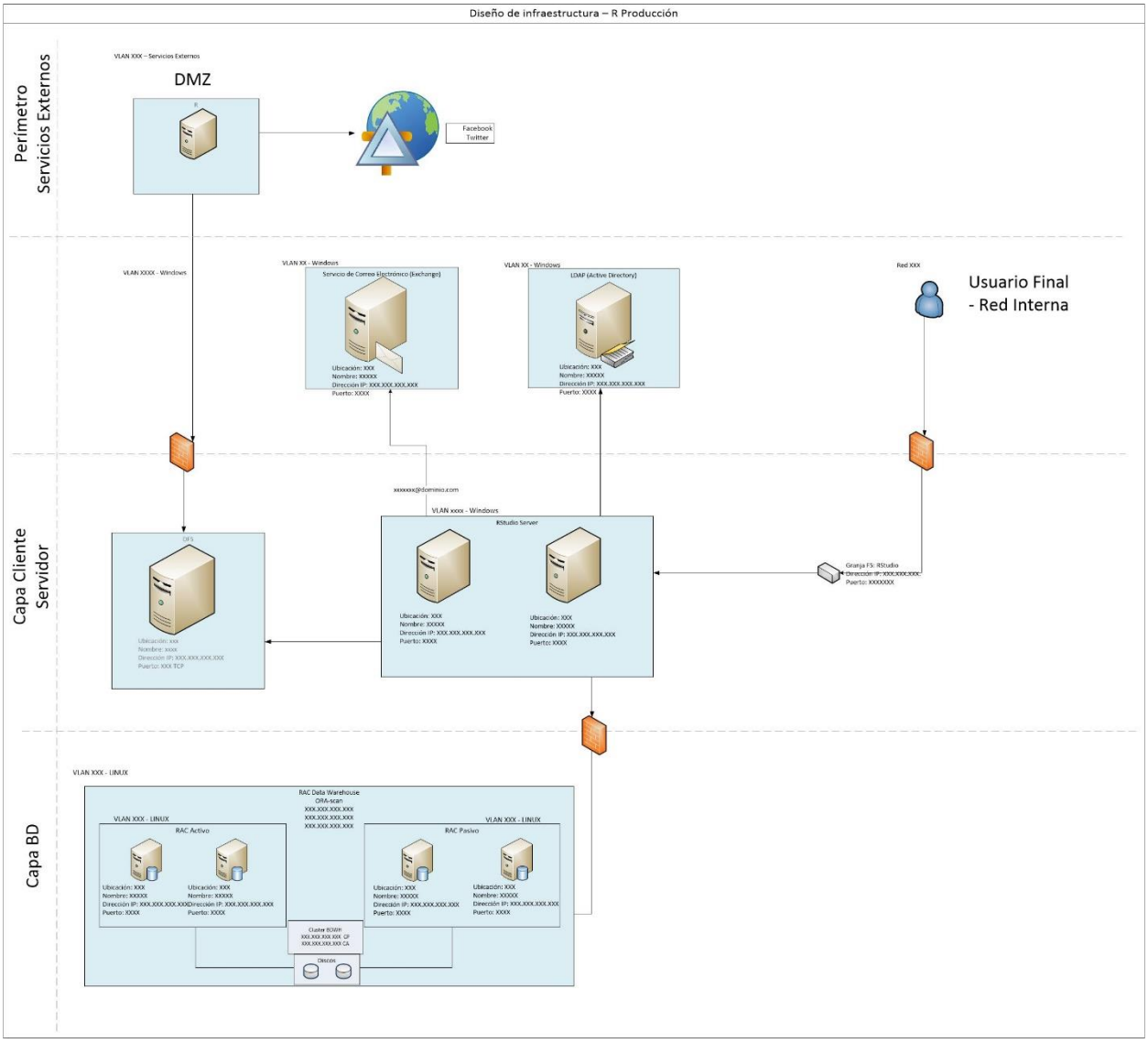


Figura 39. Arquitectura propuesta para la implementación
Fuente: Elaboración propia

Capítulo VI Conclusiones y recomendaciones

Conclusiones

Durante el desarrollo de este proyecto se realizan diferentes técnicas de minería de datos, acoplándose a lo propuesto por la metodología CRISP-DM. El proceso es bastante complejo y en muchas ocasiones es necesario regresarse a alguna de las fases, es decir, hacer un proceso de forma cíclica tal y como se aprecia en la figura 8.

