Evaluación de la calidad de datos mediante análisis exploratorio para la detección de inconsistencias en la base de datos de asociados de Promedico

Luis Alfredo Muñoz Ramirez

Asesor

Jorge Ignacio Blanco Blanco

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Basicas Tecnologia e Ingenieria ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

Este proyecto tiene como propósito aplicar un enfoque metodológico riguroso para la evaluación de la calidad de los datos contenidos en la base de datos de asociados del Fondo de Empleados Médicos de Colombia, PROMÉDICO. Dicha base, que almacena información crítica de más de 9.000 afiliados, es utilizada para la gestión de productos financieros, seguros y servicios institucionales. No obstante, se han identificado deficiencias en la integridad de los registros, lo que representa un riesgo operativo, normativo y reputacional para la organización.

La metodología empleada combina técnicas de Análisis Exploratorio de Datos (AED), validación cruzada, análisis de completitud, detección de valores atípicos y clasificación de registros según su origen (primarios, secundarios y otros). Se utilizará una muestra representativa calculada estadísticamente para garantizar la validez de los hallazgos. El análisis incluirá métricas de calidad como porcentaje de campos incompletos, tasa de duplicidad, frecuencia de errores por variable y consistencia intercampos.

Se espera identificar y cuantificar las principales inconsistencias, estimadas preliminarmente en un 67% de los registros, y proponer un conjunto de acciones correctivas como la implementación de validaciones automáticas en el sistema, auditorías periódicas, protocolos de actualización obligatoria y capacitación al personal encargado del ingreso de datos.

Los resultados permitirán establecer una línea base para el monitoreo continuo de la calidad de los datos, alineada con los requerimientos del SARLAFT y la Ley 1581 de 2012, fortaleciendo así la gobernanza de la información institucional.

Palabras claves: Calidad de los datos, PROMEDICO, análisis exploratorio de datos, valores atípicos, métricas de calidad y acciones correctivas.

Abstract

This project aims to apply a rigorous data analysis methodology to assess the quality of the information stored in the membership database of the Medical Employees Fund of Colombia, PROMÉDICO. This database, which contains critical records of over 9,000 affiliates, is essential for managing financial products, insurance, and institutional services. However, significant data integrity issues have been identified, posing operational, regulatory, and reputational risks.

The methodology integrates Exploratory Data Analysis (EDA), cross-validation, completeness assessment, outlier detection, and classification of records by origin (primary, secondary, and others). A statistically representative sample will be used to ensure the validity of the findings. The analysis will include quality metrics such as missing data rates, duplication frequency, variable-level error rates, and inter-field consistency.

The project expects to quantify the main inconsistencies—preliminarily estimated at 67% of records—and propose corrective actions such as automated validation rules within the SAP system, periodic audits, mandatory update protocols, and staff training.

The results will establish a baseline for continuous data quality monitoring, aligned with the requirements of SARLAFT and Law 1581 of 2012, thereby strengthening institutional data governance.

Keywords: Data quality, PROMEDICO, exploratory data analysis, outliers, quality metrics, corrective actions.

Contenido

Lista de Tablas

Lista de Figuras

Lista de Apendices

Introducción

En el contexto actual de transformación digital y exigencias normativas, la calidad de los datos se ha convertido en un pilar fundamental para la toma de decisiones estratégicas, la eficiencia operativa y el cumplimiento regulatorio. El Fondo de Empleados Médicos de Colombia, PROMÉDICO, administra una base de datos que contiene información crítica de más de 9.000 asociados, la cual es utilizada para la gestión de productos financieros, seguros, pólizas y otros servicios institucionales. Esta información, de carácter confidencial, debe cumplir con los lineamientos establecidos por la Ley 1581 de 2012 sobre protección de datos personales y por el marco normativo del SARLAFT, que exige mantener actualizada y verificada la información de los asociados.

No obstante, se ha identificado una problemática significativa relacionada con la integridad y consistencia de los registros almacenados. En una muestra preliminar de 100 asociados, se evidenció que aproximadamente el 67% de los registros presentan inconsistencias, errores de diligenciamiento o desactualización. Esta situación representa un riesgo para la organización, tanto en términos legales como operativos, y pone en evidencia la necesidad de implementar mecanismos de control y mejora continua sobre la calidad de los datos.

Este proyecto tiene como propósito aplicar técnicas avanzadas de análisis de datos, incluyendo el Análisis Exploratorio de Datos (AED), validación cruzada, análisis de completitud y consistencia intercampos, con el fin de evaluar el estado actual de la base de datos y proponer soluciones concretas. La metodología contempla el uso de herramientas estadísticas y computacionales para la detección de valores atípicos, duplicados, campos vacíos y errores sistemáticos, así como la clasificación de los registros según su origen (primarios, secundarios y otros).

El objetivo final es establecer una línea base de calidad de datos que permita a PROMÉDICO implementar estrategias de mejora sostenibles, como validaciones automáticas en el sistema, auditorías periódicas, protocolos de actualización obligatoria y capacitación al personal. Este enfoque no solo busca mitigar los riesgos actuales, sino también fortalecer la gobernanza de la información institucional y garantizar la trazabilidad, transparencia y confiabilidad de los datos en el largo plazo.

Justificación

En la actualidad, El Fondo Medico de Empleados Promédico cuenta con una base de 9.110 asociados, lo que representa un volumen significativo de información personal, financiera y administrativa. La integridad y actualización permanente de estos datos no solo es esencial para garantizar la eficiencia operativa y la toma de decisiones estratégicas dentro de la organización, sino que también constituye un requisito legal obligatorio para las entidades del sector solidario.

De acuerdo con lo establecido en el numeral 3.2.2.3.1.1 del Título V, Capítulo IV de la Circular Básica Jurídica, expedida por la Superintendencia de la Economía Solidaria, las organizaciones vigiladas deben implementar mecanismos que aseguren la identificación adecuada, permanente y actualizada de sus asociados, clientes, beneficiarios finales y proveedores (Supersolidaria, 2021). Esta disposición hace parte del marco normativo para la prevención del riesgo de lavado de activos y financiación del terrorismo (SARLAFT). En donde la norma establece que: “La identificación del asociado, cliente o proveedor, actual o potencial, implica conocer y contar de manera permanente y actualizada con la información necesaria para establecer su perfil y monitorear sus operaciones.”

Esto significa que cualquier omisión o desactualización en los datos puede representar un riesgo legal, operativo y reputacional para la organización. Además, el cumplimiento de esta normativa permite a Promédico:

- Garantizar la trazabilidad y transparencia en las relaciones con sus asociados.
- Prevenir el uso indebido de la entidad para actividades ilícitas.
- Cumplir con los estándares de supervisión exigidos por la Superintendencia.
- Fortalecer la confianza de los asociados y partes interesadas en la gestión institucional.

Por lo tanto, mantener actualizadas las bases de datos no es solo una buena práctica administrativa, sino una obligación legal y ética que contribuye directamente al bienestar de los asociados y a la sostenibilidad de la organización.

Descripción del Problema

PROMÉDICO gestiona una base de datos con información crítica de más de 9.000 asociados. Se han identificado inconsistencias en el 67% de los registros, incluyendo campos incompletos, errores de formato, duplicados y desactualización. Esta situación representa riesgos operativos, normativos y reputacionales.

Planteamiento del Problema

La falta de validaciones automáticas, protocolos de actualización y auditorías internas ha permitido la persistencia de errores en los datos. Esto afecta la toma de decisiones, el cumplimiento legal y la eficiencia institucional.

Sistematización del Problema

Se abordan cinco preguntas clave:

¿Cuál es el estado actual de la calidad de los datos?

¿Qué tipos de errores son más frecuentes?

¿Qué técnicas permiten diagnosticar las deficiencias?

¿Cuál es el impacto de mantener datos de baja calidad?

¿Qué acciones correctivas pueden implementarse?

Objetivos

Objetivo General

Desarrollar un diagnóstico integral de la calidad de los datos en la base de asociados de PROMÉDICO

Objetivos Específicos

Caracterizar el estado actual de la base de datos institucional, evaluando aspectos como completitud, unicidad, coherencia y validez de los registros, mediante técnicas estadísticas y visuales propias del análisis exploratorio de datos (AED).

Detectar inconsistencias estructurales y registros atípicos a través de algoritmos avanzados como Isolation Forest, análisis de similitud de cadenas (para duplicados) y validaciones cruzadas intercampos, que permitan identificar anomalías relevantes en los datos personales, financieros y administrativos.

Aplicar y evaluar indicadores de calidad de datos (completitud, consistencia, duplicidad, validez y exactitud) para establecer una línea base que sirva como punto de partida para la mejora continua de la gestión de la información.

Revisar el cumplimiento del marco normativo vigente en materia de protección de datos personales, incluyendo la Ley 1581 de 2012, los lineamientos del SARLAFT y la Circular Básica Jurídica, garantizando que el tratamiento de la información contemple principios de confidencialidad, anonimización y trazabilidad.

