Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Analítica de big data, calidad de datos y machine learning aplicados a la gestión de riesgos

Caso de una entidad financiera en Perú

AUTOR: ENMA RAYZA GOMEZ CCAPA
MENTOR: RODRIGO DEL ROSSO

2022

Resumen

Hoy en día, las áreas encargadas de brindar soporte analítico dentro de las entidades financieras se enfrentan a un desafío constante. Dicho desafío, consiste en el tratamiento de grandes volúmenes de datos e implementación de metodologías analíticas, pues los procesos y métodos tradicionales resultan insuficientes. Esta situación, ha dado lugar a que las tecnologías de *big data* y el uso de modelos de aprendizaje automático o *machine learning* surjan como alternativa de solución para las organizaciones y el sector bancario no es la excepción.

Este trabajo se desarrolla en el área de gestión de riesgo de crédito en una entidad financiera ubicada en Perú. Se va a construir un repositorio de datos que reúna las variables de riesgo más importantes. Seguido de esto, se establecerán expectativas que verifiquen la calidad de los datos que forman parte del repositorio. Finalmente, se va a plantear un modelo de árbol de decisión como alternativa frente al modelo logit, aplicado al análisis de créditos que fueron castigados por esta entidad durante el 2020.

Los datos utilizados corresponden al periodo diciembre 2020 hasta diciembre 2021 y un *dataset* con clientes con situación incobrable (castigados). Para verificar la eficacia de los modelos planteados, se van a realizar análisis de brechas y uso de métricas para evidenciar las mejoras que traen las propuestas planteadas en esta investigación.

Palabras clave: gestión de riesgos**,** *big data*, arquitectura de datos, análisis de brechas, calidad de datos, árbol de decisión, modelo logístico, crédito castigado

Contenido

Introducción

Con el paso del tiempo, la cantidad de información que ingresa a las áreas que se encargan de brindar soporte analítico en las entidades financieras se ha incrementado exponencialmente y así los sistemas tradicionales de administración y herramientas para el análisis resultan insuficientes. Esta situación ha dado lugar al ingreso de las tecnologías de big data y el uso de modelos de aprendizaje automático como alternativa para el logro de objetivos dentro de las organizaciones dedicadas al rubro financiero. La presente investigación se desarrolla en el contexto de una entidad financiera ubicada en Perú, poniendo énfasis en el área de seguimiento de portafolio de créditos, cuyo objetivo principal es velar por mantener niveles adecuados de riesgo y anticipar eventos de pérdida que representen peligro para los intereses de la entidad. Sin embargo, existen inconvenientes para llevar a cabo estos objetivos de manera óptima. Para ser más preciso, el manejo poco eficiente de los datos se evidencia en la utilización de tiempos cuya principal consecuencia resulta en una mayor asignación de recursos al trabajo operativo, descuidando el verdadero enfoque del área, que es controlar, analizar y medir el riesgo haciendo uso de métricas y modelos Frente a esta problemática, surge la iniciativa de seleccionar las más adecuadas técnicas de big data y machine learning a fin de revertir la ausencia de procesos analíticos relacionados con la gestión de datos y modelos predictivos.

1. Big data y analítica aplicada a la gestión de riesgos en la entidad

En la actualidad, el *big data* se ha convertido en una herramienta usada por las entidades financieras para la detección de riesgos. De acuerdo con un artículo publicado por McKinsey & Company (2016), existen seis iniciativas que los bancos deben tener en cuenta para mantenerse a la vanguardia en lo que respecta a gestión de riesgos: regulación, expectativas del cliente, **tecnología y analítica avanzada**, surgimiento de nuevos riesgos, el riesgo como mitigador de sesgos y ahorro de costos. La iniciativa de tecnología y analítica avanzada es capaz de combinar diferentes recursos de información y organizar grandes cantidades de datos que permitan tener conocimiento integral del cliente. En este sentido, la entidad podrá identificar operaciones que corren peligro de caer en incumplimiento y tomar acciones preventivas que minimicen posibles pérdidas que afecten la rentabilidad del negocio.

El objetivo de este apartado es evidenciar el alcance que tiene la creación de un repositorio de datos en el área de gestión de riesgos de una entidad financiera ubicada en Perú. Para hacerlo, se parte del análisis de la arquitectura actual hasta el establecimiento de una arquitectura objetivo de los datos. En consecuencia, se tendrá como resultado, la producción de informes relacionados con variables de riesgo de crédito.

1.1. Manejo actual de los datos

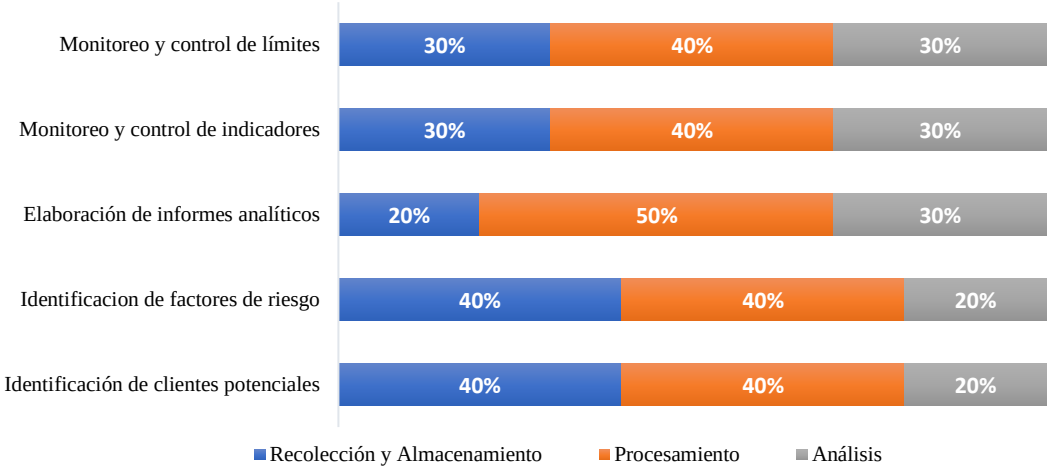
Seguimiento de portafolio forma parte de las áreas encargadas de brindar soporte analítico a unidades clave del negocio como finanzas y productos. Pardo Ramírez (2018), asegura que uno de los principios con los que debe estar comprometida la gestión de riesgos es que forma parte de la toma de decisiones y debe basarse en la mejor información disponible.

Entonces, es importante poner foco a como se vienen administrando las diferentes fuentes de datos. Actualmente, el analista consulta un servidor asignado al equipo. Esto implica tareas de recolección, almacenamiento y transformación de datos cuyo tiempo de ejecución consume aproximadamente el 30% de lo que se dispone para cumplir con los entregables.

El procesamiento es la etapa donde los datos se convierten en información relevante para la toma de decisiones. Sin embargo, en esta área concentra una

carga operativa que trae como consecuencia menor cantidad de tiempo para analizar los resultados y elaborar los informes de seguimiento y gestión de riesgo (Gráfico 1).

Gráfico 1: Distribución del tiempo empleado en el procesamiento de datos



Fuente: Elaboración propia

La mayor parte de los recursos se dedica a hacer consultas, hojas de cálculo y tablas dinámicas, que en algunos casos son formatos que repiten los mismos datos; lo cual resulta poco eficiente dado el crecimiento del volumen de operaciones que ingresan cada día. Este crecimiento está en línea con el crecimiento en conjunto que ha tenido el sistema financiero peruano en los últimos años. Para ilustrar este punto, el cuadro 1 muestra la estructura del sistema financiero desde el 2019 por número de clientes y créditos. Se puede observar que solo la banca múltiple supera los 4 millones de deudores, lo cual evidencia la necesidad de adoptar otras metodologías para el control y seguimiento de estos.

Tabla 1: Evolución de estructura del Sistema Financiero Peruano

Dic-19/Dic-21

	2019		2020		2021	
	Deudores	Créditos (S/M)	Deudores	Créditos (S/M)	Deudores	Créditos (S/M)
Banca Múltiple	4,563,638	286,086	4,313,802	326,022	4,236,611	346,040
Empresas Financieras	2,657,628	13,840	2,423,408	13,341	2,136,527	12,038
Cajas Municipales	1,847,739	23,577	1,761,338	26,455	1,883,009	28,099
Cajas Rurales	582,891	2,400	502,590	2,394	430,252	2,097
Entidades de desarrollo MYPE	267,359	2,638	217,122	2,550	161,458	2,718

Fuente: Superintendencia de Banca y Seguros[1]

Por lo expuesto, para cumplir con el objetivo de este apartado, se va a proponer cambio en la gestión de los datos. En primer lugar, la propuesta consiste en revisar procesos actuales, seguido de la elección de un modelo a seguir y finalmente proponer una arquitectura de datos que permita aplicar analítica de *big data* en el área de seguimiento de portafolio.

1.2.Definiciones sobre arquitectura de datos y analítica de big data

1.2.1. Arquitectura de datos

Las empresas poseen varios orígenes de datos que utilizan para tomar decisiones. En este caso, dado el contexto donde se desarrolla la investigación, dichas decisiones están vinculadas con la gestión de riesgo de crédito. La diversidad de datos que existe se relaciona directamente con la problemática que afronta esta área del banco y en base a ello, surge el desafío de construir un esquema que gire en torno a las principales variables de riesgo.

[1] Información obtenida de https://www.sbs.gob.pe/estadisticas-y-publicaciones/estadisticas-/sistema-financiero_

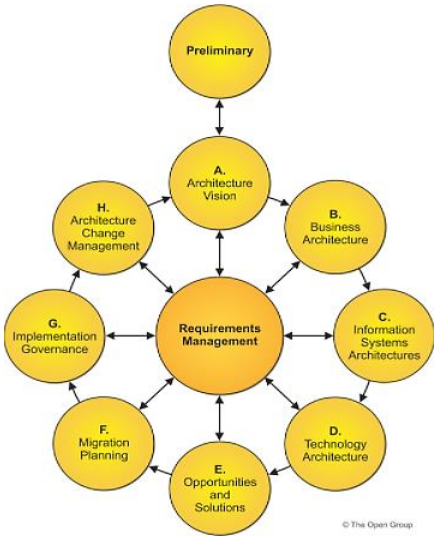
En esa línea, se tomará como referencia los pasos a seguir para la construcción de una arquitectura de datos, sin embargo, no es posible hablar de arquitectura de datos sin mencionar el concepto de arquitectura empresarial.

De acuerdo con *The system and software consortium* (2005), "Arquitectura empresarial relaciona la misión, metas y objetivos de la organización con el desarrollo de procesos e infraestructura tecnológica requeridos", compuesta por cuatro capas o dimensiones: arquitectura de negocio, arquitectura de aplicaciones, arquitectura de datos y arquitectura de infraestructura; en la práctica, se suele tomar un marco de referencia o *framework* como punto de partida.

Framework es el modelo para el desarrollo de la arquitectura empresarial. Dentro de los más conocidos están los propuestos por Zachman: DoDAF[2], FEAF[3], TEAF[4] y TOGAF. Se va a tomar como referencia el *framework* TOGAF, uno de los más reconocidos y usados a nivel mundial cuyo objetivo es aumentar la eficiencia de los procesos de negocio.