Este proceso permite cumplir con los objetivos planteados, generando mucho valor para el negocio porque encuentra patrones que con anterioridad se desconocen y no son evidentes con los datos crudos, evitando tomar acciones y decisiones preventivas.

Después de reunirse con el negocio y comprender sus necesidades, se inicia con el proceso de comprensión de los datos para alinearlos con los requerimientos y posteriormente entrar a una de las fases más complejas de un proyecto de minería de datos: la preparación de los datos. Esta fase es de las que consume más dedicación porque para encontrar una estructura de los datos hay que llevar a cabo diversos esfuerzos y pruebas, en muchos casos se debe pasar por las distintas fases para conocer si los datos finales son los más adecuados para el modelo predictivo.

Al final, las pruebas y análisis de cada una de las variables empleadas en el modelo posibilitan cumplir con el primer objetivo de este proyecto, el cual consiste en *"Identificar las variables representativas que se van a utilizar en el modelo de minería de datos"***.** En esta etapa es necesario ir un poco más allá de lo pensado, se deben aplicar más técnicas, como la de discretización.

Este proceso lleva un gran esfuerzo porque también se utiliza la validación cruzada para determinar por medio del método de kmeans cuáles son los mejores algoritmos de dicho método

para clusterizar los datos. El resultado es satisfactorio y da una mejor estructura a las variables cuantitativas usadas en el proceso de validación cruzada. Es más útil emplear los atributos numéricos, por lo tanto, es importante hacer esa transformación antes, al ayudar a conocer la estructuración y clasificación que debe tener cada variable.

El proceso de preparación de los datos también permite cumplir con el segundo objetivo de este proyecto, el cual es *"Descubrir la correlación de las variables disponibles para determinar cuáles inciden más sobre la rentabilidad del cliente".* Mediante la funcionalidad de R se analiza la correlación de las variables y se grafican las que tienen mayor correlación entre sí, para ser manejadas en el conjunto de datos final que se utiliza en el proceso de validación cruzada, descartando una gran cantidad que se tiene al inicio, gracias a los análisis y estructuración dados a la información.

En cuanto a trabajo y tiempo, los dos primeros objetivos de este proyecto son los que mayor esfuerzo implican porque conocer y preparar los datos es una de las tareas más complejas en un proyecto de minería de datos, no es simple encontrar una estructura general adecuada para los datos, también se requiere de un trabajo arduo para hacer una escogencia de las variables y tomar la decisión de cuáles se deben descartar en el proceso. Cada una tiene un significado distinto y requiere de un análisis por separado para comprender por qué tienen valores de nulidad o ceros y así al final decidir si se deben mantener hasta el final del proceso o ser desechadas.

El tercer objetivo del proyecto es *"Identificar el mejor algoritmo predictivo al crear y ejecutar el modelo de validación cruzada con todos los algoritmos",* el cual se cumple en la fase de modelado, donde se demuestra que durante el proceso de validación cruzada el mejor algoritmo para la detección de clientes rentables es el de redes neuronales porque es el que mejor predice a los clientes no rentables.

El cuarto objetivo del proyecto consiste en *"Probar el algoritmo obtenido en la validación cruzada con un conjunto de datos que cumpla las mismas características del set de datos usado en el proceso"* y se cumple en la fase de evaluación, en la cual Evaluación.se ejecuta para un conjunto de datos suministrado con las mismas características -en cuanto a variables- que el set de datos original usado en el proceso de validación cruzada obtenido como resultado final durante la ejecución de las fases del proceso de CRISP-DM.

El objetivo cinco, *"Determinar los segmentos de clientes y zonas geográficas donde se concentran los clientes rentables",* permite conocer que solo ciertos segmentos y provincias son las más predominantes en cuanto a la rentabilidad de los clientes, lo cual posibilita a las gerencias de la institución bancaria tomar acciones ya sea para ofrecerles mejores condiciones en los productos, brindarles mejor atención o conocer cuáles son las necesidades que tienen para estudiar financieramente si es factible cumplirlas, además para dirigir a grupos de personas que les brinden atención en sus zonas geográficas.

