UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS

FACULTAD DE INGENIERÍA

PROGRAMA ACADÉMICO DE INGENIERÍA DE SISTEMAS DE INFORMACIÓN

Modelo de análisis predictivo para abandono de clientes en una empresa administradora de fondos colectivos

TESIS

Para optar el título profesional de Ingeniero de Sistemas de Información

AUTORES

Estrada Valderrama, Alvaro Antonio (0000-0003-0780-5197)

Cortez Acosta, Leandro Jesus (0000-0003-2785-1166)

ASESOR

Barrientos Padilla, Alfredo (0000-0002-0029-4913)

Lima, 23 de enero de 2023

DEDICATORIA

Esta investigación se la dedicamos a nuestras familias por todo el apoyo y comprensión del tiempo en la elaboración de este trabajo

AGRADECIMIENTOS

Se agradece plenamente a la organización que se usó de base para el presente estudio que confió en nosotros y nos abrió sus puertas para este proyecto. Asimismo, a cada uno de los profesores que nos guiaron y reforzaron nuestros conocimientos para la elaboración de este trabajo.

RESUMEN

En la actualidad, el rubro de los fondos colectivos está tomando cada vez más posicionamiento en Sudamérica como una alternativa al crédito vehicular de los bancos, una modalidad en la cual las personas a través de aportes mensuales y de modalidades de sorteo y remates mensuales, pueden lograr el sueño del auto o la casa propia. Sin embargo, este sector se está viendo afectado por el gran porcentaje de clientes que dejan de pagar sus cuotas durante 3 meses y la organización los considera como clientes que han abandonado al fondo colectivo. Esta situación va aumentando cada vez más y las entidades en este sector no se han arriesgado a optar por soluciones soportadas en tecnologías y técnicas de machine learning por lo que actualmente toman decisiones basadas en los datos que proporciona una entidad externa llamada Equifax, la cual brinda información acerca del segmento de cada cliente, sin embargo este proceso no está trayendo buenos resultados ya que cada vez más aumenta el porcentaje de clientes que ingresan a la empresa y luego abandonan, perjudicando así a los grupos en los que se encuentran y a los ingresos mensuales de la organización. Es por ello, que el presente proyecto busca desarrollar un modelo de análisis predictivo basado en machine learning, que permita identificar el comportamiento de los clientes desertores y así poder predecir el porcentaje de probabilidad que tiene cada cliente de abandonar a la organización, con el fin de que los autores involucrados en el proceso de negocio de retención de clientes puedan tomar decisiones y orientas sus campañas comerciales basándose en datos.

Palabras clave**:** abandono de clientes; algoritmos supervisados de clasificación; aprendizaje de máquina; modelo predictivo; servicios financieros

ABSTRACT

Currently, the category of collective funds is gaining more and more position in South America as an alternative to vehicle credit from banks, a modality in which people, through monthly contributions and raffle and monthly auctions modalities, can achieve the dream of owning a car or home. However, this sector is being affected by the large percentage of clients who stop paying their installments for 3 months and the organization considers them as clients who have abandoned the collective fund. This situation is going to increase more and more and the entities in this sector have not risked opting for solutions supported by machine learning technologies and techniques, which is why they currently make decisions based on the data provided by an external entity called Equifax, which provides information about the segment of each client, however this process is not bringing good results since the percentage of clients that enter the company and then leave is increasing more and more, thus harming the groups in which they belong and the organization monthly income. For this reason, this project seeks to develop a predictive analysis model based on machine learning, which allows identifying the behavior of deserting customers and thus being able to predict the percentage probability that each customer has to abandon the organization, in order to that the authors involved in the customer retention business process can make decisions and guide their commercial campaigns expanded on data.

Keywords: customer churn; supervised classification algorithms; machine learning; predictive model; financial services

N°5591_Modelo de análisis predictivo para abandono de clientes en una empresa administradora de fondos colectivos

INFORME DE ORIGINALIDAD

17%	17%	7%	6%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	upc.aws.openrepository.com Fuente de Internet	4%
2	repositorioacademico.upc.edu.pe Fuente de Internet	3%
3	www.scribd.com Fuente de Internet	1%
4	hdl.handle.net Fuente de Internet	1%
5	www.mdpi.com Fuente de Internet	1%
6	isidore.science Fuente de Internet	<1%
7	www.analytics10.com Fuente de Internet	<1%
8	journalofbigdata.springeropen.com Fuente de Internet	<1%
9	expeditiorepositorio.utadeo.edu.co Fuente de Internet	

		<1 %
10	www.researchgate.net Fuente de Internet	<1 %
11	mdpi-res.com Fuente de Internet	<1 %
12	link.springer.com Fuente de Internet	<1 %
13	repositorio.unicauca.edu.co:8080 Fuente de Internet	<1 %
14	Jan Kozak, Krzysztof Kania, Przemysław Juszczuk, Maciej Mitręga. "Swarm intelligence goal-oriented approach to data-driven innovation in customer churn management", International Journal of Information Management, 2021 Publicación	<1 %
15	repositorio.furg.br Fuente de Internet	<1 %
16	eprints.ucm.es Fuente de Internet	<1 %
17	Reyes García Carlos Tomás. "La minería de datos como herramienta para la toma de decisiones en el proceso de calendarización de cursos de cómputo", TESIUNAM, 2007 Publicación	<1 %

18	reunir.unir.net Fuente de Internet	<1 %
19	www.americaspg.com Fuente de Internet	<1 %
20	Submitted to University of East London Trabajo del estudiante	<1 %
21	www.jove.com Fuente de Internet	<1 %
22	Submitted to Universidad Internacional de la Rioja Trabajo del estudiante	<1 %
23	dspace.espol.edu.ec Fuente de Internet	<1 %
24	pure.ewha.ac.kr Fuente de Internet	<1 %
25	bibliotecadigital.econ.uba.ar Fuente de Internet	<1 %
26	www.igi-global.com Fuente de Internet	<1 %
27	cybertesis.unmsm.edu.pe Fuente de Internet	<1 %
28	ouci.dntb.gov.ua Fuente de Internet	<1 %
29	www.emerald.com Fuente de Internet	

		<1 %
30	www.smv.gob.pe Fuente de Internet	<1 %
31	digibuo.uniovi.es Fuente de Internet	<1 %
32	www.fonbienes.com.pe Fuente de Internet	<1 %
33	repositorio.cuc.edu.co Fuente de Internet	<1 %
34	www.cacic2016.unsl.edu.ar Fuente de Internet	<1 %
35	qdoc.tips Fuente de Internet	<1 %
36	t21.pe Fuente de Internet	<1 %
37	www.slideshare.net Fuente de Internet	<1 %
38	Nhi N.Y. Vo, Shaowu Liu, Xitong Li, Guandong Xu. "Leveraging unstructured call log data for customer churn prediction", Knowledge-Based Systems, 2021 Publicación	<1 %
39	Submitted to Universidad Andina del Cusco Trabajo del estudiante	<1 %

40	Submitted to UTEC Universidad de Ingeniería & Tecnología Trabajo del estudiante	<1%
41	repositorio.uchile.cl Fuente de Internet	<1%
42	Submitted to California Southern University Trabajo del estudiante	<1%
43	rcs.cic.ipn.mx Fuente de Internet	<1%
44	www.pcsignos.com.ar Fuente de Internet	<1%
45	Submitted to Ajou University Graduate School Trabajo del estudiante	<1%
46	Sebastiaan Höppner, Eugen Stripling, Bart Baesens, Seppe vanden Broucke, Tim Verdonck. "Profit driven decision trees for churn prediction", European Journal of Operational Research, 2018 Publicación	<1%
47	Submitted to University of Bradford Trabajo del estudiante	<1%
48	repositorio.uniandes.edu.co Fuente de Internet	<1%
49	dokumen.pub Fuente de Internet	<1%

50	www.businessempresarial.com.pe Fuente de Internet	<1 %
51	www.scitechnol.com Fuente de Internet	<1 %
52	Submitted to Eynesbury Institute of Business and Technology Trabajo del estudiante	<1 %
53	dehesa.unex.es:8443 Fuente de Internet	<1 %
54	fdocuments.us Fuente de Internet	<1 %

Excluir citas	Apagado	Excluir coincidencias	< 20 words
Excluir bibliografía	Activo		

TABLA DE CONTENIDOS

1 DESCRIPCIÓN DEL PROYECTO 18

1.1 ANTECEDENTES 18
1.2 DOMINIO DEL PROBLEMA 18
1.3 PLANTEAMIENTO DE LA SOLUCIÓN 19
1.4 OBJETIVOS DEL PROYECTO 19
1.4.1 Objetivo General 19
1.4.2 Objetivos específicos 19
1.5 PLANIFICACIÓN DEL PROYECTO 19
1.5.1 Gestión del alcance 20
1.5.2 Gestión del tiempo 20
1.5.3 Gestión de recursos humanos 21
1.5.4 Gestión de comunicaciones 22
1.5.5 Gestión del riesgo 22

2 LOGROS DE LOS STUDENT OUTOMES 24

2.1 STUDENT OUTCOME (1) 24
2.1.1 Descripción 24
2.1.2 Evidencia 24
2.2 STUDENT OUTCOME (2) 24
2.2.1 Descripción 24
2.2.2 Evidencia 24
2.3 STUDENT OUTCOME (3) 26
2.3.1 Descripción 26
2.3.2 Evidencia 26
2.4 STUDENT OUTCOME (4) 26
2.4.1 Descripción 26
2.4.2 Evidencia 26
2.5 STUDENT OUTCOME (5) 26
2.5.1 Descripción 26
2.5.2 Evidencia 27
2.6 STUDENT OUTCOME (6) 27
2.6.1 Descripción 27
2.6.2 Evidencia 27
2.7 STUDENT OUTCOME (7) 27
2.7.1 Descripción 27
2.7.2 Evidencia 27

3 MARCO TEÓRICO 28

3.1 EMPRESAS ADMINISTRADORAS DE FONDOS COLECTIVOS (EAFC) 28
3.2 ABANDONO DE CLIENTES 28
3.3 PREDICCIÓN DEL ABANDONO DE CLIENTES 29
3.3.1 Decision tree 30
3.3.2 Random forest 31

3.3.3 Logistic Regression 32
3.4 CRISP-DM 32
3.5 SUPERINTENDENCIA DEL MERCADO DE VALORES 33
3.6 REGLAMENTO DEL SISTEMA DE FONDOS COLECTIVOS Y DE SUS EMPRESAS ADMINISTRADORAS 33
4 ESTADO DEL ARTE 36
4.1 REVISIÓN DE LA LITERATURA 36
4.2 METODOLOGÍA 36
4.2.1 Planificación 36
4.2.2 Desarrollo 39
4.2.3 Resultados 42
4.3 CONCLUSIONES 116
4.3.1 Conclusiones de artículos de la categoría de abandono de clientes 116
4.3.2 Conclusiones de artículos de la categoría de modelos predictivos 117
4.3.3 Conclusiones de artículos de la categoría de modelos predictivos para abandono de clientes 117
4.3.4 Conclusiones generales 118
5 DESARROLLO DEL PROYECTO 120
5.1 ANÁLISIS 120
5.1.1 Análisis de la necesidad 120
5.1.2 Análisis de herramientas para análisis predictivo 120
5.1.3 Análisis de algoritmos para análisis predictivo 123
5.1.4 Análisis de variables predictoras de abandono de clientes en empresas financieras 126
5.2 DISEÑO DEL MODELO 127
5.2.1 Subprocesos y roles 130
5.2.2 Entradas 130
5.2.3 Análisis y desarrollo 132
5.2.4 Salidas 141
5.2.5 Arquitectura de capas 143
6 VALIDACIÓN DE LA PROPUESTA 145
6.1 VALIDACIÓN DE FACTIBILIDAD TÉCNICA 145
6.1.1 Requerimientos de hardware 145
6.1.2 Recursos humanos 146
6.2 VALIDACIÓN DE FACTIBILIDAD ECONÓMICA 147
6.2.1 Costo de RRHH 147
6.2.2 Costo de servicio de enriquecimiento de datos de Equifax 149
6.2.3 Costo de licencias 151
6.2.4 Costo total del proyecto 152
6.2.5 Estructura de ganancias por cliente 152
6.2.6 Validación de viabilidad económica y rentabilidad mediante VAN y TIR 153
7 CONCLUSIONES 156

8 RECOMENDACIONES 157
9 BIBLIOGRAFÍA 158

ÍNDICE DE TABLAS

Tabla 1 Gestión del tiempo del proyecto. 21
Tabla 2 Gestión del tiempo del proyecto. 22
Tabla 3 Gestión de comunicaciones. 23
Tabla 4 Tabla de artículos científicos revisados 40
Tabla 5 Listado de algoritmos referenciados en los artículos científicos analizados. 124
Tabla 6 Listado de variables predictivas referenciadas en los artículos científicos analizados. 126
Tabla 7 Lista de subprocesos y roles asignados. 130
Tabla 8 Tabla de variables propuestas por el modelo predictivo. 131
Tabla 9 Tabla de variables con el tipo de gráfico propuesto para su análisis. 133
Tabla 10 Requerimientos de hardware del servidor 145
Tabla 11 Requerimientos mínimos para computadora de escritorio o laptop 146
Tabla 12 Requisitos de tiempo de experiencia de los roles 147
Tabla 13 Listado de roles y subprocesos en los que intervienen. 147
Tabla 14 Cuadro resultante de estimación de costos. 149
Tabla 15 Necesidad y costo de enriquecimiento de datos por Equifax 151
Tabla 16 Costo total del proyecto. 152
Tabla 17 Ingresos y ganancia neta por mes producto de la inversión 153
Tabla 18 Flujo de caja de los próximos seis bimestres 154
Tabla 19 Parámetros para cálculo de VAN y TIR 154

ÍNDICE DE FIGURAS

Figura 1 Porcentaje de clientes resueltos por año en una empresa de fondos colectivos... 19
Figura 2 Organigrama RRHH del proyecto..... 21
Figura 3 Problemática de abandono de clientes en el sector de fondos colectivos. 24
Figura 4 Arquitectura de capas de la solución propuesta. 25
Figura 5 Diseño del modelo de análisis predictivo propuesto..... 25
Figura 6 Ejemplificación del ciclo de funcionamiento de los fondos colectivos. 28
Figura 7 Fórmula de tasa de abandono anual en empresas administradoras de fondos colectivos. 29
Figura 8 Ejemplificación de cálculo de tasa anual de abandono de clientes. 29
Figura 9 Ejemplo de proceso de trabajo de un modelo para abandono de clientes. 30
Figura 10 Formula asociado al algoritmo Decision Tree 31
Figura 11 Fórmula de cálculo para el índice Gini 31
Figura 12 Modelo CRISP-DM 32
Figura 13 Componentes del framework de búsqueda PICOC 37
Figura 14 Desarrollo de los componentes del framework de búsqueda PICOC para la RQ1. 38
Figura 15 Desarrollo de los componentes del framework de búsqueda PICOC para la RQ2. 38
Figura 16 Desarrollo de los componentes del framework de búsqueda PICOC para la RQ3. 38
Figura 17 Desarrollo de los componentes del framework de búsqueda PICOC para la RQ4. 38
Figura 18 Estructura del framework propuesto. 43
Figura 19 Listado de variables seleccionadas. 44
Figura 20 Distribución de los resultados con el conjunto de datos experimental. 45
Figura 21 Ejemplo de variables TF-IDF. 47
Figura 22 Ejemplificación de la relación existente entre los términos. 47
Figura 23 Modelo esemble multi-stacking basado en machine learning. 49
Figura 24 Modelo propuesto por los autores. 50
Figura 25 Tabla de resultados obtenidos en el sistema propuesto LOGIT_ACT 51
Figura 26 Gráfica de resultados obtenidos en el sistema propuesto LOGIT_ACT. 51
Figura 27 Listado de acciones preventivas propuestas por los autores. 52
Figura 28 Listado de resultados post implementación del sistema preventivo y de realización de acciones preventivas. 53
Figura 29 Variables de un conjunto de datos de una empresa Sur Coreana de telecomunicaciones..... 56
Figura 30 Gráfico de árboles para la predicción de abandono construidos con EvTree and ProfTree. 57
Figura 31 Ranking promedio de cada clasificador sobre los diferentes conjuntos de datos para varias métricas de rendimiento. 58
Figura 32 Metodología CRISP-DM 59
Figura 33 Parámetros de rendimiento de los modelos testeados en la investigación. 60
Figura 34 Análisis de costo – beneficio. 60
Figura 35 Framework del trabajo experimental de la investigación. 62

Figura 36 Descripción de los conjuntos de datos para la predicción de abandono de clientes. 63
Figura 37 Rendimiento de los modelos de predicción en el conjunto de datos 1. 64
Figura 38 Rendimiento de los modelos de predicción en el conjunto de datos 2. 64
Figura 39 Rendimiento de los modelos de predicción en el conjunto de datos 1 balanceado. 65
Figura 40 Rendimiento de los modelos de predicción en el conjunto de datos 2 balanceado. 65
Figura 41 Arquitectura del sistema de predicción de abandono propuesto. 67
Figura 42 Arquitectura de configuración de sistema Apache Flume. 68
Figura 43 Distribución de algunas variables categóricas de la investigación. 69
Figura 44 Periodos de datos históricos y futuros. 70
Figura 45 Comparación de resultados AUC antes y después de adicionar SNA a las variables estadísticas. 71
Figura 46 Resultados AUC para cada algoritmo de clasificación en el conjunto de datos "NotOffered". 72
Figura 47 Proceso propuesto por la investigación. 73
Figura 48 Proceso general de trabajo CCPBI-TAMO. 74
Figura 49 Estructura de LSTM. 75
Figura 50 Diagrama de flujo del algoritmo SFO. 76
Figura 51 Framework propuesto por la investigación para la predicción de abandono de clientes. 78
Figura 52 Características de los conjuntos de datos aplicados al estudio. 79
Figura 53 Primeros resultados de la investigación para cada algoritmo. 80
Figura 54 Segundos resultados de la investigación para cada algoritmo. 81
Figura 55 Terceros resultados de la investigación para cada algoritmo de clasificación híbrido. 81
Figura 56 Arquitectura de sistema propuesta en la investigación. 82
Figura 57 Framework modelo multi fase para desarrollar y gestionar el desarrollo de un modelo predictivo de abandono de clientes. 83
Figura 58 Resultados de validación cruzada k-fold para todos los modelos. 84
Figura 59 Comparación de modelos machine learning. 85
Figura 60 Modelo propuesto basado en mecanismo de atención LSTM. 86
Figura 61 Propuesta de configuración de hiperparámetros. 87
Figura 62 Comparación de resultados obtenidos por cada algoritmo. 88
Figura 63 Ejemplo de árbol de decisión entrenado en el modelo random forest. 89
Figura 64 Resultados de las principales métricas de rendimiento usadas para evaluar los modelos predictivos. 90
Figura 65 Gráfica de resultados para la curva AUC. 91
Figura 66 Listado de variables con su importancia. 92
Figura 67 Listado de Factores o variables y los papers en los que han sido referenciados según la literatura. 93
Figura 68 Listado de parámetros y sus descripciones. 95
Figura 69 Fases del modelo propuesto por la investigación. 97
Figura 70 Cortes establecidos para el score obtenido con el modelo de machine learning. 98

Figura 71 Hiperparámetros óptimos para cada método. 99
Figura 72 Comparación de resultados por cada modelo machine learning de la investigación. 100
Figura 73 Metodología propuesta por los autores. 104
Figura 74 Enfoque basado en CNN. 105
Figura 75 Proceso aplicativo en la investigación. 106
Figura 76 Definición de periodo independiente y periodo dependiente para la predicción. 107
Figura 77 Variaciones del modelo. 108
Figura 78 Parámetros del modelo. 109
Figura 79 Número de variables promedio incluidas en el modelo final. 110
Figura 80 Resultados en términos de AUC y TDL para las diferencias variaciones de modelos. 110
Figura 81 Proceso para cumplir con los objetivos propuestos por los autores. 112
Figura 82 Listado total de variables. Variable target (primera fila) y atributos seleccionados (filas restantes) utilizados en la tarea supervisada de predicción de abandono de clientes. 114
Figura 83 Estrategia de modelado para construir el modelo de predicción de abandono de clientes. 115
Figura 84 Diagrama de caja de las transacciones de la cuenta de cheques de atributo frente a la variable binaria de destino que indica si el cliente abandonó la entidad financiera. 116
Figura 85 Benchmarking de herramientas de análisis predictivo 122
Figura 86 Cuadrante Gartner 2018 de mejores herramientas para data science. 123
Figura 87 Benchmarking de algoritmos seleccionados. 125
Figura 88 Diseño del modelo de análisis predictivo propuesto. 129
Figura 89 Estrategia para construir el modelo predictivo en base al momento de predicción establecido. 133
Figura 90 Metodología propuesta para un correcto subproceso de calidad de datos. 134
Figura 91 Subproceso propuesto para la transformación de variables. 135
Figura 92 Ejemplificación de conversión de la variable Edad. 136
Figura 93 Proceso de división del conjunto de datos. 137
Figura 94 Ejemplificación de matriz de confusión resultante. 138
Figura 95 Ejemplificación de gráfico de curva AUC. 139
Figura 96 Proceso general para la implementación del modelo predictivo propuesto. 141
Figura 97 Prototipo de dashboard para la visualización de los datos resultantes del modelo. 142
Figura 98 Arquitectura de capas de la solución propuesta. 144
Figura 99 Cuadro de estimación de horas de desarrollo. 148
Figura 100 Capas de datos del servicio Equifax. 150
Figura 101 Tabla de la propuesta enviada por Equifax 151
Figura 102 Captura de pantalla del precio de paquete Designer de Alteryx. 152

1 DESCRIPCIÓN DEL PROYECTO

1.1 Antecedentes

El sector de fondos colectivos está tomando cada vez más presencia en Perú y en Sudamérica en general. La Asociación Automotriz resalta que este tipo de financiamiento de fondos colectivos representa el 7% de la venta total de vehículos livianos nuevos en Perú en el año 2019. Además, Los fondos colectivos que operan en Perú cuentan con más de 56,129 clientes activos al cierre del 2019 distribuidos en las 7 empresas que conforman este rubro. Sin embargo, el rubro viene atravesando una problemática desde ya hace algunos años, el excesivo porcentaje de clientes que abandonan el fondo colectivo. En 2021, la empresa de fondos colectivos sufrió el abandono de 6876 clientes, que representan el 49% de su cartera activa. Numerosos estudios demuestran que evitar la pérdida de clientes ahorra dinero, ya que adquirir nuevos clientes puede costar hasta 5 veces más que retener un cliente existente (Xiao J et.al., 2014).

1.2 Dominio del problema

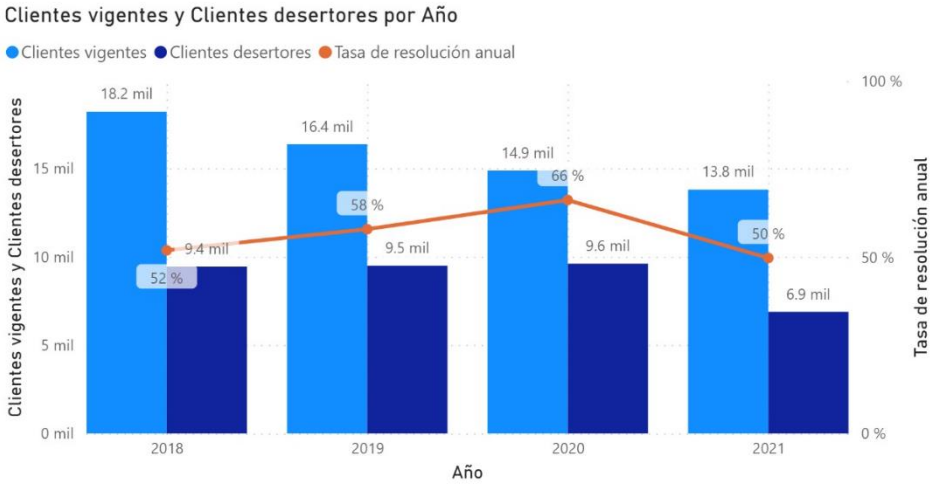
Se demuestra la relevancia del sector de fondos colectivos de estudio en reportes de la Asociación Automotriz donde se resalta que este tipo de financiamiento de fondos colectivos representa el 7% de la venta total de vehículos livianos nuevos en Perú en el año 2019. Los fondos colectivos que operan en Perú cuentan con más de 56,129 clientes activos al cierre de 2019 distribuido en las 7 empresas que conforman este rubro.

Numerosos estudios demuestran que evitar la pérdida de clientes ahorra dinero, ya que adquirir nuevos clientes puede costar hasta 5 veces más que retener a un cliente existente (Xiao J et.al., 2014).

Se realizaron las medidas necesarias para determinar la evolución del abandono en una empresa de fondos colectivos y se halló que el abandono promedio anual es del 49% de clientes de la cartera total activa. En 2021, la empresa sufrió el abandono de 6876 clientes, que representan el 49% de su cartera activa.

Figura 1

Porcentaje de clientes resueltos por año en una empresa de fondos colectivos



1.3 Planteamiento de la solución

Se plantea la solución de modelo de análisis predictivo para el abandono de clientes en una empresa administradora de fondos colectivos, la cual busca brindarle a la empresa las herramientas necesarias para poder reducir el porcentaje anual de abandono de clientes.

1.4 Objetivos del proyecto

1.4.1 Objetivo General

Proponer la implementación de un modelo de análisis predictivo de abandono de clientes en una empresa administradora de fondos colectivos.

1.4.2 Objetivos específicos

- OE1: Analizar las diferentes herramientas de predicción y algoritmos de machine learning, así como también las variables predictoras que permitan la implementación de un modelo de análisis predictivo de abandono de clientes.
- OE2: Diseñar un modelo de análisis predictivo que permita pronosticar los clientes con alta probabilidad de abandono.
- OE3: Validar la factibilidad técnica y económica de la implementación del modelo de análisis predictivo en una empresa administradora de fondos colectivos.

1.5 Planificación del proyecto

Con el fin de tener una eficiente gestión del desarrollo del proyecto, se realiza la planificación de este para tener definidos desde un principio las necesidades y entregables

que se van a abarcar, así como también los hitos principales para cada entregable junto con su respectiva prioridad. A su vez, contemplar las responsabilidades de cada Stakeholder del proyecto, cómo será la comunicación con ellos y los riesgos a los que se podrían enfrentar.

1.5.1 Gestión del alcance

El alcance de la solución propuesta tiene las siguientes consideraciones:

- Análisis de modelos predictivos implementados para el abandono de clientes en distintas entidades.
- Definición de las preguntas de investigación a través del método PICOC.
- Análisis de comparativo de herramientas predictivas, algoritmos de Machine Learning y variables predictoras determinantes para la predicción del abandono de clientes.
- Diseño del modelo de la solución de análisis predictivo.
- Diseño del dashboard propuesto para el consumo de los datos predictivos generados por el modelo.
- La información se obtendrá de reportes reales de la organización, ocultando cualquier dato sensible y alterando las cifras y valores multiplicándolos por un factor aleatorio.

1.5.2 Gestión del tiempo

En esta fase se busca definir los principales hitos del proyecto, con el fin de tener las fechas estimadas de los entregables y actividades para su respectivo seguimiento.

Tabla 1

Gestión del tiempo del proyecto.

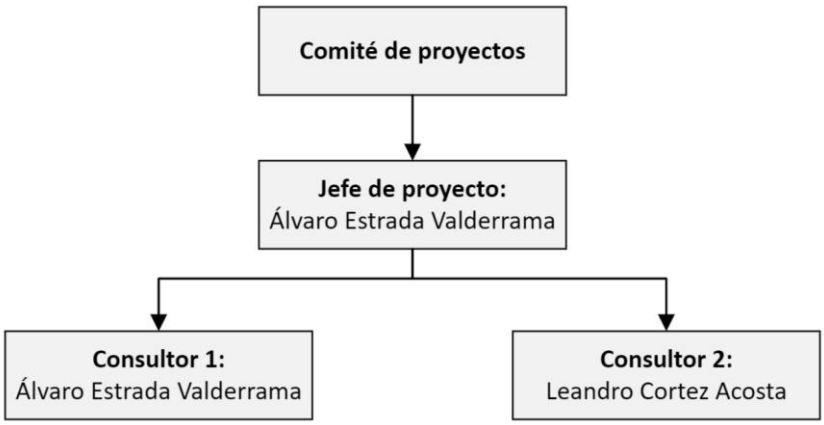
Hito del Proyecto	Fecha Estimada	Entregables Incluidos	Prioridad
Investigación de las soluciones predictivas para abandono de clientes en organizaciones del sector	30/11/2022	Informe de estado del arte de soluciones relevantes para el problema de investigación de predicción de abandono de clientes.	Alta
Diseño del modelo de la solución predictiva	15/12/2022	Informe del diseño de los componentes que conforman el modelo de análisis predictivo propuesto	Alta
Arquitectura de capas de la solución propuesta	01/01/2023	Informe de diseño de la arquitectura de capas de la solución propuesta	Alta
Validación técnica y económica de la solución de modelo de análisis predictivo	27/01/2023	Informe de validación técnica y económica del modelo de análisis predictivo.	Alta

1.5.3 Gestión de recursos humanos

En esta fase se identifican y definen los roles y responsabilidades de las actividades a lo largo del desarrollo del proyecto.

Figura 2

Organigrama RRHH del proyecto.



Para cada uno de los roles especificados previamente, se establecen sus siguientes responsabilidades:

Tabla 2

Gestión del tiempo del proyecto.

Rol	Responsabilidades
Comité de Proyecto	Realizar una revisión de los proyectos de acuerdo a los criterios establecidos por la escuela. Evaluar la calidad del proyecto. Monitorear el cumplimiento de los entregables.
Asesor de tesis	Asesorar el desarrollo de los entregables. Validar la calidad de los entregables.
Jefe de Proyecto	Llevar a cabo la ejecución del proyecto. Registrar los resultados de las investigaciones en documentación. Garantizar el cumplimiento del cronograma de entregas.
Consultor	Desarrollar la investigación pertinente para el proyecto. Desarrollar la propuesta de diseño e implementación de la solución.

1.5.4 Gestión de comunicaciones

En esta etapa, el objetivo es establecer las directrices necesarias para lograr una comunicación clara y definida entre todos los involucrados.

- Para todas las reuniones, se utilizará ya sea la herramienta Google Meet o Blackboard Collaborate, previa coordinación con al menos 24 horas de anticipación.
- Toda la coordinación de trabajos y progresos se llevará a cabo exclusivamente a través del correo electrónico de la universidad.
- Se creará un resumen que recopile los temas tratados en cada reunión.

1.5.5 Gestión del riesgo

En esta etapa se identifican de manera detallada los riesgos potenciales que pueden afectar el desarrollo del proyecto. Cada riesgo se describe junto con su probabilidad de ocurrencia, impacto esperado y la estrategia específica para mitigarlo.

Tabla 3

Gestión de comunicaciones.

#	Riesgo	Probabilidad	Impacto	Acciones
1	Modificaciones en los contratos o políticas de los servicios y aplicaciones empleadas.	Baja	Media	**Aceptar** La empresa asume cualquier impacto de cambios en las políticas, términos o condiciones de las herramientas empleadas ya que no tiene un marco contractual que permita rechazar cualquier cambio.
2	Cambios en la definición y alcance del proyecto.	Media	Media	**Mitigar** Informar regularmente a los interesados sobre el progreso del proyecto con el fin de detectar oportunamente los cambios necesarios para garantizar el cumplimiento de los requisitos establecidos por la institución educativa.
3	Retraso en las entregas de los avances del proyecto.	Media	Alta	**Mitigar** Supervisar el progreso del proyecto para asegurarse de cumplir con los plazos establecidos para la entrega.
4	Pérdida de información relevante para el proyecto.	Baja	Alta	**Mitigar** Guardar los documentos asociados al proyecto en servicios de almacenamiento en la nube (como OneDrive) que cuenten con la capacidad de controlar versiones y recuperar archivos.
5	Falta de disponibilidad de los stakeholders o asesores.	Media	Media	**Mitigar** Informar de manera temprana a las autoridades académicas sobre la escasez de recursos o expertos disponibles.

2 LOGROS DE LOS STUDENT OUTOMES

2.1 STUDENT OUTCOME (1)

2.1.1 Descripción

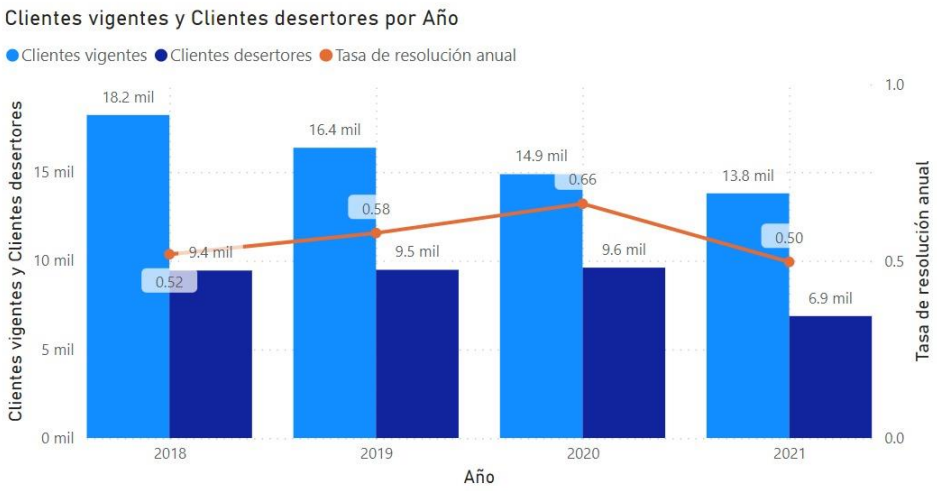
La capacidad de identificar, formular y resolver problemas complejos de ingeniería aplicando los principios de ingeniería, ciencia y matemática.

2.1.2 Evidencia

- Durante la realización de este proyecto, se analizó una problemática real en el sector de fondos colectivos, para la cual se planteó una propuesta de solución basada en principios estadísticos, matemáticos y la ciencia de los datos.

Figura 3

Problemática de abandono de clientes en el sector de fondos colectivos.



2.2 STUDENT OUTCOME (2)

2.2.1 Descripción

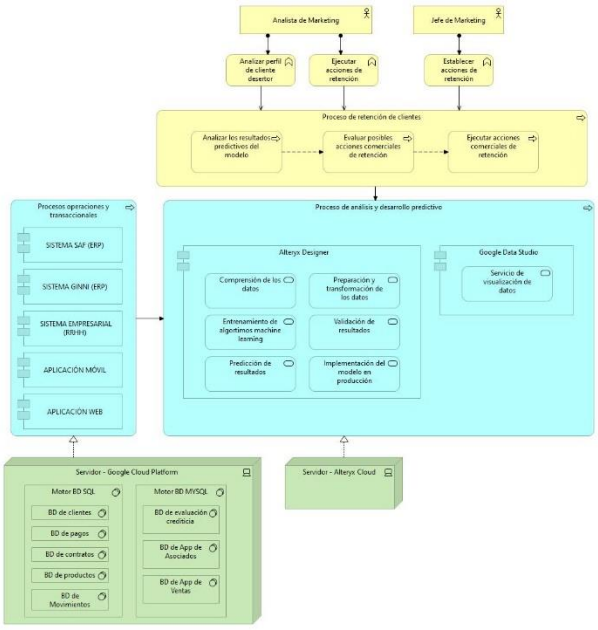
La capacidad de aplicar el diseño de ingeniería para producir soluciones que satisfagan necesidades específicas con consideración de salud pública, seguridad y bienestar, así como factores globales, culturales, sociales, ambientales y económicos.

2.2.2 Evidencia

- Se diseñó una esquematización de la arquitectura de la solución integrada con los componentes actuales de la organización y sus procesos de negocio.

Figura 4

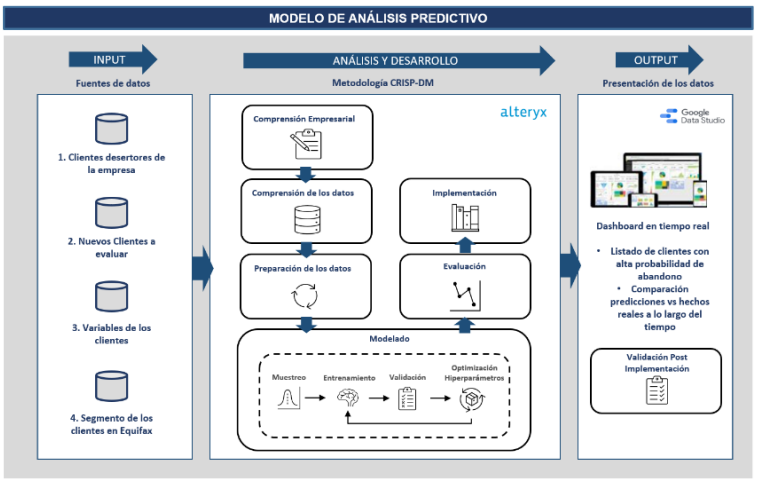
Arquitectura de capas de la solución propuesta.



- Se diseñó un modelo de análisis predictivo para solucionar las necesidades de la organización.

Figura 5

Diseño del modelo de análisis predictivo propuesto.



- Se validó que la propuesta favorece económicamente al crecimiento de las empresas de fondos colectivos a fin de que brinden una mejor oferta de servicios a la comunidad.

2.3 STUDENT OUTCOME (3)

2.3.1 Descripción

Capacidad de comunicarse efectivamente con un rango de audiencias.

2.3.2 Evidencia

- El proyecto fue presentado progresivamente ante distintos perfiles de expertos. Los hitos expuestos ante ellos fueron: la investigación científica, el planteamiento del problema, el diseño de la solución y la validación de esta.
- La totalidad del proyecto fue presentado ante un jurado de sustentación.

2.4 STUDENT OUTCOME (4)

2.4.1 Descripción

Capacidad de reconocer responsabilidades éticas y profesionales en situaciones de ingeniería. Hacer juicios informados, que deben considerar el impacto de las soluciones de ingeniería en contextos globales, económicos, ambientales y sociales.

2.4.2 Evidencia

- Las fuentes empleadas para este proyecto han sido adecuadamente citadas y referenciadas.
- La información sensible compartida por la organización fue empleada de forma responsable; sin crear copias mal intencionadas o filtrarla.
- La validación se hizo con datos generados a partir de los datos reales. Ello con el fin de mantener la confidencialidad de las organizaciones y participantes que figuran en dichos formularios.
- No se grabaron las reuniones virtuales sin la autorización de los participantes de la llamada.

2.5 STUDENT OUTCOME (5)

2.5.1 Descripción

Capacidad de funcionar efectivamente en un equipo cuyos miembros juntos proporcionan liderazgo, crean un entorno de colaboración e inclusivo, establecen objetivos, planifican tareas y cumplen objetivos.

2.5.2 Evidencia

- Se establecieron objetivos específicos para el desarrollo del modelo, los cuales fueron validados con cada uno de los indicadores de éxito.
- Haciendo uso de la gestión del tiempo, se establecieron hitos para los principales entregables del modelo.
- El modelo ha sido desarrollado definiendo roles y responsabilidades claros para cada uno de los autores y participes.

2.6 STUDENT OUTCOME (6)

2.6.1 Descripción

Capacidad de desarrollar y llevar a cabo la experimentación adecuada, analizar e interpretar datos, y usar el juicio de ingeniería para sacar conclusiones.