TOGAF utiliza el método *Arquitecture Development Method* (ADM), el cual consta de una serie de pasos interactivos y cíclicos para el desarrollo de una arquitectura empresarial completa. Como muestra la figura 1, la arquitectura de datos forma parte del ciclo de desarrollo (Fase C).

Figura 1: Método ADM



Fuente: The Open Group

[2] Department of Defense Architecture Framework
[3] Federal Enterprise Architecture Framework
[4] Treasury Enterprise Architecture Framework

Con respecto a la Fase C correspondiente a la arquitectura de datos, Arango, Londoño y Zapata (2010) la definen como aquella que "…describe los activos lógicos y físicos de los datos como un activo de la empresa, y la administración de los recursos de la información". Asimismo, según Zachman (1987), el éxito del negocio y los costos que ello conlleva dependen cada vez más de sus sistemas de información, los cuales requieren de un enfoque y disciplina para la gestión de estos. TOGAF indica que dentro de la fase C se desarrollan los siguientes pasos:

i. Seleccionar modelos de referencia
ii. Desarrollar la descripción de la arquitectura de datos de línea base
iii. Desarrollar la descripción de la arquitectura de datos de destino
iv. Realizar un análisis de brechas
v. Definir los componentes candidatos que conforman el plan de itinerario
vi. Resolver los impactos al panorama de arquitectura
vii. Conducir una revisión formal con los interesados
viii. Finalizar la arquitectura de datos
ix. Crear el documento de definición de arquitectura

Cabe mencionar que, en este trabajo, no se van a desarrollar todos los pasos de la fase C, ya que este implica que los resultados queden documentados y difundidos a toda la organización, lo cual no está dentro de los objetivos de este apartado.

1.2.2. **Analítica de Big Data**

La analítica de Big Data (*Big Data analytics*) es el uso de técnicas analíticas aplicadas a conjuntos de grandes volúmenes de datos. En otras palabras, es la conjugación de analítica avanzada y *big data* para descubrir patrones ocultos, correlaciones e información útil (Russon,2011). Esto trae consigo beneficios económicos y hace que las organizaciones sean más competitivas y eficientes. La analítica es usada para identificar cambios y cómo reaccionar ante estos, por ejemplo, en el área de seguimiento de portafolio, será una herramienta potente para descubrir segmentos de clientes, comprender el comportamiento de la

tendencia del indicador de mora, etcétera. Está compuesta por un conjunto de técnicas que incluyen analítica predictiva, análisis estadístico, minería de datos, programación SQL (*quering and reporting*) y visualización de datos (*dashboards*).

La analítica predictiva mira los datos desde una perspectiva utilizada para construir modelos e incluye análisis estadístico avanzado, mientras que minería de datos consiste en encontrar patrones y estructuras ocultas. Estas dos técnicas son las más utilizadas en el análisis de datos y están adaptadas para ser aplicadas en *big data* (Joyanes,2013).

Por otro lado, la programación SQL enfocada en consultas y reportería utiliza procesos OLAP para el análisis de información sobre determinados patrones de interés. "El análisis OLAP realiza los reportes sumarios que los ejecutivos necesitan para la toma de decisiones, cálculos complejos, enfoques de detalles operativos y consultas no programadas" (Reinosa, Maldonado, Muñoz, Damiano y Abrutsky, 2012). Las herramientas de visualización sirven para que los gerentes participen activamente sobre los reportes generados, los más representativos son los *dashboard* y *balanced score card*, que son interfaces entre el usuario e informes.

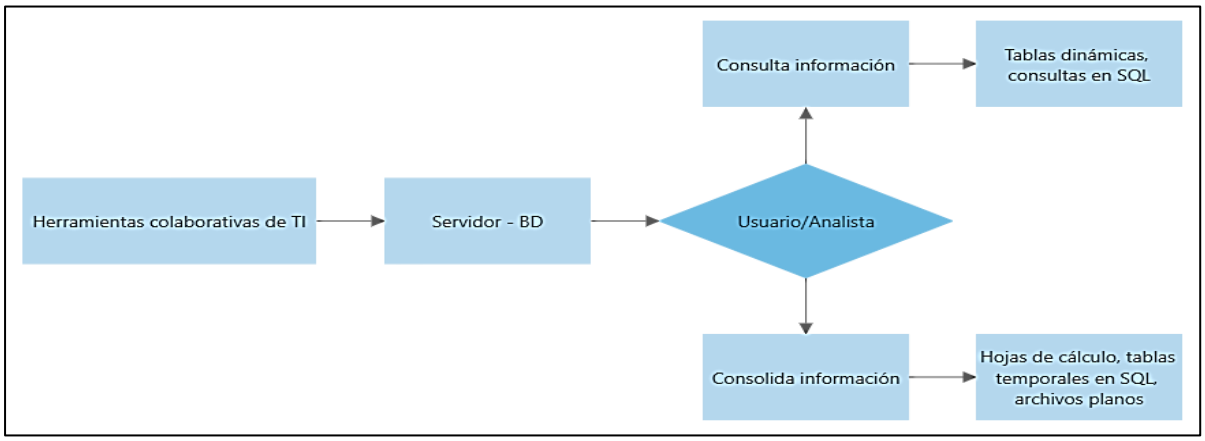
Como parte del desarrollo de este primer apartado, se realizará la propuesta de un repositorio de datos bajo el enfoque de arquitectura de datos, seguido del uso de analítica de big data: *reporting* con análisis *OLAP*, minería de datos y la elaboración de un *dashboard*. Las herramientas tecnológicas de apoyo serán *SQL server*, R-*Studio, Python y Power BI.*

1.3. Creación del repositorio de datos

1.3.1. Revisión de formatos y nueva arquitectura

Desarrollado el punto 1.2.1, se van a seguir los pasos i,ii, iii y v de la Fase C (Arquitectura de datos). El modelo de referencia será el *framework* TOGAF, seguido de esto, se debe describir la arquitectura base, que está representada de la siguiente manera:

Figura 2: Diagrama de gestión de datos actual – Arquitectura base



Fuente: Elaboración propia

Los orígenes de datos son administrados por el área de TI con ayuda de herramientas colaborativas. Estas recogen información en línea de las transacciones que ingresan diariamente al banco como pagos, desembolsos, cancelaciones, etc., para luego ser distribuidas a las diferentes unidades del negocio.

Cada unidad tiene habilitado un servidor que concentra en tablas estructuradas los datos como insumo principal para realizar los análisis e informes que forman parte de las actividades diarias. Como se mencionó antes, los analistas toman lo que necesitan del servidor para llevarlo a diferentes formatos (hojas de trabajo). El siguiente cuadro muestra en qué consiste cada proceso y los formatos que asociados a cada uno de estos:

Tabla 2: Formatos relacionados con los procesos

Proceso	Nombre de Formato	Descripción
Control de límites	Semáforo para el control de límites	Registra la evolución y desviación de los principales indicadores de riesgo sobre los límites y apetito fijados por el directorio.
Monitoreo y control de indicadores de riesgo	Cuadro evolutivo de indicadores de riesgo	Muestra la evolución de los indicadores de riesgo.
Identificación de factores de riesgo	Base de datos con información de cierre mensual	Procesa los atributos contenidos en las tablas de cierre mensual a fin de encontrar patrones que indiquen alerta sobre el portafolio.
Identificación de clientes potenciales	Base de datos de clientes	Filtra clientes que cumplen con determinadas condiciones de elevado riesgo.
Elaboración de informes analíticos	Base de datos con información de cierre mensual	Consolida información procesada de diferentes tablas en forma de cuadros y gráficos.

Fuente: Elaboración propia

Arquitectura objetivo (To-Be)

Esta parte de la fase C sirve para realizar un análisis que parte de los formatos descritos en el cuadro 2. De cada formato se seleccionan los datos más relevantes que serán almacenados en un repositorio de datos (Tabla SQL) que llevará el nombre de "Repositorio_Riesgos".

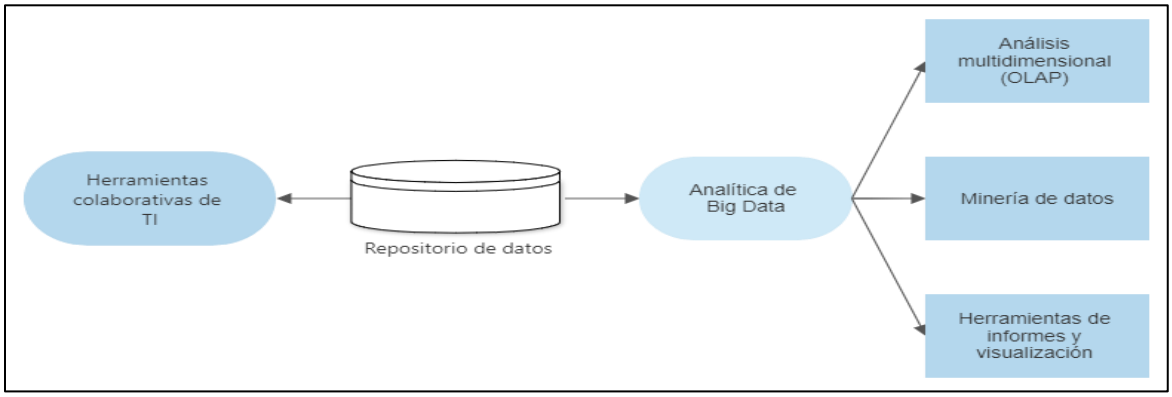
Tabla 3: Datos utilizados en los procesos de seguimiento de portafolio

Campos para almacenar en "Repositorio_Riesgos"	
1 FECHA_CIERRE	13 SALDO_MORA
2 ID_CLIENTE	14 SALDO_JUDICIAL
3 ID_OPERACION	15 SALDO_REFINANCIADO
4 ACTIVIDAD_ECONÓMICA	16 SALDO_PROVISIÓN_GENÉRICA
5 PRODUCTO	17 SALDO_PROVISIÓN_ESPECÍFICA
6 TIPO_EXPOSICIÓN	18 GASTO DE PROVISIÓN
7 SEGMENTO_SBS	19 SALDO_APR
8 CLASIFICACIÓN	20 INDICADOR_RIESGO_CAMBIARIO
9 BANCA_INTERNA	21 INDICADOR_SOBRE_ENDEUDAMIENTO
10 DÍAS_MORA	22 INDICADOR_REPROGRAMACIÓN
11 SALDO_CAPITAL	23 NIVEL_RIESGO_INTERNO
12 SALDO_PROVISIONABLE	24 COD_NIVEL_RIESGO_INTERNO

Fuente: Elaboración propia

La propuesta desde un inicio fue encontrar la mejor arquitectura que se adecue a las necesidades de información del área. Por ese motivo, se construyó un repositorio de datos en lugar de hacer consultas por separado. La figura 3 muestra cómo el repositorio de datos toma el lugar que el servidor tenía en la arquitectura base (Figura 2), seguido de las aplicaciones de analítica de *big data*, que en conjunto darán lugar al cambio en la gestión de datos con miras a mejorar la calidad de los informes.

Figura 3: Diagrama de gestión de datos propuesto – Arquitectura *to be*



Fuente: Elaboración propia

1.3.2. Aplicaciones OLAP y data mining con herramientas de visualización

Se crearon vistas en cubos *OLAP* con el repositorio de datos. Las herramientas de apoyo fueron *SQL Server* y *Power BI* para minería y visualización de datos respectivamente.