Debido a que las condiciones económicas de un cliente y una empresa varían en el tiempo, se debe tener claro que los patrones también cambian con frecuencia, por lo tanto se debe contar con el apoyo y la participación de las áreas expertas para la formulación de reglas, con el fin de monitorear periodicamente posibles comportamientos que permitan enriquecer el modelo. Por consiguiente, es fundamental considerar que los modelos que proveen las técnicas de minería de datos deben ser revisados y reentrenados con cierta frecuencia, para actualizarlos con los nuevos datos y variables en caso de ser necesario.

Se debe entender que en los procesos de minería de datos se puede hacer uso de distintas tecnologías para apoyarse en la identificación de patrones que faciliten descubrir comportamientos ya sea en los clientes para medir su rentabilidad o para cualquier otro tipo de problema según la

investigación llevada a cabo y muchas veces depende en gran medida de las características de los datos disponibles para el estudio.

En minería de datos se pueden hacer diversos trabajos con herramientas de analítica avanzada, por ejemplo, se puede realizar un análisis de serie de tiempo y predecir un comportamiento de los datos preferiblemente para uno, dos, tres o hasta siete días; no se recomiendan más porque puede dar un resultado sesgado en la predicción.

Recomendaciones

La empresa debe efectuar un proceso de licitación para adquirir la licencia servidor de RStudio y tener un ambiente robusto para la carga, manipulación y procesamiento de la información, máxime si quiere aplicar algún análisis descriptivo o predictivo en tiempo real sobre los datos.

En cuanto a la implementación en producción, en la sección "Implementación" del capítulo V se diseña y propone una arquitectura acorde también a las capacidades que en la actualidad se poseen en la empresa, para no incurrir en la adquisición de otros equipos innecesarios; por eso se aconseja usar las bases de datos ya existentes, solo se debe adquirir la herramienta para minería, que en este caso se plantea utilizar R y RStudio.

Se debe contar con procesos de mejora continua en la calidad de los datos en las fuentes que alimentan el DWHC y en los datos contenidos en este; estos procesos deben automatizarse y revisarse periódicamente. Además cuando existan datos que no se puedan resolver por medio de las herramientas tecnológicas, se debe asignar a uno o más colaboradores las tareas de revisión de las inconsistencias generadas.

Es recomendable extraer los datos por utilizar para probar el modelo en un archivo o hacer una tabla en la base de datos que los contenga y no usar los que están directamente en el almacén de datos general de la empresa, porque pueden afectar los resultados y el rendimiento si se requieren llevar a cabo cambios o transformaciones a los datos, o bien, afectar a otros usuarios que hacen consultas constantes de las estructuras de la base de datos para obtener otro tipo de información.

Por último, realizar un análisis de importancia sobre las variables del conjunto de datos y distintas ejecuciones para probarlas en el proceso con cada método, o sea si el algoritmo ***"nnet"*** genera que las variables por usar son ***"x,y,z"*** y el modelo ***"svm"*** genera que son ***"y,z,x"***, entonces se usen con el orden que cada algoritmo indica, para analizar y probar los resultados por separado y tomar la decisión de cuál es el modelo predictivo más óptimo para los datos.

Los valores de aceptabilidad que se definen apegados a la metodología CRISP-DM se alcanzan en un 95 %, lo cual convierte a este proyecto y el modelo predictivo resultante del análisis en un algoritmo óptimo, por consiguiente se considera factible realizar su implementación dado que cumple con todas las fases explicadas en el capítulo V.

Capítulo VII Reflexiones finales

Los proyectos de minería de datos son muy complejos. Al inicio se llega a subestimar la fase de preparación de los datos, sin embargo durante el desarrollo del proyecto se encuentra que es de las fases más engorrosas porque requiere de muchas pruebas para generar los datos finales del modelo.

Para crear un proyecto de minería de datos, es necesario contar con un grupo interdisciplinario que posibilite establecer los objetivos, analizar los datos e interpretar los resultados.

El gran volumen de información que poseen las empresas del sector financiero demanda la utilización de un sistema altamente robusto, eficiente y escalable, usando tecnologías como Cloudera o Hortonworks.