2.6.2 Evidencia

- El proyecto utilizó herramientas de análisis de datos y visualización de datos para realizar la validación del modelo.
- Los datos brindados por la organización fueron analizados adecuadamente; convirtiéndolos en información digerible, a través de gráficos, para los usuarios de negocio involucrados en el proceso de retención de clientes.
- La sección de conclusiones involucra un análisis detallado de todo el desarrollo de proyecto y de los hallazgos identificados.

2.7 STUDENT OUTCOME (7)

2.7.1 Descripción

Capacidad de adquirir y aplicar nuevos conocimientos según sea necesario, utilizando estrategias de aprendizaje apropiadas.

2.7.2 Evidencia

- Se realizó una investigación y aprendizaje de la metodología CRISP-DM para la propuesta de desarrollo del proyecto.
- Según el análisis realizado, se identificó la herramienta Alteryx como la más idónea para utilizarse en este proyecto, tomando en cuenta las necesidades y las capacidades técnicas de los trabajadores de la organización. Esto llevo al aprendizaje completo de la herramienta, adquiriendo así nuevos conocimientos.

3 MARCO TEÓRICO

El desarrollo del presente capítulo presentará la definición de todos los términos y conceptos esenciales en el ámbito de la inteligencia de negocios. Dentro de la información presentada se detallan los conceptos básicos y específicos del contexto tecnológico del proyecto

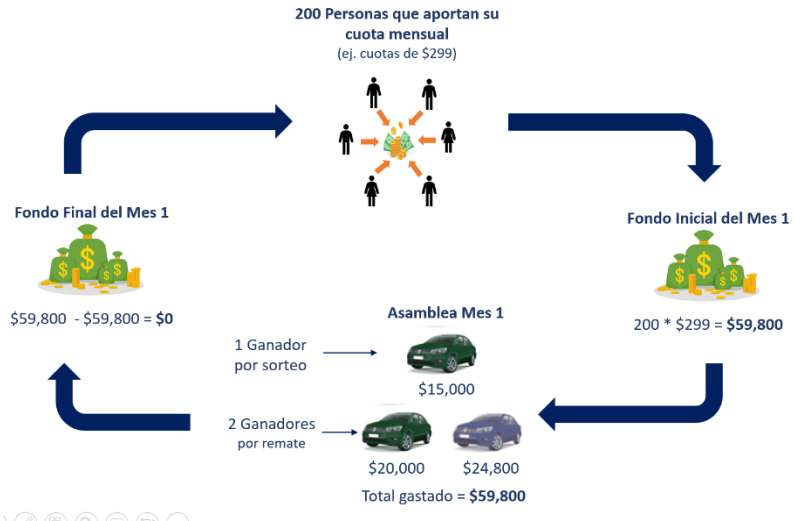
3.1 Empresas administradoras de fondos colectivos (EAFC)

Según la definición de la SMV:

> El sistema de fondos colectivos es una modalidad bajo la cual se adquieren bienes y/o servicios a través de los aportes mensuales de un determinado número de cuotas a pagar periódicamente por las personas naturales o jurídicas asociadas en un Grupo, quienes someten sus intereses individuales respecto de los intereses del Grupo. La adquisición de los bienes y/o servicios objeto del Contrato se da dentro del plazo de vigencia del Grupo, mediante adjudicaciones periódicas a efectuar por sorteo o remate u otra modalidad, las cuales son financiadas con el fondo colectivo constituido por las denominadas "Cuotas Capitales" aportadas por los asociados. (SMV, 2022a, p. 1)

Figura 6

Ejemplificación del ciclo de funcionamiento de los fondos colectivos.



3.2 Abandono de clientes

El abandono de clientes es un problema para la mayoría de las empresas porque afecta directamente los ingresos de la empresa (Kassem et.al., 2020). Los gerentes de marketing se ven obligados a centrar más atención y recursos en la retención de clientes (Tamaddoni et.al.,

2017). Una pequeña mejora en la retención de clientes puede conducir a un aumento significativo en las ganancias (Van den Poel & Lariviere, 2004). Además, varios estudios han demostrado que adquirir un nuevo cliente suele ser de cinco a seis veces más costoso que retener a un cliente existente (Athanassopoulos, 2000).

En el sector específico de empresas administradoras de fondos colectivos el abandono se mide con una tasa anual, la cual se calcula dividiendo el número de clientes desertores en todo el año sobre el promedio mensual de número de clientes de la cartera vigente.

Figura 7

Fórmula de tasa de abandono anual en empresas administradoras de fondos colectivos.

$$tasa\ de\ abandono\ anual = \frac{número\ de\ clientes\ desertores\ anuales}{promedio\ mensual\ de\ clientes\ de\ la\ cartera\ vigente}$$

A continuación, se muestra una figura que ejemplifica el cálculo de la tasa anual de abandono de clientes para un año en específico en una empresa administradora de abandono de clientes.

Figura 8

Ejemplificación de cálculo de tasa anual de abandono de clientes.

	Mes													
	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SET	OCT	NOV	DIC		
Clientes desertores	782	667	747	740	517	769	701	689	743	515	737	730	Suma	8,337
Cartera vigente	16,580	16,529	15,134	15,627	15,481	15,049	16,896	16,463	15,538	16,354	15,140	15,529	Promedio	15,860
													Tasa anual	52.57%

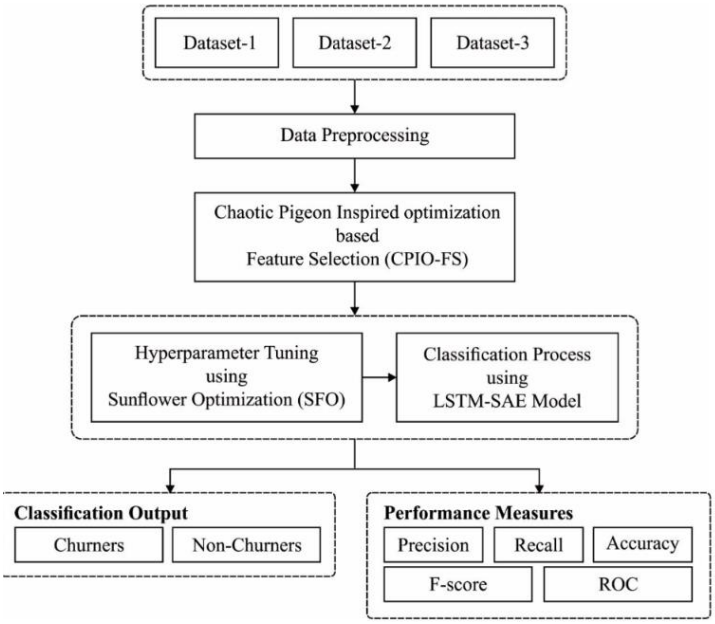
3.3 Predicción del abandono de clientes

La predicción de abandono de clientes (o CCP por sus siglas en inglés) es un tema de investigación con amplio abordaje, ya que tiene una gran relevancia tanto a nivel académico como empresarial. Especialmente a nivel de negocios, parte de la gestión de la relación con el cliente (o CRM por sus siglas en inglés) consta de estrategias de retención del cliente a fin de reducir el abandono, los modelos CCP posibilitan estimar a futuro la probabilidad de abandono de un cliente en base a su información histórica, la cual puede ser estructurada o no estructurada (De Caigny et.al., 2018). Existen diversos algoritmos de machine learning que se han usado en la literatura investigada, como support vector machines (SVM), deep

neural networks (DNN), ensemble gradient boosted trees, entre otros (Munkhdalai et.al., 2019), no obstante los algoritmos decision tree, random forest y logistic regression son los más conocidos e incluso adoptados como estándar para CCP en algunas industrias (De Caigny et.al., 2020) debido a su alto nivel de robustez, facilidad de entender e interpretabilidad de resultados.

Figura 9

Ejemplo de proceso de trabajo de un modelo para abandono de clientes.



De "Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms", por Pustokhina et al., 2021.

3.3.1 Decision tree

Los árboles de decisiones son clasificadores construidos en base a divisiones recursivas en el espacio de instancias. Este consta de un nodo raíz del cual nace el resto del árbol. Los demás nodos siempre tienen una arista entrante, si tiene además arista saliente entonces se trata de un nodo intermedio, si ya no tiene arista saliente entonces es un nodo terminal o también llamado "hoja". Cada nodo intermedio divide el espacio de la instancia en dos o más partes de acuerdo a la función o criterio seleccionada, si se trata de una variable categórica entonces se divide según los valores del atributo, y si se trata de una continua entonces la condición de división será un rango (Rokach & Maimon, 2005).

En 1993, Quinlan usó el nombre “arboles de decisiones” para describir una técnica de “divide y vencerás”. La siguiente formula se utiliza para determinar la entropía y la ganancia de información de cada característica.

Figura 10

Formula asociado al algoritmo Decision Tree

$$h(D) = \sum_{x_{\epsilon} i} -p(i) log_2 p(i)$$

De “An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms”, por Liu et al., 2022.

Donde D es el conjunto de datos, *i* es el conjunto de clases en D, y *p(i)* es la probabilidad de cada clase.

3.3.2 Random forest

Los bosques aleatorios son un tipo de técnica de aprendizaje en conjunto (ensemble) para realizar clasificaciones y regresiones y consta de un modelo entrenado con una gran cantidad de árboles de decisiones independientes para luego producir una predicción en base a la predicción promedio de los árboles individuales (Idris & Khan, 2012).

Para calcular el Índice de Gini, comience con uno y deduzca la suma de las probabilidades al cuadrado para cada clase de ese número. El índice Gini se puede representar formalmente de la siguiente manera:

Figura 11

Fórmula de cálculo para el índice Gini

$$Gini = 1 - \sum_{x-1}^{c} (p_x)^2$$

De “An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms”, por Liu et al., 2022.

Donde *px* representa la probabilidad de que un elemento se asigne a una categoría particular

3.3.3 Logistic Regression

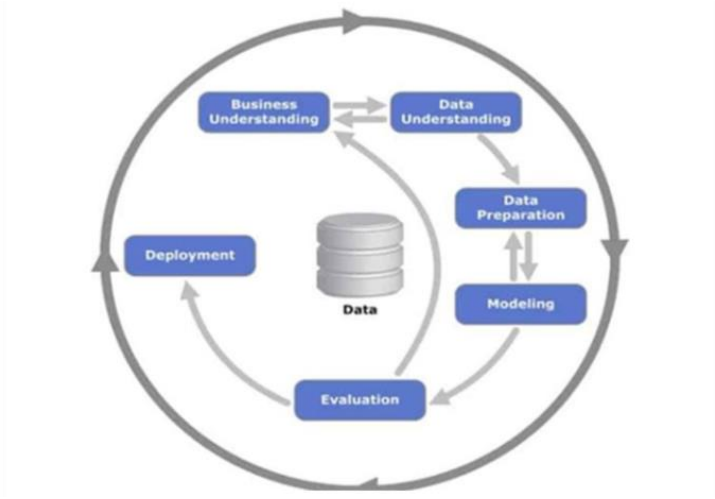
Existen dos tipos de regresión logística, binaria y multinomial, es binaria cuando la variable dependiente es dicotómica. Estos modelos estiman su predicción luego de analizar la relación entre una o más variables independientes, asigna probabilidades a resultados discretos mediante la función Sigmoide, que convierte los resultados numéricos en una expresión de probabilidad entre 0 y 1.

3.4 CRISP-DM

CRISP-DM es una metodología utilizada para el desarrollo de modelos predictivos basada en un enfoque de seis fases: (1) entendimiento de negocio, (2) comprensión de los datos, (3) preprocesamiento de los datos, (4) modelado, (5) evaluación y (6) despliegue (Martínez et al, 2019). La primera fase de entendimiento de negocio ayuda a comprender los objetivos y los requisitos del proyecto, incluyendo la definición del problema. La segunda fase de comprensión de datos ayuda a la familiarización y recopilación de los datos, aquí se identifican problemas relacionados con la calidad de los datos. En la tercera fase de preprocesamiento de los datos se lleva a cabo la selección de variables y registros. En la cuarta fase de modelado, se ejecutan las técnicas, algoritmos y herramientas de minería de datos. La quinta fase de evaluación ayuda a determinar si los resultados cumplieron con los objetivos de negocio e identifica problemas comerciales que deberían haberse abordado antes. Finalmente, en la sexta fase, se despliega el modelo en producción (Nadali et al, 2011).

Figura 12

Modelo CRISP-DM



De “The use of knowledge extraction in predicting customer churn in B2B”, por Jamjoom, 2021.

3.5 Superintendencia del mercado de valores

La Superintendencia del Mercado de Valores (SMV) es una entidad especializada y técnica que forma parte del Ministerio de Economía y Finanzas. Su principal objetivo es “velar por la protección de los inversionistas, la eficiencia y transparencia de los mercados bajo su supervisión, la correcta formación de precios y la difusión de toda la información necesaria para tales propósitos” (SMV, 2022b, p. 1).

Entre las funciones de la SMV se encuentran las siguientes:

- “Dictar las normas legales que regulen materias del mercado de valores, mercado de productos y sistema de fondos colectivos” (SMV, 2022b, p. 1).
- “Supervisar el cumplimiento de la legislación del mercado de valores, mercado de productos y sistemas de fondos colectivos por parte de las personas naturales y jurídicas que participan en dichos mercados” (SMV, 2022b, p. 1).
- “Promover y estudiar el mercado de valores, el mercado de productos y el sistema de fondos colectivos” (SMV, 2022b, p. 1).

3.6 Reglamento del sistema de fondos colectivos y de sus empresas administradoras

La regulación bajo la cual están sujetas las empresas del presente sector es el reglamento del sistema de fondos colectivos y de sus empresas administradoras, emitida y gestionada por la Superintendencia del Mercado de Valores. Este establece las normas bajo las cuales opera el sistema de fondos colectivos y a las que deben sujetarse las empresas administradoras de fondos colectivos, así como las personas que se relacionan directa o indirectamente con las referidas empresas (SMV, 2022a).

Se toma este reglamento como base para la presente propuesta, ya que toda solución que impacte en los procesos internos debe pasar por un control que asegure que se están cumpliendo con todas las normas establecidas.

La SMV especifica en el reglamento las definiciones de algunos términos utilizados en el presente estudio:

- Administradora: “Empresa Administradora de Fondos Colectivos” (SMV, 2022a, p. 1).

- Asociado: “Persona, natural o jurídica, que ha celebrado un Contrato de Administración de Fondos Colectivos y ha efectuado los pagos iniciales previstos en el Contrato” (SMV, 2022a, p. 1).
- Certificado de Compra: “Documento que representa el derecho del Asociado a la aplicación del importe que se consigna en él para la adquisición del bien y/o servicio” (SMV, 2022a, p. 1).
- Contrato: “Contrato de Administración de Fondos Colectivos” (SMV, 2022a, p. 1).
- Cuota Capital: “Importe que el Asociado aporta en forma periódica al Fondo Colectivo, destinado a cubrir el valor del bien y/o servicio materia del Contrato” (SMV, 2022a, p. 1).
- Cuota de Administración: “Importe que el Asociado paga a la Administradora en forma periódica durante la vigencia de cada Contrato, por el servicio de administración de Fondos Colectivos” (SMV, 2022a, p. 1).
- Cuota de Inscripción: “Importe que el Asociado paga por única vez a la Administradora, por el derecho de ingreso a un Grupo por cada Contrato del que sea titular” (SMV, 2022a, p. 1).
- Cuota de Fondo de Reserva: “Importe que abona el Asociado o la Administradora para cubrir los incrementos de precio del bien y/o servicio a adquirir u otro concepto contemplado en el Programa y en el Contrato” (SMV, 2022a, p. 1).
- Cuota de Seguro: “Importe que paga el Asociado por seguro de desgravamen, seguro de vida o cualquier otro riesgo asegurable vinculado al Contrato, de haberse contemplado en el Programa y en el Contrato” (SMV, 2022a, p. 1).
- Cuota Total: “Sumatoria de la Cuota de Administración, la Cuota Capital y, en los casos en que el Programa y el Contrato lo contemple, la Cuota de Fondo de Reserva y/o la Cuota de Seguro” (SMV, 2022a, p. 1).
- Fondo Colectivo:

  Patrimonio autónomo constituido por las Cuotas Capital pagadas por los Asociados en el marco del Contrato, con el objeto de adquirir bienes y/o servicios por acción conjunta de los asociados que lo integran. Igualmente, comprende los resultados de las inversiones de los recursos del Fondo Colectivo, las penalidades y la Cuota de Fondo de Reserva, en los casos

establecidos en el Reglamento y siempre que su existencia haya sido prevista en el Programa y en el Contrato. (SMV, 2022a, p. 1)

- Grupo: “Número determinado de Contratos que conforman un Fondo Colectivo, bajo el alcance de un Programa” (SMV, 2022a, p. 1).
- Ley: “Decreto Ley N° 21907 y Decreto Ley N° 22014 y sus modificatorias” (SMV, 2022a, p. 1).

4 ESTADO DEL ARTE

En este capítulo, se proporcionará una explicación exhaustiva de las fuentes de investigación empleadas para ampliar y enriquecer la información relativa a los temas abordados en el proyecto de tesis.

4.1 Revisión de la literatura

La exploración de la literatura revela la amplia variedad de artículos científicos utilizados como respaldo para la investigación. A través de un exhaustivo análisis sistemático de artículos científicos, se lograron identificar modelos, resultados previos, implementaciones y otras propuestas relevantes relacionadas con la investigación del proyecto.

4.2 Metodología

La estrategia utilizada para la selección de artículos científicos relacionados con el proyecto titulado "Modelo de análisis predictivo para el abandono de clientes en una empresa de fondos colectivos" se divide en tres etapas:

- Planificación: En esta primera etapa, se formulan preguntas de investigación y se utiliza un método de búsqueda para identificar los artículos científicos relevantes.
- Desarrollo: En la segunda etapa, se aplican criterios de selección y exclusión para elegir los estudios primarios pertinentes. Además, se procede a responder las preguntas de investigación con base en las investigaciones encontradas.
- Resultados: La última etapa presenta el análisis de las preguntas de investigación en función de los hallazgos obtenidos de las investigaciones seleccionadas.

Esta metodología permite llevar a cabo un proceso sistemático y estructurado para seleccionar y analizar los artículos científicos que respaldan el proyecto de investigación.

4.2.1 Planificación

4.2.1.1 Palabras Claves

Los términos clave que empleamos para la búsqueda de los artículos científicos son los siguientes:

- “Predictive model”
- “Customer churn”
- “Financial services”
- “Fariness and bias”

- “Collective funds”
- “Crowdfunding”

4.2.1.2 Preguntas de Investigación

- RQ1: ¿Cuáles son los impactos que genera el abandono de cliente en las empresas de servicios financieros?
- RQ2 ¿Qué técnicas y tipos de modelos predictivos se han realizado respecto al abandono de clientes?
- RQ3: ¿Qué metodologías existen para elaborar modelos predictivos en el sector financiero?
- RQ4: ¿Qué tipo de estudios se han realizado sobre modelos predictivos para el abandono de clientes en el sector financiero?

4.2.1.3 Método de búsqueda PICOC

La metodología aplicada para la búsqueda es el PICOC. Es un framework que permite determinar el alcance de la investigación (Petticrew M, Roberts H, 2006). Los criterios de búsqueda se basaron en la importancia de implementar de un modelo de análisis predictivo de abandono de clientes en una empresa administradora de fondos colectivos para tomar decisiones operativas y comerciales basadas en datos. Se encontró diversos artículos científicos que muestran modelos y metodologías, probadas y validadas en otras empresas de similares características.

Figura 13

Componentes del framework de búsqueda PICOC

	Research Question	PICOC		Artículos encontrados SCOPUS
RANSOMWARE + HEALTH INFORMATION	RQ1: ¿CUALES SON LOS RIESGOS Y AMENAZAS QUE EL RANSOMWARE GENERA EN LA INFORMACIÓN DE SALUD MÉDICA?	P (POPULATION)	PATIENTS	49
		I (INTERVENTION)	ATTACKS, VULNERABILITY, RISK, THREATS, COST	
		C (COMPARISON)	-	
		O (OUTCOME)	PRIVACY, PROTECTION, SECURITY HEALTH INFORMATION	
		C (CONTEXT)	HOSPITAL, CLINIC	

QUERY EN SCOPUS:
TITLE-ABS-KEY ("RANSOMWARE" AND ("HEALTH INFORMATION" OR "HOSPITAL" OR "CLINIC") AND ("ATTACKS" OR "VULNERABILITY" OR "RISK" OR "THREATS" OR "COST"))

De “Componentes del framework de búsqueda PICOC”, por Roberts, 2006.

En el presente estudio se utilizó este método de búsqueda para identificar los artículos más relevantes que ayuden a responder las preguntas de investigación planteadas.

Figura 14

Desarrollo de los componentes del framework de búsqueda PICOC para la RQ1.

	Research Question	PICOC		Artículos encontrados Scopus
Customer churn + Financial services	¿Cuáles son los impactos que genera el abandono de cliente en las empresas de servicios financieros?	P (Population)	customer	10
		I (Intervention)	**risk, costs, economic losses, issue**	
		C (Comparison)	-	
		O (Outcome)	Customer retention	
		C (Context)	bank, insurance, finance	

Figura 15

Desarrollo de los componentes del framework de búsqueda PICOC para la RQ2.

	Research Question	PICOC		Artículos encontrados Scopus
Customer churn + Predictive model	¿Qué técnicas y tipos de modelos predictivos se han realizado respecto al abandono de clientes?	P (Population)	Customer	99
		I (Intervention)	-	
		C (Comparison)	**Techniques, models and frameworks**	
		O (Outcome)	Predictive and client retention	
		C (Context)	Financial companies	

Figura 16

Desarrollo de los componentes del framework de búsqueda PICOC para la RQ3.

	Research Question	PICOC		Artículos encontrados Scopus
Predictive model + Financial services	RQ3: ¿Qué metodologías existen para elaborar modelos predictivos en el sector financiero?	P (Population)	Customer	286
		I (Intervention)	-	
		C (Comparison)	**Methodology, standard, best practice**	
		O (Outcome)	Studies	
		C (Context)	**Financial services, finance, bank**	

Figura 17

Desarrollo de los componentes del framework de búsqueda PICOC para la RQ4.

	Research Question	PICOC		Artículos encontrados Scopus
Predictive model + Customer churn + Financial services	¿Qué tipo de estudios se han realizado sobre modelos predictivos para el abandono de clientes en el sector financiero?	P (Population)	Customer	13
		I (Intervention)	Quantitative studies	
		C (Comparison)	-	
		O (Outcome)	Customer retention	
		C (Context)	**bank, finance**	

4.2.2 Desarrollo

Los papers han sido agrupados en tres tipologías. La primera está catalogada como "Problema" y está comprendida por 3 papers. En este subgrupo indagamos cuál es el impacto del problema escogido en otros contextos y cómo se intentan solucionar. La segunda está catalogada como "Técnica" y está comprendida por 8 papers. En este subgrupo ahondamos en los aspectos técnicos de machine learning, sus beneficios, limitaciones y sus potenciales aplicaciones. La tercera está catalogada como "Técnica aplicada al problema" y está comprendida por 9 papers, donde se discuten distintas propuestas técnicas que pueden dar solución al problema. A continuación, se muestra una tabla con todos los artículos científicos revisados, indicando a qué pregunta responde cada uno.

Tabla 4

Tabla de artículos científicos revisados

TIPOLOGÍA	ID	TÍTULO	AUTORES	AÑO	FUENTE	RANKING	PREGUNTA
Problema (Abandono de Clientes)	1	Behavioral attributes and financial churn prediction	Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., & Pentland, A. "Sandy."	2018	EPJ Data Science	Q1	RQ4
	2	Behavior analysis of customer churn for a customer relationship system: An empirical case study	Cheng, L. C., Wu, C., & Chen, C.	2019	Journal of Global Information Management	Q2	RQ1
	3	Leveraging unstructured call log data for customer churn prediction	Vo, n., Liu, S., Li, X. & Xu. G	2021	Knowledge-Based Systems	Q1	RQ1
Técnica (Modelo predictivo)	4	Data mining for modeling students performance: A tutoring action plan to prevent academic drop out	Burgos, C., Campanario, M., de la Peña, D., Lara, J. A., Lizcano, D. & Martínez, M.	2019	Computers & Electrical Engineering	Q1	RQ2
	5	A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM	Wang, C., Han, D., Liu, Q. & Luo, S.	2019	IEEE Access	Q1	RQ3
	6	Dropout early warning systems for high school s tudents using machine learning	Chung, J. & Lee, S.	2019	Children and Youth Services Review	Q1	RQ3
	7	Predicting University Students' Academic Succe ss and Major Using Random Forests	Beaulac, C. & Rosenthal, J. S.	2019	Research in Higher Education	Q1	RQ2
	8	Profit driven decision trees for churn prediction	Höppner, S., Stripling, E., Baesens, B., Broucke, S. & Verdonck, T.	2020	European Journal of Operational Research	Q1	RQ2
	9	A machine learning approach to predict the success of crowdfunding fintech project	Yeh, J. & Chen, C.	2020	Journal of Enterprise Information Management	Q1	RQ3
	10	The use of knowledge extraction in predicting customer churn in B2B	Jamjoom, A. A.	2021	Journal of Big Data	Q1	RQ2
	11	Intelligent Decision Forest Models for Customer Churn Prediction	Usman-Hamza, F.E., Balogun, A.O.; Capretz, L.F., Mojeed, H.A., Mahamad, S., Salihu, S.A., Akintola, A.G., Basri, S., Amosa, R.T. & Salahdeen, N.K.	2022	Applied Sciences (Switzerland)	Q2	RQ2
Técnica aplicada al problema (Modelo predictivo para	12	A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees	De Caigny, A., Coussement, K., & De Bock, K. W.	2018	European Journal of Operational Research	Q1	RQ2
	13	Customer churn prediction in telecom using machine learning in big data platform	Ahmad, A.K., Jafar, A. & Aljoumaa, K.	2019	Journal of Big Data	Q1	RQ2

abandono de clientes)	14	A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector	Ullah, I., Raza, B., Malik, A. K., Imran M., Islam, S. U. & Kim, S. W.	2019	IEEE Access	Q1	RQ2
	15	An empirical comparison of machine-learning methods on bank client credit assessments	Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J.Y., Ryu, K.H.	2019	Sustainability (Switzerland)	Q1	RQ3
	16	Incorporating textual information in customer churn prediction models based on a convolutional neural network	De Caigny, A., Coussement, K., De Bock, K. W. & Lessmannd, S.	2020	International Journal of Forecasting	Q1	RQ4
	17	Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms	Pustokhina, I., Pustokhin, D., RH, A., Jayasankar, T., Jeyalakshmi, C., García, V. & Shankar, K.	2021	Information Processing and Management	Q1	RQ2
	18	Propension to customer churn in a financial institution: a machine learning approach	Lima Lemos, R.A., Silva, T.C. & Tabak, B.M.	2022	Neural Computing and Applications	Q2	RQ4
	19	An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms	Liu, R., Ali, S., Bilal, S.F., Sakhawat, Z. Imran, A., Almuhaimeed, A., Alzahrani, A. & Sun, G.	2022	Applied Sciences (Switzerland)	Q2	RQ2
	20	Customer churn prediction system: a machine learning approach	Lalwani, Praveen; Mishra, Manas Kumar; Chadha, Jasroop Singh; Sethi, Pratyush Lalwani, P., Mishra, M. K., Chadha, J. S. & Sethi, P.	2022	Computing (Vienna/New York)	Q2	RQ2

4.2.3 Resultados

A continuación, se presentará el resumen por cada artículo respondiendo las 4 preguntas de investigación alineados al aporte del autor que se describe en los artículos, el proceso que se empleó para la solución de los artículos, así como su principal resultado.

Pregunta 1: ¿Cuáles son los impactos que genera el abandono de cliente en las empresas de servicios financieros?

Artículo N° 02

Título: Behavior analysis of customer churn for a customer relationship system: An empirical case study (Análisis del comportamiento del abandono de clientes para un sistema de relación con el cliente: un caso de estudio empírico)

Aporte: Por un lado, los autores nos aportan conocimientos que nos permiten entender el impacto que tiene el abandono de clientes en la industria bancaria. A las empresas les cuesta de cinco a seis veces más adquirir un nuevo cliente que retener a un cliente existente (Athanassopoulos 2000; Slater & Narver 2000). La gestión del abandono de clientes tiene como objetivo minimizar las pérdidas causadas por el desgaste de los clientes, en la actualidad las empresas vienen muy interesados en identificar los posibles abandonos (Keramati et al., 2016).

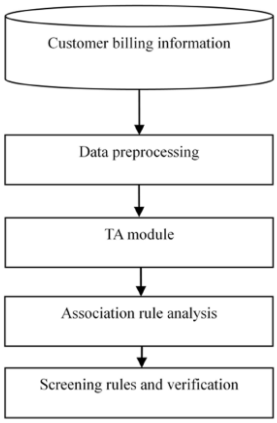
Por otro lado, los autores nos aportan la propuesta de un nuevo modelo que los bancos locales de Taiwán pueden utilizar para detectar potenciales clientes que abandonaran y reducir los impactos de este problema en las empresas. El modelo incorpora una base de datos de gestión de relaciones con el cliente con un factor de tiempo incorporado y aplica abstracción temporal para representar datos durante un período de tiempo específico según lo definido por expertos.

Proceso: Este estudio utiliza la base de datos de clientes de un banco específico para explorar las características de comportamiento de los clientes que abandonan la empresa. La base de datos de gestión de relaciones (CRM) incluye datos demográficos, relaciones con los clientes y los datos de la cuenta de facturación mensual de la tarjeta de crédito del banco.

El marco que proponen los autores consiste en los siguientes subprocesos: Captura de los atributos, preprocesamiento de los datos, modulo TA y captura de reglas y validación como se ilustra en la Figura 1.

Figura 18

Estructura del framework propuesto.



De “Behavior analysis of customer churn for a customer relationship system: An empirical case study”, por Cheng et al., 2019.

Según datos publicados por la Dirección Bancaria de la Comisión Supervisora Financiera en agosto de 2011, aproximadamente el 36,7% de los clientes de tarjetas de crédito no han utilizado su tarjeta durante más de 6 meses. En este estudio los autores consideran que un cliente ha abandonado cuando llevan 6 meses sin pagar con tarjeta de crédito. Los atributos del cliente se definen en las siguientes 3 categorías:

1. Clientes Sobrevivientes (Transacciones Normales): Clientes que tienen una tarjeta de crédito y lo han utilizado por lo menos 1 vez en los últimos 6 meses (según lo definido por el Negociado Bancario de la Comisión de Supervisión Financiera).
2. Clientes potenciales de abandono: dichos clientes pueden tener una tarjeta de crédito válida para el consumo normal pero no lo han usado por más de 6 meses.
3. Clientes de abandono voluntario: estos clientes han cancelado todas sus tarjetas de crédito para el banco especificado por voluntad propia

Preprocesamiento de datos: para evitar la desviación de los datos, los autores excluyen cierto tipo de datos, como tarjetas de particulares. Asimismo, establecen los filtros de la data que, si será parte del conjunto de datos final, como los clientes con patrones de consumo normales. Luego de discutir con los expertos, se eligieron 15 variables del cliente relacionadas a su información demográfica y el estado de sus consumos y pagos.

Figura 19

Listado de variables seleccionadas.

Data category	Item	Attributes	Remarks
	1	Customer identification no.	
Demographic data	2	Age	
	3	Gender	1 (male) 2 (female)
	4	Place of residence	Indicated by the area code of the residential phone
	5	Customer affiliation	1 financial customer groups 2 customer groups with development value 3 general customer groups
Customer relationship data	6	Frequency of customer calls to the service center	unit: number of times
	7	Customer satisfaction with their service center calls	H: satisfied N: no interview L: dissatisfied
	8	Frequency of return calls from the customer service center for care (or marketing)	unit: number of times
Spending account data	9	Closing day	date bills are generated; YYYYMMDD (Republic of China date), for the use of time interval
	10	Perpetual credit line	unit: NTD
	11	Credit interest rate per cycle	2 integers and 3 digits after decimal point
	12	Consumption amount in the current period	unit: NTD
	13	Balance of installment	unit: NTD
	14	Balance of cycle credit	unit: NTD
	15	Code showing payment status in the prior period (amount)	1 entirely paid the unpaid balance 2 paid the minimum payment 3 (did not pay the minimum payment 4 X (no unpaid balance)

De “Behavior analysis of customer churn for a customer relationship system: An empirical case study”, por Cheng et al., 2019.

TA Module: El modelo TA se aplica primero utilizando la base de datos de gestión de relaciones con el cliente que contiene la serie temporal de características. Los campos de atributos de datos se dividen en atributos de abstracción temporales y atributos de abstracción no temporales, las tendencias de los cambios en los campos a lo largo del tiempo y los valores de referencia utilizados en la práctica se seleccionan en base a las sugerencias dadas por el supervisor de servicio al cliente. El proceso básico de TA se lleva a cabo para los elementos de atributo que tienen la abstracción temporal.

Análisis de reglas de asociación: La técnica de minería de datos relacional se utiliza para comprender la asociación entre los datos demográficos del cliente, sus datos de facturación y el abandono de clientes. La minería de reglas de asociación se aplica a los datos transaccionales de los clientes.

Data experimental y reglas: Como etapa final del proceso, el conjunto de datos experimental quedara proporcionado de la siguiente manera:

Figura 20

Distribución de los resultados con el conjunto de datos experimental.

Category code	Customer category	Number of transactions	Proportion
1	Surviving customers	95,323	84.82%
2	Potential churn customers	15,179	13.51%
3	Volunteer churn customers	1,875	1.67%

De "Behavior analysis of customer churn for a customer relationship system: An empirical case study", por Cheng et al., 2019.

Las reglas de asociación se utilizarán para categorizar los patrones de comportamiento del cliente en 3 categorías: cliente sobreviviente, cliente potencial de abandono y cliente voluntario de abandono. Data la distribución desequilibrada en la cantidad de los distintos tipos de cliente, se establecen diferentes valores de apoyo, confianza e importancia para ayudar a detectar reglas útiles.

Principal Resultado: Los resultados de este artículo indican que el sistema es relativamente eficaz en la detección temprana de abandono de clientes y, por lo tanto, útil para ayudar a los bancos a abordar los problemas antes de que se agraven. Este estudio proporciona un sistema experto para que los bancos evalúen la calidad de su estrategia de marketing y sus campañas para restablecer las relaciones con los clientes. Además, el estudio brinda a los gerentes una guía para ayudarlos a seleccionar el método apropiado para identificar a los potenciales clientes que abandonaran. El centro de llamadas podría llamarlos para informales de algún nuevo servicio o información.

Artículo N° 03

Título: Leveraging unstructured call log data for customer churn prediction (Aprovechamiento de los datos de registro de llamadas no estructurados para la predicción de abandono de clientes)

Aporte: Los autores proponen un modelo de predicción de abandono de clientes para el sector de servicios financieros utilizando datos no estructurados, que son las comunicaciones con el cliente mediante llamadas telefónicas grabadas, es decir, adoptaron una estrategia como convencional al utilizar este tipo de información no estructurada. Los autores

recolectaron una gran cantidad de datos del centro de llamadas, con más de dos millones de llamadas pertenecientes a más de doscientos mil clientes y llevaron a cabo varios experimentos.

La retención de clientes es importante en la industria de servicios financieros. Estudios han demostrado que adquirir un nuevo cliente suele ser cinco veces más caro que retener a un cliente existente. Por el alto costo de adquisición de clientes, las empresas establecidas se centran más en Retención de clientes en lugar de adquisición (Rudin, 2019). Por lo tanto, predecir el abandono de clientes puede conducir a un ahorro sustancial de ingresos, especialmente para el sector de servicios financieros

Proceso: En primer lugar, los autores se enfocan en analizar una base de datos integrada de registro de llamadas de clientes, ya que la hipótesis es que los modelos de predicción que utilizan características combinadas del cliente pueden mejorar la predicción de abandono a comparación de los modelos que solo utilizan datos estructurados básicos del perfil del cliente. Los autores incorporan técnicas de minería de texto múltiple en los conjuntos de datos para extraer cuatro diferentes características: Importancia del término, incrustación de frases, información léxica y rasgos de personalidad:

Importancia del término: La técnica más común para derivar características de importancia de términos es la bolsa de palabras (Bag of words). Sin embargo, en nuestro contexto la técnica más adecuada seria la "frecuencia de termino – frecuencia de documento inversa", ya que hay muchas palabras con un significado menos perspicaz como "Hola" o "los" (Scherer et.al., 2015). La metodología captura el significado textual principal mediante el uso de la expresión TFIDF. La tabla nos muestra algunos ejemplos de registros de llamadas con el valor que tienen para cada termino identificado, así como también indicarnos si esa llamada significo un abandono o no.

Figura 21

Ejemplo de variables TF-IDF.

Sample TF-IDF features.

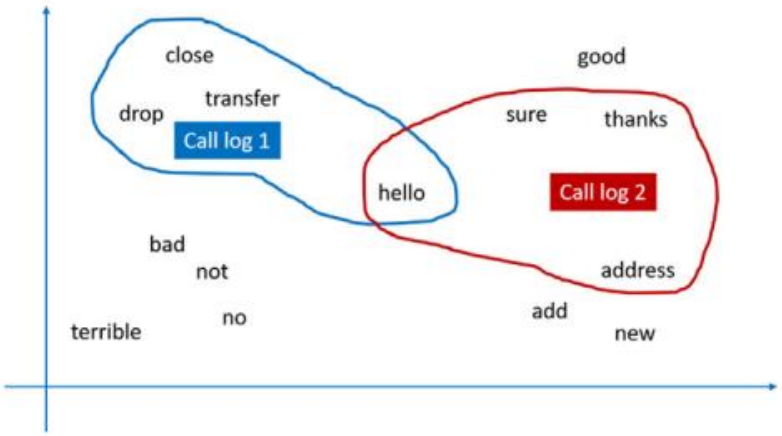
Call ID	Term					
	"close"	"transfer"	"hello"	"yes"	"no"	Churn
Call 1	1.63	0.24	0.07	0.20	0.73	1
Call 2	0	2.19	0.07	0.27	0.15	1
Call 3	1.63	0.97	0.07	0.40	0	0
Call 4	0	0	0.07	0.27	0	0
Call 5	0	0.49	0.07	0.20	0.15	0
Call 6	0	0.24	0.07	0	0.15	0
Call 7	0	0	0	0	0.15	0

De "Leveraging unstructured call log data for customer churn prediction", por Vo et al., 2021.

Incrustación de frases: Se aprovecha un algoritmo bastante utilizado, modelo Word2Vec, para extraer un total de 50 funciones de incrustación de palabras. En la siguiente imagen el modelo Word Embedding captura las relaciones entre los términos. El registro de llamadas que se muestra en azul es un cliente que abandonó y en rojo es un cliente normal.

Figura 22

Ejemplificación de la relación existente entre los términos.



De "Leveraging unstructured call log data for customer churn prediction", por Vo et al., 2021.

Información léxica: Si bien la técnica más utilizada es el análisis de sentimientos, los autores creen que hay ideas más significativas en otras dimensiones y temas de la lengua inglesa, con el fin de poder explicar suficientemente el comportamiento del cliente. Los autores aprovechan la indagación lingüística y Word Count 2015 (LIWC) para extraer conceptos latentes y características de texto relacionadas con el tema.