Figura 4: Cubo *OLAP* para el índice de vencidos por producto

	CARTERA	PRODUCTO	CAPITAL	RATIO_VENCIDO	INDICADOR_CARTERA	INDICADOR_PRODUCTO
1	CONSUMO	CONVENIO	1362046657.75999	0.00885428405208503	0	0
2	NULL	CONVENIO	1362046657.75999	0.00885428405208503	1	0
3	HIPOTECARIO	HIPOTECARIO	514648545.900001	0.0732433767282503	0	0
4	NULL	HIPOTECARIO	514648545.900001	0.0732433767282503	1	0
5	HIPOTECARIO	MI VIVIENDA	235729322.37	0.0427429575527525	0	0
6	NULL	MI VIVIENDA	235729322.37	0.0427429575527525	1	0
7	CONSUMO	PRESTAMO PERSONAL	27788834.26	0.0949149440858913	0	0
8	NULL	PRESTAMO PERSONAL	27788834.26	0.0949149440858913	1	0
9	CONSUMO	PRESTAMO VEHICUL...	1118298.28	0.24484238677359	0	0
10	NULL	PRESTAMO VEHICUL...	1118298.28	0.24484238677359	1	0
11	CONSUMO	TARJETA DE CREDITO	13537470.66	0.0636909502265913	0	0
12	NULL	TARJETA DE CREDITO	13537470.66	0.0636909502265913	1	0
13	NULL	NULL	2154869129.22999	0.0295163681762557	1	1
14	CONSUMO	NULL	1404491260.95999	0.011273508978032	0	1
15	HIPOTECARIO	NULL	750377868.270001	0.0636617466612319	0	1

Fuente: Elaboración propia

Figura 5: Cubo *OLAP* para el índice de vencidos de la banca empresa

	CARTERA	EQUIPO	CAPITAL	RATIO_VENCIDO	INDICADOR_CARTERA	INDICADOR_EQUIPO
1	COMERCIAL	BANCA EMPRESA	593557792.16	0.00210747613210139	0	0
2	NULL	BANCA EMPRESA	593557792.16	0.00210747613210139	1	1
3	COMERCIAL	BANCA NEGOCIOS	1378326.09	0	0	0
4	NULL	BANCA NEGOCIOS	1378326.09	0	1	1
5	COMERCIAL	INSTITUCIONAL	236579968.75	0	0	0
6	NULL	INSTITUCIONAL	236579968.75	0	1	1
7	COMERCIAL	PROVINCIAS	65684618.62	0.0134037941986607	0	0
8	NULL	PROVINCIAS	65684618.62	0.0134037941986607	1	1
9	COMERCIAL	PROYECTOS CO...	1666771.62	0.0419658513264103	0	0
10	NULL	PROYECTOS CO...	1666771.62	0.0419658513264103	1	1
11	COMERCIAL	RECUPERACION...	139774362.76	0.48561199721974	0	0
12	NULL	RECUPERACION...	139774362.76	0.48561199721974	1	1
13	NULL	NULL	1038641840	0.0674702137360459	1	1
14	COMERCIAL	NULL	1038641840	0.0674702137360459	0	0

Fuente: Elaboración propia

Los cubos[5] están diseñados para el índice de vencido por producto y/o unidad de negocio (Figuras 4 y 5). Las celdas cuyo valor es *NULL* agrupa totales de cada categoría que puede tomar una variable, campo o atributo. Por ejemplo, el cubo de vencidos para la banca personal (Fig. 4) toma valor *NULL* para cartera y producto en la fila 13. Es decir, esta fila muestra el índice de vencimiento de toda la banca personal sin discriminar el tipo de cartera (consumo o hipotecario) o el producto. Por otro lado, las filas 14 y 15 toman el valor de consumo e hipotecario respectivamente en la columna cartera y *NULL* para producto, siendo los valores calculados para esas filas un subtotal por tipo de cartera. Otra característica importante del cubo es que genera una columna de indicador por cada dimensión o variable que se desea analizar, en este caso se generó las columnas "INDICADOR_CARTERA" e "INDICADOR_PRODUCTO", donde 1 indica que se incluyen todas las categorías de la dimensión, esto a la vez coincide con el valor NULL.

La misma lectura se le puede dar al cubo para la banca empresa (Fig.5), la diferencia es que este segmento del banco no cuenta con otros tipos de cartera. Sin embargo, la organización está en pleno crecimiento y no descarta entrar a nuevos mercados empresariales.

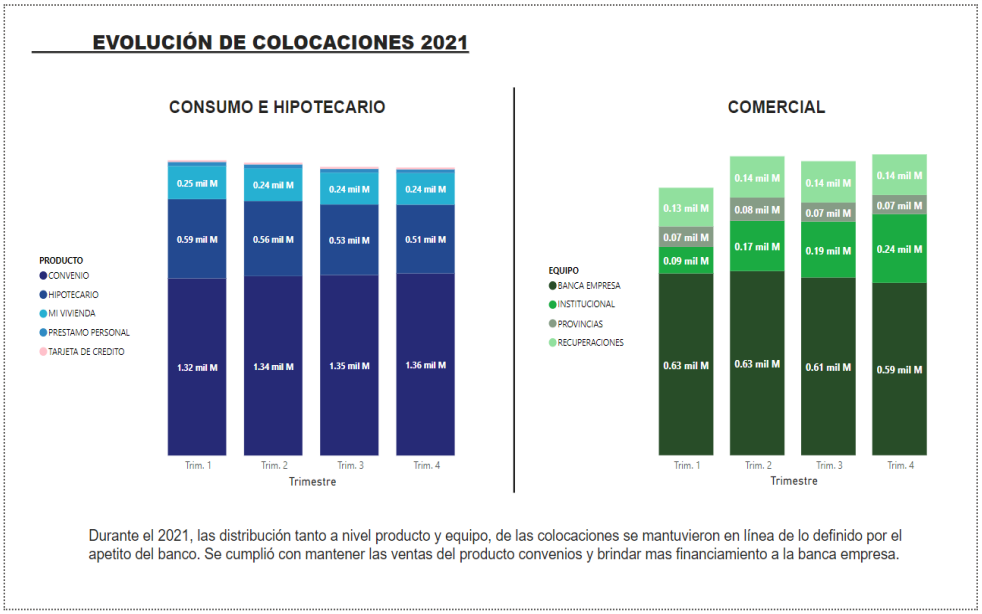
1.3.3. Data Mining y visualización de informes

Los cubos se diseñarán a demanda del analista con los datos del repositorio. Esto es solo una parte, ya que la información debe ser mostrada a los ejecutivos para la toma de decisiones mediante informes. Para esto será necesario utilizar herramientas como Power BI, la cual está habilitada para conectarse con diferentes aplicaciones como SQL, Oracle, Python, R-Studio entre otras. En vista de que los datos están en SQL, se establecerá conexión directa con el repositorio de datos.

Las figuras que se muestran a continuación pertenecen a un informe trimestral del área de gestión de riesgos, donde se muestra la evolución de las colocaciones y el indicador de vencidos.

[5] Para ver el código ir al apéndice 1

Figura 6: Evolución de colocaciones por tipo de cartera

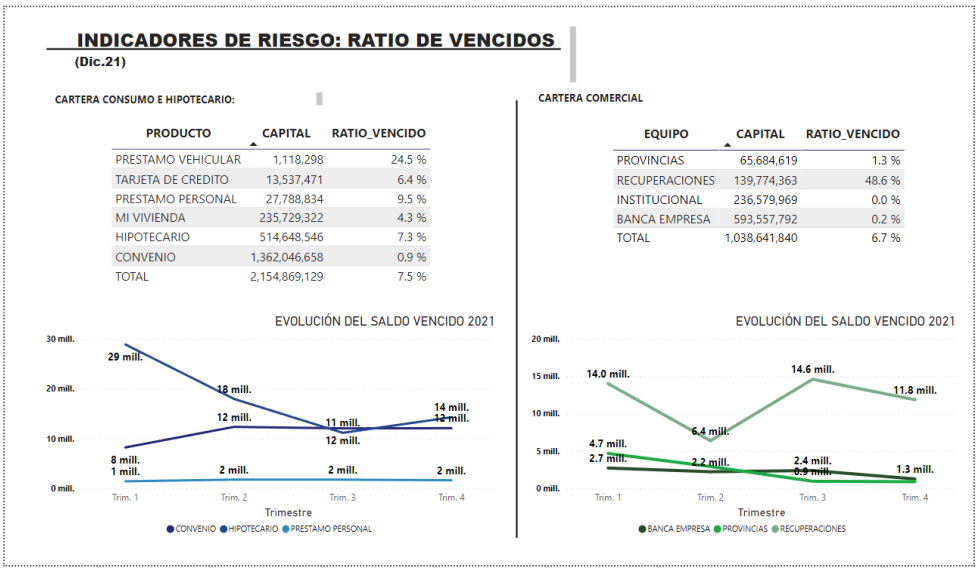


Repositorio de Datos Riesgos

Fuente: Elaboración propia

La figura 6 muestra la evolución trimestral de las colocaciones por tipo de cartera y las líneas de negocio que la conforman. La conexión directa de los datos con la herramienta de visualización hizo posible generar el informe de forma automática. Con la arquitectura de datos base (Fig. 4), esto no hubiera sido posible sin antes pasar por el uso de los formatos tradicionales.

Figura 7: Indicador de vencido por tipo de cartera



Repositorio de Datos Riesgos

Fuente: Elaboración propia

En esta parte del informe, se muestra el indicador de vencimiento por cartera y línea de negocio al cierre del 2021, así como la evolución trimestral del mismo. Para llegar a esto, se diseñaron dos cubos *OLAP:* uno para cartera de consumo e hipotecario y otro para la cartera comercial. La herramienta SQL permite guardar los cubos como si fueran vistas (*view*) de tal manera que se puedan conectar con el entorno de *Power BI.*

1.4. Análisis de Brechas

En el *framework* TOGAF, el análisis de brechas parte de la construcción de una matriz donde los encabezados de cada fila son los elementos que rigen la arquitectura actual y los encabezados de columna los elementos de la arquitectura de destino. En la intersección, se debe indicar qué elementos han sido omitidos o incluidos en el nuevo esquema.

La figura 8 muestra el resultado del análisis de brechas aplicado al problema de gestión de datos que afronta el área de seguimiento de portafolio. La última fila describe las mejoras que se lograron (brechas) y la última columna indica qué actividades o procesos relacionados con la gestión de datos se dejaron de lado.

Figura 8: Análisis de brechas

Arquitectura base	Arquitectura objetivo				Actividades eliminadas
	Repositorio de Riesgos	Repositorio de Riesgos	Cubos OLAP	Informes en Power BI	
Analista consulta datos	Omitido				X
Analista consolida datos en diferentes formatos		Omitido			X
Analista procesa los datos: cálculo de indicadores			Omitido		X
Analista elabora informes de seguimiento de riesgos				Omitido	
Entrega de resultados a la gerencia				Incluido	
Nuevo	Brecha: Se reemplazó como fuente de datos a diferentes tablas del servidor por una fuente de datos única: "Repositorio Riesgos"		Brecha: Las consultas (*querys*), tablas dinámicas son reemplazadas por los cubos OLAP y vistas directas en SQL.	Brecha: Los cubos *OLAP* y otras visualizaciones se obtienen directamente de "Repositorio Riesgos", permitiendo la elaboración de informes.	