Capítulo VIII Trabajos a futuro

Con este proyecto se pueden generar nuevas ideas para llevar a cabo diferentes proyectos de minería de datos descriptivos y predictivos, por ejemplo, se puede crear un modelo predictivo para conocer si las oficinas nuevas que se quieren abrir en alguna zona geográfica de Costa Rica -que cumplan con las condiciones similares a las existentes- van a ser o no rentables, con el objetivo de tomar decisiones oportunas para decidir si es conveniente abrirlas y darles el impulso económico.

Crear un modelo analítico que ayude a efectuar una detección proactiva de fuga de clientes, para tomar acciones que permitan sostenerlos dentro de la empresa, complementado con un modelo predictivo que determine el momento en que un cliente cambie de segmento superior o inferior para apoyar campañas de fidelización o mejorar la atención a clientes pertenecientes a segmentos con niveles preferenciales.

Gestionar un análisis aglomerativo de clientes y productos mediante los datos históricos, usando algoritmos y reglas de asociación, para sustituir el método de distribución por quintiles; además de gestionar la clasificación de los clientes nuevos según el análisis de clúster creado, para ubicarlos correctamente en el segmento que pertenecen.

Crear procesos de análisis de datos no estructurados para estudiarlos mediante series de tiempo y así determinar dónde se presentan con anterioridad o a futuro los mayores cambios de indicadores financieros de la empresa.

Crear modelos predictivos que posibiliten detectar en cuáles períodos existe un mayor retiro de dinero en colones y dólares de los cajeros automáticos para abastecerlos en un tiempo oportuno y evitar molestias en los clientes por falta de dinero.

Medir por medio de series de tiempo los períodos donde hay mayor afluencia y concurrencia de clientes a los cajeros de las oficinas comerciales, para establecer cuáles servicios se utilizan en esos períodos y si es conveniente poner más personal y/o automatizar servicios.

Glosario

Big data: Son aquellos datos cuya escala, distribución, diversidad y oportunidad requieren del uso de nuevas técnicas y arquitecturas para facultar de ideas y nuevas fuentes de valor al negocio. No es un término asociado solo a volumen, se debe tener una visión más amplia de lo que se puede obtener analizando los tipos de datos. En la actualidad es frecuente encontrar muchos artículos donde hablan de las tres "*Vs*" del *Big Data*, refiriéndose a volumen, variedad y velocidad, pero existen muchas personas pioneras en la implementación que amplían el concepto agregando nuevas características como la veracidad y el valor; entonces ya no solo se debe hablar de las tres "*Vs*", sino de las cinco "*Vs*" del *Big Data*. Marr (2014) habla de la importancia de la veracidad que consiste en el orden o confiabilidad de los datos y a pesar de tenerse acceso a Big Data, si no se puede convertir en valor, es inútil.

Calibración: Consiste en buscar mejorar un algoritmo minimizando los errores globales y maximizando la asertividad o el área bajo la curva (AUC) ROC.

Conglomerado del Banco de Costa Rica: Es el Banco de Costa Rica más las subsidiarias, BCR SAFI, BCR Valores Puesto de Bolsa, BCR Seguros y BCR Pensiones.

CRISP-DM: Por sus siglas en inglés que significan *Cross-Industry Standard Process for Data Mining*, es un método de proceso integral probado para orientar trabajos de minería de datos. Está dividido en fases con sus respectivas tareas y las relaciones entre ellas.

Curva ROC: Compara la tasa de falsos positivos con la de verdaderos positivos. Las curvas ROC (*Receiver Operating Characteristic*) contrastan la especificidad contra la sensibilidad de un modelo. Se puede clasificar de cuatro formas: si la curva está en la misma función identidad, es como tirar una moneda a la suerte, es decir el modelo no sirve; si está por debajo de la identidad,

es muy malo; si está un poco más arriba de la función identidad, es regular y si está bastante arriba, es considerado como bueno o muy bueno. Si se evalúa el área bajo la curva "AUC" (*Area Under Curve*), se puede determinar la calidad de un modelo, porque se dibuja sobre un rectángulo de área 1. Si el área del AUC es 1, es un modelo perfecto; sin embargo es muy difícil que suceda. En la figura 40 se observa e interpreta como la probabilidad de que ante un par de datos sobre correo electrónico sea uno considerado *spam* y el otro no, el AUC es casi perfecta asegurando que ante una prueba se clasifican correctamente.