El diccionario LIWC 2015 contiene alrededor de 6.400 términos y emociones. Cada término tiene una entrada de diccionario correspondiente separada que identifica una o más categorías. Por ejemplo, el término "decepcionado" pertenece a cinco categorías diferentes: "Verbo", "Afecto general", "Enfoque pasado", "Emoción negativa" y “triste". Si el cliente dijo el término "decepcionado", todos estos las puntuaciones de cinco categorías léxicas aumentarían respectivamente.

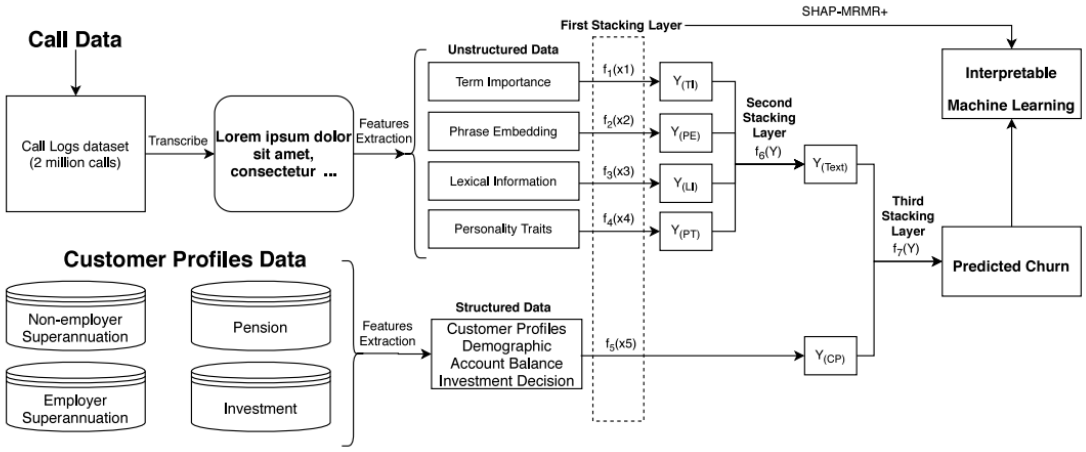
Rasgos de personalidad: Personality Mining es una técnica avanzada de minería de datos para encontrar los rasgos de una persona por la forma en que habla y actúa. El modelo que proponen los autores incluye minería de personalidad de cinco factores. El modelo Big Five contiene cinco rasgos humanos fundamentales: apertura a la experiencia, escrupulosidad, extroversión, amabilidad y neuroticismo (Goldberg, 1990). En este caso, los autores utilizan un algoritmo de Random Forest para entrenar el modelo de predicción en cuanto a personalidad.

La justificación del uso de rasgos de personalidad en la predicción de abandono es que los clientes con personalidades variadas pueden hacer diferentes decisiones financieras Desde la perspectiva del modelo, la minería de personalidad podría tratarse como una tarea de aprendizaje supervisado, y los resultados puede explicar suficientemente el comportamiento del cliente.

En segundo lugar, los autores proponen un método multi-stacking para construir el modelo predictivo final ya que este enfoque reduce el tiempo de entrenamiento del modelo significativamente a comparación de los modelos que combinan todas las características linealmente.

Figura 23

Modelo esemble multi-stacking basado en machine learning.



De “Leveraging unstructured call log data for customer churn prediction”, por Vo et al., 2021.

Finalmente, para evaluar el desempeño de los modelos, los autores construyeron cuatro líneas base que son los modelos de predicción usando solo las características básicas de los cuatro diferentes datasets de clientes. Los experimentos se ejecutan con 10 veces la validación cruzada y los rendimientos se promedian para todos los pliegues. Se usarán puntajes para el área bajo la curva (AUC) como métrica para la evaluación.

Principal Resultado: Los resultados de la investigación realizada en la industria de servicios financieros de Australia, muestran que el modelo propuesto por los autores puede predecir con precisión el abandono de clientes y sus riesgos. Además, los resultados de los experimentos muestran que los datos no estructurados de los registros de llamadas de los clientes puede mejorar la precisión de los modelos de pronóstico de abandono de clientes, asimismo se pueden utilizar para generar conocimientos significativos utilizando el aprendizaje automático interpretable con características de personalidad y segmento de clientes, estos conocimientos pueden ayudar a los gerentes a desarrollar estrategias de retención personalizadas para diferentes segmentos de clientes y ahorrar millones de dólares beneficiándose de la temprana identificación de clientes con alto riesgo de abandono.

Pregunta 2: ¿Qué técnicas y tipos de modelos predictivos se han realizado respecto al abandono de clientes?

Artículo N° 04

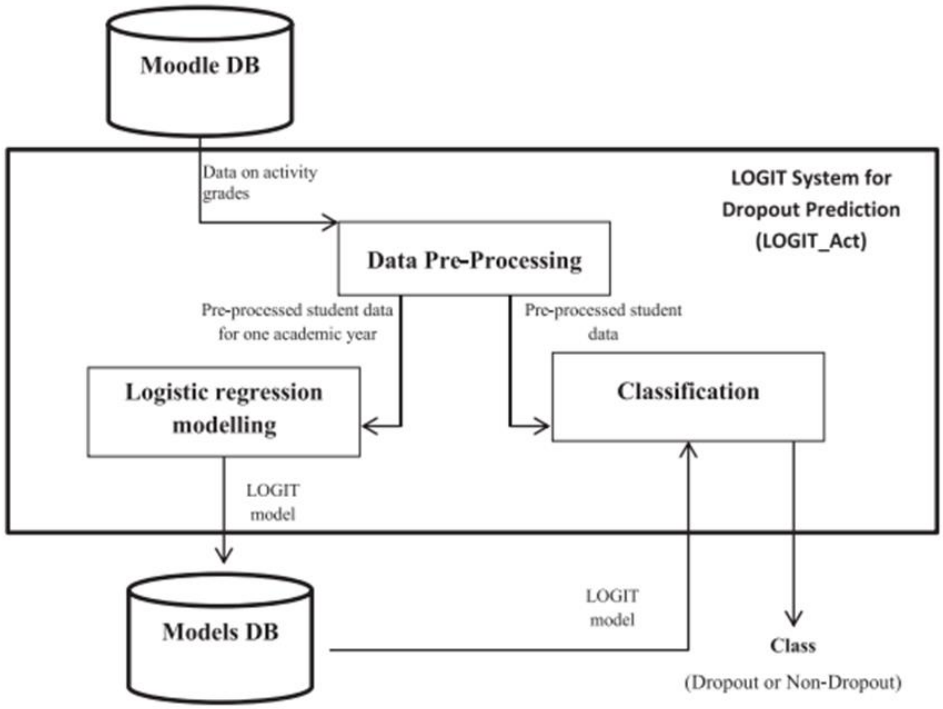
Título: Data mining for modeling students' performance: A tutoring action plan to prevent academic droput (Minería de datos para modelar el desempeño de los estudiantes: Un plan de acción de tutoría para prevenir la deserción académica)

Aporte: El principal aporte de los autores es una herramienta y un plan de tutoría que pueda ser usado por la institución educativa en estudio, y en otras, a fin de reducir el índice de deserción en cursos e-learning. Esto lo logran gracias al uso de técnicas predictivas de minería de datos, específicamente regresión logística, basado en las notas obtenidas por los alumnos en las actividades de los cursos.

Proceso: El primer paso de los autores es realizar el descubrimiento de conocimiento en base a las notas de las actividades desarrolladas por los estudiantes. Para ello, los autores realizan minería de datos en las notas de las actividades. Esta minería de datos está compuesta por los siguientes pasos: preprocesamiento, modelar la regresión lineal y clasificar los estudiantes a través del sistema LOGIT_Act.

Figura 24

Modelo propuesto por los autores.



De "Data mining for modeling students performance: A tutoring action plan to prevent academic dropout", por Burgos et al., 2019.

Luego comparan métricas de precisión, sensibilidad, especificidad e índice de éxito de los modelos obtenidos versus los principales sistemas existentes en este campo.

Resultado: Las distintas métricas evaluadas en el sistema propuesto (LOGIT_Act) muestran un desempeño superior a las de las otras soluciones propuestas en la semana 10 y en general, como se puede ver en la siguiente tabla y figura, respectivamente.

Figura 25

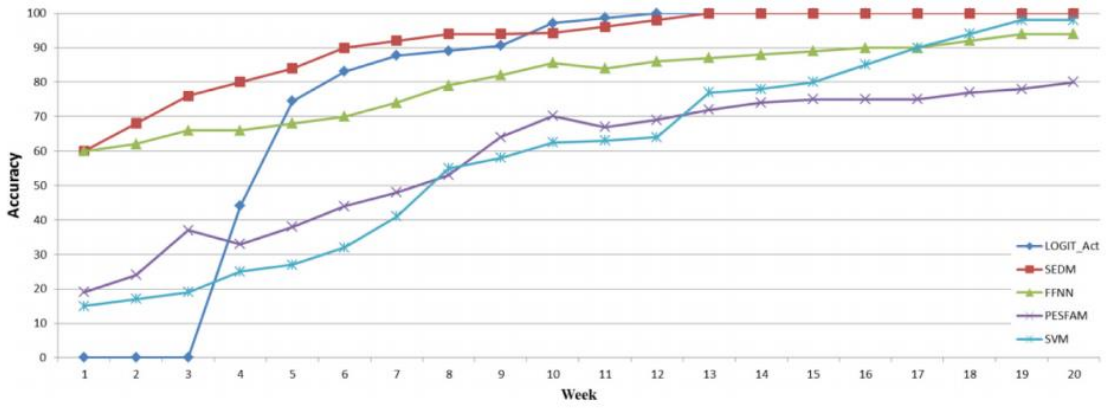
Tabla de resultados obtenidos en el sistema propuesto LOGIT_ACT

Proposal	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)
LOGIT_Act	98.95	96.73	97.14	97.13
SEDM	85.71	92.31	94.87	94.23
FFNN	68.97	76.92	88.46	85.58
PESFAM	43.24	61.54	73.08	70.19
SVM	36.73	69.23	60.26	62.50

De “Data mining for modeling students performance: A tutoring action plan to prevent academic dropout”, por Burgos et al., 2019.

Figura 26

Gráfica de resultados obtenidos en el sistema propuesto LOGIT_ACT.



De “Data mining for modeling students performance: A tutoring action plan to prevent academic dropout”, por Burgos et al., 2019.

En ningún caso, alguno de los parámetros para la propuesta es inferior al 96% en la semana 10 y, en algunos casos, supera sustancialmente a las otras propuestas.

Los valores de los indicadores de rendimiento del método de clasificación sugieren que el método es capaz de predecir correctamente qué estudiantes son posibles desertores.

Teniendo en cuenta que el modelo predictivo cumple con altos estándares de precisión, este no es suficiente para reducir el índice de deserción. Para ello, es necesario llevar al mismo tiempo un programa de tutoría.

Los autores proponen acciones preventivas a realizar por los roles tutor e instructor en semanas específicas del curso.

Figura 27

Listado de acciones preventivas propuestas por los autores.

Week	Facilitator	Action
1	Tutor	Courtesy call at the start of the academic year
1	Instructor	Public message of welcome to the course via virtual classroom
2	Instructor	Videoconferenced welcome session
4	Instructor	Email to potential dropouts
7	Instructor	Telephone call to potential dropouts
10	Tutor	Telephone call to potential dropouts (from one or more courses)

De “Data mining for modeling students performance: A tutoring action plan to prevent academic dropout”, por Burgos et al., 2019.

Luego de implementar el sistema de predicción de deserción y realizar las actividades del plan de acción de tutoría, se tuvo el resultado mostrado en la siguiente tabla.

Figura 28

Listado de resultados post implementación del sistema preventivo y de realización de acciones preventivas.

	Academic year							
	2013–2014				2014–2015			
Course	Students	Dropout	Non-dropout	% Dropout	Students	Dropout	Non-dropout	% Dropout
COMP	26	7	19	26.92	25	3	22	12
INF-SYST	17	3	14	17.65	16	1	15	6.25
NETS	15	6	9	40	13	2	11	15.38
DM	7	2	5	28.57	8	1	7	12.5
DB	39	8	31	20.51	38	4	34	10.53
Total	**104**	**26**	**78**	**25**	**100**	**11**	**89**	**11**

De "Data mining for modeling students performance: A tutoring action plan to prevent academic dropout", por Burgos et al., 2019.

Como se puede apreciar, hubo una reducción sustancial en la deserción académica en los 5 cursos en estudio, logrando una reducción total promedio de 25% a 11% en 1 año.

Artículo N° 07

Título:

Predicting University Students' Academic Success and Major Using Random Forests (Pred ecir el éxito académico y la especialización de los estudiantes universitarios utilizando bosques aleatorios)

Aporte: El autor menciona que hay pocos estudios en el paradigma de Deep Learning, ya que aún se encuentra en sus etapas iniciales. Su estudio se encuentra dentro de los primeros que buscan presentar la importancia de Deep Learning para mejorar la predicción de los estudiantes que podrían abandonar cursos. Así mismo, este estudio tiene la intención de analizar el impacto de la participación en línea de los estudiantes en su desempeño y la implementación de técnicas para la predicción temprana del retiro de los estudiantes.

Proceso: Analizando la cantidad de datos recopilados, para ello primero tiene que pasar por un procesamiento inteligente y pueda ir agrupando la información por semana, la primera de ellas tiene relación con la cantidad de clic por parte de los estudiantes, se puede verificar que a partir de la semana 10 esta cantidad de clic disminuye, esto se representa como la interacción entre la plataforma con el estudiante va descendiendo y a su vez, el compromiso también.

La información base para obtener la tasa de deserción debe ser procesada y transformada. Siguiendo esta secuencia:

Convertir los datos en una secuencia semanal, se organiza de esta forma para tener una visión de las actividades realizadas por el alumno en esa semana, luego pasa a reordenarse según estas acciones.

Se aplica el modelo LSTM, al tener toda la información organizada esta puede ser interpretada por el sistema para obtener el rendimiento de los estudiantes.

Todos los datos obtenidos son importantes, se trata de recopilar la mayor cantidad de ellos como: números de clics, evaluaciones y el comportamiento que han tenido los estudiantes a lo largo de los módulos de los cursos tomados en línea.

Luego de realizar las pruebas correspondientes se comprueba que los resultados son mejores con el paso de las semanas, esto se debe a que con el incremento de los días se obtiene una mayor cantidad de datos, es decir, los datos en la semana 8 son mayores a la de la semana 1, logrando obtener mejores resultados de precisión.

Se realizó la comparación con otras técnicas como: ANN y LR, en comparación con ellas, LSTM no realiza sus resultados en función al tiempo, busca recopilar información por semana de cada uno de los estudiantes, obteniendo como resultado una mayor precisión en los porcentajes arrojados por la deserción de estudiantes, esto se debe a que las técnicas convencionales predicen erróneamente si un alumno no está activo en las semanas iniciales.

En el punto experimental en LSTM con 3 capas, donde cada una de ellas recibe como información de entrada, la salida de la anterior, esto logra tener un mejor resultado, los datos han sido calculados para un tiempo aleatorio y todo es colocado a la red neural, tomando exclusivamente solo la información correspondiente a la cantidad de clics.

En resumen, el autor presenta las contribuciones obtenidas para su estudio. En primer lugar, se realiza el procesamiento de los datos por un determinado tiempo, ayudando a recopilar todas las actividades realizadas por el estudiante en una semana. Luego, se realiza las comparaciones de RNAs y la regresión logística. Aproximadamente en la semana 10 se podrá tener un 80% de precisión. Por último, se analiza los resultados obtenidos en el punto anterior para poder realizar la toma de decisiones y generar estrategias para aumentar el porcentaje de retención en los estudiantes.

Resultado: El principal resultado encontrado de los autores comprueba que la técnica LSTM brinda los resultados más claros y precisos sobre el tema, obteniendo una exactitud del 97%. Además, sugieren tomar otros factores distintos a los clics para obtener patrones que ayuden

en el estudio de los alumnos en riesgo de abandonar sus estudios. Por otro lado, se debe considerar que aún falta mucho por investigar y este campo es muy amplio, sin embargo, con el análisis realizado por el autor en este artículo se puede decir que si se puede predecir la deserción estudiantil en cursos en línea.

Artículo N° 08

Título: Profit driven decision trees for churn prediction (Árboles de decisión basados en beneficios para la predicción de abandono)

Aporte: Los modelos de abandono se seleccionan comúnmente en función de las medidas de rendimiento relacionadas con la precisión, como por ejemplo el área bajo la curva ROC (AUC). Sin embargo, estos modelos a menudo no están bien alineados con el requisito comercial central de maximizar las ganancias, en el sentido de que los modelos no tienen en cuenta los costos de clasificación errónea. Por lo tanto, el aporte de los autores es construir un modelo de predicción de abandono que sea rentable e interpretable.

La técnica que proponen los autores, llamada ProfTree, utiliza un algoritmo evolutivo para aprendizaje de árboles de decision impulsados por las ganancias. Los autores realizan un estudio con un conjunto de datos de la vida real de una empresa de servicio de telecomunicaciones para evaluar ProfTree comparado con los métodos clásicos de árboles de decision impulsados por la precisión. La medida de beneficio máximo esperada (EMPC) ha sido propuesta con el fin de seleccionar el modelo de predicción de abandono más rentable.

Proceso: El proceso para el caso de estudio incluye la recopilación de un conjunto de datos de abandono de clientes de un operador de telecomunicaciones de Corea del Sur. El conjunto de datos contiene una muestra de 889 clientes y 10 variables explicativas como se muestra en la tabla. Los clientes etiquetados como abandono conforman 277 de las observaciones, es decir, el 31,16%. La meta es que se construya un modelo que prediga posibles abandonos teniendo en cuenta la rentabilidad. Además, que el modelo sea fácilmente interpretable de modo que los resultados puedan ser comunicados al departamento de marketing.

Figura 29

Variables de un conjunto de datos de una empresa Sur Coreana de telecomunicaciones

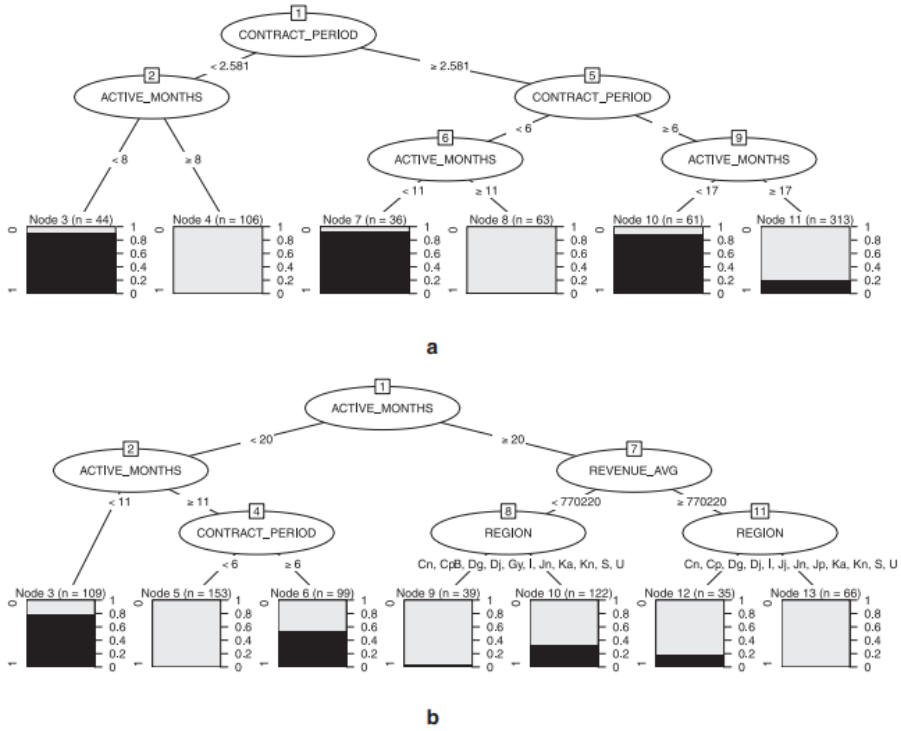
Variable	Description
churn	Did the customer leave the telecom operator?
region	Region where the customer lives.
prod_num	Number that identifies the customer's product.
active_months	Time since the customer joined the operator (in months).
contract_period	Length of the contract period.
revenue_avg	Average revenue.
nonpay_period	How long did the customer not pay the bills?
overdue_amt	Amount that the customer is overdue.
count_disconnect	Number of times the service was disconnected.
count_complaint	Number of filed complaints.
autopay	Did the customer use the automatic payment option?

De "Profit driven decision trees for churn prediction", por Hoppner et al., 2020.

Se estableció construir un modelo de árbol con una profundidad máxima de tres niveles. Adicional a ProfTree, se utilizaron otros métodos de árbol de clasificación como EvTree, CART y CTree. En segunda instancia, se dividió aleatoriamente el conjunto de datos asignando un 70% para el entrenamiento y un 30% para las pruebas, estratificados según el indicador de abandono para obtener distribuciones de abandono similares al conjunto de datos original. Todos los árboles están obligados a tener un mínimo de 10 observaciones por nodo terminal, 20 observaciones por nodo interno y una profundidad máxima de 3. Adicionalmente, el árbol de inferencia condicional se construyó con un nivel de significación de 1% en lugar del predeterminado de 5% ya que parece ser más apropiado para un total de 623 observaciones.

Figura 30

Gráfico de árboles para la predicción de abandono construidos con EvTree and ProfTree.



De "Profit driven decision trees for churn prediction", por Hoppner et al., 2020.

Podemos apreciar que Evtree, usa las mismas dos variables (active_months y contract_period) que ProfTree para predecir abandonos, pero cada método lo utiliza con diferentes valores de corte. Por otro lado, ProfTree incorpora dos variables adicionales (revenue_avg y región) en su árbol de clasificación. ProfTree emplea variables como el ingreso promedio y la región para determinar la probabilidad de abandono del cliente.

Principal Resultado: Se comparó el EMPC (métrica para evaluar el modelo más rentable) promedio de ProfTree para cada conjunto de datos de abandono. ProfTree tiene el mejor rendimiento general en términos de EMPC y MPC. El modelo propuesto por los autores ocupa el primer lugar en 6 de 9 conjuntos de datos, lo que resultado en un ranking promedio de 1.56 en una escala del 1 al 7. ProfTree ofrece en 6 de 9 conjuntos de datos el modelo de abandono más rentable, y en 4 de 6 el rendimiento es significativamente superior.

Figura 31

Ranking promedio de cada clasificador sobre los diferentes conjuntos de datos para varias métricas de rendimiento.

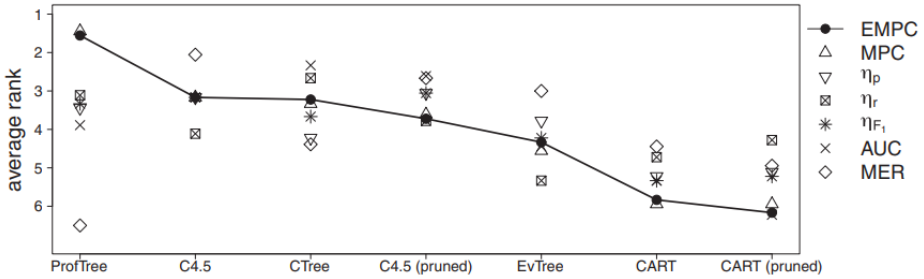


Fig. 7. Average rank of the classifier over the different datasets for various performance metrics.

De "Profit driven decision trees for churn prediction", por Hoppner et al., 2020.

Artículo N° 10

Título: The use of knowledge extraction in predicting customer churn in B2B (El uso de la extracción de conocimiento para predecir el abandono de clientes en B2B)

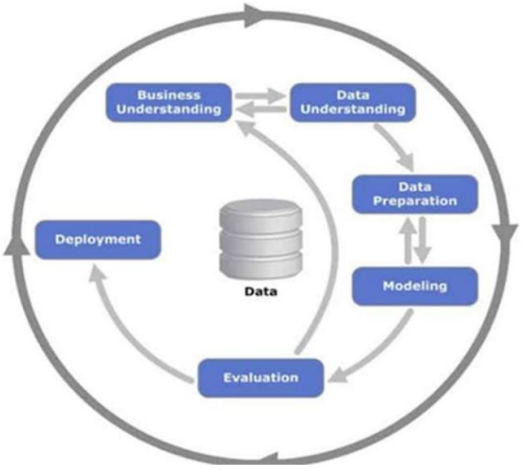
Aporte: El aporte de los autores es investigar el uso de la extracción de conocimiento en predecir el abandono de clientes en compañías de seguro, utilizando técnicas de minería de datos. Además, no se han realizado muchos estudios para averiguar cómo se puede controlar el abandono de clientes en compañías de seguros ya que el campo de la predicción del abandono de clientes recibe mucho menos énfasis en contextos de empresa a empresa, mientras que está bien investigado en el contexto de cliente a empresa (Tamaddoni et.al., 2017). Por lo tanto, los autores buscan con este estudio pronosticar que clientes abandonaran y comprender la razón del abandono. Un cliente ofrece información valiosa sobre posibles decisiones a través de los datos de comportamiento del cliente. Por ejemplo, los modelos de predicción de abandono determinan los clientes que dejan de utilizar un servicio o producto **(**Keramati et.al., 2016**)**

Proceso:

El estudio adopto K-Means y Arboles de decisión para construir un modelo de predicción de abandono de clientes utilizando la metodología CRISP-DM. La metodología CRISP-DM se basa en seis fases: 1) comprensión empresarial o de negocio, 2) Comprensión de los datos, 3) Procesamiento de los datos, 4) Modelado, 5) Evaluación y 6) Implementación (Martínez et al., 2019)

Figura 32

Metodología CRISP-DM



De “The use of knowledge extraction in predicting customer churn in B2B”, por Jamjoom, 2021.

Se crearon cinco conjuntos de datos de entrenamiento para generar los modelos predictivos a través del algoritmo árbol de decisión, variando la distribución entre los clientes que abandonaría y los que no abandonaron.

Posteriormente, ya se tiene construido el conjunto de datos de prueba y los conjuntos de datos de entrenamiento previamente creados con árboles de decisión. Los conjuntos de datos de entrenamiento fueron utilizados en cuatro diferentes técnica o algoritmos para predecir que clientes abandonaran. Se observaron diferentes resultados basados en diferentes configuraciones técnicas.

Para la elección del modelo, los autores indican que el algoritmo más usado en la mayoría de los artículos científicos es el árbol de decisión, sin embargo, las redes neuronales se utilizaron mayormente para predecir abandono de clientes. En el estudio, los autores identifican que el peor desempeño lo tiene el algoritmo de árboles de decisión, sin embargo, la regresión logística y las redes neuronales pasan a un siguiente nivel para evaluar su costo beneficio de cada una.

Figura 33

Parámetros de rendimiento de los modelos testeados en la investigación.

Performance parameter	Logistic regression	Decision tree	Neural networks
AUC	0.036	0	0.075
AUK	0.769	0.5	0.629
Kappa	0.291	0	0.181
Threshold	0.8	0.1	0.5
Precision	0.220	–	0.117
Sensitivity	0.103	1	0.591

De "The use of knowledge extraction in predicting customer churn in B2B", por Jamjoom, 2021.

Figura 34

Análisis de costo – beneficio.

Performance parameter	Logistic regression E9	Logistic regression E10	Logistic regression M9	Logistic regression M10
AUC	0.045	0.081	0.066	–
AUK	0.527	0.552	0.816	–
Performance parameter	**Neural Network E9**	**Neural Network E10**	**Neural Network M9**	**Neural Network M10**
AUC	0.061	0.043	0.069	0.071
AUK	0.215	0.531	0.926	0.052

De "The use of knowledge extraction in predicting customer churn in B2B", por Jamjoom, 2021.

Finalmente, después de evaluar el costo beneficio, los autores concluyen que ninguno de estos dos algoritmos parece ser mejor que el otro, por lo tanto, se puede interpretar que los algoritmos de redes neuronales y la regresión logística son las técnicas más importantes para predecir el abandono de clientes.

Principal Resultado:

Los procedimientos de minería de datos pueden ser muy exitosos en la extracción de información oculta que nos permita conocer el comportamiento del cliente. El estudio dio como principal resultado que la regresión logística y la red neuronal son las mejores técnicas para predecir el abandono de clientes. Una distribución 50-50 (50% clientes que no

abandonaron – 50% clientes que abandonaron) del conjunto de datos de entrenamiento resulto efectivo el algoritmo de regresión logística, sin embargo, una distribución 70-30 funciono de manera efectiva para el algoritmo de red neuronal.

En conclusión, cada técnica funciona efectivamente con una distribución distinta del dataset. Este estudio ha demostrado que los modelos de predicción pueden ser utilizados a lo largo de la estrategia de marketing de una compañía de seguros de salud y en cualquier contexto en general con un enfoque basado en la solución de problemas de negocio

Artículo N° 11

Título: Intelligent Decision Forest Models for Customer Churn Prediction (Modelos de bosques de decisiones inteligentes para la predicción de abandono de clientes)

Aporte: En este estudio se desarrolló un modelo de bosque de decision inteligente (DF) para el abandono de clientes en el rubro de las telecomunicaciones. Específicamente se investigaron los rendimientos en la predicción del modelo de árbol logístico (LMT), el bosque aleatorio (RF) y los árboles funcionales (FT). Además, los autores realizaron una amplia experimentación para determinar la eficacia del modelo propuesto, utilizando conjuntos de datos públicos sobre el abandono de clientes en las telecomunicaciones.

Proceso: El procedimiento experimental utilizado en este estudio tiene como objetivo analizar y corroborar empíricamente la eficacia de los modelos propuestos. En concreto se diseñaron e investigaron dos fases de experimentación y los rendimientos de predicción de los modelos resultantes se compararon en un método justo y coherente.

Figura 35

Framework del trabajo experimental de la investigación.

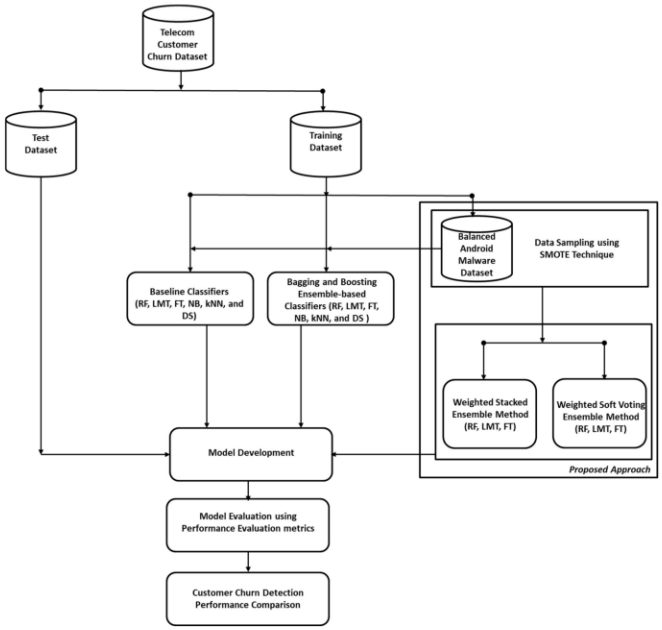


Figure 1. Experimental Framework.

De "Intelligent Decision Forest Models for Customer Churn Prediction", por Usman-Hamza, 2022.

En la primera fase, se investigaron los diferentes modelos de predicción existentes para el abandono de clientes (LMFT, FT y RF), estos modelos fueron evaluados en el conjunto de datos original ya que el propósito de esta primera fase de experimentación era evaluar la predicción y la eficacia de los modelos con conjunto de datos desequilibrados. A partir de entonces, el problema del desequilibrio se resolvió implementando el muestreo de datos con el método SMOTE. Posteriormente se vuelven a evaluar los modelos en el conjunto de datos equilibrado. Los hallazgos de esta primera fase indicaran la eficacia de los modelos investigados en conjuntos de datos equilibrados por el método de muestreo sobre los conjuntos de datos originales.

En la segunda fase, de igual manera se evaluaron ciertos modelos en el conjunto de datos original y el conjunto de datos equilibrado, estos modelos fueron Weighted Soft Voting Ensemble Decision Forest Method (WSVEDFM) y Weighted Stacking Ensemble Decision Forest Method (WSEDFM). Esta fase tiene como objetivo validar la eficacia de estos modelos con o sin el problema de desequilibrio de clases.

Para la fase de experimentación de este trabajo de investigación, los autores recopilaron dos conjuntos de datos sobre el abandono de clientes con diversas variables y estos se utilizaron para entrenar y probar los modelos. El primer conjunto de datos se obtuvo del repositorio de Kaggle ML y el segundo conjunto de datos se descargó del repositorio de UCI ML. Estos conjuntos de datos seleccionados están disponibles públicamente y se utilizan regularmente en estudios existentes de abandono de clientes. El conjunto de datos 1 describe información sobre una empresa de telecomunicaciones que proporciona servicios de voz e internet a los clientes, consta de 3333 observaciones, de las cuales 2850 son no abandonos y 483 son abandonos, con 21 variables. Este conjunto de datos tiene una tasa de abandono del 14,49% y una relación de desequilibrio IR de 5,9. El conjunto de datos 2 tiene 5000 observaciones de las cuales 4493 son de no abandono, mientras que 507 son abandonos, esto significa que este conjunto de datos tiene una tasa de abandono del 10,14% y un IR de 8,86

Figura 36

Descripción de los conjuntos de datos para la predicción de abandono de clientes.

Dataset	Features	Instances	Churners	Non-Churner	Churn Rate	IR
Dataset 1	20	3333	483	2850	14.49%	5.9
Dataset 2	18	5000	507	4493	10.14%	8.86

De “Intelligent Decision Forest Models for Customer Churn Prediction”, por Usman-Hamza, 2022.

Las métricas evaluación del modelo fueron el área bajo la curva (AUC) y el coeficiente de correlación de Mather (MCC), con estas métricas se evaluaron las capacidades de predicción de los diferentes modelos para el abandono de clientes. La selección de estos indicadores de desempeño se basa en el uso generalizado y constante de estas métricas en distintos estudios existentes. Específicamente MCC es una métrica confiable porque considera todos los cuadrantes de la matriz de confusión para cada modelo desarrollado (Jain, H.; Khunteta, A.; Srivastava, S, 2020)

Principal Resultado: Los hallazgos mostraron que los modelos propuestos proporcionaron un rendimiento de predicción superior. Además, se propone soluciones óptimas para el abandono de clientes en la industria de las telecomunicaciones. A continuación, se muestran los resultados obtenidos por los modelos evaluados en los 4 distintos conjuntos de datos: El

dataset 1 original, el dataset 2 original, el dataset 1 balanceado (SMOT) y el dataset 2 balanceado (SMOT).

Figura 37

Rendimiento de los modelos de predicción en el conjunto de datos 1.

	Accuracy (%)	AUC	F-Measure	MCC
NB	88.24	0.834	0.834	0.465
kNN	83.38	0.603	0.821	0.237
DS	86.56	0.603	0.841	0.317
* LMT	**94.75**	**0.905**	**0.945**	**0.777**
* FT	94.42	**0.905**	0.942	0.763
* RF	90.97	0.896	0.895	0.581

* The superior CCP performances are bold and type-faced, and the proposed DF models are designated by an asterisk.

Nota. Información del rendimiento de los modelos. De "Intelligent Decision Forest Models for Customer Churn Prediction", por Usman-Hamza, 2022.

Figura 38

Rendimiento de los modelos de predicción en el conjunto de datos 2.

	Accuracy (%)	AUC	F-Measure	MCC
NB	89.86	0.503	?	?
kNN	81.90	0.510	0.820	**0.020**
DS	89.86	0.496	?	?
* LMT	**89.86**	0.500	?	?
* FT	**89.86**	0.500	?	?
* RF	89.49	**0.508**	**0.850**	0.003

* The superior CCP performances are bold and type-faced, and the proposed DF models are designated by an asterisk.

De "Intelligent Decision Forest Models for Customer Churn Prediction", por Usman-Hamza, 2022.

Figura 39

Rendimiento de los modelos de predicción en el conjunto de datos 1 balanceado.

	Accuracy (%)	AUC	F-Measure	MCC
NB	78.33	0.866	0.883	0.567
kNN	88.27	0.881	0.883	0.767
DS	64.84	0.65	0.863	0.346
* LMT	93.60	0.971	0.966	0.872
* FT	**94.83**	**0.975**	**0.968**	**0.897**
* RF	92.14	0.943	0.945	0.843

* The superior CCP performances are bold and type-faced, and the proposed DF models are designated by an asterisk.

De "Intelligent Decision Forest Models for Customer Churn Prediction", por Usman-Hamza, 2022.

Figura 40

Rendimiento de los modelos de predicción en el conjunto de datos 2 balanceado.

	Accuracy (%)	AUC	F-Measure	MCC
NB	77.21	0.825	0.772	0.545
kNN	83.91	0.839	0.839	0.678
DS	87.84	0.499	0.500	0.257
* LMT	91.59	0.896	0.876	0.752
* FT	**94.26**	**0.973**	**0.942**	**0.883**
* RF	90.32	0.880	0.503	0.806

* The superior CCP performances are bold and type-faced, and the proposed DF models are designated by an asterisk.

De "Intelligent Decision Forest Models for Customer Churn Prediction", por Usman-Hamza, 2022.

Artículo N° 12

Título: A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees (Nuevo algoritmo de clasificación híbrido para la predicción de abandono de clientes basado en regresión logística y árboles de decisión.)

Aporte: Se propone un nuevo modelo de clasificación híbrido llamado logit leaf model (LLM) el cual mitiga las limitaciones que tienen los modelos de árbol, problema para manejar relaciones lineales entre variables, y los modelos de regresión logística, dificultades con los efectos de interacción entre variables. Para ello el LLM construye distintos modelos

en distintos segmentos de los datos, en lugar de usar el dataset completo para predecir, y al mismo tiempo manteniendo la comprensibilidad de los modelos construidos en las hojas.

Proceso: Configuración de experimento:

- Datos y diseño de experimento: Se utilizaron catorce conjuntos de datos de abandono de diferentes fuentes. El diseño del experimento consiste en hacer un benchmark del desempeño de LLM en comparación con los cuatro algoritmos mencionados anteriormente.
- Preprocesamiento de datos: Se aplica imputación de valores faltantes, dependiendo de la variable se imputa cero, promedio o moda. Se crean variables dummy para las variables categóricas y se imputa cero a los datos faltantes. Para atributos con menos de 5% de valores faltantes, se eliminan esas observaciones a fin de limitar el impacto de los procedimientos de imputación. Se aplica detección y tratamiento de valores atípicos a través de una técnica llamada Winsorization para transformar los valores atípicos valores aceptables que estén dentro de tres desviaciones estándar. Finalmente se aplica la técnica undersampling, donde se reduce el número de observaciones de la clase mayoritaria, a fin de que la proporción de ambas clases sea pareja.
- Selección de variables: Se debe tener en cuenta que un clasificador tiene mejor desempeño cuando es entrenado con un conjunto de datos pequeño y con variables de alto nivel predictivo, en comparación de un conjunto de datos grande que tiene datos redundantes o sucios. Los procedimientos de selección de datos input previene de que el modelo se sobreajuste a datos sucios y aumenta la estabilidad de predicciones del modelo a través de reducir la colinealidad.

Resultados: Se comparó el enfoque híbrido propuesto con árboles de decisión, regresión logística, random forests y modelos de árboles logísticos a nivel de rendimiento predictivo y comprensibilidad. Se usaron las métricas AUC y TDL para medir el rendimiento y se identificó que este fue superior al de los cuatro modelos anteriormente señalados. Además, se valida la comprensibilidad a través de un caso de estudio donde se demuestra que hay mayores beneficios de usar LLM en comparación con árbol de decisiones o regresión logística.