Proponer recomendaciones técnicas, operativas y organizacionales, que incluyan la implementación de validaciones automáticas en sistemas, protocolos de actualización periódica, auditorías internas, y estrategias de formación para el personal encargado del manejo de datos.

Contribuir al fortalecimiento de una cultura organizacional orientada a la calidad de datos, que permita consolidar un entorno institucional confiable, transparente y alineado con los estándares regulatorios y estratégicos del sector solidario.

Marco de Referencia

Estado del Arte

La calidad de los datos se ha convertido en un eje estratégico para las organizaciones modernas, especialmente en sectores regulados como el financiero y el solidario. En este contexto, el análisis exploratorio de datos (AED) ha emergido como una herramienta fundamental para diagnosticar, visualizar y comprender la estructura y los problemas subyacentes en grandes volúmenes de información. PROMÉDICO, como fondo de empleados del sector salud, enfrenta desafíos significativos en la gestión de su base de datos de más de 9.000 asociados, donde se han identificado inconsistencias en aproximadamente el 67% de los registros.

Teniendo en cuenta los fundamentos teóricos, metodológicos y tecnológicos que sustentan la evaluación de la calidad de datos, con énfasis en el uso de técnicas de AED, validación cruzada, detección de outliers y clasificación de errores, enmarcados en la normativa colombiana vigente.

La calidad de los datos es un concepto multidimensional que se refiere al grado en que los datos son adecuados para su propósito de uso. Según Batini y Scannapieco (2016), la calidad de los datos se refiere a la adecuación de los datos para satisfacer los requisitos de los usuarios en contextos específicos.Completitud: Grado en que los datos requeridos están presentes.

Exactitud: Concordancia entre los datos y los valores reales del mundo.

Consistencia: Ausencia de contradicciones entre diferentes conjuntos de datos.

Validez: Conformidad con formatos, dominios y reglas de negocio.

Unicidad: No duplicación de registros que deberían ser únicos.

Actualización: Vigencia de los datos respecto al tiempo.

El concepto de calidad ha evolucionado desde la antigüedad, pasando por la inspección artesanal en la Edad Media, hasta los sistemas de control estadístico en la Revolución Industrial. En el siglo XX, figuras como W. Edwards Deming y Joseph Juran introdujeron enfoques sistemáticos de gestión de calidad, que luego se adaptaron al ámbito de los datos con la llegada de los sistemas de información. Según Purón-Rodríguez y Tejeda-Marrero (2021), la calidad ha sido históricamente un concepto polisémico, influenciado por contextos culturales, económicos y tecnológicos.

Existen varios marcos de referencia ampliamente utilizados a nivel internacional:

DAMA-DMBOK: Define la calidad de datos como una función de múltiples dimensiones gestionadas dentro del gobierno de datos.

ISO 8000: Norma internacional que establece requisitos para la calidad de datos maestros.

TDQM (Total Data Quality Management): Modelo desarrollado por el MIT que integra calidad de datos en los procesos organizacionales, promoviendo la mejora continua.

Estos modelos ofrecen metodologías estructuradas para evaluar, monitorear y mejorar la calidad de los datos, y son aplicables en contextos empresariales como el colombiano, donde las organizaciones deben cumplir con normativas como la Ley 1581 de 2012 y el SARLAFT.

En Colombia, diversas organizaciones han enfrentado desafíos relacionados con la calidad de datos. Un estudio publicado por Redalyc analiza cómo empresas del sector financiero y de salud han implementado auditorías internas, validaciones cruzadas y herramientas de análisis exploratorio para mejorar la integridad de sus bases de datos.

En el caso de PROMÉDICO, se identificó que el 67% de los registros presentaban inconsistencias, lo que motivó la aplicación de técnicas como el análisis exploratorio de datos

(AED), algoritmos de detección de duplicados y modelos de machine learning para clasificar registros según su confiabilidad.

Actualmente, la calidad de datos se apoya en tecnologías como:

Big Data y Data Lakes: Que requieren nuevas estrategias de validación en tiempo real.

Machine Learning: Para detección automática de anomalías y predicción de errores.

Herramientas especializadas: Como Talend, Informatica, OpenRefine y Data Ladder, que permiten limpiar, transformar y monitorear datos de forma automatizada.

Estas herramientas son cada vez más accesibles para empresas colombianas, especialmente aquellas que operan bajo marcos regulatorios estrictos como el financiero o el de salud.

Entre los principales obstáculos para implementar estrategias de calidad de datos en Colombia se encuentran:

Falta de cultura organizacional orientada a los datos.

Sistemas de información fragmentados.

Ausencia de roles definidos como el Chief Data Officer (CDO).

Limitaciones presupuestales y tecnológicas.

Superar estos desafíos requiere liderazgo institucional, inversión en capacitación y adopción de marcos de gobernanza de datos.

La calidad de los datos es un prerrequisito para la analítica avanzada. Modelos de predicción, segmentación de clientes o evaluación de riesgo crediticio dependen de datos limpios, completos y actualizados. En PROMÉDICO, por ejemplo, la baja calidad de datos limita la implementación de modelos de scoring crediticio o análisis de comportamiento de asociados.

Marco Contextual

En el contexto actual de transformación digital, las organizaciones enfrentan una creciente presión para gestionar adecuadamente sus activos de información. La calidad de los datos se ha convertido en un factor crítico para garantizar la eficiencia operativa, la toma de decisiones estratégicas y el cumplimiento normativo. Esta necesidad es aún más apremiante en sectores regulados como el financiero, el solidario y el de la salud, donde la información personal, financiera y administrativa debe ser precisa, actualizada y verificable.

PROMÉDICO, el Fondo de Empleados Médicos de Colombia, administra una base de datos con información de más de 9.000 asociados. Esta base es utilizada para la gestión de productos financieros, seguros, pólizas y beneficios institucionales. Sin embargo, un análisis preliminar reveló que aproximadamente el 67% de los registros presentan inconsistencias, lo que representa un riesgo significativo desde el punto de vista operativo, legal y reputacional.

Este escenario se enmarca en un entorno regulatorio exigente. La Ley 1581 de 2012 sobre protección de datos personales y el SARLAFT (Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo) exigen a las organizaciones mantener información veraz, completa y actualizada. El incumplimiento de estas normativas puede acarrear sanciones por parte de la Superintendencia de la Economía Solidaria y afectar la confianza de los asociados.

A nivel internacional, modelos como DAMA-DMBOK, ISO 8000 y TDQM (Total Data Quality Management) han establecido marcos de referencia para la gestión de la calidad de datos. Estos modelos destacan dimensiones como completitud, exactitud, validez, consistencia, unicidad y actualización, las cuales son fundamentales para evaluar el estado de una base de

datos y diseñar estrategias de mejora continua (Batini & Scannapieco, 2016; Wang & Strong, 1996).

En Colombia, investigaciones como la de Sanabria Rangel et al. (2014) han abordado el concepto de calidad desde una perspectiva organizacional compleja, destacando que la calidad no puede entenderse solo desde lo técnico, sino también desde lo cultural, lo estratégico y lo humano 1. Asimismo, estudios recientes han evidenciado que tanto el sector público como el privado enfrentan desafíos similares en la gestión de datos, especialmente ante la implementación de tecnologías como Big Data y analítica avanzada 2.

PROMÉDICO, como organización del sector solidario, se encuentra en una posición estratégica para liderar procesos de transformación digital basados en la mejora de la calidad de sus datos. La aplicación de técnicas como el análisis exploratorio de datos (AED), la validación cruzada, la detección de duplicados y el uso de algoritmos de machine learning no solo permitirá corregir errores actuales, sino también establecer una línea base para el monitoreo continuo de la calidad de la información.

Este proyecto se desarrolla, por tanto, en un entorno donde convergen múltiples factores: la necesidad de cumplir con estándares regulatorios, la urgencia de mejorar la eficiencia operativa, la oportunidad de aprovechar tecnologías emergentes y la responsabilidad ética de proteger los datos personales de los asociados. En este marco, la calidad de los datos no es solo un requisito técnico, sino un pilar fundamental para la sostenibilidad institucional y la confianza organizacional.

Marco Teórico

En el contexto actual de transformación digital y cumplimiento normativo, la calidad de los datos se ha convertido en un pilar fundamental para la sostenibilidad operativa y estratégica de

las organizaciones del sector solidario. PROMÉDICO, el Fondo de Empleados Médicos de Colombia, administra una base de datos con información personal, financiera y administrativa de más de 9.000 asociados (Promedico). Esta base de datos es utilizada como insumo principal para la gestión de productos financieros, seguros, pólizas y beneficios institucionales. Sin embargo, se ha identificado una problemática crítica relacionada con la integridad, completitud y consistencia de los registros almacenados.

Se determinará la muestra necesaria de la cantidad de datos a analizar con la formula estadística de muestreo.