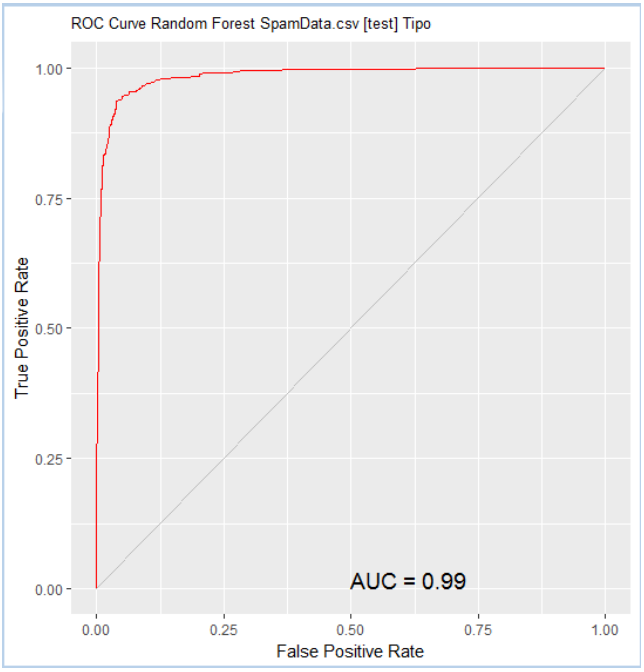


Figura 40. Curva ROC sobre correos spam
Fuente: Elaboración propia

Data Warehouse Corporativo: Es el repositorio de datos generales del Banco Costa Rica, ahí se almacena información histórica desde el año 2014. Se denota por DWHC.

Gobierno de datos: Es una disciplina encargada de la orquestación de gente, procesos y tecnología que permite habilitar a una compañía apalancar la información como un recurso de valor empresarial y, al mismo tiempo, es la encargada de mantener a los usuarios, auditores y reguladores satisfechos, usando la mejora de la calidad de los datos para retener clientes, constituyendo y guiando a nuevas oportunidades en el mercado (IBM, 2012).

Individuos: Es una colección de variables que lo describen, también se le conoce como registro, objeto, entidad, caso, punto o ejemplo de observación.

Modelo determinístico: Un modelo determinístico es un modelo matemático donde las mismas entradas producen invariablemente las mismas salidas, no contemplándose la existencia del azar ni el principio de incertidumbre (Rojas, 2017).

Modelo probabilístico: Un modelo probabilístico es un modelo matemático que ayuda a predecir la conducta de futuras repeticiones de un experimento aleatorio mediante la estimación de una probabilidad de ocurrencia de dicho evento concreto (Rojas, 2017).

Tabla de aprendizaje: Es la tabla con la que se entrena (aprende) el modelo matemático de predicción, o sea, a partir de esta tabla se calcula la función f de la definición anterior.

Tabla de prueba (*Testing*): Es la tabla con la cual se valida el modelo matemático de predicción, es decir, se verifica que los resultados en individuos que no participan en la construcción del modelo son buenos o aceptables.

Variable: Es una propiedad o característica de un individuo, por ejemplo color de cabello, temperatura, estado civil.

Variable continua: Puede tomar cualquier valor en un intervalo dado (son números reales).

Variable cualitativa: Es aquella que clasifica las unidades en categorías. Las categorías pueden tener un orden natural (ordinales) o no (nominales).

Variable cuantitativa: Es cuando los valores de la variable son "números", cada valor posible es menor o mayor que otro valor. Por ejemplo: edad, ingresos, nota en un examen, número de años de educación, kilómetros de distancia entre lugares.

Variable discreta: Es aquella en la cual se puede contar el número posible de valores (son números enteros).

Variables nominales: Aquí los valores son "categorías" y las categorías son valores diferentes por una cualidad, no por una cantidad. Sin embargo, ningún "valor" se puede decir que sea mayor o menor que otro. Por ejemplo: partido político al que vota, región en que vive, sexo, estado civil.