Fuente: Elaboración propia

2. Expectativas de calidad de datos para el repositorio “Riesgos”

Los grandes volúmenes de datos representan un reto y a la vez una oportunidad para las organizaciones. Un artículo de BBVA (2018), señala que el desafío está en “separar el grano de la paja y que la calidad supere a la cantidad de datos”. Por un lado, se tiene el almacenamiento en cantidad y por otro la calidad del dato como activo de la empresa.

Rastreador.com es una compañía nacida en internet que empezó almacenando sus datos en Excel y que en la actualidad trabaja con herramientas de *machine learning*. También ha implementado la plataforma *Beconomy,* que utiliza datos de usuarios en línea para brindarle consejos de salud financiera de acuerdo con su perfil. Tener este vínculo con los usuarios les genera aporte de valor, pues a mayor beneficio que les aporta la entidad que les brinda servicio, más confianza tienen en los aplicativos o productos basados en datos.

En esta parte de la investigación, se va a tocar el tema de calidad de datos, utilizando como insumo principal el repositorio “Riesgos” construido en el apartado anterior. El objetivo, será verificar que los datos cumplan con ciertas características que aseguren la veracidad de informaciones futuras, generadas a partir de dicho repositorio.

2.1. Importancia de los procesos de calidad de datos

Las bases de datos son uno de los principales activos para las empresas y suelen tener problemas de calidad, siendo los más frecuentes: valores duplicados, valores perdidos, valores mal registrados e inconsistencias.

En el contexto de la banca, los procesos que forman parte de las unidades de riesgos no son ajenos a estos problemas. Sumado a esto, son auditables y la información que depende de los estos es utilizada por otras áreas. Así mismo, existen reportes e informes que son remitidos de manera formal a diferentes instancias como los entes reguladores, gerencia general, casa matriz o directorio.

Los analistas consultan las bases de datos que necesitan para generar sus reportes y antes de seguir con su elaboración. Seguido de esto, cotejan las cifras

más importantes (saldo cartera total, saldo de cartera vencida) con cifras oficiales que salen del balance general*. Si estos coinciden, se continúa con el trabajo. Sin embargo, no existe un esquema establecido que asegure la validez y consistencia de dichas cifras.

En base a la situación que atraviesa la unidad de gestión de riesgos de esta entidad, surge una iniciativa de mejora para la gestión de los datos. En este sentido, la propuesta consiste en desarrollar un modelo de calidad de datos que tiene como referente principal el modelo *GQM*.

2.2. Calidad de datos y Modelo GQM

El término calidad de datos es extenso. Algunos autores como Wang (1997) y Batini (2006), enfocan la definición de calidad en diferentes contextos de aplicación. Echverry (2007) en cambio, propone medirla en forma práctica desde un sistema de *data warehousing*. Por otro lado, Naumann (1996) y Sastre,Peralta y Ruggia (2008) se centran en las consultas de datos hechas por usuarios y sus exigencias en términos de calidad.

Bobrowski, Marré y Yankelevich (1998) sostienen que "… Para obtener una medida precisa de la calidad de los datos, se debe elegir qué atributos considerar y cuánto contribuye cada uno a la calidad en conjunto". Este trabajo va a tomar como referencia el concepto multidimensional. Con respecto a esta mirada, Pipino sostiene:

> Las empresas deben lidiar tanto con las percepciones subjetivas de las personas involucradas con los datos, como con las mediciones objetivas sobre el conjunto de datos en forma de pregunta. Las evaluaciones subjetivas de la calidad de los datos reflejan las necesidades y experiencias de partes interesadas: los recopiladores, custodios y consumidores de productos de datos. (Pipino et al,2002)

Tabla 4: Criterios de calidad de datos

Atributo	Definición
Accesibilidad	La medida en que los datos están disponibles o se pueden recuperar fácil y rápidamente
Cantidad adecuada de datos	La medida en que el volumen de datos es apropiado para la tarea en cuestión
Credibilidad	La medida en que los datos se consideran verdaderos y creíbles
Completitud	La medida en que los datos no faltan y son lo suficientemente amplios y profundos para la tarea en cuestión.
Representación concisa	La medida en que los datos se presentan en el mismo formato
Facilidad de manipulación	La medida en que los datos son fáciles de manipular y aplicar a diferentes tareas
Libre de errores	La medida en que los datos son correctos y confiables.
Interpretabilidad	La medida en que los datos están en los idiomas, símbolos y unidades apropiados, y las definiciones son claras
Objetividad	La medida en que los datos son imparciales y sin prejuicios
Pertinencia	La medida en que los datos son aplicables y útiles.
Reputación	La medida en que los datos son muy apreciados en términos de su fuente o contenido
Seguridad	La medida en que el acceso a los datos se restringe adecuadamente para mantener su seguridad
Oportunidad	La medida en que los datos están suficientemente actualizados para la tarea en cuestión
Comprensibilidad	La medida en que los datos se comprenden fácilmente
Valor añadido	La medida en que los datos son beneficiosos y proporcionan ventajas a partir de su uso

Fuente: Data Quality Assessment, Pipino (2002)

Lo más importante es elegir los criterios de calidad que se van a aplicar a los datos en función al uso que se les dará. Se debe apuntar a un equilibrio entre calidad y cantidad pues al evaluar las fuentes de datos, la atención debe centrarse en cómo la calidad afecta negativamente al análisis y las decisiones que pueden tomarse a partir de los mismos. En este sentido, lo que se propone es diseñar una herramienta o modelo de evaluación de datos (MED) utilizando el enfoque GQM.

GQM es un método orientado a generar una métrica que mide un objetivo de una manera determinada a través de la utilización de preguntas. Proporciona

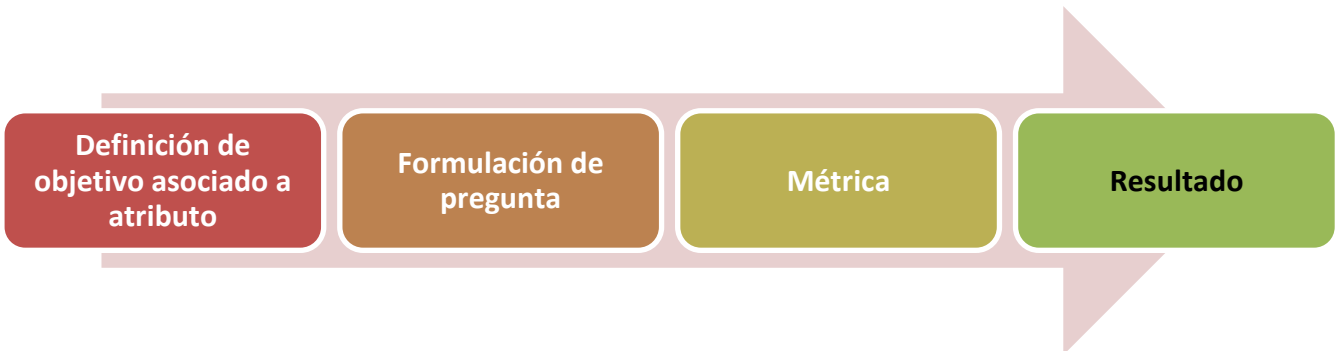
una manera útil para definir mediciones, tanto del proceso como de los resultados de un proyecto. (Calabrese, Esponda, Pasini y Pesado, 2020)

Las organizaciones que persiguen objetivos deben trazarse metas para sí mismas y sus procesos. El resultado de aplicar el enfoque GQM, es la especificación de un sistema de medición de datos, relacionados a una problemática en específico dentro de la organización. En primer lugar, se define un objetivo y se plantean preguntas en torno al mismo, generando métricas en base a las respuestas (Basili y Caldiera,1994). El modelo consta de tres niveles:

- Nivel conceptual: se identifica a lo que se aspira respecto a los productos, procesos o recursos relativos a un entorno particular dentro de la organización.
- Nivel operativo: se ajustan las preguntas con el fin de verificar el cumplimiento del objetivo. Las preguntas en este nivel deben hacer referencia al área, proceso o producto que se está evaluando.
- Nivel cuantitativo: en esta parte se asocia un conjunto de datos para cada pregunta formulando métricas y obtener una respuesta de forma cuantitativa.

El siguiente esquema resume el modelo de evaluación de calidad de datos que se va a utilizar:

Figura 9: Esquema de calidad de datos



Fuente: Elaboración propia

2.3. Desarrollo del modelo de evaluación de datos

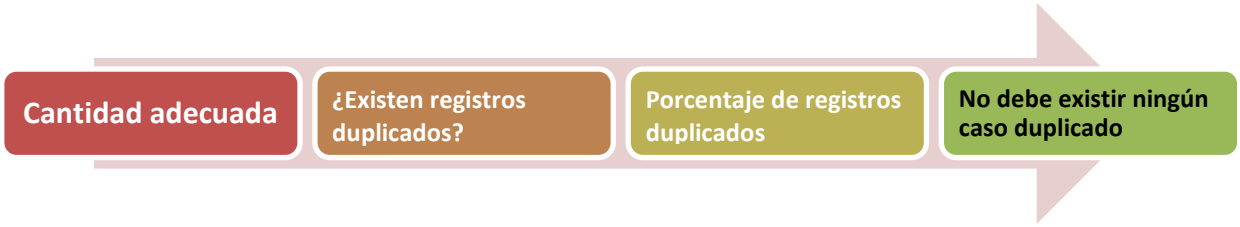
A continuación, se plantean los objetivos de calidad de datos planteados para el repositorio de datos "Riesgos", construido en el apartado 1. Se utilizó la librería *Great expectations* disponible en Python. Según el portal oficial de esta librería "*Great Expectations* [6]es la herramienta líder para validar, documentar y perfilar datos con el fin de mantener la calidad y mejorar la comunicación entre los equipos".

Siguiendo el esquema planteado en el punto anterior (Figura 9), se definen las siguientes expectativas:

2.3.1. Cantidad adecuada del número de operaciones

El repositorio "Riesgos", está compuesto por filas con información de cada operación crediticia. Por este motivo, no puede existir códigos de operación duplicados, ya que esto traería errores a la hora de sumar y contabilizar.

Figura 10: Objetivo cantidad adecuada



Fuente: Elaboración propia

Los resultados de esta librería se dan en formato *json.* El parámetro *mostly* permite establecer un umbral que representa el máximo de valores duplicados permitidos. En este caso, dicho parámetro toma el valor de cero, ya que el número de operación debe ser único.