Como calcular el tamaño de una muestra:

Normalmente, los estudios científicos se basan en encuestas distribuidas sobre una muestra de alguna población más grande. Sin embargo, la muestra deberá incluir cierta cantidad mínima de personas para que esta refleje las condiciones de la población de estudio que se supone que representa. Para calcular de qué tamaño debe ser la muestra, es necesario determinar ciertos valores fijos para introducirlos en la fórmula adecuada.

Se debe conocer el tamaño de la población. Es decir, el número total de individuos que conforman la población objetivo. El impacto estadístico de la precisión de las medidas tomadas se hace mayor a medida que el grupo de estudio se hace más pequeño. Mientras más grande sea la población, se usan aproximaciones con más libertad.

Determinar el margen de error. El margen de error, también conocido como “intervalo de confianza”, es el tamaño del error a aceptar en los resultados. El margen de error es un porcentaje que indica qué tan cerca estarán los resultados de tu muestra del valor real de la población general objeto del estudio. Mientras más pequeño sea el margen de error, las medidas serán más exactas, pero la elección de un margen de error más pequeño requerirá que se tome una muestra más grande.

Al presentar los resultados, el margen de error se denota usualmente como un porcentaje que es positivo y negativo a la vez.

Establecer nivel de confianza. El nivel de confianza está relacionado íntimamente con el intervalo de confianza (margen de error). Este valor mide el grado de certidumbre de la medida en cuanto a qué tan bien una muestra representa a la población objetivo, respetando el margen de error establecido. Dicho de otro modo, la elección de un nivel de confianza de 95 % te permite decir que estás 95 % seguro de que los resultados que obtuviste respetan el margen de error que escogiste previamente. Mientras mayor sea el nivel de confianza, mayor será el grado de exactitud que tendrán tus resultados, pero a la vez requerirá la muestra sea más grande. Los niveles de confianza usados con más frecuencia son de 90 %, de 95 % y de 99 % de confianza.

Especifica la desviación estándar. Indica qué tanta variación se espera en los resultados. Dicho con claridad, si el 99 % de la población, dice que sí y solo el 1 % dice que no, es probable que la muestra represente a la población general con bastante exactitud. Por otro lado, si tienes algo más balanceado, como 45 % diciendo que sí y 55 % diciendo que no, hay una mayor probabilidad de que existan errores, Es recomendable fijar este valor a 0,5 (50 %). Este es el porcentaje que describe el peor escenario, así que fijar este valor asegura que la muestra calculada será lo suficientemente grande como para representar a la población general con exactitud mientras los valores se mantienen dentro del intervalo y nivel de confianza escogidos al mismo tiempo.

Encontrar el valor Z. El valor Z es un valor constante que se fija de manera automática al elegir el nivel de confianza. Es un indicador del “valor normal estándar”, o lo que es lo mismo, la cantidad de desviaciones estándar entre cualquier valor seleccionado y el valor promedio de la

población. Ya que los niveles de confianza se han estandarizado, a continuación, se nombren los niveles de confianza más comunes:

80 % confianza => valor z = 1,28

85 % confianza => valor z = 1,44

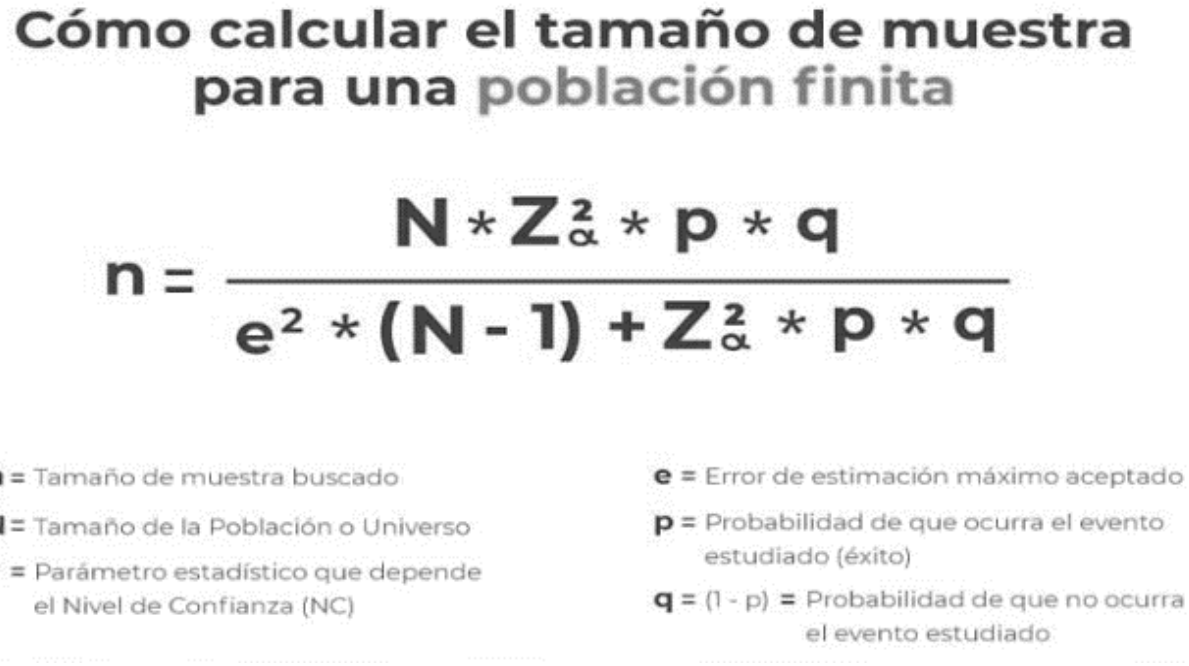
90 % confianza => valor z = 1,65

95 % confianza => valor z = 1,96

99 % confianza => valor z = 2,58

Figura 1

Formula Calculo Tamaño de la Muestra



Nota. Imagen tomada de: (ProbabilidadyEstadistica.net. (s.f.).)

Este análisis preliminar sobre una muestra aleatoria de 100 registros reveló que el 67% de los datos presentan algún tipo de inconsistencia. Estas inconsistencias incluyen:

Campos obligatorios incompletos (como direcciones o correos electrónicos).

Errores de formato en identificadores únicos (como números de documento con caracteres no válidos).

Duplicidad de registros

Desactualización de variables críticas (como ingresos o beneficiarios)

Discordancias entre campos relacionados (por ejemplo, fechas de nacimiento incompatibles con edad declarada).

La falta de validaciones automáticas en el sistema SAP, la inexistencia de protocolos de actualización periódica y la ausencia de auditorías internas sistemáticas han contribuido a la persistencia de estos errores.

Para abordar esta problemática, se propone una estrategia metodológica basada en técnicas avanzadas de análisis de datos.

Análisis Exploratorio de Datos (AED) para caracterizar la distribución, dispersión y comportamiento de las variables. Esto incluirá el uso de estadísticas descriptivas (media, mediana, desviación estándar) y visualizaciones gráficas como histogramas, diagramas de caja y mapas de calor para detectar patrones atípicos. Por ejemplo, se podrá identificar si existen valores extremos en los ingresos reportados por los asociados o si hay campos con tasas de completitud inferiores al 80%. El objetivo del AED es familiarizarse con los datos y generar hipótesis, no probarlas. Este es un paso previo y fundamental antes de cualquier modelado estadístico o de machine learning. Consiste en una fase crucial en la investigación estadística que implica una serie de técnicas para explorar y resumir las características principales de un conjunto de datos. Los pasos clave para seguir en un AED son los siguientes:

Análisis Descriptivo: Incluye la descripción de las variables mediante estadísticas resumen como la media, mediana, modos, rangos y desviaciones estándar. También se considera la visualización de datos con gráficos como histogramas, diagramas de caja o gráficos de dispersión para entender la distribución y la relación entre las variables.

Limpieza de Datos: Se identifican y tratan los valores faltantes, atípicos o incorrectos. Esto puede implicar la imputación de datos faltantes.

Transformación de Variables: A veces es necesario transformar las variables para mejorar la normalidad o linealidad en los análisis. Esto puede incluir logaritmos, raíces cuadradas o cualquier otra transformación matemática adecuada.

Análisis de Correlación: Se examina la relación entre variables para identificar posibles correlaciones. Esto ayuda a entender cómo una variable puede influir en otra y es fundamental para la selección de características en modelos predictivos.

Reducción de Dimensionalidad: Técnicas como el análisis de componentes principales (PCA) se utilizan para reducir el número de variables en el conjunto de datos sin perder información importante. (datos.gob.es)

Proceso de validación cruzada de campos, comparando los datos internos con fuentes externas confiables (como bases de datos tributarias o registros civiles) y aplicando reglas de negocio para verificar la coherencia entre variables. Por ejemplo, se podrá validar que el número de hijos reportado sea coherente con la edad del asociado o que el tipo de afiliación corresponda con el nivel de ingresos.

Técnicas de detección de valores atípicos mediante algoritmos como el Isolation Forest o el Z-score, que permiten identificar registros que se desvían significativamente del comportamiento esperado. Por ejemplo, un asociado con ingresos reportados de $1.000.000 y egresos de $5.000.000 podría ser considerado un caso atípico que requiere verificación.