Variables ordinales: Cuando las variables son ordinales, se puede contar con número de casos, comparar entre categorías, pero no se pueden realizar operaciones numéricas, porque son categorías, no números. Tienen un orden, pero no existe una distancia o intervalo definido entre los valores. Por ejemplo: la profesión (bachiller, licenciado, máster, doctor). El tratamiento estadístico que se les da en ocasiones es como variables cualitativas y en otros casos como variables cuantitativas.

Bibliografía

Adkins, T. (2012). Applying Analytics to High Performance Customer Profitability Models in Financial Services. *SAS Global Forum 2012* (págs. 1-9). SAS Institute Inc.

Ahern, M. y Bercher, M. (1982). Customer Profitability Model for Use in Retail Banking. *Financial Review, 17*(2), 79.

Aluja, T. (2001). La mineria de datos, entre la estadistica y la inteligencia artificial. *Universitat Plitécnica de Catalunya, 25*(3), 479-498.

Banco de Costa Rica. (s.f.). *Acerca del BCR.* Recuperado de: http://www.bancobcr.com/acerca%20del%20bcr/index.html

Chavarría, C. (2011). La dicotomía cuantitativo/cualitativo: falsos dilemas en investigación social. *Actualidades en Psicología, 25* (112). Recuperado de: https://revistas.ucr.ac.cr/index.php/actualidades/article/view/70

CIMA. (2001). *Activity-based Management–An Overview*. Recuperado de: http://www.cimaglobal.com/Documents/ImportedDocuments/ABM_techrpt_0401.pdf

Clarke, B., Fokoue, E. y Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning.* Estados Unidos: Springer Science+Business Media.

Cokins, G. y Charles, R. (2011). Increase Customer Proftability Using Data Mining and Advanced Analytics. *SAS*, 1-23.

Correa Uribe, S., Puerta Zapata, A. y Restrepo Gómez, B. (2002). *Módulo seis: Investigación evaluativa. Especialización en teoría, métodos y técnicas de investigación social*. Bogotá, Colombia: Instituto Colombiano para el Fomento de la Educación Superior.

Garrido, L. (2001). Aplicaciones empresariales de Data Mining. *Universitat de Barcelona, 25*(3), 499-508.

Giordano, A. (2010). *Data Integration Blueprint and Modeling: Techniques for a Scalable and Sustainable Architecture.* Estados Unidos: Pearson Education.

Graham, W. (2011). *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge*. Australia: Springer Science+Business Media.

Gruber, T. (1993). *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. Estados Unidos: Universidad de Stanford.

Haenlein, M., Kaplan, A. M. y Beeser, A. (2007). A Model to Determine Customer Lifetime Value in a Retail Banking Context. *European Management Journal, 25*(3), 221-234.

Han, J., Kamber, M. y Pei, J. (2011). *Data Mining Concepts and Techniques,* (3° ed.). Estados Unidos: Morgan Kaufmann.

Hernández, R., Fernández, C. y Baptista, P. (2004). *Metodología de la investigación*. México: McGrawHill.

Holm, M. (2012). *Customer Proftability Measurement Models.* Recuperado de: http://openarchive.cbs.dk/bitstream/handle/10398/8437/Morten_Holm.pdf?sequence=1

IBM. (2012). Seis pasos para el gobierno de datos. Recuperado de: https://www.ibm.com/developerworks/ssa/data/library/techarticle/gobierno-datos/index.html

IBM Corporation. (2010). *Aplicación del análisis predictivo para mejorar la gestión de las relaciones con los clientes.* Nueva York, Estados Unidos: Autor.

Kaplan, R. S. y Cooper, R. (1998). *Cost and effect: Using integrated cost systems to drive profitability and performance.* Massachusetts, Estados Unidos: Harvard Business School Press.

Knight, R. (1975). Customer Profitability Analysis-Part I: Alternative Approaches Toward Customer Profitability. *Economic Review*, 11-20.

Kotler, P. y Keller, K. (2012). *Direccion de Marketing,* (14° ed.). México: Pearson Educación.