Artículo N° 13

Título: Customer churn prediction in telecom using machine learning in big data platform (Predicción de abandono de clientes en telecomunicaciones utilizando aprendizaje automático en plataforma de big data)

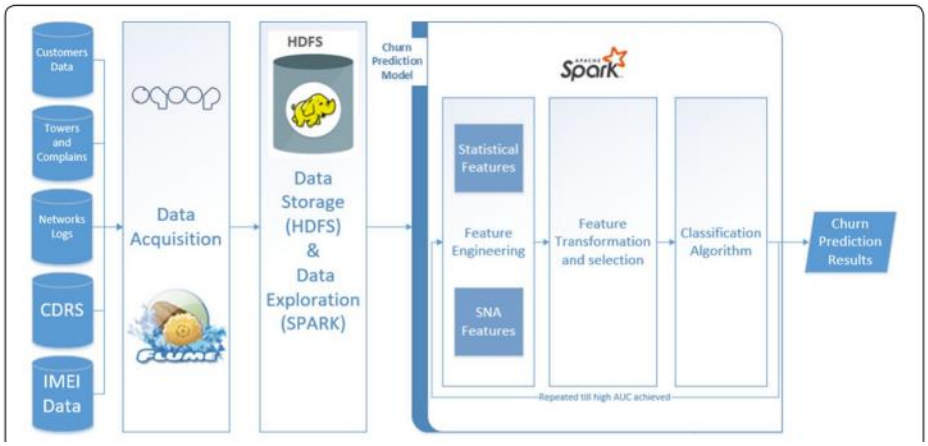
Aporte: La rotación de clientes es un problema importante y una de las preocupaciones más importantes para las grandes compañías, debido al efecto directo sobre los ingresos de las empresas. El principal aporte de este trabajo es desarrollar un modelo de predicción de abandono de clientes que ayude a los operadores de telecomunicaciones a predecir los clientes que probablemente estén sujetos a abandono. Este modelo utiliza técnicas de machine learning en una plataforma de big data. Por otro lado, otro aporte de los autores es utilizar las redes sociales de los clientes en la predicción, mediante la extracción de características de análisis de redes sociales (SNA).

Proceso: En primer lugar, se recopilo un conjunto de datos de nueve meses consecutivos. Este conjunto de datos se utilizará para extraer las variables de predicción de abandono. El ciclo de vida de los datos pasará por varias etapas como se muestra en la Fig.1

El motor Spark se usa en la mayoría de las fases del modelo, como el procesamiento de datos, la función de ingeniería, el entrenamiento y prueba del modelo, esto ya que se utiliza el procesamiento en memoria RAM. Otra ventaja es que este motor contiene una gran variedad de bibliotecas para implementar todo el ciclo de vida del machine learning.

Figura 41

Arquitectura del sistema de predicción de abandono propuesto.

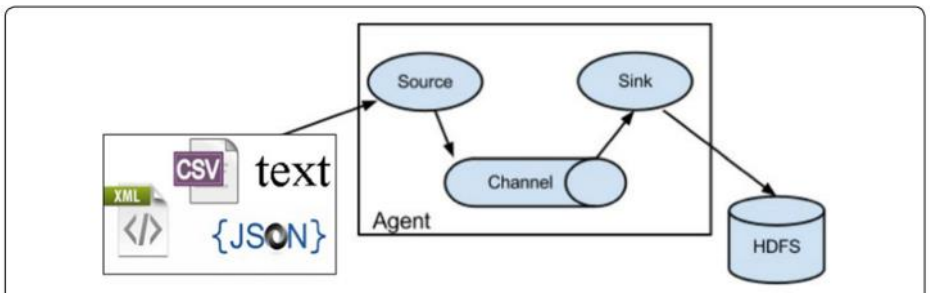


De “Customer churn prediction in telecom using machine learning in big data platform”, por Ahmad, Jafar & Alijoumaa, 2019.

Los datos estructurados, semiestructurados y no estructurados, fueron recopilados mediante Apache Flume, un sistema distribuido que se utiliza para recopilar y mover archivos no estructurados (CSV y texto) y archivos semiestructurados (JSON, XML) a HDFS. La siguiente figura muestra la arquitectura diseñada por los autores

Figura 42

Arquitectura de configuración de sistema Apache Flume.

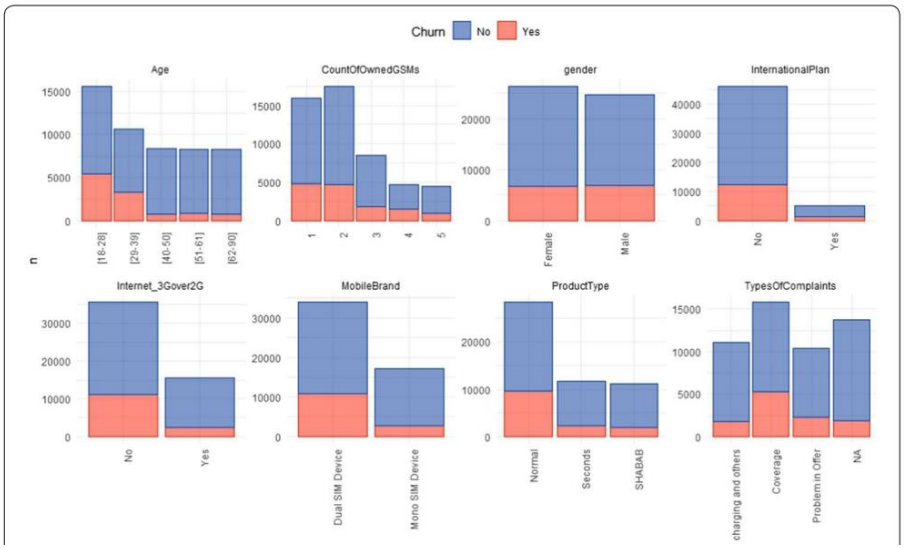


De “Customer churn prediction in telecom using machine learning in big data platform”, por Ahmad, Jafar & Alijoumaa, 2019.

En segundo lugar, se procesaron los datos para convertirlos de su estado natural a variables que puedan ser utilizadas en algoritmos de machine learning. Este es el proceso que llevo más tiempo a los autores debido al gran número de columnas. Dado que se tiene datos relacionados con las acciones de los clientes en redes sociales, además de datos relacionados con llamadas, SMS, MMS y uso de internet para cada cliente por día, semana y mes para cada acción durante los nueve meses, las columnas aumentaron más de 3 veces. Además, se agregaron variables relacionadas con las denuncias presentadas de los clientes. Otras variables estaban relacionadas con el número de quejas, el porcentaje de cobertura de quejas sobre el total de quejas presentadas, la duración promedio entre cada dos quejas secuencialmente, entre otras. En la siguiente imagen se muestra la distribución con respecto a la variable target de algunas de las variables predictoras.

Figura 43

Distribución de algunas variables categóricas de la investigación.



De “Customer churn prediction in telecom using machine learning in big data platform”, por Ahmad, Jafar & Alijoumaa, 2019.

En tercer lugar, se realizó una etapa de limpieza de datos, donde se tomaron las siguientes acciones:

- Se eliminaron los registros que tienen más del 90% de las variables faltantes.
- Se eliminó las variables que tienen más del 70% de valores faltantes.
- Se imputó datos.
- Se categorizo con “otros” algunos valores faltantes en variables categóricas.
- Los valores numéricos faltantes fueron reemplazados con el promedio de la variable.
- Las variables categóricas eran 78, solo se escogió las 31 más frecuentes y las demás fueron reemplazadas por una sola categoría, por lo que el número final fue de 32.

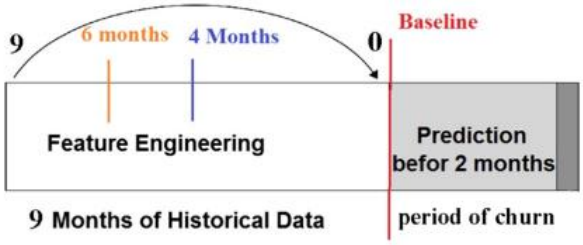
Posteriormente en esta misma etapa, se calcula la correlación entre variables numéricas usando Pearson y se eliminó las variables muy correlacionadas, aunque esta eliminación no tuvo efecto en el resultado final. El número de variables después de estos procesos excedía a las 2000 variables. Para finalizar esta etapa, se etiqueto a cada cliente con el valor “Si abandonó” o “No abandonó”. Para este caso de estudio, se consideraba "Si abandonó” a un cliente con una fase de inactividad de 2 de los 9 meses. El dataset final de la muestra cuenta

con 5 millones de clientes donde 4 700 000 clientes son activos (“No abandono”) y 300 000 clientes que si abandonaron.

La siguiente figura muestra los periodos de datos históricas y el periodo futuro en el que el cliente puede dejar la empresa. Con apoyo de marketing se estableció predecir el abandono antes de los 2 meses de la acción de abandono real, con el fin de tener suficiente tiempo para la acción proactiva que se pueda tomar como negocio con estos potenciales clientes que abandonaran.

Figura 44

Periodos de datos históricos y futuros.



De “Customer churn prediction in telecom using machine learning in big data platform”, por Ahmad, Jafar & Alijoumaa, 2019.

En cuarto lugar, la solución propuesta divide los datos en dos grupos: el grupo de entrenamiento y el grupo de prueba. El grupo de entrenamiento consta del 70% del conjunto de datos y el grupo de prueba consta de un 30%. Los hiperparámetros de los algoritmos se optimizaron mediante validación cruzada k-fold (valor de k=10). La variable target esta desequilibrada y esto puede causar un cambio significativo y un impacto negativo en los modelos finales, por lo que se solucionó este problema tomando una muestra de datos para hacer las dos clases equilibradas (Li Y, Luo P & Wu C, 2014).

En quinto lugar, se realizó el entrenamiento de distintos algoritmos. Se comenzó a entrenar el algoritmo de árbol de decisiones, experimentando con varias configuraciones de hiperparámetros y número máximo de nodos. El numero optimizado de nodos fue de 398 nodos y el valor de profundidad fue de 20.

También se entrenó el algoritmo de Random Forest, se experimentó cambiando el hiperparámetro de número máximo de árboles con los siguientes valores: 100,200,300,400 y 500 árboles. Los mejores resultados se obtuvieron con una cantidad máxima de árboles de

200. Se entrenó el algoritmo de GBM, optimizando el hiperparámetro de número de árboles con valores de hasta 500 árboles, teniendo los mejores resultados también con 200 árboles. El algoritmo de GBM dio mejores resultados que random forest y decision tree.

Finalmente se instaló el algoritmo XGBOOST en el framework spark 2.3 y se integró con la biblioteca ML en Spark, se aplicaron los mismos pasos de los últimos 3 algoritmos, optimizando los hiperparámetros y obteniendo mejores resultados con 180 árboles.

Principal Resultado: El método de preparación y selección de variables y el ingreso de las variables provenientes de redes sociales fueron fundamentales para el éxito de este modelo, ya que el valor de AUC en la empresa donde se realizó el caso de estudio, SiriaTel, alcanzó el 93.301% mediante el algoritmo modelo de árbol XGBOOST, el cual logro los mejores resultados. EL algoritmo GBM, el bosque aleatorio y el árbol de decisión quedaron en los lugares siguientes.

Figura 45

Comparación de resultados AUC antes y después de adicionar SNA a las variables estadísticas.

Features	XGBOOST (%)	GSM (B) (%)	Random Forest (%)	Decision Tree (%)
Statistical features	84	82	79.1	76
SNA features	75.3	71	69	67.2
Statistical and SNA features	93.3	90.89	87.76	83

De “Customer churn prediction in telecom using machine learning in big data platform”, por Ahmad, Jafar & Alijoumaa, 2019.

Se ha evaluado el modelo con un nuevo conjunto de datos relacionados con diferentes periodos de tiempo y sin ninguna acción proactiva de marketing, XGBOOST también dio el mejor resultado con 89% de AUC, la disminución se puede deber al fenómeno del modelo de datos no estacionario, por lo que el modelo necesita entrenamiento cara período de tiempo.

Figura 46

Resultados AUC para cada algoritmo de clasificación en el conjunto de datos "NotOffered".

Algorithm	XGBOOST	GSM (B)	Random Forest	Decision Tree
SyriaTel New Data "NotOffered"	89%	85.5%	83.4%	79.1%

De "Customer churn prediction in telecom using machine learning in big data platform", por Ahmad, Jafar & Alijoumaa, 2019.

Artículo N° 14

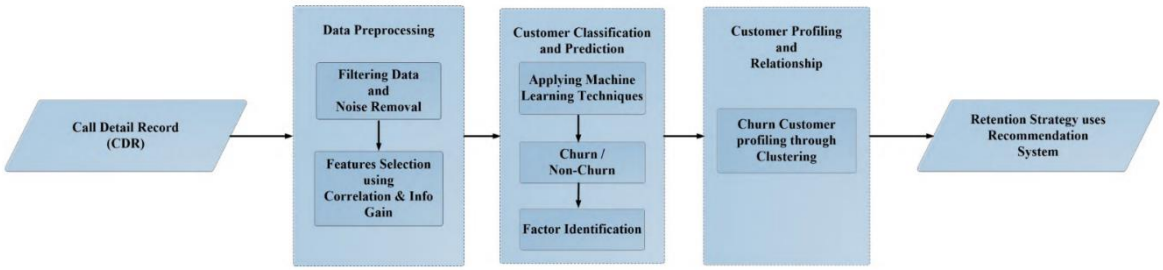
Título: A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector (Un modelo de predicción de abandono utilizando Random Forest: análisis de técnicas de machine learning para la predicción de abandono e identificación de factores en el sector de las telecomunicaciones).

Aporte: Los autores proponen un modelo de predicción de abandono que utiliza clasificación, así como técnicas de agrupamiento para identificar los clientes de abandono y proporciona los factores detrás del abandono de clientes en el sector de las telecomunicaciones.

Proceso: La selección de características se realiza utilizando la ganancia de información y el filtro de clasificación de atributos de correlación. El modelo propuesto primero clasifica los datos de abandono de clientes mediante algoritmos de clasificación, en los que el algoritmo Random Forest (RF) tuvo buen desempeño con un 88,63 % de clasificaciones acertadas. Después de la clasificación, el modelo propuesto segmenta los datos de los clientes que se retiran al categorizar a los clientes en grupos utilizando la similitud del coseno para proporcionar ofertas de retención basadas en grupos. Este documento también identificó factores de abandono que son esenciales para determinar las causas fundamentales del abandono.

Figura 47

Proceso propuesto por la investigación.



De "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", por Ullah et al., 2019.

Resultado: El modelo de predicción de abandono propuesto se evalúa utilizando métricas, como exactitud, precisión, recuperación, medida f y área de características operativas de recepción (ROC). Los resultados revelan que el modelo de predicción de abandono propuesto produjo una mejor clasificación de abandono utilizando el algoritmo de RF y la creación de perfiles de clientes utilizando k-means clustering.

Artículo N° 17

Título: Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms (Estrategia dinámica de predicción de abandono de clientes para inteligencia de negocios usando análisis de texto con algoritmos evolutivos y optimizados)

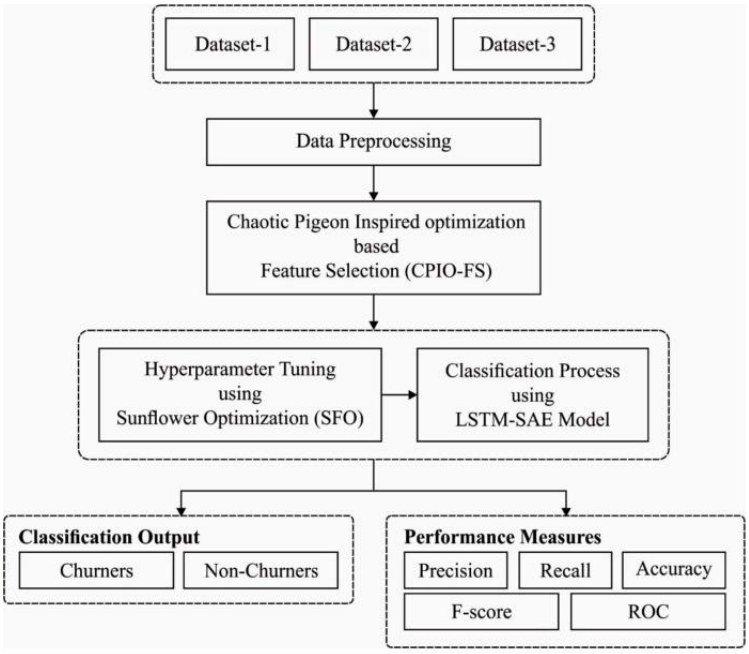
Aporte: Los autores presentan un trabajo de investigación que propone un nuevo modelo predictivo de abandono de clientes para promover la inteligencia empresarial en el sector de las telecomunicaciones. Debido a que las características entre los clientes que abandonan y los que no abandonan son similares, este trabajo de investigación diseña un modelo de estrategia dinámica utilizando un algoritmo de análisis de texto con optimizaciones metaheurísticas (CCPBI-TAMO). Además, utilizan técnicas de selección de características basadas en optimizaciones inspiradas en palomas caóticas (CPIO-FS) que es empleada para el proceso de selección de características y reduce la complejidad del cálculo. Por otro lado, nos aportan utilizando la optimización de girasoles (SFO) para optimizar y ajustar los

hiperparámetros para mejorar más aun el rendimiento del modelo en cuanto a la predicción de abandono de clientes (CPP)

Proceso: En este estudio se introduce un nuevo modelo predictivo de abandono de cliente para promover la inteligencia empresarial en el sector de las telecomunicaciones. Para lograr esto, los autores proponen el siguiente modelo.

Figura 48

Proceso general de trabajo CCPBI-TAMO.



De “Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms”, por Pustokhina, 2019.

Este modelo CCPBI-TAMO realiza un preprocesamiento de los datos comerciales en diferentes etapas para mejorar la calidad de los datos. Seguido por un proceso de selección de características utilizando CPIO-FS, lo cual da como resultado subconjuntos de características óptimas. Luego los datos comerciales se pueden clasificar en usuarios que abandonan y usuarios que no abandonan mediante el uso del modelo LSTM-SAE, donde los hiperparámetros se ajustan mediante el algoritmo SFO.

Para la primera etapa de preprocesamiento, los datos recopilados del sector comercial se procesan previamente para convertir los datos en datos significativos, para lograr esto se realizaron tres subprocesos: transformación de datos (incluye la conversión de variables

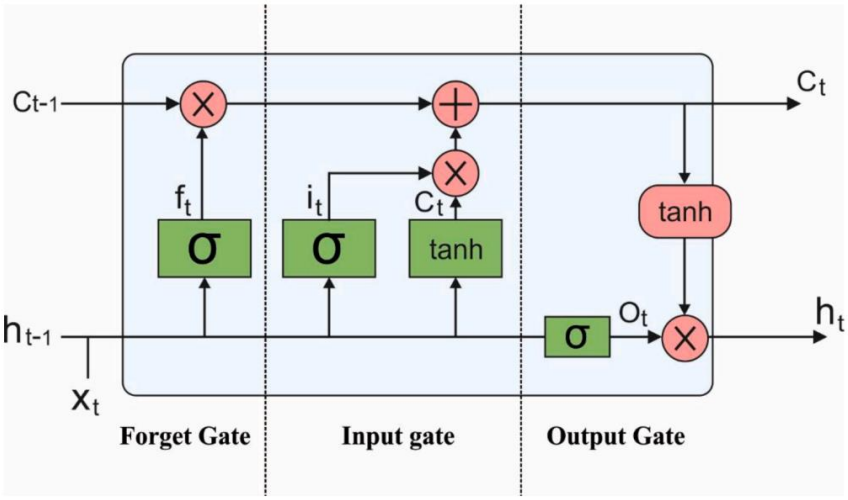
categóricas en valores numéricos), etiquetado de clases y normalización “min-max" para mejorar la calidad de los datos y estén en un nivel uniforme.

Para la segunda etapa de elección de características, se utiliza CPIO-FS, un enfoque metaheurístico novedoso que es estimulado por el comportamiento de las palomas que regresan a casa (Tian et al., 2020).

Una vez se genera el subconjunto reducido de características utilizando el algoritmo CPIO-FS, la tercera etapa es la clasificación de datos utilizando el modelo LSTM-SAE, para determinar presencia de clientes que abandonan y no abandonan. La figura ilustra la estructura de LSTM capaz de minar dependencias de corta distintas con una seria de características básicas y se representa como una compacta y potente representación

Figura 49

Estructura de LSTM.

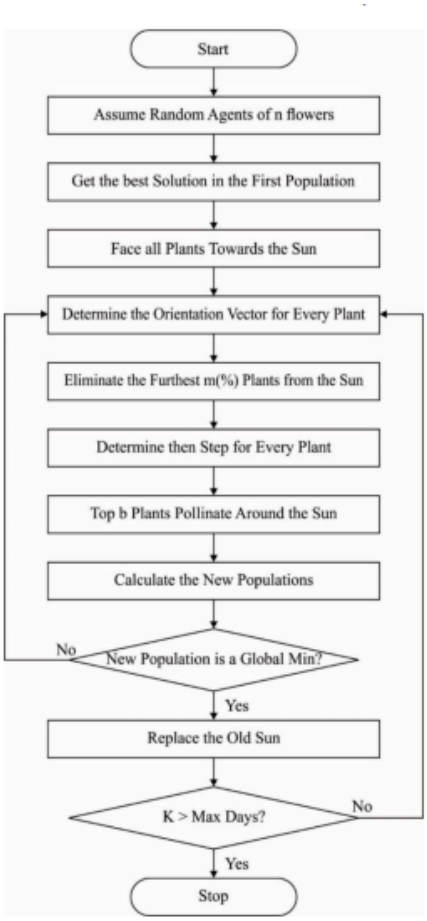


De “Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms”, por Pustokhina, 2019.

Como cuarta etapa del proceso, se busca optimizar y cambiar los hiperparámetros con el fin de mejorar la precisión del modelo, para esto utilizan el algoritmo SFO de tal manera que la interpretación de la clasificación se puede improvisar.

Figura 50

Diagrama de flujo del algoritmo SFO.



De “Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms”, por Pustokhina, 2019.

Finalmente, se validaron los resultados experimentales del modelo CCPBI-TAMO en los conjuntos de datos de referencia de tres abandonos (Amín et al., 2019). El modelo se simula utilizando la herramienta Python 3.6.5 y los resultados se inspeccionan y evalúan con medidas diferentes.

Principal Resultado: Se realiza un análisis de simulación detallado en el conjunto de datos de predicción de rotación de clientes de referencia y los valores experimentales destacaron el rendimiento superior del modelo prepuesto sobre los otros métodos comparados con la precisión máxima de 95.56%, 93.44% y 92.74% en los conjuntos de datos aplicados 1-3 respectivamente. Por otro lado, el modelo propuesto ayuda a extraer, examinar y visualizar indicadores clave de rendimiento KPIs del gran volumen de datos empresarial, ayudando en

la eficacia del proceso de toma de decisiones. Como siguientes pasos, los autores proponen realizar el modelo en internet de las cosas (IoT) y en un entorno basado en la nube para predecir abandonos en aplicaciones en tiempo real como telecomunicaciones, comercio electrónico, planificación de viajes, etc.

Artículo N° 19

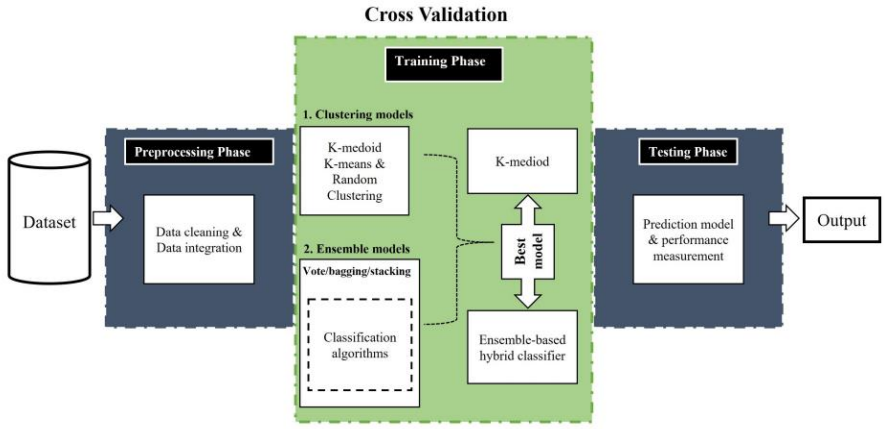
Título: An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms (Un esquema híbrido inteligente para la predicción de abandono de clientes que integra algoritmos de agrupación y clasificación)

Aporte: En este estudio los autores proponen predecir el abandono de clientes basada en un sistema que incorpora el aprendizaje de agrupamiento y técnicas de clasificación. El aporte es la propuesta de un modelo de predicción de abandono utilizando algoritmos de clasificación y un conjunto de agrupamiento para mejorar el rendimiento del modelo. La mayoría de los estudios utilizan los más conocidos algoritmos de agrupamiento como k-means, k-medoids y random para probar conjuntos de datos de predicción de abandono, sin embargo, en este estudio con el objetivo de mejorar los resultados, se utiliza una técnica hibrida utilizando diferentes algoritmos de conjuntos para evaluar el rendimiento del modelo propuesto. Los algoritmos de agrupamiento mencionados se integran con diferentes clasificadores: Gradient Boosted Tree (GBT), Decision Tree (DT), Random Forest (RF), Deep Learning (DL), and Naive Bayes (NB). Esto se evalúa en un conjunto de datos de telecomunicaciones que se adquirieron de Orange y Cell2Cell.

Proceso: Los autores proponen un enfoque que mejore la efectividad de los resultados de predicción de abandono mediante el empleo de una técnica hibrida que combina varios agrupamientos, clasificaciones y conjuntos, incluyendo embolsado y apilamiento. En la siguiente imagen se presenta la arquitectura propuesta por los autores:

Figura 51

Framework propuesto por la investigación para la predicción de abandono de clientes.



De "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms", por Liu et al., 2022.

Los conjuntos de datos incluidos en este estudio son de fácil acceso a través de internet. Orange Telecom ofrece su información en su sitio web (Miller et al, 2009). El otro conjunto de datos fue proporcionado por Cell2Cell y está disponible en el sitio web del centro de gestión de relaciones con los clientes de la universidad de Duke (Verbeke et al, 2012). El primer conjunto de datos está disponible en línea con la distribución de su clase desequilibrada, en cambio, el segundo conjunto de datos ha sido preprocesado y se tiene una versión balanceada. Posteriormente se realiza cierta limpieza de datos ya que hay variables que solo tienen un solo valor o tienen demasiados nulos. Las características de los conjuntos de datos son detalladas por los autores en la siguiente tabla:

Figura 52

Características de los conjuntos de datos aplicados al estudio.

Attributes	Cell2Cell	Orange
Complete examples	40,000	50,000
Complete features	76	260
Numerical features	68	190
Nominal features	8	70
Data sharing	Balanced	Imbalanced
Missing values	No	Yes

De "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms", por Liu et al., 2022.

Una vez obtenido el dataset, viene la fase de preprocesamiento de datos, ya que este conjunto de datos contiene irregularidades, como valores faltantes, duplicados, etc. La eliminación de este ruido o de variables vacías se maneja a través de la herramienta WEKA. Además, todas las variables de la muestra se cambian a una representación numérica para clasificarles en clases pequeñas, medianas y grandes según el número de observaciones en cada clase (Sorokina, 2009).

La fase de entrenamiento se divide en dos subprocesos, el entrenamiento que se realizara con algoritmos de clustering (k-means, k-medoid, random clustering) y el entrenamiento con los algoritmos de clasificación (Random Forest, Decision Tree, Gradient Boosted Tree, etc.). Para la agrupación en clústeres en primera instancia se proporciona al algoritmo una gran cantidad de datos sin etiquetar para que el programa descubra cualquier patrón que pueda en los datos.

Finalmente se realiza una etapa de testing según el proceso propuesto por los autores, donde se compara las distintas métricas de cada algoritmo con el objetivo de elegir al mejor, tanto para los algoritmos de clustering como para los algoritmos de clasificación. Las medidas de evaluación de rendimiento son las siguientes:

- **Accuracy:** La precisión es el porcentaje de ejemplos verdaderamente predichos para todos los tipos de predicciones hechas por el modelo.
- **Precision:** estiman el total de ejemplos identificados como positivos por el algoritmo verdaderamente pertenecen a la clase positiva.

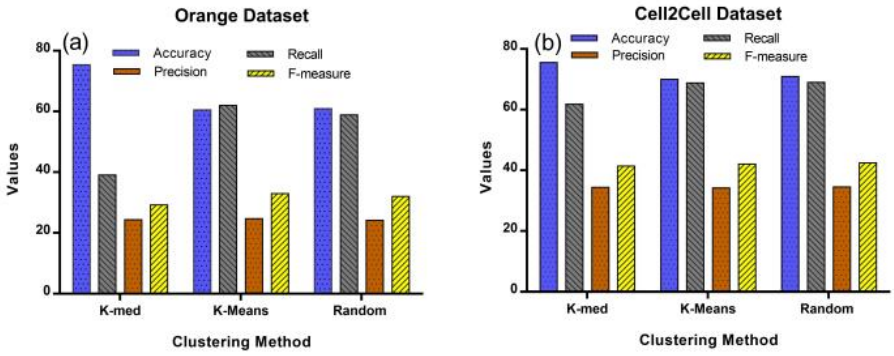
- **Recall:** as medidas de recuerdo estiman a qué sección de muestras pertenecen realmente la clase positiva es verdaderamente predicha positivamente por el modelo.
- **F-measure:** La medida F se calcula tomando la media armónica de precisión y recuerdo.

Principal Resultado: Se compararon varios enfoques no supervisados empleando los modelos más populares de agrupamiento (k-means, k-med y Random). El rendimiento de las técnicas de agrupamiento se muestra a través de un gráfico en la siguiente imagen. El eje "x" indica el algoritmo de agrupamiento y el eje "y" los valores del rendimiento para las 4 medidas que se establecieron como métricas para la evaluación.

El grafico muestra que la técnica k-medoids alcanza el 75.44% de accuracy en el conjunto de datos Orange, y el 75.56% de accuracy para en conjunto Cell2Cell. El resultado principal para este experimento es que el algoritmo k-medoids supera a los demás en cuanto a métricas de evaluación.

Figura 53

Primeros resultados de la investigación para cada algoritmo.

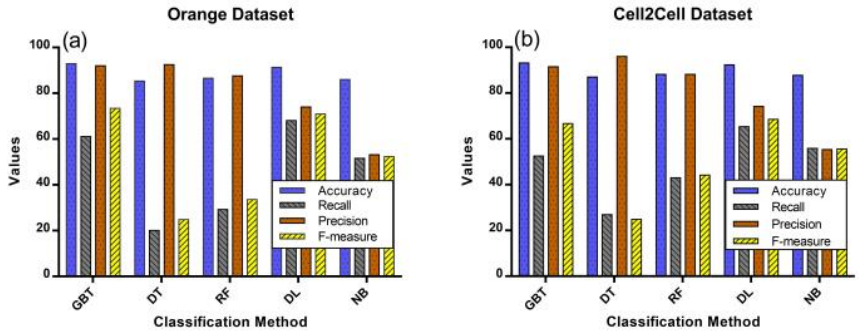


De "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms", por Liu et al., 2022.

Para el segundo experimento, los resultados que comparan los distintos algoritmos de clasificación se muestran en el siguiente gráfico. El algoritmo Gradient Boosted Tree (GBT) demostró el mayor accuracy en ambos conjuntos de datos, 92,98% para el conjunto de datos de Orange y 93.19% para el conjunto de datos de Cell2Cell, mientras que el algoritmo de decision tree (DT) demuestra los resultados más bajos en accuracy.

Figura 54

Segundos resultados de la investigación para cada algoritmo.

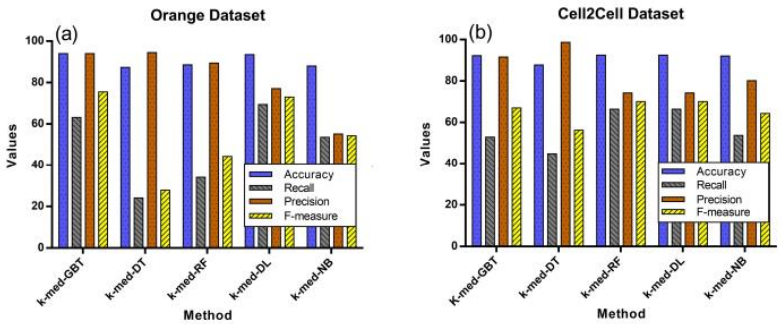


De "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms", por Liu et al., 2022.

Finalmente, en el tercer experimento los autores construyen una red hibrida fusionando el método de agrupamiento k-medoids con cada clasificador (DT, RF, GBT, DL y NB). Esta propuesta hibrida funciona empleando el algoritmo de agrupamiento k-medoids para generar conjuntos de entrenamiento y de prueba. Los resultados de la prueba indica que el modelo hibrido de k-medoids con GBT superan a las otras combinaciones, ya que esta combinación logra accuracy de 94% y 92,25% respetivamente.

Figura 55

Terceros resultados de la investigación para cada algoritmo de clasificación híbrido.



De "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms", por Liu et al., 2022.

Artículo N° 20

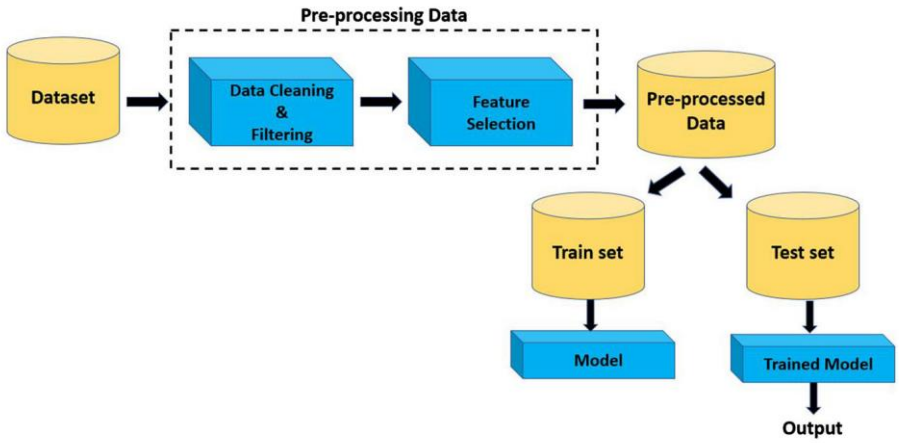
Título: Customer churn prediction system: a machine learning approach (Sistema de predicción de abandono de clientes: un enfoque de aprendizaje automático)

Aporte: El aporte de los autores es haber aplicado un algoritmo de búsqueda gravitacional para realizar la selección de variables y para reducir las dimensiones del conjunto de datos. Además, después del procesamiento previo de los datos, han aplicado distintas técnicas de aprendizaje automático que se utilizan para predicciones, como la regresión logística, SVM, entre otras. Adicionalmente, realizaron una validación cruzada k-fold para evitar el sobreajuste. Por otro lado, los autores han podido utilizar el poder del machine learning para optimizar los algoritmos y mejorar los resultados. Para la validación, han evaluado los algoritmos en el conjunto de prueba usando la matriz de confusión y el área bajo la curva AUC.

Proceso: La metodología propuesta por los autores consta de seis fases. En las dos primeras fases se realiza el preprocesamiento de datos y el análisis de variables. En la tercera fase, se realiza la selección de variables mediante el algoritmo de búsqueda gravitacional. A continuación, los datos se dividen en dos partes: conjunto de entrenamiento 80% y conjunto de prueba 20%.

Figura 56

Arquitectura de sistema propuesta en la investigación.

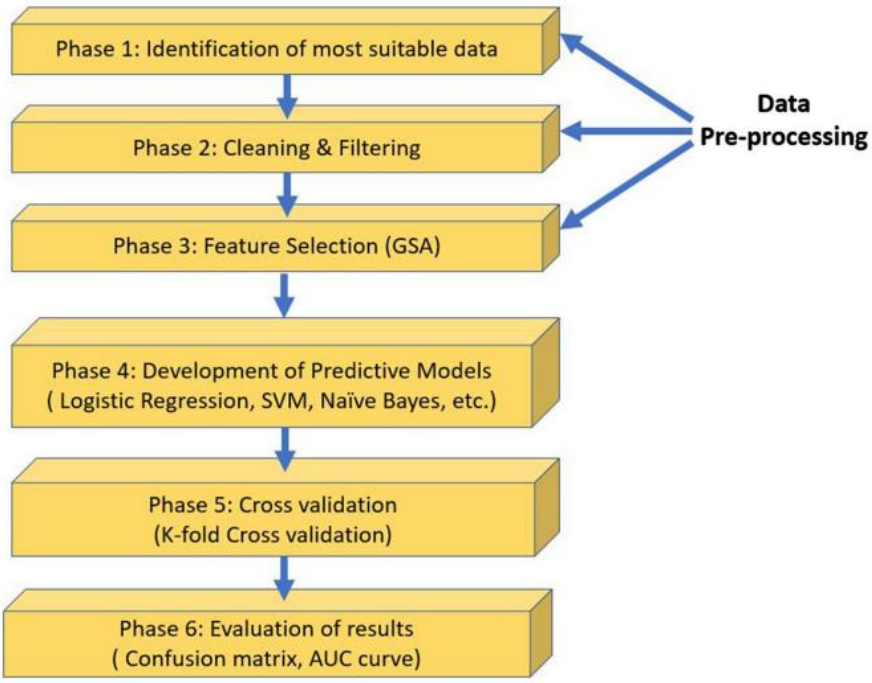


De “Customer churn prediction system: a machine learning approach”, por Lalwani et al., 2022.

El modelo propuesto a detalle consta de 6 fases: Identificación de los datos más adecuados (análisis de varianza, matriz de correlación, eliminación de valores atípicos, etc.). Fase 2: Limpieza y filtrado (manejo de valores nulos y faltantes). Fase 3: Selección de variables (usando GSA). Fase 4: Desarrollo del modelo predictivo (Regresión logística, SVM, Naive Bayes, etc.). Fase 5: validación cruzada (utilizando k-fold). Fase 6: Evaluación de los modelos predictivos sobre el conjunto de testing (usando matriz de confusión y la curva AUC).

Figura 57

Framework modelo multi fase para desarrollar y gestionar el desarrollo de un modelo predictivo de abandono de clientes.



De “Customer churn prediction system: a machine learning approach”, por Lalwani et al., 2022.

En el proceso de predicción, se aplicaron los modelos predictivos más populares: logistic regression, naive bayes, support vector machine, random forest, decision trees, etc. La validación cruzada de K-folds se utiliza sobre el conjunto de entrenamiento para el ajuste de hiperparámetros y para evitar el sobreajuste de modelos. El procedimiento de re-muestreo

es utilizado para evaluar modelos de machine learning en un conjunto de datos limitado. El procedimiento tiene un solo parámetro llamada “k”, que se refiere al número de grupo divididos en una muestra de datos determinada. La validación cruzada por “k-fold” divide el conjunto de entrenamiento en “k” grupos. De los grupos divididos un grupo se elige aleatoriamente como conjunto de prueba y los de más se quedan como conjuntos de entrenamiento. A partir de ahí, se ajusta el modelo y se valida la puntuación obtenida con los datos. Los resultados obtenidos en esta fase del proyecto se muestran en la siguiente tabla:

Figura 58

Resultados de validación cruzada k-fold para todos los modelos.