[6] https://greatexpectations.io/

Figura 11: Cantidad adecuada del número de operaciones

```
gedf[gedf['FECHA']=='31/12/2021'].expect_column_values_to_be_unique('OPERACIÓN',mostly=0,result_format={'result_format': 'BASIC'})
```


```
{
  "success": true,
  "exception_info": {
    "raised_exception": false,
    "exception_traceback": null,
    "exception_message": null
  },
  "result": {
    "element_count": 50313,
    "missing_count": 0,
    "missing_percent": 0.0,
    "unexpected_count": 0,
    "unexpected_percent": 0.0,
    "unexpected_percent_total": 0.0,
    "unexpected_percent_nonmissing": 0.0,
    "partial_unexpected_list": []
  },
  "meta": {},
  "expectation_config": {
    "expectation_type": "expect_column_values_to_be_unique",
    "kwargs": {
      "column": "OPERACI\u00d3N",
      "mostly": 0,
      "result_format": {
        "result_format": "BASIC"
      }
```


Fuente: Elaboración propia

La figura 11 muestra que, durante el mes de diciembre de 2021, no se registraron valores duplicados en la columna “Operación”. Esto se puede verificar en la línea 1 (*success*).

2.3.2. Columnas libres de errores (Free-of-error)

Se va a ejemplificar con los siguientes atributos:

Clasificación de cartera: según el tipo de cliente y los créditos que toman, la cartera está dividida en tres categorías:

Clase 1: Consumo

Clase 2: Hipotecario

Clase 3: Comercial

Calificación de riesgo: Es la calificación que obtiene el cliente en base a los días de mora:

Clase 1: 0-Normal

Clase 2: 1-CPP

Clase 3: 2-Deficiente

Clase 4: 3-Dudoso

Clase 5: 4-Pérdida

Segmento: segmento del cliente en función con los lineamientos establecidos por el regulador. Pueden ser:

Clase 1: Corporativos

Clase 2: Grandes empresas

Clase 3: Medianas empresas

Clase 4: Pequeñas empresas

Clase 5: Microempresas

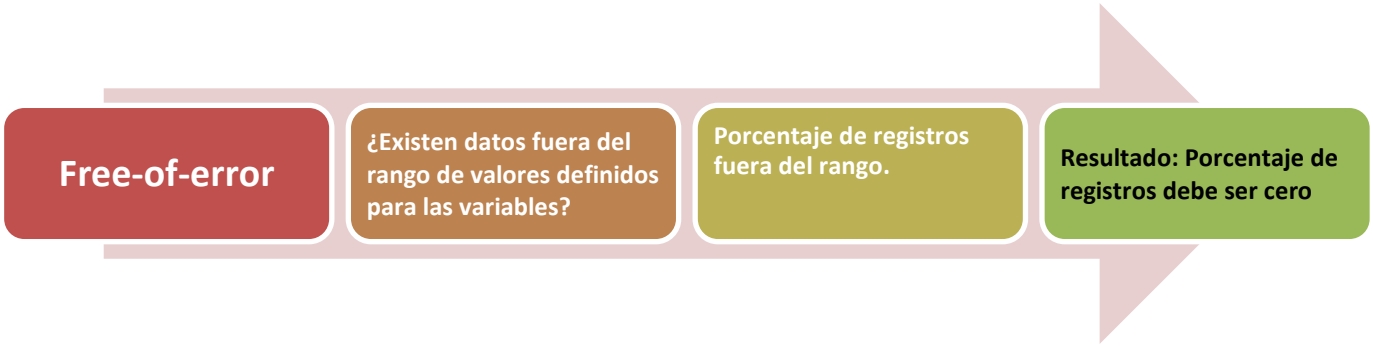
Clase 6: Consumo revolvente

Clase 7: Consumo no revolvente

Clase 8: Hipotecarios

Para estas columnas, se busca identificar registros que están tomando valores fuera de las clases definidas. La métrica utilizada será el porcentaje de valores distintos a las categorías correspondientes al tipo de cartera, clasificación de riesgo o segmento. Dado a que son clases definidas tanto por la entidad y el regulador, el único umbral será cero, es decir no se aceptan valores diferentes a las categorías establecidas.

Figura 12: Objetivo *Free-of-error*



Fuente: Elaboración propia

Figura 13: Calificación de riesgo free-of-error

```
gedf.expect_column_values_to_be_in_set('CALIFICACION',[0,1,2,3,4],mostly=0,result_format={'result_format': 'BOOLEAN_ONLY'})
```


```
{
  "meta": {},
  "exception_info": {
    "raised_exception": false,
    "exception_traceback": null,
    "exception_message": null
  },
  "expectation_config": {
    "meta": {},
    "expectation_type": "expect_column_values_to_be_in_set",
    "kwargs": {
      "column": "CALIFICACION",
      "value_set": [
        0,
        1,
        2,
        3,
        4
      ],
      "mostly": 0,
      "result_format": {
        "result_format": "BOOLEAN_ONLY"
      }
    }
  },
  "result": {},
  "success": true
}
```


Fuente: Elaboración propia

Figura 14: Clasificación de cartera free-of-error

```
[18] gedf[gedf['FECHA']=='31/12/2021'].expect_column_values_to_be_in_set('CARTERA',['CONSUMO', 'HIPOTECARIO', 'COMERCIAL'],mostly=0,result_format={'result_format': 'BOOLEAN_ONLY'})
```


```
{
  "meta": {},
  "exception_info": {
    "raised_exception": false,
    "exception_traceback": null,
    "exception_message": null
  },
  "expectation_config": {
    "meta": {},
    "expectation_type": "expect_column_values_to_be_in_set",
    "kwargs": {
      "column": "CARTERA",
      "value_set": [
        "CONSUMO",
        "HIPOTECARIO",
        "COMERCIAL"
      ],
      "mostly": 0,
      "result_format": {
        "result_format": "BOOLEAN_ONLY"
      }
    }
  },
  "result": {},
  "success": true
}
```


Fuente: Elaboración propia

Figura 15: Segmento de cliente free-of-error


```
  "meta": {},
  "expectation_type": "expect_column_values_to_be_in_set",
  "kwargs": {
    "column": "SEGMENTO",
    "value_set": [
      [
        "CONSUMO NO REVOLVENTE",
        "HIPOTECARIOS",
        "PEQUE\u00d1AS EMPRESAS",
        "GRANDES EMPRESAS",
        "MEDIANAS EMPRESAS",
        "CONSUMO REVOLVENTE",
        "CORPORATIVOS",
        "MICROEMPRESAS"
      ]
    ],
    "mostly": 0,
    "result_format": {
      "result_format": "BOOLEAN_ONLY"
    }
  }
},
"result": {},
"success": true
}
```

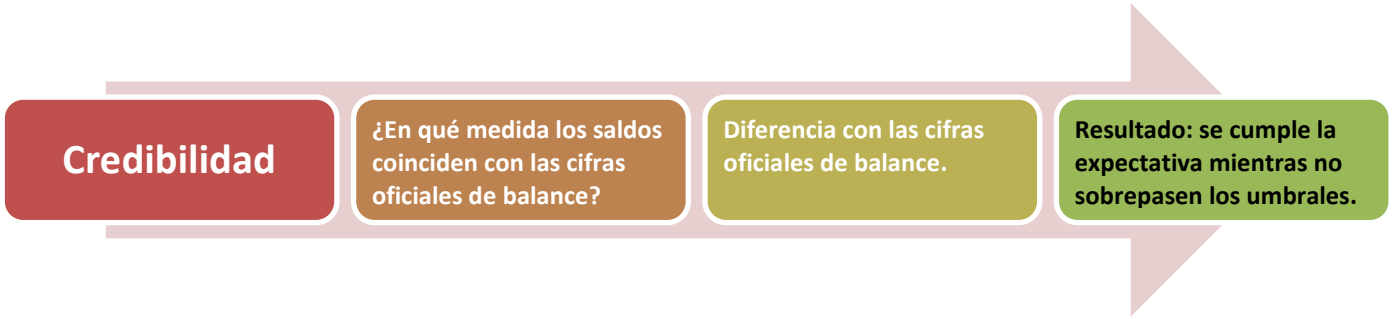

Fuente: Elaboración propia

Al igual que en el caso anterior, se tomó diciembre para generar las expectativas de validez. El resultado es exitoso en los tres casos, pues todos los clientes se encuentran bien agrupados en función al tipo de cartera, tienen la clasificación de riesgo que le corresponde y segmento bien asignado.

2.3.2. Credibilidad de saldos

Se van a evaluar las columnas “Saldo Capital” y “Saldo Vencido” con el objetivo de que los saldos totales, es decir la suma de cada columna sea comparada con cifras oficiales del balance*. Esto es importante, dado que los informes de evolución de deuda o indicadores de vencidos son fundamentales para establecer estrategias de negocio.

Figura 16: Objetivo de credibilidad



Fuente: Elaboración propia

Figura 17: Saldo Capital vs cifras de balance

```
[44] gedf[gedf['FECHA']=='31/12/2021'].expect_column_sum_to_be_between('SALDO_CAPITAL',min_value=min,max_value=max,result_format={'result_format': 'BOOLEAN_ONLY'})

{
  "meta": {},
  "exception_info": {
    "raised_exception": false,
    "exception_traceback": null,
    "exception_message": null
  },
  "expectation_config": {
    "meta": {},
    "expectation_type": "expect_column_sum_to_be_between",
    "kwargs": {
      "column": "SALDO_CAPITAL",
      "min_value": 3626390971,
      "max_value": 3626410971,
      "result_format": {
        "result_format": "BOOLEAN_ONLY"
      }
    }
  },
  "result": {},
  "success": true
}
```

Fuente: Elaboración propia

Figura 18: Saldo Vencido vs cifras de balance

```
✓ [48] gedf[gedf['FECHA']=='31/12/2021'].expect_column_sum_to_be_between('SALDO_VENCIDO',min_value=min,max_value=max,result_format={'result_format': 'BOOLEAN_ONLY'})
```

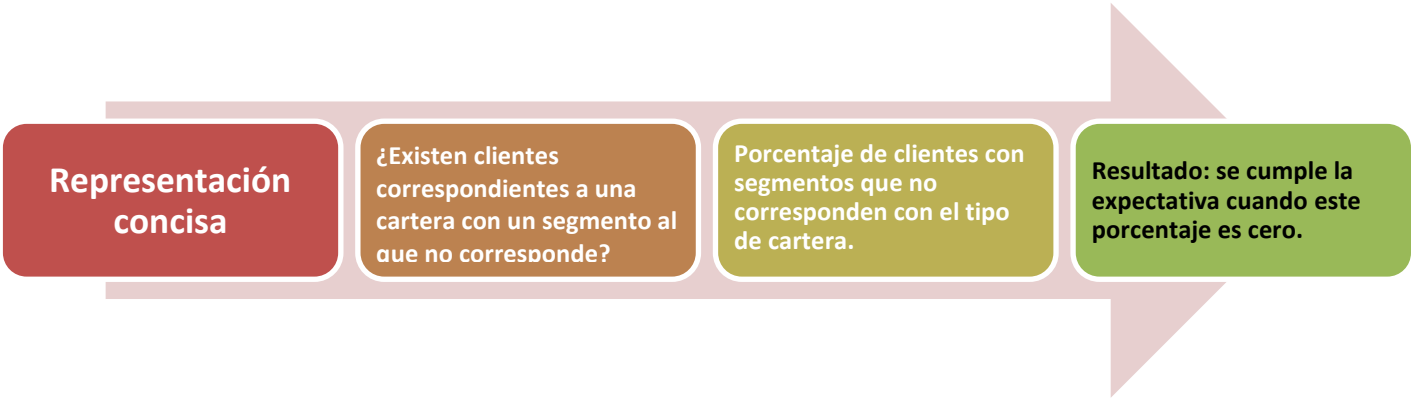

```
{
  "meta": {},
  "exception_info": {
    "raised_exception": false,
    "exception_traceback": null,
    "exception_message": null
  },
  "expectation_config": {
    "meta": {},
    "expectation_type": "expect_column_sum_to_be_between",
    "kwargs": {
      "column": "SALDO_VENCIDO",
      "min_value": 48108364,
      "max_value": 48118364,
      "result_format": {
        "result_format": "BOOLEAN_ONLY"
      }
    }
  },
  "result": {},
  "success": true
}
```