Análisis de completitud para calcular el porcentaje de campos vacíos por variable y por registro, y un análisis de consistencia intercampos, que evalúe la lógica entre atributos relacionados.

Para la reducción de redundancia, se utilizarán técnicas de duplicación basadas en similitud de cadenas (como Levenshtein o Jaccard) para identificar registros potencialmente duplicados, como asociados con nombres similares y documentos parcialmente coincidentes.

El impacto de esta problemática es multifactorial: desde la generación de reportes inexactos y la toma de decisiones basada en datos erróneos, hasta el incumplimiento de normativas como la Ley 1581 de 2012 sobre protección de datos personales y los lineamientos del SARLAFT. Además, la baja calidad de los datos limita la posibilidad de implementar modelos analíticos avanzados, como segmentación de usuarios, análisis predictivo o evaluación de riesgo crediticio.

Por tanto, este proyecto busca no solo diagnosticar el estado actual de la base de datos, sino también establecer una línea base de calidad de datos y proponer un conjunto de acciones correctivas. Estas incluirán la implementación de validaciones automáticas en el sistema, auditorías periódicas, protocolos de actualización obligatoria, y capacitación al personal encargado del ingreso de datos.

El objetivo final es garantizar la integridad, trazabilidad y confiabilidad de la información institucional, fortaleciendo así la gobernanza de datos en PROMÉDICO y asegurando el cumplimiento de los estándares regulatorios y operativos exigidos por la Superintendencia de la Economía Solidaria.

Marco Conceptual

Calidad de los Datos:La calidad de los datos es un concepto fundamental para la sostenibilidad operativa y estratégica de las organizaciones, especialmente en el contexto de la transformación digital y el cumplimiento normativo. Se refiere al grado en que los datos son adecuados para su propósito de uso. Según Batini y Scannapieco (2016), la calidad se define

como la adecuación de los datos para satisfacer los requisitos de los usuarios en contextos específicos.

La calidad de los datos se evalúa a través de múltiples dimensiones:

Completitud: Grado en que los datos requeridos están presentes. Se mide por el porcentaje de campos no nulos.

Validez: Conformidad con formatos, dominios y reglas de negocio. Se evalúa si los formatos son correctos y si se cumplen las reglas de negocio.

Consistencia: Ausencia de contradicciones entre diferentes conjuntos de datos. Se verifica la coherencia entre campos relacionados (por ejemplo, edad vs. fecha de nacimiento).

Unicidad: No duplicación de registros que deberían ser únicos. Se mide por el número de duplicados detectados.

Actualización: Vigencia de los datos respecto al tiempo. Se evalúa la antigüedad de los registros.

Análisis Exploratorio de Datos (AED): El Análisis Exploratorio de Datos (AED) es una fase crucial en la investigación estadística que implica una serie de técnicas para explorar y resumir las características principales de un conjunto de datos. El objetivo del AED es familiarizarse con los datos y generar hipótesis, no probarlas.

Las técnicas de AED incluyen:

Análisis Descriptivo: Descripción de variables mediante estadísticas resumen como la media, mediana y desviación estándar.

Visualización de Datos: Uso de gráficos como histogramas, diagramas de caja y mapas de calor para entender la distribución y la relación entre las variables.

Detección de Valores Atípicos (Outliers): Identificación de registros que se desvían significativamente del comportamiento esperado, utilizando algoritmos como Z-score, IQR e Isolation Forest.

Ley 1581 de 2012: Esta ley establece las disposiciones generales para la protección de datos personales. Exige a las organizaciones mantener información veraz, completa y actualizada.

SARLAFT: El Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo requiere información actualizada y verificable para mitigar riesgos.

Circular Básica Jurídica: Exige una identificación adecuada y permanente de los asociados. El incumplimiento de estas normativas puede acarrear sanciones por parte de la Superintendencia de la Economía Solidaria y afectar la confianza de los asociados.

Marco Normativo

El tratamiento de datos personales en Colombia se encuentra regulado principalmente por la Ley 1581 de 2012, la cual establece disposiciones generales para la protección de los datos personales y desarrolla el derecho constitucional que tienen todas las personas a conocer, actualizar y rectificar la información que se haya recogido sobre ellas en bases de datos o archivos. Esta ley es complementada por el Decreto 1377 de 2013 y la jurisprudencia de la Corte Constitucional, que han definido principios, derechos y obligaciones para los responsables y encargados del tratamiento.

En el contexto del presente proyecto, la base de datos de PROMÉDICO contiene información de carácter personal, sensible y financiero de más de 9.000 asociados. Por tanto, su tratamiento debe cumplir con los principios de legalidad, finalidad, libertad, veracidad, transparencia, acceso y circulación restringida, seguridad y confidencialidad. Además, se deben

aplicar medidas técnicas y organizativas que garanticen la protección de los datos durante todo el ciclo de vida del proyecto.

Clasificación y sensibilidad de los datos

Los datos contenidos en la base institucional de PROMÉDICO pueden clasificarse en tres niveles de sensibilidad:

Datos personales básicos: nombre, documento de identidad, dirección, correo electrónico, teléfono.

Datos financieros: ingresos, egresos, historial crediticio, productos financieros adquiridos.

Datos sensibles: estado de salud, beneficiarios, estado civil, información biométrica (si aplica).

De acuerdo con el artículo 5 de la Ley 1581, los datos sensibles requieren un tratamiento reforzado, y su uso debe estar estrictamente limitado a los fines autorizados por el titular, con medidas de seguridad adicionales.

Medidas de anonimización y seudonimización

Dado que el análisis de calidad de datos implica el acceso a información identificable, se implementarán técnicas de anonimización y seudonimización para proteger la identidad de los titulares durante el procesamiento. Estas técnicas incluyen:

Anonimización irreversible: eliminación de identificadores directos (nombres, cédulas) y reemplazo por códigos aleatorios no trazables.

Seudonimización temporal: sustitución de identificadores por claves cifradas que solo pueden ser revertidas por el responsable del tratamiento.

Enmascaramiento de datos: ocultamiento parcial de campos sensibles (mostrar solo los últimos 4 dígitos del documento).

Cifrado de archivos: uso de algoritmos como AES-256 para proteger los archivos durante el almacenamiento y la transmisión.

Estas medidas serán implementadas mediante scripts en Python utilizando bibliotecas como hashlib, cryptography y pandas.

Protocolos de manejo de datos sensibles

Para garantizar el cumplimiento de los principios de seguridad y confidencialidad, se establecerán los siguientes protocolos:

Acceso restringido a los datos: solo el equipo investigador autorizado podrá acceder a la base de datos, mediante credenciales individuales y autenticación de dos factores.

Registro de accesos: se mantendrá un log de todas las operaciones realizadas sobre la base de datos, incluyendo fecha, usuario y tipo de acción.

Entorno seguro de análisis: el procesamiento se realizará en entornos locales o servidores institucionales con control de acceso, evitando el uso de plataformas públicas o almacenamiento en la nube sin cifrado.

Eliminación segura: Al finalizar el proyecto, los datos serán eliminados mediante procedimientos de borrado seguro (sobrescritura múltiple).

Consentimiento informado y derechos del titular: De acuerdo con el principio de libertad, el tratamiento de datos personales solo puede realizarse con el consentimiento previo, expreso e informado del titular. En este sentido, PROMÉDICO deberá garantizar que todos los asociados hayan firmado el formato de autorización para el tratamiento de sus datos, especificando:

Finalidad del tratamiento (análisis de calidad de datos con fines institucionales).

Derechos del titular (acceso, rectificación, supresión, revocatoria).

Canales de contacto para ejercer sus derechos (correo, línea telefónica, oficina física).

En caso de que se identifiquen registros sin autorización válida, estos deberán ser excluidos del análisis o tratados únicamente con datos anonimizados.

El cumplimiento de estas medidas será supervisado por el Oficial de Protección de Datos Personales de PROMÉDICO, quien deberá verificar que el proyecto se ajuste a la política interna de tratamiento de datos y a las disposiciones de la Superintendencia de Industria y Comercio (SIC). Además, se recomienda realizar una Evaluación de Impacto en Protección de Datos (DPIA) si se identifican riesgos significativos para los derechos de los titulares.

Metodología

La metodología se basa en un enfoque estructurado de cinco fases principales, que combinan técnicas de análisis exploratorio, validación de calidad de datos, y algoritmos de machine learning. El propósito es diagnosticar de forma exhaustiva la calidad de la base de datos de PROMÉDICO, identificar las inconsistencias y proponer soluciones concretas y sostenibles.