Model	k-fold cross validation (cv=5)%
Logistic regression	79.85
Decision tree	79.56
Adaboost classifier	80.72
Adaboost classifier (Extra Tree)	80.41
KNN classifier	78.51
Random forest	79.28
Random forest (adaboost)	80.39
Naive bayes (gaussian)	75.86
SVM classifier linear	78.65
SVM classifier poly	79.75
SVM (adaboost)	73.48
XGboost classifier	79.5
CatBoost classifier	80.34

De “Customer churn prediction system: a machine learning approach”, por Lalwani et al., 2022.

Finalmente, los resultados obtenidos del conjunto de prueba se evaluaron utilizando la matriz de confusión y la curva AUC.

Principal Resultado: El resultado del estudio es que los algoritmos de Adaboost y XGboost Classifier obtuvieron la mejor precisión con 81.71% y 80.8% respectivamente. La puntuación más alta de AUC de 84% la obtuvieron los algoritmos Adaboost y XGboost Classifiers, superando a los demás algoritmos predictivos.

Figura 59

Comparación de modelos machine learning.

Model	Accuracy(%)	Recall(%)	Precision(%)	F-Measure(%)	AUC Score %
Logistic Regression	80.45	80.23	79.11	78.89	82
Logistic Regression (Adaboost)	76.57	75.57	56.61	64.71	78
Decision Tree	80.14	80.1	78.81	78.89	83
Adaboost Classifier	81.71	81.21	80.14	80.28	84
Adaboost Classifier (Extra Tree)	81.14	81.64	80.57	80.60	72
KNN Classifier	79.64	79.71	78.38	77.00	80
Random Forest	78.04	78.68	77.54	77.91	82
Random Forest (Adaboost)	81.21	81.28	80.19	80.29	82
Naive Bayes (Gaussian)	77.07	77.12	77.60	77.31	80
SVM Classifier Linear	79.14	79.89	78.67	78.86	79
SVM Classifier Poly	80.21	80.64	79.66	78.11	80
SVM (Adaboost)	74.07	74.43	54.91	63.17	80
XGBoost	80.8	80.7	80.3	78.7	84
CatBoost	81.8	82.2	81.2	79.6	82

De “Customer churn prediction system: a machine learning approach”, por Lalwani et al., 2022.

Pregunta 3: ¿Qué metodologías existen para elaborar modelos predictivos en el sector financiero?

Artículo N° 5

Título: A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM (Un enfoque de deep learning para la calificación crediticia de los préstamos entre pares mediante el mecanismo de atención LSTM)

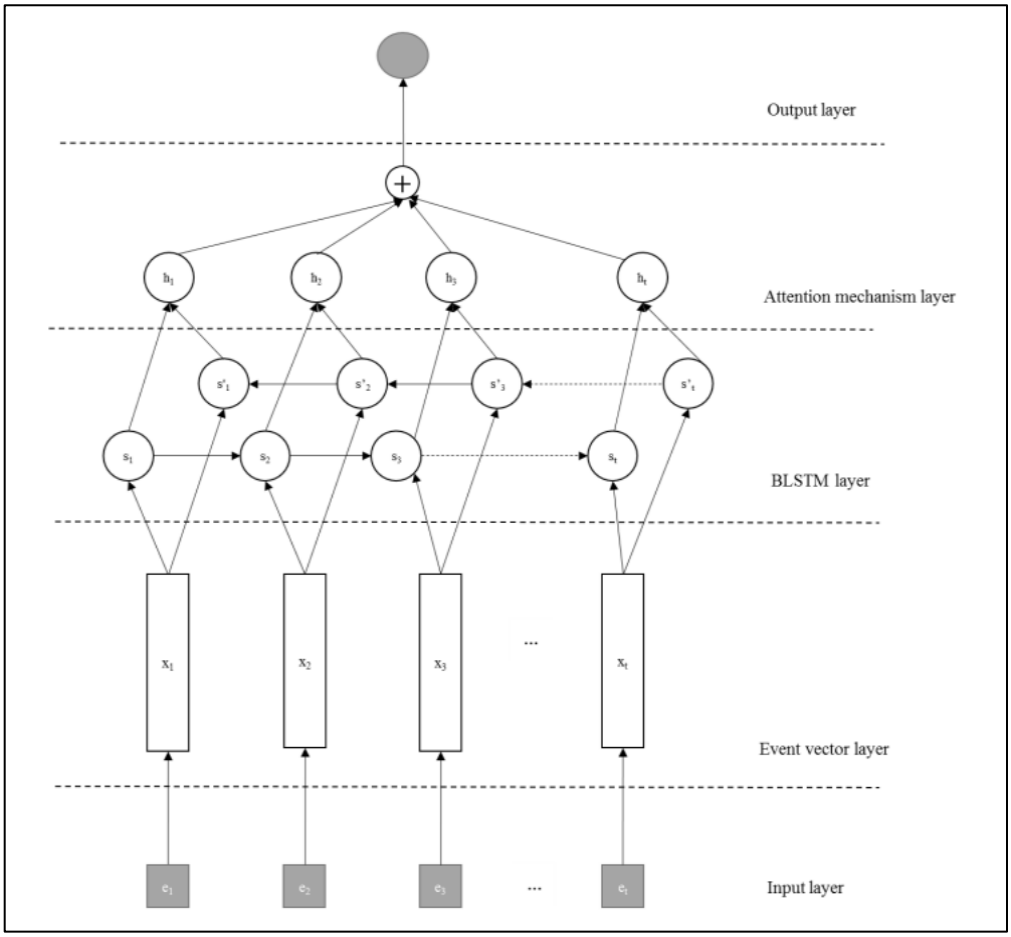
Aporte: Los autores proponen un método de evaluación de crédito mediante el mecanismo de atención LSTM usando los datos de comportamiento de operaciones online de los prestatarios, con la finalidad de predecir si el préstamo terminará en morosidad. El experimento se realiza con datos reales y se demuestra que la solución propuesta puede

mejorar la exactitud de la precisión comparado con los métodos tradicionales artificiales de extracción de variables y el modelo LSTM estándar.

Proceso: Con el modelo propuesto se plantea tratar la secuencia de comportamientos como oraciones y los eventos como palabras, usando un método de evaluación de crédito basado en mecanismo de atención LSTM (AM-LSTM), este consiste en cinco capas: capa de entrada, capa de vectores de evento, capa BLSTM, capa de mecanismo de atención y capa de salida.

Figura 60

Modelo propuesto basado en mecanismo de atención LSTM.



De “A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM”, por Wang et al., 2019.

Es importante mencionar que la capa de entrada recibe una secuencia de eventos cronológicos del prestatario, y que para recibir la salida en la capa final se usa una función

sigmoide para obtener la probabilidad de la morosidad. Para evaluar el modelo se considerarán las métricas de desempeño curva ROC, AUC y KS. Para demostrar la ventaja del método AM-LSTM se realizó un análisis comparativo mediante tres métodos: BOA-XGBoost, LSTM, BLSTM y BLSTM-Meanpool.

Configuración de experimento:

- Conjunto de datos: Los datos fueron provistos de forma anónima por una plataforma de préstamos P2P en China y la base consta de un total de 100 mil prestatarios. Además, se cuenta con una etiqueta llamada morosidad que puede tener el valor de 1 o 0, si el prestatario cayó en morosidad o no respectivamente.
- Configuración de parámetros y entrenamiento del modelo: Se usó la librería de Python Keras como framework de deep learning y Tensorflow para el backend. Se empleó 80% de los datos para entrenamiento y 20% para validación, además 10% de los datos de entrenamiento se separaron para seleccionar los mejores hiperparámetros a través de búsqueda de grilla (grid search), quedando finalmente la siguiente configuración

Figura 61

Propuesta de configuración de hiperparámetros.

Parameter	*Parameter Description*	*Parameter Value*
maxlen	event sequence length	100
n_events	Number of _events type	200
event_dim	dimension of event vector	16
lstm_units	neurons number of LSTM model	50
lstm_dropout	Dropout ratio	0.2
mini_batch_size	Mini-batch size	64

De “A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM”, por Wang et al., 2019.

Resultado: Se realizaron múltiples experimentos usando espacios vectoriales de eventos de 2, 4, 8, 16, 32, 48, 64 y 128 dimensiones y se determinó que, a partir de 16 dimensiones, el efecto de la predicción se mantiene estable, de esa forma se puede obtener el mejor desempeño del modelo al mismo tiempo que previniendo el sobreajuste.

Se obtuvieron las siguientes métricas de KS y AUC para los cinco métodos usados, y se puede comprobar que BLSTM, comparado con el estándar LSTM, tiene una mejora de 9.69% en KS y 3.21% en AUC. Por otro lado, BLSTM-Meanpool comparado con BLSTM tiene una mejora de 3.72% en KS y 0.93% en AUC. Sin embargo, el modelo propuesto AM-LSTM fue el que obtuvo las mejores métricas con 2.46% en KS y 66.9% en AUC.

Figura 62

Comparación de resultados obtenidos por cada algoritmo.

Models	KS	AUC
BOA-XGBoost	0.099	0.561
LSTM	0.196	0.623
BLSTM	0.215	0.643
BLSTM-Meanpool	0.223	0.649
AM-LSTM	0.246	0.669

De "A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM", por Wang et al., 2019.

Artículo N° 6

Título:

Dropout early warning systems for high school students using machine learning (Sistemas de alerta temprana de deserción de alumnos de secundaria usando machine learning)

Aporte: Los autores han identificado que los modelos predictivos usando machine learning tienen un gran potencial para desarrollar un sistema de alerta temprana para identificar a alumnos con riesgo de desertar a fin de ayudarlos proactivamente. El sistema que desarrollan se basa en la técnica de machine learning, random forest.

Proceso: El primer paso de los autores es obtener los datos a usar en el modelo predictivo. Estos fueron obtenidos de los datos del Sistema de Información de Educación Nacional (NEIS) del 2014. NEIS es un sistema nacional de información de administración educativa conectado a través de Internet con alrededor de 12,000 escuelas, 17 oficinas de educación de la ciudad / provincia y el Ministerio de Educación. NEIS conecta las escuelas a través de Internet y procesa toda la administración educativa, incluidos los datos de los estudiantes,

los recursos humanos, el presupuesto y las funciones de contabilidad, que deben ser manejadas por todas las escuelas primarias y secundarias en Corea.

El segundo paso es definir el resultado del modelo predictivo. En este caso es la deserción de estudiantes. El resultado binario que representa la deserción del alumno es creado en base a las variables “cambio de registro escolar” y “razón de la deserción”. Los valores “expulsión” y “abandono” son considerados como deserción. El tercer paso es configurar y establecer los parámetros, basados en el set de datos de NEIS, para predecir la deserción de estudiantes. En el estudio se usaron 12 parámetros.

El cuarto paso es entrenar el modelo random forest. Para ello se usó el paquete Caret incluido en el software R. Los 165715 registros en el set de datos original fue dividido en datos de entrenamiento (80%) y datos de prueba (20%). El número de árboles de decisión en el modelo random forest fue establecido en 500.

Resultado: La validación de 10 pliegues cruzados demostró que el óptimo número de parámetros elegidos aleatoriamente es 7. La siguiente figura muestra un ejemplo de árboles de decisión de los 500 árboles de decisión entrenados en el modelo random forest.

Figura 63

Ejemplo de árbol de decisión entrenado en el modelo random forest.

De “Dropout early warning systems for high school students using machine learning”, por Chung & Lee, 2019.

A continuación, se presentan los resultados de las principales métricas de rendimiento usadas para evaluar modelos predictivos.

Figura 64

Resultados de las principales métricas de rendimiento usadas para evaluar los modelos predictivos.

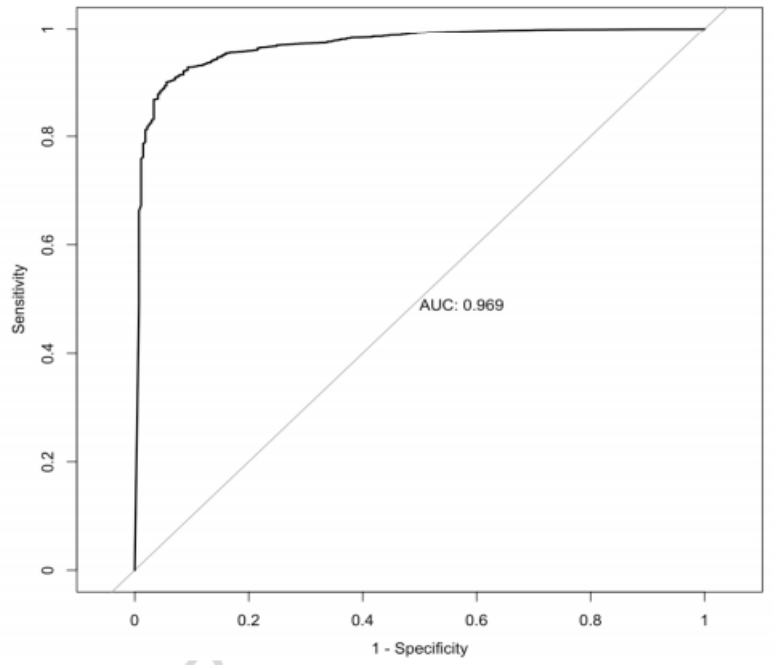
		True	
		Dropout	Non-dropout
Predicted	Dropout	228	1711
	Non-dropout	41	31162

Performance Metrics	Accuracy	Sensitivity	Specificity	AUC
Values	0.95	0.85	0.95	0.97

De “Dropout early warning systems for high school students using machine learning”, por Chung & Lee, 2019.

Figura 65

Gráfica de resultados para la curva AUC.



De “Dropout early warning systems for high school students using machine learning”, por Chung & Lee, 2019.

También se definen los valores escalares de la importancia de las variables en el modelo random forest.

Figura 66

Listado de variables con su importancia.

Variables	Importance of variables
Unauthorized absence	100.0000
Unauthorized lateness	43.1058
Time of self-regulated activity	32.6835
Time of career development	31.3415
Unauthorized early leave	29.0357
Time of club activity	19.5394
Unauthorized lateness in first 4 weeks	17.7885
Time of volunteer work	13.7303
Unauthorized class absence	11.6224
Unauthorized absence in first 4 weeks	9.0998
Unauthorized class absence in first 4 weeks	0.3937
Unauthorized early leave in first 4 weeks	0.0000

De "Dropout early warning systems for high school students using machine learning", por Chung & Lee, 2019.

Artículo N° 9

Título: A machine learning approach to predict the success of crowdfunding fintech project (Un enfoque de machine learning para predecir el éxito de proyectos fintech crowdfunding)

Aporte: Los autores proponen un método de redes neuronales artificiales (ANN) basado en ensemble machine learning para de esta forma generar distintas redes neuronales y prevenir el problema de sobreajuste, usando variables relacionadas a la teoría de capital social, teoría de capital humano y teoría de nivel de procesamiento (LOP) en mira de ofrecer un modelo que ayude a predecir la probabilidad de éxito de proyectos de crowdfunding.

Proceso: Se hace un repaso de la literatura en las teorías mencionadas (capital social, capital humano y LOP).

- Capital social: Los individuos dentro de una red hacen inversiones en mutuo acuerdo y reconocidas a fin de preservar la dominancia del grupo y la solidaridad entre sus miembros. Se puede observar desde tres dimensiones: estructural, relacional y cognitiva, que pueden ser sintetizadas como el nivel de inter-conectabilidad entre los miembros, la fuerza y calidad de las relaciones, y las interpretaciones subjetivas de pensamientos compartidos respectivamente.
- Capital humano: El capital humano tiene componentes genéricos y específicos. El componente genérico se refiere a conocimiento general poseído por los emprendedores acumulados durante su educación y experiencia profesional, mientras que el componente específico consiste en las habilidades y capacidades que los fundadores pueden aplicar

directamente a su emprendimiento. Estudios demuestran que el capital humano es un activo invaluable y puede compensar la falta de experiencia en el negocio y falta de recursos. Además, si la firma tiene múltiples dueños tendrá más probabilidades de sobrevivir comparado con una firma de un solo dueño.

- Nivel de procesamiento: Depende de la cantidad de trabajo cognitivo realizado, las personas procesan la información en distintos niveles de codificación mental. Según la teoría, el nivel de intensidad del procesamiento de información está influenciada por el tipo de infracción o el tipo de carga.

  En base a la literatura revisada se adaptan 15 factores relacionados a la probabilidad de éxito de proyectos de crowdfunding, como se muestra en la siguiente tabla:

Figura 67

Listado de Factores o variables y los papers en los que han sido referenciados según la literatura.

Factors	Supporting references
Tagline number	Lagazio and Querci (2018), Marelli and Ordanini (2016), Zhou *et al.* (2018)
The number of videos	Reyes and Bahm (2016), Kunz *et al.* (2017), Bi *et al.* (2017), Zhou *et al.* (2018)
The number of photos	Beier and Wagner (2014); Younkin and Kuppuswamy (2017)
The length of videos	Dahlhausen *et al.* (2016), Guo *et al.* (2014)
The number of updates	Zheng *et al.* (2014), Kraus *et al.* (2016), Gleasure and Morgan (2018)
The number of comments	Bannerman (2013), Zheng *et al.* (2014), Kromidha and Robson (2016)
The number of backers	Kromidha and Robson (2016), Gleasure and Morgan (2018), Giudici *et al.* (2018)
Facebook fans	Lin and Lu (2011), Mollick (2014), Zheng *et al.* (2014)
Past contributions (the number of projects in which an entrepreneur make contribution for someone else)	Kuti and Madarász (2014), Colombo *et al.* (2015)
Goal	Greenberg *et al.* (2013), Cordova *et al.* (2015), Calic and Mosakowski (2016), Kuppuswamy and Bayus (2017)
The duration of the project	Mollick (2014), Cordova *et al.* (2015), Lukkarinen *et al.* (2016)
The number of owners	De Buysere *et al.* (2012), Cordova *et al.* (2015), Zvilichovsky *et al.* (2015)
Past campaigns (the number of others' projects in which an entrepreneur raised before)	Dresner (2014), Beaulieu *et al.* (2015), Hong *et al.* (2016)
Past raised amount (the amount of others' projects in which an entrepreneur raised before)	Dresner (2014), Beaulieu *et al.* (2015), Hong *et al.* (2016)
Pledged on another platform	Lagazio and Querci (2018), Marelli and Ordanini (2016), Zhou *et al.* (2018)

De "A machine learning approach to predict the success of crowdfunding fintech project", por Yeh & Chen, 2020.

Respecto al método, se desarrolló un modelo predictivo basado en algoritmos de machine learning para predecir el éxito de un proyecto de crowdfunding. Se removieron variables irrelevantes para mejorar la capacidad de generalización, al inicio se contaron con todas las variables, pero una por una se fue eliminando tomando como criterio la que tenga la menor significancia hasta que no se notara ninguna mejora con las eliminaciones. Los datos fueron recolectados de Indiegogo, teniendo 4474 muestras en total entre 2015 y 2017, además cabe mencionar que se obtuvieron de dos fuentes diferentes, una de ellas fue directamente de las bases de datos de Indiegogo, y la otra fue del proyecto Facebook. Adicionalmente, se aplica muestreo sin reemplazamiento para crear los conjuntos de entrenamiento y pruebas.

Se propone el uso de redes neuronales ensemble (ENN) que genera múltiples redes neuronales artificiales (ANN) para hacer la predicción final, además las estructuras y parámetros de estas redes son diferentes. El experimento consta de dos partes: Entrenamiento y pruebas. En la fase de entrenamiento se realizan los siguientes pasos:

- Configuración de parámetros: Se establece el número de ANNs, el máximo número de capas ocultas, el máximo número de neuronas en cada capa oculta, la tasa de datos de entrenamiento, el umbral de exactitud, entre otros. En la tabla a continuación se muestran todos los parámetros y su notación:

Figura 68

Listado de parámetros y sus descripciones.

Notation	Description
m	The number of ANN models
$h_{\max}$	The maximum number of hidden layers
$c_{\max}$	The maximum number of hidden nodes
r	The percentage of training data for testing neural networks in training phase
$w_{threshold}$	The threshold level of the accuracy
h_m	The hidden layers of the m-th ANN
c_m	The hidden nodes of the m-th ANN
d_{TR}	Total training data in training phase
d_{TRTR}, m	Training dataset of training data in training phase
d_{TETR}, m	Testing dataset of training data in training phase
d_{TE}	Testing data in testing phase
$\omega_{m,j,g}$	The weight associated with the link between neuron j and neuron g of the m-th ANN
$\theta_{m,g}$	The constant of neuron g of the m-th ANN
$o_{m,g}$	The output of neuron g of the m-th ANN
f(net)	Activation function
$net_{m,g}$	The g-th result of the network inputs $x_{m,i}$ of the m-th ANN
$x_{m,i}$	The i-th input vector of the m-th ANN
$e_{m,g}$	The computed output error at neuron g of the m-th ANN
$t_{m,g}$	The g-th target output of the m-th ANN
η	The parameter of learning rate with $0 < \eta < 1$
$T_{threshold}$	The threshold level for the convergence of ANN
$a_{m,l}$	The accuracy of the l-th calculated input testing data from the m-th ANN
a_m	The average accuracy of all $a_{m,l}$ in the m-th ANN
c_k	The output of calculated input of testing data in testing phase from the k-th ANN

De “A machine learning approach to predict the success of crowdfunding fintech project”, por Yeh & Chen, 2020.

- Obtener datos de entrenamiento de la base de datos: Los datos recolectados se almacenaron en una base de datos y se dispone de ellos.
- Crear múltiples ANNs de forma aleatoria: Se usan los valores máximos configurados previamente para generar distintas combinaciones de ANNs aleatoriamente.
- Entrenar los modelos ANN y ajustar los pesos en cada uno en base a combinaciones aleatorias.
- Revisar sesgo en las variables y actualizar pesos.
- Repetir los dos pasos previos hasta que las ANN hagan convergencia, cuando la diferencia entre la salida del modelo actual y el anterior sea menor que el umbral definido.

- Usar el conjunto de datos de prueba para validar el modelo.
- Filtrar y quedarse solo con las ANNs con mayor exactitud.

En la fase de pruebas se realizan los siguientes pasos:

- Obtener los datos de prueba de la base de datos.
- Predecir la probabilidad de éxito en base a las variables de entrada.

Para evaluar el desempeño del modelo se usa la métrica exactitud que se calcula dividiendo el número de clasificaciones correctas por el número total de observaciones.

Resultados: Se incluyeron 15 variables independientes al modelo. Se identificó que el número promedio de respaldadores es 483, la duración promedio 34.49 días, la cantidad de actualizaciones promedio 7.42, los fans de Facebook promedio 1453, entre otras. Se ejecutaron tres pruebas diferentes. En la primera se fuerza el uso de las 15 variables definidas, se comparó el modelo propuesto con regresión lineal, regresión logística y ANN y en todos los casos se obtuvo una exactitud superior, de 95.75% comparado con 91.03%, 92.07% y 94.79% respectivamente. La segunda prueba usa los resultados del proceso de selección de variables, donde se descartaron las que tenían un coeficiente de correlación de Pearson con la variable target no estadísticamente significante ($p < 0.05$), obteniendo como resultado que nuevamente el modelo ENN propuesto tuvo mayor exactitud que los modelos de regresión lineal, regresión logística y ANN con 96.23% en comparación de 91.19%, 91.91% y 94.87% respectivamente, y también desempeñó mejor que el modelo ENN de la primera prueba. En la tercera y última prueba se emplea el enfoque de eliminación por iteraciones y se obtiene como resultado una exactitud menor que las dos primeras pruebas, siendo de 89.18% para el ENN, esto se debe a que al usar menos variables el modelo tiene menos información para generalizar óptimamente, en tal sentido se concluye que la mejor versión del modelo es usando un proceso de selección de variables por correlación de Pearson, como se evidencia en la segunda prueba.

Artículo N° 15

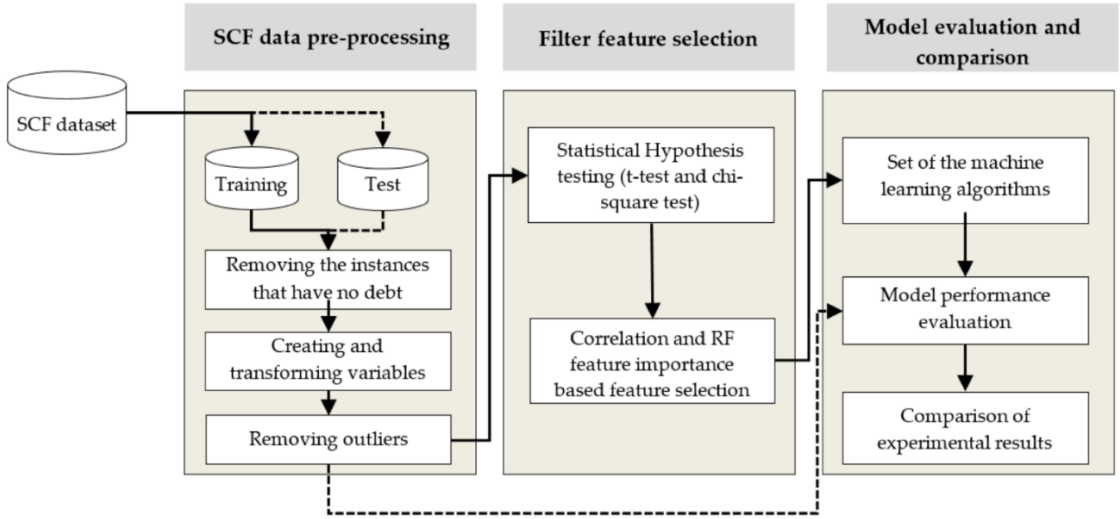
Título: An empirical comparison of machine-learning methods on bank client credit assessments (Una comparación empírica de métodos de machine learning en evaluaciones de crédito de clientes bancarios)

Aporte: Los autores proponen un enfoque de machine learning para abordar la evaluación de riesgo de los clientes en instituciones financieras, en donde datos reales de clientes comparan dos métodos de selección de variables, por un lado prueba de hipótesis, correlación y medición de importancia de variables basado en random forest, y por otro lado

un nuevo enfoque de random forest (NAP), y también se comparan distintos algoritmos de machine learning como regresión logística, máquinas de vectores de soporte (SVM), redes neuronales profundas (DNN) y ensemble gradient boosted trees como XGBoost. Además, se identificó que hay pocos estudios que aborden el problema de gestión de riesgo, ya que la mayoría se enfoca en clasificación binaria de asignación de crédito o abandono, entre otras.
Proceso: Se propone un framework para el sistema propuesto el cual consiste en tres fases como se ve en la siguiente figura.

Figura 69

Fases del modelo propuesto por la investigación.



De “An empirical comparison of machine-learning methods on bank client credit assessments”, por Munkhdalai et al., 2019.

El conjunto de datos SCF proviene de la encuesta de finanzas de consumidores y consta de 345 variables, se divide en conjunto de entrenamiento y de pruebas, cada uno con una muestra de 4113 y 4245 observaciones respectivamente. La variable dependiente es la amortización de la deuda morosa (LATE), de tal forma que, si el cliente no tiene deuda de pagos morosos el valor será “no”, 0, caso contrario “si”, 1.
Los autores proponen una estrategia para comparar las evaluaciones de crédito FICO (modelo de evaluación basado en experiencia humana) en la que se toma el porcentaje de la población en cada categoría de crédito de FICO de forma ordenada y se establecen cortes

para el score obtenido con el modelo de machine learning, estos cortes se pueden apreciar en la siguiente tabla.

Figura 70

Cortes establecidos para el score obtenido con el modelo de machine learning.

Credit Rating	FICO Score	The Percent of Population (%)	The Probability of Default (%)	Interest Rate
C1	800 or more	13	1	5.99
C2	750–799	27	1	5.99
C3	700–749	18	4.4	6.21
C4	650–699	15	8.9	6.49
C5	600–649	12	15.8	7.30
C6	550–599	8	22.5	8.94
C7	500–549	5	28.4	9.56
C8	Less than 499	2	41	-

De “An empirical comparison of machine-learning methods on bank client credit assessments”, por Munkhdalai et al., 2019.

Respecto a los algoritmos de selección de variables se utiliza TSFFS (siglas de two-stage filter feature selection) y una adaptación del método NAP. La implementación de TSFFS consta de dos pasos, en el primero para eliminar variables irrelevantes se evalúa la significancia de cada una con dos pruebas de hipótesis: t-test para variables continuas y test chi-cuadrado para variables categóricas. Ambos tests evalúan si las variables independientes proveen información estadísticamente significante sobre la solvencia de los clientes. En el segundo paso se eliminan variables poco importantes basado en la medida de importancia y correlación de variables en random forest, esta aplica tanto para variables categóricas como continuas. Por otro lado, la implementación de NAP usa un framework de pruebas de permutación para evaluar la hipótesis nula de independencia entre la variable dependiente y vectores multidimensionales de las variables independientes a fin de identificar cuáles son las más o menos relevantes.

Los enfoques de machine learning elegidos son los algoritmos LR, MARS, SVM, RF, XGBoost y ANN, y se usan para comparar el score con el sistema FICO. Para prevenir sobreajuste en algunos modelos, se utilizó la técnica grid search CV de 10 pliegues para encontrar los hiperparámetros óptimos, el espacio de búsqueda se muestra en la siguiente tabla.

Figura 71

Hiperparámetros óptimos para cada método.

Method	Parameters	Symbol	Search Space
Support Vector Machine	Gamma	γ	0.001, 0.01, 0.1
	Cost	C	10, 100, 1000
	Epsilon	ε	0.05, 0.15, 0.3, 0.5
Random Forest	Number of features randomly sampled	$mtry$	3, 6, 9, 12, 15, 18, 21
	Minimum size of terminal nodes	$nodesize$	50, 80, 110
	Number of tree	$ntree$	500, 1500, 2500
XGBoost	Maximum tree depth	D_{max}	2, 4, 6, 8
	Minimum child weight	w_{mc}	1, 2, 3, 4
	Early stop round		100
	Maximum epoch number	$epoch$	500
	Learning rate	τ	0.1
	Number of boost	N	60
	Maximum delta step	δ	0.4,0.6,0.8,1
	Subsample ratio	r_s	0.9,0.95,1
	Column subsample ratio	r_c	0.9,0.95,1
	Gamma	γ	0, 0.001

De “An empirical comparison of machine-learning methods on bank client credit assessments”, por Munkhdalai et al., 2019.

Respecto a las redes neuronales se crearon 6 con arquitecturas distintas, las tres primeras usan la función de activación sigmoide con una, tres y cinco capas ocultas con ocho nodos respectivamente, los tres restantes usan la función de activación ReLU en cada capa oculta y la función softmax en la capa output, de igual forma que las tres primeras, con una, tres y cinco capas ocultas con ocho nodos respectivamente.

Resultado: De las 345 variables solo se mantuvieron 222 al seleccionarlas a través del algoritmo TSFFS, para las variables continuas se usó t-test y test chi-cuadrado y se identificó que, por ejemplo, la variable de balance total de préstamo para mejoras en el hogar no tiene diferencia estadísticamente significativa con respecto a que tan buen o mal pagador es un cliente, al tener un p-value de 0.127, por el lado de variables categóricas se usó el test chi-cuadrado de independencia para comprar frecuencias entre buenos y malos pagadores, por ejemplo la variable de fuentes de información para investigar decisiones no tiene diferencia estadísticamente significativa al tener un p-value de 0.067. Como segundo paso del algoritmo TSFFS se usó la correlación e importancia de variables en random forest, y como resultado se obtienen solo 116 variables. Por otro lado, se aplicó el algoritmo NAP y se obtuvieron 76 variables con alto poder predictivo y que no están correlacionadas entre sí.

Finalmente se calcularon distintas métricas como AUC, true positive rate, false positive rate, accuracy, para los distintos algoritmos planteados usando las dos técnicas de selección de

variables descritas, teniendo como resultado que, para la selección de variables con TSFFS, MLP con sigmoide tuvo el mejor AUC, con 86.81%, lo cual es 0.09% más que MLP con softmax, respecto al accuracy y FPR el mejor resultado lo tuvo XGBoost con 84.87% y 13.61% respectivamente y el mejor TPR lo tuvo el random forest. Respecto a la selección de variables mediante NAP, el mejor AUC lo tuvo Deeper MLP con sigmoide con 87.48%, el mejor H-measure MLP con softmax con 43.11%, el mejor TPR SVM con 87.88% y el mejor FPR y accuracy XGBoost con 11.63% y 86.50% respectivamente. En líneas generales se identificó que las redes neuronales MLP con activación sigmoide y XGBoost tuvieron mejores resultados y usando la técnica de selección de variables NAP pueden lograr aún mejores métricas.

Figura 72

Comparación de resultados por cada modelo machine learning de la investigación.

FS	Machine Learning Models	AUC	H-Measure	TPR	FPR	Accuracy
TSFFS	Logistic	0.8507	0.3880	0.7668	0.2031	0.7950
	MARS	0.8283	0.3591	0.7005	0.1834	0.8094
	SVM	0.7841	0.2429	0.4793	0.1396	0.8368
	RF	0.8544	0.4039	**0.8534**	0.2709	0.7368
	XGBoost	0.8587	0.3897	0.6192	**0.1361**	**0.8487**
	MLP with sigmoid	**0.8681**	**0.4336**	0.8259	0.2108	0.7914
	Deep MLP with sigmoid	0.8657	0.4232	0.8135	0.2103	0.7911
	Deeper MLP with sigmoid	0.8581	0.3952	0.7528	0.1915	0.8051
	MLP with softmax	0.8672	0.4243	0.8389	0.2324	0.7720
	Deeper MLP with softmax	0.8637	0.4155	0.7917	0.2115	0.7887
	Deep MLP with softmax	0.8631	0.4128	0.8135	0.2175	0.7844
NAP	Logistic	0.8667	0.4151	0.7762	0.2090	0.7901
	MARS	0.8462	0.3868	0.7166	0.1815	0.8122
	SVM	0.8083	0.3097	**0.8788**	0.4394	0.5803
	RF	0.8682	0.4214	0.8497	0.2765	0.7313
	XGBoost	0.8633	0.3987	0.5824	**0.1163**	**0.8650**
	MLP with sigmoid	0.8726	0.4256	0.8171	0.2337	0.7695
	Deep MLP with sigmoid	0.8718	0.4233	0.8228	0.2303	0.7730
	Deeper MLP with sigmoid	**0.8748**	0.4298	0.8306	0.2318	0.7720
	MLP with softmax	0.8742	**0.4311**	0.8358	0.2417	0.7631
	Deeper MLP with softmax	0.8664	0.4126	0.8140	0.2285	0.7742
	Deep MLP with softmax	0.8682	0.4172	0.8161	0.2303	0.7725

De “An empirical comparison of machine-learning methods on bank client credit assessments”, por Munkhdalai et al., 2019.

Pregunta 4: ¿Qué tipo de estudios se han realizado sobre modelos predictivos para el abandono de clientes en el sector financiero?

Artículo N° 1

Título: Behavioral attributes and financial churn prediction (Atributos de comportamiento y predicción de abandono financiero)

Aporte: Propuesta de variables y patrones espaciotemporales y de comportamiento dinámico en base a transacciones, que dan como resultado una mejora significativa en el desempeño de los modelos de predicción de abandono en el sector financiero y que además se puede aplicar en otros sectores si se cuenta con dichos datos. Este tema es de gran relevancia ya que los autores plantean que retener clientes es mucho menos costoso que adquirir nuevos, hasta 16 veces menor, además que reducir la tasa de abandono en solo 5% puede aumentar la rentabilidad entre un 25 y 125%.

Proceso:

Datos: Se obtuvieron dos sets de datos de muestras obtenidas entre julio de 2014 y julio de 2015 que fueron donados por una institución financiera importante de un país de la OECD. Las muestras contienen información demográfica, transacciones de tarjetas de crédito, transferencias de dinero y transferencias de fondos electrónicos. Las dos muestras contienen 100 mil y 60 mil observaciones de clientes respectivamente, y 45 millones y 22 millones de transacciones respectivamente. Los datos personales de los clientes (como nombres, datos de contacto, datos sensibles) fueron anonimizados.

Variables: Se caracterizan las variables de comportamiento y se introducen nuevas variables llamadas entropía de decisión las cuales explican los patrones variables de gasto y transferencia de los clientes.

- Variables demográficas: Se emplearon género, estado civil, estado educativo, tipo de empleo, ingresos y edad. Todos los datos estuvieron completos a excepción de ingresos, en la cual había un 2% de datos faltantes y fueron llenados con el promedio.
- Patrones espaciotemporales y de elección: Se emplearon variables implícitas de patrones de gasto espaciotemporal y patrones de elección financiera. Para los patrones de gasto espaciotemporal se armaron tres variables, diversidad, lealtad y regularidad, las cuales miden que tan diversos o leales son los clientes en sus patrones de gasto desde la perspectiva de tiempo y ubicación, mientras que los patrones de elección financiera indican cómo los clientes distribuyen sus actividades financieras respecto a comercios, categorías de gasto y las direcciones de transferencias de fondos.

Etiquetado: En la literatura existen diversas definiciones de abandono de clientes, desde si dejan de usar una tarjeta de crédito hasta si realizaron cierta acción en un periodo determinado de tiempo. Entonces la definición de abandono es subjetiva y se debe definir acorde a las heurísticas establecidas por la industria. En el presente estudio se establecieron

tres definiciones para evaluación en base a información de segmentación, patrones de gasto en tarjetas de crédito, y operaciones en cuenta.

Propiedades del experimento: Para evaluar el desempeño de las variables, se hizo un análisis con cuatro conjuntos de datos diferentes al separar cada base en dos, teniendo ventanas de tiempo de etiquetado diferentes. Específicamente los conjuntos de datos B1 y B2 están más sesgados hacia un mayor uso de tarjetas de crédito, mientras que los conjuntos de datos A2 y B2 están orientadas para una predicción a más corto plazo. La misma metodología se aplica de forma independiente a cada uno de estos conjuntos de datos. Antes de la generación de modelos predictivos, a los valores faltantes de cada característica se les asignan los valores medios y todas las características numéricas se estandarizan eliminando la media y escalando los valores a la varianza unitaria. Se ha aplicado codificación dummy sobre las variables categóricas para que cada una de ellas se represente con tantas variables binarias como el número de sus niveles menos uno, esto para evitar redundancia y correlación entre dos variables generadas.

Se emplearon Random Forests como la técnica de entrenamiento de clasificación. La configuración de híper-parámetros consistió de 500 árboles y un máximo de dos variables por árbol. Se empleó validación cruzada estratificada de 8 pliegues para que, en cada iteración, casi la misma proporción de abandonos y no abandonos estuvieran involucradas en el proceso de evaluación.

Para mitigar el riesgo de desbalanceo entre el número de abandonos y no abandonos en los conjuntos de datos, se aplicó SVM-SMOTE con una proporción de 0,25, lo que significa que la clase minoritaria se sobre muestrea hasta que su proporción alcanza una cuarta parte de la de la clase mayoritaria. Todas las implementaciones de pre procesamiento y clasificación se realizaron en lenguaje Python, principalmente con el paquete Scikit-learn, y para SVM-SMOTE, se empleó el paquete de código abierto Imbalanced-learn de Lemaître et al. (2016).