Fuente: Elaboración propia

2.3.3. Representación concisa entre atributos (consistencia)

Los informes analíticos contemplan diferentes vistas del portafolio. En este sentido, se pueden elaborar tablas cruzadas y gráficos, pero antes se debe asegurar que las columnas que hacen referencia a un mismo aspecto estén bien representadas. Para esto, se van a tomar las columnas “Cartera” y “Segmento”. Por ejemplo, la cartera comercial no puede tener clientes del segmento hipotecario o consumo revolvente, así como las carteras correspondientes a créditos hipotecarios o de consumo no pueden registrar créditos corporativos o grandes empresas. El esquema muestra la secuencia que va a seguir la evaluación de ambos atributos:

Figura 19: Representación concisa entre atributos



Fuente: Elaboración propia

Figura 19: Representación concisa entre atributos cartera y segmento

```
[52] gedf.expect_column_pair_values_to_be_in_set('CARTERA','SEGMENTO', CARTERA_SEGMENTO,
                        meta = {"notes": {"content": [ "Cada cartera debe tener un determinado segmento de clientes" ]}},result_format= {'result_format':'BOOLEAN_ONLY'})

    "raised_exception": false,
    "exception_traceback": null,
    "exception_message": null
  },
  "expectation_config": {
    "meta": {
      "notes": {
        "content": [
          "Cada cartera debe tener un determinado segmento de clientes"
        ]
      }
    },
    "expectation_type": "expect_column_pair_values_to_be_in_set",
    "kwargs": {
      "column_A": "CARTERA",
      "column_B": "SEGMENTO",
      "value_pairs_set": [
        [
          "CONSUMO",
          "CONSUMO NO REVOLVENTE"
        ],
        [
          "CONSUMO",
          "CONSUMO REVOLVENTE"
        ],
        [
          "HIPOTECARIO",
          "HIPOTECARIOS"
        ],
        [
          "COMERCIAL".


        ],
        [
          "COMERCIAL",
          "CORPORATIVOS"
        ],
        [
          "COMERCIAL",
          "GRANDES EMPRESAS"
        ],
        [
          "COMERCIAL",
          "MEDIANAS EMPRESAS"
        ],
        [
          "COMERCIAL",
          "PEQUE\u00d1AS EMPRESAS"
        ],
        [
          "COMERCIAL",
          "MICROEMPRESAS"
        ]
      ],
      "result_format": {
        "result_format": "BOOLEAN_ONLY"
      }
    }
  },
  "result": {},
  "success": true
}
```

Fuente: Elaboración propia

Se observa que, para el periodo analizado, los datos de las columnas “Cartera” y “Segmento” son consistentes. Los clientes de cada cartera del banco tienen el segmento que corresponde.

3. Modelo de machine learning aplicado a la gestión de riesgo

Los bancos pueden mejorar la gestión de su portafolio con ayuda de modelos basados en matemáticas y datos. La digitalización y el incremento del número de clientes han generado la necesidad de desarrollar metodologías capaces de clasificar o predecir con la mayor precisión posible.

Dentro del rubro financiero se están aplicando algoritmos de machine learning, los cuales se han adaptado con éxito al gran número de datos con información financiera de clientes. Este tipo de algoritmos tienen la capacidad de ser aplicables a diferentes áreas del negocio. Desde la detección del riesgo hasta la identificación de identidad con apoyo de biometría e imágenes (Leo, Sharma y Maddulety, 2019). Dada esta versatilidad, surge la inquietud de aplicar este tipo de modelos en el marco de la gestión de riesgo de crédito de la entidad que se está tomando como referencia.

En este apartado se va a desarrollar un modelo de machine learning con árbol de decisión, cuyos resultados van a ser comparados con un modelo de regresión

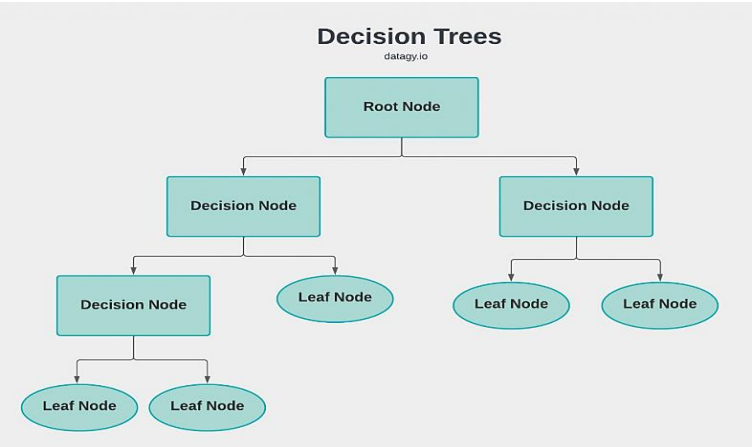
logística. A través de métricas que indiquen el nivel de precisión de cada uno de estos, se va a evaluar el alcance que traería su aplicación.

3.1. Árbol de decisión (*Decision tree*)

El árbol de decisión es un método no paramétrico que no requiere supuestos distribucionales, permite detectar interacciones, modela relaciones no lineales y no es sensible a la presencia de datos faltantes y *outliers* (Breiman, Friedman, Olshen & Stone, 1984). Operan con diferentes metodologías, entre las que están los Cart, Chaid y Chaid exhaustivo. Estas difieren en la forma de asignación, reglas de partición y criterios de parada. Cualquiera sea el método, se generan *n* nodos terminales y una escala de probabilidades con n posibles valores que es el resultado final (Cardona,2004).

Este apartado se va a centrar los árboles de decisión de clasificación, los cuales funcionan como diagramas de flujo. Cada nodo de un árbol de decisión representa un punto de decisión que se divide en dos nodos de hoja. Cada uno de estos nodos representa el resultado de la decisión y cada una de las decisiones también puede convertirse en nodos de decisión. Eventualmente, las diferentes decisiones conducirán a una clasificación final. El siguiente diagrama ilustra como es que funcionan los árboles de decisión para llegar a un resultado final:

Figura 20: Diagrama de árbol de árbol de decisión



Supongamos que se tienen 10 variables predictoras con dos categorías o clases cada una. Las posibles combinaciones serían más de mil. Es ahí donde cobra

importancia el árbol de decisión, ya que el algoritmo devolverá el mejor árbol que tome la decisión más acertada haciendo uso de las probabilidades.

3.2.Criterios de evaluación de modelos

Los modelos se evalúan en función de su capacidad predictiva a la hora de discriminar entre una u otra clase en cuestión. La evaluación se da comparando las clases predichas por el modelo con respecto a la clase real.
Una ves seleccionadas las variables que se utilizarán para modelar, es necesario dividir el conjunto de datos en entrenamiento y prueba. Es importante realizar este proceso a fin de aislar variables que el modelo no haya "visto", para saber si realmente aprendió a desarrollar la tarea que buscaba aprender o simplemente memorizó los datos que se usaron en el entrenamiento (Martinez, 2022).

3.2.1. Matriz de confusión

Una manera de comparar las predicciones del modelo con la clase real a la que pertenece cada individuo es la matriz de confusión:

Figura 21: Matriz de confusión

		PREDICCIÓN	
		NEGATIVO	POSITIVO
OBSERVACIÓN	NEGATIVO	Verdaderos negativos (VN)	Falsos Negativos (FN)
	POSITIVO	Falsos Positivos (FP)	Verdaderos Positivos (VP)

Elaboración propia.

- Verdaderos Positivos (VP): Número de observaciones que se clasificaron correctamente como "positivos".
- Verdaderos Negativos (VN): Número de observaciones que se clasificaron correctamente como "negativos".

- Falsos Positivos (FP): También conocido como error tipo I, es el número de observaciones que se clasificaron incorrectamente como "positivos".
- Falsos Negativos (FN): También conocido como error tipo II, es el número de observaciones que se clasificación incorrectamente como "negativos".

3.2.2. Métricas de evaluación

- Exactitud (Accuracy): Proporción de predicciones correctas.

$$Exactitud = \frac{VP + VN}{Total\ de\ obs.} = \frac{VP + VN}{VP + FP + FN + VN}$$

- Tasa de Error: Proporción de observaciones clasificadas incorrectamente.

$$Tasa\ de\ error = 1 - Exactitud = \frac{FP + FN}{Total\ de\ obs.}$$

- Sensibilidad (*Precisión*): También conocido como tasa de verdaderos positivos, es la proporción de casos positivos que fueron correctamente identificados.

$$Sencibilidad = \frac{VP}{Total\ positivos} = \frac{VP}{VP + FN}$$

- Especificidad (*Recall*): También conocido como tasa de verdaderos negativos, es la proporción de casos negativos correctamente identificados.

$$Especificidad = \frac{VN}{Total\ negativos} = \frac{VN}{VN + FP}$$

- Tasa de Falsos positivos: También conocido como Error tipo I, es la probabilidad de que se dé un resultado positivo cuando el valor verdadero es negativo.

$$TFP = 1 - Especificidad = \frac{FP}{VN + FP}$$

- Tasa de Falsos negativos: También conocido como Error tipo II, es la probabilidad de que la prueba pase por alto un verdadero positivo, es decir, que se dé un resultado negativo cuando el verdadero valor es positivo.