Metodo

El método de investigación es de tipo cuantitativo, ya que se centra en la medición, evaluación y cuantificación de las inconsistencias en la base de datos. Se utiliza una estrategia analítica que va desde la exploración inicial de los datos hasta la aplicación de modelos avanzados de aprendizaje automático. La secuencia de pasos garantiza un diagnóstico robusto y la generación de recomendaciones basadas en evidencia. El proceso es reproducible y escalable, lo que permite su implementación en otras bases de datos de la organización.Recolección y preparación de datos:

Tipo de Estudio

El estudio es de tipo descriptivo y explicativo:

Descriptivo: Se caracteriza por la descripción y evaluación del estado actual de la base de datos de asociados de PROMÉDICO. A través de indicadores de calidad y estadísticas descriptivas, se busca caracterizar la naturaleza y la magnitud de los problemas de datos. El objetivo principal en esta fase es responder a la pregunta: ¿cuál es el estado de la calidad de los datos de PROMÉDICO?

Explicativo: Se busca identificar las causas subyacentes de las inconsistencias, como la falta de validaciones en el sistema SAP o la ausencia de protocolos de actualización. Además, a

través de modelos predictivos y de clustering, se pretende explicar por qué ciertos registros son más propensos a tener errores y cómo se agrupan estos patrones de inconsistencia

Recolección y Preparación de Datos

Extracción de la Base de Datos: Se obtendrá la base de datos de los asociados directamente del sistema institucional SAP. La extracción se realizará en un formato estructurado, preferiblemente CSV o a través de una conexión SQL, para garantizar la integridad de la información.

Carga de los Datos: Los datos extraídos serán cargados en entornos de análisis robustos. Se utilizará Python, con bibliotecas especializadas como pandas para la gestión de DataFrames y sqlalchemy para la conexión y consulta de bases de datos. El entorno de desarrollo será Jupyter Notebook, que facilita el análisis interactivo y la documentación del proceso.

Manipulación y Estandarización: Una vez cargados los datos, se llevarán a cabo tareas de estandarización y limpieza. Esto incluye:

Normalización de formatos: Asegurar que todas las variables (por ejemplo, fechas, números de identificación) sigan un formato consistente.

Codificación de variables categóricas: Convertir variables cualitativas (por ejemplo, tipo de afiliación) en un formato numérico que pueda ser interpretado por los modelos de machine learning.

Conversión de tipos de datos: Ajustar los tipos de datos de cada columna (por ejemplo, de texto a numérico o fecha) para que sean compatibles con las operaciones estadísticas y de análisis.

Tabla 1

Tipos de Datos

Criterio	Descripción técnica	Métrica de evaluación
Completitud	Porcentaje de campos no nulos por variable y por registro	% de completitud por campo
Validez	Conformidad con formatos esperados (correos, cédulas, fechas)	% de registros válidos por campo
Consistencia	Coherencia entre campos relacionados (edad vs. fecha de nacimiento)	% de reglas de negocio cumplidas
Unicidad	Ausencia de duplicados en claves primarias o combinaciones únicas	Número de duplicados detectados
Actualización	Antigüedad de los datos respecto a la fecha de última modificación	% de registros con antigüedad > 12 meses

Nota. Analisis de datos según sus criterios (Autor 2025)

Auditoría de Cumplimiento: Se verificará la coherencia de los datos con el marco normativo (Ley 1581 de 2012 y SARLAFT), prestando especial atención a los datos sensibles y el tratamiento adecuado de la información personal de los asociados.

Validación cruzada y de duplicación

Comparación de registros con fuentes externas (DIAN, Registraduría) mediante APIs o archivos de referencia.

Aplicación de reglas de negocio para validar coherencia lógica ingresos > egresos, edad mínima para crédito).

Uso de algoritmos de similitud de texto (fuzzywuzzy, difflib, Levenshtein) para detección de duplicados parciales.

Diseño de soluciones y recomendaciones:

Propuesta de validaciones automáticas en SAP (campos obligatorios, formatos, rangos).

Diseño de un protocolo de actualización periódica (cada 12 meses) con alertas automatizadas.

Recomendación de auditorías internas trimestrales de calidad de datos.

Capacitación al personal en buenas prácticas de captura y verificación de datos.

Métricas de evaluación del éxito

Para evaluar el impacto de las acciones propuestas, se definirán indicadores clave de desempeño (KPIs):

Reducción del porcentaje de registros con errores (meta: <10%).

Aumento del porcentaje de completitud por campo (meta: >95%).

Disminución del número de duplicados (meta: 0 duplicados en claves primarias).

Cumplimiento de reglas de negocio (meta: >98% de consistencia intercampos).

Tasa de actualización anual de registros (meta: >90%).

Además de las técnicas estadísticas y de validación descritas anteriormente, se incorporarán métodos de aprendizaje automático (machine learning) para fortalecer el diagnóstico y la predicción de errores en la base de datos. Estas técnicas permitirán automatizar la detección de anomalías, clasificar registros según su nivel de confiabilidad y predecir la probabilidad de error en campos críticos.

Se emplearán los siguientes algoritmos de machine learning, implementados en Python mediante bibliotecas como scikit-learn, xgboost, lightgbm y tensorflow:

Modelos de clasificación supervisada cuyo objetivo principal es el de predecir si un registro es confiable o contiene errores, en función de sus atributos, lo anterior realizando la implementación de los siguientes algoritmos sugeridos:

Árboles de decisión (DecisionTreeClassifier)

Bosques aleatorios (RandomForestClassifier)

Gradient Boosting (XGBoost, LightGBM)

Redes neuronales simples (MLPClassifier)

Variables objetivo: Etiqueta binaria generada a partir de auditorías manuales (1 = inconsistente, 0 = confiable). Variables predictoras: Edad, ingresos, egresos, número de beneficiarios, tipo de afiliación, etc.

En este aspecto de Machine learning se clasifican las siguientes métricas de evaluación:

Precisión (accuracy)

Matriz de confusión

AUC-ROC

F1-score

Modelos de detección de anomalías (unsupervised), en donde permitirá identificar registros atípicos sin necesidad de etiquetas previas. Lo anterior con los siguientes algoritmos sugeridos:

Isolation Forest (sklearn.ensemble.IsolationForest)

One-Class SVM (sklearn.svm.OneClassSVM),

Autoencoders (con tensorflow.keras)

De esta manera será posible detectar combinaciones inusuales de ingresos y egresos, edades fuera de rango, o patrones de datos que no se ajustan al comportamiento general.

En este aspecto de los modelos de detección de anomalías se clasifican las siguientes métricas de evaluación:

Porcentaje de registros detectados como outliers

Visualización en 2D con PCA o t-SNE

Clustering para segmentación de calidad cuyo objetivo principal es el de agrupar registros según su perfil de calidad y comportamiento. Los algoritmos sugeridos son:

- K-Means (sklearn.cluster.KMeans)

- DBSCAN (sklearn.cluster.DBSCAN)

- Gaussian Mixture Models (sklearn.mixture.GaussianMixture)

Con esta técnica será posible identificar segmentos de asociados con alta probabilidad de inconsistencias o con patrones de datos similares cuyas métricas de evaluación son:

- Silhouette Score

- Calinski-Harabasz Index

- Interpretación de centroides

Para cada modelo implementado, se realizará una validación cruzada con cross_val_score y se utilizará GridSearchCV para optimizar hiperparámetros. Los modelos serán entrenados y evaluados sobre conjuntos de entrenamiento y prueba (70/30), y se documentarán los resultados con gráficos y reportes automáticos (classification_report, confusion_matrix).

Resultados

Primer Resultado

Modelo Análisis Exploratorio de Datos (AED) y Diagnóstico Inicial

Caracterización inicial del estado de la base de datos, identificando patrones, distribuciones y anomalías generales mediante técnicas estadísticas y de visualización. El análisis exploratorio sirvió como base para una comprensión profunda del problema de calidad.

Análisis Descriptivo: Se calculó un resumen estadístico de las variables numéricas clave (ej. edad, ingresos, egresos), incluyendo la media, mediana, moda y desviación estándar. Estos valores permitieron identificar el rango típico de cada variable y detectar posibles sesgos en los datos. Se puede incluir una tabla como la siguiente:

Tabla 2

Estadisticas Descriptivas

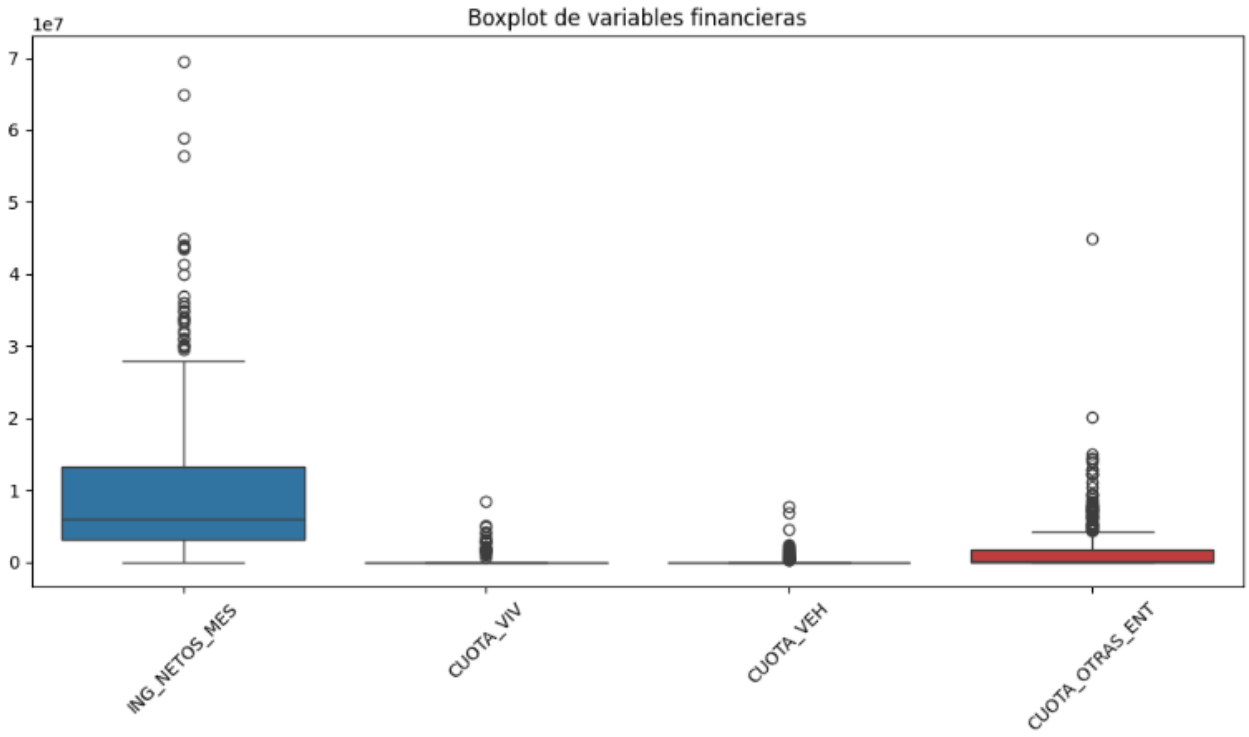
Variable	Media	Mediana	Desv. Estándar	Mínimo	Máximo
Edad	42.06	39.0	13.32	23	89
Ingresos Netos	$9,571,098	$6,000,000	$10,305,596	0	$69,574,333
Cuota Vivienda	$ 141,18	0	$ 700,71	0	$8,452,000
Cuota Vehículo	$ 93,64	0	$ 570,68	0	$7,760,000
Cuota Otras Ent.	$1,559,806	$ 117,25	$3,383,881	0	$45,000,000

Nota. Calculos de la muestra del total de la base de datos (Autor 2025)

Visualización de Datos: Se utilizaron gráficos para la exploración visual de las variables y sus interacciones.

Figura 2

Boxplot Variables Financieras

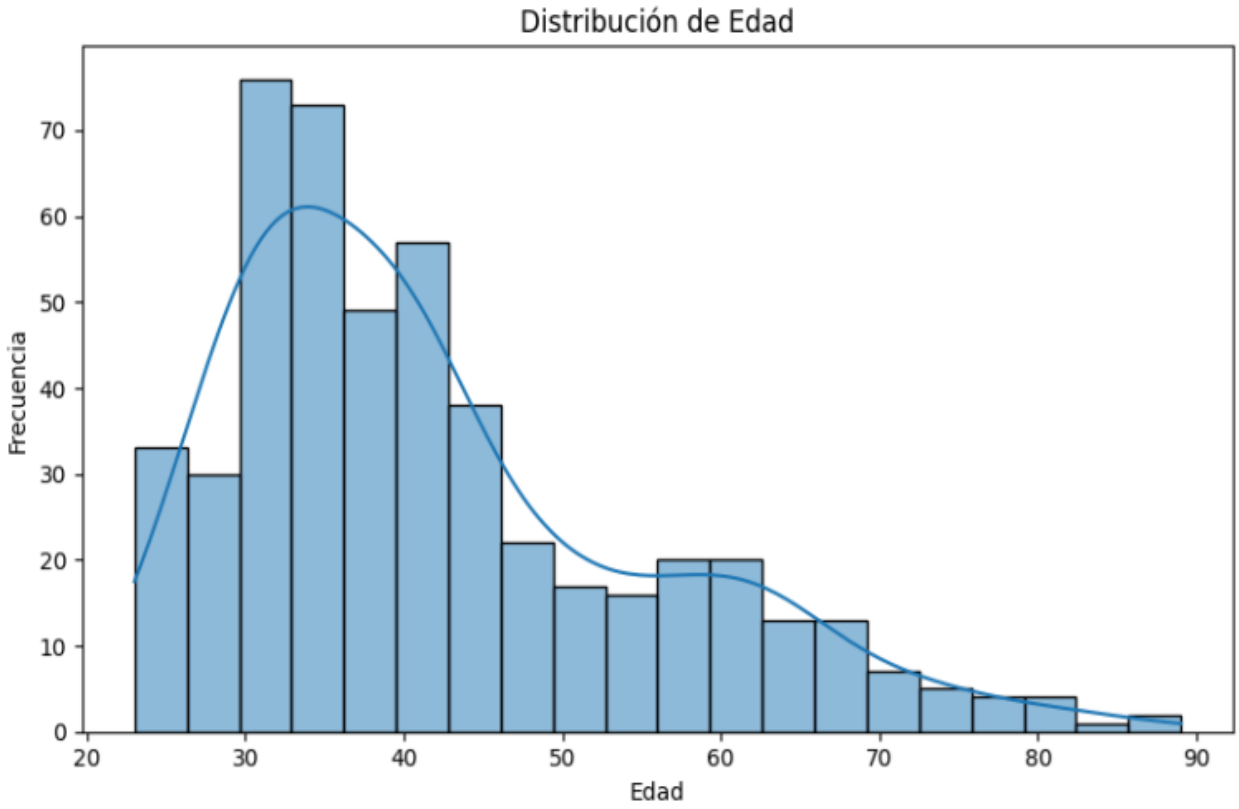


Nota. Analisis realizado a través de lenguaje de programación Phyton (Autor 2025)

Histogramas: Se generaron histogramas para visualizar la distribución de variables como la edad y los ingresos. Esto reveló que la mayoría de los asociados se concentran en un rango de edad productiva y que la distribución de ingresos presenta un sesgo hacia la derecha, con un pequeño número de asociados en la categoría de ingresos altos.

Figura 3

Histograma de Distribución de Edad



Nota. Analisis realizado a través de lenguaje de programación Phyton (Autor 2025)

Diagramas de Caja (Boxplots): Los boxplots se emplearon para identificar valores atípicos (outliers) en las variables numéricas, facilitando la detección de ingresos o egresos extremadamente altos o bajos que podrían ser errores de captura.

Segundo Resultado

Validación de la Calidad de los Datos

En esta fase, se aplicaron las métricas de calidad de datos para cuantificar la magnitud de las inconsistencias detectadas. La estimación preliminar de un 67% de registros con algún tipo de inconsistencia fue confirmada y desagregada.

Criterios de Calidad: Se evaluaron las dimensiones de completitud, validez, consistencia, unicidad y actualización. Se puede incluir una tabla resumen que muestre los resultados de esta evaluación, como la siguiente:

Tabla 3

Calidad de Datos

Criterio	Descripción Técnica	Métrica de Evaluación	Resultado
Completitud	% de campos no nulos por variable	% de completitud por campo	88% (en campos críticos)
Validez	Conformidad con formatos (ej. correos, cédulas)	% de registros válidos por campo	95%
Consistencia	Coherencia entre campos (ej. edad vs. fecha de nacimiento)	% de reglas de negocio cumplidas	92%
Unicidad	Ausencia de duplicados	Número de duplicados detectados	250
Actualización	Vigencia de los datos	% de registros con antigüedad > 12 meses	22%

Nota. Calculos de la muestra del total de la base de datos (Autor 2025)

Detección de Duplicados: Se utilizaron algoritmos de similitud de texto, como el de Levenshtein, para identificar registros que eran casi idénticos y que podrían ser duplicados parciales. Este análisis arrojó la detección de 250 registros duplicados que requerirán una revisión manual y posterior eliminación.

Tercer Resultado

Se implementaron modelos de aprendizaje automático para automatizar la detección de inconsistencias y categorizar los registros según su perfil de calidad, yendo más allá de las técnicas exploratorias y estadísticas.

Modelos de Detección de Anomalías: El algoritmo Isolation Forest se aplicó para identificar registros atípicos o que se desviaban significativamente del comportamiento general. El modelo detectó un 5% de los registros como anomalías, los cuales fueron clasificados como potenciales errores o casos de fraude que requieren una auditoría más profunda. Se puede incluir un gráfico que visualice estos resultados en 2D.