Principal resultado: Se demuestra que los patrones espaciotemporales y la entropía de elección están significativamente relacionados con la decisión de abandono de los clientes. Debido al alto nivel de desbalance entre clases a predecir, se utilizó la métrica área bajo la curva ROC, sin embargo, se debe tener en cuenta que esta no tiene en cuenta el costo malas predicciones para cada clase en particular, por lo tanto, es más general y asume que ambas tienen el mismo peso. Se sugiere abordar el problema con un enfoque de evaluación basado en ganancias, por ejemplo, ganancia máxima, para seleccionar el modelo más rentable respecto al abandono de clientes. Adicionalmente se emplea una técnica de evaluación del

modelo bajo distintos grupos en base a variables que pueden tener sesgo implantado, como sexo y rango etario, por el lado del sexo de la persona se determinó que no hay superioridad significante en términos de predictibilidad, sin embargo, con respecto al rango etario si se identificó que las personas con más de 50 años eran más difíciles de predecir.

Artículo N° 16

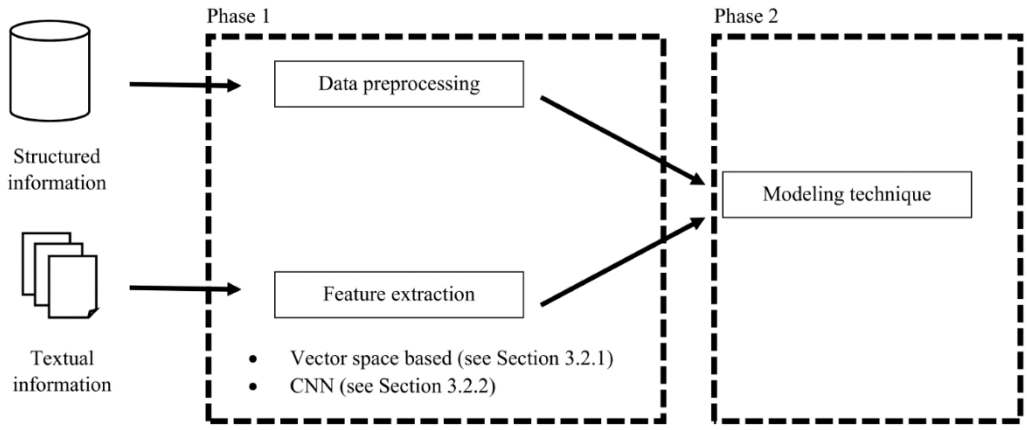
Título: Incorporating textual information in customer churn prediction models based on a convolutional neural network (Incorporación de información textual en los modelos de predicción de abandono de clientes basados en una red neuronal convolucional)

Aporte: Los autores proponen el uso de datos textuales no estructurados procesados a través de redes neuronales convolucionales (CCNs por sus siglas en inglés Convolutional neural networks) en modelos de predicción de abandono (CCP por sus siglas en inglés Customer churn prediction) a fin de mejorar su desempeño predictivo. Además, se estudia por primera vez el uso de CNNs en minería de texto aplicado a CCP, donde se demuestra que estas brindan mejores resultados comparado con las mejores prácticas actuales. De esta forma aportan un precedente que puede ser aprovechado por las empresas, instituciones y profesionales para mejorar sus modelos CCP y finalmente tener mejores campañas de retención.

Proceso: La metodología empleada abarca (1) el uso de un framework para integrar los datos textuales en modelos CCP, el cual establece que los datos estructurados se deben preprocesar completamente separados de los datos textuales debido a la amplia diferencia en complejidad de tratamiento de ambos tipos de datos, posteriormente se generarán variables a partir de los datos no estructurados y se unirán al conjunto de datos de entrada del modelo que puede existir desde antes o ser nuevo.

Figura 73

Metodología propuesta por los autores.

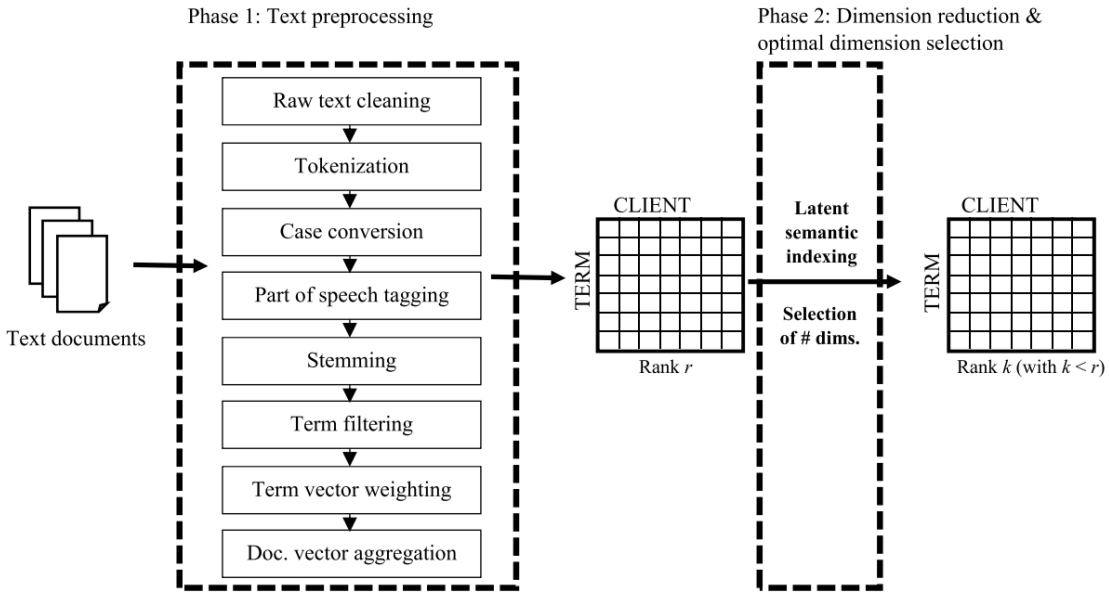


De “Incorporating textual information in customer churn prediction models based on a convolutional neural network”, por De Caigny et al., 2020.

(2) También abarca la técnica de procesamiento de datos textuales para lo cual se proponen dos enfoques: basado en el espacio vectorial, que aborda el pre procesamiento y la reducción de dimensionalidad, y CNN, donde los datos son transformados al pasar por las diferentes capas de la red. Por el lado del enfoque basado en el espacio vectorial, el pre procesamiento incluye un cleansing básico (como remover caracteres especiales y tokenización), etiquetado de sintáctica (etiquetado de palabras en su categoría sintáctica (por ejemplo: sustantivos, verbos), stemming (reemplazar palabras por su palabra raíz, por ejemplo: se reemplazan discussion, discussed y discussing por discuss),y finalmente un filtrado donde se quitan palabras irrelevantes o no informativas, de tal forma que solo se mantienen los sustantivos, verbos, adverbios y adjetivos, y se quitan palabras con frecuencia muy pequeña, como palabras que aparecen solo una o dos veces. Basado en el corpus del texto ya limpio, se genera un vector de término ponderado para cada mensaje y se calcula como el producto de la frecuencia de los términos y la frecuencia inversa del documento. Finalmente, los vectores son agregados por cada cliente, ya que un cliente puede tener varios mensajes en un periodo de tiempo. Seguido de ello, es necesario un proceso de reducción ya que la matriz de término-por-cliente generada anteriormente será muy grande y no será adecuada para usar directamente en un modelo predictivo debido a alto nivel de ruido y dispersión en las variables, para ello se la técnica de reducción de dimensionalidad.

Figura 74

Enfoque basado en CNN.



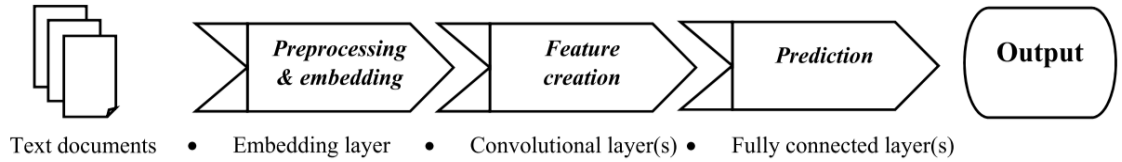
De “Incorporating textual information in customer churn prediction models based on a convolutional neural network”, por De Caigny et al., 2020.

Por el lado del enfoque basado en CNN se divide en tres partes principales. Primero el texto debe ser transformado en un formato numérico para lo cual se aplica codificación one-hot, luego, a diferencia del enfoque basado en vectores espaciales, no se requiere mayor preprocesamiento ya que el segundo paso consiste en embeber los términos como vectores en un espacio de dimensiones reducidas. Esto trae dos ventajas principales para el uso de redes neuronales artificiales (ANNs por sus siglas en inglés Artificial neural networks) profundas. Por un lado, las ANNs usualmente tienen dificultades computacionales con vectores dispersos de alta dimensionalidad, y esto se resuelve normalmente usando términos embebidos. Por otro lado, este proceso mejora el poder de generalización del modelo. En segundo lugar, se crean nuevas variables en base a los vectores de términos embebidos en la capa convolucional del CNN donde se utilizan stages convolucionales para mejorar la eficiencia computacional y stages detectores y de pooling que ayudan a orquestar el proceso, estos distintos stages se pueden unir de distintas maneras y crear modelos más complejos.

En tercer lugar, la predicción requiere que las variables creadas en la capa convolucional sean combinadas en un clasificador y para ello se utilizan una o más capas completamente conectadas.

Figura 75

Proceso aplicativo en la investigación.



De “Incorporating textual information in customer churn prediction models based on a convolutional neural network”, por De Caigny et al., 2020.

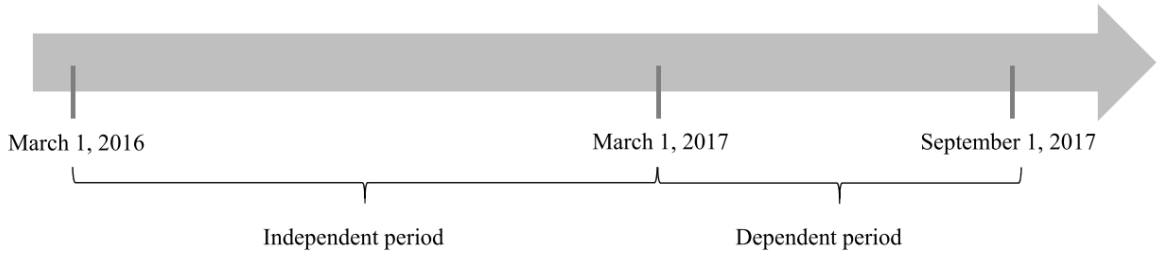
(3) Finalmente se proponen cuatro preguntas de investigación que ayudarán a validar el modelo planteado. Estas son: ¿La adición de data textual, en forma de comunicaciones electrónicas escritas entre el cliente y el asesor, mejora el desempeño predictivo de un modelo CCP?, ¿cuál es el mejor enfoque de extracción de variables para integrar información textual no estructurada en un modelo de CCP?, ¿hay alguna diferencia en patrones de abandono entre clientes con información textual y clientes sin información textual? y ¿un modelo entrenado únicamente con datos textuales extrae suficiente información para lograr un desempeño predictivo competitivo?

Se realizó un experimento del estudio el cual se explicará en cuatro partes: presentación de datos, preprocesamiento de datos estructurados y no estructurados, modelamiento y métricas de evaluación.

Respecto a los datos, estos fueron obtenidos de un gran proveedor de servicios financieros europeo, se separan en dos periodos de tiempo, independiente y dependiente, debido a que las predicciones se hacen con 1 año de información histórica y calcula la probabilidad de abandono en los próximos seis meses.

Figura 76

Definición de periodo independiente y periodo dependiente para la predicción.



De "Incorporating textual information in customer churn prediction models based on a convolutional neural network", por De Caigny et al., 2020.

El conjunto de datos comprende 607125 clientes, de los cuales el 2% abandonaron en el periodo dependiente (se considera abandono cuando el cliente cierra todas sus cuentas con la institución). Por el lado de los datos no estructurados, 66642 clientes enviaron mensajes a su asesor durante el periodo independiente, de los cuales el 1.77% abandonó en el periodo dependiente. Se utilizaron 37 variables estructuradas provistas por la propia compañía. Se dividió el conjunto de datos en sets de entrenamiento, selección y validación, cada uno conteniendo un tercio de los datos.

Respecto al preprocesamiento, por el lado de los datos estructurados se aplicó imputación de cero, mediana o moda a variables con más del 5% de valores faltantes, dependiendo de la variable, y se crearon variables dummy. Normalmente clientes con valor faltante en una variable suelen tener valores faltantes en varias variables, así que, para minimizar el impacto de los procedimientos de imputación, directamente se eliminarán clientes con valores faltantes en variables con menos de 5% de valores faltantes. Es importante mencionar que para las variables categóricas transformadas a dummy la cantidad de variables creadas debe ser el número de categorías únicas de la variable original menos uno ya que de lo contrario se estaría redundando información para el modelo. Los valores atípicos, considerados como valores alejados a más de tres desviaciones estándar del promedio de la variable, son tratados usando winsorización. Finalmente, se realiza un muestreo de datos de tipo undersampling, en el cual se reduce la cantidad de observaciones de clientes que no abandonaron a la misma cantidad de observaciones de clientes que si abandonaron a través de muestreo aleatorio. Esta técnica es usada ampliamente en modelos CPP.

Por el lado de datos textuales, se deben procesar 189665 mensajes textuales bajo el enfoque basado en espacio vectorial, para ello se identifica que el corpus completo consiste en 79914 términos únicos, los cuales son reducidos a 60737 después de aplicar stemming. Posteriormente, el proceso de filtrado deja menos de 1000 términos únicos trabajables. Luego se agregan los documentos por cada cliente y se reducen un poco más a través de indexación semántica latente. Con ello se obtienen ocho nuevas variables basadas en texto "z".

Respecto al modelamiento, se muestra un resumen de los tipos de modelos probados y su configuración.

Figura 77

Variaciones del modelo.

Model name	Data		Type of information		Feature extraction approach of unstructured data		Modeling technique
	Full dataset	Only customers involved in email communications	Structured	Text	Vector space	Deep learning	
Mod_FD	X		X		n.a.	n.a.	Logit
Mod_S		X	X		n.a.	n.a.	Logit
Mod_SU_VS		X	X	X	X		Logit
Mod_SU_DL_P		X	X	X		X[a]	Logit
Mod_SU_DL_L		X	X	X		X[b]	Logit
Mod_U_VS		X		X	X		Logit
Mod_U_DL_P		X		X		X	n.a.

[a]Deep learning method using probabilities.
[b]Deep learning method using the output of the last hidden layer.

De "Incorporating textual information in customer churn prediction models based on a convolutional neural network", por De Caigny et al., 2020.

Se construye un CNN no estático, la capa de embebido es seguida de capas convolucionales de una dimensión y capas de pooling máximo para cada tamaño de filtro. Los resultados de esas capas son concatenados y usados como entrada para la capa completamente conectada. Esta última capa usa una función de activación "softmax" para clasificación que genera una probabilidad de abandono para cada cliente, en el modelo Mod_SU_DL_P esas probabilidades sirven como una nueva variable para el modelo CCP, indicando la probabilidad de abandono basándose netamente en las comunicaciones escritas. El modelo Mod_SU_DL_L usa directamente las variables creadas por las campas convolucionales en el modelo CCP, estas variables normalmente servirían como una entrada para la capa completamente conectada y resume la información textual con respecto al abandono. La

configuración de parámetros del CNN se resume en la siguiente tabla y fueron tomados en base a estudios previos que usaron datos textuales similares para problemas de clasificación.

Figura 78

Parámetros del modelo.

Parameter	Value	Reference
Word embeddings	200D, French Wikipedia	Fauconnier (2015)
Optimizer	Adam	Kingma and Ba (2015) and Kvamme et al. (2018)
Filter windows	3, 4, 5	Kim (2014)
Number of filters	10	Kim (2014)
Dropout rate	0.5	Kim (2014)
Batch size	50	Kim (2014)
Hidden dimensions	100	Kim (2014)
L2 constraint	3	Kim (2014)

De “Incorporating textual information in customer churn prediction models based on a convolutional neural network”, por De Caigny et al., 2020.

Se construye el modelo CCP usando regresión logística, técnica que demuestra buen funcionamiento para modelos CCP ya que las probabilidades posteriores son estimadas directamente en una regresión logística de forma que se tiene un output comprensible y porque los resultados de modelos logísticos son robustos y tienen buen desempeño predictivo acorde a la comparación de diversos experimentos en CCP. Las variables son elegidas usando “forward selection”, la primera variable que entra al modelo es la que tenga la estadística x cuadrado más grande, el resto de las variables son consideradas para el modelo final hasta que se cumpla cierta regla que detenga las iteraciones. Para determinar el número de variables se emplea un enfoque de 3 x 5 CV. En la tabla a continuación se muestra el número de variables promedio por cada modelo.

Figura 79

Número de variables promedio incluidas en el modelo final.

	Mod_FD	*Mod_S*	*Mod_SU_VS*	*Mod_SU_DL_P*	*Mod_SU_DL_L*
Avg. # variables[a]	17.0	16.8	19.0	17.5	22.6
Avg. # textual variables	/	/	2.6	1.0	6.9

[a]Textual variables included.

De "Incorporating textual information in customer churn prediction models based on a convolutional neural network", por De Caigny et al., 2020.

Respecto al proceso final de evaluación, el desempeño predictivo de los modelos se mide con la métrica AUC (en español área bajo la curva de características operativas del receptor) y TDL (en español elevación del decil superior) ya que son usadas con frecuencia en modelos CCP.

Resultados: Las métricas obtenidas en los distintos modelos propuestos se muestran en la siguiente tabla:

Figura 80

Resultados en términos de AUC y TDL para las diferencias variaciones de modelos.

Results in terms of the AUC and TDL for different model variations.

	AUC	TDL
Mod_FD	86.938% (0.008)	5.867 (0.225)
Mod_S	87.810% (0.008)	6.188 (0.251)
Mod_SU_VS	87.885% (0.007)	6.211 (0.224)
Mod_SU_DL_P	89.666% (0.010)	6.789 (0.282)
Mod_SU_DL_L	**89.875% (0.010)**	**6.868 (0.335)**
Mod_U_VS	54.102% (0.018)	1.269 (0.206)
Mod_U_DL_P	73.056% (0.036)	3.642 (0.624)

De "Incorporating textual information in customer churn prediction models based on a convolutional neural network", por De Caigny et al., 2020.

Dando respuesta a las cuatro preguntas de investigación planteadas: (1) se realizó una comparación por pares y se determinó que la integración de datos textuales en los modelos

CCP pueden mejorar significativamente el desempeño predictivo a nivel de métricas AUC y TDL, sin embargo, solo cuando estos datos no estructurados fueron procesados a través de una CNN profunda, teniendo un F-value de 17.741 y p-value de 0.006 en la métrica AUC y 6.940 y 0.057 respectivamente en la métrica TDL para el modelo Mod_SU_DL_P, y 18.904 y 0.006 en la métrica AUC y 7.295 y 0.057 en la métrica TDL para el modelo Mod_SU_DL_L, al compararlos con el modelo Mod_S, (2) el enfoque de deep learning para procesar los datos no estructurados tienen mejor desempeño que el enfoque basado en espacio vectorial, teniendo un F-value de 13.388 y p-value de 0.010 para la métrica AUC y 6.702 y 0.046 respectivamente para la métrica TDL para el modelo Mod_SU_DL_L al compararlo con el modelo MOD_SU_VS, (3) el patrón de deserción de clientes con y sin comunicaciones por email es diferente, teniendo un F-value de 4.466 y p-value de 0.054 para la métrica AUC y 6.923 y 0.021 respectivamente para la métrica TDL para el modelo Mod_S al compararlo con el modelo MOD_FD, (4) los modelos que solamente usan datos textuales no tienen comparación con modelos que usan datos estructurados, se comparó el modelo Mod_S con Mod_U_VS y Mod_U_DL_P y se obtuvo un F-value de 320.648 y p-value de 0.000 para la métrica AUC y 133.999 y 0.000 respectivamente para la métrica TDL en el primer modelo y un F-value de 19.172 y p-value de 0.002 para la métrica AUC y 14.354 y 0.004 respectivamente para la métrica TDL en el segundo modelo.

Respecto a la importancia de las variables, se halló que la integración de datos textuales representa una importancia relativa de entre 5% a 25% en los modelos CCP.

Usando una fórmula de beneficio se determinó que la ganancia de una campaña de retención que contiene datos textuales aumenta en 192.914 euros, tomando en cuenta el CLV, el costo del incentivo, el costo administrativo, entre otros.

Artículo N° 18

Título: Propension to customer churn in a financial institution: a machine learning approach (Propensión al abandono de clientes en una institución financiera: un enfoque de aprendizaje automático)

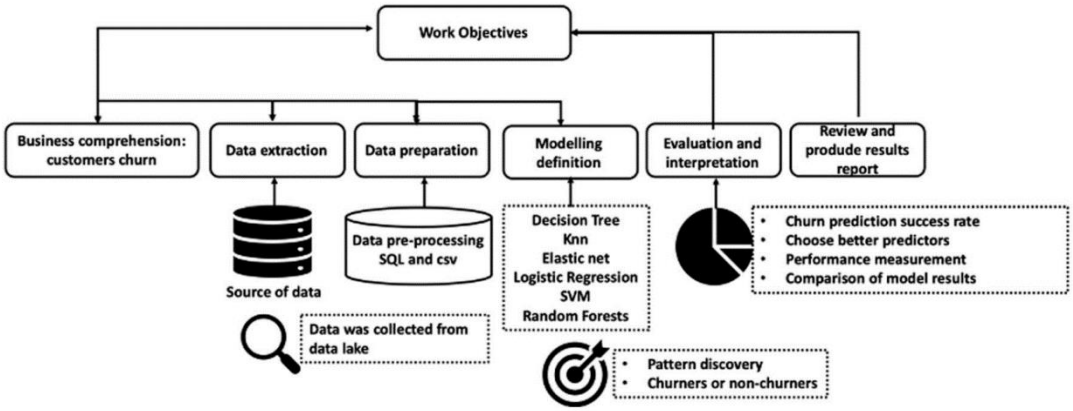
Aporte: En este artículo, los autores exploran un conjunto de datos “rico” en variables relacionadas al comportamiento del cliente a lo largo del tiempo. Los autores logran entrenar distintos modelos de machine learning supervisado con este gran conjunto de datos y concluir que el modelo Random Forest es el que mejor cumple con el objetivo de predecir el abandono de clientes (80,2% recall). Gracias al estudio que han realizado, nos aportan con

la documentación de nuevos conocimientos sobre los principales determinantes que predicen el abandono de clientes. Encontraron que la frecuencia con la que los clientes utilizan los servicios financieros, el volumen del crédito y la posesión de productos, eran las variables de mayor poder predictivo, de tal forma que una estrategia potente es la de fortalecer la relación con los clientes a través de la venta de más productos para retener al cliente.

Proceso: En este artículo, los autores plantean la siguiente metodología o proceso a seguir para atacar el requerimiento de negocio, basándose en lo que dice Chapman et al. (2000).

Figura 81

Proceso para cumplir con los objetivos propuestos por los autores.



De “Propension to customer churn in a financial institution: a machine learning approach”, por Lima Lemos et al., 2022.

Como primer subproceso, realizan la etapa de entendimiento de negocio y procesamiento de datos, incluye la extracción de los datos de un CRM que es alimentado por un gran datalake que, a su vez, es alimentado por datos provenientes de los sistemas transaccionales de la entidad financiera, los cuales proporcionan datos y comportamientos del día a día en la prestación de los servicios. Además, procesaron los datos y armaron un conjunto de datos a nivel de cliente para reducir los tiempos computaciones del análisis. Las variables fueron escogidas basadas en el conocimiento experto del negocio bancario relacionándolas con el problema del abandono de clientes. Al no saber que atributos serán verdaderamente significativos, escogen un amplio conjunto de variables. Con el conocimiento necesario de negocio y la exploración de los datos logran seleccionarse 35 atributos (Figura 2) y 500 000

observaciones de clientes de cuenta corriente observados durante 12 meses de relación con la institución. Posteriormente, realizaron la limpieza de datos atípicos y en algunos casos la imputación de variables. Como etapa final de este primer subproceso, eliminaron los atributos de varianza cercanos a cero, y estandarizaron todas las variables numéricas, esto da como resultado la eliminación de algunos atributos para que finalmente queden 28 (Rajeswari, 2015).

Figura 82

Listado total de variables. Variable target (primera fila) y atributos seleccionados (filas restantes) utilizados en la tarea supervisada de predicción de abandono de clientes.

Class/Attribute	Data type	Description
Churned	Binary (Yes or No)	Customer closed their current accounts or stopped moving them for six months (churned)
Segment	Nominal (4 segments)	Customer segment (Basic income, middle class, high income, and very high income)
Automatic_Debt	Binary (Yes or No)	Use of the direct debit service - at least once in the last 60 days
Salary_Credit	Binary (Yes or No)	Receipt of salary - at least once in the last 60 days
Accreditation	Binary (Yes or No)	Membership to the accreditation service/card make-up
Insurance	Binary (Yes or No)	Ownership of insurance product consortium, capitalization or pension plan
Portability_Request	Binary (Yes or No)	Request for salary credit portability to another financial institution
Complaint_Request	Binary (Yes or No)	A registered complaint in channels managed by OUVID (Ombudsman, SAC, Procon, BACEN)
Automatic_Debt_DIFF	Real value in $[-1, 1]$	Evolution of the use of the automatic debit service - at least once in the last 60 days
Salary_Credit_DIFF	Real value in $[-1, 1]$	Evolution of salary receipt - at least once in the last 60 days
Insurance_DIFF	Real value in $[-1, 1]$	Evolution of the insurance company's product ownership - insurance, consortium, capitalization or pension
Qualified_Products	Integer	Number of products that the customer owns, and that is indicated for the if segment
Qualified_Products_Previous	Integer	Quantity of customer products, and which is indicated for the Position segment: 6 months before
Qualified_Products_DIFF	Integer	Number of products that the customer owns, and that is indicated for the if segment - Absolute variation between 6 months
Qualified_Products_PERC	Percentage	Number of products that the customer owns, and that is indicated for the if segment - Percentage change between 6 months
Products	Integer	Number of products the customer owns
Products_Previous	Integer	Number of products the customer owns - Position: 6 months before
Products_DIFF	Integer	Number of products that the customer has - Absolute change between 6 months
Products_PERC	Percentage	Number of products that the customer has - Percentage change between 6 months
Transactions	Value in R$	Number of spontaneous movements carried out in the current account
Transactions_Previous	Value in R$	Number of spontaneous movements carried out in current account - Position: 6 months before
Transactions_DIFF	Value in R$	Number of spontaneous movements performed in the current account - Absolute variation between 6 months
Transactions_PERC	Percentage	Number of spontaneous movements performed in the current account - Percentage change between 6 months
Investment	Value in R$	Volume invested in investments, savings or deposit account
Investment_Previous	Value in R$	Volume invested in investments, savings or deposit account - Position: 6 months before
Investment_DIFF	Value in R$	Volume invested in investments, savings or deposit account - Absolute change between 6 months
Investment_PERC	Percentage	Volume invested in investments, savings or deposit account Percentage variation between 6 months
Credit	Value in R$	The volume of commercial and housing loans active
Credit_Previous	Value in R$	The volume of commercial and active housing credit - Position: 6 months before
Credit_DIFF	Value in R$	The volume of commercial and active housing credit - Absolute change between 6 months
Credit_PERC	Percentage	The volume of commercial and active housing credit - Percentage change between 6 months
Profitability	Value in R$	Profitability (financial return indicator) of the client, accumulated 12 months
Profitability_Previous	Value in R$	Profitability (financial return indicator) of the client, accumulated 12 months - Position: 6 months before

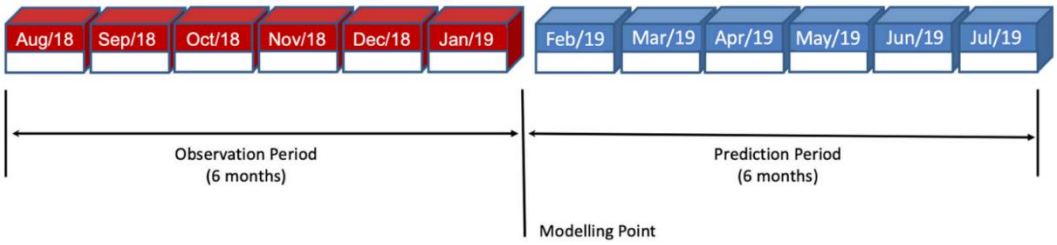
De "Propension to customer churn in a financial institution: a machine learning approach", por Lima Lemos et al., 2022.

Como segundo subproceso, validan que el conjunto de datos sea una muestra representativa de toda la población en terminamos de características observables, edad y dispersión geográfica. Calculan algunas estadísticas básicas sobre el contenido de la muestra y las comparan a las estadísticas evaluadas sobre toda la población para determinar la representatividad de la muestra. Este análisis es crítico para garantizar que la muestra de datos refleja con precisión el comportamiento de los clientes.

En el tercer subproceso, realizan el modelado del proyecto. La figura 2 muestra como los primeros seis meses se utilizan para construir predictores (atributos, variables), y los últimos seis meses se utilizan para definir la variable objetivo. En este caso de estudio, los atributos se componen de los rasgos financieros de clientes extraídos de agosto de 2018 a enero 2019 (color rojo), teniendo como objetivo determinar si el cliente abandono durante los seis meses siguientes, es decir de febrero a julio 2019 (color azul) (De Lima Lemos et al, 2022).

Figura 83

Estrategia de modelado para construir el modelo de predicción de abandono de clientes.

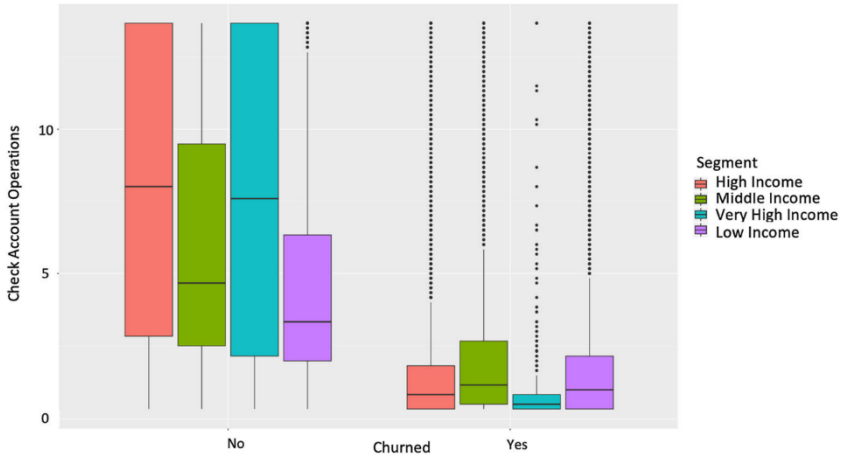


De “Propension to customer churn in a financial institution: a machine learning approach”, por Lima Lemos et al., 2022.

Finalmente, en el cuarto subproceso realizan una exploración y análisis de datos de ciertos atributos específicos para indagar un poco más de su poder predictivo y su entendimiento de negocio mediante gráficos de caja.

Figura 84

Diagrama de caja de las transacciones de la cuenta de cheques de atributo frente a la variable binaria de destino que indica si el cliente abandonó la entidad financiera.



De “Propension to customer churn in a financial institution: a machine learning approach”, por Lima Lemos et al., 2022.

Principal Resultado: Se evaluó los distintos algoritmos de machine learning supervisado, el Random Forest logró los mejores resultados, pudiendo identificar al 80,2% de los clientes que abandonarían el negocio en los meses siguientes (recall). Otro hallazgo importante fue identificar que la frecuencia con la que los clientes utilizan los servicios financieros, el volumen de créditos y la posesión de productos son las variables con mayor poder predictivo. Finalmente, el ultimo resultado nos indica que incluso en el escenario más conservador del modelo que incluye únicamente el 20% de éxito de cualquier campaña lanzada basándose en los datos predichos, estaría significando preservar 290 millones de dólares de ingreso anual en este gran banco de Brasil.

4.3 Conclusiones

A continuación, se presentan un resumen detallado de cada uno de los artículos seleccionados, respetando el orden del cuadro presentado previamente y agrupados según las tipologías descritas en el apartado de prefacio:

4.3.1 Conclusiones de artículos de la categoría de abandono de clientes

Dentro de la primera categoría seleccionada, la cual agrupa diferentes trabajos de investigación relacionados al abandono de clientes, se puede concluir que, de manera general, todas las empresas, independientemente del rubro, son afectadas por la rotación o abandono de sus clientes, ya que esto impacta directamente a sus ingresos. Además, hemos validado que distintos estudios científicos indican que conseguir un nuevo cliente, puede costar hasta cinco veces más que retener a un cliente existente. Por otro lado, la correcta gestión del abandono de clientes tiene como objetivo minimizar estas pérdidas causadas por el desgaste de los clientes, lo que conlleva a que en la actualidad las empresas estén muy interesadas en identificar los posibles abandonos antes que sucedan. Poder predecir el abandono de clientes puede conducir a un ahorro sustancial de ingresos, especialmente para el sector de servicios financieros, ya que estarían ahorrando este alto costo de adquisición de clientes nuevos.

4.3.2 Conclusiones de artículos de la categoría de modelos predictivos

En esta categoría, los artículos se centran en demostrar la utilidad de emplear modelos predictivos en diversos proyectos que ayuden a resolver problemas de negocio. En la mayoría de los casos, hemos analizado artículos que no se alejen mucho del problema de negocio que queremos resolver como equipo, por ende, las aplicaciones de estos modelos han girado en torno al abandono de los clientes y a la segmentación de clientes. La conclusión de esta categoría es la metodología o la seria de pasos que siguen la gran mayoría de los artículos estudiados para diseñar, analizar, modelar y validar los modelos predictivos. En este caso, las fases que encontramos en común en todos los artículos son las siguientes: 1) comprensión empresarial o de negocio, 2) Comprensión de los datos, 3) preprocesamiento de los datos, 4) Modelado, 5) Evaluación e 6) Implementación. Por último, hemos concluido con esta sección que las métricas más adecuadas para evaluar los modelos predictivos son el accuracy, la precisión, el área bajo la curva AUC y el recall. Asimismo, los algoritmos que usualmente son utilizados y posteriormente comparados en los artículos estudiados son: Arboles de decisión, Random Forest, Regresión Logística, Kmeans, K-medoids, XGBoost, entre otros.

4.3.3 Conclusiones de artículos de la categoría de modelos predictivos para abandono de clientes

En esta categoría, se abordaron artículos netamente delimitados a utilizar modelos predictivos aplicados a la resolución del problema de abandono de clientes. En conclusión, los artículos demuestran muy buenos resultados en la predicción del abandono de clientes, mostrando valores de alrededor del 90% en la métrica de accuracy. El algoritmo que más resalta en los resultados de los artículos al obtener los resultados es el XGBOOST, sin embargo, este algoritmo es muy poco interpretable y explicable al negocio, por lo que los siguientes algoritmos con mejores resultados serían el árbol de decisión y el random forest, los cuales si son interpretables y se puede realizar una explicación del modelo a los actores de negocio como por ejemplo el departamento de marketing. Por otro lado, concluimos de todos estos artículos estudiados en esta categoría, que los datos no estructurados pueden aportarnos variables con alto potencial predictivo, ya que denotan un alcance más cercano al comportamiento del cliente, asimismo las variables relacionadas a las redes sociales complementadas con las variables propias de negocio son variables con alto potencial predictivo.

4.3.4 Conclusiones generales

Tras un exhaustivo proceso de investigación, análisis y levantamiento de información, se concluye lo siguiente:

- El abandono o rotación de clientes es un problema actual que impacta a muchas empresas debido a que afecta directamente a los ingresos. Según estudios, la adquisición de un cliente nuevo puede costar de 5 a 6 veces más que la retención de un cliente ya existente, especialmente en entidades de servicios financieros.
- La mayoría de los modelos desarrollados por distintos autores, siguen metodologías muy similares las cuales tienen en común las siguientes fases: 1) comprensión empresarial o de negocio, 2) Comprensión de los datos, 3) pre procesamiento de los datos, 4) Modelado, 5) Evaluación e 6) Implementación.
- Existen diversos tipos de métodos para determinar los conjuntos de datos que serán utilizados para el entrenamiento y pruebas del modelo. Sin embargo, la mayoría de los estudios comparten la idea de segmentar estos conjuntos en un 70% para el entrenamiento y un 30% para la validación.
- Según los estudios revisados, hemos concluido que la distribución de la variable objetivo puede ser un problema al momento de entrenar y evaluar los modelos. Por ejemplo, si el conjunto de datos está conformado por 90% de observaciones

etiquetadas como “no abandono” y 10% de observaciones etiquetadas como “abandono”, esto puede generar problemas y se tendría que utilizar ciertas técnicas para distribuir de manera uniforme el conjunto de datos.

- Todos los artículos científicos relacionados a modelos predictivos realizan el proceso de entrenar y validar más de un algoritmo de clasificación. Es decir, es un paso por realizarse si o si para obtener los mejores resultados ya que cada entidad de negocio funciona distinta. Entre los algoritmos más utilizados por los autores para su comparación están: Arboles de decisión, random forest, regresión logística, XGBoost y SVM.

5 DESARROLLO DEL PROYECTO

En el presente capítulo se mostrará los procedimientos seguidos para poder concluir de manera exitosa el desarrollo de nuestro proyecto en su totalidad. Asimismo, se comparten gráficos y tablas para mostrar lo desarrollado a lo largo del proyecto.

5.1 Análisis

5.1.1 Análisis de la necesidad

La solución propuesta basada en machine learning busca soportar la toma de decisiones en el proceso de retención de clientes mediante la predicción de clientes con alta probabilidad de abandono de sus productos y por lo tanto de la organización. Actualmente el proceso de retención se emplea a todos los clientes de la organización sin distinción alguna y sin enfocar esfuerzos en cierto segmento especifico. Con esta solución, los principales actores del proceso de retención podrán tener a su disposición un listado de potenciales clientes en los cuales enfocar sus esfuerzos y acciones comerciales. Además, se busca brindarle a la organización una ventaja competitiva al permitir tomar decisiones de manera ágil y soportadas en datos. Asimismo, esta solución tecnológica les permitirá abordar de una manera más eficiente su objetivo de reducir la tasa de abandono de clientes en la organización.

5.1.2 Análisis de herramientas para análisis predictivo

Para el desarrollo de la solución es necesario la selección de una herramienta de análisis predictivo que satisfaga las necesidades de la organización de fondos colectivos.

Estas herramientas tienen como principal función analizar y procesar los datos. Posteriormente, la herramienta también debe permitir realizar el apartado predictivo del proyecto, es decir, el entrenamiento, validación y predicción del algoritmo estadístico escogido. Para efectos de solución. Se centrará el análisis de las herramientas en aquellas que realicen las principales funciones que requiere el proyecto. Dentro de los criterios a analizar se tendrán los siguientes:

- **Usabilidad de la herramienta:** En este criterio se evalúa la facilidad de uso de la herramienta por parte de usuarios envueltos en el ámbito de tecnologías de información, sin necesidad de consultar la documentación de la herramienta.
- **Capacidad de procesamiento:** En este criterio se evalúa la capacidad de la herramienta para procesar grandes volúmenes de información.