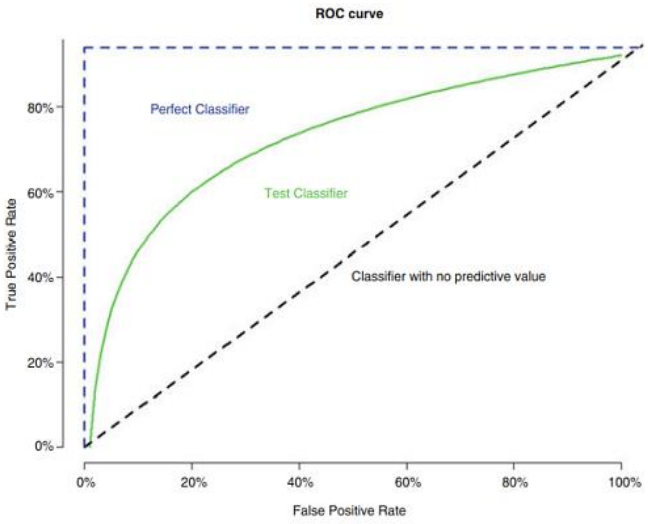
$$TFN = 1 - Sencibilidad = \frac{FN}{VP + FN}$$

3.2.3. Curva ROC (Receirver Operating Characteristic)

La curva ROC es una representación gráfica de la fracción de falsos positivos (abscisas) frente a la fracción de verdaderos positivos (ordenadas).

$$ROC = \begin{cases} y = FP \\ x = VP \end{cases}$$

Figura 22: Curva ROC



Fuente: Ivo Dinov, 2018

La línea azul representa una clasificación perfecta donde se tiene 0 % de falsos positivos y 100 % de verdaderos positivos, la curva de color verde representa un ejemplo de curva de ROC para un modelo de clasificación, y la línea diagonal color negro, corresponde a un modelo incapaz de discriminar entre las clases positivas y negativas. Una manera de cuantificar el rendimiento de este tipo de modelos es el área bajo la curva (AUC[7]). Esta medida determina que tan bueno es el modelo para discriminar entre una clase u otra. Según Ivo D. Dinov (2018), la siguiente tabla muestra la valoración del modelo en función al valor del AUC bajo la curva ROC:

[7] Área Under the Curve

Tabla 5: Valoración AUC

AUC	Desempeño
0.5-0.6	Sin discriminación
0.6-0.7	Malo
0.7-0.8	Regular
0.8-0.9	Bueno
0.9-1.0	Excelente

Fuente: Ivo Dinov, 2018

3.3. Modelo de árbol de decisión para el análisis de créditos castigados

Se van a tomar datos de una cartera de créditos por convenio que fue castigada[8] por esta entidad financiera durante el año 2020. La variable de estudio será el motivo de castigo "cese" que toma valores 0 y 1, donde 1 representa al cliente que dejó de pagar por haber sido cesado en su centro de labores y 0 caso contrario. El análisis tiene una parte exploratoria de las variables predictoras, seguido del entrenamiento del modelo de árbol de decisión. Para verificar la eficiencia que tiene este método de aprendizaje automático, se va a contrastar con los resultados de un modelo logístico que será aplicado al mismo set de datos.

3.3.1. Análisis Exploratorio

El set de datos está compuesto por 1386 clientes, no registra valores perdidos ni filas duplicadas.

Figura 23: Descripción del set de datos

Number of observations	1386
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%

Fuente: Elaboración propia

[8] Cartera castigada: créditos clasificados como pérdida, íntegramente provisionados, que han sido retirados de los balances de las empresas. Para castigar un crédito, debe existir evidencia real de su irrecuperabilidad o debe ser por un monto que no justifique iniciar acción judicial o arbitral (SBS, 2015).

Se tienen cuatro variables cualitativas: edad, ingresos, género y fuerzas armadas. Todas son del tipo cualitativo que describen características personales del cliente.

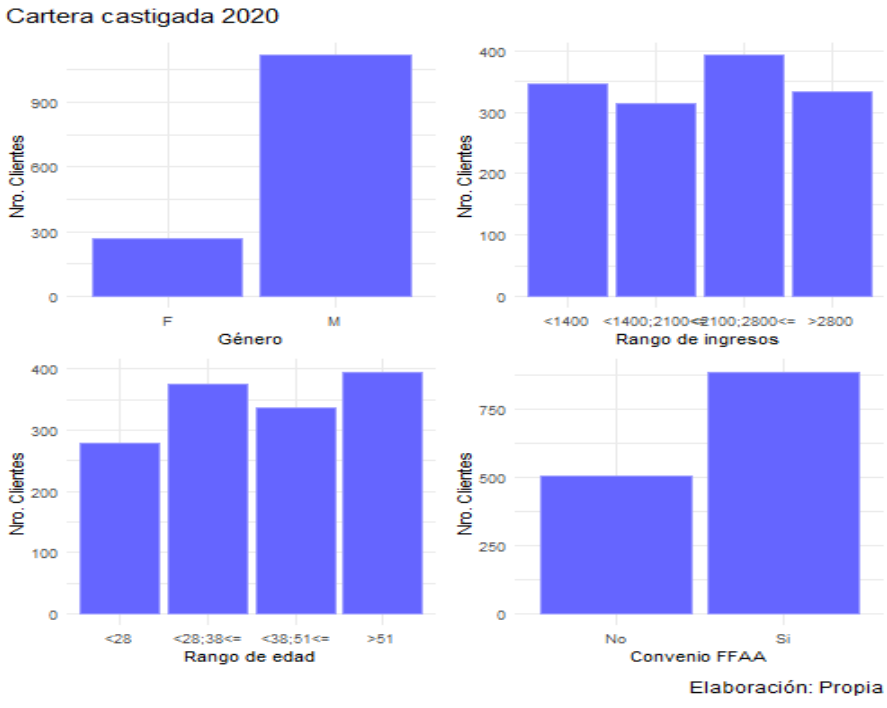
Para la edad se tienen cuatro rangos que agrupan a clientes menores de 21 años hasta los que superan los 51. Se observa que, en esta cartera, la mayor parte supera los 51 años, seguido de los que tienen entre 28 y 38.

Los ingresos también están representados en cuatro clases diferentes. No existe concentración de una en específico. La clase con mas individuos es la que posee ingresos netos mensuales entre S/2100 y S/2800, seguido de los que ganan menos de S/1400. Predominando así clientes con niveles de ingresos que están cerca al sueldo mínimo (S/1025).

Con relación al género, predominan los varones, representando el 80.7% del total.

La variable Fuerzas Armadas, es un indicador que marca si el cliente trabajaba en una entidad correspondiente a instituciones gubernamentales de las fuerzas armadas del Perú. Se consideró esta variable dado que el sector público, en específico las fuerzas armadas forman parte del mercado objetivo de créditos por convenios.

Gráfico 1: Variable predictoras

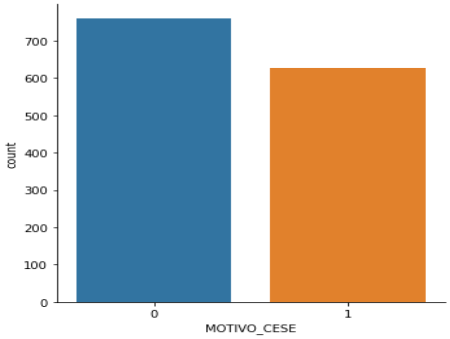


Fuente: Elaboración propia

Variable respuesta:

La variable en estudio es el motivo de castigo “cese”. Este motivo, agrupa aquellos clientes fueron despedidos y como consecuencia, tuvieron falta de ingresos para cumplir con sus obligaciones. En esta cartera, se observa que, de los castigados durante el 2020, 626 clientes (45%) fueron despedidos de su centro de labores.

Gráfico 2: Variable respuesta

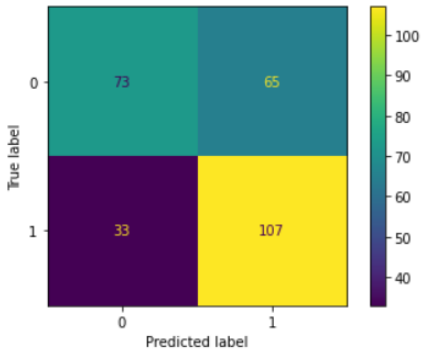


Fuente: Elaboración propia con librería *sklearn*

3.3.2. Modelo de árbol de decisión

Este modelo tiene una precisión del 65%. Si se pone atención a la clase de interés, que es el grupo de clientes castigados porque dejaron de pagar dada su situación de despido (cese), la sensibilidad o tasa de acierto para este grupo es del 74%. Por otro lado, la clase correspondiente a otros motivos tiene una precisión del 53%. Esto pondera a la baja la precisión total del modelo. La curva ROC (Gráfico 3), tiene un AUC de 73%. De acuerdo con la tabla 5 (Punto 3.2.3), este modelo tiene un rendimiento regular, ya que no supera el 80%.

Gráfico 3: Matriz de confusión -Modelo árbol de decisión



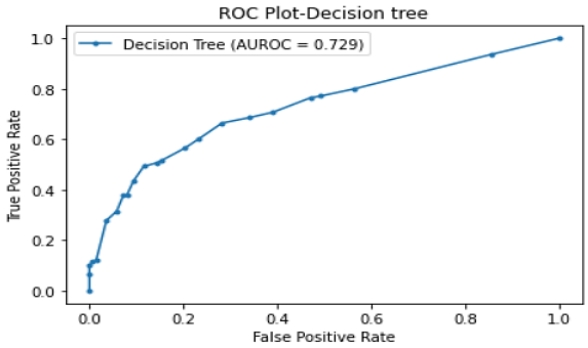
Fuente: Elaboración propia con librería *sklearn*

Figura 24: Reporte de clasificación

	precision	recall	f1-score	support
0	0.69	0.53	0.60	138
1	0.62	0.76	0.69	140
accuracy			0.65	278
macro avg	0.66	0.65	0.64	278
weighted avg	0.66	0.65	0.64	278

Fuente: Elaboración propia con librería *sklearn*

Gráfico 4: Curva ROC -Modelo árbol de decisión

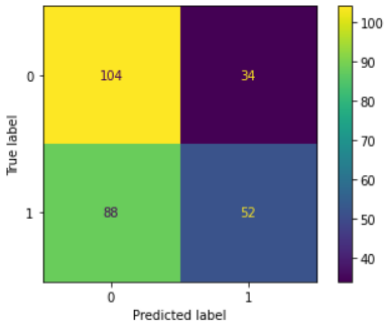


Fuente: Elaboración propia con librería *sklearn*

3.3.3. Modelo de regresión logística

Este modelo predice probabilidades de ocurrencia para un evento en específico. Las mismas se calculan en función a un grupo de variables predictoras. El contexto de esta investigación plantea que, a través de este modelo, se calcule la probabilidad de que un cliente haya sido castigado por el motivo cese. El resultado indica que, a nivel global este modelo es significativo, considerando los ingresos, edad, género y si el cliente cumple o no con la condición de haber trabajado en una entidad de fuerzas armadas. Sin embargo, las métricas de evaluación muestran que el modelo tiene baja capacidad predictiva. Solo el *accuracy* es del 56%; a nivel de grupo, el *recall* de la clase de interés (1) es del 37%, teniendo mayor acierto para los otros casos donde el cliente no fue castigado por despido, es decir dejó de pagar por otros motivos. El área bajo la curva es cercana al modelo base (sin variables predictoras), tomando un valor de 59.7%.

Gráfico 4: Matriz de confusión -Modelo reg. logística



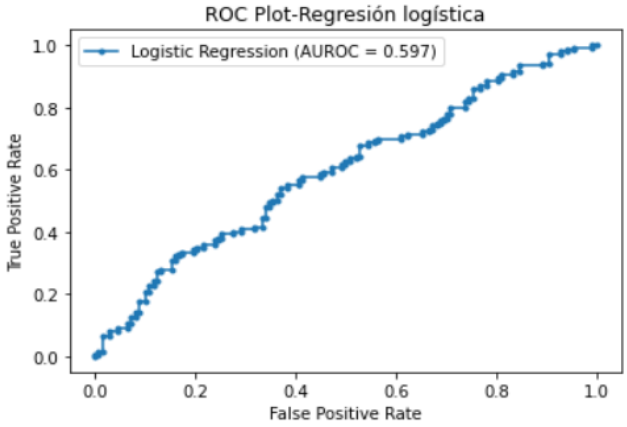
Fuente: Elaboración propia con librería *sklearn*

Figura 25: Reporte de clasificación

	precision	recall	f1-score	support
0	0.54	0.75	0.63	138
1	0.60	0.37	0.46	140
accuracy			0.56	278
macro avg	0.57	0.56	0.55	278
weighted avg	0.57	0.56	0.54	278

Fuente: Elaboración propia con librería *sklearn*

Gráfico 5: Curva ROC - Modelo de reg. logística



Fuente: Elaboración propia con librería *sklearn*

3.3.4. Comparación de modelos

Se puede tomar la tabla 6 para evaluar la capacidad predictiva de cada modelo. Se observa que el mejor desempeño del modelo de árbol de decisión o *decisión tree*, en tanto que el modelo de regresión logística tiene tasas bajas de precisión.