Marr, B. (2014). *Big Data: Los 5 Vs que todo el mundo debe saber*. Recuperado de: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/

McNab, D. (2005). *What is customer profitability and why should we measure it ?* Recuperado de: https://www.the-cma.org/disciplines/analytics/archive/what-is-customer-profitability

Medina Veloz, G., Luna Rosas, F. J., Tavarez Avendaño, J. F. y Narváez Murillo, R. (2015). Calibración y selección del modelo de aprendizaje no supervisado K-Medias, de una encuesta sobre factores de riesgo en el consumo de drogas entre estudiantes. *Revista de Análisis Cuantitativo y Estadístico, 3*(7), 1-9. Recuperado de: http://www.ecorfan.org/bolivia/researchjournals/Analisis_Cuantitativo_y_Estadistico/vol3num7/Revista_de_Analisis_Cuantitativo_V3_N7_1.pdf

Mellman, M. (1963). Marketing Cost Analysis-Development and Current Practices. *The Accounting Review, 38*(1), 118-123.

Morelo Tapias, K. (2014). *Sistema para caracterización de perfiles de clientes de la empresa Zona T*. (Tesis de Ingeniería de Sistemas). Universidad de Cartagena. Clombia.

Núñez Vásquez, M. (2013). *Herramientas de inteligencia de negocios aplicadas al análisis de la información educativa del sistema PIAD, en Costa Rica.* (Trabajo de investigación para

optar por la maestría en Tecnologías de Bases de Datos). Universidad Cenfotec. Costa Rica.

Padrón, J. (1992). *Paradigmas de investigación en ciencias sociales*. Venezuela: Universidad Nacional Abierta de Venezuela.

PowerData. (2017). *Integración de datos: concepto e importancia en la empresa actual.* Recuperado de: https://www.powerdata.es/integracion-de-datos

Rajan, R., Simonson, E. y Jain, A. (2014). *Customer Lifetime Value Analytics in Retail Banking.* Recuperado de: http://www.genpact.com/docs/default-source/resource-/customer-lifetime-value-analytics-in-retail-banking---staying-competitive-and-profitable

Rakotomamonjy, A. y Gasso, G. (s.f.). *Introduction au Data-Mining*. INSA Rouen -Département ASI.

Real Academia Española. (2017). *Diccionario de la lengua española*. España: Autor.

Research, J., Shetha, J. y Sisodia, R. (2001). *Marketing productivity Issues and analysis.* Recuperado de: http://rajsisodia.com/raj/wp-content/uploads/2012/07/Marketing-Productivity-Issues-and-Analysis.pdf

Rojas Rodríguez, O. (2017). *Conceptos básicos de minería de datos*. Costa Rica.

Serrahima Formosa, R. (2011). *La rentabilidad financiera.* Recuperado de: https://raimon.serrahima.com/wp-content/uploads/2011/03/053_Rentabilidad.pdf

Sevin, C. (1965). *Marketing Productivity Analysis.* Misuri, Estados Unidos: McGraw-Hill.

The CRISP-DM consortium. (2000). *CRISP-DM 1.0.*

Thomas, C. (1992). *Data Quality Management and Technology.* Estados Unidos: Bantam Books.

Villanueva, J. y Hanssens, D. M. (2007). Customer Equity: Measurement, Management and Research Opportunities. *Foundations & Trends in Marketing, 1*(1), 1-96.

Wikipedia. (2018). *Validación cruzada*. Recuperado de: https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada

Witten, I., Frank, E. y Hall, M. (2011). *Data Mining Practical Machine Learning Tools and Techniques*, (3° ed.). Estados Unidos: Morgan Kaufmann.

Apéndices

Como se explica durante todo este documento, para la realización del proyecto se usa la herramienta de *software* libre R y RStudio. En dichas herramientas se desarrollan todas las fases de CRISP-DM que abarcan desde la preparación de los datos hasta el modelado y evaluación, generando varios archivos con formato HTML para analizar los resultados, los cuales se pueden observar en la tabla 12.

Tabla 12. Archivos HTML creados en el desarrollo del proyecto

Preparación	Discretización	ACP	Modelado	Evaluación
 TFG_1_PrepararDatos.html	 TFG_2_Discretizar.html	 TFG_3_ACP.html	 TFG_4_Paralelismo_V1.html  TFG_4_Paralelismo_V4.html	 TFG_5_Evaluacion.html

Fuente: Elaboración propia