- **Soporte y Comunidad:** En este criterio se evalúa el soporte brindado por la empresa dueña de la herramienta ante cualquier consulta o incidente. Asimismo, la cantidad de usuarios que conforman la comunidad online de la herramienta, en la cual se resuelven gran cantidad de dudas.
- **Seguridad:** En este criterio se evalúa el nivel de seguridad de la herramienta, se analiza las políticas de seguridad de las empresas que proveen las herramientas.
- **Relación calidad – precio:** En este criterio se evalúa la relación entre los beneficios de la herramienta y su precio.
- **Visualización:** En este criterio se evalúa la capacidad que tiene la herramienta para generar gráficos básicos y complejos que permitan el correcto análisis de los datos.
- **Robustez:** En este criterio se evalúa el nivel de carga de procesamiento que soporta la herramienta, es decir, como responde ante altos niveles de procesamiento de datos.
- **Interacción con aplicaciones y datos:** En este criterio se evalúa la diversidad en las fuentes de datos que permiten la integración con distintas aplicaciones.

El cuadro comparativo resultante de las herramientas más importantes y en tendencia según Gartner a lo largo de los últimos años fueron comparadas y evaluadas según los criterios mencionados anteriormente.

Figura 85

Benchmarking de herramientas de análisis predictivo

		Herramienta											
		Alteryx Analytics	SAP Predictive Analytics	IBM Analytics	Amazon Web Services	SAS Enterprise Miner	RStudio	Jupyter Notebooks	DataRobot	RapidMiner	Google Cloud Platform	KNIME Analytics Platform	Orange
Criterio	Usabilidad de la herramienta	5	4	4	3	3	4	3	4	3	4	1	3
	Capacidad de procesamiento	4	5	4	4	4	4	4	4	4	3	4	2
	Soporte y Comunidad	4	4	4	5	4	4	4	2	2	3	3	2
	Seguridad	4	4	4	4	4	3	3	4	4	3	3	2
	Relación calidad-precio	4	4	4	4	3	4	5	3	4	3	4	5
	Visualización	5	5	5	4	5	4	3	3	4	3	3	0
	Robustez	4	4	4	4	4	3	3	4	3	3	2	1
	Integración con aplicaciones y datos	5	4	4	4	4	4	4	3	3	3	2	1
	Total	35	34	33	32	31	30	29	27	27	25	22	16

La herramienta con mayor puntaje, elegida para la realización del modelo de análisis predictivo, es Alteryx, obteniendo 35 puntos. Según las Reseñas y calificaciones de plataformas de aprendizaje automático y ciencia de datos multi-persona de Gartner, Alteryx obtiene una calificación de 4.6/5 con 346 calificaciones, superando a IBM SPSS, RapidMiner, DataRobot, KNIME, SAS.

En el informe del Cuadrante Mágico de Gartner (2019), Alteryx fue reconocido como un Challenger y logró la posición más destacada en términos de capacidad de ejecución. Esta clasificación valida la habilidad de Alteryx para facilitar a los usuarios la exploración, compartición y preparación de datos, así como la realización de análisis estadísticos, predictivos, prescriptivos y espaciales. Alteryx también ofrece la capacidad de implementar y gestionar modelos analíticos utilizando su plataforma analítica integral de extremo a extremo.

Fortalezas de Alteryx según Gartner (2019):

> "Gran experiencia del cliente: Alteryx obtuvo una calificación de primer nivel en la experiencia del cliente en nuestra encuesta de clientes de referencia. Las puntuaciones fueron consistentemente altas para la experiencia general del cliente, los planes para realizar inversiones adicionales, la inclusión de mejoras de productos,

las funciones solicitadas en versiones posteriores y las capacidades generales del producto.

Habilitación colaborativa de una amplia base de usuarios: Enfoque sin código de Alteryx es atractivo para un amplio espectro de usuarios, desde analistas de negocios y datos hasta científicos de datos ciudadanos. Un enfoque en la facilidad de uso y la cohesión de su plataforma permite la colaboración entre los usuarios.” (p. 1)

Figura 86

Cuadrante Gartner 2018 de mejores herramientas para data science.

De “Cuadro mágico de gartner 2019 para data science & machine learning platforms”, por Gartner, 2019.

5.1.3 Análisis de algoritmos para análisis predictivo

Para las fases de entrenamiento y validación del modelo propuesto, es necesario seleccionar que algoritmos deben ser evaluados y comparados, ya que existe un gran número de algoritmos utilizados en modelos predictivos. Para esta selección, hemos analizado distintos artículos científicos relacionados específicamente a predecir el abandono de clientes.

Tabla 5

Listado de algoritmos referenciados en los artículos científicos analizados.

N°	Algoritmo	Tipo Algoritmo	Cantidad Referencias	Referencias
1	Decision Tree	Classification	5	(De Lima, Silva & Miranda, 2022), (Liu et al., 2022), (Jamjoom, 2021), (Kasem, Jafar & Aljoumaa, 2019), (Lalwani et al.,2021)
2	K-nearest Neighbors	Classification	2	(De Lima, Silva & Miranda, 2022), (Liu et al., 2022)
3	Logistic Regression	Classification	3	(De Lima, Silva & Miranda, 2022), (Jamjoom, 2021), (Vo et al.,2020), (Lalwani et al.,2021)
4	Random Forests	Classification	6	(De Lima, Silva & Miranda, 2022), (Liu et al., 2022), (Vo et al.,2020), (Kasem, Jafar & Aljoumaa, 2019), (Lalwani et al.,2021), (Usman-Hamza et al., 2022)
5	Support Vector Machines	Classification	2	(De Lima, Silva & Miranda, 2022), (Lalwani et al.,2021)
6	XGBoost	Classification	4	(Liu et al., 2022), (Vo et al.,2020), (Kasem, Jafar & Aljoumaa, 2019), (Lalwani et al.,2021)
7	Deep Learning	Classification	1	(Liu et al., 2022)
8	Naive Bayes	Classification	3	(Liu et al., 2022), (Vo et al.,2020), (Lalwani et al.,2021)
9	Neural Networks	Classification	1	(Jamjoom, 2021)
10	Extra Tree Classifier	Classification	1	(Lalwani et al.,2021)
11	CatBoost Classifier	Classification	1	(Lalwani et al.,2021)
12	Logistic Model Tree	Classification	1	(Usman-Hamza et al., 2022)
13	ProfTree	Classification	1	(Höppner et al., 2018)

En base al análisis realizado, hemos seleccionado los algoritmos basándonos en la cantidad de referencias en los papers: Decision tree, Random Forest, Logistic Regression, XGBoost y Naive Bayes. Asimismo, según el análisis de los resultados y conclusiones de los artículos científicos, dichos algoritmos seleccionados son los que obtienen mejores métricas en su evaluación. Por otro lado, Logistic Model Tree y ProfTree, son algoritmos evolucionados, que fueron originados de la combinación de 2 o más algoritmos tradicionales, es por ello que tampoco han sido escogidos para la propuesta debido a su alto nivel de complejidad y poca información en la literatura.

Adicional a ello, existen otros criterios que consideramos deben tomarse en cuenta para la elección de los algoritmos:

- **Interpretabilidad:** Este criterio busca evaluar el nivel de interpretabilidad de los resultados que emite el algoritmo como salida. Por ejemplo, un árbol de decisión tiene mayor nivel de interpretabilidad de sus resultados al momento de exponerlos

ante un nivel gerencial dentro de la organización, en comparación a un algoritmo XGBoost.

- **Configuración de hiperparámetros:** Este criterio busca evaluar la flexibilidad de configuración del algoritmo mediante sus hiperparámetros, ya que esto ayuda a optimizar y tener la posibilidad de mejorar las métricas resultantes.
- **Factibilidad técnica con la herramienta elegida:** Estos criterios buscan evaluar la factibilidad de utilizar el algoritmo en la herramienta elegida para el desarrollo del proyecto

Estos criterios se evaluarán asignando una nota a cada algoritmo entre 0 y 5. Siendo 0 cumplimiento débil, 1 cumplimiento parcial, 2 cumplimiento normal, 3 cumplimiento por encima de lo normal, 4 cumplimiento sobresaliente y 5 cumplimiento excelente. Esta puntuación se llevó a cabo basándose en el análisis previo de cada artículo científico.

Figura 87

Benchmarking de algoritmos seleccionados.

		Algoritmo				
		Decision tree	Random Forest	Logistic Regression	Naive Bayes	XGBoost
Criterio	Interpretabilidad de resultados	5	3	3	2	2
	Configuración de hiperparámetros	5	4	0	3	3
	Factibilidad técnica con la herramienta elegida	5	5	5	2	2
	Total	**15**	**12**	**8**	**7**	**7**

El rubro de los fondos colectivos en Sudamérica en un rubro en crecimiento según los antecedentes detallados al inicio del documento. Los criterios planteados en esta comparativa buscan la selección de los algoritmos más adecuados para la implementación de este proyecto en este tipo de empresas. Por ejemplo, la interpretabilidad de los resultados será muy importante para que los roles correspondientes sustenten sus hallazgos ante un nivel gerencial de negocio. Tomando en cuenta estos criterios y las puntuaciones, se eligen

los algoritmos Decision Tree, Random Forest y Logistic Regression como los propuestos por el modelo de análisis predictivo para abandono de clientes en una empresa administradora de fondos colectivos.

5.1.4 Análisis de variables predictoras de abandono de clientes en empresas financieras

En base a la literatura se han recopilado diversas variables más usadas en modelos predictivos de abandono de clientes en distintas industrias.

Tabla 6

Listado de variables predictivas referenciadas en los artículos científicos analizados.

N°	Variable	Tipo de dato	Referencias
1	Segmento	Categórica	(De Lima, Silva & Miranda, 2022), (De Caigny, Coussement, De Bock & Lessmann 2019)
2	Crédito salarial	Numérico continuo	(De Lima, Silva & Miranda, 2022)
3	Acreditación	Binaria	(De Lima, Silva & Miranda, 2022)
4	Seguro	Categórica	(De Lima, Silva & Miranda, 2022)
5	Solicitud de portabilidad de crédito	Binaria	(De Lima, Silva & Miranda, 2022)
6	Reclamos	Binaria	(De Lima, Silva & Miranda, 2022), (Abdelrahim, Assef & Kadan, 2019)
7	Productos	Numérico discreto	(De Lima, Silva & Miranda, 2022), (Hoppner, Stripling, Baesens, Broucke & Verdonck, 2018), (De Caigny, Coussement, De Bock & Lessmann 2019)
8	Transacciones	Valor en dólares	(De Lima, Silva & Miranda, 2022), (Cheng, Wu & Chen, 2019), (Lalwani, Kumar, Singh & Sethi, 2021), (De Caigny, Coussement, De Bock & Lessmann 2019)
9	Inversión	Valor en dólares	(De Lima, Silva & Miranda, 2022)
10	Línea de crédito	Valor en dólares	(De Lima, Silva & Miranda, 2022), (Cheng, Wu & Chen, 2019), (De Caigny, Coussement & De Bock, 2018), (De Caigny, Coussement, De Bock & Lessmann 2019)
11	Rentabilidad	Valor en dólares	(De Lima, Silva & Miranda, 2022)
12	Demografía de residencia	Categórica	(Cheng, Wu & Chen, 2019), (Abdelrahim, Assef & Kadan, 2019), (Hoppner, Stripling, Baesens, Broucke & Verdonck, 2018)
13	Edad	Numérico discreto	(Cheng, Wu & Chen, 2019), (Kaya et. al, 2018), (De Caigny, Coussement, De Bock & Lessmann 2019)
14	Frecuencia de llamadas	Categórica	(Cheng, Wu & Chen, 2019), (Abdelrahim, Assef & Kadan, 2019), (Ullah et. al, 2019), (De Caigny, Coussement & De Bock, 2018)

15	Género	Categórica	(Cheng, Wu & Chen, 2019), (Abdelrahim, Assef & Kadan, 2019), (Lalwani, Kumar, Singh & Sethi, 2021), (Kaya et. al, 2018), (De Caigny, Coussement, De Bock & Lessmann 2019)
16	Satisfacción del cliente en llamadas	Categórica	(Cheng, Wu & Chen, 2019), (Abdelrahim, Assef & Kadan, 2019)
17	Frecuencia de devolución de llamadas al cliente	Categórica	(Cheng, Wu & Chen, 2019), (Abdelrahim, Assef & Kadan, 2019), (De Caigny, Coussement & De Bock, 2018)
18	Método de pago	Categórica	(Lalwani, Kumar, Singh & Sethi, 2021), (Hoppner, Stripling, Baesens, Broucke & Verdonck, 2018), (De Caigny, Coussement, De Bock & Lessmann 2019)
19	Periodo de contrato	Numérico discreto	(Hoppner, Stripling, Baesens, Broucke & Verdonck, 2018)
20	Ingreso promedio	Valor en dólares	(Hoppner, Stripling, Baesens, Broucke & Verdonck, 2018), (Kaya et. al, 2018), (De Caigny, Coussement, De Bock & Lessmann 2019)
21	Periodos de no pago	Numérico discreto	(Hoppner, Stripling, Baesens, Broucke & Verdonck, 2018)
22	Estado marital	Categórica	(Kaya et. al, 2018), (De Caigny, Coussement, De Bock & Lessmann 2019)
23	Grado de estudios	Categórica	(Kaya et. al, 2018)
24	Tipo de empleo	Categórica	(Kaya et. al, 2018), (De Caigny, Coussement, De Bock & Lessmann 2019)
25	Tiempo de llamadas en total	Numérico discreto	(Ullah et. al, 2019), (De Caigny, Coussement & De Bock, 2018)

Adaptándonos al contexto y particularidad del rubro de fondos colectivos, se seleccionan las variables más referenciadas en los estudios. Además, las variables que sean factibles de obtener para el rubro de fondos colectivos. En consecuencia, seleccionamos las variables: Segmento, Seguro, Reclamos, Productos, Transacciones, Línea de crédito, Demografía de residencia, Edad, Frecuencia de llamadas, Género, Satisfacción del cliente en llamadas, Método de pago, Periodo de contrato, Ingreso promedio, Periodos de no pago, Estado marital, Grado de estudios, Tipo de empleo.

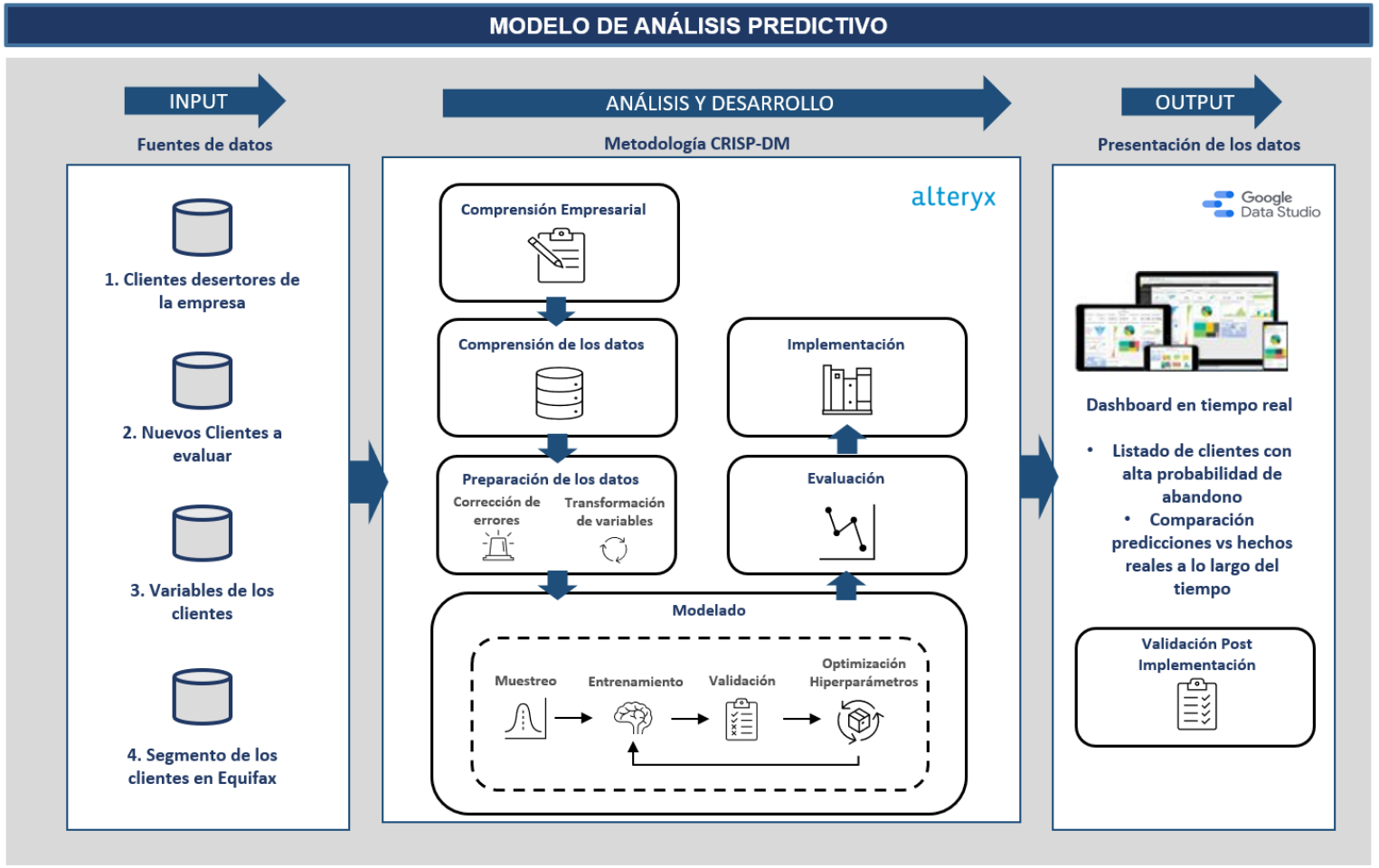
5.2 Diseño del modelo

El diseño del modelo propuesto esquematiza la agrupación y secuencia de los componentes más importantes para el correcto desarrollo e implementación del modelo de análisis predictivo en empresas de fondos colectivos. Los componentes que conforman esta propuesta fueron extraídos de los mejores modelos de análisis predictivo en los artículos científicos que se analizaron en la primera parte del estudio, los cuales están relacionados específicamente problema de abandono de clientes, sin embargo, este proyecto orienta todos estos componentes al rubro de los fondos colectivos. Se identifican 3 grandes componentes en el diseño de la solución: Las entradas, el análisis y desarrollo y las salidas. A continuación,

se irá detallando punto por punto cada paso realizado para el desarrollo del presente proyecto. En el siguiente gráfico, se busca resumir de manera sencilla el diseño de la solución y sus componentes.

Figura 88

Diseño del modelo de análisis predictivo propuesto.



5.2.1 Subprocesos y roles

En la siguiente tabla se muestra una lista de los subprocesos que conforman el proceso propuesto. Además, se indican los roles principales que intervienen en el subproceso.

Tabla 7

Lista de subprocesos y roles asignados.

Subprocesos	Rol
Entendimiento del negocio/problema	Data Analyst
Diseño de arquitectura	Data Architect
Exploración de los datos	Data Analyst / Data Scientist
Preparación de los datos	Data Engineer / Data Scientist
Modelado	Data Scientist / Machine Learning Engineer
Evaluación	Data Scientist
Despliegue	Data Architect / Machine Learning Engineer
Visualización de resultados y operaciones	Data Analyst

5.2.2 Entradas

Las entradas del modelo de análisis predictivo propuesto son las siguientes:

- **Clientes desertores de la empresa:**

  Contiene información histórica de los clientes desertores y no desertores de la empresa. En este conjunto de datos se encontrará la variable objetivo, la cual nos indicara si el cliente ha desertado o no ha desertado de la entidad financiera.

- **Nuevos Clientes por evaluar:**

  Es el listado de clientes nuevos que deseamos evaluar a través del modelo predictivo, es decir, son los clientes que queremos predecir si abandonaran o no abandonaran a la organización.

- **Variables de los clientes:**

  Según el análisis realizado previamente en la literatura sobre las diferentes variables predictoras que se han utilizado en los distintos modelos predictivos para abandono de clientes, se detalla las variables propuestas para el modelo orientado al rubro de los fondos colectivos. Cabe mencionar que dichas variables se

obtienen tanto de fuentes internas de la organización (sistemas transaccionales) como de fuentes externas, (entidades externas).

Tabla 8

Tabla de variables propuestas por el modelo predictivo.

N°	Variable en literatura	Variable equivalente enfocada a fondos colectivos	Descripción
1	Segmento	Segmento Equifax	Segmentación del cliente según Equifax (Segmento 1, 2, 3, 4, 5)
2	Seguro	Seguro vehicular	Seguro contratado para el vehículo (Seguro contratado con la empresa, seguro contratado de forma externa)
3	Reclamos	Reclamos en libro	Indicador de si el cliente registró al menos un reclamo en el libro de reclamaciones (Si, No)
4	Productos	Contratos	Número de contratos poseídos por el cliente
5	Transacciones	Movimientos	Importe en dólares de los movimientos realizados por el cliente
6	Línea de crédito	Línea de crédito adicional	Línea de crédito adicional en dólares asignada al cliente para financiamientos, diferencia de precio, trámites producto de la evaluación crediticia
7	Demografía de residencia	Demografía de residencia	Región, provincia y distrito de residencia del cliente
8	Edad	Edad	Edad del cliente
9	Frecuencia de llamadas	Frecuencia de llamadas al call center	Número de llamadas del cliente al call center
10	Género	Género	Género del cliente
11	Satisfacción del cliente en llamadas	Satisfacción del cliente en llamadas al call center	Satisfacción del cliente en llamadas al call center (1 -> poco satisfecho con la atención, a 5 -> muy satisfecho con la atención)
12	Método de pago	Método de pago de cuotas	Método de pago empleado para las cuotas (pago por app/web de banco, pago presencial en banco, débito automático)
13	Periodo de contrato	Meses del contrato	Número de meses del contrato (60, 72, 120)
14	Ingreso promedio	Ingreso promedio en dólares	Ingreso promedio del cliente en dólares declarado en el ingreso de venta
15	Periodos de no pago	Mensualidades no pagadas	Número de mensualidades no pagadas por el cliente hasta la fecha
16	Estado marital	Estado marital	Estado marital del cliente (soltero, casado, viudo, divorciado, conviviente)
17	Grado de estudios	Grado de estudios	Grado de estudios del cliente (primaria, secundaria, egresado, bachiller, titulado, magister, doctor, sin grado)
18	Tipo de empleo	Tipo de empleo	Tipo de empleo del cliente (dependiente, independiente, sin empleo)

- **Segmento de los clientes en Equifax:**

  La entidad Equifax es la fuente externa propuesta por el modelo, la literatura previamente nos indica que una variable con alta capacidad predictiva en el abandono de clientes es el nivel socioeconómico. Sin embargo, esta variable no se captura de manera interna en la organización por lo que se deberá obtener dicha variable de un servicio pagado ante la entidad Equifax, la cual nos proporcionará dicha variable para todos nuestros clientes.

5.2.3 Análisis y desarrollo

El proceso de análisis y desarrollo del modelo predictivo para abandono de clientes consta de 6 subprocesos que se rigen por la metodología CRISP-DM. A continuación, se detallará cada subproceso tomando en consideración que el modelo propuesto es orientado al rubro de los fondos colectivos.

5.2.3.1 Comprensión empresarial

La fase de comprensión empresarial requiere que se tenga claridad de los aspectos de negocio involucrados en el proyecto. Para la correcta implementación del proyecto propuesto, se requiere definir los siguientes aspectos:

- **Definir el abandono de un cliente:**

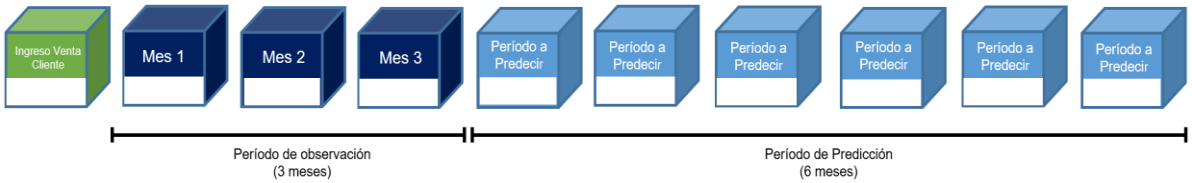
  En el rubro de fondos colectivos se considera abandono de cliente cuando este aún no ha recibido el bien (vehículo, inmueble) y deja de pagar su contrato asociado con la empresa durante 3 meses seguidos. Si este escenario ocurre cuando el bien esta entregado, no se considera abandono de cliente. Esta comprensión nos permitirá posteriormente categorizar la variable objetivo de cada cliente: abandonó o no abandonó.

- **Definir el momento de predicción:**

  Tomando en cuenta las variables seleccionadas en la fase de análisis, se requiere tener un periodo de observación del cliente en el cual podamos obtener el total de las variables propuestas, ya que muchas de ellas requieren de un periodo de tiempo para poder obtenerse. En este caso, se propone un periodo de observación de 3 meses.

Figura 89

Estrategia para construir el modelo predictivo en base al momento de predicción establecido.



5.2.3.2 Comprensión de los datos

En esta fase se debe comprender la distribución de los datos. Para ello debemos analizar mediante gráficos cada variable, utilizaremos distintos gráficos dependiendo del tipo de dato de la variable, es decir, se debe realizar un análisis exploratorio general de los datos.

Tabla 9

Tabla de variables con el tipo de gráfico propuesto para su análisis.

Tipo de variable	Variable	Tipo de Gráfico
Categórica	Segmento Equifax Seguro vehicular Reclamos en libro Grado de estudios Tipo de empleo Género Satisfacción del cliente en llamadas al call center Método de pago de cuotas Estado marital Demográfica de residencia	Gráfico de barras, Gráfico de sectores o circular
Numérica	Edad Frecuencia de llamadas al call center Meses del contrato Ingreso promedio en dólares Mensualidades no pagadas Contratos Movimientos Línea de crédito adicional	Histograma, Diagrama de tallo y hoja, Diagrama de cajas

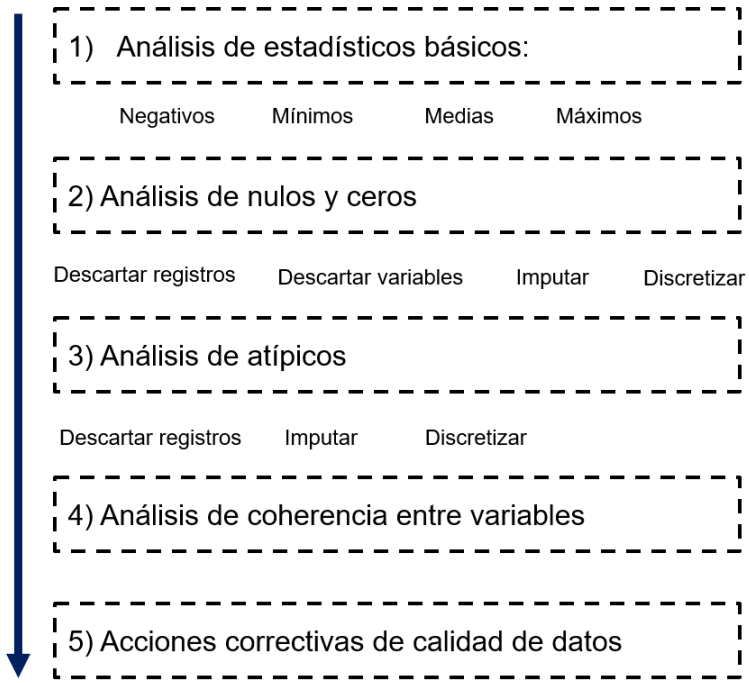
Mediante el análisis independiente de cada variable, se identificarán datos que no hagan sentido a nivel de negocio, datos faltantes, datos atípicos, etc. Se deberán identificar todas estas casuísticas para poder corregirlas en la fase de preparación de los datos.

5.2.3.3 Preparación de los datos

Esta fase es la más extensa y fundamental en el proceso de desarrollo del modelo predictivo. Tiene como objetivo realizar dos principales tareas: corregir cualquier error en los datos y transformar las variables. Todo error de datos que no se corrija en esta fase, implicara un retroceso en el proyecto si se detecta posteriormente. Para este primer objetivo se propone utilizar la siguiente secuencia de procesos de calidad de datos para conocer exactamente lo que se requiere revisar.

Figura 90

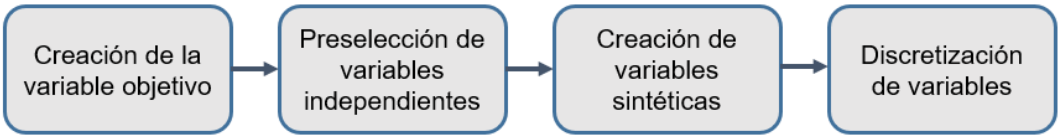
Metodología propuesta para un correcto subproceso de calidad de datos.



Una vez terminado el proceso propuesto para la calidad de los datos, la fase de preparación también debe incluir una etapa que consista en poder dar forma a nuestras variables predictivas con el objetivo de ser más entendibles por nuestros algoritmos de aprendizaje automático. Esta transformación de datos consta de ciertas reglas base según indican en los artículos científicos.

Figura 91

Subproceso propuesto para la transformación de variables.



En primer lugar, debemos crear nuestra variable objetivo o target, ya que esta será la variable más importante para que los algoritmos puedan saber si un cliente abandonó o no abandonó el fondo colectivo. El algoritmo se va a basar justamente en la variable target para realizar sus predicciones, por lo que es muy importante definir bien la definición de cliente desertor o customer churn.

En segundo lugar, deberíamos realizar una preselección de variables independientes, es decir, esta acción permite reducir la cantidad de variables y así contrarrestar ciertos problemas como la cantidad del procesamiento necesario. Sin embargo, para este caso particular de los fondos colectivos, ya se ha realizado una preselección de variables al momento de analizar y recolectar de cada artículo científico, las mejores variables para predecir el abandono de un cliente, las cuales se detallan en la sección 5.2.2 variables de cliente como entrada al modelo.

En tercer lugar, creamos nuevas variables a partir de variables existentes que aportan una lógica de negocio adicional, adoptan la forma que necesitan los algoritmos y permiten extraer el máximo de información recogido en las variables originales. Para crear estas nuevas variables, se utilizarán las variables originales preparadas en la fase anterior, solo en las que se pueda aplicar las siguientes reglas de comportamiento:

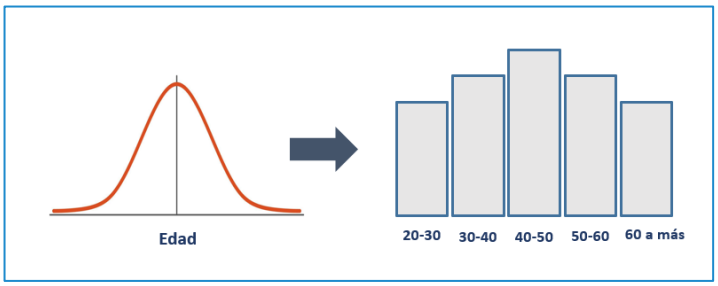
- **Tenencia:** transformar las variables relacionadas a saldos o consumos en indicadores 0 y 1, es decir, para indicar si tiene o no tiene saldos o consumos en cierto producto en específico.
- **Contratación:** Crear una variable que me permita saber si el cliente ha adquirido un nuevo producto en los últimos 3 meses, en el sector de los fondos colectivos, seria poder saber si esa persona ha comprado algún nuevo certificado en el periodo que tenemos de evaluación del cliente.

- **Cancelación:** Crear una variable que permita saber si el cliente ha cancelado cierto producto. En el sector de fondos colectivos, podría interpretarse como haber realizado una proforma de un segundo vehículo en el periodo de observación y luego no haya concretado la venta.
- **Medias:** Si se tiene variables relacionadas a importes o saldos de cada mes, sacar el promedio de los meses de observación y simplemente utilizar la variable del promedio.
- **Tendencia:** Busca crear una variable que refleje la tendencia de cierta variable a lo largo del tiempo: en aumento, se mantiene, disminuye. En el rubro de fondos colectivos, según las variables escogidas en este modelo, ninguna aplica para crear una variable de tipo tendencia.

En cuarto lugar, se busca transformar variables continuas en variables categóricas. Esto incrementa la capacidad de generalización, facilita gestionar los valores nulos y valores atípicos que se hayan podido saltar de la fase de preparación, y mejora la interpretación de los resultados hacia las áreas de negocio. A continuación, se ejemplifica la conversión de la variable continua "Edad", a una nueva variable categórica "Edad_Disc" en la cual agrupamos las edades en tramos: clientes de 20 a 30 años, clientes de 30 a 40 años, clientes de 40 a 50 años, clientes de 60 años a más.

Figura 92

Ejemplificación de conversión de la variable Edad.



5.2.3.4 Modelado

La fase del modelado predictivo es un proceso que consta de 4 etapas que en conjunto tienen como objetivo utilizar los datos preparados y transformados para que los algoritmos de aprendizaje automático tomen esos datos como modelo y puedan predecir si un cliente

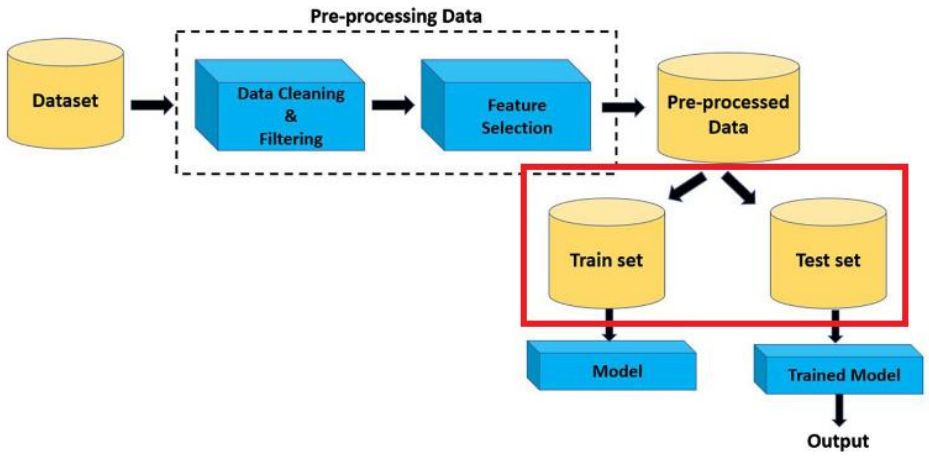
abandonará o no abandonará el fondo colectivo. Las etapas dentro del modelado son: Muestreo, Entrenamiento, Validación y Configuración de hiperparámetros.

5.2.3.4.1 Muestreo

Según los distintos modelos predictivos que hemos analizado en el estado del arte, los estudios recomiendan realizar una partición del conjunto de datos antes de la fase de entrenamiento. Se propone dividir de forma aleatoria el 70% del conjunto de datos para la fase de entrenamiento (train set), y el 30% restante para la fase de validación (test set), ya que con este porcentaje de datos se podrá obtener métricas de evaluación con datos que el modelo no ha visto anteriormente, lo cual simularía a un ambiente productivo.

Figura 93

Proceso de división del conjunto de datos.



Adaptado de "Customer churn prediction system: a machine learning approach", por Lalwani et al., 2021.

5.2.3.4.2 Entrenamiento

En esta fase se aplicarán los algoritmos de aprendizaje automático propuestos, es decir, el conjunto de datos de entrenamiento (train set) preparado servirá como entrada para cada uno de los algoritmos, y estos nos brindarán uno o más resultados como salidas. En este caso, los algoritmos mínimos a evaluar propuestos por el modelo predictivo de abandono de clientes en una empresa administradora de fondos colectivos son: Decision Tree, Random Forest, Logistic Regression. El modelo propone evaluar como mínimo esos 3 algoritmos y ponerlos a competir con las métricas que se detallaran en las etapas de validación y evaluación.

5.2.3.4.3 Validación

Según la literatura, esta fase del proceso es clave para analizar el desempeño del modelo propuesto. Para la validación del modelo se utilizará el conjunto de datos de validación (test set), el cual representa el 30% del conjunto de datos original. Además, se propone utilizar la matriz de confusión y la curva AUC como los medios para obtener métricas resultantes de los algoritmos y poder documentar los resultados.

Las métricas de evaluación que propone el modelo predictivo se basan en el análisis de gran cantidad de artículos científicos que exponen las métricas utilizadas para validar y evaluar los algoritmos, además en la situación actual tecnológica y de negocio de las empresas administradoras de fondos colectivos.

A continuación, se ejemplificará como sería el resultado de la matriz de confusión y que datos deberíamos evaluar:

- **Verdaderos positivos:** Es el número de clientes que sabemos que abandonó el fondo colectivo, y que el modelo predijo correctamente que abandonarían.
- **Verdaderos Negativos:** Es el número de clientes que sabemos que no abandonó el fondo colectivo, y que el modelo predijo correctamente que no abandonarían.
- **Falsos positivos:** Es el número de clientes que sabemos que no abandonó el fondo colectivo, pero que el modelo predijo erróneamente que si abandonarían.
- **Falsos negativos:** Es el número de clientes que sabemos que abandonó el fondo colectivo, pero que el modelo predijo que no abandonarían.

Figura 94

Ejemplificación de matriz de confusión resultante.

		Resultados predicción del modelo	
		Abandonará	No Abandonará
Realidad Histórica	Abandona	Verdaderos positivos	Falsos negativos
	No Abandona	Falsos positivos	Verderos Negativos

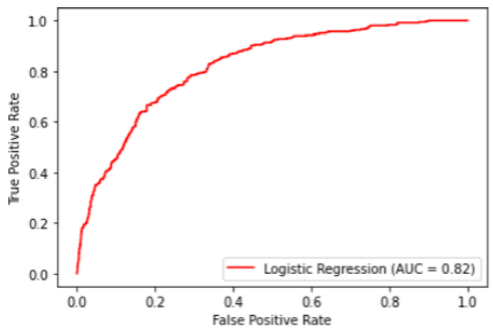
El modelo de análisis predictivo propuesto recomienda analizar esta matriz basándose en las estrategias o acciones que el negocio tenga planeado realizar, las cuales deberían detallarse

en la etapa de comprensión empresarial. Con este dato sabremos que sección de la matriz de confusión impacta más al negocio. Por ejemplo, si la acción de negocio es invertir dinero en regalos para cada uno de los clientes que el modelo prediga que abandonarán, la sección que tiene más impacto son los falsos positivos, porque significa el porcentaje de clientes para los cuales gastaré dinero en sus regalos, y no iban a abandonar el fondo colectivo. En cambio, si la estrategia no contempla una inversión de dinero, se podría revisar la sección de verdaderos positivos, porque así sabremos el porcentaje de clientes que si abandonarían que logramos abarcar correctamente con el modelo.

Por otra parte, el análisis de los resultados de la curva AUC, es orientarlo a cuantificar el rendimiento de los algoritmos en las clases positivas y negativas del conjunto de prueba, Cuanto mayor sea el valor de la puntuación AUC, mejor se desempeñará el modelo en las clases positivas y negativas. Las puntuaciones de AUC deben documentarse para cada algoritmo entrenado. En la siguiente ejemplificación, tomamos un gráfico resultante del artículo científico de Lelwani, el cual nos muestra un resultado de 82% de AUC para el algoritmo Logistic Regression para el conjunto de datos aplicado en ese estudio.

Figura 95

Ejemplificación de gráfico de curva AUC.