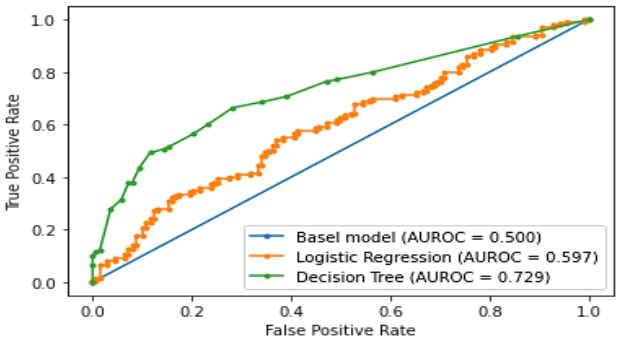
Tabla 6: Métricas de evaluación

MODELO	Accuracy	Sensibilidad	Especificidad	TFP	TFN
Decisión Tree	64.7%	76.4%	68.9%	31.1%	23.6%
Regresión logística	56.1%	37.1%	54.2%	45.8%	62.9%

Fuente: Elaboración propia

El gráfico 6, compara la curva ROC para ambos modelos, en este se evidencia de forma conjunta que el modelo de árbol de decisión tiene un AUC superior al de la regresión logística.

Gráfico 6: Curva ROC



Fuente: Elaboración propia con librería *sklearn*

4. Conclusiones

- Las entidades financieras, en específico aquellas que están en proceso de crecimiento, cada vez necesitan más de los datos para la toma de decisiones. Esto va de la mano con la cantidad de información que ingresa a la organización, en ese sentido, trabajar bajo el enfoque de *big data* es una alternativa que permite gestionar los datos de manera eficiente.
- Los cubos *OLAP* a partir del repositorio de datos hicieron posible la creación de vistas de las dimensiones, cómo el producto y tipo de cartera. Seguido de esto se pudo habilitar una conexión directa del repositorio con una herramienta de visualización; en este caso Power BI. En suma, la aplicación de analítica de big data hizo posible mejorar la eficiencia de los procesos de recolección y procesamiento de datos, para dar lugar a resultados analíticos más sofisticados.

- La creación del repositorio de datos deja abierta la posibilidad de crear otro repositorio con datos abiertos. Los datos abiertos son información pública que proporciona el regulador. Entonces, se pueden consolidar variables de riesgos de otras entidades financieras, lo cual va a permitir hacer informes de competencia, perfil de riesgos y vistas de comparación entre bancos.
- El modelo de calidad de datos creado para el repositorio de riesgos es un proceso que asegura información confiable. Esto es importante, ya que los datos son el principal activo que tiene el área para generar informes analíticos para la toma de decisiones.
- Se establecieron cinco expectativas: unicidad, validez, precisión y consistencia. En todos los casos se cumplieron las metas establecidas. Esto es positivo pues denota que se viene trabajando con información confiable a pesar de la ausencia de un proceso formal.
- La creación de expectativas para datos no debe limitarse a ser un paso previo para la generación de informes. El modelo que se ha propuesto en esta investigación también se puede incluir como parte del perfilamiento de datos para modelos estadísticos o de aprendizaje automático.
- El modelo de árbol de decisión planteado en el apartado 3 mostró mayor precisión que el modelo logístico. De acuerdo con las métricas de evaluación, este predice mejor aquellos clientes que fueron castigados debido a que fueron cesados.
- Emplear modelos de machine learning como el árbol de decisión, representa una alternativa que permite aprovechar al máximo la información de los clientes. Esto en el sentido de que este tipo de algoritmos tienen la capacidad computacional de hacer múltiples combinaciones entre las variables predictoras. Cosa que el modelo de regresión logística simple no hace.
- Los resultados obtenidos en el modelo de árbol de decisión permite caracterizar e identificar patrones entre los clientes que fueron castigados por motivo de cese. Conocer características en específico, permite hacer un *feedback* de cara a mejorar políticas de aceptación de riesgo y poner énfasis a determinados clientes a la hora de hacer seguimiento de cartera. Así mismo, se puede adoptar alguna estrategia que impida que el crédito llegue a la instancia de ser castigado.

Bibliografía

Arango, L. y. (2010). Arquitectura empresarial, una visión general. *Revista Ingenierías Universidad de Medellín*.

Basili, & Caldiera. (1994). Goal question metric paradigm. En *Encyclopedia of software engineering* (págs. 528-532).

Batini, Cappiello, Francalaci, & Maurino. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*.

Bobrowski, Marré, & Yankelevich. (1998). A Software Engineering View of Data Quality.

Breiman, Friedman, Olshen, & Stone. (1984). *Classification and regresion trees.*

Brissett, P. R. (2018). *La importancia para la gestión de riesgos entidades financieras.* Bogotá,Colombia.

Calabrese, Esponda, Pasini, & Pesado. (2020). Modelo de evaluación de datos utilizando el enfoque GQM.

Hernández, C. (2004). Aplicación de árboles de decisión en modelos de riesgo de crédito. *Revista colombiana de estadística*.

Ivo, D. (2018). Data Science and predictive analytics.

Leo, Sharma, & Maddulety. (2019). Machine Learning in Banking Risk Management:A literature review. *MDPI*.

Luis, J. A. (2013). *Análisis de grandes volúmenes de datos en organizaciones* . Mexico: Omega.

Martínez Fernandez, T. C. (2022). Comparación de modelos machine learning aplicados al riesgo de crédito.

Naumann. (2002). Quality-driven query answering for integrated information systems. *Computer Science*.

Philipp Härle, A. H. (2016). The future of bank risk. *McKinsey & Company*.

Pipino, Lee, & Wang. (2002). Data Quality Assessment.

Reinosa, Maldonado, Muñoz, Damiano, & Abrutsky. (2012). *Bases de datos.* Buenos Aires: Alfaomega.

Russon. (2011). *Big Data Analytics*. Obtenido de IBM: http://www.tdwi.org

Sastre, Peralta, & Ruggia. (2008). Evaluación de Calidad en una Aplicación de Data Warehousing: de la Definición de Metas a la Especificación de Métricas.

SBS. (2015). Glosario de términos e indicadores financieros. Obtenido de https://intranet2.sbs.gob.pe/estadistica/financiera/2015/Setiembre/SF-0002-se2015.PDF#:~:text=13.,iniciar%20acci%C3%B3n%20judicial%20o%20arbitral.

Strong, Lee, & Wang. (1997). Data quality in context.

Zachman. (1987). A framework for information systems architecture. *IBM System Journals*.

Apéndices

Apéndice I – Consultas para generar vistas y cubos en SQL

```
CREATE VIEW TF_A_1
AS
SELECT CARTERA,
        PRODUCTO,
        SUM(SALDO_CAPITAL) CAPITAL,
        SUM(SALDO_VENCIDO+SALDO_JUD)/SUM(SALDO_CAPITAL)
RATIO_VENCIDO,
                GROUPING(CARTERA) as INDICADOR_CARTERA,
                GROUPING(PRODUCTO) as INDICADOR_PRODUCTO
FROM BD_BANK_CRED
WHERE FECHA='2021-12-31' AND PRODUCTO NOT IN ('CARTA CREDITO','CARTA
FIANZA') AND CARTERA IN ('CONSUMO','HIPOTECARIO')
GROUP BY CUBE(CARTERA, PRODUCTO)


CREATE VIEW TF_A_2
AS
SELECT CARTERA,
        EQUIPO,
        SUM(SALDO_CAPITAL) CAPITAL,
        SUM(SALDO_VENCIDO+SALDO_JUD)/SUM(SALDO_CAPITAL)
RATIO_VENCIDO,
                GROUPING(CARTERA) as INDICADOR_CARTERA,
                GROUPING(CARTERA) as INDICADOR_EQUIPO
FROM BD_BANK_CRED
WHERE FECHA='2021-12-31' AND PRODUCTO NOT IN ('CARTA CREDITO','CARTA
FIANZA') AND CARTERA LIKE 'COMERCIAL'
GROUP BY CUBE(CARTERA, EQUIPO)
```

Apéndice II - Librerías utilizadas en Python

Apartado	Librería	Código de instalación
Apartado 2: Calidad de datos	Great Expectations	!pip install great_expectations import great_expectations as ge
Apartado 2: Calidad de datos	Acceso a archivos y directorio	import warnings import glob
Apartado 3: Modelos de machine learnig	Procesamiento de datos	import pandas as pd import seaborn as sb import numpy as np import statsmodels.api as sm import matplotlib.pyplot as plt from pandas_profiling import ProfileReport %matplotlib inline
Apartado 3: Modelos de machine learnig	Modelo de regresión logística	import statsmodels.api as sm
Apartado 3: Modelos de machine learnig	Modelo de árbol de decisión	from sklearn.datasets import load_boston from sklearn.ensemble import RandomForestClassifier from sklearn.metrics import accuracy_score from sklearn.metrics import confusion_matrix from sklearn.metrics import plot_confusion_matrix from sklearn.metrics import classification_report from sklearn import model_selection from sklearn.compose import ColumnTransformer from sklearn.preprocessing import OneHotEncoder from sklearn import tree from sklearn.model_selection import cross_val_score from sklearn.model_selection import KFold from sklearn.model_selection import train_test_split from sklearn.model_selection import RepeatedKFold from sklearn.model_selection import GridSearchCV from sklearn.model_selection import ParameterGrid from sklearn.tree import DecisionTreeClassifier from sklearn.ensemble import RandomForestClassifier from sklearn.metrics import roc_curve, roc_auc_score from sklearn.linear_model import LogisticRegression from sklearn import metrics

Reporte de mentoría

Trabajo Final de Especialización de Enma Gomez:

Mentor: Rodrigo del Rosso

El tema planteado en el trabajo final de la especialización es adecuado y se circunscribe a la problemática que enfrentan las distintas organizaciones, como en este caso vinculado al soporte analítico que reciben las entidades financieras, puntualmente una organización radicada en la República del Perú. Un aspecto vital es la construcción de un repositorio de datos que reúna las variables de riesgo más importantes para el análisis de los factores de riesgo más relevantes.

En mi opinión, el problema planteado en este trabajo final se encuentra debidamente justificado y es de relevancia para los fines buscados en la especialización. Rayza ha planteado en forma adecuada tanto objetivos como hipótesis, son consistentes y coherentes a la pregunta y tema de investigación planteados. El tema se encuentra totalmente articulado respecto al campo disciplinar de la especialización, se articula adecuadamente con las distintas materias de los programas y el tema propuesto atiende una problemática de gestión de grandes volúmenes de datos en organizaciones. Creo que el trabajo tendrá un aporte muy significativo y que será factible de ser realizado en el plazo previsto.

Ezequiel Rodrigo Javier Del Rosso