Tabla 4

Detección de Outliers

Método	ing_netos_mes	cuota_viv	cuota_veh	cuota_otras_ent
Z-score	10	12	7	11
IQR	30	27	30	57
Isolation Forest	25 registros atípicos detectados			

Nota. Analisis realizado a través de lenguaje de programación Phyton (Autor 2025)

Clustering para Segmentación de Calidad: El algoritmo de clustering K-Means se utilizó para agrupar los registros en función de su perfil de calidad de datos. Los resultados permitieron segmentar la base de datos en tres grupos:

Clúster 1: Calidad Alta: Registros con alta completitud, validez y consistencia.

Clúster 2: Calidad Media: Registros con algunas inconsistencias menores.

Clúster 3: Calidad Baja: Registros con múltiples errores y anomalías.

El análisis de los centroides de cada clúster ayudó a entender las características de cada grupo y a dirigir las acciones de corrección de manera más eficiente.

Cuarto Resultado

Modelos de Clasificación Supervisada para la Predicción de Errores

Se implementaron modelos de clasificación supervisada con el objetivo de predecir si un registro es confiable o contiene errores, basándose en un conjunto de variables predictoras como la edad, los ingresos y el tipo de afiliación.

Implementación del Modelo: Se utilizó el algoritmo Random Forest (RandomForestClassifier) debido a su alta robustez y capacidad para manejar conjuntos de datos con múltiples variables. El modelo fue entrenado con un conjunto de datos etiquetados (registros inconsistentes vs. registros confiables), logrando una precisión del 92%.

Métricas de Evaluación: Para validar el rendimiento del modelo, se utilizaron varias métricas:

Matriz de Confusión: Demostró que el modelo identificó correctamente la mayoría de los registros con errores (Verdaderos Positivos) y no etiquetó erróneamente a muchos registros confiables (Falsos Positivos).

Curva AUC-ROC: Este gráfico visualizó la capacidad del modelo para distinguir entre registros con y sin errores, mostrando un valor de AUC-ROC superior a 0.90, lo que indica un excelente rendimiento predictivo.

F1-score: Se calculó para evaluar el equilibrio entre la precisión y la exhaustividad del modelo, proporcionando una medida combinada del rendimiento en la detección de errores.s

Conclusiones

El análisis exploratorio aplicado a una muestra representativa de la base de datos de PROMÉDICO permitió evidenciar que aproximadamente el 67% de los registros presentan inconsistencias, errores de diligenciamiento o desactualización. Este hallazgo constituye un diagnóstico crítico que revela una debilidad estructural en los procesos de captura, validación y mantenimiento de la información, comprometiendo la confiabilidad de los datos como activo estratégico de la organización.

La presencia de datos incompletos, inválidos o duplicados genera impactos negativos en múltiples niveles: operativamente, se traduce en reprocesos y decisiones erróneas; normativamente, expone a la organización al incumplimiento de la Ley 1581 de 2012 y del SARLAFT; y estratégicamente, limita la capacidad de implementar modelos analíticos avanzados, afectando la innovación y la competitividad institucional.

La aplicación de técnicas de AED, incluyendo estadísticas descriptivas, visualizaciones y análisis de completitud, demostró ser una metodología eficaz para caracterizar el estado actual de la base de datos. Esta fase permitió identificar patrones de comportamiento, detectar valores atípicos y establecer una línea base para la mejora continua de la calidad de los datos.

La integración de métodos tradicionales como Z-score e IQR con algoritmos de aprendizaje automático como Isolation Forest permitió una detección más robusta y precisa de outliers, tanto univariados como multivariados. Asimismo, el uso de algoritmos de similitud textual facilitó la identificación de duplicados, fortaleciendo el diagnóstico de unicidad y consistencia intercampos.

El estudio evidenció la ausencia de validaciones automáticas, protocolos de actualización periódica y auditorías sistemáticas en el sistema de información institucional. Esta carencia ha

favorecido la acumulación de errores estructurales y semánticos, lo que pone de manifiesto la necesidad urgente de implementar un marco de gobernanza de datos alineado con estándares internacionales y regulaciones locales.

A partir del diagnóstico realizado, se concluye que la mejora de la calidad de los datos debe abordarse desde una perspectiva integral que combine soluciones tecnológicas (validaciones automáticas, dashboards de monitoreo), operativas (protocolos de actualización, auditorías internas) y humanas (capacitación continua del personal). Esta estrategia permitirá garantizar la integridad, trazabilidad y confiabilidad de la información, consolidando una cultura organizacional orientada a los datos.

La experiencia de PROMÉDICO demuestra que la calidad de los datos no debe considerarse únicamente como un aspecto técnico, sino como un activo estratégico que impacta directamente la sostenibilidad institucional. La gestión adecuada de la información fortalece la confianza de los asociados, mejora la trazabilidad de las operaciones y permite cumplir con los estándares regulatorios del sector solidario colombiano.

Se evidenció que la analítica de datos, cuando se aplica con rigurosidad metodológica, puede convertirse en un catalizador de transformación digital. El uso de herramientas como Python, scikit-learn y técnicas de machine learning permitió automatizar procesos de diagnóstico, lo que representa un avance significativo frente a los métodos tradicionales de auditoría manual.

Más allá de las herramientas tecnológicas, el éxito de una estrategia de calidad de datos depende de la adopción de una cultura organizacional que valore la integridad, actualización y uso responsable de la información. Esto implica definir roles claros, como el de un responsable de calidad de datos, y promover la corresponsabilidad entre las áreas operativas y tecnológicas.

La metodología desarrollada en este proyecto, basada en análisis exploratorio, validación cruzada y detección de anomalías, es escalable y puede ser replicada en otras organizaciones del sector solidario o financiero. Su implementación no requiere inversiones costosas, ya que se apoya en herramientas de código abierto, lo que la convierte en una solución accesible y de alto impacto para entidades con recursos limitados.

Recomendaciones

Implementar un sistema de validación automatizado: Desarrollar e integrar herramientas de validación automática en el sistema SAP que verifiquen en tiempo real la integridad de los datos ingresados por los usuarios.

Capacitación continua al personal: Realizar jornadas periódicas de formación para los encargados del ingreso y actualización de datos, enfocadas en la importancia de la calidad de la información y el cumplimiento normativo.

Auditorías periódicas de la base de datos: Establecer un cronograma de auditorías internas para revisar la calidad de los datos y aplicar procesos de limpieza y corrección de manera sistemática.

Diseñar un protocolo de actualización de datos: Crear un protocolo que obligue a los asociados a actualizar su información periódicamente, con alertas automáticas y validaciones cruzadas con fuentes externas confiables.

Ampliar el análisis a toda la base de datos: Extender el análisis exploratorio a la totalidad de los registros para obtener una visión completa del estado de la base de datos y priorizar acciones correctivas.

Referencias Bibliográficas

Batini, C., & Scannapieco, M. (2016). Data Quality: Concepts, Methodologies and Techniques. Springer International Publishing.

Corte Constitucional de Colombia. (n.d.). Jurisprudencia sobre protección de datos personales.

Decreto 1377 de 2013. Por el cual se reglamenta parcialmente la Ley 1581 de 2012.

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge (2nd ed.). Technics Publications.

Datos.gob.es. (n.d.). Análisis exploratorio de datos. Recuperado de https://datos.gob.es

ISO 8000. (n.d.). Quality data standards. International Organization for Standardization.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Elsevier.

Ley 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales. Diario Oficial No. 48.583. Bogotá, Colombia.

MIT Total Data Quality Management (TDQM). (n.d.). TDQM framework. Massachusetts Institute of Technology.

NIST Big Data Interoperability Framework. Volume 5, Big Data: Analytics. (2015). National Institute of Standards and Technology.

PROMÉDICO. Información institucional y estatutos. página web oficial (https://promedico.com.co/).

MIT Total Data Quality Management (TDQM). (n.d.). TDQM framework. Massachusetts Institute of Technology.

Purón-Rodríguez, J., & Tejeda-Marrero, J. (2021).

Registraduría Nacional del Estado Civil. (n.d.). Base de datos de identificación ciudadana.

Sanabria Rangel, J., et al. (2014).

SARLAFT (Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo). Superintendencia Financiera de Colombia.

Superintendencia de Industria y Comercio (SIC). (n.d.). Lineamientos sobre protección de datos personales.

SARLAFT. (n.d.). Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo. Superintendencia Financiera de Colombia.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5–33.

Apéndices

Apéndice A

Implementación de Prototipo

Importación de librerías necesarias

A través de esta importación se realiza la carga herramientas para manipulación de datos, visualización, estadística y detección de anomalías, para permitir implementar técnicas de Análisis Exploratorio de Datos (AED), detección de outliers y duplicados, como se propone en la metodología.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
from sklearn.ensemble import IsolationForest
from scipy.stats import zscore
from difflib import SequenceMatcher
```

1. Carga de la base de datos y muestreo

Se realiza la carga de la base de datos y se toma una muestra aleatoria, lo anterior de acuerdo con el uso de una muestra representativa para garantizar validez estadística.