(a) LR

De “Customer churn prediction system: a machine learning approach”, por Lalwani et al., 2021.

5.2.3.4.4 Configuración de hiperparámetros

En la fase anterior se obtuvieron los primeros resultados por cada algoritmo. Sin embargo, estos resultados pueden ser volverse a obtener luego de realizar ciertas acciones de mejora en los hiperparámetros de los algoritmos. Esta configuración manual y a demanda que realiza el científico de datos, tiene como objetivo mejorar y alinear los resultados a las estrategias comerciales que el negocio propone como acciones a tomar ante la predicción de clientes que abandonarían el fondo colectivo.

A continuación, se detalla cuáles son los hiperparámetros configurables en cada algoritmo propuesto por nuestro modelo de análisis predictivo.

- **Random Forest:** Número de variables candidatas en cada Split, número de árboles a construir
- **Decision tree:** Split o medida de corte, parámetro de complejidad (regula la poda del árbol), profundidad del árbol en niveles, mínimo de casos en los que el nodo no se divide.
- **Logistic Regression:** No tiene hiperparámetros a configurar.

5.2.3.5 Evaluación

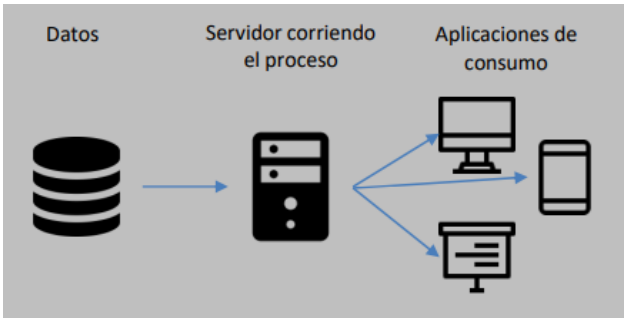
En esta fase se utilizarán todos los resultados documentados en las fases de validación, y también los cambios que realizamos en los hiperparámetros con sus respectivos nuevos resultados. El proceso de evaluación simplemente consta de comparar los resultados de cada algoritmo, interpretarlos, analizar sus ventajas y desventajas en cuanto a la interpretación de los resultados, y poder decidir por la elección de un algoritmo ganador, el cual será posteriormente implementado en un ambiente productivo.

5.2.3.6 Implementación

La implantación consiste en la puesta del modelo generado en algún sistema de producción. El modelo propuesto sugiere seguir los siguientes pasos generales para la puesta en producción.

Figura 96

Proceso general para la implementación del modelo predictivo propuesto.



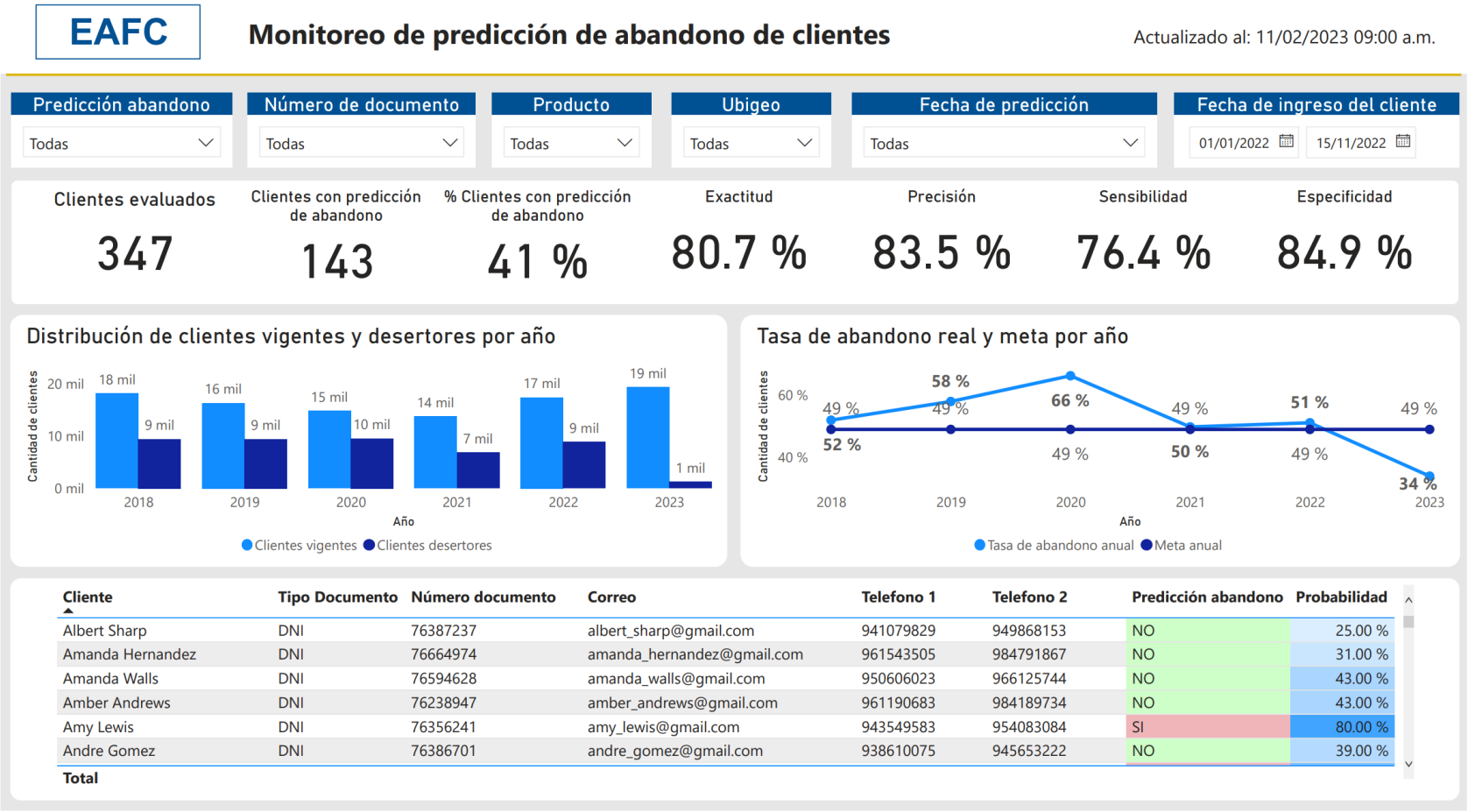
5.2.4 Salidas

Las principales salidas del modelo de análisis predictivo propuesto son las siguientes:

- Dashboard en tiempo real con el objetivo que el proceso de negocio de retención de clientes tenga la información disponible y actualizada al momento, ya que constantemente se estarán lanzando acciones comerciales y preventivas a los clientes con alta probabilidad de abandono. Además, el dashboard también ira reconociendo a lo largo del tiempo que clientes predichos por el modelo efectivamente llegan a abandonar o no abandonar la organización, lo cual servirá para el proceso de validación post implementación.
- El dashboard propuesto facilita el acceso a los resultados del modelo, la predicción del abandono y la probabilidad de ocurrencia. Asimismo, se muestran datos del cliente para su rápida identificación y contactabilidad. El tablero cuenta con distintos filtros que permiten navegar entre los distintos criterios como: Producto, Ubigeo y Fecha de predicción. Estos datos soportarían la toma de decisiones y orientación de las campañas comerciales del proceso de retención de clientes.

Figura 97

Prototipo de dashboard para la visualización de los datos resultantes del modelo.



Cliente	Tipo Documento	Número documento	Correo	Telefono 1	Telefono 2	Predicción abandono	Probabilidad
Albert Sharp	DNI	76387237	albert_sharp@gmail.com	941079829	949868153	NO	25.00 %
Amanda Hernandez	DNI	76664974	amanda_hernandez@gmail.com	961543505	984791867	NO	31.00 %
Amanda Walls	DNI	76594628	amanda_walls@gmail.com	950606023	966125744	NO	43.00 %
Amber Andrews	DNI	76238947	amber_andrews@gmail.com	961190683	984189734	NO	43.00 %
Amy Lewis	DNI	76356241	amy_lewis@gmail.com	943549583	954083084	SI	80.00 %
Andre Gomez	DNI	76386701	andre_gomez@gmail.com	938610075	945653222	NO	39.00 %
Total							

5.2.5 Arquitectura de capas

La arquitectura de capas de la solución propuesta nos muestra como esta se integra con las distintas capas dentro de la organización: capa de negocio, capa de aplicaciones y capa tecnológica.

- **Capa de negocio**

  Actualmente, la empresa de fondos colectivos tiene un pequeño proceso de retención de clientes, el cual se ejecuta mensualmente por el área de marketing. El proceso actual consta de la ejecución de envío de correos preventivos a la nueva cartera de clientes indicando y recomendando el pago puntual. Sin embargo, con la implementación de la solución propuesta, este proceso se convierte en estratégico y consta de dos actividades: el análisis del dashboard resultante como salida del modelo de análisis predictivo y la toma de decisiones estratégicas y comerciales basadas en los datos resultantes del modelo. Los actores involucrados en el proceso de negocio se mantienen siendo los mismos: El analista de Marketing y el jefe de Marketing.

- **Capa de aplicaciones**

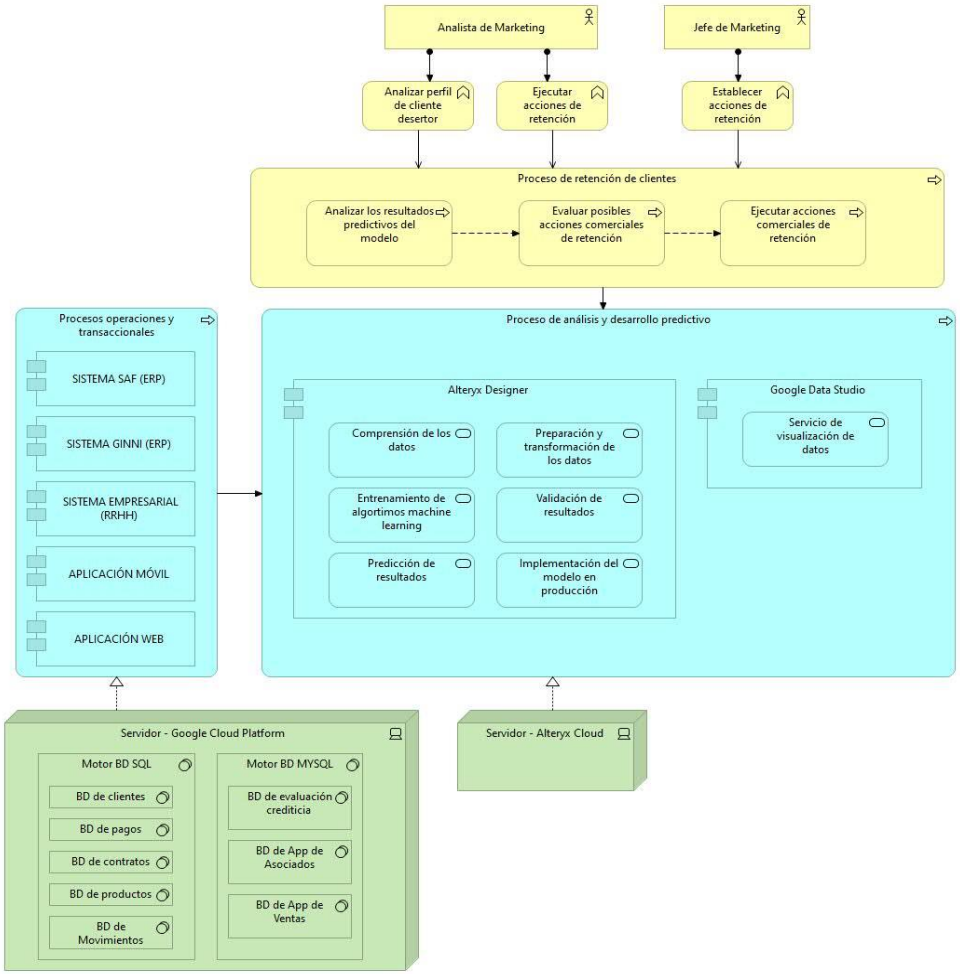
  En esta capa se puede visualizar las distintas tareas que se requieren realizar para el proceso de análisis predictivo, y la aplicación que va a soportar todas estas tareas, en este caso, Alteryx Designer. Por otro lado, la salida del modelo incluye la propuesta de desarrollo y puesta en producción de un dashboard en tiempo real, el cual esta soportado por la aplicación Google Data Studio. Asimismo, la capa de aplicaciones detalla las tablas involucradas en la propuesta, las cuales serán utilizadas para todo el proceso de análisis predictivo en Alteryx Designer.

- **Capa tecnológica**

  La capa tecnológica nos muestra como los componentes necesarios para la implementación del modelo predictivo propuesto. En este caso se decidió por dos instancias una SQL y una MYSQL, alojadas en Google Cloud Platform, debido a que la organización tiene toda su estructura y arquitectura de datos en esta plataforma, por lo tanto, con este diagrama buscamos incluir nuestra solución en la situación actual de la empresa.

Figura 98

Arquitectura de capas de la solución propuesta.



6 VALIDACIÓN DE LA PROPUESTA

6.1 Validación de factibilidad técnica

6.1.1 Requerimientos de hardware

A continuación, se detallan los requerimientos de hardware necesarios para la implementación del proyecto, los cuales deben soportar el software Alteryx Designer y el uso del servicio Looker.

6.1.1.1 Servidor

Según la página web oficial de Alteryx Designer, el software requiere de los siguientes requisitos mínimos de hardware:

Tabla 10

Requerimientos de hardware del servidor

Cantidad	Descripción	Requisito mínimo	Requisito recomendable
1	Núcleos de procesador	4	-
1	Velocidad de procesador	2.5 GHz	-
1	Memoria RAM	8 GB DDR3	16 GB DDR3
1	Espacio libre en disco	500 GB	1 TB

La empresa ya cuenta con servidores virtualizados con estas características, además de tener espacio para habilitar nuevos de ser necesario.

6.1.1.2 Computadora de escritorio o laptop

Se requiere de una computadora de escritorio o laptop para la implementación ya que se necesita para los siguientes fines:

- Acceder al servidor remoto donde se encuentra instalado Alteryx Designer.
- Acceder al servicio de Google Looker para construir el dashboard.

Dado que este equipo se usará principalmente para trabajar con aplicaciones y servicios cuyo alojamiento y cómputo se ejecutan en otra máquina, las especificaciones mínimas necesarias son básicas. Se proponen las siguientes especificaciones como mínimo:

Tabla 11

Requerimientos mínimos para computadora de escritorio o laptop

Cantidad	Descripción	Requisito mínimo	Requisito recomendable
1	Procesador	Core i3 7ma gen	Core i5 7ma gen
1	Memoria RAM	4 GB DDR3	8 GB DDR3
1	Espacio libre en disco	500 GB	1 TB
1	Tamaño de pantalla	13”	15”

Las estaciones de trabajo de todos los colaboradores del área de TI, en su totalidad laptops, si cumplen con los requisitos mínimos recomendados.

6.1.1.3 Infraestructura de red

La empresa tiene una infraestructura de red que le permite tener integradas todas las estaciones de trabajo y los servidores de desarrollo y producción de los distintos sistemas que gestiona la empresa.

Respecto al servidor donde se alojará Alteryx Designer, puede ser uno ya existente configurado con las características mínimas especificadas, o crear uno nuevo (virtual) y se le daría el mismo trato que los existentes en la empresa, con lo cual estaría dentro de la red corporativa y estaría sujeto al mismo marco de seguridad que se maneje.

Se concluye que no será necesario adquirir ningún hardware adicional o servicio de infraestructura en nube a la infraestructura de red ya gestionada por la empresa actualmente.

6.1.2 Recursos humanos

Se requiere que cada rol involucrado en la implementación del proyecto cumpla un requisito de tiempo de experiencia mínimo.

Tabla 12

Requisitos de tiempo de experiencia de los roles

Rol	Tiempo mínimo de experiencia (años)	Tiempo ideal de experiencia (años)
Analista de datos	2	4
Arquitecto de datos	1	3
Científico de datos	3	5
Ingeniero de datos	2	4
Ingeniero de machine learning	1	3
Scrum master	1	3
Product owner	3	6

6.2 Validación de factibilidad económica

6.2.1 Costo de RRHH

Se realizó una estimación de horas hombre para cada rol involucrado en el proceso de desarrollo, a continuación, se listan los roles y los subprocesos en los que intervienen.

Tabla 13

Listado de roles y subprocesos en los que intervienen.

Rol abrebiatura	Rol	Subproceso
D_An	Data analyst	Fuentes de datos, Comprensión empresarial, Comprensión de datos, Presentación de datos
D_Ar	Data architect	Comprensión empresarial, Implementación
DS	Data scientist	Comprensión de datos, Preparación de datos, Modelado, Evaluación
DE	Data engineer	Preparación de datos
MLE	Machine learning engineer	Modelado, Implementación
SM	Scrum master	Seguimiento y control
PO	Product owner	Seguimiento y control

Las horas se estimaron en base a un desarrollo consistente en 4 sprints de dos semanas bajo la metodología scrum, en la siguiente figura se detallan las horas por rol en cada sprint. Se estimó un total de 560 horas hombre de desarrollo con los siete roles previamente definidos.

Figura 99

Cuadro de estimación de horas de desarrollo.

Fase	Subproceso	Sprint 1							Sprint 2							Sprint 3							Sprint 4						
		D_An	D_Ar	DS	DE	MLE	SM	PO	D_An	D_Ar	DS	DE	MLE	SM	PO	D_An	D_Ar	DS	DE	MLE	SM	PO	D_An	D_Ar	DS	DE	MLE	SM	PO
Input	Fuentes de datos	40																											
Análisis y desarrollo	Comprensión empresarial	40	20																										
Análisis y desarrollo	Comprensión de datos								40		40																		
Análisis y desarrollo	Preparación de datos										40	40																	
Análisis y desarrollo	Modelado																	40		40									
Análisis y desarrollo	Evaluación																	40											
Análisis y desarrollo	Implementación																							40			40		
Output	Presentación de datos																						20						
Gestión	Seguimiento y control						10	10						10	10						10	10						10	10
Total horas		**120**							**180**							**140**							**120**						
		560																											

Posteriormente se realizó el cálculo de costo por horas de cada rol, esto dividiendo el salario de cada uno, acorde a las estimaciones propuestas por Glassdoor (s.f.) https://www.glassdoor.com/, sobre la cantidad de horas trabajadas por semana, estimadas en 40 horas, dándonos resultado el costo por hora del rol, que multiplicándolo por las horas requeridas nos da el costo total.

Tabla 14

Cuadro resultante de estimación de costos.

Rol abr	Rol	Horas dedicadas	Salario según Glassdoor	Costo por hora	Costo total
D_An	Data analyst	140	S/ 5,000.00	S/ 31.25	S/ 4,375.00
D_Ar	Data architect	60	S/ 9,000.00	S/ 56.25	S/ 3,375.00
DS	Data scientist	160	S/ 6,000.00	S/ 37.50	S/ 6,000.00
DE	Data engineer	40	S/ 8,000.00	S/ 50.00	S/ 2,000.00
MLE	Machine learning engineer	80	S/ 4,750.00	S/ 29.69	S/ 2,375.00
SM	Scrum master	40	S/ 6,000.00	S/ 37.50	S/ 1,500.00
PO	Product owner	40	S/ 5,750.00	S/ 35.94	S/ 1,437.50
Total		**560**			**S/ 21,062.50**

La estimación concluye que el costo por horas hombre en total es de S/ 21,062.50.

6.2.2 Costo de servicio de enriquecimiento de datos de Equifax

Equifax, a través de su Servicio de Información para Evaluación Crediticia (SIEC) provee información relevante de personas naturales y jurídicas a empresas para que puedan realizar una evaluación y análisis crediticio que permita tomar decisiones más adecuadas y precisas. Equifax declara que la información contenida en su base de datos es obtenida a partir tanto de sus fuentes de información como Central de Riesgos (CEPIR), como de fuentes accesibles al público, ambos supuestos exentos de consentimiento para el tratamiento de dicha información conforme a ley.

En la figura siguiente se detallan todas las capas de datos que brinda Equifax en su servicio y que pueden ser consumidos según la necesidad de la empresa.

Figura 100

Capas de datos del servicio Equifax.

		CATEGORÍAS		Obligatorio
		Persona Natural	Persona Jurídica	
1RA DIMENSIÓN: DATO ENRIQUECIDO	Nivel 1 (Capa básica)	- SBS obligatorio (RCC)	- SBS obligatorio (RCC)	SI
	Nivel 2 (Capa de valor agregado)	- SBS opcional (tamaño empresa, Essalud, flag bancarizados) - Sistema No Regulado (Microfinanzas) - Otras deudas (Sicom, Protestos, Sunat)	- SBS opcional (tamaño empresa, Essalud, flag bancarizados) - Sistema No Regulado (Microfinanzas) - Otras deudas (Sicom, Protestos, Sunat)	NO
	Nivel 3 (Capa analítica)	- Scores de Riesgo: Risk Predictor, Score Microfinanzas - Scores de Cobranza: preventiva, reactiva temprana, reactiva tardía - Otros Scores: Sobrendeudamiento, Income Predictor	- Scores de Riesgo: Score Empresas	NO
2DA DIMENSIÓN: CONTACTO	Nivel 1 (Fuentes públicas)	- Datos: dirección, teléfono, e-mail - Fuentes: Páginas Blancas, Sunat, CCL	- Datos: dirección, teléfono, e-mail - Fuentes: Páginas Blancas, Sunat, CCL	NO
	Nivel 2 (Fuentes exclusivas)	- Datos: dirección, teléfono, e-mail - Fuentes crediticias: Sicom, Decomicro	- Datos: dirección, teléfono, e-mail - Fuentes crediticias: Sicom, Decomicro	NO
	Nivel 3 (Fuentes consentidas)	- Datos: dirección, teléfono, e-mail - Fuentes: Base de Datos con Consentimiento	- NA	NO
3RA DIMEN_ SIÓN: GEO	Nivel Geo	- 1: direcciones normalizadas - 2: direcciones normalizadas y georreferenciadas	- 1: direcciones normalizadas - 2: direcciones normalizadas y georreferenciadas	NO

Para el caso de estudio en cuestión, se determinó que las capas recomendadas para obtener variables que ayuden a la evaluación crediticia y a la predicción del abandono son: Nivel 1 y 2, Income predictor, Segmento de riqueza y Estilos de vida, mientras que la capa mínima necesaria es Segmento de riqueza. Se solicitó una cotización a Equifax para tener una proyección de costos del servicio, se muestran los resultados en la siguiente tabla.

Figura 101

Tabla de la propuesta enviada por Equifax

DETALLE DE LA PROPUESTA			
Cantidad de Registros	Datos a Enriquecer	Precio Unitario SOLES	Precio Total Soles
10,000	Nivel 1 y 2- Capa básica y valor agregado	0.424	4,240.00
10,000	Income predictor	0.347	3,470.00
10,000	Segmento de Riqueza	0.347	3,470.00
10,000	Estilos de vida	0.463	4,630.00
		TOTAL	15,810.00

Tomando en consideración el requisito mínimo, entonces se necesitarán enriquecer 20 mil registros para entrenar el modelo y 10 mil registros para predicciones de ventas de un año.

Tabla 15

Necesidad y costo de enriquecimiento de datos por Equifax

Registros	Uso de datos	Data a enriquecer	Precio unitario	Precio total
20,000	Entrenamiento del modelo	Segmento de riqueza	S/. 0.347	S/. 6,950
10,000	Nuevas predicciones			S/. 3,470
Total				**S/. 10,420**

El análisis concluye que el costo de enriquecimiento de datos para una bolsa de 30 mil personas asciende a S/ 10,420.00, siendo 20 mil clientes los mínimos recomendables para realizar el entrenamiento y validación del modelo.

6.2.3 Costo de licencias

Se realizó una cotización de la licencia del software Alteryx Designer, toda la información de costos se encuentra en su página web oficial (https://www.alteryx.com/es-419/products/pricing). El costo por una licencia para un usuario, lo cual es suficiente para el desarrollo del modelo, asciende a US$ 5,195.00, que al tipo de cambio actual es equivalente a S/ 20,060.32.

Figura 102

Captura de pantalla del precio de paquete Designer de Alteryx.

De “Precio de productos”, por Alteryx, 2022.

6.2.4 Costo total del proyecto

Se concluye entonces la estimación final de costos para la realización del proyecto, contemplando el costo de los RRHH, enriquecimiento de datos con Equifax y licencia Alteryx, por un total de S/ 59,778.62.

Tabla 16

Costo total del proyecto.

Tipo de Costo	Importe
Costo de recursos humanos	S/. 21,062.50
Costo de variables Equifax	S/. 10,420.00
Costo de licencia Alteryx Cloud	S/. 20,060.32
Total	**S/. 51,542.82**

6.2.5 Estructura de ganancias por cliente

Las empresas del rubro tienen tres principales fuentes de rentabilidad:

- Cuota de administración del contrato.
- Interés de financiamiento (diferencia de precio, accesorios, trámites).
- Margen a través de facturación de vehículo y seguro

De las tres la única que es obligatoria y se cobra desde el inicio hasta el fin del ciclo de vida del cliente es la cuota de administración de contrato. En promedio, esta cuota corresponde a una tasa del 18% del valor del contrato que el cliente paga todos los meses. Por otro lado, se conoce que el promedio de valor de contrato es de US$ 19,000.00.

Tomando estos datos en consideración podemos estimar el ingreso promedio por cliente que completa su ciclo de vida dentro de la empresa multiplicando el valor de contrato promedio con la tasa de cuota de administración promedio, dándonos como resultado que el ingreso promedio por cliente es de US$ 3,420.00 o S/ 12,996.00.

Este ingreso está repartido de manera proporcional por mes en todo ciclo de vida del contrato, el cual actualmente tiene una duración de 60 meses, es decir, el ingreso mensual por cliente corresponde a un aproximado de US$ 57.00 o S/ 216.60.

6.2.6 Validación de viabilidad económica y rentabilidad mediante VAN y TIR

Para calcular el flujo de caja producto de la inversión, tomamos la hipótesis de que el modelo permitirá reducir un 5% la tasa existente de abandono anual. Esto significa que mes a mes habría una retención de 29 clientes, y que estos se irán acumulando y cada mes pagarán su cuota al igual que el resto de los clientes.

Tabla 17

Ingresos y ganancia neta por mes producto de la inversión

Mes	Clientes retenidos por mes	Clientes retenidos acumulado	Ingreso producto de la inversión (S/ 216.60 por cliente por mes)	Ganancia neta sobre ingreso (43.6% del ingreso bruto)	Ganancia neta por bimestre
1	29	29	S/ 6,281.40	S/ 2,738.69	S/ 8,216.07
2	29	58	S/ 12,562.80	S/ 5,477.38	
3	29	87	S/ 18,844.20	S/ 8,216.07	S/ 19,170.83
4	29	116	S/ 25,125.60	S/ 10,954.76	
5	29	145	S/ 31,407.00	S/ 13,693.45	S/ 30,125.59
6	29	174	S/ 37,688.40	S/ 16,432.14	
7	29	203	S/ 43,969.80	S/ 19,170.83	S/ 41,080.36
8	29	232	S/ 50,251.20	S/ 21,909.52	
9	29	261	S/ 56,532.60	S/ 24,648.21	S/ 52,035.12
10	29	290	S/ 62,814.00	S/ 27,386.90	
11	29	319	S/ 69,095.40	S/ 30,125.59	S/ 62,989.88
12	29	348	S/ 75,376.80	S/ 32,864.28	

Tomando una tasa de oportunidad anual de 10%, que corresponde a una tasa de oportunidad bimestral de 1.6%, y considerando que la inversión inicial es de S/ 51,542.82, se tienen los siguientes datos para el cálculo de la VAN y la TIR.

Tabla 18

Flujo de caja de los próximos seis bimestres

Periodo	Flujo de Caja
2023 – Bimestre 2	S/. 8,216
2023 – Bimestre 3	S/. 19,170
2023 – Bimestre 4	S/. 30,125
2023 – Bimestre 5	S/. 41,080
2023 – Bimestre 6	S/. 52,035
2024 – Bimestre 1	S/. 62,989

Tabla 19

Parámetros para cálculo de VAN y TIR

Parámetro	Valor
Número de periodos (bimestres)	6
Tasa de oportunidad anual	10 %
Tasa de oportunidad bimestral	1.6 %
Inversión inicial	S/. 51,542.82

Calculamos la VAN con la siguiente fórmula:

$$VAN = \sum_{n}^{6} \frac{FC_n}{(1+i)^n}$$

Reemplazando en la fórmula obtenemos:

$$VAN = -I + \frac{FC_1}{(1+i)^1} + \frac{FC_2}{(1+i)^2} + \frac{FC_3}{(1+i)^3} + \frac{FC_4}{(1+i)^4} + \frac{FC_5}{(1+i)^5} + \frac{FC_6}{(1+i)^6}$$

$$VAN = -S/.51{,}542.82 + \frac{S/.8{,}216}{(1+0.016)^1} + \frac{S/.19{,}170}{(1+0.016)^2} + \frac{S/.30{,}125}{(1+0.016)^3} + \frac{S/.41{,}080}{(1+0.016)^4} + \frac{S/.52{,}035}{(1+0.016)^5} + \frac{S/.62{,}989}{(1+0.016)^6} VAN = S/147{,}725.82$$

Para resolver la TIR primero llevamos la VAN a cero mediante prueba y error variando la tasa de oportunidad:

1.6% es a VAN = S/ 147,725.82

43% es a VAN = - S/ 227.85

Interpolamos la VAN menor a cero en la siguiente fórmula para hallar la TIR exacta:

$$\frac{i_1 - i_2}{i_1 - TIR} = \frac{VAN > 0 - VAN < 0}{VAN > 0 - (0)}$$

$$\frac{0.016 - 0.43}{0.016 - TIR} = \frac{147{,}725.82 - (-227.85)}{147{,}725.82 - (0)}$$

$$\frac{-0.414}{0.016 - TIR} = \frac{147{,}953.67}{147{,}725.82}$$

$$TIR = 42.93\%$$

Se comprueba que la tasa de oportunidad bimestral de 42.93% hace cero a la VAN o está próxima. Por lo tanto, la inversión es rentable ya que el costo de oportunidad del dinero de 1.6% bimestral es menor a la TIR de 42.93% bimestral.

7 CONCLUSIONES

La propuesta de implementación de un modelo de análisis predictivo de abandono de clientes en empresas administradoras de fondos colectivos se lleva a cabo con el fin de reducir el porcentaje anual de abandono de clientes, brindando los lineamientos y pasos necesarios para el desarrollo e implementación de una solución que permita predecir los clientes con alta probabilidad de abandono:

- Se concluye que el modelo de análisis predictivo propuesto integra los mejores componentes para lograr una predicción precisa de clientes con alta probabilidad de abandono. El modelo propuesto sugiere la herramienta más adecuada para el rubro de estudio, los algoritmos más apropiados para este tipo de proyecto, y las variables con mayor poder predictivo según la literatura, con el valor agregado de estar adaptadas al contexto de los fondos colectivos.
- Se concluye que la propuesta es viable económicamente. El VAN indica que se recupera la inversión inicial de S/. 51,542.82 y adicionalmente se genera una ganancia neta de S/. 147,725.82.
- Se concluye que el dashboard propuesto con los resultados del modelo de análisis predictivo permitirá conocer el listado de clientes con alta probabilidad de abandono y permitirá a los actores de negocio involucrados tomar decisiones basadas en datos precisos.
- Al implementar la propuesta será viable para la organización reducir la tasa de abandono anual de clientes en un 5%.

8 RECOMENDACIONES

- Se recomienda a la empresa administradora de fondos colectivos desarrollar e implementar un modelo de análisis predictivo de abandono de clientes basándose en los lineamientos que detalle el presente proyecto, con la finalidad de obtener un beneficio económico.
- Se recomienda continuar la investigación acerca de la implementación de distintas soluciones tecnológicas actuales, aplicadas al rubro de los fondos colectivos, ya que es un rubro que está tomando cada vez más importancia en Perú y Sudamérica para la compra de vehículos bajo esa modalidad.
- Se recomienda consultar el foro oficial de la comunidad Alteryx ante cualquier duda sobre la herramienta.

9 BIBLIOGRAFÍA

Ahmad, A.K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data, 6*, 28. doi:10.1186/s40537-019-0191-6.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research, 94*, 290–301.

Athanassopoulos, A. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research*, *47*(3), 191–207. doi:10.1016/S0148-2963(98)00060-5.

Beaulac, C., & Rosenthal, J. S. (2019). Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education, 60,* 1048-1064. doi:10.1007/s11162-019-09546-y.

Burgos, C., Campanario, M., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. (2019). Data mining for modeling students performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering, 66*. doi:10.1016/j.compeleceng.2017.03.005.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., (2000). CRISP-DM 1.0: step-by-step data mining guide. *SPSS inc.*, *9*, 13.

Cheng, L. C., Wu, C., & Chen, C. (2019). Behavior analysis of customer churn for a customer relationship system: An empirical case study. *Journal of Global Information Management, 27*(1), 111-127. doi:10.4018/JGIM.2019010106.

Chung, J., & Lee, S. (2019). Dropout early warning systems for high school students using machine laearning. *Childen and Youth Services Review, 96*, 346-353. doi:10.1016/j.childyouth.2018.11.030.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760-772. doi:10.1016/j.ejor.2018.02.009.

De Caigny, A., Coussement, K., De Bock, K. W., & Lessmannd, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. *International Journal of Forecasting, 36*(4), 1563-1578. doi:10.1016/j.ijforecast.2019.03.029.

Gartner (2019). *Alteryx Reviews*. Gartner. Recuperado el 10 de diciembre de 2022, de https://www.gartner.com/reviews/market/multipersona-data-science-and-machine-learning-platforms/vendor/alteryx

Höppner, S., Stripling, E., Baesens, B., Broucke, S., & Verdonck, T. (2020). Profit driven decision trees for churn prediction. *European Journal of Operational Research, 284* (3), 920-933. doi:10.1016/j.ejor.2018.11.072.

Idris, A., & Khan, A. (2012). Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers. *15th International Multitopic Conference*, INMIC 2012. 23-27. doi:10.1109/INMIC.2012.6511498.

Jamjoom, A. A. (2021). The use of knowledge extraction in predicting customer churn in B2B. *Journal of Big Data, 8*, 110. doi:10.1186/s40537-021-00500-3.

Kassem, E., Hussein, S., Abdelrahman, A., & Alsheref, F. (2020). Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content. *International Journal of Advanced Computer Science and Applications, 11*(5). doi:10.14569/IJACSA.2020.0110567.

Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., & Pentland, A. (2018). Behavioral attributes and financial churn prediction. *EPJ Data Science*, *7*, 41. doi:10.1140/epjds/s13688-018-0165-5.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation, 2*(1), 10. doi:10.1186/s40854-016- 0029-6.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing (Vienna/New York), 104*, 271–294. doi:10.1007/s00607-021-00908-y".

Lima Lemos, R.A., Silva, T.C., & Tabak, B.M. (2022). Propension to customer churn in a financial institution: a machine learning approach. *Neural Computing and Applications, 34*, pages11751–11768. doi:10.1007/s00521-022-07067-x.

Liu, R., Ali, S., Bilal, S.F., Sakhawat, Z., Imran, A., Almuhaimeed, A., Alzahrani, A., & Sun, G. (2022). An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms. *Applied Sciences (Switzerland), 12*, 9355. doi:10.3390/app12189355.

Martínez, F., Contreras, L., Ferri, C., Hernandez, J., Kull, M., Lachiche, N. Ramirez, M., & Flach, P. (2019). CRISP-DM twenty years later: from data mining processes to data science trajectories. *IEEE Trans Knowl Data Eng*. https://doi.org/10.1109/TKDE.2019.2962680.

Munkhdalai, L., Munkhdalai, T., Namsrai, O.E., Lee, J.Y., & Ryu, K.H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability (Switzerland), 11*(3), 699. doi:10.3390/su11030699.

Nadali A, Kakhky E.N., & Nosratabadi H.E. (2011). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. In: 2011 3rd International Conference on Electronics Computer Technology. *New York: IEEE, 6*, 161–5

Pustokhina, I., Pustokhin, D., RH, A., Jayasankar, T., Jeyalakshmi, C., García, V., & Shankar, K. (2021). Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. *Information Processing and Management, 58*(6), 102706. doi:10.1016/j.ipm.2021.102706.

Rajeswari, M., & Devi, T. (2015). Design of modified ripper algorithm to predict customer churn. *Int J Eng Technol, 4*(2), 408.

Rokach, L., & Maimon, O. (2005). Data Mining and Knowledge Discovery Handbook. *Springer*, 321–352. doi: 10.1007/0-387-25465-X_15

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Mach. Intell. 1* (5) 206.

Scherer, A., Wünderlich, N.V., & Von Wangenheim, F. (2015). The value of self-service: Long-term effects of technology-based self-service usage on customer retention. *MIS Quart*, *39* (1).

Superintendencia del Mercado de Valores. (2022a). *Sistema De Fondos Colectivos*. SMV. Recuperado el 10 de diciembre de 2022, de https://www.smv.gob.pe/Frm_VerArticulo.aspx?data=A01ACE41D475DBBFE553A0B77266A32748078C45EA48C40E6D141A10881827675A42524295E1665D3B1AB45880

Superintendencia del Mercado de Valores. (2022b). *Finalidad y funciones.* SMV. Recuperado el 10 de diciembre de 2022, de https://www.smv.gob.pe/Frm_VerArticulo?data=17B15B848FCE8F37FA86E13166C6752043C6DCB32142B823F43909D41274C8008858C8

Tamaddoni, A., Stakhovych, S., & Ewing, M. (2017).The impact of personalized incentives on the proftability of customer retention campaigns. *J Mark Manag*, *33*, 327–47. https://doi.org/10.1080/0267257x.2017.1295094.

Tian, A. Q., Chu, S. C., Pan, J. S., Cui, H., & Zheng, W. M. (2020). A Compact Pigeon-Inspired Optimization for Maximum Short-Term generation mode in cascade hydroelectric power station. *Sustainability*, *12*, 767.

Ullah, I., Raza, B., Malik, A. K., Imran M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access, 7*, 60134-60149. doi:10.1109/ACCESS.2019.2914999.

Usman-Hamza, F.E., Balogun, A.O., Capretz, L.F., Mojeed, H.A., Mahamad, S., Salihu, S.A., Akintola, A.G., Basri, S., Amosa, R.T., & Salahdeen, N.K. (2022). Intelligent Decision Forest Models for Customer Churn Prediction. *Applied Sciences (Switzerland), 12*(16), 8270. doi:10.3390/app12168270.

Van den Poel, D., & Larivière, B (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 157(1), 196-217. doi: 10.1016/S0377-2217(03)00069-9

Vo, N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. *Knowledge-Based Systems, 212*. doi:10.1016/j.knosys.2020.106586

Wang, C., Han, D., Liu, Q., & Luo, S. (2019). A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM. *IEEE Access, 7*, 2161-2168. doi: 10.1109/ACCESS.2018.2887138.

Xiao, J., Xiao, Y., Huang, A., Liu, D., & Wang, S. (2014). Feature-selection-based dynamic transfer ensemble model for customer churn prediction. *Knowledge and Information Systems*. 43. 29-51. 10.1007/s10115-013-0722-y.

Yeh, J., & Chen, C. (2020). A machine learning approach to predict the success of crowdfunding fintech project. *Journal of Enterprise Information Management*. doi:10.1108/JEIM-01-2019-0017.