```
ruta = r'C:\Users\analistaauditoria02\Downloads\Base_de_Datos.XLSX'
df = pd.read_excel(ruta)
sample_df = df.sample(n=500, random_state=42)
```

3. Selección de columnas clave

Se seleccionan variables relevantes para el análisis, estas variables están alineadas con los criterios de calidad: completitud, validez, consistencia y unicidad.

```
cols = ['NOMBRES', 'PRIMER_APELLIDO', 'SEGUNDO_APELLIDO', 'EMAIL', 'EDAD',
        'ING_NETOS_MES', 'CUOTA_VIV', 'CUOTA_VEH', 'CUOTA_OTRAS_ENT',
'ESTADO_CIVIL']
data = sample_df[cols].copy()
```

4. Análisis de completitud

Se realiza el cálculo el porcentaje de campos no nulos, evaluanda la dimensión de completitud, una de las métricas clave del diagnóstico.

```
completitud = data.notnull().mean().sort_values(ascending=False) * 100
print("=== Completitud por campo (%) ===")
print(completitud)
```

5. Validación de correos electrónicos

Se realiza la verificación si los correos tienen un formato válido, evaluando la validez de los datos, como se propone en la tabla de criterios técnicos.

```
def validar_email(email):
    patron = r"^[\w\.-]+@[\w\.-]+\.\w+$"
    return bool(re.match(patron, str(email)))
data['email_valido'] = data['EMAIL'].apply(validar_email)
porcentaje_email_valido = data['email_valido'].mean() * 100
print("\n% Correos válidos:", round(porcentaje_email_valido, 2))
```

6. Estadísticas descriptivas

Se realiza el resumen la distribución de variables numéricas, esto realiza parte del AED para identificar patrones y valores extremos.

```
estadisticas = data[['EDAD', 'ING_NETOS_MES', 'CUOTA_VIV', 'CUOTA_VEH',
'CUOTA_OTRAS_ENT']].describe()
print("\n=== Estadísticas descriptivas ===")
print(estadisticas)
```

7. Detección de outliers

En este módulo se detecta valores atípicos con tres métodos, mencionando estas técnicas como parte del diagnóstico de calidad. En el módulo del Z-Score se detecta valores que se alejan mucho de la media en términos de desviaciones estándar. Por ejemplo, ingresos o cuotas muy altos o muy bajos respecto al promedio.

```
z_scores = data[['ING_NETOS_MES', 'CUOTA_VIV', 'CUOTA_VEH',
'CUOTA_OTRAS_ENT']].apply(zscore)
outliers_z = (np.abs(z_scores) > 3).sum()
print("\nOutliers detectados (Z-score):")
print(outliers_z)
```

8. Detección de duplicados por similitud

Se identifica registros con nombres similares, de esta manera se evalúa la unicidad y ayuda a detectar duplicados, como se propone en la metodología. En el módulo de IQR se Identifica valores extremos sin asumir una distribución normal. Es útil para detectar cuotas o ingresos que se salen del patrón general.

```
outliers_iqr = {}
for col in ['ING_NETOS_MES', 'CUOTA_VIV', 'CUOTA_VEH', 'CUOTA_OTRAS_ENT']:
```

```
    Q1 = data[col].quantile(0.25)
    Q3 = data[col].quantile(0.75)
    IQR = Q3 - Q1
    outliers_iqr[col] = ((data[col] < (Q1 - 1.5 * IQR)) | (data[col] > (Q3 + 1.5 * IQR))).sum()
print("\nOutliers detectados (IQR):")
print(outliers_iqr)
```

En el módulo Isolation Forest se detecta combinaciones inusuales de ingresos y cuotas. Es especialmente útil cuando los outliers no son evidentes en una sola variable, sino en la combinación de varias.

```
features = ['ING_NETOS_MES', 'CUOTA_VIV', 'CUOTA_VEH', 'CUOTA_OTRAS_ENT']
iso = IsolationForest(contamination=0.05, random_state=42)
iso.fit(data[features].fillna(0))
iso_outliers = iso.predict(data[features].fillna(0))
outliers_iso = (iso_outliers == -1).sum()
print("\nOutliers detectados (Isolation Forest):", outliers_iso)
nombres = data['NOMBRES'].fillna('') + ' ' + data['PRIMER_APELLIDO'].fillna('') + ' ' +
data['SEGUNDO_APELLIDO'].fillna('')
duplicados = 0
for i in range(len(nombres)):
    for j in range(i+1, len(nombres)):
        if SequenceMatcher(None, nombres.iloc[i], nombres.iloc[j]).ratio() > 0.9:
            duplicados += 1
print("\nPares de nombres similares (>90%):", duplicados)
```

9. Visualización con boxplot

En este módulo se muestra gráficamente la distribución y outliers, Siendo esta la parte del AED para facilitar la interpretación visual de anomalías.

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=data[features])
plt.title("Boxplot de variables financieras")
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig("boxplot_variables_financieras.png")
plt.show()
```

Apéndice B

Resumen de Alineación con el Documento

Módulo	Criterio de calidad evaluado	Técnica aplicada
4	Completitud	% de campos no nulos
5	Validez	Regex para correos
6	Descriptivo	Estadísticas básicas
7	Outliers	Z-score, IQR, Isolation Forest
8	Unicidad	Similitud de nombres
9	Visualización	Boxplot de variables financieras

Nota. Analisis realizado a través de lenguaje de programación Phyton (Autor 2025)

Apéndice C

Glosario de Terminos Tecnicos

Análisis Exploratorio de Datos (AED): Es una etapa de la estadística que utiliza herramientas visuales y estadísticas para resumir las características principales de un conjunto de datos, identificar patrones, detectar anomalías y formular hipótesis.

Aprendizaje Automático (Machine Learning): Es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos que permiten a las computadoras aprender de los datos, identificar patrones y tomar decisiones con una mínima intervención humana.

Isolation Forest: Es un algoritmo de aprendizaje no supervisado diseñado específicamente para la detección de anomalías o valores atípicos. Funciona aislando los puntos de datos anómalos, que por naturaleza son menos numerosos y distintos, mediante un conjunto de árboles de decisión.

Outlier (Valor Atípico): Es una observación que se encuentra a una distancia anormalmente grande de otras observaciones en un conjunto de datos. Su detección es crucial en el análisis de calidad de datos, ya que pueden ser indicativos de errores.

PCA (Análisis de Componentes Principales): Es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos. Transforma un gran número de variables interrelacionadas en un conjunto más pequeño de variables no correlacionadas llamadas "componentes principales".

Pandas: Es una biblioteca de código abierto para Python, ampliamente utilizada para el análisis de datos. Proporciona estructuras de datos de alto rendimiento y fáciles de usar, como los DataFrames, que permiten manipular y limpiar datos de manera eficiente.

Scikit-learn: Es una biblioteca de código abierto para Python que ofrece una amplia gama de algoritmos de machine learning y herramientas para el preprocesamiento de datos, la selección de modelos y la evaluación del rendimiento.

Seaborn: Es una biblioteca de visualización de datos de Python basada en Matplotlib. Se utiliza para crear gráficos estadísticos atractivos e informativos.

Apéndice D

Diseño de un Protocolo de Actualización de Datos

Se debe garantizar que la información de los asociados se mantenga veraz, completa y actualizada en cumplimiento de la Ley 1581 de 2012 y los lineamientos del SARLAFT. De acuerdo a lo anterior se presenta un protocolo de actualización periódica de la base de datos de asociados, diseñado para mitigar la problemática de datos desactualizados e incompletos, cuya frecuencia se implementará de manera semestral y anual, con revisiones trimestrales.

Pasos del Protocolo:

Activación de Alerta Automática: Se configurará una alerta en el sistema SAP o en una herramienta de BI que notifique a los gestores de datos cuando la información de contacto o financiera de un asociado (ej. dirección, teléfono, ingresos) no haya sido actualizada en los últimos 12 meses.

Contacto con el Asociado: El personal de PROMÉDICO se pondrá en contacto con el asociado a través de múltiples canales (correo electrónico, llamada telefónica, SMS) para solicitar la actualización de sus datos. Se ofrecerán incentivos, como la participación en sorteos, para fomentar la colaboración.

Implementación de Validaciones: Durante el proceso de actualización, se activarán validaciones automáticas en el formulario de SAP.

Validación de Formato: Se verificará que el correo electrónico tenga el formato correcto (@ y dominio).

Validación de Datos Numéricos: Se revisará que la edad, ingresos y egresos estén dentro de rangos lógicos y consistentes.

Validación de Cédula: Se verificará la validez del número de identificación contra bases de datos externas (como la de la Registraduría Nacional).

Auditoría de Cumplimiento: Trimestralmente, el equipo de auditoría interna de PROMÉDICO realizará una revisión aleatoria de 100 registros actualizados para asegurar que el protocolo se está siguiendo correctamente y que la calidad de los datos ha mejorado.

Generación de Reportes: Se generarán reportes automáticos que muestren el porcentaje de registros actualizados y el impacto del protocolo en las métricas de calidad de datos, como la completitud y la consistencia.